Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

Detecting individual preferences and erroneous verdicts in mixed martial arts judging using Bayesian hierarchical models

Benjamin Holmes[a,b,*], Ian G. McHale[a], Kamila Żychaluk[b]

[a] *Centre for Sports Business, University of Liverpool Management School, L69 7ZH, UK*
[b] *Department of Mathematics, University of Liverpool, L69 7SH, UK*

ARTICLE INFO

Article history:
Received 16 September 2022
Accepted 3 July 2023
Available online 17 July 2023

Keywords:
OR in sports
Biases
Mixed martial arts
Bayesian statistics
Hierarchical model

ABSTRACT

In this paper, we use Bayesian hierarchical models to investigate the decision-making of judges of mixed martial arts (MMA) contests. Whilst there has been research into the judging of various sports in the past, none have explicitly modelled the judges' behaviours at an individual level. We progress the literature by demonstrating that judges have personal preferences towards the different actions that they must assess during a fight. The preferences themselves may be the deciding factor in a bout, as demonstrated using a historical case study. We apply the concept of variable significance to the predictions of scores, to assess whether a judge's verdict was within reason. Finally, we develop a model that predicts a bout's fair outcome, which could be used in various ways in MMA.

Crown Copyright © 2023 Published by Elsevier B.V.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

1. Introduction

Every day humans make decisions and judgements, be they conscious or subconscious, which for the most part, are inconsequential. However, every weekend sporting officials across the globe determine the fortunes of individual athletes, teams, stakeholders, and loyal supporters, sometimes resulting in controversy. Accurate and fair officiating is of paramount importance to the integrity of sport, and it is for this reason that public scrutiny of fairness within sport is a pursuit worthy of study.

Whilst referees within sports such as football and ice-hockey, or umpires within cricket and tennis, can indirectly influence outcomes, sports in which judges decide the final result are particularly vulnerable to spurious outcomes. Perhaps out of all these sports, mixed martial arts (MMA) judges face the most challenging task.

Many other sports have features making the judges' work more straightforward. For example, athletes compete independently in numerous Olympic sports, such as figure skating. Or in some judged sports, athletes have a set number of attempts to perform their best techniques, for instance, ski jumping. Finally, it is often the case that judges are assessing techniques which are the same, or at least very similar, and simple to rank. For example, in diving, a dive consisting of two somersaults is deemed better than a dive consisting of one; or in boxing, a judge only has to assess one type of offensive action: a punch.

However, MMA judges do not have things so simple. Fighters compete against one another, meaning judges have to assess the performance of two athletes simultaneously. Bouts can be 25 minutes long, resulting in hundreds of actions to assess, which can occur at any given second. Typically these actions are of many different types (e.g. punches, kicks, or throws). MMA comprises the full spectrum of martial arts, allowing athletes to implement strikes from sports such as boxing or taekwondo, throws from judo and wrestling, or chokes and joint-locks from Brazilian jiu-jitsu. These techniques are not always similar in their function, and can have varying degrees of impact, so it is often unclear how to score each. For instance, how does one score a punch aiming for the opponent's head that is partially blocked, versus a kick that lands flush on the opponent's leg?

Not only is their job extremely difficult, but they are under some of the most intense scrutiny of any judges. We believe two reasons contribute to this scrutiny: the sport's popularity, and the

[*] Corresponding author at: Centre for Sports Business, University of Liverpool Management School, L69 7ZH, UK.

E-mail addresses: B.Holmes@liverpool.ac.uk (B. Holmes), ian.mchale@liverpool.ac.uk (I.G. McHale), zychaluk@liverpool.ac.uk (K. Żychaluk).

https://doi.org/10.1016/j.ejor.2023.07.004
0377-2217/Crown Copyright © 2023 Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

consequences of winning and losing. First, compared to other subjectively judged sports, other than boxing, there is a much larger audience. The largest live crowd attendance for the top tier of MMA, the UFC (Ultimate Fighting Championship), was 57,127 fans in 2019 for UFC 243: Whittaker vs. Adesanya.[1] The largest pay-per-view event saw 2,400,000 buys in 2018 for UFC 229: Khabib vs. McGregor. The popularity of MMA means many fans will be scrutinising the judges' verdicts.

The second reason for the scrutiny experienced by UFC judges is that the consequences of winning and losing can be significant. Given the large sums of money on offer for an athlete to win a contest, the immediate impact on an individual can be staggering. For example, consider the title fight between Jon Jones and Dominick Reyes in August 2020. This fight resulted in a highly contentious decision made by the judges. All three judges scored the bout in favour of Jones, despite 76.4% of the public believing Reyes was victorious.[2] Whilst the payout to both athletes was in the hundreds of thousands, Reyes reportedly lost an estimated $150,000 win-bonus.[3]

The consequences of winning and losing are not limited to short-term financial gain. Losing a fight can drastically affect an athlete's future career prospects. Since Jones is widely regarded as the best MMA athlete of all time, had Reyes won, he could have begun to build a legacy as one of the sport's greatest competitors. His future fight(s) as the champion would have certainly been accompanied by larger payouts. MMA is an unforgiving sport: given the ever-changing rankings and increasing pool of talent within each organisation, and the limited amount of fights an athlete can compete in each year, losing a single bout can set an athlete back years as they have to begin climbing the rankings to achieve a title-shot again.

Given this background, and the importance placed on judges making good and fair decisions, in this paper, we develop a Bayesian hierarchical model to investigate judging within the UFC. We use our model to show that individual judges have different preferences regarding the techniques that athletes may attempt. These preferences can be the difference between winning and losing a fight. We believe this is an important finding to the sport. In the aftermath of controversial verdicts, athletes, fans, and stakeholders need to understand how an official may have come to their conclusion. But ultimately, our model can be used to homogenise judges' preferences such that unfair or controversial decisions are less commonplace. Our judging model can help train new judges, identify current judges who may be performing poorly, or even provide a benchmark score to assist when giving verdicts on fights.

In gymnastics Heiniger & Mercier (2021) developed tools to assess individual judges' scores objectively. Gymnastics is scored by penalising athletes for various errors, and it is a judge's task to accurately detect such errors during a given routine. Control scores can be derived post-competition by an outside judging panel using video reviews. Judges' scores can then be compared with the median of all other panel and control scores to assess their skill.

Judging in figure skating has come under much scrutiny. Accusations of corruption during the 1998 and 2002 Olympics led to a new scoring system being introduced[4], whilst the scoring system itself came under fire in Frederiksen & Machol (1988) who showed the system had paradoxical properties such as intransitivity.

Boxing judges have a long and notorious history of poor decisions. One of the most famous was a draw between Lennox Lewis and Evander Holyfield–in which the media and public believed Lewis clearly won the fight. This particular fight was the subject of research in Lee et al. (2002). The authors used exact tests, logistic regression, and a direct Bayesian model to demonstrate that two of the three judges scored the bout significantly different from other professionals. Interestingly, the authors acknowledge the key point we address in this paper: that judges may weigh the various criteria of boxing differently. However, they do not explicitly investigate or include this aspect.

Despite its popularity, the relative youth of MMA means that its judges have been the subject of just a handful of papers. Collier et al. (2012) and Feldman (2020) explore the effect of the various actions on the judges' decisions (finding that knockdowns are the most influential); Gift (2018) extends these models to include variables possibly indicative of bias. These three papers have a commonality: they model the population-level effects of actions on the judges' decisions. We progress the literature by implementing hierarchical models, allowing each judge to have their own effects.

Although identifying judges' preferences is clearly an important issue, not only has this not been discussed in the MMA judging literature, we are yet to find examples from any sport.

2. Data

We collected the scores within each round of UFC fights submitted by judges (and fans) from mmadecisions.com. Only fights which ended via a decision are available, limiting the dataset to fights where the judges' scores were used to decide on the outcome.

To model the judges' decisions as a function of the events occurring in a fight, and to identify how each judge valued each type of action, in-round fight statistics were scraped from ufcstats.com. The in-round statistics covered a variety of actions, including: 'strikes', the location of the strike (e.g. head, body, or leg), and the strength of the strike (split into two categories: significant or non-significant); 'takedowns' (actions used by a fighter to bring an opponent to the ground); 'control time' (how long a fighter was in a dominating grappling position in the round); 'submission attempts' (e.g. chokes and joint locks); 'reversals' (actions used to take a fighter from being controlled to being in control of a grappling exchange); and 'knockdowns' (strikes which cause an opponent to fall to the ground).

