

HOME ADVANTAGE IN NORTH AMERICAN PROFESSIONAL
SPORTS BEFORE AND DURING COVID-19: A BAYESIAN
PERSPECTIVE

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

By

Nicholas Higgs

©Nicholas Higgs, August 2021. All rights reserved.

Unless otherwise noted, copyright of the material in this thesis belongs to
the author.

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

Home advantage in professional sports is a widely accepted phenomenon despite the lack of any controlled experiments at the professional level. The return to play of professional sports during the COVID-19 pandemic presents a unique opportunity to analyze the hypothesized effect of home advantage in neutral settings. While recent work has examined the effect of COVID-19 restrictions on home advantage in European football, comparatively few studies have examined the effect of restrictions in the North American professional sports leagues. In this work, we infer the effect of and changes in home advantage prior to and during COVID-19 in the professional North American leagues for hockey, basketball, baseball, and American football. We propose a Bayesian multilevel regression model that infers the effect of home advantage while accounting for relative team strengths. We also demonstrate that the Negative Binomial distribution is the most appropriate likelihood to use in modelling North American sports leagues as they are prone to overdispersion in their points scored. We further demonstrate that multilevel regression provides better model fit to the datasets considered in this thesis as compared to traditional regression modelling and simple averaging often employed in related work. Our model gives strong evidence that home advantage was negatively impacted in the NHL and NBA during their strongly restricted COVID-19 playoffs, while the MLB and NFL showed little to no change during their weakly restricted COVID-19 seasons.

Acknowledgements

I would like to thank my supervisor for giving me the freedom to pursue my research interests and the guidance to ensure my work achieves a high standard.

I dedicate this thesis to my wife, without whom I would not have been able to pursue and achieve my goals.

Contents

Permission to Use	i
Abstract	iii
Acknowledgements	iv
Contents	vi
List of Tables	vii
List of Figures	viii
List of Abbreviations	x
1 Co-Authorship Statement	1
2 Introduction	2
2.1 Contribution	3
2.2 Thesis Organization	5
3 Background on Bayesian Inference	6
3.1 Introduction	6
3.2 Multilevel Modelling	8
3.3 Markov Chain Monte Carlo	13
3.4 Model Evaluation and Selection	19
4 Related Work on Sports Analytics	26
5 Methods	28
5.1 Multilevel Model	28
5.1.1 Model Fit in PyMC3	30
5.2 Experiments	32
5.2.1 Data	32
5.2.2 Complete pooling, No pooling, and Partial pooling	34
5.2.3 Negative Binomial Regression	35
5.2.4 Inferring Home Advantage	36
6 Results	37
6.1 Complete pooling, No pooling, Partial pooling	37
6.2 Negative Binomial Regression	40
6.3 Inferring Home Advantage	40
7 Discussion and Conclusions	46
7.1 Discussion	46
7.2 Limitations	48
7.3 Conclusions	48
References	50
Appendix A Appendix	53

List of Tables

6.1 Comparison of estimated negative log-likelihood of leave-one-out cross-validation (LOO) for each model across each league. The differences between the Poisson, Negative Binomial (NB), and Normal models are reported relative to the best fitting model (dLOO) for each league; along with the standard error of the estimated differences (dSE). The dispersion statistic, σ_p , indicates how much greater the variance is than the mean for point totals in each league and signals overdispersion when $\sigma_p > 2$. The NB model noticeably outperforms the Poisson model for leagues with greater overdispersion (MLB and NFL) while being nearly identical for leagues with little to no overdispersion (NHL and NBA). The NB model also outperforms the Normal model in each league except the NFL where they are close to one another while both vastly outperforming the Poisson model.	42
A.1 The estimated probabilities that the home advantage parameter during the 2020 COVID-19 restricted games (β_{20}) is less than 0, the previous four seasons (2016-2019) mean ($\bar{\beta}_{16-19}$), and the previous seasons individual means ($\bar{\beta}_{19}, \bar{\beta}_{18}, \bar{\beta}_{17}, \bar{\beta}_{16}$).	53

List of Figures

3.1	Comparison of county parameter estimates between traditional regression (no pooling) and multilevel regression (partial pooling). Notice how the partial pooling estimates “shrink toward the mean”. Further note how this shrinkage is greater for the counties with fewer observations and lesser for counties with more observations.	10
3.2	Under ideal circumstances a Markov chain will first converge to the typical set (a) and then explore it efficiently (b). Unfortunately, in higher dimensions most MCMC algorithms struggle to explore the typical set and inefficiently sample a small portion (c, green). We desire algorithms that make use of the geometry of the target distribution to properly explore the typical set during sampling (d). Images are from Figures 7, 10, 11 of [7]. Permission to use was granted by the author under a CC BY-NC 4.0 license (https://creativecommons.org/licenses/by-nc/4.0/).	14
3.3	The gradient and corresponding vector field of a probability distribution points to its mode which is often away from the typical set in higher dimensions (a). Ideally we want to twist the vector field to align with the typical set (b). The mode, gradient, and typical set of a probabilistic system are mathematically equivalent to a planet, gravitational field, and orbit in a physical system (c). Adding momentum to the system to cause a satellite to enter a stable orbit (d) is equivalent to twisting a vector field to align with the typical set of a probabilistic system. Images are from Figures 12, 13, 14, 17 of [7]. Permission to use was granted by the author under a CC BY-NC 4.0 license (https://creativecommons.org/licenses/by-nc/4.0/).	17
3.4	Example of how increasing model complexity leads to better model fit on the train-set, but can come at the cost of increasingly worse performance on the test-set. Model fit here is measured visually and in terms of mean-squared-error (MSE: lower is better) and R-squared (R^2 : closer to 1 is better). The dataset in (a) is generated by a degree-2 polynomial with some added noise and is split into train and test sets. A degree-1 polynomial underfits the data (b). More complex polynomials improve the fit on the train-set (c, d, e). However, increasingly complex polynomials become overfit as evidenced by increasingly worse test-set performance (d, e). The overall trend of increasing model complexity, how it relates to underfitting and overfitting, and where the tradeoff is optimal is captured in (f).	20
5.1	An example of how a Bayesian model would be defined in an academic textbook or paper (a) and how a probabilistic programming language such as PyMC3 would create the model in Python code (b). Note how the priors have to be defined first because the code will be executed procedurally. The two definitions are essentially identical otherwise.	31
6.1	Comparison of models via their Log-Score (higher is better) on train and test sets, as well as the PSIS-LOO estimated Log-Score, for each league. The complete-pooling model under-fits, the no-pooling model over-fits, and the partial-pooling model provides the best trade-off in fitting the data while protecting against over-fitting. The PSIS-LOO estimates consistently predict how the models would rank on an unseen test-set.	38
6.2	Comparison of models via their PSIS-LOO estimated Log-Score for each league, ranked from best (highest) to worst (lowest) on the y-axis. The black points and lines represent the point estimate and its standard error. The grey triangle and lines represent the estimated difference and the standard error of the difference for each model relative to the best model. The standard error of the difference is generally much smaller than the standard error of the estimate because errors in the estimates for each model are highly correlated.	39
6.3	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.	41

6.4	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.	43
6.5	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.	45
A.1	Offensive and Defensive team ratings for the 2020 NHL season. The points with the team labels next to them are ratings generated by traditional regression, and the corresponding ratings are generated from multilevel regression to highlight the effect of shrinkage to the mean.	54
A.2	Offensive and Defensive team ratings for the 2020 NHL season. The points with the team labels next to them are ratings generated by traditional regression, and the corresponding ratings are generated from multilevel regression to highlight the effect of shrinkage to the mean.	55
A.3	Offensive and Defensive team ratings for the 2020 MLB season. The points with the team labels next to them are ratings generated by traditional regression, and the corresponding ratings are generated from multilevel regression to highlight the effect of shrinkage to the mean.	56
A.4	Offensive and Defensive team ratings for the 2020 NFL season. The points with the team labels next to them are ratings generated by traditional regression, and the corresponding ratings are generated from multilevel regression to highlight the effect of shrinkage to the mean.	57

List of Abbreviations

COVID-19	Coronavirus Disease 2019
NHL	National Hockey League
NBA	National Basketball Association
MLB	Major League Baseball
NFL	National Football League
MCMC	Markov Chain Monte Carlo
HMC	Hamiltonian Monte Carlo
LOO-CV	Leave-One-Out Cross-Validation
KL divergence	Kullback-Leibler divergence
lppd	log-pointwise-predictive-density
AIC	Akaike Information Criterion
WAIC	Widely Applicable Information Criterion or Watanabe Akaike Information Criterion
PSIS-LOO	Pareto-Smoothed Importance-Sampling Leave-One-Out cross-validation
PPC	Posterior Predictive Check
PPL	Probabilistic Programming Language
NUTS	No U-Turn Sampler
xG	expected Goals
RAPM	Regularized Adjusted Plus-Minus

1 Co-Authorship Statement

Portions of this thesis were published in [33]. I collected the data, carried out the experiments, and jointly conceived of the experiments. To remain consistent with wording used in [33] I have elected to use the third person “we” opposed to the first person “I” for the abstract and main body of the thesis.

2 Introduction

In professional sports, home teams tend to win more on average than visiting teams [52] [17] [45]. This phenomenon has been widely studied across several fields including psychology [1] [58], economics [21] [19], and statistics [12] [36] among others [6]. While home advantage is now a widely accepted phenomenon, the magnitude of the advantage and its cause are not as clearly understood or widely accepted as its existence. Part of the difficulty in analysing the specifics of home advantage is due to the lack of controlled experiments, because nearly every professional game is played in one of the team’s home stadium in their home city. While there have existed some show matches at neutral sites, their relative sample sizes are too small from which to draw any reasonable conclusions. For example, the National Football League only plays about 4-5 neutral site games out of a total 256 games each regular season.

The return to play of professional sports during the COVID-19 pandemic presents a unique opportunity to analyse teams playing in situations where home advantage may genuinely no longer apply. The leagues have restricted travel and fan attendance or even created a bubble where only one or two stadiums are used and only the players and necessary staff are present for the games. We consider this restricted return to play as a control group where travel, home stadium familiarity, and home crowd have been controlled (i.e. removed) for enough games to provide a reasonable sample to analyse. There has been considerable academic work analysing the effect of COVID-19 restrictions on home advantage in European football [6]. However, comparatively there has been a lack of work analysing the effect in the North American professional sports leagues. In fact, to the authors knowledge there has only been one work focused on home advantage during COVID-19 across the big four North American professional leagues; and it only investigated the NBA [39]. In this work, we aim to fill this gap by inferring the effect of and changes in home advantage prior to and during COVID-19 in the big four North American leagues: the National Hockey League (NHL), the National Basketball Association (NBA), Major League Baseball (MLB), and the National Football League (NFL).

Previous works analysing home advantage tend to pool the results of all teams within a league into one overarching group statistic to analyse, such as overall-league-home-win-percentage. However, more recent work tries to better account for differences amongst teams within a league, such as differences in teams offensive and defensive strengths, by utilizing multiple regression models that account for the effect of other variables while inferring the effect of home advantage. The majority of modern work utilizing more sophisticated regression modelling has been applied to sports outside of the four North American professional leagues considered in this thesis. We fill this gap by developing a Bayesian multilevel regression model and show how the model fits the datasets, of the four professional leagues considered, better than simple averaging and

traditional regression modelling.

Professional sports leagues adopted different restrictions in response to the COVID-19 pandemic. The NHL and NBA had the strongest restrictions where they both created a COVID-19 bubble where all games were played at the same consistent location with players quarantined together separate from their families and the outside world. While this proved to be extremely effective in terms of player safety [16] [61] it seems likely that it was the most extreme in terms of its effect on players performance and psychology. In contrast, teams in the MLB and NFL still travelled to their opponents home stadiums. These leagues restricted fan attendance and media access, with some NFL stadiums allowing small amounts of fans to attend. Thus, all leagues lacked a potential home crowd effect, but only the NHL and NBA restrictions removed the additional factors of travel and home city familiarity. Because the restrictions for the NHL and NBA were more strict than those of the MLB and NFL, analysing all four leagues brings the potential to see similarities and differences across leagues as well as within each individual league. Thus similarities in NHL and NBA as compared to similarities and differences with the MLB and NFL can potentially shed light on the differing effects contributing to home advantage, in particular the differences in the effects of home crowds, familiarity with home cities, and travel. This is noteworthy because of the implications in relation to previous work investigating the causes of home advantage [58] [12] [17] [15] [39] [23] [42]. In McHill & Chinoy [39], the authors argue that home advantage in the NBA’s COVID-19 bubble arose from either circadian disruption or the general effect of travel. Our work builds upon such previous works by considering the NBA’s COVID-19 bubble and its effects on home advantage while also comparing and contrasting to other similar COVID-19 bubbles in the NHL and different COVID-19 restrictions seen in the MLB and NFL.

2.1 Contribution

We adopt a Bayesian framework to develop a Negative Binomial multilevel regression model that adjusts for relative team strengths while inferring home advantage. We are motivated to 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 analyse 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.

Beyond the motivation for adopting a Bayesian framework, we further show that multilevel regression modelling provides better fit to the datasets considered in this thesis as measured by out-of-sample predictive fit. We also show that several of the professional sports considered are prone to overdispersion in their points

scored. We then demonstrate that the Negative Binomial distribution is a better likelihood function to use than the Poisson and Normal distribution that are more commonly used for regression modelling in related work. With the efficacy of our model established, we then fit the final model and examine the resulting distributions of likely values for the home advantage parameter.

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. We further show that a multilevel model that pools information for teams offensive and defensive strengths provides better model fit as measured by estimated out-of-sample predictive fit as compared to traditional regression modelling and simple averaging used in many other works of sports modelling and home advantage inference.

We believe our model would potentially benefit coaching staffs as well as stakeholders in the gambling industry. Coaching staffs could benefit from our model when determining roster changes. This is relevant when coaches need to balance developing younger players and giving their backups play-time without hurting their teams overall performance and place in the league standings. By using our model to better quantify relative differences in team strengths and home advantage, coaches could more precisely determine for which games they could play their younger players and backups more without sacrificing too many game wins. Our model also benefits both bookmakers and bettors in the gambling industry by better quantifying the effect of, the change in, and the uncertainty of home advantage both in times of relative stability (e.g. pre-COVID-19) and in less stable times (e.g. the return to play of COVID-19). This is useful for anyone trying to set more accurate betting lines (i.e. bookmakers), or someone trying to identify less accurate betting lines (i.e. bettors).

The main contributions of this thesis can be summarized as follows:

1. First study to provide concrete analysis of home advantage in a controlled setting for the NHL, NBA, MLB, and NFL.
2. Corroberates results from similar studies analysing professional European soccer leagues finding a drop in home advantage in some but not all leagues.
3. Organized a dataset comprised of game results across the NHL, NBA, MLB, and NFL for the years 2016-2020.
4. Developed a Bayesian multilevel model, provided background on multilevel modelling and evaluation,

demonstrated efficacy of model on the datasets used in this thesis.

5. Demonstrated how North American professional sports are prone to overdispersion in point totals.
6. Proposed a Negative Binomial multilevel regression model to account for overdispersion, evaluated with respect to Poisson and Normal regression models commonly performed in related work.
7. Developed a model that potentially benefits coaching staffs as well as stakeholders in the gambling industry.

2.2 Thesis Organization

The rest of the thesis is organized as follows. The Background chapter provides an overview of related work as well as an introduction to Bayesian statistics, multilevel modelling, fitting Bayesian models via Markov chain Monte Carlo sampling, and evaluating models. The Methods chapter describes in-depth the Bayesian multilevel regression that was developed to infer home advantage, and how the various data experiments for the main contributions of this thesis are set up. The Results chapter presents and discusses the results of the experiments introduced in the Methods chapter. The Conclusions chapter contains a discussion of the results and their implications of the findings and contributions of the thesis.

3 Background on Bayesian Inference

In this chapter we introduce Bayesian inference and multilevel modelling. We then explore how to fit and evaluate Bayesian models.

3.1 Introduction

Bayes theorem, and by extension Bayesian statistics and inference, is named after an amateur mathematician and Presbyterian minister from the 18th century, the Reverend Thomas Bayes. After Bayes' death, his friend Richard Price found an essay Bayes wrote titled "an imperfect solution of one of the most difficult problems in the doctrine of chances". Price saw the value in Bayes' work and submitted it to the Royal Society for publication. In Bayes' time (circa the 17-18th centuries) probability and statistical theory as we know it today was in its infancy and not widely studied as it is now. The leading probability thinkers of the time conceived of the subject through the lens of gambling and games of chance [18] [41]. The thinkers of the time had reasoned about probability from cause to effect in these contexts (e.g. what are the odds of getting four aces in a poker hand?). What Bayes had shown in his essay was a potential solution to a yet unsolved problem: the so-called inverse-probability problem of reasoning from effect to cause (e.g. if a player deals himself four aces in three hands in a row, what are the odds his deck is loaded?). The legendary mathematician Pierre Simone Laplace fused Bayes' ideas with his own and published what we now know as Bayes theorem in 1825 [56]. The legacy of Bayes and Laplace's work is that we now refer to the general approach of using data (effect) to estimate parameters (cause) through the use of Bayes' theorem as Bayesian statistics and inference.

Bayes theorem can be stated simply as:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \quad (3.1)$$

This equation can be derived 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 substitute and isolate $p(A|B)$ to arrive at equation 3.1.

