

HARVARD UNIVERSITY  
Graduate School of Arts and Sciences




DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the  
Department of Statistics,  
have examined a dissertation entitled

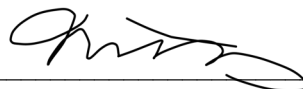
A Tale of Two Multi-Phase Inference Applications

presented by Kathryn McKeough,

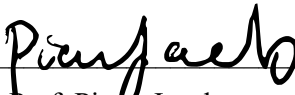
candidate for the degree of Doctor of Philosophy and hereby  
certify that it is worthy of acceptance.

Signature 

Typed name: Dr. Mark E. Glickman

Signature 

Typed name: Prof. Xiao-Li Meng

Signature 

Typed name: Prof. Pierre Jacob

Date: April 21, 2020



# A Tale of Two Multi-Phase Inference Applications

A dissertation presented

by

Kathryn McKeough

to

The Department of Statistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Statistics

Harvard University

Cambridge, Massachusetts

April 2020

© 2020 Kathryn McKeough  
All rights reserved.

## A Tale of Two Multi-Phase Inference Applications

# Abstract

Multi-phase inference refers to any sequential procedure where the results, or some realization of the output of one phase, is fed into another phase. Multi-phase models are becoming more prevalent in applied statistical analyses as data gets bigger and more complicated. They offer a solution for complex statistical problems where modeling all parameters jointly has its limitations. We explore two applications, one in sports analytics and astronomy, where we choose multi-phase models to explore our data.

### **Part 1 - Predicting Athlete Performance:**

It is often the goal of sports analysts, coaches, and fans to predict athlete performance over time. Methods such as Elo, Glicko, and Plackett-Luce based ratings measure athlete skill based on results of competitions over time but have no predictive strength on their own. Growth curves are often applied in the context of sports to predict future ability, but these curves are too simple to account for complex career trajectories. We propose a non-linear, mixed-effects growth curve to model the ratings as a function of time and other athlete-specific covariates. The mixture of growth curves allows for flexibility in the estimated shape of career trajectories between athletes as well as between sports. We use the fitted growth curves to make predictions of an athlete's career trajectory in two ways. The first is a model of how athlete performance progresses over time in a multi-competitor scenario as an extension to the Plackett-Luce model. The second is a method that applies the

growth curve as a second step to existing rating systems of multi-competitor and head-to-head sports. We show this method can be applied to different sports by using examples from men's slalom and women's luge, respectively.

### **Part II - Defining Regions that Contain Complex Astronomical Structure:**

Astronomers are interested in delineating boundaries of extended sources in noisy images. Examples include finding outlines of a jet in a distant quasar or observing the morphology of a supernova remnant over time. Analyzing the morphology of these objects is particularly challenging for X-ray images of high redshift sources where there are a limited number of high-energy photon counts. Low-counts Image Reconstruction and Analysis (LIRA), a Bayesian multi-scale image reconstruction, has been tremendously successful in analyzing low count images and extracting noisy structure. However, we do not always have supplementary information to predetermine ROI, and the size and shape can significantly affect flux/luminosity. To group similar pixels, we impose a multi-phase model using the output of LIRA to build a distribution for the shape of the ROI. We adopt the Ising model as a prior on assigning the pixels to either the background or the ROI. This Bayesian post-process step informs the final boundary. This method is applied to observed data as well as simulations to show it is capable of picking out meaningful ROIs.

# Table of Contents

Title Page	i
Copyright	ii
Abstract	iii
Table of Contents	v
Acknowledgements	ix
Foreword	xi
<b>I Predicting Athlete Performance</b>	<b>1</b>
<b>1 Plackett-Luce Model with a Parametric Growth Curve</b>	<b>3</b>
1.1 Introduction . . . . .	3
1.2 Related Work . . . . .	4
1.2.1 Predicting Athlete Strength over Time . . . . .	5
1.2.2 Growth Curves . . . . .	7
1.3 Model Definition . . . . .	8
1.3.1 Adding Covariates . . . . .	11

*Table of Contents*

---

1.4	Model Fitting & Selection . . . . .	12
1.5	Results . . . . .	14
1.5.1	Case Study: Women’s Luge . . . . .	15
1.5.2	Evaluation . . . . .	21
1.6	Conclusion . . . . .	23
<b>2</b>	<b>Two-Step Model for Predicting Athlete Strength</b>	<b>25</b>
2.1	Introduction . . . . .	26
2.2	Methods . . . . .	28
2.2.1	Ratings . . . . .	28
2.2.2	Growth Curve . . . . .	33
2.3	Model Fitting . . . . .	35
2.4	Results . . . . .	36
2.4.1	Case Study: Men’s Slalom . . . . .	36
2.4.2	Coverage . . . . .	42
2.5	Clustering Career Trajectories . . . . .	44
2.6	Conclusions . . . . .	48
<b>II</b>	<b>Defining Regions that Contain Complex Astronomical Structures</b>	<b>51</b>
<b>3</b>	<b>Defining Regions that Contain Complex Astronomical Structures</b>	<b>53</b>
3.1	Introduction . . . . .	53
3.2	Model & Inference . . . . .	56
3.2.1	Step 1: Image Reconstruction Using LIRA . . . . .	57
3.2.2	Step 2: Distribution of Pixel Assignments . . . . .	59



3.2.3	Step 3: Optimal Boundary . . . . .	66
3.3	MCMC Implementation . . . . .	68
3.4	Validation . . . . .	69
3.4.1	2D Gaussian Simulation . . . . .	70
3.5	Results . . . . .	80
3.5.1	Extragalactic Jets . . . . .	80
3.5.2	Source 87A . . . . .	81
3.6	Conclusion . . . . .	85
<b>Afterword</b>		<b>87</b>
<b>A Orthogonal Polynomials</b>		<b>93</b>
<b>B Hierarchical Clustering</b>		<b>95</b>
<b>C Model Compatibility for the LIRA + Ising Model</b>		<b>97</b>
<b>Bibliography</b>		<b>99</b>

This page is intentionally left blank.

# Acknowledgements

The work presented in this dissertation would not have been possible without the help and support of those around me. I would like to take this opportunity to acknowledge the people who have directly and indirectly influenced myself and my work that eventually led to the writing of this dissertation.

I would first like to thank my two advisors Mark Glickman and Xiao-Li Meng. Both took me on as a student early in my Statistics career at Harvard and have helped me grow since the beginning. Mark, for introducing me to the field of sports analytics and providing support within and outside of our research projects together. Xiao-Li, for being an inspiration to come to Harvard and warmly welcoming me into the CHASC astrostatistics community. Pierre Jacob, for serving on my committee and for many interesting discussions involving multi-phase inference. Vinay Kashyap, Aneta Siemiginoska, and Peter Freeman; they are the reason I began studying statistics at my time at Carnegie Mellon University and in the REU programs at the Harvard Smithsonian Center for Astrophysics. The CHASC group, in particular Andreas Zezas, David van Dyk, Luis Campos, and Shihao Yang were instrumental in brainstorming and validating the decisions made in my astrostatistics projects. They were very generous with their time and resources which granted me endless educational opportunities. I would also like to remember Alanna Connors who was the impetus for me to begin my first project as an astro-statistician. It was an honor continuing in her footsteps while working with LIRA. The faculty, staff, and fellow students in the Department of Statistics at Harvard. Friends, family, and teachers I met at Carnegie Mellon through the Physics and Statistics departments. A final thank you to those who helped directly with the editing process who did not have to. Writing does not come naturally to me and I was grateful to have help and fresh perspectives. (Phil Jones, Sanqian Zhang, Dan Violette, and Kyle Henson)

“Things are only impossible until they are not.”

— Jean-Luc Picard

*To my family, friends, and mentors who helped make this possible.*

# Foreword

Multi-phase models are becoming increasingly necessary in a world where data is increasingly larger and more complex. We define multi-phase to refer to any sequential procedure where the results, or some realization of the output of one phase, is fed into another phase. This encompasses pre-processing techniques including, but not limited to dimension reduction (e.g., [Jacques and Preda, 2014](#)), meta-analysis (e.g., [Lunn et al., 2013](#)), or recovering missing data (e.g., [Xie and Meng, 2017](#)). When faced with the task of performing inference on a network of parameters, it is common for a Statistician with a Bayesian mindset to immediately jump to a completely joint model, where all parameters are estimated at the same time and influence the inference of one another. This mindset is not unprecedented as it is the ideal in a perfectly specified model ([Gelman et al., 2013](#)). However, in an applied setting, *all models are misspecified* to a varying degree. It is also sometimes physically impossible due to computational or scientific constraints to construct or fit such a model. Multi-phase inference is often necessary when analyzing data in real-world applications. We see multi-phase models in many domains such as econometrics ([Murphy and Topel, 1985](#)), environmental studies ([Blangiardo et al., 2011](#)), genetics ([Li and Stephens, 2003](#)), and physical-biological models ([Béal et al., 2010](#)). In this dissertation, we explore two different applications in the domains of sports analytics and astronomy. These two applications are connected by the choice to use multi-phase inference

Part I consists of two chapters that focus on estimating and predicting athlete performance over time. A common way of characterizing athlete performance is through the use of rating systems.

Ratings quantify an athlete’s performance based on the results of a game, match, or competition. They make it simple to compare athletes’ abilities to one another objectively and to determine the odds of one athlete defeating another. However, rating systems cannot inherently make predictions about future athlete performance. Knowledge of future ratings is useful to coaches or managers for tasks such as recruitment, training, or resource allocation. For example, it might be useful to know if two athletes with the same ability this season will perform similarly the following year. In Chapter 1, we develop a modification to the classic Plackett-Luce model for rank ordered data that incorporates a novel growth curve to characterize the time-varying nature of athlete performance. In Chapter 2, we apply the growth curve model as a post-processing step to already existing, estimated athlete ratings. Implementing a multi-phase model here allows us to generalize the application of the growth curve to ratings in other types of sports. In both cases, incorporating the growth curve as an assumption on how ratings change over time gives us not only the ability to predict future ratings, but also make exploratory observations such as clustering career trajectories.

Part II discusses a boundary algorithm built to apply to sparse, diffuse, and irregularly shaped astronomical objects. The technique uses a multi-phase model which first reconstructs details in the image, builds a distribution of the boundary, and finally optimizes over the distribution to get an estimate of the boundary. Image reconstruction is done using a Bayesian technique called LIRA, which is known for eliciting previously unobserved detail in sparse images (Esch et al., 2004; Connors and van Dyk, 2007). The second step is a novel method to model the boundary in the form of pixel assignments, assigning pixels within the reconstructed image to the source and to the background. The Ising distribution is imposed a priori in the second step to induce cohesiveness between pixel assignments. Although the applications we discuss in Part II apply only to astronomy, the technique can be applied to any low-count, low-resolution image. A multi-phase technique is implemented to simplify the process of adding the new pixel assignment structure, while utilizing the results from an already proven reconstruction method, LIRA.

Liu et al. (2009) presents several reasons why a researcher would choose to implement a multi-

phase model. The first is when some assumptions in the model are unreliable. Breaking the models into phases and addressing them in a step-wise manner rather than all at once leads to weaker interactions between modules than a fully joint approach. [Liu et al.](#) highlights a pedagogical example of a simple random effects model where a multi-phase modeling decision provides more robust inference. Let's say we are regressing  $Y_{ij}$  of individuals,  $j = 1, \dots, n$  within-groups,  $i = 1, \dots, N$ , given the random intercept  $b_i$ :

$$\begin{aligned} Y_{ij} &= b_i + \epsilon_{ij} \\ \epsilon_{ij} &\sim N(0, \sigma_i^2) \\ b_i &\sim N(0, \tau^2), \end{aligned}$$

where  $\tau^2$  and the set of  $\sigma_i^2$  are unknown variance parameters. When modeled jointly, [Liu et al.](#) shows that when we break the assumption that  $b_i$  is normal, for example, if one of the  $b_i$  is an order of magnitude higher than all the others, then  $\sigma_i^2$  is inflated rather than  $\tau^2$ . To mitigate this unexpected outcome, we can infer the  $\sigma_i^2$  before incorporating it as given to infer  $\tau^2$  and  $b_i$ . A common application for multi-phase model is Pharmacokinetic / Pharmacodynamic (PKPD) models ([Bennett and Wakefield, 2001](#)). PK models predict the concentration of a drug in the bloodstream over time based on scientific knowledge and known parameters. Therefore, these estimates are believed to be close to the truth and trustworthy. PD models rely on observations of biological responses recorded over time as a product of the concentration of the drug, thus are more subject to error. If we want to model both cases jointly to share information, we run the risk of the estimation of the PD parameters negatively influencing the PK parameters. By first estimating the PK model parameters and then using the results to estimate the PD model parameters, we are cutting the model so the less reliable PD data do not influence the PK model parameters. [Jacob et al. \(2017\)](#) formalizes the benefits of using a multi-phase model over a fully joint model in different cases of misspecification. The authors show in cases of model misspecification, a two-step model will lead to better predictive inference than a single model that incorporates all data sources

jointly.