We choose to include the official rankings of the athletes to detect whether any judges are prone to expectation bias. The rankings of fighters were obtained from kaggle.com/datasets/martj42/ufc-rankings. The rankings are given for the top 15 fighters in each weight category, with all other fighters classed as 'unranked'. We assign unranked fighters a rank one more than the weight category maximum (i.e. we give them a ranking of 16 if the maximum rank was 15). It is convention in the UFC that the current champion is ranked at rank 0 (the other ranked fighters are effectively challengers to the champion). To make interpretations simpler, we reverse the order of the rankings, so that better rankings have larger numerical values.

We scraped each fighter's height, reach, and date of birth from ufcstats.com as these may affect how the judges view the fight.

Finally, we obtained information on the attendances at each event from wikipedia.org/wiki/List_of_UFC_events. During the COVID-19 pandemic, there were 51 events (281 fights) with no fans. These fights provide a unique opportunity to assess whether,

[1] https://www.tapology.com/search/mma-event-figures/ppv-pay-per-view-buys-buyrate.

[2] http://www.mmadecisions.com/decision/10877/Jon-Jones-vs-Dominick-Reyes.

[3] https://www.sportekz.com/mma/jon-jones-vs-dominick-reyes-purse-payouts.

[4] See, for instance, https://en.wikipedia.org/wiki/2002_Winter_Olympics_figure_skating_scandal.

and to what extent, the crowd influences individual judges, and whether judges favour fighters competing in their home country. We note there were 24 events (109 fights) where the UFC did not release the attendance. We omitted these events from the analysis.

Data from these separate sources needed to be merged, resulting in a final dataset of 17,105 unique judge's scores from 5,800 rounds in 1,840 fights spanning from 16/02/2013 to 18/06/2022. This included 309 unique judges who scored a median of 12 (mean of 55.36) rounds, with minimum and maximum rounds of 3 and 1,573, respectively. A brief explanation of the scoring criteria is given in Appendix A.

There were 38 rounds in which a fighter was deducted one point, and three rounds where a fighter was deducted two points. Since these deductions were applied at the referee's discretion, we chose to model the "adjusted" score of the judges and unapply these deductions. Of these adjusted scores, 16,359 (95.64%) rounds were scored 10-9. There were 728 (4.26%) scored as 10-8, just two (0.01%) scored as 10-7, and finally, 16 (0.09%) draws.

In addition to the main dataset, we were able to obtain the judgemental scores of fans for 1,832 of the fights. We found a median of 32 (mean of 94.35) fans submitted scores for a fight, the maximum was 4,030 (interestingly, for the aforementioned Jon Jones vs. Dominick Reyes fight), whilst the minimum was four. Again, we modelled the adjusted scores, and found 93.49% of scores were 10-9, 4.77% were 10-8, 0.20% were 10-7, and 1.54% were draws. This immediately shows that the fans are much more likely to submit rarer scores, particularly in the case of 10-10. We will use the fans' scores in a separate model from the judges' scores to compare how fans value actions to how judges do.

The variables used to predict which fighter won a round are as follows:

- *In-round statistics*: knockdowns, significant head/body/leg strikes landed/missed, takedowns landed/missed, reversals, control-time, submission attempts, non-significant strikes landed/missed.[5]
- *Fighter information*: age, height, reach, stance, official ranking, and whether the fighter is the weight category champion. 'Stance' describes how the fighter stands in a fight. The different categories of fighter stance are left foot forward (orthodox, 77.09% of fighters), right foot forward (southpaw, 18.21% of fighters), a mix of the two (switch, 4.55% of fighters), or neither (open, 0.15% of fighters). We used a single variable, 'orthodox stance', to assess how judges interpreted different styles.
- *Crowd information*: we include a binary indicator representing whether the athlete is fighting in their home country, as well as the interaction of this term with an indicator representing whether a live audience was present.
- *Judge identity*: we know the name of the judge awarding the scores, and use these to identify differences between judges' valuations of the different variables with regards to the variable's contribution to the judge's round score.

In the remainder of the paper, we will broadly refer to any variables that are not in-round statistics as 'bias' variables, since they should not directly influence a judge's verdict.

3. Methodology

We fit a hierarchical ordered-logit model in a Bayesian framework using the STAN software Stan Development Team (2021a) within the R statistical programming language R Core Team (2020). We use proportional odds model here as unlike continuation-ratio model, it is reversible and collapsible (Greenland, 1994), see Appendix C. Utilising a hierarchical Bayesian framework means that the common prior distributions will more heavily influence the coefficients of judges with limited data. Intuitively this makes sense since judges with a small number of observations will have their coefficients shrunk towards the common prior mean. A frequentist approach could result in large and unrealistic coefficient estimates.

Let $y_{rj}^{ab} \in \{$7-10, 8-10, 9-10, 10-10, 10-9, 10-8, 10-7$\}$ denote the score given by a judge j in round r from the perspective of fighter a facing opponent b. Suppose we have $n = 1, \ldots, N$ observations of judges' scores of unique rounds within fights, and for brevity, we will refer to these as y_n. Suppose there exists $j = 1, \ldots, J$ judges and $k = 1, \ldots, K$ predictors.

For each variable, we use the difference between the opposing athletes' values in that variable. This applies to any binary variables as well. For instance, if a home fighter is fighting an away fighter, they will have +1 and -1, respectively. Since the observations for opposing fighters are now mirror-images of each other, we randomly sample one observation to be used for model fitting. Further, we rescaled each variable by dividing by its maximum absolute value (this ensures differences of zero still have zero effect).

We model y_n using an ordered-logit regression with mean λ_n and thresholds indicating the cutoffs of each category denoted by $t = (t_1, \ldots, t_6)$.

To ensure our model is realistic, the probability of a fighter getting a 10-9 must be identical to the probability their opponent receives a 9-10. This is one issue not discussed by Gift (2018). To implement this in our ordered logit, we do not directly estimate the cutoffs of each threshold, but instead estimate the spacing. Imagine both fighters having not attempted any techniques start at 10-10. Any subsequent actions shift the predicted score probabilities away from 10-10 in either direction. Consequently, s_1 denotes the spacing between zero and the threshold of a fighter winning 10-9. Then, s_2 denotes the space between the 10-9 and 10-8 thresholds. Finally, s_3 denotes the space between winning 10-8 and 10-7. The vector $t = (-s_1 - s_2 - s_3, -s_1 - s_2, -s_1, s_1, s_1 + s_2, s_1 + s_2 + s_3)$ denotes the six cutoffs. We place a weakly informative half-normal prior on these spacings, $s_i \sim \text{Half-Normal}(0, 5)$, for $i = 1, 2, 3$. Consequently, we ensure the spacings are positive, the cutoffs are ordered correctly, and there is the required symmetry.[6]

Each judge has an individual set of parameters, representing the value they attribute to each action, denoted by $\beta_j = (\beta_{j1}, \ldots, \beta_{jK})$. We place a multivariate-normal (MVN) prior on the $J \times K$ parameters–as suggested by Gelman & Hill (2006, ch. 13)–enabling correlation between judge's preferences. One can imagine such correlations exist as judges may favour grappling or strikes, perhaps due to their background in martial arts.

A weakly-informative hyper-prior is placed on the mean of the MVN prior, such that $\mu_k \sim N(0, 5)$. The covariance matrix, Σ, is decomposed into a correlation matrix, Ω, and vector of coefficient scales $\sigma_{1,\ldots,K} \sim \text{Half-Normal}(0, 2.5)$ (Barnard et al., 2000). The cor-

[5] The non-significant strikes in the data are not as granular as the significant counter-parts and only split by whether they landed or not.

[6] Whilst we are surely not the first to implement a symmetrical ordered-logit model such as this, we note that we found no statistical packages implementing it within R. For instance, the most widely used function for implementing ordered logistic regression, `polr` within the `MASS` package, has no such feature.

Table 1
Summary of population-level effects in the model. Recall that each variable was rescaled by dividing by its maximum absolute value. Consequently, we display the “unit effect”, that is, the effect of a one unit increase in each variable, and order the table by these values.

