Bayesian Analysis (2018) **13**, Number 2, pp. 335–358

A Comparison of Truncated and Time-Weighted Plackett–Luce Models for Probabilistic Forecasting of Formula One Results

Daniel A. Henderson[*] and Liam J. Kirrane[†]

Abstract. We compare several variants of the Plackett–Luce model, a commonly-used model for permutations, in terms of their ability to accurately forecast Formula One motor racing results. A Bayesian approach to forecasting is adopted and a Gibbs sampler for sampling from the posterior distributions of the model parameters is described. Prediction of the results from the 2010 to 2013 Formula One seasons highlights clear strengths and weaknesses of the various models. We demonstrate by example that down weighting past results can improve forecasts, and that some of the models we consider are competitive with the forecasts implied by bookmakers odds.

Keywords: Bayesian inference, Gibbs sampling, latent variable models, permutations, ranks, sport.

1 Introduction

Formula One is the premier class of motor racing that comes under the governance of the Fédération Internationale de l'Automobile (FIA). Since its inception in 1950, the FIA Formula One World Championship has grown into a multi-billion pound industry, underpinned by lucrative television rights and sponsorship deals, with a global audience in the hundreds of millions each year; Smith (2013) provides a comprehensive historical account. Despite its popularity, published statistical analyses of Formula One results are relatively rare; three relatively recent, relevant examples are Eichenberger and Stadelmann (2009), Phillips (2014) and Bell et al. (2016). Eichenberger and Stadelmann (2009) fit a linear regression model to results from 1950 to 2006 with finishing position as the response variable and with driver effects and car–year effects, as well as other covariates such as the weather and race distance. Phillips (2014) uses race points as the response variable (standardised to the scale in use between 1991 and 2002, and with hypothetical fractional points awarded to drivers who fail to score points) in a nonlinear regression model with individual driver effects, team effects and competition effects (which capture the competitiveness of racing). The model is fitted to results from 1950 to 2013. The results in Phillips (2014) are comprehensively compared to subjective driver rankings and the historical context of the results is discussed. Bell et al. (2016) also take a historical perspective as they aim to rank the best Formula One drivers of all

[*] School of Mathematics & Statistics, Newcastle University, Newcastle upon Tyne, NE1 7RU, U.K., daniel.henderson@ncl.ac.uk

[†] School of Mathematics & Statistics, Newcastle University, Newcastle upon Tyne, NE1 7RU, U.K.

https://doi.org/10.1214/17-BA1048

time. They adopt a Bayesian approach with (standardised and normalised) race points as the response variable in a multilevel model which also accounts for team effects; they find the team effects to be more important than driver effects. Bell et al. (2016) also provide a discussion of the comparison of their modelling approach and their results to those of Phillips (2014) and Eichenberger and Stadelmann (2009). All three papers focus primarily on quantifying driver performance over several eras. Our motivation is different; we take a more forward looking perspective and focus on forecasting future results.

The finishing order of the drivers in a Formula One race is of paramount importance since drivers score points in each race corresponding only to their finishing position. As such, the result of a race can be viewed as a permutation of the set of drivers in the race, and so we focus on models for permutations; see Marden (1995) for an overview. In particular, we focus on the Plackett–Luce model (Plackett, 1975; Luce, 1959), a popular model for permutations. The Plackett–Luce model specifies the probability that the winner of the race beats all rivals, multiplied by the probability that the second placed driver beats all remaining rivals conditional on being beaten by the winner, and so on until we have the probability that the second from last placed driver beats the last placed driver. Directly modelling the results as permutations allows a probabilistic interpretation of the results which provides a natural framework for forecasting future results. We adopt a Bayesian approach to prediction as it provides a natural mechanism for updating predictions as new data become available and it can fully account for parameter uncertainty when forming predictions. This quantification of parameter uncertainty may be particularly useful when making predictions based on only a small number of past results. Bayesian inference for the Plackett–Luce model has been considered by several authors including Guiver and Snelson (2009) and Caron and Doucet (2012). See also Glickman and Hennessy (2015) for an application to the Bayesian analysis of rank ordered data in multi-competitor sports. Frequentist (maximum likelihood) inference for the parameters of the Plackett–Luce model has also been well-studied; Hunter (2004) provides a thorough review of the field together with some novel algorithms.

In this paper we compare the basic Plackett–Luce model with three variants of the basic model in terms of their ability to forecast the results from the 2010–2013 Formula One seasons. These four seasons were chosen for analysis because the same points scoring system was in place during this period. Race results in the form of a full finishing order of all the drivers who competed in each race were downloaded from www.formula1.com in June 2014. The data comprise 77 races involving 42 drivers. The first variant of the basic Plackett–Luce model that we consider is a truncated Plackett–Luce model, in which only the finishing positions of the top-r drivers in the race are used; all drivers not finishing in the top-r places are grouped together. The motivation for this model comes from the fact that for forecasting purposes the results in the top-r places are usually of primary interest, and we might not want some spurious poor results for some drivers to impact on our forecasts for their likely performance in future races. Furthermore, because the points system in 2010–2013 gave zero points to all drivers finishing lower than 10th place, there might be a case for supposing that, for the top drivers at least, the effort expended for the lower places is probably less than that for the top places.

Censoring the observations by pooling these lower placed drivers together may be better at forecasting the winner or other high placed finishing positions.

The second variant of the basic model that we consider is a reverse Plackett–Luce model. In the reverse Plackett–Luce model we simply model the reverse finishing order of the drivers via the basic Plackett–Luce model. In other words, we propose a model for the probability that the last placed driver is beaten by all rivals, multiplied by the probability that the second from last placed driver is beaten by all remaining rivals conditional on beating the last placed driver, and so on. Graves et al. (2003) proposed such a model in the context of modelling National Association for Stock Car Auto Racing (NASCAR) results and argued that it overcomes some of the drawbacks of the Plackett–Luce model such as being strongly influenced by poor results and therefore penalising the good drivers too harshly for the occasional lowly finish. Such lowly finishes for the good drivers, Graves et al. (2003) argue, are commonplace in motor racing due to reliability issues with the cars which are beyond the driver's control, crashes and so on. Graves et al. (2003) called the reverse Plackett–Luce model the *attrition* model, because it focuses on the process of drivers dropping out, with the best drivers the ones who are able to 'survive' the longest in the race. The results in Graves et al. (2003) suggest, unequivocally, that the reverse Plackett–Luce ('attrition') model is far superior to the Plackett–Luce model when analysing the full finishing order of drivers in NASCAR races over several seasons. Due to the similarity of F1 and NASCAR it is likely that the results of Graves et al. (2003) will transfer to F1. However, it is of interest to see whether the attrition model is better at forecasting aspects of the race results other than the full finishing order.

For the purposes of forecasting it seems natural to assume that results become less important the further in the past that they are. With that in mind, the third variant that we consider is a time-weighted version of both the attrition and Plackett–Luce models in which past data is down-weighted when making forecasts for future races. We adopt a Bayesian approach to inference and forecasting, with our forecasts based on the sequential prior predictive distributions of the relevant quantities. In order to implement time-weighting, we utilise power prior distributions (Ibrahim and Chen, 2000) in a similar manner to the power-weighted densities method of McCarthy and Jensen (2016). This allows us to easily update our forecasts as new data become available through minimal changes to the basic models and an efficient computational scheme. Specifically, we utilise the latent variable formulation of the Plackett–Luce model as discussed in Caron and Doucet (2012) and modify the Gibbs sampler of Caron and Doucet (2012) to accommodate both the time-weighted and truncated versions of the Plackett–Luce model.

For modelling/forecasting purposes we attribute the result of a race solely to the drivers who competed in the race; no team/car or other information is taken into account. This allows us to focus on relatively simple models which are easy to use and for which computation is efficient. It also allows us to update our forecasts easily as new data becomes available. We acknowledge that in analysing the results of a race what we are really analysing is the combination of the driver's ability and other factors such as the team/car performance. Previous attempts to additionally model team performance and other factors can be found in Eichenberger and Stadelmann (2009), Phillips (2014) and Bell et al. (2016).

The paper is structured as follows. The Plackett–Luce model and its variants are described in Section 2. Section 3 describes our approach to forecasting, discusses a time-weighted version of the Plackett–Luce model, and outlines an efficient Gibbs sampling algorithm for the time-weighted (truncated) Plackett–Luce model. Results are presented in Section 4 and the paper concludes, in Section 5, with a summary.

2 Models

2.1 Basic notation

Suppose we have data on n races involving K drivers represented by the set $\mathcal{K} = \{1, 2, \ldots, K\}$. Suppose that $n_i \leq K$ drivers are involved in race i with the set of these n_i drivers being denoted $\mathcal{K}_i$. The finishing order of the ith race is denoted $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{in_i})$. This is a permutation of the elements in $\mathcal{K}_i$, such that X_{i1} is the driver who finished first, X_{i2} is the driver who finished second, and so on.

