This is the Title of My Thesis

A thesis submitted to the
College of Graduate and Postdoctoral Studies
in partial pulfillment of the requirements
for the degree of Master of Science
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Author's Name

1

Permission to Use

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Disclaimer

Reference in this thesis to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the University of Saskatchewan. The views and opinions of the author expressed herein do not state or reflect those of the University of Saskatchewan, and shall not be used for advertising or product endorsement purposes.

Requests for permission to copy or to make other uses of materials in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building, 110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan S7N 5C9 Canada

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9 Canada

Abstract

This is the abstract of my thesis.

Acknowledgements

Acknowledgements go here. Typically you would at least thank your supervisor.

This is the thesis dedication (optional)

Contents

Permission to Use i

Abstract iii

Acknowledgements iv

Contents vi

List of Tables vii

List of Figures viii

List of Abbreviations ix

1 Introduction 1

2 Background 2
- 2.1 Home Advantage . . . 2
  - 2.1.1 Related Work . . . 2
  - 2.1.2 Contribution . . . 4
- 2.2 Bayesian Inference . . . 5
  - 2.2.1 Introduction . . . 5
  - 2.2.2 Multilevel Modeling . . . 6
  - 2.2.3 Markov Chain Monte Carlo . . . 8

3 Methods 9
- 3.1 Multilevel Model . . . 9
- 3.2 Negative Binomial . . . 12
- 3.3 Experiments . . . 12

4 Results 13
- 4.1 Experiments . . . 13
- 4.2 Home Advantage . . . 13

5 Conclusion 20

References 21

Appendix A Sample Appendix 21

Appendix B Another Sample Appendix 22

List of Tables

List of Figures

4.1 Distributions of the estimated home advantage for the NHL, NBA, MLB, and NFL for pre and post COVID adjusted seasons. Home advantage for playoffs are reported for NHL and NBA because that is when their COVID restricted games took place. Home advantage for regular season is reported for MLB and NFL as their respective playoff seasons are too small for stable results. Red distributions represent COVID-19 bubble adjusted seasons. 14
4.2 Distributions of the estimated home advantage for the NHL, NBA, MLB, and NFL over the past 5 seasons from 2016-2020. Home advantage for playoffs are reported for NHL and NBA because that is when their COVID restricted games took place. Home advantage for regular season is reported for MLB and NFL as their respective playoff seasons are too small for stable results. Red distributions represent COVID-19 bubble adjusted seasons. 16
4.3 Comparison of distribution of home points in the models and the observed data for each league. The Negative Binomial model noticeably provides a better overall fit across each league. . . . 18

List of Abbreviations

SCUBA	Self Contained Underwater Breathing Apparatus
LOF	List of Figures
LOT	List of Tables

1 Introduction

testing

2 Background

2.1 Home Advantage

2.1.1 Related Work

The initially most well known and cited work on home advantage in sports was done in 1977 by Schwartz and Barsky [**?**] who analyzed and found home advantage to exist in professional hockey, basketball, baseball and football. In [**?**] the authors accept home advantage as a real phenomena after reviewing the relevant literature and argue for a framework that focuses on game location, psychological states, behavioral states, and performance outcomes to try to understand the underlying causes of home advantage. Follow up work a decade later by Carron et al. [**?**] reviewed the literature and concluded that home advantage was still present in both amateur and professional sports, in both individual and team sports, across genders, and across time. More recent works [**?**] [**?**] confirm the continued existence of home advantage in the North American professional leagues we are considering in this study: the NHL, NBA, NFL, and MLB. In general, older studies on home advantage tend to use correlation methods of aggregated full season statistics (e.g. combining all teams home wins into one home win percentage to see if it is above 50%), whereas more recent studies more often build statistical regression models from game level data that adjust for additional factors, such as relative team strengths, and try to infer the effect of the home advantage parameter on the regression model.

There have been several studies analyzing home advantage in the context of COVID-19 adjusted seasons; however, nearly all of them have focused exclusively on European Soccer leagues. In [**?**] thirteen such works are summarized, of which only two used correlation methods and the other eleven made use of regression analysis to infer the change in home advantage. Benz and Lopez themselves use a bivariate Poisson regression model to infer home advantage, thus making for twelve of the fourteen studies making use of regression analysis. Ten of these studies found a drop in HA during the COVID-19 adjusted seasons, with the other four reporting mixed results where HA dropped in some leagues but not in others. We are only aware of one academic article looking at home advantage in the COVID-19 adjusted seasons for the NBA [**?**] where the authors found presence of home advantage prior to the NBA's bubble and argue for teams travel schedules having the most notable impact. As of this writing there are no academic papers examining home advantage during the COVID-19 adjusted seasons for the NHL, NFL, or MLB, although several online blog articles exist (cite?? or leave out?) most of which take a quick cursory glance at raw home win percentages and do not account for

team performance relative to strength of opponents as is done in the work summarized in [**?**]. This paper is a first look at using regression to infer home advantage through team performance while adjusting for quality of opponents instead of only looking at aggregated statistics such as win percentage.

There is a growing body of work in sports analytics that turns to building statistical models to measure relative team strengths while accurately predicting game outcomes. These works have their roots found in Bradley-Terry models [**?**] and Bayesian state-space models [**?**]. Further advancements and examples from the NHL, NBA, NFL, and MLB are comprehensively summarized in [**?**] and follow a form similar to the model in [**?**] as Bayesian methods generally offer more flexibility to be able to extend and customize these models and are generally more stable when fitting the models to data [**?**] while better capturing the uncertainty in estimating parameters opposed to classical point estimates and p-values which are increasingly under criticism in modern science. While most of this work was developed with a focus on predicting game outcomes and measuring team strengths, they often include a term to adjust for home advantage and as such can be re-purposed to be used to infer home advantage as is done in the majority of works summarized by [**?**]. In this paper we aim to take the first attempt to use these methods to infer home advantage during the COVID-19 adjsuted seasons of the NHL, NBA, NFL, and MLB.