Bayes theorem is usually derived from the basic rules of probability and introduced as equation 3.1, and then counter-intuitive examples are used to show the efficacy of the theorem (e.g. a test for a rare disease is 99% accurate, if you test positive what is the probability that you have the disease?). However, Bayesian statistics extends 3.1 to the context of data and model parameters. In this extension, Bayes theorem instead describes a joint probability distribution over all observed and unobserved parameters in a statistical model

[51]. With a data set x and parameters θ , we can rewrite Bayes theorem as:

$$P(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \quad (3.2)$$

In equation 3.2, each part of the equation is referred to and interpreted differently than in 3.1. The conditional probability $P(\theta|x)$ is referred to as the *posterior* distribution and represents the probability of the model parameters, θ , conditional on the data, x . The conditional probability of the data given the model parameters, $P(x|\theta)$, is referred to as the *likelihood*. The probability of particular model parameter values existing in the population, $p(\theta)$, is referred to as the *prior* distribution. The denominator, $p(x)$, functions as merely a normalizing factor to ensure that the posterior probabilities sum to 1, but it does not change their relative values. Notice that $p(x)$ does not explicitly depend on θ . Thus, we can simplify 3.2 by dropping $p(x)$ and re-interpret Bayes theorem recognizing that the posterior distribution is proportional to the likelihood function multiplied by the prior distribution:

$$p(\theta|x) \propto p(x|\theta)p(\theta) \quad (3.3)$$

Intuitively, Bayesian statistics starts with a prior belief (i.e. prior distribution over parameters of a model) and then updates that belief with new information (i.e. the likelihood of the data) resulting in an updated posterior belief (i.e. the posterior distribution of model parameters). This updated posterior will then serve as the new prior distribution when more information is available in the future for us to yet again update our belief. In this way our beliefs are continually updated with data in order to make them increasingly more accurate. While this intuition is generally appealing on its own, more importantly Bayesian statistics has been successful in solving challenging problems in applied statistics both historically and more recently [51]. Despite its intuitive appeal and real world successes, Bayesian statistics actually declined in popularity during the first half of the 20th century. Understanding this decline and how it has been overcome shows why Bayesian techniques are becoming increasingly popular in the modern scientific era.

The Decline and Resurgence of Bayesian Statistics

The primary reasons for the decline in the popularity of Bayesian statistics were an objection to the “subjective” use of prior distributions, and the difficulty in actually computing the posterior distribution in 3.2. Most prominent statisticians of the early 20th century did not like the “subjectivity” of specifying a prior that could potentially influence the results of inference for what was supposed to be objective science. In particular, many of the most prominent statisticians of the 20th century, including R.A. Fisher and Karl Pearson, were vocal opponents of subjectivity and Bayesian statistics which they saw as being synonymous. This “smear campaign” combined with computational difficulties led much of statistics and science to turn to judging the probability of an event according to how frequently it occurs among many observations. This preferred view of statistics, lauded by its theorists of the early 20th century as being “objective”, led to

the widespread adoption of traditional statistical methods often referred to as *frequentist* statistics. Despite these challenges, Bayesian statistics still saw some use and success in real-world contexts including but not limited to actuarial science and military applications [38]. Meanwhile traditional frequentist methods have come under increasing criticism in recent decades [34] [5]. Scientists in the modern era are becoming increasingly aware that all statistical methods are subjective in the sense that all statistical techniques make assumptions. Bayesian statistics is merely more transparent about its assumptions in its model formulation than most other methods. While subjectivity in the form of using priors was originally seen as a negative, it is instead now seen as a strength. Since “all models are wrong, but some are useful” [9], it is more responsible to be upfront and clear about the subjective assumptions of any statistical model or methodology rather than blindly optimizing a misunderstood technique, such as seeking a p-value < 0.05 .

The second primary reason for the decline of Bayesian statistics in the 20th century is the difficulty in actually computing 3.2 for most problems. The computational difficulties arise specifically from the normalization factor ($p(x)$) in 3.2. The normalization factor can also be thought of as the probability of the dataset, which is something we generally don’t know a priori. Thus, we have to turn to the law of total probability in order to compute it which means that the calculation of this normalization factor requires integrating over all possible parameters (θ) as follows:

$$p(x) = \int_{\theta} p(x|\theta)p(\theta)d\theta \quad (3.4)$$

The integral in 3.4 can sometimes be computed analytically in low dimensions, most often in situations known as *conjugate priors* where the prior and likelihood conveniently combine into another known distribution. However, in higher dimensions where the number of parameters making up θ is larger and when using distributions for the prior and likelihood that do not form convenient conjugate pairings, the integral in 3.4 becomes mathematically intractable. This means the integral can not be computed exactly and instead can at best be approximated using numerical techniques. However, many of the numerical techniques and computing devices we have today did not exist in the first half of the 20th century which meant Bayesian methods were out of reach for most scientists. The advances in computing as well as the theory behind numerical techniques, most notably Markov chain Monte Carlo, have made approximating the posterior in 3.2 computationally feasible which has greatly contributed to the resurgence of Bayesian techniques in recent decades.

3.2 Multilevel Modelling

The process of having a prior belief, updating it with new information, and then taking the resulting posterior to be your updated belief, or your new prior belief moving forward, is a process that often resonates with people and how their views and beliefs about the world are constructed and updated. While not only intuitively appealing, there are also many practical examples where the use of a prior distribution in Bayesian

inference leads to better results [51]. This makes Bayesian inference an appealing option; however, actually constructing or deciding on a prior distribution can be difficult. Multilevel modelling is a method that sets up the prior distribution to be learned from the data. This makes specifying a prior easier as the data is primarily determining the prior, and it generally also leads to better out-of-sample predictive performance, making multilevel modelling an attractive option. This section provides background to understand how multilevel modelling works and describes the motivation for using multilevel modelling. The effectiveness of multilevel modelling is further explored in the data experiments described in the Methods chapter.

Multilevel modelling 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 or categories within the overall data are ignored, thus the data are “completely pooled”. No-pooling is when separate models for each individual group or 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 as a trade-off between the complete-pooling (e.g. overall group mean) and no-pooling (e.g. individual group means) extremes. For groups with fewer data points the multilevel model produces estimates more similar to the complete-pooling estimate, and for groups with more data points the model produces estimates more similar to the no-pooling estimates. 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 the mean”. The amount of shrinkage depends on the samples-sizes, the variation within groups, and the variation between groups.

It is helpful to consider an estimate from a simple partially pooled model in order to understand how partial-pooling works. Consider a model that has one categorical predictor indicating which group an observation belongs to (e.g. which province a voter lives in), and no other predictors. We denote this model by $y = \alpha_j$. In this case the partially pooled model will generate predictions for each group by constructing a weighted average of the group level means and the overall mean. Mathematically the estimates would be computed as follows:

$$\hat{\alpha}_j^{multilevel} \approx \frac{\frac{n_j}{\hat{\sigma}_y^2} \bar{y}_j + \frac{1}{\hat{\sigma}_\alpha^2} \bar{y}_{all}}{\frac{n_j}{\hat{\sigma}_y^2} + \frac{1}{\hat{\sigma}_\alpha^2}} \quad (3.5)$$

where $\hat{\alpha}_j^{multilevel}$ is the estimate from the multilevel model for the j -th group. It is the weighted average of the j -th groups average (\bar{y}_j) and the average of all groups combined (\bar{y}_{all}). The weights are determined by the within-group variance ($\hat{\sigma}_y^2$), the sample size of the j -th group (n_j), and the variance among the groups ($\hat{\sigma}_\alpha^2$). In this way, larger (smaller) sample sizes for the j -th group and the lower (greater) within-group variance leads to larger (smaller) weight placed on the j -th group average for the final estimate. Smaller (larger) variance among the groups leads to a larger (smaller) weight placed on the overall average for the estimate of the j -th group. This view makes it clear that the estimates from a multilevel model will compute a group estimate in a similar way to a more traditional regression model but will then shrink that estimate toward the overall mean weighted by the groups sample size, the within group variance, and the among group variances.

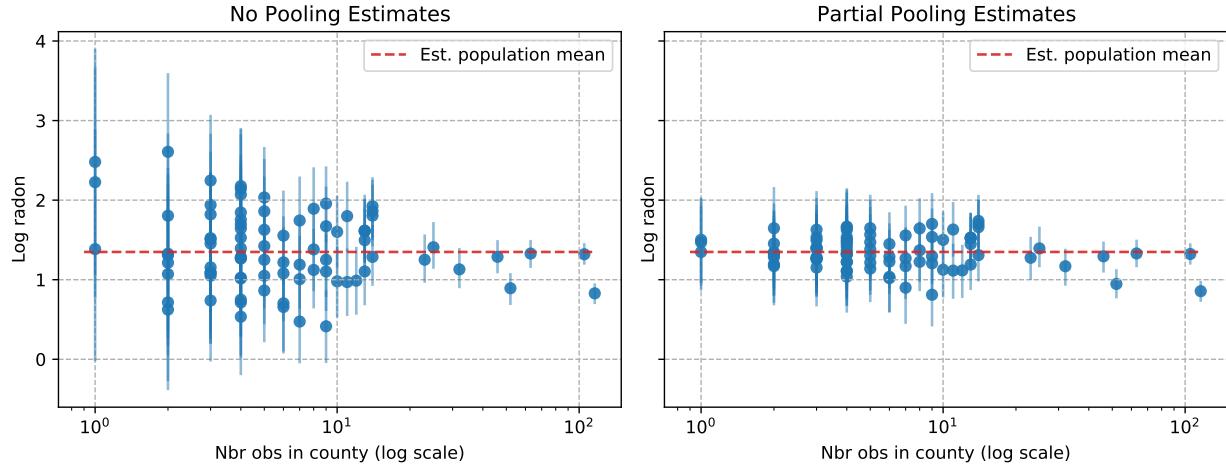


Figure 3.1: Comparison of county parameter estimates between traditional regression (no pooling) and multilevel regression (partial pooling). Notice how the partial pooling estimates “shrink toward the mean”. Further note how this shrinkage is greater for the counties with fewer observations and lesser for counties with more observations.

Here the overall mean and how much the estimates should shrink toward it is determined by the data and represents the prior distribution over the parameters which is then conditioned by the data. It is in this manner that the prior is learned from the data by pooling information across groups.

Equation 3.5 has an \approx symbol rather than an $=$ symbol because it is only in a few mathematically convenient cases, such as conjugate priors, that the group level estimate would precisely reduce to the formula in equation 3.5. Including more predictors, more mathematically complex transformations and other engineered features, and using more varied probability distributions that do not result in conjugate priors, all lead to estimates that are no longer mathematically tractable and instead require approximation methods such as Markov chain Monte Carlo to generate the estimates. However, even in such complex cases where the estimates cannot be computed analytically they still in practice function the same way as outlined by equation 3.5 [27].

House Radon Contamination Example

For a more concrete example of the regularizing shrinkage to the mean effect of multilevel modelling we consider the canonical example from Andrew Gelman’s work [27] [24], which is now often considered as the introductory tutorial of multilevel modelling [57]. In [24] the strengths and limitations of multilevel modelling are illustrated through an example of the prediction of home radon levels in U.S. counties. To identify areas of high radon exposure, the Environmental Protection Agency coordinated the collection of radon measurements in a random sample of more than 80,000 houses in the United States. In addition to these measurements were predictors for indicating if the measurement was on the first floor of the home or in the basement, and what the county uranium levels are for each county in which the homes were located (approximately 3,000

counties total). Gelman showed that multilevel modelling outperformed traditional regression modelling as measured by out-of-sample performance estimated by cross-validation prediction errors.

A simplified example of the work in [24] is shown in Figure 3.1 where we compare the results of fitting a traditional regression model and a multilevel regression model that estimates house radon levels using the county that the houses belong to as the only predictor for the counties in the state of Minnesota. This simplified model can be expressed as $y_i = \alpha_{j[i]} + \epsilon_i$, where y_i is the radon level for house i ($i = 1, \dots, 919$), $\alpha_{j[i]}$ is the average radon level for the j -th county ($j = 1, \dots, 85$) of which the i -th house belongs, and ϵ_i represents the random errors due to measurement error, temporal within-house variation, or variation among houses. As previously discussed, the traditional regression model will model radon in each county independently resulting in no-pooling estimates, while the multilevel model will pool information across counties to make estimates similar to equation 3.5 resulting in partial-pooling estimates. We note how the counties with fewer observations result in greater shrinkage towards the overall mean, while counties with more observations stay closer to their no-pooling estimates. The overall mean across counties acts as a regularizing prior distribution, but this prior was also learned from the data. This regularizing effect of shrinking individual group parameter estimates towards the overall group mean is the essential idea behind multilevel modelling. The appendix contains figures showing the shrinkage effect of multilevel modelling on team ratings from the model used in this thesis.

Advantages of Multilevel Modelling

The advantages of multilevel modelling are thoroughly explored in the works of [26] [27] [37] and are summarized here to provide further motivation for the use of multilevel modelling in this thesis. Multilevel modelling is useful because it 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 [47]. This makes multilevel models useful beyond predictions alone. In contrast, many machine learning methods such as neural networks and ensemble decision trees are not interpretable.

Many datasets have an inherent multilevel structure for which multilevel modelling can provide more efficient inference of regression parameters (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 begin sub-dividing them into more and more sub-groups. For example, opinion polls trying to predict who voters will vote for based on age, race, income, location, interests etc. Each split leaves smaller and smaller groupings that have the potential for better model fit, since there are more predictors, at the risk of over-fitting, since sample sizes of the sub-groups become increasingly small.

Multilevel models allow for including predictors at two different levels of a regression model. 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 [24].

Multilevel modelling avoids problems in classical regression such as collinearity when trying to include group-level indicators as well as group-level predictors in the same model by using *index variables* instead of *dummy variables*. This is most noticeable when considering the “reference group” that results from using dummy variables to encode categorical variables in traditional regression. N many categories can be encoded with $N - 1$ many dummy variables. For example, consider three categories only requiring two dummy variables we can refer to as α and β . The first category is encoded with $\alpha = 1$ and $\beta = 0$. The second category is encoded with $\alpha = 0$ and $\beta = 1$. The third category is encoded with $\alpha = 0$ and $\beta = 0$; there is no need for a third dummy variable. If a third dummy variable is added the resulting system of equations used to solve for the coefficients for each indicator variable becomes singular [22], which means you cannot get estimates for the regression coefficients using linear algebra. Furthermore, because the third category is encoded by the absence of the first two, it does not get its own regression coefficient and instead the intercept and all other regression coefficients of the model are changed to reflect this. As a result, the intercept will now represent the third category and the coefficients for the other categories will represent differences relative to the third category. This makes the third category the reference group and can be a source of confusion when interpreting a model. In Bayesian modelling this requires specifying a prior distribution for the difference of each category from the reference category, as well as a prior distribution for the reference category. By contrast, an index variable gives an index to each category (a unique integer for each category, starting at 1 and increasing up to the number of categories) and does not require a reference group. This allows for assigning the same prior distribution to each category, which cannot be done with dummy variables, and makes scaling a model to include more or new categories seamless. It also makes interpreting the coefficients for each category easier as they no longer represent differences from one of the categories. Leveraging index variables allows multilevel models to avoid issues with collinearity while being more interpretable.

Multilevel modelling aids in inferring the right standard error by 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, and predictors (e.g. forecasting state-by-state outcomes in the U.S. election, one must account for correlation of outcome between states in a given year). This becomes more useful in cases where the uncertainty in estimation is of interest rather than the estimate itself.

Sometimes predictions require multilevel modelling, 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 by using group level predictors to estimate where the new school would fall relative to the other groups.

Multilevel modelling is attractive because it comes with all the benefits of regression modelling while generally outperforming classical regression modelling in predictive accuracy. The primary source of improvement over classical regression is due to the shrinkage of parameter estimates generally improving out-of-sample predictive fit. The improved performance is due to using all the data to perform inferences for groups, especially those with small sample sizes. At one extreme classical estimation can be useless if the sample size is small in a group or category. At the other extreme classical regression ignores group-level variation which can be misleading especially when some groups have small sample sizes. Multilevel modelling compromises between the overall noisy within-group estimates (no-pooling) and the oversimplified regression estimate that ignores group variation (complete-pooling). The shrinkage effect of multilevel modelling acts as a form of regularization that protects from over-fitting to produce more accurate predictions on unseen data. Overfitting in regression modelling and how multilevel models help prevent it is explored further in the first data experiment described in 5.2.2.

While the benefits of multilevel modelling are bountiful, they are mathematically more complex than classical regression and are generally mathematically intractable. This means we can not compute the parameter estimates directly and instead need to approximate them with more advanced numerical approximation techniques.

3.3 Markov Chain Monte Carlo

For most models of practical interest, exact inference is intractable, and so we have to resort to some form of approximation [8]. The primary end goal of Bayesian inference is computing the posterior distribution. It is with the posterior distribution that we can perform inference and answer questions about quantities of interest. The issue is that computing the posterior distribution for nearly all but the simplest of models is not only difficult but often impossible. That is to say that we can not derive a closed form mathematical expression that represents the posterior distribution. We can, however, approximate the posterior. Researchers have developed many different methods of numerical approximation which can be employed to approximate the posterior distribution in Bayesian inference.