2.2 The Plackett–Luce model

Let $\lambda_k > 0$ be a parameter representing the potential performance of driver $k \in \mathcal{K}$. In what follows we will refer to λ_k as the "ability" of driver k, but, as discussed in Section 1, these parameters do not necessarily reflect a driver's true ability but rather an amalgam of all factors which affect the driver's potential performance in future races. A commonly used model for the finishing positions of the drivers in race i specifies this joint probability through a product of marginal and conditional probabilities,

$$p(\boldsymbol{x}_i|\boldsymbol{\lambda}, \mathcal{K}_i) = \prod_{j=1}^{n_i-1} \frac{\lambda_{x_{ij}}}{\sum_{m=j}^{n_i} \lambda_{x_{im}}}, \tag{1}$$

where $\boldsymbol{\lambda} = \{\lambda_k\}_{k \in \mathcal{K}}$ is the collection of all driver abilities. In the remainder of the paper we drop the explicit conditioning on the set of drivers $\mathcal{K}_i$ for economy of notation. Here, λ_k is proportional to the probability that driver k wins a race. This model is commonly called the Plackett–Luce model (Plackett, 1975; Luce, 1959) though it has been suggested independently by several other authors, such as Harville (1973). The model is composed of the probability that x_{i1} beats all rivals, multiplied by the probability that x_{i2} beats all remaining rivals conditional on being beaten by x_{i1}, and so on until we have the probability that x_{in_i} came last, given that all the other drivers were ranked higher.

Clearly, the likelihood of the ability parameters under the Plackett–Luce model (1) is unaltered if we re-scale $\boldsymbol{\lambda}$. As such, the data only provides information about the *relative* standard of the drivers, not their absolute standard. Some implications of this non-identifiability are addressed later in Section 3.2.

It is well-known (Diaconis, 1988; Marden, 1995) that the Plackett–Luce model can also be represented as the marginal probability

$$p(\boldsymbol{x}_i|\boldsymbol{\lambda}) = \int p(\boldsymbol{x}_i|\boldsymbol{Z}_i,)p(\boldsymbol{Z}_i|\boldsymbol{\lambda})\mathrm{d}\boldsymbol{Z}_i, \tag{2}$$

where $\boldsymbol{Z}_i = \{Z_{ij}; j \in \mathcal{K}_i\}$ are exponentially distributed latent variables, one for each driver in race i,

$$Z_{ij}|\lambda_j \sim \mathrm{Exp}(\lambda_j), \qquad i = 1, 2, \ldots, n, \quad j \in \mathcal{K}_i,$$

and

$$p(\boldsymbol{x}_i|\boldsymbol{Z}_i) \equiv \Pr(\boldsymbol{X}_i = \boldsymbol{x}_i|\boldsymbol{Z}_i) = \Pr(Z_{ix_{i1}} < Z_{ix_{i2}} < \cdots < Z_{ix_{in_i}}) \tag{3}$$

is the *rank likelihood* (Pettitt, 1982). In other words, the kth ranked driver is the one with the kth smallest value of their latent variable. One interpretation of the latent variables $\boldsymbol{Z}_i$ is as the unobserved finishing times of the n_i drivers; the individual with the fastest time wins, and so on. This latent variable representation gives a generative model which can be used to simulate realisations from the Plackett–Luce model. This representation corresponds to a Thurstonian order-statistics model with Gumbel (extreme-value) latent variables (Yellott, 1977).

2.3 Partial rankings: the truncated Plackett–Luce model

As discussed in Section 1, we could argue that there is less information in the lower rankings; good drivers can have mechanical failures beyond their control, for example, or their effort may decrease when the chance of a points finish is out of their reach. The specification of the Plackett–Luce model (1) through conditional probabilities makes it simple to analyse partial rankings. For race i, taking only the results down to r_ith place where $1 \le r_i \le n_i$ we define

$$p(\boldsymbol{x}_i|\boldsymbol{\lambda}, r_i) \equiv p(x_{i1}, \ldots, x_{ir_i}|\boldsymbol{\lambda}) = \prod_{j=1}^{r_i^*} \frac{\lambda_{x_{ij}}}{\sum_{m=j}^{n_i} \lambda_{x_{im}}}, \tag{4}$$

where $r_i^* = \min(r_i, n_i - 1)$. Drivers who participate in the race but who do not finish in the top r_i are all treated equally, as not finishing in the top r_i.

We note that the vase models of Silverberg (1980), of which Benter (1994) is a special case, account for variable effort for the lower places. The truncated Plackett–Luce model can be seen as a special case of Benter's model in which there is no information in the rankings after position r_i.

2.4 The reverse Plackett–Luce model or attrition model

In the reverse Plackett–Luce model, as in the Plackett–Luce model, let $\boldsymbol{Z}_i = \{Z_{ij}; j \in \mathcal{K}_i\}$ be exponentially distributed latent variables, one for each driver in race i, such that

$$Z_{ij}|\lambda_j \sim \mathrm{Exp}(\lambda_j), \qquad i = 1, 2, \ldots, n, \quad j \in \mathcal{K}_i.$$

The data are connected to these latent variables through the rank likelihood

$$p(\boldsymbol{x}_i|\boldsymbol{Z}_i) \equiv \Pr(\boldsymbol{X}_i = \boldsymbol{x}_i|\boldsymbol{Z}_i) = \Pr(Z_{ix_{in_i}} < Z_{ix_{in_{i-1}}} < \cdots < Z_{ix_{i1}}). \tag{5}$$

Under this generative model, the kth ranked driver is the one with the kth *largest* value of their latent variable. This model is alluded to in several places in Marden (1995), as the "backwards Plackett" model, for example. The $\boldsymbol{Z}_i$ can be interpreted as the "failure" times of the n_i drivers in the ith race, with the winner being the individual that fails last or, in other words, lasts the longest. Based on this interpretation, Graves et al. (2003) call this model the attrition model as it represents the attrition that is common in motor racing, in that cars fail due to mechanical problems that are beyond a driver's control; a poor driver is unlikely to win a race, but a good driver can very easily come last. As such the attrition model naturally deals with some of the issues that the truncated Plackett–Luce model of (4) was designed to address.

Note that the marginal likelihood under the attrition model can be obtained by simply reversing the data for each race in the basic Plackett–Luce model, that is, by replacing $\boldsymbol{X}_i$ by the reverse finishing order. Therefore, the attrition model can be fitted using the same methodology as for the (truncated) Plackett–Luce model and so in Section 3 we focus on describing computation for the truncated Plackett–Luce model. Under the attrition model, however, the driver "ability" parameters have a different interpretation. Now λ_k is proportional to the probability that driver k finishes last. The driver with the smallest λ_k value indicates the "least worst" driver.

The factorisation of the marginal finishing order under the attrition model does not lend itself to a simple truncated version as did the Plackett–Luce model. Deriving the marginal probability $p(x_{i1},\ldots,x_{ir_i}|\boldsymbol{\lambda})$ would involve summing out the random variables X_{ir_i+1} to X_{in_i} from the joint probability given by (1) but with reversed data. This brute force summation is possible but extremely computationally intensive. For these computational reasons, and the fact that the attrition model naturally deals with the occasional poor result, we have not pursued a truncated attrition model here.

3 Bayesian predictive inference and forecasting

Our main aim is to use the models described in the previous section to forecast the results of future races based on observed results. We let $D_t = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t\}$ denote the results of the first t races, and set $D \equiv D_n$ to be the whole set of results. We adopt a Bayesian approach and, for a particular model $M = m$ with parameters $\boldsymbol{\lambda}$, we base forecasts for the outcome of the next race $\boldsymbol{x}_t$ on the predictive distribution

$$p(\boldsymbol{x}_t|D_{t-1},m) = \int p(\boldsymbol{x}_t|\boldsymbol{\lambda},m)p(\boldsymbol{\lambda}|D_{t-1},m)\mathrm{d}\boldsymbol{\lambda}, \tag{6}$$

where we assume that race outcomes are conditionally independent given the parameters, so that $p(\boldsymbol{x}_t|D_{t-1},\boldsymbol{\lambda},m) = p(\boldsymbol{x}_t|\boldsymbol{\lambda},m)$. In (6), $p(\boldsymbol{\lambda}|D_{t-1},m) \propto p(\boldsymbol{\lambda}|m)p(D_{t-1}|\boldsymbol{\lambda},m)$ is the current conditional density of $\boldsymbol{\lambda}$ given by Bayes' Theorem, $p(\boldsymbol{\lambda}|m)$ is the initial prior density of $\boldsymbol{\lambda}$ and

$$p(D_{t-1}|\boldsymbol{\lambda},m) = \prod_{i=1}^{t-1} p(\boldsymbol{x}_i|\boldsymbol{\lambda},m) \tag{7}$$

is the current likelihood of $\boldsymbol{\lambda}$, as specified by the (truncated) Plackett–Luce model (4).

3.1 Time-weighted forecasts

For the purposes of forecasting it seems reasonable to assume that race results become less important the further in the past they are. Essentially, we are questioning the exchangeability of the data caused by changes in driver abilities $\boldsymbol{\lambda}$ over time or by unmodelled covariates (such as changes to the driver's car, changes in his team personnel, the positive effects of several good results, or the negative effects of a few bad results).

Rather than explicitly modelling changes in $\boldsymbol{\lambda}$ over time through a state-space model for example, an alternative pragmatic approach, which can be seen as a preliminary step to a more sophisticated modelling approach such as that in Glickman and Hennessy (2015), is to represent the non-exchangeability of the data through the likelihood function rather than the prior. We do so by down-weighting information from past results more heavily as time progresses. McCarthy and Jensen (2016) give details of such an approach in a general time series context. They show that simple models with so-called power-weighted densities can be competitive with more sophisticated state-space models for forecasting financial time series. Our time-weighted modification of the (truncated) Plackett–Luce model assumes that at time τ the likelihood contribution of race i, which took place at time τ_i, is

$$p(\boldsymbol{x}_i|\boldsymbol{\lambda}, m)^{\psi(\tau-\tau_i)}. \tag{8}$$

Here $\psi(\cdot)$ is a non-decreasing function such that $0 \leq \psi(x) \leq 1$ for $x \geq 0$, implying that a result from the distant past is weighted no more heavily than a recent result. We assume that $\psi(0) = 1$, and that the functional form of ψ depends on a single parameter ξ, but leave discussion of the precise form of $\psi(\cdot)$ till Section 3.3.