The second reason for multi-phase inference is the decision of domain experts for scientific understanding or future scientific developments. An early example of multi-phase inference is multiple imputation (Rubin, 1987). In multiple imputations, we employ different techniques to use observed data to build a model to fill in missing values that we were unable to observe. Frequently this is done by the statistician who is also doing further analysis (e.g., Tu et al., 1993). Even so, this is often done as a pre-processing step, thus falling into the realm of multi-phase inference (Xie and Meng, 2017). In some cases, as Rubin suggests, the scientist doing the imputation may not even know who will use the results. As data grows in size and complexity, it becomes less common that a single person has complete vertical control of the data from its collection to its final analysis. More often than not, people of different expertise are in charge of pre-processing the data than those doing the analysis. Therefore, we must be conscious of how we deal with the data and results in each step. Blocker and Meng (2013) outlines two scientific examples of multi-phase inference from the fields of genetics and astronomy. In both cases, we see scenarios where the raw, observed data is processed before making inference on parameters of interest, but how the data is processed can affect the final results. For example, when studying microarrays, the relationship between the observed probe-level intensities and the true gene expressions is affected by an unknown noise variable. Scientists are typically interested in parameters that are dependent on the true gene expression but must use the pre-processed intensities to make inferences about them.

Other reasons for modularization or cuts might be purely computational. Sometimes it helps in the speed of mixing or convergence of MCMC algorithms if the inference is made in multiple phases. Multi-phase inference should also be considered if there is a lack of identifiability in the parameters of the fully joint model (Liu et al., 2009; Plummer, 2014). In the random effects example from Liu et al. (2009), the multi-phase solution has a much simpler sampling procedure than the fully joint model. Li and Stephens (2003) introduces a “product of approximate conditional models”, which simplifies the likelihood of haplotype groups by breaking them into products of conditional



probability models. The authors suggest that a fully joint model is preferred, but infeasible due to computational limitations. [Murphy and Topel \(1985\)](#) suggest an improved two-step approach for econometrics, replacing unobserved components with estimated values, with asymptotically correct uncertainty. Using this approach can be robust to situations where joint estimation methods are computationally impossible. In a world where the amount and complexity of data are growing, it is apparent that statisticians must understand multi-phase models to make inference in these real-world scenarios.

This page is intentionally left blank.

## Part I

# Predicting Athlete Performance

This page is intentionally left blank.

# Chapter 1

## Plackett-Luce Model with a Parametric Growth Curve

### 1.1 Introduction

The ability to predict an athlete's performance is highly sought after by sports analysts, coaches, and fans. For example, knowing when an athlete's ability may peak could allow coaches to make more intelligent recruiting and resourcing decisions. Predicting an athlete's skill is not a particularly novel idea in the field of sports analytics. However, most researchers place less emphasis on analyzing multi-competitor sports and instead focus on head-to-head sports. The primary difference between multi-competitor sports, such as races, or judge-based competitions, and head-to-head sports is that the results are a rank ordering of all competitors rather than the identification of a binary winner or loser (e.g., track events, diving or ice skating). Head-to-head sports can be conceptualized as a specific case of rank-ordered sports in which only two athletes compete at a time. As a result, the results from multi-competitor analyses can apply to both rank-ordered and head-to-head sports.

Relative player performance is often measured by player "strength". Player strength is a quan-

titative measure of player ability based on past performances. If an athlete has a greater strength than another athlete, the higher-rated athlete will defeat the lower-rated athlete in direct competition with higher probability. One approach to measuring these strengths over time is to use a growth-curve model. A growth curve is a parametric representation of how a numerical quantity changes over time. Growth-curve-based methods of prediction are advantageous when compared to nonparametric methods because they can more easily capitalize on built-in assumptions that are made when modeling player strength over time. Conversely, unless cautiously addressed, nonparametric methods can result in high prediction errors when extrapolating out of sample, limiting their utility to description, rather than prediction.

Growth-curve models are well-proven and have been utilized for prediction in many fields, including sports-analytics literature. However, the growth curves currently used to predict athlete strengths are too simple. They do not account for complicated changes in performance over time, such as multi-modal career trajectories due to injury. We present a novel model for player strengths in multi-competitor sports that incorporates a flexible growth-curve model to estimate the evolution of athlete performance over time. This growth curve has a more flexible shape than any that have been used in sports literature. In Section 1.2 we introduce current, including growth-curve-based, methods for estimating performance over time. In Section 1.3, we introduce the Plackett-Luce model and discuss the construction of the growth curve to estimate performance over time. We summarize the model fitting and selection techniques in Section 1.4. In Section 1.5, we discuss the application of our methods to the results of professional women’s luge events.

## 1.2 Related Work

Every sport can be categorized as either a head-to-head or multi-competitor competition. Head-to-head sports are defined as games or matches that determine a winner and a loser. Popular sports that fall under this definition include basketball, tennis, and soccer. For head-to-head competitions,

we focus on estimating athlete performance through paired-comparison models. The most commonly used model is the Bradley-Terry model (Bradley and Terry, 1952), which is the basis of the widely used Elo ratings and more recently popularized for its use in chess leagues, Glicko ratings (Elo, 1978; Glickman, 1999). Another approach for paired comparisons is the Thurstone-Mosteller model (Mosteller, 1951), which differs from the Bradley-Terry model by setting the probability of winning in the probit space rather than the logit space.

The majority of multi-competitor sports are races, where players compete for the fastest time, or point-based systems, where judges award players points based on performance. Sports that fall under this definition include marathons, Formula 1 racing, and diving. The modeling of the performance of athletes in these sports is less studied than in their head-to-head counterparts. The Plackett-Luce model, also known as the rank-order logit model, is a commonly used approach to modeling rank-ordered results. This model was created as an extension to the multinomial logit choice model (Luce, 1959) for rank orderings (Plackett, 1975). The Plackett-Luce is the primary focus of the model presented in this chapter. The Plackett-Luce model without modification provides no way of accounting for changes in athlete strength over time.

### 1.2.1 Predicting Athlete Strength over Time

Using past performance to predict an athlete's strength changes over time is often more helpful than modeling performance in a single competition. For example, if two athletes have a similar strength at a single time point, it would be essential to know whether one athlete is on the decline of their career versus an upswing. This section looks at recent work in tracking strength over time and predicting future performance based on these longitudinal models.

Several nonparametric approaches capture estimates of player performance over time using the Plackett-Luce model for multi-competitor sports. Baker and McHale (2013) track golfer strength over time by fitting splines through golfer ability via barycentric rational interpolants. Caron

and Teh (2012) use a nonparametric method for ranked, multi-competitor events where strengths progress over time using a gamma process. Although both methods offer a solution to modeling the time varying nature of athlete strength, neither offer a means for extrapolating and predicting future performance.

Bayesian approaches using stochastic updates can be used to estimate performance over time. TrueSkill is a Bayesian skill rating system that ranks head-to-head competitions in online gaming. The rating system is generalizable to team sports as it can model the performance of a team by aggregating individual skill. The authors model skill throughout time using normally distributed updates to the ability parameter (Herbrich et al., 2007). Glickman and Hennessy (2015), henceforth denoted as GH, is another Bayesian technique that utilizes a Plackett-Luce model to estimate the skill of athletes in multi-competitor sports over time. This model propagates performance through time via a Gaussian random walk. None of these models accounts for player-specific covariates, including how performance might change due to an athlete's age or other time-varying traits.

A separate approach for estimating athlete skill are dynamic models such as Markov transition models. Markov transition models use a hierarchical Bayesian model to evaluate the evolution of performance over time. Jensen et al. (2009) and Glynn and Tokdar (2017) have used this technique to predict the number of home runs a Major League Baseball player will hit based on their past performance. Although these models have demonstrated predictive power, they are only applied to a measured performance (home runs) and are not used to estimate skill from rank-ordered results.

What is missing is modification to the Plackett-Luce model that models the systematic changes in athlete strength over time while providing a means for predicting future performance and accounting for athlete specific observations. A solution is to use a parametric approach by modeling athlete strength over time using a growth curve.



### 1.2.2 Growth Curves

A growth curve is any parametric function that describes how a quantity changes over time. Shapes can range widely and may take the form of linear, exponential, logistic, or S-shaped models. Growth curves can also include more complicated relationships like mixed-effects or nonlinear regression (Panik, 2014; Pinheiro and Bates, 2000). Growth curves are powerful tools in longitudinal analyses across many different disciplines. Wishart (1938), was an early adopter of a growth curve, using the technique to fit a different quadratic function to the weight of pigs eating three different diets over 16 weeks. The use of a growth curve allowed Wishart to utilize all recorded measurements across the 16-week trial to determine whether weight gain was significantly different between the three groups. Just looking at the difference between the beginning and end of the trial was insufficient in delineating the diets. Growth curves can be used to measure rates of decay. This is demonstrated in compartmental models that are used in pharmacology to determine the expected amount of a drug or substance left in the bloodstream (Perrier and Gibaldi, 1973). Growth curves can also be used for prediction as in Airolidi et al. (2009), where gene expressions are used to predict the growth rate of cellular cultures.

Growth curves have many use cases in sports analytics. In particular, growth curves with mixed effects or varying parameters per athlete are commonly used to describe performance over time. One example is Brander et al. (2014), which looks at National Hockey League data to determine the effect of aging on scoring for hockey players. They fit a quadratic or cubic function of age to predict performance in terms of scoring to determine the age at which player performance peaks, using a fixed intercept term to control for variation between players. Malcata et al. (2014) models the expected times of a triathlete for swimming, cycling, and running events as a function of age. They choose to model performance as a quadratic function of age and as a linear trend for the calendar year with coefficients for random effects so that each athlete has a different set of parameters. This model is similar to Brander et al., but the shape of the curve changes per athlete. Bell et al. (2016) use multilevel modeling on Formula 1 race results to account for team and driver effects and identify

the best racers over time.

Career trajectories do not always take a simple polynomial shape. Because of this, [Moudud et al. \(2008\)](#) use the nonlinear logistic growth curve with a mixed-effects growth curve to predict the speed of youth cross country skiers. While effective, this model is monotonically increasing and therefore limited to youth athletes. Although not a sports-related example, a closely related exercise by [Bradlow and Fader \(2001\)](#) uses the generalized gamma curve to model the ranking of songs on Billboard charts. Both of these growth curves use an exponential decay term with the expectation that, over time, performance converges to a specific value or decays altogether, a characteristic we expect to be shared with a career trajectory. This model is also too limited to use on athletes since the guarantee that a song has a single peak on Billboard charts is a safer assumption than an athlete having solely one peak performance in their career.

We can improve the Plackett-Luce model by extending it to take advantage of the growth-curve-based model assumptions on the time variation of athlete strength. We create a novel growth curve that has a flexible enough shape to account for any type of career shape. No sport specific assumptions were made in creating the growth curve so that it may be generalizeable to all sports. Estimating parameters of the growth curve and the ability to project future athlete strength allows one to easily compare and contrast career trajectories between athletes even if they are at different stages of their career.

### 1.3 Model Definition

During  $T$  discrete time periods we observe  $n$  athletes that participate in observed competitions. Each athlete  $i = 1, \dots, n$  has an ability parameter  $\theta_{it}$  that indicates the competitor strength at a given time period  $t = 1, \dots, T$ . Within each time period  $K_t$  competitions take place where in competition  $k = 1, \dots, K_t$ ,  $m_{kt}$  competitors participate.

We use a Plackett-Luce model on the athletes' latent performance at a particular moment in time. The winner to a given game or match has the highest latent performance, the second-place athlete would have the second-highest, and so on until the final-place athlete. However, this performance is unobserved, so we must infer it based on the observed rank ordering. Let  $Y_{it}$  be the latent performance by competitor  $i$  at time  $t$ . We specify the distribution to be an extreme value distribution:

$$Y_{it}|\theta_{it} \sim \text{Gumbel}(\theta_{it}) . \quad (1.1)$$

in which it follows that the likelihood conditional on  $\boldsymbol{\theta}_t = \{\theta_{1t}, \dots, \theta_{nt}\}$ , for a given competition  $k$  within time period  $t$  as

$$L_{kt} = P(Y_{(1)t} > Y_{(2)t} > \dots > Y_{(m_{kt})t} | \boldsymbol{\theta}_t) = \prod_{i=1}^{m_{kt}-1} \frac{\exp(\theta_{(i)t})}{\sum_{\ell=i}^{m_{kt}} \exp(\theta_{(\ell)t})} , \quad (1.2)$$

where  $\theta_{(i)t}$  is the parameter that corresponds to the ordered, from largest to smallest, latent performance  $Y_{(i)t}$  of the  $m_{kt}$  participating competitors.