Most methods that attempt to approximate the posterior distribution (or approximate integrals more generally) work well in low-dimensional settings but struggle or outright fail in high-dimensional settings due to a phenomenon known as *concentration of measure*. Concentration of measure refers to the fact that in low dimensions the probability mass of a distribution is concentrated around its mode, but in higher dimensions the probability mass of a distribution is surprisingly not concentrated around its mode and becomes increasingly further away from the mode as the dimensionality increases. The probability mass of a distribution concentrates into a density band into which almost all random draws from a distribution will

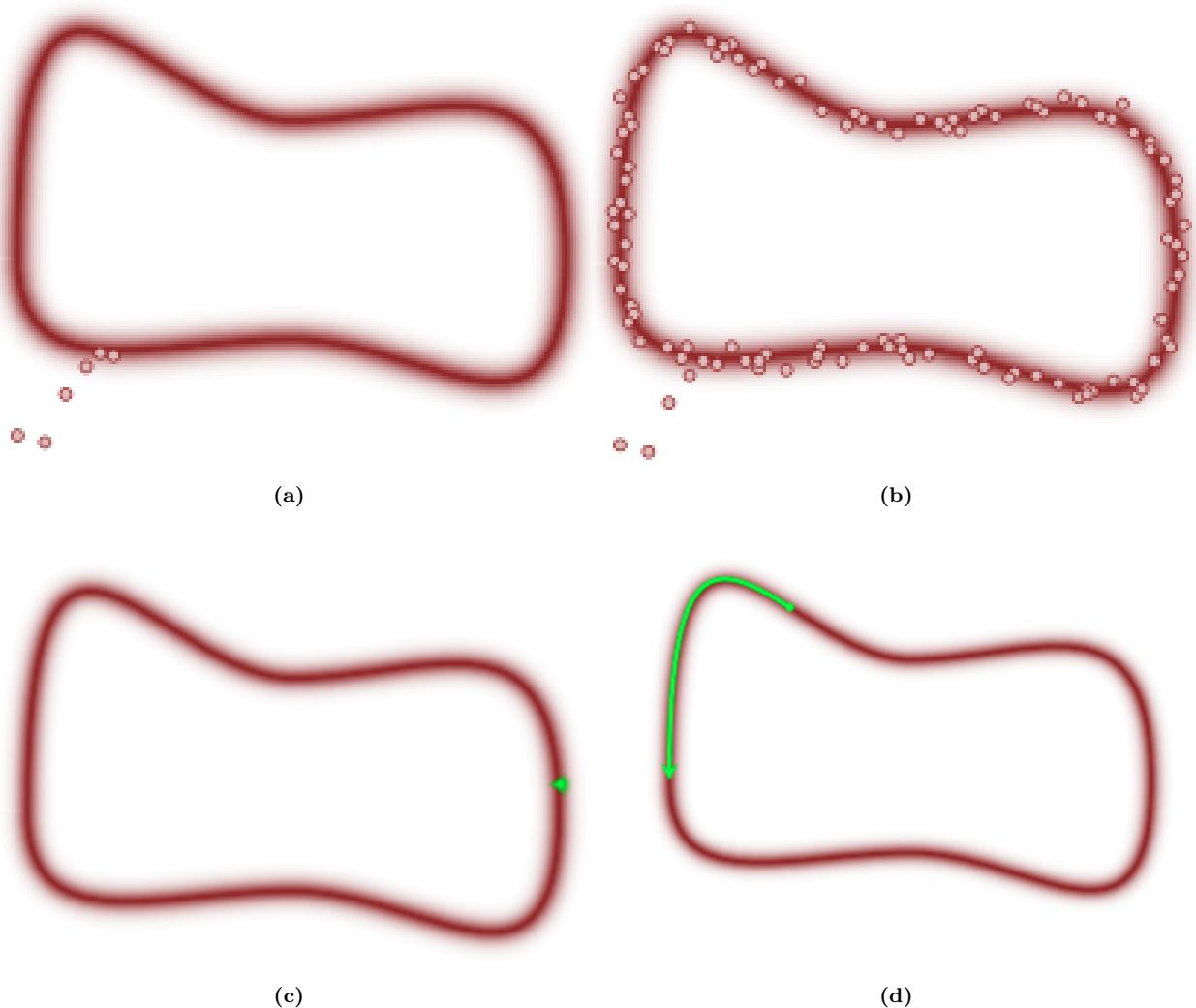


Figure 3.2: Under ideal circumstances a Markov chain will first converge to the typical set (a) and then explore it efficiently (b). Unfortunately, in higher dimensions most MCMC algorithms struggle to explore the typical set and inefficiently sample a small portion (c, green). We desire algorithms that make use of the geometry of the target distribution to properly explore the typical set during sampling (d). Images are from Figures 7, 10, 11 of [7]. Permission to use was granted by the author under a CC BY-NC 4.0 license (<https://creativecommons.org/licenses/by-nc/4.0/>).

fall, and is referred to as the *typical set*. The typical set is therefore precisely where we want to sample from in order to generate accurate approximations of the posterior distribution. For a deeper treatment of these concepts see [7] [14]. The models considered in this thesis are relatively high-dimensional, therefore we employ the current state of the art, Hamiltonian Monte Carlo (HMC), for approximating high-dimensional posterior distributions by sampling efficiently from the typical set. In order to understand what makes HMC so effective we first explore its simpler foundational method, Markov chain Monte Carlo (MCMC).

The most common method to approximate computing a desired probabilistic quantity is to repeatedly draw independent samples from the probability distribution and to then average over those samples to approximate the quantity of interest. This is known as Monte Carlo sampling. Just as statisticians traditionally aim to draw independent samples in order to estimate desired quantities, such as the mean, variance, or specific quantiles, about a target population, Monte Carlo sampling aims to draw independent samples from a probability distribution in order to approximate the distribution or a specific property of that distribution.

Drawing independent samples from a known distribution to then only be able to approximate said distribution appears counter-intuitive at best and wholly wasteful at worst. In practice, however, we do not actually know the distribution that we want to sample from. The most powerful and surprising insight of statistical computing, and MCMC in particular, is that we can sample from a distribution that we do not know and then use those samples to approximate the unknown distribution. We can do this by drawing samples from, or “visiting” each part of, the distribution in proportion to its relative probability. Sampling in proportion to the relative probability of a distribution is done by making Markov transitions via the use of a Markov chain. We then draw samples in this manner enough times to generate a sequence of samples that closely approximates the distribution of interest.

A Markov chain is a probabilistic model that describes a sequence of possible states in which the probability of each state depends only on the previous state. This means that no matter how the process arrived at the current state, the possible future states are fixed based on the current state. This allows you to go from one state to another repeatedly as many times as you desire or need to. The entire sequence of states you visit then represents a chain. For our purposes, we can think of states as locations in the parameter space of the distribution which we are trying to sample from, and the chain is the sequence of samples. Future states can then be determined by the relative probability density of other locations in the parameter space, computed as in 3.3. This forms the basis of one of the most well-known MCMC algorithms, the Metropolis algorithm.

The simplest and most well known MCMC algorithm, the Metropolis algorithm, begins by randomly selecting a starting location in the parameter space, generates a new “proposal” location to move to in the parameter space, but only moves to this new location if it has a higher density relative to the previous location or by random chance proportional to the difference in relative densities of the current location to the proposed location. As the algorithm runs for more samples, it will visit each location in the parameter space proportional to the probability density of each location, thus the sequence of samples will approximate the probability distribution more accurately as more samples are drawn. The drawback of this algorithm is

trying to determine how many samples is enough. While it has been shown that the samples will tend toward the correct proportions and thus correct probability densities [40] [32], there is no rigorous general theory to determine how many samples is enough and how accurate your approximation actually is so that you can know if you have sampled enough.

There have been many advances upon and extensions of the Metropolis algorithm that attempt to improve the generalizability and efficiency of the sampling. These include but are not limited to the Metropolis-Hastings algorithm and Gibbs sampling [32] [28]. These algorithms can broadly be grouped together as “guess and check” algorithms. They “guess” a random proposal of where to move, they then “check” the posterior probability at that location and compare it to the current location. The consequence is that the quality of proposals becomes the primary bottleneck. If the algorithm makes poor proposals then much of the compute time of the algorithm is wasted when it could be touring the parameter space collecting more samples instead of rejecting proposals.

Many of the extensions of the Metropolis algorithm do try to overcome this by having a tunable step-size parameter. While it does help in some cases, this step-size parameter leads to a trade-off between improving the acceptance rate of proposals at the cost of exploring the parameter space and vice versa. A smaller step-size will improve the acceptance rate of proposals and will lead to more samples being accepted and thus more efficient sampling; however, this comes at the cost of not being able to explore or tour the full parameter space as efficiently and thus more samples are needed to get a representative sample of the parameter space. These small steps from one proposal to the next will often result in the samples staying in the same area and often “re-exploring” the same areas opposed to exploring the full parameter space. Increasing the step-size will improve the exploration but will come at the cost of a lower acceptance rate of proposals as proposals will more often be from low probability areas of the distribution. Furthermore, as the dimensionality of the parameter space increases so too does the concentration of measure which only exacerbates the challenge of efficiently exploring the parameter space of the typical set for these algorithms. Figure 3.2 illustrates the idealized scenario for MCMC algorithms.

The fundamental issue with guess and check algorithms is that proposals are generally bad when they are random and don’t know anything about the target distribution. This issue is further exacerbated when trying to estimate distributions that have high-dimensional parameter spaces, because of the previously mentioned phenomenon concentration of measure [7] [14]. This makes the random proposals from “guess and check” methods increasingly inefficient and ultimately poor estimators of distributions with many parameters. To overcome this researchers have turned to creating algorithms that try to incorporate more information from the target distribution when making proposals in order to explore the typical set more efficiently. The current state of the art is known as Hamiltonian Monte Carlo and is the method employed in this thesis to fit Bayesian models.

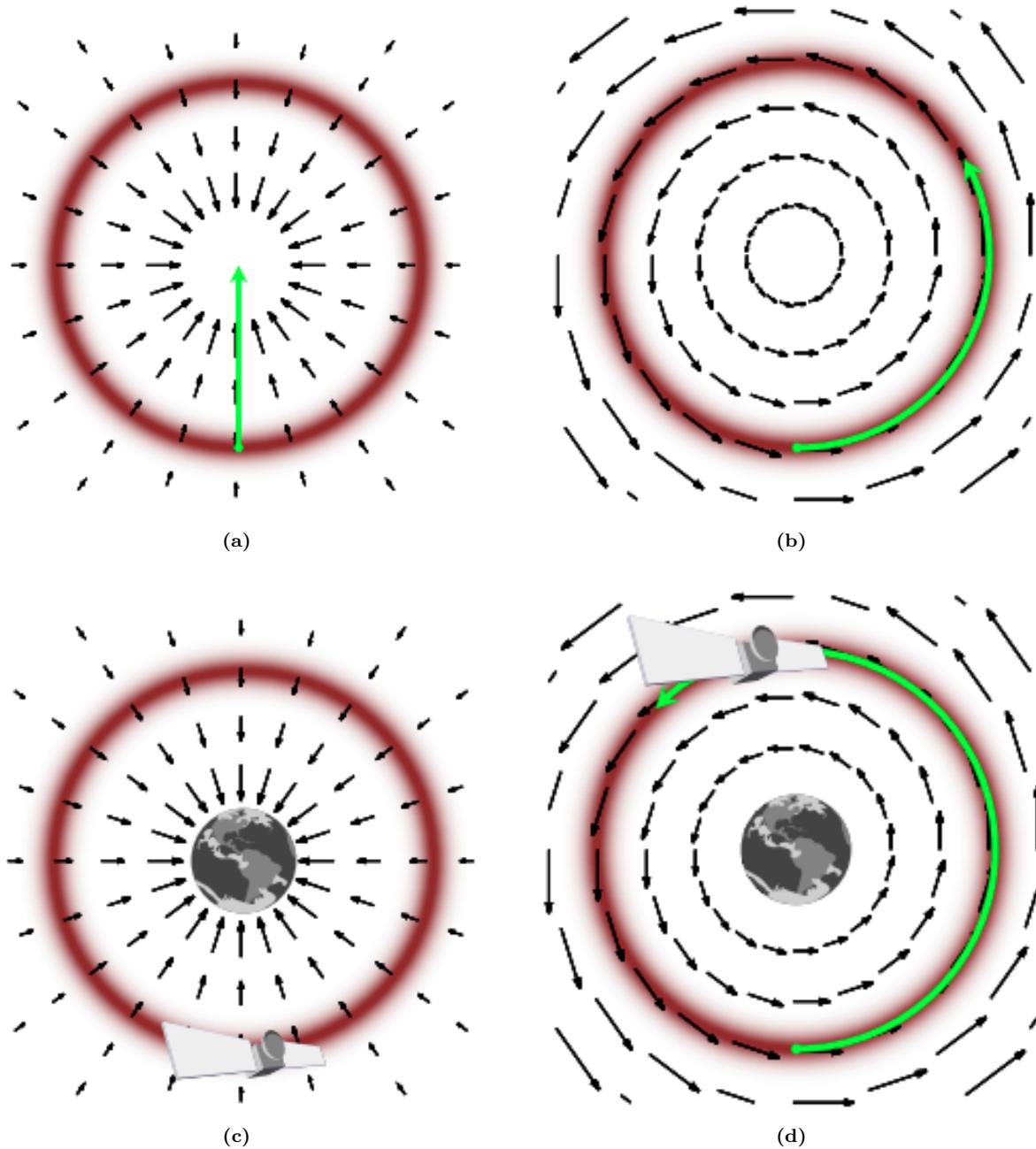


Figure 3.3: The gradient and corresponding vector field of a probability distribution points to its mode which is often away from the typical set in higher dimensions (a). Ideally we want to twist the vector field to align with the typical set (b). The mode, gradient, and typical set of a probabilistic system are mathematically equivalent to a planet, gravitational field, and orbit in a physical system (c). Adding momentum to the system to cause a satellite to enter a stable orbit (d) is equivalent to twisting a vector field to align with the typical set of a probabilistic system. Images are from Figures 12, 13, 14, 17 of [7]. Permission to use was granted by the author under a CC BY-NC 4.0 license (<https://creativecommons.org/licenses/by-nc/4.0/>).

Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) exploits information about the geometry of the typical set to greatly improve the efficiency at accurately sampling the parameter space of the target distribution. The insight of HMC is that for every probabilistic system there is a mathematically equivalent *physical* system, with equivalent differential geometry, about which we can reason and solve for in the exact same way a physicist would compute the conservative dynamics of a physical system via phase space and Hamilton's equations [7]; hence the name Hamiltonian Monte Carlo.

HMC works by exploiting the *gradient* of the target probability density function. That gradient defines a vector field that we can manipulate to be aligned with the typical set. Then we can follow this vector field in order to explore and sample from the typical set more efficiently. By itself, the gradient of the target probability density function points towards the mode of the distribution, and thus away from the typical set. Additional structure is required to twist the vector field generated by the gradient into a vector field aligned with the typical set. This additional structure can be thought of as adding *momentum* in such a way as to keep the corresponding dynamics of the system *conservative*. That is to say that the conservative dynamics of the physical system requires volumes to be preserved in accordance with Hamilton's equations. A rigorous derivation and exposition of conservative dynamics and Hamilton's equations is beyond the scope of this thesis but can be found in [7]. Here we give an intuitive explanation of how conservative dynamics in physical systems works and how it relates to the probabilistic systems considered in this thesis. The intuitive relation between a probabilistic system and a physical system is illustrated in Figure 3.3.

Intuitively, a mode, a gradient, and a typical set in a probabilistic system can be equivalently related to a planet, a gravitational field, and an orbit in a physical system. Exploring this physical system with a satellite is mathematically equivalent to exploring and sampling our probabilistic system. A satellite at rest will fall to the planet due to the planet's gravitational pull. Adding momentum to the satellite allows it to enter a stable orbit and not be pulled into the planet. However, adding too much momentum causes the satellite to leave the stable orbit and fly out to the depths of space. Conversely, adding too little momentum causes the satellite to again be pulled into the planet. Adding just the right amount of momentum to the satellite for it to remain in a stable orbit is the mathematical equivalent of the corresponding dynamics of the system remaining conservative, and is computed by ensuring the preservation of volume in position-momentum phase space [7]. For our purposes, the above analogy means that the same mathematics used to compute how much momentum to add to a physical system in order to ensure the corresponding dynamics are conservative (i.e. putting a satellite into a stable orbit) can be used to twist the gradient of a target probability density function and its vector field into one that corresponds to the typical set. We can then make proposals by taking steps proportionally random to the vector field that follows the typical set. This will ensure that our sample proposals will be attracted toward the typical set, and will then stay in and efficiently explore the typical set.

For this thesis we make use of the Probabilistic Programming library PyMC3 [50] and its implementation

of HMC to fit our models. PyMC3 and our use of it to build and fit our models is explored more in 5.1.1.

3.4 Model Evaluation and Selection

It is not enough to simply create a model and fit it to a dataset. There are infinitely many models that could be created and fit to a dataset, and some models will be better than others depending on the context or objective of the model. Thus, it is important we evaluate how well a given model fits a dataset and gauge its predictive performance. This allows us to understand how effectively a given model fits our dataset and it gives us a way to select the “best” model from among several models.

Model selection in a Bayesian context separates itself from more traditional null hypothesis testing by considering the existence of many possible models rather than assuming one model and evaluating its likelihood. Null hypothesis testing has a single (null) model and seeks data such that the model can be judged as sufficiently likely or unlikely. In contrast, Bayesian model selection assumes that there exists many potential models that could have generated the dataset we have, and instead tries to reason about which of those models was more likely to have produced the dataset. Thus we need tools to compare models in order to select the “best fitting” one, or the one that was most likely to have produced the data. This section describes how we evaluate the efficacy of Bayesian models via cross validation, information theory, and posterior predictive checks.

Cross Validation

It is natural to desire a model that fits the dataset as well as possible. However, it is possible to create models that fit a specific dataset so well that they fail to generalize to the larger population of which the dataset is only a sample. When this happens we say that a model is “over-fit”. While fitting a given dataset as well as possible seems desirable, it often comes at the cost of the model only fitting and retrodicting the dataset it was trained on and then performing much worse on unseen data or future scenarios for which the model was ideally created for. Thus it is important that we do not evaluate a model only on the basis of how well it fits and retrodicts the dataset it was trained on, but that we try to estimate how well the model fits the larger population and its ability to predict unseen data that it was not trained on.

Because a models performance on unseen data is more desirable than its performance on the data it was trained on, researchers have developed methods that actually make a model fit worse to the data it was trained on so that it fits unseen data better. This counter-intuitive idea is known as *regularization* and is a vast area of research in statistical inference and machine learning. Bayesian models perform regularization through the use of priors and through a process in multilevel modelling known as shrinkage to the mean which has previously been discussed in section 3.2. In order to see the effect of regularization and to compare various models we need a method of model evaluation. This section explores how we evaluate models through estimating their evaluation on unseen data.