The time-weighted likelihood at time $\tau \geq \tau_{t-1}$, based on results up to and including those of race $t-1$, is then simply

$$p(D_{t-1}|\boldsymbol{\lambda}, m, \boldsymbol{\tau}_{t-1}, \tau, \xi) = \prod_{i=1}^{t-1} p(\boldsymbol{x}_i|\boldsymbol{\lambda}, m)^{\psi(\tau_i)}, \tag{9}$$

where $\boldsymbol{\tau}_{t-1} = (\tau_1, \dots, \tau_{t-1})$ denotes the dates of the $t-1$ races. Clearly taking $\psi(\tau_i) = 1$ for all i gives the original likelihood from (7), that is, $p(D_{t-1}|\boldsymbol{\lambda}, m, \boldsymbol{\tau}_{t-1}, \tau, \xi) = p(D_{t-1}|\boldsymbol{\lambda}, m)$.

Equation (9) does not correspond to a generative model for the Plackett–Luce models described in Section 2. We can, however, view this time-weighted likelihood as forming the basis of a *power prior* distribution (Ibrahim and Chen, 2000) $p(\boldsymbol{\lambda}|D_{t-1}, m, \boldsymbol{\tau}_{t-1}, \tau, \xi)$, which is obtained, via Bayes' Theorem, as $p(\boldsymbol{\lambda}|D_{t-1}, m, \boldsymbol{\tau}_{t-1}, \tau, \xi) \propto p(\boldsymbol{\lambda}|m)p(D_{t-1}|\boldsymbol{\lambda}, m, \boldsymbol{\tau}_{t-1}, \tau, \xi)$ where $p(\boldsymbol{\lambda}|m)$ represents the initial beliefs about the parameters. This is simply a form of power prior distribution in which $\psi(\tau - \tau_i)$ quantifies the uncertainly in the historical data D_{t-1} and therefore down-weights this historical data when forming the prior distribution; see Ibrahim and Chen (2000) for a review of power priors in terms of regression models.

Forecasts about the next race $\boldsymbol{x}_t$ at time τ_t are based on the predictive distribution

$$p(\boldsymbol{x}_t|D_{t-1}, m, \boldsymbol{\tau}_{t-1}, \tau_t, \xi) = \int p(\boldsymbol{x}_t|\boldsymbol{\lambda}, m)p(\boldsymbol{\lambda}|D_{t-1}, m, \boldsymbol{\tau}_{t-1}, \tau_t, \xi)\mathrm{d}\boldsymbol{\lambda}. \tag{10}$$

The only modification from (6) is that the current conditional distribution of the ability parameters is based on the time-weighted data. It turns out that the time-weighted conditional distributions $p(\boldsymbol{\lambda}|D_{t-1}, m, \boldsymbol{\tau}_{t-1}, \tau, \xi)$ for the (truncated/reverse) Plackett–Luce models are relatively easy to sample from by using Gibbs sampling; details are provided in Section 3.2.

3.2 Bayesian computation for the time-weighted Plackett–Luce model

Likelihood

At time τ, the weighted likelihood contribution of the ith race under the truncated Plackett–Luce model (4) is

$$p(\boldsymbol{x}_i|\boldsymbol{\lambda}, r_i)^{\psi(\tau-\tau_i)} = \left\{ \prod_{j=1}^{r_i^*} \frac{\lambda_{x_{ij}}}{\left(\sum_{m=j}^{n_i} \lambda_{x_{im}}\right)} \right\}^{\psi(\tau-\tau_i)} = \prod_{j=1}^{r_i^*} \frac{\lambda_{x_{ij}}^{\psi(\tau-\tau_i)}}{\left(\sum_{m=j}^{n_i} \lambda_{x_{im}}\right)^{\psi(\tau-\tau_i)}}. \quad (11)$$

The time-weighted likelihood based on the first $t-1$ races $D_{t-1} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{t-1}\}$ can be written

$$p(D_{t-1}|\boldsymbol{\lambda}, \boldsymbol{r}_{t-1}, \boldsymbol{\tau}_{t-1}, \tau, \xi) = \prod_{i=1}^{t-1} p(\boldsymbol{x}_i|\boldsymbol{\lambda}, r_i)^{\psi(\tau-\tau_i)} = \prod_{k\in\mathcal{K}} \lambda_k^{w_k} \prod_{i=1}^{t-1} \prod_{j=1}^{r_i^*} \left(\sum_{m=j}^{n_i} \lambda_{x_{im}} \right)^{-\psi(\tau-\tau_i)}, \quad (12)$$

where $\boldsymbol{r}_{t-1} = (r_1, \ldots, r_{t-1})$ and $w_k = \sum_{i=1}^{t-1} \sum_{j=1}^{r_i^*} \mathbb{I}(x_{ij} = k)\psi(\tau - \tau_i)$ is the (time-weighted) number of races in which the kth driver competes and finishes in the top r_i^* positions. Here $\mathbb{I}(x)$ denotes an indicator function which equals 1 if x is true and equals 0 otherwise. Clearly, taking $\psi(\tau - \tau_i) = 1$ for all i gives the likelihood under the truncated Plackett–Luce model. Also, taking $r_i = n_i$ (and thus $r_i^* = n_i - 1$) for all i, gives the likelihood under the Plackett–Luce model. Reversing the data and taking $r_i = n_i$ gives the attrition model. The model described above could be fitted by using a modification of the MM algorithm of Hunter (2004) to maximise the likelihood. We will, however, adopt a Bayesian approach to inference and prediction.

Prior specification

Beliefs about the relative abilities of the drivers are expressed through independent gamma prior distributions, $\lambda_k \sim \text{Gamma}(a_k, b)$ for $k \in \mathcal{K}$, with $a_k, b > 0$. The scale parameter b can be set equal to 1, without loss of generality, due to the scale invariance of $\boldsymbol{\lambda}$. The choice of a_k allows flexibility in specifying beliefs about the abilities of the drivers. Taking $a_k = a$ for all $k \in \mathcal{K}$ leads to an exchangeable prior specification in which we believe that some drivers are better than others but we have no beliefs about who the stronger and the weaker drivers are. This specification has the attractive property that each possible permutation of finishing orders in the first race, and therefore each possible result of this first race, is equally likely *a priori*. In other words this prior induces a uniform distribution on the set of permutations $\sigma(\mathcal{K}_1)$. With the exchangeable specification $\lambda_k \sim \text{Gamma}(a, 1)$ for $k \in \mathcal{K}$, the prior probability that driver i beats driver j has a $\text{Beta}(a, a)$ distribution under the Plackett–Luce and reverse-Plackett–Luce models. Taking $a = 1$ clearly leads to uniform $U(0, 1)$ beliefs about these head-to-head probabilities, which has some appeal in the absence of specific prior information.

Posterior computation

In order to make predictions within the Bayesian paradigm we need the conditional (power prior) distribution $p(\boldsymbol{\lambda}|D_{t-1}, \boldsymbol{r}_{t-1}, \boldsymbol{\tau}_{t-1}, \tau, \xi) \propto p(\boldsymbol{\lambda})p(D_{t-1}|\boldsymbol{\lambda}, \boldsymbol{r}_{t-1}, \boldsymbol{\tau}_{t-1}, \tau, \xi)$, which is not available in closed form. In the special case of the Plackett–Luce model ($\psi(\tau - \tau_i) = 1$ and $r_i = n_i$ for all i) Caron and Doucet (2012) showed that by introducing some carefully chosen latent variables an efficient Gibbs sampler can be constructed for sampling from the joint distribution of parameters and latent variables given the data. We modify their approach to accommodate time-weighting and truncation, and propose the Gibbs sampler of Algorithm 1. Full details are provided in the Supplementary material (Henderson and Kirrane, 2017).

Algorithm 1 Gibbs sampler for time-weighted truncated Plackett–Luce model

```
1. Initialise $\boldsymbol{\lambda}^{(0)}$ arbitrarily.
2. For iteration $\ell = 1, 2, \ldots$
    (a) for $i = 1, \ldots, t-1$ and for $j = 1, \ldots, r_i^*$, sample
            $Y_{ij}^{(\ell)}|D_{t-1}, \boldsymbol{\lambda}^{(\ell-1)}, \boldsymbol{r}_{t-1}, \boldsymbol{\tau}_{t-1}, \tau, \xi \sim \text{Gamma}\left(\psi(\tau - \tau_i), \sum_{m=j}^{n_i} \lambda_{x_{im}}^{(\ell-1)}\right);$
    (b) for $k \in \mathcal{K}$ sample
            $\lambda_k^{(\ell)}|D_{t-1}, \boldsymbol{Y}_{t-1}^{(\ell)}, \boldsymbol{r}_{t-1}, \boldsymbol{\tau}_{t-1}, \tau, \xi \sim \text{Gamma}\left(a_k + w_k, b + \sum_{i=1}^{t-1} \sum_{j=1}^{r_i^*} \delta_{ijk} Y_{ij}^{(\ell)}\right),$
        where $w_k = \sum_{i=1}^{t-1} \sum_{j=1}^{r_i^*} \mathbb{I}(x_{ij} = k)\psi(\tau - \tau_i)$ and $\delta_{ijk} = \mathbb{I}(k \in \{x_{ij}, \ldots, x_{in_i}\})$;
    (c) (Optional) sample $\Lambda^{(\ell)} \sim \text{Gamma}\left(\sum_{j \in \mathcal{K}} a_j, b\right)$ and for $k \in \mathcal{K}$ set
            $\lambda_k^{\star(\ell)} = \frac{\lambda_k^{(\ell)}}{\sum_{j \in \mathcal{K}} \lambda_j^{(\ell)}} \Lambda^{(\ell)}.$
        Then for $k \in \mathcal{K}$ set $\lambda_k^{(\ell)} = \lambda_k^{\star(\ell)}$.
```

Algorithm 1 entails a trivial modification of the Gibbs sampler for the Plackett–Luce model given in Caron and Doucet (2012) and their algorithm can be seen as a special case of Algorithm 1 when $\psi(\tau - \tau_i) = 1$ and $r_i = n_i$ for all i.

