A Ubiquitous Framework for Statistical Ranking Systems

Harry Spearing, B.Sc.(Hons.), M.Res

Submitted for the degree of Doctor of Philosophy at Lancaster University.

June 2023

STOR-i
excellence with impact

Abstract

Ranking systems are everywhere. The thesis will often select sports as its motivating applications, given their accessibility; however, schools and universities, harms of drugs, quality of wines, are all ranked, and all with arguably far greater importance. As such, the methodology is kept necessarily general throughout. In this thesis, a novel conceptual framework for statistical ranking systems is proposed, which separates ranking methodology into two distinct classes: *absolute systems*, and *relative systems*.

Part I of the thesis deals with absolute systems, with a large portion of the methodology centred on extreme value theory. The methodology is applied to elite swimming, and a statistical ranking system is developed which ranks swimmers, based initially on their personal best times, across different swimming events. A challenge when using extreme value theory in practice is the small number of extreme data, which are by definition rare. By introducing a continuous data-driven covariate, the swim-time can be adjusted for the distance, gender category, or stroke, accordingly, and so allowing all data across all 34 individual events to be pooled into a single model. This results in more efficient inference, and therefore more precise estimates of physical quantities, such as the fastest time possible to swim a particular event.

Further increasing inference efficiency, the model is then expanded to include data comprising all the performances of each swimmer, rather than just personal bests. The data therefore have a *longitudinal* structure, also known as *panel data*, containing repeated measurements from multiple independent subjects. This work serves as the first

attempt at statistical modelling of the extremes of longitudinal data in general and the unique forms of dependence that naturally arise due to the structure of the data. The model can capture a range of extremal dependence structures (asymptotic dependence and asymptotic independence), with this characteristic determined by the data. With this longitudinal model, inference can be made about the careers of individual swimmers - such as the probability an individual will break the world record or swim the fastest time next year.

In Part II, the thesis then addresses relative systems. Here, the focus is on incorporating *intransitivity* into statistical ranking systems. In transitive systems, an object A ranked higher than B implies that A is expected to exhibit preference over B. This is not true in intransitive systems, where pairwise relationships can differ from that which is expected from the underlying rankings alone. In some intransitive systems, a single underlying and unambiguous ranking may not even exist. The seminal Bradley-Terry model is expanded on to allow for intransitivity, and then applied to baseball data as a motivating example. It is found that baseball does indeed contain intransitive elements, and those pairs of teams exhibiting the largest degree of intransitivity are identified. Including intransitivity improves prediction performance for future pairwise comparisons.

The thesis ultimately concludes by harmonising the two parts - acknowledging that in reality, there is always some relative element to an absolute system. Forging the armistice between these system types could enflame research into the areas connecting them, which until now remains barren.

Acknowledgements

It didn't take long before my supervisor, the well-respected distinguished professor Jonathan Tawn, became, well, Jon. Your inappropriateness, risky jokes and general disregard for social norms helped keep me at ease. Thanks for the levelling mockery, and for the exquisite hiking tips. You've taught me so much, both technical and most certainly not. On a more serious note, for your integrity and generosity I can only express the utmost respect. Thank you for going above and beyond what can reasonably be expected from a supervisor. It's been a pleasure.

To my supervisors at ATASS: Dave, Tim and Grace. Thanks for your helpful feedback, your welcoming and supportive spirit, and your visible passion throughout the project.

A PhD can seem a lonesome endeavour, so a sense of community can make a big difference. Accordingly, I thank those involved at STOR-i, for the invaluable source of friendship, laughter, and unpredictability. The uniqueness on display was frankly and refreshingly eye-opening.

Thanks to the admin team, for all your hard work in organising away days and events, and for your masterful command of the daily trials and tribulations, to which I'm sure I made my contribution... Especially, thanks to: Kim, for your quiet sarcasm and grounding incredulity; Jen, for the many cherishable and oh so meaningfully-meaningless chats - I hope you're still triple checking the microwave settings; and to Wendy, for the friendly shoulder, the piss taking, and your infectious confidence. I wish

you all the best.

I've been lucky to make important friendships along the way. In particular, I would like to mention:

Jess Gillam, for your inexorable kindness and acceptance, at my lowest points, and at all the other points, too. I never doubt whether I can be myself in your presence. If the world would have more Jess Gillam's, a better world it would be; though I concede that there can only ever be one. Perhaps that's why it's special.

Flexi-styles pays their gratitude to Kathryn Turnbull - a source of silliness, wisdom, fond memories, and good food. And, from a rain splattered windshield from a passenger seat, thanks for unexpectedly illuminating the shadows at the very darkest of times.

To my live-in tartan pal, Carla Pinkney. For flying the flag of our authentic and musical home of ambiance, porridge, and shite TV.

An Sophia, mein Neunmalklug, danke, dass du mir zur Seite gestanden hast, für die Zweisamkeit. Thanks for patching me up after the inevitable, with an undiluted disinfectant in your hand, and an I-told-you-so smile on your face; for the dances, the hikes and the swims.

To Jack Baker, Harjit Hoolay, David Sanchez, and Jake Clarkson: you showed me the lighter, brighter notes; that a PhD mustn't always be a serious venture. Jack, your flatulence regularly redefines my notion of pungency. Harjit, your unquenchable curiosity and outrageous honesty has led me to question whether it really is the best policy. David, you showed me that a well-balanced diet can be found in a wheelie bin. Jake, what started out as a car-share to our favourite budget supermarket, became an endless stream of recounting dreams, of ceaseless stories, bottomless rabbit-holes, infinite tangents, that go on and on, and on, forevermore. Thanks for listening to the problems I'd be too embarrassed to tell anyone else; for seriously discussing my latest theories, which others dismiss as the ramblings of delirium; and for sharing the passion for my ever-changing crazes, where others simply miss the point. Your unique humour

rarely fails to amuse me; your delicate balance of sheer logic and insanity, to bemuse me. Your peculiarities are your strengths, and they're a privilege to be a part of. As a pack, you offered a home away from home, and you remain an ever-stable supply of surprise.

Those with an experience of PhD life will be aware that the journey can be just as much personal as academic. As such, it wouldn't feel right to miss out my queer pals and allies I have met along the way. Thank you for helping give me the courage to be me.

Dressing up at the Christmas meal, and singing karaoke at the summer ball, make up just some of the many highlights, some too shocking to be written. Thanks to all those who have accompanied me: for a quiet drink or a mountain bike trail; for a heated debate or a swim in a lake; for those impulsive Monday night antics, and those peacefully shared silences; for a comforting word, or a coffee conversation; for the hugs, and the tickles. You know who you are.

If I'd have got nothing else out of my time here at Lancaster, then it'd still have been worth it. As it turns out, I wrote a thesis too.

Despite the hardships, I remain incredibly lucky. In fact, from a purely statistical standpoint, I'm lucky to have been born at all... It's fair to say that none of this thesis would exist at all without my family. But I mean that in more than the obvious biological sense, and for that I express my unbounded gratitude. You've been my ever-present safety net. In particular, to my big brother, Joe, and to my mum and dad: you've been with me on this mathematical journey from the very beginning. From desperately attempting Joe's maths homework, to calculating the time it will take to get to Cornwall (before the invention of Google Maps). Thank you for the gentle encouragement, the loving guidance, the unwavering loyalty.

And although here my acknowledgements formally conclude, my thanks, of course, go on.

Dedication

To Mum and Dad, for giving me the support and the freedom to find my own way.

It would be a present to you, my sweet, if it weren't your gift to me.

- A.A.Milne

Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

Chapter 3 has been published as Spearing, H., Tawn, J. A., Irons, D., Paulden, T., and Bennett, G. (2021). Ranking, and other properties, of elite swimmers using extreme value theory. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(1):368–395.

Chapter 4 has been submitted for publication as Spearing, H., Tawn, J. A., Irons, D., and Paulden, T. (2023). A framework for statistical modelling of the extremes of longitudinal data, applied to elite swimming.

Chapter 7 has been published as Spearing, H., Tawn, J. A., Irons, D., and Paulden, T. (2023). Modeling intransitivity in pairwise comparisons with application to baseball data. *Journal of Computational and Graphical Statistics*, 0(0):1–10.

The word count for this thesis is approximately 52968.

Harry Spearing

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Pick your favourite sport. I'm sure you could stew over the world's best competitor at this current moment, or of all time, or who would claim victory if A played B. But is it conceivable that a system could be developed which returns an objective answer to these seemingly subjective puzzles? In developing such systems, it is crucial to capture as much information as is possible from our dynamic world. Athletes' injuries, weather events, even economic factors all impact the outcome of these events and the implied abilities of the athletes or teams. Ranking systems in sport use a wide range of strategies in order to capture these signals, from graph theory to extreme value theory, so that the "best" system of ranking sports teams or athletes is formulated. Ranking systems in sport are not only interesting to the inquisitive fan, as particular motivation arises from companies such as ATASS Sports, who develop innovative methods of ranking and rating sports teams and athletes. After all, a fair and accurate system is at the core of all sports organisational bodies and the multi-billion pound conglomerates they represent.

But these systems are not exclusive to sports. Methodological advances in the field of ranking systems have far-reaching consequences: ranking systems prioritise certain web-pages, influence schools and hospitals, and even declare the most essential medical treatments. Poor methodology here leads to far harsher repercussions than incorrectly seeding a tennis tournament...

Progress league tables were introduced for schools in the UK as a fairer and more meaningful way to compare the effectiveness of schools (Leckie and Goldstein, 2017). The measure of school progress has changed several times: the initial "value-added" measure was criticised because it failed to take differences in pupils' socio-economic backgrounds into effect (Office, 2003) and there was no attempt at uncertainty quantification.

These issues were then rectified by the introduction of the "contextual-value-added" measure. This included age, gender, ethnicity, socio-economic status, and other factors as covariates in a random-effects model. From this statistical model, uncertainty quantification was then available. However, the contextual-value-added measure was subsequently dropped as it was considered too difficult to understand, and was "a poor predictor of success" (Department for Education, 2010).

This highlights an important point - the *aim* of a ranking system completely shapes its design, and the methodology used to construct it. In our example, it is clear that the aim of ranking schools is not agreed upon: is it to promote the mental, physical, and spiritual well-being and development of all pupils? Or is it to correctly forecast pupils' qualifications? The result is a system which should be accurate, a good predictor of future events and have reliable uncertainty quantification yet fair and easy-to-understand. Not an easy task.

Due to funding or sanctions that often accompany these rankings, schools tend to

change their behaviour by, for example, "teaching for exams." Initially, pupils' qualifications may have contributed to evidence of their development; however, it now merely indicates their ability to memorise facts. Hospital ranking systems are also incentive inducing, yet the introduction of these rankings within the National Healthcare Service (NHS) is negatively correlated with the quality of care (Propper et al., 2004). Incentive-fuelled ranking systems can therefore cause a change in behaviour through interactions between the rankings and the "players", thus rendering the rankings inaccurate.

NHS ranking systems use a combination of metrics, such as: hospital standardised mortality ratio, a measure of in-hospital deaths; Summary Hospital-level Mortality Indicator, which measures deaths occurring in the 30 days after being discharged; deaths after surgery; and deaths in low-risk conditions. The aim is to provide a better view of overall performance than any single metric; however, the rankings can be sensitive to the choice of aggregation. In fact, hospitals can move almost half of the league table as a result of subtle differences in the aggregation (Jacobs et al., 2005).

It may appear logical, that one common primary aim of a ranking system is to provide a single order of preference for a set of objects; yet, a single ordering of preference may not exist at all. Consider an object A, which is preferred to B, which is preferred to an object C and C is in turn preferred to A. This cycle of preference is an example of *intransitivity*, and in this scenario the interpretation of the rankings is not as transparent. Intransitivity arises in artificial constructions, such as dice games (De Schuymer et al., 2003), but also emerges throughout the natural world, for example, in competition between bacteria (Reichenbach et al., 2007) and in mating choices of lizards (Sinervo and Lively, 1996). It is still possible to form an overall ranking where better ranked objects perform better on average, but there may be specific pairs for which the worse ranked object is expected to express preference over the better ranked object. Alternatively, if the objects are indistinguishable as being statistically significantly different, then enforcing some illusory ranking only engenders a fictional

ordering. The bulk of ranking methodology pays no attention to either of these possibilities.

The appropriate ranking methodology depends on the system at hand. It may be helpful to introduce two distinct categories: *absolute* systems, and *relative* systems.

An *absolute system* is context-free, that is, isolated events are meaningful. For example, the 100m sprint event could be treated as an absolute system, because the knowledge that a person runs it in, say, 9.6 seconds conveys information about this person's ability, irrespective of the context or the opponents. This judgement arises due to some vague perception of the bio-mechanics at play and a loose idea of the typical behaviour of the population.

The surface temperature of stars can also be deemed an absolute system. The astrophysics which govern this system inform us that a star such as *WR 102* with a surface temperature of $210,000K$ (Sander et al., 2012) is, technically speaking, "bloody hot", without the need for direct comparison with any other star. Alternatively, "bloody hot" can be quantified via knowledge of the typical surface temperature, or the distribution of surface temperature across stars. Absolute systems are therefore understood through the physical or statistical properties of the system. Theoretically, the same is true of sprinting. It is possible that an omniscient understanding of bio-mechanics and physics is sufficient to quantify a "fast" sprint time, though with our current level of knowledge, a statistical framework feels more appropriate.

On the face of it, ranking objects in an absolute system, such as runners in a given competition, may seem banal; simply compare their times. But this simple comparison no longer holds when comparing between: different distances; people born of differing sexes, which can lead biological biases, for example, disparate levels of testosterone; fairly adjusting for disabilities; hurdles vs. no-hurdles; let alone differences in equipment such as Eliud Kipchoge's staggering world record set with the controversial Alphafly

shoes. The dynamic component of these systems thickens the plot. Due to improvements in training methods and changes in technology, a fair comparison between run times recorded many years apart is infeasible without some adjustment for the era of the run (Stephenson and Tawn, 2013).

Even when comparing times under identical conditions, applying ranking methodology is beneficial because comparisons between runs have a more tangible interpretation as they can be described in terms of probabilities. For example, it is natural to compare the relative quality of two run times t_1 and t_2 in an event by $\Pr\{T < t_1\}/\Pr\{T < t_2\}$, where T represents the random variable corresponding to a run time for an event, rather than, say, the metric $t_1 - t_2$. A by-product of a suitable ranking system is that other features of interest can be estimated. The ultimate possible run time for any given event can be estimated (Arderiu and de Fondeville, 2022), or the probability of observing a star with some extreme surface temperature $\tau > 210,000K$. In Chapter 3, with the aid of extreme value theory, a statistical framework to rank elite swimmers is presented.

A *relative system*, on the other hand, requires context for any sense of meaning. Armed with the knowledge that a player wins a tennis match, no amount of aerodynamics and general relativity can help infer the ability of this victor. This information is irrelevant without the context - the opponent. Context in relative systems stems from variation of adversaries. An absolute system could be viewed as a competition against nature. Certain natural laws make running a marathon under two hours, observing a star hotter than $210,000$K or an earthquake with magnitude larger than 9.5 unlikely. In all these systems, the adversary - nature - is the same for all objects in the system. Inversely, in relative systems the adversary can be different for each object. Omitting the context of the adversary then is unfair at best, and at worst renders comparisons illogical. Differing adversaries also motivate differing strategies, complicating matters further. The competition is therefore different for each object, thus stipulating context.

The distinction between relative and absolute systems aids the modelling process,

but of course, there is a blending of the two in reality. Knowledge of someone's marathon time alone conveys a lot about their ability, but even in marathons there is some level of strategy. In the absence of context - their competitors - the time cannot tell the whole story. This thesis explores ranking methodology for both classes of system.

1.2 Outline of Thesis

Both relative and absolute systems are deliberated in this thesis, which consists of two parts, labelled Part I and Part II. The former is dedicated to absolute systems, using extreme value theory as a framework, and applied to elite swimming. The latter concerns relative systems, specifically paired-comparison methodology applied to baseball.

Whilst a common output in both systems is a global ranking, the mathematical modelling and application-specific considerations diverge significantly. The strategy of objects can be critical when modelling relative systems, but in an absolute system this is obsolete by definition. Accordingly, each part contains its own literature review and background research, conclusions, and further work, and appendices. Concluding remarks relevant to both parts and a shared bibliography are located at the end of the thesis.

The outline of Part I is as follows. As described above, absolute systems benefit from a statistical framework when quantifying the meaning of observations, and here we utilise extreme value theory; a background is provided in Chapter 2.

Chapter 3 applies univariate extreme value theory to elite swimming. The work knits together multiple systems by consolidating all elite swim events over distance, stroke and gender into a single unified system, forming one unique ranking over all swimmers. The International Swimming Federation (FINA) uses a very simple points system with the aim to rank swimmers across all swimming events. The points acquired is a function of the ratio of the recorded time and the current world record for that event. But with some world records considered "better" than others, bias is introduced between events, with some being much harder to attain points where the world record is hard to beat. A model based on extreme value theory is introduced, where swim-times are modelled through their rate of occurrence, and with the distribution of the best times following a generalised Pareto distribution. Within this framework, the strength of a particular swim is judged based on its position compared to the whole distribution

of swim-times, rather than just the world record. This work embraces the dynamical aspect of absolute systems, where identical events occurring at different points in time have radically different interpretations. As training methods improve over the years, as well as changes in technology, such as full body suits, the date of the swim must be accounted for. The parameters of the generalised Pareto distribution, for each of the 34 individual long course events, will be shown to vary with covariates, leading to a novel single unified description of swim quality over all events and time. This structure, which allows information to be shared across all strokes, distances, and genders, improves the predictive power as well as the model robustness compared to equivalent independent models. A by-product of the model is that it is possible to estimate other features of interest, such as the ultimate possible time, the distribution of new world records for any event, and to correct swim-times for the effect of full body suits. The methods are illustrated using a dataset of the fastest 500 personal-best swim-times for each event in the period 2001-2018.

Chapter 4 extends this work into the multivariate domain in order to capture not just personal-best data, but the whole history of each swimmer's measurements. Using more data allows for more precise forecasting and prediction of future world records and the general state of the system. The methods developed are kept general, using multivariate extreme value theory for analysing observations in the tails of longitudinal data, i.e., a data set consisting of a large number of short time series, which are typically irregularly and non-simultaneously sampled, yet have some commonality in the structure of each series and exhibit independence between time series. Although extreme value theory has been developed for ever more rich data structures, the unique features of longitudinal data have not been considered previously. Across time series the data are assumed to follow a common generalised Pareto distribution, above a high threshold. To account for temporal dependence of such data we require a model to describe (i) the variation between the different time series properties, (ii) the changes in distribution

over time, and (iii) the temporal dependence within each series. Our methodology has the flexibility to capture a range of dependence structures within the extremes of such data (asymptotic dependence and asymptotic independence), with this characteristic determined by the data. Bayesian inference is used with MCMC techniques to help address the need for inference of parameters that are unique to each of the time series. The novel methodology is illustrated through the analysis of data from elite swimmers in the men's 100m breaststroke. Unlike previous analyses of personal-best data in this event, inference about the careers of individual swimmers is available - such as the probability an individual will break the world record or swim the fastest time next year.

Chapter 5 summarises Part I and proposes extensions to the works in Chapters 3 and 4.

The outline of Part II follows thusly. In the relative systems domain, ranking methodology is dominated by paired-comparison approaches. Chapter 6 reviews the paired-comparison literature. Particular attention is paid to the Bradley-Terry model and its subsequent adaptations.

Chapter 7 extends the Bradley-Terry model. The seminal Bradley-Terry model imposes a deeply restricted form of transitivity, namely, that the probabilities of object A beating B and B beating C designate the probability of A beating C, with these probabilities determined by a skill parameter for each object. Such transitive models do not account for different strategies of play between each pair of objects, which gives rise to *intransitivity*. Various intransitive parametric models have been proposed but they lack the flexibility to cover the different strategies across n objects, with the $O(n^2)$ values of intransitivity modelled using $O(n)$ parameters, whilst they are not parsimonious when the intransitivity is simple. Their lack of adaptability is overcome by allocating each pair of objects to one of a random number of K intransitivity levels, each level representing a different strategy. This novel approach for the skill parameters

involves having the n objects allocated to a random number of $A < n$ distinct skill levels, to improve efficiency and avoid false rankings. Although up to $O(n^2)$ unknown parameters may have to be estimated for (A, K), it is anticipated that in many practical contexts $A + K < n$. The Bradley-Terry class of model is revealed to be a special case, when $(A = n - 1, K = 0)$, of this broader class of semi-parametric model which encompasses intransitivity - the Intransitive Clustered Bradley-Terry model. This new model is shown to have an improved fit relative to the Bradley-Terry, and the existing intransitivity models, in out-of-sample testing when applied to simulated and American League baseball data.

Chapter 8 suggests an array of extensions for this broader class of model, in particular the optimal choice of constraint when applied to sports with complex tournament structure.

Chapter 9 provides concluding statements from both parts, and suggests further work which combines the methodology for absolute and relative systems into one unifying framework.

Part I

Absolute Systems

Only a Sith deals in absolutes.

— Obi-Wan Kenobi

Chapter 2

Extreme Value Theory

Max-stability and threshold-stability form the bedrock of univariate extreme value theory, by providing unique limiting distributions of random variables. The two most commonly used approaches for modelling univariate extreme values are the *block maxima* approach and the *peaks-over-threshold* approach. The former models the maxima of the data using the generalised extreme value distribution, while the latter models exceedances above a threshold using the generalised Pareto distribution. A combination of these approaches can also be formed, using point processes. For a thorough review of statistical methods for univariate extremes, see Coles (2001).

2.1 Block maxima

Let $X_1, \ldots, X_n$ be a set of n independent and identically distributed random variables, with each variable having the continuous distribution function F, and consider $M_n = \max\{X_1, \ldots, X_n\}$, the maximum of a block of length n. From Fisher and Tippett (1928), the Extremal Types Theorem follows:

Theorem 2.1.1 (Extremal Types Theorem)**.** *If there exist norming sequences* $\{a_n >$

$0\}_{n=1}^{\infty}$ *and* $\{b_n\}_{n=1}^{\infty}$*, such that*

$$\Pr\left\{\frac{M_n - b_n}{a_n} \leq x\right\} \to G(x) \text{ as } n \to \infty$$

for $x \in \mathbb{R}$ *and where the limiting distribution* G *is non-degenerate, then* G *must be the distribution function of one of Fréchet, Gumbel, or negative Weibull random variables.*

The three aforementioned distributions form the extreme value families of distributions, defined as follows:

- **Fréchet:** $G(x) = \begin{cases} 0, & x \leq b, \\ \exp\left[-\left(\frac{x-b}{a}\right)^{\alpha}\right], & x > b, \end{cases}$ for $\alpha < 0$;
- **Gumbel:** $G(x) = \exp\left\{-\exp\left[-\left(\frac{x-b}{a}\right)\right]\right\}, \quad x \in \mathbb{R}$;
- **Negative Weibull:** $G(x) = \begin{cases} \exp\left\{-\left[-\left(\frac{x-b}{a}\right)\right]^{\alpha}\right\}, & x < b, \\ 1, & x \geq b, \end{cases}$ for $\alpha < 0$,

for $a > 0$, $b \in \mathbb{R}$. If Theorem 2.1.1 holds, then F is said to be in the maximum domain of attraction (MDA) of G, and its normalised maxima must therefore converge in distribution to one of the three extremal families. Without norming M_n a degenerate distribution is obtained. This is because a point mass is obtained at the end-point of the distribution x_F, since $\Pr\{M_n < x\} = F^n(x)$, $x \in \mathbb{R}$, and $F^n(x) \to 0$ as $n \to \infty$ for all $x < x_F$, whereas $F^n(x) = 1 \; \forall x > x_F$. Moreover, using $F^n(x)$ is often impractical as F is unlikely to be known. Thus, the norming of M_n is paramount.

The extreme value families can in fact be shown to be special cases of a single family: the generalised extreme value (GEV) distribution, with distribution function

$$G(x) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_{+}^{-1/\xi}\right\}, \; x \in \mathbb{R}, \tag{2.1.1}$$

with $y_+ = \max(y, 0)$, and where μ, $\xi \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$, are the location, shape and scale parameters, respectively. Together, Theorem 2.1.1 and result (2.1.1) are powerful,

as Theorem 2.1.1 holds as the limit distribution for a very broad class of continuous distributions F, and then result (2.1.1) implies that whatever F is in this class, the maxima must follow a single specific class of distribution, determined by only three parameters.

The extreme value family is dictated by the shape parameter: for $\xi > 0$, we have the Fréchet family; for $\xi < 0$, the negative Weibull family; and $\xi = 0$ is interpreted as the limit $\xi \to 0$, leading to the Gumbel family. Consequently, the shape parameter also determines the lower- and upper-end-points of G, x_G and x^G, respectively, with $x_G = \mu - \sigma/\xi$ if $\xi > 0$ and $x^G = \mu - \sigma/\xi$ if $\xi < 0$, and $(x_G, x^G) = (-\infty, \infty)$ otherwise.

A critical feature in the analysis of extremes is max-stability. A distribution satisfies the max-stability property if and only if sample maxima of independent draws lead to equivalence in distribution subject to a change in location and scale. More exactly, if a distribution G is max-stable then there exist constants $A_n > 0$ and B_n such that $G^n(A_n x + B_n) = G(x),\ \forall n \in \mathbb{N}, x \in \mathbb{R}$. The GEV distribution is max-stable, meaning the maxima of GEV-distributed random variables also follows a GEV distribution. Uniquely, the GEV family is the only family of distributions which satisfy this property.

Assuming that equation (2.1.1) holds exactly for large n, then

$$\Pr\{M_n \leq x\} = G\left(\frac{x - b_n}{a_n}\right) = \tilde{G}(x),$$

with $\tilde{G}$ representing a GEV distribution with a different location and scale parameter to that of G. This result allows for modelling of maxima in practice, as data can be blocked into sections of equal length n, with the maxima of each block considered realisations of $\tilde{G}$. This approach is termed the block maxima approach, see Figure 2.1.1 (left). Here, only the maxima (in red) are used for inference of the GEV parameters.

Estimating the quantiles of the GEV distribution typically forms an important aspect of extreme value analysis since they describe the kinds of values which can reasonably be expected to be exceeded over any given period. These quantiles can be

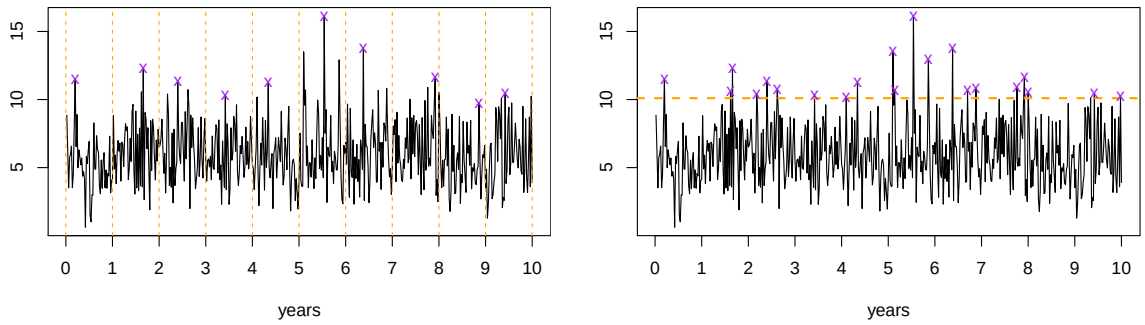


Figure 2.1.1: Illustration of the block maxima (left) and threshold exceedance (right) approaches for the modelling of univariate extremes. The purple crosses denote those data used for inference from the same simulated data.

estimated by inverting equation (2.1.1), whereby the solution to $G(x_p) = 1 - p$ gives

$$x_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left\{1 - [-\log(1-p)]^{-\xi}\right\}, & \text{for } \xi \neq 0, \\ \mu - \sigma \log[-\log(1-p)], & \text{for } \xi = 0. \end{cases} \tag{2.1.2}$$

The quantile x_p can be interpreted as the value which is exceeded in a single block with probability p. It is commonly termed that x_p is the *return level* of the *return period* $1/p$. For example, if each block represents a year of data, and $p = 0.1$, then x_p is the value we can expect to be exceeded once every $1/0.1 = 10$ years.

Practically, the maximum likelihood estimates of (μ, σ, ξ), $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ respectively can be calculated (Coles, 2001), and are then used to calculate the MLE of the return level $\hat{x}_p$ via equation (2.1.2) and the invariance property of MLE. The variance of this estimate can be found via standard results of the asymptotic normality of maximum likelihood estimators.

Of course, the extreme phenomena of interest may be small, such as when applied to low-temperature physics (Fyodorov and Bouchaud, 2008). Here, since

$$\max\{X_1, \ldots, X_n\} = -\min\{-X_1, \ldots, -X_n\}, \tag{2.1.3}$$

theory for modelling the lower tail of a distribution immediately follows from the theory required to model the upper tail of a distribution. Thus, analysis of minima can be studied under the same framework.

2.2 Threshold exceedances

A criticism of the block maxima approach is that only a single value per block is used for inference. On the other hand, the peaks over threshold approach considers all observations above a suitably high threshold as being extreme. This allows all of the most extreme data to be analysed, unlike the block maxima approach, and typically leads to more efficient inference, see Figure 2.1.1 (right). See Pickands (1975) and Balkema and De Haan (1974) for a full justification of the threshold exceedance approach. Let

$$N_n(x) = \sum_{i=1}^{n} \mathbb{1}\left(X_i > a_n x + b_n\right),$$

with $\mathbb{1}(A)$ being an indicator of event A occurring, then $N_n(x)$ is the random variable corresponding to the number of random variables $X_1, \ldots, X_n$ exceeding a threshold $a_n x + b_n$, with a_n and b_n as in limit (2.1.1). So $N_n(x)$ has a Binomial distribution with

$$N_n(x) \sim \text{Binomial}(n, 1 - F(a_n x + b_n)).$$

Under the same conditions behind the GEV limit from equation (2.1.1) where, for $X \sim F$ and F being in the maximum domain of attraction of a GEV distribution, i.e., $F^n(a_n x + b_n) \to G(x)$ for suitable $a_n > 0, b_n$ as $n \to \infty$, then for all x

$$n \log F(a_n x + b_n) \to \log G(x).$$

Thus, using standard Taylor series approximation, for all x

$$n[1 - F(a_n x + b_n)] \to -\log G(x) = [1 + \xi(x - \mu)/\sigma]_+^{-1/\xi}, \quad \text{as } n \to \infty. \tag{2.2.1}$$

Using property (2.2.1), then the classic Poisson limit from a Binomial gives that as $N_n(x) \to N(x)$, then

$$N(x) \sim \text{Poisson}\left(\lambda(x)\right), \quad \lambda(x) = [1 + \xi(x - \mu)/\sigma]_+^{-1/\xi}. \tag{2.2.2}$$

Furthermore, it follows that for $x > u$ and X distributed as X_i, that as $n \to \infty$

$$\Pr\{X > a_n x + b_n | X > a_n u + b_n\} \to \log G(x)/\log G(u) =: \bar{H}_u(x),$$

where $\bar{H}_u(x) := 1 - H_u(x)$, and the distribution function H_u is of the form

$$H_u(x) = 1 - \left[1 + \xi\left(\frac{x - u}{\tilde{\sigma}_u}\right)\right]_+^{-\frac{1}{\xi}}. \tag{2.2.3}$$

Formally, H_u is termed the generalised Pareto distribution (GPD), denoted as $\text{GPD}(\tilde{\sigma}_u, \xi)$ with threshold u, shape parameter ξ and scale parameter $\tilde{\sigma}_u \in \mathbb{R}_+$, giving the distribution of the excess $X - u$ of exceedances of u, i.e., conditioned on $X > u$. The GPD limit distribution (2.2.3) gives an asymptotic model for the distribution of exceedances above a threshold u, no matter the distribution F. The GPD scale parameter $\tilde{\sigma}_u$ is linked to the GEV scale parameter via $\tilde{\sigma}_u = \sigma + \xi(u - \mu)$. Importantly, the GPD shape parameter ξ is equivalent to the GEV shape parameter ξ, see Coles (2001) for the interpretation of this. For $\xi < 0$, there exists a finite upper-end-point

$$x^H = u - \tilde{\sigma}_u/\xi : \; H_u(x) = 1, \; \forall x > x^H.$$

In contrast, for $\xi \geq 0, \; H_u(x) < 1, \; \forall x < \infty$, so $x^H = \infty$ in this case.

Like max-stability, *threshold-stability* is another important property in the analysis of extremes. A family of distributions has threshold-stability if, given the family is valid for excesses over some threshold u_0, then is also valid for excesses over all thresholds u, such that $x^H > u > u_0$, albeit with a change in scale. It turns out that the only family of distribution with this property is the GPD. Therefore, if it is appropriate to model $(X - u_0)|\{X > u_0\}$ as a GPD$(\tilde{\sigma}_{u_0}, \xi)$, then $(X - u)|\{X > u\}$ can also be modelled as a GPD, but with GPD$(\tilde{\sigma}_u, \xi)$ if $x < x^H$. This new scale parameter is given as $\tilde{\sigma}_u = \tilde{\sigma}_{u_0} + \xi(u - u_0)$, that is, the GPD scale parameter at higher thresholds is linearly dependent on the higher threshold value. Moreover, note that if $\xi < 1$

$$\mathbb{E}\left[(X - u)|\{X > u\}\right] = \frac{\tilde{\sigma}_u}{1 - \xi} = \frac{\tilde{\sigma}_{u_0} + \xi(u - u_0)}{1 - \xi}, \ \forall u > u_0, \tag{2.2.4}$$

and thus the expectation is also linear in u for all $u > u_0$. By reparametrising as

$$\tilde{\sigma}_u^* := \tilde{\sigma}_u - \xi u, \tag{2.2.5}$$

then the modified scale parameter $\tilde{\sigma}_u^*$ becomes independent of the threshold, which can help with threshold selection and inference. Another useful reparametrisation is to perform inference on the orthogonal parameters ξ and $\nu := \tilde{\sigma}_u(1 + \xi)$ (Chavez-Demoulin and Davison, 2005), which can improve efficiency, especially when performing MCMC in a Bayesian framework.

The choice of threshold u is often user specified, based on bias-variance trade-off. If u is too high, then less data are available for inference, thus inducing high variance and uncertainty around the parameter estimates. If u is selected too low, then the approximation to the GPD limit becomes poor, creating bias in the parameters' estimates.

The optimum choice of threshold has been the subject of much historical focus. See Scarrott and MacDonald (2012) for a comprehensive overview. Davison and Smith

(1990) suggest a graphical method, the mean residual life plot. From (2.2.4), the expectation $\mathbb{E}[X - u|X > u]$, which can be estimated by the sample mean of the threshold excesses, is a linear function of u, for $u > u_0$, if the GPD is a valid model for $X > u_0$. Therefore, by plotting the sample mean excess of a threshold for a range of candidate thresholds, the selected threshold is then the point above which linearity is observed in the plot, although identifying this in practice is difficult due to sampling variation which changes with threshold. Alternatively, by reparametrising as in (2.2.5) the maximum likelihood estimates $\hat{\tilde{\sigma}}^*$ and $\hat{\xi}$ of $\tilde{\sigma}^*$ and ξ, respectively, should be constant (apart from sampling variation) for a valid GPD. Using *parameter stability plots* - plotting a range of values of u against corresponding estimates of $\hat{\tilde{\sigma}}^*$ and $\hat{\xi}$ - a suitable threshold u is selected as the lowest value above which the maximum likelihood estimates remain constant as a function of u.

Extreme value theory is primarily concerned with the tail behaviour, and specifically in this model the upper-tail behaviour for $X > u$. However, an appropriate model for the "bulk" of the distribution of X is also of interest. Typically, this is chosen to be the empirical distribution $\tilde{F}(x)$ (Coles, 2001), such that the resulting distribution function is given as

$$F(x) := \begin{cases} \tilde{F}(x) & \text{if } x \leq u, \\ 1 - \lambda_u \left[1 + \xi \frac{x-u}{\tilde{\sigma}_u}\right]_+^{-1/\xi} & \text{if } x > u, \end{cases}$$

where $\lambda_u = 1 - \tilde{F}(u)$.

In the absence of a parametric model for the domain below the threshold, the GPD parameters, which govern the distribution above the threshold, are uninformed by those data below the threshold. In this case, the data below the threshold can be assumed censored from below at the level of the threshold. Then, the contribution to the log-

likelihood of an observation x is

$$\ell(x|\lambda_u, \tilde{\sigma}_u, \xi_: = \begin{cases} 1 - \lambda_u & \text{if } x \leq u, \\ \frac{\lambda_u}{\tilde{\sigma}_u}\left[1 + \xi\frac{x-u}{\tilde{\sigma}_u}\right]_+^{-1/\xi-1} & \text{if } x > u. \end{cases}$$

The threshold exceedance approach leads to a model for the extreme tail with two components: a model for the number of exceedances above the threshold, which from equation (2.2.2) is Poisson with mean $\lambda = [1 + \xi(x - \mu)/\sigma]_+^{-1/\xi}$; and a model for threshold exceedances, $H_u(x)$ which is a GPD. Next, we explore a unification of these two components.

2.3 Point process framework

As can be seen from the derivations above, both the rate of exceedances of a high threshold (2.2.2) and the GPD parameters (2.2.3) are functions of the GEV parameters (2.1.1). In fact, the block maxima and threshold exceedance approaches can be combined using a point process limit (Pickands, 1971) which exploits this property. The key result is formulated in Coles (2001):

Theorem 2.3.1 (Extremal Point Process Theorem)**.** *Let $X_1, \ldots, X_n$ be a sequence of independent and identically distributed random variables and suppose that there exists appropriate norming sequences $\{a_n > 0\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ such that, for any $x \in \mathbb{R}$*

$$\Pr\left\{\frac{M_n - b_n}{a_n} \leq x\right\} \to G(x), \quad n \to \infty,$$

as in limit (2.1.1) with G non-degenerate, and with lower- and upper-end-points $x_G := \sup\{x \in \mathbb{R} : G(x) = 0\}$, and $x^G := \inf\{x \in \mathbb{R} : G(x) = 1\}$, respectively, so $x_G < x^G$.

Then the sequence of point processes

$$P_n = \left\{ \left(\frac{i}{n+1}, \frac{X_i - b_n}{a_n} \right) \; : \; i = 1, \ldots, n \right\},$$

converges on regions in the form $(0,1) \times [x_G, \infty)$ *to a non-homogeneous Poisson process* P *with integrated intensity measure* Λ*. It follows that* Λ *of* P *on* $\mathcal{A}_{t,x} := [0,t] \times [x, x^G]$*, where* $0 < t \leq 1, \; x > x_G$ *is*

$$\Lambda\left(\mathcal{A}_{t,x}\right) = t \left[1 + \xi \left(\frac{x-\mu}{\sigma} \right) \right]_+^{-\frac{1}{\xi}}. \tag{2.3.1}$$

The scaling in P_n enforces that, as $n \to \infty$, the first component to be continuous on $[0,1]$, and the maximum of the second component to be non-degenerate with limiting distribution (2.1.1). Theorem 2.3.1 implies that the intensity function λ for P is, for $t \in [0,1]$ and $x_G < x < x^G$,

$$\lambda(t,x) = \frac{\partial^2 \Lambda\left(\mathcal{A}_{t,x}\right)}{\partial x \, \partial t} = \frac{1}{\sigma} \left[1 + \xi \left(\frac{x-\mu}{\sigma} \right) \right]_+^{-\frac{1}{\xi}-1}. \tag{2.3.2}$$

From standard Poisson process properties, the number of points of P falling within the set $S \subseteq [0,1] \times (x_G, \infty)$ follows a Poisson distribution with mean $\Lambda(S) = \int_S \lambda(t,x) \, \mathrm{d}x \, \mathrm{d}t$, with $\lambda(t,x)$ given by expression (2.3.2).

Statistical application of the point process model assumes that for large enough n, the limit $P_n \to P$ holds exactly. After absorbing norming constants into the limiting intensity, it is assumed that P, with intensity (2.3.2), applies to the points $\{(i/(n+1), X_i); i = 1, \ldots, n\}$ on the set $\mathcal{A}_{1,u} = [0,1] \times (u, \infty]$. If $\boldsymbol{x} = \{(t_1, x_1), \ldots, (t_m, x_m)\}$ denote the m of these points that fall in $\mathcal{A}_{1,u}$, then the likelihood of the parameters $\theta := (\mu, \sigma, \xi)$ based on a realisation $(x_1, \ldots, x_n)$ from the random variables $(X_1, \ldots, X_n)$ is

$$L(\theta | \boldsymbol{x}) \propto \exp\left\{ -\Lambda(\mathcal{A}_{1,u}) \right\} \prod_{i=1}^{m} \left\{ \left[1 + \xi \left(\frac{x_i - \mu}{\sigma} \right) \right]_+^{-\frac{1}{\xi}} \right\}.$$

Inference using this likelihood gives information about both the mean number of exceedances of the threshold u and the distribution of the threshold exceedances.

The link between the first two approaches - block maxima and threshold exceedances - now becomes apparent. If $N_n(A)$ is the number of points in A, where $A \subset \mathcal{A}_{t,x}$ where $N_n(A) \to N(A) \sim \text{Poisson}\left[\Lambda(A)\right]$, as $n \to \infty$, then the event $N_n(A_x) = 0$ for $A_x = (0,1) \times (x, \infty)$ is equivalent to $\{(M_n - b_n)/a_n \leq x\}$. As such,

$$\begin{aligned}
\Pr\left\{\frac{M_n - b_n}{a_n} \leq x\right\} &= \Pr\{N_n(A_x) = 0\} \\
&\to \Pr\{N(A_x) = 0\} \\
&= \exp\left(-\Lambda\left(A_{1,x}\right)\right) \\
&= \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\},
\end{aligned}$$

as $n \to \infty$, thus equivalence with limit distribution (2.1.1) is reached.

In a similar vein, an equivalence with threshold exceedances can be shown. As the integrated intensity (2.3.1) can be written as

$$\Lambda\left([0,t] \times [x,\infty)\right) = \Lambda_t\left([0,t]\right) \times \Lambda_x\left([x,\infty)\right),$$

where

$$\Lambda_t\left([0,t]\right) = t \text{ and } \Lambda_x\left([x,\infty)\right) = \left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi},$$

then for any $u > x_G$ and $x + u < x^G$,

$$\begin{aligned}
\Pr\left\{\frac{X_i - b_n}{a_n} \leq x + u \;\middle|\; \frac{X_i - b_n}{a_n} > u\right\} &= 1 - \Pr\left\{\frac{X_i - b_n}{a_n} > x + u \;\middle|\; \frac{X_i - b_n}{a_n} > u\right\} \\
&\to 1 - \frac{\Lambda_x\left([x+u, \infty)\right)}{\Lambda_x\left([u, \infty)\right)}, \quad \text{as } n \to \infty \\
&= 1 - \frac{\left[1 + \xi\left(\frac{x+u-\mu}{\sigma}\right)\right]_+^{-1/\xi}}{\left[1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right]_+^{-1/\xi}} \\
&= 1 - \left[1 + \frac{\xi x}{\tilde{\sigma}_u}\right]_+^{-1/\xi} \\
&=: H_u(x),
\end{aligned}$$

where $H_u(x)$ is the GPD derived in equation (2.2.3).

In estimating the GEV parameters, the point process approach uses all data larger than u for inference rather than only block maxima (like the peaks over threshold approach). Therefore, using point processes can result in more efficient inference than that of the block maxima approach. Unlike the GPD, the parameters in the point process parametrisation are invariant to the choice of threshold u.

While still assuming independence, the assumption of the random variables being identically distributed can be relaxed somewhat by, for example, including a covariate structure. In the most general case, suppose the parameters $(\mu(t), \sigma(t), \xi(t))$ vary with the covariate time, denoted t. The non-homogeneous Poisson process then allows for time-dependent rates of occurrences and excess distributions, see Smith (1989), with intensity

$$\lambda(t, x) = \frac{1}{\sigma(t)}\left[1 + \xi(t)\left(\frac{x - \mu(t)}{\sigma(t)}\right)\right]_+^{-\frac{1}{\xi(t)}-1},$$

where $\sigma(t) \in \mathbb{R}_+$, $\forall t$ and so the integrated intensity is

$$\Lambda(\mathcal{A}_{1,u}) = \int_0^1 \left[1 + \xi(t)\left(\frac{u - \mu(t)}{\sigma(t)}\right)\right]_+^{-\frac{1}{\xi(t)}} \mathrm{d}t.$$

2.4 Extreme value theory in sports

Within sports, Robinson and Tawn (1995) use extreme value theory (EVT) to model athletics data, and Strand and Boes (1998) use EVT to determine the peak age and deterioration of competitive 10K road race runners. Stephenson and Tawn (2013) fit a GEV distribution to yearly maxima of athletics times across different distance and eras. The location and scale parameters vary as a parametric function of the distance, and an exponential trend allows for a smooth adjustment for era. Since the best run-times are the smallest, the negated data are analysed, as in formulation (2.1.3); however, the GEV distribution for maxima could also be used by modelling run-*speeds*, by taking the reciprocal of the data. This choice is not trivial. Non-linear transformations of the measurement scale in extreme value modelling lead to disparate inference, as Wadsworth et al. (2010) depict with a motivating example from North Sea wave data. They analyse extreme wave height H and wave drag force F - a vital variable for offshore structural design - where $F \propto H^2$ (Tromans and Vanderschuren, 1995), and found that separately studying these two variables generates inconsistent inference. In spite of sharing the same underlying physical phenomena, the estimates for the shape parameters had opposite signs, implying a finite upper limit for wave height, but a non-zero probability of an arbitrarily large drag force. In the context of sports, this could mean that the analysis of event-times concludes that the distribution of run-times has a light tail, whilst analysis of run-speeds concludes there is no limit to the speed a human can run, or vice versa. Thus, attending to data transformation is critical, and getting it wrong can result in physically impossible conclusions. Despite this, Gomes and Henriques-Rodrigues (2019) use peaks-over-threshold methodology to analyse the best swim-speeds by transforming data of the best swim-times. This allows them to analyse the right-hand tail of the distribution, but the effect of this transformation on the inference is unknown. The effect of this transformation will be discussed in Chapter 5.

2.5 Multivariate extreme value theory

2.5.1 Copula-based modelling

Utilising *copulae* to characterise the dependence between univariate variables is a common approach Joe (1997); Nelsen (2007). Typically, the margins are modelled first, with the dependence then modelled separately, which often leads to faster inference. Through Sklar's Theorem (Sklar, 1959), all the dependence can be ascertained through the copular alone.

Theorem 2.5.1 (Sklar's Theorem)**.** *For the random vector* $X = \{X_i : i \in \{1, \dots, d\}\}$ *having joint distribution* F*, and each component with continuous marginal distribution* $X_i \sim F_i$ *for* $i \in \{1, \dots, d\}$*, there exists a unique copula* C *such that*

$$F(x) = C\{F_1(x_1), \dots, F_d(x_d)\},$$

for all $x_i \in dom(F_i)$, $\forall i \in \{1, \dots, d\}$.

The copula C is a multivariate distribution function on standard uniform margins, i.e., $C : [0,1]^d \to [0,1]$. The copula can be used for any marginal distribution, by first transforming the marginal distribution through the probability integral transform; for a random variable X_i with distribution function F_i, then $F_i(X_i) \sim \text{Unif}(0,1)$, no-matter the distribution of X_i. So, the Copula can be used to model dependence structure regardless of the marginal distribution (as long as the inverse F_i^{-1} exists) and crucially, the dependence structure is invariant to the marginal transformation. Hence, marginal modelling and dependence modelling can be approached separately. An oft used choice of copula is the Gaussian copula, given as

$$C(X) = \int_{-\infty}^{\Phi^{-1}(x_1)} \cdots \int_{-\infty}^{\Phi^{-1}(x_d)} \phi_d(s; \Sigma)\, \mathrm{d}s,$$

with $\phi_d(s; \Sigma)$ denoting the standard d-dimensional Gaussian density, with depedence structure determined by the correlation matrix Σ.

Chapter 3

Ranking, and other Properties, of Elite Swimmers using Extreme Value Theory

3.1 Introduction

On the face of it, comparing the performances of two swimmers in a given competition appears straightforward, simply compare their swim-times. But this simple comparison no longer holds when we compare between different distances, strokes or genders, let alone swimmers under different regulations for full body suits. In addition, due to the improvement in training methods, as well as changes in technology, a fair comparison between swim-times recorded many years apart is infeasible without some adjustment for the era of the swim.

The International Swimming Federation (FINA) uses a very simple points system to tackle this issue. The points acquired for a particular swim is a function of the ratio of the swim-time and the current world record for that event, specifically the points $p_{i,j}$ given to swimmer i in event j is $p_{i,j} \propto (b_j/t_{i,j})^3$ where b_j is the current world record in

event j, and $t_{i,j}$ is the time of swimmer i in event j. With some world records considered better than others however, bias is introduced between events, with some being much harder to attain points where the world record is hard to beat. Furthermore, the ranking method has high sensitivity as it is determined only by the set of current world records, so rankings can change substantially when a single record is broken. Importantly, FINA rankings are used by many countries and organisations for selection for regional and international competitions, so the ranking must be an accurate representation of the swimmer's true ability.

The aim, is to produce a global model that can fairly compare between strokes, gender and distance, as well as considering the improvement over time of elite sporting performance. This paper utilises extreme value theory to model the very best swim-times as being observations from a generalised Pareto distribution (GPd) so that the strength of a particular swim is judged on its position compared to the whole distribution of swim-times across all events, rather than just the world record for that event. This ensures a more efficient comparison between events. Moreover, comparison between swim-times within the same event has a more tangible interpretation since it can be described in terms of probabilities. For example, by considering swim-times t_1 and t_2 in an event, then it is natural to compare the relative quality of these by $\Pr(T > t_1)/\Pr(T > t_2)$, where T represents the random variable corresponding to a swim-time for an event, rather than, say, the metric $t_1 - t_2$. A by product of this global model is that other features of interest can be estimated, for example the ultimate possible swim-time for any given event. The distribution of the next world record swim-time for each event can be estimated, and even the distribution of the waiting time, and therefore the expected waiting time, until the next world record is broken and the probability of that record being in a particular event. In addition, swim-times can be corrected for the effect of full body suits, to allow for fair comparison between those swimmers wearing suits and those not.

The data to be studied comprise the top 500 swim-times, with at most one time per swimmer per event, in all 34 individual long course (LC) swimming events, i.e., in a 50m pool, from all major competitions between the start of 2001 and last quarter of 2018. Any data not officially accepted by FINA are removed, for example observations that were later rescinded due to the use of performance enhancing drugs. For the remainder of this article, *negative swim-times* will be analysed, and simply referred to as *swim-times*, so that if a swim-time is faster than another it has the larger negative swim-time of the two. So, for the best swim-times we are interested in the biggest negative swim-times. Therefore the paper focuses on methods for largest values, which is the typical methodological approach to extreme values (Coles, 2001). Results for actual swim-times are obtained by simply negating the results we obtain for negated swim-times. Additionally, independence is assumed between all swim-times across different years, strokes and distances, even if they are achieved by the same swimmer. Both of these two points will be discussed further in Section 3.5.

The past use of extreme value theory for sports modelling is varied. In athletics, work has been done to create a model which pools information between different distances and over time (Stephenson and Tawn, 2013). The threshold exceedance model of Smith (1989) is used by Strand and Boes (1998) to model times of long distance runners. Specifically, the typical change of time taken to run 10 kilometres with respect to the age of the athlete is modelled via a Gumbel distribution, where times within ages and across ages are assumed to be independent, and men's and women's times are modelled separately. More generally, Riegel (1981) finds a linear relationship between log world record time and log distance over many sports. Modelling men's and women's data separately is a common theme in sports data.

In swimming, Gomes and Henriques-Rodrigues (2019) use extreme value theory to model the distribution of swim-times across all LC events using independent fits for each event. Adam and Tawn (2012) explore the progression of the top performances in

swimming events over time by modelling the times of the gold medallist swimmers in the Olympic Games. Dependence due to the same swimmer winning two events at an Olympic Games is included via a bivariate extreme value distribution (Tawn, 1988). We are unaware of any previous publication that models swim data globally across gender, distance, stroke, and considers the improvements over time.

The article is set out as follows. Section 3.2 introduces extreme value theory, and the point process representation of Smith (1989), see also Coles (2001), which forms the basis of our model. Section 3.3 describes the full global model and the justification for the shared fit. In Section 3.4 the features of interest discussed above will be estimated based on the final fitted model, such as the ultimate possible swim-time for each event, examples of the best swimmers of all time under this model, the distribution of new world records, the expected time until the next world record is broken, the probability of the next world record being in a given event, and the result of adjusting for regulations of full body suits on current world records. Section 3.5 discusses the possible impacts of any major assumptions made in the modelling process, as well as investigating further improvements and applications to the proposed model.

3.2 Theory

3.2.1 Extremes of identically distributed variables

Univariate extreme value theory (EVT) provides the framework for our modelling strategy. In its simplest form it applies to an independent identically distributed (IID) random sample $X_1, \ldots, X_n$ with each variable having a continuous distribution function F. The two main approaches in EVT are the block maxima method and the peaks over threshold methods. The asymptotic theory behind these two methods is as follows. Let $M_n = \max\{X_1, \ldots, X_n\}$ be the maximum of a block of length n. We seek the distribution of M_n for large n, and in particular appropriate choices of norming sequences

$a_n > 0$ and b_n are sought such that, as $n \to \infty$,

$$\begin{aligned} \Pr\left\{\frac{M_n - b_n}{a_n} \leq x\right\} &= \Pr(X_1 \leq a_n x + b_n, \ldots, X_n \leq a_n x + b_n) \\ &= F^n(a_n x + b_n) \\ &\to G(x) \end{aligned}$$

where the limiting distribution $G(x)$ is non-degenerate. The only possible non-degenerate limiting distribution of equation (3.2.1) is the generalised extreme value distribution function (GEVd). The exact form is given by

$$G(x) = \exp\left(-[1 + \xi(x - \mu)/\sigma]_+^{-1/\xi}\right), \tag{3.2.2}$$

where $\mu,\ \xi \in \mathbb{R},\ \sigma \in \mathbb{R}^+$, are the location, shape and scale parameters respectively and $y_+ = \max(y, 0)$. Figure 3.2.1 (left) illustrates the density of the GEVd for different values of ξ, while $\mu = 0$, $\sigma = 1$. For $\xi < 0$, there exists a finite value $x_G = \mu - \sigma/\xi$: $G(x) = 1,\ \forall x > x_G$. In contrast, for $\xi \geq 0,\ G(x) < 1,\ \forall x < \infty$. The GEVd result is powerful as it holds as the limit distribution for a very broad class of continuous distributions F and implies that whatever F is in this class, the maxima must follow a single class of distributions, determined by only three parameters.

The block maxima method of Coles (2001) assumes that limit (3.2.1) holds exactly for a large enough block size n, for example all observations in a month or a year. Given a sample of length kn the approach is to split the series into k blocks with n values in each block. Then the k values of the block maxima are used to estimate the parameters (μ, σ, ξ) of the model, assuming that each of these variables is IID and follows a GEVd.

The peaks over threshold (POT) approach considers only the observations above a suitably high threshold. This allows all of the most extreme data to be analysed, unlike

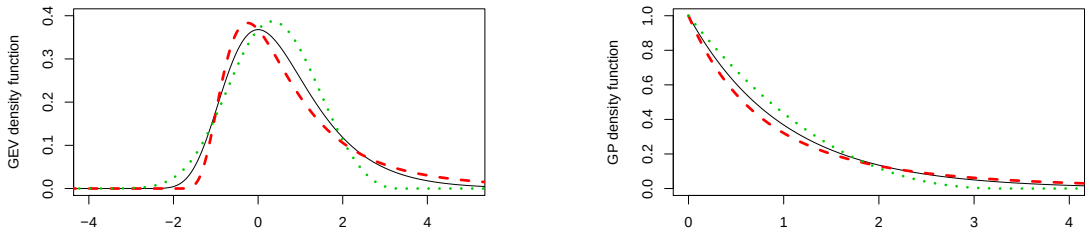


Figure 3.2.1: Density functions for the GEVd ($\mu = 0, \sigma = 1, \xi$) (left) and GPd ($u = 0, \tilde{\sigma}_u = 1, \xi$) (right) for three different shape parameters: $\xi = 0$ (solid line), $\xi = -0.3$ (dotted line) and $\xi = 0.3$ (dashed line).

the block maxima approach, and typically leads to more efficient inference. Let

$$N_n(x) = \sum_{i=1}^{n} \mathbb{1}\left(X_i > a_n x + b_n\right),$$

with $\mathbb{1}(A)$ be an indicator of event A occurring, then $N_n(x)$ is the random variable corresponding to the number of $X_1, \ldots, X_n$ exceeding a threshold $a_n x + b_n$, with a_n and b_n as in limit (3.2.1). So $N_n(x)$ has a Binomial distribution with $N_n(x) \sim \mathrm{B}(n, 1 - F(a_n x + b_n))$. Under the same conditions behind the GEVd limit from equation (3.2.1), as $n \to \infty$,

$$n \log F(a_n x + b_n) \to \log G(x),$$

and so, using standard Taylor series approximation, for all x

$$n[1 - F(a_n x + b_n)] \to -\log G(x) = [1 + \xi(x - \mu)/\sigma]_+^{-1/\xi}. \tag{3.2.3}$$

Using property (3.2.3), then the classic Poisson limit from a Binomial gives that as $n \to \infty$, $N_n(x) \to N(x)$, where $N(x)$ is a Poisson random variable with mean $[1 + \xi(x - \mu)/\sigma]_+^{-1/\xi}$. Furthermore, it follows that for $x > u$ and X distributed as X_i, that

$$\Pr\{X > a_n x + b_n | X > a_n u + b_n\} \to \log G(x)/\log G(u) = \bar{H}_u(x), \tag{3.2.4}$$

where $\bar{H}_u(x) = 1 - H_u(x)$, where the distribution function H_u is of the form

$$H_u(x) \equiv 1 - \left[1 + \xi\left(\frac{x-u}{\tilde{\sigma}_u}\right)\right]_+^{-\frac{1}{\xi}}, \tag{3.2.5}$$

is the generalised Pareto distribution function (GPd) with threshold u, shape parameter ξ and scale parameter $\tilde{\sigma}_u \in \mathbb{R}^+$ is linked to the GEVd parameters via $\tilde{\sigma}_u = \sigma + \xi(u - \mu)$. Limit distribution H_u gives an asymptotic model for the distribution of exceedances above a threshold u, no matter what the distribution F. Figure 3.2.1 (right) illustrates the density of the GPd for different values of ξ. For $\xi < 0$, there exists a finite value $x_H = u - \tilde{\sigma}_u/\xi : \ H_u(x) = 1, \ \forall x > x_H$. In contrast, for $\xi \geq 0, \ H_u(x) < 1, \ \forall x < \infty$.

The POT approach leads to a model for the extreme tail with two components: a model for the number of exceedances of the threshold, which is Poisson with mean $\lambda = [1 + \xi(x - \mu)/\sigma]_+^{-1/\xi}$, and a model for threshold exceedances, $H_u(x)$ which is a GPd. The choice of threshold u is user-specified, with the choice based on the usual bias-variance trade-off, the subject of much historical focus (Scarrott and MacDonald, 2012).

As can be seen from the derivation above both the rate and GPd parameters are functions of the GEVd parameters. In fact, the block maxima and POT approaches can be combined using a point process limit which exploits this property. Consider, the point process model of extremes, defined on a sequence

$$P_n = \left\{\left(\frac{i}{n+1}, \frac{X_i - b_n}{a_n}\right) \ : \ i = 1, \ldots, n\right\},$$

where the scaling here enforces that, as $n \to \infty$, the first component is continuous on

$[0, 1]$, and the maximum of the second component to be non-degenerate with limiting distribution (3.2.2). In particular as $n \to \infty$, $P_n \to P$ where P is a non-homogeneous Poisson process on $(0, 1] \times (b_l, \infty)$, where $b_l = \max\{x \in \mathbb{R} : G(x) = 0\}$ where G is the limit distribution (3.2.2) (Smith, 1989). It follows that the integrated intensity Λ of P on $\mathcal{A}_{t,x} = [0, t] \times [x, \infty]$, where $0 < t \leq 1,\ x > b_l$ is

$$\Lambda\left(\mathcal{A}_{t,x}\right) = t\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}},$$

which implies that the intensity function λ for P is, for $t \in (0, 1]$ and $x > b_l$,

$$\lambda(t, x) = \frac{\partial^2 \Lambda\left(\mathcal{A}_{t,x}\right)}{\partial x\, \partial t} = \frac{1}{\sigma}\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}-1}. \tag{3.2.6}$$

From standard Poisson process properties we have that the number of points of P in any set $S \subseteq [0, 1] \times (b_l, \infty)$ follows a Poisson distribution with mean $\Lambda(S) = \int_S \lambda(t, x)\, \mathrm{d}x\, \mathrm{d}t$, with $\lambda(t, x)$ given by expression (3.2.6).