Up to this point, we have established the set up for a Plackett-Luce model (Plackett, 1975). To account for the competitor's ability to change over time, we assume a novel parametric growth-curve model on  $\theta_{it}$  as a function of the current period  $t$  and the period in which the athlete first competed  $t_{0i}$ . The model we assume on  $\theta_{it}$  is the growth-curve model

$$\theta_{it} = \alpha_i + \left[ \beta_{i0} + \beta_{i1}t_1^* + \beta_{i2}t_2^* + \beta_{i3}t_3^* + \dots + \beta_{ip}t_p^* \right] e^{-\omega(t-t_{0i})} , \quad (1.3)$$

where the bracketed term is a  $p$ -th order polynomial. Intercepts and coefficients  $\alpha_i$  and  $\boldsymbol{\beta}_i = \{\beta_{i0}, \beta_{i1}, \beta_{i2}, \dots, \beta_{ip}\}$  vary per individual and the rate,  $\omega > 0$  is fixed across individuals. The  $\boldsymbol{\beta}_i$  and  $\alpha_i$  are fit via random effects. Sharing a common distribution allows for information may be shared across athletes, assisting in the estimate of the parameters for the athletes with few observations. The order of the polynomial is a free parameter  $p$  that needs to be chosen via model selection as described in Section 1.4. To avoid correlation between polynomial orders of time, we

instead use orthogonal polynomials of  $t - t_{0i} = 0, \dots, T - t_{0i}$ , denoted as  $t_1^*, t_2^*, \dots, t_p^*$ . Orthogonal polynomials are calculated across the entire data set by creating an orthogonal basis from a QR decomposition done via the Gram-Schmidt process. Since orthogonal polynomials are calculated using a time up to  $T$ , when used for projection, we apply the same Gram-Schmidt coefficients used in the initial basis up to  $T$ , but on times  $T + 1, T + 2$ . See Appendix A for details on the creation of the orthogonal polynomials in fitting and prediction. The raw time is used in the exponent since orthogonal polynomials are not guaranteed to be non-negative. To ensure identifiability of the strength parameters we center the  $\theta_{it}$  at every time  $t$  by requiring:

$$\sum_{i=1}^n \theta_{it} = 0.$$

This growth curve form is useful because it is constructed in such a way that it offers flexibility while relying on intuitive coefficients. The examples of growth curves used in sports that are mentioned throughout this chapter only fit the data up to a quadratic polynomial. While a quadratic describes the typical athlete trajectory, increasing performance until reaching their peak and then declining in performance, it does not allow for more complex trajectories. To encourage flexibility in the model, the growth curve in Equation 1.3 allows for a polynomial of varying degrees. The intercepts  $\alpha_i$  not only adjust for different starting strength, but they also represent the limit of decay over a long career. The coefficients  $\beta_{i0}, \beta_{i1}, \dots, \beta_{ip}$  and intercept  $\alpha_i$  parameters differ by athlete since we do not expect each athlete to have the same career trajectory. However, the parameters are assumed to come from a common normal distribution. For the  $\alpha_i$  this is a normal distribution with a mean of 0 and a variance of  $\sigma_\alpha^2$ . The  $\beta_{ib}$  coefficients are drawn from independent normal distributions with different means  $\eta_{\beta_b}$  and variances  $\sigma_{\beta_b}^2$  corresponding to the different polynomial orders  $b = 0, 1, \dots, p$ .

The decay rate of  $\omega$  shows the performance decaying over time. This concept is intuitive because of the physical nature of sports, an athlete's performance decays with age. The parameter  $\omega$  remains

fixed across athletes because we believe the overall decay of an athlete’s performance should be consistent between athletes in the same sport.

The use of vague and flat priors while not imposing any sport specific knowledge are recommended to generalize the model to all multi-competitor sports. Thus it is recommended to keep priors non-informative or weakly informative. For the intercept and coefficient parameters, the support of the prior must be a continuous distribution that spans all real numbers, e.g., a normal distribution. The distribution of the prior on  $\omega$  should span all non-negative or positive numbers, and could take the form of an exponential or gamma distribution.

### 1.3.1 Adding Covariates

In some cases, (e.g. the case study in Section 1.5) we may be able to collect more player-specific observations. Shown here are two examples of how such covariates may be incorporated into the growth-curve model described in Equation 1.3.

If available, variables such as weight, height, and equipment specifications that change over time could be useful in predicting performance. For example, youth athletes who grow taller more quickly may increase in performance more quickly than expected compared to their slower-growing counterparts. Time-varying, player-specific covariates  $X_{it}$  can be added to the growth curve linearly in the form,

$$\theta_{it} = \alpha_i + \left[ \beta_{i0} + \beta_{i1}t_1^* + \beta_{i2}t_2^* + \beta_{i3}t_3^* + \dots + \beta_{ip}t_p^* \right] e^{-\omega(t-t_{0i})} + \gamma X_{it}, \quad (1.4)$$

where  $\gamma$  is the coefficient on the covariate describing linear changes over time.

Another variable we might include is player age. The older an athlete is, the less time they have in their career and the more quickly their performance decays. However, the impact varies from sport to sport. For example, in target shooting, skill is usually a far superior predictor of performance than age, which is why these sports boast the oldest Olympic athletes with the longest

careers. Conversely, gymnastics is a sport that requires an enormous amount of flexibility and dexterity that tends to be only attainable by young athletes. For simplicity we keep the coefficient for age fixed within each sport. The final model form of  $\theta_{it}$  includes the age  $z_i$  of the athletes first professional appearance:

$$\theta_{it} = \alpha_i + \left[ \beta_{i0} + \beta_{i1}t_1^* + \beta_{i2}t_2^* + \beta_{i3}t_3^* + \dots + \beta_{ip}t_p^* \right] e^{-(\omega_0 + \omega_1 z_i)(t - t_{0i})}, \quad (1.5)$$

where  $\omega_1 \geq 0, \omega_2 \geq 0$  are the linear coefficients describing the decay rate. This form of growth curve is used to fit women's luge data in Section 1.5. Another option may be to make our growth curve a function of age rather than a function of time. Although directly incorporating age into the growth curve would be simple, the length of an athlete's career has a greater impact on performance and age should be incorporated as a covariate in a different way. Also, indexing by length of career instead of age makes it easier to compare career trajectories across athletes of different ages. Our preferred alternative is to include age as a variable in the decay rate of the growth curve.

## 1.4 Model Fitting & Selection

We choose use a Markov Chain Monte Carlo (MCMC) sampling process for model fitting. At a time point  $t$ , we sample from the posterior distribution of our estimate of  $\theta_{it}$  and corresponding growth curve parameters. Rather than a using a point estimate by maximizing the posterior, we choose to use information about the entire posterior. Capturing this uncertainty is vital to making predictions as it informs the user of the reliability of the estimates.

There are several ways to fit this model via MCMC. The most straightforward is through Gibbs sampling where we iteratively sample each of the growth curve parameters,  $\beta_0 = \{\beta_{0i} ; i = 1, \dots, n\}, \beta_1, \dots, \beta_p, \alpha = \{\alpha_i ; i = 1, \dots, n\}, \omega$ , and the intercept and coefficients respective means and variances, conditional on the remaining, current values of the parameters and the observed

outcomes (Geman and Geman, 1984).

1.  $p(\omega|\beta_0, \dots, \beta_p, \alpha, Y)$
2.  $p(\sigma_\alpha^2, \eta_{\beta_1}, \sigma_{\beta_1}^2, \dots, \eta_{\beta_p}, \sigma_{\beta_p}^2|\beta_0, \dots, \beta_p, \alpha, Y)$
3.  $p(\alpha|\beta_0, \dots, \beta_p, \omega, \sigma_\alpha^2 Y)$
4.  $p(\beta_0|\beta_1, \dots, \beta_p, \alpha, \omega, \eta_{\beta_1}, \sigma_{\beta_1}^2, Y), \dots, p(\beta_p|\beta_0, \dots, \beta_{p-1}, \alpha, \omega, \eta_{\beta_p}, \sigma_{\beta_p}^2, Y)$

Each conditional distribution is complicated to draw from directly given the Plackett-Luce likelihood, but can be sampled using Metropolis-Hastings with a normal jump distribution (e.g., Gelman et al., 2013). For our application in Section 1.5, the approach we use to obtain posterior draws is No U Turns Sampling (NUTS; Hoffman and Gelman, 2011). NUTS is an extension to the Hamiltonian Monte Carlo (HMC) sampler which is easily implemented using the RStan software (Neal, 2012).

We select the order of the polynomial via cross-validation (CV) procedures. We focus on CV approaches to optimize our model’s ability to predict athlete strength. In this sense, CV is one approach that focuses on model fitting in sample and lead to over-fitting and poor out-of-sample predictive performance. One approach is leave-one-out cross-validation (LOO-CV) which has historically been applied in the Bayesian context for model selection (e.g., Alqallaf and Gustafson, 2001). The LOO-CV method developed by Vehtari et al. (2016) maintains a Bayesian framework by using the full posterior to select the best model for predictive accuracy. Bayesian LOO-CV has shown in various numerical experiments to outperform other metrics such as the Watanabe-Akaike Information Criterion (WAIC), likelihood ratio test (LRT) and information-based approaches (AIC, BIC, and DIC) in some models. It can perform better when optimizing over prediction accuracy in a Bayesian setting since it utilizes the full posterior (Luo et al., 2017; Piironen and Vehtari, 2017). Bayesian LOO-CV is a technique that proved to perform well in recovering the correct degree polynomial, but other methods may be considered when performing model selection. For example, an alternative to using LOO-CV is the Leave Future Out Cross Validation (LFO-CV).

Instead of optimizing error between the model fit and the data the model is trained on, this method optimizes error between new, future projections (Bürkner et al., 2019). This CV is particularly useful for time-series data with the intent of extrapolating strength parameters into the future.

Starting at a linear model ( $p = 1$ ) we use LOO-CV calculate the difference in expected log pointwise predictive density ( $\Delta$ ELPD) between  $p$  and  $p + 1$  and corresponding standard error. We choose to accept the higher-order polynomial if the ELPD for  $p + 1$  is greater than 2 standard errors away from that of the ELPD for  $p$ . Model selection is made separately per sport because we do not expect every sport to have the same polynomial order in the growth curve. This forward stepwise procedure ensures that we have the simplest model while still capturing the desired flexibility of athlete career trajectories. The forward stepwise procedure remains the same no matter which CV procedure is used.

## 1.5 Results

We apply the model detailed in Section 1.3 to women’s luge data. We observe the athletes’ birth dates so that we can incorporate age as a covariate in our model. The model is fit using NUTS in R Stan. We fit the model with 15,000 iterations with a burn-in of 13,000, for a total of 2,000 draws used for inference. To ensure convergence of the NUTS, we fit the model using three independent chains to calculate the split  $\hat{R}$ . The split- $\hat{R}$  is a modification to the traditional  $\hat{R}$  suggested by Gelman and Rubin (1992) that compares variation between beginning and end of chain to ensure stationary of the convergence (Gelman et al., 2013). As recommended by Gelman et al. (2013), if any parameter is larger than 1.1 then we should not accept the posterior samples as valid. Figure 1.1 shows the split- $\hat{R}$  values for the model fit to women’s luge data. Although some values are moderately high ( $> 1.05$ ) all fall under 1.1. To ensure stability of the NUTS chains in the long term and ensure that there are no hidden modes we run five independent chains for 50,000 iterations. The split- $\hat{R}$  values for the longer chains remain  $\approx 1$  and retain the same posterior for the primary



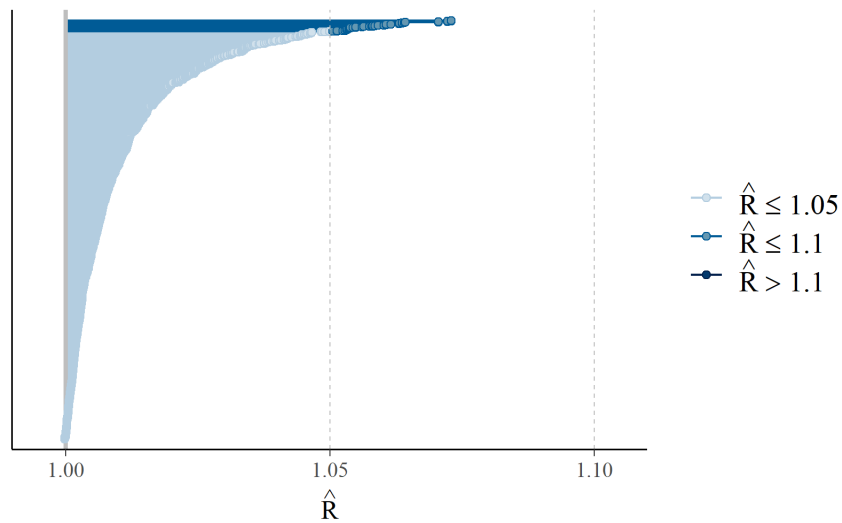


Figure 1.1: The  $\hat{R}$  for posterior draws of all parameters while fitting the model described in Section 1.5.

parameters in our analysis.

### 1.5.1 Case Study: Women’s Luge

We apply this method to women’s luge athletes. Women’s luge has been a participating sport in the Winter Olympics since 1964. Luge can be raced as a duo, but for the sake of simplicity, we only focus on individual athlete events. During a single women’s luge event, each athlete gets a pre-determined number of chances to sled down the same designated track. The women are timed from the time they cross the starting line at the top of the track to the time they cross the finish line at the bottom. The final ranking is determined by the cumulative time of all runs. Since the final result is a rank ordering of all participating athletes, luge is considered a multi-competitor sport.