In [**?**] the authors show the improved efficacy of the Poisson distribution instead of the more common Normal distribution [**?**] for modelling points scored by each team in each game. In [**?**] the authors follow the work in ntzoufras arguing for the use of a bivariate Poisson distribution that accounts for small correlation between two teams scoring and show its efficacy over ordinary least squares regression in inferring home advantage via simulations. However, as is shown in [**?**] there is no need of the bivariate Poisson when working within the Bayesian framework because hierarchical models of two conditionally independent Poisson variables mix the observable variables at the upper level which results in correlations already being taken into account. In [**?**] the authors argue for more complex methods to limit the shrinkage of their hierarchical model as their data was from leagues with a large range of team strengths. We follow [**?**] who showed that the "big four" North American Professional leagues are very close in team strength and thus do not reduce the shrinkage from our hierarchical model.

This next part may need to be taken out... The challenge with methods that look at correlations among raw statistics such as home win percentage is that they fail to account for other factors such as relative team strengths. For example, a weaker team may have poor home win percentage because they have a poor overall win percentage. That same team; however, may perform better at home than they do at other stadiums whilst still losing to stronger opponents and vice versa. This discrepancy can be further impacted by imbalanced schedules. In the professional leagues we consider, teams generally do not face each each opponent the same number of times and do not face the same strength of opponents at home and away in a perfectly balanced manner. While studies often recognize this discrepancy, they often claim that it is a small effect that can be ignored [**?**] without showing evidence. We argue that these issues and any debate over how much of an effect they have is most reliably mitigated by accounting for other factors, most notably

team strengths, when trying to infer home advantage. Regression analysis methods are most often used for precisely their ability to account for multiple factors when performing inference, and as such they are most appropriate for our focus of analyzing home advantage.

To aid in showing the benefit of using regression to adjust for relative team strengths when trying to infer home advantage, we present a simulation of game outcomes and the difference between raw home win percentage and our models inferred results. We use a similar data generating process to [**?**]. We generate team attacking and defensive strengths ($\alpha_t \sim N(0, 0.5)$ and $\delta_t \sim N(0, 0.5)$) for $t = 1, ..., 20$ teams. A game is then simulated by using model (cite model number) to randomly generate the number of points scored by each team with the home advantage parameter set to $\beta = 0.25$ and intercept $\mu = 0$. We simulate a regular season by simulating 4 games between each pair of teams, two home games each, resulting in 76 game appearances for each team for a total of 760 games, which most closely resembles the NHL and NBA regular season schedules. Then we simulate the top 16 teams playing a knock-out tournament bracket consisting of best 4 out of 7 matches, again closely following the NHL and NBA formats. We then compare using raw home win percentage to predict existence of home advantage compared to model (cite model). The results can be seen in figure (??).

There does exist work in baseball analytics analyzing the stability of home win percentages (cite, those articles found on your phone); however, they look at the stability over decades at a time. In our context we simply do not have decades worth of covid restricted professional games. Given the "small data" of our problem, we maintain that Bayesian inference is best suited for this task.

- high level explanation/transition to why bayesian methods are what I used

2.1.2 Contribution

We adopt a Bayesian framework to develop a Negative Binomial regression model that adjusts for relative team strengths while inferring home advantage. We choose this approach for two main reasons. First, alternative methods that rely on correlations among raw statistics, such as home win percentage, fail to account for other factors such as relative team strengths. Our regression approach can infer changes in team performance while adjusting for quality of opponents. Second, the Bayesian framework gives more interpretable results and more flexibility in model building than classical regression methods. The Bayesian framework results in distributions for the estimates of each parameter in our model. This allows us to analyze these distributions directly to determine the probability a parameter is greater (less) than a certain value or that it exists in a specific interval, avoiding the confusion that often arises interpreting p-values and confidence intervals.

By examining the resulting home advantage parameter estimates of our model from before and during the COVID-19 pandemic, we can draw conclusions about the existence of the home advantage phenomenon and provide new evidence for its potential causes. We hypothesize that home advantage is a real phenomenon, thus we expect its parameter estimate to drop during the COVID-19 seasons relative to before the COVID-19

seasons. We are also interested in examining if any differences in relative changes in home advantage exist across the leagues as some leagues had different COVID-19 restrictions which could affect home advantage differently. We also show that point totals in North American professional sports are prone to overdispersion, thus, the Negative Binomial distribution allows for better model fit than the more common Poisson and Normal distributions used in regression analyses.

2.2 Bayesian Inference

2.2.1 Introduction

- basic intro and explanation of bayes theorem - p(a—b) = p(b—a)p(a)/p(b) from p(a and b) ¡- probably should switch to theta's and x's - perhaps the breast cancer example

Bayes theorem is named such after revered thomas bayes. It was actually named posthumously after his friend found some of his work and had it published after his death. Thomas Bayes was trying to solve an inverse probability problem...

Bayes theorem can be stated simply as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.1}$$

This equation can be derived simply from basic rules of probability and conditional probability, namely that $P(A \cap B) = P(A|B)P(B)$ and equivalently $P(A \cap B) = P(B|A)P(A)$. One can then simply substitue and isolate $P(A|B)$ to arrive at equation 2.1.

In equation 2.1, each part of the equation is often referred to and interpreted differently. $P(B|A)$ is referred to as the *likelihood*, and is understood in the same way likelihood is understood in traditional frequentist statistics. $P(A)$ is referred to as the *prior*, and can be understood as the prior belief we have for the value of A. $P(B)$ is referred to as the *data*...