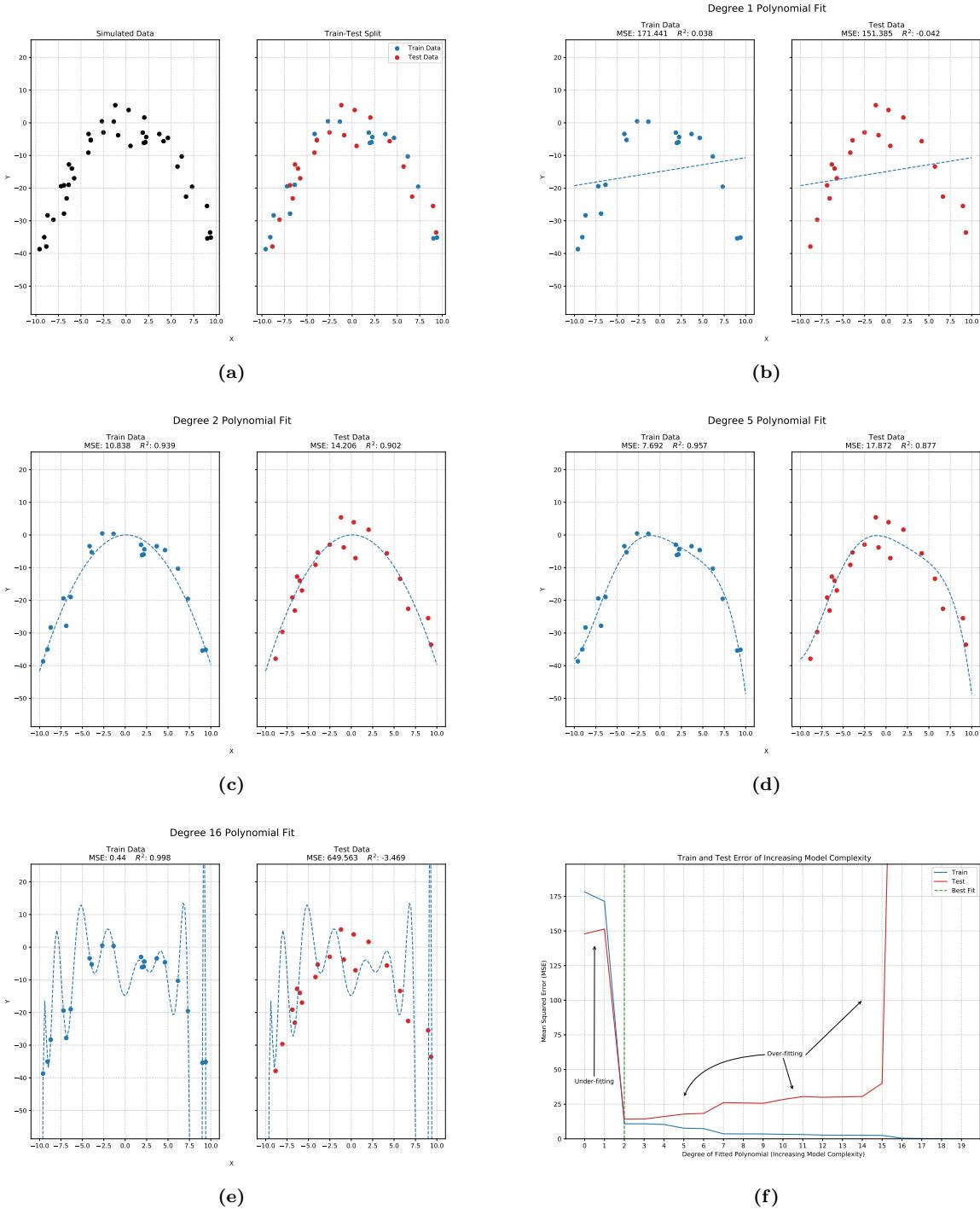


Figure 3.4: Example of how increasing model complexity leads to better model fit on the train-set, but can come at the cost of increasingly worse performance on the test-set. Model fit here is measured visually and in terms of mean-squared-error (MSE: lower is better) and R-squared (R^2 : closer to 1 is better). The dataset in (a) is generated by a degree-2 polynomial with some added noise and is split into train and test sets. A degree-1 polynomial underfits the data (b). More complex polynomials improve the fit on the train-set (c, d, e). However, increasingly complex polynomials become overfit as evidenced by increasingly worse test-set performance (d, e). The overall trend of increasing model complexity, how it relates to underfitting and overfitting, and where the tradeoff is optimal is captured in (f).

The simplest way to approximate how well a model will perform on unseen data is to “hold-out” a portion of your dataset referred to as a *test-set*, fit your model to the rest of the dataset referred to as the *train-set*, and then check the fit and predictive performance of the model on the test-set. The objective is to create a model that has the best fit and predictive performance on the test-set rather than the train-set. If a model performs notably worse on the test-set, then you conclude the model is likely over-fit and you potentially reject it even though it may be the highest performing model on the train-set. This process is illustrated in Figure 3.4. While this method is straightforward and generally effective it does have some drawbacks. The size of the train-set used to train the model is now smaller and may be too small to accurately reflect the model if it were trained on a larger dataset. The selection of which data is divided into the train and test sets may also bias the model and thus bias the estimate of its test-set evaluation. For example, if an essential group or cluster of similar data points are all put into the test-set then the models poor performance on the test-set could be misleading. Cross-validation is an attempt to alleviate these concerns and improve upon this method.

Cross-validation partitions the dataset randomly into K -many sets, with $2 \leq K \leq N$, where N is the size of the dataset. It then uses one of the sets as the test-set, and combines the rest into the train-set. The model is fit on the train-set and then evaluated on the test-set. This process is then repeated using each of the K -many sets as the train and test splits and the average performance across all test-sets becomes the estimate for out-of-sample performance. In this way the entire dataset is used in both training and testing, helping to alleviate small data or biased train-test splitting concerns.

As you increase the value of K you also increase the size of the train-set which gives not only a better model fit but is more likely to over-fit if the model itself is prone to over-fitting, something we desire to find out. However, increasing the value of K also means you need to train and test the model more times. This is most noticeable when considering the extreme case where K is the size of the dataset, known as leave-one-out cross-validation (LOO-CV). In this case you train the model on all but one data point and then test on the one data point that was left out, and repeat for each data point. While this gives the best estimate of out of sample performance, it requires you to train the model a large number of times. In our modern big-data era, where datasets often number in the thousands or more, this can become computationally infeasible. A common way to deal with this is to reach a compromise by using a smaller value of K , usually 5 or 10.

While measuring the out-of-sample predictive performance of a model is of highest importance in evaluating a model, we have not discussed the specific measure itself. The example in Figure 3.4 uses traditional measures of model fit based on point-predictions the models make, but Bayesian models produce entire distributions of estimates not just single point-predictions. Using a single point-prediction from an entire distribution greatly reduces and misses the vast majority of information that the entire distribution represents. Applied Bayesian statisticians have instead turned to the field of information theory to measure differences in distributions rather than point-estimates, and have tied these theories to the approximation of estimating LOO-CV performance without requiring refitting the model more times than is practical. We now discuss

the development of these techniques as they are the primary way we evaluate models in this thesis.

Information Criterion

Information theory is a field of mathematics concerned with representing data in a compact fashion (i.e. data compression) as well as transmitting data over a noisy channel and storing it in a way that is robust to errors [43]. Intuitively, quantifying information is viewed as measuring how much surprise there is in an event, where surprise relates to how likely or probable an event is. Thus, a surprising (lower probability) event is one that gives us more information than an unsurprising (higher probability) event. Claude Shannon first formalized these ideas in his foundational work on information theory [53] where he conceived of measuring information as the number of binary digits (bits) required to represent an event (or distribution). He extended this notion from discrete bits to a theoretical measure of a continuous amount of bits. In his work information is formally defined mathematically as information entropy:

$$H(p) = -E\log(p_i) = -\sum_{i=1}^n p_i \log(p_i) \quad (3.6)$$

where p_i is the probability of event i (or the i -th data point) occurring according to probability distribution p . In this way, information entropy can be thought of as measuring the uncertainty contained in a probability distribution as the average log-probability of the events (i.e. the data) [37]. We can then use this concept of information to reason mathematically about how much one probability distribution differs from another with respect to a dataset. Specifically, we can compute the average number of extra bits required to represent the data when using our models distribution as compared to the true distribution. This is captured mathematically in the form of Kullback-Leibler divergence (KL divergence) defined as:

$$D_{KL}(p, q) = \sum_i p_i (\log(p_i) - \log(q_i)) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right) \quad (3.7)$$

The KL divergence can be thought of as the average difference in log probability between the true distribution (p) and our models distribution (q) [37]. It gives us a way to compare how similar two distributions are. Note that it does not satisfy some specific mathematical constraints, in particular it is generally not symmetric ($D_{KL}(p, q) \neq D_{KL}(q, p)$), which means KL divergence is not a “measure” or a “distance” in the strict mathematical sense. KL divergence instead represents how much effort is needed to turn one distribution into another, or how much one distribution diverges from another. It is useful because it gives us a rigorous way to compare how similar two distributions are; however, there is one glaring issue of practical significance. The issue is that we do not actually know what the true distribution is (p) and therefore we are left with approximating the divergence. It turns out that the true distribution (p) is not entirely necessary for comparing models because it is just an additive term which means you can compute the relative difference between models without actually knowing it [37]. Thus, we can use the sum of log probabilities of each

observation known as the *log-score*:

$$S(q) = \sum_i \log(q_i) \quad (3.8)$$

Relative differences in log-scores will match relative differences in KL divergence [37]. This means that we can compare models via their relative log-scores even though the magnitude of the log-score is not easily interpretable. In practice, for Bayesian modelling, we need to average over the posterior distribution (i.e. we need to consider the entire distribution of possible model parameters proportional to how likely they are) and compute a Bayesian log-score for a given dataset y and model parameters θ_s known as the *log-pointwise-predictive-density (lppd)*:

$$lppd(y, \theta) = \sum_i \log \left(\frac{1}{S} \sum_s p(y_i | \theta_s) \right) \quad (3.9)$$

The log-score has traditionally been scaled by -2 and referred to as the *deviance*. This is because it has been shown that under general conditions and for many model types a difference between two deviances has a chi-squared distribution [20]. Thus, the factor of -2 would be there to scale the log-score to have this property to make more traditional computations such as likelihood ratio tests more convenient and interpretable. Modern users now tend to use just the log-score, or lppd, itself especially in a Bayesian context where traditional methods such as likelihood ratio tests are generally not used as often.

We have derived a way of measuring statistical distance between a models distribution and the target distribution via information theory through the rigorously defined KL divergence. This statistical distance gives us a rigorous way of comparing models by evaluating which models distribution is closest to the target distribution. In practice we can not compute KL divergence precisely but instead approximate relative differences in KL divergence via the log-score, or equivalently the lppd in the Bayesian context. While the log-score is a way to measure the distance of our model from its target, it has the same flaw that nearly all methods for evaluating models have: *the log-score will generally always improve as the model becomes more complex and thus has the potential to over-fit the dataset*. Researchers have developed a method of evaluation known as *information criteria* which essentially adds a penalty term to the log-score in order to account for increasing model complexity. Thus, as increasingly complex models will generally have a better (higher) log-score, they will need to improve the log-score by more than the penalty otherwise it is likely they are fitting to random noise and not actually improving general model fit. This adjusted log-score can then be used to determine the best fitting model.

The first notable contribution to the development of information criterion was made by Japanese statistician Hirotugu Akaike [2]. The insight was that for each additional parameter added to a regression model, the test-set deviance becomes worse by a factor about twice the number of parameters. Akaike then showed that this is the case under some general conditions, and derived what he called *an information criteria* which

is now more commonly referred to as Akaike Information Criterion (AIC):

$$AIC = D_{train} + 2k \approx E(D_{test}) \quad (3.10)$$

where D_{train} is the deviance on the train-set, k is the number of parameters, and $E(D_{test})$ is the expectation of the deviance on the test-set. Sumio Watanabe then generalized AIC to apply to a wider class of models and conditions to develop the Widely Applicable Information Criterion (WAIC) [60], sometimes referred to as Watanabe Akaike Information Criterion:

$$WAIC(y, \theta) = -2 \left(lppd - \sum_i var_\theta(\log p(y_i | \theta)) \right) \quad (3.11)$$

where y is the dataset and θ are the parameters of the model we are evaluating. The $lppd$ is the same as defined in equation 3.9, and $\sum_i var_\theta(\log p(y_i | \theta))$ is the penalty term which is an estimate of the sum of the variances of the log-likelihood for each observation (i) in the dataset. The penalty term is computed by computing the log-likelihood for each sample of parameters from the posterior distribution for the same observation y_i and computing the variance of these log-likelihoods; then repeating this variance computation for each observation y_i in the dataset and summing to yield the penalty term. The -2 transforms the $lppd$ into a deviance measure as deviance was used for AIC rather than log-score. The negative sign before the 2 turns the $lppd$ into a positive measure that we wish to minimize instead of maximize, and then the penalty term is a positive value that a more complex model needs to “overcome” in order to justify its use.

The WAIC provides a method for estimating the relative KL divergence by estimating the out-of-sample deviance by computing the negative $lppd$ and adding a generalized penalty term that will make the estimate worse (larger) the more complex the model is. WAIC and information criterion in general attempt to estimate the out-of-sample deviance. More recent work attempts to bridge the gap between out-of-sample deviance and cross-validation resulting in the current state of the art method for estimating out-of-sample predictive fit that we use for model evaluation in this thesis.

Pareto-Smoothed Importance Sampling

Vethari et al. [59] introduced an efficient computation for estimating LOO-CV from MCMC samples without requiring the repeated re-fitting of the model, referred to as Pareto-Smoothed Importance-Sampling Leave-One-Out cross-validation (PSIS-LOO). Their method also computes the $lppd$, but instead of adjusting it with the penalty term as in 3.11, they use importance sampling to re-weight the log-score of the $lppd$ to more accurately reflect what the log-score for each data point would have been if that point had not been used to fit the model, hence the estimation of LOO-CV by re-weighting rather than re-fitting the model. The insight of Vehtari et al. is that the weights needed in order to re-weight the posterior samples for a given data point (e.g. the weights w_s for re-weighting data point y_1 : $\frac{1}{\sum_s w_s} \sum_s w_s p(y_1 | \theta_s)$) in a manner that resembles what the probability would have been if that data point had not actually been observed in the

dataset the model was trained on, turn out to be the inverse of the probability that the posterior draw gave to the held out data point (e.g. the weights for y_1 are $w_s = \frac{1}{p(y_1|\theta_s)}$). Vehtari et al. further improved the stability of the importance sampling estimates by showing the upper tail of these importance weights fit a generalized Pareto distribution. Instead of using the estimated importance weights themselves, they instead use the weights to fit a generalized Pareto distribution and then use the weights implied by the fitted Pareto distribution resulting in Pareto-smoothed weights. PSIS-LOO is thusly computed as:

$$PSIS - LOO(y, \theta) = \sum_i \log \left(\frac{1}{\sum_s w_s^i} \sum_s w_s^i p(y_i|\theta_s) \right) \quad (3.12)$$

where w_s^i is the Pareto-smoothed importance weight for data point y_i with sampled parameters θ_s . The rest is the same as in 3.9 with the key change being the re-weighting performed by multiplying by the weights (w_s^i) and re-averaging proportional to the re-weighting ($\sum_s w_s^i$).

The PSIS-LOO estimate is also an estimate of the out-of-sample relative KL divergence. As a result, the PSIS-LOO estimates are often similar to WAIC estimates in practice, but were shown to be more stable and consistent than WAIC estimates [59]. The models that are fit and compared in this thesis all gave near identical results for both PSIS-LOO and WAIC. We have chosen to use just PSIS-LOO as it has become the modern standard.

Posterior Predictive Checks

Posterior predictive checks (PPCs) give us a way to visually check how well a model fits the dataset. Since Bayesian models are generative, we can use the fitted model to simulate values and then observe how closely these generated values resemble the observed values of the dataset. If the distribution of simulated values closely resembles the observed dataset then we say that the model is “well-specified” or is a good fit to the dataset. If the distribution of simulated values does not closely resemble the observed dataset then we say that the model is “misspecified” or is a bad fit to the dataset. It is notable that PPCs do not only reveal that a model is potentially misspecified but often also give insight into how to potentially improve the model by visually seeing for which data points the model is struggling to fit well. In this way PPCs are not only useful for model evaluation but for model building as well. PPCs informed the models chosen in the second data experiment discussed in section 5.2.3.

4 Related Work on Sports Analytics

The initially most well known and cited work on home advantage in sports was done in 1977 by Schwartz and Barsky [52] who analysed and found home advantage to exist in professional hockey, basketball, baseball and football. In [17] 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, behavioural states, and performance outcomes to try to understand the underlying causes of home advantage. Follow up work a decade later by Carron et al. [15] 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 [48] [30] 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 generally 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 analysing home advantage in the context of COVID-19 adjusted seasons; however, nearly all of them have focused exclusively on European Soccer leagues. In [6] 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 home advantage during the COVID-19 adjusted seasons, with the other four reporting mixed results where home advantage 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 [39] 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.. 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 [10] and Bayesian state-space models [29]. Further advancements and examples

from the NHL, NBA, NFL, and MLB are comprehensively summarized in [36] and follow a form similar to the model in [4] 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 [3] while better capturing the uncertainty in estimating parameters opposed to classical point estimates and p-values which are increasingly under criticism in modern science [34] [5]. 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 [6]. In this paper we aim to take the first attempt to use these methods to infer home advantage during the COVID-19 adjusted seasons of the NHL, NBA, NFL, and MLB.

In [36] the authors show the improved efficacy of the Poisson distribution instead of the more common Normal distribution [3] for modelling points scored by each team in each game. In [6] the authors follow the work in [35] 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 [4] there is no need of the bivariate Poisson when working within the Bayesian framework because multilevel (sometimes referred to as 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 [4] the authors argue for more complex methods to limit the shrinkage of their multilevel model as their data was from leagues with a large range of team strengths. We follow [36] 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 multilevel model.

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 often do not face each other 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 [48] 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 we argue that they are most appropriate for our focus of analysing and inferring home advantage.