The data set we used was provided by the US Olympic Committee and consists of results from 142 events from the 2004 to the 2017 winter seasons of 166 women. All recorded events are professional-level, single sled women’s luge events. Along with complete ranking results from each



Figure 1.2: Erin Hamlin participating in the single women’s luge event in the 2014 Winter Olympics at Sochi.

event, we have participating athletes’ birth dates. Figure 1.3 shows the distribution of the observed length of the athlete’s careers and the age of their first appearance in the data set.

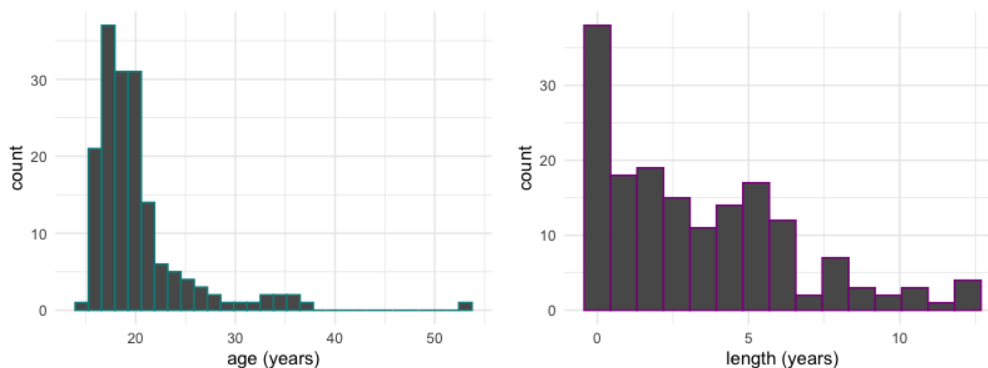


Figure 1.3: The distribution of women luge athlete’s age at the start of their career (top) and the length of their career (bottom).

We separate the events by dividing them into periods. The dates are divided into four evenly spaced periods per year starting on January 1, April 1, July 1, and October 1. The data set spans 50 periods total starting with October 1, 2004 and ending with March 31, 2017. It also features roughly 5-6 events per period in the winter periods (October - March) and no events per period in the off-seasons (April - September). We choose three month periods as they are long enough to observe an athlete perform at least once within the competition but short enough that athletes’ ability does not change significantly within the period. This selection enables us to see details of

changes in ability across a typical athlete's career. We calculate this orthogonal bases  $t_1^*, \dots, t_p^*$  using the **poly** function in **R** (Chambers and Hastie, 1992).

We include age as a covariate thus fitting the model in Equation 1.5. As mentioned in Section 1.3, each polynomial parameter varies per athlete, but is drawn from the same distribution as the parameters of the same degree. The decay rate parameters  $\omega_0, \omega_1$  priors are Exponential distributions since they are non-negative.

$$\begin{aligned}\alpha_i | \sigma_\alpha^2 &\sim \text{Normal}(0, \sigma_\alpha^2) \\ \beta_{bi} | \eta_{\beta_b}, \sigma_{\beta_b}^2 &\sim \text{Normal}(\eta_{\beta_b}, \sigma_{\beta_b}^2) \\ \omega_0, \omega_1 &\sim \text{Exp}(10)\end{aligned}$$

The hyper-priors are kept vague and flat. The mean parameters  $\eta_\beta$  can take on any value. However, we suspect the variance parameters to be relatively small so we make their uniform parameters sufficiently wide to encompass any feasible value:

$$\begin{aligned}\sigma_\alpha &\sim \text{Uniform}(0, 10^4) \\ \eta_{\beta_b} &\sim \text{Normal}(0, 100) \\ \sigma_{\beta_b} &\sim \text{Uniform}(0, 10^3).\end{aligned}$$

Model selection is conducted via LOO-CV. The results of the model selection are shown in Table 1.1. The final model is determined to have  $p = 3$ :

$$\theta_{it} = \alpha_i + \left[ \beta_{i0} + \beta_{i1}t_1^* + \beta_{i2}t_2^* + \beta_{i3}t_3^* \right] e^{-(\omega_0 + \omega_1 z_i)t}.$$

Figure 1.4 shows the fitted player strength parameters of four of the top athletes in luge: Alex Gough (CAN), Natalie Geisenberger (GER), Summer Britcher (USA) and Tatjana Hüfner (GER).

Table 1.1: Results of LOO-CV on Women's Luge Data

Model 1	Model 2	$\Delta$ ELPD	SE
$p = 1$	$p = 2$	18.8	7.1
$p = 2$	$p = 3$	29.6	8.2
$p = 3$	$p = 4$	-2.8	6.0

These four athletes have started competing at different dates so they are at different stages of their career. Hűfner, the eldest, is beginning to decline in performance, whereas Britcher is the youngest athlete and has a sharp increase in performance over the past three years. Geisenberger has had a steady career and always performed at the top, whereas Gough took several years to reach a similar strength.

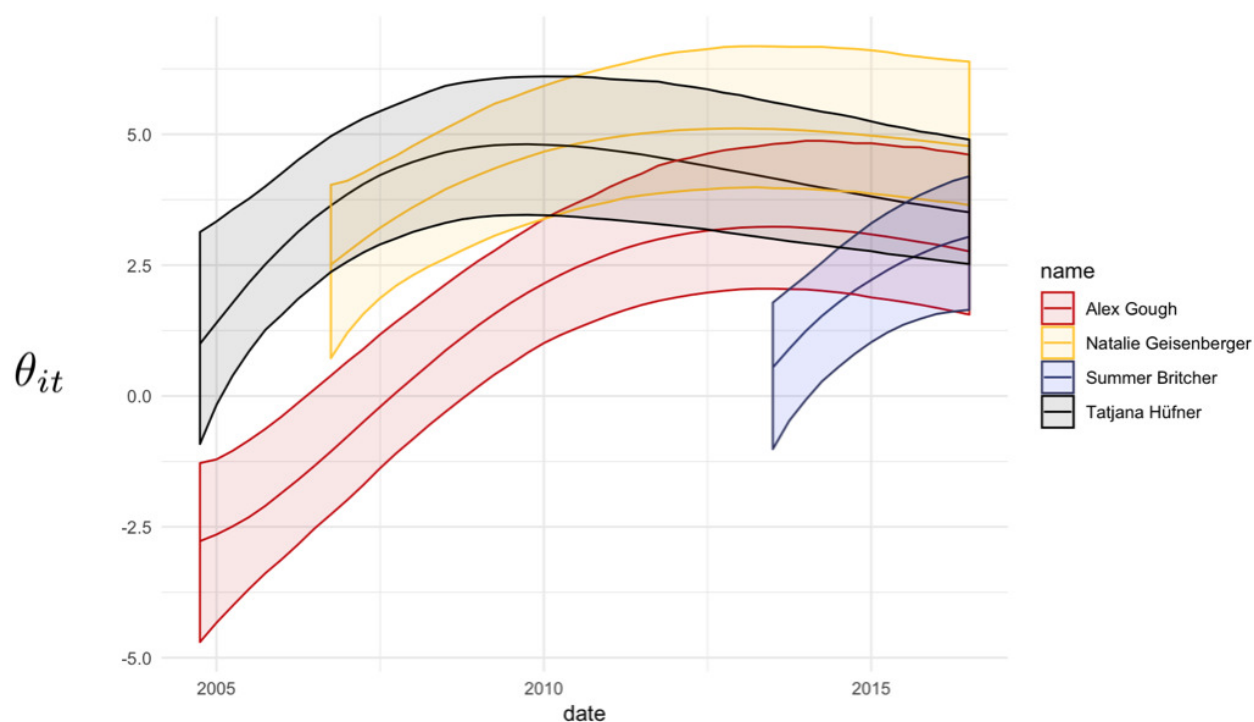


Figure 1.4: Growth-curve model fit to observed time range for four of the top women's luge athletes. Plotting the growth curves against one another lets us compare the different levels of performance between the athletes at a given time.

Figure 1.5 shows the same trajectories but aligned by years since the start of each athlete's

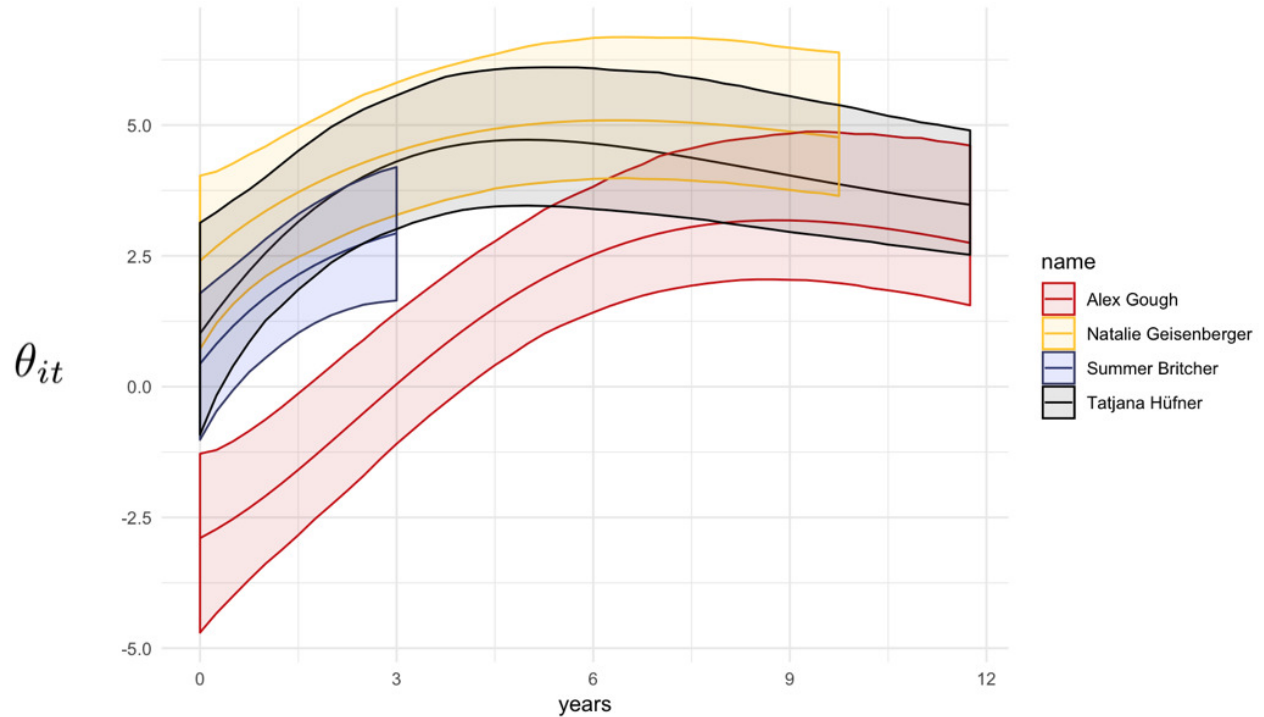


Figure 1.5: The same growth curve fits as in Figure 1.4 but aligned at the beginning of the career.

career  $t - t_{0i}$ . This view allows us to compare trajectories between athletes while they are at the same points in their career. As we can see, the top luge athletes share similarly shaped trajectories. One point of interest for coaches or team managers would be comparing younger players' career trajectories to the trajectories of their more established counterparts. In Figure 1.6, we show Britcher's career projected out seven years compared to Geisenberger's current career trajectory. By plotting the trajectories side by side, aligned at the beginning of their careers, we can see that they have similarly shaped trajectories, but that Britcher never reaches the level of performance of Geisenberger.