Variable	Mean, μ	Unit effect	SD	HDI (2.5%)	HDI (97.5%)
Knockdowns	6.603	1.651	0.456	5.687	7.474
Submissions	3.997	0.799	0.311	3.387	4.606
Reversals	0.762	0.381	0.161	0.443	1.069
Takedowns landed	3.121	0.347	0.264	2.626	3.660
Home*Crowd	0.330	0.330	0.109	0.111	0.536
Champion	0.174	0.174	0.125	−0.070	0.419
Significant head landed	10.088	0.163	0.287	9.515	10.634
Significant body landed	4.407	0.110	0.330	3.759	5.061
Significant leg landed	2.845	0.102	0.158	2.537	3.156
Non-significant missed	2.285	0.051	0.603	1.108	3.456
Ranking	0.605	0.038	0.120	0.364	0.834
Significant body missed	0.596	0.030	0.190	0.226	0.959
Non-significant landed	3.021	0.021	0.369	2.292	3.749
Height	0.162	0.018	0.111	−0.055	0.385
Control-time	2.470	0.008	0.122	2.223	2.699
Significant leg missed	0.038	0.003	0.160	−0.294	0.339
Significant head missed	0.225	0.003	0.170	−0.115	0.554
Age	−0.138	−0.008	0.090	−0.324	0.026
Reach	−0.317	−0.026	0.120	−0.538	−0.072
Orthodox	−0.059	−0.059	0.039	−0.141	0.014
Takedowns missed	−1.058	−0.106	0.151	−1.366	−0.765
Home	−0.174	−0.174	0.110	−0.375	0.053

relation matrix is given a prior of LKJ(2), as recommended in Stan Development Team (2021b, ch. 1.13). The LKJ(η) distribution is a probability distribution over positive definite symmetric matrices with unit diagonals (Lewandowski et al., 2009). Thus, it provides a prior over the correlation matrices. When $\eta = 1$, the density is uniform over all matrices. When $0 < \eta < 1$, stronger correlations (which can be positive or negative) are favoured, and approach perfect correlation as η approaches 0. Alternatively, when $\eta > 1$ weaker correlations are favoured, and approach 0 as η approaches infinity.

The model in full is thus as follows:

$$\begin{aligned}
y_n &\sim \text{Ordered-Logit}(\lambda_n, t)\\
\lambda_n &= \beta_{j_n} x_n\\
t &= (-s_1 - s_2 - s_3, -s_1 - s_2, -s_1, s_1, s_1 + s_2, s_1 + s_2 + s_3)\\
\beta_j &\sim \mathcal{N}_K(\mu, \Sigma)\\
\Sigma &= \text{Diag}(\sigma)\Omega\text{Diag}(\sigma)\\
\mu_k &\sim \mathcal{N}(0, 5)\\
\sigma_k &\sim \text{Half-Normal}(0, 2.5)\\
\Omega &\sim \text{LKJ}(2)\\
s_{1,2,3} &\sim \text{Half-Normal}(0, 5).
\end{aligned}$$

To increase efficiency and improve the likelihood of convergence, we re-parameterise the model using the ‘non-centred parameterisation’ (Stan Development Team, 2021b, ch. 23.7, p. 276). A summary of the non-centred parameterisation is given in Appendix B.

We run four chains, each with 2,000 samples and 2,000 warm-up iterations. The smallest effective sample size was 1,482.13, and the largest $\hat{R}$ (the potential scale reduction factor) was 1.002; both indicate convergence.

We also have information on how fans have scored each round. The data is in the form of the number of fans awarding each score. We thus fit a model with y as the round score but weight the observation according to the proportion of fans who scored that particular round in that way. The full fans model has been included in Appendix D, and we discuss comparisons with the judges model in the next section.

To ensure robustness of our results, we experimented with two other models. First, we incldued wider hyper-priors: Normal(0, 10), Half-Normal(0, 5), LKJ(1). Second, we used alternative distributions with similar shapes as the original priors: a multivariate student-t distribution, with 5 degrees of freedom and scale of 5, instead of the MVN; and Gamma, with shape 1.73 and scale 1.15, instead of the half-Normal. In both experiments, none of the results presented in the following changed significantly.

4. Results

4.1. Population effects

Table 1 presents a summary of the latent population-level effects of each variable on the judges’ scores, indicated by μ in the model. These parameters indicate the average effect of each action. We also report the 2.5% and 97.5% Highest Density Intervals (HDI) for each coefficient.

All in-round actions, other than *Takedowns missed*, have a positive effect, in-line with the Unified Rules and past literature. As one would expect, “big” moves such as *Knockdowns* and *Submissions*–which both have the potential to finish a fight immediately–have the largest unit effects.

We see there is a non-significant effect for the home fighter main effect (*Home*), but a significant positive effect for the interaction with a live audience (*Home**Crowd*). This would suggest that the crowd influences the judges.

We also find a significant positive effect from the official rankings (*Ranking*), suggesting higher ranked fighters are overly favoured (recall, we reversed the order of rankings).

4.2. Individual preferences

The focus of this research is to investigate judges’ preferences at an individual level and discern whether significant differences in judges’ decision-making exist. Fig. 1 displays the posterior densities of each coefficient for the 25 judges who scored the most

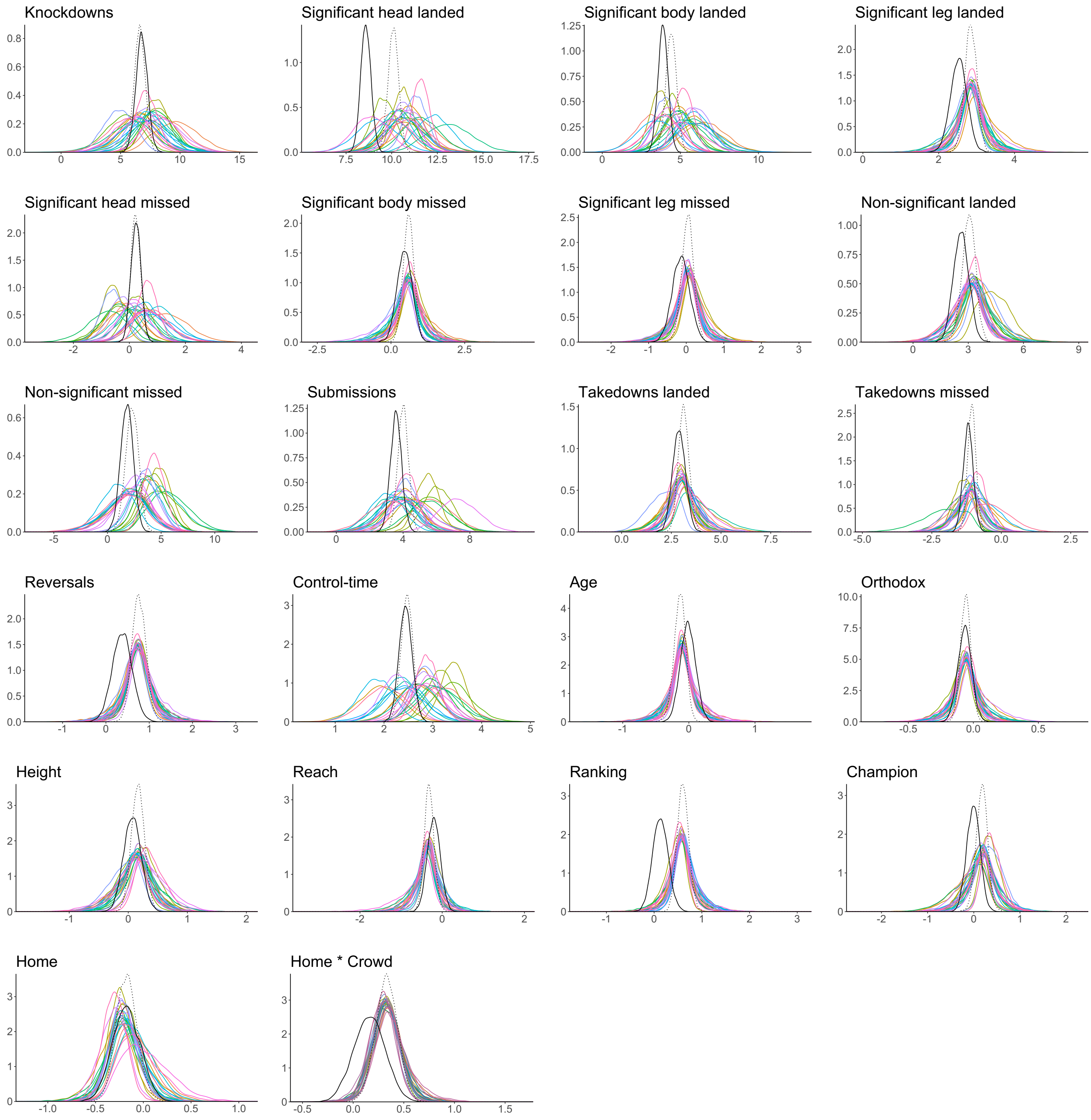


Fig. 1. Plots of the posterior densities for the 25 judges who scored the most rounds, the latent population density (dotted black), and the fans (solid black).

rounds within the data. The latent population-level effect is shown as the dotted black line for reference. Further, the corresponding density based on the model fitted to the fans' scores for rounds is displayed as a solid black line.