As if often the case in science, the name of a discovery is usually not the actually discoverer or first user. In the case of Bayes theorem, it is known that (a century earlier?) Piere Simone Laplace was using what we now refer to as Bayes theorem to solve an inverse probability problem of his own. In his work...

- bayes vs frequentist - mention thomas bayes and laplace - ra fischer and the 'smear' campaign

Despite its early roots, bayesian statistics took a back seat during the 20th century. Most prominent statisticians of the time did not like the "subjectivity" of specifying a prior that could potentially influence the results of inference. In particular, the most prominent statistician of the 20th centry, RA Fischer, was a vocal opponent of "subjectivity". Many other prominent statisticians such as ... also comdemned the use of bayesian statistics.

The other issue with bayesian statistics that prevented it from becoming a more mainstream technique for inference is its computational difficulties. Although Bayes theorem is rather simple to state, it often leads to requiring the computation of integrals that is either exceedingly difficult or outright impossible analytically.

In order to overcome this, a user is left with essentially two options to approximate the posterior distribution. The first is to change the prior and likelihood distributions into ones that are solvable analytically; this is known as choosing a conjugate prior. The issue with this method is that you are forced to use different prior and likelihood distributions than you original wanted and thus are no longer accurately representing the problem. The other method for approximation is to use a sampling procedure, such as a markov chain monte carlo (MCMC) procedure, to generate enough samples of the target posterior distribution to reasonably approximate it. The issue with this method is that MCMC sampling can be very slow taking a long time to converge and the results can be noisy and do not have any guarantees for the accuracy of the approximation.

- brief mention of why bayes is making a combeback now - real world success (mostly military) - advances in computing power - advances in mcmc theory (hmc)

Despite bayesian statistics being pushed to obscurity in the early-mid 20th century via a smear campaign and the difficulties in its computations, that latter part of the 20th century in to the 21st centruy is seeing a rise in the popularity and use of bayesian statistics through its real world successes, advances in computing power, and advances in MCMC theory.

- need real world examples from 'the theory that wouldn't die' book - need some hard stats on advances in computing power and sampling - need to cite beatancourt's work on HMC and future developments

2.2.2 Multilevel Modeling

MOTIVATION - a generalization of regression methods with many use cases such as prediction, data reduction, and causal inference from experiments and observational studies. - called multilevel or hierarchical for two reasons. 1) the structure of the data (e.g. students clustered within schools within districts within states within countries...). 2) the model itself has its own hierarchy with parameters of the within-groups regression at the bottom, controlled by the hyperparameters of the upper-level model. - multilevel modeling can be viewed as a trade-off between two extremes: complete-pooling and no-pooling. Complete-pooling is when an overall average is used and variations among groups/categories within the overall data are ignored, thus the data are completely pooled. No-pooling is when separate models/averages for each individual group/category are used and any correlations or dependencies among the groups are ignored, thus the data is not pooled at all. In this view, multilevel models are seen as partially-pooled where their estimates can be thought of in a simplified way as a weighted average of the complete-pooling and no-pooling extremes. For groups/categories with fewer data points the multilevel model weighs the complete-pooling more, and for groups/categories with more data points the model weighs the no-pooling estimates more. This results in what is commonly referred to as shrinkage whereby partially pooled estimates are essentially the no-pooling estimates that have been shrunk toward to complete-pooling estimate, or shrunk toward to the mean. The amount of shrinkage depends on the samples-sizes, the variation within groups, and the varition between groups. - outperforms classical regression in predictive accuracy since multilevel modeling includes least squares regression as a simple case. Generally considered essential for prediction, useful for data reduction, and helpful for causal

inference. - multilevel modeling can be viewed as a white-box method whereby each part of the model can be fully interpreted, understood, and customized. This makes it ideal for inference. Furthermore, multilevel models are Bayesian graphs which means that Judea Pearls causal calculus (or do-calculus) can be used to infer causality. This makes multilevel models useful beyond predictions alone. In contrast, this is something that many machine learning methods such as neural networks and decision trees, can not do. - multilevel modeling allows separating estimates of predictive effects of an individual predictor and its group-level mean (usually referred to as direct and contextual effects of the predictor). - several motivations for their use: - learning about treatment effects that vary: how does y change when some x is varied (with all other inputs held constant)? Often its not the overall effect of x but rather how this effect varies in the population or among groups/categories of interest. - using all the data to perform inferences for groups with small sample size: at one extreme classical estimation can be useless if the sample size is small in a group/category, and at the other a classical regression ignoring group-level variation can be misleading as well. Multilevel modeling (partial pooling) compromises between the overall noisy within-group estimates (no-pooling) and the over-simplified regression estimate that ignores group indicators (complete-pooling). - predictions: as discussed above, the partial-pooling or shrinkage effect of multilevel modeling is a form of regularization that protects from underfitting (complete-pooling) and overfitting (no-pooling) to produce more accurate predictions on unseen data (the holy grail of machine learning). - analysis of structured data and more efficient inference for regression parameters: many datasets have an inherent multilevel/hierarchical structure (e.g. students within schools, patients within hospitals, laboratory assays on plates, elections in districts within states, or data from cluster sampling etc.). Even simple cross-sectional data can be placed in a larger multilevel context. For example, many datasets initially thought to be big data often become small data once you being sub-dividing them into more and more groups/categories. For example, opinion polls trying to predict who you vote for based on age, race, income, location, interests etc. Each split leaves you with smaller and smaller groupings that have the potential for better model fit(more predictors) at the risk of overfitting (small samples within groups/categories). - including predictors at two different levels: You can specify models that have individual level predictors and group level predictors. For example, in estimating radon levels in houses you could have measurements at the individual level (individual houses, indicator if the sensor is in the basement, etc.) and then predictors at the group level (county-level uranium readings) and using both together provides better model fit than separating them. Multilevel modeling avoids problems in classical regression such as colinearity when trying to include group-level indicators as well as group-level predictors in the same model. - getting the right standard error (accurately accounting for uncertainty in prediction and estimation): To get an accurate measure of predictive uncertainty, one must account for correlation of the outcome between groups/categories/predictors (e.g. forecasting state-by-state outcomes in US election, one must account for correlation of outcome between states in a given year). Sometimes the uncertainty in estimation is of interest rather than the estimate itself. Sometimes predictions require multilevel modeling, such as when making predictions for a new group. For example, consider a model of test scores for students

within schools. You could model school-level variability in classical regression (or another machine learning model such as decision trees or neural nets) with an indicator for each school. But it is impossible in this framework to make a prediction for a new student in a new school, because there is no indicator in the model for this new school. This type of problem is handled seamlessly when using the multilevel framework.

