New order-statistics-based ranking models and faster computation of outcome probabilities

Rose Baker
School of Business
University of Salford, UK
r.d.baker@salford.ac.uk

January 14, 2019

Abstract

In sport, order-statistics-based models such as Henery's gamma model and the Thurstone-Mosteller type V model are useful in estimating competitor strengths from observed performance of players in competitions between 2 or more players. They can also be applied in many other areas, such as analysis of consumer preference data, which would be useful to marketing management. Two new families of such models derived from the exponentiated exponential and Pareto distributions are introduced. Use of order statistics-based models when there are more than 2 competitors has been hampered by lack of an efficient method of computation of outcome probabilities as a function of competitor strengths, and a fast method of computation of outcome probabilities is presented, that exploits the fact that the integral to be evaluated is an iterated integral.

Keywords

Ranking model; Plackett-Luce model; Thurstone model; Numerical integration; exponentiated exponential distribution; Pareto distribution.

1 Introduction

Rating and ranking are ubiquitous human activities, occurring everywhere from the daily life of the individual who must prioritize many possible undertakings and consumer products, to organizations, who must rank opportunities and threats, and rate available resources, including 'human resources'. Ranking models can thus be applied in many areas, such as sport, marketing, computing (ranking web pages for search engines), voting etc. (e.g. Alvo and Yu, 2014). The focus is on sport here, but the results are generally applicable.

In statistically-based models of sporting performance, each competitor or team has an (unknown) rating or strength. Similarly, in marketing, consumer preference data can be used to assign a rating to each brand. This article is concerned with the situation where results of competitions are available only as ranks, and not as ratings. Each possible ranking (ordering of scores) of competitors then occurs with a probability that is a function of the competitor strengths, as specified by the ranking model. By fitting the model to data from all available contests, the competitor strengths can be estimated, e.g. by using likelihood-based inference. Thus competitor strengths or ratings can be estimated from whatever performance data are available, and can then be used to predict future results, or to give an overall ranking for the competitors.

This article addresses the problem of calculating the probability of particular rankings in this situation, where ratings or strengths must be deduced from the results of contests. The contribution of the new methodology introduced here is twofold: some new order-statistics-based ranking models are presented, and a faster method of computing the probability of an observed ranking given the competitor ratings is also presented. The two aims of this article are related: the fast method of computation paves the way for the introduction of new ranking models, which would otherwise be unusable. Faster computation is badly needed, as existing methods for evaluating the integral that arises with this class of models, such as Monte-Carlo integration, are far too slow to be usable. Figure 1 shows this graphically: integration beats Monte-Carlo computations for up to about 10 competitors, and then Monte-Carlo methods are faster, but the new method of computation is the only one that is feasible for more than very few players.

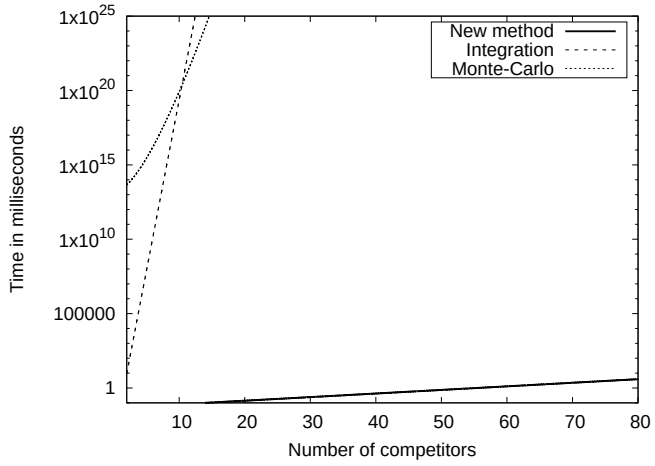

Figure 1: Approximate time in milliseconds on a desktop computer needed for multivariate integration, Monte-Carlo integration, and the new integration method against number of competitors. The x-scale starts at 2.

The relevance to management is obvious for sports managers, who can use this type of methodology to rate players. Sport is indeed the main area where this type of work is being done, partly because a lot of data is available, and also because the model predictions are of interest to bookmakers and bettors (see e.g. Barnett and Clarke, 2005).

Outside sport, marketing managers could use this methodology to rate their own and competitors' products using consumer preference data. One example could be combining product rankings done by various groups. Here for example mobile 'phones might be ranked by several different consumer groups, and the rankings might well not be complete. What rating for one's own and rival products could be deduced from such data? In general, as more and more data become available to decision-makers, the methodology described here will become increasingly relevant to management in general.

The next section introduces the useful class of order-statistics-based ranking models, after which the literature on the topic of ranking models is cited. Next individual models in this class are discussed, the improved method of computation is described, and some examples are given.

1.1 Order-statistics models

A useful class of models is that of order-statistics models, where one could think of a race of runners or horses, the probability of a ranking being the probability of a particular order of finishing times. There is a pdf $f_i(x_i|\alpha_i)$ for the ith player to finish at time x_i, where α_i is the corresponding strength parameter. Equivalently, in golf the lowest score wins, in shot-putting the longest distance wins, so the negative of the distance putted would be used in the model. The variable is referred to as 'time' in this article.

Omitting the strength parameter for simplicity, the probability of the ranking $1, 2, 3 \cdots n$ is the iterated integral

$$p_{123\cdots n} = \int_0^\infty f_1(x_1)\,\mathrm{d}x_1 \int_{x_1}^\infty f_2(x_2)\,\mathrm{d}x_2 \cdots \int_{x_{n-1}}^\infty f_n(x_n)\,\mathrm{d}x_n. \qquad (1)$$

As (1) shows, an iterated integral is one in which the integrand can be integrated over each of the n variables in turn.