We can immediately see the disagreement in how judges value several actions. The most striking difference is for *Significant head missed*, where some judges deem this as a positive effect, yet others see it as a negative. This itself is not surprising: whilst landing strikes should clearly have a positive impact on a fighter's score, it is harder to definitively say who, if anyone, should benefit from missing strikes. Should the defending fighter benefit from good defence in dodging the incoming strike? Or should the attacking fighter be awarded for being aggressive despite missing? The Unified Rules state: “No scoring is given for defensive manoeuvres. Using smart, tactically sound defensive manoeuvres allows the fighter to stay in the fight and to be competitive”. This would suggest that whilst neither benefits in the ‘effective striking’ criteria; perhaps the attacker would benefit through ‘aggressiveness’. Consequently, by the official rules, we would argue that those who value missed significant head strikes as a negative are incorrect.

Table 2
Tables showing: summaries of the Jensen–Shannon divergences calculated for the 600 individual pairs of judges selected from the 25 judges who scored the most rounds in the data (left); and the Jensen–Shannon divergence between the fans opinions and the overall population of judges for each variable included in the model (right).

Judges					*Fans*	
Variable	Median	Mean	Max	SD	Variable	JSD
Control-time	0.219	0.297	0.987	0.258	Significant head landed	0.978
Significant head missed	0.169	0.250	0.883	0.221	Ranking	0.826
Significant head landed	0.161	0.240	0.903	0.221	Reversals	0.480
Significant body landed	0.175	0.236	0.784	0.197	Significant leg landed	0.378
Submissions	0.124	0.235	0.918	0.238	Significant body landed	0.345
Non-significant missed	0.163	0.202	0.763	0.185	Champion	0.271
Knockdowns	0.121	0.180	0.821	0.171	Home*Crowd	0.257
Takedowns missed	0.093	0.138	0.717	0.142	Submissions	0.248
Takedowns landed	0.052	0.089	0.609	0.103	Age	0.238
Champion	0.051	0.089	0.465	0.097	Non-significant landed	0.169
Home	0.045	0.075	0.425	0.079	Significant leg missed	0.158
Non-significant landed	0.035	0.072	0.471	0.092	Reach	0.122
Height	0.040	0.070	0.440	0.079	Takedowns landed	0.116
Significant leg landed	0.028	0.052	0.344	0.060	Takedowns missed	0.114
Significant body missed	0.032	0.050	0.290	0.055	Height	0.106
Reach	0.030	0.048	0.279	0.050	Non-significant missed	0.105
Orthodox	0.029	0.047	0.324	0.053	Significant body missed	0.085
Ranking	0.024	0.037	0.204	0.039	Orthodox	0.032
Significant leg missed	0.020	0.037	0.228	0.043	Home	0.030
Reversals	0.017	0.028	0.207	0.031	Knockdowns	0.025
Age	0.017	0.025	0.132	0.024	Control-time	0.012
Home*Crowd	0.009	0.016	0.107	0.018	Significant head missed	0.002

Control-time also has a wide spread of densities; at least in this case, all judges agree on the sign of the effect. Control-time is one of the more complex and subjective actions to assess. The Unified Rules state that “top and bottom position fighters are assessed more on the impactful/effective result of their actions, more so than their position”. So merely being in control of an opponent should not weigh more than establishing an offence from a dominant position.

Looking at *Submissions*, there appear to be a few judges who are far from the others. In particular, one judge is almost entirely separate from the population-level density.

There are several actions which are largely agreed upon both in size and sign of effect: *Significant leg landed*, *Takedowns landed*, *Takedowns missed*, and the ‘bias’ variables *Height*, *Reach*, *Ranking*, and the *Home*Crowd* interaction.

We can examine the differences between opinions in a more robust way by examining the Jensen–Shannon divergence (JSD) between the densities of the judges. JSD is a normalised and symmetric version of the Kullback–Leibler divergence (Lin, 1991). We want to measure the disagreement between two judges, so the symmetry between divergences is an attractive property of JSD. For the top 25 judges displayed in Fig. 1, we calculate the JSD between each pair of unique judges; consequently, we obtain 600 JSD values. In Table 2, we report the median, mean, maximum and standard deviation over these values to observe the overall measure of disagreement between judges.

We see that *Control-time* is the most controversial of the variables, followed by *Significant head missed*, which were both clear from the discussed posterior densitites. The least controversial in-round statistic was *Reversals*. Of the bias variables, the most controversial was *Champion*, and least was *Home*Crowd*.

4.3. Comparison with the fans

In this section, we look at how fans value each in-round action when judging the winner of the round, and compare the fans' valuations with the judges'. The fans effectively act as a crowd, and there have been several studies on the wisdom of crowds in sports. Brown & Reade (2019), for example, look at the wisdom of amateur crowds, like ours is, in predicting the outcome of sporting events, including martial arts. As is mostly the case in such studies, the crowd proves to be ‘wise’. Here, we do not know the ground truth (what the round should have been scored), but we can compare the crowd (the fans) with the judges.

For all variables, the fans' opinions are within the observed densities of the individual judges, and for most variables, there is an overlap with the overall population effects. Several of the variables are almost identically weighted: *Knockdowns*, *Significant head missed*, and *Control-time* for instance. Some variables have more obvious differences: *Significant head landed* and *Reversals*, but, for the most part, it appears the fans and judges value actions very similarly to the judges.

One interesting finding is that the fans appear to be *less* influenced than the judges by the bias variables. The effects for the official rankings and being a home fighter are of particular interest. We see notable differences between the fans and the judges, whereby the coefficients for the fans are closer to zero than the judges. The most likely reason for bias towards home fighters in front of a live audience is that noise sways the judges. The fans who submit scorecards to mmadecisions will (most likely) not be present in the audience and thus will be less exposed to the noise. Perhaps the powers that govern UFC might consider having one or more judges away from the arena, similar to how Video Assisted Referees (VAR) operate in football.

We can also calculate the JSD for each variable between the fans and the overall population of judges (recall this is the latent variable μ in the model). This allows us to more robustly say which variables the fans and judges disagree most on. Table 2 displays the JSD between the fans and judges for each variable.

Interestingly, several bias variables rank highly in terms of disagreement: the difference in rankings, interaction between an au-

Table 3
Probabilities of each scoreline predicted by the judge and fan models for several different pseudo-rounds. For each q we calculate the q'th quantile of the absolute value of each variable and include these values as the variables for the observation. Consequently, $q = 0$ represents a round in which both fighters were exactly even across all variables, whilst $q = 1$ represents the most one-sided round possible in which the fighter scored the maximum in each variable. We ensured that variables with an estimated negative coefficient were accounted for.

	Judges				*Fans*			
q	10-10	10-9	10-8	10-7	10-10	10-9	10-8	10-7
0.00	0.002	0.498	0.001	0.000	0.029	0.483	0.002	0.000
0.25	0.002	0.672	0.001	0.000	0.026	0.632	0.004	0.000
0.50	0.001	0.914	0.006	0.000	0.011	0.870	0.020	0.000
0.75	0.000	0.884	0.112	0.000	0.001	0.780	0.209	0.002
0.85	0.000	0.487	0.513	0.000	0.000	0.371	0.613	0.014
0.90	0.000	0.114	0.886	0.000	0.000	0.097	0.823	0.071
0.95	0.000	0.001	0.970	0.028	0.000	0.002	0.216	0.782
0.98	0.000	0.000	0.491	0.509	0.000	0.000	0.013	0.987
1.00	0.000	0.000	0.000	1.000	0.000	0.000	0.000	1.000

dience and home fighter, and whether one athlete is the champion, are all within the top ten. We have seen from Fig. 1 that the fans are less influenced than the judges with regards to each of these variables.

We will now compare the differences in the scoreline thresholds, s_i (which have not been plotted). We will create several pseudo-fights to assess the probabilities of the different scorelines. For a given value q, we will find the q'th quantile of the absolute value of each variable.[7] These quantiles become the variables in the associated pseudo-observation. Consequently, $q = 0$ represents the closest round possible, in which the athletes were even across all variables, whilst $q = 1$ is the most one-sided round possible, in which the dominant fighter scored the maximum of each variable. We note that whilst we explicitly allowed correlations between the variables in the model, we have not included such correlations in this set-up.

The predicted probabilities for each scoreline for several different values of q are shown in Table 3.

One "oddity" to note from Table 3 is that in close rounds ($q = 0$), it is more likely that the judge picks one of the fighters to win the round than give a draw (10-10). It would seem more natural to score a close round as a tie. However, this is an effect of the Unified Rules which actively discourage the use of 10-10 scorelines: "A 10-10 round in MMA should be extremely rare and is not a score to be used as an excuse by a judge that cannot assess the differences in the round... If there is any discernible difference between the two fighters during the round the judge shall not give the score of 10-10". Even when examining a fake fight in which each fighter was exactly even across all variables (that is, $q = 0$), the mean posterior predictive probability of a 10-10 round is just 0.002.