Statistical application of the point process model assumes that for large enough n, the limit $P_n \to P$ holds exactly. After absorbing norming constants into the limiting intensity, it is assumed that P, with intensity (3.2.6), applies to the points $\{(i/(n + 1), X_i); i = 1, \ldots, n\}$ on the set $\mathcal{A}_{1,u} = [0, 1] \times (u, \infty]$. If $\boldsymbol{x} = \{(t_1, x_1), \ldots, (t_m, x_m)\}$ denote the m of these points that fall in $\mathcal{A}_{1,u}$, then the likelihood for the parameters $\theta = (\mu, \sigma, \xi)$ is

$$L(\theta; \boldsymbol{x}) = \exp\left\{-\Lambda(\mathcal{A}_{1,u})\right\} \prod_{i=1}^{m} \lambda(t_i, x_i). \tag{3.2.7}$$

Inference using this likelihood gives information about both the mean number of exceedances of the threshold u and the distribution of the threshold exceedances (the GPd). When a datum x_i has been recorded to some precision s such that the true value x_i' is unknown but $x_i' \in [x_i - s/2, x_i + s/2)$, interval censoring is introduced, which

can be factored into the likelihood via

$$L(\theta;\boldsymbol{x}) \propto \exp\{-\Lambda(\mathcal{A}_{1,u})\}\prod_{i=1}^{m}\int_{x_i-s/2}^{x_i+s/2}\lambda(t_i,x)\,\mathrm{d}x$$
$$= \exp\{-\Lambda(\mathcal{A}_{1,u})\}\prod_{i=1}^{m}\left\{\left[1+\xi\left(\frac{x_i-s/2-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}} - \left[1+\xi\left(\frac{x_i+s/2-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}}\right\}.$$

3.2.2 Extreme values of non-identically distributed variables

The derivations so far have assumed IID variables, however this need not be the case. Whilst still assuming independence, the assumption of identically distributed data is relaxed by including a covariate structure. In order to take the date of the swim into consideration, time is introduced as a covariate such that, in the most general case, all parameters of θ are allowed to vary with time, for example $\theta(t) = (\mu(t), \sigma(t), \xi(t))$. The non-homogeneous Poisson process allows for time-dependent rates of occurrences and excess distributions, see Smith (1989). Under this relaxation, equation (3.2.6) becomes

$$\lambda(t,x) = \frac{1}{\sigma(t)}\left[1+\xi(t)\left(\frac{x-\mu(t)}{\sigma(t)}\right)\right]_+^{-\frac{1}{\xi(t)}-1}, \tag{3.2.8}$$

and so the integrated intensity is

$$\Lambda(\mathcal{A}_{1,u}) = \int_0^1 \left[1+\xi(t)\left(\frac{u-\mu(t)}{\sigma(t)}\right)\right]_+^{-\frac{1}{\xi(t)}}\mathrm{d}t. \tag{3.2.9}$$

The full likelihood function, accounting for interval censoring, can then be expressed, as in equation (3.2.7), but with $\Lambda(\mathcal{A}_{1,u})$ and $\lambda(t,x)$ given in equations (3.2.9) and (3.2.8),

such that

$$L(\theta;\boldsymbol{x}) = \exp\{-\Lambda(\mathcal{A}_{1,u})\}\prod_{i=1}^{m}\int_{x_i-s/2}^{x_i+s/2}\lambda(t_i,x)\,\mathrm{d}x$$
$$= \exp\{-\Lambda(\mathcal{A}_{1,u})\}\prod_{i=1}^{m}\left\{\left[1+\xi(t_i)\left(\frac{x_i-s/2-\mu(t_i)}{\sigma(t_i)}\right)\right]_+^{-\frac{1}{\xi(t_i)}} - \left[1+\xi(t_i)\left(\frac{x_i+s/2-\mu(t_i)}{\sigma(t_i)}\right)\right]_+^{-\frac{1}{\xi(t_i)}}\right\}, \quad (3.2.10)$$

where the parameters within $\theta(t)$ are found by maximising this likelihood. If $\{y_i : i = 1,\ldots,18\}$ is the set of start dates of years from 2001-2019, then the expected rate of exceedances of u with year $2000+i$ is given by

$$\Lambda_i(\mathcal{A}_{1,u}) = \int_{y_i}^{y_{i+1}}\left[1+\xi(t)\left(\frac{u-\mu(t)}{\sigma(t)}\right)\right]_+^{-\frac{1}{\xi(t)}}\mathrm{d}t.$$

If the change in the parameters is small over the course of each year, then the rate can be approximated as

$$\Lambda_i(\mathcal{A}_{1,u}) \approx \left[1+\xi(y_i^*)\left(\frac{u-\mu(y_i^*)}{\sigma(y_i^*)}\right)\right]_+^{-\frac{1}{\xi(y_i^*)}}(y_{i+1}-y_i), \quad (3.2.11)$$

where $y_i^* = (y_i+y_{i+1})/2$. Likewise the excess distribution at a time t is given for $x > u$ by

$$\Pr\{X_t > x | X_t > u\} = \left[1+\xi(t)\left(\frac{x-u}{\tilde{\sigma}_u(t)}\right)\right]_+^{-\frac{1}{\xi(t)}},$$

where $\tilde{\sigma}_u(t) = \sigma(t)+\xi(t)\,[u-\mu(t)]$.

3.3 Model for swimming data

3.3.1 The Data

The data are from the FINA swimming website's database, at http://www.fina.org/, which contains around the top 500 recorded swim-times for all 34 individual LC swimming events. The fastest swim time per swimmer per event is taken, irrespective of the year in which it occurs. The data includes interval censored observations which come from the rounding of recorded timings. Given that the data are rounded, in seconds to 2 decimal places, the interval censoring likelihood (3.2.10) is formally needed with $s = 0.01$. In practice using standard likelihood (3.2.7) instead would give similar results in practice, with the exception of 50m events as the rounding is a more substantial part of the variation in these data.

In order to develop a consistent approach across all events $e \in E$ where E is the set of all 34 individual LC swim events, the threshold for each event was set such that there were an identical number of exceedances in each event. From plotting PP and QQ plots for each event e independently over a range of thresholds u'_e, the thresholds were set such that there were 200 exceedances in each event, as this appropriately balances the bias and variance for the majority of events. The choice of a single threshold selection approach is discussed in Section 3.5, but is worth noting that this could lead to uncertainty estimates that are underestimated. For each event e, the threshold used for the model, u_e, was set to $u_e = u'_e - s/2$, to account for the interval censoring.

Properties of the 200 best times for the 100m men's butterfly swim-times are illustrated in Figure 3.3.1, with these being typical across all events. There is a general increasing trend in the rate of occurrences over time. In addition to this trend there is a noticeable step-increase in the frequency of observations in the top 200 swims between the introduction, in 2008, and subsequent banning, from the start of 2010, of *swim-suits* by FINA (Shipley, 2009). Swim-suits have been found to reduce drag by up to 35% in

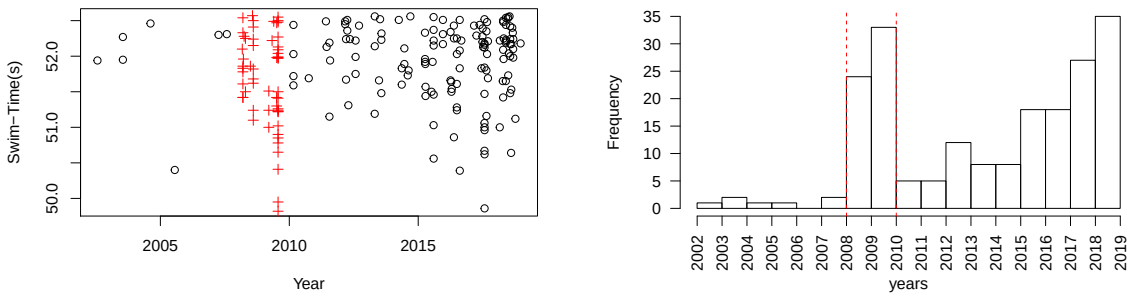


Figure 3.3.1: Data for the men's 100m butterfly. The data (left) shows the raw data for the swim-times and so the lower tail is the feature of interest. Here, the crosses indicate swims recorded within then swim-suit period. Similarly, the observed annual rates of exceeding the threshold (right) include dashed vertical lines (right) which indicate the swim-suit time period.

independent testing (Moria et al., 2011), and a significant number of world records were set during their use. Particularly in 2009, the introduction all polyurethane suits, such as the 'Arena X-Glide', saw a significant improvement in performances (Foster et al., 2012). Figure 3.3.1 shows that there appears to be differences in performances between 2008 and 2009 which illustrates an impact of changes of full-body suit technology.

There is an inconsistency in the selection of the competitions in the FINA database, with only important competitions being represented in some of the earlier years, whereas later years cover all high-level competitions. One consequence of this is that the rate per year of exceeding the threshold u_e will increase over time due to this feature, with the effect being largest in the earliest years. So, changes in the threshold exceedance rate, for each event, arise from a combination of improved swimming performance and the database formulation. Therefore, care must be taken when interpreting this feature in the analyses. There is also the potential for the distribution of swim times that exceed the threshold to change over time due to this biased selection of competitions in the database. Any such effect should be minimal on inferences given that most exceedances

are from the later years, so the likelihood is naturally most influenced by data from later years. The model we develop presumes there is no such bias to the distribution of excesses, but this assumption is tested (see Figure 3.3.4 (right)) and shown to provide a sufficiently good description of the early data.

3.3.2 Separate Event Model

The Poisson point process framework allows us to model the time varying rate of observations above threshold, as well as the distribution of these observations. To incorporate the general increasing frequency of swim-times observed in Figure 3.3.1, time was included as a covariate in the model. The swim-suit factor was included via an indicator covariate, where the assumption is made that all observations during the swim-suit epoch were by swimmers wearing a swim-suit, and initially it is assumed that the swim-suit effect is constant throughout this epoch.

Following Davison and Smith (1990) and Coles (2001) the Poisson process parameters $\mu^{(e)}(t)$, $\sigma^{(e)}(t)$ and $\xi^{(e)}(t)$ are initially assumed to vary smoothly with time t in the model for each separate event. From fitting each event independently, it was then concluded, via use of AIC, that a linear dependence on time is appropriate for the parameters $\mu^{(e)}(t)$ and $\sigma^{(e)}(t)$ to describe the increase in rates of observations. Moreover, $\xi^{(e)}(t)$ is assumed to be constant over time as is common in the literature across extreme value applications to rainfall, sea-level, and athletics amongst others, e.g, Smith (1989), Robinson and Tawn (1995), Strand and Boes (1998), which find that despite changes in the distribution due to various covariates, the shape parameter is constant and is therefore taken as some unknown fixed value $\xi^{(e)}(t) = \xi^{(e)}$ for that event.

Although the patterns in the rate of observations is noticeable from plots alone, patterns in the distribution of the observations exceeding the threshold are not so obvious. To find an appropriate model for the distribution of exceedances above the thresholds, several models for the GPd parameters were fitted and compared, which included, but

were not limited to, linear trends over time and including indicators of swim-suit effects. Interestingly, after model comparison it was concluded that for each event the distribution of observations above the threshold is independent of covariates, indicating that any improvements over time are due to an increase in quantity of exceedances above the threshold, rather than any change in the nature of the exceedances themselves. These findings in the data about the rate and the distribution of the best swims are reflected in the following parametrisations.

For a given event $e \in E$, the Poisson process is parametrised as either,

$$\begin{aligned}
\xi^{(e)}(t) &= \xi^{(e)}, \\
\mu^{(e)}(t) &= \mu_0^{(e)} + \beta^{(e)} t + \gamma^{(e)} \mathbb{1}_{\{t \in S_t\}}, \\
\sigma^{(e)}(t) &= \sigma_0^{(e)} + \xi^{(e)} \beta^{(e)} t + \xi^{(e)} \gamma^{(e)} \mathbb{1}_{\{t \in S_t\}},
\end{aligned} \tag{3.3.1}$$

or,

$$\begin{aligned}
\xi^{(e)}(t) &= \xi^{(e)}, \\
\mu^{(e)}(t) &= \mu_0^{(e)} + \beta^{(e)} t + \gamma_1^{(e)} \mathbb{1}_{\{t \in S_{t_1}\}} + \gamma_2^{(e)} \mathbb{1}_{\{t \in S_{t_2}\}}, \\
\sigma^{(e)}(t) &= \sigma_0^{(e)} + \xi^{(e)} \beta^{(e)} t + \xi^{(e)} \gamma_1^{(e)} \mathbb{1}_{\{t \in S_{t_1}\}} + \xi^{(e)} \gamma_2^{(e)} \mathbb{1}_{\{t \in S_{t_2}\}},
\end{aligned} \tag{3.3.2}$$

where $\theta^{(e)}(t)$ represents θ for event e at time t, and $\mu_0^{(e)},\ \xi^{(e)} \in \mathbb{R},\ \sigma_0^{(e)} \in \mathbb{R}^+$ are the location, shape, and scale parameters for the Poisson process, $\beta^{(e)} \in \mathbb{R}$ controls the linear trend in $\mu^{(e)}(t)$ and $\sigma^{(e)}(t)$. In the case of assuming a single swim-suit effect, $\gamma^{(e)} \in \mathbb{R}$ controls the magnitude of this effect, $\mathbb{1}$ is the indicator function and $S_t \in [2008, 2009]$ denotes the time period in which swim-suit were allowed, and in the case of allowing for the differing effects of the two major suit-types, as noted in Section 3.3.1, $\gamma_1^{(e)} \in \mathbb{R}$ and $\gamma_2^{(e)} \in \mathbb{R}$ control the effects of these two suit-types, with $S_{t_1} \in [2008]$ and $S_{t_2} \in [2009]$ denoting the approximate time periods in which these suits were

active. In particular t is linearly standardised to have zero mean and unit variance over the observed data. Both parametrisations (3.3.1) and (3.3.2) ensure that the GPd scale parameter for exceedances of the level u_e at time t is covariate-independent. For example, with parametrisation (3.3.1),

$$\begin{aligned}\tilde{\sigma}_u^{(e)}(t) &= \sigma^{(e)}(t) + \xi^{(e)}\left[u_e - \mu^{(e)}(t)\right] \\ &= \sigma_0^{(e)} + \xi^{(e)}\beta^{(e)}t \; + \; \xi^{(e)}\gamma^{(e)}\mathbb{1}_{\{t\in S_t\}} + \xi^{(e)}(u_e - [\mu_0^{(e)} + \beta^{(e)}t + \gamma^{(e)}\mathbb{1}_{\{t\in S_t\}}]) \\ &= \sigma_0^{(e)} + \xi^{(e)}(u_e - \mu_0^{(e)}) \\ &:= \tilde{\sigma}_u^{(e)}, \end{aligned} \tag{3.3.3}$$

and the same clearly holds for parametrisation (3.3.2) so that the two GPd parameters, $\xi^{(e)}$ and $\tilde{\sigma}_u^{(e)}$, and thus the distribution above the threshold is identically distributed over covariates, as required. It is common to use a log link in the scale parameter in the non-homogeneous Poisson process to ensure positivity, however this would make the covariate independence of $\tilde{\sigma}_u^{(e)}$, property (3.3.3), impossible. Instead, $\mu_0^{(e)}$, $\sigma_0^{(e)}$ and $\xi^{(e)}$ are constrained such that $\tilde{\sigma}_u^{(e)}$ in expression (3.3.3) is positive.

Figure 3.3.2 shows all the model parameter estimates from parametrisation (3.3.2), obtained by fitting independently across events: three GEV parameters, μ_0, σ_0, ξ, one trend parameter β, and two swim-suit parameters γ_1 and γ_2 for each of the 34 events, giving a total of 204 independent parameters. These parameters, after the transformation described below, are plotted against $u_{L,e} = \log(-u_e)$, recalling that the data are negative, with $u_e < 0$, so $u_{L,e}$ is the log of the 200th best swim-time for event e in the data. For each of the transformed parameters

$$\sigma_L^{(e)} = \log(\tilde{\sigma}_u^{(e)}),\; \mu_L^{(e)} = \log\left(-\mu_0^{(e)}\right),\; \beta_L^{(e)} = \log\left(\beta^{(e)}\right),\; \gamma_{L,1}^{(e)} = \sqrt{\gamma_1^{(e)}},\; \gamma_{L,2}^{(e)} = \sqrt{\gamma_2^{(e)}},$$

there is some linear or near-linear relationship with $u_{L,e}$, and $\xi^{(e)}$ is approximately constant. In the case of the location parameter $\mu_0^{(e)}$, this is a consequence of the

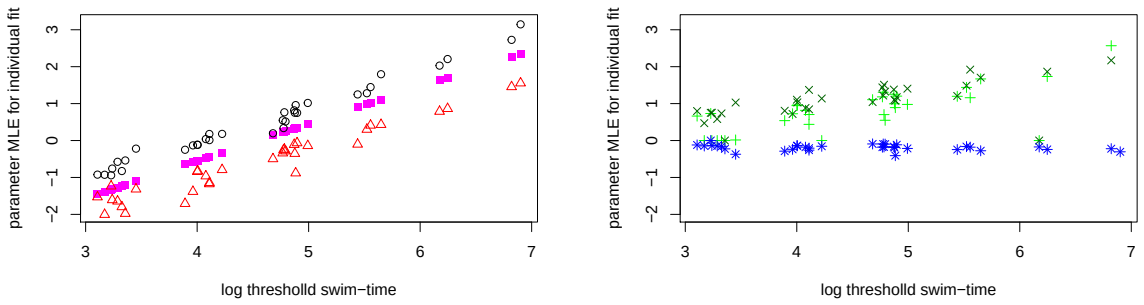


Figure 3.3.2: Transformed parameter estimates against log threshold swim-time $u_L = \log(-u)$. A linear or near-linear relationship is apparent for most of the parameters: for $\sigma_L = \log(\tilde{\sigma}_u)$ (black circle), $\mu_L = \log(-\mu_0)$ (purple square), $\beta_L = \log(\beta)$ (red triangle),$\gamma_{L,1} = \sqrt{\gamma_1}$ (light-green plus, $+$) and $\gamma_{L,2} = \sqrt{\gamma_2}$ (dark-green cross, $\times$). The shape parameter ξ (blue star) is approximately constant. Note that μ_L has been rescaled, by subtracting 5 uniformly, to be visible on the plot.

choice of threshold. More generally, power law relationships are commonly found in sports (Sylvan Katz and Katz, 1999), and the connection between u_e and $\tilde{\sigma}_u^{(e)}, \mu_0^{(e)}, \beta^{(e)}$ was hypothesised based on the prevalence of log-log relationships in sports modelling (Riegel, 1981). This relationship however, does not explain the dependence between swim-time and swim-suit effects $\gamma_1^{(e)}$ and $\gamma_2^{(e)}$ well, which is a combined result of the biomechanical and physical relationship between range of movement and flexibility, drag, buoyancy and total energy expenditure amongst other factors. The reason for this complex relationship is not explored in this article, but was chosen based on a Box-Cox transformation in the single suit case of γ to γ^*, such that

$$\gamma^* = \begin{cases} (\gamma^{\delta_\gamma} - 1)/\delta_\gamma & \delta_\gamma \neq 0, \\ \log(\gamma) & \delta_\gamma = 0, \end{cases}$$

where γ^* is assumed to come from a model which is linear in $u_{L,e}$ with a normal error distribution with constant variance. The choice of $\delta_\gamma = 1/2$ is consistent with the

Box-Cox transformation, which gives an MLE and 95% confidence interval of $\delta_\gamma = 0.52\ (0.37, 0.68)$, and also agreed with Box-Cox transformation applied to $\gamma_1^{(e)}$ and $\gamma_2^{(e)}$ in the two-suit case. Box-Cox transformations were also applied to the other parameters to confirm the log-log hypothesis, for example $\delta_\beta = 0.049\ (-0.11, 0.19)$, indicating that a log relationship is appropriate. These relationships motivate the across event model of the next section.

3.3.3 Across Event Model

Parametric Model

Now that models (3.3.1) and (3.3.2) have been shown to be suitable for each event, it is desired that information can be shared between events to reduce parameter uncertainty and improve predictive performance. By doing this we ensure that the across event model is more robust than models (3.3.1) and (3.3.2) with respect to anomalous data, which could lead to over-fitting.

A natural first step here would be to consider distance as a covariate and a log-log relationship. Distance does work well in athletics, as long as it is within the same gender (Riegel, 1981). However, distance does not work well when pooling across both genders, and across different strokes, since for example breaststroke is always slower than freestyle for a given distance and gender, and so inherent bias will be introduced due to the physical nature of the difference in strokes. Instead, the threshold swim-time is used as a covariate, since naturally slower strokes, whose corresponding scale parameters for example are likely to be larger, will also have a larger covariate, the threshold swim-time. This allows for a given parameter to vary smoothly across events, rather than to be discretised by the distance of the event. Thus, no adjustment is needed to compare between different strokes and genders.

From Figure 3.3.2 it is initially hypothesised that the shape parameter $\xi^{(e)}$ can be held constant across all events, and that the transformed parameters $\sigma_L^{(e)}$, $\mu_L^{(e)}$, $\beta_L^{(e)}$,

$\gamma_{L_1}^{(e)}$ and $\gamma_{L_2}^{(e)}$ increase linearly with $u_{L,e}$. A similar figure (not shown) exists for the single-suit parametrisation, which suggests linearity for $\gamma_L^{(e)}$ also. Thus, it is proposed that the parameters are pooled across the 34 events via the following model:

$$\begin{aligned}
\xi^{(e)} &= \xi, \\
\mu_L^{(e)} &= \alpha_1 + \vartheta_1 u_{L,e}, \\
\sigma_L^{(e)} &= \alpha_2 + \vartheta_2 u_{L,e}, \\
\beta_L^{(e)} &= \alpha_3 + \vartheta_3 u_{L,e}.
\end{aligned}$$

In the single-suit case,

$$\gamma_L^{(e)} = \alpha_4 + \vartheta_4 u_{L,e}, \tag{3.3.8}$$

and in the two-suit case,

$$\gamma_{L_1}^{(e)} = \alpha_4 + \vartheta_4 u_{L,e},\ \gamma_{L_2}^{(e)} = \alpha_4 + \varepsilon + \vartheta_4 u_{L,e}, \tag{3.3.9}$$

for some parameters $\boldsymbol{\psi} = \{\xi, \varepsilon,\ \{\alpha_i, \vartheta_i \in \mathbb{R} :\ i = 1, \ldots, 4\}\}$. Having two separate gradients, ϑ_4 and ϑ_5 such that $\gamma_{L_1}^{(e)} = \alpha_4 + \vartheta_4 u_{L,e}$, and $\gamma_{L_2}^{(e)} = \alpha_4 + \varepsilon + \vartheta_5 u_{L,e}$, was also considered, but it was found that a common gradient, such that $\vartheta_5 = \vartheta_4$, sufficed. In fact, several other models were considered (not reported), for example including a different intercept for men's and women's events in the linear model, or using separate linear models for different distances, but these were found to produce no improvement. The full likelihood of the across event parametric model then, assuming independence between events, is therefore given as

$$L(\boldsymbol{\psi}; \boldsymbol{x}) = \prod_{e \in E} \left\{ \exp\left[-\Lambda^{(e)}\left(\mathcal{A}_{1,u}\right)\right] \prod_{i=1}^{200} \int_{x_i^{(e)}-s/2}^{x_i^{(e)}+s/2} \lambda^{(e)}(t_i^{(e)}, x)\, \mathrm{d}x \right\}.$$

Model	Constraints	AIC/RIC	# ind. parameters
$\mathcal{M}_{1a}$	independent fits, single-suit (3.3.1)	0	170
$\mathcal{M}_{1b}$	independent fits, two-suits (3.3.2)	-23.7	204
$\mathcal{M}_2$	$\mathcal{M}_{1a}$ with constraint (3.3.4)	-38.7	137
$\mathcal{M}_3$	$\mathcal{M}_2$ with constraint (3.3.5)	-52.1	105
$\mathcal{M}_4$	$\mathcal{M}_3$ with constraint (3.3.6)	-29.6	73
$\mathcal{M}_5$	$\mathcal{M}_3$ with constraint (3.3.10)	-58.7	74.1
$\mathcal{M}_6$	$\mathcal{M}_5$ with constraint (3.3.7)	-87.7	42.1
$\mathcal{M}_{7a}$	$\mathcal{M}_6$ with constraint (3.3.8)	-90.6	10.3
$\mathcal{M}_{7b}$	$\mathcal{M}_{1b}$ with constraints (3.3.4), (3.3.5), (3.3.7), (3.3.9), (3.3.10)	-121.5	11.2

Table 3.3.1: Model comparison showing the AIC or RIC for each model, normalised by the independent fits model with a single suit, model $\mathcal{M}_{1a}$. The RIC, defined by expression (3.3.11), is used when a spline is fitted to a parameter over events and defines the number of effective degrees of freedom. A lower AIC or RIC indicates a better model fit.

This pooled structure was incrementally implemented as shown in Table 3.3.1. The first model fitted, $\mathcal{M}_{1a}$ pools no parameters and considers only a single suit, such that each event e has 5 independent parameters, $(\mu_0^{(e)}, \sigma_u^{(e)}, \xi^{(e)}, \beta^{(e)}, \gamma^{(e)})$, resulting in a total of 170 parameters. The AIC can be seen to improve from $\mathcal{M}_{1a}$ to $\mathcal{M}_{1b}$ by including the separate effect of two suits, despite the significant increase in the number of free parameters. From model $\mathcal{M}_{1a}$, the pooling structure begins to be implemented, and there is an improvement to $\mathcal{M}_2$, where now constraint (3.3.4) is introduced such that

all events share a common shape parameter. Again, the model fit improves from $\mathcal{M}_2$ to $\mathcal{M}_3$ by employing constraint (3.3.5), however, when trying to enforce linearity between $\sigma_L^{(e)}$ and $u_{L,e}$ across $e \in E$ via constraint (3.3.6), model $\mathcal{M}_4$, the fit was poorer. The events which mainly contributed to this worsened fit were the men's and women's 200m free and women's 50m fly, but the fit was also generally worse across the vast majority of events, which could be explained by some non-linearity observed in Figure 3.3.2.

The inadequacy of a linear relationship (3.3.6) between σ_L and u_L suggests that a fully parametric model to describe this relationship was slightly too restrictive, and motivates the need for a more flexible but parsimonious model, for which we use semi-parametric techniques. Model $\mathcal{M}_5$ was therefore introduced which relaxes the linear constraint (3.3.6) on σ_L, and instead uses the spline based non-parametric approach described in Section 3.3.3, which lets the smooth dependence of σ_L on u_L to be captured by allowing the data to govern the precise nature of this relationship, whilst keeping the dependencies of μ and ξ on u_L the same and keeping β and γ unconstrained, as in model $\mathcal{M}_4$. From here, models $\mathcal{M}_6$ and $\mathcal{M}_{7a}$ are then fitted by cumulatively employing constraints (3.3.7) and (3.3.8) respectively, and finally $\mathcal{M}_{7b}$ is fitted by the addition of an extra suit parameter to $\mathcal{M}_{7a}$, see Table 3.3.1. The best fitting model, determined via regularisation information criteria (RIC) (Shibata, 1989) which is defined by expression (3.3.11), is $\mathcal{M}_{7b}$ with only approximately 11 parameters. Critically, note the substantial improvement from models $\mathcal{M}_{7a}$ to $\mathcal{M}_{7b}$, showing a clear impact of changes in full body suit technology over the period when these suits were allowed.

Confidence intervals were found via parametric bootstrapping, such that model $\mathcal{M}_{7b}$ was re-fitted to 250 simulated datasets, to estimate the sampling distribution of parameter estimators. The number of observations from event e in simulated dataset j, $N_j^{(e)}$ is simulated directly via, $N_j^{(e)} \sim \text{Poisson}\left(\Lambda^{(e)}\left(\mathcal{A}_{1,u}\right)\right)$. For an event e and replication j, $N_j^{(e)}$ swim-times $x_1^{(e)}, \ldots, x_{N_j^{(e)}}^{(e)}$ and the time of these swims $t_1^{(e)}, \ldots, t_{N_j^{(e)}}^{(e)}$ were generated via a probability integral transform on equation (3.2.5) for the swim-

times, and the distribution function (3.2.8) integrated over x for the times respectively. Some of the resulting bootstrapped parameter estimates resulted in infeasible estimates, for example inferring that the ultimate possible swim-time is worse than some swim-times in the original data set, or that the expected next world record swim-time is worse than the current world record, and so these data sets were discarded. The remaining 240 data sets quantify the natural variation in the data and thus provide the basis for obtaining confidence intervals. All confidence intervals referred to subsequently in this article are obtained via this method.

The estimated values for ϑ_3 and ϑ_4 under model $\mathcal{M}_{7b}$, the associated gradients for the trend parameters and swim-suit parameters respectively, were $\hat{\vartheta}_3 = 0.940$ $(0.936, 0.942)$ and $\hat{\vartheta}_4 = 0.460$ $(0.432, 0.470)$. The relative confidence interval widths are smaller on ϑ_3 than ϑ_4, and this is likely due to the swim-suit parameter being dependent on less data than the trend parameter, since only data in swim-suit years effect it. In comparison, the gradient governing the linear relationship (3.3.5) is estimated at $\hat{\vartheta}_1 = 1.0016$ $(1.0010, 1.0019)$. The tight confidence intervals here indicate the strong relationship between u_L and μ_L.

Semi-Parametric model

To achieve the appropriate flexibility to model the relationship observed in Figure 3.3.2 between σ_L and u_L we use a d-degree spline function (De Boor, 1978), which is a piecewise polynomial function that is constructed to be continuous and d times continuously differentiable over a closed interval domain. It is a weighted linear sum of q, d-degree basis splines, called B-splines, with the k^{th} B-spline $B_k(x)$ centred on a *knot* at point x_k. The spline function used for σ_L is denoted by

$$\sigma_L(u_L) = \sum_{k=1}^{q} a_k B_k(u_L) \tag{3.3.10}$$

where a_k is the k^{th} element of the spline coefficient vector $\boldsymbol{a} = (a_1, \ldots, a_q)$ which is constant over all events such that, given a vector $\boldsymbol{a}$, the value of σ_L for any given event e is a function of u_L only, see Appendix A for further details.

Although function (3.3.10) can model any non-linear relationship, we wish for this relationship to be smooth and increasing. In order to enforce this smoothness, the likelihood function is extended to a penalised likelihood which contains a roughness penalty. The penalty is governed by $\phi_r p_r = \phi_r \boldsymbol{a}^T P \boldsymbol{a}$, where $P \in \mathbb{R}^{q \times q}$ is the penalty matrix, and $\phi_r > 0$ determines the amount of penalisation. The choice of P determines the nature of the penalty and is chosen based on the form of the data, or some prior belief. In this case a 2nd order penalty on the finite differences of adjacent coefficients (Eilers and Marx, 1996), and a degree $d = 4$ spline was chosen, see Appendix A. This penalises σ_L having a large second derivative, and penalises fits for σ_L that depart from linearity. Additionally, since it is believed apriori that the GPd scale parameter is an increasing function of the threshold swim-time, a hard constraint $\phi_m p_m$ ensures monotonicity in the spline function, where p_m is defined as follows: allow

$$\{z_1, \ldots, z_k\} = \left\{ \min_{e \in E} u_{L,e} \right\} \cup \left\{ x_i : \frac{\mathrm{d}\sigma_L(x_i)}{\mathrm{d}x} = 0,\ i = 2, \ldots, k-1 \right\} \cup \left\{ \max_{e \in E} u_{L,e} \right\}$$

to be a discrete set of size k containing all stationary points and end points of the spline function, then

$$p_m = -\sum_{i=1}^{k-1} \left(\sigma_L(z_{i+1}) - \sigma_L(z_i)\right) \mathbb{1}\left\{\sigma_L(z_{i+1}) - \sigma_L(z_i) < 0\right\}.$$

With the GPd scale parameter $\tilde{\sigma}_u$ for a particular event e being defined by the spline via

$$\tilde{\sigma}_u^{(e)} = \exp\left[\sum_{k=1}^{q} a_k B_k(u_{L,e})\right],$$

the full joint penalised likelihood across all events becomes

$$L_p(\boldsymbol{\varphi}, \phi_r, \phi_m; \boldsymbol{x}) = \prod_{e \in E} \left\{ \exp\left[-\Lambda^{(e)}\left(\mathcal{A}_{1,u}\right)\right] \prod_{i=1}^{200} \int_{x_i^{(e)}-s/2}^{x_i^{(e)}+s/2} \lambda^{(e)}(t_i, x)\, \mathrm{d}x \right\} \exp\left[-(\phi_r p_r + \phi_m p_m)\right],$$
$$= L(\boldsymbol{\varphi}; \boldsymbol{x}) \exp\left[-(\phi_r p_r + \phi_m p_m)\right],$$

where $\boldsymbol{\varphi}$ are the parameters of the model, and L is the unpenalised likelihood. The penalised log-likelihood for model $\mathcal{M}$ is therefore given as

$$\ell_p(\mathcal{M}) = \ell(\mathcal{M}) - \phi_r p_r - \phi_m p_m,$$

where ℓ is the unpenalised log-likelihood, and $\phi_m > 0$ is sufficiently large such that monotonicity is a hard constraint. The value of ϕ_m is found by finding a ϕ_m such that

$$\max\left(\ell_p(\mathcal{M}|\phi_m)\right) = \max\left(\ell_p(\mathcal{M}|\phi_m + \epsilon)\right),$$

for any $\epsilon > 0$. Theoretically, this can be found by allowing $\phi_m \to \infty$, however it can be difficult for optimisation routines to converge to this global maxima. Therefore, in practise ϕ_m is increased iteratively by initially setting $\phi_m = 0$ and finding the parameter that give $\max\left(\ell_p(\mathcal{M}|\phi_m = 0)\right)$. Then ϕ_m is increased iteratively, using the previous solution as the initial starting parameters, until there is no change in $\mathcal{M}$ and therefore also no change in $\ell(\mathcal{M})$. Instead of a constraint on the spline function itself to enforce monotonicity, I-splines (Ramsay, 1988) could have been used as a basis instead of B-splines, and then positivity constraints on the basis splines would have enforced monotonicity. This construction may have resulted in more efficient computation, but would yield essentially identical model fits and results.

The choice of ϕ_r is selected using 10-fold cross validation to maximise model predictive performance at data points not used for fitting (Ewans and Jonathan, 2008). The model is fitted based on a random stratified sample of 90% of the data, the training

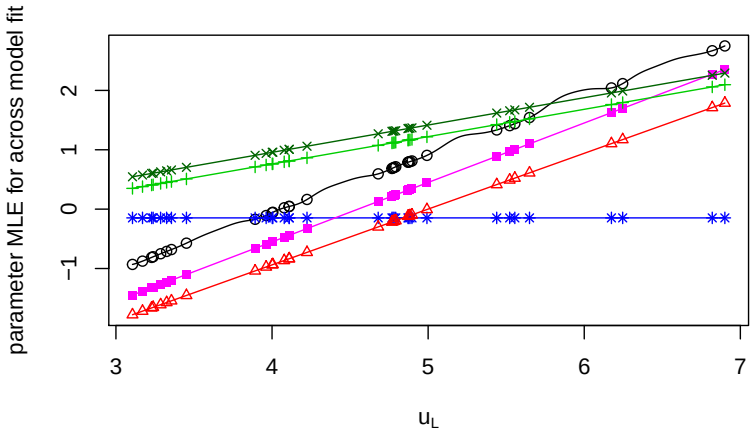


Figure 3.3.3: Fitted parameters for model $\mathcal{M}_{7b}$, as a function of u_L: $\sigma_L(u_L)$ (black circles) is governed by the spline, whilst $\beta_L(u_L)$ (red triangles), $\mu_L(u_L)-5$ (purple squares), $\gamma_{L_1}(u_L)$ light green pluses, $+$) and $\gamma_{L_2}(u_L)$ (dark green crosses, $\times$) vary linearly with u_L. The shape parameter (blue stars) has a constant value of $\hat{\xi} = -0.147$ $(-0.152, -0.143)$. Note that μ_L has been rescaled, by subtracting 5 uniformly, to be visible on the plot.

data, which is then used calculate the log-likelihood based on the remaining 10% of the data, the test data. The log-likelihood for each of the 10 non-overlapping sets of test-data is summed to obtain a 'predictive' log-likelihood based on the prediction accuracy of the model. This process is repeated 20 times at a range of different values of ϕ_r, with the value of ϕ_r which corresponds to the best average predictive performance being selected as the optimum penalty. It was found that the change in predictive log-likelihood was robust to changes in ϕ_r, and it is thought that this is due to the hard constraint on monotonicity already accounting for much of the variability in the spline fits. For model $\mathcal{M}_{7b}$, an optimum penalty of $\phi_r = 15$ was found. Given this, the full model can be fitted and the parameters as a function of u_L are shown in Figure 3.3.3.

Since models $\mathcal{M}_5$, $\mathcal{M}_6$, $\mathcal{M}_{7a}$ and $\mathcal{M}_{7b}$ are semi-parametric, AIC can no longer be used as a model comparison tool since the number of degrees of freedom is not defined. Instead, RIC is used, which uses the effective degrees of freedom g, as opposed to

degrees of freedom. Otherwise, RIC is defined identically to AIC, that is

$$\text{RIC} = -2\ell(\boldsymbol{\varphi}) + 2 \text{ tr} \left[I(\boldsymbol{\varphi}) J(\boldsymbol{\varphi}, \phi_r, \phi_m)^{-1} \right], \tag{3.3.11}$$

such that $g = \text{tr}\left[I(\boldsymbol{\varphi})J(\boldsymbol{\varphi}, \phi_r, \phi_m)^{-1}\right]$ where I is the observed Fisher information criteria of the unpenalised likelihood L, J is the negative Hessian matrix of the penalised log-likelihood L_p, and $\text{tr}(A)$ is the trace of the square matrix A.

Assessment of model $\mathcal{M}_{7b}$ fit

The rate of exceedances and the distribution above threshold must both be considered to determine the overall quality of the selected model fit. A pooled PP plot is used to determine how well the model fits the distribution of swim-times above threshold. The pooled PP plot, Figure 3.3.4 (left), allows the combined fit of all 34 events to be analysed at once. The fit generally is very good, especially considering the reduction from 204 to 11.2 parameters. The areas of weaker fit can mainly be attributed to two events, the 200m men's free, and the 50m men's fly. These two events increase the RIC by 10.4 and 9.7 respectively, both of which is significant evidence of lack of fit, so caution should be exercised when drawing conclusions from these two events. Somewhat surprisingly though, we find that removing these events from the analysis makes no substantial difference to the diagnostic shown in Figure 3.3.4 (left). Figure 3.3.4 (right) shows another pooled PP plot, using the same model fit, but only using data from the period $[2001, 2003]$. These data also appear to be fit very well, and this implies that any potential bias introduced by the early period data selection problems, highlighted in Section 3.3.1, is minimal.

A nice feature of this pooled model is that natural ordering across different strokes is preserved even for events which carry a less good fit. For example, the parameters for the 50m men's fly will always indicate that it is a faster event than the 50m men's breaststroke, i.e., by predicting a faster ultimate possible swim-time or next world

record swim-time.

Figure 3.3.5 shows the expected rate of observations exceeding u_e per year, compared to what was observed in the data, for the women's 100m freestyle. Similar plots for all 34 events were examined (not shown). It can be seen that the observed rate of observations almost always falls between (and once only marginally outside) the 95% confidence intervals, including during the swim-suit era and the early period of the database when competition selection may have induced bias as identified in Section 3.3.1. The estimated expected number of observations is not systematically above or below the observed number of observations. For a year in which the observed rate is higher than expected, often in the next year this observed rate is below the expected rate, which is due to the discrete nature of the plot.

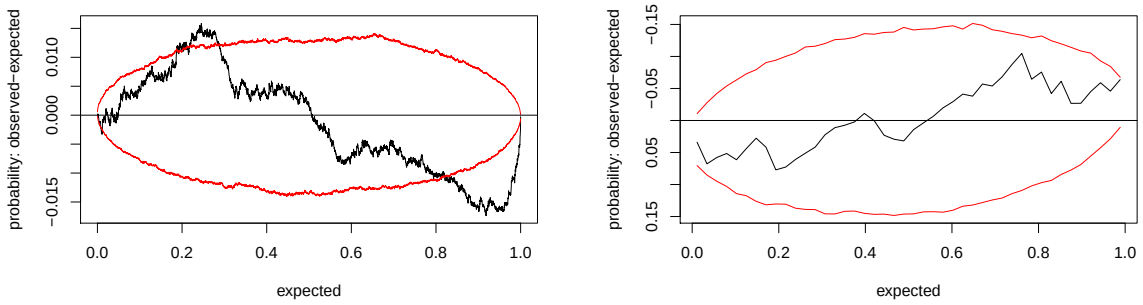


Figure 3.3.4: PP plot (plotted as observed minus expected probabilities) pooled over all events, with 95% tolerance intervals, using both the whole data set (left) and only data from $[2001, 2003]$ (right).

3.4 Results from Model

3.4.1 Rankings

From fitting model $\mathcal{M}_{7b}$, the final rankings of the best ever swim-times can be constructed. The rankings are determined by the r-value of a swim-time x, that is, the

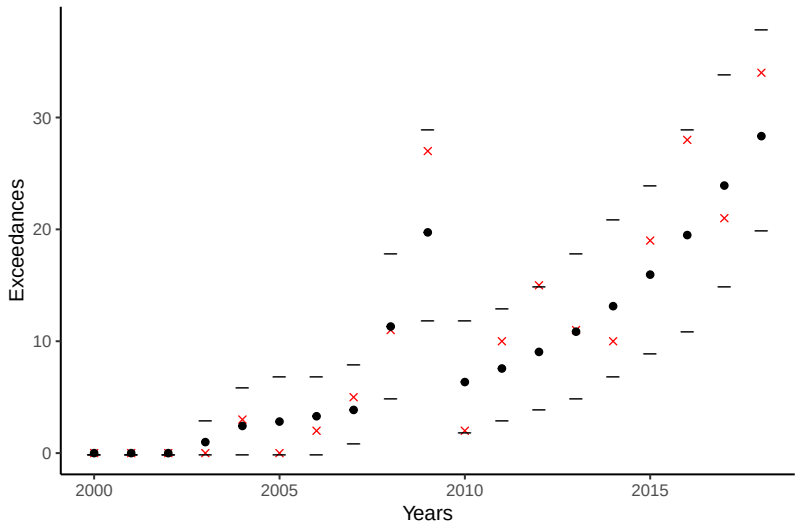


Figure 3.3.5: The estimated expected (black circles) and observed (red crosses) number of observations per year better than u_e for women's 100m freestyle, with 95% confidence intervals for the estimated values given by the lower and upper horizontal lines. The two swim-suit years, 2008 and 2009, have increased rates of exceedances relative to neighbouring years.

rate at which observations better than x occur in the given event. If $X_t^{(e)}$ is the random variable denoting a new observed negative swim-time in event e at a time t where this swim-time is better than u_e, then the expected rate R at which an observation $X_t^{(e)}$ is faster than swim-time x occurs is defined as follows:

$$\begin{aligned} R\{X_t^{(e)} > x + s/2\} &= \Pr\{X_t^{(e)} > x + s/2 | X_t^{(e)} > u_e\} \Lambda_{y(t)}^{(e)}(\mathcal{A}_{1,u}) \\ &= \bar{H}_u^{(e)}(x + s/2) \Lambda_{y(t)}^{(e)}(\mathcal{A}_{1,u}) \\ &\approx \left[1 + \xi\left(\frac{x + s/2 - u_e}{\tilde{\sigma}_u^{(e)}}\right)\right]_+^{-\frac{1}{\xi}} \left[1 + \xi\left(\frac{u_e - \mu^{(e)}(y^*(t))}{\sigma^{(e)}(y^*(t))}\right)\right]_+^{-\frac{1}{\xi}} \end{aligned} \quad (3.4.1)$$

for all $x + s/2 > u_e$, where the final approximation follows from equation (3.2.11), where $y(t)$ is the year in which $X_t^{(e)}$ occurs and $y^*(t) = y(t) + 1/2$ is the mid point of years $y(t)$ and $y(t) + 1$. An estimate of $R\{X_t^{(e)} > x + s/2\}$ gives the r-value, and therefore a measure of the 'quality' of the swim-time x. By adding $s/2$, the censoring is taken into consideration, since the true observed swim-time $X_t^{(e)}$ would need to be faster by an amount greater than the precision of the data to be recorded as being faster.

Figure 3.4.1 shows the best 20 swimmers from the 2001 to end of 2018 period, based on the r-value of their swim. Note that swimmers names can occur multiple times where they have recorded swim-times in more than one event. The error bars show the 95% confidence intervals from the parametric bootstrapping. It is also possible to quantify how much better one swimmer is than another by analysing what proportion of time the bootstrapped samples give one swimmer ranked ahead of another. For example, Adam Peaty, ranked 12th, beats Katinka Hosszu, ranked 11th, on 48% of rankings from the bootstrapped data sets. In contrast, Katie Ledecky's 1500m free performance, ranked 2nd, never beats top ranked Sarah Sjostrom's 50m fly performance, giving strong evidence for ranking Sarah Sjostrom better.

The lower confidence intervals for the ranks of both Zige Liu and Lin Zhang are much wider in comparison to the others in the top 20, and one possible reason is that they

were swam during the second swim-suit period, S_{t_2} (2009). As noted in Section 3.3.3, the relative uncertainty for ϑ_4, which controls the swim-suit effect, is comparatively large, and this added uncertainty propagates through to the rankings. Essentially, the confidence intervals are showing that, if the parameter associated with the 2009 swim-suit is overvaluing the effect of this suit, then their true ranks could be much lower. This same effect is not seen in Paul Biedermann's rank, also swam in 2009, however this was in the 200m men's free which has previously been identified as an area of weaker fit.

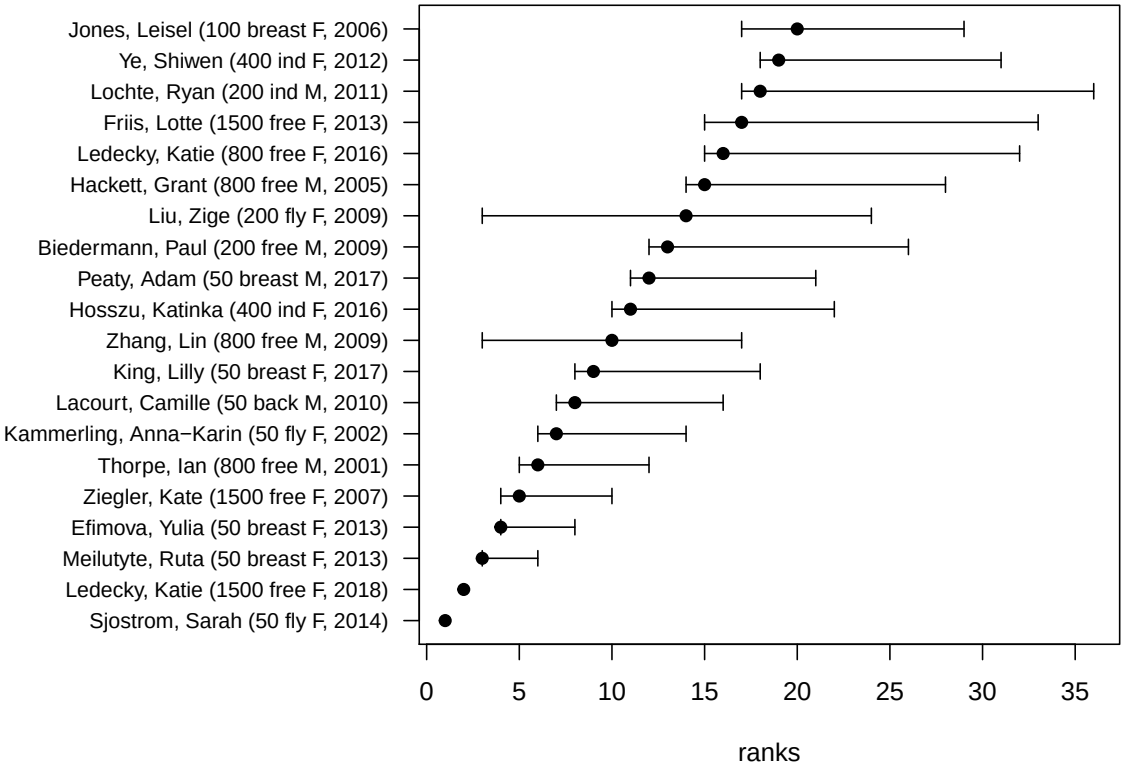


Figure 3.4.1: The ranking of the top 20 swimmers from the data set, with 95 % CIs from bootstrapped data sets. Better ranked swimmers are lower on the y-axis.

Interestingly, in some cases the time when the swim was performed can effect the

order of the rank within the same event. Ruta Meilutyte, Yulia Efimova and Lilly King hold ranks 3, 4, and 9 respectively, all from the 50m women's breaststroke, however the fastest time of the three is Lilly King's with a time of 29.40 seconds in July 2017, compared to times of 29.48 seconds and 29.52 seconds for Ruta Meilutyte and Yulia Efimova respectively, which were both swam in July 2013, five years earlier, which indicates that they achieved comparatively better results given their era.

It is worth noting that 7 of the top 20 estimated ranked swims occur in 50m races, which is approximately 50% more than the number that would be expected if the assumption that all events are equally competitive holds. In fact, this assumption is unlikely to hold in practice, since the 50m backstroke, breaststroke and fly are non-Olympic events, and as such the competitiveness of these events may be less than the Olympic events, which increases the disparity between the observed and expected number of 50m races in the top 20 ranks. Conversely, the top 20 rankings for the independent fits model $\mathcal{M}_{1a}$ and $\mathcal{M}_{1b}$ (not shown), were found to be proportionately represented by all distances. In models $\mathcal{M}_{1a}$ and $\mathcal{M}_{1b}$ fits the 50m events have larger corresponding shape parameters than other events on average, and than the common shape parameter for $\mathcal{M}_{7b}$, particularly the men's fly and women's and men's free had comparatively much larger shape parameters than other events. Therefore, it was initially thought that high rankings of swimmers in the 50m events may be due to the enforcing of a constant shape parameter across all events, and so perhaps a different modelling strategy is required for the shorter events. However, it was found from calculating profile likelihood based 95% confidence intervals that the shape parameters were -0.067 $(-0.221, -0.045)$, 0.000 $(-0.173, 0.013)$ and -0.080 $(-0.253, -0.090)$ for the men's fly and women's and men's free respectively, which all overlap with the shared shape parameter of model $\mathcal{M}_{7b}$, $\hat{\xi} = -0.147$ $(-0.152, -0.143)$. Thus, it appears that the comparatively larger shape parameters for the 50m events is mostly due to natural variation. A more formal test for a different shape parameter, common to all 50m

events, would be to evaluate the RIC under a model with indicator covariates for the shape parameters for these 50m events. This was not considered necessary given the evidence from the profile likelihood intervals given above. Notably, the ordering of the very top four ranks was the same under both $\mathcal{M}_{1b}$ and $\mathcal{M}_{7b}$.

A national rankings table can also be made by only including a given nation in the comparison, and could be used for that nation's Olympics selection, for example. This would also change the confidence intervals for the rankings as swimmer's are only compared to others from the same nation.

3.4.2 Ultimate times

Finding limits to human sports performance has interested academics for years, in athletics for example Blest (1996). In swimming, Nevill et al. (2007) attempt to determine the ultimate possible time by analysing world record swims from 1957 to 2007, and Huub and Trultens (2005) approach this from a biomechanical perspective. In this article, the ultimate time is determined from the GPd function.

It was found that the MLE for the shape parameter with 95% confidence intervals was $\hat{\xi} = -0.147$ $(-0.152, -0.143)$ which (since $\hat{\xi} < 0$) can be interpreted as there being a finite bound on the best possible time a human can achieve in any given event. In many applications getting such a narrow confidence intervals, and hence such clear evidence $\xi < 0$, is difficult to achieve. Here this has been enabled by the pooling of data from all 34 events, giving a sample from the model of 6800 observations to inform us of the value of ξ. The ultimate possible time for an event can be estimated directly from the parameter estimates since for event e there exists an end-point $x_{H,e} = u_e - \tilde{\sigma}_u^{(e)}/\xi$: $H_u^{(e)}(x) = 1,\ \forall\ x > x_{H,e}$. Note that $x_{H,e}$ is covariate independent in the selected model, which seems reasonable since we expect the gap between the ultimate possible time and the world record to shrink as world records improve, but the ultimate time is still unreachable and 'set-in-stone'. For example, the MLE for the ultimate possible time

for the men's 100m breaststroke given by the model is 53.81 (53.60, 53.97) seconds. In comparison, Adam Peaty's fastest time in the dataset is 57.10 seconds from 2018. This 3 second difference made Peaty's *Project 56* (https://www.bbc.co.uk/sport/av/swimming/40650276), his challenge to swim a sub 57s 100m breaststroke, seem more achievable than at first glance. In fact, Peaty has since succeeded in his Project 56, setting a new world record of 56.88 seconds in 2019.

For each event, Figure 3.4.2 shows these estimated ultimate times normalised by the corresponding current world records as of the beginning of 2019, ordered by increasing threshold swim-times. For the vast majority of events, the ultimate time is 93-95% of the current world record. For the women's 50m butterfly and women's 1500m freestyle however, the current world record is very close to the ultimate time. In fact, approximately a 3% improvement would see these ultimate times being reached. This finding is not so surprising as these two world records correspond to the top two ranks, Sarah Sjostrom and Katie Ledecky from Figure 3.4.1. In comparison, the world record swim-time for the men's 100m free, which does not make the top 20 ranks, would require a 7% improvement to reach the ultimate time, suggesting this is the weakest of the current world records.

3.4.3 Expected new world record time

Let $X_t^{*(e)}$ be the random variable denoting the swim-time of a new world record in event e at time t, then the distribution of $X_t^{*(e)}$ follows immediately from equations (3.2.4) and (3.2.5), i.e.,

$$\Pr\{X_t^{*(e)} > x\} = \Pr\{X_t^{(e)} > x | X_t^{(e)} > r_e\} = \bar{H}_{r_e}^{(e)}(x) = \left[1 + \xi\left(\frac{x - r_e}{\tilde{\sigma}_{r_e}^{(e)}}\right)\right]_+^{-\frac{1}{\xi}}, \text{if } x > r_e, \tag{3.4.2}$$

where $\tilde{\sigma}_{r_e}^{(e)} = \sigma_0^{(e)} + \xi(r_e - \mu_0^{(e)})$ and r_e is the world record for event e at the end of 2018 such that $r_e := \max(\boldsymbol{x}_e)$ where $\boldsymbol{x}_e$ are all the observations in event e. Note that the

right hand side of expression (3.4.2) has no time dependency, since under model $\mathcal{M}_{7b}$ the distribution of times, conditional on being above threshold u_e, is time homogeneous for any given event e, see property (3.3.3), therefore we drop the subscript t. From this the expected swim-time of the next world record in event e is

$$\mathbb{E}[X^{*(e)}] = \int_{r_e}^{x_{H,e}} x \frac{\mathrm{d}H_{r_e}^{(e)}(x)}{\mathrm{d}x}\,\mathrm{d}x = r_e + \frac{\tilde{\sigma}_{r_e}^{(e)}}{1-\xi}, \text{ if } \xi < 1.$$

Figure 3.4.2 shows the estimated expected swim-time of the next world record relative to the world record at the end of 2018, where events are ordered by increasing swim-time. Censoring is ignored in this calculation, as it would have such a negligible effect. The expected improvement varies only slightly between events, ranging from an expected improvement of 0.5% for Katie Ledecky's 1500m women's free performance, to a 0.9% for Cesar Cielo's 100m men's free performance. In events where the ultimate time is close to the current record, the expected next world record is also closer to the current record, and vice versa.

The small variation between expected improvement is at first surprising, since it might be expected that 'better' records, such as those of Katie Ledecky and Sarah Sjostrom, would be beaten by much smaller amounts. However, it is also likely that these records will take longer to be broken and so the improvements in training methods will be more significant by the time a new record is set, which may reduce the variation in the percentage improvement.

The confidence intervals here describe the confidence in the mean of the corresponding estimate, but it might also be interesting to determine the prediction interval, e.g., the 95% interval of possible swim-times that the next world record swim-time in event e will be in. The predictive distribution of $\Pr\{X^{*(e)} < x\}$ can be found as follows: if $\{\hat{\Theta}^{(i)} : \ i = 1, \ldots, n\}$ are the $n = 240$ bootstrapped parameter estimates, where $\hat{\Theta}^{(i)}$

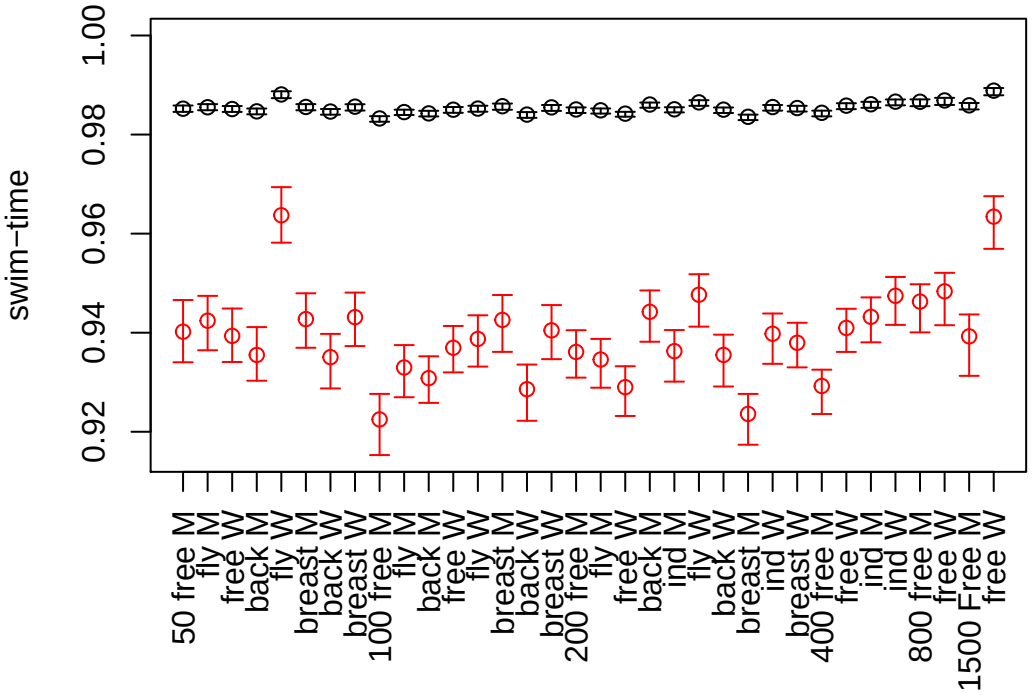


Figure 3.4.2: The estimated expected next world record swim-time (upper black) and ultimate possible time (lower red) for each event the values are rescaled by world record as at the end of 2018, with 95 % CI's from bootstrapped data sets.

corresponds to the MLE's from simulated data set i, then for large n and $x > r_e$

$$\begin{aligned}\Pr\{X^{*(e)} < x\} \approx & \frac{1}{n}\sum_{i=1}^{n} \Pr\{X^{*(e)} < x|\hat{\Theta}^{(i)}\} \\ = & \frac{1}{n}\sum_{i=1}^{n} H_{r_e}^{(e)}(x|\hat{\Theta}^{(i)}) \\ = & 1 - \frac{1}{n}\sum_{i=1}^{n}\left[1 + \xi^{(i)}\left(\frac{x - r_e}{\tilde{\sigma}_{r_e}^{(i,e)}}\right)\right]_{+}^{-\frac{1}{\xi^{(i)}}},\end{aligned}$$

where $\xi^{(i)}$ and $\tilde{\sigma}_{r_e}^{(i,e)}$ are the bootstrapped parameter estimates for ξ and $\tilde{\sigma}_{r_e}^{(e)}$ corresponding to simulated data set i. Similar predictive distributions can be found for the other features of interest in Figures 3.4.3 and 3.4.4, as described in Sections 3.4.4 and 3.4.5.

3.4.4 Time until world record is next set for an event

The distribution of time taken until a new world record is set in a particular event e is of interest. Let $T^{(e)}$ be a random variable describing the time at which a new world record is next set in event $e \in E$. The probability $F_{T^{(e)}}(t) = \Pr\{T^{(e)} < t\}$ that a world record for event e is set before some time t can be found as follows. For current time 1, until a time t ($t > 1$) there will be $N_t^{(e)}$ exceedances of the threshold u_e in event e, and for the current record to be first broken after t all of the $N_t^{(e)}$ observations need to be slower than the current record. Therefore, the following notation is introduced: let $X_{1:N_t^{(e)}}^{(e)} = \{X_i^{(e)}, i = 1, \ldots, N_t^{(e)}\}$ where $X_i \overset{\text{iid}}{\sim} H_u^{(e)}$ and $H_u^{(e)}$ has GPd. Then $N_t^{(e)}$ has a Poisson distribution with mean

$$\Lambda^{(e)}\left(\mathcal{A}_{(1,t),u}\right) = \int_1^t \left[1 + \xi\left(\frac{u_e - \mu^{(e)}(y)}{\sigma^{(e)}(y)}\right)\right]_{+}^{-\frac{1}{\xi}} \mathrm{d}y,$$

and the probability that a world record for event e is set before t is

$$\begin{aligned} F_{T^{(e)}}(t) &= 1 - \Pr\{T^{(e)} > t\} \\ &= 1 - \sum_{m=0}^{\infty} \Pr\{\max(X^{(e)}_{1:N_t^{(e)}}) < r_e | N_t^{(e)} = m\} \Pr\{N_t^{(e)} = m\} \\ &= 1 - \sum_{m=0}^{\infty} \left[H_u^{(e)}(r_e)\right]^m \left[\Lambda^{(e)}\left(\mathcal{A}_{(1,t),u}\right)\right]^m \exp\left[-\Lambda^{(e)}\left(\mathcal{A}_{(1,t),u}\right)\right]/m! \\ &= 1 - \exp\left[-\Lambda^{(e)}\left(\mathcal{A}_{(1,t),u}\right)\bar{H}_u^{(e)}(r_e)\right], \end{aligned} \tag{3.4.3}$$

where the final equality follows from the power series expression for the exponential function. The density function for $T^{(e)}$, $f_{T^{(e)}}$, follows from equation (3.4.3), as

$$f_{T^{(e)}}(t) = \left[1 + \xi\left(\frac{u_e - \mu^{(e)}(t)}{\sigma^{(e)}(t)}\right)\right]_+^{-\frac{1}{\xi}} \bar{H}_u^{(e)}(r_e) \exp\left[-\Lambda^{(e)}\left(\mathcal{A}_{(1,t),u}\right)\bar{H}_u^{(e)}(r_e)\right].$$

Then the expected time until a world record is next set in event e is

$$\mathbb{E}\left[T^{(e)}\right] = \int_1^{\infty} t f_{T^{(e)}}(t)\,\mathrm{d}t.$$

Figure 3.4.3 shows these MLE's along with 95% confidence intervals for $\mathbb{E}\left[T^{(e)}\right]$. It can be seen that almost all events are expected to have a new world record in the next 5 years. The longest estimated expected waiting times are again the times until Katie Ledecky's and Sarah Sjostrom's world records are broken, in the women's 1500m free and women's 50m fly respectively which correspond to the top two ranks of Figure 3.4.1, which both have expected waiting times of approximately 11 years.