This type of model is (naturally) applicable to all kinds of races, also to other types of competitive sport such as stroke play golf, target archery, and esports, and to consumer preference and election data. It can accommodate player/team covariates $\mathbf{z}_i$, e.g. $\alpha_i \propto \exp(\boldsymbol{\lambda}^T \mathbf{z}_i)$, where $\boldsymbol{\lambda}$ is a vector of parameters.

The probability is invariant under any monotonic transformation of the time scale or score. This means that several distributions of finishing time vcould be equivalent, e.g. if using an exponential distribution, the Weibull or Gumbel distributions give identical probabilities. Thus a transformation $x \to x^\beta$ for $\beta > 0$ gives the Weibull distribution, and $x \to \ln(x)$ gives the Gumbel.

Because the time scale can be rescaled without changing $p_{123\cdots n}$, it also follows that only the ratios of the strengths of competitors determine $p_{123\cdots n}$.

1.2 The range of available models

When the pdf $f(x)$ is exponential, a closed-form solution can be derived; this is the Plackett-Luce (PL) model (see e.g. Alvo and Yu, 2014). For a normal pdf, we have the Thurstone-Mosteller (TM) model, and a gamma model due to Henery (1983) interpolates between the PL and TM models. Note that for the TM model, the lower limit of integration in (1) is $-\infty$.

This reverts to zero on using the exponential of the random variable, which follows a lognormal distribution.

Any survival distribution can be used in (1). Hence in this work two other survival distributions were also used, the exponentiated exponential (EE) distribution (e.g. Gupta and Kundu, 2007 and Nadarajah, 2011), and the generalized Pareto (Lomax) distribution (Lomax, 1954). Both of these generalize the PL model.

1.3 Computing the probabilities

Developing and evaluating new models was one purpose of this article. The other was the development of faster methods of evaluating the probability (1).

The probability of a particular ranking of n competitors can be very small: with equal strengths it would be $1/n!$. Hence with 50 competitors, the probability is $\simeq 3.3 \times 10^{-65}$. Such tiny probabilities could never be computed by naïve Monte-Carlo methods (e.g. Christian and Casella, 2010), as one would need to generate more than $\simeq 10^{65}$ realizations of the ranking. Even our (unpublished) attempts to devise 'clever' Monte Carlo methods proved agonizingly slow. Multivariate integration is also very slow. Hence the type of integral evaluation described here is very useful in fitting ranking models to data where there are more than very few competitors. Figure 1 shows this situation, where a relative error of 10^{-8} is aimed for.

The basis of a fast new method is that because (1) is an iterated integral, it does not suffer (much) from the 'curse of dimensionality'. This 'curse' is the fact that, with N function evaluations per variable, an integral of dimension n requires N^n function evaluations. With an iterated integral

like (1) only nN evaluations are required, but we shall see that the number N needs to increase slowly with n to preserve accuracy. Hence the curse of dimensionality survives in a milder form. It was found on using the new method that good results could be obtained up to $n \simeq 70$ or 80.

The next section discusses some old and new survival models. Next, after a more detailed discussion of the integration methodology and the analysis of errors, several order-statistics models are fitted to some golf data previously analysed by Baker and McHale (2015), who used the PL model, and some women's tennis data (Baker and McHale, 2017) and the article concludes with a brief discussion.

2 Survival distribution models

2.1 Current models

The popular Plackett-Luce (PL) model can be derived analytically from (1) on using the exponential distribution and so setting $f(x_i) = \alpha_i \exp(-\alpha_i x_i)$. Henery (1983) and Stern (1990) have proposed a useful generalization of the PL model, where f is a gamma pdf. Writing the gamma pdf as

$$f_i(x_i) = \alpha_i(\alpha_i x)^{\beta_i - 1} \exp(-\alpha_i x_i)/\Gamma(\beta_i), \tag{2}$$

we would usually want to set all the shape parameters β_i to a common value, when the strength parameter would be α_i. Otherwise the inverse of the expected finishing time, α_i/β_i would be a good strength measure. However, in sport there tend to be many players and not always many matches per player, and so it is usually best to have few parameters per competitor. The methodology here does however allow more parameters per competitor, which might well be useful outside sport, e.g. in marketing.

There is no analytic solution for general values of β for $n > 2$. For 2-player games, Stern (1990b) derived and used a closed-form solution, and Baker and McHale (2014) derived a closed-form solution as the incomplete (regularised) beta function, which is widely available.

As $\beta \to \infty$ the distribution becomes lognormal, so that $\ln(X)$ is normally distributed and we obtain the Thurstone-Mosteller type V model (Thurstone, 1927), where

$$f(x_i) = \exp(-(x_i - \psi_i)^2/2)/\sqrt{2\pi},$$

where $\psi_i = -\ln \alpha_i$. This also has no closed-form solution for $n > 2$; for $n = 2$ it is $p_{12} = \Phi(\ln(\alpha_1/\alpha_2)/\sqrt{2})$, where Φ is the normal distribution function. The type IV Thurstone model can also be used, where $X \sim N[\psi_i, \sigma_i^2]$.

The gamma model probabilities for more than two competitors have to date been evaluated only for integer values of β, and summation formulae are given in Henery (1983) and Stern (1990a). Stern (1990a) fitted the $\beta = 2$ model to data on horse-races, and claimed a better fit than for the $\beta = 1$ (PL) model. In particular, the $\beta = 2$ model copes better with horses that often win, but which fall a long way behind unless they are in the forefront. In general, good competitors sometimes perform quite badly.

One can see why the $\beta > 1$ model works better in this case by rewriting the pdf using $y = x^\beta$ and $\gamma = \alpha_i^\beta$. Then $f_i(y_i) = \gamma_i \exp(-(\gamma_i y_i)^{1/\beta})/\Gamma(\beta+1)$. In this form, the longer tail of poor performance than for the exponential distribution can be discerned.

Stern (1990b) fitted gamma models to a variety of 2-player/team sports data, but found little difference in fit with β. Baker and McHale (2017) with a much larger dataset on women's tennis found the optimum value of β to be above 2.