However, the fans are much more likely to give a 10-10 round: for $q = 0$, fans have a 0.029 probability. Proportionately this is a massive difference, which perhaps demonstrates the fans are not aware of the rules actively discouraging tied rounds, or they are more willing to ignore the rules.

For a given round, generally speaking, a judge will be choosing between 10-9 and 9-10, or 10-8 and 10-9. Due to the rules, it is hard to imagine a scenario whereby a judge could choose between 10-8 and 8-10, or 10-8 and 9-10. Indeed, in all of our experiments, we found that each round was essentially a pick between two scores.

Fans are also much more likely to give big scores, i.e. 10-8 or 10-7. Recall that just 0.01% of rounds were scored 10-7. Even at $q =$ 0.90, a quite one-sided round, the judges have a zero probability of awarding 10-7, whilst for the fans, the probability is 0.071.

5. Case-studies

In this section, we will use the model to scrutinise judges' actual scores given within a round, overall scores, and overall decisions (that is, win, lose, or draw). Each of these case-studies will serve to highlight a particular use or feature of the model:

- In Section 5.1, we introduce the concept of a "significant prediction" to determine whether a judge's decision was valid based on the predicted posterior distributions of the probability for each outcome.
- In Section 5.2, we use our model to demonstrate that the judges' individual preferences can lead to different verdicts.
- Finally, in Section 5.3, we implement a "fair" model that removes the effect of the bias variables. We use this model to determine who should have won Jon Jones' and Dominick Reyes' infamous fight.

5.1. Zhang Weili defeats Joanna Jedrzejczyk (2020)

Regarded as one of the greatest and most competitive fights of all time, in 2020, the UFC's strawweight champion, Zhang Weili, defended her belt against number one contender and former champion, Joanna Jedrzejczyk. The bout was a back-and-forth affair, with each round being extremely close, but in the end, Zhang won via a split decision (48-47, 48-47, 47-48).

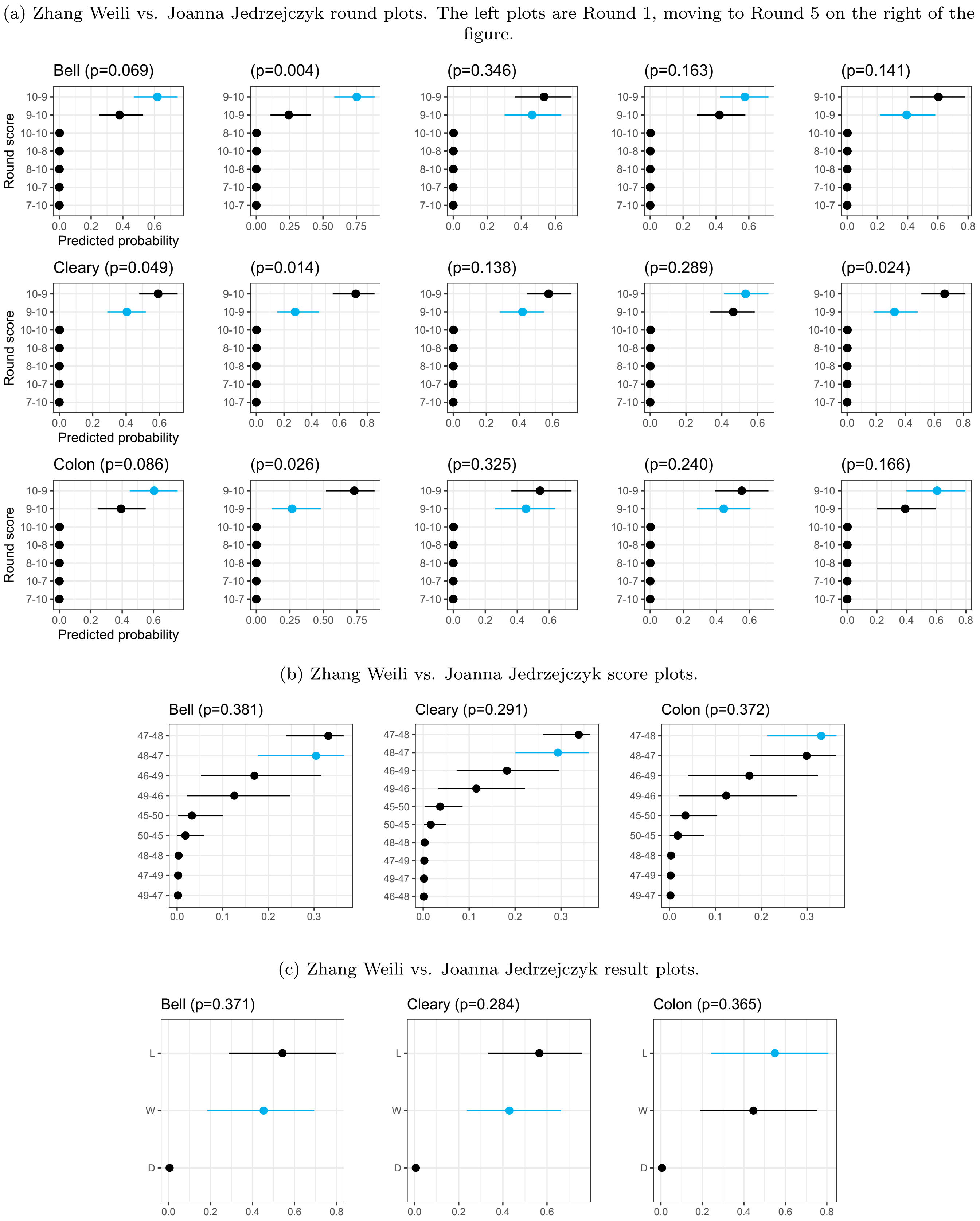
The fan scores highlight how close this fight was, as 48.6% gave the battle to Jedrzejczyk and 48.2% to Zhang (based on 1,291 scorecards[8]). The most common score was 47-48, which was given by 36.8% of fans; however, 48-47 was the verdict of 33.2%.

We will use this close fight to introduce the concept of "significant predictions". Testing the significance of a variable within a model is a staple of quantitative research, yet, to our knowledge, the concept has not been applied to predictions, despite obvious uses. In the context of MMA judging, we want to see whether a judge's decision was valid, even if it may not have been the most likely choice.

Fig. 2(a)–(c), display the posterior predictive probabilities for each score within the round, each score overall, and the overall re-

[7] For variables with a negative coefficient, we then multiply this value by −1.

[8] http://www.mmadecisions.com/decision/10984/Weili-Zhang-vs-Joanna-Jedrzejczyk.



(a) Zhang Weili vs. Joanna Jedrzejczyk round plots. The left plots are Round 1, moving to Round 5 on the right of the figure.

(b) Zhang Weili vs. Joanna Jedrzejczyk score plots.

(c) Zhang Weili vs. Joanna Jedrzejczyk result plots.

Fig. 2. Plots detailing the predictive posterior probabilities for the scores within each round, the overall scores, and the overall result, for the three judges in the bout between Zhang Weili and Joanna Jedrzejczyk. All scores are given from the perspective of Weili. Associated p-values of the predicted probabilities, introduced in Section 5.1 are also given. The chosen score is shown in blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

sult (win/draw/lose), for each of the three judges (Bell, Cleary and Colon). These plots are based on our model that accounts for their individual preferences. For brevity, we will refer to these plots as the *round*, *score*, and *result* plots, respectively. When calculating the overall score and result probabilities, we ensured the same posterior sample of the coefficients were used across the rounds. The chosen score is shown in blue.

Looking at Fig. 2(a), the distributions for each scoreline reveal much overlap in each round. Round three is a good example as the densities for Bell scoring 10-9 and 9-10 are nearly identical. Consequently, although we predict Bell has a higher probability of choosing 10-9, it is not controversial that he decided to score the round 9-10. To formalise this concept, we apply the concept of statistical significance.

Having estimated our Bayesian model through MCMC, we obtain N posterior samples for each coefficient and parameter of the model. Thus, we can obtain the seven probabilities corresponding to each score within a round for a particular sample.

Suppose we want to compare the probability of scoring the round as i to the probability of scoring it as j. For posterior sample s, denote the probability of scoring the round as i, as p_{si} (similarly, p_{sj} denotes the probability of scoring the round as j). Now, for $s = 1, \ldots, N$, we calculate the difference between these two probabilities, $p_{si} - p_{sj}$, and denote the distribution of these differences over all s, as $\mathbf{d}$.

