A Bayesian Ranking Analysis

Jackson Curtis

November 2018

1 Introduction

When one thinks of board games, the average person probably imagines the classic games they played as children like Monopoly and Sorry. To someone who grew up in a family that took their family game nights a little more serious, a subset of games known as "Eurogaming" like Settlers of Catan and Carcassonne might come to mind. However, there is another level of gaming where gaming goes from something you do on a Sunday afternoon to a major hobby. Gamers at this level often enjoy games with complexity and time requirements far surpassing those typically enjoyed by casual fans. These hobbyists find, rate, and share these more appealing, complex games on online forums such as boardgamegeeks.com.

Twilight Imperium is one of these games. Twilight Imperium can be thought of as a intergalactic version of Risk, but with far more complex strategies to achieve victory than simple world domination. When players begin Twilight Imperium they decide which of 17 factions to use. Each faction has special rules and abilities. Game makers go to great lengths to ensure "balanced" games. An ideal game is one where different factions provide a myriad of strategies to choose from, but no strategy regularly dominates the others. This paper introduces a statistical model for Twilight Imperium (which could be easily adapted to games of similar construction). Our model will be able to answer three questions of interest: (1) What are the relative abilities of the players playing the game? (2) Is the game well-balanced (meaning choosing one team over another does not provide an unfair advantage)? (3). To the extent that it is not well-balanced, which of the 17 factions is preferable?

2 Data

To answer these questions we will analyze game data collected from my siblings. One game can take up to five hours to play, so the entire dataset is nine games worth of data. This will make the Bayesian framework important, as we have very few data points to estimate our parameters of interest (some factions have appeared in as little as two games). However, by taking advantage of hierarchical structures and being careful with specifying our prior knowledge, we can get reasonable results from our data.

For each game our data consists of the players who participated, the factions they used, and their recorded scores. Each game is a race to ten points, so the winner receives ten points and everyone else will score between zero and nine. As a subjective assessment, we decided that how well someone did was not well represented by their original score (e.g. a person who scored nine points did not do 9/10ths as well as someone who won). Specifically, we wanted to more strongly reward those who finished higher, so, as a preprocessing step, we will compute finals scores as original scores plus the number of people they beat. For example, if player A won in a five player

game, his score would be 14 (his points + the number of people he beat). This had the effect of better spreading out those who did well vs. those who did poorly. As a final processing step, we divide your score by the total points awarded in the game, so that all scores in a single game sum to one. This can then be thought of as your "percentage of victory," which will be helpful in making games of different sizes comparable.

3 The Model

Our estimands of interest are each player's strength (ie ability to win games), each factions strength, and a parameter controlling the relative importance of those strengths. We can think of each faction/player combination in a game as a team, where the team's total strength will be the faction strength plus the player strength. We want to model how well each team did relative to the others, in a way that apportions the games outcome between the relative strengths. This suggests a Dirichlet distribution as our likelihood, where the random variables being modeled are the proportions of the game won by each team, and the parameters are the player and faction strengths. We can write our model as:

$$p(\underline{x_i}|\underline{\theta_p}, \underline{\theta_f}) \overset{ind}{\sim} \text{Dirichlet}(\underline{\theta_{pi}} + \underline{\theta_{fi}}) \text{ for i = 1, 2,...ngames} \tag{1}$$

$$p(\theta_{pi}|\lambda) \overset{iid}{\sim} \text{Exp}(\frac{1}{3\lambda}) \text{ for i = 1, 2,...nplayers}$$

$$p(\theta_{fi}|\lambda) \overset{iid}{\sim} \text{Exp}(\frac{1}{3(1-\lambda)}) \text{ for i = 1, 2,...nfactions}$$

$$p(\lambda) \sim \text{Beta}(4, 2)$$

In the above equations, $\underline{\theta_{pi}}$ is a vector of players strengths corresponding to the players who participated in game i, ordered identically to the scores recorded in $\underline{x_i}$. The priors on the faction strengths and player strengths are dependent on λ, a random variable with [0,1] support which characterizes the relative importance between player and faction. It can be seen that as λ approaches one, the prior on the player strengths will approach the exponential distribution with a mean of three, but the faction strengths will approach an exponential with a mean of zero. Likewise, as λ approaches zero, the players strengths will have mean zero and the faction strengths will have mean three. Because the parameters are only interpretable by comparison, the choice to use three reflects a prior belief in how variable strengths are in the population, but has little influence on the rank ordering of the players.