- worked out example leading to weighted average of normal dist estimates - conjugate priors - transition into mcmc

2.2.3 Markov Chain Monte Carlo

- basic mcmc - advances in compute - hamiltonian mc

3 Methods

- intro and abstract similar to paper

- from paper plus links to earlier discussed motivations for multi-level modeling

3.1 Multilevel Model

- similar to paper - may need to build this up so as to include all data experiments or a more general version to refer back to

We infer home advantage by fitting a regression model to predict the points scored in each game while adjusting for relative team strengths and home advantage. We adjust for relative team strengths by modelling both an offensive rating and a defensive rating for each team. We argue this better represents real differences between teams and allows the model to better infer if a team performs better or worse when playing at home by measuring its performance relative to its average offensive performance versus its opponents average defensive performance. This section describes in detail the parameters of the model, their interpretation, and how we fit the model.

We aimed to build a parsimonious model to infer home advantage for each league while adjusting for relative team strengths and accounting for uncertainty in the data and parameter estimates. We needed a method that was robust to smaller sample sizes because we only had one COVID-19 adjusted season for each league to compare to and because this sample becomes smaller as you include more parameters which splits the data into smaller groups. We also wanted to be able to quantify the uncertainty in our parameter estimates. To address these concerns we adopt a Bayesian multi-level regression model framework building upon previous work [**?**] [**?**] [**?**] [**?**] that allows for pooling results across all teams to infer home advantage. The partial-pooling of multi-level regression modelling allows us to separate the effects of individual teams offensive and defensive strengths from their group level means and helps prevent overfitting by adjusting parameter estimates through a process commonly referred to as "shrinkage to the mean" [**?**] [**?**] [**?**]. We argue the pooling of data across each teams results to better handle smaller sample sizes while preventing overfitting, and the ability to quantify the uncertainty in parameter estimates makes Bayesian multi-level regression an ideal choice for this task.

We model the response variable of the number of points scored by each team in each game as Negative Binomial:

where $y_{ij} = [y_{i1}, y_{i0}]$ is the vector of observed points scored in game i by the home ($j = 1$) and away

$(j = 0)$ teams and $\mu_{ij} = [\mu_{i1}, \mu_{i0}]$ are the goal expectations of the home and away teams in game i. The α parameter allows for the flexibility of fitting to overdispersed data where the variance is much greater than the mean. In our experiments we have found that defining α as a fraction of μ_{ij} led to better sampling and model fit. Thus, we define $\alpha_{ij} = \mu_{ij} * \lambda$ and then sample λ when fitting the model. We model the logarithm of goal expectation as a linear combination of explanatory variables:

$$\begin{aligned} \log(\mu_{i1}) &= \gamma_{sp} + \beta_{sp} + \omega_{sh[i]} + \delta_{sa[i]} \\ \log(\mu_{i0}) &= \gamma_{sp} + \omega_{sa[i]} + \delta_{sh[i]} \end{aligned} \tag{3.1}$$

where γ_{sp} is the intercept term for expected log points in season, with $s = [0, 1, 2, 3, 4]$ corresponding to the 2016, 2017, 2018, 2019, and 2020 seasons respectively. The subscript p indicates regular season $(p = 0)$ or playoffs $(p = 1)$. For the results in Figure 4.1, all previous seasons are combined $(s = 0)$ and compared to the COVID-19 adjusted season $(s = 1)$. Home advantage is represented by β_{sp} with s and p the same as the intercept. The offensive and defensive strength of the two teams are represented by ω and δ. The nested indexes $h[i]$ and $a[i]$ identify the teams playing at home and away respectively and we use this nested notation to emphasize the multi-level nature of these parameters as they are modelled as exchangeable from a common distribution [**?**] [**?**] [**?**]. This enables pooling of information across games played by all teams in a league and results in mixing of the observable variables (y_{ij}) at this higher level which accounts for correlation in home and away points scored in each game [**?**].

In this model formulation we are estimating different home advantage parameters for the regular season and playoffs as well as for each individual season. The primary motivation for this is because the NHL and NBA COVID-19 bubbles essentially only occurred during their playoffs and we therefore want to separate home advantage during the playoffs for a more direct comparison. Modelling in this way also addresses potential questions of whether home advantage changes each year or remains constant. Our results in Figure 4.1 are from estimating one home advantage parameter prior to COVID-19 and one afterwards. We then show the results of modelling home advantage separately for each season and show the results in Figure 4.2 which reveal some interesting differences as discussed in the Results section.

In (3.1) we see that the home team's goal expectation is a linear combination of the home team's offensive strength and the away team's defensive strength as well as a constant home advantage. Conversely, the away team's goal expectation is a linear combination of the away team's offensive strength and the home team's defensive strength with the home advantage parameter noticeably missing. There is no index for league because, although we use the same model consistently across each league, we fit a separate version for each league.