Identifiability issues

As discussed in Section 2.2, the likelihood (12) is invariant under scalar multiplication of $\boldsymbol{\lambda}$ and hence the combined ability level of all the drivers, $\Lambda = \sum_{j \in \mathcal{K}} \lambda_j$ is not likelihood identifiable. As pointed out in Caron and Doucet (2012), this non-identifiability can cause the Markov chain defined in Steps (a) and (b) of Algorithm 1 to mix poorly, since Λ is essentially unconstrained by the data. The rescaling step (2(c)) in Algorithm 1 improves the mixing of the Markov chain; see Caron and Doucet (2012) for full details.

We note that this non-identifiability does not cause any inferential problems, however, and no artificial parameter constraints are needed. Due to the invariance of the likelihood (12) under scalar multiplication of $\boldsymbol{\lambda}$ the average ability of the drivers, $\eta = \Lambda/K$, is also not likelihood identifiable, and so $p(\eta|D) = p(\eta)$. Therefore (our beliefs about) the average ability of the drivers η will not change over time with the inclusion of more data. In other words, this non-identifiability provides a natural mechanism for looking at differences over time, without having to artificially standardise our inferences.

3.3 Specification of the time-weighting function

So far we have assumed that the weight function $\psi(\cdot)$ is completely specified through a parameter ξ, although we have not explicitly discussed choices. There is considerable flexibility in the choice of weighting function; McCarthy and Jensen (2016) give some general guidelines and Dixon and Coles (1997) briefly discuss some possibilities in the context of modelling football scores. We assume that time is measured discretely in days, denoted x, and use a geometric weighting function,

$$\psi(x) = \xi^x, \qquad 0 \leq \xi \leq 1, \tag{13}$$

in which after 1 day we assign a weight of ξ to a race result, after two days this result is assigned a weight of ξ^2 and so on. Taking $\xi = 1$ gives the standard (non-time-weighted) models and, as ξ decreases, past results are down-weighted more heavily. We can choose a specific value of ξ by considering the number of days it takes the results to be weighted half of present results. Alternatively, rather than specifying a single specific value, we could use the data to inform our choice. For example, we can fit several models for a range of values of ξ to D_{t-1} and then choose to base predictions for race t on the value of ξ which minimises an appropriate loss function based on the previous predictions from the various models. This and other strategies have been investigated and the results are reported in the Supplementary material. We find that an approach based on sequentially choosing the value of ξ which maximises the log prior predictive probability (see Section 4.4) to work well. We note that this is essentially the method proposed in McCarthy and Jensen (2016) based on maximising the one step ahead predictive likelihood.

4 Results

In this section we compare the predictive performance of the various models described so far. In particular we compare the models on their ability to accurately forecast the outcomes of future races as results become available throughout the course of the 2010 to 2013 F1 seasons. We focus on five aspects of performance: predicting (i) the race winner, (ii) the top three finishers in a race (those with a "podium" finish), (iii) the top 10 finishers in a race (those with a "points" finish), (iv) the season champion, and (iv) the full finishing order for a race. Whilst all these aspects are of interest to motor racing fans, the first four probabilities may be useful for betting purposes as bookmakers and betting exchanges typically offer markets on these outcomes. The probability of the observed finishing order has no real relevance for betting purposes, but it provides an overall measure of the suitability/performance of the various models.

4.1 Modelling choices and computational details

For the purposes of illustration we present results for seven models. In each model we took the prior parameter $a_k = a = 1$ for all $k \in \mathcal{K}$. A sensitivity analysis reported in the Supplementary material suggests that the predictions are not particularly sensitive to this choice. The seven models are the standard Plackett–Luce model (`PL`), the attrition (reverse Plackett–Luce) model (`a`), truncated Plackett–Luce models with truncation at $r_i = 6$, $r_i = 10$ and $r_i = 14$, for all i (models `PL.t6`, `PL.t10` and `PL.t14` respectively) and time-weighted versions of both the Plackett–Luce model and the reverse Plackett–Luce (attrition) model with ξ chosen sequentially to maximise the log prior predictive probability (models `PL.tw` and `a.tw`, respectively). Further details on the choice of ξ are provided in the Supplementary material. The set of models is denoted $\mathcal{M} = \{\texttt{a}, \texttt{a.tw}, \texttt{PL}, \texttt{PL.tw}, \texttt{PL.t6}, \texttt{PL.t10}, \texttt{PL.t14}\}$.

After each race, each model was fitted by running the appropriate Gibbs sampler for $N = 10000$ iterations after a burn-in of 100 iterations. This level of burn-in was more than adequate; the Markov chains appear to converge very quickly to stationarity, and appear to mix well. We emphasise that no tuning or adaptation is necessary with these Gibbs samplers and so it is easy to fit the models efficiently as new data become available. The models are coded in R; code and data are available from `http://www.mas.ncl.ac.uk/~ndah6/F1/`.

4.2 Forecasting the winner, the podium finishers and the points finishers

For a given race, bookmakers and betting exchanges typically offer markets on who is going to win (finish first), finish on the podium (a top 3 finish) and finish in a point position (a top 10 finish), amongst many others. Therefore we initially focus on forecasting these aspects of the data. The models can be used to estimate the probability that driver j finishes in the top q places in race t, for each participating driver $j \in \mathcal{K}_t$, based on data up to but not including race t, D_{t-1}. Clearly this probability is 0 for all drivers who do not compete in race t. We use Monte Carlo simulations from the models

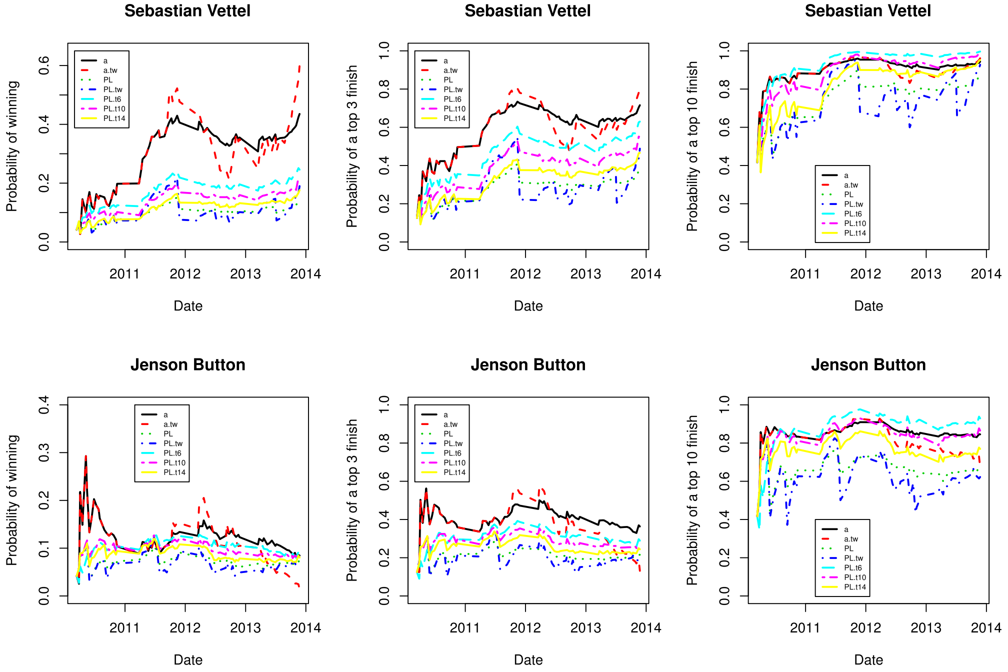

Figure 1: Estimates of the probability of winning ($\hat{p}_{jt}^{[M,1]}$, left column), finishing on the podium ($\hat{p}_{jt}^{[M,3]}$, middle column) and finishing in a points position ($\hat{p}_{jt}^{[M,10]}$, right column) for Sebastian Vettel (upper row) and Jenson Button (lower row) for each race of the 2010–2013 Formula One seasons based on the various models $M \in \mathcal{M}$.

to estimate these probabilities due to the lack of analytic expressions for most of them. Essentially, for each model M, we play out the next race, race t, N times and estimate the probability that driver j finishes in the top q places by the proportion of simulated races in which driver j finishes in the top q places, which we denote $\hat{p}_{jt}^{[M,q]}$. Algorithms for simulating race outcomes under the Plackett–Luce model and the reverse Plackett–Luce model use the exponential latent variable formulation, as described in Section 2, and are described in the Supplementary material. Precise details of how we estimate the predictive probabilities are also given in the Supplementary material.