We can also use our model fit to understand the general trajectory of athletes in a particular sport. To do so, we take advantage of the hierarchical structure of the polynomial coefficients and

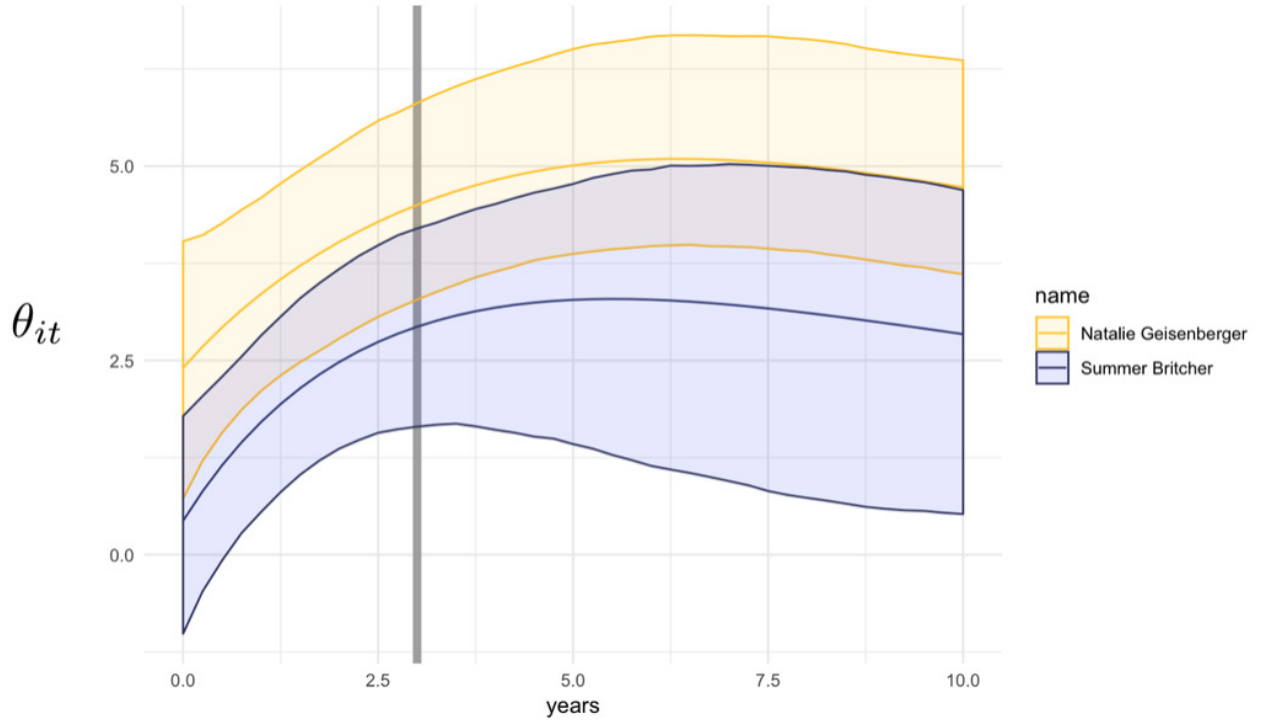


Figure 1.6: Career trajectory of Natalie Geisenberger next to the projected career trajectory of Summer Britcher aligned at the start of their careers. The grey vertical line shows the current point whereas to the right is the predicted strength for Britcher.

construct a curve using the estimated means,

$$\bar{\theta} = \left[ \eta_0 + \eta_1 t_1^* + \dots + \eta_p t_p^* \right] e^{-\omega(t-1)} . \quad (1.6)$$

In the women's luge case we incorporate the average starting age across all players  $\bar{z}$ ,

$$\bar{\theta} = \left[ \eta_0 + \eta_1 t_1^* + \eta_2 t_2^* + \eta_3 t_3^* \right] e^{-(\omega_0 + \omega_1 \bar{z})(t-1)} . \quad (1.7)$$

Figure 1.7 compares the mean curve between women's luge and a second example, men's slalom skiing. The curve tells us that the average luge athlete is expected to increase their performance from the start of their career for about 5 years, at which time the athlete reaches a plateau. This trajectory is different than that of men's slalom athletes who are expected to begin declining in

performance after 6-7 years. This difference may be because slalom skiing is a more-physical sport; therefore, longer careers are less likely. This contrasts luge, which is a more-skilled sport where age is a less of an important factor in performance. Although this curve cannot explain such phenomena, it is still useful in comparing the common shape of the trajectory between the two sports.

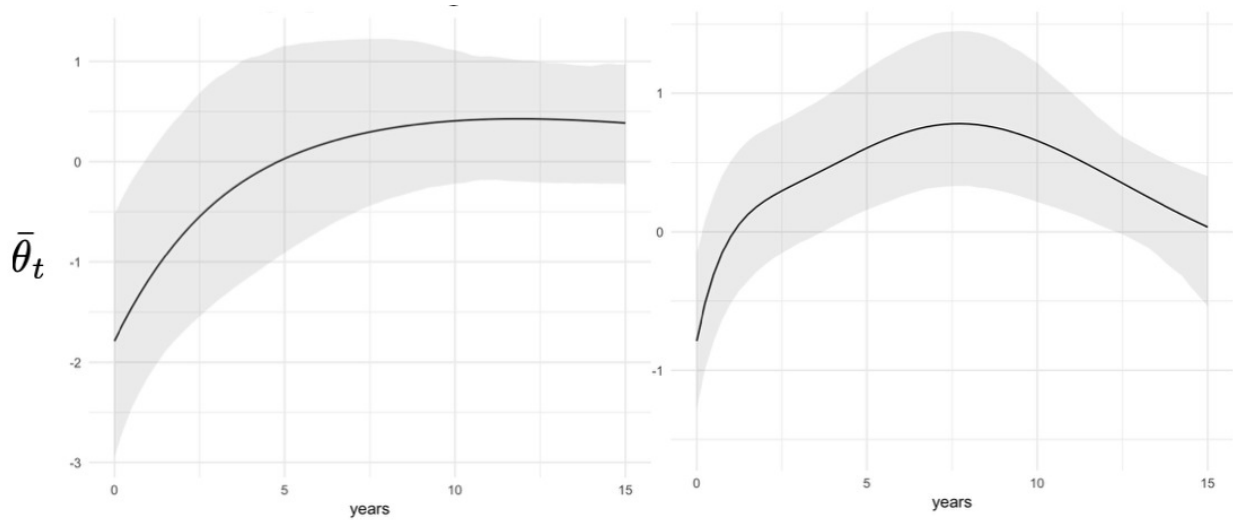


Figure 1.7: The mean curves for women’s luge (left) and men’s slalom (right). By looking at the mean curves we can compare the typical trajectory of athletes between two sports.

## 1.5.2 Evaluation

To evaluate the predictive ability of our model, we use a weighted average of the Spearman rank correlation (Spearman, 1904) between our projected latent performance and the true race results. This method is also used to evaluate the GH model in Glickman and Hennessy (2015).

The weighted average of the Spearman rank correlation at predicted time  $T$  is:

$$\rho_W = \frac{\sum_{k=1}^{K_T} (m_{kT} - 1) \hat{\rho}_k}{\sum_{k=1}^{K_T} (m_{kT} - 1)},$$

where  $K_T$  is the number of races at time  $T$  and  $m_{kT}$  is the number of athletes competing in race

<i>Projection Interval</i>	<i>Prediction Time Interval</i>	
	1 Year	2 Years
7-9 Months	0.651	0.599
10-12 Months	0.639	0.584
19-21 Months	–	0.591
22-24 Months	–	0.600

Table 1.2: Estimated weighted Spearman correlation for predictions removing the last one year and last two years. Predictions are made in each of the two periods with observations per year to calculate the estimated weighted Spearman correlation.

$k = 1 \dots K_T$ . We evaluate  $\rho_W$  at each iteration of the posterior draw. To estimate  $\hat{\rho}_k$  we first evaluate  $\theta_{iT}$  at desired time  $T$ , for each player that appears in race  $k$  using the posterior draws of the growth curve parameters. We then draw the latent performance  $Y_{iT}$  from the Gumbel distribution with location  $\theta_{iT}$ . We can then calculate the Spearman rank correlation between the true race results and the sampled latent variables to obtain  $\hat{\rho}_k$ .

Table 1.2 shows the evaluation for the women’s luge example in several settings. First, we remove only the final year of results and fit the growth-curve model. We use this model built with missing data to predict the results for matches in the two out of sample observed periods of that year<sup>1</sup>. We then remove the final two years of results to fit the data and predict the results for matches in the four observed periods across both years. Figure 1.8 shows the estimated weighted Spearman correlation and 95% credible interval across the four periods within the two years of withheld information. Notice that the correlation does not drop, even evaluated out greater than four periods, showing that our methods are not just valuable for current predictions, but can extrapolate far beyond the end of the observed data. We compare the Spearman correlation to that calculated for evaluating GH (the dotted line in Figure 1.8). Our model performs similarly across all predictions, even up to 8 periods.

<sup>1</sup>Since luge is a winter sport, athletes only compete in the winter season which takes place in two out of the four periods each year



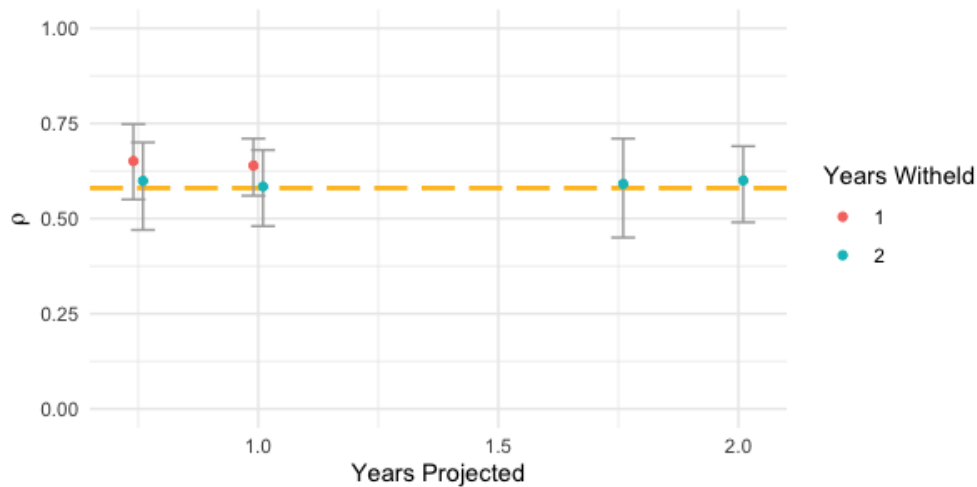


Figure 1.8: The estimate of the weighted Spearman correlation and 95% CI. The dotted line shows the performance of the GH model.

## 1.6 Conclusion

The approach presented in this chapter successfully models an athlete’s strength in multi-competitor sports, allowing the user to compare current and future athlete performance. The growth curve was constructed to be as flexible as needed to fit even the most irregular career trajectories. Because there are no sport-specific assumptions made in the creation of the Plackett-Luce likelihood or growth curve, this method is generalizable to any individual sport.

We recognize that since we only observe athletes while they perform in professional tournaments, our method is susceptible to some selection bias since we do not include those athletes who do not qualify for every event or retire early. This likely causes bias in our predictions of the athlete strength. It may also cause an underestimation of uncertainty over long-term predictions.

This method is intended to be applied to a single sport at a time, but by comparing the distributions of the random effects  $(\alpha, \beta)$  and the decay rate  $(\omega)$  we can make comparisons between sports. We also introduce the mean curve, which is an estimate of the approximate shape of a career trajectory for an average athlete in the sport. The characteristic mean curves could also be

used to compare the typical career trajectory across sports for insight about when we expect players to reach their peak performance, on average. This estimate is particularly useful for athletes who have few observed appearances in games or matches, providing little information about their career trajectory.

We can extend this basic model into a discrete mixture model of different average career trajectories. This exercise helps us find similarities between athletes and elicit potential clusters or groups of career trajectories. Clustering these trajectories could give insight into what makes specific athletes perform in a particular way throughout their careers. We visit clustering as an extension in Chapter 2.

## Chapter 2

# Two-Step Model for Predicting Athlete Strength

Athlete rating systems are methods for quantifying the relative abilities of athletes within a particular sport. Ratings are defined as an estimate of an athlete's strength in a particular moment in time. In general, athletes with higher ratings are more likely to defeat those with lower ratings in direct competition. However, many of these rating systems do not inherently model athlete's strength over time. As a result, they are poor at forecasting future athlete ability. Chapter 1 applies a probabilistic rating system to estimate ability over time in individual, multi-competitor sports. As there are a wide variety of types of sports, including both multi-competitor and head-to-head sports, devising a system to jointly model rating over time for every possible scenario would be time consuming. In this chapter, we focus on fitting a model to pre-estimated ratings from any type of sport so we can forecast athlete abilities, cluster career trajectories, or undertake a variety of other procedures that are not inherent to existing rating systems.

## **2.1 Introduction**

Methods that both quantify athlete performance and model time variation of the performance, such as the model we propose in Chapter 1, can be computationally complex and expensive. We can simplify the process by approaching it in two steps. The first step is to build a static rating system over a fixed period that quantifies relative athlete ability as measured by the outcomes of games or matches. The second step is to fit a model to the estimates on athlete's strength over time. Rating systems have become popular among sports coaches, analysts, and fans, demonstrated by the success and widespread use of the Elo rating system across many different sports and the adaptation of the Glicko rating by chess leagues. Ratings are useful in comparing players against one another. If we know the athletes' ratings at a given time, we can compute the probability of a particular outcome for a game or match. Rating systems update with current ratings over time by relying on past competition results. However, neither can rating systems alone be used to forecast future results, nor can they be used for post-processing techniques such as clustering career trajectories.

In order to use these models for forecasting and post-processing, we need to build a model to help understand how model-generated ratings vary over time. To do this, we build upon the same concept of growth curves presented in Chapter 1 while applying it to estimated ratings to make inferences about the functional relationship of ratings over time. A growth curve is a parametric model of how a quantity changes over time. Assumptions put in place by using growth curves can reduce error and over-fitting introduced by non-parametric models. Furthermore, growth curves can be defined in such a way that their parameters can describe physical attributes, increasing their interpretability. The assumption of a growth-curve model helps describe how individual skills or abilities change over time. This allows us to forecast future ratings and compare career trajectories through clustering.

In addition to comparing one athlete's future performances against another, we compare trends

across sports to gain a better idea of when we expect athletes to reach their peak performance and how quickly we expect their abilities to decay. Using a functional form of ratings over time can be helpful in a variety of post-processing analyses, including clustering athletes based on career trajectories. Clustering methods are an excellent way to explore potential types of athletes' careers and learning what characteristics athletes share within these types of careers. This type of analysis is not necessarily helpful for prediction or inference, but it can give us detailed insight into the characteristics of the athletes.