3.4.5 Probability that a record is next set in a particular event

Now suppose that we wish the find the probability that the next event to have a world record that is broken is in event e. Let $T^{(-e)}$ be the random variable denoting the time

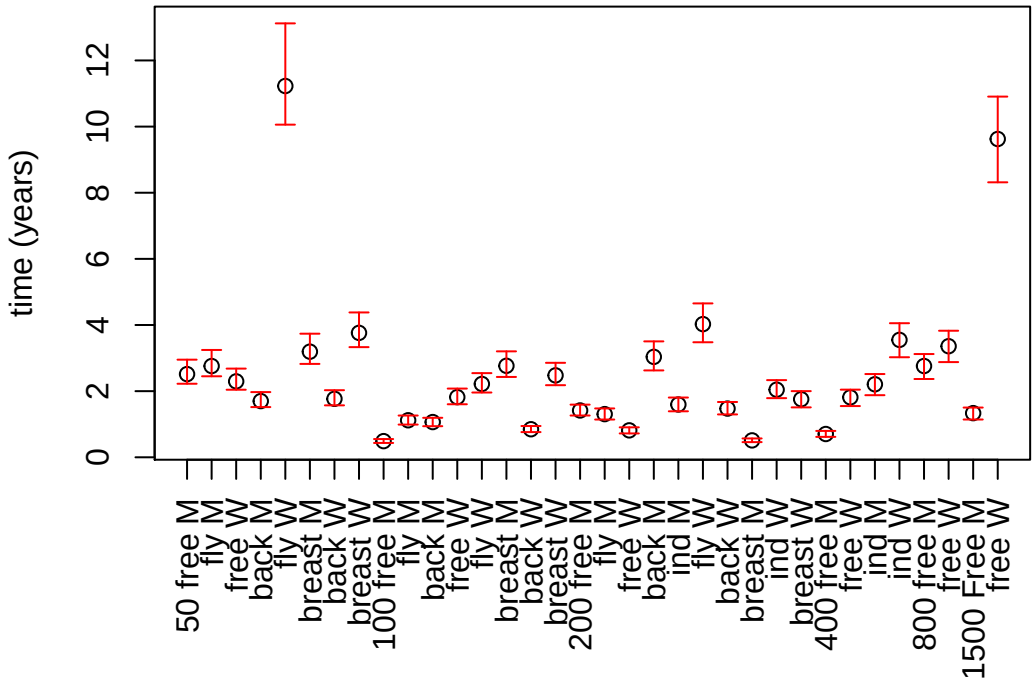


Figure 3.4.3: The estimated expected time (in years) until the world record is broken with 95% CI's from bootstrapped data sets.

taken for a world record to be set in any other event apart from e, i.e.,

$$T^{(-e)} := \min_{k \in E \setminus \{e\}} \{T^{(k)}\}.$$

Then the probability that the next world record that is set is in event e is given by

$$\begin{aligned}
&\Pr\{T^{(-e)} > T^{(e)}\} \\
&\quad = \int_1^\infty \Pr\{T^{(-e)} > T^{(e)} | T^{(e)} = t\} \Pr\{T^{(e)} = t\} \, \mathrm{d}t \\
&\quad = \int_1^\infty \prod_{k \in E \setminus \{e\}} \left\{ \exp\left[-\Lambda^{(k)}\left(\mathcal{A}_{(1,t),u}\right) \bar{H}_u^{(k)}(r_k) \right] \right\} \\
&\qquad \left[1 + \xi \left(\frac{u_e - \mu^{(e)}(t)}{\sigma^{(e)}(t)} \right) \right]_+^{-\frac{1}{\xi}} \bar{H}_u^{(e)}(r_e) \exp\left[-\Lambda^{(e)}\left(\mathcal{A}_{(1,t),u}\right) \bar{H}_u^{(e)}(r_e) \right] \mathrm{d}t \\
&\quad = \int_1^\infty \left\{ \exp\left[-\sum_{k \in E} \Lambda^{(k)}\left(\mathcal{A}_{(1,t),u}\right) \bar{H}_u^{(k)}(r_e) \right] \right\} \left[1 + \xi \left(\frac{u_e - \mu^{(e)}(t)}{\sigma^{(e)}(t)} \right) \right]_+^{-\frac{1}{\xi}} \bar{H}_u^{(e)}(r_e) \, \mathrm{d}t,
\end{aligned}$$

where the second equality follows because

$$\Pr\{T^{(-e)} > T^{(e)} | T^{(e)} = t\} = \prod_{k \in E \setminus \{e\}} \left\{ \exp\left[-(\Lambda^{(k)}\left(\mathcal{A}_{(1,t),u}\right) \bar{H}_u^{(k)}(r_k) \right] \right\}$$

due to the assumption of independence between swims in different events and the result derived in equation (3.4.3) for a single event. Figure 3.4.4 shows these estimated probabilities with the previously identified 'better' records having a lower probability of being broken next. The most likely record to be broken is the men's 100m free. The estimates of these probabilities using model $\mathcal{M}_{1b}$ was compared (not shown), and it has less variance between events.

3.4.6 Adjusting Swim-Suit Influenced Times

In 2010 Brazil's Cesar Cielo called for FINA to scrap any records set in the now-banned swim-suits, due to those records being much more difficult to break. Rather than this

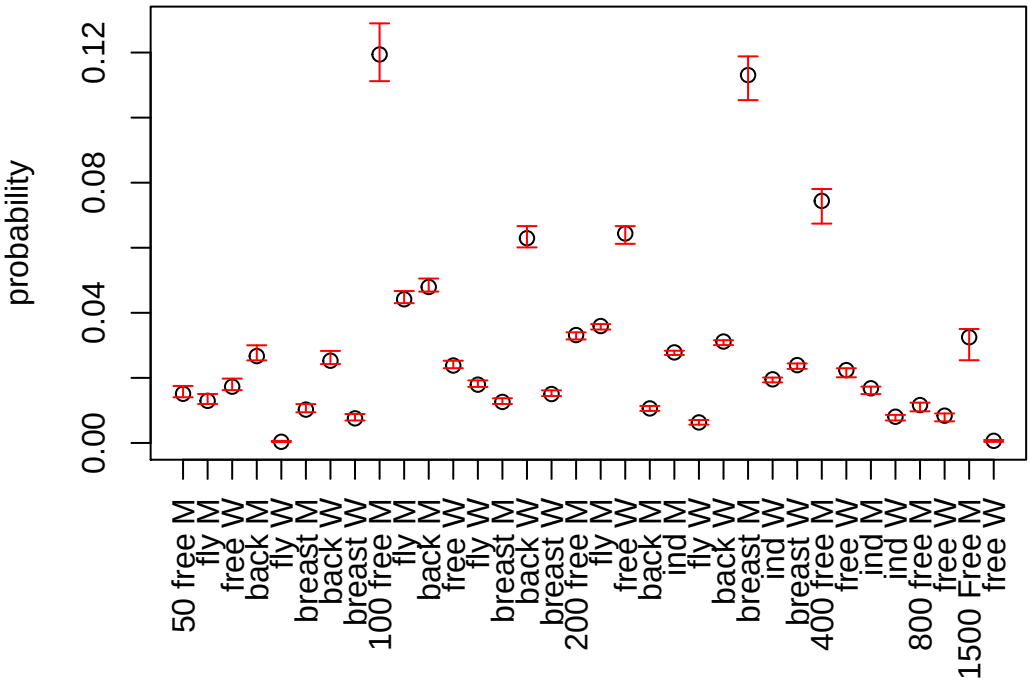


Figure 3.4.4: Estimated probabilities that the next world record is set in a particular event, with 95% CI's from bootstrapped data sets.

however, it is desirable to find a fair comparison between swim-times of those swimmers wearing a swim-suit and those not, and even construct a framework such that swim-times can be fairly compared with other future technological advancements.

Since the rank of a swim-time is based on the rate R at which better observations occur, it is possible to adjust the swim-time for the use of a swim-suit. Let $x > u$ be a swim-time occurring at time q during the swim-suit period i.e, $q \in S_{t_1} \cup S_{t_2}$, and z is a swim-time occurring at the same time but as if it were not swam using a swim-suit. Then the swim-time correction from a recorded swim-time x to an equivalent swim-time without the swim-suit z is made by selecting z such that the rate of exceeding x, R, and the corrected rate of exceeding z without a swim-suit, R_C, are equal. That is, find z as the solution to

$$R\{X_q^{(e)} > x\} = R_C\{X_q^{(e)} > z\}, \tag{3.4.4}$$

where R is defined in equation (3.4.1) and R_C is defined by

$$R_C\{X_q^{(e)} > z\} = \Pr\{X_q^{(e)} > z | X_q^{(e)} > u_e\} \Lambda_{C,q}^{(e)}\left(\mathcal{A}_{1,u}\right),$$

where

$$\Lambda_{C,q}^{(e)}\left(\mathcal{A}_{1,u}\right) = \left[1 + \xi\left(\frac{u_e - \mu_C^{(e)}(q)}{\sigma_C^{(e)}(q)}\right)\right]_+^{\frac{1}{\xi}},$$

$\sigma_C^{(e)}(q) = \sigma_0^{(e)} + \xi\beta q$, and $\mu_C^{(e)}(q) = \mu_0^{(e)} + \beta q$. Thus, the adjusted swim-time z is found via the solution to equation (3.4.4), given as

$$z = u_e + \frac{\tilde{\sigma}_u^{(e)}}{\xi}\left\{\frac{\Lambda_q^{(e)}\left(\mathcal{A}_{1,u}\right)\bar{H}_u^{(e)}(x)}{\Lambda_{C,q}^{(e)}\left(\mathcal{A}_{1,u}\right)} - 1\right\}.$$

As an example, Cesar Cielo's 6th rank swim-time of 20.91s in the 50m freestyle in 2009 gets adjusted to 21.18 once the swim-suit effect is removed. The reverse can be found, that is the time a swimmer would have got, had they been wearing a swim-suit, e.g., Adam Peaty's current 100m breaststroke world record time of 56.88s gets adjusted to

56.25s with a swim-suit from 2008, and adjusted to 55.96 with a swim-suit from 2009, indicating that a "Project 55" could be achieved with just the addition of a swim-suit. By adjusting for technology in this way, it is possible to determine which current world records would still stand, had swim-suits never played a part. Table 3.4.1 shows those current world records set using swim-suits, and their estimated adjustments. Moreover, Table 3.4.1 shows what the world record would be, and who the world record holder would be, once the effect of swim-suits is removed. Out of the 10 world records which have been set by swimmers wearing swim-suits, only 2 would still stand today, Zige Liu's 200m fly world record, and Zhang Lin's 800m free world record. It is worth noting that the assumption that the most up-to-date technology available is always being used, is occasionally violated, for example, Phelps' 100m and 200m fly world records in 2009 were swam with the LZR Speedo suits from 2008. There can be additional complications when taking technology into account, such as Phelps' 400m individual medley world record from 2008, in which only the leg suit was worn. These issues could be addressed with the addition of explicit data about which technology was being used in a given swim.

3.5 Discussion

Throughout this article, the swim-times are negated before being analysed so that we can use existing methodology for larger values. Alternatively, by analysing swim-*speed*, Gomes and Henriques-Rodrigues (2019) apply peaks-above-threshold methodology directly, since a smaller swim-time equates to a larger swim-speed. This raises the question of which transformation is best, and what classes of transformation give similar results. From limit (3.2.1), it can be seen that any linear transformations will be absorbed into the norming constants a_n and b_n so that inference is invariant for positive linear transformations. Conversely, Wadsworth et al. (2010) show that non-linear

Event	WR swim	WR	AWR	NSWR	NSWR swim
50 free M	Cielo (2009)	20.91	21.18	**21.11**	**Proud (2018)**
100 free M	Cielo (2009)	46.91	47.99	**47.04**	**McEvoy (2016)**
100 fly M	Phelps (2009)	49.82	50.83	**49.86**	**Dressel (2017)**
200 fly M	Phelps (2009)	111.51	113.33	**112.71**	**Milak (2018)**
200 back M	Peirsol (2009)	111.92	113.47	**112.96**	**Lochte (2011)**
200 free F	Pellegrini (2009)	112.98	114.99	**113.61**	**Schmitt (2012)**
200 fly F	**Zige (2009)**	121.81	**123.38**	124.06	Jiao (2012)
400 free M	Biedermann (2009)	220.07	223.13	**220.08**	**Thorpe (2002)**
400 ind M	Phelps (2008)	243.84	245.72	**245.18**	**Lochte (2012)**
800 free M	**Lin (2009)**	452.12	**455.31**	458.57	Sun (2011)

Table 3.4.1: World records (WR) set with swim-suits, the adjusted times (AWR), and the best corresponding non-swim-suit times (NSWR). "Would-be" world records and world record holders, after adjusting for swim-suits, are marked in bold.

transformations lead to different results. Wadsworth et al. (2010) consider the class of Box-Cox transformations as part of the extreme value analysis with negating of the data and inversion to swim-speed as special cases. Thus a possible route for future research is to find the best Box-Cox parameter and to see if this changes in a systematic way over distance, gender and stroke.

Only the best time is recorded from each swimmer in a given event which, for cases where swimmers in the data set are still active, could lead to poor predictive performance. For example, let $X_t^{(w,e)}$ be the random variable denoting a swim-time by the current world record holder in event e at time t, and $X_t^{(i,e)}$ be the random variable denoting a swim-time by another swimmer i in event e at time t, then the probability of a world record-holder setting a new personal best, and therefore new world record, is likely to be larger than the probability of any new swimmer setting a world record, such that $\Pr\{X_\tau^{(w,e)} > r_e\} > \Pr\{X_\tau^{(i,e)} > r_e\}$ for $\tau > t$. This could be accounted for by allowing more than one swim-time to be recorded per swimmer, however this gives rise to dependency between swim-times in the same event, and would need to be adjusted for.

Independence is assumed between swim-times for different strokes, genders and distances. This simplifying assumption may not be true when the same swimmer competes across many distances or strokes, meaning that the uncertainty of our estimates would be underestimated. Of the swim-times that exceed the thresholds u_e, i.e., for $e \in E$, the proportion of unique swimmers to total data points is around half, and so the effective sample size of independent swimmers in the data set will be less than the number of total data. In the case that there is perfect correlation between the same swimmer in separate events, then the effective sample size will be equal to the number of unique swimmers, approximately half the total data values, which means the variance could be underestimated by at most a factor of 2. This could be corrected for by estimating some inflation parameter $1 \leq \phi \leq 2$, such that the actual variance is equal to $\phi \text{var}(\hat{\theta})$, where

$\text{var}(\hat{\theta})$ is the variance obtained by assuming complete independence between observations, see Kent (1982). It may be necessary to use multivariate techniques in order to capture some of the correlation between data points resulting from the same swimmer in different competitions (Adam and Tawn, 2012).

There are extra sources of uncertainty not accounted for. Quantifying the uncertainty due to the choice of threshold is not considered, since a single threshold selection approach is used, as is common in the extreme value theory literature (Scarrott and MacDonald, 2012). However this uncertainty could be quantified by using the cross-validatory technique of Northrop et al. (2017). Also, since the analysis is performed in a frequentist framework, only parameter uncertainty is considered, however when predicting future events such as the time until a new world record is set in a particular event, it is also valuable to consider the predictive uncertainty. This could be accounted for by moving to a Bayesian framework, and carrying out parameter estimates via Markov chain Monte Carlo with a prior on the spline roughness penalty.

The constant evolution of the para-swimming classification system is testament to the challenge of creating fair competition in disability swimming. The number of classifications itself is open to debate, with too many classifications resulting in too few swimmers in each classification and therefore a drop in competitiveness, and too few classifications resulting in bias such that there is unfair differences between swimmer's physical limitations within the same class. Of course, this problem stems from the discrete nature of the classification system, but a model of the type presented in this article would allow for a continuous "classification variable" which pools across disability, to allow fair competition over all disability types and comparison between disabilities. In a similar way, this model could allow for more fair comparison with transgender swimmers. Regulations around transgender athletes in sports is a controversial topic, with the regulations being changed again for the upcoming 2020 Olympic Games. This controversy largely arises due to determining whether a transgender athlete should compete

in the men's or women's event, and is determined on a case by case basis. However, our type of covariate model can allow for a more fluid description of gender, since the adjustment or categorisation is determined simply by the threshold time u_e which can easily be modelled as continuous across events or gender status. In addition, cases of unusual testosterone levels can be dealt with in the same way. In junior swimming, because of the discretisation of age groups, some swimmers can be almost a whole year younger than others in the same competition, which creates an unfair disadvantage. The same idea of a continuous scale for age groups would allow for fair comparison of 'age-adjusted' swim-times. Ultimately, it is possible to have a global model which fairly compares swimmers of all genders and disabilities, and even junior swimmers, across different events.

Chapter 4

A Framework for Statistical Modelling of the Extremes of Longitudinal Data, Applied to Elite Swimming

4.1 Introduction

Traditional statistical techniques are designed to describe the behaviour of the "typical" data and many analyses involve the identification and removal of observations from the tails of the data to improve robustness. But what if the data of most interest *are* those observations in the tails? When considering natural disasters such as flooding, stresses or corrosion on a structure, financial crises, or sporting records, it is precisely these *extreme* values that are most pertinent. *Extreme value theory* (EVT) is a branch of statistics specifically designed to model such extreme or rare events, with the methods having a strong probabilistic framework based on asymptotic justifications. This paper presents novel methodology for the analysis of longitudinal data where the extreme

values are of primary interest.

Early EVT methods describe the extremal behaviour of independent univariate random variables, possibly in the presence of covariates, with the book of Coles (2001) an accessible introduction. More recently, the extremal properties of ever more rich data structures have been studied, with theory and associated methodology developed. For univariate stationary processes the following features have been considered: long- and short-range dependence (Ledford and Tawn, 2003), Markov structure (Winter and Tawn, 2017), and hierarchical clustered data (Smith and Goodman, 2000; Bottolo et al., 2003; Dupuis et al., 2023). For multivariate extreme value problems, structure has been identified and exploited through the use of graphical structures (Engelke and Hitz, 2020), sparsity (Engelke and Ivanovs, 2021), and models for conditional structure through asymptotic independence (Heffernan and Tawn, 2004). Most recently, various approaches have been developed for spatial, and spatial temporal extreme events, such as r-Pareto processes (de Fondeville and Davison, 2022), spatial conditional asymptotically independent processes (Wadsworth and Tawn, 2022), the associated processes in space and time (Simpson and Wadsworth, 2021), and for spatial mixture processes (Richards et al., 2023). However, no current adaptations allow for EVT to model longitudinal data (Diggle et al., 2002), sometimes called panel data, which has been so widely studied for the body of the data.

Longitudinal data comprises a number of *subjects*, with each subject recording a time series of responses. Specifically, there are a set of subjects, $\mathcal{I}$, with a subject i having a set of measurements $\mathcal{J}_i$, for all $i \in \mathcal{I}$. The measurement $X_{i,j}$ belonging to subject i, occurs at a known time $t_{i,j} \in \mathbb{R}$, for all $j \in \mathcal{J}_i$, $i \in \mathcal{I}$. The typical assumptions made about the collection $\{X_{i,j} : j \in \mathcal{J}_i, \text{ for } i \in \mathcal{I}\}$ are that the $X_{i,j}$ are independent over different $i \in \mathcal{I}$, irrespective of j, but they are potentially dependent across $j \in \mathcal{J}_i$ for any given $i \in \mathcal{I}$. An important special case is when the $X_{i,j}$ are independent over different $j \in \mathcal{J}_i$ irrespective of i, a situation we refer to as subject-

conditional independence. Usually $\max_{i\in\mathcal{I}} \mid \mathcal{J}_i \mid \ll \mid \mathcal{I} \mid$, so there are a large number of subjects relative to the number of measurements per subject, and the way that the distribution of $X_{i,j}$ varies with $t_{i,j}$, i.e., due to a varying mean value, is similar across subjects.

Even though the panel data analysis of Dupuis et al. (2023) suggests a considerable overlap with this set up, the focus of their modelling and inference is very different to ours, with their priority being marginal inference for different subjects whereas we infer the within-subject measurement dependence and a population-based marginal model. They consider a simplified setting where all $\mathcal{J}_i$ are equal to some common $\mathcal{J}$, i.e., all subjects have the same number of measurements and measurement times are identical across all $i \in \mathcal{I}$. The data are split over B blocks, $\{\mathcal{J}^b : b = 1, \ldots, B\}$ forming a partition over $\mathcal{J}$. Then, the joint behaviour of the subject block maxima $\{\max_{j\in\mathcal{J}^b} X_{i,j} : \text{ for } i \in \mathcal{I}, b = 1, \ldots, B\}$ are studied assuming these are independent over blocks. The temporal dependence structure of the within-subject behaviour, which is the focus of our analysis, is not considered.

For analysing the extremes of longitudinal data, the *sample* $\mathcal{I}$ comprises those subjects with at least one extreme observation within the observed time-frame. An important distinction is then made between this sample of subjects $\mathcal{I}$, and the *population* of subjects, which includes those subjects with extreme measurements that are exclusively outside the observed time-frame. For the observed time-frame, subjects in the *population* may have either no measurements at all, or have measurements that are exclusively non-extreme. One reason for this distinction is when making predictions about future extreme events far ahead of the observed time-frame. In applications where individual subjects exhibit non-stationarity, future extreme events change from being measurements on the observed subjects, to measurements on subjects that are not yet present in the observed time-frame, but form part of the population of subjects. We will make this distinction clear in Sections 4.3 and 4.5.

The structure of longitudinal data combines aspects of time series and multivariate data. The critical differences between longitudinal data and standard multivariate time series are that (i) the length of the time series per subject is small, (ii) the number of time series, i.e., subjects, is large and (iii) there is no restriction that the measurement times across subjects are synchronised or are to be regularly spaced. Furthermore, the longitudinal data format differs from multiple measurement types taken on the same subject at the same time points as each other (e.g., in a health data set measurements of blood sugar, blood glucose, and resting heart rate), because the measurements are assumed to be independent across the observed time series.

Longitudinal data analyses arise most commonly in designed trials (e.g., in clinical or corrosion contexts) whereby multiple subjects (e.g., patients or material coupon samples) have a single quantity (e.g., blood pressure or corrosion, respectively) measured over time. There have been very limited examples of extreme value modelling of clinical and corrosion data, with none capturing the full specification of the longitudinal data structure. Clinical extreme value examples include Southworth and Heffernan (2012) and Papastathopoulos and Tawn (2015) but neither of these look at repeated measurements on the same subject. For corrosion there have been covariate models which allow for the mean pit depth to increase with time, but they only have one observation per coupon (Laycock and Scarf, 1993). Fougères et al. (2006) do consider multiple observations per coupon but assume that observations from the same coupon are IID.

Perhaps the closest approach to our modelling of longitudinal extremes is Fougères et al. (2009), who use a latent/random-effect positive stable mixture model to produce a multivariate extreme value distribution to model dependence in repeated observations of pit depth across different coupons. Since they assume all $X_{i,j}$ are conditionally independent and identically distributed given a random effect R_i for each subject, the dependence across time per subject is exchangeable, i.e., all pairs $(X_{i,j}, X_{i,k})$, for $j \neq k \in \mathcal{J}_i$, have the same dependence structure. This limited form of temporal depen-

dence per subject is likely be too simplistic for generic longitudinal data, where pairs $(X_{i,j}, X_{i,k})$ often have dependence weakening as the time between them, $|t_{i,j} - t_{i,k}|$, increases. The use of the positive stable distribution to capture the variation between subjects - through both the mean and variance of $X_{i,j}|R_i$, for all $j \in \mathcal{J}_i$ - leads to the largest R_i values corresponding to the subjects with $X_{i,j}$ values that are much larger than for other subjects. This is highly restrictive both for the limitation on how the population is distributed, but also as it enforces a strong form of extremal dependence over time, termed *asymptotic dependence*. Asymptotic dependence, defined in Section 4.2.2, constrains that if a subject gives the largest value in the population at some time point, then they are likely to do this at all time points. Our paper aims to be the first foray into developing broadly usable EVT methods for longitudinal data, with the flexibly to model both asymptotic dependent and asymptotic independent temporal extremal dependence structures and to capture trends in the means of subjects' responses over time.

One area where extreme value analysis of longitudinal data is particularly important is sports data, e.g., athletics and swimming. Here, athletes/swimmers (subjects) all strive for the best times for completing their event, with their personal career progression having stages of improvement and decline as they age, and with them competing at irregular times which can be different from each other. Furthermore, the overall performances by the elite athletes/swimmers are improving over time, i.e., expected annual world best times are reducing, so that records are being broken more often than would be expected if the data were from with independent and identically distributed variables. Surprisingly, we have found no examples of such statistical analyses in this area, despite its clear relevance, e.g., for studying the progression of records, and predictions for who will set the best time next year. Therefore our longitudinal EVT methods will be demonstrated by analysing a dataset of elite swimmers.

The application of EVT methods is not new for sports' data. EVT is used by

Robinson and Tawn (1995) to model athletics data and by Strand and Boes (1998) to estimate the peak age of competitive 10K road race runners. Stephenson and Tawn (2013) fit a generalised extreme value (GEV) distribution to yearly maxima of athletics times across different distances and eras. There, the GEV location and scale parameters are allowed to vary as a parametric function of the distance, and an exponential trend allows for a smooth adjustment for era. Spearing et al. (2021) use EVT to model the evolution of elite swimming over time, including the effect of different swim-suit technologies, and combine data across different swimming strokes, gender categories and distances through the use of a data-based covariate.

None of these models attempt to model dependence structure - either they assume that performances from the same subject are independent of each other, or only incorporate each subject's best performance into the data set. The consequence of both of these data handling approaches for repeated observations per subject is incomplete inference: the former uses a smaller data set than is available, leading to inefficient inference; and the latter produces an underestimation of standard errors and confidence interval widths when the independence assumptions are invalidated. However, the true limitation of these simplifications runs deeper. The lack of any longitudinal structure in these models means that no statistical inference can be conducted on any facet involving individual competitors, e.g., Strand and Boes (1998) infers the peak age of the *typical* runner, but cannot draw conclusions about any *individual* runner.

We illustrate our novel EVT methodology for longitudinal data in the context of elite swimming, specifically in the mens' 100m breaststroke (long course) event. A swimmer is defined to be elite if they have ever produced a swim-time less that a certain threshold u. The selection of this threshold u is a source of debate in general (Scarrott and MacDonald, 2012), but is here taken as the 200th fastest personal-best swim-time in this mens' 100m breaststroke event, which is $u = 61.125$ seconds - see Section 4.5 for a discussion on this point. In our approach (i) all the available recorded

swims from each elite swimmer are modelled, irrespective of whether they are below or above u, (ii) the swimmer who produced each swim-time is accounted for, as is their age at which it was achieved, (iii) the dependence between swim-times from the same swimmer is captured, with this dependence allowed to weaken as the inter-swim-time increases.

Figure 4.1.1 depicts the competition-best swim-times from a subset of five of the 200 elite swimmers who epitomise the range of typical career trajectories and the different rates at which swimmers compete in competitions. Of these swimmers, Adam Peaty holds the current world record and so, by definition, the fastest personal-best (PB). Ilya Shymanovich has the second fastest PB in the data, Sakci Hueseyin 8th, Sakimoto Hiromasa 101st, and Takahashi the 196th fastest PB, which is only just faster than the threshold. There is a notable difference between the performance of the top two swimmers. Peaty is consistently fast, producing the seven fastest times of the competition-best dataset, and with all his performances faster than u. Conversely, Shymanovich is in a clear progression stage of his career, moving from being consistently slower than u, to consistently faster. Furtherly contrasting, Takahashi only once swims faster than u.

Even across these five swimmers we can see vastly differing *career trajectories* and swimmer's strategies for which, and how many, competitions in which they participate. Despite this, it is visible even in these data that for each swimmer, swim performance achieved more closely together in time are generally more alike than those a further apart. This naturally reflects training cycles and form which induce local time-dependent variations in performance.

Now consider the marginal distribution of the extreme swimming values, i.e., the values below $u = 61.125$. To motivate a possible model for these values we draw on EVT. Specifically, for a continuous random variable X, EVT gives that the generalised Pareto distribution (GPD) is the only non-degenerate limit distribution for scale-normalised

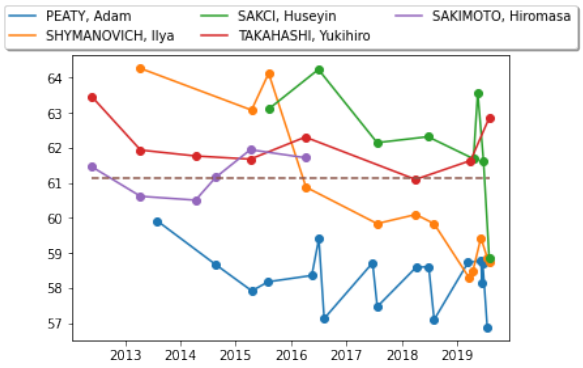


Figure 4.1.1: Data for swim-times (in seconds) plotted against the calendar time when it was achieved for the mens' 100m breaststroke (long course) event. All competition best performances are shown for five selected swimmers over time. The dashed line indicates the extremal threshold u, dictated by the PB time of the 200th fastest swimmer.

difference of X from the threshold v as v tends to the lower endpoint of the distribution of X (Pickands, 1975). In practice it is common to assume that the GPD is a sufficiently good approximation to the data relative to the threshold (Davison and Smith, 1990). In our case we take $v = u$, as Spearing et al. (2021) show that the personal-best swim-times better than the 200th top personal-best time, for each swimming event, are well modelled by a GPD. This finding encourages us to consider the GPD as a marginal model for all swimmers' available performances better than this same extreme threshold. This choice is further supported by asymptotic theory for univariate stationary processes that exhibit weak long-range dependence conditions, where the distribution of cluster maxima, and arbitrary values of the process excesses of a threshold, are identical in the limit as the threshold tends to the upper endpoint of the stationary distribution (Leadbetter, 1991)

Thus, all observations faster than u can be assumed to have a known class of marginal distribution, the GPD; however, we have no justifiable parametric model for observations slower than u. In modelling and predicting the extremes of longitudinal data, it is desirable that the extreme data be the most influential. Therefore, obser-

vations slower than u are treated as censored at the level of the threshold. That is, if $X_{i,j} > u$ we only use this knowledge, and the time $t_{i,j}$ at which it was performed, but not the precise value of $X_{i,j}$. As a consequence, all but one of Takahashi's observations are censored, whereas all of Peaty's observations can be modelled with the GPD. Critically, the values slower than the threshold are not lost as, firstly, they provide information about the rate of performing better than the threshold. Secondly, they inform us of the dependence structure for individual swimmers as, e.g., three successive swims slower than u followed by three faster than u may suggest stronger dependence than the same swimmer having 6 alternate swims faster and slower than u.

Conventional presentation of EVT theory pertains to the largest values - or equivalently the upper tail - in a sample, yet the best swim-times are the smallest - or in the lower tail. By applying our methodology to *negative* swim-times, standard EVT results can be utilised. So, throughout we present theory and methods for the upper extremes of longitudinal data. Section 4.2 presents the extensions of univariate EVT to cover the time series aspect of each subject's data. Section 4.3 contains the main contribution of the paper - a novel approach to the modelling of the extremes of longitudinal data. Section 4.4 presents the general Bayesian inference framework and Section 4.5 details how this modelling and inference framework can be applied to the elite swimming data, and provides examples of particular inferences and predictions that are available using our methodology. A discussion and future work is in Section 4.6.

4.2 Motivating Theory

4.2.1 Univariate extremes

In its simplest form, univariate extreme value theory (EVT) applies to independent and identically distributed (IID) random samples $Y_1, \ldots, Y_n$, where each variable has continuous distribution function F. The block maxima and peaks over threshold meth-

ods are the two core approaches in univariate EVT (Coles, 2001). We are interested in formulating a theoretically justified marginal extreme value model for temporally dependent variables and describing the dependence structure induced by within- and across-subject observations for longitudinal data. To keep the formulation sufficiently simple we also consider the extremes of a stationary process, observed at regular time intervals, $X_1, \ldots, X_n$ which also has the marginal distribution function F but satisfies conditions such that its long-range dependence is restricted to behave as effectively independent, see Leadbetter et al. (2012) for their precise form and discussion of the limit results (4.2.2) and (4.2.3). Under such conditions, the following results hold. If $M_{Y,n} := \max\{Y_1, \ldots, Y_n\}$ and there exist norming sequences $a_n > 0$ and b_n, such that

$$\Pr\left\{\frac{M_{Y,n} - b_n}{a_n} \leq x\right\} = F^n(a_n x + b_n) \to G(x), \text{ as } n \to \infty, \tag{4.2.1}$$

where that the limiting distribution $G(x)$ is non-degenerate, then $G(x)$ must be a generalised extreme value (GEV) distribution, which has the form

$$G(x) = \exp\left(-[1 + \xi(x - \mu)/\sigma]_+^{-1/\xi}\right), \tag{4.2.2}$$

where $\mu,\ \xi \in \mathbb{R},\ \sigma \in \mathbb{R}^+$, are the location, shape and scale parameters respectively and with the notation $y_+ := \max(y, 0)$. Then for $M_{X,n} := \max\{X_1, \ldots, X_n\}$, if $(M_{X,n} - b_n)/a_n$ has a non-degenerate limit distribution, as $n \to \infty$, it follows that

$$\Pr\left\{\frac{M_{X,n} - b_n}{a_n} \leq x\right\} \to [G(x)]^\theta, \text{ as } n \to \infty, \tag{4.2.3}$$

where $0 < \theta \leq 1$ is the extremal index; a measure of extremal temporal dependence, with $1/\theta$ being the limiting mean number of exceedances of a high threshold by the $\{Y_t\}$ process per cluster of exceedances. Ferro and Segers (2003) define a cluster and discuss inference for θ.

We are primarily interested in having an asymptotically motivated model for the upper tail behaviour of $\{X_t\}$ and $\{Y_t\}$. These models are derived directly from the limiting distribution of block maxima identified above. First, denote $D_G := \{x \in \mathbb{R} : 0 < G(x) < 1\}$. Then applying a Taylor series approximation to limit (4.2.1) gives,

$$n[1 - F(a_n x + b_n)] \to -\log G(x) = [1 + \xi(x - \mu)/\sigma]_+^{-1/\xi}$$

as $n \to \infty$, for all $x > u$ with both x and u in D_G. It follows that, for all $Y \sim F$,

$$\Pr\{Y > a_n x + b_n | Y > a_n u + b_n\} \to \log G(x)/\log G(u) =: \bar{H}_u(x), \tag{4.2.4}$$

with $\bar{H}_u(x) := 1 - H_u(x)$, and where the distribution function H_u is written as

$$H_u(x) = 1 - \left[1 + \xi\left(\frac{x - u}{\sigma_u}\right)\right]_+^{-\frac{1}{\xi}}. \tag{4.2.5}$$

where $\sigma_u = \sigma + \xi(u - \mu)$. The distribution function H_u is termed the generalised Pareto distribution (GPD), denoted GPD(σ_u, ξ), with threshold u, shape parameter $\xi \in \mathbb{R}$ and scale parameter $\sigma_u \in \mathbb{R}_+$. For $\xi < 0$, there exists a finite value $x^H = u - \sigma_u/\xi : \ H_u(x) = 1, \ \forall x > x^H$, whereas for $\xi \geq 0, \ H_u(x) < 1, \ \forall x < \infty$. This GPD result is powerful as it holds as the limit distribution for a very broad class of continuous distributions F (Leadbetter et al., 2012).

Now consider the tail behaviour of the stationary process $\{X_t\}$. The same GPD(σ_u, ξ) limit distribution holds for $\Pr\{X > a_n x + b_n | X > a_n u + b_n\}$ as $n \to \infty$. However, there is an additional result due to Leadbetter (1991) which gives that for an arbitrary cluster maxima X_C of $\{X_t\}$, then $\Pr\{X_C > a_n x + b_n | X_C > a_n u + b_n\}$ as $n \to \infty$, is also GPD(σ_u, ξ). This has motivated the use of the generalized Pareto distribution as a statistical model for cluster maxima (Davison and Smith, 1990), but critically for our purposes showed the strong connection between the distribution of the maxima

of identically distributed variables in a cluster and an arbitrarily selected threshold exceedance. This is a feature we exploit in Section 4.3.1.

In practice the limit distribution (4.2.4) is assumed to hold exactly for some finite n, or equivalently for some fixed threshold $a_n u + b_n$, corresponding to a high quantile of Y or X. A consequence of this assumption is that the limit distribution H_u gives an asymptotic model for the distribution of exceedances above a threshold u, no matter the form of marginal distribution F and implies that whatever F is within this class, values above a suitably high threshold u must follow a single class of distributions, determined by only two parameters.

To complete the description of the tail of the marginal distribution we define the marginal probability of an threshold exceedance, $\lambda_u := \Pr(X > u)$, and select the threshold u above which the GPD tail is assumed to hold. The choice of u is the subject of much historical focus, primarily relating to bias-variance trade-off (Scarrott and MacDonald, 2012; Danielsson et al., 2001; Northrop et al., 2017; Varty et al., 2021). The methods typically are based on the threshold-stability of the GPD, namely that if the GPD approximation (4.2.4) is valid for exceedances above some threshold $u \in D_G$, then it holds for excesses over all higher thresholds v, where $v \in D_G$ and $v > u$. So if u is the lowest threshold for which approximation (4.2.4) is exact, then any lower threshold will have excesses that do not follow the GPD, whereas thresholds larger than u ignore relevant observations and lead to inefficient inference.

4.2.2 Extremal dependence: measures and modelling strategies

Since there may be dependence between the measurement values of the time series for a given subject, dependence needs to be accounted for in the extreme values. Here we draw on knowledge of extremal dependence measures and the associated modelling strategies generally before considering the specific features that are unique to longitu-

dinal data.

In modelling dependence between the extremes of two variables the most published approaches first involve deciding on the *form* of extremal dependence, and then looking for an appropriate model formulation subject to that form (Coles et al., 1999). For bivariate extremes, with continuous random variables (X_1, X_2) with marginal distributions F_1 and F_2, respectively, the two forms of extremal dependence in the upper tail are determined by *the coefficient of asymptotic dependence* $\chi := \lim_{q \uparrow 1} \chi(q)$ where, for $0 < q < 1$,

$$\chi(q) := \Pr\{F_1(X_1) > q \mid F_2(X_2) > q\} = \Pr\{F_1(X_1) > q, F_2(X_2) > q\}/(1-q), \quad (4.2.6)$$

with *asymptotic dependence* given by $0 < \chi \leq 1$ and *asymptotic independence* by $\chi = 0$. In essence, asymptotic dependence allows the very largest values of X_1 and X_2 to occur together, unlike for asymptotic independence. This interpretation is made precise by looking at the limiting distribution of normalised componentwise maxima of independent and identically distributed vectors $\{(X_{1i}, X_{2i}) : i = 1, \ldots, n\}$, such that the marginal distributions are non-degenerate. Then, the two variables are termed asymptotic dependent, or asymptotic independent, if that limiting distribution exhibits dependence, or independence, respectively.

Although there exists only two extremal dependence scenarios in the bivariate case, in multivariate extremes higher order dependence structures can lead to asymptotic dependence for different subsets of the variables, so the number of different extremal model structures grows exponentially with dimension (Simpson et al., 2020). Even in the bivariate case, variables may exhibit extremal dependence in without asymptotic dependence, with this dependence measured by the *coefficient of asymptotic indepen-*

dence, $\bar{\chi} := \lim_{q \uparrow 1} \bar{\chi}(q) \in (-1, 1]$, where

$$\bar{\chi}(q) := \frac{2 \log \Pr\{F_2(X_2) > q\}}{\log \Pr\{F_1(X_1) > q, F_2(X_2) > q\}} - 1, \tag{4.2.7}$$

for $0 < q < 1$, with independent variables giving $\bar{\chi} = 0$, and $0 < \bar{\chi} < 1$ ($\bar{\chi} < 0$) corresponding to a positive (negative) extremal dependence form of asymptotic independence respectively, and $\bar{\chi} = 1$ arises when the variables are asymptotically dependent. Both χ or $\bar{\chi}$ are invariant to the marginal distributions, so in terms of models for the joint distribution it is helpful to consider different copulas. Heffernan (2000) presents $(\chi, \bar{\chi})$ for a range of copulae.

We now focus the discussion on copulae that are relevant to longitudinal data analysis. Fougères et al. (2009) use the copula of the multivariate extreme value distribution with logistic(α) dependence structure, which in the bivariate case has $(\chi, \bar{\chi}) = (2-2^{\alpha}, 1)$ for $0 \leq \alpha < 1$ and $(\chi, \bar{\chi}) = (0, 0)$ when $\alpha = 1$. In terms of extremal dependence, this copula model is restrictive as it cannot capture any positive dependence within the asymptotic independence case. It also has limitations for modelling longitudinal data: the copula is exchangeable, which is unrealistic for most time series data; and the conditional distributions for this copula are non-trivial to simulate from. The latter property complicates inference for future extreme events. Due to these features we instead consider the d-dimensional Gaussian copula

$$C(\boldsymbol{x}) = \int_{-\infty}^{\Phi^{-1}(x_1)} \cdots \int_{-\infty}^{\Phi^{-1}(x_d)} \phi_d(\boldsymbol{s}; \Sigma) \, \mathrm{d}\boldsymbol{s},$$

for $\boldsymbol{x} = (x_1, \ldots, x_d) \in [0, 1]^d$ and $\boldsymbol{s} \in \mathbb{R}^d$, with $\phi_d(\boldsymbol{s}; \Sigma)$ denoting the d-dimensional Gaussian density, with standardized margins and dependence structure determined by the $d \times d$ correlation matrix Σ. In the bivariate case this copula has the properties $(\chi, \bar{\chi}) = (0, \rho)$ for correlation parameter $-1 < \rho < 1$, and $(\chi, \bar{\chi}) = (1, 1)$ for $\rho = 1$ (Coles et al., 1999). Furthermore, as the multivariate copula is determined by its bivariate

marginals, which are all asymptotically independent (except for the pathological case when $\rho = 1$), it is not necessary to consider asymptotic dependence at any higher order. The Gaussian copula allows for considerable flexibility and parsimony in terms of the dependence structures that can be modelled and it benefits from having closed form conditional distributions for describing and simulating the time series features of longitudinal data. Given these properties, we will use the Gaussian copula structure as a building block for modelling dependence of measurements of within-subject behaviour in Section 4.3.2 but first, in Section 4.2.3, we show that we can approximate any level of asymptotic dependence at finite levels using the asymptotically independent Gaussian copulae for subjects within a longitudinal data context. Consequently our model has all the good features of the copula used by Fougères et al. (2009) but none of the disadvantages.

4.2.3 Sources of extremal dependence for longitudinal data

Measures of longitudinal data dependence

To illustrate the possible forms extremal dependence structures that can occur in longitudinal data, consider a special case of the set up of Section 4.1 with n subjects indexed by $\mathcal{I}$, with a continuous time process for each subject $i \in \mathcal{I}$ is $\{X_i(t)\}$ for all t which are observed at a set of identical and equally spaced time points across subjects. We denote $X_{i,j} = X_i(t_{i,j}) = X_i(t_j)$ where t_j is the jth time point. We also assume that the time series for each subject are stationary and with marginal distribution $F_i(\cdot) = F(\cdot; \alpha_i)$ where F is a common continuous distribution function family with parameter $\alpha_i \in \mathbb{R}$ which can vary over $i \in \mathcal{I}$. We term α_i the attribute of subject i, with the property that $F(x; \alpha_i) > F(x; \alpha_j)$ for all $x \in \mathbb{R}$ for all $\alpha_i > \alpha_j$. Increasing the attribute of a subject makes the quantiles of its measurement distribution larger.

There are a range of different extremal features of the longitudinal data that can be studied. Given the potential heterogeneity between subjects, the most basic application

of the coefficient of asymptotic dependence for within-subject dependence at time-lag τ for each subject $i \in \mathcal{I}$ is:

$$\chi_i(\tau) := \lim_{q \uparrow 1} \Pr(F(X_i(\tau); \alpha_i) > q \mid F(X_i(0); \alpha_i) > q),$$

or the equivalent asymptotic independence measure $\bar{\chi}_i(\tau)$. However, this measure does not provide a global description of the dependence across all subjects in $\mathcal{I}$. To study the extremal behaviour over subjects at each time point, consider $M_t := \max_{i \in \mathcal{I}} X_i(t)$ for different t. This leads to the lag τ dependence measure

$$\chi_\tau^{(M)} := \lim_{q \uparrow 1} \Pr(F^{(M)}(M_\tau) > q \mid F^{(M)}(M_0) > q)$$

where $F^{(M)}(x) := \prod_{i \in \mathcal{I}} F_i(x) = \prod_{i \in \mathcal{I}} F(x; \alpha_i)$, and also its equivalent asymptotic independence measure $\bar{\chi}_\tau^{(M)}$. An alternative is to consider dependence between values in the marginal tail for each time point. This corresponds to picking a random subject from the population $\mathcal{I}$ at each time point, giving the lag-τ dependence measure

$$\chi_\tau^{(R)} := \lim_{q \uparrow 1} \Pr(F^{(R)}(X_\tau^{(R)}) > q \mid F^{(R)}(X_0^{(R)}) > q)$$

where $X_\tau^{(R)}$ is a random selection from $\{X_i(\tau) : i \in \mathcal{I}\}$, so has marginal distribution function $F^{(R)}(x) := \sum_{i=1}^n F(x; \alpha_i)/n$. Again the equivalent asymptotic independence measure is $\bar{\chi}_\tau^{(R)}$. When all subjects are identically distributed and have the same temporal dependence structure, then each of these extreme dependence measures at lag-τ are identical to the measure of asymptotic dependence (asymptotic independence) (4.2.6) and (4.2.7) respectively for the associated identically distributed variables. Thus each measure has equal validity when assessing dependence for longitudinal data.

Study of how attributes determine the form of extremal dependence

To help understand the information given by the extremal dependence measures for longitudinal data of Section 4.2.3 we explore both $\chi_{i\tau}$ and $\chi_{\tau}^{(M)}$ for a particularly simple version of longitudinal data. Specifically, the variables consist of n independent subjects, all of which have measurements at two time points - which are the same across subjects - and each measurement is independent per subject except for subject n. Additionally all subjects assume the same attributes as each other except for one. So, in the notation of Section 4.1, $\mathcal{J}_i = \{1, 2\}$ for all $i \in \mathcal{I}$. The first $n-1$ subjects in $\mathcal{I}$ are identically distributed over both subjects and time points, with $X_{i,j} \sim N(0, 1)$ for $i = 1, \ldots, n-1$ and $j = 1, 2$, while subject n has a different mean, namely $X_{nj} \sim N(\alpha_n, 1)$ for $j = 1, 2$ and (X_{n1}, X_{n2}) are bivariate normal with correlation $0 \leq \rho < 1$, which with standard margins has joint distribution function denoted by $\Phi_2(\cdot, \cdot; \rho)$. Thus here $F(x; \alpha_i) = \Phi(x - \alpha_i)$, with attributes $\alpha_1 = \ldots = \alpha_{n-1} = 0$ and α_n.

We will consider two cases for α_n (i) $(2 \log n)^{1/2}/\alpha_n = o(1)$ as $n \to \infty$ and (ii) $\alpha_n/(2 \log n)^{1/2} = o(1)$ as $n \to \infty$, i.e., so the latter includes both $\alpha_n \to \infty$ as $n \to \infty$ and $\alpha_n = 0$ for all n. We will show that cases (i) and (ii) lead to results which are consistent with asymptotic dependence and asymptotic independence, respectively in this non-identically distributed setting. Given the longitudinal variables are not necessarily identically distributed over subjects we explore both the standard definitions of asymptotic dependence and some alternatives to aid transparency.

First consider case (i) where α_n grows more rapidly with n with the subject specific dependence measures at lag-1, i.e., $(\chi_{i1}, \bar{\chi}_{i1})$. We have $(\chi_{i1}, \bar{\chi}_{i1}) = (0, 0)$ for subjects $i = 1, \ldots, n-1$ due to the independence assumption and due to the bivariate Normal distribution for subject n we have $(\chi_{i1}, \bar{\chi}_{i1}) = (0, \rho)$. So there is asymptotic independence across subjects, although subject n is not independent. For comparison, Fougères et al. (2009) model this such that given their subjects attribute, i.e., the random effect,

there is subject-specific independence over time for all subjects. So the basic Gaussian formulation is more general from this perspective.

Now consider the dependence of (M_1, M_2), which is needed for the evaluation of measures $\chi_1^{(M)}$ and $\bar{\chi}_1^{(M)}$ of Section 4.2.3. As we want to consider limits $n \to \infty$ for the population n of subjects in the longitudinal data, we modify our notation to let the maximum measurement over all subjects for time point j be denoted by $M_{n,j} := \max(\{X_{i,j} : i \in \mathcal{I}\})$, for $j = 1, 2$. We explore the dependence between (M_{n1}, M_{n2}) as $n \to \infty$. First, consider the two marginal probabilities $\Pr\{M_{nj} - \alpha_n < x\}$, for some $x \in \mathbb{R}$ and $j \in \{1, 2\}$. Then,

$$\Pr\{M_{nj} - \alpha_n < x\} = [\Phi(\alpha_n + x)]^{n-1} \Phi(x) \to \Phi(x) \tag{4.2.8}$$

as $n \to \infty$, i.e., a non-degenerate Gaussian limit. This result follows from the GEV to GPD link in Section 4.2.1 since for α_n growing as described above, $n[1-\Phi(\alpha_n+x)] \to 0$ for all $x \in \mathbb{R}$ since for $y \in \mathbb{R}$, from univariate extreme value results for standard Gaussian variables $n[1 - \Phi(a_n y + b_n)] \to \exp(-y)$ for $a_n = (2 \log n)^{-1/2}$ and $b_n = (2 \log n)^{1/2} + o(1)$ (Leadbetter et al., 2012). Now consider the joint probability, for $(x, y) \in \mathbb{R}^2$, as $n \to \infty$, given by

$$\Pr\{M_{n1} - \alpha_n < x, M_{n2} - \alpha_n < y\} = [\Phi(\alpha_n + x)\Phi(\alpha_n + y)]^{n-1} \Phi_2(x, y; \rho) \to \Phi_2(x, y; \rho), \tag{4.2.9}$$

where the non-degenerate limit arises using the same logic as for the marginal convergence. The joint maxima are asymptotically dependent when $\rho > 0$, with the limit not restricted to being a bivariate extreme value distribution as the variables are not identically distributed.

Now consider case (ii) for the behaviour of α_n. Then the equivalent results to the above are that as $n \to \infty$, $\Pr\{(M_{nj} - b_n)/a_n < x\} \to G(x)$, where $G(x) =$

$\exp(-\exp(-x))$, and

$$\Pr\{(M_{n1} - b_n)/a_n < x, (M_{n2} - b_n)/a_n < y\} \to G(x)G(y).$$

These limits show both a change in the marginal limit distribution from Gaussian to Gumbel and that there is now independence of limiting componentwise maxima.

These two asymptotic regimes for longitudinal data illustrate that the nature of extremal dependence is different for this framework than for the results for distinct stationary series. The limiting behaviour in these two scenarios illustrate that it is not essential to have asymptotic dependence per subject to achieve asymptotic dependence for longitudinal data; asymptotic dependence can be achieved by having subjects with a heavy tailed attribute distribution; and that both asymptotic dependence and asymptotic independence can be achieved from a simple Gaussian copula. Critical to the form of extremal dependence is the level of between-subject variation (via the attribute variation) relative to the within-subject variation. Here in case (i) α_n dominates the maximum of the measurements over all other subjects but not in case (ii). In essence, this occurs with the model of Fougères et al. (2009) with the latent positive stable law determining an attribute and conditional independence over measurements given this attribute.

To help better understand the nature of the asymptotic dependence case the supplementary material covers another version of the measure $\chi_\tau^{(M)}$ that allows both n and the quantile to grow in combination. Specifically, we consider the conditional probability $\Pr(M_{n2} > x_n \mid M_{n1} > x_n)$, where $x_n \to \infty$ and letting $\alpha_n = x_n - \delta$ for some constant δ.

4.3 Extremal Model for Longitudinal Data

4.3.1 Population Marginal Model

When developing a marginal model for the population of longitudinal random variables $\{(X_{i,j}, t_{i,j}) : j \in \mathcal{J}_i, i \in \mathcal{I}\}$, we make a critical decision of ignoring the subject-specific nature of the data as is conventional in previous analysis of this sort of data. We refer to this characteristic as *subject-ignorant*. Instead, the information regarding specific subjects is captured through our dependence modelling in Section 4.3.2. The reasons for this strategy are three-fold. Firstly, the number of observations per subject, e.g., $|\mathcal{J}_i|$ for subject i, is likely to be small in most applications and so a separate marginal model (see Section 4.2.1) per subject for the data in the tails is an unrealistic target. Secondly, modelling the tail of a population using a single GPD enables inference to be made about the population as a whole, additionally to subject-specific inferences. For example, in elite swimming, this allows inferences about how best performances over different swimmers have evolved over time (Spearing et al., 2021). Thirdly, following the convention of modelling the tail of a population using a single GPD enables application specific structure identified from previous analyses, which ignore subject knowledge, to be exploited.

Given the strategy described above, consider a generic pair (X, t), which to simplify presentation is written as X_t. For a constant threshold u over time, with methods for this threshold selection discussed in Section 4.2.1, there are three features of the distribution of X_t we describe: the behaviour above the threshold u, the probability of X_t exceeding u, and the distribution of X_t being below u. The latter is not typically studied in extremes of a univariate variable, but keeping track of the behaviour below the threshold is important here for dependence modelling of within-subject data in Section 4.3.2.

Above the threshold u we assume that $\Pr\{X_t < x | X_t > u\}$ has a $\text{GPD}(\sigma_u(t), \xi)$, as

given by expression (4.2.5). Although X_t is potentially complex in its variation over t, we only allow temporal variation in this distribution through the scale parameter. Pragmatically, assuming the shape parameter to be homogeneous is a typical approach in the modelling of extremes, since: there is limited evidence against this assumption in almost all applications, it aids parameter identifiability, and even when it is homogeneous it is difficult to estimate well. The probability of exceeding the threshold $\Pr\{X_t > u\} =: \lambda_u(t)$ is also allowed to vary with time. There is much literature on modelling approaches for how $(\sigma_u(t), \lambda_u(t))$ vary with t, including parametric (Davison and Smith, 1990), semi-parametric (Chavez-Demoulin and Davison, 2005) and fully non-parametric such as with splines (Jonathan and Ewans, 2013), Gaussian process (Casson and Coles, 1999) or machine learning approaches (Richards and Huser, 2022). We use parametric modelling, with models for our application set out in Section 4.5.2.

The X_t, conditionally on being below u, are assumed to follow some unknown but continuous density function $h_t : (-\infty, u] \to \mathbb{R}_+$, with $\int_{-\infty}^{u} h_t(s)\,\mathrm{d}s = 1$, where h_t does not depend on $(\lambda_u, \sigma_u, \xi)$. Combining all these models gives the distribution function F_{X_t} of X_t as

$$F_{X_t}(x) = \begin{cases} 1 - \lambda_u(t)\left[1 + \xi(x-u)/\sigma_u(t)\right]_+^{-\frac{1}{\xi}}, & x > u, \\ \left[1 - \lambda_u(t)\right] \int_{-\infty}^{x} h_t(s)\,\mathrm{d}s, & x \leq u. \end{cases} \tag{4.3.1}$$

As with the vast majority of extreme value modelling we avoid imposing a structure on the distribution of $X_t < u$, i.e., the density h_t here. Even if a parametric model for h_t had no parameters in common with those in the GPD or λ_u models, there is a risk of bias from mis-specifying h_t in the longitudinal setting due to the dependence between values $X_{i,j}$ and $X_{ij'}$ for $j' \neq j$, where $X_{i,j} < u < X_{ij'}$. In such cases, errors in modelling below the threshold can induce errors above the threshold to compensate. When making inferences about the extremes values of the longitudinal data, any actual value $X_{i,j}$ below u is instead treated as censored, i.e., as a realisation of the event

$X_{i,j} < u$.

4.3.2 Dependence Structure in a Latent Space

The focus now turns to modelling the dependence structure of random variables $\{(X_{i,j}, t_{i,j}) : j \in \mathcal{J}_i, i \in \mathcal{I}\}$, with the structure identified in Section 4.1. Specifically, we need to allow for temporal dependence between within-subject variables and independence between across-subject variables, so unlike in Section 4.3.1 knowledge of each subject's contribution to the data is accounted for. The formulation of these models builds on the theoretical findings of Section 4.2.3, which showed that multivariate Gaussian distributions for within-subject variations combined with an attribute distribution that has the capacity for both heavier and shorter tails than the within-subject Gaussian distribution, provide sufficient flexibility to allow for both asymptotic dependence and asymptotic independence, respectively.

The adopted modelling strategy bears likeness to that of copula modelling (Joe, 1997; Nelsen, 2007) or more specifically as in Wadsworth et al. (2017) and Huser and Wadsworth (2019), i.e., focusing on the joint structure of variables, without concern for its implications on the marginals at that stage. Subsequently, in Section 4.3.3, the marginal distribution of this model is linked to the formulation in Section 4.3.1. In particular, a model is adopted in terms of variables $\{(Z_{i,j}, t_{i,j}) : j \in \mathcal{J}_i, i \in \mathcal{I}\}$, where $Z_{i,j} = T_t(X_{i,j})$ for a function T_t defined in Section 4.3.3, and we refer to this as modelling in the *latent space*.

Consider a model in the latent space for the dependence arising between measurements from the same subject, e.g., $\{(Z_{i,j}, t_{i,j}) : j \in \mathcal{J}_i\}$ for subject i, a characteristic we term *subject-conditional dependence*. We follow standard Gaussian modelling assumptions of longitudinal data analysis (Diggle et al., 2002), which are also coherent with results in Section 4.2.3. The subject-specific model takes $Z_{i,j}$, across $j \in \mathcal{J}_i$, as realisations of a Gaussian process $Z_i(t)$ over time $t \in \mathbb{R}$ observed at the times

$\boldsymbol{t}_i := \{t_{i,j}; j \in \mathcal{I}_i\}$. Specifically,

$$Z_i(t) \sim \mathcal{GP}\left(\mu_i(t), \nu_i^2 K_{\boldsymbol{\kappa}}(\cdot,\cdot)\right), \text{ for all } t \in \mathbb{R}, \tag{4.3.2}$$

where the *mean function* $\mu_i(t) : \mathbb{R} \to \mathbb{R}$ is a subject-specific time-dependent mean, $\nu_i > 0$ is a homogeneous subject-specific standard deviation, and $K_{\boldsymbol{\kappa}}$ is a stationary kernel, which is shared over subjects, and which dictates the subject-conditional correlation between the process at any times $t \in \mathbb{R}$ and $t' \in \mathbb{R}$ with hyper-parameters $\boldsymbol{\kappa}$. The term $\mu_i(t)$ allows for the statistical properties of individual subjects to evolve over time separately from that of the population marginal model, as is the case for many applications in longitudinal analysis. To avoid over-parametrisation over individuals it is reasonable to assume that

$$\mu_i(t; \boldsymbol{\theta}_i, \boldsymbol{\gamma}) = \alpha_i + \mu(t, \tau_i; \boldsymbol{\gamma}), \text{ for all } t \in \mathbb{R}, \tag{4.3.3}$$

for a subject-ignorant function μ with parameters $\boldsymbol{\gamma}$, subject-specific parameters $\boldsymbol{\theta}_i := (\alpha_i, \tau_i)$ and covariates (which are ignored in this formulation, but are used in Section 4.5.2). To ensure that α_i is identifiable, the maximum of the function μ, over t, is set to zero, i.e., $\alpha_i = \max_{t\in\mathbb{R}} \mu_i(t; \boldsymbol{\theta}_i)$. Then α_i is the subject *attribute* as in Section 4.2.3. When $\mu \equiv 0$ in model (4.3.2) the subject-specific dependence measures are $(\chi_{i\tau}, \bar{\chi}_{i\tau}) = (0, K_{\boldsymbol{\kappa}}(0, \tau))$, for all $i \in \mathcal{I}$.

The form of the stationary kernel is application specific. A powered exponential is used

$$K_{\boldsymbol{\kappa}}(t, t') = \exp(-\kappa_0 |t - t'|^{\kappa_1}),$$

with $\boldsymbol{\kappa} = (\kappa_0, \kappa_1) \in \mathbb{R}_+ \times [0.5, 2]$ in Section 4.5, where smaller κ_0 gives less subject-conditional dependence (with the limit $\kappa_0 \to \infty$ giving subject-conditional independence); and κ_1 influences the local smoothness of the process, with larger κ_1 giving a

smoother process, with the limit $\kappa_1 \to 2$ corresponding to a process which is infinitely differentiable, and when $\kappa_1 = 1$ the process is Markov. There are many other well-established stationary kernels, e.g., the Matérn family, see Diggle et al. (1998). Some of these were trialled in exploratory analysis for the application in Section 4.5 but were found to make no practical differences, which is a consequence of there being only a few observations per subject and the absence of observations at short time lags.

Conditioning on the parameters of this latent model, the marginal distribution of Z, an arbitrary observation from the longitudinal data set in latent space, is

$$G_Z(z) = \frac{1}{n} \sum_{i \in \mathcal{I}} \sum_{j=1}^{n_i} \Phi\left(\frac{z - \mu_i(t_{i,j})}{\nu_i} \right), \tag{4.3.4}$$

where $n_i := |\mathcal{I}_i|$ and $n = \sum_{i \in \mathcal{I}} n_i$. Thus the marginal distribution of Z is a mixture of Gaussian variables over subjects and different observation times.

Finally, consider the marginal variation across subjects. As in Section 4.2.3 this variation is captured exclusively through the distribution of the attributes $\{\alpha_i : i \in \mathcal{I}\}$. Over the whole population of subjects all α_i are taken to be independent and identically distributed with $\alpha_i \sim N(0, V_\alpha^2)$ for all $i \in \mathcal{I}$, for a given fixed value of $V_\alpha > 0$. This model for the attributes is assumed to hold over the population as a whole, as well as over set of observed subjects $\mathcal{I}$, with the former covering all potential subjects outside the time window of observations.

From Section 4.2.3, it is clear that the variation in the α_i relative to the within subject variability, i.e., ν_i for subject i, is what determines whether the longitudinal data exhibit asymptotic dependence or asymptotic independence. So, the precise value of V_α is irrelevant for differentiating between within-subject variation and population variation, as this is controlled by ν_i / V_α. Hence V_α can be fixed to any chosen value, since the $\{\nu_i\}$ are parameters to be estimated from the data, and so their values can adapt proportionally to changes in V_α. Thus the data determine the form of longitudinal

data extremal dependence, the degree of which we measure using the lag-τ coefficients $(\chi_\tau^{(M)}, \bar{\chi}_\tau^{(M)})$ or $(\chi_\tau^{(R)}, \bar{\chi}_\tau^{(R)})$ of Section 4.2.3.

4.3.3 Transforming Margins between Observed and Latent Spaces

The probability integral transform (4.3.5) is used to transform between the observation scale of X and the latent space of Z, defined in Sections 4.2.1 and 4.3.2 respectively. Specifically, the variables $X_{i,j}$ and $Z_{i,j}$, both occurring at time $t_{i,j}$ are linked by

$$G_Z(Z_{i,j}) = F_{X_{t_{i,j}}}(X_{i,j}), \text{ so } Z_{i,j} := T_t(X_{i,j}) = G_Z^{-1}\{F_{X_{t_{i,j}}}(X_{i,j})\} \tag{4.3.5}$$

where F_{X_t} and G_Z are defined by expressions (4.3.1) and (4.3.4), respectively and T_t is the transformation outlined in Section 4.3.2.

For those $X_{i,j}$ points above the threshold on the original margins, the transform is given as

$$Z_{i,j} = G_Z^{-1}\left\{1 - \lambda_u(t_{i,j})\left[1 + \xi(X_{i,j} - u)/\sigma_u(t_{i,j})\right]_+^{-\frac{1}{\xi}}\right\},$$

whereas when these points are below the threshold,

$$Z_{i,j} = G_Z^{-1}\left\{\left[1 - \lambda_u(t_{i,j})\right]\int_{-\infty}^{X_{i,j}} h_{t_{i,j}}(s)\,\mathrm{d}s\right\}.$$

The complication here is that the density function h_t is unknown and we do not want to model it, hence the censoring approach outlined in Section 4.2.1. Instead, for this range of $X_{i,j}$, the random variable $V_{i,j} := \int_{-\infty}^{X_{i,j}} h_{t_{i,j}}(s)\,\mathrm{d}s$ is uniform(0,1) distributed. So the auxiliary variable $V_{i,j} \sim \text{Uniform}(0,1)$ is introduced into the transformation when $X_{i,j} < u$, to give $Z_{i,j} = G_Z^{-1}\left\{\left[1 - \lambda_u(t_{i,j})\right] V_{i,j}\right\}$. A consequence of the transformation is that the threshold u in the observation space becomes time-varying in the latent space, i.e., $u_Z(t) = G_Z^{-1}\left\{\left[1 - \lambda_u(t)\right]\right\}$.