Figure 1 displays estimates of the probabilities of winning, finishing in a podium position and finishing in a points position, under the seven models in $\mathcal{M}$, for two drivers who competed in all F1 races during the period 2010–2013. The upper row of plots in Figure 1 gives these probabilities for Sebastian Vettel, the German driver who won the World Drivers' Championship each year from 2010 to 2013. The lower row of plots in Figure 1 gives these probabilities for the British former World Champion Jenson Button. The left-hand column of plots in Figure 1 show the probabilities of winning each race. Note that closed-form estimates of the probability of winning can be calculated for the various versions of the Plackett–Luce model. These closed-form probabilities are practically indistinguishable from the sample-based estimates, suggesting that the sample-based estimates are adequate for our purposes. Details are given in the Supplementary material.

From Figure 1 we can clearly see that the predictive probabilities change substantially over time, adapting to new data as it becomes available. For example, looking at the probability of Vettel winning (upper left plot in Figure 1) the probabilities under the time-weighted attrition model are broadly similar to those under the standard attrition model; they are, however, more reactive to the results. For Button, the attrition model and the time-weighted attrition model give fairly similar estimates for the win and top 3 probabilities up to early 2013, then the time-weighted estimates drop off much more steeply. This indicates that the time-weighted model reacts more swiftly than the non-time-weighted model to Button's lower level of achievement in 2013.

The probabilities follow similar patterns over time under the various models, but there are considerable differences in the actual probabilities under the different models, especially between the attrition models and the Plackett–Luce models. In order to assess which model gives the most accurate forecasts we next examine the predictive performance of the models.

Predictive assessment

A logarithmic scoring rule (Bernardo and Smith, 1994) can be used to assess the quality of the forecast probabilities $\hat{p}_{jt}^{[M,q]}$; we "score" an amount $S(\boldsymbol{p}, i) = \log(p_i)$ if i occurs and our forecast probabilities are given in the vector $\boldsymbol{p}$. In the binary case (for example, driver j does/does not finish in the top q) the logarithmic scoring rule is a strictly proper local scoring rule; see Gneiting and Raftery (2007) for a review of proper scoring rules for prediction. Specifically, if a successful outcome (a top q finish) for driver j in race t is indicated by $o_{jt}^{[q]}(\boldsymbol{x}_t) = \mathbb{I}\,(j \in \{x_{t1}, \ldots, x_{tq}\})$, then the score for the prediction $\hat{p}_{jt}^{[M,q]}$ is

$$S(\hat{p}_{jt}^{[M,q]}, o_{jt}^{[q]}(\boldsymbol{x}_t)) = o_{jt}^{[q]}(\boldsymbol{x}_t) \log \hat{p}_{jt}^{[M,q]} + \{1 - o_{jt}^{[q]}(\boldsymbol{x}_t)\} \log \left(1 - \hat{p}_{jt}^{[M,q]}\right).$$

We combine the scores for each driver in the tth race to give a combined score for our predictive probabilities of

$$S_t^{[M,q]} = \sum_{j \in \mathcal{K}_t} S(\hat{p}_{jt}^{[M,q]}, o_{jt}^{[q]}(\boldsymbol{x}_t)).$$

Scores under the different models can be compared, with the model with the largest score giving the best predictions for this facet of the data. We track the scores and cumulative sum of the scores over races; the sum of the scores over races 1 to t, $S^{[M,q]}(t) = \sum_{i=1}^{t} S_t^{[M,q]}$ gives an overall measure of the quality of the forecasts for model $M \in \mathcal{M}$.

Figure 2 displays several comparisons of the seven models in terms of log scores. The extreme outliers in the log scores for the winner probabilities correspond to the unfancied Venezuelan driver, Pastor Maldonado, winning the Spanish Grand Prix in May 2012. At the time of writing (November 2016), this is Maldonado's only Grand

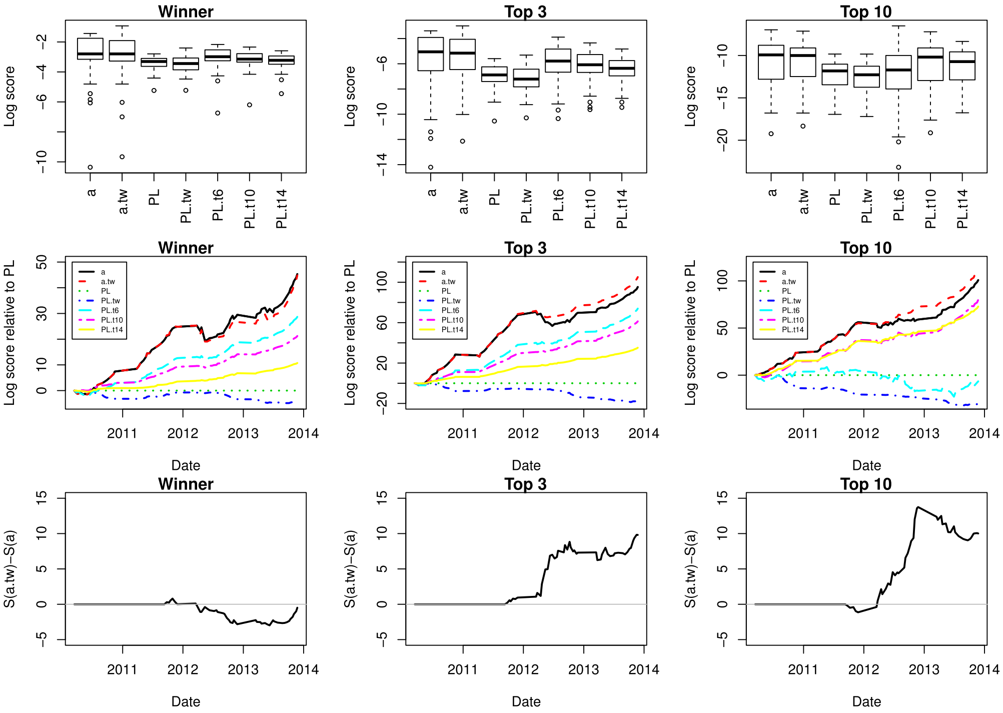

Figure 2: Comparison of log scores for forecasts of the winner (left column), the top 3 places (middle column) and the top 10 places (right columns). The upper row gives boxplots of the log scores under the different models combined over the years 2010–2013. The middle row gives the cumulative score for each model minus that for the Plackett–Luce model over the years 2010–2013. The lower row gives the cumulative score for the time weighted reverse Plackett–Luce model minus that for the reverse Plackett–Luce model.

Prix victory. This low probability event hits the log score under the attrition models hard, as can be seen by the dip in relative log score in early 2012 in the middle left plot in Figure 2.

Taken as a whole, the plots of Figure 2 suggest that the attrition models clearly outperform the Plackett–Luce models at forecasting each set of probabilities. Truncation of the Plackett–Luce model generally improves forecasts, with the more severe truncation giving more accurate forecasts for the winner and top 3 probabilities, but truncation at $r_i = 6$ not performing well for the top 10 forecasts. For top 10 forecasts, truncation at $r_i = 10$ performs the best of the Plackett–Luce models, not surprisingly; truncation at $r_i = 6$ clearly throws away too much information. The effect of down-weighting past results is not so clear cut. For the attrition model there is practically no difference for predicting the winner, but time-weighting does have a positive impact when predicting the top 3 and top 10 finishers. Note that the predictions under the time-weighted attrition model are the same as those under the attrition model with no time-weighting up to near the end of the 2011 season. This is because the optimal ξ up to that point is $\xi = 1$. The strategy we have implemented for choosing ξ for the Plackett–Luce model leads to worse forecasts than when no time-weighting is implemented; other strategies are discussed in the Supplementary material.

Comparison of observed and expected values of selected statistics

We can informally assess the goodness of fit of the models by comparing the observed number of wins, podium finishes and points finishes for each of the drivers with what we would expect under the different models. The expected number of top q finishes for driver j over the course of the four seasons from 2010 to 2013 is simply the sum of the predictive probabilities for a top q finish, $\sum_{t=1}^{n} \hat{p}_{jt}^{[M,q]}$. These values are reported in Table 1 for a selection of ten prominent/notable drivers. The results in Table 1 demonstrate that the observed and expected values match up well for the attrition model, and time-weighted attrition model but match up less well for the various Plackett–Luce models. In particular, the Plackett–Luce model massively underestimates the number of wins for Sebastian Vettel. The attrition models also underestimate Vettel's wins, but to a much lesser extent. Vettel clearly exceeded expectations in the four year period 2010–2013. Overall these results suggest that the attrition models provide a good fit to these data and are superior to the Plackett–Luce models.