The TM model has been fitted using Monte-Carlo integration (e.g. Alvo and Yu, 2014). Lack of a fast method of computation has limited the use of both models.

2.2 New models

The exponentiated exponential distribution has distribution function $F(x) = (1 - \exp(-\alpha x))^{\beta}$, where $\beta > 0$. Thus the pdf is

$$f(x) = \alpha\beta(1 - \exp(-\alpha x))^{\beta-1} \exp(-\alpha x).$$

This pdf reduces to the exponential pdf when $\beta = 1$, so we regain the PL model. The pdf looks roughly like the gamma pdf, and for small x, $f(x) \simeq \alpha\beta(\alpha x)^{\beta-1}$. It has mode $x_m = \ln(\beta)$. Its properties are broadly similar to the gamma distribution. The survival function can be written down explicitly, which is useful in computing win probabilities when there are only 2 competitors. The exponentiated exponential can be transformed into a monotonically decreasing distribution that is longer tailed than the exponential as using the same transformation for the gamma, i.e. with $y = x^{\beta}$. This distribution can be shown to have a monotonically decreasing pdf and behaves like $\exp(-\alpha x^{1/\beta})$ in the tail.

One can also generalize the PL model by using the generalized Pareto distribution, so that $S_i(x_i) = (1 + \alpha_i x_i)^{-\nu}$, where $\nu > 0$. This is also the Lomax distribution; it becomes the full generalized Pareto on applying an affine transformation to x_i, which would be redundant in this case. Then

$$f_i(x_i) = \frac{\nu\alpha_i}{(1 + \alpha_i x_i)^{\nu+1}}.$$

This pdf is longer-tailed than the exponential. The case $\nu = 1$ is interesting, as this is a special case of the log-logistic distribution, and can be regarded

as analogous to the normal limit of the gamma model. In this case, the distribution is logistic rather than normal. This is not of course a true limit, which occurs when $\nu \to 0$.

2.3 Two-player games and other special cases

When there are 2 players, the PL model reduces to the Bradley-Terry (BT) model; see e.g. Dewart and Gillard (2018). The computation of the gamma probability p_{12} as an incomplete beta function has been mentioned, and the fact that the model is computable for integer β.

The EE model gives

$$p_{12} = \alpha_2 \beta \int_0^\infty \exp(-\alpha_2 x)(1 - \exp(-\alpha_2 x))^{\beta-1}(1 - \exp(-\alpha_1 x)^\beta \, \mathrm{d}x,$$

or on changing variable to $z = \exp(-\alpha_2 x)$

$$p_{12} = \beta \int_0^1 (1 - z^r)^\beta (1 - z)^{\beta-1} \, \mathrm{d}z,$$

where $r = \alpha_1/\alpha_2$. For $\beta = 1$ this reduces to $p_{12} = r/(1 + r) = \alpha_1/(\alpha_1 + \alpha_2)$ as it must. When $\beta = 2$,

$$p_{12} = 1 - \frac{5}{r+1} + \frac{2}{2r+1} + \frac{4}{r+2},$$

or

$$p_{12} = \frac{r^2(2r+7)}{(r+1)(r+2)(2r+1)}.$$

For arbitrary β, the integral can be evaluated by standard methods, but cannot be reduced to a special function. Some fortran code is available in the online supplement for evaluating p_{12} and its first two derivatives. This uses the transformed integral

$$p_{12} = \beta\gamma \int_0^1 (1 - x^{r\gamma})^\beta (1 - x^\gamma)^{\beta-1} x^{\gamma-1} \, \mathrm{d}x,$$

which with $\gamma \simeq 3$ is zero at both limits. The program does trapezoidal integration, followed by two Richardson extrapolations. Good accuracy is obtained with 40 points.

The derivatives are computed by differentiating under the integral sign.

As $\beta \to \infty$ there is a limiting form of the EE distribution. Using the product-limit form of the exponential, we obtain the distribution function

$$F(x) = \exp(-\beta \exp(-\alpha x)).$$

This is the log-Fréchet distribution, defined on the whole real line. Computations for this distribution can be done, but suffer from numerical difficulties and are not discussed further.

Clearly, for integer β, equation (1) can be solved analytically, as the integrand is a sum of exponentials.

For the Pareto model, the probability (1) is given for $n = 2$ by

$$p_{12} = \alpha_1 \nu \int_0^\infty (1 + \alpha_1 x)^{-\nu-1} (1 + \alpha_2 x)^{-\nu} \, \mathrm{d}x,$$

which after a change of variable gives

$$p_{12} = \frac{r\nu}{(r-1)^{2\nu}} \int_1^r \frac{(y-r)^{2\nu-1}}{y^{\nu+1}} \, \mathrm{d}y.$$

If $r < 1$ the formula is still valid, and is equal to

$$p_{12} = \frac{r\nu}{(1-r)^{2\nu}} \int_r^1 \frac{(r-y)^{2\nu-1}}{y^{\nu+1}} \, \mathrm{d}y.$$

In this form one can see that when ν is a multiple of $1/2$, the integral can be evaluated analytically by expanding the numerator of the integrand in a power series.

It seems that an analytic expression for p is not possible for more than 3 competitors.