This model formulation results in the intercept representing the logarithm of the overall average of points scored with $exp(\beta_{sp}), exp(\omega_{sh[i]})$, and $exp(\delta_{sa[i]})$ representing multiplicative increases or decreases to the average points scored to determine the expected points scored for an individual game. This can be seen by

considering:

$$\begin{aligned}
\log(\mu_{i1}) &= \gamma_{sp} + \beta_{sp} + \omega_{sh[i]} + \delta_{sa[i]} \\
\mu_{i1} &= \exp(\gamma_{sp} + \beta_{sp} + \omega_{sh[i]} + \delta_{sa[i]}) \\
\mu_{i1} &= \exp(\gamma_{sp}) * \exp(\beta_{sp}) * \exp(\omega_{sh[i]}) * \exp(\delta_{sa[i]})
\end{aligned} \tag{3.2}$$

For example, a home advantage parameter of $\beta = 0.25$ would result in multiplying the average points scored by $\exp(0.25) \approx 1.28$, which can be interpreted as an increase of about 28% in expected points scored by the home team in a game between teams with relative offensive and defensive strengths $\omega_{sh[i]}$ and $\delta_{sa[i]}$ respectively.

Model Fit in PyMC3

The models are fit using PyMC3, an open source probabilistic programming language that allows us to fit Bayesian models with their implementation of a gradient based Hamiltonian Monte Carlo (HMC) No U-Turn Sampler (NUTS) [**?**]. As in other previous work [**?**] [**?**], we use Bayesian modelling and fitting approaches to allow us to incorporate some prior baseline knowledge of parameters as well as better quantifying uncertainty in the interpretation of parameter estimates.

The Bayesian approach means we need to specify suitable prior distributions for all random parameters in the model. The prior distributions for parameters in our model are:

$$\begin{aligned}
\gamma_{sp} &\sim \mathcal{N}(\theta^*, \sigma^{2*}) \\
\beta_{sp} &\sim \mathcal{N}(0, 1) \\
\lambda &\sim \text{Uniform}(0, 1000) \\
\omega_s &\sim \mathcal{N}(0, \sigma_{s\omega}) \\
\delta_s &\sim \mathcal{N}(0, \sigma_{s\delta}) \\
\sigma_{s\omega} &\sim \text{HalfNormal}(1) \\
\sigma_{s\delta} &\sim \text{HalfNormal}(1)
\end{aligned} \tag{3.3}$$

where θ^* is the logarithm of the average points scored, and σ^{2*} is the logarithm of the variance of points scored, over the regular seasons and playoffs of the league being modelled. We note that we found γ_{sp} fits close to θ^* even when using a weakly informative prior, but we keep this formulation as it maintains the spirit of using prior information in Bayesian analysis. We allow λ to potentially be large for instances where there is no overdispersion in the outcome variable because a large λ results in a large α_{ij} which makes the Negative Binomial distribution become similar to a Poisson distribution.

The model is fit using PyMC3's NUTS sampler using 4 chains of 2000 iterations with 1000 tune steps for a result of 8,000 samples from 12,000 total draws. It is standard practice to check convergence with the $\hat{R}$

statistic from [**?**] [**?**]. Each model fit produced $\hat{R}$ statistics of 1.00 with no divergences [**?**].

3.2 Negative Binomial

- tables and figures like paper for justification

3.3 Experiments

- synthetic data and then real data showing overfitting and the value of shrinkage - maybe the synthetic ha simulations - the experiments from the paper

4 Results

4.1 Experiments

4.2 Home Advantage

- similar to paper

The distributions for the estimates of the home advantage parameters from pooling the previous four pre-COVID-19 seasons/playoffs together can be seen in Figure 4.1 with the COVID-19 restricted season/playoffs coloured red. The peaks of these distributions represent the most likely values for the home advantage parameter and their width represents the uncertainty in these estimates. We can use these distributions to directly measure the probability the home advantage parameter is less than the previous seasons. The leftward shift of the distribution for the COVID-19 restricted season/playoffs suggests that home advantage decreased in the NHL, NBA, and NFL while not changing for the MLB.

Figure 4.2 shows results from estimating home advantage individually for each prior season. This more granular view of pre-COVID-19 home advantage reveals greater season-to-season variation in home advantage that is missing in Figure 4.1. Nevertheless, the year-over-year estimates in Figure 4.2 show the results of reduced home advantage in COVID-19 restricted season/playoffs holding for the NHL, NFL, and NBA, albeit with a single past season with lower home advantage in both the NFL and NBA. The remainder of this section examines these estimated distributions and their implications.

For the NHL and NBA data, Figures 4.1 and 4.2 and our analysis focus on their playoff seasons because the NHL and NBA COVID-19 seasons only took place during their playoff seasons. In contrast, the MLB and NFL had COVID-19 restrictions for their entire seasons, therefore, Figures 4.1 and 4.2, and our analysis for those leagues are focused on their regular season games. Focusing on the MLB and NFL regular seasons is not only convenient but arguably necessary as their playoff seasons consist of much fewer games than the NHL and NBA playoff seasons, resulting in high uncertainty of parameter estimates. The NHL and NBA regular season results as well as the MLB and NFL playoff results are provided in the supplementary materials.

The home advantage parameter, β, represents a multiplier of $\exp(\beta)$ applied to expected points. For example, an estimated home advantage parameter for the NBA of 0.05 represents a $\exp(0.05) \approx 1.0513$ multiplier on expected points or an increase in expected points of 5%. With average points scored in the NBA being around 107 this would translate to approximately a 5-point home advantage on average in the NBA playoffs. We provide a full description and interpretation of the model in the Methods section.