4.3 Predicting the championship winner

One of the main outcomes of interest throughout the course of a season is the identity of the eventual winner of the Driver's Championship. Drivers score points corresponding to their finishing position in each race with the driver accumulating the most points at the end of the season being declared the winner. Table 2 gives the points system that was in place throughout the 2010–2013 seasons. After race $t-1$ we estimate the probability that driver j wins that season's Drivers' Championship as follows. For $\ell = 1, \ldots, N$ we simulate race results $\boldsymbol{x}_t^{(\ell)}, \boldsymbol{x}_{t+1}^{(\ell)}, \ldots, \boldsymbol{x}_{t*}^{(\ell)}$ for the rest of the season (here $t*$ denotes the last race of the season) from the current prior predictive distribution. We then convert simulated race results into points, based on Table 2, and add these simulated points for the rest of the season $d_j^{(\ell)}$ to the driver's actual current points total after race $t-1$, d_{jt-1}. This gives N sampled values $d_{jt-1}^{(\ell)} = d_{jt-1} + d_j^{(\ell)}$ from the predictive distribution of the driver's final points total for the season. For each $\ell = 1, 2, \ldots, N$, the driver with the most predicted points is the simulated champion. We estimate the probability that driver j is the season champion, $p_{jt-1}^{[M,C]}$, by the proportion of simulated seasons in which driver j had the highest predicted points total, that is

$$\hat{p}_{jt-1}^{[M,C]} = \frac{1}{N} \sum_{\ell=1}^{N} \mathbb{I}\left(d_{jt-1}^{(\ell)} = \max_{k \in \mathcal{K}} (d_{kt-1}^{(\ell)}) \right).$$

Figure 3 plots these probabilities for eventual four-time champion Sebastian Vettel over the course of the 2010 to 2013 seasons. For the sake of clarity, the probabilities under

Model	Statistic	Vettel	Alonso	Hamilton	Button	Webber	Rosberg	Räikkönen	Maldonado	Massa	Schumacher
Observed	Wins	34	11	11	8	7	3	2	1	0	0
	Podiums	53	42	27	25	32	9	15	1	8	1
	Top 10s	68	69	62	61	65	55	33	7	58	31
`a`	Wins	23.34	13.17	10.16	9.54	11.16	1.73	2.05	0.04	2.35	0.34
	Podiums	45.13	36.48	32.19	30.99	33.47	10.69	8.71	0.15	12.67	2.98
	Top 10s	69.14	67.30	65.84	65.40	65.83	53.80	27.41	8.58	55.50	31.91
`a.tw`	Wins	23.38	13.24	9.20	8.80	9.81	2.13	3.34	0.06	2.55	0.36
	Podiums	44.37	36.00	29.75	28.34	30.22	11.56	11.08	0.44	12.80	3.01
	Top 10s	68.14	66.12	63.49	62.47	63.22	51.73	27.42	11.32	52.94	30.01
`PL`	Wins	7.61	8.26	5.05	5.59	6.72	4.67	3.85	1.04	4.83	2.01
	Podiums	21.69	23.49	15.01	16.55	19.54	13.94	10.90	3.30	14.49	6.32
	Top 10s	58.13	61.14	47.70	51.09	55.92	45.71	28.43	13.73	47.05	23.70
`PL.tw`	Wins	8.23	7.20	4.83	5.15	6.35	4.22	3.54	1.26	4.48	1.86
	Podiums	22.80	20.65	14.37	15.33	18.41	12.71	10.11	4.01	13.47	5.74
	Top 10s	57.17	56.10	45.10	47.61	52.18	42.15	26.96	16.38	43.77	21.41
`PL.t6`	Wins	13.24	9.65	8.74	7.80	8.83	4.46	3.09	0.43	4.49	1.83
	Podiums	35.50	28.11	25.83	23.48	26.00	14.07	9.44	1.42	14.16	5.89
	Top 10s	71.84	70.14	69.44	67.28	68.04	56.04	30.25	8.68	56.63	29.03
`PL.t10`	Wins	10.84	9.22	7.35	7.47	8.96	4.68	3.51	0.41	4.70	2.23
	Podiums	30.05	26.59	21.83	22.14	25.83	14.48	10.40	1.35	14.59	7.07
	Top 10s	68.57	67.56	64.11	64.43	67.31	53.13	30.22	7.21	53.62	30.15
`PL.t14`	Wins	8.79	8.89	6.21	6.69	8.41	4.84	3.61	0.72	4.45	2.30
	Podiums	24.93	25.26	18.51	19.80	24.19	14.70	10.43	2.34	13.56	7.21
	Top 10s	63.11	64.89	57.07	58.94	63.78	50.31	29.00	11.10	47.76	28.31

Table 1: Observed and expected statistics under the models in $\mathcal{M}$ for selected drivers.

Finishing position	1	2	3	4	5	6	7	8	9	10	≥ 11
Points	25	18	15	12	10	8	6	4	2	1	0

Table 2: Formula One points system for the seasons 2010–2013.

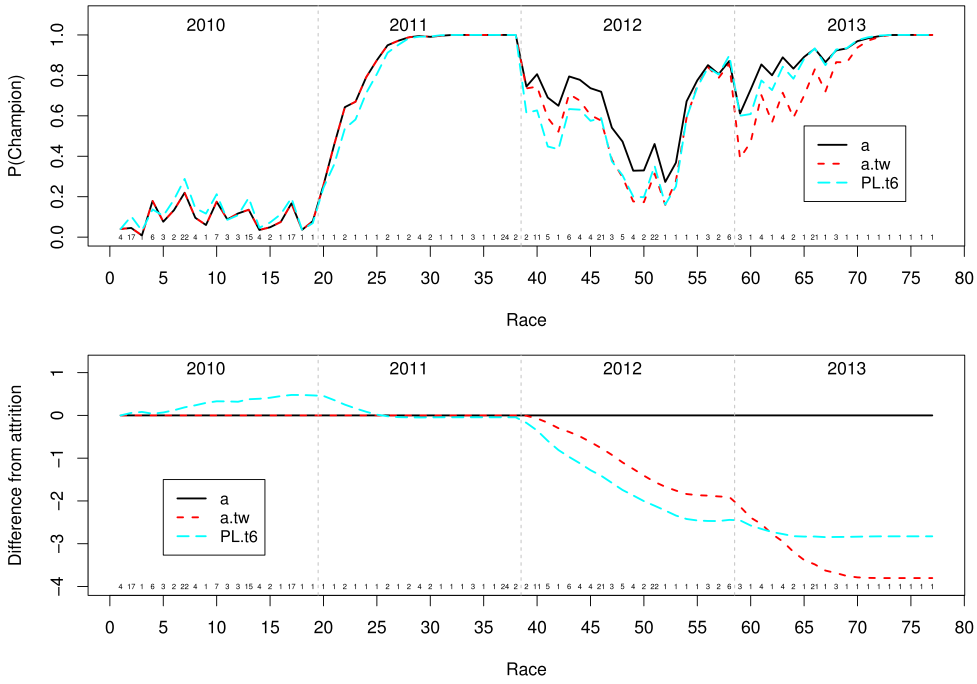

Figure 3: Upper plot: Estimated probability of Sebastian Vettel winning that season's Drivers' Championship prior to each race of the 2010–2013 F1 seasons under three models (`a`, `a.tw` and `PL.t6`). Lower plot: difference in cumulative probability of Vettel winning the Drivers' Championship relative to the attrition model. Also displayed at the foot of each plot is Vettel's finishing position for each race.

only three models are compared: the attrition model (`a`), the time-weighted attrition model (`a.tw`), and the Plackett–Luce model truncated at 6th position (`PL.t6`). We chose `PL.t6` as it was the best of the (truncated) Plackett–Luce models at predicting the winner and the top three places (see Figure 2). Being able to accurately predict these high placings is important for predicting the Championship winner as the high placings contribute the most points to a driver's total. Also displayed in Figure 3 is Vettel's finishing position for each race. The probabilities under all three models are fairly similar. The main difference between the models lies early in a season where the truncated model typically assigns a lower probability to Vettel being champion than the attrition models. A notable exception is for the 2013 season where the time-weighted attrition model gives the lowest probabilities out of these three models. In 2011 and 2013 there was little doubt from fairly early in the season as to Vettel's eventual success, with practical certainty towards the end of the season that he would be victorious. In contrast, the 2012 season was closely fought, with the lead in the Drivers' Championship changing hands seven times; see Smith (2013) for a detailed account. The 2012 Championship winner was still uncertain going into the last race, the Brazilian Grand Prix; Vettel lead the Championship race on 273 points, with Alonso, his closest rival, second on 260 points. At that point all three models made Vettel the favourite to win the title with a probability of around 0.85. Vettel only managed a 6th place finish in the Brazilian Grand Prix, but with Alonso finishing 2nd, it meant that Vettel won the title by only 3 points. The 2010 season was perhaps even closer with several drivers capable of winning the championship going into the last race, the Abu Dhabi Grand Prix. Fernando Alonso lead the points race with 246, the Australian Mark Webber (Vettel's team mate) was second on 238 points, and Vettel lay third, 15 points adrift of Alonso on 231 points. At this point, all three models had Vettel around a 7–8% chance of being champion, with Alonso approximately 70% and Webber around 20–23%. Vettel won the race — with Alonso finishing 7th and Webber 8th — to win his first World Championship and become the then youngest-ever F1 World Champion.