In an extremely close round, $\mathbf{d}$ will be centred at 0, whilst in the most extreme case, it will be close to either -1 or 1. If the majority of $\mathbf{d}$'s mass is on one side of 0, then that suggests a significant difference exists between the two sets of probabilities. We then calculate the proportion of $\mathbf{d}$ on each side of 0, and find the minimum of these to be p. Small values of p that the sets of probabilities are significantly different.

We display a p-value in each plot. In cases where the judge's decision was different to the most likely predicted by the model, we calculate the p-value associated with the difference of these two sets of probabilities. If the judge submitted the most likely score, then we give the p-value comparing that score with the second most likely score.

We see from Fig. 2(c) that although the model predicts Zhang actually should have lost,[9] this result was not significant for any of the judges. Consequently, we can conclude that all of their final verdicts were within reason.

5.2. Edson Barboza defeats Danny Castillo (2013)

We use this fight to demonstrate how the individual preferences of the judges themselves may influence the final outcome.

In this bout, most fans believed the outcome was a 28-28 draw (61.8% of 152). The first round was dominated by Castillo, with the majority thinking it was an 8-10 (87.5%). The remaining two rounds were then clearly Barboza, with 73.7% and 96.7% giving him rounds two and three as 10-9, respectively.

Judge Derek Cleary scored the bout this way, arriving at the 28-28 consensus scorecard. However, Michael Bell and Wade Viera arrived at 29-28, having given 9-10 in the first round. Fig. 3(a)–(c) display the round, score, and result plots for this fight, respectively.

Looking at Fig. 3(a), we see the model would have predicted Bell and Cleary to score the first round as 8-10, but interestingly, Vierra would have most likely given it a 9-10. There are no further disagreements in rounds two or three. Consequently, the most likely scores for Bell and Cleary were 28-28, but 29-28 for Vierra. Correspondingly, the most likely result for Bell and Cleary was a draw, whilst Vierra was a win for Barboza.

Given that we predicted the judges to predict different outcomes, this demonstrates how their individual preferences may influence the result of a fight. We believe this is important for all participants of MMA–stakeholders, athletes, fans, and even judges themselves–to understand. Given the often high-stakes nature of bouts, it is common to see judges receive backlash for their decisions. Understanding these judges have their own opinions helps everyone appreciate the complexities of judging in what is still a relatively young sport, and avoid unnecessary bad publicity.

5.3. Jon Jones defeats Dominick Reyes (2020)

We now return to the infamous bout between Jon Jones and Dominick Reyes, to introduce the concept of the “fair-score” model, which stakeholders could use in various ways to compare or calibrate judges.

The first step in establishing the fair score is removing the bias variables' effect. Recall from Section 2, these are any variables which aren't in-round statistics.

To remove the effect of the bias terms, we could fit a new model which uses only the in-round statistics as independent variables. However, we have established several significant effects from the bias variables. Consequently, removing them entirely would introduce inherent “omitted-variable” bias. Instead, we use the original model and set any bias variable to 0 when making the fair predictions. This has the desired effect of removing their effects, whilst not introducing biases.

In Section 4.2, we demonstrated that judges have individual preferences towards each action, and in Section 5.2, we showed how these preferences might determine who wins a fight. Consequently, in the fair model, we aim to remove these individual preferences, to establish an average score. With that in mind, we use the model's latent population effects, represented by μ, rather than the judge effects, β_j.

Fig. 4(a)–(c) display the round, score, and result plots. We include the posterior predicted probabilities associated with the fair model. For comparison to the fair model, we remove the effect of the bias terms from the judges' probabilities, but keep their individual preferences.

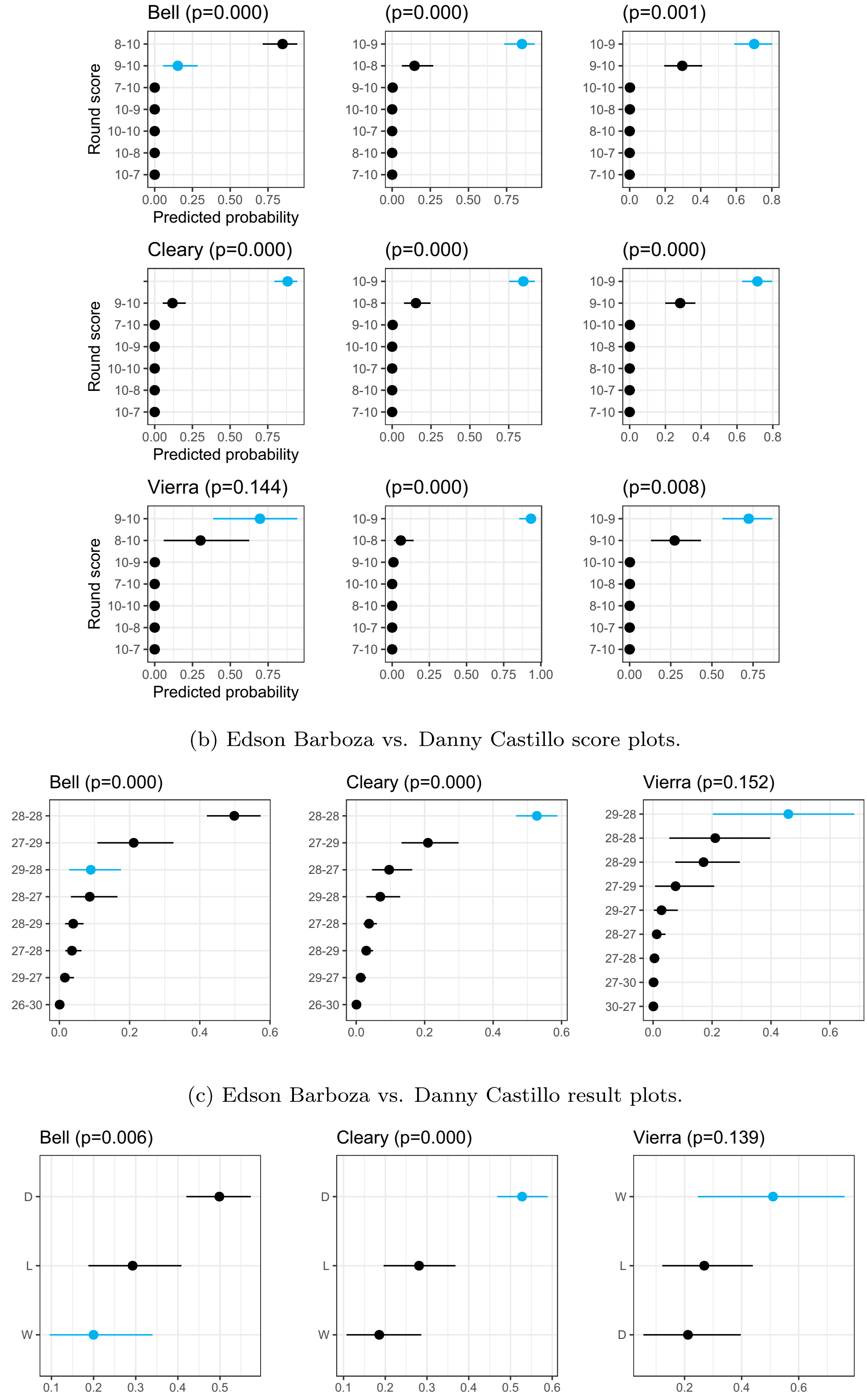
As was the consensus fan opinion, according to the fair model, Reyes should have won the first three rounds, and Jones the last two. The most likely score of the fair model in all of these rounds was significantly different from the next most likely score, with $p = .000$ in each.

From Fig. 4(b), we see that the fair model predicts the consensus score, 47-48, the score of 69.8% of fans. The 48-47 by Rosales and 49-46 by Soliz were significantly different from what we would have predicted they would score the fight. Lee's 48-47 was also significantly different, but to a lesser extent.

Finally, looking at the result plots in Fig. 4(c), the fair model would have predicted that Jones lost the fight. The model also predicts that each judge should have given the fight to Reyes overall. However, it is interesting that the decisions of Lee and Soliz are not significantly different from the predicted result. From the model's predictions of their behaviours (after removing bias terms), we see that either fighter winning would have been a just decision. The same cannot be said for Rosales, whom we would predict to side with Reyes.

[9] An interesting point to make is that although in the round plots we would have predicted Zhang to win–having been favoured in three rounds–we wouldn't predict her to win overall. This is because the rounds that Jedrzejczyk won were won with a much higher probability.