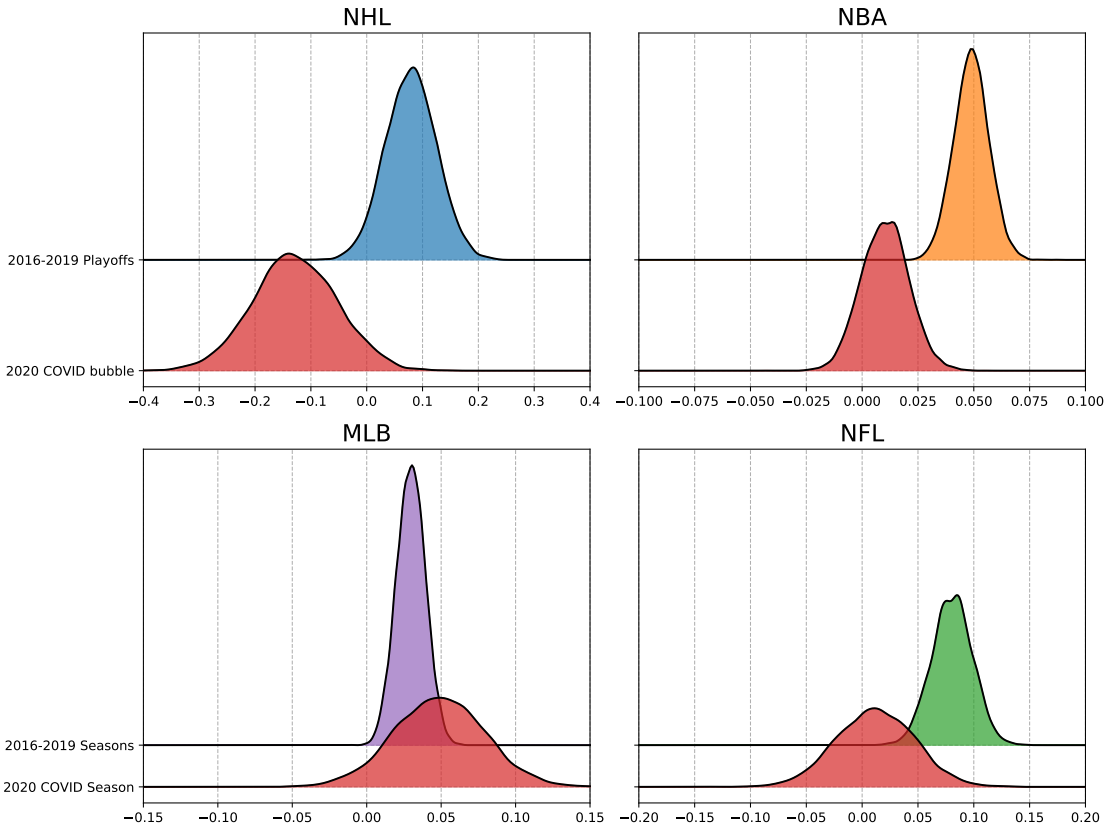

Figure 4.1: Distributions of the estimated home advantage for the NHL, NBA, MLB, and NFL for pre and post COVID adjusted seasons. Home advantage for playoffs are reported for NHL and NBA because that is when their COVID restricted games took place. Home advantage for regular season is reported for MLB and NFL as their respective playoff seasons are too small for stable results. Red distributions represent COVID-19 bubble adjusted seasons.

For the NHL data, the results in both Figures 4.1 and 4.2 show the home advantage parameter confidently above 0 for pre-COVID-19 seasons and confidently below 0 for the COVID-19 bubble. The probability the home advantage parameter (β) is less than 0 for the COVID-19 bubble is $\Pr(\beta < 0) = 0.95$. The probability the home advantage parameter is less than the previous playoff seasons mean of 0.081 is 0.998. These results give strong evidence that home advantage in the NHL was negatively impacted by the COVID-19 bubble.

For the NBA data, the pooled home advantage parameter estimate in Figure 4.1 is confidently above 0 and tightly around 0.05. For the COVID-19 affected playoffs, the probability the home advantage is less than 0 is only 0.17, but the probability that it is less than the pre-COVID-19 mean of 0.05 is 0.999, suggesting that home advantage in the NBA was negatively impacted by the COVID-19 bubble. However, when examining the year-to-year estimates of prior seasons in Figure 4.2 we see a decreasing trend in home advantage in the NBA playoffs with the estimate for the NBA playoffs in 2017 appearing as almost as much of an outlier as the COVID-19 estimate. This suggests the decreased home advantage in the COVID-19 could potentially be a random outlier. The uncertainty in these estimates means we can not make definitive conclusions in the absence of more data. We conclude that it is probable that home advantage in the NBA decreased in the COVID-19 bubble but not as definitively as the NHL results.

For the MLB data, the home advantage parameter is surprisingly likely to be slightly greater than it had been in previous seasons. The probability the home advantage parameter is less than the mean of the previous seasons is $\Pr(\beta < 0.036) = 0.26$. When comparing the COVID-19 estimate to the previous seasons in Figure 4.2 there appears to be no noteworthy difference. This gives evidence that home advantage in the MLB was unlikely to be negatively impacted by the COVID-19 restrictions and was likely unaffected by the restrictions.

For the NFL data, the pooled home advantage parameter estimate in Figure 4.1 is confidently above 0 with a mean of 0.078. For the COVID-19 affected season, the probability the home advantage is less than 0 is 0.388, but the probability that it is less than the pre-COVID-19 mean of 0.078 is 0.976, suggesting that home advantage in the NFL was negatively impacted by the COVID-19 restrictions. However, when examining the year-to-year estimates of prior seasons there is a clear pattern of home advantage decreasing in the NFL and even being lower in 2019 than it was in the 2020 COVID-19 adjusted season. We argue the results in Figure 4.2 are enough to overturn the results in Figure 4.1 and conclude that home advantage in the NFL was not impacted from its previous trend by the COVID-19 restrictions.