5 Methods

In this chapter we define the multilevel regression model used throughout the rest of this thesis. We also set up and describe three data experiments that are the main contributions of this thesis. The primary goal of our model is to infer home advantage prior to and during the COVID-19 pandemic, thus we develop a model that has a home advantage parameter that we allow to vary across seasons whilst accounting for relative differences in teams offensive and defensive strengths. To motivate why the model is built the way it is we also conduct experiments to test the efficacy of multilevel regression modelling as well as the use of the Negative Binomial distribution opposed to the Poisson and Normal distributions more commonly used in regression analyses of sports data.

5.1 Multilevel Model

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, such as offensive and defensive team strengths for each team, which split the data into smaller groups from which we estimate the model parameters. 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 [4] [29] [36] [6] that allows for pooling results across all teams to infer home advantage. The pooling occurs specifically for the parameters that represent offensive and defensive team ratings. 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 while preventing over-fitting by adjusting parameter estimates through a process commonly referred to as “shrinkage to the mean” [26] [27] [37] as discussed in

section 3.2. We argue the pooling of data across each teams results to better handle smaller sample sizes while preventing over-fitting, 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:

$$y_{ij} | \mu_{ij}, \alpha_{ij} \sim \text{NegativeBinomial}(\mu_{ij}, \alpha_{ij}) \quad (5.1)$$

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:

$$\begin{aligned} \log(\mu_{i1}) &= \gamma_{sp} + \beta_{sp} + \omega_{sh[i]} + \delta_{sa[i]} \\ \log(\mu_{i0}) &= \gamma_{sp} + \omega_{sa[i]} + \delta_{sh[i]} \end{aligned} \quad (5.2)$$

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. To instead model the situation where all previous seasons are combined and compared to the one COVID-19 adjusted season (see Figure 6.4), s is instead a binary indicator with $s = 0$ indicating all previous combined seasons and $s = 1$ indicating the COVID-19 adjusted season. The subscript p indicates regular season ($p = 0$) or playoffs ($p = 1$). Home advantage is represented by β_{sp} with s and p the same as the intercept. The offensive and defensive strength of the home and away 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 [37] [26] [27]. 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 [4]. Note that the offensive and defensive strengths represented this way could potentially lead to problems of identifiability suggested by previous works [4] [6] [35] and fully described in [37]. The issue is that a given difference in relative team strengths can be solved by multiple different team ratings, similar to a system of equations being singular. In line with previous works [4] [6] [35], we force the offensive and defensive ratings across all T teams within each league to sum to zero for each season:

$$\sum_{t=1}^T \omega_{st} = 0, \sum_{t=1}^T \delta_{st} = 0 \quad (5.3)$$

Not only does this make identifiability a non-issue, but it also improves interpretability of the fitted team ratings as zero represents an average team rating with stronger and weaker ratings being correspondingly above or below zero. Also note that in this formulation defensive ratings are strong (weak) in the negative (positive) direction. This can be seen by considering how a negative defensive rating decreases the expected number of points in 5.2, thus it represents a strong defensive team. Offensive ratings are the opposite by being strong (weak) in the positive (negative) direction.

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 6.4 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 6.5 which reveal some interesting differences as discussed in the Results section.

In (5.2) 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 we perform a separate model fit for each league. This is because each league varies greatly in their respective point totals which is explored further in the second experiment in section 5.2.3.

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:

$$\begin{aligned} \log(\mu_{i1}) &= \gamma_{sp} + \beta_{sp} + \omega_{sh[i]} + \delta_{sa[i]} \\ \mu_{i1} &= \exp(\gamma_{sp} + \beta_{sp} + \omega_{sh[i]} + \delta_{sa[i]}) \\ \mu_{i1} &= \exp(\gamma_{sp}) * \exp(\beta_{sp}) * \exp(\omega_{sh[i]}) * \exp(\delta_{sa[i]}) \end{aligned} \tag{5.4}$$

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.

5.1.1 Model Fit in PyMC3

The models are fit using PyMC3, an open source probabilistic programming language (PPL) that allows us to fit Bayesian models with their implementation of a gradient based Hamiltonian Monte Carlo (HMC) No

```

 $y_{ij} | \mu_{ij}, \alpha \sim \text{NegativeBinomial}(\mu_{ij}, \alpha)$ 
 $\log(\mu_{i1}) = \gamma + \beta + \omega_h[i] + \delta_a[i]$ 
 $\log(\mu_{i0}) = \gamma + \omega_a[i] + \delta_h[i]$ 

 $\gamma \sim \text{Normal}(\theta^*, \sigma^{2*})$ 
 $\beta \sim \text{Normal}(0, 1)$ 
 $\alpha \sim \text{Uniform}(0, 1000)$ 
 $\omega \sim \text{Normal}(0, \sigma_\omega)$ 
 $\delta \sim \text{Normal}(0, \sigma_\delta)$ 
 $\sigma_\omega \sim \text{HalfNormal}(1)$ 
 $\sigma_\delta \sim \text{HalfNormal}(1)$ 

```

```

import pymc3 as pm, theano.tensor as tt
with pm.Model() as model:
    # priors
    sigma_o = pm.HalfNormal('sigma_o', sigma=1)
    sigma_d = pm.HalfNormal('sigma_d', sigma=1)
    omega = pm.Normal('omega', mu=0, sigma=sigma_o, shape=num_teams)
    delta = pm.Normal('delta', mu=0, sigma=sigma_d, shape=num_teams)
    beta = pm.Normal('beta', mu=0, sigma=1)
    alpha = pm.Uniform('alpha', 0, 1000)
    gamma = pm.Normal('gamma', mu=np.log(avg_points), sigma=np.log(avg_points_sd))
    # likelihood
    home_mu = tt.exp(gamma + beta + omega[df.home_team] + delta[df.away_team])
    away_mu = tt.exp(gamma + omega[df.away_team] + delta[df.home_team])
    home_points = pm.NegativeBinomial('home_points', mu=home_mu, alpha=alpha, observed=df.home_points)
    away_points = pm.NegativeBinomial('away_points', mu=away_mu, alpha=alpha, observed=df.away_points)
    # call the sampler (Hamiltonian Monte Carlo)
trace = pm.sample(2000, tune=1000, cores=4)

```

(a)

(b)

Figure 5.1: An example of how a Bayesian model would be defined in an academic textbook or paper (a) and how a probabilistic programming language such as PyMC3 would create the model in Python code (b). Note how the priors have to be defined first because the code will be executed procedurally. The two definitions are essentially identical otherwise.

U-Turn Sampler (NUTS) [50]. PPLs are a tool for statistical modelling that try to bridge the gap between complex statistical definitions and easy to write code. They define a set of primitives for drawing random numbers and conditioning constructs. This enables defining random variables, how to sample them, storing their samples in memory, and how much weight to give them in a given execution of a program. For Bayesian modelling this means you can define all the variables in your model; whether they are priors, likelihoods, observed variables, or unobserved latent variables. Then you can sample all the variables in your model using any method supported by the PPL such as HMC. An example of how a model would be defined in an academic textbook or paper and then how it would be written in PyMC3 is shown in Figure 5.1. PPLs make defining and fitting models simpler and more accessible to non-experts. This allows for iteratively building and improving models as part of a workflow that would have been impossible for previous generations of scientists.

As in other previous work [4] [6], 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:

$$\begin{aligned}
\gamma_{sp} &\sim \mathcal{N}(\theta^*, \sigma^{2*}) \\
\beta_{sp} &\sim \mathcal{N}(0, 1) \\
\lambda &\sim \text{Uniform}(0, 1000) \\
\omega_{st} &\sim \mathcal{N}(0, \sigma_{s\omega}) \\
\delta_{st} &\sim \mathcal{N}(0, \sigma_{s\delta}) \\
\sigma_{s\omega} &\sim \text{HalfNormal}(1) \\
\sigma_{s\delta} &\sim \text{HalfNormal}(1)
\end{aligned} \tag{5.5}$$

where the offensive and defensive ratings for each team, ω_{st} and δ_{st} respectively for $t = 1, \dots, T$, are drawn from the same shared distributions, $\mathcal{N}(0, \sigma_{s\omega})$ and $\mathcal{N}(0, \sigma_{s\delta})$, which have their own priors, $\sigma_{s\omega}$ and $\sigma_{s\delta}$, known as hyperpriors. It is because each teams ratings are drawn from the respective same shared distribution that information is partially-pooled across teams while fitting the model. 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. PyMC3's implementation of the Negative Binomial distribution defines the parameter μ as the mean directly (a Poisson distribution parameter) and the parameter α as a Gamma distribution parameter [50] such that the variance is equal to $\mu(1 + \frac{\mu}{\alpha})$. Thus the variance will generally be greater than the mean except for when $\alpha \rightarrow \infty$. Since we have defined $\alpha = \mu * \lambda$, the variance is then equal to $\mu(1 + \frac{1}{\lambda})$. Therefore, we use a prior that allows λ to be small to model overdispersion while also allowing λ to potentially be large for instances where there is little to no overdispersion in the outcome variable and we instead want the Negative Binomial distribution to tend toward a Poisson distribution.

The model is fit using PyMC3's NUTS sampler using 4 chains of 2,000 iterations with 1,000 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 [25] [11]. Each model fit produced \hat{R} statistics of 1.00 with no divergences [7], meaning the samples converged, the model fit is valid, and we can analyze the resulting parameter estimates to draw our inferences from.

5.2 Experiments

In this section we describe how the datasets we used were curated. We then set up and describe the three data experiments that make the main contributions of the thesis. The results are shown and discussed in the Results chapter.

5.2.1 Data

For each league we gathered data from the five most recent seasons spanning the years 2016-2020, both regular season and playoffs. For our model, for each game, we need to track the teams that are playing, which teams are home and away, their respective game point totals, which season the game occurred, and whether or not the game occurred in the playoffs or regular season.

The NHL data is sourced from Natural Stat Trick [44]. A typical NHL season consists of 82 games played by each team. Prior to the Vegas Golden Knights joining the league in 2017, there were 30 teams resulting in 1230 games per season. Since 2017 there are 1271 games played with 31 teams in the league. The playoffs consist of a bracket of 16 teams playing best-of-seven series, for an average of 80-90 games total. We note that the 2020 season was shortened to 1082 games due to stopping for the initial outbreak of the COVID-19

pandemic. The 2020 playoffs occurred inside the NHL bubble when play resumed, consisting of 6 games to determine positions 1-8 and 8 best-of-five series to determine positions 9-16 before beginning the usual playoff structure. This resulted in 129 games played in the NHL’s COVID-19 bubble.

The NBA data is sourced from the basketball-reference website [54]. The structure of the regular season and playoff schedules is similar to that of the NHL. A typical NBA season consists of 30 teams each playing 82 games for a total of 1230 games. The playoffs consist of a bracket of 16 teams playing best-of-seven series, for an average of 80-90 games total. Like the NHL, the 2020 NBA season was shortened to 971 games due to stopping for the initial outbreak of the COVID-19 pandemic. The 2020 playoffs occurred inside the NBA bubble when play resumed, consisting of 8 additional games for each of the top 22 teams to determine seeding of the top 16 teams before beginning the usual playoff structure. This resulted in 172 games played in the NBA’s COVID-19 bubble.

The MLB data is sourced from retrosheet [49]. A typical MLB season consists of 30 teams each playing 162 games for a total of 2430 games. The playoffs can be viewed as an 8 team bracket, but there are 4 “wildcard” teams that play two best-of-one games to determine the last two spots for the 8 teams that make the first round called the Division Series. The Division Series consists of best-of-five series to determine who moves on to the League Championship Series. The League Championship Series and the following World Series Championship consist of best-of-7 series to determine the winner. This playoff structure usually results in an average of 30-40 games. The 2020 COVID-19 restricted season reduced the number of scheduled games to 60 for each team. This change combined with cancellations due to outbreaks within teams reduced the total number of games to 898. The playoffs replaced the best-of-one wildcard round with best-of-three series involving all top 8 seeded teams. This resulted in a total of 52 playoff games. We note that the 2020 season saw some double-header games where teams switched home and away even though both games were played at the same stadium. We found this to have essentially no impact due these games making up a relatively small portion of total games (45/898) and to home advantage being so small in the MLB. We have reported the results with home and away defined as who batted last in each inning for all games.

The NFL data is sourced from the football-reference website [55]. A typical NFL season consists of 32 teams each playing 16 games for a total of 256 games. The playoffs usually consist of a bracket of the top 12 teams playing best-of-one games (the top 4 teams getting a first round “bye”) resulting in 11 games total. Although the 2020 season had restrictions on fan attendance, the regular season schedule did not change and the playoff set-up only slightly changed by expanding to consist of the top 14 teams (only the top 2 getting a first round “bye”) resulting in 13 games total. We exclude the Super Bowl as well as international site games from our analysis for consistency, as they are generally played at neutral sites and there are very few of them (i.e. 4-5 neutral site games out of a total 256 games each season).

5.2.2 Complete pooling, No pooling, and Partial pooling

Objective

In the Background on Bayesian Inference chapter we described the theory behind multilevel modelling and gave examples to motivate and explain why multilevel modelling is more effective than traditional regression modelling or simple averaging. For this experiment we want to explore how multilevel modelling performs on the datasets used in this thesis, as well as testing the efficacy of PSIS-LOO for estimating models out-of-sample predictive performance on sports data.

Models

We create and compare three models in order to show the benefits of how multilevel modelling partially pools information across teams to improve model fit while preventing over-fitting. The first model is referred to as a *completely pooled* model where the data from all teams is completely pooled into one overall average to be used. The completely pooled model adjusts the model described in section 5.1 by modifying equation 5.2 as follows:

$$\begin{aligned}\log(\mu_1) &= \gamma_{sp} + \beta_{sp} \\ \log(\mu_0) &= \gamma_{sp}\end{aligned}\tag{5.6}$$

This is the simplest regression model where an average number of points for home teams and away teams is calculated and then used for predictions.

The second model is referred to as a *no pooling* model where essentially a separate regression fit is made for each team's offensive and defensive ratings; ignoring the information from other teams. This is a traditional regression model and is defined near identically to the model described in section 5.1, with the only difference being that the team strength parameters are not pooled. In practice this means that the priors for the team strength parameters ω_s and δ_s described in equation 5.5 are changed to $\mathcal{N}(0, 1)$. This prevents information being pooled across groups, hence the name no pooling.

Finally the multilevel model is fit as described in section 5.1, which is referred to as a *partially pooled* model for the context of this experiment. The effect of this model is shrinking the team strength estimates from the no pooling model towards the overall mean. This effect was described in section 3.2 and in theory helps to prevent over-fitting but worsening the fit to the training dataset in order to improve the out-of-sample fit. This experiment is designed to test how this theory holds up on the sports datasets considered in this thesis.

In the completely pooled model there is essentially one global average used for all groups and in this way the model has high bias and ignores the differences amongst groups. In the no pooling model each group has its own parameter fit, but completely ignores the data from other groups and how they are fit resulting in lower bias and a better fit to the data but at the risk of over-fitting. The partially pooled model is a balance

between these two extremes allowing for a better model fit than the completely pooled model while better protecting against over-fitting than the no pooled model. The differences in team ratings from the no-pooled and partially-pooled models can be seen in the appendix.

Evaluation

To compare these models we randomly split each sports dataset in half to create a train-set and test-set. The train-set is used to train each model and evaluate training fit, and then the model will also be evaluated on the test-set in order to approximate the fit on unseen data. Models are evaluated by computing their log-score as defined in 3.8. We additionally compute the PSIS-LOO estimate for each model on the train-set to evaluate how well it approximates the log-score on the test-set. The same fitting and evaluating procedures are performed for each model on each dataset.

5.2.3 Negative Binomial Regression

Objective

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 analysing European football data [35] [4] [6]. 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 analyse in this paper. Table 6.1 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. The dispersion statistic 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 $1.2 < T < 2$ [46] [13]. Table 6.1 shows the NBA, MLB, and NFL having potential overdispersion in their point totals and thus, the Poisson distribution is likely inappropriate and less effective. This suggests the use of 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. This experiment is aimed at comparing and contrasting using the Negative Binomial distribution as the likelihood for our model opposed to the Poisson and Normal distributions that are more commonly used.

Models

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 (5.1) 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.

Evaluation

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 [59] we approximate LOO using Pareto-smoothed importance sampling (PSIS) and report the results in Table 6.1. We note here that we also used the widely-applicable information criterion (WAIC) [60] but found the results to be nearly identical and the conclusions the same.

5.2.4 Inferring Home Advantage

Objective

The primary goal of this thesis is to infer the potential effect of and change in home advantage in North American professional sports prior to and during COVID-19. After establishing the efficacy of our Negative Binomial multilevel regression model in the previous experiments, our final experiment is to fit our model to the datasets and then examine the distribution of the home advantage parameter and to discuss the implications.

Models

To make inferences about home advantage prior to and during the COVID-19 pandemic we fit the multilevel model previously described in section 5.1. The fitting of this model results in parameter estimates for home advantage across each of the four leagues analysed for four seasons prior to and one season during COVID-19.

Evaluation

We examine the distributions of the parameter estimates for home advantage that result from the model fit. We analyse the trends and differences across seasons as well as leagues in order to make inferences and draw conclusions about the impact of home advantage in these sports.

6 Results

This chapter presents the results of the data experiments, making up the main contributions of this thesis, introduced and explained in Methods chapter.

6.1 Complete pooling, No pooling, Partial pooling

The results of comparing models via their Log-Score, and how well PSIS-LOO is at estimating the out-of-sample performance, can be seen in Figure 6.1. The results show the same general trends across each league: 1) The complete-pooling model under-fits the data compared to the other models, but it also over-fits less compared to the other models as seen by its test-set performance not degrading as much, 2) The no-pooling model over-fits the data the most as it generally has the best performance on the train-set but never has the best performance on the test-set, 3) The partial-pooling model consistently has the best train-set performance and is therefore the best performing model.

These results confirm the advantageous theory behind multilevel modelling explored in the Background chapter. The multilevel model (partial-pooling model) consistently outperformed the complete-pooling model on both the train-set and the test-set. Interestingly the multilevel model consistently performed worse than the no-pooling model on the train-set but outperformed the no-pooling model on the test-sets. This is regularization at work and shows the efficacy of multilevel modelling over traditional regression. but fits the train-set worse than the no-pooling model. We note that we also performed this same experiment with a 75-25 train-test split which generated the same results as the 50-50 train-test split we are reporting here.