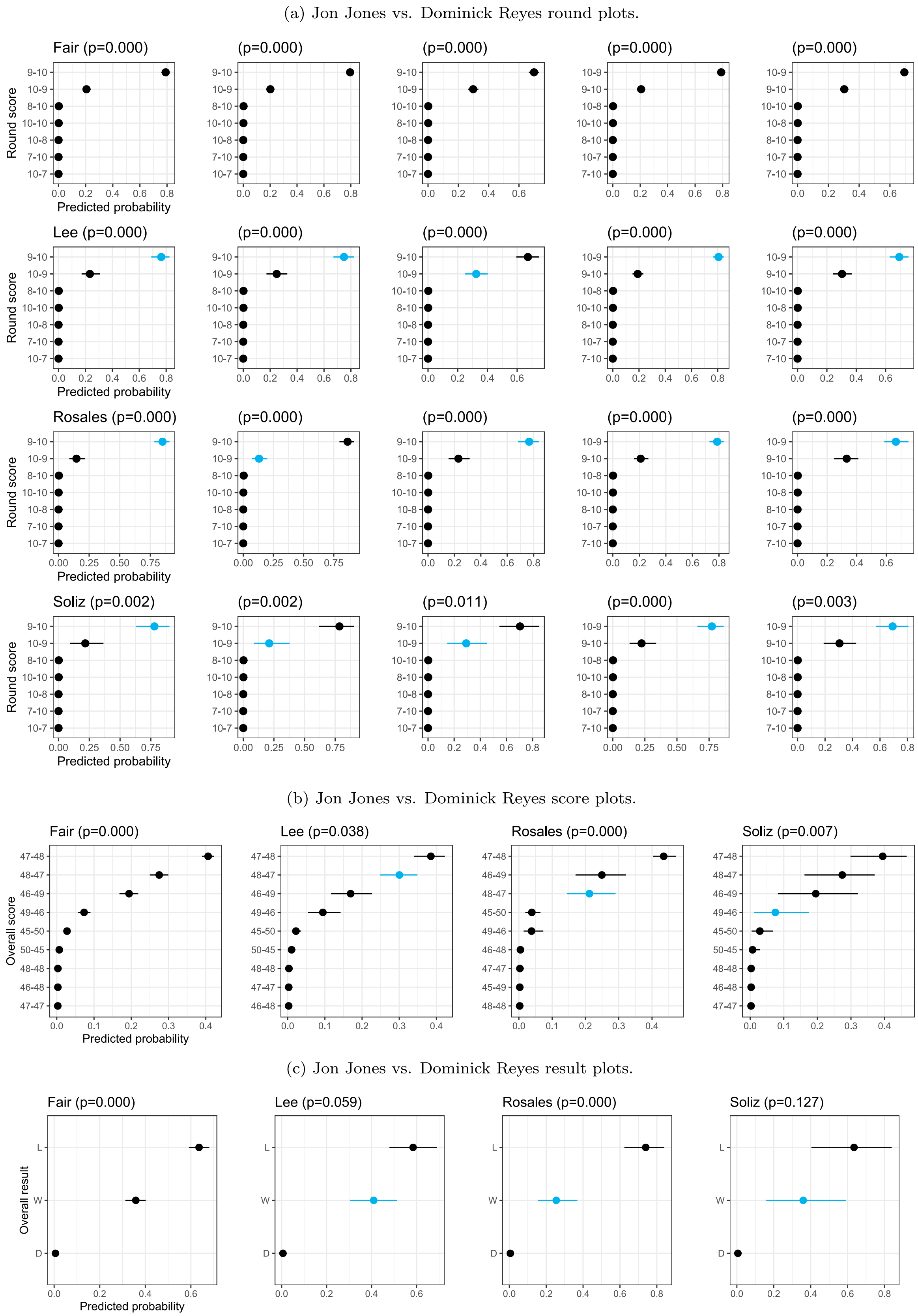
(a) Edson Barboza vs. Danny Castillo round plots. The left plots are Round 1, moving to Round 3 on the right of the figure.



(b) Edson Barboza vs. Danny Castillo score plots.

(c) Edson Barboza vs. Danny Castillo result plots.

Fig. 3. Plots detailing the predictive posterior probabilities for the scores within each round, the overall scores, and the overall result, for the three judges in the bout between Edson Barboza and Danny Castillo. All scores are given from the perspective of Barboza. Associated *p*-values of the predicted probabilities, introduced in Section 5.1 are also given. The chosen score is shown in blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(a) Jon Jones vs. Dominick Reyes round plots.

(b) Jon Jones vs. Dominick Reyes score plots.

(c) Jon Jones vs. Dominick Reyes result plots.

Fig. 4. Plots detailing the predictive posterior probabilities for the scores within each round, the overall scores, and the overall result, for the three judges in the bout between Jon Jones and Dominick Reyes. All scores are given from the perspective of Jones. Associated *p*-values of the predicted probabilities, introduced in Section 5.1 are also given. We include the predictions made by the *fair* model, which removes the effect of bias terms and the judges' individual preferences. For the judges, we keep their preferences, but remove the effect of the bias variables (that is, any variable that is not an in-round statistic). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

6. Conclusions

In this paper, we investigated whether individual preferences exist between judges of MMA contests. Whilst there has been some research into judging in sports, we believe this is the first research to directly explore the different opinions of individual judges and how these differences can influence the outcomes of competitions and contests.

Using a Bayesian hierarchical model, trained on a large set of MMA scores including several novel variables, we found various levels of disagreement between judges across the different in-round actions. The most notable was in scoring missed significant head strikes. We found that some judges deemed these as positive actions, whilst others believed they were negative. It is stated in the rules that fighters should not be rewarded for successful defensive manoeuvres. Consequently, we believe the judges who assess them as positive are correct. Using a real-life example, we demonstrated that these preferences can be the deciding factors in a fight, and may lead different judges to declare different winners.

Whilst these findings point towards different preferences amongst the many MMA judges, they perhaps also evidence rather vague and subjective judging criteria. Whilst some subjectivity is inherent in all live judging, it should not be the case that an action can be positive with one judge and negative with another.

We demonstrated the use of our models in potentially detecting erroneous decisions and establishing who should have won a fight, given the data available to us. Technology has recently been successfully implemented in football and tennis to assist the officials, namely VAR and Hawkeye. Whilst a mathematical model cannot entirely replace live judges (particularly until the level of available data is improved), we maintain there are several potential uses for our model: as a tool for training or calibrating the judges; detecting consistently problematic judges; gauging whether a fight was indeed controversial, and if so, how controversial; or demonstrating to judges that they may hold biases, as this might help to reduce them.

We introduced the concept of significant predictions in this paper. The judges' scores are particularly suited to this idea, as we want to see whether a given score is mathematically controversial or within reason.

A similar model was estimated to explain the scores submitted by fans on mmadecisions. Using this model, we could examine whether the fans and judges agree on the values of each action. Considering the recency of the sport's mainstream popularity, we were pleased to find that the fans weigh each in-round action comparably to the judges. The biggest difference is in the thresholds for giving each score: the public is much more likely to submit ties and big scores. An interesting finding was that the fans appear to be less influenced by bias variables, such as home-crowd influence and the official rankings.

Whilst investigating the fans' scores is interesting in its own right, there are real-world applications. The Professional Fighting Championship recently partnered with Verdict (who, like mmadecisions, allow fans to submit scores) so that in certain fights, the fans' scores are used as the official result. Our findings suggest that this potentially controversial approach may be a valid solution. However, further research should investigate the presence of other biases within the fan scores, for instance, biases towards more popular athletes, or disadvantages when fighting in a different timezone (as your fans may not be awake to submit scores).

Investigating the preferences and biases of judges in non-sporting environments, for instance, those in legal cases, would be a compelling avenue for further research. Unfortunately, a corresponding data set would be hard, if not impossible, to come by. MMA contests result in counts of various actions that can be used to model the judges' opinions, making it a favourable environment for research on individual preferences.

Acknowledgements

The authors would like to acknowledge the support of this work through the EPSRC and ESRC Centre for Doctoral Training on Quantification and Management of Risk and Uncertainty in Complex Systems and Environments Grant no. (EP/L015927/1)

Appendix A. Introduction to MMA and the Literature on Judging

Rounds are scored according to the criteria set out in the Unified Rules of Mixed Martial Arts (California State Athletic Commission, 2017), which were originally set in 2001 in an effort to protect fighters, whilst also legitimising MMA as a sporting spectacle. A judge must assess three criteria in sequence, if and only if they deem the preceding criteria to be exactly even.

1. The first criterion to be assessed, *efffective striking and grappling*, consists of legal blows or grappling techniques that have an immediate or cumulative impact on the potential to end the bout, where the immediate impact weighs more heavily. No scoring is given for defensive actions; the benefit of good defence is to stay in the fight and keep it competitive. This criterion should be sufficient to score the majority of rounds.
2. If the preceding criteria are deemed to be exactly even, the judge must assess whether the athletes are aggressively attempting to finish the fight, known as *effective aggression*.
3. The final criteria, to be assessed only if both the preceding criteria are exactly even, is *fighting area control*. This is evidenced by who controls the pace, place, and position of the match.