For making joint inferences across marginal and dependence structure parameters

the likelihood functions in Section 4.4 require the Jacobian terms for these transformations. These terms require the marginal density in the latent space, i.e.,

$$g_Z(z;\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\nu}) = \frac{1}{n}\sum_{i\in\mathcal{I}}\sum_{j=1}^{n_i}\frac{1}{\nu_i}\phi\left(\frac{z-\mu_i(t_{i,j};,\boldsymbol{\theta}_i,\boldsymbol{\gamma})}{\nu_i}\right),$$

where $\boldsymbol{\nu} := \{\nu_i : i \in \mathcal{I}\}$ and $\boldsymbol{\theta} := \{\boldsymbol{\theta}_i : i \in \mathcal{I}\}$. For a realisation x of X (or v of V) when the observation is above (or below) u, respectively, the associated realised value z of Z is obtained using the transformations above. The corresponding Jacobian terms, J_+ and J_- for above and below the threshold, respectively at time t are

$$\begin{aligned} J_+(x;t,\xi,\boldsymbol{\sigma},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\nu}) &= \frac{\lambda_u(t;\boldsymbol{\beta})}{\sigma_u(t;\boldsymbol{\sigma})g_Z(z;\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\nu})}[1+\xi(x-u)/\sigma_u(t;\boldsymbol{\sigma})]_+^{-\frac{1}{\xi}-1}, \\ J_-(v;t,\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\nu}) &= [1-\lambda_u(t;\boldsymbol{\beta})]/g_Z(z;\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\nu}), \end{aligned}$$

which follows from $\int g_Z(z;t)\,\mathrm{d}z = \int f_X(x;t)\,\mathrm{d}x$ for all t, and where $\boldsymbol{\sigma}$ and $\boldsymbol{\beta}$ are the parameters of the model for $\sigma_u(t)$ and $\lambda_u(t)$, respectively.

4.3.4 Predicting future extreme events in longitudinal data

A benefit of accounting for the longitudinal structure is that now inference and predictions of extreme events regarding individual subjects is ascertainable, e.g., a new record is achieved by a particular subject $i \in \mathcal{I}$. To make such inferences, each subject's mean function over time is incorporated, as well as the temporal dependence around this. Both of these aspects are described by the Gaussian process model of Section 4.3.2, which gives analytical solutions to probabilities of future events through its closed form conditional distributions. In the supplementary material, we provide an example prediction, namely, the probability of a subject $i \in \mathcal{I}$ breaking the record for the maximum measurement r in some future time period F, denoted by the event $A_i^F(r)$. The probability of $A_i^F(r)$ is derived under some assumptions for an idealised scenario, including

having independent and identically-distributed variables.

Such simplifying assumptions are realistic for some applications and thus analytical results are feasible. However, in applications where these assumptions can no longer be assumed, evaluation of the probabilities of such complex future events are most simply conducted through Monte Carlo methods, simulating over different realisations of the longitudinal process for the fitted model. A particular complication arises in applications with subject-specific mean functions - as we have in our elite swimming application, Section 4.5 - which induce non-identically-distributed variables, for there it must be recognised that in the longer-term the extreme events are more likely to be due to subjects not yet observed in $\mathcal{I}$.

In the short-term however, these future extreme events are most likely to be obtained by the current subjects in $\mathcal{I}$, followed by a transitional medium-term in which extremes arise from a mixture of these populations of subjects. In the supplementary material, we develop the outline of a simulation framework for such inferences, setting out some possible choices that need to be made in relation to the currently unobserved subjects. Going forward beyond the observed time-frame, this framework outlines three classes of subjects: (i) those subjects in $\mathcal{I}$, indexed by $\mathcal{I}^c$ with $\mathcal{I}^c \subseteq \mathcal{I}$, which are still producing at least one measurement above u in the future time window; (ii) those subjects $\mathcal{I}^f$, which produced measurements exclusively below the threshold within the observed time-frame and so $\{\mathcal{I}^f \cap \mathcal{I}\} = \emptyset$, but in the future produce a measurement above u; and (iii) those subjects $\mathcal{I}^n$ with no recordings at all within the observed time-frame but which in the future period produce at least one measurement above u. The supplementary material details a strategy of simulating from each of these three groups.

4.4 Inference

The likelihood is constructed in two steps. First, we assume that the parameters $(\xi, \boldsymbol{\sigma}, \boldsymbol{\beta})$ and the vector of auxiliary variables for the marginal variables in the observed space are known, so that only the parameters effecting the latent space need to be estimated. Secondly, we account for uncertainty in these marginal parameters and auxiliary variables. For deriving the likelihood in the latent space for a given subject i with observations $(\boldsymbol{Z}_i, \boldsymbol{t}_i) := \{(Z_{i,j}, t_{i,j}) : j \in \mathcal{J}_i\}$, we define the correlation matrix between all of subject i's observations by the correlation matrix $\Sigma^i_{\boldsymbol{\kappa}} := K_{\boldsymbol{\kappa}}(\boldsymbol{t}_i, \boldsymbol{t}_i)$, i.e., the $(j,k)^{th}$ entry $\Sigma^{i,(j,k)}_{\boldsymbol{\kappa}} := K_{\boldsymbol{\kappa}}(t_{i,j}, t_{i,k})$ is the correlation between $Z_{i,j}$ and $Z_{i,k}$. Recalling that observations for a subject are from a multivariate Gaussian distribution, and that across different subjects are independent, the likelihood in the latent space for realisations $\boldsymbol{z} := \{\boldsymbol{z}_i : i \in \mathcal{I}\}$ at times $\boldsymbol{t} := \{\boldsymbol{t}_i : i \in \mathcal{I}\}$, is given as

$$L_\ell(\boldsymbol{z}; \boldsymbol{t}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\nu}, \boldsymbol{\kappa}) \propto \prod_{i \in \mathcal{I}} \nu_i^{-n_i} |\Sigma^i_{\boldsymbol{\kappa}}|^{-1} \exp\left(-\frac{1}{2} \tilde{\boldsymbol{z}}_i^T \Sigma^i_{\boldsymbol{\kappa}} \, \tilde{\boldsymbol{z}}_i\right) \tag{4.4.1}$$

$$\text{with } \tilde{\boldsymbol{z}}_i := \left\{ \frac{z_{i,j} - \mu_i(t_{i,j}; \boldsymbol{\theta}_i, \boldsymbol{\gamma})}{\nu_i} : j \in \mathcal{J}_i \right\}, \ \forall i \in \mathcal{I}.$$

The parameters for the margins in the observational space are in practice unknown. Therefore the full likelihood requires the Jacobian terms, from expression (4.3.6), which control the transformations between the two spaces. Let the sets of observations which are below and above the threshold be $\mathcal{L}_- := \{(i,j) : X_{i,j} \leq u : j \in \mathcal{J}_i, i \in \mathcal{I}\}$ and $\mathcal{L}_+ := \{(i,j) : X_{i,j} > u : j \in \mathcal{J}_i, i \in \mathcal{I}\}$ respectively. The full likelihood of parameters $\boldsymbol{\Theta} := (\xi, \boldsymbol{\sigma}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\nu}, \boldsymbol{\kappa})$ and auxiliary variables is

$$L(\boldsymbol{x}, \boldsymbol{v}; \boldsymbol{t}, \boldsymbol{\Theta}) \propto L_\ell(\boldsymbol{z}; \boldsymbol{t}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\nu}, \boldsymbol{\kappa}) \times \left(\prod_{(i,j) \in \mathcal{L}_-} J_-(v_{i,j}; t_{i,j}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\nu}) \right) \left(\prod_{(i,j) \in \mathcal{L}_+} J_+(x_{i,j}; t_{i,j}, \xi, \boldsymbol{\sigma}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\nu}) \right),$$

where $\boldsymbol{v} := \{v_{i,j} : (i,j) \in \mathcal{L}_-\}$ and $\boldsymbol{z}$ is a function of $\boldsymbol{x}$ and $\boldsymbol{v}$, as identified in Section 4.3.3.

There are likely to be complications with using an entirely likelihood-based approach to inference for this model. Firstly, with two parameters per subject, limited data per subject, and many subjects, an asymptotic-based inference justification and its associated uncertainty evaluation is not supported. There are issues with dealing with the auxiliary variables, with the need to integrate over these making likelihood evaluation difficult. In many multivariate extreme value models with complex dependence structure, inference is conducted via pseudo likelihoods and using bootstrap procedures which avoid the need to model the dependence structure (Davison et al., 2012). Such approaches are not suitable here because of the large influence of a small number of large values on the parameter estimates, so biased inference is likely (Healy et al., 2023). Furthermore, with longitudinal data there are problems of identifiability under bootstrap sampling given that data from subjects with more limited data are more likely to be omitted in replicate samples than for other subjects. Instead, we adopt a Bayesian inference framework, which provides full uncertainty quantification of all parameters and auxiliary variables simultaneously. The Bayesian framework also allows for easier uncertainty quantification in the prediction of future events, see Section 4.5.

Let the parameters $\boldsymbol{\Theta}$ have prior distribution $\pi_{\boldsymbol{\Theta}}(\boldsymbol{\Theta})$, and let the prior $\pi_{V_{i,j}}(v)$ for all $(i,j) \in \mathcal{L}_-$ be uniform $(0,1)$ distributed and to be independent across these variables. Then, the full posterior distribution can be written as

$$\pi\left(\boldsymbol{\Theta}, \boldsymbol{v} | \boldsymbol{x}, \boldsymbol{t}\right) \propto \pi_{\boldsymbol{\Theta}}\left(\boldsymbol{\Theta}\right) L(\boldsymbol{x}, \boldsymbol{v}; \boldsymbol{t}, \boldsymbol{\Theta}). \tag{4.4.2}$$

In Section 4.5.3 we present the prior $\pi_{\boldsymbol{\Theta}}$ for our analysis of elite swimming data.

Before detailing our Markov chain Monte Carlo (MCMC) methods for simulating from the joint posterior, we identify a computational issue that influences our choice of MCMC strategy. Specifically, in order to transform the data from the observed space into the latent space, the inverse of the Gaussian mixture distribution (4.3.4)

is required, but that has no analytical solution. Numerical solution of this inverse is required for each likelihood evaluation, and for each data point for each subject. Exact numerical solution on this scale is computationally infeasible. Instead, for likelihood evaluation we use a grid search algorithm, searching over a finite regular grid $\mathcal{Z}_G$ in the latent space, for each data point $x_{i,j}$, such that

$$z_{i,j} = \begin{cases} \text{argmin}_{z \in \mathcal{Z}_G} \left(|G_Z(z) - F_X\left(x_{i,j}, t_{i,j}\right)| \right), & x_{i,j} > u, \\ \text{argmin}_{z \in \mathcal{Z}_G} \left(|G_Z(z) - [1 - \lambda_u(t_{i,j})] v_{i,j}| \right), & x_{i,j} \leq u, \end{cases} \tag{4.4.3}$$

where both F_X and G_Z depend on the parameter values of each likelihood evaluation. This grid search approach slows down inference significantly since it requires a factor of $|\mathcal{Z}_G|$ more evaluations relative to there being an exact solution to equation (4.3.5). Moreover, the discrete nature of the grid search, with no gradient information, rules out our use of a range of popular Bayesian inference algorithms, e.g., Hamiltonian Monte Carlo (Duane et al., 1987) and the No U-Turn Sampler (Hoffman and Gelman, 2014). Section 4.6 discusses this point further.

Given these constraints and the slow likelihood evaluation, a Metropolis-Hastings (MH) algorithm is implemented that utilises the Python package *PyMC* (Salvatier et al., 2016), which enables efficient inference through automated optimisation of the algorithms' tuning parameters. In the case of MH, this provides well-tuned proposal distributions for optimal exploration of the joint posterior distribution. For a further speed-up, we sample a large number (in our case 40) of shorter MCMC chains (2000 samples each, including 1000 'burn-in' samples) in parallel using high-performance computing, which then undergo standard diagnostics for checking of convergence (Gelman and Rubin, 1992). By randomly drawing all realisations $\boldsymbol{v}$ from its prior distribution at each step of the MCMC algorithm, and then considering only the marginal distribution for $\boldsymbol{\Theta}$, in essence pseudo-marginal MCMC (Andrieu and Roberts, 2009) is performed, and the posterior $\pi\left(\boldsymbol{\Theta}|\boldsymbol{x}, \boldsymbol{t}\right)$ is recovered.

In order to attain inference for future predictions, see Section 4.3.4, the full prediction uncertainty is propagated through the model. Given the set of simulated timestamps $\boldsymbol{t}_i^*$ of future observations by a subject i, which are randomly generated by the process described in Section 4.3.4, the variables $Z_i(t) \sim \mathcal{GP}\left\{\mu(t;\boldsymbol{\theta}_i,\boldsymbol{\gamma}), K_{\boldsymbol{\kappa}}(\cdot,\cdot)\right\}$, jointly for t over the set $\boldsymbol{t}_i^*$, are sampled from the Gaussian process, with the parameter values being a random sample s from the joint posterior $\pi\left(\boldsymbol{\Theta},\boldsymbol{v}|\boldsymbol{x},\boldsymbol{t}\right)$. The sample is then transformed back to its original margins. For those samples above the time-varying threshold on the latent scale u_Z, the GPD parameter values used in the transformation are the same sample s from the posterior.

For an observation below the threshold - which is by definition not extreme - the actual value on the original margins is unimportant for inference of extreme events. Only the time of occurrence and the knowledge that they are below the threshold are relevant, in order to characterise the dependence structure, and which we already have from the simulation process. However, for visualisation purposes it is useful to have some estimate of non-extreme values on the original scale, see Figure 4.5.4. In this case the empirical CDF is used, though it is acknowledged that this does not include the uncertainty in the distribution on the original margins.

4.5 Application

4.5.1 Data

The data analysed constitutes mens' 100m breaststroke results in FINA competitions in the period 2012-2019, obtained from the FINA website. A few strategic decisions were made about which data to analyse. Firstly, only data of each swimmer's best time swam per competition was selected, i.e., one swim per competition. This removes much of the tactical element, e.g., weaker swimmers may need to swim to full capacity during the heats of competitions, whereas a top swimmer can typically afford to save their

best performances for the finals. Using exclusively these *competition maxima* helps to ensure that each observation is a good approximation of the swimmer's best ability at that time. It also has the benefit of avoiding the need to capture performance strategy or to deal with issues of dependence at very short time lags.

Secondly, in extreme value analysis, the scale on which the data analysis is performed can impact the results (Wadsworth et al., 2010). Following the discussion in Spearing et al. (2021) minimum swim-times are modelled, but modelling the maximum swim-speed (Gomes and Henriques-Rodrigues, 2019), i.e., the reciprocal of the times swam, is also an option. For analysing minimum swim-times, results exist for the behaviour of the lower tails of a distribution, however they are rarely applied (Robinson and Tawn, 1995) and give identical results to our strategy. We therefore analyse negative swim-times, and then negate any estimated quantiles in order to provide results for actual swim-times that make use of the more commonly-used methodological frameworks for upper tails. Finally, the threshold must be selected. The analysis of Spearing et al. (2021) identified the 200th fastest personal best (PB) over the period 2001-18 as a suitable (negative) extreme threshold $u = -61.125$ seconds, so this threshold is adopted here despite now focusing on the larger data set of all competition maxima per swimmer.

Our model has two subject-specific parameters $\boldsymbol{\theta}_i = (\alpha_i, \tau_i)$ per swimmer. Unless swimmer i has undertaken sufficient swims in the data set then the posterior for the parameters for such swimmers will be weakly informed by the data, or even unidentifiable from the data if swimmer i has only one recording. Here the standard Bayesian approach, and perhaps the most obvious, is to carry out analysis regardless and acknowledge that the marginal posterior distributions for such $\boldsymbol{\theta}_i$ will be almost identical to the associated prior distributions. However, the prior on α_i is necessarily vague to allow for variation over swimmers, see Section 4.5.3, so the posterior information about these parameters adds little value to the overall inference. Moreover, it comes at a large computational cost from the many uninformative parameters, which requires the

MCMC to do approximately twice as many of the slow likelihood evaluations, see Section 4.4. Our analysis is instead restricted to only those swimmers with a "sufficient" number, i.e., more than m, of recordings in the data set. So, for the set of swimmers $\mathcal{I}_m$ that have recorded m or fewer swims, i.e., for all $i \in \mathcal{I}$, with $n_i := |\mathcal{J}_i| \leq m$, these data are ignored. The analysis is therefore conducted on the swimmers $\mathcal{I} \setminus \mathcal{I}_m$. A potential consequence of restricting the data set is that the GPD may no longer be a good fit to the tails of the data; however, we show in Section 4.5 that this does not appear to be the case. Section 4.6 discusses alternative approaches that do use the data for swimmers $\mathcal{I}_m$ and which do not suffer from computational complications, but they require additional modelling assumptions.

If m is chosen to be too small, some $\boldsymbol{\theta}_i$ will have marginal posteriors with only minor differences from their priors and at the computational cost of needing more MCMC samples for convergence given the two additional variables per extra swimmer included. With m too large, too much data are excluded so posteriors are less well informed than necessary. The strategy for choosing m is to observe how the number of swimmers that have swum less than or equal to m, i.e., $k_m := \sum_{i \in \mathcal{I}} \mathbb{1}\{n_i \leq m\}$, varies with m. An abrupt increase was found when $m = 7$. Therefore, by selecting $m = 7$, a relatively large proportion of those swimmers with only few observations (40%) are discarded, whilst only losing 20% of the total observations. The final dataset used for analysis contained 120 swimmers, with 1435 total observations. In an early analysis the model was fitted using only the 10 most prolific swimmers, i.e., $m = 18$. Using these data the posterior means of the GPD parameters were very similar to those in the final analysis, reported in Section 4.5.4, indicating that there is very little bias, or sensitivity, introduced through the choice of m.

4.5.2 Modelling applied to swimming

From Spearing et al. (2021), the conditional distribution of extreme swim-times $\Pr\{X < x|X > u\}$ for large u can be treated as identically distributed over time, and so we take $\sigma_u(t) =: \sigma_u \in \mathbb{R}_+, \ \forall t$, i.e., $\boldsymbol{\sigma} = \sigma_u$. The common temporal trend across the population of elite breaststroke swimmers can then be captured through the probability of exceeding the threshold λ_u, through a smooth monotonically increasing function for λ_u. A logit-linear functional form for λ_u was found appropriate for the change in λ_u over t. Specifically, for a swim-time in year $t \in \{2012, \ldots, 2020\}$ and parameters $\boldsymbol{\beta} := (\beta_0, \beta_1) \in \mathbb{R}^2$, we denote the model

$$\lambda_u(t; \boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 t)/[1 + \exp(\beta_0 + \beta_1 t)]. \tag{4.5.1}$$

We next turn attention to the *subject-specific* trends. In elite swimming, the subject trend captures a swimmer's *career trajectory* - the tendency for athletes to enter elite sports as relatively inexperienced, improve until some individual *peak* ability, and then decline before leaving the sport. Swimmers tend to improve rapidly towards their peak mean performance α_i, at an age of τ_i, as they mature physically, and then stop competing within a few years of reaching this peak. Here we allow the time at which peak mean performance is achieved to vary over swimmers to allow for their differences in maturity. The lack of data in the decline of the career trajectory enables the parsimonious assumption of a symmetric career trajectory about the peak. From what can be identified from the data, after transformation to the latent space, a quadratic mean trend in age of swimmer, with curvature $\gamma < 0$, seems a reasonable approximation to this mean performance progression. Hence, we introduce swimmers' age as a covariate, by first including the covariate $b_i \in \mathbb{R}$, swimmer i's birth date, so that $t - b_i$, for $t > b_i$, is the age at which swimmer i recorded the swim at time t. Thus, for a swim recorded

at a time t by a swimmer i, the mean function is

$$\mu_i(t; b_i, \boldsymbol{\theta}_i, \gamma) = \alpha_i - \gamma(t - b_i - \tau_i)^2, \text{ for all } t \in \mathbb{R},$$

for all $i \in \mathcal{I}$, where $\boldsymbol{\theta}_i =: (\alpha_i, \tau_i) \in \mathbb{R} \times \mathbb{R}_+$, and here $\boldsymbol{\gamma} = \gamma > 0$. We do not attempt to have a swimmer-specific parameter for γ given the limited number of swims per swimmer. Furthermore, it was found that the variance across swims from a swimmer i, ν_i^2 could be assumed the same across swimmers, i.e., $\nu_i =: \nu \in \mathbb{R}_+, \ \forall i \in \mathcal{I}$. It seems that different latent mean values per swimmer is sufficient to capture the across-swimmer effects.

4.5.3 Prior specification

The DAG in Figure 4.5.1 illustrates the full model specification for this swimming application, and in particular, the formulation of $\pi_{\boldsymbol{\Theta}}$, defined in Section 4.4. For simplicity the priors are assumed to be mutually independent across all components of $\boldsymbol{\Theta}$, i.e., the full prior can be written as

$$\pi_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}) = \pi(\xi)\,\pi(\sigma_u)\pi(\boldsymbol{\beta}) \left(\prod_{i \in \mathcal{I}} \pi(\alpha_i)\,\pi(\tau_i) \right) \pi(\gamma)\,\pi(\nu)\,\pi(\boldsymbol{\kappa}). \tag{4.5.2}$$

We now explain our choices of these marginal priors in the sequence shown in expression (4.5.2).

Considerable discussion on priors for GPD parameters goes back to Coles and Tawn (1996). Selecting the shape parameter prior to be $\text{logit}(\xi + 1) \sim \mathcal{N}(\text{logit}(0.8), 0.3)$ approximately restricts the domain of the shape parameter to be $-1 < \xi < 0$. The constraint $-1 < \xi$ is reasonable as the likelihood is infinite otherwise when the upper endpoint of the GPD is set to the sample maximum (Smith, 1985), and so if violated it could lead to estimates of the GPD which imply that there is no possible improvement on the best time already achieved. The constraint $\xi < 0$ implies that there exists some

finite limit on the fastest possible performance, which is sensible. Furthermore, analysis of elite swimmers' PB data from 2001-2019 (Spearing et al., 2021) found strong evidence of a common negative shape parameter over all swimming distances, strokes and gender categories. For determining the prior distribution for the GPD scale parameter we exploited knowledge from Spearing et al. (2021) that an estimated value of this parameter using PB data for this event was close to 1. So taking $\sigma_u \sim \text{Gamma}(25, 25)$ enforces positivity, has the required mean value, and a standard deviation of 0.2. Lastly for the threshold exceedance rate parameters $\boldsymbol{\beta}$, as given in expression (4.5.1), the priors $\beta_0 \sim N(0, 0.5)$, and $\beta_1 \sim \text{Gamma}(0.1, 0.1)$ are imposed. The latter prior is selected to reflect our knowledge of an improvement on the ability of swimmers (Spearing et al., 2021), and combined with the former gives a wide range of likely exceedance rates of approximately $(0.1, 0.9)$.

Now consider the prior choices for the parameters that determine the distribution of the process in the latent space. First consider the marginal parameters. As discussed in Section 4.3.2, we take $\alpha_i \sim N(0, V_\alpha^2)$, where here we select $V_\alpha = 6$, to allow for considerable variation in the skill level of the elite swimmers. The priors on the peak ages of the swimmers were taken to be $\tau_i \sim N(25, 2.5^2)$ to reflect that a swimmer's typical might peak at roughly 25 years old and with a reasonable probability that it falls somewhere in the region $(17.5, 32.5)$ years of age. The prior for the rate of quadratic decay from the peak performance in the latent space is given by $\gamma \sim \text{Gamma}(0.5, 0.5)$ distribution. This is weakly informative with a small preference for γ to be arbitrarily close to 0, so that a posterior suggesting $\gamma > 0$ cannot simply be an artefact of the prior. As we expect a much greater difference between subjects than within a subject's performances, the variance of the prior for ν is taken to be much smaller than V_α, with a prior $\nu \sim \text{Gamma}(1, 1)$. Finally, consider the kernel parameters: $\kappa_0 \sim \text{Gamma}(0.5, 0.5)$ enforces $\kappa_0 > 0$, where the hyper-parameters choice allows κ_0 to be arbitrarily close to 0, i.e., the special case of subject-specific independence; and for $\kappa_1 \in (0.5, 2)$, the prior

was selected be to $\text{logit}(\kappa_1 - 0.5)/1.5) \sim N(\text{logit}(1), 2)$, so the prior mean of κ_0 is the Markov special case, and the larger standard deviation allows an exploration of the full domain.

4.5.4 Results

Subject-specific Inference

We first focus on the within-subject features of the model that provides information about individual swimmers as well as playing a key role in determining the dependence structure across the longitudinal data of elite breaststroke swimmers. As identified in Section 4.2.3, there are two features of the subject-specific behaviour which affect the extremal dependence of these data: the subject-specific variation in the attributes, here captured by the $\{\alpha_i : i \in \mathcal{I}\}$; and the within-subject dependence, given by the Gaussian process.

The marginal posterior distributions of the parameters $\boldsymbol{\theta}_i$ are shown in Figure 4.5.2 for the top ten swimmers, as defined in Section 4.5.4, a ranking that strongly correlates with the swimmers with the ten largest posterior mean α_i values. With the exception of the posterior for Adam Peaty's α_i, there is considerable overlap between the other nine posteriors, with Peaty's having both a larger mean and 50.5% of the variation of the others. The larger mean value is not too surprising as Peaty holds the 7 fastest times, and 11 of the top 20, for the competition-best data analysed, together with all the top 20 times over all swims in this event. The posteriors for the τ_i for these top ten swimmers are broadly more self-consistent across swimmers, with almost all posterior mass for the peak performance age in the range $(25, 35)$ years, though both Peaty and Andrew Michael appear to have lower peak ages, with Peaty almost certainly having peaked before the age of 30 (he is 29 at the time of writing).

What is possibly most intriguing about these posteriors is that the posterior of α_i for Nicolo Martinenhi has upper quantiles which exceed the same quantiles for Peaty's

α_i, with his mean and median α_i being notably smaller than Peaty's values. There are three possible compounding causes for this which we explored as follows. Firstly, it could be that Martinenhi produced some high quality swims, but also has much variability in these, which suggests he may be capable of getting much better swims; however, this is unlikely since only two of Peaty's swim-times are slower than Martinenhi's PB. Secondly, the greater posterior uncertainty of Martinenhi's α_i could be as he has much less swims in the database relative to Peaty, but in fact Martinenhi has 14 better than the threshold, which is comparable to Peaty's 17. The third cause, and seemingly the most likely, is that Martinenhi is relatively young - five years younger than Peaty - being 20 years old when he produced his most recent time in the database. For younger swimmers it is difficult to disentangle between peak age and attributed, which is evidence by Martinenhi having the largest posterior correlation, of 0.89, between his (α_i, τ_i) of the top 10 swimmers, e.g., for Peaty this is 0.50. Martinenhi's large uncertainty in peak age is contributing greatly to the uncertainty in his attribute; his peak is still to come - but we are uncertain in its level.

The posterior 95% highest posterior density interval (HPDI) for the subject-specific quadratic trend curvature γ is $(0.015, 0.029)$, showing that there is concrete evidence of a rising and falling career trajectory, especially given the prior favours γ being arbitrarily close to 0. The 95% HPDI for the ratio of within-subject to across subject variation, i.e., ν/V_α, is $(0.17, 0.18)$, showing that the majority of the variation in the extremes of these longitudinal data is explained by swimmer identification. Furthermore, with Peaty having much the latest α_i, Section 4.2.3 indicates there will be asymptotic dependence, irrespective of the within-subject dependence $\rho(\tau)$ at lag τ. The posterior mean and pointwise 95% (HPDI) are shown in Figure 4.5.2 (right) for the measure of subject-specific asymptotic independence $\bar{\chi}_{i,\tau} = \rho(\tau)$, for lag $\tau \in [5, 365]$ days. This inference indicates that at 50 days there is reasonable dependence per swimmer and even at 6 months lag there is non-negligible subject-conditional dependence. The other measures

of extremal dependence are difficult to ascertain due to the underlying variables being non-identically distributed as evidenced by the posterior for $\gamma > 0$.

Subject-ignorant Marginal Inference

Now consider the joint posterior inferences for the subject-ignorant marginal distribution parameters for the GPD and tail exceedance probabilities, i.e., $\{\sigma_u, \xi, \boldsymbol{\beta}\}$, which are derived from the joint posterior (4.4.2) for all the model parameters. The posterior mean for ξ and its 95% HPDI are -0.22 $(-0.25, -0.20)$, providing strong evidence for a negative shape parameter. Similarly, for β_1 these values are 0.13 $(0.09, 0.16)$, showing that the rate of achieving extreme elite performances in this event is increasing over the time window from 2012-2019, with the posterior mean and 95% HPDI for $\lambda_u(t)$ being 0.34 $(0.30, 0.38)$ for 2012 and 0.55 $(0.51, 0.59)$ for 2019, showing that the value of β_1 corresponds to a substantial difference in the behaviour of the population over the observed time period.

We now adapt this inference for informing us about extreme marginal events. As described in Section 4.2, when $\xi < 0$ there is an estimated upper endpoint $x_H = u - \sigma_u/\xi$, which in the context of swimming is interpreted as the best possible performance humanly possible, given the current technology, a quantity that has been widely studied in sports (Huub and Trultens, 2005; Nevill et al., 2007). Figure 4.5.3 shows the posterior distribution of x_H, and the closeness of Peaty's current world record to the ultimate possible time. The posterior places the endpoint closer to the current record than similar analysis of PB data (Spearing et al., 2021), with the earlier analysis pooling information across events.

The expected value of the next world record swim-time is obtained by exploiting the *threshold-stability* property of a GPD (Coles, 2001). Since the (negative) current world record $r = -56.88 > u$, exceedances above r follow a GPD. Setting the threshold at r, then exceedances of r, denoted $X_{r_+} := \{X : X > r\}$, follow a GPD$(\sigma_r = \sigma_u + \xi(r-u), \xi)$

and the expected next world record time is $\mathbb{E}[X_{r_+}] = r + \sigma_r/(1-\xi)$. Figure 4.5.3 (left) shows the posterior distribution of $\mathbb{E}[X_{r_+}]$. Although having some overlap with the posterior of x_H, the posterior of $\mathbb{E}[X_{r_+}]$ is much nearer Peaty's current record than x_H. The simplicity of the result for the expected record value arises as both ξ and σ_u are constant over time and the expectation is not conditional on the current swimmers' performances, with the latter considered in Section 4.5.4. Furthermore, this posterior for $\mathbb{E}[X_{r_+}]$ provides no information about when this record is likely to be achieved. An indication of this time-scale is given in Figure 4.5.3 (right), where we present the posterior for the rate $\lambda_r(t)$ per future year t of swims by elite swimmers beating Peaty's record r. Here $\lambda_r(t) = s_t\lambda_u(t)[1+\xi(r-u)/\sigma_u]^{-1/\xi}$, where s_t is number of total swims per year by elite swimmers. The posterior mean and 95% HPDI are shown for $\lambda_r(t)$ over the window $2023-30$, assuming that the total number of swims per year remains the same as in 2019.

Model Diagnostics

Diagnostics for the marginal GPD element of our model are well-established, so here novel diagnostics that focus on the subject-specific characteristics of the data are presented. Rather than the latent scale as in Section 4.3.2, diagnostics are shown on the observed scale, so observations can be compared with predictive distributions for the associated swim-dates. Figure 4.5.4 shows the observations over time for six top swimmers, identified in Figure 4.5.5. All these swimmers have performances that are generally improving over time, and some have performances which are worse than the threshold, i.e., slower than the threshold. As such slow swims are treated as censored at the threshold, modelling these precise values is not of great importance, with the prime focus concerning swim-times better than the threshold.

A sample of size 400 was generated from the posterior predictive sample for each past date of a swim for each of these swimmers, i.e., ignoring in the simulation stage

the actual observation values by not conditioning on them. These distributions are shown on Figure 4.5.4 under-laying the corresponding observations. In a well-fitting model, each observation should appear to represent a sample from these simulated distributions. The model seems to have captured the improving career trajectory to date. The posterior predictive distributions indicate that the model fits well, as most observations are reasonably central to their associated distribution for all swims better than the threshold, and even for the swims not as good as the threshold. Maybe to be expected, Peaty's three best swim-times, each world records when achieved, are into the tails of their associated predictive distributions. For weaker swims, Martinenghi and Shymanovich have performances which are unexpectedly slow relative to what our model would anticipate.

Figure 4.5.4 also shows samples for these predictive distributions in the future, as the points from 2020-32, obtained under a stochastic model for the number and dates of future swims assuming that the swimmers continue to compete at current rates (see the supplementary material for details). As most of these future samples improve, or stay reasonably static, over time this illustrates that these swimmers are early in their careers. In contrast, for Peaty there is a clear decay of performances from 2024. To help see how these plots link to the earlier inferences, we also present on this figure the posterior mean and 95% HPDI for each swimmer's τ_i, which captures the period where the predictive samples seem to plateau. Section 4.5.4 uses these future subject-specific predictive samples to draw a range of inferences for future extreme events.

Subject-specific Predictions for Current Swimmers

Here we make predictive inference for future extreme events linked to specific swimmers, thus illustrating the novelty of inferences that are possible using our longitudinal extreme value model. Section 4.3.4 identified three groups ($\mathcal{I}^c, \mathcal{I}^f, \mathcal{I}^n$) of swimmers to consider when predicting future extreme events, and the supplementary material sets

out the Monte Carlo strategies for the evaluation of the corresponding posterior distributions. To avoid the extra assumptions that are required to study groups $\mathcal{I}^f$ and $\mathcal{I}^n$, only swimmers in $\mathcal{I}^c$ deemed actively competing at the start of the future period are studied, which we take to be those swimmers in $\mathcal{I}$ who have recordings in the most recent year of data. From our model and posterior predictive inference, standard extreme value properties, such as the distribution of the annual maxima, are simple to derive; however in sport, extreme events are mostly concerned with breaking records. We therefore focus on results for events linked to beating the current world record and setting PB times. Throughout, the future behaviour of swimmers is assumed consistent with the past data, so illness or sudden retirement are not accounted for, e.g., we ignore knowledge that Peaty has had some absences from the sport since 2021.

First consider the beating of the current world record. The joint posterior predictive distribution samples, illustrated in Figure 4.5.4, provide samples of future longitudinal data for the swimmers. The probability that the world record is beaten by a swimmer in $\mathcal{I}^c$ in the next 12 years has a posterior predictive probability of 0.53. This doesn't mean the record will be broken with this probability, as we do not consider swimmers in groups $\mathcal{I}^f$ or $\mathcal{I}^n$. Figure 4.5.5 (left) splits this probability up to show the posterior predictive probability for swimmer i beating the record, for the 10 most likely swimmers in $\mathcal{I}^c$. This gives a novel *ranking* method for swimmers within an event, as it focuses on the future potential of swimmers (through taking their stage of career trajectory into account) more than their past achievement (which is the exclusive focus of typical ranking methods). Perhaps unsurprisingly, Figure 4.5.5 (left) shows that Peaty is ranked the highest, i.e., the most likely to first beat his own world record of the swimmers in $\mathcal{I}^c$, with a predictive probability of 0.24. Martinenghi is ranked second, which is expected given findings in Figure 4.5.2 (middle), with a predictive probability of 0.092.

To assess how soon these swimmers can first beat the current record, Figure 4.5.5 (middle) shows the predictive distribution of the year in which a swimmer will be the

first of the current swimmers to beat the record. These posteriors are shown for the top six ranked swimmers in Figure 4.5.5 (left). These results show that if Peaty does break his record, it is most likely to happen within the next four years, primarily as a consequence of his age exceeding his peak age. In contrast, Martinenghi is most likely to beat the current record in 4-10 years. Figure 4.5.5 (right) shows the posterior distribution of the future PB time for each swimmer. These distributions show that there is a reasonable chance of each swimmer beating their current PB, though with Peaty less likely to do this than the other five swimmers shown, who all have a high posterior probability of beating their current PBs. This finding is not surprising, as swimmers that are currently near their peak have a limited chance of beating their PBs, and the younger swimmers have the largest chance of setting future new PBs as they are still improving.

4.6 Discussion

This article proposes the first analysis for extreme values of data arising from a longitudinal structure comprising multiple subjects, each with a time series of measurements. Although much new asymptotic theory remains to be developed, as the number of subjects and the lengths of their time series tend to infinity at potentially different rates, our focus has been in terms of putting down the framework for statistical modelling and associated inference. Furthermore, we have exhibited that this framework provides a basis for novel analysis of elite swimming data, and have illustrated the additional challenges that arise in practice, e.g.: non-stationarity over subjects, subjects with very limited data, and the need to model subjects not in the data. Our analyses also show how the models can be used for making important inference about future extreme events involving the current subjects.

This generic framework for longitudinal data analysis involving extreme values con-

tains a set of modelling decisions which are application specific. Core examples are the choice of functional forms for (i) the subject-specific mean function μ_i for all $i \in \mathcal{I}$, (ii) the threshold exceedance rate function λ_u, and (iii) the GPD scale parameter function σ_u. In our swimming application, fully parametric functional forms were established from prior application-specific knowledge. For the period of data we analysed, λ_u was modelled to be monotonically increasing, reflecting knowledge that the quality of swimmers has been improving generally in this period. However, if data prior to 2010 were used, a monotonic form would not be appropriate due to the phasing out of performance-enhancing full-body swim-suits, see Spearing et al. (2021) for estimates of how these suits enhanced elite swimming across events. In the absence of such knowledge we could use non-parametric approaches, e.g., as identified in Section 4.3.2.

As discussed in Section 4.5, for swimmers with very few observations, a tactical decision must be made between including them all at a high computational inefficiency, or discarding them from the analysis, at the cost of bias. Although we developed a pragmatic compromise in the selecting of which swimmers to include, here we briefly outline other possibilities. The most naive approach would be to incorporate all swimmers by pooling together all data (measured swim-times and the corresponding age of swimmer on the date of the swim) by the swimmers with less that m swims, so that they share a common (α_i, τ_i) pair. Although this would avoid much added computational complexity by constraining the additional parameters and it would use all data for the GPD inference, it risks bias as the swimmers being pooled together may be very different in their skills (attributes). Alternatively, a more specific clustering approach could be developed in which each of the subjects with less than m observations is pooled with a subject with more than m observations, so that they share the same (α_i, τ_i). This approach could be more effective if a suitable cluster metric could be developed that draws parallels with our clustering of intranstivities in Spearing et al. (2023).

The initial reasoning for parameter reduction was the computational intensity of the

model fit, where the bottleneck comes from inverting the Gaussian mixture distribution G_Z. Improvements to the current grid search approach (4.4.3) used to approximate G_Z^{-1} will allow the model to be scaled to large data sets. One possibility is to approximate the derivatives of G_Z^{-1} with respect to each element of $(z, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\nu})$ to enable more efficient gradient-based samplers (Duane et al., 1987). However, due to approximating such a large number of gradients, there is still no guarantee that overall inference speed will be improved.

An entirely novel aspect of our inference has been the subject-specific features such as the marginal Gaussian distribution and the variation across subjects is modelled through the different subject-specific attributes $\{\alpha_i : i \in \mathcal{I}\}$. Although Gaussian marginals are leveraged on the grounds of the parsimony of conditional and unconditional Gaussian processes, this choice is rather unimportant to the outcomes of the inference. This is due to the weak common prior across attributes, resulting in a posterior which is driven by the data. The resulting posterior for a new subject's α_i is a Gaussian mixture model; where it is recognised that this reflects only subjects capable of achieving measurements above a high threshold, and is not applicable to the population as a whole. Despite this restriction to the extreme subjects, our swimming analysis shows that the variation between attributes for swimmers is substantially larger than natural variation of extreme times for any selected swimmer. Hence the analysis has disentangled the variations of the longitudinal data to better inform future inference for extremes and records both unconditionally and conditionally for the current field of elite swimmers.

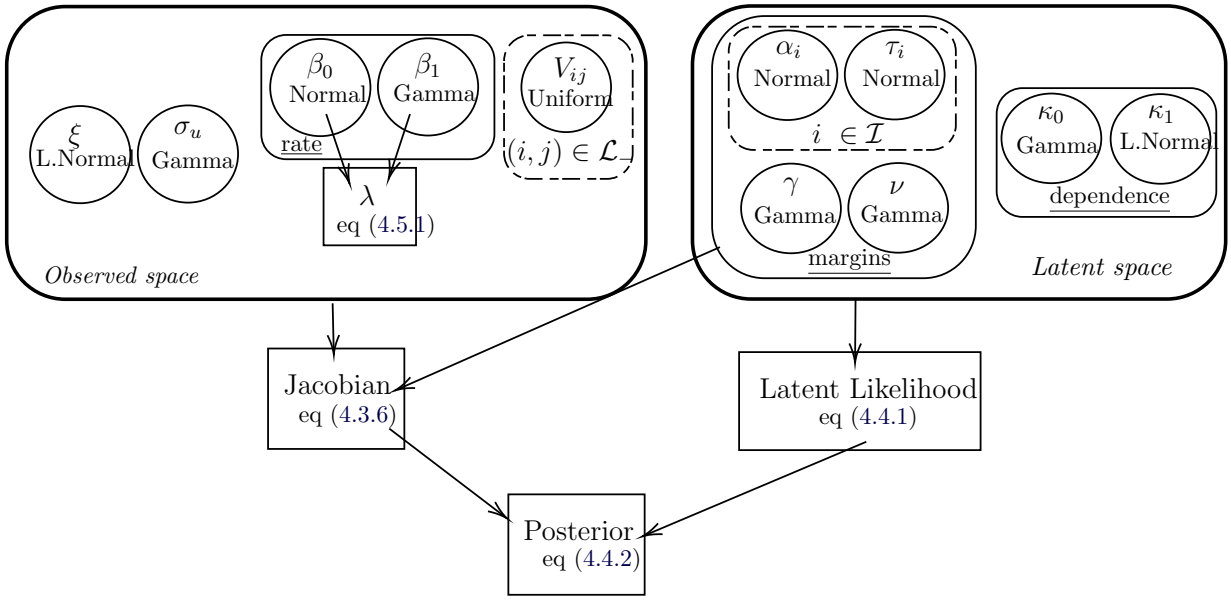


Figure 4.5.1: DAG illustrating the model flow with associated priors. The *observed space* (left) shows the parameters for the extreme margins: (ξ, σ_u), the GPD parameters; (β_0, β_1), of the rate function λ_u for exceeding the threshold u; and the auxiliary variables $V_{i,j} : (i,j) \in \mathcal{L}_-$ corresponding to the censored observations below u. The parameters of the *latent space* (right), $(\boldsymbol{\theta} := \{\boldsymbol{\theta}_i := (\alpha_i, \tau_i) : i \in \mathcal{I}\}, \gamma, \nu)$, determining the marginal distribution of the Gaussian mixture, and the kernel parameters $\boldsymbol{\kappa} := (\kappa_0, \kappa_1)$ which dictate the dependence structure. Both the observed space and latent space parameters determine the Jacobian (4.3.6), whilst the only the latent space parameters determine the latent-likelihood (4.4.1). The posterior distribution then contains the Jacobian, the latent-likelihood, and the prior distributions (4.5.2).

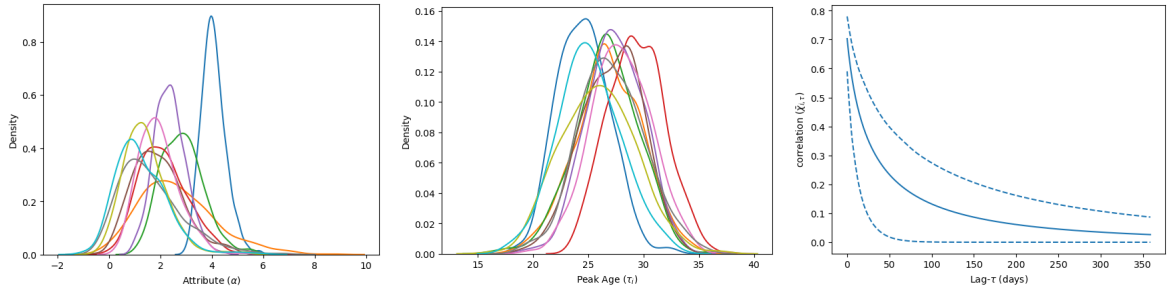


Figure 4.5.2: Posterior inferences for subject-specific features of the model. For the top 10 swimmers as defined in Section 4.5.4 (which correlates strongly with those swimmers with the largest posterior mean values of α_i over $i \in \mathcal{I}$), the posterior distribution of these swimmers' attributes α_i (left) and peak ages τ_i (middle) is shown. The line colours in these plots identifies the different swimmers, with the color coding being explained in Figure 4.5.5 (left). The right panel shows the mean posterior and 95% HPDI for the subject-specific asymptotic independence measure $\bar{\chi}_{i,\tau}$ against time lag τ in days (right).

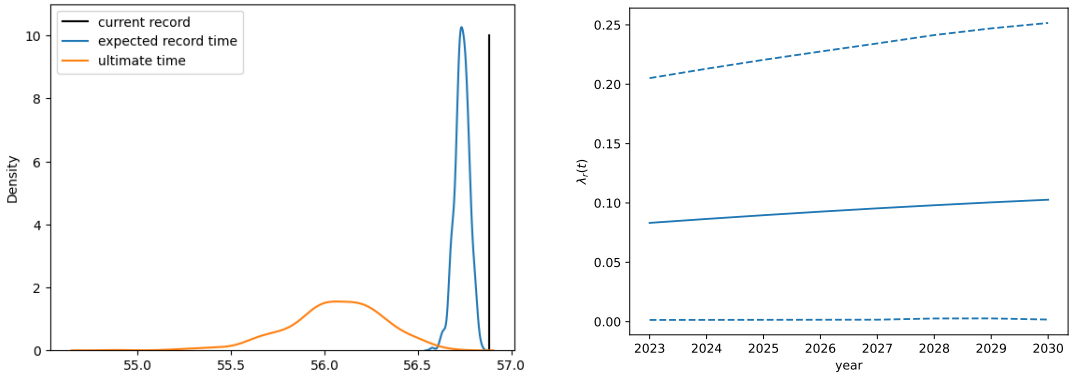


Figure 4.5.3: Mens' 100m breaststroke inference. The current record time in seconds (left) (shown by a black vertical line) is held by Peaty at the time of this analysis. The posterior distributions for expected next record swim-time (blue) and ultimate swim-time for this event (orange). Future predictions (right) show the posterior mean (solid line) and 95% HPDI (dashed lines) of the rate $\lambda_r(t)$ of swims by elite swimmers of beating Peaty's current record in year t.

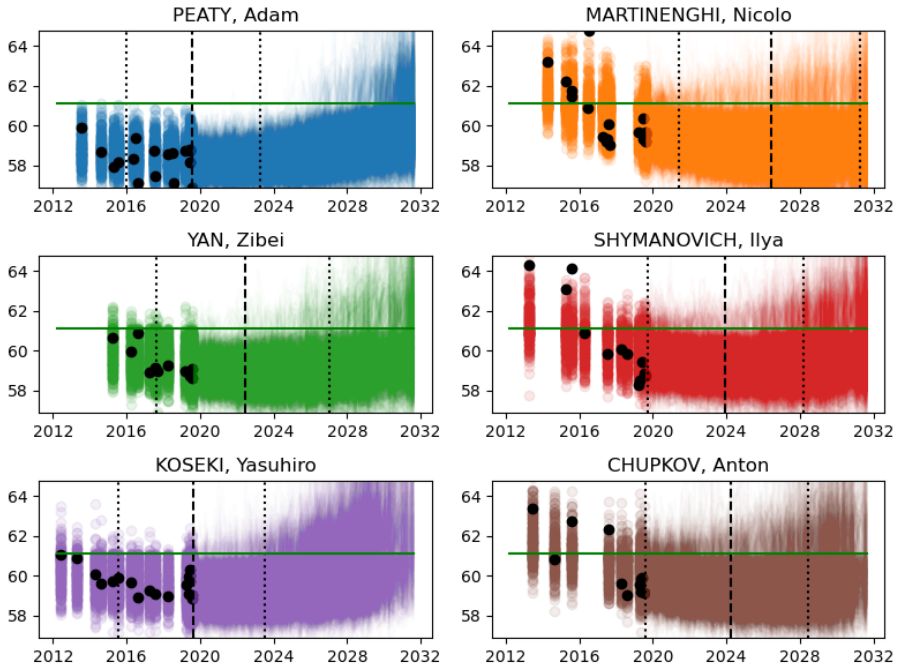


Figure 4.5.4: Within-subject diagnostics for six top swimmers: observed swim-dates and swim performance in seconds (shown as black dots); samples from the posterior predictive distributions (coloured points) for these swimmers for the dates of their swims in the past, and for simulated swim times in the future. The threshold u is shown by the horizontal line and the posterior mean and 95% HPDIs for the peak age τ_i are shown by vertical lines.

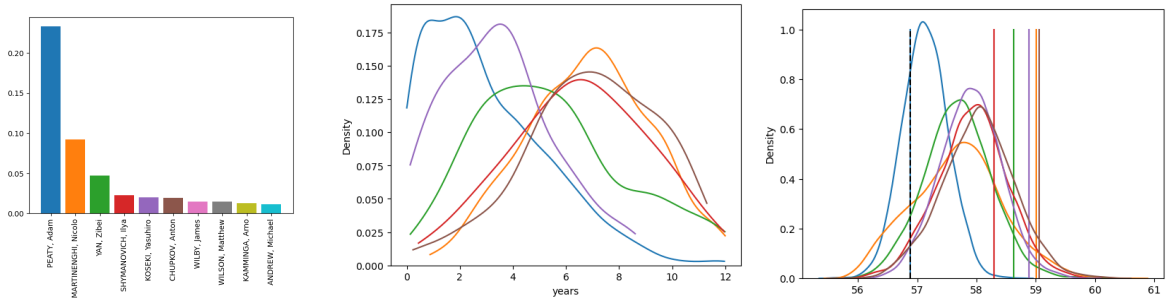


Figure 4.5.5: Inference for individual swimmers: probability that each swimmer will be the next swimmer in $\mathcal{I}^c$ to beat the current world record (left) for the 10 most likely; the posterior distributions for each swimmer for the time at which they are the first the swimmers in $\mathcal{I}^c$ to beat the current record (middle); and the posterior distributions of the expected personal-best of all future times, with the vertical lines showing their current PBs (right). The swimmers shown in middle and right panels are the six top swimmers in the left panel. The colours on the middle and right plots identifies the swimmers, with colors identified in the left panel.

Chapter 5

Conclusions and Further Work

5.1 Conclusions

The literature review in Chapter 2 fuses together the backbones of univariate extreme value theory: the block maxima approach; the peaks-over-threshold approach; and the extremal point process, which unifies the first two within a common framework. But extreme value theory is a deeply researched area of statistics, of which Chapter 2 only scrapes the surface. Subsequent methodological developments expand into the multivariate domain (Barnett, 1976; Heffernan and Tawn, 2004; Coles and Tawn, 1994), spatial frameworks (Ribatet, 2013), and can even model both spatial and temporal aspects (Simpson and Wadsworth, 2021), and in combination with machine learning techniques (Farkas et al., 2021).

In Chapter 3, the point process methodology is used to model data of elite swimmers. These data display a smooth increase in the rate of swim performances which exceed the extremal threshold, reflecting improvements in sports science, such as nutrition and training methodology. This exceedance rate also shows step changes, correlating with the introduction of specific technologies such as full-body swim-suits. The data also suggests that the conditional distribution of swim-times, given this threshold is

exceeded, is not evolving over time. A novel parametrisation is then introduced which allows for exactly this: an exceedance rate which is time-dependent, with a conditional distribution above the threshold which is identically distributed over time.

The unified model pooled data over all 34 individual events, which helped reduce parameter uncertainty. Previous attempts to pool data across events use distance as a covariate, for example in athletics (Stephenson and Tawn, 2013), but this would introduce bias in swimming, since swim-times in, say freestyle, are generally faster than in, say, breaststroke, over the same distance. Using distance as a covariate also induces bias if pooling across gender, as athletes in the men's category tend to record faster times than those in the women's category in all but longer distance events (Bam et al., 1997). This was resolved by using the threshold itself in each event as a covariate in the pooled model, a technique which is novel in extreme value modelling in general, and which can adjust the swim-time for the distance, gender category, or stroke, accordingly. Section 5.2 discusses possibilities for further applications of this.

Chapter 4 developed a first attempt at combining the areas of extreme value theory and longitudinal data analysis. Here, dependence structures that are specific to longitudinal analysis are explored, and the resulting forms of extremal dependence that arise. Unlike the majority of multivariate extreme value analysis the model developed in Chapter 4 has the flexibility to capture both asymptotic dependence and asymptotic independence, with this being determined by the data.

5.2 Further Work

Both Chapters 3 and 4 negate the swim-time data in order to use the more commonly utilised methodology for the upper-tail of a distribution. Alternatively, Gomes and Henriques-Rodrigues (2019) invert the swim-time and apply peaks-above-threshold methodology directly to swim-*speed*, since a smaller swim-time equates to a larger swim-

speed. However, Wadsworth et al. (2010) show that such non-linear transformations lead to differing conclusions, and so conclusions drawn from swim-time analysis are likely incompatible with those drawn from swim-speed analysis, with no intuitive way of discerning which conclusions are most reliable. Analysing a Box-Cox transformation of the data, and estimating this transformation parameter jointly as part of the modelling procedure, would remove the guess-work from the choice of data scale. Moreover, it allows for the uncertainty in the scale of the data to be accounted for and to be quantifiable. It could also be interesting to see if the Box–Cox parameter changes in a systematic way over distance or stroke.

The use of the extremal threshold as a covariate in Chapter 3 is a novel approach in modelling of extreme values, and helps pool data across dimensions with no obvious physical links (in the swimming examples, across strokes and gender categories.) In fact, the only constraint is that the data can be assumed independent across the pooled dimension, as is done in Chapter 3 between swim-times for different strokes, gender categories and distances. This simplifying assumption may not be true when the same swimmer competes across different distances, for example, meaning that the uncertainty of the estimates would be underestimated. Multivariate techniques could be employed in order to capture some of the correlation between data points resulting from the same swimmer in different competitions (Adam and Tawn, 2012). Outside of swimming, this approach could even be used to pool across different sports, for example, elite marathon running and pole-vault.

The constant evolution of the para-swimming classification system is testament to the challenge of creating fair competition in disability sports, with athletes currently grouped into discrete categories based on their disability type, and which inevitably leads to some athletes having an unfair advantage or disadvantage within their category. The number of classifications itself is open to debate, with too many classifications resulting in too few swimmers in each classification and therefore a drop in compet-

itiveness, and too few classifications resulting in large differences between swimmers' disability types within the same group. Of course, this problem stems from the over-simplification and discretisation of a variable, in this case disability, which in reality exists on a continuous spectrum. However, a pooled model of the type presented in Chapter 3 would allow for a continuous 'classification variable' which pools across disability, to allow fairer competition over all disability types and comparison between disabilities. In junior swimming, because of the discretisation of age groups, some swimmers can be almost a whole year younger than others in the same competition, which creates an unfair disadvantage. The same idea of a continuous scale for age groups would allow for fair comparison of 'age-adjusted' swim times.

The inclusion of transgender (and in particular MtF) athletes has led to controversy, with the regulations being changed again for the most recent Olympic Games. A large part of the change in regulation comes from a gradual shift away from irrelevant and outright transphobic policies, such as the "The Stockholm Consensus" (Committee et al., 2003), which requires transgender athletes to have completed "anatomic changes consistent with their professed gender" (Genel, 2017). Despite these changes there is still little indication of an overall increase in the acceptance of gender variance within sport (Sykes, 2006). Due to the sex assigned at birth, allowing such athletes to compete in the women's category has the potential for an unfair advantage, e.g., this sometimes leads to higher levels of testosterone; however, denying access to sport only increases the isolation and social stigmatisation often experienced by the transgender community (Cromwell, 1999). Rather than these current binary over-simplifications however, our covariate could allow for sports to reflect the underlying continuous and fluid definition of gender. In the same vein, cases of unusual testosterone levels can easily be dealt with. Ultimately, it is possible to have a global model which fairly compares swimmers across gender and of different disabilities, and even junior swimmers, across different events.

Of course, the overall goal in adjusting for disability, age or gender is to promote perceived equitable competition, but we know these are but three variables of many that impact fairness in sport. For example, having to overcome discrimination due to 'race' can negatively impact the likelihood of succeeding in a given sport (Hylton, 2008). In this way, to fully level the playing field in sports requires a continuous description of the complexity and variety in social inequality more broadly.

Part I Appendices

Appendix A

Spline Construction

Let $B_k^d(x)$ be the value of the k^{th} d degree B-spline basis function at a point x, where $k = \{1, \ldots, q\}$, $q \in \mathbb{Z}^+$, and x_k denotes the k^{th} knot, such that $B_k^d(x)$ is strictly positive within the region $x_k < x < x_{k+d}$. The exact form of the splines can be formed recursively from 0 degree basis splines. Note that 0 degree splines are trivial to form, described as step-functions over the region of each knot such that

$$B_k^0(x) = \begin{cases} 1, & x_k \leq x < x_{k+1}, \\ 0, & \text{otherwise.} \end{cases}$$

Then, using the formula (De Boor, 1978) for $d \geq 1$

$$B_k^{d+1}(x) = \frac{x - x_{k-(d+1)}}{x_{k-1} - x_{k-(d+1)}} B_k^d(x) + \frac{x_k - x}{x_k - x_{k-d}} B_{k+1}^d(x),$$

higher degree B-splines are formed. Figure A.0.1 shows splines of degrees $d = 1, 2, 3$. It can be seen that as the degree of the basis function increases, the function becomes smoother and has a larger range. The spline function $Y(x)$ is then constructed as

$$Y(x) = \sum_{k=1}^{q} a_k B_k^d(x)$$

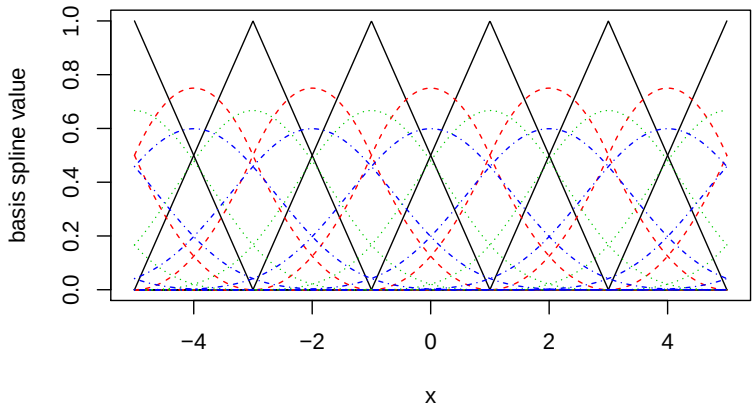


Figure A.0.1: Basis spline functions $B_k^d(x)$ with degree d: degree 1 (black solid), 2 (red dashed), 3 (green dotted), and 4 (blue dot-dashed), and knots are spaced at integer values.

where a_k is the k^{th} B-spline coefficient, and $\boldsymbol{a} = \{a_i : i = 1, \ldots, q\}$ is the coefficient vector. Generally, q is chosen to be large, such that the fitted curve shows more variation than can be justified by the data. To reduce this variation, a penalty on the finite differences of adjacent coefficients of Eilers and Marx (1996) is used. The penalty is governed by $\phi \boldsymbol{a}' P \boldsymbol{a}$, where $P \in \mathbb{R}^{q \times q}$ is the penalty matrix, and $\phi > 0$ determines the amount of penalisation. The choice of P is based on some prior belief of the shape of the data. The penalty matrix used was a second order, such that

$$P = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & \dots & 0 & 0 & 0 \\ -2 & 5 & -4 & 1 & 0 & \dots & 0 & 0 & 0 \\ 1 & -4 & 6 & -4 & 1 & \dots & 0 & 0 & 0 \\ 0 & 1 & -4 & 6 & -4 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 & -4 & \dots & -4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & \dots & 6 & -4 & 1 \\ 0 & 0 & 0 & 0 & 0 & \dots & -4 & 5 & -2 \\ 0 & 0 & 0 & 0 & 0 & \dots & 1 & -2 & 1 \end{bmatrix},$$

which penalises a large second derivative, thus penalising fits that depart from linearity.

Appendix B

Supplementary Material for Chapter 4

This document accompanies Chapter 4. Section B.1 includes further investigations into the nature of the extremal dependence of the scenarios derived in Section 4.2.3. Section B.2 shows analytical results for probabilities of future extreme events in longitudinal data using the model of Section 4.5.2, under some simplifying assumptions. In reality, many applications will require to full flexibility of our novel model, as seen in Section 4.5, and in this case Monte Carlo simulation provides computational solutions. A strategy for this is set out in Section B.3.

B.1 Further limit results for studying extremal dependence of longitudinal data

Building on the results from Section 4.2.3, here we explore further the nature of extremal dependence in longitudinal data. To help better understand the asymptotic dependence case we consider a version of measure $\chi_\tau^{(M)}$ which allows both n and the quantile to grow in combination. Specifically, consider the conditional probability $\Pr(M_{n2} > x_n \mid$

$M_{n1} > x_n$), where $x_n \to \infty$ and letting $\alpha_n = x_n - \delta$ for some constant δ. The marginal probability is then

$$\Pr(M_{n1} > x_n) = 1 - \Pr(M_{n1} < x_n) = 1 - [\Phi(x_n)]^{n-1} \Phi(x_n - \mu_n).$$

Now consider the joint probability

$$\begin{aligned}\Pr(M_{n1} > x_n, M_{n2} > x_n) &= 1 - \Pr(M_{n1} < x_n) - \Pr(M_{n1} < x_n) + \Pr(M_{n1} < x_n, M_{n2} < x_n) \\ &= 1 - 2\,[\Phi(x_n)]^{n-1} \Phi(x_n - \alpha_n) + [\Phi(x_n)]^{n-1} \Phi_2(x_n - \alpha_n, x_n - \alpha_n; \rho).\end{aligned}$$

Then, in case (i), consider setting α_n as above with x_n, this gives the limit

$$\Pr(M_{n2} > x_n \mid M_{n1} > x_n) \to \frac{1 - 2\Phi(\delta) + \Phi_2(\delta, \delta; \rho)}{1 - \Phi(\delta)}.$$

The above limit is non-zero for all finite δ and when $\rho = 0$ this limit is $1 - \Phi(\delta)$, which is positive for all $\delta < \infty$. So, when $\rho = 0$, despite the independence of within-subject observations, the longitudinal structure induces asymptotic dependence. This is different from the findings for $\rho = 0$ in limit (4.2.9), showing the limits that give identical findings about the form of extremal dependence for identically distributed variables can give contrary results for longitudinal data. For case (ii) we have that $\Pr(M_{n2} > x_n \mid M_{n1} > x_n) \to 0$, i.e. asymptotic independence.