Multi-phase inference is a common way to simplify complicated processes across a variety of applications. Using a two-step procedure, we extract the benefits of imposing a growth-curve model to describe the time-varying nature of player ratings, while gaining the ability to use it with rating systems across head-to-head and multi-competitor sports. Multi-phase inference is useful in sports analytics because those who create and utilize sports analytics models come from a wide variety of backgrounds, from amateur fans to expert statisticians. Furthermore, not everyone has access to all sources of data since professional teams own some data, and some data requires complicated pre-processing. Using a multi-phase method can help alleviate these issues and lead to better inferences on the parameters of interest. [Bornn et al. \(2017\)](#) discuss several multi-step procedures that first pre-process raw player tracking data and use the results for more advanced inference, such as clustering athlete movements. Frequently, forecasting techniques use a combination of raw observations and pre-estimated statistics. In [Baker and McHale \(2013\)](#), the authors use the point-spread and over-under from the betting line. In [Silver \(2019\)](#), the author uses defensive/offensive ratings and other homegrown statistics in their forecasting. In this use case, we will use the same growth curve model in Chapter 1 but extend its use by applying it to a variety of pre-existing rating systems using a multi-phase process.

Section 2.2 describes how to fit the proposed growth curve method as a post-processing step to rating data. The model inference and model selection process is discussed in Section 2.3. Section 2.4 applies this two-step method to men's slalom data and discusses how fitting the growth-curve

model is used for prediction and Section 2.5 discusses how these growth curves can be used for clustering career trajectories.

## 2.2 Methods

We suggest a two-step process to build a model that estimates and predicts athlete performance over time. The first step is to use an existing rating system to model athlete rating at equal time intervals. The second step is to fit a nonlinear mixed-effects regression model to these rating estimates and corresponding estimated uncertainty. In this section, we first describe the Elo, Glicko, and GH rating systems and then show that we can model the time-varying nature of these estimated ratings using a growth curve.

### 2.2.1 Ratings

During  $T$  discrete time periods we observe  $n$  athletes that participate in at least one observed competition during this time period. Each athlete  $i = 1, \dots, n$  has an ability represented by the parameter  $\theta_{it}$  that indicates the competitor strength at a given time period  $t = 1, \dots, T$ . For each period  $K_t$  competitions take place wherein competition  $k = 1, \dots, K_t$ ,  $m_{kt}$  competitors participate.

In head-to-head sports, we are interested in the expected outcome of a particular game or match, or the probability that one athlete defeats another athlete. Let  $Y_{ij,tk}$  be a binary variable that indicates the outcome between athletes  $i$  and  $j$  for competition  $k$  at time  $t$ :  $Y_{ij,tk} = 1$  if  $i$  defeats  $j$  and 0 otherwise. Let  $\theta_{it}$  be the relative latent ability or skill of athlete  $i$  in time  $t$ . At a certain time  $t$  and competition  $k$  the likelihood of athlete  $i$  defeating athlete  $j$  is distributed as:

$$Y_{ij,tk} | p_{ij,tk} \sim \text{Bernouli}(p_{ij,tk}), \quad (2.1)$$

where  $p_{ij,tk} = P(Y_{ij,tk} = 1)$ . The Bradley-Terry model assumes a logistic function such that  $\text{logit}(p_{ij,tk}) = \theta_{it} - \theta_{jt}$ . Thus it follows that,

$$P(Y_{ij,tk} = 1 | \theta_{it}, \theta_{jt}) = \frac{e^{\theta_{it}}}{e^{\theta_{it}} + e^{\theta_{jt}}} = \frac{1}{1 + e^{\theta_{jt} - \theta_{it}}}. \quad (2.2)$$

(Bradley and Terry, 1952). The Bradley-Terry model is the basis of both the Elo and Glicko rating systems.

### Elo

Elo ratings are a Bradley-Terry based system that was originally used for ranking chess players by skill based on their performance (Elo, 1978). The Elo rating system is designed to estimate athlete ability,  $\theta_{it}$ , through an estimated rating  $\mu_{it}$ . Conventionally the Elo system uses a different scale for player strengths in which the estimated probability of athlete  $i$  defeating athlete  $j$  is,

$$E_{ij} = \frac{1}{1 + 10^{(\mu_{jk} - \mu_{ik})/400}}. \quad (2.3)$$

Elo accounts for change in estimated ratings by discrete updates after each competition <sup>1</sup>. The updating procedure is

$$\mu_{i,k+1} = \mu_{ik} + \kappa(Y_{ij,k} - E_{ij}) \quad (2.4)$$

where  $\kappa$  is a user-defined tuning parameter. It could be a factor that correlates with the difficulty of the tournament. The ratings in the Elo system are balanced over time, that is if one player rating increases by  $\delta$ , then their opponent's rating decreases by  $\delta$ .

### Glicko

Glicko was proposed by Glickman (1999) to incorporate variability in parameter estimates over time to estimate the relative strength of athletes. The Glicko model assumes a Bradley-Terry

<sup>1</sup>Typically Elo ratings are updated after each performance  $k$  rather than over time  $t$

implementation but assumes the evolution of  $\theta_{it}$  over time through a stochastic process. The model assumes the future ratings are related to the past ratings by the following relationship:

$$\theta_{i,t+1}|\theta_{it}, \nu_{i,t+1}^2 \sim \text{Normal}(\theta_{it}, \nu_{i,t+1}^2). \quad (2.5)$$

The Glicko system is an improvement on the Elo system because it accounts for “reliability” (Glickman, 1999). That is, if a player competes more frequently, their rating is estimated with less error: i.e. we are more confident in the estimate than in an estimate for a player who competes infrequently. Similar to the Elo rating system, we estimate the athlete’s  $i$  ability parameter at time  $t$  by the rating  $\mu_{it}$ . Glicko also updates an estimated uncertainty on our rating estimates through the colloquially named “ratings deviation” (RD). We can utilize this uncertainty in our projections of skill to give a more reliable prediction interval. We denote the RD at time  $t$  for athlete  $i$  as  $\nu_{it}$ . The expected score of player  $i$  defeating player  $j$  is

$$E_{ij} = \frac{1}{1 + 10^{g(\sqrt{\nu_{it}^2 + \nu_{jt}^2})(\mu_{jk} - \mu_{ik})/400}}, \quad (2.6)$$

where  $g(\nu) = (1 + 3q^2\nu^2/\pi^2)^{-1/2}$  and  $q = \ln 10/400$ . With Glicko, we update both the estimated rating  $\mu_{it}$  and the RD  $\nu_{it}$  at each rating period  $t$  after recording results against  $m_{tk}$  opponents across different competitions: <sup>2</sup>

$$\mu_{i,t+1} = \mu_{it} + \frac{1}{\nu_{it}^{-2} + d^{-2}} \sum_{j=1}^m g(\nu_j)(Y_{ij,t} - E_{ij}) \quad (2.7)$$

$$\nu_{i,t+1} = (\nu_{it}^{-2} + d^{-2})^{-1/2} \quad (2.8)$$

$$d^2 = \left[ q^2 \sum_{j=1}^m (g(\nu_j))^2 E_{ij}(1 - E_{ij}) \right]^2, \quad (2.9)$$

where  $d^2$  is an approximation to the likelihood-based variation when moving from period to period. Unlike the Elo system, the Glicko system’s ratings are not balanced over time. An increase in one

---

<sup>2</sup>The length of time in a period is determined by the user. Glickman recommends to set periods where athletes average about 5 to 10 performances per period (*The Glicko System* by Mark Glickman : <http://www.glicko.net/glicko/>).



player's ratings does not indicate a mirrored decrease in their opponent's ratings.

### Glickman & Hennessey (GH)

A generalization of this method to more than two players is to estimate the skill parameter for multi-competitor sports. The Plackett-Luce model assumes an extreme value distribution on the [Luce \(1959\)](#) logit model for rank orderings ([Plackett, 1975](#)). The fully Bayesian Plackett-Luce model is discussed in more detail in Chapter 1. We use the Plackett-Luce-based rating system presented by [Glickman and Hennessey \(2015\)](#), now referred to as the GH model. Here we assume  $Y_{itk}$  is the latent performance, distributed conditionally on athlete ability  $\theta_{it}$ , by the maximum value distribution, the Gumbel distribution. For a particular observed rank ordering of  $m_{kt}$  athletes in competition  $k$  during period  $t$  we formulate the likelihood as

$$L_{kt} = P(Y_{1tk} > Y_{2tk} > \dots > Y_{m_{kt}tk} | \boldsymbol{\theta}_t) = \prod_{i=1}^{m_k-1} \frac{\exp(\theta_{it})}{\sum_{\ell=i}^{m_k} \exp(\theta_{\ell t})}. \quad (2.10)$$

In Chapter 1, we perform inference using a fully Bayesian version of the GH model. However, this can be too computationally expensive to do when updating ratings at every period. Instead, [Glickman and Hennessey](#) describe a rating system similar to Glicko and Elo, that uses a Newton-Raphson approach to update the estimated ratings  $\boldsymbol{\mu}_t = (\mu_{1t}, \dots, \mu_{nt})$  and estimated standard deviation  $\boldsymbol{\nu}_t = (\nu_{1t}, \dots, \nu_{nt})$ . We assume that at a given time  $t$ , the prior distributions for the ability of competitors  $\boldsymbol{\theta}_t = (\theta_{it}, \dots, \theta_{nt})$  are independently distributed as,

$$\theta_{it} \sim \text{Normal}(\mu_{it}, \nu_{it}^2) \quad (2.11)$$

with the likelihood on the contribution from competition  $k$  given Equation 2.10. Given the estimated ratings and standard deviations, we can write the log posterior of  $\boldsymbol{\theta}_t$  as

$$\log p(\boldsymbol{\theta}_t | \boldsymbol{\mu}_t, \boldsymbol{\nu}_t^2) = c_0 + \log p(\boldsymbol{\theta}_t | \boldsymbol{\mu}_t, \boldsymbol{\nu}_t^2) + \sum_{k=1}^{K_t} \log L_{kt}, \quad (2.12)$$

where  $c_0$  is a normalizing constant. Implementing optimization through the Newton-Raphson method <sup>3</sup>, we obtain the mode which is the approximate normal posterior mean,  $\boldsymbol{\mu}_t^*$ . We also find the second derivative matrix evaluated at the mode, then take the negative of the matrix inverse to approximate the posterior covariance matrix to obtain  $\boldsymbol{\nu}^{2*}$ . Moving between periods  $t$  to  $t + 1$  we assume,

$$\theta_{i,t+1} | \theta_{it}, \tau^2 \sim \text{Normal}(\theta_{it}, \tau^2) \quad (2.13)$$

$$\Rightarrow \theta_{i,t+1} | \mu_{it}^*, \nu_{it}^{2*}, \tau^2 \sim \text{Normal}(\mu_{it}^*, \nu_{it}^{2*} + \tau^2) \quad (2.14)$$

$$\sim \text{Normal}(\mu_{i,t+1}, \nu_{i,t+1}^2). \quad (2.15)$$

Ties are accommodated using an approximation that incorporates the outcomes in the likelihood as if all athletes of the same rank outperform each other (Breslow and Crowley, 1974). Instead of estimating through a fully Bayesian process, its quicker to treat  $\tau$  as fixed in advance and estimate the parameter by optimizing a predictive fit criterion (e.g., Glickman and Hennessy, 2015).

Estimating the ratings at each time interval is the first step in the two step method. The appropriate rating method to use depends on the context. The GH rating system should be used in multi-competitor competitions and the Glicko or Elo rating system should be used in head-to-head competitions. We recommend using Glicko over Elo in the second step since it also gives estimates of the uncertainty on the strength estimates, whereas Elo does not. However, Elo is a more widely used rating system thus may be the only system available or more interpretable depending on the source of the ratings and the audience. The ratings procedure for Glicko and GH can be adapted to smooth parameters based on earlier results using Rauch-Tung-Streibel (RTS), a Kalman smoother (Rauch et al., 1965). These are shown to give the best estimate of measuring strength at a given time point (Glickman and Hennessy, 2015). The same cannot be done for Elo ratings.

---

<sup>3</sup>Described in detail in Appendix A of Glickman and Hennessy (2015)

## 2.2.2 Growth Curve

In each of the cases mentioned in Section 2.2.1, we end up with an estimate of athlete rating measured at different time points,  $\mu_{it}$ . The rating systems provide sequential estimates of relative athlete performance over defined periods within a fixed time interval. We propose fitting a non-linear mixed-effects regression model to the estimated ratings to predict future ratings. We treat our estimated ratings as fixed and assume the distribution

$$\mu_{it} = g_i(t) + \epsilon_{it} , \quad (2.16)$$

where  $\epsilon_{it} \sim \text{Normal}(0, \sigma_{it}^2)$ . This application of the growth curve is similar to the autoregressive process assumed in the Glicko and GH models shown in Equation 2.5. The Elo rating system does not provide a way to estimate the uncertainty on the rating estimates over time. The variance does not change by individual or over time:

$$\sigma_{it}^2 = \sigma_{\theta}^2 . \quad (2.17)$$

With the GH and Glicko models, we estimate the uncertainty of our rating estimates over time by updating the RDs. We incorporate this into the variance term as,

$$\sigma_{it}^2 = \sigma_{\theta}^2 + \nu_{it}^2 , \quad (2.18)$$

where  $\nu_{it}$  is the RD estimate and  $\sigma_{\theta}^2$  is the regression error, consistent across athletes and time.