When using integral expressions for win probabilities with 2 competitors, it is sufficient to compute $\partial p/\partial \alpha_1$, and $\partial^2 p/\partial \alpha_1^2$, from which the corresponding derivatives w.r.t. α_2 can be found: writing $p(\alpha_1/\alpha_2) = g\{\ln(\alpha_1) - \ln(\alpha_2)\}$, clearly

$$\partial p/\partial \ln(\alpha_2) = -\partial p/\partial \ln(\alpha_1)$$

and

$$\partial^2 p/\partial \ln(\alpha_2)^2 = \partial^2 p/\partial \ln(\alpha_1)^2.$$

From this we have that

$$\partial p/\partial \alpha_2 = -(\alpha_1/\alpha_2)\partial p/\partial \alpha_1,$$

$$\partial^2 p/\partial \alpha_2^2 = (\alpha_1/\alpha_2)^2 \partial^2 p/\partial \alpha_1^2 + 2(\alpha_1/\alpha_2^2)\partial p/\partial \alpha_1.$$

Finally, from $\partial^2 p/\partial \ln(\alpha_1)\partial \ln(\alpha_2) = -\partial^2 p/\partial \ln(\alpha_1)^2$, it follows that

$$\partial^2 p/\partial \alpha_1 \partial \alpha_2 = -\alpha_2^{-1}\partial p/\partial \alpha_1 - (\alpha_1/\alpha_2)\partial^2 p/\partial \alpha_1^2.$$

The next topic is the computation of probabilities for these models in the general case.

3 Integration Method for the general n-player case

An algorithm for computing the probability of a ranking will be presented. It has been adapted to cope with some generalizations of the ranking problem, where for example the lower-placed competitors are not ranked. It must be stated at the outset that bookmakers and bettors usually only need to compute odds for various events that occur with a probability much higher than that of a particular ranking. For example, in golf, that a player wins, is in the first 10, or 'makes the cut'. In this case, Monte Carlo simulation will be adequate.

3.1 The algorithm

We can rewrite (1) as:

$$S_n(x) = \int_x^\infty f_n(u)\,\mathrm{d}u,$$

where S denotes a survival function. For i from $n-1$ to 1,

$$S_i(x) = \int_x^\infty f_i(u) S_{i+1}(u)\,\mathrm{d}u,$$

and finally

$$p_{123\cdots n} = S_1(0).$$

Then an algorithm for computing $p_{123\cdots n}$ is as follows:

1. Compute $S_n(x) \quad \forall x$;
2. for i from $n-1$ down to 1 compute $S_i(x) = \int_x^\infty f_i(u) S_{i+1}(u)\,\mathrm{d}u \quad \forall x$;
3. read off $p_{123\cdots n} = S_1(0)$.

The extended trapezoidal rule was used to approximate this procedure numerically, using a transformed distribution $g_i(x)$ defined on the range $0 < X < 1$, exploiting the fact that monotonic transformations of timescale do not change the value of the integral. The rule is

$$\int_0^1 g(x)\,\mathrm{d}x \simeq S(0) = h\{g(0)/2 + g(h) + g(2h) + \cdots g(N-1)h) + g(Nh)/2\},$$

and of course

$$\int_{mh}^1 g(x)\,\mathrm{d}x \simeq S(mh) = h\{g(mh)/2 + g((m+1)h) + g((m+2)h) + \cdots g(N-1)h) + g(Nh)/2\},$$

where $hN = 1$. The algorithm proceeds by computing S_n at transformed times Nh down to 0, then S_{n-1} and so on. At time mh, the sum $T((m+$

$1)h) = h\{\sum_{m+1}^{N-1} S(mh) + S(1)/2\}$ is used with the just computed $S(mh)$, and $T(mh)$ cumulated.

The whole procedure takes Nn operations, so computing time is linear in the number of competitors; we shall see however that N needs to increase with n to preserve accuracy.

Although the algorithm is simple, a consideration of errors is necessary to achieve accurate results. The Euler-Maclaurin summation formula (e.g. Press *et al*, 2007) gives the error of the approximation as an asymptotic series: to second order this is

$$\int_{mh}^{1} g(x)\,\mathrm{d}x - S(mh) \simeq -(h^2/12)(g'(1)-g'(mh)) + (h^4/720)(g'''(1)-g'''(mh)),$$

where primes denote derivatives. Thus the error of a single integration is $O(h^2)$ if f is transformed so that $g(0)$ and $g(1)$ have finite first derivatives. This order of error is preserved through the n multiple (iterated) integrations, because if $S_{i+1}(mh)$ is accurate to $O(h^2)$, a further error of $O(h^2)$ is incurred in evaluating the integrand at each grid point, and so the total further error incurred is $O(Nh^3) = O(h^2)$. Apart from roundoff error, there is an additional error arising from the impossibility of adding very small numbers to large numbers in a computer, resulting from the finite size of the mantissa (e.g. Press *et al* 2007). This last error increases with N.

The procedure is then to compute the required probability, to obtain a result with error $O(h^2)$. Repeating the procedure with $2N$ grid values enables a Richardson extrapolation to be carried out, which reduces the error to $O(h^4)$. All this methodology is quite standard, but has not been applied to iterated integrals.

3.2 Variable transformation

The transformation of the random variable, carried out to keep the derivatives of the pdf finite, must map it into $[0, 1]$. For the gamma distribution pdf (2) a finite pdf at zero and unity is required. The transformation $z = (1 - \exp(-\alpha_0 x))^{1/\gamma}$ was used, so that $x = -\ln(1 - z^\gamma)/\alpha_0$. The pdf (2) becomes

$$f(z) = \Gamma(\beta)^{-1}\gamma(\alpha/\alpha_0)^\beta z^{\gamma-1}(-\ln(1 - z^\gamma))^{\beta-1}(1 - z^\gamma)^{\alpha/\alpha_0-1}. \qquad (3)$$

This is shown in figure 2. This has the required properties; $f(z) \to 0$ as

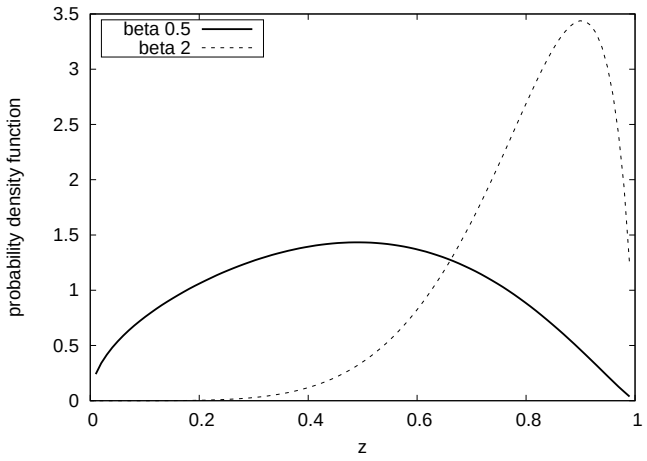

Figure 2: The transformed gamma pdf from (3) with $\alpha = 1, \gamma = 3, \alpha_0 = 1/2$ for $\beta = 1/2$ and $\beta = 2$.