Underlying all these limiting results is the fact that subject n will be the componentwise maximum with probability 1 in case (i) and 0 in case (ii) for how α_n grows. This is shown through the following limit, which for case (i) explores the probability

that the same subject gives a large measurement value at each time point, i.e.,

$$
\begin{aligned}
&\Pr(X_{n1} = M_{n1}, X_{n2} = M_{n2} \mid M_{n1} > \alpha_n) \\
&= \Pr\{\max(X_{11}, \ldots, X_{(n-1)1}) < X_{n1}, X_{n1} > \alpha_n, \max(X_{12}, \ldots, X_{(n-1)2}) < X_{n2} \mid M_{n1} > \alpha_n\} \\
&= \int_{y=-\infty}^{\infty} \int_{x=0}^{\infty} \Pr\{\max(X_{11}, \ldots, X_{(n-1)1}) < \alpha_n + x, \max(X_{12}, \ldots, X_{(n-1)2}) < \alpha_n + y \\
&\qquad\qquad \mid X_{n1} = x, X_{n2} = y\} \phi_2(x, y; \rho)\, \mathrm{d}x\, \mathrm{d}y / \Pr(M_{n1} > \alpha_n) \\
&= \int_{y=-\infty}^{\infty} \int_{x=0}^{\infty} [\Phi(\alpha_n + x)\Phi(\alpha_n + y)]^{(n-1)} \phi_2(x, y; \rho)\, \mathrm{d}x\, \mathrm{d}y / \Pr(M_{n1} > \alpha_n) \\
&\to 2 \int_{y=-\infty}^{\infty} \int_{x=0}^{\infty} \phi_2(x, y; \rho)\, \mathrm{d}x\, \mathrm{d}y = 1
\end{aligned}
$$

as $n \to \infty$, as the powered terms tend to 1, as in limit (4.2.8), and that limit with $x = 0$ explains the denominator tending to $1/2$, and the double integral is $1/2$ due to symmetry of the standard bivariate normal density about $x = 0$. Similarly, for case (ii) this limit is 0.

B.2 Evaluation of probabilities of future extreme events for longitudinal data

A benefit of accounting for the longitudinal structure is that now inference and predictions of extreme events regarding individual subjects is ascertainable, e.g., the probability that a new record is achieved by a particular subject $i \in \mathcal{I}$. To make such inferences, each subject's mean function over time is incorporated, as well as the temporal dependence around this. Both of these aspects are described by the Gaussian process model of Section 4.3.2, which gives analytical solutions to probabilities of future events through its closed form conditional distributions. Let $\mathcal{J}_i^F$ be the set of future measurements for subject i, with the future schedule of measurement points defined as $\{t_{i,j} : j \in \mathcal{J}_i^F, i \in \mathcal{I}^F\}$, where all such $t_{i,j} > t_{\max}$ for a current time $t_{\max}$.

In practice the evaluation of the probabilities of such complex events are most simply conducted through Monte Carlo methods, simulating over different realisations of the longitudinal process for the fitted model, with evaluation achieved empirically over a large sample of replicates. We present results and various assumptions of this type in Section B.3, but here we derive the analytical expression for one such event under an idealised set-up to illustrate the complexity even in this simplified scenario.

Consider the event $A_i^F(r)$, corresponding to the subject $i \in \mathcal{I}$ breaking the record for the maximum measurement in some future time period identified by F and holding that record at the end of period, given that the current maximum measurement is r. Consider the case where (i) all parameters of the model in the latent space are known; (ii) no subjects outside $\mathcal{I}$ produce measurements in time period F; (iii) the observed subjects have a constant mean function over time, i.e., $\mu_i(t) = \alpha_i$ in expression (4.3.3); and (iv) that there is subject-conditional independence for each subject. A benefit of assumption (iv) is that it removes the need to consider the history of each subject's measurements including which subject holds the current record.

To derive $P(A_i^F)$ it is most easy to work in the latent space, recognising that the current record transforms to the value $r_Z := G_Z^{-1}\left[F_Z\left(r, t_r\right)\right]$ in the latent space. First define $M_i^F := \max\left(\{Z_{i,j} : j \in \mathcal{J}_i^F\}\right)$, the maximum measurement for subject i in the future time period, then this distribution has the survivor function of $P(M_i^F > z) = 1 - [\Phi(z; \alpha_i, \nu_i)]^{|\mathcal{J}_i^F|}$, given assumptions (iii) and (iv). Also let $M_{-i}^F = \max\left(\{Z_{kj} : j \in \mathcal{J}_k^F, k \in \mathcal{I} \setminus \{i\}\}\right)$ be the maximum of all other subjects' measurements in this future period. Then $P(A_i)$ is given as

$$\begin{aligned} P(A_i) =& P\{M_i^F > \max[r_Z, M_{-i}^F]\} = P\{M_i^F > r_Z > M_{-i}^F]\} + P\{M_i^F > M_{-i}^F > r_Z\} \\ =& P(M_i^F > r_Z) \left(\prod_{k \in \mathcal{I} \setminus \{i\}} \Phi(r_Z; \mu_k, \nu_k)^{|\mathcal{J}_k|} \right) + \int_{r_Z}^{\infty} P(M_i^F > z) f_{M_{-i}^F}(z)\, \mathrm{d}z, \end{aligned} \tag{B.2.1}$$

where

$$f_{M^F_{-i}}(z) = \left(\prod_{h \in \mathcal{I} \setminus \{i\}} \Phi(z; \mu_h, \nu_h)^{|\mathcal{J}_h|} \right) \sum_{k \in \mathcal{I} \setminus \{i\}} |\mathcal{J}_k| \frac{\phi(z; \mu_k, \nu_k)}{\Phi(z; \mu_k, \nu_k)}.$$

B.3 Adapting predictions for new subjects

For making inferences about the future behaviour of extreme values for longitudinal data there are a number of substantial challenges linked to the subject-specific characteristics of the data structure. Analytical results such as result (B.2.1) are available in simple cases, but with the mean functions inducing non-identically-distributed variables, it must be recognised that, in the longer-term, the extreme events are more likely to be due to subjects not yet observed in $\mathcal{I}$. In the short-term however, these future extreme events are most likely to be obtained by the current subjects in $\mathcal{I}$, followed by a transitional medium-term in which extremes arise from a mixture of these populations of subjects. Here we develop the outline of a framework for such inferences, setting out some possible choices that need to be made in relation to the currently unobserved subjects. The model parameters here are treated as known, and Section 4.4 presents how to account for that additional uncertainty.

For the observed data there are n subjects, indexed $\mathcal{I}$, each with at least one measurement above the threshold u. Going forward beyond the observed time-frame, there are then three types of subject to consider: (i) those subjects in $\mathcal{I}$, indexed by $\mathcal{I}^c$ with $\mathcal{I}^c \subseteq \mathcal{I}$, which are still producing at least one measurement above u in the future time window; (ii) those subjects $\mathcal{I}^f$, which produced measurements exclusively below the threshold within the observed time-frame and so $\{\mathcal{I}^f \cap \mathcal{I}\} = \emptyset$, but in the future produce a measurement above u; and (iii) those subjects $\mathcal{I}^n$ with no recordings at all within the observed time-frame but which in the future period produce at least one measurement above u. To help remember the terminology the superscripts here denote c for *current* subjects with a future threshold exceedance, f for subjects in the

population which are active in the observed time-frame and which record their *first* exceedance of u in the future time period, and n for an entirely *new* subject which records an exceedance of u in the future time period.

For each subject in each of the groups $\mathcal{I}^c, \mathcal{I}^f$ and $\mathcal{I}^n$ measurement series are simulated over a time window of $(t_{\max}, t_{\max} + T)$ where $t_{\max}$ is the maximum time in the observed database and T is the length of the future period of interest. As membership of these three groups depends on a subject achieving a measurement larger than u in the future time-period, the number in each group is random. In practice it is easiest to first generate a time series for each individual that could be in the three groups and then a random number of these will meet the criteria to be in the respective groups. For groups $\mathcal{I}^c$ and $\mathcal{I}^f$ the maximum number of potential subjects there could be is known from the observed numbers in the database, but in practice, computational time is saved by omitting previously measured subjects which have no measurements in the latter part of the observation window. That is, $t_{\max} - t_{i,n_i}$ being sufficiently large suggests that subject i has stopped generating measurements that have potential to be extreme. In contrast, for $\mathcal{I}^n$ assumptions must be made about the arrival rate of new potential subjects. We propose that the rate of first measurements per subject in the database per unit time-period, denoted by r_{data}, is used to estimate this rate. Then, the number of potential new subjects for the future is generated by a Poisson(Tr_{data}) random variable.

For each of the potential subjects in the three groups, the number and the times of the future measurements in $(t_{\max}, t_{\max} + T)$ and simulated realisations of the associated measurement $X_{i,j}$ are generated according to the models in Sections 4.3.1 and 4.3.2 for these times-points, conditional on any information already present about these subjects. The subject is then identified as being from a group if their maximum measurement exceeds u. These steps are discussed below, identifying the features that change across the three groups.

For each potential subject in any of the three groups, measurement time points are generated independently over subjects from a homogeneous Poisson process with rate ω_i per unit time for subject i. That is, a subject i has N_i future observations, with $N_i \sim \text{Poisson}(T\omega_i)$, with these measurement time-points uniformly distributed on $(t_{\max}, t_{\max} + T)$. The times for the future measurements for subject i are denoted by $\boldsymbol{t}^*_{i,j} := \{t^*_{i,j} : t_{\max} < t^*_{i,j} < t_{\max} + T, i = 1, \dots n^*_i\}$ where n^*_i is the realisation of N_i. For a potential subject $i \in \{\mathcal{I}^c \cup \mathcal{I}^f\}$, an estimate of ω_i is based on the empirical rate of measurements up to $t_{\max}$ for the subject in the database. For each potential subject $i \in \mathcal{I}^n$, ω_i is estimated from the observed population of subjects $\mathcal{I}$. Specifically, a subject j is randomly drawn from $\mathcal{I}$ with associated rate ω_j, and then we take $\omega_i \sim \log N((\log(\omega_j) - \psi^2/2, \psi^2)$. This choice ensures that the expected value of ω_i is an existing rate ω_j, and where the choice of ψ can be selected based on how representative the subjects in $\mathcal{I}$ are believed to be relative to the entire population. So, ψ can be taken larger if an under-representation of $\mathcal{I}$ is anticipated.

Next, the measurement values for each potential subject are simulated in the latent space, given the simulated future measurement times. For each potential subject i, measurements are simulated from the Gaussian process $Z_i(t) \sim \mathcal{GP}\{\mu(t; \boldsymbol{\theta}_i, \boldsymbol{\gamma}), K_{\boldsymbol{\kappa}}(\cdot, \cdot)\}$, at time-points $\boldsymbol{t}^*_i = (t^*_{i,1}, \dots, t^*_{i,n^*_i})$. These simulated processes are generated conditionally on the previous data when appropriate for the group, see below. For future realisations in the upper tail of the latent variable space we can transform back to the observed space using transformation (4.3.5). Only those potential subjects with their maximum measurement in the time interval $(t_{\max}, t_{\max} + T)$ exceeding u are included as a subject in their respective group. For deriving future scenarios we are only interested in those simulated measurements above u in original space.

We have different existing knowledge at time $t_{\max}$ for each subject depending on which of the three groups they are from, in the form of past measurement values, covariates, and information about $\boldsymbol{\theta}_i$. For a potential subject $i \in \mathcal{I}^c$, the posterior

distribution of $\boldsymbol{\theta}_i$ and the subject's covariates that determine how μ_i varies with t are available. That potential subject's Gaussian process is the simulated given the past values of $(Z_i(t_{i,1}), \ldots, Z_i(t_{i,n_i}))$. Although some of these past $Z_i(t)$ values are non-extreme in the original space, i.e., the associated $X_{i,j} < u$, our inference methods of Section 4.4 provide estimates for all of these values from which to condition on for each of the generated posterior samples for the model parameters.

Now consider a potential subject $i \in \mathcal{I}^f$. Although past observational data are available for them, as of time $t_{\max}$ these data are not included in inference, and so no estimates of subject-specific parameters $\boldsymbol{\theta}_i$ are available. Likewise for any potential subject in group $\mathcal{I}^n$. For both cases $\boldsymbol{\theta}_i$ can be drawn from the joint posterior from a randomly selected subject in $\mathcal{I}$. In both cases the Gaussian process is simulated forward from $t_{\max}$ independent of any past measurement data information, so for potential subjects in $\mathcal{I}^f$ the past data is ignored. To be able to use the Gaussian process, the relevant covariates for the potential subject are required. For a potential subject $i \in \mathcal{I}^f$ their actual covariates are used, whereas for $i \in \mathcal{I}^n$ the covariates at drawn randomly from a subject in the database (not just from subjects in $\mathcal{I}$).

Part II

Relative Systems

There's always a bigger fish

— Qui-Gon Jinn

Chapter 6

Ranking Via Paired Comparison

Favourite desserts, difficulty of yoga positions, harms of drugs (Nutt et al., 2010), "the most livable neighborhoods in New York" (Silver, 2010) - anything can be ranked. But in general, for relative systems it is easier to express a preference between a pair of objects than to rank the whole set. Optometrists, for example, often use paired comparison techniques ("better or worse?") in order to find the best prescriptions for prescription glasses (Olkin et al., 2015). Relative systems exist outside of the pairwise comparison domain, such as three-way chess, where three players compete simultaneously on the same board. But in general the literature on relative systems is dominated by paired comparison methodology, which uses knowledge of pairwise preferences to infer a global rank.

The rankings are then informative of the outcomes of future comparisons. For example, Boulier and Stekler (1999) use rankings from the Association of Tennis Professionals (ATP) and US collegiate basketball for predicting future outcomes via a generalised linear model, such that the probability of person/team a beating another b is given via the probit model

$$p_{ab} = \int_{-\infty}^{\lambda\delta_{ab}} \phi(t)\,\mathrm{d}t,$$

where ϕ is the standard Gaussian probability density function (PDF), $\lambda \in \mathbb{R}$ is an

estimateable parameter, and $\delta_{ab} \in \mathbb{Z} := R_a - R_b$ is the difference in ranks where $R_a, R_b \in \mathbb{Z}_+$ are the ranks of a and b respectively. Note that the better object has a lower rank (with the best object having rank one), and so λ is typically negative. Here the raw difference in the ranks in each pairwise match-up are used; however, the difference in rankings is in general not linear in terms of ability: the difference in ability between objects ranked one and two tends to be larger than the ability difference of objects ranked 1001 and 1002 (Lebovic and Sigelman, 2001). Consequently, the probability of preference between two objects is not a linear function of the difference of their ranks, as is assumed in the above model. Therefore, Klaassen and Magnus (2003) use transformed ranks to predict outcomes in ATP tennis matches, so that if a player a has the official ATP ranking R_a, then the transformed rank is given as $\tilde{R}_a := 8 - \log_2(R_a)$. Then, the probability a beats b, p_{ab}, is modelled again using a generalised linear model. Specifically, they use logistic regression, modelling

$$p_{ab} = \frac{\exp\left[\lambda\left(\tilde{R}_a - \tilde{R}_b\right)\right]}{1 + \exp\left[\lambda\left(\tilde{R}_a - \tilde{R}_b\right)\right]}.$$

Already, then, there's a whiff that the rankings are obscuring a more fundamental description of object preference. It is generally more accurate to ascribe to each object (or in the aforementioned example, player) some true *rating*, which may be latent. The rankings are then simply the upshot of ordering these more meaningful ratings. By only using the rankings, we lose all information about the separation between the objects. Accurate forecasting and meaningful inference, therefore, tend to use the objects' ratings rather than their ranks. A large part of the paired comparison literature centres on uncovering these ratings.

But there is more to a system than the objects and their ratings - the probability of an outcome can be affected by the state of the system itself. In wine-tasting, for example, preference between two wines may be inverted due to a change in state: we

could imagine a chilled white wine to be preferred to a red on a mid-summer's day, even if the red's rating is objectively higher. In sports, the state of the system is regularly altered via the *home-advantage* effect, which is commonly found over a range of sports (Schwartz and Barsky, 1977). Equally, a change in weather or choice of referee changes the system's state - or any exogenous factor which is external to the objects' ratings (Glickman and Stern, 2017).

The order in which objects are compared can also influence the outcomes of paired comparison. For example, the order that wines are presented to a wine-taster confounds the order of preference (Scheffé, 1952). Similarly, in sports, *tournament structure* influences the rankings. This is further complicated because the objects' strengths, or ratings, may not be stationary. Therefore tournament structure can cause both a change in state, i.e., it can be an exogenous factor, and a change in ratings for a given comparison, i.e., an endogenous factor. The time, and therefore order, of comparison is crucial, and some *dynamic ranking systems* aim to capture this. From the bountiful paired comparison approaches, the most well-known are highlighted below.

6.1 Statistical approaches

Amongst the most established methods are the Bradley-Terry (Bradley and Terry, 1952) type models. These are statistical paired comparison models which infer a rating μ_i for each object i amongst a set of objects $\mathcal{I}$ based on the outcomes of paired comparisons between the objects. The basis of these models is that the probability p_{ij} of $i \in \mathcal{I}$ beating another object $j \neq i \in \mathcal{I}$ is a monotonically increasing function of the difference of the objects' ratings, such that $p_{ij} = f(\mu_i - \mu_j)$, where $f : \mathbb{R} \to [0, 1]$, $f(x') > f(x)\ \forall x' > x$. If f is chosen to be the logistic function i.e.,

$$f(x) = \frac{1}{1 + \exp(-x)},\ x \in \mathcal{R},$$

then this is the Bradley-Terry model. When f is the Gaussian CDF (probit) then this gives the Thurston-Mosteller model (Thurstone, 1927; Mosteller, 1951a,b). For the parameters $\mu := \{\mu_i : i \in \mathcal{I}\}$ to be identifiable a constraint is required, typically on either a single parameter, for example setting $\mu_k = 0$ for some $k \in \mathcal{I}$, or on the sum, for example $\sum_{i\in\mathcal{I}} \mu_i = 0$. Mathematically, the constraint is required because any constant $c \in \mathbb{R}$ could be added to all the values $\mu_i^* = \mu_i + c$, $\forall i \in \mathcal{I}$, resulting in identical inference, since only the difference between the parameters is informative. Intuitively, the constraint is required because we are modelling a *relative* system, that is, the system requires *context*, and each object rating is meaningless in isolation.

Home advantage is widely recognised as an important feature in pairwise comparison, and particularly when applied to sport (Cattelan et al., 2013). In Bradley-Terry type models, this is naturally achieved by changing p_{ij} to a probability

$$p_{ij}^{(i)} = f(\mu_i + \gamma - \mu_j) \; \forall i \neq j \in \mathcal{I}, \tag{6.1.1}$$

if i is the *home object*, where $\gamma \in \mathbb{R}$ determines the effect of being at home, which here is common over all pairs of objects. If $\gamma > 0$ ($\gamma < 0$) then the probability of a home preference is increased (decreased) relative to the objects' ratings. This effect can be extended to vary over objects by replacing γ by γ_i in expression (6.1.1), giving a set of home advantage parameters $\{\gamma_i : i \in \mathcal{I}\}$. To ensure these parameters are all identifiable, some $\gamma_k = 0$, $k \in \mathcal{I}$ can be fixed, though no additional constraints are needed if there is a common γ. Alternatively, placing a lasso penalty (Tibshirani, 1996) on the home advantage parameters $\{\gamma_i : i \in \mathcal{I}\}$ will ensure that only a subset of distinct home-advantage parameters are considered. Masarotto and Varin (2012) cast a lasso penalty on the ratings themselves. The penalty constrains the ratings such that objects with statistically insignificant differences in rating are rated identically, which avoids over-interpretation of misinformed rankings. This can also be achieved via directly clustering the objects' ratings into distinct levels. See Chapter 7 for a novel approach

to clustering in pairwise comparison.

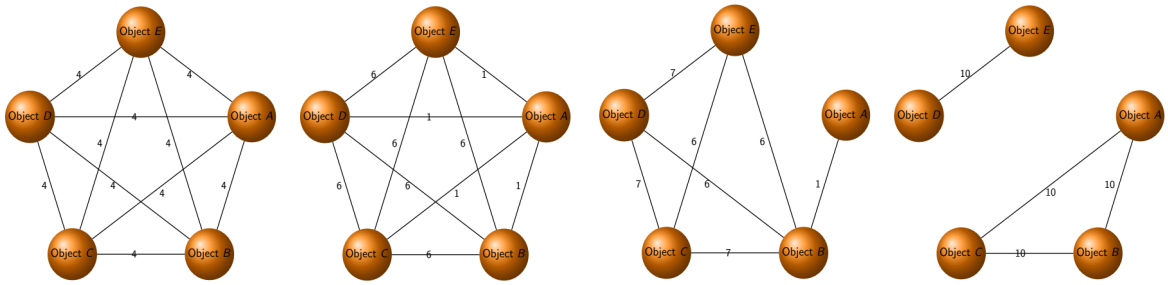


Figure 6.1.1: Illustration of four possible tournament structures, created using the same total number of comparisons.

Both clustering and penalising via lasso alter the parameter identifiability and therefore the required constraints. Identifiability is also affected by the order of comparison, or *tournament structure*. It's helpful to consider the graphical view of a paired comparison system here, with objects as nodes and their comparisons as edges. Figure 6.1.1 depicts four graphs resulting from different tournament structures, but with the total number of comparisons equal within the system. Assuming that all objects are identical, then if all objects are compared to all others an equal number of times - a tournament structure known as *round robin*, Figure 6.1.1 (far left) - this creates maximal connection between all pairs of teams. As such, the total uncertainty across all the objects' parameter estimates - the system uncertainty - is minimised. If there is imbalance but the graph is still fully connected as in Figure 6.1.1 (middle left), then information is transferred less efficiently between the objects and the system uncertainty increases. In Figure 6.1.1 (middle right) the connectivity is further reduced, with object A's rating only determined via the single comparison it has with object B, making object A's rating difficult to estimate. Finally, in Figure 6.1.1 (far right), the graph is disconnected, and none of the objects' ratings are identifiable without further constraints, despite the same total number of comparisons. Precisely, if a tree cannot be formed on the graph using the available pairwise connections, then more constraints are required in

order for any objects' ratings to be identifiable. This is proven in Chapter 7, which explores the combined effect of both clustering and tournament structure on parameter identifiability in paired comparison.

Some systems allow for a comparison between two objects to result in "no-preference", known in sports as a *draw*, or *tie*. Extensions of the Bradley-Terry model have been proposed for handling no-preference. Two distinct extensions are Cattelan et al. (2013) and Hankin (2020). The former uses ordinal logistic regression, treating preference for either object and no-preference, as outcomes of an ordered multinomial random variable, which can then be analysed via an ordered link model. In contrast, the latter treats the problem as a competition between the two objects and a third theoretical object (called "draw monster"), such that when preference is expressed for the draw monster, the outcome of the match corresponds to no-preference between the two actual objects. The strength of the draw monster therefore reflects the proclivity to draw.

6.2 Exploring Intransitivity

The models considered thus far use a single parameter to describe each object's rating. The direct implication of this is the imposition of transitivity. Transitivity is defined over a set of three of objects, and constrains that, if A is preferred to B, and B is preferred to C, then A must be preferred to C. Further, a set of objects $\mathcal{I}$ with cardinality $|\mathcal{I}| \geq 3$ exhibit transitivity if all triplet-subsets $\tilde{\mathcal{I}} \subseteq \mathcal{I}$, $|\tilde{\mathcal{I}}| = 3$, exhibit transitivity. A system exhibits *intransitivity* if at least one triplet-subset of its set of objects violates the transitivity constraint. The most famous example of an intransitive system is the game Rock-Paper-Scissors: a paired comparison system in which Paper is preferred to Rock, which is preferred to Scissors, which is in turn preferred to Paper. Consider the "deterministic" version, where a player r always picks Rock, a player p always picks Paper, and a player s always picks Scissors. There is clearly no set of

three ratings available to objects Rock, Paper and Scissors which reflects this behaviour in this formulation. Modelling such a system therefore requires a deviation from the classical one-object-one-rating structure.

The Rock-Paper-Scissors game depicts the violation of *deterministic transitivity* within a relative system. Through our statistical lens however, we are more interested in *stochastic transitivity*. For a set of 3 objects i, j, k, where $i \succ j$ denotes that i is preferred to j for all $i \neq j \in \{i, j, k\}$, there are three definitions of stochastic transitivity: weak stochastic intransitivity whereby

$$\Pr\{i \succ j\} > 0.5,\ \Pr\{j \succ k\} > 0.5 \Rightarrow \Pr\{i \succ k\} > 0.5,$$

i.e., we can deduce preferences only; strong stochastic transitivity, whereby

$$\Pr\{i \succ j\} > 0.5,\ \Pr\{j \succ k\} > 0.5 \Rightarrow \Pr\{i \succ k\} \geq \max\{\Pr\{i \succ j\}, \Pr\{j \succ k\}\}, \tag{6.2.1}$$

i.e., we can also deduce something about the probability; and linear stochastic transitivity, whereby

$$\Pr\{i \succ j\} = F(\mu_i - \mu_j),\ F : \mathbb{R} \to [0, 1],\ \frac{\mathrm{d}F(x)}{\mathrm{d}x} > 0,\ \forall x,$$

where $\mu_a \in \mathbb{R}$ is the rating of object a, $\forall a \in \{i, j, k\}$. If $F(x) = [1 + \exp(-x)]^{-1}$ then this is the Bradley-Terry model, so $\operatorname{logit}(\Pr\{i \succ j\}) = \mu_i - \mu_j,\ \forall i \neq j \in \mathcal{I}$. Of these three definitions, linear stochastic transitivity is the strictest, and implies that given any three objects $i \neq j \neq k \in \mathcal{I}$, and two probabilities $\Pr\{i \succ j\}$, $\Pr\{i \succ k\}$, the probability $\Pr\{j \succ k\}$ is completely determined by the other two probabilities. This will be proven in Chapter 7. From hereon in, stochastic transitivity will simply be termed *transitivity*.

If *intransitivity* is the violation of transitivity, then intransitivity is not well defined,

since it is not clear which of the three aforementioned forms is violated. Moreover, intransitivity can form part of an arbitrarily large cycle. For example, the $n > 3$ objects $\{x_i : i \in \{1, \ldots, n\}\}$ may have preference probabilities $\Pr\{x_i \succ x_{i+1}\} > 0.5,\ \forall i \in \{1, \ldots, n-1\}$ and $\Pr\{x_n \succ x_1\} > 0.5$, resulting in an intransitive cycle of length n. In this scenario It is not clear which of the n objects is causing the intransitivity. Whether a cycle of length 3 has more or less intransitivity than a cycle of length n is also ambiguous without some means of quantifying the intransitivity of a system. Of course, in the stochastic scenario a worse ranked object could express preference over a better ranked object from any number of independent comparisons purely due to chance, i.e., an "upset", and distinguishing between an upset and intransitivity can be challenging. In fact, intransitivity is commonly assumed as being due to inference variation or as arising due to errors in the dataset (Skinner and Freeman, 2009; Kéri, 2011), such as underlying incompleteness of preferences, and it is therefore treated as a nuisance which should be removed in these contexts. However, in the case of Rock-Paper-Scissors this is clearly not the case - the intransitivity is built into the inherent structure of the competition. Indeed, it has been shown experimentally that intransitivity can be a real feature of a system which cannot be accounted for by errors or natural variation (Tversky, 1969; Montgomery, 1977; Lindman and Lyons, 1978). In recommender systems, some view intransitivity as arising due to the aggregation process from different judges' underlying preferences, but that each judge's underlying preferences are still transitive (Rendle et al., 2009; Pan and Chen, 2013). Other work recognizes that intransitivity may be systematic even in a *one judge system*, which is where Chapter 7 sits, and some model intransitivity as arising from both sources (Chen et al., 2017).

Systematic intransitivity even under a one judge system is observable not just in artificial constructions, such as dice games (De Schuymer et al., 2003) or quantum games in physics (Makowski and Piotrowski, 2006), but also occurs naturally, for example, in

competition between bacteria (Reichenbach et al., 2007) or mating choices of lizards (Sinervo and Lively, 1996). Pahikkala et al. (2010) contains many more examples, and argues that violation of weak stochastic transitivity can occur in any situation where the best strategy in a given comparison depends on the strategy of the opponent. Given this, it would not be surprising to find cases of intransitivity in sports. In fact, by drawing parallels with social choice theory, Smead (2019) provides a philosophical argument as to why intransitivity is not only unsurprising, but is particularly likely to occur in sports.

Poisson models handle some amount of intransitivity without directly modelling it. They differ from the approaches presented thus far, in that they model preference indirectly by, for example, modelling a *score-line.* This can be more efficient since it utilises more information. Maher (1982) model score-lines in football (soccer). For a given pair of teams (i, j), $\forall i \neq j \in \mathcal{I}$ in a league comprising the set of teams $\mathcal{I}$, the goals scored by each team in a comparison between them is assumed Poisson distributed. Letting X_{ij} and X_{ji} be the goals scored by team i and j respectively, then

$$X_{ij} \sim \text{Poisson}(\alpha_i \beta_j) \perp\!\!\!\perp X_{ji} \sim \text{Poisson}(\alpha_j \beta_i),$$

where $\alpha_k > 0$ and $\beta_k > 0$ are interpreted as the attacking and defensive abilities of a team $k \in \mathcal{I}$. A home-advantage is included by modelling

$$X_{ij}^{(i)} \sim \text{Poisson}(\gamma \alpha_i \beta_j), \qquad X_{ji}^{(i)} \sim \text{Poisson}(\alpha_j \beta_i),$$

where the superscript denotes the home team, and $\gamma \in \mathbb{R}$ dictates the advantage (or disadvantage) of playing at home. Various extensions exists: modelling correlation between $X_{ij}^{(i)}$ and $X_{ji}^{(i)}$ and weighting the importance of data based on its recency (Dixon and Coles, 1997); and by framing the model in a Bayesian context with ratings that update after each comparison (Rue and Salvesen, 2000). Other examples of modelling

score-line include McHale and Morton (2011), who model the number of games in a tennis match to exploit the knowledge that losing 0-6, 0-6 is worse than losing 6-0, 6-7, 6-7. Note that in this example the losing player actually scores more total points, and given this datum alone the losing player would actually be ranked as better.

The Poisson models' use of attacking and defensive parameters for each object equips it with the support for some intransitivity. Although more degrees of freedom often provide greater flexibility, it is not simply the quantity of parameters that determines this support. Chapter 7 shows that it is possible to develop a model which violates even weak intransitivity using fewer parameters than one-object-one-parameter models, e.g., Bradley-Terry. It turns out that Poisson models have the flexibility to violate strong transitivity, but not weak intransitivity. Carroll and De Soete (1991) argue that in order to be realistic, a paired-comparison model should violate strong stochastic transitivity, but not necessarily weak intransitivity. But both approaches necessarily fail to model Rock-Paper-Scissors. Support for the violation of weak stochastic transitivity, and therefore intransitive triplets (or cycles), is less common. Early examples include the work of Tsai and Böckenholt (2006), which uses a random effects model to capture intransitivity, and that of Pahikkala et al. (2010), which tackles the problem from a machine learning standpoint using kernel methods to estimate the intransitive relations.

Chapter 7 presents a novel framework which encapsulates all three forms of intransitivity; allowing the flexibility of the model to be determined by the data. When modelling truly transitive objects, the model becomes the special case of the Bradley-Terry model, exhibiting linear transitivity. When there is evidence that the data violates strong transitivity, the model can capture this parsimoniously, potentially using less degrees of freedom than the Poisson models. When the data violate weak transitivity the model adapts again, allowing for modelling the "simple" deterministic game of Rock-Paper-Scissors.

6.3 Heuristic ranking methods

Being statistical approaches, the ranking methods depicted above quantify uncertainty around their parameters, be that objects' ratings or attacking and defensive abilities. Though lacking this functionality, heuristic methods for ranking offer speed. This allows for assumptions found in their statistical counterparts to be relaxed, such as stationarity. Heuristics can therefore be dynamic whilst remaining computationally affordable.

In a *connected* relative system, see Figure 6.1.1 (far left, middle left and middle right), all pairwise relationships between objects $\mathcal{I}$ are liable to change after any single comparison between a pair $(i, j) : i \neq j \in \mathcal{I}$, though the relationship between (i, j) is generally impacted the most.

The Elo rating system (Elo, 1978) is a dynamic heuristic introduced by the physicist and chess player Arpad Elo for ranking chess. It is *dynamic* in the sense that the objects' ratings update with each comparison, and is a *heuristic* in the sense that in a given comparison, only the two compared objects have their ratings updated rather than the full set, thus only approximating the full system change. This enhances its computational efficiency, and allows for the Elo rating system to be used even for systems containing any arbitrarily large number of objects. The premise of the Elo system is that the change in ratings of two objects engaged in pairwise comparison should depend on the expectation of the outcome. Let $X_{ij} \sim \text{Binomial}(p_{ij})$ be the random variable corresponding to a comparison between objects $i \in \mathcal{I}$ and $j \neq i \in \mathcal{I}$, with X_{ij} being 1 if object i is preferred, and X_{ij} being 0 if object j is preferred, and where p_{ij} is the probability of i expressing preference over j. Then, the *Elo rating* μ_i of object i is updated to a rating μ_i' after a comparison with object j via

$$\mu_i' = \mu_i + K(x_{ij} - \mathbb{E}[X_{ij}]),\ \forall i \neq j \in \mathcal{I},$$

where $x_{ij} \in \{0, 1\}$ is a realisation from X_{ij}, and $K \in \mathbb{R}_+$ is known as the K-factor

and reflects the dynamic aspect of the system, by controlling the weighting of recent events on the overall ratings. The K-factor can be thought of as a measure of the 'forgetfulness' of the system. If K is too large, then the ratings become too volatile and sensitive to recent events, increasing the variance in the ratings. Take K too small, and the objects' ratings are too heavily weighted on early events which may no-longer be a true reflection of the object's rating, thus inducing bias.

The Elo system is in fact a special case of the Glicko system (Glickman, 1999), a Bayesian construction which allows for uncertainty in the object abilities, although it still only updates the ratings of the compared pair in a given comparison. The uncertainty in an object's ability increases the longer the object remains dormant. The Glicko2 system (Glickman, 2001) extends this by introducing a volatility measure, which indicates the expected variation in an object's performances given its rating.

Both methods commonly use a logistic link function between the object ratings and the win probability as in Bradley-Terry, so that

$$p_{ij} = \mathbb{E}[X_{ij}] = \frac{1}{1 + \exp\left[(\mu_j - \mu_i)/\lambda\right]},$$

where $\lambda > 0$ is a parameter that determines the spread of the ratings, which does not effect the order, or rank, of the objects. Alternatively, a Gaussian link function can be used, as in the Thurstone-Mosteller model, such that

$$p_{ij} = \Phi\left(\frac{(\mu_i - \mu_j)}{\lambda}\right).$$

Chapter 7

Modelling Intransitivity in Pairwise Comparisons with Application to Baseball Data

7.1 Introduction

The seminal Bradley-Terry model (Bradley and Terry, 1952) is commonly used to rank objects from paired comparison data. Given a set $\mathcal{I}$ of n objects with each object $i \in \mathcal{I}$ having skill $r_i \in \mathbb{R}$, then the Bradley-Terry model gives, for $i \neq j \in \mathcal{I}$,

$$p_{ij}^{(\mathrm{BT})} = \Pr\{i \succ j\} := \{1 + \exp[-(r_i - r_j)]\}^{-1}, \tag{7.1.1}$$

where $a \succ b$ denotes preference for object a over b, and $r_1 = 0$ to avoid identifiability issues. A ranking of the objects is given by sorting estimates of $r := \{r_i \in \mathbb{R} : i \in \mathcal{I}\}$. This model is transitive, i.e., $p_{jk}^{(\mathrm{BT})}$ is given by $p_{ij}^{(\mathrm{BT})}$ and $p_{ik}^{(\mathrm{BT})}$, for all $i \neq j \neq k \in \mathcal{I}$, see Section 7.3.

Now consider the game of Rock-Paper-Scissors, a zero-sum game in which Rock beats Scissors, Scissors beats Paper, and Paper beats Rock, and specifically consider

the deterministic scenario where players (r,p,s) always pick (Rock, Paper, Scissors) respectively. In this scenario, all win probabilities in a game are either 0 or 1 depending on the opponent, and each player wins their next game with probability 1/2 if their next opponent is to be selected at random. Whatever way the skill of a player is defined, the symmetry of this game set-up unquestionably leads to the conclusion that the three players have equal skill levels.

Conclusions drawn from a Bradley-Terry model fitted to data from this simple game are surprisingly poor. Given a round-robin tournament, where each player plays all other players an equal number of times, the model will correctly estimate that all players are equally ranked in terms of skills; however, it would also estimate all pairwise win probabilities to be 1/2, which couldn't be more wrong. Even worse, is that any illusory ranking can result when the tournament is not round-robin, e.g., if the most common pairing of players is (r,s) and the other two pairings occur equally often then the Bradley-Terry model will rank player r as top. The key reason for the failure of the Bradley-Terry model is its transitive nature, a trait shared by almost all commonly used ranking systems.

Here we develop a novel pairwise comparison model, and an associated ranking system, which accounts for *intransitivity*. Thus, it describes how specific pairwise probabilities differ from probabilities given by overall skill levels alone, i.e., how probabilities differ from those given by the Bradley-Terry model. The Rock-Paper-Scissors game also illustrates that ranking can involve ties, where subsets of players can have equal skill levels, and that tournament structure can effect the subsequent inference. We also address some aspects associated with these issues.

The concept and associated modelling of intransitivity is not new. Makowski and Piotrowski (2006) present many examples of competitions exhibiting intransitivity and argue that it can occur whenever the best strategy in a given comparison depends on the strategy of the opponent, and Smead (2019) provides a philosophical argument

as to why intransitivity is particularly likely to occur in sports. Given this, it is not surprising to find cases of intransitivity in e-sports (Makhijani and Ugander, 2019; Chen and Joachims, 2016; Duan et al., 2017). Other applications include social choice, real sensory analysis, and election data-sets.

With n competitors there are $n(n-1)/2$ interactions, or intransitivities, so even in round-robin competitions, with m rounds, there are too many terms to estimate efficiently using empirical methods, unless m/n is large. Causeur and Husson (2005) proposed an $O(n^2)$ parameter extension of the Bradley-Terry model to address intransitivity. Subsequently $O(nd)$ parametric models have been studied for some fixed $d \in \mathbb{N}$ ($d \ll n$), see all the models in Section 7.2, but they lack the flexibility to cover the potentially $O(n^2)$ different intransitivities across n players, leading to bias; whilst they are not parsimonious when the intransitivity is simple, leading to inefficiency.

Although the *concept* of intransitivity is quite clear, there is no established *measure* of the amount of intransitivity in a dataset. In this work, we propose a definition of intransitivity through a distance metric between the assumed probability of paired comparisons under a Bradley-Terry model, and the empirical or model-based probability estimate, such that for any given dataset the magnitude of the intransitivity present is unambiguous. A *flexible* model then, is one which is capable of exploring the space of all possible combinations of intransitivity, as defined by this measure. Any parametric model is restricted to a subset of this space by definition, with this restriction being most obviously revealed when assessing predictive performance.

We then develop a novel semi-parametric extension of the Bradley-Terry model, allocating the $n(n-1)/2$ pairs of objects to a random number K, with $0 \leq K \leq n(n-1)/2$, of distinct intransitivity levels, each level representing a different strategy. We term this model the *Intransitive Clustered Bradley-Terry* (ICBT) model. Relative to the aforementioned parametric models, this ICBT model provides greater flexibility to enable the incorporation of varying structures, and degrees of, intransitivity. As

many of these strategies will have similar effects, we anticipate that K should be small, yet the random property of K provides the potential for it to be large when required. This flexibility ensures that our model is parsimonious, whatever the complexity of the data. For our Rock-Paper-Scissors illustration $K = 1$.

Moreover, our novel approach for the objects' skills is to allocate the n objects into a random number of $A+1 \leq n$ distinct skill levels, to improve efficiency and avoid false rankings. This constraint recognises that from paired comparison data there will be objects that are indistinguishable as having statistically significantly different skill levels, e.g., for our Rock-Paper-Scissors illustration $A = 0$. So clustering skills avoids over-interpretation of misinformed rankings, a feature Masarotto and Varin (2012) address by clustering skills via a lasso procedure.

The basis of our model is the belief that in practice there are likely to small subsets of skill and intransitivity levels, namely $A \leq n - 1$ and $K \ll n(n-1)/2$ respectively. As we have little prior knowledge about the skills of the objects or the intransitivities of the pairs of objects, we allow the clustering of objects into different skill levels, and of the pairs of objects into separate intransitivity levels, to be determined entirely through a Bayesian hierarchical model. We take each of (A, K), the allocations of objects to skill levels, and the allocations of the pairs of objects to intransitivity levels as unknown, with inference being conducted via a reversible jump Markov chain Monte Carlo (RJMCMC) algorithm. This formulation does offer computational challenges; however, we anticipate that typically the posterior will give a high probability that $A + K < n$ and that many of the cluster allocations also will be strongly identified. Our inference framework offers the opportunity to select a highly simplified model, with the values of A, K and allocations fixed at values given by posterior means/modes if these are found to align with known structure about the paired comparison. In the absence of such knowledge our results allow for the full uncertainty of these features to be accounted for.

In certain circumstances our model has the potential to identify and correct for imbalanced tournament structure on overall rankings since teams are not penalised if they (unfairly) compete most frequently against those whom they perform systematically worse to relative to what is expected based on respective skills alone.

We use American League Baseball data to illustrate the performance of our methods in comparison to existing models for a range of reasons. Firstly, each game results in a win or a loss for a team. Secondly, it is known to be a highly strategic game, see Section 7.5, so we anticipate that the level of intransitivity will be high. Finally, although the tournament structure is not round robin, each team plays each other team often, and so the existence of intransitivity should become apparent in inference. Indeed this is found in Section 7.5, where our model is shown to have an improved fit over the Bradley-Terry model and existing parametric intransitivity models in out of sample testing for each of the nine seasons we study.

The layout is as follows. Section 7.2 introduces other approaches to modelling intransitivity. Section 7.3 then introduces our novel measure of intransitivity, the ICBT model, and the ranking formulation. Section 7.4 contains details of the inference, including prior specification, our full Bayesian hierarchical modelling strategy, an overview of the RJMCMC algorithm and its novel features, and an overview of a simulation study. Section 7.5 compares this model with the Bradley-Terry model and other competitor models, using American League baseball data. Section 7.6 is a discussion. Full details of the RJMCMC algorithm, simulation study, and extended analysis of the baseball application are in the supplementary material.

7.2 Literature on Intransitive Modelling

The blade-chest model of Chen and Joachims (2016), extends the Bradley-Terry model into d-dimensions by incorporating so-called *blade* and *chest* vectors $b_i, c_i \in \mathbb{R}^d$ for each

object $i \in \mathcal{I}$. There are two versions: the *-dist* and *-inner* variants, given respectively by

$$\text{logit}\left(p_{ij}^{(BCD)}\right) := ||b_j - c_i||_2^2 - ||b_i - c_j||_2^2 + r_i - r_j, \text{ and logit}\left(p_{ij}^{(BCI)}\right) := b_i^T \cdot c_j - b_j^T \cdot c_i + r_i - r_j.$$

The blade and chest parameters of all the objects can be viewed as features on a d-dimensional latent space. Then, if an object i's blade is close to object j's chest, and simultaneously object i's chest is far from object j's blade, then object i has an additional advantage over object j. If $d = 2$ this model can represent a deterministic Rock-Paper-Scissors game, by placing the blade of Rock at the chest of Scissors, the blade of Scissors at the chest of Paper, and the blade of Paper at the chest of Rock. By increasing d, ever more complex relationships can be captured between the pairs of objects. Given n objects and $b_i, c_i \in \mathbb{R}^d$ for each object i, the model contains $2d(n-1)$ identifiable parameters. The r parameters can be absorbed into the blade and chest parameters; however, the above parametrisation makes it clear that the Bradley-Terry model is a special case of the blade-chest model, when $b_i = b_j = c_i = c_j, \ \forall i, j \in \mathcal{I}$.

Duan et al. (2017) introduce a generalised model for intransitivity, with

$$\text{logit}\left(p_{ij}^{(G)}\right) = \mu_i^T \Sigma \mu_j + \mu_i^T \Gamma \mu_i - \mu_j^T \Gamma \mu_j,$$

with $\mu_i \in \mathbb{R}^d$, where d is even, being a d-dimensional *strength* vector, for an object $i \in \mathcal{I}$, and $\Sigma, \Gamma \in \mathbb{R}^{d \times d}$ are so-called *transitive matrices*. The first matrix Σ represents the interactions between objects, and Γ controls how an individual object's strength components form the object's overall strength. The number of identifiable parameters is $d(3d/2 + n - 1)$, since there are two $d \times d$ matrices (Σ, Γ), of which Σ is skew-symmetric, and n d-dimensional vectors. They show that their model is a generalisation of the blade-chest and Bradley-Terry models. Specifically, the blade-chest-inner model arises when $\mu_i = (b_i, c_i)$, $\mu_j = (b_j, c_j)$, and $||\mu_i||_2^2 = ||\mu_j||_2^2$, that is, the objects all have equal

skill, and Σ is a block diagonal matrix with two $(d/2) \times (d/2)$ matrices of zeros on the diagonal and matrices $I_{d/2}$ and $-I_{d/2}$ on the off-diagonals where I_m is the $m \times m$ identity matrix, then $\mu_i^T \Sigma \mu_j = b_i^T \cdot c_j - b_j^T \cdot c_i$. The degrees of freedom are restricted by regularization, using an L_2 norm for the object strength vectors and Frobenius norm for both transitivity matrices. The tuning parameters are selected via cross-validation.

Makhijani and Ugander (2019) introduced a majority vote model with object i having a vector of d skill attributes, $(\mu_{i,1}, \ldots, \mu_{i,d})$, where d is odd. Then, given a suitable choice of mapping function f, e.g., logistic or Gaussian, define $q_{ij}^l = f(\mu_{i,l} - \mu_{j,l})$, $\forall l \in \{1, \ldots, d\}$ to be the probability of i beating j based only on their lth attribute. Then, majority vote model says that the probability of i being preferred to j overall, is the probability that it wins across the majority of attributes. For $d = 1$ the model is linearly transitive, but not when $d = 3$, as

$$\Pr\{i \succ j\} = q_{ij}^1 q_{ij}^2 q_{ij}^3 + (1 - q_{ij}^1) q_{ij}^2 q_{ij}^3 + q_{ij}^1 (1 - q_{ij}^2) q_{ij}^3 + q_{ij}^1 q_{ij}^2 (1 - q_{ij}^3).$$

7.3 Modelling

7.3.1 Measure of Intransitivity

From the model definition (7.1.1), the Bradley-Terry model assumes linear transitivity. This assumption constrains the pairwise probabilities of the model such that, given $p_{ij}^{(\mathrm{BT})}$ and $p_{jk}^{(\mathrm{BT})}$ from (7.1.1) for any $i \neq j \neq k \in \mathcal{I}$, the probability $p_{ik}^{(\mathrm{BT})}$ is completely determined. It is straightforward to show that the form of $p_{ik}^{(\mathrm{BT})}$ is given as

$$p_{ik}^{(\mathrm{BT})} = \frac{p_{ij}^{(\mathrm{BT})} p_{jk}^{(\mathrm{BT})}}{1 + 2 p_{ij}^{(\mathrm{BT})} p_{jk}^{(\mathrm{BT})} - \left(p_{ij}^{(\mathrm{BT})} + p_{jk}^{(\mathrm{BT})} \right)}, \ \forall j \neq i, k,$$

noting it is independent of the choice of *bridge* object j. Therefore, there can be no interaction that is specific to the pair $\{i, k\}$, that is not already captured between all other pairs.

Including intransitivity, however, allows for some pairwise probabilities to depart from those assumed by the Bradley-Terry model. This can be formalised by supposing that for all $i \neq k \in \mathcal{I}$ the true probability of preference $i \succ k$ is given as some function $f : \{[0, 1], \mathbb{R}\} \to [0, 1]$ of the Bradley-Terry probability and the intransitivity, θ_{ik}, of the pair $\{i, k\}$, then we can write

$$p_{ik} := f\left(p_{ik}^{(\mathrm{BT})}, \theta_{ik}\right), \ \forall i \neq k \in \mathcal{I}, \tag{7.3.1}$$

where we identify the form of f in Section 7.3.2. We define the intransitivity to be the displacement of the true probabilities from the Bradley-Terry probabilities on the log-odds scale, so that

$$\theta_{ik} := \log\left(\frac{p_{ik}/(1 - p_{ik})}{p_{ik}^{(BT)}/(1 - p_{ik}^{(BT)})}\right), \ \forall i \neq k \in \mathcal{I}, \tag{7.3.2}$$

is the amount of intransitivity between the pair of objects $\{i, k\}$. A value of $\theta_{ik} = 0$ indicates that the comparison is transitive, i.e., the pairwise probabilities could be modelled by the Bradley-Terry model. As a consequence we require $f(x, 0) = x, \ x \in [0, 1]$. The choice of log-odds ratio in equation (7.3.2) reflects the non-linearity of probabilities. For example, if $\epsilon = 0.099$, then a small linear shift in probability from 0.5 to $0.5 + \epsilon$ has little impact on the odds, which remain at approximately 1:2. However, a linear shift in probability from 0.9 to $0.9 + \epsilon$ has a huge impact on the odds, which move from 1:10 to 1:1000. Moreover, our definition (7.3.2) for the intransitivity θ_{ik} also

imposes rotational symmetry for pairs of objects, that is

$$\theta_{ki} = \log\left(\frac{p_{ki}/(1-p_{ki})}{p_{ki}^{(BT)}/(1-p_{ki}^{(BT)})}\right) = \log\left(\frac{(1-p_{ik})/p_{ik}}{(1-p_{ik}^{(BT)})/p_{ik}^{(BT)}}\right) = -\theta_{ik},\ \forall i \neq k \in \mathcal{I}, \tag{7.3.3}$$

so we need to find $\{\theta_{ik},\ \forall i > k \in \mathcal{I}\}$ only, in order to completely define $\{\theta_{ik},\ \forall i \neq k \in \mathcal{I}\}$.

7.3.2 Model formulation

To find the function f in equation (7.3.1), equation (7.3.2) can be simply rearranged which gives

$$p_{ik} = \frac{p_{ik}^{(BT)}\exp(\theta_{ik})}{p_{ik}^{(BT)}\exp(\theta_{ik}) + 1 - p_{ik}^{(BT)}},\ \forall i \neq k \in \mathcal{I}, \tag{7.3.4}$$

and so for any pair $\{i, k\}$, equation (7.3.4) can be re-written as

$$p_{ik} = \frac{1}{1+\exp[-(\theta_{ik} + r_i - r_k)]},\ \forall i \neq k \in \mathcal{I}. \tag{7.3.5}$$

Here the effect of θ_{ik} is clear, positive (negative) θ_{ik}, increases (decreases) the probability of team i beating team k relative to their skills alone, i.e., relative to the Bradley-Terry model.

Thus far, the model contains the flexibility to describe $P := \{p_{ik} \in [0,1], \forall i \neq k \in \mathcal{I}\}$ completely. Here P contains $n(n-1)/2$ degrees of freedom, because $p_{ki} = 1 - p_{ik}$; however, the model contains $n(n-1)/2+n$ parameters: $n(n-1)/2$ values of intransitivity between pairs, and n skill parameters from r, and thus the model parameters are not identifiable. One way of ensuring identifiability in the standard Bradley-Terry model is to fix one object's skill level, and here it is chosen that $r_1 = 0$. As well as this constraint on the objects' skill parameters, the intransitivity parameters require constraints for parameter identifiability. The minimal set of required constraints is identified in Proposition 1, see the Appendix for the proof.

Proposition 1. *Consider a round-robin tournament with pairs of objects (i, j) being compared, with $i, j \in \mathcal{I}$, with $i \neq j$ where $|\mathcal{I}| = n$. Suppose that the probabilities of i beating j are p_{ij} where these probabilities are given by expression (7.3.5), with $r_1 = 0$ and intransitivity values θ_{ij}. If a set of $n-1$ pairs of objects, indexed by $\mathcal{J}_{n-1}$, have their intransitivity values set to arbitrary specified values, then all the rest of the $\{r_i\}$ and $\{\theta_{ij}\}$ parameters in expression (7.3.5) are identifiable if $\mathcal{J}_{n-1}$ forms a connected graph over $\mathcal{I}$. Furthermore, if less than $n-1$ pairs' intransitivity values are specified or if $\mathcal{J}_{n-1}$ is not a connected graph over $\mathcal{I}$ then identifiability is not achievable.*

We choose the $n-1$ constraints to be $\theta_{ij} = 0,\ \forall (i,j) : i = 1, j \in \mathcal{I} \setminus \{1\}$, that is, all pairs involving object 1 have intransitivity set to 0. Proposition 1 gives that if any further constraints are imposed on the intransitivity values the flexibility of model (7.3.5) will be compromised.

With the above constraints, the minimal conditions for parameter identifiability are satisfied, but the model is still likely to overfit with so many parameters. To rectify this we restrict the total number of degrees of freedom, by restricting both the number of intransitivity values to only $K \leq n(n-1)/2$ unique values and restricting the number of unknown skill values to be $A < n$, where $A + K \leq n(n-1)/2$ and ideally $A + K \ll n(n-1)/2$. In this fashion our ICBT model embraces intransitivity in a parsimonious way.

Firstly, consider the $A+1$ unique *skill values*, which ensures parsimony in the model by clustering the objects' skills r into distinct values which are sufficiently statistically significantly different. Since $r_1 = 0$ is fixed, there are only A unknown *skill levels*, $\phi \in \mathbb{R}^A$. By defining the labels of the set of skill levels to be $\mathcal{A} := \{-A_-, \ldots, 0, \ldots, A_+\}$ with A_+ being the number of skill levels which are greater than 0 and A_- the number of skill levels less than 0 such that $A_- + A_+ + 1 = A + 1 = |\mathcal{A}| \leq n$, we impose the equivalent condition in our model by fixing the skill level with label $\{0\}$ to be $\phi_0 = 0$, and fixing object 1 to always be allocated to this cluster. The possible skill values an

object can take are therefore defined as

$$\{\phi_0 = 0, \phi := \{\phi_a \in \mathbb{R}, \forall a \in \mathcal{A} \setminus \{0\}\} : \phi_{-A_-} < \cdots < \phi_0 < \cdots < \phi_{A_+}\},$$

where ϕ are the unknown *skill levels*, and the ordering helps with label switching problems in the inference. The *skill cluster allocation* of object i, denoted $y_{\{i\}} \in \{0,1\}^{A+1}$, is a binary vector which takes the value 1 at position $s \in \mathcal{A}$ and 0 everywhere else, if object $i \in \mathcal{I}$ belongs to cluster s. The set $Y := \{y_{\{i\}} : i \in \mathcal{I} \setminus \{1\}\}$ then contains all the objects' skill cluster allocations except object $\{1\}$ which has fixed cluster allocation. Therefore, by defining $\mathcal{S}_{\{i\}}(Y) := \text{argmax}_s y_{\{i\},s},\ \forall i \in \mathcal{I} \setminus \{1\}$, then the objects' skills can be written as

$$r_i = \begin{cases} \phi_{\mathcal{S}_{\{i\}}(Y)}, & i \in \mathcal{I} \setminus \{1\} \\ 0, & i = 1 \end{cases} := f_r(\phi, Y, i),\ \forall i \in \mathcal{I}. \tag{7.3.6}$$

Now consider the K unique values of intransitivity to describe the different inter-object strategies. Of the $n(n-1)/2$ pairs of objects, many will adopt similar strategies depending on their opponents. These similar strategies are translated by the model as having similar departures from transitivity, and are thus clustered together. For example, suppose some group of objects $\mathcal{V} : j \notin \mathcal{V}$ competed against object j in the same way. Then it would be reasonable to assume that θ_{ij} is the same for all $i \in \mathcal{V}$. This creates clusters of pairs of objects, such that the pairs are clustered according to them having identical intransitivity.

In order to measure the departure from a Bradley-Terry model, a *linearly transitive cluster* is imposed, which contains the set of pairs $\mathcal{J}_T \subseteq \{\{i,k\} : i \neq k \in \mathcal{I}\}$, which have an intransitivity level $\theta_0 = 0$. Thus, the Bradley-Terry modelling assumption (7.1.1) holds for these pairs, such that $p_{ik} = p_{ik}^{(\text{BT})}$, for all $\{i,k\} \in \mathcal{J}_T$. Given the existence of this cluster, there must be strong evidence from the data to produce an additional

cluster with an intransitivity level close to 0. This choice does not impose transitivity of pairs as the linearly transitive cluster $\mathcal{J}_T$ may be empty, except for all pairs with object 1 which are classified as transitive due to our imposition of constraints for identifiability from Proposition 1. Let the distinct set of intransivity levels be

$$\theta_{\mathcal{K}} := \{\theta_0 = 0,\ \theta = \{\theta_k \in \mathbb{R}_+,\ \forall k \in \mathcal{K}\} : 0 < \theta_1 < \cdots < \theta_K\},$$

where $\mathcal{K} = \{1, \ldots, K\}$ and the levels of intransitivity are ordered from smallest to largest. The *levels of intransitivity*, θ, contain the set of positive values of intransitivity which, due to symmetry and the completely transitive cluster with intransitivity value θ_0, then define the full $2K + 1$ possible values of intransitivity between any pair of objects.

We define the intransitivity cluster allocation of a given pair $\{i, k\}$ to be another binary matrix $z_{\{i,k\}}$, which takes the value 1 at position $s \in \{-K, \ldots, K\}$ and 0 everywhere else, if the pair $\{i, k\}$ belongs to cluster s. The clusters are therefore labelled from $-K$ to K, where a cluster labelled $k \in \{1, \ldots K\}$ has cluster level θ_k, a cluster labelled $k \in \{-K, \ldots, -1\}$ has cluster level $-\theta_{-k}$, and a cluster with label 0 has cluster level $\theta_0 = 0$. The set $Z := \{z_{\{i,k\}},\ \forall i > k \in \mathcal{I} \setminus \{1\}\}$ then defines all the cluster allocations for all the free pairs $i \neq k \in \mathcal{I} \setminus \{1\}$, because of the rotational symmetry. For example, if the Kth index of $z_{\{i,k\}}$ has value $z_{\{i,k\},K} = 1$, then this indicates that the pair $\{i, k\}$ belongs to the cluster with label K, whose cluster level is the largest level of intransitivity θ_K, and this enforces that the pair $\{k, i\}$ belongs to cluster $-K$ and has the smallest level of intransitivity $-\theta_K$. If the cluster allocation of the pair $\{i, k\} \in \mathcal{I} \setminus \{1\}$ is

$$\mathcal{S}_{\{i,k\}}(Z) := \begin{cases} \text{argmax}_s z_{\{i,k\},s} & \text{if } i > k, \\ -\text{argmax}_s z_{\{k,i\},s} & \text{if } i < k, \end{cases}$$

then the level of intransitivity for a pair $\{i,k\}$, θ_{ik} can be redefined as

$$\begin{aligned}\theta_{ik} &:= f_\theta\left(\theta, Z, \{i,k\}\right) \qquad (7.3.7)\\ &= \begin{cases} \theta_{\mathcal{S}_{\{i,k\}}(Z)}\mathbb{1}\{\mathcal{S}_{\{i,k\}}(Z) \geq 0\} - \theta_{-\mathcal{S}_{\{i,k\}}(Z)}\mathbb{1}\{\mathcal{S}_{\{i,k\}}(Z) < 0\}, & \{i,k\} \in \mathcal{I} \setminus \{1\} \\ 0, & \text{otherwise} \end{cases}\end{aligned}$$

where $\mathbb{1}$ is the indicator function, and remembering that $\theta_0 = 0$.

The full model can be written either in terms of equation (7.3.5), noting that the parameters will be clustered, or can be written in terms of the levels and the cluster allocations,

$$p_{ik} = \left(1 + \exp\left\{-\left[f_\theta\left(\theta, Z, \{i,k\}\right) + f_r\left(\phi, Y, i\right) - f_r\left(\phi, Y, k\right)\right]\right\}\right)^{-1}.$$

So the ICBT model is defined by $\psi = \{\phi = \{\phi_a : a \in \mathcal{A} \setminus \{0\}\},\ \theta = \{\theta_k : k \in \mathcal{K}\}\}$.

Due to the intransitivity levels being fixed to 0 for all pairs of objects involving object 1, an adjustment is required to get a more interpretable value of intransitivity between the pairs. We define the adjusted intransitivity to be

$$\theta^*_{ij} := \text{logit}\left(p_{ij}\right) - \text{logit}\left(p^{(BT)}_{ij}\right) = \theta_{ij} + r_i - r_j - \left(r^{(BT)}_i - r^{(BT)}_j\right), \qquad (7.3.8)$$

that is, the difference between the logits of the pairwise probability between our ICBT model and the Bradley-Terry model. Note that the rotational symmetry of $\{\theta_{ij}\}$ (7.3.3) also imposes rotational symmetry on $\{\theta^*_{ij}\}$, that is, $\theta^*_{ij} = \theta^*_{ji}, \forall i \neq j \in \mathcal{I}$.

To help see the value of this reparametrisation, consider then the earlier example of a deterministic game of Rock-Paper-Scissors. Take Rock as the constrained object, then Rock has fixed skill level $r_r = 0$, and that pairs involving Rock have intransitivity 0, that is $\theta_{rp} = \theta_{rs} = 0$, where the p and s subscripts denote Paper and Scissors. To maintain that Rock always beats Scissors $p_{rs} = 1$, then from the constraints, we get

an excellent approximation from the ICBT model when $r_s = -M$ for some large M, with the approximation improving as $M \to \infty$. Likewise $r_p = M$, and $\theta_{ps} = -3M$. With this model there is only one skill level M, and one non-zero intransitivity level $-3M$. This parametrisation somewhat hides the symmetry of the intransitivity over pairs. However, with definition (7.3.8), then $\theta^*_{rs} = \theta^*_{sp} = \theta^*_{pr} = M$, resulting in an intuitive and easy interpretation of the intransitivity, reflecting the symmetry of the game, no-matter the choice of the fixed parameters.

7.3.3 Model Ranking

In the Bradley-Terry model, the skill parameters can simply be ordered to give a rank since a greater skill always results in higher win probabilities against all other objects. In our ICBT model this is not the case, because both the intransitivity parameters of each pair and the skill parameters of the objects impact the win probability between any pair. However, below we present two intuitive methods for determining overall ability, and therefore ranking.

Firstly, if $p_{ij} = \Pr\{i \succ j\}$ is the probability of an object i beating object j according to our model, then we can rank the objects by ordering

$$p_{\cdot} := \left\{ p_{i\cdot} := \frac{1}{n-1} \sum_{j \in \mathcal{I}:\, j \neq i} p_{ij} : i \in \mathcal{I} \right\}, \tag{7.3.9}$$

that is, $p_{i\cdot}$ is the average probability of object i beating any other object $j \neq i \in \mathcal{I}$.