In summary, results for pooled (Figure 4.1) and individual (Figure 4.2) past seasons give strong evidence that home advantage in the NHL was negatively impacted during the COVID-19 restricted playoff season and that home advantage in the MLB was unaffected by the restrictions. Pooled past season results also suggest home advantage was negatively impacted by the COVID-19 restricted seasons for the NBA and NFL, however a closer examination of the individual past season results reveals a trend of decreasing home advantage over the past few seasons, which may partly account for the lower home advantage found during NBA and NFL COVID-19 restrictions.

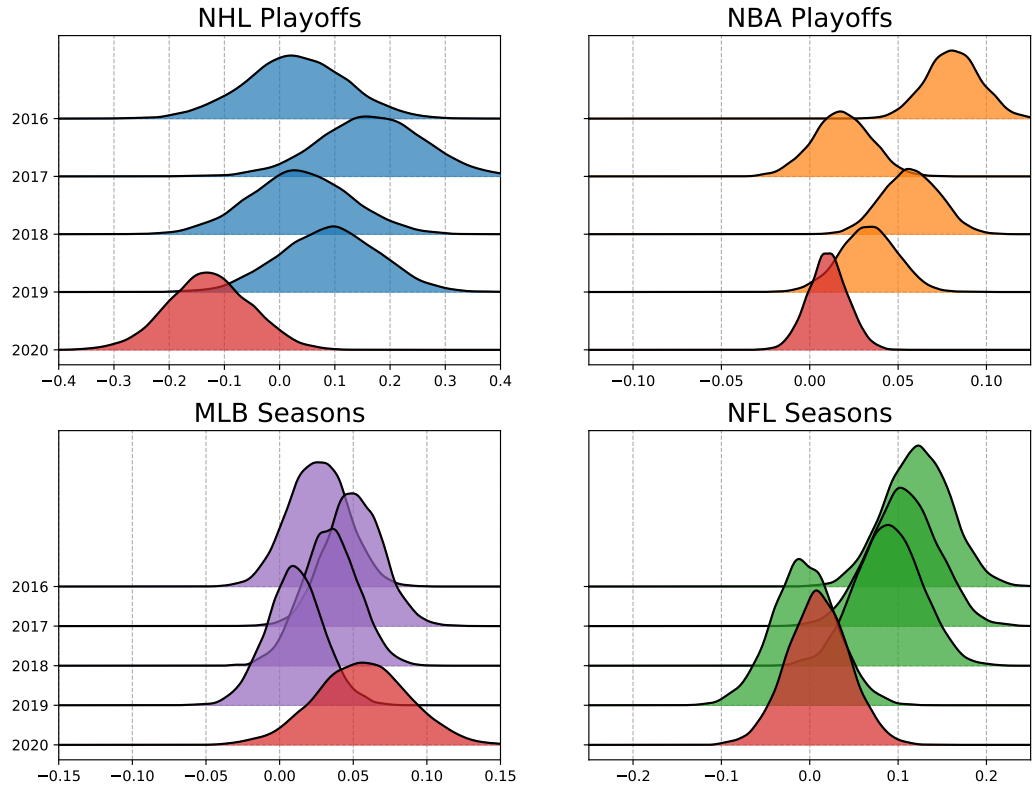

Figure 4.2: Distributions of the estimated home advantage for the NHL, NBA, MLB, and NFL over the past 5 seasons from 2016-2020. Home advantage for playoffs are reported for NHL and NBA because that is when their COVID restricted games took place. Home advantage for regular season is reported for MLB and NFL as their respective playoff seasons are too small for stable results. Red distributions represent COVID-19 bubble adjusted seasons.

Model Fit Comparisons

In this section we present the results of fitting our model with the Negative Binomial distribution as the likelihood for point totals, as compared to the more commonly used Normal and Poisson distributions.

Since point totals in sports are positive integers, the Poisson distribution is a natural choice for modelling their outcomes. The effectiveness of the Poisson distribution for modelling point totals has been shown in several works analyzing European football data [**?**] [**?**] [**?**]. One shortcoming of the Poisson distribution is that it only has one parameter and this leads to the strong assumption that the mean is equal to the variance. For low scoring sports like European football and hockey, this is usually a fine assumption. However, this is an invalid assumption for several of the sports we analyze in this paper. Table **??** reports the dispersion statistic σ_p. The dispersion statistic represents how much greater the variance is than the mean while adjusting for sample size and model complexity and is computed as $\chi^2/(n-p)$ for each league, where χ^2 is the Pearson chi-squared statistic of the point totals data, and $n-p$ are the degrees of freedom with n representing the sample size of the point totals data and p representing the number of predictors in our model. The commonly suggested threshold, $\sigma_p > T$, for determining when a Poisson model is no longer appropriate is around $1.2 < T < 2$ [**?**] [**?**]. Table **??** shows the NBA, MLB, and NFL having potential overdispersion in their point totals and thus, the Poisson distribution is likely inappropriate and less effective. We instead opt for using the Negative Binomial distribution because it has an extra parameter α that gives greater flexibility and better model fit to data that is overdispersed while still adequately fitting models without overdispersion.

To establish the efficacy of the Negative Binomial distribution in our model, we fit and compare models using the Poisson and Normal distributions across each league. We fit Poisson and Normal regression models by changing the likelihood of the model in (**??**) to $y_{ij}|\mu_{ij} \sim \text{Pois}(\mu_{ij})$ for the Poisson regression (and subsequently drop α from the rest of the model as it is not needed), and $y_{ij}|\mu_{ij}, \sigma^2 \sim \mathcal{N}(\mu_{ij}, \sigma^2)$ for the Normal regression (and use a weakly informative prior $\sigma^2 \sim \text{HalfNormal}(50)$). Otherwise the models are identical and their interpretation remains the same as is discussed in the Methods section.