The growth curve follows the same form as in Chapter 1 as it was built to be representative of how ratings change over time:

$$g_i(t) = \alpha_i + \left[ \beta_{i0} + \beta_{i1}t_1^* + \beta_{i2}t_2^* + \beta_{i3}t_3^* + \dots + \beta_{ip}t_p^* \right] e^{-\omega(t-t_{i0})} , \quad (2.19)$$

where  $\alpha_i$  and  $\beta_i = \{\beta_{i0}, \beta_{i1}, \beta_{i2}, \dots, \beta_{ip}\}$  vary per individual and  $\omega > 0$  is fixed across individuals.

The polynomial component is meant to provide flexibility by modeling a variety of career shapes. The decay parameter reflects the assumption that an athlete’s strength will decay over time. The hierarchical parameters will give each individual their own trajectory shape. The parameters will be related, which is useful for estimating the performance of athletes for whom we do not have many observations, but they will not be identical. To prevent correlation between polynomial terms, we instead use orthogonal polynomials  $t^*$  of  $t - t_{0i}$  where  $t_{0i}$  is the period of first appearance for athlete  $i$ . We use the raw  $t - t_{0i}$  for the exponent since orthogonal transformations are not guaranteed to be non-negative. Details on orthogonalization are in Appendix A.

The intercept and coefficient terms of the growth curve are drawn from independent normal distributions:

$$\alpha_i | \sigma_\alpha^2 \sim \text{Normal}(0, \sigma_\alpha^2) \tag{2.20}$$

$$\beta_{ib} | \eta_{\beta_b}, \sigma_{\beta_b}^2 \sim \text{Normal}(\eta_{\beta_b}, \sigma_{\beta_b}^2) . \tag{2.21}$$

The decay rate prior is set to have a mean of 10, a value much larger than we expect the parameter to take thus keeping the prior relatively flat:

$$\omega_0, \omega_1 \sim \text{Exp}(10) .$$

The priors on the mean and variances of the parameter are flat and vague. The bounds are chosen to be large enough to contain any feasible value for the mean and variance terms:

$$\sigma_\alpha \sim \text{Uniform}(0, 10^4) \tag{2.22}$$

$$\eta_{\beta_b} \sim \text{Normal}(0, 10^2) \tag{2.23}$$

$$\sigma_{\beta_b} \sim \text{Uniform}(0, 10^3) . \tag{2.24}$$

The constraint we impose on  $\sigma_\theta^2$  incorporates the value of  $\sigma_\alpha^2$  to ensure the scales are roughly the

same. This helps ensure the identifiability of the variance of the intercept in the growth curve and the variance of the rating, as the terms are inversely proportional to one another:

$$\frac{\sigma_{\theta}^2}{\sigma_{\alpha}^2} \sim \text{Gamma}(100, 100) .$$

As mentioned in Chapter 1, we can also easily add covariates, such as the athlete age at the start of their career  $z_i$ , as part of the decay coefficient:

$$g_i(t, z_i) = \alpha_i + \left[ \beta_{i0} + \beta_{i1}t_1^* + \beta_{i2}t_2^* + \beta_{i3}t_3^* + \dots + \beta_{ip}t_p^* \right] e^{-(\omega_0 + \omega_2 z_i)(t - t_{0i})} . \quad (2.25)$$

We assume that the older the athlete, the more quickly their physical abilities will decay. If we are able to record athlete specific, time-varying, covariates (e.g., weight, height ,etc.) we can incorporate them into the model additively:

$$g_i(t, X_{it}) = \alpha_i + \left[ \beta_{i0} + \beta_{i1}t_1^* + \beta_{i2}t_2^* + \beta_{i3}t_3^* + \dots + \beta_{ip}t_p^* \right] e^{-\omega(t - t_{0i})} + \gamma X_{it} . \quad (2.26)$$

## 2.3 Model Fitting

Although we are fitting a non-linear model we can use MCMC techniques similar to classic, linear regression (e.g., [Gelman et al., 2013](#)). Using a Gibbs sampler we can iteratively sample from the conditional distributions for the growth curve parameters  $\boldsymbol{\beta}_0 = \{\beta_{i0}; i = 1, \dots, n\}, \dots, \boldsymbol{\beta}_p, \boldsymbol{\alpha} = \{\alpha_i; i = 1, \dots, n\}, \omega$ , the growth curve variance  $\sigma_{\theta}^2$  and the estimated ratings  $\boldsymbol{\mu}_{it} = \{\mu_{it}; i = 1, \dots, n, t = 1, \dots, T\}$  ([Geman and Geman, 1984](#)).

1.  $p(\omega | \boldsymbol{\beta}_0, \dots, \boldsymbol{\beta}_p, \boldsymbol{\alpha}, \boldsymbol{\mu}, \sigma_{\theta}^2)$
2.  $p(\boldsymbol{\alpha}, \boldsymbol{\beta}_0, \dots, \boldsymbol{\beta}_p | \omega, \boldsymbol{\mu}, \sigma_{\theta}^2)$
3.  $p(\sigma_{\theta}^2 | \boldsymbol{\beta}_0, \dots, \boldsymbol{\beta}_p, \omega, \boldsymbol{\alpha}, \boldsymbol{\mu})$

The conditional distribution at steps 1 and 3 can be sampled using Metropolis-Hastings using a Normal jump distribution for step 3, and a Gamma distribution for step 1 (e.g., [Gelman et al., 2013](#)). In step 2 all intercept and coefficient parameters can be combined into one step by drawing them jointly using a multivariate normal distribution. This sampler could be implemented directly, but as in Chapter 1, we found using No U Turns Sampling (NUTS) via the RStan software to be an effective way to sample from the posterior ([Hoffman and Gelman, 2011](#)).

Model selection to determine the maximum order or polynomial for the growth curve should be performed in the same way as in Chapter 1. We use forward stepwise model selection to encourage simpler models to avoid overfitting. We found using Bayesian leave-one-out cross validation (LOO-CV) for model selection to be sufficient in selecting the correct model in our case though other methods are possible ([Vehtari et al., 2016](#)). Refer to Section 1.4 in Chapter 1 for more details on these CV methods.

## 2.4 Results

The model is applied to men’s skiing data to compare athletes’ career trajectories based on observed outcomes of races. We validate our methods by seeing how well we predict future, out-of-sample ratings using the estimates of the parameters of our growth curves. Using the functional form of the estimated career trajectories, we perform a clustering analysis as a post-processing step to fit the model as a tool for describing the results.

### 2.4.1 Case Study: Men’s Slalom

Slalom skiing is a type of alpine skiing in which a skier navigates turns around a sequence of poles or gates along a 600 to 700 foot-long slope. During a single event, each athlete has one chance to ski the course, and the athletes are ranked according to how quickly they complete the course.

Thus, slalom skiing is a multi-competitor sport. Slalom skiing has been an alpine skiing event in the Winter Olympics since 1936. We apply our model to 403 male skiers to estimate ratings from 2004 to 2017. We only include events at the professional level for individual men’s slalom skiing. The average number of seasons in which a skier competes professionally is around 3. Figure 2.1 (right) shows the distribution of the length of the athletes’ careers. We use age as a covariate in our growth curve in the same way we use age in the Luge example in Chapter 1. The distribution of the athletes’ ages at the start of their career is in Figure 2.1 (left). The average age at the start of a skier’s career is around 23. The United States Olympic Committee provided race results and player birth dates.

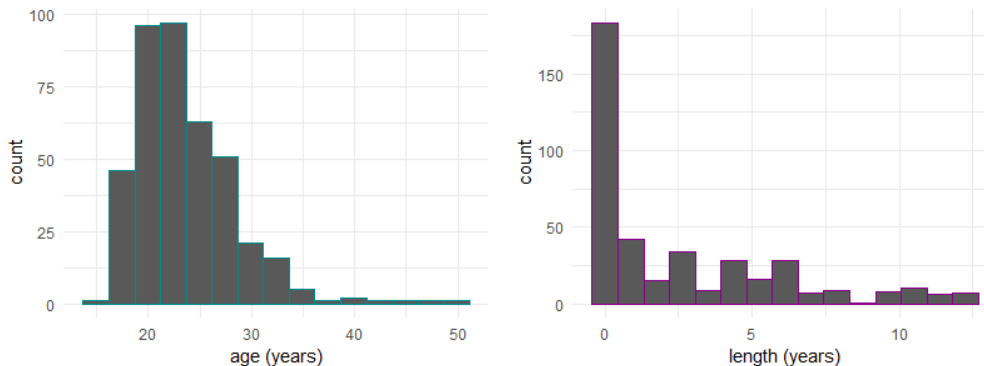


Figure 2.1: The distribution of men slalom athlete’s age at the start of their career (left) and the length of their career (right).

Before estimating the ratings, the data is divided into 50 periods  $t = 1, \dots, 50$ , each of length 3 months. The periods range from October 2004 to March 2017. An average of 5-6 events occurs during each winter period (October-March) and no events occur during the summer periods (April-September). We choose a period length of 3 months to be large enough to have multiple competitions within each period and short enough where the athletes’ ratings should not change significantly from competition to competition. We then observe the details of changes in the career trajectory over time.

Ratings and RDs are estimated via the GH rating system. Along with estimated ratings we also have the athletes’ birth dates so we include age at the start of their career ( $z_i$ ) as a covariate.

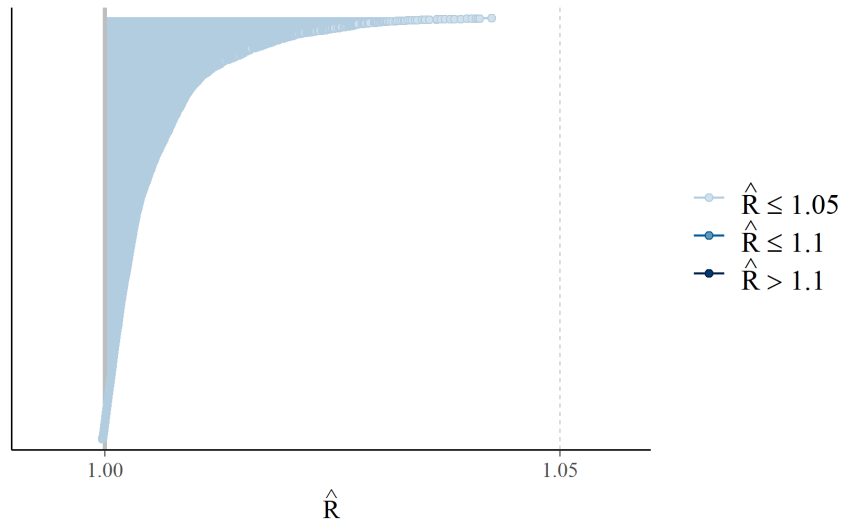


Figure 2.2: The  $\hat{R}$  for posterior draws of all parameters while fitting the model described in Section 2.4

After model selection, the final growth-curve model is,

$$\mu_{it} \sim \text{Normal}(g_i(t, z_i), \sigma_\theta^2 + \nu_{it}^2) \quad (2.27)$$

$$g_i(t, z_i) = \alpha_i + [\beta_{i0} + \beta_{i1}t_1^* + \beta_{i2}t_2^* + \beta_{i3}t_3^*]e^{-(\omega_0 + \omega_1 z_i)(t - t_{0i})}. \quad (2.28)$$

We fit the model as described in Section 2.2 in RStan using NUTS. We run the sampler for 5,000 iterations with a burn-in of 3,000, leaving us with 2,000 posterior draws to use for inference. The model reached adequate convergence according to diagnostics on the  $\hat{R}$  diagnostics described in Chapter 1. Figure 1.1 shows the split- $\hat{R}$  values for the model fit to men’s slalom data. The convergence for the posterior samples is ideal with the split- $\hat{R}$  value under 1.05 for all parameters. After running five chains for 50,000 iterations, we calculated  $\hat{R}$  values of approximately one and observed no other modes were reached during the longer run. We perform model selection using the technique described in Section 1.4 using LOO-CV to determine the number of polynomial terms that optimize prediction performance to be  $p = 3$ .



Figures 2.3 and 2.4 show the model fit for four top-performing men’s slalom athletes: André Myhrer (Sweden), Marcel Hirscher (Austria), Patrick Thaler (Italy) and Alexis Pinturault (France). The points represent the estimated rating calculated via the GH model, and the lines show the predicted rating estimated by the draws from the posterior distributions of the growth-curve parameters. The shaded area shows the 95% credible interval indicating where we expect 95% of the true ratings to lie.