Secondly, if we consider the intransitivity between an object i and an opposing object $j \neq i$ as some "boost" which contributes to the overall ability (which could be negative), then the overall ability a_i of object i could be defined by

$$a_i := r_i + \frac{1}{n} \sum_{j \in \mathcal{I}} \theta_{ij}, \text{ where } \theta_{ii} = 0, \ \forall i \in \mathcal{I}, \tag{7.3.10}$$

that is, the object skill plus its average intransitivity level. Definition (7.3.10) is equivalent to the Bradley-Terry definition of 'ability'. Defining logit $\left(p_{ii}^{(BT)}\right) = 0,\ \forall i \in \mathcal{I}$, a Bradley-Terry gives

$$\frac{1}{n}\sum_{j\in\mathcal{I}} \text{logit}\left(p_{ij}^{(BT)}\right) = r_i - \frac{1}{n}\sum_{j\in\mathcal{I}} r_j, \tag{7.3.11}$$

where the sum on the right hand side does not depend on i, so the skill of object i is entirely determined by r_i. Similarly, in our model

$$\frac{1}{n}\sum_{j\in\mathcal{I}} \text{logit}\left(p_{ij}\right) = r_i + \frac{1}{n}\sum_{j\in\mathcal{I}} \theta_{ij} - \frac{1}{n}\sum_{j\in\mathcal{I}} r_j = a_i - \frac{1}{n}\sum_{j\in\mathcal{I}} r_j, \tag{7.3.12}$$

then given definition (7.3.10), both (7.3.11) and (7.3.12) have the same form but with a_i replacing r_i. Then a ranking can be formed by ordering the set of abilities $a := \{a_i : i \in \mathcal{I}\}$. We argue that the first method, using the probabilities $p_{\cdot}$ to rank the objects, is more meaningful since it is directly associated with the pairwise probabilities, the modelling of which is our ultimate aim. The application to baseball data of both methods is discussed in the supplementary material.

7.4 Inference

7.4.1 Likelihood

The data, $x := \{x_c : c \in \mathcal{C}\}$, are binary, and $i \succ j$ denotes that i is preferred to j. Then $x_c = 1$ if $i_c \succ j_c$, and $x_c = 0$ otherwise, where $i_c, j_c \in \mathcal{I}$ are the objects being compared in comparison c. Then, the log likelihood for the ICBT model is

$$\ell(x|\phi, Y, A, \theta, Z, K) = \sum_{c\in\mathcal{C}} \left[x_c \log\left(p_{i_c j_c}\right) + (1 - x_c)\log\left(1 - p_{i_c j_c}\right)\right], \tag{7.4.1}$$

where $p_{i_c j_c}$ is given by the ICBT model for all $c \in \mathcal{C}$ and is calculated from the set of parameters $(\phi, Y, A, \theta, Z, K)$. All pairs' intransitivities $\{\theta_{ik} : i \neq j \in \mathcal{I}\}$ can be found

from the intransitivity levels θ and the cluster allocations Z, using equation (7.3.7), so it is only necessary to do inference on these parameters, rather than the full $2K+1$ separate clusters. Therefore from here onwards the term *intransitivity levels* refers only to those K values which have positive intransitivity. Similarly, any individual object's skill $r_i\ \forall i \in \mathcal{I}$ can be found from knowing the ability levels ϕ and the cluster allocations Y, using equation (7.3.6). We formulate a Bayesian hierarchical model, which treats both K and A as unknown parameters, thus accounting for uncertainty in the number of clusters. The posterior is therefore written as

$$\pi\left(\phi, Y, A, \theta, Z, K|x\right) \propto L\left(x|\phi, Y, A, \theta, Z, K\right)\pi\left(\phi, Y, A, \theta, Z, K\right)$$

where $L(\cdot) = \exp[\ell(\cdot)]$ is the likelihood and $\pi\left(\phi, Y, A, \theta, Z, K\right)$ is the prior.

7.4.2 Prior Specification

Formulating the prior, we make the assumption that $Z \perp\!\!\!\perp \theta|K$ that is, the intransitivity level allocations and intransitivity levels are independent from one another given the number of intransitivity levels K. Likewise, it is assumed that $Y \perp\!\!\!\perp \phi|A$. Furthermore, we assume that the clustering of the objects' skills and the clustering of the pairs' intransitivities are independent systems, that is, $A \perp\!\!\!\perp K$, $\phi \perp\!\!\!\perp \theta$, and $Y \perp\!\!\!\perp Z$. This means that the prior specification for the two features we are clustering, skills and intransitivities, can be approached separately.

Consider first the prior specification for the clustering of the intransitivity values of the pairs. Remember that labels $z_{\{i,j\}}, \forall i > j \in \mathcal{I}\backslash\{1\}$ have domain $\{-K, \ldots, K\}$, that is, $z_{\{i,j\}} : \{-K, \ldots, K\} \to \{0, 1\},\ \forall i > j \in \mathcal{I} \setminus \{1\}$, and also that $z_{\{i,1\}}, \forall i \in \mathcal{I}\backslash\{1\}$ (and by symmetry $z_{\{1,j\}}, \forall j \in \mathcal{I}\backslash\{1\}$ too) are fixed in the transitive level $\{0\}$ for identifiability

purposes, see Section 7.3.2. Let the prior on the cluster allocation be

$$z_{\{i,j\}}|\omega_K \sim \text{multinomial}\left(1, \omega_K\right), \ \forall i > j \in \mathcal{I} \setminus \{1\},$$

where ω_K is on $\{-K, \ldots, K\}$ such that

$$\omega_K = \{\omega_{K,s} \in [0,1] : s \in \{-K, \ldots, K\}, \ \sum_{s=-K}^{K} \omega_{K,s} = 1\}.$$

The distribution of $Z|\left(\omega_K, K\right)$ is assumed independent over all pairs $i > j \in \mathcal{I} \setminus \{1\}$, i.e.,

$$f\left(Z|\omega_K, K\right) = \frac{\left(\sum_{k=-K}^{K} |b_k|\right)!}{\prod_{k=-K}^{K} |b_k|!} \prod_{s=-K}^{K} \omega_{K,s}^{|b_s|},$$

where $b_k = \{(i,j) : i > j \in \mathcal{I} \setminus \{1\} : z_{\{i,j\},k} = 1\} \ \forall k \in \{-K, \ldots, K\}$ is the set of allocated pairs of objects belonging to cluster k. We set $\omega_K|K \sim \text{Dirichlet}\left(\bar{\gamma}_K\right)$ to come from $2K+1$ dimensional Dirichlet prior distribution, and $\bar{\gamma}_K \in \mathbb{R}_+^{2K+1}$ is the hyper-parameters vector. We use an uninformative prior, setting $\bar{\gamma}_K = \gamma_K 1_{2K+1}$ where 1_{2K+1} is a vector of ones of length $2K+1$ and $\gamma_K \in \mathbb{R}_+$. In this case, the ω_K parameter can be marginalised out, by

$$\begin{aligned} f(Z|\gamma_K, K) &= \int_{\omega_K} f(Z|\omega_K, K) f(\omega_K|\gamma_K, K) \, \mathrm{d}\omega_K \\ &= \frac{\left(\sum_{k=-K}^{K} |b_k|\right)!}{\prod_{k=-K}^{K} |b_k|!} \frac{\Gamma((2K+1)\gamma_K)}{\Gamma(\gamma_K)^{2K+1}} \frac{\prod_{k=-K}^{K} \Gamma\left(\gamma_K + |b_k|\right)}{\Gamma\left((2K+1)\gamma_K + \sum_{k=-K}^{K} |b_k|\right)} \end{aligned} \tag{7.4.2}$$

where integration on ω_K is taken over the $2K+1$ simplex. This is referred to as a Dirichlet-multinomial allocation prior. The prior for K is a Poisson(λ_K) distribution with probability mass function denoted $g_0\left(k|\lambda_K\right)$ so that $\mathbb{E}\left[K|\lambda_K\right] = \lambda_K$, with $\lambda_K > 0$. Note that $K = 0$ is feasible, as this corresponds to the Bradley-Terry model since $\theta|(K=0) = \emptyset$ and so only the transitive cluster exists, that is $\theta_{\mathcal{K}}|(K=0) = \theta_0$, and

$\{i,j\} \in \mathcal{J}_T,\ \forall i \neq j \in \mathcal{I}$, so all pairs belong to the transitive cluster $\mathcal{J}_T$. Formally $K < n(n-1)/2 - n$ but as this is large relative to our prior beliefs on K, for simplicity we ignore this constraint in the inference.

As the θ elements are ordered in increasing order and are positive, the prior on the θ parameters is taken to be the joint distribution of K order statistics drawn from independent gamma random variables, such that

$$h_0(\theta|K) = K! \prod_{i=1}^{K} h_0(\theta_i|\alpha,\beta), \text{ with } 0 < \theta_1 < \cdots < \theta_K,\ K \geq 1, \tag{7.4.3}$$

and where $h_0(x|\alpha,\beta)$ is the Gamma(α,β) density with shape and scale $\alpha,\beta > 0$ respectively.

Consider the prior for the skill levels clustering. The set of skill cluster allocations has distribution $Y = \{y_{\{i\}}|\omega_A, A \sim \text{multinomial}(1,\omega_A), \forall i \in \{2,\ldots,n\}\}$, where ω_A has domain on $\{-A_-,\ldots,A_+\}$. The distribution of $y_{\{i\}}|(\omega_A,A)$ is assumed independent over all objects $i \in \mathcal{I}\setminus\{1\}$ such that

$$f(Y|\omega_A,A) = \prod_{i\in\mathcal{I}\setminus\{1\}} f(y_{\{i\}}|\omega_A,A),$$

where

$$\omega_A = \{\omega_{A,s} \in [0,1] : s \in \{-A_-,\ldots,A_+\},\ \sum_{s=-A_-}^{A_+} \omega_{A,s} = 1\}.$$

Again, $\omega_A|(A,\gamma_A) \sim$ Dirichlet$(\bar{\gamma}_A)$ is modelled to come from an $A+1$ dimensional Dirichlet prior distribution, with $\bar{\gamma}_A = \gamma_A 1_{A+1}$ where $\gamma_A \in \mathbb{R}_+$. Marginalising out as in derivation (7.4.2), another Dirichlet-multinomial allocation prior is obtained by integrating ω_A over the $A+1$ dimensional simplex. The prior density for the skill allocations is therefore given as

$$f(Y|\gamma_A,A) = \frac{n!}{\prod_{a=-A_-}^{A_+}|c_a|!} \frac{\Gamma(A\gamma_A)}{\Gamma(\gamma_A)^A} \frac{\prod_{a=-A_-}^{A_+}\Gamma(\gamma_A+|c_a|)}{\Gamma(A\gamma_A+n)}, \tag{7.4.4}$$

because $\sum_{a=-A_-}^{A_+} |c_a| = n$, where $c_a := \{i,\ \forall i \in \mathcal{I} \setminus \{1\} : y_{\{i\},a} = 1\}$ is the set of objects belonging to skill cluster $a \in \mathcal{A}$.

The prior for the number of unknown skill levels A is taken to be a truncated Poisson distribution with parameter (λ_A), $\lambda_A > 0$ with probability mass function

$$g_A\left(a|\lambda_A\right) = \frac{\lambda_A^a}{a!}\left(\sum_{i=0}^{n-1}\frac{\lambda_A^i}{i!}\right)^{-1} \quad a = 0, 1, \ldots, n-1.$$

Similarly to θ, the prior choice for ϕ is taken to be the joint distribution of order statistics of independent and identically distributed $A+1$ Gaussian random variables such that

$$\pi(\phi|A) = (A+1)! \prod_{a\in\mathcal{A}\setminus\{0\}} \pi(\phi_a) \text{ for } \phi_{A_-} < \cdots < \phi_0 < \ldots \phi_{A_+},$$

where $\phi_a \sim \mathcal{N}\left(0, \nu_A^2\right)\ \forall a \in \mathcal{A} \setminus \{0\}$, and with $\nu_A \in \mathbb{R}_+$. The $(A+1)!$ term arises as ϕ_0 can occur anywhere in the sequence of ϕ.

In summary, the prior $\pi\left(\phi, Y, A, \theta, Z, K\right)$ is equal to

$$(A+1)!\left[\prod_{a\in\mathcal{A}_{\{-0\}}} \pi\left(\phi_a\right)\right] f(Y|\gamma_A, A) g_A\left(A|\lambda_A\right) K!\left[\prod_{i=1}^{K} h_0\left(\theta_i|\alpha,\beta\right)\right] f(Z|\gamma_K, K)\ g_0\left(K|\lambda_K\right),$$

where, $\lambda_K, \lambda_A, \gamma_K, \gamma_A, \nu_A$, and α, β are the hyper-parameters.

7.4.3 Reversible jump Markov chain Monte Carlo sampler

Inference is made via a reversible jump Markov chain Monte Carlo sampler (Green, 1995), which provides samples from the posterior distribution $\pi\left(\phi, Y, A, \theta, Z, K|x\right)$, that is, the intransitivity and skill levels, the allocations to the these levels, and the number

of levels. Since the number of skill and intransitivity levels (A, K) are assumed to be unknown, the uncertainty in these parameters must be accounted for, thus motivating the use of a reversible jump sampler. In a sense, the reversible jump sampler mixes over models as well as parameters, and thus fully accounts for this uncertainty in the final inference.

The ICBT model is structured to try to favour the Bradley-Terry model as a special case, and this is reflected in our sampler, by explicitly incorporating the completely transitive cluster θ_0 as an ever present cluster, even if no pairs are allocated to this cluster at a given iteration of the sampler. To ensure the skill and intransitivity levels both remain ordered, the updates to these levels occur in a transformed space such that no update can lead to a change in order.

The reversible jump algorithm used is a split-merge sampler (Green and Richardson, 2001), which is adapted from the work of Ludkin (2020). The sampler comprises three separate moves: a standard Markov chain Monte Carlo Metropolis-Hastings move, which samples parameters ϕ, θ, and reallocates clusters Y, Z; splitting or merging clusters; and adding or deleting empty clusters. For full details of the construction of the algorithm and its implementation using base R (R Core Team, 2020), see the supplementary material.

7.4.4 Model assessment

The inferences produced by any model are only meaningful if the model itself is accurate. This accuracy is measured here by how well the model fits out of sample. If $\mathcal{C}$ is the set of total observed pairwise comparisons, then let $\mathcal{C}_s$ be the set of comparisons on which the model is fitted and $\mathcal{C}_t$ be the set of comparisons on which the model performance is analysed, such that $\mathcal{C}_s \cup \mathcal{C}_t = \mathcal{C}$ and $\mathcal{C}_s \cap \mathcal{C}_t = \emptyset$. We use log-loss $l(x^*)$ of the test dataset x^* to measure model performance, which we take to be the average negative

log-likelihood per observation in x^*, i.e.,

$$l(x^*) = -\frac{1}{|\mathcal{C}_t|} \sum_{c \in \mathcal{C}_t} \left[x_c \log(\hat{p}_{i_c j_c}) + (1 - x_c) \log(1 - \hat{p}_{i_c j_c}) \right], \tag{7.4.5}$$

where $\hat{p}_{ij}$ is the point estimate of p_{ij} based on the training dataset of comparisons $\mathcal{C}_s$ and $x^* := \{x_c : c \in \mathcal{C}_t\}$ is the set of test data, where the notation is as used in the expression for the likelihood (7.4.1).

7.4.5 Simulation study

The model was tested using simulated datasets where the number of objects, the number of round-robin tournaments, and the amount of intransitivity varied between the datasets. The sensitivity of our model to these parameters was then tested by comparing out of sample prediction accuracy with a standard Bradley-Terry model. This provided insights into the amount of data, and the amount of intransitivity, required for our more complex model to outperform the Bradley-Terry model. A full analysis is provided in the supplementary material.

7.5 Baseball Data

7.5.1 Data

Baseball was chosen to illustrate the methodology due to the high frequency of games, with accessible data for the American League Baseball obtained from www.retrosheet.org. The data are from the 2010-2018 seasons, with the 2010-2012 seasons involving 14 teams, and the 2013-2018 seasons involving 15 teams due to the Houston Astros moving from the National League to the American League. We analyse each season's data separately here, and jointly over years in the supplementary material.

The tournament structure is not as simple as the round robin tournament we con-

sidered in the simulation study. The American League is split into three divisions based on location: East, Central and West, with five teams in each (since 2013). Within the same division, pairs of teams play each other approximately 20 times, and pairs of teams from different divisions play each other around 5-7 times, as well as any Playoffs and World Series matchups, totalling around 140-160 matches per team every season, depending on the season and the team. Baseball is known to be a highly strategic game, with issues such as player selection, handedness of the of batters, strength and speed of players, and tactics such as "small ball" vs "long ball" all considered of great importance. So we anticipate that the level of intransitivity will be high.

The vast majority (at least 99.5%) of all matches are played at the home of one of the two teams competing in the game, with the rest played at neutral venues. Playing at home is well known to have the potential to increase the probability of the home team winning the match across a range of sports (Dixon and Coles, 1997). Although prediction and model interpretation could be improved by incorporating this effect, we decided not to address home advantage here. Our reason was that none of the existing intransitivity models have such a feature, as they were developed for applications devoid of home advantage, such as e-sports, so a comparison of the different models would only be fair if we did not include this property. However, in Section 7.6 we formulate the home advantage adaptation given its potential interest.

If pairs of teams do not play equally home and away, then ignoring home advantage could lead to misinterpretation of the estimated ICBT model parameters, e.g., if team i mostly played team k with team i at home, the home advantage would feed into θ_{ik}. We do not believe this is problematic due to the near perfect balance of home to away matches per team, and the maximum home percentage within pairs of teams is 70% for 2010-11 and only 57% subsequently.

7.5.2 Inference

The baseball data are analysed using the ICBT model, and its results are compared with those of the Bradley-Terry model and with the existing models of Section 7.2 except for the model of Duan et al. (2017) due to the subjective choices required for some parameters.

The ICBT model incorporates uncertainty in the choice of model itself, that is, the number of clusters and therefore how many parameters. Our prior distributions for number of intransitivity levels K and skill levels A are shown in Figure 7.5.1. We used a Poisson($\lambda_K = 2$) prior for K, with the hyper-parameter to give a 95% prior chance that $K \in [0, 5]$, as it was thought that there would only be a few different pairwise strategies. Similarly, the prior for A was taken to be Poisson($\lambda_A = 7$) as this hyper-parameter choice gave a 95% prior chance that $A \in [2, 13]$. The justification for our choice of the other hyper-parameters $(\gamma_K, \gamma_A, \alpha, \beta, \nu_A)$ and a sensitivity analysis to hyper-parameter choice is reported in the supplementary material.

Now consider the posterior distributions for K and A based on the 2018 season data, also shown in Figure 7.5.1. Despite the prior only providing vague information across values of $K \leq 5$, the data clearly favours having a single intransitivity level, meaning three possible clusters for each pair: a positive level, the completely transitive level, and the mirrored negative level. Further, although the prior gave a 14% probability to the Bradley-Terry model ($K = 0$), the posterior probability for that model is estimated to be zero, showing strong evidence of intransitivity in the dataset. For the distinct skill levels the change from prior to posterior is relatively small, with a mean (and 95% credible interval) of 7.94 $(4, 12)$, with the number of distinct skills levels favouring $A \in [6, 9]$.

The posterior estimated values of the teams' skills, and the variance of these values in particular, provides a helpful summary of how competitive the tournament is in a season, with the smaller the variance the more closely contested the tournament. For

our model the skills' variance ranged from 0.027 (2012 season) to 0.32 (2018 season) and for the standard Bradley-Terry model, 0.031 (2015 season) to 0.23 (2018 season) - both models suggesting that 2018 was the least competitive season. In the 2013 season the variance of skill levels according to Bradley-Terry is almost 3 times that of our model. However, the 2013 season was found to contain a particularly large amount of intransitivity, indicating that the large range of skills in the Bradley-Terry model could be the result of compensating for an inability to express the intransitivity. This has perhaps resulted in the Bradley-Terry model concluding that the 2013 season was less competitive than it was in reality.

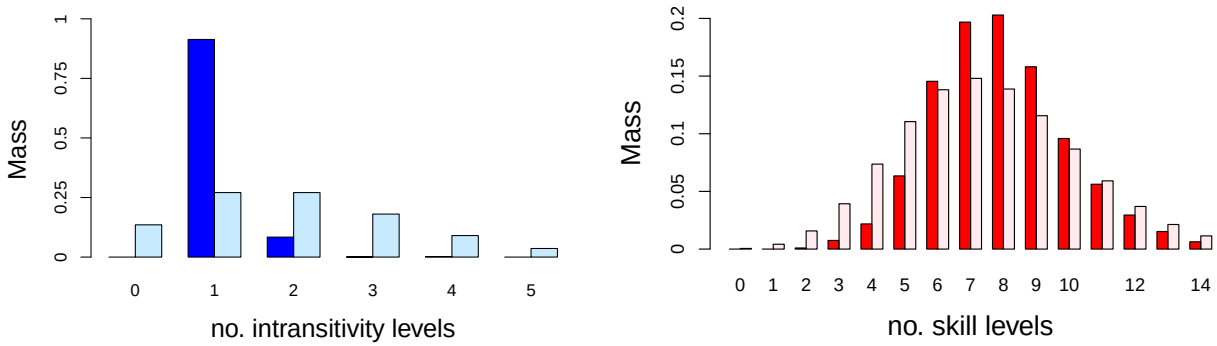


Figure 7.5.1: Posterior distributions of the K intransitivity levels (left) and the A skill levels (right) for the 2018 season: with the associated prior distributions in a lighter colour.

Now consider the pairwise interactions between teams. These interactions could be inferred from either the *intransitivity of the posterior mean* $\hat{\theta}^*_{ij}$, $\forall i \neq j$, or by the *posterior mean of the intransitivity parameter* $\hat{\theta}_{ij}$, $\forall i \neq j$. The supplementary material contains a comparison for both and concludes that $\hat{\theta}^*_{ij}$ is more meaningful and interpretable here, so we focus on that. For the 2018 season, Figure 7.5.2 (left) shows $\hat{\theta}^*_{ij}$, for each pair of teams $i > j \in \mathcal{I}$: recall that intransitivity has rotational symmetry, i.e., $\theta^*_{ij} = -\theta^*_{ji}$, $\forall i \neq j$. The teams are sorted by their rank according to $p_.$, given by

definition (7.3.9), see Figure 7.5.2 (right). Reading from the teams on the y-axis to x-axis there is a large positive value of intransitivity from Baltimore (BAL) to Tampa Bay (TBA) of 0.78 with 95% credible interval $(0.36, 1.22)$, indicating that Baltimore played better against Tampa Bay than expected, given their overall abilities. This is consistent with the data, with Baltimore winning 11 out of 19 matches between the two teams, despite being ranked lower.

The analysis of these intransitivities between pairs, and that of the skills of each team, can be combined to produce an overall ranking of the teams. As discussed in Section 7.3.3, with further details in the supplementary material, $p_{\cdot}$ provides a suitable ranking of the teams. For the 2018 season Figure 7.5.2 (right) shows the ranks according to $p_{\cdot}$ compared to the Bradley-Terry ranks. Both have been linearly scaled to help with a visual comparison, such that the best and worst teams have abilities 1 and 0 respectively. The two sets of estimated rankings using $p_{\cdot}$ are clearly correlated; however, there is some difference in the ordering of the ranks, indicating that intransitivity may have been masking the true ranks of some teams. For example, consider Tampa Bay, ranked 6th by the Bradley-Terry model. Tampa Bay's good record against Kansas City (KCA) has a much lower weighting than their poor record against Baltimore in the Bradley-Terry model due to the differing frequency of these match-ups, and therefore impacts the overall rank of Tampa Bay. The ICBT model however, recognises that good or bad records against particular teams could be due to the presence of intransitivity, and therefore penalises Tampa Bay less overall, ranking them 5th, thus illustrating our point in Section 7.1 that the ICBT model makes adjustments for tournament imbalance. Similar plots and inferences are drawn from the other seasons (2010-2017) but with different team rankings in each year.

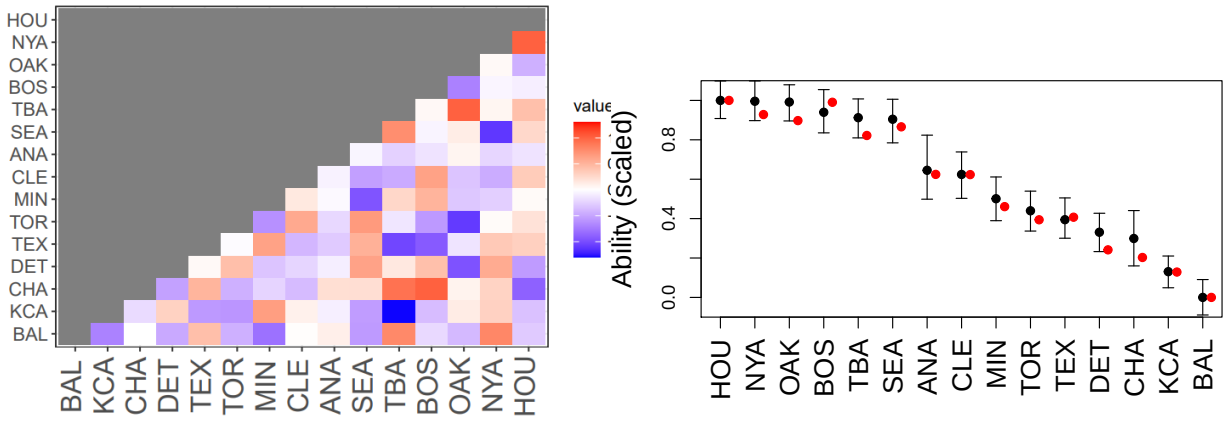


Figure 7.5.2: Analysis of 2018 season: the posterior mean of the intransitivity parameter, $\hat{\theta}^*_{ij}$ across all pairs of teams $i > j \in \mathcal{I}$ (left); ranking according to definition (7.3.9) (black) and Bradley-Terry model (red) for all teams $i \in \mathcal{I}$ (right).

7.5.3 Model Performance

To test the model performance, 70% of games from each season are randomly selected to be training data, on which the model is fitted, with the remaining 30% used as test data, on which the log-loss score is calculated. This random selection is appropriate as none of the models compared take time-dependency into consideration, a feature discussed in the supplementary material. The variation due to this random sampling in the training-test split is accounted for by taking 100 separate random training-test splits for each season. For each replicate of training data the model is fitted separately to each season's data. Relative log-loss is then calculated by subtracting the log-loss of a baseline coin tossing model.

Table 7.5.1 shows these negative relative log-loss scores for all years of baseball data, along with 95% confidence intervals, with this measure evaluated for the ICBT, Bradley-Terry, blade-chest and majority vote models. Since a larger value of negative relative log-loss indicates better model performance, a positive value indicates an improvement on the coin tossing model. So all four models improve on simply using coin tossing, showing that there is information to be exploited for inference and pre-

diction. The Bradley-Terry, blade chest and majority vote models all have somewhat similar performance to each other across the years, with the most improved fit being in 2018. In contrast our model is the best performing out of the four models in terms of out-of-sample prediction on all years of data. When assessed as the cumulative improvement over years, relative to coin tossing, the ICBT model is 2.8 times better than the Bradley-Terry model, showing that we have substantially improved predictive performance. The difference in log-loss scores relative to the Bradley-Terry model is largest for the 2013 season, which in Section 7.5.2 has been identified as the season with the largest intransitivity.

year	ICBT	BT	blade-chest	majority vote
2010	44(38, 46)	17(15, 18)	17(2, 27)	20(13, 25)
2011	46(33, 49)	15(13, 17)	17(-1, 26)	20(13, 26)
2012	49(44, 53)	14(11, 16)	24(8, 33)	22(14, 31)
2013	64(36, 69)	23(21, 25)	33(13, 42)	31(22, 39)
2014	39(29, 45)	9(7, 11)	10(-12, 21)	13(6, 19)
2015	34(12, 45)	5(2, 7)	9(-14, 17)	9(1, 16)
2016	42(32, 55)	10(8, 12)	18(2, 30)	18(10, 28)
2017	36(10, 50)	13(11, 15)	13(-4, 22)	16(9, 22)
2018	73(65, 79)	46(44, 47)	48(19, 56)	51(43, 57)

Table 7.5.1: Negative relative log-loss $\times 10^3$ (compared to a coin-tossing model) for each year of baseball data for the ICBT, Bradley-Terry, blade-chest and majority vote models. 95% confidence intervals, in parentheses, come from random training-test splits of the data.

7.6 Conclusions and Discussion

We have proposed a new model and inference structure for paired comparison data. We frame this in the context of sport competitions, baseball in particular, with *teams* competing against each other, though the potential applications of the model are much broader. Our proposed model, the Intransitive Clustered Bradley-Terry (ICBT) model, extends the standard Bradley Terry model, which is widely considered as the baseline model for such data. The extension allows for intransitivity so that the difference in *skill* levels between two objects being compared is not the only factor affecting the probabilities of the outcomes. There are a number of models which already allow for intransitivity, but each of these are quite restricted in the parametric form of intransitivity relative to our semi-parametric approach, which recognises that certain patterns of interaction between pairs of objects can be common over multiple pairs. Our model also allows for objects' skills to be clustered, a feature that is novel to paired comparisons, with this inducing parsimony and avoiding obtaining distinct rankings for some items when there is no evidence from the data that they are not equally good. We have shown evidence from American League baseball that our model provides a distinct improvement on existing models.

The ICBT model has complete flexibility, in the sense that cluster allocation to skill and intransitivity levels is not predetermined. In order that the data identify the appropriate structure of clustering, and for the inference to account for the uncertainty in this choice, the model is fitted via RJMCMC.

Based on the clusters with the highest posterior probabilities, we anticipate that experts in the particular sport may be able to identify some patterns of clustering that are interpretable, e.g., associated with different styles of play. In such cases, these clustering features could be hard wired into the model as the only options, resulting in more efficient inference. A referee made the helpful suggestion that if accounting for clustering uncertainty was not an issue then the inference could be simplified by

estimating the ICBT model with group lasso penalties to induce clusters. We feel that our model works sufficiently well for the current applications but agree that it presents an exciting springboard for the consideration of various extensions to the model and its inference. We finish by illustrating a few such possible extensions.

In Section 7.5 we did not attempt to account for home advantage, which is widely recognised as an important feature in sport, e.g., Cattelan et al. (2013) incorporate it in a Bradley-Terry model, though to the best of our knowledge it has not been accounted for in the existing intransitivity models. The most natural way to achieve this is to change p_{ik} given by expression (7.3.5) to a probability $p_{ik}^{(i)}$ of the home team i beating the away team k, with

$$p_{ik}^{(i)} = \frac{1}{1 + \exp[-(\theta_{ik} + \gamma + r_i - r_k)]}, \text{ and } p_{ki}^{(i)} = 1 - p_{ik}^{(i)} \; \forall i \neq k \in \mathcal{I}, \tag{7.6.1}$$

where $\gamma \in \mathbb{R}$ determines the effect of playing at home, which here is common over all pairs of teams. If $\gamma > 0$ ($\gamma < 0$) then the probability of a home win is increased (decreased) relative to the other factors of skill and intransitivity. This effect can be extended to vary over teams by replacing γ by γ_i in expression (7.6.1). To ensure these γ_i parameters are all identifiable, we fix $\gamma_1 = 0$, though no additional constraints are needed if there is a common γ, but that is all that is required under the conditions of Proposition 1 on the other parameters, as we are able to exploit data that distinguishes which team is at home.

This article only considered win-loss scenarios. Extensions of the Bradley-Terry have been proposed for handling draws. Two distinct methods for handling draws are given by Cattelan et al. (2013) and Hankin (2020). The former use ordinal logistic regression, treating win, loss, draw as outcomes of an ordered multinomial random variable, which can then be analysed via an ordered link model. In contrast, the latter treats the problem as a competition between the two teams and a third theoretical team, such that when the theoretical team wins the outcome of the match corresponds

to a draw between the two actual teams. The ICBT model can be adapted similarly, with the use of the clustering strategy extended to pooling teams to account for their similar cautiousness, leading to them drawing more often than would be expected.

We have assumed that all teams play each other. If this is not the case we cannot improve on the prior inference for the θ_{ik} parameters for pairs (i, k) that do not play each other. This is not a restriction for Bradley-Terry or the existing intransitivity models, where the associated p_{ik} are determined by the observed pairs. This raises issues about identifiability of the ICBT model parameters. Our approach, through Proposition 1, is no longer sufficient leaving the open problem of which parameters to fix in order to give the most efficient inference.

Chapter 8

Conclusions and Further Work

8.1 Chapter Summary and Conclusions

Chapter 6 was a whistle-stop tour of the most commonly utilised ranking methods, and included: one-object-one-rating methods, with the seminal Bradley-Terry model being the prime example; indirect approaches like the Poisson models; and heuristics like the well-known Elo system. A full exploration would also include Markov models, such as PageRank (Brin and Page, 1998), and the closely linked Massey (Massey, 1997) and Colley (Colley, 2002) methods.

The paired comparison literature is extraordinarily broad, and each of the above methods have had numerous extensions proposed. Often motivated by sports, many of these extensions are application specific, encompassing, for example, injuries, transfers or referee bias. But this thesis is not concerned with the application specific angle, rather, it strives for a ubiquitous framework for statistical ranking methods. Accordingly, the literature review in Chapter 6 was kept general, presenting the core methods. More depth was permitted to discussing the often-unknowingly-surmised transitive assumption, and the seminal works are presented in this vein, highlighting the three common forms of stochastic transitivity. The Bradley-Terry type models are shown to

obey linear transitivity. The Poisson models flout both linear and strong stochastic transitivity, adhering only to weak stochastic transitivity, while few models contain the flexibility to defy weak stochastic transitivity.

Chapter 7 lays out the blueprint for modelling intransitivity-rife data via a generalisation of the Bradley-Terry model, the Intransitive Clustered Bradley-Terry (ICBT) model. The ICBT model unifies all three forms of transitivity within a single class of semi-parametric model, from adhering to the strictest form to violating the weakest, and with this flexibility governed by the data. From this blueprint, similar extensions are trivial for other one-object-one-rating systems. The same initiative could be translated for the Poisson models, providing them too with the opportunity to violate weak transitivity.

Next, some thoughts on further work are presented, both for the ICBT model, and more general adaptations for the Bradley-Terry model.

8.2 Further Work

8.2.1 ICBT model improvements

For the application at hand, it may be known that some forms of transitivity must be upheld. In these cases the ICBT model could be constrained to these forms, but remain flexible within this domain. For example, it may be known a priori that a system comprising objects $\mathcal{I}$ should adhere to strong stochastic intransitivity. In this case, the ICBT model can be fitted to data from this system, subject to prior distributions on the ratings and intransitivity parameters $\mu := \{\mu_i \in \mathbb{R} : i \in \mathcal{I}\}$ and $\theta := \{\theta_{ij} \in \mathbb{R} : i \neq j \in \mathcal{I}\}$, respectively, which are selected to enforce the strong stochastic transitivity constraint (6.2.1) for all $(i \neq j \neq k) \in \mathcal{I}$, i.e.,

$$p_{ik} - \max\{p_{ij}, p_{jk}\} > 0 | p_{ij}, p_{jk} > 0.5, \ \forall (i \neq j \neq k) \in \mathcal{I}$$

with $p_{ij} = f(\mu_i, \mu_j, \theta_{ij})$ as defined in Chapter 7. This would result in an ICBT model which is able to explore a range of intransitivities but subject to satisfying strong stochastic transitivity. This extension hinges on finding joint prior distributions for all of μ, θ which satisfy this constraint.

Chapter 7 touches upon connectivity, or tournament structure, and its impact on the parameter uncertainty. Within a symmetry structure, i.e., round-robin, parameter uncertainty must be invariant to the choice of constraint, given the constraints are sufficient. But in the absence of symmetry it is not clear which constraints minimise the total parameter uncertainty. The constraint on the intransitivity parameters are selected as $\theta_{1j} = 0, \forall j \in \mathcal{I} \setminus \{1\}$, i.e., all comparisons involving object one have intransitivity 0. So long as the choice of parameters forms a tree on the graph, any choice of constraint is valid; however, the uncertainty in the free parameters is not necessarily the same.

This motivates the need to determine the optimal choice of intransitivity constraints $\theta_c := \{(i,j) : \theta_{ij} = 0, i > j \in \mathcal{I}\}$ for a given tournament structure, such that uncertainty in the system is minimised. Uncertainty can be quantified via the variance in the estimates $\{\hat{\mu}_i : i \in \mathcal{I}\}$, $\{\theta_{ij} : (i,j) : i \neq j \in \mathcal{I}^2\}$ of the parameters μ and θ, respectively. Alternatively, since the pairwise preference probabilities are arguably more interpretable than the parameters themselves, the uncertainty can be quantified more directly via the variance in estimates $\hat{p}_{ij} = f(\hat{\mu}_i, \hat{\mu}_j, \hat{\theta}_{ij})$ of the preference probabilities p_{ij}, for all pairs $(i,j) : i \neq j \in \mathcal{I}^2$. Denoting n_{ij} to be the number of comparisons between objects (i,j), then the variance in $\hat{p}_{ij}$ is known to be approximated by

$$\text{Var}\left(\hat{p}_{ij}\right) = \frac{\hat{p}_{ij}(1-\hat{p}_{ij})}{n_{ij}},$$

from the binomial assumption. The contribution to the total variance, or uncertainty, of each pair (i,j) comprises the contributions from the variances of the parameters,

$\text{Var}(\hat{\mu}_i)$, $\text{Var}(\hat{\mu}_j)$, and $\text{Var}\left(\hat{\theta}_{ij}\right)$. As the θ_{ij} parameter depends only on comparisons between the pair (i, j), then the expected value of $\text{Var}\left(\hat{\theta}_{ij}\right)$ is inversely proportional to the number of comparisons n_{ij}. Regard the graphical interpretation of the system, Figure 6.1.1, with $\mathcal{I}$ the set of nodes, and comparisons between objects representing the edges where the weight of an edge between two nodes (i, j) is equal to the number of comparisons n_{ij} for all $(i, j) : i \neq j \in \mathcal{I}$. Then, for a fixed tournament structure, it is here hypothesised that the choice of constraints θ_c which minimise the total system uncertainty system corresponds to the minimum spanning tree of the graph. Intuitively, only pairs which are compared more frequently, and therefore who's relationships are better informed by the data, have free intransitivity parameters.

8.2.2 Bradley-Terry Reparametrisation

Some systems naturally contain more stochasticity than others. Consider your chances of getting a higher return on the stock market than an expert stockbroker, versus beating a chess grand master. The ratio of skill to sheer luck varies greatly from one system to the next.

For a system comprising the set of objects $\mathcal{I}$, the Bradley-Terry model uses ratings $\mu := \{\mu_i \in \mathbb{R} : i \in \mathcal{I}\}$ to model the pairwise-preference probabilities, and it reflects the *stochasticity* in the system by the *spread* of the ratings μ. For a highly stochastic system, such as a system of stockbrokers, the ratings will be close together, so that even objects with large differences in rankings have preference probabilities that are close to 0.5. Conversely, in near deterministic systems, the ratings will have a large spread, meaning preference probabilities are close to 0 or 1, even for similarly ranked pairs of objects. The spread of the ratings is rarely considered, but offers a useful insight into an important property of the system at hand, which the following reparametrisation attempts to formalise.

The Bradley-Terry model often constrains either $\sum_{i \in \mathcal{I}} \mu_i = 0$, or $\mu_1 = 0$ for pa-

rameter identifiability. Here the former is selected, and further, it is proposed to also constrain $\sum_{i\in\mathcal{I}}\mu_i^2 = 1$, i.e., fixing

$$\mu_n = -\sum_{i=1}^{n-1}\mu_i, \quad \mu_{n-1}^2 = 1 - \sum_{i\in\mathcal{I}\setminus\{n-1\}}\mu_i^2,$$

so that the ratings μ are then standardised with mean zero and variance one. Then, the Bradley-Terry model is reparametrised as

$$p_{ij} = \left\{1 + \exp\left(-\left[\frac{\mu_i - \mu_j}{\beta}\right]\right)\right\}^{-1}, \ \beta \in \mathbb{R}_+, \ i \neq j \in \mathcal{I}, \tag{8.2.1}$$

where β is termed the *stochasticity* parameter. As $\beta \to \infty$, then $p_{ij} \to 0.5$, and so the system noise completely dominates any possible difference in the ratings. When $\beta \to 0$, then $p_{ij} \to \mathbb{1}\{\mu_i > \mu_j\}$, where $\mathbb{1}$ is the indicator function, meaning that the system is completely deterministic, with the better rated object always being preferred. This reparametrisation could help mixing when using MCMC for inference in the Bayesian setting. The traditional parametrisation of the Bradley-Terry model requires each parameter update to reflect changes in both pairwise preference and system stochasticity, whereas model (8.2.1) orthogonalises the stochasticity, allowing for potentially more efficiency in estimating the μ parameters.

8.2.3 Volatility

What this reparametrisation reveals, is that the *source* of the stochasticity is assumed to be only due to the system, i.e., it is exogenous, and that it is common for all objects. It assumes the objects to be perfectly consistent, and that each comparison reflects the true fixed ability of each object.

But for many applications, in each paired comparison, the objects have some *performance* which is some noisy realisation of a latent ability. This variation around the

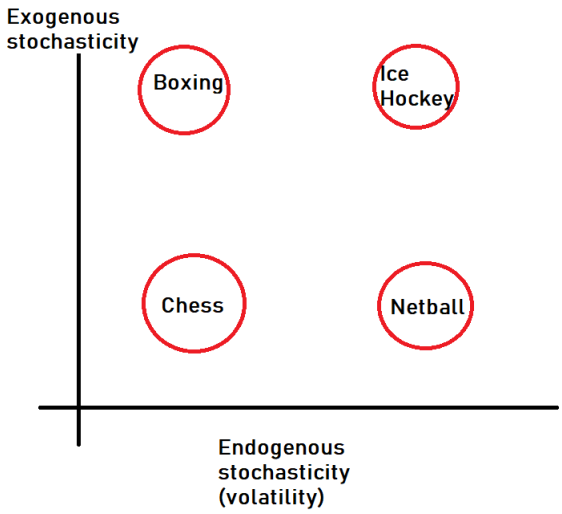


Figure 8.2.1: A toy example of the sources of stochasticity in different sports. The team element of ice hockey and netball introduces volatility, as the team changes from one game to the next, whereas boxers may exhibit more consistent performances. The low-scoring nature of ice hockey introduces exogenous stochasticity, as one freak goal can have a large impact. Similarly, one lucky punch can end a boxing match. Chess has no team aspect and is highly skill based, making it the least stochastic system overall.

latent ability introduces *endogenous* stochasticity, which is termed *volatility*, and which contributes to the total stochasticity of the system. Importantly, the volatility can be different for each object.

Figure 8.2.1 considers the total system stochasticity of four popular sports. Fitting reparametrisation (8.2.1) to data from these four systems may result in similar stochasticity parameters β, for both boxing and netball; however, this neglects the source of the stochasticity. In the toy example, boxing has high exogenous and low endogenous stochasticity, whereas netball has low exogenous and high endogenous stochasticity.

Volatility can be dealt with formally by defining the *performance* of any object $i \in \mathcal{I}$ in any given comparison $c \in \mathcal{C}_i$, where $\mathcal{C}_i$ is the set of comparisons involving object i, to be a random quantity given as $r_{i,c} := \mu_i + Z_{i,c}$, where $Z_{i,c}$ is some zero-mean random

variable with variance $\epsilon_i^2 \in \mathbb{R}_+$, and assuming

$$Z_{i,c} \perp\!\!\!\perp Z_{j,c},\ \forall c \in \{\mathcal{C}_i \cap \mathcal{C}_j\},\ i \neq j \in \mathcal{I}, \qquad Z_{i,c} \perp\!\!\!\perp Z_{i,c'},\ \forall c \neq c' \in \mathcal{C}_i, i \in \mathcal{I}.$$

That is, the noise term is independent over all objects and from one comparison to the next. For an object i, ϵ_i is then the *volatility* of object i. For the above example, fitting this model to results of boxing matches between a set of boxers $\mathcal{I}$ might give estimates of $\{\epsilon_i \approx 0 : i \in \mathcal{I}\}$, but large β, whereas, if using data from a set of netball teams, the model may estimate β to be small, but the estimates of $\{\epsilon_i : i \in \mathcal{I}\}$ to be large. Despite this, the model still allows for the estimates of pairwise probabilities to be similar between the two datasets.

Consider a particular comparison $c \in \{\mathcal{C}_i \cap \mathcal{C}_j\}$ involving objects i and j, and for ease of notation, the subscript c here is dropped without loss of generality, so that the noise terms of objects i and j in comparison c are denoted $Z_{i,c} = Z_i$ and $Z_{j,c} = Z_j$, respectively. Then the expectation of the probability that i is preferred to j in this comparison is given as

$$\mathbb{E}\left[p_{ij}|\mu_i, \mu_j, \epsilon_i, \epsilon_j, \beta\right] = \int_{z_i}\int_{z_j}\left(\left\{1 + \exp\left(-\left[\frac{\mu_i + z_i - (\mu_j + z_j)}{\beta}\right]\right)\right\}^{-1} f_{Z_i}(z_i) f_{Z_j}(z_j)\right) \mathrm{d}z_i\, \mathrm{d}z_j,$$

for all $i \neq j \in \mathcal{I}$, given suitable assumptions for the form of the random variables $Z_{i,c}$, such as

$$Z_{i,c} \sim \mathcal{N}(0, \epsilon_i^2),\ \forall c \in \mathcal{C}_i, i \in \mathcal{I}.$$

An interesting avenue for further work is to compare various relative systems in terms of their properties of endogenous and exogenous stochasticity, by comparing the fitted values of β and $\{\epsilon_i\}_{i\in\mathcal{I}}$ between systems. It could also be interesting to explore the properties of $\mathbb{E}\left[p_{ij}|\mu_i, \mu_j, \epsilon_i, \epsilon_j, \beta\right]$. Clearly, with increasing volatilities ϵ_i, ϵ_j, then for a

fixed β,

$$\mathbb{E}\left[p_{ij}|\mu_i, \mu_j, \epsilon_i, \epsilon_j, \beta\right] \to 0.5, \text{ as } \epsilon_i, \epsilon_j \to \infty.$$

The rate of this convergence is less obvious.

8.2.4 Statistical Learning

Section 8.2.1 considered how to minimise uncertainty in the ICBT model, given the graph structure as dictated from the data, as a function of the choice of parameter constraint. Now considering the standard Bradley-Terry model with fixed constraints as in parametrisation (8.2.1) we can consider the same problem from the opposite angle. Namely, given objects $\mathcal{I}$, how can the system uncertainty be minimised as a function of the graph structure? *Minimising uncertainty* has multiple interpretations, i.e., minimising the uncertainty in the ranking, or minimising the total uncertainty in the parameters. This depends on the choice of loss function. Two viable examples are provided.

Define the set of random variables $X(\mathcal{C}_m) \in \{0,1\}^m$ to represent the possible outcomes of a set of $m \geq |\mathcal{I}| - 1$ paired comparisons $\mathcal{C}_m \in (i,j) : i \neq j \in (\mathcal{I} \times \mathcal{I})^m$ from the set of objects $\mathcal{I}$. If $x(\mathcal{C}_m) \in \{0,1\}^m$ are then the realisations from $X(\mathcal{C}_m)$, which has probability mass function governed by the Bradley-Terry model given μ, then let $\mathcal{G}_{\mathcal{I},X(\mathcal{C}_m)}$ define the random graph representing all possible graphs from this model, and $\mathcal{G}_{\mathcal{I},x(\mathcal{C}_m)}$ is a realisation of this system. Furthermore, $\mathcal{C}_m$ is constrained to form a tree on any graph $\mathcal{G}_{\mathcal{I},x(\mathcal{C}_m)}$, thus satisfying the minimal identifiability constraints for the Bradley-Terry model.

Then, define $\hat{R}_i|x(\mathcal{C}_m) \in \{1,\ldots,|\mathcal{I}|\}$ to be the estimated rank of object i given the observed graph $\mathcal{G}_{\mathcal{I},x(\mathcal{C}_m)}$, which is given by ordering the estimates of the ratings

$$\hat{\mu}|x(\mathcal{C}_m) := \{\hat{\mu}_j|x(\mathcal{C}_m) : j \in \mathcal{I}\},$$

which are given from, for example, maximum likelihood estimates. Assuming some true latent ranking $R_i \in \{1, \ldots, |\mathcal{I}|\},\ \forall i \in \mathcal{I}$ of the objects, then

$$\Pr\left\{\left(\hat{R}_i|x(\mathcal{C}_m)\right) = R_i\right\}$$

is the probability that the estimated rank $\hat{R}_i|x(\mathcal{C}_m)$ of object i is equal to its true rank given the observed comparisons. One choice of valid loss function is

$$\mathcal{R}(x(\mathcal{C}_m)) := -\sum_{i\in\mathcal{I}} \log\left(\Pr\{(\hat{R}_i|x(\mathcal{C}_m)) = R_i\}\right),$$

and then the optimal set of comparisons $\mathcal{C}_m^*$ can be selected as

$$\mathcal{C}_m^* = \text{argmin}_{\mathcal{C}_m}\mathbb{E}_X\left[\mathcal{R}(X(\mathcal{C}_m))\right].$$

That is, the comparisons are selected so that in expectation the sums of the log probability that all objects are correctly ranked is maximised. The joint probability that all objects are correctly ranked could also be maximised.

For real applications, the true rank is likely unknown. In this case, the loss function could be constructed via the notion of entropy. Define the *entropy* of an object i, and of the whole system, respectively, to be

$$S_i(x(\mathcal{C}_m)) := \sum_{k=1}^{|\mathcal{I}|} \log\left(\Pr\{(\hat{R}_i|x(\mathcal{C}_m)) = k\}\right), \quad \mathcal{S}(x(\mathcal{C}_m)) = \sum_{i\in\mathcal{I}} S_i(x(\mathcal{C}_m)).$$

Then, the optimal set of comparisons is selected as

$$\mathcal{C}_m^* := \text{argmin}_{\mathcal{C}_m}\mathbb{E}_X\left[\mathcal{S}\left(X(\mathcal{C}_m)\right)\right].$$

Hence, the expected system entropy after observing the set of comparisons $\mathcal{C}_m^*$ is minimised.

In the absence of observed comparisons between objects or prior information about the objects' abilities, it seems intuitive that a symmetric comparison structure (round robin) would minimise this expected ranking uncertainty under both definitions from arguments of symmetry. But what about when observations from prior comparisons $x(\mathcal{C}_m)$, $m > 0$ are available? For some $m' \geq \max(|\mathcal{I}| - 1 - m, 1)$ further comparisons $\mathcal{C}_{m'}$, the optimal choice $\mathcal{C}^*_{m'}$ is defined as

$$\mathcal{C}^*_{m'}|x(\mathcal{C}_m) = \text{argmin}_{\mathcal{C}_{m'}} \mathbb{E}_X \left[\mathcal{S}(X(\mathcal{C}_{m'}), x(\mathcal{C}_m))\right],$$

where

$$\mathcal{S}\left[X(\mathcal{C}_{m'}), x(\mathcal{C}_m)\right] = \sum_{i \in \mathcal{I}} \sum_{k=1}^{|\mathcal{I}|} \log\left(\Pr\left\{\left[\hat{R}_i | X(\mathcal{C}_{m'}), x(\mathcal{C}_m))\right] = k\right\}\right),$$

i.e., $\mathcal{C}_{m'}$ is the optimal choice of m' comparisons which maximally reduces the system entropy in expectation, given the observations $x(\mathcal{C}_m)$.

It is clear that the choice of objects to compare will be dependent on the current estimates of μ, since objects with more similar rating estimates are more likely to be incorrectly ranked, therefore having a larger entropy gradient.

Glickman and Jensen (2005) approach this problem through Bayesian optimal design, defining the optimal pair to compare next, from the $n(n-1)/2$ possible comparisons, as that which maximises the gain in Kullback–Leibler information from the prior to the posterior distribution in expectation. Although Glickman and Jensen (2005) solves the problem for $m = 1$, this quickly becomes infeasible when m is large, due to the explosion of possible combinations to consider. The optimal choice of the set of next comparisons may also depend on the size of the horizon m, since a larger m allows for more exploration, that is, learning of the parameters μ, before exploitation of this information. This exploration-exploitation trade-off is a common theme in the *multi-armed bandit* literature, with Thompson sampling (Thompson, 1933) and the de-

terministic upper-confidence-bound algorithm (Auer et al., 2002) being two heuristic solutions. A foray into the realm of bandit literature could present Bradley-Terry-type ranking models with an array of untapped research.

Part II Appendices

Appendix C

Supplementary Material for Chapter 7

C.1 Introduction

This document accompanies Chapter 7. Section C.2 briefly outlines the main novelties in our inference method, Sections C.3, C.4 and C.5 contain full details of the implementation of the inference, specifically of the split-merge steps, the add-delete steps and the standard MCMC steps in the algorithm, respectively. Section C.6 describes the steps taken to help convergence of the algorithm. Section C.9 details the simulation experiments, and Section C.10.1 contains less central findings from the analysis of baseball data in Section 7.5 of the main article. Note that in this supplementary document, all vectors are denoted in bold for ease of reading, and correspond to the equivalent unbolded notation in the main article.

C.2 Algorithm: areas of key interest

Inference is made via a reversible jump Markov chain Monte Carlo sampler (Green, 1995), which provides samples from the posterior distribution $\pi(\boldsymbol{\phi}, \boldsymbol{Y}, A, \boldsymbol{\theta}, \boldsymbol{Z}, K|\boldsymbol{x})$,

that is, the intransitivity and skill levels, the allocations to the these levels, and the number of levels.

Since the number of skill and intransitivity levels (A, K) are assumed to be unknown, the uncertainty in these parameters must be accounted for, thus motivating the use of a reversible jump sampler. In a sense, the reversible jump sampler mixes over models as well as parameters, and thus fully accounts for this uncertainty in the final inference.

The ICBT model is structured to try to favour the Bradley-Terry model as a special case, and this is reflected in our sampler, by explicitly incorporating the completely transitive cluster θ_0 as an ever present cluster, even if no pairs are allocated to this cluster at a given iteration of the sampler. To ensure the skill and intransitivity levels both remain ordered, the updates to these levels occur in a transformed space such that no update can lead to a change in order.

The reversible jump algorithm used is a split-merge sampler (Green and Richardson, 2001), which is adapted from the work of Ludkin (2020). The sampler comprises three separate moves: a standard Markov chain Monte Carlo Metropolis-Hastings move, which samples parameters $\boldsymbol{\phi}$, $\boldsymbol{\theta}$, and reallocates clusters $\boldsymbol{Y}, \boldsymbol{Z}$; splitting or merging clusters; and adding or deleting empty clusters.

Let $(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s)$ be the current values of the parameters at step s in the sampler, and for any move a proposal $(\boldsymbol{\phi}', \boldsymbol{Y}', A', \boldsymbol{\theta}', \boldsymbol{Z}', K')$ is made. To be concrete,

$$\begin{aligned}
\boldsymbol{\phi}^s &:= \{\phi_a^s : a \in \{-A_-^s, \dots, A_+^s\}\} \\
\boldsymbol{Y}^s &:= \{\boldsymbol{y}_{\{i\}}^s : i \in \mathcal{I} \setminus \{1\}\} \\
\boldsymbol{\theta}^s &:= \{\theta_k^s : k \in \{1, \dots, K^s\}\} \\
\boldsymbol{Z}^s &:= \{\boldsymbol{z}_{\{i,j\}}^s : i > j \in \mathcal{I} \setminus \{1\}\},
\end{aligned}$$

for all steps $s \in \{1, \dots, S\}$ in the sampler where $S \in \mathbb{N}$, and where $\boldsymbol{y}_{\{i\}}^s$ is defined here

formally as

$$\boldsymbol{y}_{\{i\}} := \begin{cases} \left\{ y^s_{\{i\},a} \in \{0,1\} : a \in \mathcal{A},\ \sum_{a\in\mathcal{A}} y^s_{\{i\},a} = 1 \right\}, & \forall i \in \mathcal{I} \setminus \{1\}, \\ \left\{ y^s_{\{1\},a} \in \{0,1\} :\ a \in \mathcal{A} : y^s_{\{1\},0} = 1, y^s_{\{1\},a} = 0 : a \neq 0 \right\}, & \text{if } i = \{1\}, \end{cases}$$

that is, $y^s_{\{i\},a} = 1$ if object $i \in \mathcal{I}$ is allocated to skill cluster $a \in \mathcal{A}$ at step s, and is equal to 0 otherwise; and the skill of object $\{1\}$ is fixed to always be in cluster 0, i.e., $y^s_{\{1\},0} = 1$, for all $s \in \{1, \ldots, S\}$,. Remembering the identifiability constraints for the intransitivities outlined in the main text, see Proposition 1, the intransitivity allocations are formally defined as

$$\boldsymbol{z}^s_{\{i,j\}} := \begin{cases} \left\{ z^s_{\{i,k\},a} \in \{0,1\} : a \in \{-K, \ldots, K\},\ \ \sum_{a=-K}^{K} z^s_{\{i,k\},a} = 1, \right. \\ \qquad\qquad \left. \forall i > k \in \mathcal{I} \setminus \{1\} \right\}, \\ \left\{ z^s_{\{i,1\},a} \in \{0,1\} : a \in \{-K, \ldots, K\},\ \ z^s_{\{i,1\},0} = 1, z^s_{\{i,1\},a} = 0 : a \neq 0, \right. \\ \qquad\qquad \left. \forall i \in \mathcal{I} \setminus \{1\},\ \text{if } k = 1 \right\}, \end{cases}$$

that is $z^s_{\{i,k\},a} = 1$ if the pair $\{i, k\}$ is in the cluster labelled s, and $z_{\{i,k\},s} = 0$ otherwise.

As discussed in Section 7.4.2, the clustering of our two features - the skills of the objects and the intransitivity of the pairs of objects - are treated as independent systems. Therefore we approach them separately: all of Sections C.3.1, C.4.1 and C.5.1 outline the intransitivity components of the algorithm, and Sections C.3.2, C.4.2 and C.5.2 outline their skill component counterparts.

There are two key characteristics in our Intransitive Clustered Bradley-Terry (ICBT) model which require adaptations to be made to the standard split-merge sampler. Firstly, the vector of parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ must be ordered in order to maintain identifiability. Secondly, symmetry is imposed on the intransitivity parameters $\boldsymbol{\theta}$, such that $\theta_{ik} = -\theta_{ki}\ \forall i \neq k \in \mathcal{I}$. Additionally, to ensure efficient converge of the algorithm, a targeted initialisation routine was introduced as a precursor to the main RJMCMC algorithm. Novel aspects of these three elements are discussed below.

Ordering

To preserve the order of the parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, a suitable *matching function* y, see (C.3.1), appropriately bounds each parameter. The matching function then serves as a transformation to the real numbers, for example, for some $\theta_k \in \boldsymbol{\theta}$ with $\theta_{k-1} < \theta_k < \theta_{k+1}$, then $y : [\theta_{k-1}, \theta_{k+1}] \to \mathbb{R}$. In this transformed space, parameter updates can be made without concern for the bounds, before being transformed back to their bounded space via the inverse of the matching function, $y^{-1} : \mathbb{R} \to [\theta_{k-1}, \theta_{k+1}]$. In the case of the standard MCMC updates for θ_k, a random walk is used on the transformed space.

For the reversible jump moves, for example, a split of an intransitivity level, an *auxiliary variable* $u \sim \sigma\chi^2_1$, for $\sigma > 0$, where χ^2_1 has unit-degree Chi-squared distribution, is introduced, which is then added and subtracted to the transformed intransitivity level to give two new proposed levels on the transformed scale. After retransforming back to the original scale via inversion of the matching function, lower and upper proposed intransitivity levels are proposed from the split move with order preserved.

Note that the transitive cluster can be split, where the bounds for the matching function (C.3.1) are given as $\theta_0 \in [\theta_{-1}, \theta_1]$; however, in this case one of the proposed intransitivity levels must be 0, to maintain the generalisation to a Bradley-Terry model. That is, if the $k = 0$ cluster, the transitive cluster, is proposed to be split, then the lower proposed intransitivity level must be $\theta'_{0'} = 0$, but the upper proposed level $\theta'_{1'}$ follows the same rules as for the general case, equation (C.3.5).

Once the new levels have been proposed, allocations to these levels are then considered. Let $\boldsymbol{b}_k := \{\{i, j\},\ \forall i > j \in \mathcal{I} \setminus \{1\} : z_{\{i,j\},k} = 1\}$ denote the set of pairs $i > j \in \mathcal{I} \setminus \{1\}$ of objects which belong to the cluster-to-be-split, which for now we assume is not the transitive cluster. At this point target information could be included by using the approach of Zanella (2020); however, we found no significant efficiency gains. Therefore for simplicity the cluster allocations to the two proposed split levels are selected uniformly at random. The probability of the cluster allocation, given that

cluster $k \neq 0$ has been split, is given as

$$q(\boldsymbol{Z}') = \left(\frac{|\boldsymbol{b}_k|!}{|\boldsymbol{b}'_{k'}|!\ |\boldsymbol{b}'_{k'+1}|!} \right) 2^{-(|\boldsymbol{b}_k|)} \ \left(\frac{|\boldsymbol{b}_{-k}|!}{|\boldsymbol{b}'_{-k'}|!\ |\boldsymbol{b}'_{-(k'+1)}|!} \right) 2^{-(|\boldsymbol{b}_{-k}|)}$$

where the factorial terms appear since all possible combinations of allocations must be considered, and remembering that cluster $-k$ will also be split into clusters $-k'$ and $-k'-1$ with intransitivity levels $\theta'_{-k'} = -\theta'_{k'}$ and $\theta'_{-k'-1} = -\theta'_{k'+1}$ due to the symmetry of the intransitivity levels. The proposed number of clusters is of course $K' = K + 1$.

To split a *skill* level, the same form of matching function (C.3.1) can be used to maintain ordering. The skill level $\phi_0 = 0$ can be split, although in this case one of the proposed skill levels must also be 0, and object 1 must stay in this level. The proposed number of skill levels is $A' = A + 1$.

Symmetry

As eluded to above, our intransitivity levels require symmetry to be maintained, see equation (7.3.3), and so updating intransitivity level θ_k also impacts θ_{-k}. Our adaptation of the split-merge algorithm handles this by reflecting all updates to the intransitivity values. For example, updating parameter θ_k to θ'_k also updates parameter $\theta_{-k} = -\theta_k$ to $\theta'_{-k} = -\theta'_k$. Likewise, any reallocation of a pair (i, j) from, say, cluster k to cluster k', also reallocates pair (j, i) from cluster $-k$ to cluster $-k'$. Considering the reversible jump moves, splitting an intransitivity level k into k' and $(k+1)'$ also results in splitting cluster $-k$ into $-k'$ and $-(k+1)'$.