Bookmakers' odds can be considered a gold-standard in terms of forecasting the outcomes of sporting events. It is therefore interesting to compare the forecasts under the various Plackett–Luce type models with odds given by bookmakers. We note that this may be an unfair comparison since bookmakers can use all available information to inform their odds, such as qualifying grid position, reports from testing, and so on, whereas the models discussed here only use the finishing positions from races starting with the 2010 season, and simple exchangeable prior beliefs about driver abilities. Bookmakers odds against an outcome occurring of $a : b$ convert to an implied probability of the outcome occurring of $p = b/(a + b)$, and typically include an in-built edge; for disjoint events, the bookmakers implied probabilities typically add up to $1 + x$, with $x > 0$, where $100x\%$ is their expected profit. Odds from one particular bookmaker (Sky BET) were obtained for the 2013 Drivers' Championship winner. The data from Sky BET consist of a time series of odds against each driver winning the Championship. Figure 4 plots the (unscaled) implied probabilities calculated from these odds for three drivers: eventual 2013 Champion Sebastian Vettel, runner up Fernando Alonso from Spain, and the Finnish driver Kimi Räikkönen, who won the first Grand Prix of the 2013 season, and in addition finished second in three of the first five races. Also included in Figure 4 are estimated championship probabilities $\hat{p}_{jt-1}^{[M,C]}$ under the attrition model and the Plackett–Luce model. Informally, if the ratio of our model-based probability to the unscaled implied bookmakers probability is greater than 1 then it represents a potential betting opportunity in that if our model-based probabilities are more accurate than the bookmakers probabilities then our expected gain will be positive in the long run. Clearly, the probabilities $\hat{p}_{jt-1}^{[M,C]}$ under the attrition model for Sebastian Vettel being champion are greater (and, from mid-March to July 2013, much greater) than the bookmaker's probabilities, suggesting several attractive betting opportunities. If we were to base predictions on the Plackett–Luce model, which has become somewhat of a standard model for data on permutations, then a bet on Vettel winning would not have looked particularly attractive for most of the season. It is also reassuring to see that at no point during the season would either Alonso or Räikkönen have presented an

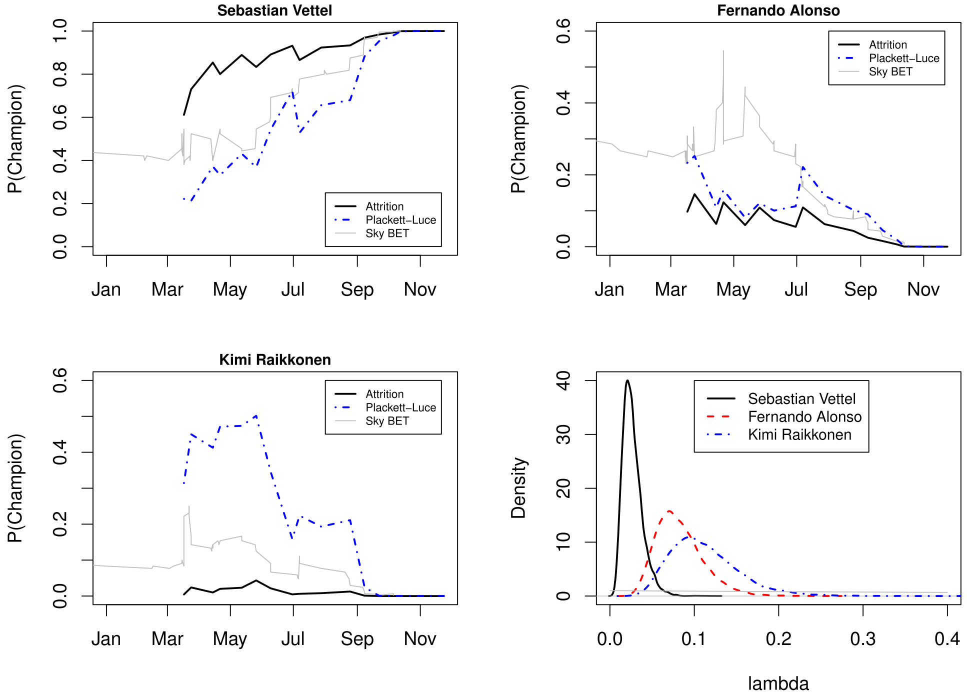

Figure 4: Estimated probability of winning the 2013 Drivers' Championship over time for Sebastian Vettel (upper left), Fernando Alonso (upper right) and Kimi Räikkönen (lower left), based on the attrition model, on the Plackett–Luce model and on unscaled implied probabilities from a bookmaker. The lower right plot shows estimated posterior densities for the driver ability parameters for the three drivers after the last race of the 2013 season under the time-weighted attrition model.

attractive betting opportunity under the attrition model. Figure 4 illustrates that despite its apparent simplicity, the attrition model (and its time-weighted variants) can be competitive with possibly more sophisticated systems and expert judgements employed by bookmakers. For completeness, Figure 4 also includes estimated posterior densities of the driver ability parameters for Vettel, Alonso and Räikkönen based on all the races from the 2010 to 2013 seasons. These posteriors are based on the time-weighted attrition model with $\xi = 0.9970$ which is the optimal value based on maximising the log marginal likelihood. The plot clearly shows Vettel's dominance over these other two drivers during this period; recall that under the attrition model smaller values of the parameter are better.

4.4 Full finishing order

So far we have concentrated on assessing the models in terms of their predictive performance on aspects of the data that are of interest to the typical motor racing fan or bettor, such as who is going to win or finish on the podium. It is also of interest for statistical reasons to compare the models on their ability to forecast the whole finishing order for each race. From (10), the predictive probability of the exact finishing order of the tth race $\boldsymbol{x}_t$, based on the data so far D_{t-1}, under the time-weighted truncated Plackett–Luce model, can be estimated consistently by the sample average

$$\hat{p}(\boldsymbol{x}_t|D_{t-1},\boldsymbol{r}_{t-1},\boldsymbol{\tau}_{t-1},\tau_t,\xi)=\frac{1}{N}\sum_{\ell=1}^{N}\prod_{j=1}^{n_t-1}\frac{\lambda_{x_{tj}}^{(\ell)}}{\sum_{m=j}^{n_t}\lambda_{x_{tm}}^{(\ell)}},$$

where $\boldsymbol{\lambda}^{(\ell)}$, for $\ell=1,2,\ldots,N$, are sampled from $p(\boldsymbol{\lambda}|D_{t-1},\boldsymbol{r}_{t-1},\boldsymbol{\tau}_{t-1},\tau_t,\xi)$.

The sum of the log transformed prior predictive probabilities, for example,

$$\log p(D_t|\boldsymbol{\tau}_{t-1},\xi)=\sum_{i=1}^{t}\log p(\boldsymbol{x}_i|D_{i-1},\boldsymbol{\tau}_{i-1},\xi)$$

gives an estimate of the log marginal likelihood, or model evidence, based on the data up to race t. (Note that if greater accuracy is required, Chib's method (Chib, 1995) can be used to estimate the log marginal likelihood under the non-time-weighted models. An alternative method such as the power posterior method of Friel and Pettitt (2008) may be required for the time-weighted versions). The upper left panel of Figure 5 displays the log prior predictive probabilities under the seven models computed immediately before each race of the 2010 to 2013 Formula One seasons. Clearly, the attrition models outperform the Plackett–Luce models, and (unsurprisingly) the truncated variants of the Plackett–Luce model are not as good as the untruncated model at predicting the full finishing order. Based on the upper left panel of Figure 5, the time-weighted versions of both the attrition model and the Plackett–Luce model look to be marginally better than their non-time-weighted counterparts.

Bayes factors (Kass and Raftery, 1995) are commonly used for choosing between models and the log marginal likelihood for each model minus that for the Plackett–Luce model gives the log Bayes factor for that model relative to the Plackett–Luce model. These log Bayes factors relative to the Plackett–Luce model are displayed in the upper right panel of Figure 5. Kass and Raftery (1995) provide guidelines for interpreting the log Bayes factor of model A relative to model B in terms of the strength of evidence against model B; a log Bayes factor of 5 or more constitutes very strong evidence against model B. This strength of evidence against the Plackett–Luce model is achieved by the attrition models after only a few races, and over time the strength of evidence grows. The differences between the time-weighted attrition model and the attrition model are less evident from the upper right plot, and so in the lower left plot we plot the log Bayes factor for the time weighted attrition model against the attrition model. Kass and Raftery's guidelines are also indicated by the horizontal lines. We see that there is no preference early on, but by late 2011 the time-weighted attrition model is deemed superior and by mid 2012 the evidence against the attrition model (in favour of the time-weighted version) is very strong, and generally, apart from a slight dip in 2013, gets stronger. The prior predictive probabilities can also be used

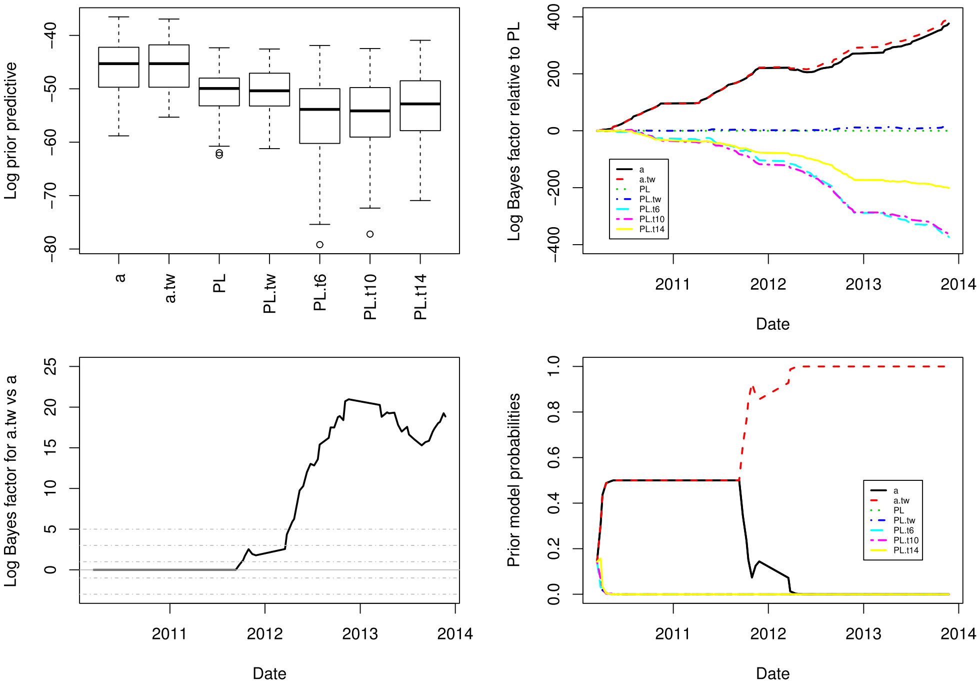

Figure 5: Comparison of models in terms of (i) log prior predictive probabilities (upper left), (ii) log Bayes factors relative to the Plackett–Luce model (upper right), (iii) the log Bayes factor of the time-weighted attrition model relative to the attrition model (lower left), and (iv) prior model probabilities (lower right), prior to each race of the 2010–2013 F1 seasons. The horizontal lines on the lower left plot represent the strengths of evidence suggested by Kass and Raftery (1995).