We set an informative prior on λ based on our belief that the game creators would be at least somewhat successful in balancing the teams (otherwise the game would not have received so much acclaim). Our prior assumes that, on average, the player will have twice as large an effect on the outcome as his faction, but the prior is still quite diffuse to allow uncertainty. We will tweak the prior specification on λ as a robustness check.

4 Posterior Computation

The posterior of our parameters can be written as follows:

$$p(\underline{\theta_p}, \underline{\theta_f}, \lambda, |\text{data}) \propto p(\underline{\theta_p}|\lambda) * p(\underline{\theta_f}|\lambda) * p(\lambda) * \prod_{i=1}^{ngames} p(\underline{x_i}|\underline{\theta_p}, \underline{\theta_f}) \tag{2}$$

This joint posterior clearly does not have a known form, so we will solve for the complete conditional distributions in order to perform MCMC to get posterior estimates. The notation can be challenging, but will be explained below:

$$p(\theta_{pi}|\underline{x_i}, \underline{\theta_f}, \underline{\theta_{p(-i)}}, \lambda) \propto \exp(\frac{-1}{3\lambda}) \prod_{j=1}^{\text{i's games}} \frac{1}{B(\underline{\theta_{pj}} + \underline{\theta_{fj}})} x_i^{\theta_{pj}} \tag{3}$$

In the above equation, the product is over all games that player i participated in. x_i is the proportion of the game won by player i. The beta function is defined as:

$$B(\underline{\alpha}) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)} \tag{4}$$

We can further simplify the conditional by dropping all terms in the product of $\Gamma(\alpha_i)$ that don't use θ_i. The conditional for the faction parameters are almost the same, with the swapping of the prior contribution:

$$p(\theta_{fi}|\underline{x_i}, \underline{\theta_p}, \underline{\theta_{f(-i)}}, \lambda) \propto \exp(\frac{-1}{3 * (1-\lambda)}) \prod_{j=1}^{\text{i's games}} \frac{1}{B(\underline{\theta_{pj}} + \underline{\theta_{fj}})} x_i^{\theta_{fj}} \tag{5}$$

Finally, the complete conditional for λ is:

$$p(\lambda|\underline{x_i}, \underline{\theta_p}, \underline{\theta_f}) \propto \lambda^{3-n_p}(1-\lambda)^{1-n_f} \exp(\frac{-1}{3\lambda} \sum_{i=1}^{n_p} \theta_{pi} - \frac{1}{3(1-\lambda)} \sum_{i=1}^{n_f} \theta_{fi}) \tag{6}$$

4.1 MCMC

We used the above equations to run the Metropolis-Hastings algorithm. We used univariate proposals and saved a new iteration after all parameters had accepted or rejcted one proposal. We ran four chains, for 50,000 iterations each, and then discarded the first 5,000 of each chain as burn in. Table 1 reports the effective sample size and R-hat diagnostics for each parameter.

	n-Effective	$\hat{R}$		n-Effective	$\hat{R}$
Doug	14015	1.001	Letnev	5749	1.003
Jared	8167	1.000	Mentak	7421	1.003
Kyle	4500	1.001	Muaat	5820	1.002
Landen	6491	1.001	Naalu	6376	1.003
Phil	7406	1.001	Nekro	5650	1.002
Sam	5071	1.001	Norr	7471	1.001
Tyler	5236	1.001	Saar	4610	1.001
Tyrel	4780	1.001	Sol	5998	1.002
Arborec	5798	1.001	Winnu	5093	1.002
Ghosts	5838	1.002	Xxcha	7556	1.002
Hacan	4792	1.002	Yin	6935	1.001
Jol-nar	7085	1.002	Yssaril	4068	1.003
L1z1x	4362	1.002	λ	1594	1.004

Table 1: Effective sample size and $\hat{R}$ statistics for each of our 26 parameters.