The PSIS-LOO estimates of out-of-sample performance consistently ranked the models in the correct order measured by test-set performance, despite only having the train-set available to make these estimates. This shows how effective PSIS-LOO is at estimating out-of-sample performance, why it has become the current state of the art for model evaluation in Bayesian statistics, and why we opt for using PSIS-LOO estimates for ranking models in this thesis. We note that the exact magnitude of the PSIS-LOO estimates were sometimes over or under estimated relative to the test-set results, but that for a given dataset they were consistently over or under estimated. This is interesting because one of the additional contributions of Vehtari et el. [59] was showing errors in PSIS-LOO estimates are highly correlated for the same dataset. Instead of more naive computations for the standard error, they instead derived an estimate for the standard error of the difference between models trained and evaluated on the same dataset. This new standard error generally provides tighter bounds that more accurately reflect the correlation in errors of PSIS-LOO estimates on the

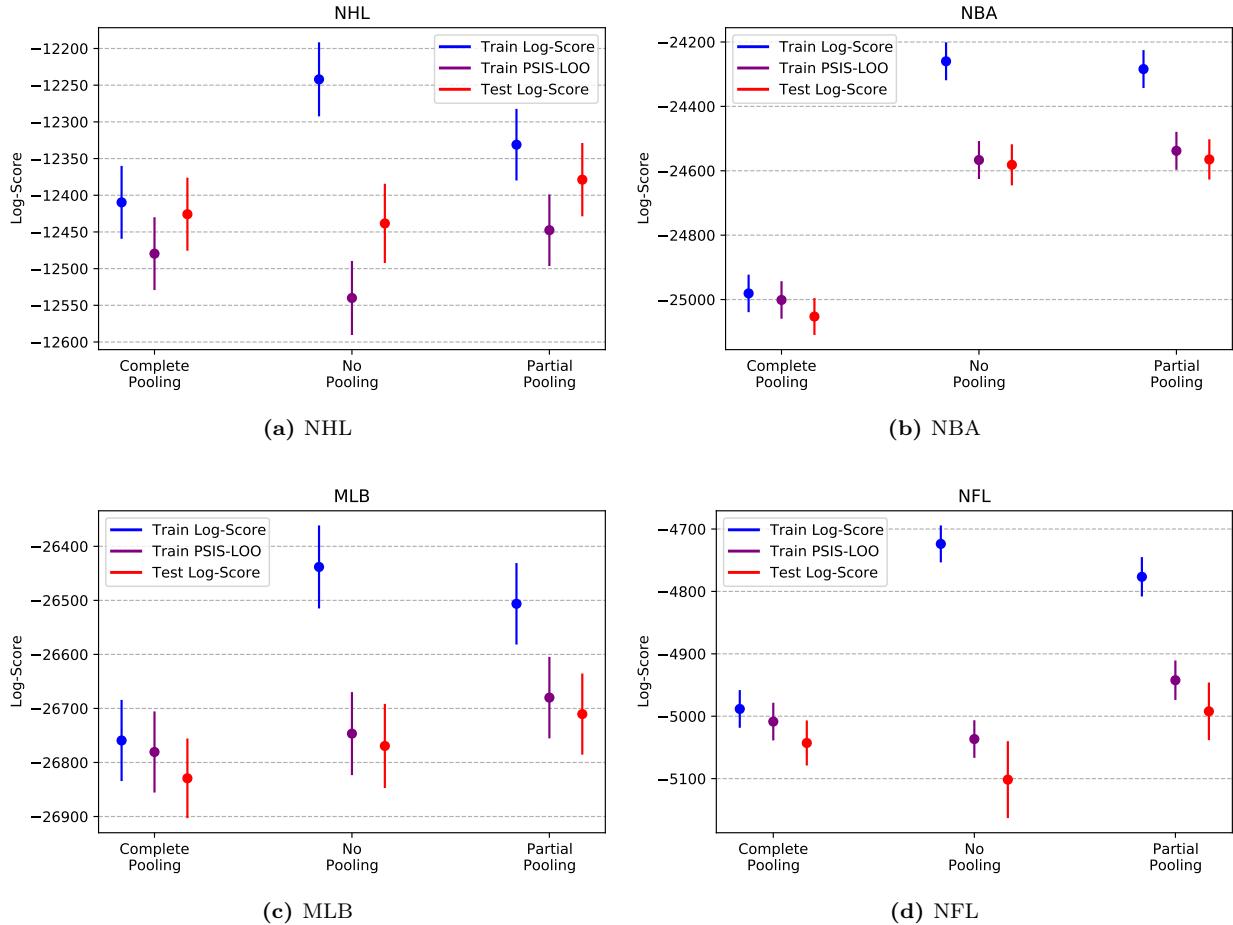


Figure 6.1: Comparison of models via their Log-Score (higher is better) on train and test sets, as well as the PSIS-LOO estimated Log-Score, for each league. The complete-pooling model under-fits, the no-pooling model over-fits, and the partial-pooling model provides the best trade-off in fitting the data while protecting against over-fitting. The PSIS-LOO estimates consistently predict how the models would rank on an unseen test-set.

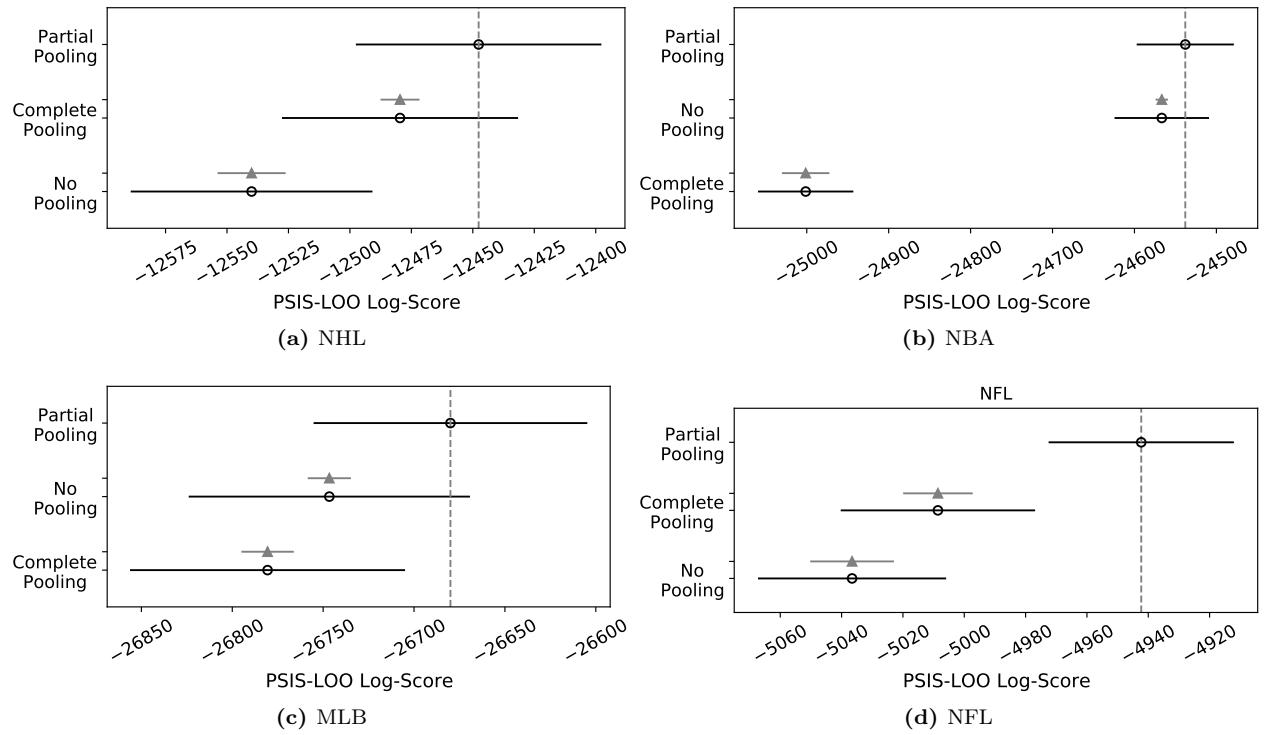


Figure 6.2: Comparison of models via their PSIS-LOO estimated Log-Score for each league, ranked from best (highest) to worst (lowest) on the y-axis. The black points and lines represent the point estimate and its standard error. The grey triangle and lines represent the estimated difference and the standard error of the difference for each model relative to the best model. The standard error of the difference is generally much smaller than the standard error of the estimate because errors in the estimates for each model are highly correlated.

same dataset. A graphical representation of this can be seen in Figure 6.2.

6.2 Negative Binomial Regression

The differences in model fit for the various likelihood distributions considered can be seen visually in Figure 6.3. Visually inspecting the distributions in Figure 6.3 is known as a Posterior Predictive Check (PPC) as described in section 3.4. The PPCs in Figure 6.3 are performed by plotting the distribution of observed home point totals in black along with 2000 sampled model fits in green for Poisson, blue for Negative Binomial, and red for Normal; with the respective mean model fits across the 2000 samples 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. The precise model comparisons are depicted in Table 6.1 and reveal the same patterns seen in the models PSIS-LOO estimated out-of-sample predictive fit. Because the point totals of the sports we are considering are positive integers prone to overdispersion and based on the results in Table 6.1 and Figure 6.3, we conclude that the Negative Binomial distribution is the most appropriate for regression modelling professional hockey, basketball, baseball, and American football.

6.3 Inferring Home Advantage

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 6.4 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 6.5 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 6.4. Nevertheless, the year-over-year estimates in Figure 6.5 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. A table of the estimated probabilities of the home advantage parameter being less than 0, the previous seasons grouped together, and the previous seasons individually can be seen in Table A.1. The remainder of this section examines these estimated distributions and their implications.

For the NHL and NBA data, Figures 6.4 and 6.5 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

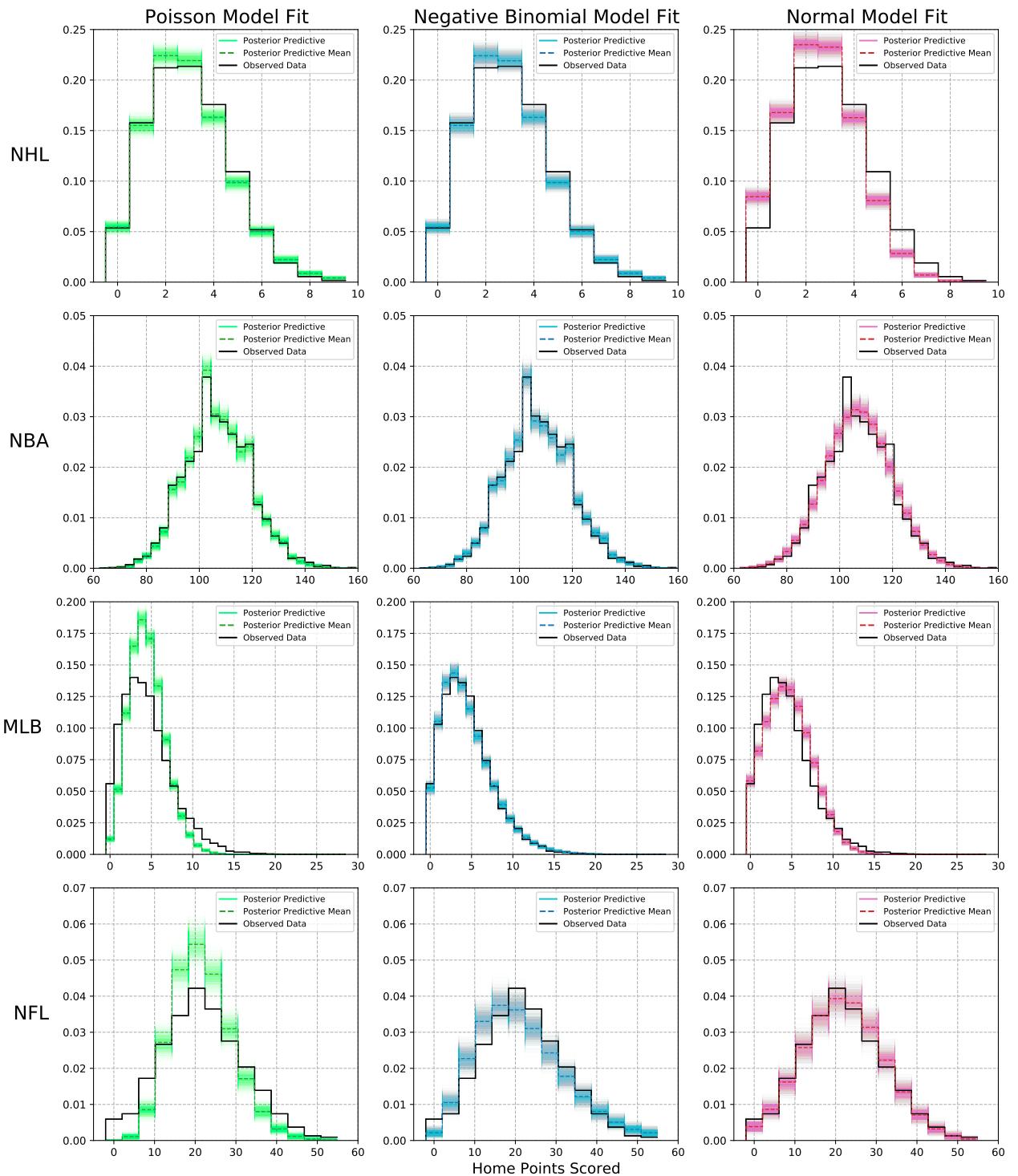


Figure 6.3: Comparison of distribution of home points in the models and the observed data for each league. The Negative Binomial model noticeably provides a better overall fit across each league.

	σ_p	Model	PSIS-LOO	dLOO	dSE
NHL	0.99	Poisson	-24761.3	-	-
		NB	-24761.5	0.2	0.2
		Normal	-25140.9	379.5	23.4
NBA	1.50	Poisson	-49018.3	53.5	11.0
		NB	-48964.8	-	-
		Normal	-48981.9	16.6	7.5
MLB	2.27	Poisson	-57458.7	4115.8	120.9
		NB	-53342.9	-	-
		Normal	-55696.8	2353.17	65.1
NFL	4.56	Poisson	-11751.2	2042.5	119.0
		NB	-9841.7	133.0	22.1
		Normal	-9708.7	-	-

Table 6.1: Comparison of estimated negative log-likelihood of leave-one-out cross-validation (LOO) for each model across each league. The differences between the Poisson, Negative Binomial (NB), and Normal models are reported relative to the best fitting model (dLOO) for each league; along with the standard error of the estimated differences (dSE). The dispersion statistic, σ_p , indicates how much greater the variance is than the mean for point totals in each league and signals overdispersion when $\sigma_p > 2$. The NB model noticeably outperforms the Poisson model for leagues with greater overdispersion (MLB and NFL) while being nearly identical for leagues with little to no overdispersion (NHL and NBA). The NB model also outperforms the Normal model in each league except the NFL where they are close to one another while both vastly outperforming the Poisson model.

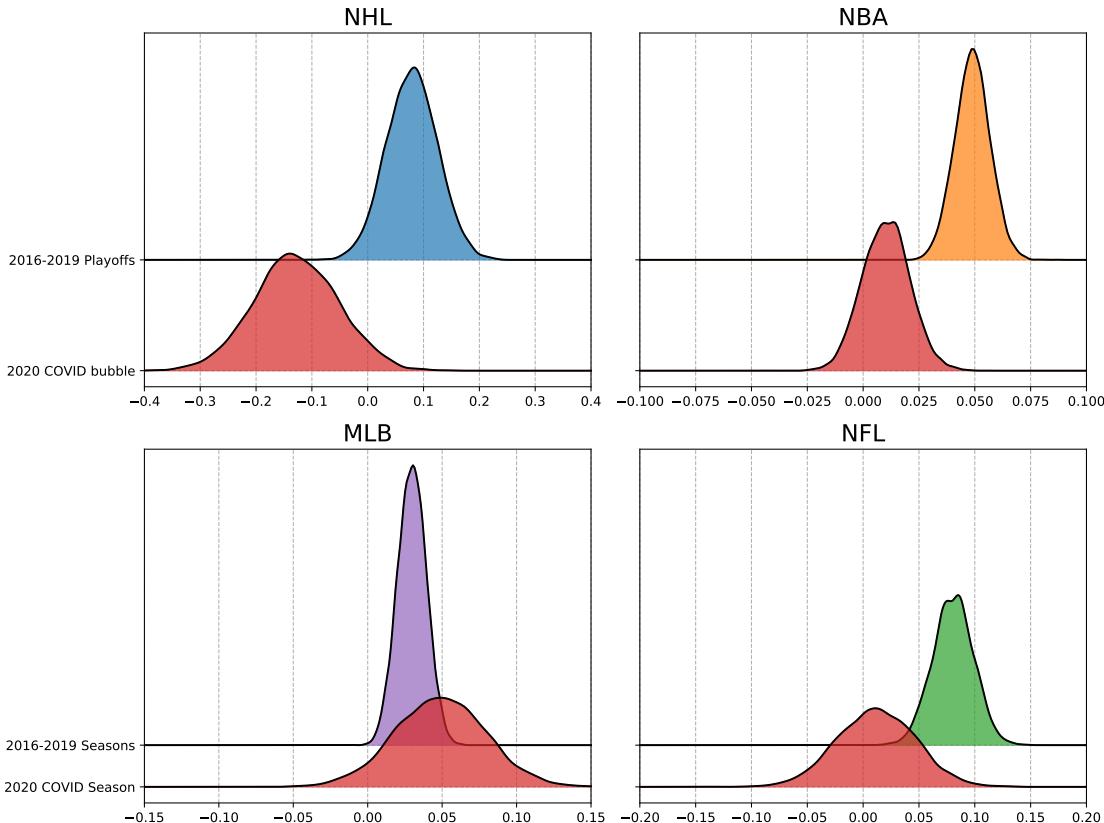


Figure 6.4: Distributions of the estimated home advantage for the NHL, NBA, MLB, and NFL for pre and post COVID adjusted seasons. Home advantage for playoffs are reported for NHL and NBA because that is when their COVID restricted games took place. Home advantage for regular season is reported for MLB and NFL as their respective playoff seasons are too small for stable results. Red distributions represent COVID-19 bubble adjusted seasons.