This symmetry also impacts the add and delete moves. Empty clusters can occur after a split, if all object pairs in the split cluster move to one of the new proposed clusters. Additionally, within the Gibbs reallocation move it is possible for all object pairs in a given cluster to move to other clusters. An empty cluster could exist for a long time before being merged, and even then the proposed merge may not be accepted. To deal with this more efficiently, *delete cluster* moves are included, and therefore *add*

cluster moves are included in order to satisfy detailed balance. If an intransitivity cluster k' is added, then the symmetry imposed on the intransitivity system means a cluster $-k'$ must also be added. The impact of this symmetry on the intransitivity delete move however is more subtle. The set of empty intransitivity clusters must be defined as $\mathcal{K}_e := \{k \in \{1, \ldots, K\} : \boldsymbol{b}_k = \boldsymbol{b}_{-k} = \emptyset\}$, such that a level $k_e \in \mathcal{K}_e$ is only considered empty, and therefore available for deletion, if both clusters k_e and k_{-e} are empty, that is $\boldsymbol{b}_{k_e} = \boldsymbol{b}_{k_{-e}} = \emptyset$. This is a necessary consideration because $\boldsymbol{Z}'$ was defined to contain only allocations for the 'upper triangle', that is $\boldsymbol{Z}' := \{\boldsymbol{z}'_{\{i,j\}} : i > j \in \mathcal{I}\backslash\{1\}\}$. Therefore it is possible that there are no pairs $\{i, j\}$ $\forall i > j \in \mathcal{I} \setminus \{1\}$ in cluster, say $k > 0$; however, there may exist a pair $\exists\{h, m\} \in -a : h > m \in \mathcal{I} \setminus \{1\}$ in cluster $-k$, and thus by symmetry $\{m, h\} \in a$. Note also that cluster 0 is not included in $\mathcal{K}_e$, as there must always be exactly one transitive cluster.

Initialisation

Convergence of the algorithm is not guaranteed in finite time. In the baseball example of Chapter 7 10000 samples were drawn from algorithm, with an additional 2000 burn-in samples. Four such chains were deployed with different random starting parameters, and convergence of the chains assessed here via visual assessment. There exists a variety of standard metrics, such as Gelman-Rubin statistics (Gelman and Rubin, 1992) which aim to formalise this process.

To ensure the best possible chance of convergence of the RJMCMC algorithm, good initial parameter estimates are useful. A standard Bradley-Terry model is first fitted to the data, which gives the skills $\boldsymbol{r}^{(BT)} := \{r_i^{(BT)} : i \in \mathcal{I}\}$, and the associated pairwise win probabilities $\{p_{ik}^{(BT)} : i > k \in \mathcal{I} \setminus \{1\}\}$. Then, from the data, if w_{ik} is the number of comparisons in which object i is favoured to object k, out of a possible n_{ik} comparisons, then a set of naïve pairwise probabilities can be formed by using these, and regularised with a prior. If $\{p_{ik}^{(n)} := w_{ik}/n_{ik} : i > k \in \mathcal{I} \setminus \{1\}\}$, which can be considered as $w_{ik} \sim$

Binomial$(n_{ik}, p_{ik}^{(n)})$, then by using a beta prior $\bar{p} \sim \text{Beta}\,(\alpha_p, \beta_p)$, with $\alpha_p = \beta_p = 2$, the initial naïve posterior probabilities are set to

$$\left\{ \hat{p}_{ik}^{(n)} = \left(\frac{w_{ik} + \alpha_p}{n_{ik} + \alpha_p + \beta_p} \right) : i > k \in \mathcal{I} \setminus \{1\} \right\}.$$

This choice of hyper-parameters provides a weakly informative prior with expectation 1/2. The regularisation provided by this prior prevents values of $\hat{p}_{ik}^{(n)}$ becoming too close to, or exactly, 0 or 1, which result in unrealistic and therefore poor starting values for the intransitivity. This is no concern in the main RJMCMC algorithm because the prior helps to regularise this. From these two sets of probabilities, the initial estimates for the intransitivity are found as

$$\left\{ \theta_{ik} = \log \left(\frac{\hat{p}_{ik}^{(n)} / (1 - \hat{p}_{ik}^{(n)})}{p_{ik}^{(BT)} / (1 - p_{ik}^{(BT)})} \right) : i > k \in \mathcal{I} \setminus \{1\} \right\}.$$

These values of intransitivity are clustered into their initial levels $\boldsymbol{\theta}$ using k-means clustering. This is done across a range of number of levels K to produce several separate initial models. The BIC is then calculated for each model, and the best is returned. Conditional on this model, the skills $\boldsymbol{r}^{(BT)}$ are also clustered again by using several k-means algorithms across a range of A, and the best model according to the BIC is selected as the initial estimates for the ICBT model. In general, these initial estimates are very good and correspond to a likelihood which is in the top 50^{th} percentile of the likelihoods provided by samples of the RJMCMC algorithm after convergence.

Next, from these initial starting parameters an MCMC algorithm is implemented, with only updates on the skill levels $\boldsymbol{\phi}$ and the intransitivity levels $\boldsymbol{\theta}$, with the other parameters $(\boldsymbol{Z}, \boldsymbol{Y}, K, A)$ fixed. After convergence of this algorithm, a second MCMC algorithm then allows additionally the allocations $\boldsymbol{Y}$ and $\boldsymbol{Z}$ to update, where the initial parameter values correspond to the highest posterior of the previous algorithm. Finally, only after convergence of these two MCMC algorithms, is the full split-merge RJMCMC

algorithm implemented, where samples are taken from all of $(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{Z}, \boldsymbol{Y}, K, A)$.

C.3 Split-Merge components

C.3.1 Split-Merge components : intransitivity

Intransitivity split move

To split an intransitivity cluster, first a cluster must be chosen to split, and this is chosen at random. A potentially more efficient proposal could be made by proposing to split a cluster, say, proportional to cluster size or as a function of its vicinity to neighbouring clusters. For simplicity, in this article all clusters in split and merge moves are randomly selected with equal probability.

Let the cluster to be split be denoted as $k \in \{0, \ldots, K\}$, with intransitivity level $\theta_k^s \in \boldsymbol{\theta}_{\mathcal{K}}$ and the proposed new clusters be k' and $k'+1$ with intransitivity levels $\theta'_{k'}, \theta'_{k'+1} \geq 0$, respectively, so that the proposed intransitivity levels are given as $\boldsymbol{\theta'} = \{\theta_0^s, \ldots, \theta_{k-1}^s, \theta'_{k'}, \theta'_{k'+1}, \theta_{k+1}^s, \ldots, \theta_K^s\}$. This notation is chosen because the ordering of the clusters must be maintained, for example $\theta_{k-1}^s < \theta'_{k'} < \theta'_{k'+1} < \theta_{k+1}^s$, to avoid label switching issues in the inference. Note that with this notation, the transitive cluster can be split, however in this case, one of the proposed intransitivity levels must be 0, to maintain the generalisation to a Bradley-Terry model. The proposed number of clusters is of course $K' = K^s + 1$.

Next, the new cluster parameters are proposed. To preserve ordering, consider the transformation of $\theta_k^s,\ k \neq 0$ such that the allowed range of θ'_k changes from $(\theta_{k-1}^s, \theta_{k+1}^s)$ to $(-\infty, \infty)$. Then, a new intransitivity level is proposed on this scale, and then transformed back to the its original range $(\theta_{k-1}^s, \theta_{k+1}^s)$. If $k = K^s$, then the range $(\theta_{K^s-1}^s, \theta_{K^s+1}^s)$ is currently undefined because $\theta_{K^s+1}^s$ is undefined. Therefore we define

$$\theta_{K^s+1}^s := \theta_{K^s}^s + (\theta_{K^s}^s - \theta_{K^s-1}^s) = 2\theta_{K^s}^s - \theta_{K^s-1}^s,$$

which acts an upper bound for the proposed new intransitivity levels $\theta'_{K'}, \theta'_{K'+1}$. Thus $\theta'_{K'+1}$ cannot increase by more than $\theta^s_{K^s} - \theta^s_{K^s-1}$, which is also the maximum $\theta'_{K'}$ can decrease by without violating the ordering constraint. If $k = 0$, i.e., the completely transitive cluster has a split proposed, then the lower proposed level must be $\theta'_{0'} = 0$, but the level $\theta'_{1'}$ follows the same rules as for a general $k \neq 0$, described below, therefore upper and lower bounds are defined for all values in $\boldsymbol{\theta}^s$.

The transformation onto the range $(-\infty, \infty)$ is achieved using a *matching function* $y(\theta^s_k|\theta^s_{k-1}, \theta^s_{k+1})$, defined as

$$y(x|\theta^s_{k-1}, \theta^s_{k+1}) := \text{logit}\left(\frac{x - \theta^s_{k-1}}{\theta^s_{k+1} - \theta^s_{k-1}}\right),\ k \in \{1, \ldots, K\},\ x \in (\theta^s_{k-1}, \theta^s_{k+1}), \tag{C.3.1}$$

where $\text{logit}(a) := \log(a/1-a),\ a \in [0,1]$. Then, the *auxiliary variable* $u \sim \sigma\chi^2_1$, for $\sigma > 0$, is added (and subtracted) to the transformed intransitivity level to give the proposed upper (and lower) transformed intransitivity levels, respectively, such that

$$y(\theta'_{k'+1}|\theta^s_{k-1}, \theta^s_{k+1}) = y(\theta^s_k|\theta^s_{k-1}, \theta^s_{k+1}) + u \tag{C.3.2}$$

$$y(\theta'_{k'}|\theta^s_{k-1}, \theta^s_{k+1}) = y(\theta^s_k|\theta^s_{k-1}, \theta^s_{k+1}) - u. \tag{C.3.3}$$

After a retransformation back to the original scale, this gives values of the lower and upper proposed intransitivity levels $\theta'_{k'}$ and $\theta'_{k'+1}$, with $\theta_{k-1} < \theta'_{k'} < \theta'_{k'+1} < \theta_{k+1}$, as

$$\theta'_{k'} = \begin{cases} \frac{\theta^s_{k-1} + \theta^s_{k+1}\exp(-u)\left(\theta^s_k - \theta^s_{k-1}\right)/\left(\theta^s_{k+1} - \theta^s_k\right)}{1 + \exp(-u)\left(\theta^s_k - \theta^s_{k-1}\right)/\left(\theta^s_{k+1} - \theta^s_k\right)} & \text{if } k' \neq 0' \\ 0 & \text{if } k' = 0' \end{cases} \tag{C.3.4}$$

$$\theta'_{k'+1} = \frac{\theta^s_{k-1} + \theta^s_{k+1}\exp(u)\left(\theta^s_k - \theta^s_{k-1}\right)/\left(\theta^s_{k+1} - \theta^s_k\right)}{1 + \exp(u)\left(\theta^s_k - \theta^s_{k-1}\right)/\left(\theta^s_{k+1} - \theta^s_k\right)}. \tag{C.3.5}$$

Equations (C.3.4) and (C.3.5) are the main contributions to ensure that ordering and labelling is preserved in RJMCMC type algorithms.

The next state of the sampler $(\boldsymbol{\phi}^{s+1}, \boldsymbol{Y}^{s+1}, A^{s+1}, \boldsymbol{\theta}^{s+1}, \boldsymbol{Z}^{s+1}, K^{s+1})$ is then accepted

to be $(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}', \boldsymbol{Z}', K')$ with some probability A_{split}, and otherwise

$$(\boldsymbol{\phi}^{s+1}, \boldsymbol{Y}^{s+1}, A^{s+1}, \boldsymbol{\theta}^{s+1}, \boldsymbol{\theta}^{s+1}, \boldsymbol{Z}^{s+1}, K^{s+1}) = (\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s).$$

The form of the acceptance probability A_{split} is computed as:

$$\begin{aligned} A_{split} =& \frac{\pi\left(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}', \boldsymbol{Z}', K^s+1|\boldsymbol{x}\right)}{\pi\left(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s|\boldsymbol{x}\right)} \frac{q(K^s, \boldsymbol{Z}^s, \boldsymbol{\theta}^s|K^s+1, \boldsymbol{Z}', \boldsymbol{\theta}')}{q(K^s+1, \boldsymbol{Z}', \boldsymbol{\theta}'|K^s, \boldsymbol{Z}^s, \boldsymbol{\theta}^s)} \frac{1}{q_u(u)} J_{split} \\ =& \frac{\pi\left(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}', \boldsymbol{Z}', K^s+1|\boldsymbol{x}\right)}{\pi\left(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s|\boldsymbol{x}\right)} \frac{q(merge|K^s+1)}{q(split|K^s)} \frac{q_k(k', k'+1)}{q_k(k)} \frac{1}{q_u(u)} \frac{1}{q(\boldsymbol{Z}')} J_{split}, \end{aligned}$$

which is the ratio of the posterior densities, the ratio of the proposal densities, and the ratio of the densities of the auxiliary variables and the Jacobian J_{split}, see Appendices C.7.1 and C.7.3 for details. If $K^s = 0$, then the probability of proposing a *merge*, see Section C.3.1, is $q(merge|K^s = 0) = 0$ and the probability of proposing a split $q(split|K^s = 0) = 1$. Otherwise, $q(merge|K^s \neq 0) = q(split|K^s \neq 0) = 1/2$, and therefore the ratio of is simply

$$\frac{q(merge|K^s+1)}{q(split|K^s)} = \frac{1}{1 + \mathbb{1}\{K^s = 0\}}.$$

The probability of selecting cluster k to split is $q_k(k) = \frac{1}{K^s}$, $\forall k \in \{1, \ldots, K^s\}$. The probability of selecting clusters $k', k'+1$, $\forall k' \in \{1, \ldots, K'-1\}$ to be merged back again is also $q_k(k', k'+1) = \frac{1}{K'-1} = \frac{1}{K^s}$, so these terms will cancel. The density of the auxiliary variables is the density of u, $q_u(u)^{-1}$, which is given as $\chi_1^2\left(u/\sigma\right)$, where $\chi_1^2(x)$ is the density of a chi-squared distribution with 1 degree of freedom.

The probability of the allocations is $q\left(\boldsymbol{Z}'\right)$, and this is dependent on the cluster being split and the current state of the model, for which there are three different situations to consider:

The general case: $(k \neq 0)$

Since the transitive level $\theta_0 = 0$ adds additional complications, we first depict the acceptance probability for the general case that $k \neq 0$. Denote the set of pairs $i > j \in \mathcal{I} \setminus \{1\}$ of objects which belong to the cluster-to-be-split, i.e., $k \neq 0$, as $\boldsymbol{b}_k^s = \{\{i,j\},\ \forall i > j \in \mathcal{I} \setminus \{1\} : z^s_{\{i,j\},k} = 1\}$. Then, the probability of the cluster allocation, given that cluster $k \neq 0$ has been split,

$$q(\boldsymbol{Z}') = \left(\frac{|\boldsymbol{b}_k^s|!}{|\boldsymbol{b}'_{k'}|!\ |\boldsymbol{b}'_{k'+1}|!}\right) 2^{-(|\boldsymbol{b}_k^s|)} \left(\frac{|\boldsymbol{b}_{-k}^s|!}{|\boldsymbol{b}'_{-k'}|!\ |\boldsymbol{b}'_{-(k'+1)}|!}\right) 2^{-(|\boldsymbol{b}_{-k}^s|)}$$

since the new cluster allocations are selected at random and all possible combinations of allocation must be considered, and remembering that cluster $-k$ will also be split into clusters $-k'$ and $-k'-1$ with intransitivity levels $\theta'_{-k'} = -\theta'_{k'}$ and $\theta'_{-k'-1} = -\theta'_{k'+1}$. See Appendix C.7.1 for details of the Jacobian J_{split}. Combining these components, the acceptance probability of splitting a cluster $k \neq 0$ can be computed as

$$A_{split}\,|(k \neq 0) = \Lambda_{split} \left(\frac{|\boldsymbol{b}_k^s|!\ |\boldsymbol{b}_{-k}^s|!\ 2^{-(|\boldsymbol{b}_k^s|+|\boldsymbol{b}_k^s|)}}{|\boldsymbol{b}'_{k'}|!\ |\boldsymbol{b}'_{k'+1}|!\ |\boldsymbol{b}'_{-k'}|!\ |\boldsymbol{b}'_{-(k'+1)}|!}\right)^{-1} J_{split},$$

where

$$\Lambda_{split} = \frac{\pi\left(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}', \boldsymbol{Z}', K^s + 1 | \boldsymbol{x}\right)}{\pi\left(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s | \boldsymbol{x}\right)} \frac{1}{1 + \mathbb{1}\{K^s = 0\}} \frac{\sigma}{\chi_1(u/\sigma)}.$$

Splitting from the transitive cluster: $(k = 0,\ K^s \geq 1)$

We now consider the special case that we propose to split the transitive cluster $(k = 0)$, when there are other intransitive clusters $(K^s \geq 1)$. The acceptance probability for this proposal differs slightly, because $\theta_0 = 0$ is fixed. If a split is proposed on level 0 with intransitivity level θ_0^s, then the proposed clusters are $0'$ and $1'$ with intransitivity levels $\theta'_{0'}, \theta'_{1'}$; however, the completely transitive level must remain so, in order that the model remains an exact generalisation of the Bradley-Terry model, that is $\theta'_{0'} = \theta_0 = 0$. The level $1'$ has intransitivity level $\theta'_{1'} > 0$ governed by equation (C.3.5), remembering

that $\theta^s_{-1} = -\theta^s_1$. Additionally, a separate Jacobian $J_{split,0}$ is required, see Appendix C.7.3. The cluster allocation also differs. Instead of allocating pairs in k to either $k', k'+1$ and pairs in $-k$ to either $-k', -k'+1$, the pairs in 0 are allocated to one of $0', -1', 1'$. The probability of the cluster allocation is therefore

$$q_0(\boldsymbol{Z}') = \left(\frac{|\boldsymbol{b}^s_0|!}{|\boldsymbol{b}^s_{-1}|!\ |\boldsymbol{b}^s_0|!\ |\boldsymbol{b}^s_1|!}\right) 3^{-|\boldsymbol{b}^s_0|},$$

since the allocation is at random. This gives an acceptance probability of

$$A_{split}|\,(k=0, K^s \geq 1) = \Lambda_{split}\left(\frac{|\boldsymbol{b}^s_0|!}{|\boldsymbol{b}^s_{-1}|!\ |\boldsymbol{b}^s_0|!\ |\boldsymbol{b}^s_1|!}\right)^{-1}\left(\frac{1}{3}\right)^{-|\boldsymbol{b}^s_0|} J_{split,0},$$

if $k=0$, $K^s \geq 1$.

Splitting from the Bradley-Terry case: ($K^s = 0$)

Consider the special case that only the transitive level exists and therefore the model in this state is identical to the Bradley-Terry. The transformation via the matching function, equation (C.3.1), is not possible in this case since the bounds $\theta^s_{K^s+1}$ (and $\theta^s_{-(K^s+1)} = -\theta^s_{K^s+1}$) are not defined when $K^s = 0$. However, ordering is not an issue in this case, and so the levels are simply proposed as

$$\theta'_{0'} = \theta_0, \quad \theta'_{1'} = \theta_0 + u = u,$$

which gives a Jacobian of 1, and the pairs in 0 are again allocated to one of $0', 1', -1'$. Therefore the acceptance probability, for $K^s = 0$, is given as

$$A_{split}|\,(K^s = 0) = \Lambda_{split}\left(\frac{|\boldsymbol{b}^s_0|!}{|\boldsymbol{b}^s_{-1}|!\ |\boldsymbol{b}^s_0|!\ |\boldsymbol{b}^s_1|!}\right)^{-1}\left(\frac{1}{3}\right)^{-|\boldsymbol{b}^s_0|}.$$

Intransitivity merge move

So as to not confuse notation, we will consider the merge in terms of a reversed split move. Therefore let $(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}', \boldsymbol{Z}', K') = (\boldsymbol{\phi}^{s+1}, \boldsymbol{Y}^{s+1}, A^{s+1}, \boldsymbol{\theta}^{s+1}, \boldsymbol{Z}^{s+1}, K^{s+1})$ be the current values of the parameters at step s, and we aim to propose parameters $(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s)$. Two clusters are sampled to merge, however, since these must be consecutive clusters, in order to preserve order, we can simply sample with equal probability from the set $k' \sim \{0, \ldots, K'-1\}$ and then merge clusters k' and $k'+1$ into a new cluster k. The proposed number of clusters must be $K^s = K' - 1$. Next, the cluster membership $\boldsymbol{Z}^s$ is updated which is simple because all pairs of objects from both clusters k' or $k'+1$ get assigned to the new cluster k. To find a suitable cluster mean θ_k^s, the inverse of the split move is used, so inversely to the split move, the matching function

$$y(x|\theta'_{k'-1}, \theta'_{k'+2}) = \text{logit}\left(\frac{x - \theta'_{k'-1}}{\theta'_{k'+2} - \theta'_{k'-1}}\right), \; x \in (\theta'_{k'-1}, \theta'_{k'+2}),$$

is used. Then the proposed merged transformed cluster mean is given from equations (C.3.2) and (C.3.3), as

$$y(\theta_k^s|\theta'_{k'-1}, \theta'_{k'+2}) = \frac{\left[y(\theta'_{k'}|\theta'_{k'-1}, \theta'_{k'+2}) + y(\theta'_{k'+1}|\theta'_{k'-1}, \theta'_{k'+2})\right]}{2},$$

which results in a proposed merged cluster mean of

$$\theta_k^s = \theta'_{k'-1} + \frac{\left(\theta'_{k'+2} - \theta'_{k'-1}\right)}{1 + \left[\left(\frac{\theta'_{k'} - \theta'_{k'-1}}{\theta'_{k'+2} - \theta'_{k'}}\right)\left(\frac{\theta'_{k'+1} - \theta'_{k'-1}}{\theta'_{k'+2} - \theta'_{k'+1}}\right)\right]^{-1/2}}.$$

Then, the next state is accepted to be $(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s) = (\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}', \boldsymbol{Z}', K')$, with probability A_{merge}, whose precise form again depends on the situation.

The general case: $(k' \neq 0)$

In the most general case, that $k' \neq 0$, then this gives an acceptance probability of

$$A_{merge}|\,(k' \neq 0) = \Lambda_{merge} \left(\frac{\boldsymbol{b}_k^s!\, \boldsymbol{b}_{-k}^s!\, 2^{-(|\boldsymbol{b}_k^s|+|\boldsymbol{b}_k^s|)}}{\boldsymbol{b}'_{k'}!\, \boldsymbol{b}'_{k'+1}!\, \boldsymbol{b}'_{-k'}!\, \boldsymbol{b}'_{-(k'+1)}!} \right) \left(J'_{split}\right)^{-1},$$

where

$$\Lambda_{merge} = \frac{\pi\left(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K'-1|\boldsymbol{x}\right)}{\pi\left(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}', \boldsymbol{Z}', K'|\boldsymbol{x}\right)} \left(1 + \mathbb{1}\{K'=1\}\right) \frac{\chi_1(u'/\sigma)}{\sigma},$$

and where $u' = y(\theta'_{k'+1}|\theta'_{k'-1}, \theta'_{k'+2}) - y(\theta^s_k|\theta'_{k'-1}, \theta'_{k'+2})$ is the the value required for the inverse split move to get back to intransitivity levels $\theta'_{k'}$ and $\theta'_{k'+1}$, and J'_{split} is the Jacobian for the inverse split move.

Merging to the transitive cluster: $(k'=0,\ K' \geq 2)$

If $k'=0$, and therefore clusters $0', 1'$ are being merged to cluster 0 with intransitivity level $\theta^s_k = \theta^s_0 = 0$, then, similarly to the equivalent split move, the allocation probability changes because pairs from $0', -1', 1'$ are proposed to move into 0. The Jacobian also changes to $J'_{split,0}$, see Appendix C.7.3, and the acceptance probability is given as

$$A_{merge}|\,(k'=0, K' \geq 2) = \Lambda_{merge} \left(\frac{|\boldsymbol{b}_0^s|!}{|\boldsymbol{b}_{-1}^s|!\ |\boldsymbol{b}_0^s|!\ |\boldsymbol{b}_1^s|!} \right) \left(\frac{1}{3}\right)^{|\boldsymbol{b}_0^s|} \left(J'_{split,0}\right)^{-1}.$$

Merging to the Bradley-Terry case: $(K'=1)$

Similarly to the equivalent split move, if $K'=1$, then the proposed merge move would be to a Bradley-Terry model (or generalisation there-of). In this case no transformation is required, and thus no Jacobian term, and so

$$A_{merge}|(K'=1) = \Lambda_{merge} \left(\frac{|\boldsymbol{b}_0^s|!}{|\boldsymbol{b}_{-1}^s|!\ |\boldsymbol{b}_0^s|!\ |\boldsymbol{b}_1^s|!} \right) \left(\frac{1}{3}\right)^{|\boldsymbol{b}_0^s|},$$

where in this case $u' = \theta'_{k'+1}$.

C.3.2 Split-Merge components : skill

Skill split move

The split and merge components are now considered for the objects skill levels. The reversible jump moves here are similar to those of the intransitivity levels, but a few changes are needed to accommodate for the fact that levels can have either positive or negative values, and that symmetry in the skill levels is not enforced as it is for the intransitivity levels. A cluster $a \in \mathcal{A}^s$ is selected at random to be split, with a skill level $\phi_a^s \in \boldsymbol{\phi}^s$, and the proposed new clusters are labelled $a', a'+1$, and have skill levels $\phi'_{a'}, \phi'_{a'+1} \in \mathbb{R}$. Note that the skill level $\phi_0 = 0$ can be split, although in this case one of the proposed skill levels must also be 0, and object 1 must stay in this level. The proposed number of skill levels is of course $A' = A^s + 1$. Next, the new skill level parameters are proposed.

From the current allocations $\boldsymbol{Y}^s$, the proposed allocations $\boldsymbol{Y}'$ are found by randomly reallocating those objects in cluster a to either cluster a' or $a'+1$ with equal probability. The acceptance probability A_{split_A} is then computed as:

$$\begin{aligned} A_{split_A} =& \frac{\pi\left(\boldsymbol{\phi}', \boldsymbol{Y}', A^s+1, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s|\boldsymbol{x}\right)}{\pi\left(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s|\boldsymbol{x}\right)} \frac{q(A^s, \boldsymbol{Y}^s, \boldsymbol{\phi}^s | A^s+1, \boldsymbol{Y}', \boldsymbol{\phi}')}{q(A^s+1, \boldsymbol{Y}', \boldsymbol{\phi}' | A^s, \boldsymbol{Y}^s, \boldsymbol{\phi}^s)} \frac{1}{q_{u_A}(u_A)} J^r_{split} \\ =& \frac{\pi\left(\boldsymbol{\phi}', \boldsymbol{Y}', A^s+1, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s|\boldsymbol{x}\right)}{\pi\left(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s|\boldsymbol{x}\right)} \frac{q(merge|A^s+1)}{q(split|A^s)} \frac{q_a(a', a'+1)}{q_a(a)} \frac{1}{q_{u_A}(u_A)} \frac{1}{q(\boldsymbol{Y}')} J^r_{split}, \end{aligned}$$

which is the ratio of the posterior densities, the ratio of the proposal densities, and the ratio of the densities of the auxiliary variables, which is in this case just $q_{u_A}(u_A)^{-1} = \chi_1\left(u_A/\sigma_A\right)$, and J^r_{split} is the Jacobian. The probability of proposing a particular level to be split is simply

$$q_a(a) = \frac{1}{|\mathcal{A}^s|} = \frac{1}{A^s+1} \; \forall a \in \mathcal{A}^s.$$

For the inverse merge move, see Section C.3.2, the probability of proposing clusters a'

and $a'+1$ to merge is simply

$$q_a(a', a'+1) = \frac{1}{A'} \; \forall a' \in \mathcal{A}' \setminus \{A'_+\},$$

therefore $q_a(a', a'+1)/q_a(a) = 1$.

Let $\boldsymbol{c}_a^s := \{i, \; \forall i \in \mathcal{I} \setminus \{1\} : \boldsymbol{y}^s_{\{i\},a} = 1\}$ denote the set of objects which belong to the cluster-to-be-split, i.e., a. Then, the probability of the cluster allocation, given that cluster a has been split, is

$$q(\boldsymbol{Y}') = \frac{|\boldsymbol{c}_a^s|!}{|\boldsymbol{c}_{a'}^s|! \; |\boldsymbol{c}_{a'+1}^s|!} 2^{-|\boldsymbol{c}_a^s|}$$

since the new cluster allocations are selected at random.

Again there are three variants for both the splitting and merging of skill levels that must be considered:

The general case: $(a \neq 0)$

To preserve ordering, the same transformation is made as in Section C.3.1, and so the set of levels must be extended as before, but now in both positive and negative directions, such that

$$\phi^s_{A^s_+ + 1} := \phi^s_{A^s_+} + (\phi^s_{A^s_+} - \phi^s_{A^s_+ - 1})$$
$$\phi^s_{A^s_- - 1} := \phi^s_{A^s_-} - |\phi^s_{A^s_- + 1} - \phi^s_{A^s_-}|.$$

Using the same logit transformation to preserve order, new skill levels are proposed via the equations

$$\phi'_{a'} = \frac{\phi^s_{a-1} + \phi^s_{a+1} \exp(-u_A) \left(\phi^s_a - \phi^s_{a-1}\right) / \left(\phi^s_{a+1} - \phi^s_a\right)}{1 + \exp(-u_A) \left(\phi^s_a - \phi^s_{a-1}\right) / \left(\phi^s_{a+1} - \phi^s_a\right)}$$
$$\phi'_{a'+1} = \frac{\phi^s_{a-1} + \phi^s_{a+1} \exp(u_A) \left(\phi^s_a - \phi^s_{a-1}\right) / \left(\phi^s_{a+1} - \phi^s_a\right)}{1 + \exp(u_A) \left(\phi^s_a - \phi^s_{a-1}\right) / \left(\phi^s_{a+1} - \phi^s_a\right)}.$$

where the *auxiliary variable* $u_A \sim \sigma_A \chi^2_1$, for $\sigma_A > 0$.

Combining these components, the acceptance probability of splitting skill cluster a can be computed as

$$A^r_{split}|\,(a \neq 0) = \Lambda^r_{split} J^r_{split},$$

where

$$\Lambda^r_{split} = \frac{\pi\left(\boldsymbol{\phi}', \boldsymbol{Y}', A^s + 1, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s | \boldsymbol{x}\right)}{\pi\left(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s | \boldsymbol{x}\right)} \frac{1}{1 + \mathbb{1}\{A^s = 0\}} \frac{\sigma_A}{\chi(u_A/\sigma_A)} \left(\frac{|\boldsymbol{c}^s_a|!}{|\boldsymbol{c}^s_{a'}|!\ |\boldsymbol{c}^s_{a'+1}|!} 2^{-|\boldsymbol{c}^s_a|} \right)^{-1}.$$

Splitting from the zero skill: ($a = 0$, $A^s \geq 0$)

Consider the special case that the 0 level is split, but other skill levels are present. One of the split levels must remain at 0, specifically, $\phi'_{0'} = \phi^s_0 = 0$. The other proposed skill level is

$$\phi'_{(2X-1)'} = X\phi'_{a'+1} + (1 - X)\phi'_{a'},$$

where $X \sim$ Bernoulli(0.5), such that the non-zero proposed skill level can be either above or below the 0 level, and it has label $(2X - 1)'$, that is, either $1'$ or $-1'$. The Jacobian $J^r_{split,0}$ is calculated as in Appendix C.7.3, and the acceptance probability is given as

$$A^r_{split}|\,(a = 0, A^s \geq 0) = 2\Lambda^r_{split} J^r_{split,0},$$

the only difference to the general case being in the Jacobian term, and the factor of 2 which comes from the Bernoulli variable.

Splitting from the single skill model: ($A^s = 0$)

In the unlikely case that only a single skill level is present in the model, the new skill level is simply proposed as

$$\phi'_{(2X-1)'} = Xu_A - (1 - X)u_A,$$

since no transformation is needed due to no ordering issues. In this case the acceptance probability is given as

$$A^r_{split}|\,(A^s = 0) = 2\Lambda^r_{split},$$

where the Jacobian term is 1.

The next state of the sampler $(\boldsymbol{\phi}^{s+1}, \boldsymbol{Y}^{s+1}, A^{s+1}, \boldsymbol{\theta}^{s+1}, \boldsymbol{Z}^{s+1}, K^{s+1})$ is then accepted to be $(\boldsymbol{\phi}', \boldsymbol{Y}', A', \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s)$ with probability A^r_{split}, and with probability $1 - A^r_{split}$ it remains as $(\boldsymbol{\phi}^{s+1}, \boldsymbol{Y}^{s+1}, A^{s+1}, \boldsymbol{\theta}^{s+1}, \boldsymbol{Z}^{s+1}, K^{s+1}) = (\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s)$.

Skill merge move

Like the intransitivity merge move, the merge move is considered as a reverse split move, i.e., the current state is $(\boldsymbol{\phi}', \boldsymbol{Y}', A', \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s) = (\boldsymbol{\phi}^{s+1}, \boldsymbol{Y}^{s+1}, A^{s+1}, \boldsymbol{\theta}^{s+1}, \boldsymbol{Z}^{s+1}, K^{s+1})$, and we aim to propose parameters $(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s)$. Two levels $a', a'+1 \in \mathcal{A}' \setminus \{A'_+\}$ are sampled to merge with probability

$$q_a(a', a'+1) = \frac{1}{A'} \; \forall a' \in \mathcal{A}' \setminus \{A'_+\},$$

and are inverse split move is proposed with probability

$$q_a(a) = \frac{1}{A^s + 1} \; \forall a \in \mathcal{A}^s \setminus \{A^s_+\},$$

so again $q_a(a)/q_a(a', a'+1) = 1$. The three variants for the merging of skill levels are then as follows:

The general case: $(a', a'+1 \neq 0)$

In the general case that the sampler proposed to merge two non-zero levels, $a', a'+1 \neq 0$, then

$$A^r_{merge}|(a', a'+1 \neq 0) = \Lambda^r_{merge}/J'^r_{split},$$

where

$$\Lambda^r_{merge} = \frac{\pi\left(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A'-1, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s|\boldsymbol{x}\right)}{\pi\left(\boldsymbol{\phi}', \boldsymbol{Y}', A', \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s|\boldsymbol{x}\right)}\left(1+\mathbb{1}\{A'=1\}\right)\frac{\Phi(u'_A/\sigma_A)}{\sigma_A}\frac{|\boldsymbol{c}^s_a|!}{|\boldsymbol{c}^s_{a'}|!\;|\boldsymbol{c}^s_{a'+1}|!}2^{-|\boldsymbol{c}^s_a|},$$

and $u'_A \sim \sigma_A\chi^2_1$ is the the value required for the inverse split move to get back to skill levels $\phi'_{a'}$ and $\phi'_{a'+1}$.

Merging to the zero skill: $(\min(|a'|,|a'+1|)=0,\ A'\geq 2)$

If either a' or $a'+1$ are 0 (and therefore have skill levels either $\phi'_{a'}=0$ or $\phi'_{a'+1}=0$), then the merged skill level must be $\phi_a=\phi_0=0$, and u'_A is the value required to invert the merge, such that

$$u'_A = \begin{cases} |y(\phi'_{a'+1}|\phi'_{a'-1},\phi'_{a'+2})| & \text{if } a'=0 \\ |y(\phi'_{a'}|\phi'_{a'-1},\phi'_{a'+2})| & \text{if } a'+1=0 \end{cases},$$

where $y(\cdot)$ is defined in equation (C.3.1). This gives an acceptance probability of

$$A^r_{merge}|\left[\min(|a'|,|a'+1|)=0, A'\geq 2\right] = \Lambda^r_{merge}/2J'^r_{split,0},$$

so the Jacobian $J^r_{split,0}$, see Appendix C.7.4, differs from the general case, and also the factor of 1/2, which comes from the Bernoulli variable in the inverse split move.

Merging to the single skill model: $(A'=1)$

If $A'=1$, then the merged skill level must be 0, and u'_A is simply the absolute value of whichever level is not 0, that is $u'_A=\max(|\phi'_{a'}|,|\phi'_{a'+1}|)$. In this case the acceptance probability is given as

$$A^r_{merge}|(A'=1) = \Lambda^r_{merge}/2.$$

Then, the next state is accepted to be $(\boldsymbol{\phi}^s,\boldsymbol{Y}^s,A^s,\boldsymbol{\theta}^s,\boldsymbol{Z}^s,K^s)=(\boldsymbol{\phi}',\boldsymbol{Y}',A',\boldsymbol{\theta}^s,\boldsymbol{Z}^s,K^s)$, with probability A^r_{merge}.

C.4 Add-Delete components

C.4.1 Add-Delete components : intransitivity

Empty clusters

It is possible that $\exists\ a \in \mathcal{A} : \mathcal{S}_{\{i\}}(\boldsymbol{Y}) \cap a = \{\emptyset\},\ \forall i \in \mathcal{I}$, that is, there exists empty clusters. Empty clusters can occur after a split, if all object pairs in the split cluster move to one of the new proposed clusters. Additionally, within the Gibbs reallocation move, see Section C.5, it is possible for all object pairs in a given cluster to be reallocated to other clusters. An empty cluster could exist for a long time before being merged, and even then the proposed merge may not be accepted. To deal with this more efficiently, a *delete cluster* move is included. To satisfy detailed balance, an *add cluster* move must also be included. Define the set of empty clusters $\mathcal{K}_e^s$ to be $\mathcal{K}_e^s := \{a \in \{1, \ldots, K^s\} : \boldsymbol{b}_a = \boldsymbol{b}_{-a} = \emptyset\}$, such that a cluster k_e is only considered empty if both $k_e = \emptyset$ and $k_{-e} = \emptyset$. This is a necessary consideration because $\boldsymbol{Z}'$ was defined to contain only allocations for the 'upper triangle', that is $\boldsymbol{Z}' := \{\boldsymbol{z}'_{\{i,j\}} : i > j \in \mathcal{I} \setminus \{1\}\}$. Therefore it is possible that there are no pairs $\{i, j\}\ \forall i > j \in \mathcal{I} \setminus \{1\}$ in cluster, say $a > 0$; however, there may exist a pair $\exists\{h, m\} \in -a : h > m \in \mathcal{I} \setminus \{1\}$ in cluster $-a$, and thus by symmetry $\{m, h\} \in a$. Thus, to delete a cluster a, both clusters a and $-a$ must be empty, that is, we require $\boldsymbol{b}_a = \boldsymbol{b}_{-a} = \emptyset$, hence the definition of $\mathcal{K}_e^s$. Note also that cluster 0 is not being considered in $\mathcal{K}_e^s$, as there must always be exactly one transitive cluster.

Intransitivity add move

When a cluster is added, the proposed cluster mean $\theta^*|\alpha, \beta \sim h_0$ is drawn from the prior distribution, see equation (7.4.3), which is gamma distributed, and it is inserted such that correct ordering of intransitivity levels is maintained. Thus, if θ^* is, say the kth largest value in the vector $(\boldsymbol{\theta}, \theta^*)$, then the proposed cluster allocation changes such that

$\boldsymbol{b}'_l = \boldsymbol{b}^s_l,\ \boldsymbol{b}'_{-l} = \boldsymbol{b}^s_{-l}\ \forall l < k, \boldsymbol{b}'_{l+1} = \boldsymbol{b}^s_l,\ \boldsymbol{b}'_{-(l+1)} = \boldsymbol{b}^s_{-l}\ \forall l \geq k$ and $\boldsymbol{b}'_k = \boldsymbol{b}'_{-k} = \emptyset$, which defines $\boldsymbol{Z}'$. The proposed intransitivity levels become $\boldsymbol{\theta}' = \{\theta^s_1, \ldots, \theta^s_{k-1}, \theta^*, \theta^s_k, \ldots, \theta^s_{k+1}\}$.

The cluster allocation is then reordered by increasing the index of cluster allocations which belong to intransitivity levels which are larger than the added cluster. The probability of proposing to add a cluster is an increasing function of an algorithm parameter $\rho_K > 0$, and the probability of proposing to delete a cluster, see Section C.4.1, is an increasing function of the current number of empty clusters, $N^s_\emptyset := |\mathcal{K}^s_e|$, which excludes the completely transitive set $\mathcal{J}_T$. Therefore, the probability of proposing to add and delete an intransitivity level is constructed as,

$$q(add\ |N^s_\emptyset) = \frac{\rho_K}{\rho_K + N^s_\emptyset}, \qquad q(delete\ |N^s_\emptyset) = \frac{N^s_\emptyset}{\rho_K + N^s_\emptyset},$$

respectively.

Since an add or delete move in no way effects the likelihood, only the prior density changes, and the acceptance probability of adding is given as

$$\begin{aligned} A_{add} =& \frac{\pi(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A, \boldsymbol{\theta}', \boldsymbol{Z}', K^s + 1)}{\pi(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s)} \frac{1}{(K^s + 1)h_0(\theta^*|\alpha, \beta)} \frac{(N^s_\emptyset + 1)/(\rho_K + N^s_\emptyset + 1)}{\rho_K/(\rho_K + N^s_\emptyset)} \\ =& \frac{f(\boldsymbol{Z}'|\gamma_K, K^s + 1)}{f(\boldsymbol{Z}|\gamma_K, K^s)} \frac{g_0(K^s + 1|\lambda)}{g_0(K^s|\lambda)} \frac{(N^s_\emptyset + 1)(\rho_K + N^s_\emptyset)}{\rho_K(\rho_K + N^s_\emptyset + 1)}, \end{aligned}$$

with prior $f(\boldsymbol{Z}|\gamma_K, K^s)$ defined in equation (7.4.2) and $\gamma_K > 0$

Intransitivity delete move

During a delete empty cluster move, an empty cluster $k_e \in \mathcal{K}^s_e$ is selected at random. Given the sampler is in state s with $N^s_\emptyset \in \mathbb{N}_+$ empty clusters, i.e., there is at least 1 empty cluster, the acceptance probability of deleting an empty cluster, k_e with cluster

mean θ^*, where $\boldsymbol{Z}'$ is the proposed allocation after the cluster is removed, is given as

$$A_{delete} = \frac{f(\boldsymbol{Z}'|\gamma_K, K^s - 1)}{f(\boldsymbol{Z}^s|\gamma_K, K^s)} \frac{g_0(K^s - 1|\lambda)}{g_0(K^s|\lambda)} \frac{\rho_K(\rho_K + N^s_\emptyset)}{N^s_\emptyset(\rho_K + N^s_\emptyset - 1)}.$$

Again, intransitivity levels are re-indexed appropriately. Choosing $\rho_K = 1$ means that when one empty cluster exists, there is an equal chance of proposing to add or delete an empty cluster. For $0 < \rho_K < 1$, as long as there exists at least one cluster, there is always a greater chance of proposing to delete an empty cluster than to add an empty cluster.

C.4.2 Add-Delete components : skills

Similarly to the add intransitivity move, for the add skill cluster move a new skill level is drawn from the prior, that is $\phi^*|\nu_A \sim \mathcal{N}(0, \nu_A^2)$, with $\nu_A > 0$ and indexed such that the proposed vector of skill levels is ordered. The allocations are altered such that all objects maintain the same skills.

Define the set of empty skill clusters $\mathcal{A}^s_e$ to be $\mathcal{A}^s_e := \{a \in \mathcal{A}_{-0} : \boldsymbol{c}_a = \emptyset\}$, noting that cluster 0 is not being considered. During a delete empty cluster move, an empty cluster, say $a_e \in \mathcal{A}^s_e$, is selected at random to be deleted. The cluster allocations are then reordered such that, if the deleted level is positive, the index of cluster allocations which belong to skill levels which are larger than the removed cluster are decreased by one, and if the deleted level is negative, the index of cluster allocations which belong to skill levels which are smaller than the removed cluster are increased by one. That is, $\boldsymbol{c}'_l = \boldsymbol{c}^s_l - 1$ if $l > \max(a_e, 0)$, $\boldsymbol{c}'_l = \boldsymbol{c}^s_l + 1$ if $l < \min(a_e, 0)$ and $\boldsymbol{c}'_l = \boldsymbol{c}^s_l$ otherwise. Similarly to the intransitivity add and delete moves, the probability of proposing a skill cluster is an increasing function of an algorithm parameter $\rho_A > 0$, and the probability of proposing to delete a cluster is increasing function of the current number of empty clusters, $N^s_{\emptyset_A} := |\mathcal{A}^s_e|$. The probability of proposing to add or delete an empty skill

cluster is, respectively,

$$q(add\ A\ |N^s_{\emptyset_A}) = \frac{\rho_A}{\rho_A + N^s_{\emptyset_A}}, \quad q(delete\ A\ |N^s_{\emptyset_A}) = \frac{N^s_{\emptyset_A}}{\rho_A + N^s_{\emptyset_A}}.$$

Given the sampler is in state s with $N^s_{\emptyset_A} \in \mathbb{N}$ empty clusters, the acceptance probability of adding is given as

$$A_{add\ A} = \frac{f(\boldsymbol{Y}'|\gamma_A, A^s + 1)}{f(\boldsymbol{Y}|\gamma_A, A^s)} \frac{g_0(A^s + 1|\lambda_A)}{g_0(A^s|\lambda_A)} \frac{(N_{\emptyset^s_A} + 1)(\rho_A + N_{\emptyset^s_A})}{\rho_A(\rho_A + N_{\emptyset^s_A} + 1)},$$

with $f(\boldsymbol{Y}|\gamma_A, A^s)$ given as in equation (7.4.4) and $\gamma_A > 0$. Likewise, given the sampler is in state s with $N^s_{\emptyset_A} \in \mathbb{N}_+$ empty clusters, the acceptance probability of deleting an empty cluster, a with cluster mean ϕ^*, where $\boldsymbol{Y'}$ is the proposed allocation after the cluster is removed, is given as

$$A_{delete\ A} = \frac{f(\boldsymbol{Y}'|\gamma_A, A^s - 1)}{f(\boldsymbol{Y}|\gamma_A, A^s)} \frac{g_0(A^s - 1|\lambda_A)}{g_0(A^s|\lambda_A)} \frac{\rho_A(\rho_A + N^s_{\emptyset_A})}{N^s_{\emptyset_A}(\rho_A + N^s_{\emptyset_A} - 1)}.$$

C.5 Standard MCMC updates

C.5.1 Standard MCMC updates : intransitivity

Intransitivity levels update

Moves which do not propose any change to K^s or A^s can follow standard MCMC updates. The moves follow a Metropolis-Hastings-within-Gibbs procedure, where the skill levels $\boldsymbol{\phi}$, intransitivity levels $\boldsymbol{\theta}$, and cluster allocations $(\boldsymbol{Y}, \boldsymbol{Z})$ are updated sequentially.

The intransitivity levels are updated sequentially from smallest to largest, using the matching function (C.3.1) to ensure ordering is preserved. First, the proposed transformed intransitivity level $y(\theta'_1|\theta^s_0, \theta^s_2)$ is proposed via a symmetric random walk with standard deviation $\tau > 0$, such that $y(\theta'_1|\theta^s_0, \theta^s_2) \sim \mathcal{N}\left(y(\theta^s_1|\theta^s_0, \theta^s_2), \tau^2\right)$, and accepted

with probability $A^{(1)}_{levels}$, where

$$A^{(1)}_{levels} = \frac{\pi(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}', \boldsymbol{Z}^s, K^s|\boldsymbol{x})}{\pi(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s|\boldsymbol{x})} = \frac{L(\boldsymbol{x}|\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}', \boldsymbol{Z}^s, K^s)}{L(\boldsymbol{x}|\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s)} \frac{h_0\left(\theta'_1|\alpha, \beta\right)}{h_0\left(\theta^s_1|\alpha, \beta\right)}.$$

and if accepted then $\theta_1^{s+1} = \theta'_1$. The other intransitivity levels are then drawn similarly, but using the latest updated values of the intransitivity levels, such that

$$y(\theta'_k|\theta^{s+1}_{k-1}, \theta^s_{k+1}) \sim \mathcal{N}\left(y(\theta^s_k|\theta^{s+1}_{k-1}, \theta^s_{k+1}), \tau^2\right),$$

and θ'_k is accepted as the new parameter value θ_k^{s+1} with probability $A^{(k)}_{levels}$, where

$$A^{(k)}_{levels} = \frac{\pi(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}'_k, \boldsymbol{Z}^s, K^s|\boldsymbol{x})}{\pi(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s_k, \boldsymbol{Z}^s, K^s|\boldsymbol{x})} = \frac{L(\boldsymbol{x}|\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}'_k, \boldsymbol{Z}^s, K^s)}{L(\boldsymbol{x}|\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s_k, \boldsymbol{Z}^s, K^s)} \frac{h_0\left(\theta'_k|\alpha, \beta\right)}{h_0\left(\theta^s_k|\alpha, \beta\right)},$$

where $\boldsymbol{\theta}^s_k := \{\theta_1^{s+1}, \ldots, \theta^{s+1}_{k-1}, \theta^s_k, \ldots, \theta^s_{K^s}\}$ and $\boldsymbol{\theta}'_k := \{\theta_1^{s+1}, \ldots, \theta^{s+1}_{k-1}, \theta'_k, \theta^s_{k+1}, \ldots, \theta^s_{K^s}\}$.

Pair reallocation

Then the intransitivity cluster allocation $\boldsymbol{Z}$ is updated. This can be achieved by either a naïve approach, or using the conditional posterior approach. For the naïve approach, a new allocation is drawn from the prior $f(\boldsymbol{Z}|\gamma_K, K)$, using a two-step process. First, a new $\boldsymbol{\omega}'_{K^s}$ is drawn from the prior $\boldsymbol{\omega}'_{K^s}|K^s \sim \text{Dirichlet}\left(\gamma_K \mathbf{1}_{2K+1}\right)$. Then, the allocations are drawn from

$$\boldsymbol{Z}' = \{\boldsymbol{z}'_{\{i,j\}}|\boldsymbol{\omega}'_{K^s} \sim \text{multinomial}\left(1, \boldsymbol{\omega}'_{K^s}\right),\ \forall i > j \in \mathcal{I} \setminus \{1\}\},$$

with $\boldsymbol{z}'_{\{i,j\}} \perp\!\!\!\perp \boldsymbol{z}'_{\{h,m\}}|\boldsymbol{\omega}'_{K^s}$ for all $\{i,j\} \neq \{h,m\} \in (\mathcal{I} \setminus \{1\})^2$, and so by mixing over $\boldsymbol{\omega}'_{K^s}$ we are sampling from $f(\boldsymbol{Z}|\gamma_K, K)$. This will be fast, but may mix over iterations poorly since the proposal could be far away from the maximum posterior and so the acceptance rate could be very low.

Instead, the data can be used to inform the next proposal using the conditional

posterior. Consider the pair $\{i,j\} : i > j \in \mathcal{I} \setminus \{1\}$ at step s. The conditional posterior $p^s_{\{i,j\}}(y)$ that the pair $\{i,j\}$ is in cluster $y \in \{-K^s, \ldots, K^s\}$ is computed as

$$p^s_{\{i,j\}}(y) \propto L(\boldsymbol{x}_{\{i,j\}}|\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \tilde{\boldsymbol{z}}_{\{i,j\}}(y), K^s) f(\tilde{\boldsymbol{Z}}^s_{\{i,j\}}(y)|\gamma_K, K^s),\ \forall y \in \{-K^s, \ldots, K^s\}, \tag{C.5.1}$$

for all $i > j \in \mathcal{I} \setminus \{1\}$, where $\tilde{\boldsymbol{z}}_{\{i,j\}}(y)$ indicates that the pair $\{i,j\}$ is in cluster y, i.e.,

$$\tilde{\boldsymbol{z}}_{\{i,j\}}(y) := \{z_t,\ t \in \{-K^s, \ldots, K^s\} : z_y = 1, z_t = 0 \text{ otherwise}\},$$

and $\boldsymbol{x}_{\{i,j\}}$ comprises only data of comparisons between i and j, and the full allocations used in the prior $f(\tilde{\boldsymbol{Z}}^s_{\{i,j\}}(y)|\gamma_K, K^s)$ are given as

$$\tilde{\boldsymbol{Z}}^s_{\{i,j\}}(y) = \{\tilde{\boldsymbol{z}}_{\{i,j\}}(y), \{\boldsymbol{z}^s_{\{a,b\}}, \forall a > b \in \mathcal{I} \setminus \{1,i,j\}\}\},$$

which is just the usual set of allocations, except for the pair $\{i,j\}$, which is assigned to cluster y. Only the data $\boldsymbol{x}_{\{i,j\}}$ need be considered in the likelihood in equation (C.5.1) instead of the all the data, because the cluster allocation of all pairs is assumed to be independent, so only θ_{ij} will change, see definition (7.3.7), which in turn means that only the pairwise probabilities p_{ij} and p_{ji} will change, see equation (7.3.5).

Since K^s is finite, for all $s \in \{1, \ldots, S\}$, the probability $q^s_{\{i,j\}}(y)$ at iteration s that the pair $\{i,j\}$ belongs to a cluster y is found by simply normalising, to give

$$q^s_{\{i,j\}}(y) := \frac{p^s_{\{i,j\}}(y)}{\sum_{c=-K^s}^{K^s} p^s_{\{i,j\}}(c)},\ \forall y \in \{-K^s, \ldots, K^s\}.$$

Then, for the pair $\{i,j\}$, the proposed allocation is drawn from a multinomial distribution with probability $\boldsymbol{q}^s_{\{i,j\}} := \{q^s_{\{i,j\}}(y) : y \in \{-K^s, \ldots, K^s\}\}$, so that

$$\boldsymbol{z}'_{\{i,j\}} \sim \text{multinomial}\left(1, \boldsymbol{q}^s_{\{i,j\}}\right). \tag{C.5.2}$$

Note that it is is possible for $\boldsymbol{z}'_{\{i,j\}} = \boldsymbol{z}^s_{\{i,j\}}$.

So far we have explained the proposal for a given pair $\{i, j\}$. In order to reallocate all pairs, this proposal is made sequentially for all $i > j \in \mathcal{I} \setminus \{1\}$ as follows. All free pairs are initially placed into a *holding set* $\mathcal{H} = \{\{a, b\} : a > b \in \mathcal{I} \setminus \{1\}\}$. Also define an *allocated set* $\mathcal{L}$ which is initially empty. From the holding set $\mathcal{H}$, a pair $\{i, j\} \in \mathcal{H}$ is selected at random to have a new allocation proposed $\boldsymbol{z}'_{\{i,j\}}$, which is drawn from equation (C.5.2). This is then accepted with probability $A^{\{i,j\}}_{alloc} = 1$, $\forall\{i, j\} : i > j \in \mathcal{I} \setminus \{1\}$, see appendix C.8. Therefore $\boldsymbol{z}^{s+1}_{\{i,j\}} = \boldsymbol{z}'_{\{i,j\}}$ with probability 1 for all $i > j \in \mathcal{I}$. Then, the pair $\{i, j\}$ are placed in the allocated set, such that now $\mathcal{H} = \{\{a, b\} : a > b \in (\mathcal{I} \setminus \{1\})^2 \setminus \{i, j\}\}$, and the allocated set becomes $\mathcal{L} = \{\{i, j\}\}$. Pairs from the holding set continue to be reallocated until the holding set is empty, and all pairs have been reassigned.

C.5.2 Standard MCMC updates : skills

Skill levels update

Next the skill levels are updated sequentially from smallest to largest, and the new values are again proposed on the transformed scale, all of which ensures that order is preserved. Specifically, for any $a \in \mathcal{A}_{\{-0\}}$, then the new transformed parameter is proposed via a symmetric random walk with standard deviation $\tau_A > 0$, such that $y(\phi'_a|\phi^{s+1}_{a-1}, \phi^s_{a+1}) \sim \mathcal{N}\left(y(\phi^s_a|\phi^{s+1}_{a-1}, \phi^s_{a+1}), \tau^2_A\right)$. The proposed parameter is accepted with probability

$$A^{(a)}_{levels} = \frac{\pi(\boldsymbol{\phi}', \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s|\boldsymbol{x})}{\pi(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s|\boldsymbol{x})} = \frac{L(\boldsymbol{x}|\boldsymbol{\phi}', \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s)}{L(\boldsymbol{x}|\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s)} \frac{\pi\left(\phi'_a|\nu_A\right)}{\pi\left(\phi^s_a|\nu_A\right)}.$$

where π is the prior distribution of the skill levels $\boldsymbol{\phi}$, namely, normal with standard deviation $\nu_A > 0$. If accepted then $\phi^{s+1}_a = \phi'_a$.

Object reallocation

The skill cluster allocation is updated in the same way as the intransitivity cluster allocation, using the conditional posterior. By considering the object $i \in \mathcal{I} \setminus \{1\}$ at step s, the conditional posterior $p^s_{\{i\}}(w)$ that the object i is in cluster $w \in \{-A^s_-, \ldots, A^s_+\}$ is

$$p^s_{\{i\}}(w) \propto L(\boldsymbol{x}_{\{i\}}|\boldsymbol{\phi}^s, \tilde{\boldsymbol{Y}}_{\{i\}}(w), A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s) f(\tilde{\boldsymbol{Y}}^s_{\{i\}}(w)|\gamma_A, A^s),$$
$$\forall y \in \{-A^s_-, \ldots, A^s_+\},\ i \in \mathcal{I} \setminus \{1\}$$

where

$$\tilde{\boldsymbol{Y}}^s_{\{i\}}(w) := \{\tilde{\boldsymbol{y}}_{\{i\}}(w), \{\boldsymbol{y}^s_{\{i\}}, \forall i \in \mathcal{I} \setminus \{1, i\}\}\},$$

and $\tilde{\boldsymbol{y}}_{\{i\}}(w)$ indicates that object i is in cluster w, i.e.,

$$\tilde{\boldsymbol{y}}_{\{i\}}(w) := \{y_t,\ t \in \{-A^s_-, \ldots, A^s_+\} : y_w = 1, y_t = 0 \text{ otherwise}\}.$$

Here, $\boldsymbol{x}_{\{i\}}$ are only the data involving comparisons with object i, which is possible because all skill cluster allocations and skill clusters are assumed to be independent and so the likelihood of comparisons not involving object i will not be effected.

The probability $q^s_{\{i\}}(w)$ at iteration s that the object i belongs to a cluster w is again found by normalising, to give

$$q^s_{\{i\}}(w) := \frac{p^s_{\{i\}}(w)}{\sum_{c=-A^s_-}^{A^s_+} p^s_{\{i\}}(c)},\ \forall w \in \{-A^s_-, \ldots, A^s_+\},\ \forall i \in \mathcal{I} \setminus \{1\}.$$

Then the object i's the proposed allocation is drawn from a multinomial distribution with probability $\boldsymbol{q}^s_{\{i\}} := \{q^s_{\{i\}}(w) : w \in \{-A^s_-, \ldots, A^s_+\}\}$, so that

$$\boldsymbol{y}'_{\{i\}} \sim \text{multinomial}\left(1, \boldsymbol{q}^s_{\{i\}}\right),\ \forall i \in \mathcal{I} \setminus \{1\}. \tag{C.5.3}$$

Note that it is is possible for $\boldsymbol{y}'_{\{i\}} = \boldsymbol{y}^s_{\{i\}}$. Again, the full proposal for all objects is made sequentially for all $i \in \mathcal{I} \setminus \{1\}$ as follows. All objects, expect for the fixed object 1, are initially placed into a *holding set* $\mathcal{H}_A = \{i : i \in \mathcal{I} \setminus \{1\}\}$. Also define an *allocated set* $\mathcal{L}_A = \emptyset$. From the holding set $\mathcal{H}_A$, an object $i \in \mathcal{H}_A$ is selected at random to have a new allocation proposed $\boldsymbol{y}'_{\{i\}}$, which is drawn from equation (C.5.3). This is then accepted with probability $A^{\{i\}}_{alloc\ A} = 1,\ \forall\{i\} : i \in \mathcal{I} \setminus \{1\}$, see appendix C.8. The object i is then placed in the allocated set, such that now $\mathcal{H}_A = \{i : i \in \mathcal{I} \setminus \{1, i\}\}$, and the allocated set becomes $\mathcal{L}_A = \{i\}$. Objects from the holding set continue to be reallocated until the holding set is empty.

C.6 Initialisation

Based on the initialisation routine described in Section C.2, the full algorithm is presented as follows.

Algorithm 1 ICBT model: Reversible Jump algorithm.

Inputs: paired comparison data, prior parameters $(\lambda_K, \lambda_A\gamma_K, \gamma_A, \alpha, \beta)$.

Find maximum likelihood estimates for the Bradley-Terry skills $\{r_i : i \in \mathcal{I}\}$, and the associated pairwise probabilities $\{p^{(BT)}_{ik} : i \neq k \in \mathcal{I}\}$.

Find naive estimates for pairwise win probabilities $\{p^{(n)}_{ik} = w_{ik}/n_{ik} : i \neq k \in \mathcal{I}\}$.

Find initial estimates for intransitivity

$$\left\{ \theta_{ik} = \log\left(\frac{p^{(n)}_{ik}/(1 - p^{(n)}_{ik})}{p^{(BT)}_{ik}/(1 - p^{(BT)}_{ik})} \right) : i \neq k \in \mathcal{I} \right\}.$$

Cluster into intransitivity levels and skill levels using k-means clustering.

Pick model with the best BIC as initial model, which gives initial parameter estimates $(\hat{\boldsymbol{\phi}}_{\text{initial}}, \hat{\boldsymbol{Y}}_{\text{initial}}, A_{\text{initial}}, \hat{\boldsymbol{Z}}_{\text{initial}}, \hat{\boldsymbol{\theta}}_{\text{initial}}, K_{\text{initial}})$.