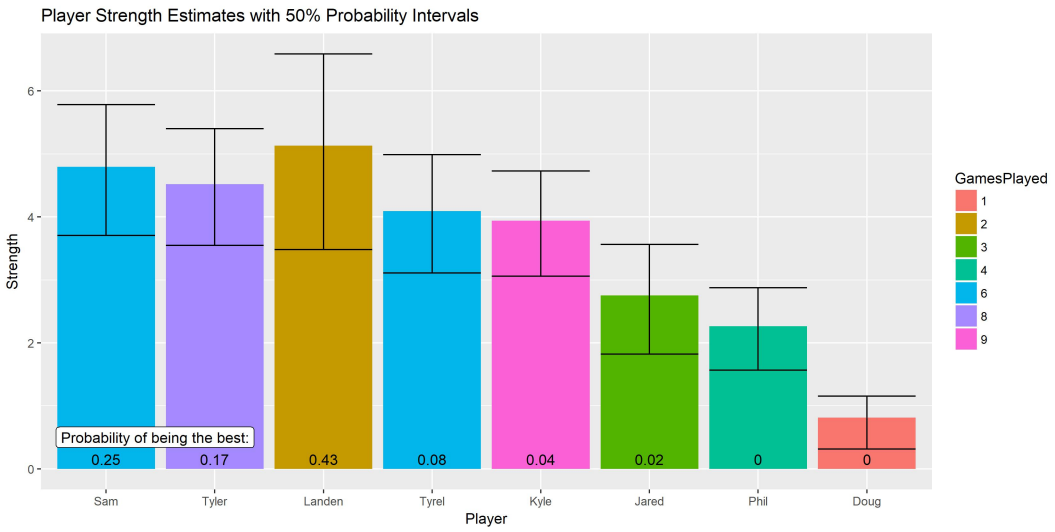

Figure 1: Mean strength estimates for the eight players with 50% probability intervals, ordered by the lower bound of the probability interval

Examining the table, it is clear that λ is the parameter with the most auto-correlation between samples, which makes sense given that it is a hyperparameter effected by all other parameters in the model. Although 180,000 iterations results in only a effective sample size of 1,594, because we are interested primarily in the central tendency of λ we are satisfied that we can make good inferences from our sample size.

5 Results

The major results are summarized in Plots 1 and 2. The images include 50% probability intervals. This may seem out of the ordinary, but our main inference is not about the absolute magnitude of the parameters. Instead, we only are interested in interpreting these relatively and probabilistically. For example, if two players played five games and the first beat the second every time, there might still be high, overlapping uncertainty about the absolute magnitude of the parameters, but there would also be high confidence that the first is better than the second. Table 2 was created to answer the question "What is the probability that player A is stronger than player B?" While we clearly don't have enough data to ascertain with certainty that one player is better than another, we can use our model to make probabilistic statements. Additionally, along the bottom of Figure 1 is the probability that each player is the overall strongest of the entire group. We see that although Landen has only played two games, his dominance in those games gives him a 43% probability of being the best overall.

Plots 2 and 3 lend credence to our original hypothesis that the game is well-balanced. The faction strengths are much more tightly grouped and close to zero than the player strengths. From the posterior distribution of λ we see that the data has shifted the prior distribution towards higher values of λ, suggesting that faction strengths are even less consequential than our original hypothesis. The posterior of λ has a mean of 0.770 and a 95% credible interval of (0.500, 0.950).