to compute prior model probabilities. For example, if we assume *a priori* that the various models have equal probability, then the initial prior model probabilities are $p(M = m) = |\mathcal{M}|^{-1}$ for $m \in \mathcal{M}$. The probability of model $m \in \mathcal{M}$ immediately before race t is $p(M = m|D_{t-1}) \propto p(M = m)p(D_{t-1}|M = m)$. These prior probabilities for each model computed immediately prior to each race of the 2010 to 2013 Formula One seasons are shown in the lower right panel of Figure 5. The plot shows that, until late in the 2011 season, the time-weighted attrition model is no different from the attrition model; this is because the optimal ξ up to this point is $\xi = 1$. The plot also shows that from late 2011 the time-weighted attrition model is the most likely model out of the ones considered, and by mid 2012 it has a probability close to 1. Bayesian model selection theory (Bernardo and Smith, 1994) would lead us to base inferences on the model which has the maximum probability. For most of the period under study this was the time-weighted attrition model. We stress, however, that this reflects the model's ability to predict the whole finishing order, which may not be of primary interest to most F1 fans or bettors. The log marginal likelihood is simply a log scoring rule and it may be that the log scores of Section 4.2 may be more suitable for choosing an appropriate model, depending on which features of the data we are most interested in forecasting; see Dawid and Musio (2014) for discussion.

It is very common to use a composite forecast resulting from a weighted combination of the forecasts from various models. In the Bayesian paradigm, the weights are usually provided by the posterior model probabilities, which are our "prior" model probabilities (see Hoeting et al. (1999) for a review of Bayesian model averaging), but they need not be. For computational reasons, however, it may be preferable to choose a single model on which to base forecasts. If we were forced to pick a single model to recommend, then on the basis of this analysis of the 2010–2013 data and our initial exchangeable prior beliefs it would be the time-weighted attrition model with ξ chosen sequentially to maximise the log prior predictive probability; over the course of the four years it is preferred in terms of nearly all the aspects of the data that we have looked at. We note that the standard (non-time-weighted) attrition model performs admirably at forecasting all aspects of the data that we have considered, and, due to its computational simplicity, it may also be worthy of recommendation.

5 Summary

We have described and implemented time-weighted versions of several variants of the Plackett–Luce model for permutations and applied these models in a case study involving probabilistic forecasting of Formula One motor racing results. The results in Section 4 confirm the findings of Graves et al. (2003), that the attrition model (reverse Plackett–Luce model) is generally to be preferred to the Plackett–Luce model for modelling motor racing results. Some improvement in forecasts over the standard Plackett–Luce model for features of interest to motor racing fans and bettors (for example the race winner, podium finishers, points finishers and Championship winner) can be gained if the Plackett–Luce model is truncated at results down to rth place. Naturally, the truncated model is poorer than the untruncated model at predicting the full finishing order, and even with truncation, the Plackett–Luce model is inferior to the attrition model. We have also demonstrated that down weighting past results can improve forecasts under the attrition model. The dataset that we have looked at here only consists of four years worth of results and it may be that time-weighted forecasts will have a greater impact when the data covers a longer time period, because non-stationarity is more likely to be an issue.

Supplementary Material

Supplementary Material for "A comparison of truncated and time-weighted Plackett–Luce models for probabilistic forecasting of Formula One results" (DOI: 10.1214/17-BA1048SUPP; .pdf). The Supplementary material contains further details on the Gibbs sampling algorithm of Section 3, details of predictive simulations, an analysis of sensitivity of predictions to prior assumptions and an investigation into optimal choices of the time weighting parameter.

References

Bell, A., Smith, J., Sabel, C. E., and Jones, K. (2016). "Formula for success: multilevel modelling of Formula One driver and constructor performance, 1950–2014." *Journal of Quantitative Analysis in Sports*, 12: 99–112. 335, 336, 338

Benter, W. (1994). "Computer-based horse race handicapping and wagering systems: a report." In Hausch, D. B., Lo, V. S. Y., and Ziemba, W. T. (eds.), *Efficiency of Racetrack Betting Markets*, 183–198. London: Academic Press. 339

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley. MR1274699. doi: https://doi.org/10.1002/9780470316870. 347, 355

Caron, F. and Doucet, A. (2012). "Efficient Bayesian inference for generalized Bradley-Terry models." *Journal of Computational and Graphical Statistics*, 21: 174–196. MR2913362. doi: https://doi.org/10.1080/10618600.2012.638220. 336, 337, 343, 344

Chib, S. (1995). "Marginal likelihood from the Gibbs output." *Journal of the American Statistical Association*, 90: 1313–1321. 354

Dawid, A. P. and Musio, M. (2014). "Theory and applications of proper scoring rules." *Metron*, 72: 169–183. MR3233147. doi: https://doi.org/10.1007/s40300-014-0039-y. 356

Diaconis, P. (1988). *Group Representations in Probability and Statistics*, volume 11 of *IMS Lecture Notes*. Institute of Mathematical Statistics. 338

Dixon, M. J. and Coles, S. G. (1997). "Modelling association football scores and inefficiencies in the football betting market." *Applied Statistics*, 46: 265–280. 344

Eichenberger, R. and Stadelmann, D. (2009). "Who is the best Formula One driver? An economic approach to evaluating driver talent." *Economic Analysis & Policy*, 39: 389–406. 335, 336, 338

Friel, N. and Pettitt, A. N. (2008). "Marginal likelihood estimation via power posteriors." *Journal of the Royal Statistical Society, Series B*, 70: 589–607. MR2420416. doi: https://doi.org/10.1111/j.1467-9868.2007.00650.x. 354

Glickman, M. E. and Hennessy, J. (2015). "A stochastic rank ordered logit model for rating multi-competitor games and sports." *Journal of Quantitative Analysis in Sports*, 11: 131–144. 336, 341

Gneiting, T. and Raftery, A. E. (2007). "Strictly proper scoring rules, prediction, and estimation." *Journal of the American Statistical Association*, 102: 359–378. 347

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: 282–291. MR1995707. doi: https://doi.org/10.1198/016214503000053. 337, 340, 356

Guiver, J. and Snelson, E. (2009). "Bayesian inference for Plackett–Luce ranking models." Proceedings of the 26th Annual International Conference on Machine Learning. 336

Harville, D. A. (1973). "Assigning probabilities to the outcomes of multi-entry competitions." *Journal of the American Statistical Association*, 68: 312–316. 338

Henderson, D. A. and Kirrane, L. J. (2017). "Supplementary Material for "A comparison of truncated and time-weighted Plackett–Luce models for probabilistic forecasting of Formula One results"." *Bayesian Analysis*. doi: https://doi.org/10.1214/17-BA1048SUPP. 343

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). "Bayesian model averaging: a tutorial." *Statistical Science*, 14: 382–417. 356

Hunter, D. R. (2004). "MM algorithms for generalized Bradley–Terry models." *Annals of Statistics*, 32: 384–406. 336, 342

Ibrahim, J. G. and Chen, M.-H. (2000). "Power prior distributions for regression models." *Statistical Science*, 15: 46–60. MR1842236. doi: https://doi.org/10.1214/ss/1009212673. 337, 341

Kass, R. E. and Raftery, A. E. (1995). "Bayes Factors." *Journal of the American Statistical Association*, 90: 773–795. 354, 355

Luce, R. D. (1959). *Individual Choice Behavior*. New York: Wiley. 336, 338

Marden, J. I. (1995). *Analysing and Modeling Rank Data*. London: Chapman and Hall. MR1346107. 336, 338, 340

McCarthy, D. and Jensen, S. T. (2016). "Power-weighted densities for time series data." *Annals of Applied Statistics*, 10: 305–334. 337, 341, 344

Pettitt, A. N. (1982). "Inference for the linear model using a likelihood based on ranks." *Journal of the Royal Statistical Society, Series B*, 44: 234–243. 339

Phillips, A. J. K. (2014). "Uncovering Formula One driver performances from 1950 to 2013 by adjusting for team and competition effects." *Journal of Quantitative Analysis in Sports*, 10: 261–278. 335, 336, 338

Plackett, R. L. (1975). "The analysis of permutations." *Applied Statistics*, 24: 193–202. MR0391338. 336, 338

Silverberg, A. R. (1980). "Statistical models for q-permutations." Ph.D. thesis, Department of Statistics, Princeton University. 339

Smith, R. (2013). *Formula 1: All the Races. The World Championship Story Race-by-Race 1950–2012*. Yeovil, Somerset: Haynes Publishing, second edition. 335, 352

Yellott, J. I. (1977). "The relationship between Luce's choice axiom, Thurstone's theory of comparative judgement, and the double exponential distribution." *Journal of Mathematical Psychology*, 15: 109–144. 339

Acknowledgments

Part of this work stemmed from Liam Kirrane's MMathStat project at Newcastle University. The authors are grateful to Sky BET (www.skybet.com) for providing their odds for the 2013 Formula One Drivers' Championship. The authors would also like to thank the Editor, Associate Editor and Reviewer for comments which improved this paper.