NFL had COVID-19 restrictions for their entire seasons, therefore, Figures 6.4 and 6.5, 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.

For the NHL data, the results in both Figures 6.4 and 6.5 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 6.4 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 6.5 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 6.5 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 6.4 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

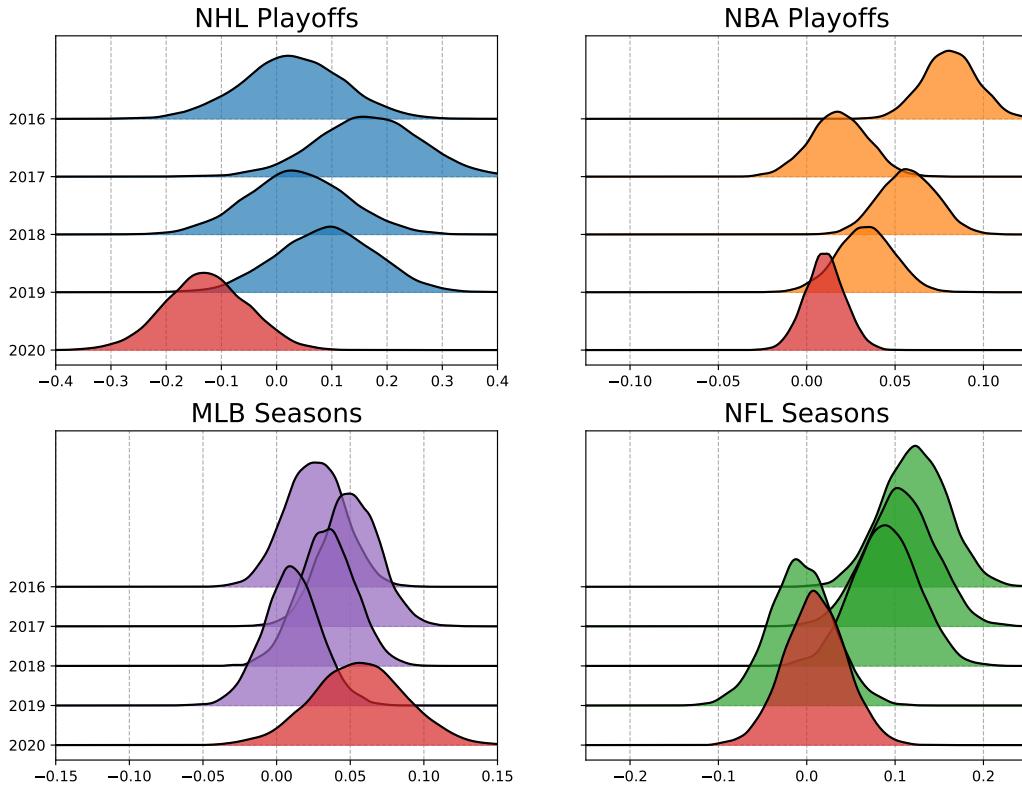


Figure 6.5: Distributions of the estimated home advantage for the NHL, NBA, MLB, and NFL over the past 5 seasons from 2016-2020. Home advantage for playoffs are reported for NHL and NBA because that is when their COVID restricted games took place. Home advantage for regular season is reported for MLB and NFL as their respective playoff seasons are too small for stable results. Red distributions represent COVID-19 bubble adjusted seasons.

even being lower in 2019 than it was in the 2020 COVID-19 adjusted season. We argue the results in Figure 6.5 are enough to overturn the results in Figure 6.4 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 6.4) and individual (Figure 6.5) 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.

7 Discussion and Conclusions

In this chapter we discuss the implications of the results from 6.1, 6.2, and 6.3. We then recognize and discuss the limitations of the methods and inferences of this thesis. We finish with our conclusions.

7.1 Discussion

The results of our model in 6.3 show strong evidence of home advantage being a real positive phenomenon in the NHL and NBA prior to the COVID-19 bubble seasons and that the COVID-19 bubble negatively impacted home advantage by removing it almost entirely. In contrast, the MLB and NFL showed a trend of relatively smaller home advantage parameter estimates in recent years and showed little to no evidence of COVID-19 restrictions having an impact on home advantage in these sports. If we contrast the COVID-19 restrictions in the NHL and NBA to the MLB and NFL, there are two notable differences. First, the NHL and NBA had much stricter COVID-19 bubbles where teams did not travel to each others stadiums, whereas the MLB and NFL did travel to the various stadiums and only restricted fans attending. This suggests that the lack of travel and home city familiarity contributes to home advantage more than a home crowd effect, and therefore results in a greater drop in home advantage in the leagues that had a strict bubble compared to the leagues that allowed travel and play at home stadiums. This agrees with McHill & Chinoy [39] and gives further evidence to the cause of home advantage being more attributable to the general effect of travel. The second difference is the relatively small to no home advantage that the model infers for the MLB and NFL relative to the strongly positive home advantage in recent years found in the NHL and NBA. While we can not fully tease out which of these two differences is stronger, this opens up potential for future work as these leagues continue to play through the COVID-19 pandemic. It will be interesting to see if home advantage returns in the NHL and NBA as they shift toward fewer restrictions similar to the MLB and NFL.

The strongest result for a decrease in home advantage due to COVID-19 restrictions was seen in the NHL. We note that this is particularly interesting because the NHL is somewhat unique to the other leagues, because the home team has an extra difference; they get the last change during stoppages of play, meaning they get to decide player match-ups. An analysis of this effect has been carried out by Meghan Hall [31] who concluded that home teams benefit when they get to control match-ups and argued that this benefit should not be discounted during the 2020 COVID-19 bubble season. The results of our model, however, seem to indicate that no home advantage existed during the NHL's COVID-19 bubble and suggests the effect of last change in the NHL is potentially not as impactful as previously thought.

Our Bayesian regression model has three key advantages over traditional methods for inferring home advantage. First, methods that rely on correlations among raw statistics fail to account for factors such as relative team strengths. For example, a weaker team may have a 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 where teams do not face the same opponents as each other in a perfectly balanced manner. While some studies recognize this discrepancy, they often claim that it is a small effect that can be ignored [48] without showing evidence. We argue that while these claims may hold up for analyses spanning decades they are not appropriate for the short COVID-19 restricted seasons we are considering. Furthermore, these issues and any debate over how much of an effect they have is most reliably mitigated by adjusting for varying team strengths when trying to infer home advantage. Regression analysis methods are primarily used for their ability to account for multiple factors when performing inference, and as such they are most appropriate for our focus of analysing home advantage. Second, the Bayesian framework gives more interpretable results and more flexibility in model building than classical regression methods. This can be seen in the results of the Bayesian framework being distributions for the estimates of each parameter in our model. In this way the implied probability and corresponding uncertainty of parameter estimates are still rigorously defined while being directly measurable and more intuitive to understand than traditional frequentist methods of confidence intervals and p-values. Third, with advancements in computational Bayesian statistics, such as probabilistic programming languages [50] and Hamiltonian Monte Carlo [7], we are able to easily define and compute flexible and complex models using various likelihood functions with ease instead of being limited to traditional methods like Normal and Poisson regressions more traditionally used in sports modelling [36] [29] [35] [4] [6]. These theoretical advantages were corroborated with the empirical results (6.1, 6.2, and 6.3) from our experiments (5.2.2, 5.2.3, and 5.2.4).

With the efficacy of Bayesian multilevel regression modelling established for use in sports analytics, we believe that the flexibility in Bayesian model building can easily transfer to other team-based competitive domains such as competitive gaming or eSports. The viewership and prize pools for eSports are rapidly growing and even eclipsing some traditional professional sports. For example, one of the largest eSports tournaments *The International* for the game *Dota 2* boasts a \$40 million prize pool; in comparison the 2021 U.S open golf tournament had a \$12.5 million prize pool and the 2021 U.S. open tennis tournament has around a \$40 million prize pool. Multilevel regression modelling is an opportunity to model and infer different variables that can attribute to team performance while controlling for other factors such as relative team strengths and should be able to see similar success in other competitive arenas.

7.2 Limitations

While our model has produced some interesting results, it is worth discussing some of its limitations and areas for future work and improvement. The most notable limitation is that the COVID-19 lockout and restricted seasons are unprecedented and come with additional caveats such as protocols for testing, impact of positive tests, reduced practices, and players being away from their families, that extrapolating all results to home advantage or fan impact alone does not address all the possible factors influencing player and team performance. The model also does not account for travel or rest before games as a potential confounding factor for home advantage. This was ignored primarily due to it being irrelevant for the NHL and NBA COVID-19 bubbles, but for the less restrictive MLB and NFL seasons as well as future COVID-19 restricted seasons this could be a potential factor worth exploring.

Our model could benefit by including group level factors when estimating the offensive and defensive strengths of teams. The multilevel structure of the Bayesian framework we have adopted naturally allows for such inclusions where team estimates would be shrunk towards a regression line fit to group level factors rather than being shrunk towards the overall mean [27] [26] [37]. For example, we hypothesize that advanced analytics metrics such as expected goals (xG) and corsi in hockey, regularized adjusted plus-minus (RAPM) in basketball, hitter splits and park factors in baseball, yards gained/allowed above/below expected in football, could all be leveraged to improve team strength estimates. This could also include personnel differences such as the effect of star players being injured, back-up goalies starting, or starting pitchers being included in the estimates of a teams relative strength for a given game. These inclusions are beyond the scope of this work as these analytics and personnel changes and their effect differ greatly across different sports. In future work, we hope to focus on an individual sport and include such factors, using the current model as a baseline to compare against.

Our model is also limited by focusing on only point totals to infer home advantage, while some previous works also analyse differences in penalties to assess a home advantage in the officiating of games [6] [58] [12] [19]. This was excluded from this work because of how much penalties and their effect differ across the various sports we considered, but is something we hope to explore in the future when analysing a single sport in more depth.

7.3 Conclusions

With the results in 6.1, we have shown that a multilevel regression model outperforms traditional regression and simple averaging as measured by out-of-sample predictive fit. Our experiments showed that simple averaging (complete-pooling of, or ignoring, relative team strengths) under-fits the data, and that traditional regression is poised to over-fit the data. Our multilevel model provides the best trade-off in providing a better fit to the datasets while preventing over-fitting.

In 6.2 we have also shown how using the Negative Binomial distribution as the likelihood function for our regression model outperforms the Poisson distribution for sports with overdispersion in their point totals such as the MLB and NFL, while still performing just as well as the Poisson distribution when there is little to no overdispersion such as in the NHL and NBA. We showed the Negative Binomial distribution also outperforms the Normal distribution across all leagues except for the NFL where both models vastly outperformed the Poisson distribution. We argue this is because the Negative Binomial distribution effectively represents positive integers like the Poisson distribution while having an extra parameter, like the Normal distribution, to account for overdispersion which represents a greater spread in the data due to greater variance. In conclusion, we have introduced a Negative Binomial multilevel regression model that fits the data better than previous works and therefore gives better opportunity to infer home advantage.

In 6.3 we found the results of our model when pooling the previous seasons prior to COVID-19 (Figure 6.4) show a noticeable decrease in home advantage for the NHL, NBA, and NFL with no noticeable change in home advantage for the MLB. However, while the year-over-year estimates (Figure 6.4) corroborate these findings to be significant for the NHL and MLB, they show the results are potentially weaker for the NBA and NFL. We argue that the results in Figure 6.5 reveal that home advantage in the NFL was already decreasing leading up to the 2020 season and that the 2020 COVID-19 restricted season had no significant impact on home advantage in the NFL. We further argue that the NBA COVID-19 restricted season may potentially be an outlier similar to the 2017 playoffs. This means we can not be as confident in our conclusions about home advantage decreasing in the NBA as we are with the NHL. We argue the results give evidence that it is likely home advantage decreased in the NBA but we can not be certain with the limited sample we have.

References

- [1] G. A. Agnew and A. V. Carron. Crowd effects and the home advantage. *International Journal of Sport Psychology*, 25(1):53–62, 1994.
- [2] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.
- [3] Jim Albert, Mark E. Glickman, Tim B. Swartz, and Ruud H. Koning. *Handbook of Statistical Methods and Analyses in Sports*. Taylor and Francis Group, LLC, 2017.
- [4] Gianluca Baio and Marta Blangiardo. Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, 37(2):253–264, 2010.
- [5] C. G. Begley and J. P. A. Ioannidis. Reproducibility in science: improving the standard for basic and preclinical research. *Circulation Research*, 116:116–126, 2015.
- [6] Luke S. Benz and Michael J. Lopez. Estimating the change in soccer’s home advantage during the covid-19 pandemic using bivariate poisson regression. *arXiv*, pages 1–23, 2020.
- [7] Michael Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv e-prints*, page arXiv:1701.02434, January 2017.
- [8] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.
- [9] G. E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.
- [10] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324, 1952.
- [11] S. P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–455, 1997.
- [12] B. Buraimo, D. Forrest, and R. Simmons. The 12th man?: refereeing bias in english and german soccer. *Journal of the Royal Statistical Society: Series A*, 173(2):431–449, 2010.
- [13] A. C. Cameron and P. K. Trivedi. Regression-based tests for overdispersion in the poisson model. *Journal of Econometrics*, 46:347–364, 1990.
- [14] Bob Carpenter. Typical sets and the curse of dimensionality, 2017. <https://mc-stan.org/users/documentation/case-studies/curse-dims.html>.
- [15] Albert V. Carron, Todd M. Loughhead, and Steven R. Bray. The home advantage in sport competitions: Courneya and carron’s (1992) conceptual framework a decade later. *Journal of Sports Sciences*, 23(4):395–407, 2005.
- [16] Nicholas J. Costonika. Nhl postseason bubbles successful because ‘everybody bought into this’. *NHL.com*, 2020.
- [17] Kerry S. Courneya and Albert V. Carron. The home advantage in sport competitions: A literature review. *Journal of Sport and Exercise Psychology*, 14(1):13–27, 1992.
- [18] F.N. David. *Games, Gods & Gambling: A History of Probability and Statistical Ideas*. Dover Publications, 1998.

- [19] T. Dohmen and J. Sauerman. Referee bias. *Journal of Economic Surveys*, 30(4):679–695, 2016.
- [20] P. K. Dunn and G. K. Smyth. *Generalized linear models with examples in R*. Springer, 2018.
- [21] D. Forrest, J. Beaumont, J. Goddard, and R. Simmons. Home advantage and the debate about competitive balance in professional sports leagues. *Journal of Sports Sciences*, 23(4):439–445, 2005.
- [22] J. Fox. *Applied Regression Analysis and Generalized Linear Models*. Sage Publications, Inc., 2008.
- [23] L. Garicano, I. Palacios-Huerta, and C. Prendergast. Favoritism under social pressure. *Review of Economics and Statistics*, 87(2):208–216, 2005.
- [24] A. Gelman. Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics*, 48(3):432–435, 2006.
- [25] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–511, 1992.
- [26] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis: Third Edition*. Taylor and Francis Group, LLC, 2014.
- [27] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006.
- [28] S. German and D. German. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.
- [29] Mark E. Glickman and Hal S. Stern. A state-space model for national football league scores. *Journal of the American Statistical Association*, 93(441):25–35, 1998.
- [30] Miguel A. Gómez, Richard Pollard, and Juan Carlos Luis-Pascual. Comparison of the home advantage in nine different professional team sports in spain. *Perceptual and Motor Skills*, 113(1):150–156, 2011.
- [31] Meghan Hall. Examining the effect of the last change in hockey, 2020. <http://meghan.rbind.io/post/last-change/>.
- [32] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.
- [33] N. Higgs and I. Stavness. Bayesian analysis of home advantage in north american professional sports before and during covid-19. *Sci Rep*, 11(14521 (2021)), 2021.
- [34] J. P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2:696–701, 2005.
- [35] D. Karlis and Ntzoufras I. Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society: Series D*, 52(3):381–393, 2003.
- [36] Michael J. Lopez, Gregory J. Matthews, and Benjamin S. Baumer. How often does the best team win? a unified approach to understanding randomness in north american sport. *Annals of Applied Statistics*, 12(4):2483–2516, 2018.
- [37] Richard McElreath. *Statistical Rethinking*. Chapman and Hall, 2020.
- [38] S.B. McGrawne. *The Theory That Would Not Die*. Yale University Press, 2011.
- [39] Andrew W. McHill and Evan D. Chinoy. Utilizing the national basketball association’s covid-19 restart bubble to uncover the impact of travel and circadian disruption on athletic performance. *Scientific Reports*, 10(1), 2020.
- [40] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

- [41] A. Moivre. *The Doctrine of Chances*. W. Pearson, 1718.
- [42] T. Moskowitz and L. J. Werheim. Scorecasting: The hidden influences behind how sports are played and games are won. *Three Rivers Press (CA)*, 2012.
- [43] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.
- [44] NaturalStatTrick. National hockey league data, 2020. data retrieved from Natural Stat Trick, <http://www.naturalstattrick.com/>.
- [45] A. M. Nevill and R. L. Holder. Home advantage in sport. *Sports Medicine*, 28(4):221–236, 1999.
- [46] E. H. Payne, M. Gebregziabher, J. W. Hardin, V. Ramakrishnan, and L. E. Egede. An empirical approach to determine a threshold for assessing overdispersion in poisson and negative binomial models for count data. *Commun Stat Simul Comput.*, 47(6):1722–1738, 2018.
- [47] Judea Pearl. *Causality: Models, reasoning, and inference*. Cambridge University Press, 2000.
- [48] R. Pollard and G. Pollard. Long-term trends in home advantage in professional team sports in north america and england (1876-2003). *Journal of Sports Sciences*, 23(4):337–350, 2005.
- [49] Retrosheet. Major league baseball data, 2020. data retrieved from retrosheet, <http://www.retrosheet.org/>.
- [50] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck. Probabilistic programming in python using pymc3. *Peer J Computer Science*, 2, 2016.
- [51] R. Schoot, S. Depaoli, R. King, B. Kramer, K. Martens, T.G. Mahlet, M. Vannucci, A. Gelman, D. Veen, J. Willemse, and C. Yau. Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1) (2021), 2021.
- [52] Barry Schwartz and Stephen F. Barsky. The Home Advantage. *Social Forces*, 55(3):641–661, 1977.
- [53] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.
- [54] SportsReferenceLLC. National basketball association data, 2020. data retrieved from basketball-reference, <http://www.basketball-reference.com/>.
- [55] SportsReferenceLLC. National football league data, 2020. data retrieved from pro-football-reference, <http://www.pro-football-reference.com/>.
- [56] S.M. Stigler. *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press, 1986.
- [57] PyMC Development Team. A primer on bayesian methods for multilevel modeling, 2018. https://docs.pymc.io/notebooks/multilevel_modeling.html.
- [58] C. Unkelbach and D. Memmert. Crowd noise as a cue in referee decisions contributes to the home advantage. *Journal of Sport and Exercise Psychology*, 32(4):483–498, 2010.
- [59] A. Vehtari, A. Gelman, and J. Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27:1413–1432, 2016.
- [60] Sumio Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594, 2010.
- [61] Jeff Zillgitt. Meet the eight key figures who helped make the nba bubble a success. *USA Today*, 2020.