		B						
		Jared	Kyle	Landen	Phil	Sam	Tyler	Tyrel
	Doug	0.07	0.01	0.02	0.09	0.00	0.00	0.01
	Jared	-	0.17	0.15	0.64	0.10	0.10	0.18
	Kyle	-	-	0.29	0.93	0.24	0.30	0.45
A	Landen	-	-	-	0.91	0.55	0.60	0.68
	Phil	-	-	-	-	0.04	0.03	0.08
	Sam	-	-	-	-	-	0.58	0.71
	Tyler	-	-	-	-	-	-	0.63

Table 2: The probability that Player A is better than Player B based on comparison of random draws from the joint posterior.

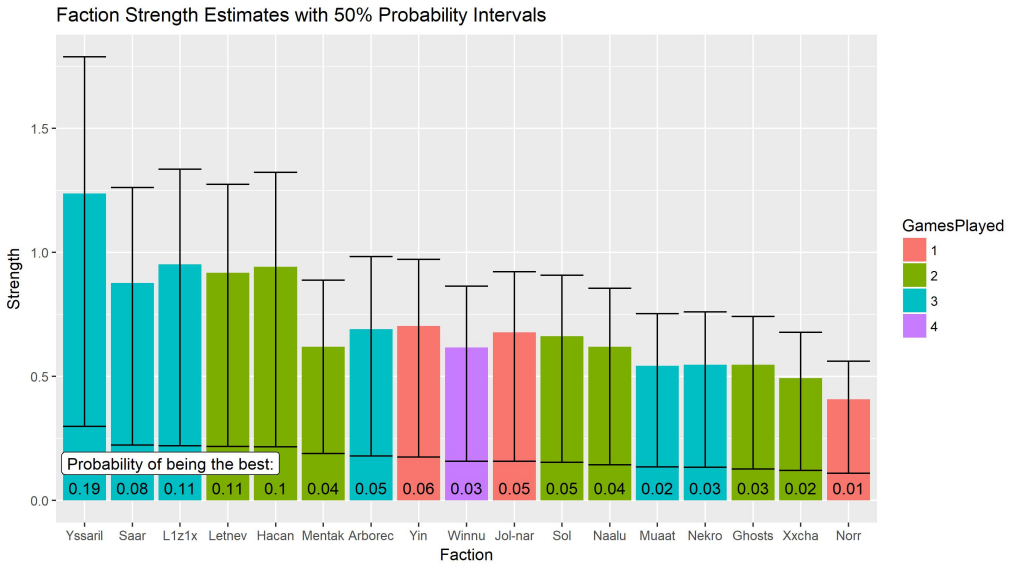

Figure 2: Mean strength estimates and intervals for the 17 factions. Note the y-axis scale difference between this and Figure 1