$z \to 1$ if $\alpha > \alpha_0$, and for $z \ll 1$, $f(z) \sim z^{\gamma\beta-1}$. Thus we require α_0 to be (say) half the minimum value of α, and $\gamma > 1/\beta$.

For the TM model, the distribution is $N[-\ln(\alpha_i), 1)]$. The logistic transformation $z = 1/(1 + \exp(-x))$ gives a distribution with support on $[0, 1]$,

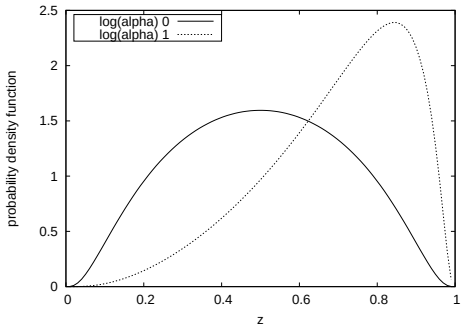

Figure 3: The transformed normal pdf from (4) with $\delta = 1$, and $\ln(\alpha) = 0$, $\ln(\alpha) = 1$.

so that

$$f(z) = \frac{1}{\sqrt{2\pi}} \frac{\exp(-(\ln(z/(1-z)) + \ln \alpha_i)^2/2)}{z(1-z)}. \tag{4}$$

This is shown in figure 3. For the exponentiated exponential distribution, the transformation $x = -\ln(1 - z^\gamma)/\alpha_0$ is used as for the gamma distribution. The pdf is then

$$f(z) = \beta\gamma(\alpha/\alpha_0)z^{\gamma-1}(1-z)^{\alpha/\alpha_0-1}(1-(1-z^\gamma)^{\alpha/\alpha_0})^{\beta-1}.$$

For the Pareto distribution, we take the transformation $x = \alpha_0^{-1} z^\gamma/(1-z)^\delta$. The pdf becomes

$$f(z) = \nu\alpha/\alpha_0) \frac{\{\delta z^\gamma + \gamma(1-z)z^{\gamma-1}\}(1-z)^{\delta\nu-1}}{\{(1-z)^\delta + (\alpha/\alpha_0)z^\gamma\}^{\nu+1}}.$$

3.3 Reducing errors

3.4 Logic errors

There are two kinds of computational error: logical/programming errors, and numerical errors arising from roundoff, etc. Programming errors are dealt with first. The error can be studied for the case where all α_i are equal,

when $p = 1/n!$, or when $\beta = 1$ for the gamma and EE models, when the PL model probability can be computed from the analytic solution. For integer values of β, Henery (1983) and Stern (1990a) give summation formulae for the gamma model. In this work a program was written that kept track of the coefficients of powers of x in the integrand, as the iterative integral was done symbolically. Finally, the required probability is the coefficient of x^0. This enabled the gamma model computation accuracy to be checked for integer values of β. Log-Likelihood differentials were checked for all models by also computing them numerically as differences.

3.5 Numerical errors

Turning to numerical errors, for the gamma and exponentiated exponential models, suitable parameter values are α_0 half the minimum of the α_i, $\gamma = 3/\beta$. For the Pareto model, $\gamma \simeq 3$, α_0 is as before, and $\delta \simeq 3/\nu$. These give low errors; the pdf is zero at $z = 0, z = 1$, and its derivative $\mathrm{d}f(z)/\mathrm{d}z$ is zero when possible.

We use the integral with Richardson extrapolation, to further reduce error. After Richardson extrapolation, the error should be $O(h^4)$. A second extrapolation based on this often works, but sometimes gives no improvement. Hence for reliable accuracy, further extrapolations are not currently recommended. An additional source of error is the rounding error arising from adding a very small number to a much larger number, caused by the finite size of the mantissa. This error gets larger when N increases, and may be a contributory reason why further extrapolations cannot be usefully done.

The proportional error σ_p on the probability p, i.e. $\sigma_p = |p - p_{\mathrm{true}}|/p_{\mathrm{true}}$,

was a useful measure of error. A constant relative error of course is desirable in giving a constant error for the log-likelihood, since $\ell = \ln(p)$, and $\delta\ell = \delta p/p$, so $\mathrm{var}(\ell) \simeq \sigma_p^2$. The proportional error was found to decrease to on average 0.006% of its original value on Richardson extrapolation, so this is definitely worthwhile. The ratio of Richardson-extrapolated error to original error increases slowly with sample size n. The proportional error afrer extrapolation itself increases with n, so that N must be increased to keep this error constant. The rule of thumb $N = 200\exp(n/18)$ (strictly $N = 2[100\exp(n/18)]$, where [] denotes nearest integer) gives an acceptable error up to $n \sim 70$ or 80, and was arrived at by regressing logged relative error on n.

Figure 4 shows the results of applying this to some golf data (discussed later). Each point represents a match. The golf dataset is used because it has a spread of numbers of players and has estimated strengths from the PL model, so it gave a realistic dataset for computing errors. However, randomly generated strengths and rankings could have been used instead with similar results. A test of error can be done using the PL model. Here the probabilities p in (1) can be computed exactly, as the integral can be evaluated analytically.