The judges follow the "10-point must" system used in boxing, whereby the winner of the round receives 10 points, whilst their opponent receives nine or less. Ties (10-10) are allowed but are actively discouraged in the rules. The rules state a 10-10 is not to be used "as an excuse by a judge that cannot assess the differences in the round". Any discernible difference (perhaps just one extra effective strike) should remove the 10-10 option within a round.

By far, the most common scoreline is a 10-9. This score can reflect an extremely close round or a round of marginal domination.

Winning a round by a large margin will result in a 10-8. Judges should always give a 10-8 when one fighter "has dominated the action of the round, had duration of the domination and also impacted their opponent with either effective strikes or effective grappling manoeuvres that have diminished the abilities of their opponent". A 10-8 should be considered when a fighter shows either dominance (e.g. when the opponent continually has to defend) or impact (e.g. by visibly hurting their opponent, causing a lack of control or ability).

Finally, a 10-7 score can be given when a fighter completely overwhelms their opponent, so the judge considers the fight should be stopped. An athlete must display immense dominance and impact to warrant this score.

To decide the winner of the fight, each judge's round scores are summed, with the winner according to each judge being the fighter with the most points. A majority verdict across the three judges is then used to identify the winner. The fight has several possible outcomes, as detailed in Table 4.

There are numerous fouls which can lead to a point deduction. Examples include headbutts, eye gouging, groin attacks, and hair pulling. The referee calls fouls, and deducting a point is entirely at the discretion of the referee, and this deduction to all of the judges' scorecards within the round.

Table 4
Different decisions which can be given based on the verdicts of the individual judges. The fight is between two fighters: Blue and Red.

Judges' overall winner			Blue	Red	Draw	Result	Decision
Blue	Blue	Blue	3	0	0	Blue	Unanimous win
Blue	Blue	Draw	2	0	1	Blue	Majority win
Blue	Red	Red	1	2	0	Red	Split win
Blue	Red	Draw	1	1	1	Draw	Split draw
Draw	Draw	Red	1	0	2	Draw	Majority draw
Draw	Draw	Draw	0	0	3	Draw	Unanimous draw

Appendix B. Non-centered parameterisation

Often, Stan can struggle to efficiently sample from the full state-space, particularly when estimating hierarchical models. A well-documented example is Neal's funnel (Neal, 2003), where the scale of the density changes over the state-space. Consequently, the optimal step-size changes as you move around the density.

In the so-called "centred" parameterisation, one may wish to model

$$\beta \sim \mathcal{N}(\mu, \sigma),$$
$$\mu \sim \mathcal{N}(0, 2.5),$$
$$\sigma \sim \text{Half-Normal}(0, 2.5).$$

Depending on the amount of data available, there will be a high correlation in the posterior between β, μ, and σ, thus leading to similar problems as Neal's funnel.

We can remove the dependencies between the parameters and hyper-parameters by parameterising β as a deterministic transformation of μ and σ. To do this, we introduce an offset term $\alpha \sim \mathcal{N}(0, 1)$, such that

$$\beta \sim \mu + \sigma\alpha.$$

The remainder of the model is as defined originally.

This line of thinking extends to a multivariate prior on β, for instance

$$\beta \sim \mathcal{N}(\mu, \Sigma),$$

where μ is a vector of mean values and Σ a covariance matrix. In this case, $\alpha \sim \mathcal{N}(0, 1)$ is a vector of independent identically distributed standard normal variables, such that

$$\beta \sim \mu + L\alpha,$$

where $LL^T = \Sigma$ is the Cholesky decomposition of Σ.

The Stan user guide recommends modelling the covariance as a correlation matrix multiplied from both sides by a diagonal matrix of standard deviations (Stan Development Team, 2021b, ch. 21.7). Suppose our covariance matrix is Σ, the correlation matrix is Ω, and the vector *sigma* denotes the standard deviations. Then, if $LL^T = \Omega$, the Cholesky factor of Σ is equal to $\text{Diag}(\sigma)L\text{Diag}(\sigma)$. Thus, if α is as defined before

$$\beta \sim \mu + \text{Diag}(\sigma) \cdot (L \cdot \alpha).$$

Appendix C. Proportional odds logistic regression

Proportional odds logistic regression for a response Y with levels y_j, $j = 1, 2, \dots, k$ is given by Greenland (1994)

$$P(Y \le y_j | \boldsymbol{x}) = \frac{\alpha_j - \exp(\boldsymbol{x\beta})}{1 + \exp(\alpha_j - \boldsymbol{x\beta})},$$

where $\boldsymbol{x}$ is a vector of covariates, $\boldsymbol{\beta}$ is a vector of unknown parameters, and the unknown intercepts α_j are such that $\alpha_1 \le \alpha_2 \le \ldots \le \alpha_k$.

Appendix D. Fan model

Here, we explicitly state the fan-score model used in Section 4.1.

$$\begin{aligned}
y_n &\sim \text{Ordered-Logit}(\lambda_n, t)\\
\lambda_n &= \beta x_n\\
t &= (-s_1 - s_2 - s_3, -s_1 - s_2, -s_1, s_1, s_1 + s_2, s_1 + s_2 + s_3)\\
\beta &\sim \mathcal{N}_K(\mu, \Sigma)\\
\Sigma &= \text{Diag}(\tau)\Omega\text{Diag}(\tau)\\
\mu_k &\sim \mathcal{N}(0, 5)\\
\tau_k &\sim \text{Half-Normal}(0, 2)\\
\Omega &\sim \text{LKJ}(2)\\
s_{1,2,3} &\sim \text{Half-Normal}(0, 5).
\end{aligned}$$

For observation i and score s, we then include the number of fans who scored the round as s as the observation weight.

References

Barnard, J., McCulloch, R., & Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica, 10*(4), 1281–1311. http://www.jstor.org/stable/24306780

Brown, A., & Reade, J. J. (2019). The wisdom of amateur crowds: Evidence from an online community of sports tipsters. *European Journal of Operational Research, 272*(3), 1073–1081. https://doi.org/10.1016/j.ejor.2018.07.015.

California State Athletic Commission (2017). The unified rules of mixed martial arts. Accessed: 27/04/2021, https://www.dca.ca.gov/csac/forms_pubs/publications/unified_rules_2017.pdf.

Collier, T., Johnson, A., & Ruggiero, J. (2012). Aggression in mixed martial arts: An analysis of the likelihood of winning a decision. *Violence and Aggression in Sporting Contests: Economics, History and Policy*, 97–109. https://doi.org/10.1007/978-1-4419-6630-8_7.

Feldman, T. (2020). The way of the fight: An analysis of MMA judging. *Journal of Applied Sport Management, 12*, 51–63. https://doi.org/10.7290/jasm120205.

Frederiksen, J. S., & Machol, R. E. (1988). Reduction of paradoxes in subjectively judged competitions. *European Journal of Operational Research, 35*(1), 16–29. https://doi.org/10.1016/0377-2217(88)90375-X.

Gelman, A., & Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models. *Analytical methods for social research*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790942.

Gift, P. (2018). Performance evaluation and favoritism: Evidence from mixed martial arts. *Journal of Sports Economics, 19*(8), 1147–1173. https://doi.org/10.1177/1527002517702422.

Greenland, S. (1994). Alternative models for ordinal logistic regression. *Statistics in Medicine, 13*(16), 1665–1677. https://doi.org/10.1002/sim.4780131607.

Heiniger, S., & Mercier, H. (2021). Judging the judges: Evaluating the accuracy and national bias of international gymnastics judges. *Journal of Quantitative Analysis in Sports, 17*(4), 289–305. https://doi.org/10.1515/jqas-2019-0113.

Lee, H. K. H., Cork, D. L., & Algranati, D. J. (2002). Did lennox lewis beat evander holyfield?: Methods for analysing small sample interrater agreement problems. *Journal of the Royal Statistical Society. Series D (The Statistician), 51*(2), 129–146. http://www.jstor.org/stable/3650314

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis, 100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory, 37*(1), 145–151. https://doi.org/10.1109/18.61115.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics, 31*(3), 705–767. https://doi.org/10.1214/aos/1056562461.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical ComputingVienna, Austria. https://www.R-project.org/.

Stan Development Team (2021a). Stan modeling language users guide and reference manual. https://mc-stan.org.

Stan Development Team (2021b). Stan user's guide. https://mc-stan.org/docs/stan-users-guide/index.html.