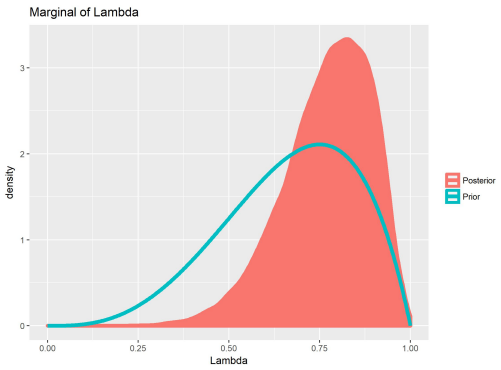

Figure 3: Posterior distribution of λ

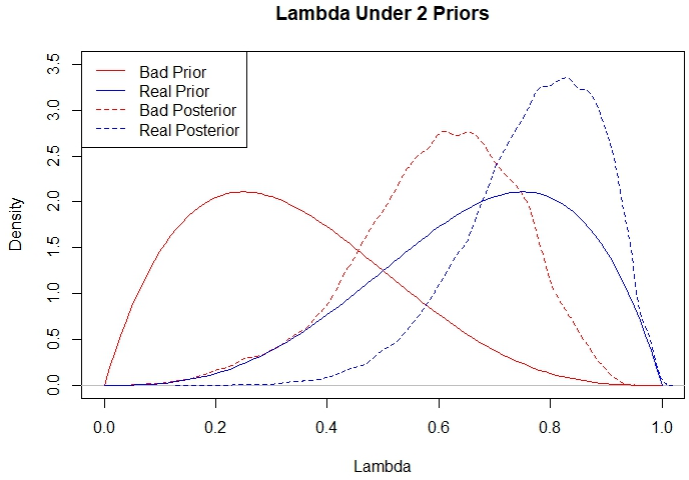

Figure 4: Posterior distribution of λ for new and original prior

6 Alternative Priors

One major concern of this model is identifiability. For example, if each player always used the exact same faction, and no two players doubled up on a faction, our model would not be able to learn about λ at all and all inferences would be based on our prior. The opposite scenario would be a perfect crossing where each player used every faction. Because the games were played without any designed structure, our data falls somewhere in the middle of the two extremes. One way in which we can see if identifiability is an issue is by repeating the analysis with a different prior on λ. We did this by altering the prior to be Beta(2,4), which expresses the opposite of our original belief, namely that faction effects are twice as important as player effects. Figure 4 shows our posterior with the new prior, and gives us good evidence that our prior is not having undue influence on the posterior. Despite the prior belief of strong faction effects, the data has clearly indicated that the faction effect is less than our stated belief. While our priors are not uninformative, the data is clearly helpful in updating our belief about λ.

Another way in which we explored robustness of our model to decisions about the prior is by changing the mean of the exponential distribution of strengths. We changed the player mean from 3λ to 10λ (likewise with the factions). This had almost no influence on our inference about λ and did not change the rank ordering of any of the players. However, the comparisons between players became more exaggerated, so the best players had a higher probability of being the best and the worse players had a higher probability of being the worst. This is reflective of the fact that the standard deviation and the mean of the exponential is the same, so as we increase the mean we are increasing the expected difference in scores. Therefore, there is less overlap in the credible intervals.

One model checking procedure we performed produced an interesting insight into the way in which we set up the model. This insight hints at limitations of the model and possible future improvements. The model checking procedure was to simulate the win rate of players in our dataset competing against hypothetical players that could be encountered in the future.

Table 3 clearly indicates a problem. Although we have no reason to think the players in our dataset are abnormal, all but one would be expected to win well over 50% of their games against a random opponent. This can be explained by the prior specification of an exponential distribution of strengths. The mode of the exponential distribution is 0, so no matter what the mean is we

	Win Rate
Doug	0.32
Jared	0.61
Kyle	0.72
Landen	0.78
Phil	0.56
Sam	0.78
Tyler	0.76
Tyrel	0.72

Table 3: Win rates in a 1 vs. 1 game simulation where both teams used an identical faction.

expect many scores to cluster around 0. However, the observed data does not seem to agree with that prior belief, so our parameters conditional on the data are not exponentially distributed. We experimented fitting a Gamma(6λ, 2) to the prior player strengths, which has the benefit of a non-zero mode but the same expectation. This however, resulted in similar problems in the opposite direction (players losing most simulated games). The root of the problem was that we are not modeling the variation in strengths between players and factions. While we could do that by putting a prior on the gamma rate, we were hesitant to introduce more parameters with so little data. Additionally, as we assessed the situation we recognized that predicting future performance was not a stated goal, and that our stated goals were well answered by the model we have, with one caveat: the exponential distribution seems to have undue weight in the predictions of the players who have played only one or two games. One player, Doug, seems irrationally punished for his one bad performance, and Landen seems irrationally rewarded for his two good performances. Both these results are tempered using the gamma model.