With this rule of thumb, computation time $\propto Nn$ on a typical desktop computer was 1.4 milliseconds for $n = 20$. For $n = 80$, it has risen to 157 milliseconds. It must be noted that for n much higher than 80 computation time becomes large, and the golf example given later is at the limit of what is feasible on a desktop computer.

Although the 'curse of dimensionality', the exploding number of integrand evaluations needed for a given accuracy, does not apply to iterative integral evaluation, it does still appear in a weakened form, in that the

number of points used N must increase with the sample size n to attain the same relative error on the probability. Also, Richardson extrapolation only reliably worked once; if this problem could be removed, computation could be speeded up further.

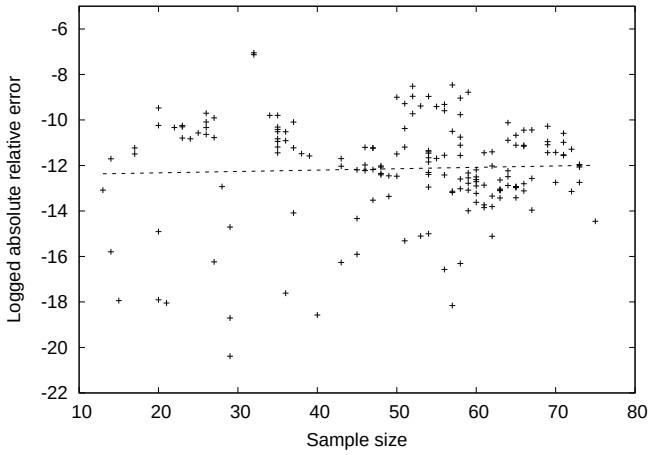

Figure 4: Logarithm of absolute relative error against sample size using the rule of thumb $N = 200 \exp(n/18)$ for the PL model, with fitted line.

3.6 Related computations

3.6.1 When some competitors are not ranked

The m competitors who performed worst may not be ranked. This happens in golf, where a number of players may not 'make the cut'. In consumer preference studies, one might also ask consumers to rank only their n favourite brands. This situation is easy to cope with, because the probability that players $1 \cdots m$ all have score $\geq x$ is simply $\prod_{i=1}^{m} S_i(x)$, where S_i is the ith

survival function. The integral required is

$$p = \int_0^\infty \prod_{i=1}^{m} S_i(x)\,\mathrm{d}x \int_x^\infty f_{m+1}(x_{m+1})\,\mathrm{d}x_{m+1} \cdots \int_{x_{n-1}}^\infty f_n(x_n)\,\mathrm{d}x_n.$$

The computer program requires only an extra array of dimension m.

3.6.2 Computing log-likelihood differentials

Differentials are readily obtained by differentiating (1) outside and under the integral sign. For the gamma model, differentiating the log-likelihood $\ell = \ln(p)$ with respect to α_i yields

$$\partial\ell/\partial\alpha_i = \frac{\beta}{\alpha_i}(1 - p^{(1)}/p),$$

where p denotes the required probability, and $p^{(1)}$ the probability when the ith competitor has a gamma distribution with shape parameter $\beta + 1$. In practice, it is easier to work with $q^{(1)}$, the value of p computed when $\beta \to \beta + 1$ in the integral for the ith competitor, the multiplying constant being unchanged. Then $\partial\ell/\partial\alpha_i = \beta/\alpha_i - q^{(1)}/\alpha_0$. For the TM model, $q^{(m)}$ is the value of p computed with a factor of $(x_i - \psi_i)^m$ in the integrand. The $\partial\ell/\alpha_i = -q^{(1)}/\alpha_i p$.

Higher derivatives may be found similarly, the computation of n partial derivatives thus requiring $n + 1$ evaluations of iterated integrals. It is of course possible to code this more tightly by reusing function values where possible; this would greatly reduce computing time.

3.6.3 Tied observations

When the worst-performing competitors are not ranked, we have the situation described in section 3.6.1 which is easily dealt with. However, when there are ties among competitors but not at the bottom of the list, this

is a difficult computational problem. One can take a tie as meaning that we do not know the ordering of the m tied competitors. The probability that m finishing times for competitors s to $s+m-1$ are $\in (x,u)$ is $G(x,u) = \prod_{i=s}^{s+m-1}\{F_i(u) - F_i(x)\}$, where F denotes distribution function, and hence with scores s to $s+m-1$ tied, the probability integral is

$$p = \int_0^\infty f_1(x_1)\,\mathrm{d}x_1 \cdots \int_{x_s}^\infty \{\,\mathrm{d}G(u,x_s)/\,\mathrm{d}u\}\,\mathrm{d}u \int_u^\infty f_{s+m+1}(x_{s+m+1})\,\mathrm{d}x_{s+m+1} \cdots \int_{x_{n-1}}^\infty f_n(x_n)\,\mathrm{d}x_n.$$

Because one of these factors contains x_s and u and so is bivariate, the iterated integral approach cannot be used.

One can instead compute p for each of the $m!$ orderings and sum. When there are many tied observations,this is computationally expensive and the best that can be suggested is to use 'maximum simulated likelihood', i.e. to generate a large number M of realizations of the ordering of the tied observations, and give each resulting contribution to the log-likelihood a weight of $1/M$.

It is also possible to use a simple approximation. When applied to the PL model, it is an approximation of Efron (1977). Here we replace each of the tied α_i by their average. With this approximation, each possible breaking of the tie yields the same probability p, so for m ties the probability is simply $m!p$. Note that the formula for log-likelihood derivatives will then change. This approximation makes it possible to treat games like golf, where the discrete score causes many ties.