Set $(\boldsymbol{\phi}^0, \boldsymbol{Y}^0, A^0, \boldsymbol{\theta}^0, \boldsymbol{Z}^0, K^0) = (\hat{\boldsymbol{\phi}}_{\text{initial}}, \hat{\boldsymbol{Y}}_{\text{initial}}, A_{\text{initial}}, \hat{\boldsymbol{Z}}_{\text{initial}}, \hat{\boldsymbol{\theta}}_{\text{initial}}, K_{\text{initial}})$

for $s = 1, \ldots, S_1$, **do**

Update $\boldsymbol{\theta}$

Update $\boldsymbol{\phi}$

Store sample $(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s)$

end for

Set $(\boldsymbol{\phi}^0, \boldsymbol{Y}^0, A^0, \boldsymbol{\theta}^0, \boldsymbol{Z}^0, K^0) = \text{argmax}_s \pi(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s | \boldsymbol{x})$.

for $s = 1, \ldots, S_2$, **do**

Update $\boldsymbol{\theta}$

Update $\boldsymbol{Z}$

Update $\boldsymbol{\phi}$

Update $\boldsymbol{Y}$

end for

Store sample $(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s)$

Set $(\boldsymbol{\phi}^0, \boldsymbol{Y}^0, A^0, \boldsymbol{\theta}^0, \boldsymbol{Z}^0, K^0) = \text{argmax}_s \pi(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s | \boldsymbol{x})$

for $s = 1, \ldots, S$, **do**

Update $\boldsymbol{\theta}$

Update $\boldsymbol{\phi}$

Let $K^s = K^{s-1}, A^s = A^{s-1}$

end for

if $K^s = 1$ **then**

propose to split an intransitivity cluster

else

with probability 1/2 propose to split or merge an intransitivity cluster

end if

if there are no empty intransitivity clusters **then**

propose adding empty intransitivity cluster

else

with probability $\frac{N_\emptyset}{N_\emptyset+\rho_K}$ propose deleting empty intransitivity cluster

with probability $\frac{\rho_K}{N_\emptyset+\rho_K}$ propose adding intransitivity empty cluster

end if

if $A^s = 1$ **then**

propose to split a skill level with probability 1/2

else

propose to split or merge a skill level

end if

if there are no empty skill clusters **then**

propose adding empty skill cluster

else

with probability $\frac{N_\emptyset}{N_\emptyset+\rho_A}$ propose deleting empty skill cluster

with probability $\frac{\rho_A}{N_\emptyset+\rho_A}$ propose adding empty skill cluster

end if

Update $\boldsymbol{Z}$

Update $\boldsymbol{Y}$

Store sample $(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s)$

Set $(\boldsymbol{\phi}^0, \boldsymbol{Y}^0, A^0, \boldsymbol{\theta}^0, \boldsymbol{Z}^0, K^0) = \text{argmax}_s \pi(\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{Z}^s, K^s | \boldsymbol{x})$

C.7 Jacobian terms

C.7.1 Jacobian for intransitivity split move: $k \neq 0$

The Jacobian for the intransitivity split move J_{split}, for the general case where $k \neq 0$ is given as

$$J_{split} = \begin{vmatrix} \frac{\partial \theta'_{k'+1}}{\partial \theta_k} & \frac{\partial \theta'_{k'}}{\partial \theta_k} \\ \frac{\partial \theta'_{k'+1}}{\partial u} & \frac{\partial \theta'_{k'}}{\partial u} \end{vmatrix}.$$

Let $\alpha = \theta_{k-1},\ \beta = \theta_{k+1}$. From equations (C.3.4) and (C.3.5), we have

$$\begin{aligned}\theta'_{k'} =& \frac{\alpha + \beta \exp(-u)\,(\theta_k - \alpha)\,/\,(\beta - \theta_k)}{1 + \exp(-u)\,(\theta_k - \alpha)\,/\,(\beta - \theta_k)} \\ \theta'_{k'+1} =& \frac{\alpha + \beta \exp(u)\,(\theta_k - \alpha)\,/\,(\beta - \theta_k)}{1 + \exp(u)\,(\theta_k - \alpha)\,/\,(\beta - \theta_k)}.\end{aligned}$$

Therefore, using the quotient rule twice,

$$\begin{aligned}\frac{\partial \theta'_{k'}}{\partial \theta_k} &= \frac{\left[1 + \exp(-u)\left(\frac{\theta_k - \alpha}{\beta - \theta_k}\right)\right]\beta \exp(-u)\frac{\beta-\alpha}{(\beta-\theta_k)^2} - \left[\alpha + \beta\exp(-u)\left(\frac{\theta_k-\alpha}{\beta-\theta_k}\right)\right]\exp(-u)\frac{\beta-\alpha}{(\beta-\theta_k)^2}}{\left[1 + \exp(-u)\left(\frac{\theta_k-\alpha}{\beta-\theta_k}\right)\right]^2} \\ &= \frac{\beta\exp(-u)\frac{\beta-\alpha}{(\beta-\theta_k)^2}\left[1 - \frac{\alpha}{\beta}\right]}{\left[1 + \exp(-u)\left(\frac{\theta_k-\alpha}{\beta-\theta_k}\right)\right]^2}, \qquad \text{(C.7.1)}\end{aligned}$$

and similarly

$$\frac{\partial \theta'_{k'+1}}{\partial \theta_k} = \frac{\beta\exp(u)\frac{\beta-\alpha}{(\beta-\theta_k)^2}\left[1 - \frac{\alpha}{\beta}\right]}{\left[1 + \exp(u)\left(\frac{\theta_k-\alpha}{\beta-\theta_k}\right)\right]^2}, \qquad \text{(C.7.2)}$$

$$\begin{aligned}\frac{\partial \theta'_{k'}}{\partial u} &= \frac{\left[1 + \exp(-u)\left(\frac{\theta_k - \alpha}{\beta - \theta_k}\right)\right]\beta \exp(-u)(-1)\frac{\theta_k-\alpha}{\beta-\theta_k} - \left[\alpha + \beta\exp(-u)\left(\frac{\theta_k-\alpha}{\beta-\theta_k}\right)\right]\exp(-u)(-1)\frac{\theta_k-\alpha}{\beta-\theta_k}}{\left[1 + \exp(-u)\left(\frac{\theta_k-\alpha}{\beta-\theta_k}\right)\right]^2} \\ &= -\frac{\beta\exp(-u)\frac{\theta_k-\alpha}{\beta-\theta_k}\left[1 - \frac{\alpha}{\beta}\right]}{\left[1 + \exp(-u)\left(\frac{\theta_k-\alpha}{\beta-\theta_k}\right)\right]^2}, \qquad \text{(C.7.3)}\end{aligned}$$

$$\frac{\partial \theta'_{k'+1}}{\partial u} = \frac{\beta\exp(u)\frac{\theta_k-\alpha}{\beta-\theta_k}\left[1 - \frac{\alpha}{\beta}\right]}{\left[1 + \exp(u)\left(\frac{\theta_k-\alpha}{\beta-\theta_k}\right)\right]^2}. \qquad \text{(C.7.4)}$$

Equations (C.7.1), (C.7.2), (C.7.3), (C.7.4) therefore give a Jacobian of

$$J_{split} = \frac{2\beta^2\frac{(\beta-\alpha)(\theta_k-\alpha)}{(\beta-\theta_k)^3}\left[1 - \frac{\alpha}{\beta}\right]^2}{1 + 4\cosh(u)\left(\frac{\theta_k-\alpha}{\beta-\theta_k}\right) + (4 + 2\cosh(2u))\left(\frac{\theta_k-\alpha}{\beta-\theta_k}\right)^2 + 4\cosh(u)\left(\frac{\theta_k-\alpha}{\beta-\theta_k}\right)^3 + \left(\frac{\theta_k-\alpha}{\beta-\theta_k}\right)^4},$$

and $J_{merge} = 1/J_{split}$.

C.7.2 Jacobian for skill split move: $a \neq 0$

For the corresponding split of a skill cluster, assuming it is not the 0 cluster, i.e., $a \neq 0$, let $\alpha = \phi_{a-1},\ \beta = \phi_{a+1}$. Then similarly to Appendix C.7.1, the Jacobian for the general split move for the skill levels is given as,

$$J^A_{split} = \frac{2\beta^2 \frac{(\beta-\alpha)(\phi_a-\alpha)}{(\beta-\phi_a)^3}\left[1-\frac{\alpha}{\beta}\right]^2}{1+4\cosh(u_A)\left(\frac{\phi_a-\alpha}{\beta-\phi_a}\right)+(4+2\cosh(2u_A))\left(\frac{\phi_a-\alpha}{\beta-\phi_a}\right)^2+4\cosh(u_A)\left(\frac{\phi_a-\alpha}{\beta-\phi_a}\right)^3+\left(\frac{\phi_a-\alpha}{\beta-\phi_a}\right)^4},$$

where $J^A_{merge} = 1/J^A_{split}$.

C.7.3 Jacobian for intransitivity split move: $k = 0$

In the case that the transitive level has a split proposed, i.e., $k = 0$, then

$$\theta'_{k'} = \theta_k = \theta_0 = 0,$$

and so

$$\frac{\partial\theta'_{k'}}{\partial u} = 0, \quad \frac{\partial\theta'_{k'}}{\partial\theta_k} = 1.$$

Therefore, with α and β defined as in Appendix C.7.1, the Jacobian is given as

$$\begin{aligned} J_{split}|(k=0) &= \left|\frac{\partial\theta'_{k'+1}}{\partial u}|(k=0)\right| \\ &= \frac{\beta\exp(u)\frac{\theta_0-\alpha}{(\beta-\theta_0)^2}\left[1-\frac{\alpha}{\beta}\right]}{\left[1+\exp(u)\left(\frac{\theta_0-\alpha}{\beta-\theta_0}\right)\right]^2} \\ &= \frac{2\beta\exp(u)}{(1+\exp(u))^2}, \end{aligned}$$

where the last equality follows since $\alpha = -\beta$.

C.7.4 Jacobian for skill split move: $a = 0$

The Jacobian here is derived very similarly to Appendix C.7.3, except that, since the levels can be negative and are not reflected around 0, it is not necessarily true that $\alpha = -\beta$. Moreover, if a split occurs on cluster a with $\phi_a = 0$, then either $\phi'_{a'} = 0$ or $\phi'_{a'+1} = 0$, with equal probability, as explained in the algorithm. In the case that $\phi'_{a'} = 0$, then

$$\phi'_{a'} = \phi_a = \phi_0 = 0,$$

and

$$\frac{\partial \phi'_{a'}}{\partial u} = 0, \quad \frac{\partial \phi'_{a'}}{\partial \phi_a} = 1.$$

Therefore the Jacobian is given as

$$\begin{aligned} J_{split}|(a=0) &= \left| \frac{\partial \phi'_{a'+1}}{\partial u} \right| \\ &= \frac{\exp(u)^{\frac{-\alpha}{\beta}} \left[1 - \frac{\alpha}{\beta}\right]}{\left[1 + \exp(u)\left(\frac{-\alpha}{\beta}\right)\right]^2} \end{aligned}$$

However, if $\phi'_{a'+1} = 0$, then $\phi'_{a'+1} = \phi_a = \phi_0 = 0$, which leads to the same result.

C.8 Allocation acceptance probability

The probability of accepting the proposed allocation for a pair $\{i, j\}$, is

$$A^{\{i,j\}}_{alloc} = \frac{L(\boldsymbol{x}_{\{i,j\}}|\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{z}'_{\{i,j\}}, K^s) f(\boldsymbol{Z}'|\gamma_K, K^s)}{L(\boldsymbol{x}_{\{i,j\}}|\boldsymbol{\phi}^s, \boldsymbol{Y}^s, A^s, \boldsymbol{\theta}^s, \boldsymbol{z}^s_{\{i,j\}}, K^s) f(\boldsymbol{Z}^s|\gamma_K, K^s)} \frac{q_z(\boldsymbol{Z}^s|\boldsymbol{q}^s_{\{i,j\}})}{q_z(\boldsymbol{Z}'|\boldsymbol{q}^s_{\{i,j\}})},$$

where

$$\boldsymbol{Z}' = \{\boldsymbol{z}'_{\{i,j\}}, \{\boldsymbol{z}^s_{\{h\}} : h \in \mathcal{H}\}, \{\boldsymbol{z}^{s+1}_{\{l\}} : l \in \mathcal{L}\}\},$$

is the set comprising the proposed allocation, the allocations from pairs in the holding set, and the new allocations of any pairs in the allocated set, and

$$\frac{q_z(\boldsymbol{Z}^s|\boldsymbol{q}'_{\{i,j\}})}{q_z(\boldsymbol{Z}'|\boldsymbol{q}^s_{\{i,j\}})} = \frac{p^s_{\{i,j\}}\left(y = \text{argmax}\left[\boldsymbol{z}'_{\{i,j\}}\right]\right)}{p^s_{\{i,j\}}\left(y = \text{argmax}\left[\boldsymbol{z}^s_{\{i,j\}}\right]\right)},$$

which is precisely the conditional posterior, and thus the acceptance probability reduces to $A^{\{i,j\}}_{alloc} = 1,\ \forall\{i,j\} : i > k \in \mathcal{I}$. Similar logic can be followed for the skill allocations, showing that $A^{\{i\}}_{alloc\ A} = 1,\ \forall\{i\} : i \in \mathcal{I} \setminus \{1\}$.

C.9 Simulation studies

To evaluate the model, some simple simulation studies were conducted and outputs compared to a standard Bradley-Terry model. Data are simulated from our model, on which both our model and a standard Bradley-Terry model are fitted. Four scenarios are considered: one in which there is zero intransitivity and therefore could be modelled equally well by the Bradley-Terry model; and scenarios with one, two and three levels of intransitivity, i.e., $K \in \{0, \ldots, 3\}$. No matter the number of intransitivity levels used to simulate the data, when fitting the model the prior mean of the number of intransitivity levels was fixed at $\lambda_K = 2$ since in practice the true number of levels would not be known. In all cases a simple m-round robin tournament structure is simulated in which each object is compared to every other object m times. The four scenarios are tested for round robins of $m \in \{4, 8, 12, 16, 20, 40\}$, and in all cases had $n = 20$ objects, with the same $n/2$ skill levels simulated from independent normal variables with mean 0, such that the range of the levels were similar to that found in real data, see Section 7.5.2. Table C.9.1 shows the four scenarios with the value of the intransitivity levels. The values of the intransitivity levels were selected such that intransitivity in the simulated data would be noticeable, but not unrealistically large.

Moreover, these values of intransitivity were similar to values found in data of American League baseball, see Section 7.5.2. To give some intuition into these values, consider a paired comparison between two objects i, j, where $p_{ij}|\,(\theta_{ij} = 0) = 0.6$ when there is no intransitivity. Then our model gives $p_{ij}|\,(\theta_{ij} = 0.4) = 0.69$ and $p_{ij}|\,(\theta_{ij} = 1.2) = 0.85$. Similarly, if $p_{ij}|\,(\theta_{ij} = 0) = 0.9$, then $p_{ij}|\,(\theta_{ij} = 1.2) = 0.97$ and $p_{ij}|\,(\theta_{ij} = -1.2) = 0.73$.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4
K	0	1	2	3
$\boldsymbol{\theta}$	NA	0.7	$(0.5, 0.9)$	$(0.4, 0.8, 1.2)$

Table C.9.1: Scenarios for the simulation experiments. All four scenarios were tested on round robins of $m \in \{4, 8, 12, 16, 20, 40\}$.

Figure C.9.1 shows the posterior distribution of K for each of the four scenarios, with the different colours showing increasing round robins in m, from left to right. In scenario 1, Figure C.9.1 (top left), for $m = 4$ (black) the model indicates there is some probability that there are more than 0 clusters, but for $m = 8$ (red) and upwards, the model correctly finds no intransitivity levels. Figure C.9.1 (bottom left) indicates that the model has more difficulty finding the correct number of levels in scenario 2, although generally as the amount of data increases, the posterior distribution agrees more with the truth. In scenario 3 the model does well in identifying the true number of clusters, Figure C.9.1 (top right), and in scenario 4, Figure C.9.1 (bottom right), as the number of round robins becomes large ($m = 40$), the model again has no trouble identifying there being truly $K = 3$ intransitivity levels.

The discrepancy between the posterior for K and the truth in scenario 2 could be due to the prior on the number of intransitivity levels having mean $\lambda_K = 2$, instead of at the truth ($K = 1$ in this scenario). This possibility was explored via sensitivity analysis, by deliberate misspecification of the prior $K|\lambda_K$ in the inference, for a range of

$\lambda_K \in \{1, \ldots, 5\}$. In the most extreme misspecification, data $\boldsymbol{x}$ were simulated with $K = 2$ (and $m = 20$), but with the model then fitted with $\lambda_K = 5$. In this case, the posterior probability of the true number of intransitivity levels was $\pi(K = 2|\lambda_K = 5, \boldsymbol{x}) = 0.79$, indicating the model is robust to prior misspecification of the number of intransitivity levels. Moreover, the effect of prior misspecification on out of sample prediction was found to be minimal. The difference in the proportion of correctly predicted out of sample data between the model with the correctly specified prior and most misspecified prior was 5.4×10^{-3}, with the worst fitting model still predicting 1.8×10^{-3} better than a standard Bradley-Terry model. The difference in log-loss, see equation (7.4.5), was 3.8×10^{-4}. In all cases, whatever the choice of λ_K, the model always performed significantly better than a standard Bradley-Terry model.

Similar histograms were plotted (not shown) for the posterior probability of the number of skill levels for each scenario. In scenario 1 ($K = 0$) the posterior density of the number of skill levels agrees very closely with the truth even in the small data case ($m = 4$), whereas for the other scenarios it takes from $m = 8$ upwards for the posterior probability to start to agree with the truth. In general, for small m the posterior mean of the number of skill clusters is smaller than the truth, that is, the prior described in Section 7.4.2 leads to simplicity in these cases. As m increases, the posterior mean then increases and tends towards the truth. This suggests the model is fitting appropriately complex models given the data available and is therefore not over-fitting as a Bradley-Terry model may do. The posterior variance of A decreases as m increases.

Our assessment of the quality of fit is based on out of sample prediction accuracy - for each of the four scenarios an additional 1000 round robin tournament was simulated to be used as a test dataset, on which out of sample accuracy was assessed. Figure C.9.2 (left) shows log-loss and percentage of correct predictions (right), for the out of sample test dataset. For the transitive case (black) there is very little difference between the

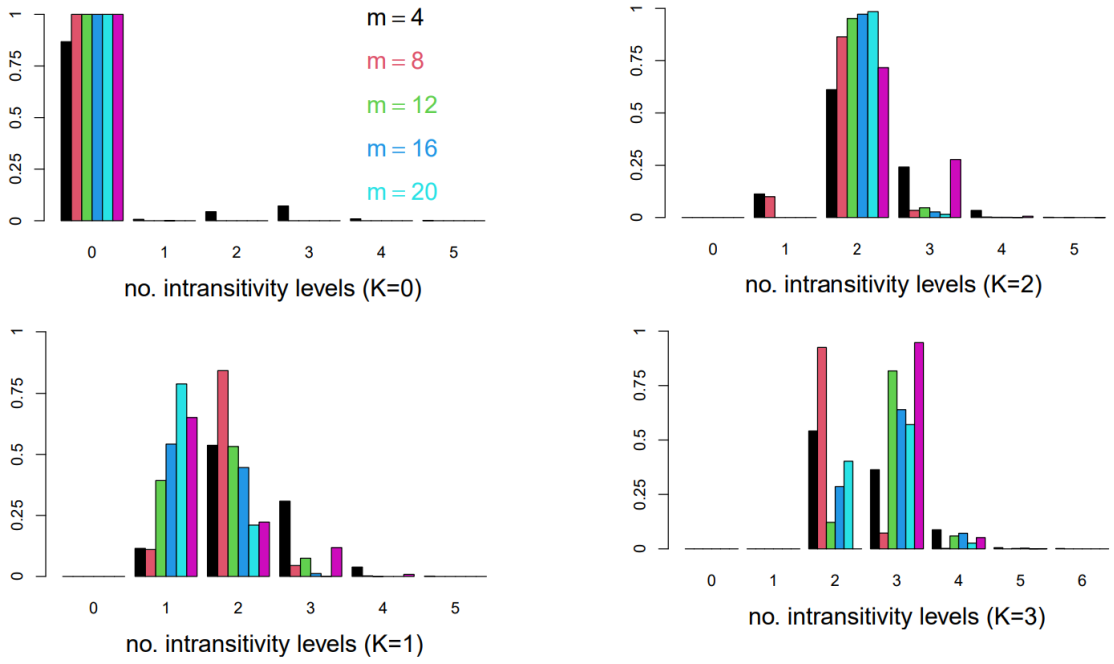


Figure C.9.1: Posterior probability of the number of intransitive levels for the four simulated scenarios: scenario 1 (top left), scenario 2 (bottom left), scenario 3 (top right) and scenario 4 (bottom right). For each scenario, the colours represent the increasing number of round robins from left to right: $m = 4$ (black), $m = 8$ (red), $m = 12$ (green), $m = 16$ (blue), $m = 20$ (cyan), and $m = 40$ (magenta).

Bradley-Terry fit and our model in both criteria, although perhaps our model performs slightly better when fitted to data from the $m = 4$ round robin case. This could be due to our model clustering the objects' skill levels and thus having less parameters (given our model also chooses $K = 0$ with a high probability in this scenario), and so the Bradley-Terry model may be over-fitting in this small data case. Particularly for scenario 2, our model is clearly over-fitting when the dataset is small, and a simpler Bradley-Terry model would perform better. However, as the amount of data increases, our model begins to perform better at around $m = 16$ round robins in terms of log loss, and by $m = 12$ in terms of percentage of correct predictions. A similar pattern occurs with scenario 4 in terms of log-loss, but here the improvement over the Bradley-Terry becomes very large as the number of round robins increases due to a larger presence of intransitivity. In this scenario our model always predicts a higher percentage of correct results, regardless of the number of round robins the models were fitted on. This discrepancy between log-loss and percentage of correct predictions in the small data case ($m = 4$), may indicate that when our model makes an incorrect prediction, the probabilities are far from the truth. This could be due to allocations of some pairs into incorrect intransitivity levels with limited data. For scenario 3 our model always has a better log-loss.

C.10 Baseball extra analysis

C.10.1 Hyper-parameter Selection

For running the sampler on all seasons of the baseball data, the same hyper-parameters were selected. The hyper-parameters are considered to belong to two categories: firstly, those that dictate the parameter values (the levels and cluster allocation), and secondly, those which dictate the model choice, i.e., the *number* of skill and intransitivity levels. Considering the first category, the *concentration parameters* for the (symmetric) *Dirich-*

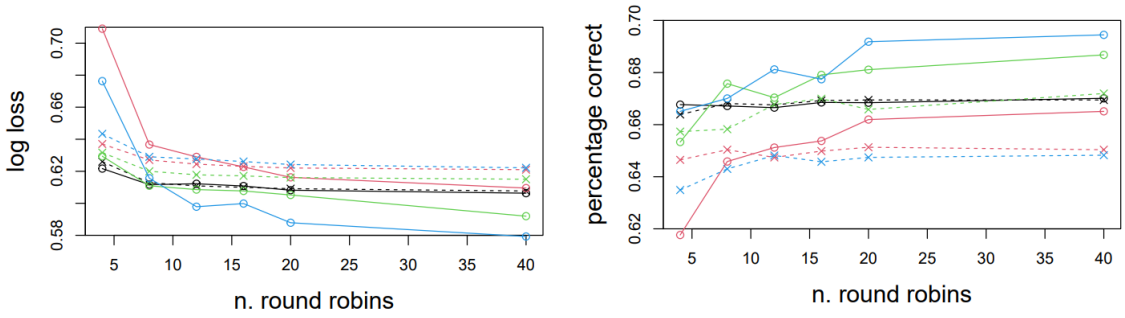


Figure C.9.2: Out of sample prediction performance assessed through: log-loss (left), percentage of correct predictions (right). Assessments are shown for all four scenario of Table C.9.1: $K = 0$ (black), $K = 1$ (red), $K = 2$ (green), and $K = 3$ (blue), for Bradley-Terry (crosses with dotted lines), and Intransitive Clustered Bradley-Terry (dots with solid lines) models.

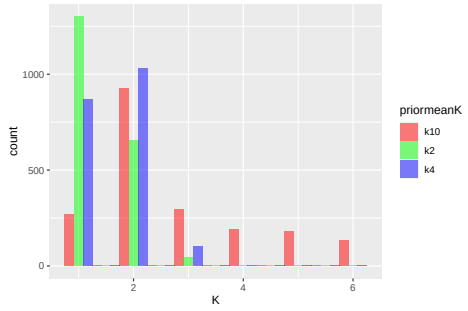
let priors, which informs the concentration of the clustering allocation, were both set to $\gamma_A = \gamma_K = 1$. This parameter choice allows for the data to be able to dictate whether the cluster allocations are dispersed equally across levels or highly concentrated in a small number of levels. The hyper-parameter which feeds into the prior on the team skill levels, $\nu_A = 1$ was selected due to the likely maximum win probability. In the small data scenario, the parameter $\nu_A = 1$ is likely to restrict the win probability in the transitive case to remain between $\approx (0.05, 0.95)$, so we felt this choice of prior did not restrict the team skill ratings. The hyper-parameters $\alpha = 1.5$, $\beta = 2$ for the gamma prior on the intransitivity levels was selected such that the mean is $\alpha/\beta = 0.75$. This was selected to reflect a reasonable shift in pairwise probability. For example, if two teams' i, j skill levels differ such that $r_i - r_j = 1$, then an intransitivity of $\theta_{ij} = 0.75$ increases the win probability from the transitive (Bradley-Terry) scenario $p_{ij}^{(BT)} = 0.73$ to $p_{ij} = 0.85$. We felt this would be a noticeable and yet still realistic change in win probability due to some strategic advantage.

The second class of hyper-parameters - those that effect the number of levels (or model selection) - were initially selected based on intuition as $\lambda_K = 2$ and $\lambda_A = n/2 = 7$;

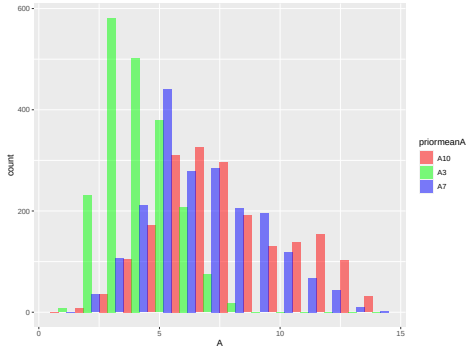
however, since the choice here is not so obvious, sensitivity analysis was also conducted. The hyper-parameter λ_K, the prior mean for the number of intransitivity levels, was chosen to be $\lambda_K = 2$, as it was thought that the different strategies of the pairs of baseball teams could perhaps be summarized by (on average) 2 different levels of extra advantage (disadvantage) (on top of Bradley-Terry) and with 95% chance in the range $[0, 5]$. This is achieved with a Poisson(2) distribution. The hyper-parameter $\lambda_K = 2$ was selected as it was thought that the different strategies of the pairs of baseball teams could be perhaps be summarized by two different levels of extra advantage (or disadvantage). The hyper-parameter λ_A, which dictates the prior mean of the number of skill levels A was set to $\lambda_A = n/2 = 7$ as we expected that the $n = 14$ teams could possibly be summarized by half the number of skill levels, as this Poisson(7) distribution has 95% interval $[2, 13]$. To test the sensitivity of these hyper-parameters, we analyse a season (2016) of the baseball data for values $\lambda_K = (2, 4, 10)$ and $\lambda_A = (3, 7, 10)$, see Figure C.10.1. Figure C.10.1a shows that even a value of $\lambda_K = 10$ doesn't pull the posterior too far from the other choices - the posterior probability of observing $K = 1$ is reduced, but the probability of observing $K = 2$ is similar to the two other options, even though $K = 2$ it is very far from the prior mean. This is likely because with the choice of Poisson distribution, the increased mean also increases variance, and so the prior is less informative. Figure C.10.1b shows that the number of skill clusters seems to be robust to the choice of λ_A, since the posterior distribution of A between models with $\lambda_A = 7$ and $\lambda_A = 10$ is very small. For $\lambda_A = 3$, the posterior distribution does appear to be quite different, however a value of $\lambda_A = 3$ is unlikely to be selected in practice.

C.10.2 2018 season

A main feature of our model is the capturing of pairwise interactions between teams. For the 2018 season, Figure C.10.2 (left) shows the posterior mean of the intransitivity



(a) Effect of choice of hyperparameter λ_K on posterior distribution of K: $\lambda_K = 10$ (red), $\lambda_K = 4$ (blue) and $\lambda_K = 2$ (green).



(b) Effect of choice of hyperparameter λ_A on posterior distribution of A: $\lambda_A = 10$ (red), $\lambda_A = 7$ (blue) and $\lambda_A = 3$ (green).

Figure C.10.1: hyperparameter sensitivity

parameter for each pair of teams $\hat{\theta}_{ij}$, $\forall i > j$, as intransitivity has rotational symmetry, i.e., $\theta_{ij} = -\theta_{ji}$, $\forall i \neq j$. The teams are sorted by their rank according to $\boldsymbol{p}_\cdot$, defined as in equation (7.3.9). Note that in Figure C.10.2 (left) the Los Angeles Angels (ANA) have an intransitivity level of 0 against all opponents; however, these levels were fixed at 0 for identifiability purposes rather than being a real interpretable feature of the team such as "ANA always play exactly as expected given their overall ability". A more interpretable representation is the intransitivity of the posterior mean $\{\hat{\theta}^*_{ij} : i \neq j\}$ given by definition (7.3.8), rather than the posterior mean of the intransitivity parameter. The reason for this being interpretable is due to

$$\theta^*_{ij} := \text{logit}(p_{ij}) - \text{logit}\left(p^{(BT))}_{ij}\right)$$

being defined as a function of pairwise probabilities. In the Bradley-Terry model, although a constraint on the parameters must be introduced in order to maintain identifiability, the resulting pairwise probabilities are invariant to this choice of constraint. The same is true of our model, that the pairwise probabilities are invariant to the choice

of constraints on the parameters. Therefore, by defining θ^*_{ij} as a function of pairwise probabilities, we can be sure that it too is invariant to the choice of identifiability constraints. Of course, the pairwise probabilities themselves are functions of underlying parameters since

$$\theta^*_{ij} = \text{logit}\left(p_{ij}\right) - \text{logit}\left(p^{(BT)}_{ij}\right) = \theta_{ij} + r_i - r_j - \left(r^{(BT)}_i - r^{(BT)}_j\right), \ \forall i \neq j \in \mathcal{I},$$

however, we need not worry about this because we know that θ^*_{ij} can be written as a function of pairwise probabilities, and therefore whatever combination of underlying parameters are involved in the expression must not depend on the choice of constraint. Clearly the two measures $\{\hat{\theta}_{ij} : i \neq j\}$ and $\{\hat{\theta}^*_{ij} : i \neq j\}$ correlate, but we believe the measure $\hat{\theta}^*_{ij}$ to be more interpretable. Considering Figure C.10.2 (right), pairs involving ANA still have intransitivity values closer to 0 than other pairs, albeit not exactly 0 since the values are not fixed as is the case in Figure C.10.2 (left). By redoing the analysis but fixing the intransitivity of all pairs involving a different team, namely Tampa Bay (TBA), we find that pairs involving ANA still have intransitivity values closer to 0 than other pairs. Moreover, neither the rankings nor model performance, see Section 7.5.3, differed significantly, indicating that the choice of fixed parameters does not impact the model.

The analysis of these intransitivities between pairs, and that of the skills of each team, can be combined to produce an overall ranking of the teams. For the 2018 season Figure C.10.3 (left) shows the ranking according to $\boldsymbol{p}_{\cdot}$. To help with a visual comparison, both ranking methods in Figure C.10.3 have been linearly scaled such that the best and worst teams have abilities 1 and 0 respectively. The ratings between $\boldsymbol{p}_{\cdot}$ and the Bradley-Terry ratings are clearly correlated, however, there is some difference in the ordering of the ranks, indicating that intransitivity may have been masking the true ranks of some teams. For example, consider TBA, ranked 6th by the Bradley-Terry model. TBA's good record against Kansas City (KCA) has a much lower weighting than their poor

record against BAL in the Bradley-Terry model due to the differing frequency of these match-ups, and therefore impacts the overall rank of TBA. Our Intransitive Clustered Bradley-Terry model however, recognises that good or bad records against particular teams could simply be due to the presence of intransitivity, and therefore penalises TBA less overall, ranking them 5th. This demonstrates that the model also accounts for *tournament structure*, such that teams are not penalised so heavily if they (unfairly) compete most frequently against those whom they perform systematically worse than expected based on skill alone. There is high correlation between the overall team abilities according to $\boldsymbol{p}_{\cdot}$ and the Bradley-Terry abilities, indicating that, although the Intransitive Clustered Bradley-Terry model is capturing more complex relationships between pairs, the overall rankings are not too dissimilar. Moreover, it manages to capture this information whilst using less parameters - on average 9.03 $(5, 13)$ with 95% credible interval shown in parentheses, as opposed to the $n - 1 = 14$ parameters required by the Bradley-Terry model. Figure C.10.3 (right) shows overall team abilities according to $\boldsymbol{a}$, see equation (7.3.10), which produces rankings which are more different to those of the Bradley-Terry.

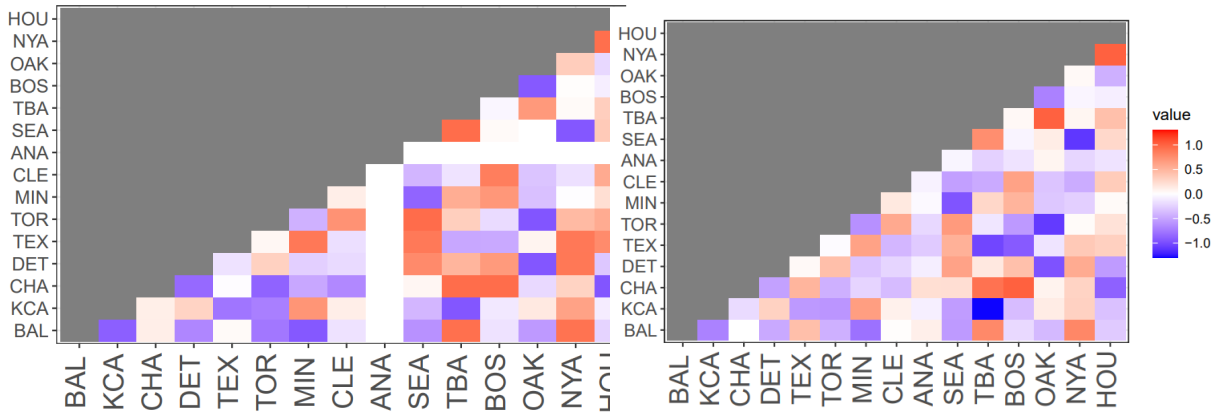


Figure C.10.2: Intransitivity between pairs of teams for the 2018 season: Posterior mean of the intransitivity, $\hat{\theta}_{ij} := \mathbb{E}\left[\theta_{ij}|\boldsymbol{x}\right], \ \forall i > j$ (left), and Intransitivity of the posterior mean, $\hat{\theta}^{*}_{ij}, \ \forall i > j \in \mathcal{I}$ (right).

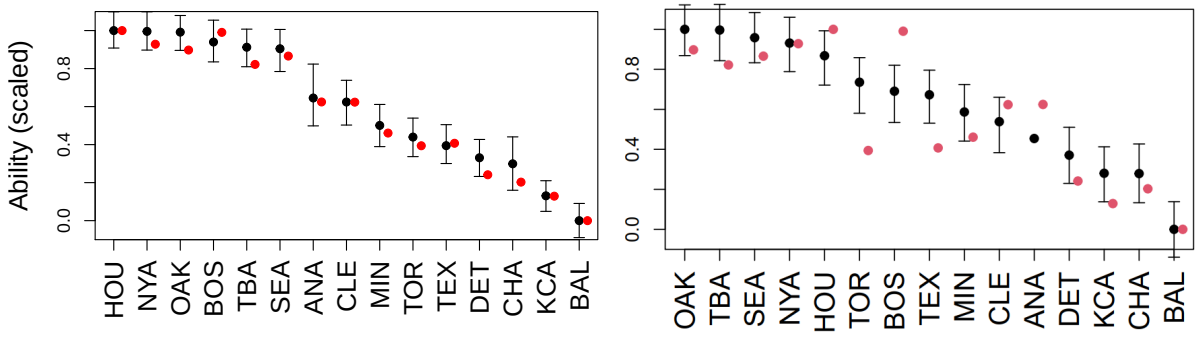


Figure C.10.3: Posterior mean of the overall abilities for the 15 American League baseball teams from the 2018 season with 95% credible intervals (black), according to $\boldsymbol{p}_{\cdot}$ (left) and $\boldsymbol{a}$ (right), and the corresponding scaled Bradley-Terry abilities (red). In each plot, teams are sorted in order of decreasing ability (by $\boldsymbol{p}_{\cdot}$ or $\boldsymbol{a}$ respectively). Uncertainties in the Bradley-Terry model (not shown) can be calculated via profile likelihood.

Of course drawing meaningful inference from the rankings is dependent on the rankings being accurate. To analyse this, we check the *ranking accuracy*, that is, the proportion of times that the better ranked team beats the worse ranked team. Maximising the ranking accuracy is not our ultimate aim, and could be done very simply by analysing all possible permutations of teams within the rankings. Note that this measure disregards all information about the probabilities, and must therefore be used with caution, but nevertheless provides a good sanity check for a ranking system. Across all seasons, ranking according to $\boldsymbol{p}_{\cdot}$, gave a higher ranking accuracy than the ranking accuracy according to $\boldsymbol{a}$, with the 2018 season giving ranking accuracies of 0.64 and 0.60 respectively. Similar plots and inferences are drawn from the other seasons (2010-2017) but with different team rankings in each year.

C.10.3 Pooled data

In addition to a season-by-season analysis, characteristics of the teams, or pairs of teams, which carry from one season to the next can be identified, by pooling data across seasons. Since in 2013 the Houston Astros moved from the National League to the American League, data are pooled from 2013-2018 American League.

Firstly, consider the intransitivity between pairs observed from the pooled data. Figure C.10.4 (left) shows the posterior pairwise intransitivity between pairs, but now for the pooled data from 2013-2018. The values are much closer to 0 in comparison to the 2018 data alone, see Figure C.10.2, or any other single year of data, indicating that there are significant differences in the strategies between pairs of teams from year to year, which when aggregated brings the intransitivity towards 0. By comparing Figures C.10.2 and C.10.4, we see that over seasons Texas (TEX) remains strong against Houston (HOU), and Houston remains weak against Cleveland (CLE).

An overall ranking can also be produced for the pooled data, which portrays the best and worst teams etc. over the 2013-2018 era. Figure C.10.5 indicates these rankings according to $\boldsymbol{p}_{\cdot}$. The model indicates that there is virtually no statistically significant difference between the teams based on these pooled data, even though significant differences are present for any individual season of data. This high variability in team ability was already suspected due to the previously identified significant changes in the intransitivities of the pairs from one year to the next. Considering intransitivity to be a marker of the combination of two teams' *strategies*, then we might expect a significant changes in strategy to accompany significant changes in overall abilities.

The statistically insignificant differences in overall rankings from these pooled data could be due in part to the aggregation of highly variable team abilities across the years. This is because time-dependency is not considered in our model, and so any significant differences in team ability that exists at a given point in time are blurred by even stronger trends in the teams' abilities over those 6 seasons. Thus when aggregated,

any significant differences in team ability at a given point in time are lost. Essentially, season-to-season form is being blurred by even stronger trends in the teams' abilities over those 6 seasons. To confirm this finding, the rankings are compared between a given year and pooled data excluding that year, for example, between the ranks from the 2015 season and the ranks based on pooled data of 2013, 2014, 2016, 2017, and 2018 seasons, using Spearman's rank hypothesis tests. The results from the Spearman's rank hypothesis tests indicate that every season produced a ranking which was statistically significantly different to the rankings produced from the pooled data excluding that year.

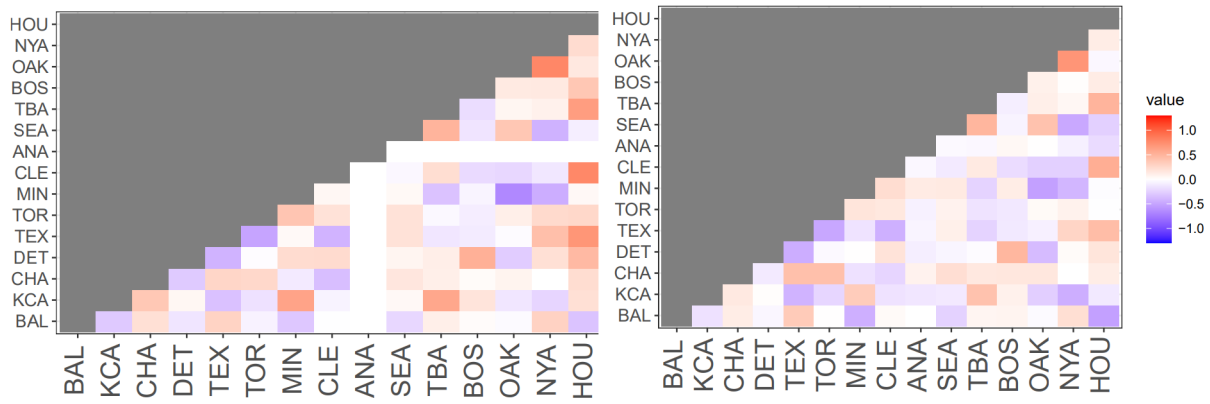


Figure C.10.4: Intransitivity between pairs of teams from pooled data of 2013-2018 seasons: posterior mean of the intransitivity, $\hat{\theta}_{ij},\ \forall i > j \in \mathcal{I}$ (left), and intransitivity of the posterior mean, $\hat{\theta}^*_{ij},\ \forall i > j \in \mathcal{I}$ (right).

In American League Baseball then, any team has the potential to improve or decline significantly over the years, a result unlikely to be found in many other sports, for example football (soccer), and which arguably makes American League Baseball a more competitive sport overall. This feature of the sport is likely due to Major League Baseball's "competitive balance tax", which levels teams' spending. For example, the largest spending of any team (https://www.spotrac.com/mlb/payroll) in American League Baseball this year was only a factor of 5 larger than the smallest. Contrastingly, in English Premier League football, this factor is over 200 times. It is

no surprise then that in football a small handful of teams seem to be consistently top ranked.

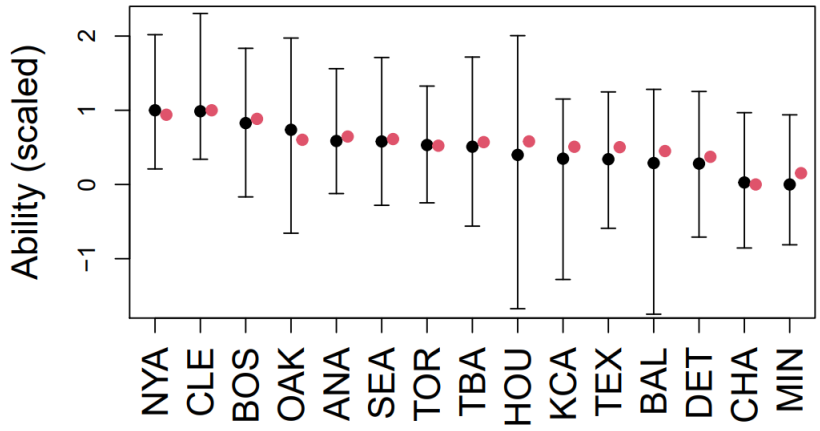


Figure C.10.5: Overall abilities, defined by $\boldsymbol{p}_{\cdot}$, from data pooled across seasons 2013-2018 (black). Corresponding Bradley-Terry rankings are shown in red.

Chapter 9

Concluding Remarks for Parts I and II

By expressing absolute systems and relative systems as two sides of the same coin, this thesis has unified two previously unconnected areas of statistical ranking methodology. The building of this *conceptual* bridge through the system-wide view now allows for the construction of its *mathematical* underpinnings, which have so far remained detached throughout. With a motivating example based on elite swimming, the first steps in unifying the underlying mathematics are eluded to.

In Chapter 4 only the competition minima was used for analysis, despite having data from all swimmers' event results. This filtering ensured that the remaining data were a good reflection of the swimmers' true abilities and were locally time identically distributed. But there is another interpretation of the necessity for this filtering, one which stems from the concept of systems.

With a view to conserve energy, the optimum performance in the heats can be considered as the slowest swim which still allows for advancement to the next stage of the competition. For slower swimmers, this may still translate as maximal effort, but the very best swimmers may deliberately and significantly under-perform. All that

is required, is to swim faster than their opponents. So, especially in the heats, it is clear that the time recorded is dependent on the context - the opponents - and so this cannot be a purely absolute system. So, the reason for filtering by competition minima - thereby discarding all but one data point per swimmer per competition - is because the method cannot handle the relativity that bleeds into this assumedly absolute system. To use all the available data would therefore require some unification of the mathematics presented in Part I and Part II: assuming that a swimmer's best times require only extreme value methodology, whilst acknowledging that the times recorded in the heats may need adjusting to reflect the strength of opposition. And so the ubiquity is revealed: rather than a binary categorisation of system - absolute or relative - perhaps a fluid description of the *relativity* of the system is more appropriate. This relativity would describe the importance of *context* in understanding a given system.

It is hoped that this ubiquitous system view of statistical ranking methodology helps in the future to bridge the gap between presently unconnected areas of statistics.

Bibliography

Adam, M. B. and Tawn, J. A. (2012). Bivariate extreme analysis of Olympic swimming data. *Journal of Statistical Theory and Practice*, 6(3):510–523.

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37:697–725.

Arderiu, A. and de Fondeville, R. (2022). Influence of advanced footwear technology on sub-2 hour marathon and other top running performances. *Journal of Quantitative Analysis in Sports*, 18(1):73–86.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multi-armed bandit problem. *Machine learning*, 47(2):235–256.

Balkema, A. A. and De Haan, L. (1974). Residual life time at great age. *The Annals of probability*, 2(5):792–804.

Bam, J., Noakes, T. D., Juritz, J., and Dennis, S. C. (1997). Could women outrun men in ultramarathon races? *Medicine and science in sports and exercise*, 29(2):244–247.

Barnett, V. (1976). The ordering of multivariate data. *Journal of the Royal Statistical Society: Series A (General)*, 139(3):318–344.

Blest, D. C. (1996). Lower bounds for athletic performance. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 45(2):243–253.

Bottolo, L., Consonni, G., Dellaportas, P., and Lijoi, A. (2003). Bayesian analysis of extreme values by mixture modeling. *Extremes*, 6:25–47.

Boulier, B. L. and Stekler, H. O. (1999). Are sports seedings good predictors?: an evaluation. *International Journal of Forecasting*, 15(1):83–91.

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

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117.

Carroll, J. D. and De Soete, G. (1991). Toward a new paradigm for the study of multiattribute choice behavior: Spatial and discrete modeling of pairwise preferences. *American Psychologist*, 46(4):342.

Casson, E. and Coles, S. G. (1999). Spatial regression models for extremes. *Extremes*, 1:449–468.

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

Causeur, D. and Husson, F. (2005). A 2-dimensional extension of the Bradley–Terry model for paired comparisons. *Journal of Statistical Planning and Inference*, 135(2):245–259.

Chavez-Demoulin, V. and Davison, A. C. (2005). Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):207–222.

Chen, J., Wang, C., Wang, J., Ying, X., and Wang, X. (2017). Learning the personalized intransitive preferences of images. *IEEE Transactions on Image Processing*, 26(9):4139–4153.

Chen, S. and Joachims, T. (2016). Modeling intransitivity in matchup and comparison data. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 227–236. ACM.

Coles, S., Heffernan, J., and Tawn, J. A. (1999). Dependence measures for extreme value analyses. *Extremes*, 2:339–365.

Coles, S. G. (2001). *An Introduction to Statistical Modeling of Extreme Values*, volume 208. Springer London.

Coles, S. G. and Tawn, J. A. (1994). Statistical methods for multivariate extremes: an application to structural design. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43(1):1–31.

Coles, S. G. and Tawn, J. A. (1996). A Bayesian analysis of extreme rainfall data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 45(4):463–478.

Colley, W. (2002). *Colley's bias free college football ranking method*. PhD thesis, Princeton University Princeton, NJ, USA.

Committee, I. O. et al. (2003). Statement of the stockholm consensus on sex reassignment in sports. *Retrieved on August*, 14:2019.

Cromwell, J. (1999). *Transmen and FTMs: Identities, bodies, genders, and sexualities*. University of Illinois Press.

Danielsson, J., de Haan, L., Peng, L., and de Vries, C. G. (2001). Using a bootstrap method to choose the sample fraction in tail index estimation. *Journal of Multivariate analysis*, 76(2):226–248.

Davison, A. C., Padoan, S. A., and Ribatet, M. (2012). Statistical modeling of spatial extremes. *Statistical Science*, 27(2):161–186.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds (with discussion). *Journal of the Royal Statistical Society: Series B*, 52(3):393–425.

De Boor, C. (1978). *A Practical Guide to Splines*, volume 27. Springer-Verlag New York.

de Fondeville, R. and Davison, A. C. (2022). Functional peaks-over-threshold analysis. *Journal of the Royal Statistical Society Series B*, 84(4):1392–1422.

De Schuymer, B., De Meyer, H., De Baets, B., and Jenei, S. (2003). On the cycle-transitivity of the dice model. *Theory and Decision*, 54(3):261–285.

Department for Education (2010). The importance of teaching: The schools white paper 2010.

Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002). *Analysis of Longitudinal Data*. Oxford University Press.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics (with discussion). *Journal of the Royal Statistical Society Series C: Applied Statistics*, 47(3):299–350.

Dixon, M. J. and Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2):265–280.

Duan, J., Li, J., Baba, Y., and Kashima, H. (2017). A generalized model for multi-dimensional intransitivity. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 840–852. Springer.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222.

Dupuis, D. J., Engelke, S., and Trapin, L. (2023). Modeling panels of extremes. *The Annals of Applied Statistics*, 17(1):498–517.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 1:89–102.

Elo, A. E. (1978). *The rating of chessplayers, past and present*. Arco Pub.

Engelke, S. and Hitz, A. S. (2020). Graphical models for extremes (with discussion). *Journal of the Royal Statistical Society Series B*, 82(4):871–932.

Engelke, S. and Ivanovs, J. (2021). Sparse structures for multivariate extremes. *Annual Review of Statistics and its Application*, 8:241–270.

Ewans, K. and Jonathan, P. (2008). The effect of directionality on northern North Sea extreme wave design criteria. *Journal of Offshore Mechanics and Arctic Engineering*, 130(4):041604.

Farkas, S., Lopez, O., and Thomas, M. (2021). Cyber claim analysis using generalized pareto regression trees with applications to insurance. *Insurance: Mathematics and Economics*, 98:92–105.

Ferro, C. A. T. and Segers, J. (2003). Inference for clusters of extreme values. *Journal of the Royal Statistical Society: Series B*, 65(2):545–556.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical proceedings of the Cambridge philosophical society*, volume 24, pages 180–190. Cambridge University Press.

Foster, L., James, D., and Haake, S. (2012). Influence of full body swimsuits on competitive performance. *Procedia engineering*, 34:712–717.

Fougères, A.-L., Holm, S., and Rootzén, H. (2006). Pitting corrosion: Comparison of treatments with extreme-value–distributed responses. *Technometrics*, 48(2):262–272.

Fougères, A.-L., Nolan, J. P., and Rootzén, H. (2009). Models for dependent extremes using stable mixtures. *Scandinavian Journal of Statistics*, 36(1):42–59.

Fyodorov, Y. V. and Bouchaud, J.-P. (2008). Freezing and extreme-value statistics in a random energy model with logarithmically correlated potential. *Journal of Physics A: Mathematical and Theoretical*, 41(37):372001.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Genel, M. (2017). Transgender athletes: how can they be accommodated? *Current sports medicine reports*, 16(1):12–13.

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

Glickman, M. E. (2001). Dynamic paired comparison models with stochastic variances. *Journal of Applied Statistics*, 28(6):673–689.

Glickman, M. E. and Jensen, S. T. (2005). Adaptive paired comparison design. *Journal of statistical planning and inference*, 127(1-2):279–293.

Glickman, M. E. and Stern, H. S. (2017). Estimating team strength in the nfl. In *Handbook of Statistical Methods and Analyses in Sports*, pages 129–152. Chapman and Hall/CRC.

Gomes, D. T. and Henriques-Rodrigues, L. (2019). Swimming performance index based on extreme value theory. *International Journal of Sports Science & Coaching*, 14(1):51–62.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Green, P. J. and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, 28(2):355–375.

Hankin, R. K. (2020). A generalization of the Bradley–Terry model for draws in chess with an application to collusion. *Journal of Economic Behavior & Organization*, 180:325–333.

Healy, D., Tawn, J. A., Thorne, P., and Parnell, A. (2023). Inference for extreme spatial temperature events in a changing climate with application to Ireland. *arXiv preprint arXiv:2111.08616*.

Heffernan, J. E. (2000). A directory of coefficients of tail dependence. *Extremes*, 3:279–290.

Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society: Series B*, 66(3):497–546.

Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Huser, R. and Wadsworth, J. L. (2019). Modeling spatial processes with unknown extremal dependence class. *Journal of the American Statistical Association*, 114(525):434–444.

Huub, T. and Trultens, M. (2005). Biomechanical aspects of peak performance in human swimming. *Animal Biology*, 55(1):17–40.

Hylton, K. (2008). *'Race'and sport: critical race theory*. Routledge.

Jacobs, R., Goddard, M., and Smith, P. C. (2005). How robust are hospital ranks based on composite performance measures? *Medical care*, pages 1177–1184.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. CRC press.

Jonathan, P. and Ewans, K. (2013). Statistical modelling of extreme ocean environments for marine design: a review. *Ocean Engineering*, 62:91–109.

Kent, J. T. (1982). Robust properties of likelihood ratio tests. *Biometrika*, 69(1):19–27.

Kéri, G. (2011). On qualitatively consistent, transitive and contradictory judgment matrices emerging from multiattribute decision procedures. *Central European Journal of Operations Research*, 19(2):215–224.

Klaassen, F. J. and Magnus, J. R. (2003). Forecasting the winner of a tennis match. *European Journal of Operational Research*, 148(2):257–267.

Laycock, P. and Scarf, P. (1993). Exceedances, extremes, extrapolation and order statistics for pits, pitting and other localized corrosion phenomena. *Corrosion Science*, 35(1-4):135–145.

Leadbetter, M. R. (1991). On a basis for 'peaks over threshold' modeling. *Statistics & Probability Letters*, 12(4):357–362.

Leadbetter, M. R., Lindgren, G., and Rootzén, H. (2012). *Extremes and Related Properties of Random Sequences and Processes*. Springer Science & Business Media.

Lebovic, J. H. and Sigelman, L. (2001). The forecasting accuracy and determinants of football rankings. *International Journal of Forecasting*, 17(1):105–120.

Leckie, G. and Goldstein, H. (2017). The evolution of school league tables in england 1992–2016:'contextual value-added','expected progress' and 'progress 8'. *British Educational Research Journal*, 43(2):193–212.

Ledford, A. W. and Tawn, J. A. (2003). Diagnostics for dependence within time series extremes. *Journal of the Royal Statistical Society: Series B*, 65(2):521–543.

Lindman, H. R. and Lyons, J. (1978). Stimulus complexity and choice inconsistency among gambles. *Organizational Behavior and Human Performance*, 21(2):146–159.

Ludkin, M. (2020). Inference for a generalised stochastic block model with unknown number of blocks and non-conjugate edge models. *Computational Statistics & Data Analysis*, 152:107051.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118.

Makhijani, R. and Ugander, J. (2019). Parametric models for intransitivity in pairwise rankings. In *The World Wide Web Conference*, pages 3056–3062. ACM.

Makowski, M. and Piotrowski, E. W. (2006). Quantum cat's dilemma: an example of intransitivity in a quantum game. *Physics Letters A*, 355(4-5):250–254.

Masarotto, G. and Varin, C. (2012). The ranking lasso and its application to sport tournaments. *The Annals of Applied Statistics*, 6(4):1949–1970.

Massey, K. (1997). Statistical models applied to the rating of sports teams. *Bluefield College*, 1077.

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

Montgomery, H. (1977). A study of intransitive preferences using a think aloud procedure. In *Decision Making and Change in Human Affairs*, pages 347–362. Springer.

Moria, H., Chowdhury, H., Alam, F., and Subic, A. (2011). Aero/hydrodynamic study of Speedo LZR, TYR Sayonara and Blueseventy pointzero 3 swimsuits. *Jordan Journal of Mechanical and Industrial Engineering*, 5(1):83–88.

Mosteller, F. (1951a). Remarks on the method of paired comparisons: I. the least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, 16(1):3–9.

Mosteller, F. (1951b). Remarks on the method of paired comparisons: Ii. the effect of an aberrant standard deviation when equal standard deviations and equal correlations are assumed. *Psychometrika*, 16(2):203–206.

Nelsen, R. B. (2007). *An Introduction to Copulas*. Springer Science & Business Media.

Nevill, A. M., Whyte, G. P., Holder, R. L., and Peyrebrune, M. (2007). Are there limits to swimming world records? *International Journal of Sports Medicine*, 28(12):1012–1017.

Northrop, P. J., Attalides, N., and Jonathan, P. (2017). Cross-validatory extreme value threshold selection and uncertainty with application to ocean storm severity. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 66(1):93–120.

Nutt, D. J., King, L. A., Phillips, L. D., et al. (2010). Drug harms in the uk: a multicriteria decision analysis. *The Lancet*, 376(9752):1558–1565.

Office, N. A. (2003). Making a difference: Performance of maintained secondary schools in england.

Olkin, I., Lou, Y., Stokes, L., and Cao, J. (2015). Analyses of wine-tasting data: A tutorial. *Journal of Wine Economics*, 10(1):4–30.

Pahikkala, T., Waegeman, W., Tsivtsivadze, E., Salakoski, T., and De Baets, B. (2010).

Learning intransitive reciprocal relations with kernel methods. *European Journal of Operational Research*, 206(3):676–685.

Pan, W. and Chen, L. (2013). Gbpr: Group preference based Bayesian personalized ranking for one-class collaborative filtering. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

Papastathopoulos, I. and Tawn, J. A. (2015). Stochastic ordering under conditional modelling of extreme values: drug-induced liver injury. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(2):299–317.

Pickands, J. (1971). The two-dimensional poisson process and extremal processes. *Journal of applied Probability*, 8(4):745–756.

Pickands, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119.

Propper, C., Burgess, S., and Green, K. (2004). Does competition between hospitals improve the quality of care?: Hospital death rates and the nhs internal market. *Journal of Public Economics*, 88(7-8):1247–1272.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*, 3(4):425–441.

Reichenbach, T., Mobilia, M., and Frey, E. (2007). Mobility promotes and jeopardizes biodiversity in rock–paper–scissors games. *Nature*, 448(7157):1046–1049.

Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-Thieme, L. (2009). BPR: Bayesian personalized ranking from implicit feedback. In *Proc. of Uncertainty in Artificial Intelligence*, pages 452–461.

Ribatet, M. (2013). Spatial extremes: Max-stable processes at work. *Journal de la Société Française de Statistique*, 154(2):156–177.

Richards, J. and Huser, R. (2022). A unifying partially-interpretable framework for neural network-based extreme quantile regression. *arXiv preprint arXiv:2208.07581*.

Richards, J., Tawn, J. A., and Brown, S. (2023). Joint estimation of extreme spatially aggregated precipitation at different scales through mixture modelling. *To appear in Spatial Statistics*.

Riegel, P. S. (1981). Athletic records and human endurance: A time-vs.-distance equation describing world-record performances may be used to compare the relative endurance capabilities of various groups of people. *American Scientist*, 69(3):285–290.

Robinson, M. E. and Tawn, J. A. (1995). Statistics for exceptional athletics records. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 44(4):499–511.

Rue, H. and Salvesen, O. (2000). Prediction and retrospective analysis of soccer matches in a league. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(3):399–418.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55.

Sander, A., Hamann, W.-R., and Todt, H. (2012). The galactic wc stars-stellar parameters from spectral analyses indicate a new evolutionary sequence. *Astronomy & Astrophysics*, 540:A144.

Scarrott, C. and MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT–Statistical Journal*, 10(1):33–60.

Scheffé, H. (1952). An analysis of variance for paired comparisons. *Journal of the American Statistical Association*, 47(259):381–400.

Schwartz, B. and Barsky, S. F. (1977). The home advantage. *Social forces*, 55(3):641–661.

Shibata, R. (1989). Statistical aspects of model selection. In *From Data to Model*, pages 215–240. Springer, Berlin, Heidelberg.

Shipley, A. (2009). FINA opts to ban all high-tech swimsuits. *Reach for the Wall. com*, 24.

Silver, N. (2010). The most livable neighborhoods in new york. *New York*, pages 32–43.

Simpson, E. S. and Wadsworth, J. L. (2021). Conditional modelling of spatio-temporal extremes for Red Sea surface temperatures. *Spatial Statistics*, 41:100482.

Simpson, E. S., Wadsworth, J. L., and Tawn, J. A. (2020). Determining the dependence structure of multivariate extremes. *Biometrika*, 107(3):513–532.

Sinervo, B. and Lively, C. M. (1996). The rock–paper–scissors game and the evolution of alternative male strategies. *Nature*, 380(6571):240–243.

Skinner, G. K. and Freeman, G. (2009). Soccer matches as experiments: how often does the 'best'team win? *Journal of Applied Statistics*, 36(10):1087–1095.

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. In *Annales de l'ISUP*, volume 8, pages 229–231.

Smead, R. (2019). Sports tournaments and social choice theory. *Philosophies*, 4(2):28.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, 72(1):67–90.

Smith, R. L. (1989). Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statistical Science*, 4(4):367–377.

Smith, R. L. and Goodman, D. (2000). *Bayesian Risk Analysis. Chapter 17 of Extremes and Integrated Risk Management, edited by P. Embrechts.* Risk Books, London.

Southworth, H. and Heffernan, J. E. (2012). Extreme value modelling of laboratory safety data from clinical studies. *Pharmaceutical Statistics*, 11(5):361–366.

Spearing, H., Tawn, J. A., Irons, D., and Paulden, T. (2023). Modeling intransitivity in pairwise comparisons with application to baseball data. *Journal of Computational and Graphical Statistics*, 0(0):1–10.

Spearing, H., Tawn, J. A., Irons, D., Paulden, T., and Bennett, G. (2021). Ranking, and other properties, of elite swimmers using extreme value theory. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(1):368–395.

Stephenson, A. G. and Tawn, J. A. (2013). Determining the best track performances of all time using a conceptual population model for athletics records. *Journal of Quantitative Analysis in Sports*, 9(1):67–76.

Strand, M. and Boes, D. (1998). Modeling road racing times of competitive recreational runners using extreme value theory. *The American Statistician*, 52(3):205–210.

Sykes, H. (2006). Transsexual and transgender policies in sport. *Women in Sport & Physical Activity Journal*, 15(1):3.

Sylvan Katz, J. and Katz, L. (1999). Power laws and athletic performance. *Journal of Sports Sciences*, 17(6):467–476.

Tawn, J. A. (1988). Bivariate extreme value theory: models and estimation. *Biometrika*, 75(3):397–415.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4):273.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tromans, P. and Vanderschuren, L. (1995). Response based design conditions in the north sea: application of a new method. In *Geology, Earth Sciences and Environmental Factors 1995 Offshore Technology Conference*, pages 387–397.

Tsai, R.-C. and Böckenholt, U. (2006). Modelling intransitive preferences: A random-effects approach. *Journal of Mathematical Psychology*, 50(1):1–14.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, 76(1):31.

Varty, Z., Tawn, J. A., Atkinson, P. M., and Bierman, S. (2021). Inference for extreme earthquake magnitudes accounting for a time-varying measurement process. *arXiv preprint arXiv:2102.00884*.

Wadsworth, J. L. and Tawn, J. A. (2022). Higher-dimensional spatial extremes via single-site conditioning. *Spatial Statistics*, 51:100677.

Wadsworth, J. L., Tawn, J. A., Davison, A. C., and Elton, D. (2017). Modelling across extremal dependence classes. *Journal of the Royal Statistical Society. Series B*, 79:149–175.

Wadsworth, J. L., Tawn, J. A., and Jonathan, P. (2010). Accounting for choice of measurement scale in extreme value modeling. *The Annals of Applied Statistics*, 4(3):1558–1578.

Winter, H. C. and Tawn, J. A. (2017). kth-order Markov extremal models for assessing heatwave risks. *Extremes*, 20:393–415.

Zanella, G. (2020). Informed proposals for local mcmc in discrete spaces. *Journal of the American Statistical Association*, 115(530):852–865.