Appendix A

Appendix

	$P(\beta_{20} < 0)$	$P(\beta_{20} < \bar{\beta}_{16-19})$	$P(\beta_{20} < \bar{\beta}_{19})$	$P(\beta_{20} < \bar{\beta}_{18})$	$P(\beta_{20} < \bar{\beta}_{17})$	$P(\beta_{20} < \bar{\beta}_{16})$
NHL	0.95	0.99	0.97	0.92	0.99	0.91
NBA	0.17	0.99	0.88	0.99	0.66	0.99
MLB	0.04	0.26	0.39	0.39	0.96	0.63
NFL	0.39	0.96	0.29	0.49	0.67	0.92

Table A.1: The estimated probabilities that the home advantage parameter during the 2020 COVID-19 restricted games (β_{20}) is less than 0, the previous four seasons (2016-2019) mean ($\bar{\beta}_{16-19}$), and the previous seasons individual means ($\bar{\beta}_{19}$, $\bar{\beta}_{18}$, $\bar{\beta}_{17}$, $\bar{\beta}_{16}$).

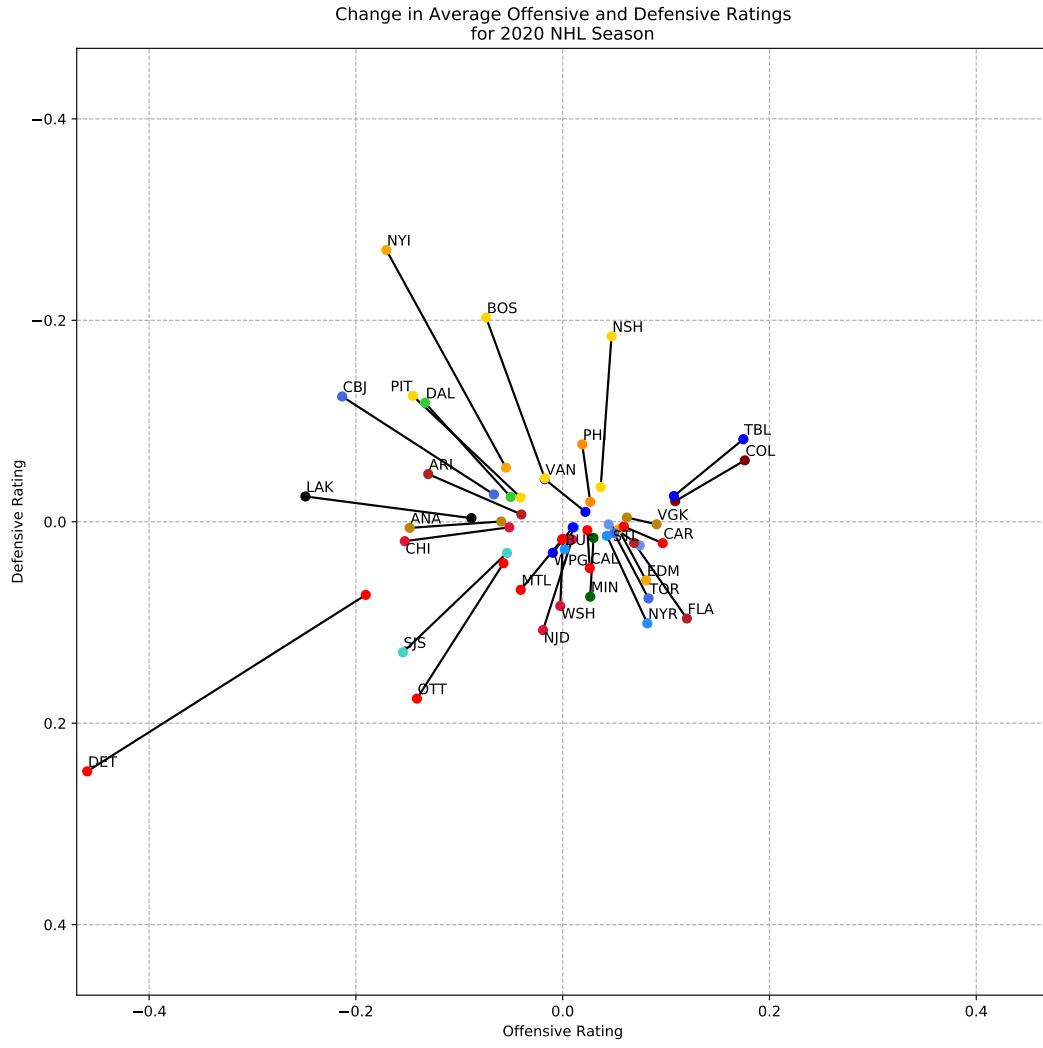


Figure A.1: Offensive and Defensive team ratings for the 2020 NHL season. The points with the team labels next to them are ratings generated by traditional regression, and the corresponding ratings are generated from multilevel regression to highlight the effect of shrinkage to the mean.

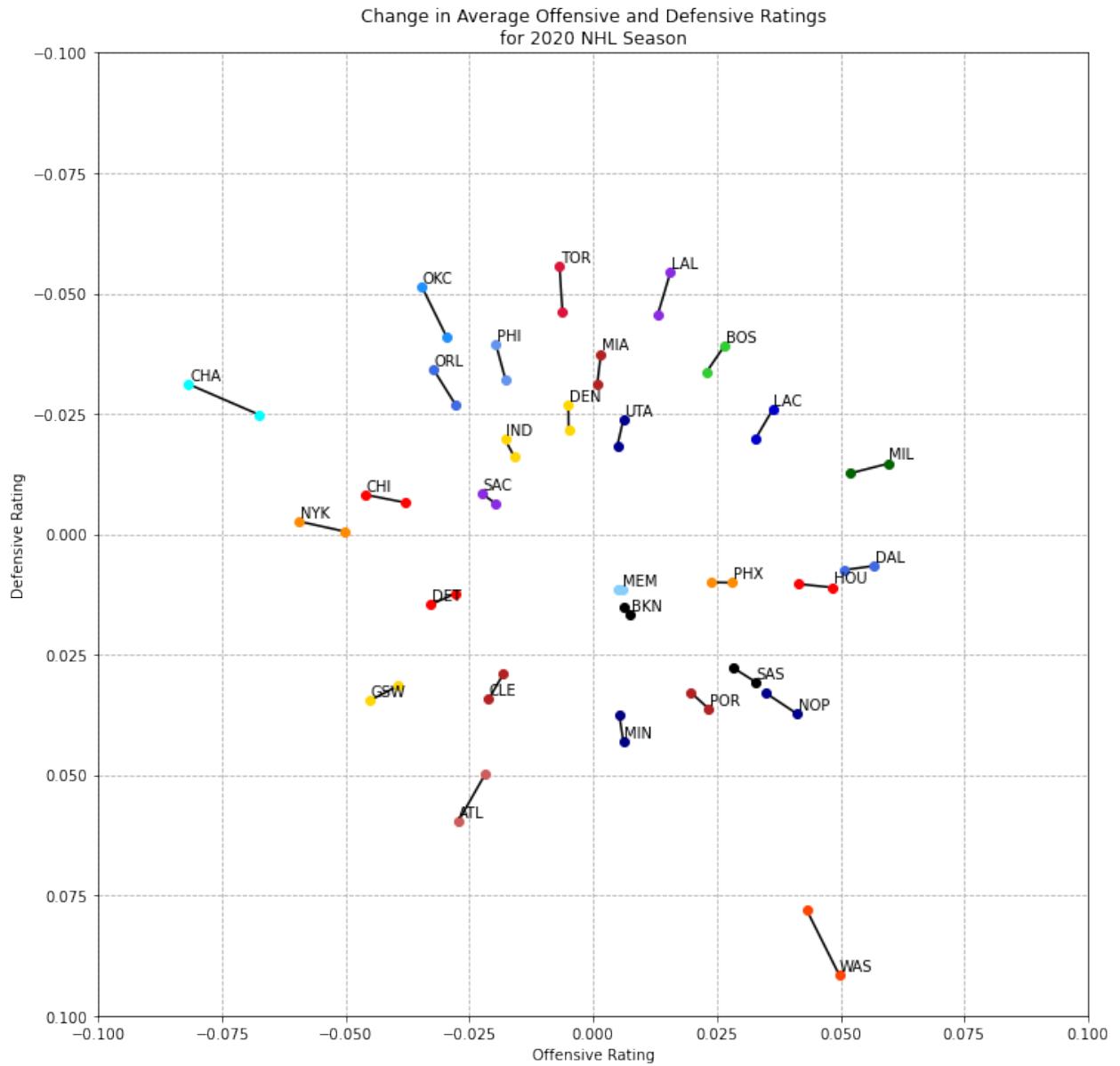


Figure A.2: Offensive and Defensive team ratings for the 2020 NHL season. The points with the team labels next to them are ratings generated by traditional regression, and the corresponding ratings are generated from multilevel regression to highlight the effect of shrinkage to the mean.

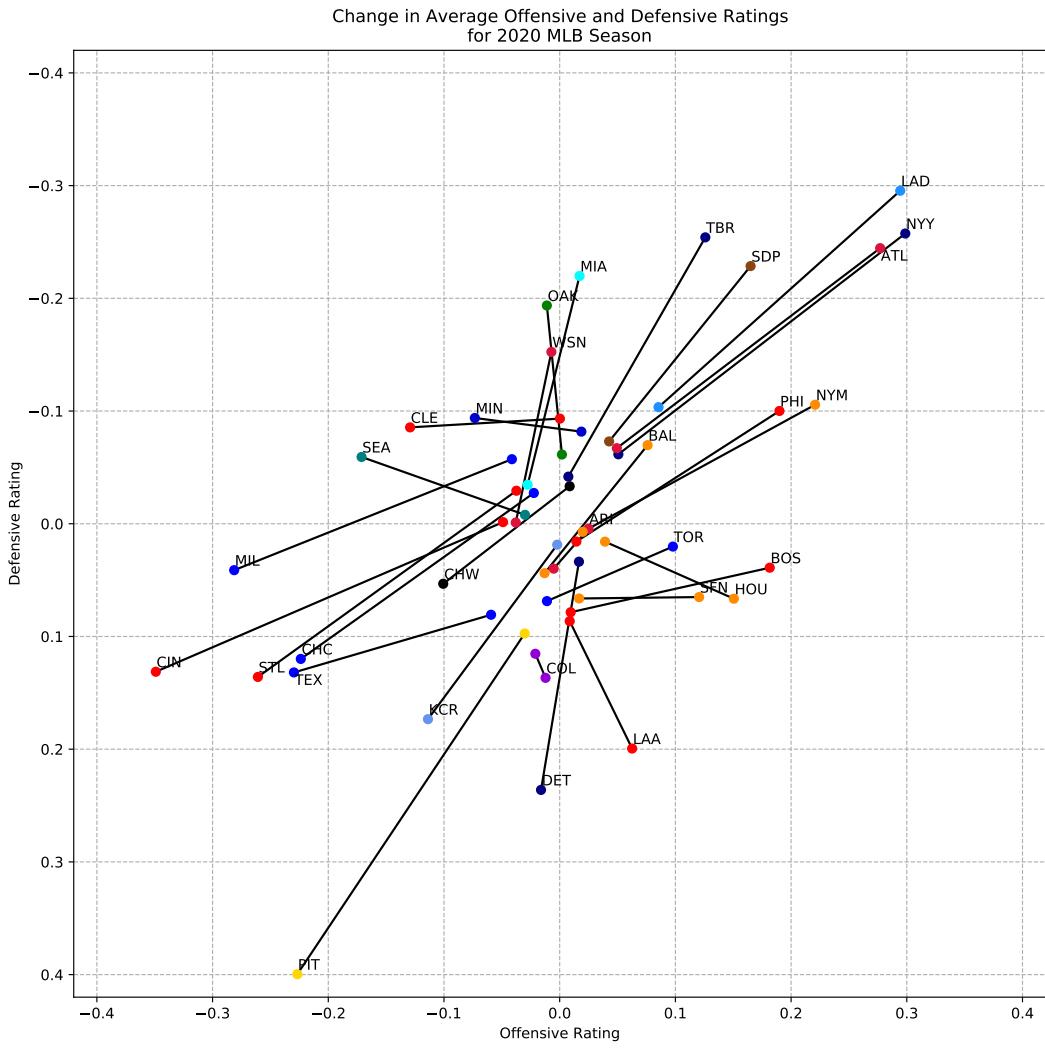


Figure A.3: Offensive and Defensive team ratings for the 2020 MLB season. The points with the team labels next to them are ratings generated by traditional regression, and the corresponding ratings are generated from multilevel regression to highlight the effect of shrinkage to the mean.

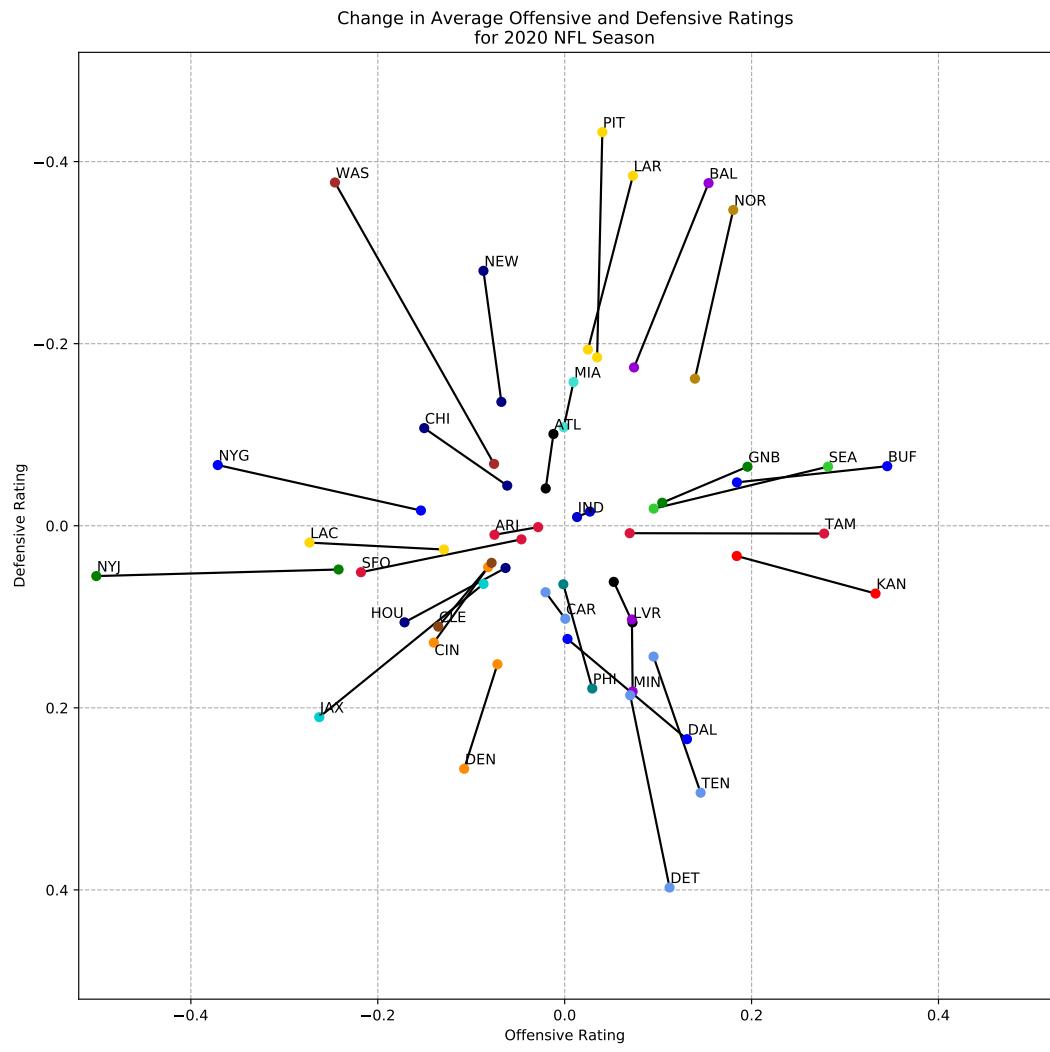


Figure A.4: Offensive and Defensive team ratings for the 2020 NFL season. The points with the team labels next to them are ratings generated by traditional regression, and the corresponding ratings are generated from multilevel regression to highlight the effect of shrinkage to the mean.