3.6.4 Code available

Probabilities for the 4 models (PL, gamma, EE and Pareto) are computed in a prototype program effprog.f90, a fortran program, available in the online materials. This program elicits model type and parameter values from the

user, and computes probabilities for a dataset of competitor strengths. The computation is done by calling `organize_ints`, which organizes the computation, and calls `prelim_calcs` to set up parameters and arrays that are used n times but need only be computed once. Routine `getint` is called to compute the integral, and it calls function `eff` to compute the pdf. Finally, `organize_ints` calls `getint` again n times, to compute quantities needed for the first and second log-likelihood derivatives for each competitor, and calls `get_diffs` to compute them. These are computed by differentiating under the integral sign, and the differentials are compared with the (less accurate) differentials derived by differencing. Many users will not require the log-likelihood differentials, and this section of code is easily omitted. The program keeps widely-used variables and arrays in a module that is invoked by all routines.

Routine `organize_ints` also calls `getpl` for the Plackett-Luce probability if the model reduces to the PL model, and prints out errors.

4 Examples

The speeding up of computation has already been demonstrated, and this section explores the fitting of different models to sports data. One model, the gamma model, is an old model that could not hitherto be fitted when there are many competitors, and the EE and Pareto models are new. The aim was to assess the various models in realistic situations, a task impossible before, when there was no feasible way of fitting them to large datasets. Readers may wonder what effect fitting a more flexible model has on the actual results of the exercise, i.e. the player rankings, but it is not feasible for many reasons (e.g., of space) to present such results. Many small changes

in rankings, mainly of middle-rank players, follow from changing the ranking model.

Baker and McHale (2015) rated male golf players using data from the four major competitions (the US masters, the US Open, the British Open and the USPGA Championship). The PL model was used to model the probabilities of the observed rankings, using time-dependent strengths; there was no computationally-feasible alternative. There were in all 279 competitions and 822 players. The maximum number of players per match was 46. The data were refitted, using the gamma, EE and Lomax models instead of the PL model. Computations had to be somewhat chopped down, even so taking hours of computer time. Ties were dealt with using the simple approximation from Efron (1977) discussed in section 3.6.3.

The aim here was to see how the model fit changed as the extra parameter varied, for the gamma, EE and Pareto models. A subsidiary aim was to test the computation methodology by using it in a real application, and it was improved and speeded up as a result.

The new programming required was to change the routine that delivered the log-likelihood and its derivatives. The data-fitting program maximises the log-likelihood for player strength parameters using first and second derivatives of the log-likelihood with respect to each of the strength parameters. The formulae mentioned in section 3.6.2 enable these to be computed. Some players 'fail the cut' and effectively tie at the bottom of the rankings. This was dealt with as described in section 3.6.1. The optimum profile log-likelihood was computed at several values of the model parameter, from which the curves in figure 5 showing the profile likelihood were drawn. For the gamma model curve, the TM model was also fitted; this corresponds to infinite β. In figure 5 the 'inverse parameter' is used as

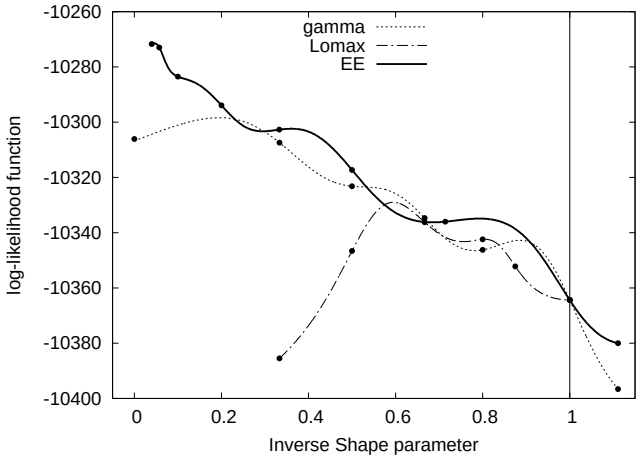

Figure 5: The profile log-likelihood for the golf data example, showing points where the log-likelihood was evaluated for the gamma, exponentiated exponential and Pareto models, linked with a smooth curve.

the abscissa. This is $1/\beta$ for the gamma and EE distributions, and $1 - 1/\nu$ for the Lomax model. This ensures that the PL model, where $\beta \to 1$ or $\nu \to \infty$ has unit abscissa. The roughness of the curves arises from the huge amount of iterative computation required to maximise likelihood functions when there are hundreds of parameters to be estimated.

The profile log-likelihood of the model fits plotted against $1/\beta$ are shown in figure 5, along with an interpolatory curve. When $\beta = 1$, the gamma, EE and Pareto models all reduce to the PL model. The statistical significance of 'floating' β can be judged because twice the increase in log-likelihood at maximum likelihood be distributed as a chi-squared with 1 degree of freedom.

From this argument it can be seen that all three models can significantly improve the fit to the data. Thus the Lomax model increases log-likelihood

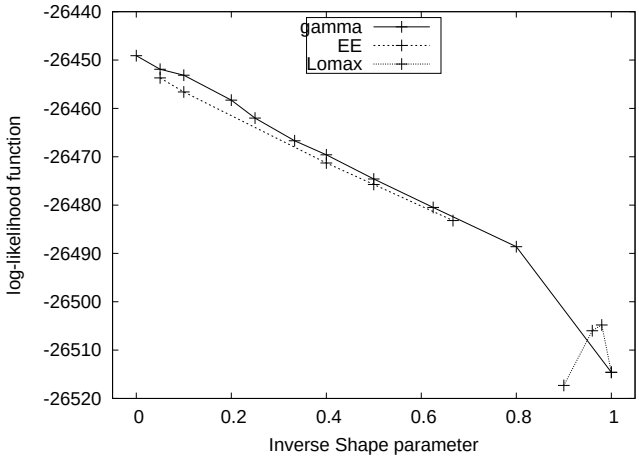

Figure 6: The profile log-likelihood for the women's tennis data example, showing points where the log-likelihood was evaluated for the gamma, exponentiated exponential and Lomax models, linked with a smooth curve.

by 28 at maximum, so that a chi-squared test would give a chi-squared of 56 with one degree of freedom—7.5 standard deviations.