We evaluate the models across each league by estimating the out-of-sample predictive fit via leave-one-out cross-validation (LOO). Following the work of Vehtari [**?**] we approximate LOO using Pareto-smoothed importance sampling (PSIS) and report the results in Table **??**. We note here that we also used the widely-applicable information criterion (WAIC) [**?**] but found the results to be nearly identical and the conclusions the same. Examining Table **??** we see that for the NHL and NBA, where there is little to no overdispersion, the Poisson and Negative Binomial models fit almost identically with the Negative Binomial model starting to show small improvement for the slightly overdispersed NBA data. As overdispersion increases for the MLB and NFL data we see the fit of the Negative Binomial model become noticeably better. The Negative Binomial model also outperforms the Normal model across each league except for the NFL where we see it fit only slightly worse while both models greatly outperform the Poisson model.

These differences in fit can be seen visually in Figure 4.3 where we plot the distribution of observed home point totals in black along with 2000 sampled model fits in green for Poisson, blue for Negative Binomial,

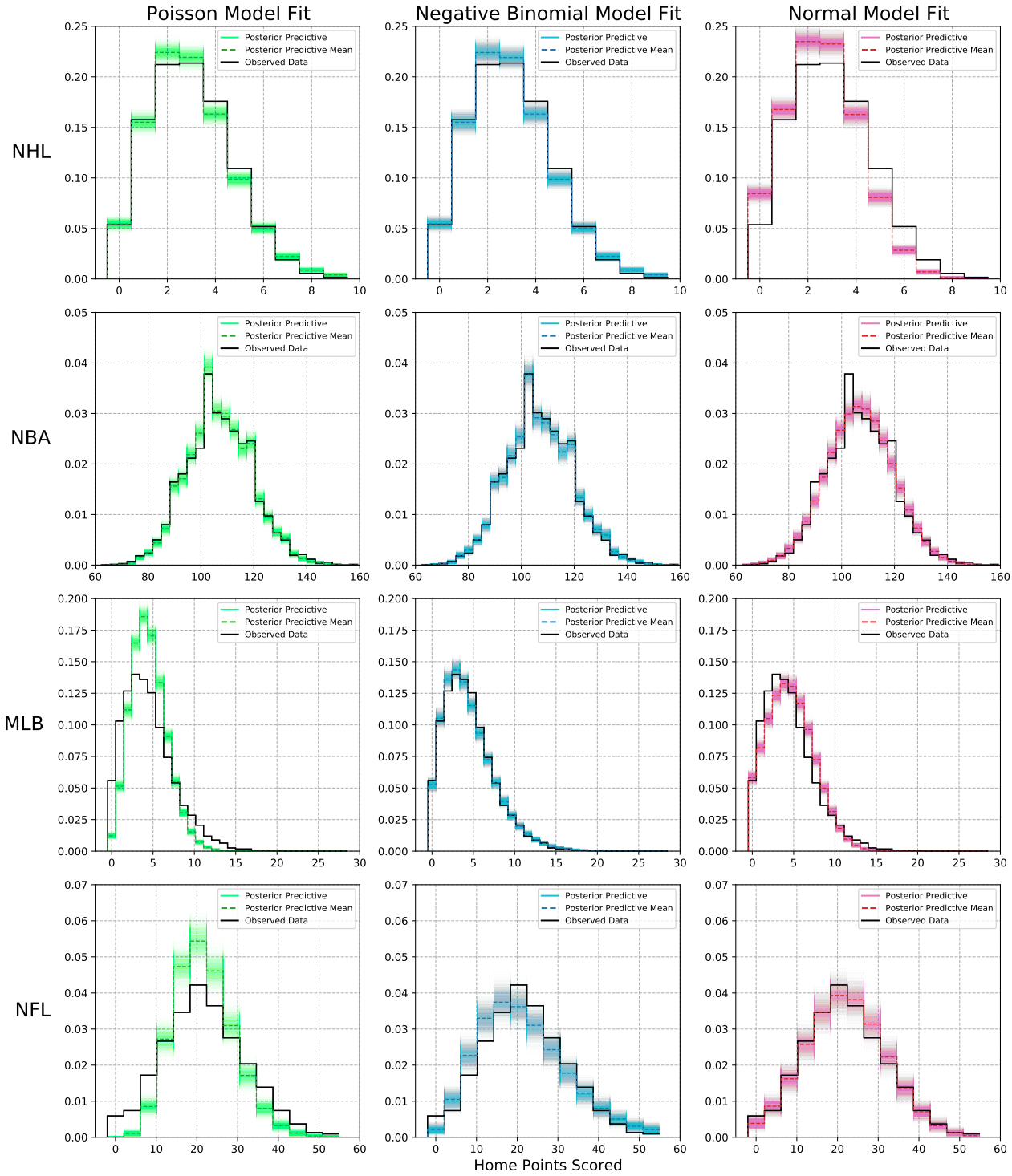

Figure 4.3: Comparison of distribution of home points in the models and the observed data for each league. The Negative Binomial model noticeably provides a better overall fit across each league.

and red for Normal; with the respective mean model fits as dashed lines. The differences between the Poisson and Negative Binomial models becomes increasingly apparent for the leagues with greater overdispersion, while the Normal model comparatively struggles for each league except the NFL where both the Normal and Negative Binomial greatly outperform the Poisson model. Because the point totals of the sports we are considering are positive integers prone to overdispersion and based on the results in Table **??** and Figure 4.3, we conclude that the Negative Binomial distribution is the most appropriate for regression modelling professional hockey, basketball, baseball, and American football.

5 Conclusion

Appendix A

Sample Appendix

Stuff for this appendix goes here.

Appendix B

Another Sample Appendix

Stuff for this appendix goes here.