7 Frequentist Approach

Frequentist approaches are going to struggle without the flexibility provided by the Bayesian framework in a small data, hierarichical setting. A simple way to get frequentist estimates of the parameters is to use the Dirichlet likelihood, and find the parameters which maximize the likelihood. We do this using optim in R and present the results in Figure 5. Assessing uncertainty in this estimate is difficult. A typical bootstrapping approach would need to be performed on game-level resampling (because the games are conditionally independent). However, resampling only nine games virtually guarantees that some of the faction or player parameters will be left out of the resulting likelihood, and thus not estimable by maximum likelihood.

Another issue is that without the hierarchical model, answering the question about relative importance of strengths is difficult. One way we might approach this problem is to compare the average faction strength to the average player strength. However, this leads to the surprising conclusion that factions are more important in determining outcomes than players ($\overline{\theta}_p = 8.5, \overline{\theta}_f = 9.4$). We suspect this is a result of overfitting, Because we are estimating 17 faction effects and only 8 player effects, they naturally have a greater ability to fit the data. We can compensate for this by creating a shrinkage constraint. If we penalize the coefficients for how variable they are, player strengths become more important. Specifically, our penalty was 5*(var(θ_{ps}) + var(θ_{fs})) subtracted from the log likelihood. This resulted in player strengths that were twice the magnitude of faction strengths, but are unsatisfactory in answering the question because varying the coefficient of the penalty arbitrarily changes the relative importance.

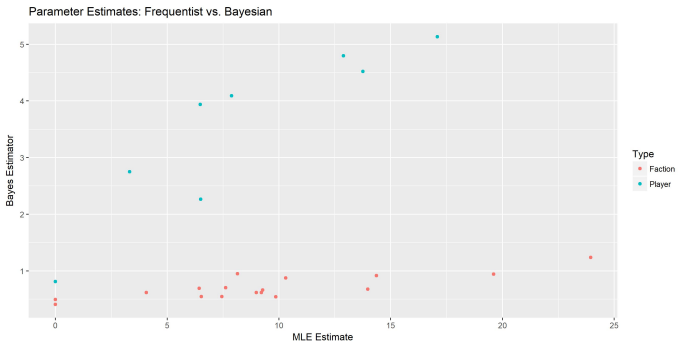

Figure 5: Unconstrained MLEs for parameters, plotted against the Bayes Estimator.

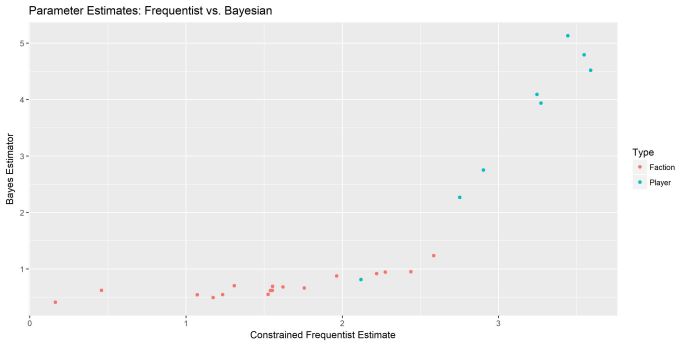

Figure 6: Constrained frequentist estimator, plotted against Bayes Estimator.

8 Conclusion and Future Work

This project has provided a rational way to estimate skill from multiple sources. We have shown that there is little evidence that factions substantially affect the outcome of a game of Twilight Imperium. The analysis of our model suggests two weaknesses of our model: our prior specification results in estimates that we don't necessarily believe for people with very few data points (one or two), and our model is not designed to be predictive of hypothetical players or factions. However, we believe our rankings are fair and accurate with moderate sample sizes and appropriately account for the lack of independence among players' results.

Future work could take several forms. If more data was available, the model could be expanded to estimate variance among player strengths in the population. This is also possible for factions, but less important as there are no unobserved factions in the population. An additional concern that a more complex model could address is strengths that vary over time, as we expect most players to do worse in their first initial games. Finally, an equally rational approach to modeling would be to assume that factions have a multiplicative effect instead of an additive effect, so that a good player got a bigger advantage from a good faction than a bad player did. Comparing models fits from the additive and multiplicative models would be enlightening.