For this example, the gamma model did well, but the EE model even better. The parameter β certainly becomes very large. As discussed in section 2.3, the limit of the EE model is the log-Fréchet distribution, which is difficult to compute.

Baker and McHale, (2017) obtained data on the results of womens tennis matches in the four Grand Slams: the Australian Open, the French Open, Wimbledon and the US Open, for the Open Era of tennis, from 1968 to the Australian Open in 2016. The results of a reanalysis are shown in figure 6. Again, the inverse β parameter is used, and the point at zero for the gamma model is the fit from the normal (Thurstone type V) model. The EE line is slightly below the gamma line, while the Lomax curve is the small peak on

the right.

It can be seen that the gamma and EE models give very similar fits, with the best fit coming from the Thurstone model. However, the improvement in log-likelihood from the BT model is only 65.5. The chi-squared improvement would be 131 with 1 degree of freedom, so there is no doubt that increasing β leads to better fits. However, there are 20552 matches and 85132 sets, so the improvement in player win probability per set is tiny. A crude calculation shows the increase in predicted probability when there is a win to be 0.0004. These results in fact bear out Stern's finding (Stern, 1992) that all paired comparison models behave similarly. With a large dataset, one can now see that for tennis, the TM model is slightly better than the BT model.

The Lomax distribution performs poorly, giving only a very modest improvement over the BT model for β around 25.

5 Conclusions

Two new order statistics-based models have been introduced, and a fast method of accurate computation has been given for order-statistics-based ranking models. Code to compute the probability of a ranking for these models is available online in the supplementary materials for this paper.

The two innovations together make a substantial advance in the state of the art of ranking in sport. The latter should certainly prove useful to practitioners, because in estimating player strengths from rank data, the gamma model with general shape parameter has hardly been used, and the TM type V model has been used only via Markov-chain Monte-Carlo (MCMC). Future work on the fast computation method could include a more detailed error analysis, that might enable further Richardson extrapolations to be

done for the case of more than 2 players, so speeding up the computation. The provision of an R package would of course also be useful.

The new models and the method of computation would be of only academic interest if the Plackett-Luce model were always the best-fitting one, because the PL model computations can be done explicitly without evaluating an integral. However, it has been shown here that in golf tournaments, the TM model fits the data better than the PL model, and the exponentiated exponential model with high β fits best of all. The reanalysis of women's tennis data from Baker and McHale (2017) shows that the gamma model and the exponentiated exponential models perform very similarly, both fitting better than the Bradley-Terry model (the 2-player version of the PL model), with the TM model giving the best fit of all. These results confirm those of Stern (1990a), that because usually good players can sometimes perform quite poorly, the gamma model with $\beta > 1$ or the TM model fit the data better. Of course, what really matters is the accuracy of the predicted ranking, not how well the model fits, but as ever in Statistics, the best-fitting model, allowing for the number of fitted parameters, is expected to give the most accurate predictions.

The model based on the exponentiated exponential distribution offers a useful alternative to the gamma model, which could be used for sensitivity analysis etc.. The Lomax model performed poorly for both tennis and golf data, but cannot be completely ruled out as a potentially useful model. Much further work on such models could be done.

References

References

[1] Alvo, M. and Yu, P. L. H. (2014), Statistical methods for Rank data, Springer, New York.

[2] Baker R. D. and McHale I. G. (2015), Deterministic evolution of strength in multiple comparisons models: who is the greatest golfer?, *Scandinavian Journal of Statistics*, **42** (1), 180-196.

[3] Baker R. D. and McHale I. G. (2014). A dynamic paired comparisons model: who is the greatest tennis player?, *European Journal of Operational Research*, **236** (2), 677–684.

[4] Baker, R. D. and McHale, I. G. (2017) An empirical Bayes model for time-varying paired comparisons ratings: who is the greatest women's tennis player? *European Journal of Operational Research*, **258** (1), 328-333.

[5] Barnett, T. and Clarke, S. R. (2005), Combining player statistics to predict outcomes of tennis matches, IMA Journal of Management Mathematics, **16**, 113120.

[6] Christian, R. and Casella, G. (2010). Monte Carlo Statistical Methods, Springer, New York.

[7] Dewart, N. and Gillard, J. (2018), Using Bradley-Terry models to analyse test cricket, IMA Journal of Management mathematics, published online, dpx013.

[8] Efron, B. (1977), The efficiency of Cox's likelihood function for censored data, *Journal of the American Statistical Association*, 72, 557-565.

[9] Gupta, R. D. and Kundu, D. (2007). generalized exponential distribution: existing results and some recent developments, *Journal of Statistical Planning and Inference*, **137**, 3537–3547.

[10] Henery, R. J. (1983), Permutation probabilities for gamma random variables, *Journal of Applied Probability*, **20** (4),822–834.

[11] Lomax, S. K. (1954), Business Failures: another example of the analysis of failure data *Journal of the American Statistical Association*, **49** (268), 847-852.

[12] Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis.* Wiley, New York.

[13] Nadarajah, S. (2011). The exponentiated exponential distribution: a survey, *AStA Advances in Statistical Analysis*, **95** (3), 219251.

[14] Plackett, R. L. (1975). The analysis of permutations, *Applied Statistics*, 24, 193-202.

[15] Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (2007). *Numerical Recipes, 3rd ed.*, Cambridge University Press, Cambridge.

[16] Stern, H. (1990a). Models for Distributions on Permutations, *Journal of the American Statistical Association*, 85, 410, 558-564.

[17] Stern, H. (1990b) A continuum of paired comparison models, *Biometrika*, **77** (2), 365-373.

[18] Stern, H. (1992) Are all linear paired comparison models empirically equivalent? *Mathematical Social Sciences* **23** 103-l17.

[19] Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Reviews* **34**, 273-286.