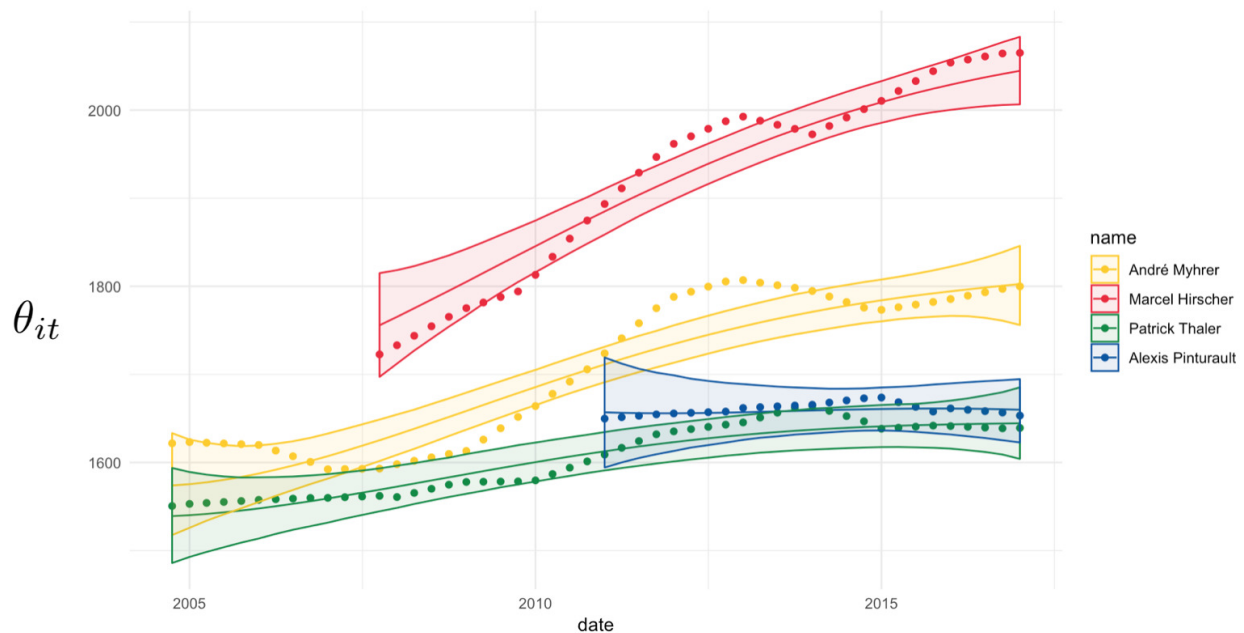


Figure 2.3: Career trajectories for four of the top men’s slalom athletes. The dotted lines are the ratings estimated by the GH ratings, and the lines with shading show the growth curve fit and 95% credible intervals. Here the ratings are aligned with the date.

Figure 2.4 shows the same information as Figure 2.3 aligned at the beginning of their careers ( $t_{0i}$ ) rather than by the date. This alignment can give us an idea of whether or not the athletes have similar trajectories and can be particularly useful if we are comparing an early-career athlete to a more developed athlete. These athletes are all in different stages of their careers. Myhrer and Thaler both have more-developed careers and have plateaued in their performance after a gradual increase. Although beginning at the same level of performance, Myhrer has had a more rapid increase and is now outperforming Thaler. Hirscher and Pinturault are at earlier stages in their

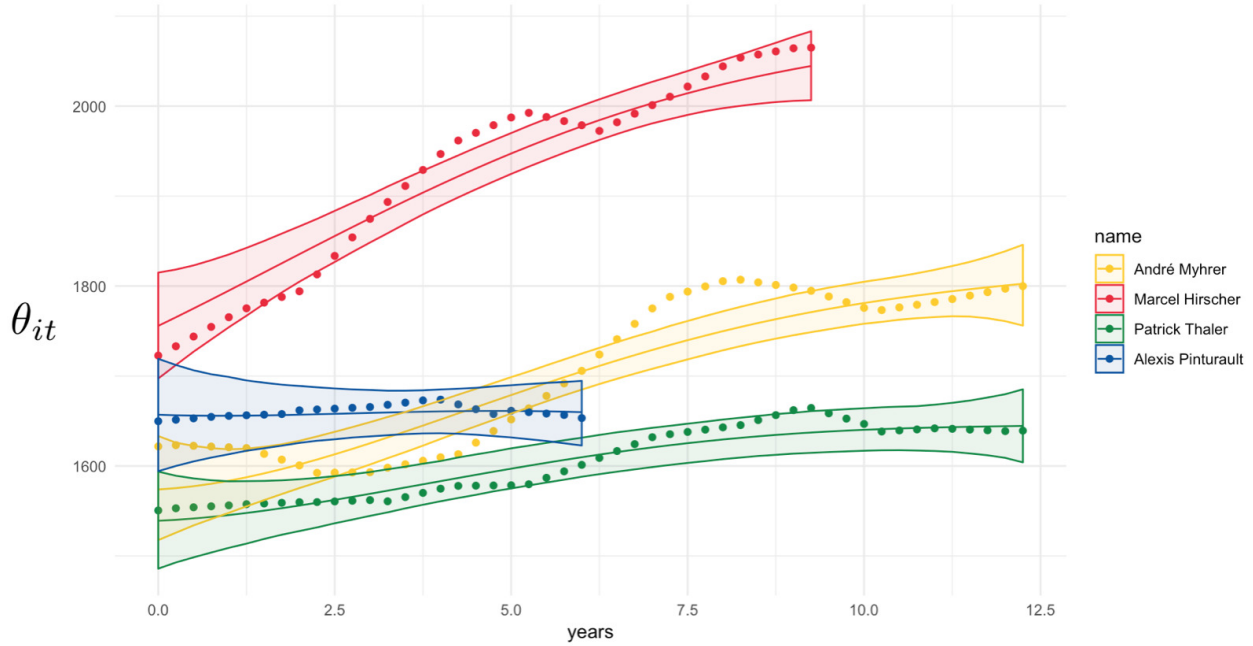


Figure 2.4: The same information is shown in 2.3 but aligned at the beginning of the athlete’s career.

career but have drastically different trajectories. Hirscher’s career is defined by a large increase in performance, whereas Pinturault has shown very little change.

Figure 2.5 shows the projected ratings of Pinturault against the in-sample fit of Thaler. We see that although Pinturault started with a greater rating than Thaler, their careers are likely to end in roughly the same way. Although this is only an example, including four of the athletes, this type of analysis can be done the same way with any combination of athletes.

Another result we can extract from this model fit is the “mean curve”, described in detail in Chapter 1. The purpose of the mean curve is to extract the general shape across the entire sport by fitting the growth curve using estimates for the polynomial means  $\eta_{\beta_b}$  and decay parameters  $\omega_0, \omega_1$ . The trajectories shown in Figure 2.6 are evaluated by,

$$\bar{\theta}_t = \left[ \eta_0 + \eta_1 t_1^* + \eta_2 t_2^* + \eta_3 t_3^* \right] e^{-(\omega_0 + \omega_1 \bar{z})(t - t_{0i})} ,$$

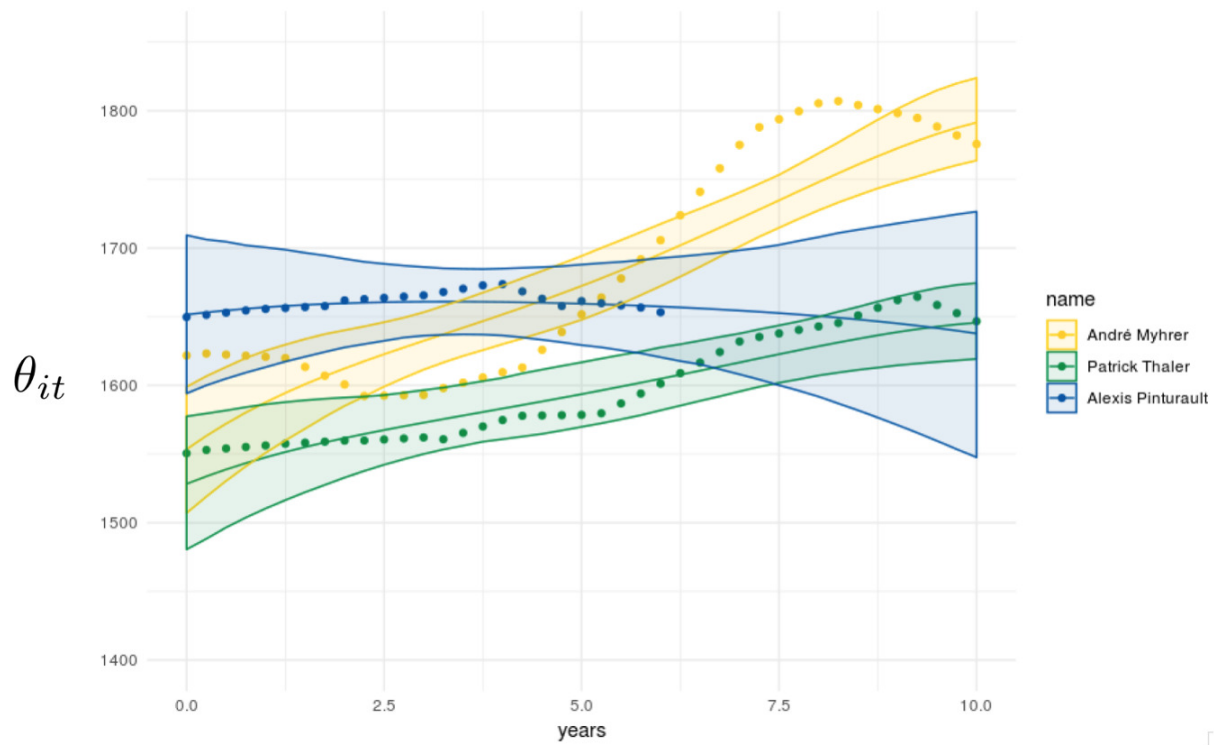


Figure 2.5: Comparing Myherer and Thaler’s fitted, full career trajectories to Pinturault’s projected career trajectory.

where  $\bar{z}$  is the average age at the start of the athletes’ careers. This result enables us to capture the typical career trajectory of any athlete. The mean curve can give us an idea of what to expect for athletes for whom we do not have many observed competition results. Figure 2.6 compares the mean curve of men’s slalom skiing, women’s luge, and women’s judo. Women’s luge and men’s slalom are multi-competitor sports, thus ratings are estimated using GH ratings. Women’s judo is a head-to-head sport thus ratings are estimated using Glicko ratings. By looking at these mean curves side by side, we can compare the typical career path of athletes in each sport. Both men’s slalom and women’s luge have similar shapes, but women’s luge has more variation in ratings. This difference could be attributed to women’s luge players having potentially longer careers because luge is a less physically demanding and a more skillful sport. In women’s judo, a modern martial art, we see a sharper drop in performance after 10 years, indicating it might be a more physically

demanding sport than men’s slalom, thus favoring younger athletes.

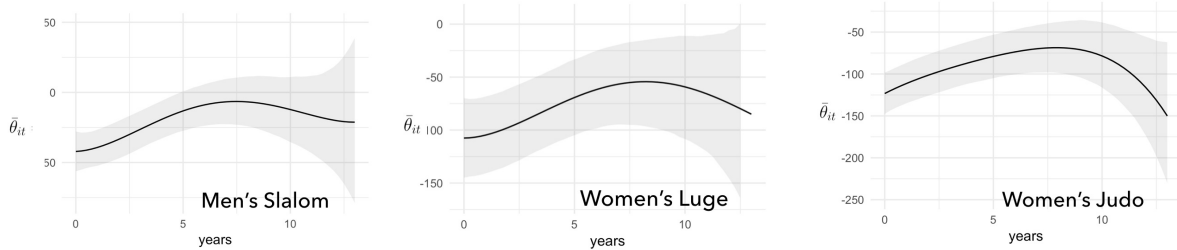


Figure 2.6: The mean curve for three Olympic sports. Women’s luge and men’s slalom are fit using GH ratings. Women’s judo are fit using Glicko ratings.

## 2.4.2 Coverage

We use the men’s slalom data to determine the estimated coverage of our predictions. Based on the posterior samples, we determine the 95% credible interval (CI), the range for which we are 95% confident contains the true athletic strength. An interval is a better descriptor of future ratings than a point estimate because it accounts for the reliability in our prediction. A correct prediction is an interval that contains the truth at any given time.

To validate the model, we estimate the coverage of our projections to a future time  $T$ . For athlete  $i$ , we evaluate the growth curve at time  $T$  using parameters sampled at each iteration of our MCMC sampler, giving us a distribution for  $\theta_{iT}$ . The 95 % CI is the interval that contains 95%, and the 50% CI is the interval that contains 50% of the evaluated samples, centered on the mean. The coverage is the percentage of times we correctly captured an athlete’s true rating within our projected interval. Figure 2.7 shows the estimated coverage at 95% and 50% intervals for four different time periods (1/4 year, 1 year, 2 years, and 4 years). If we have proper coverage, we would expect our coverage estimates to lie roughly along the dotted lines. In the 50% case, we consistently see high coverage; that is, we capture the true rating in our interval too often. In the 95% interval, we see high coverage for a quarter year, correct coverage for 1 and 2 years, and low coverage for 4 years. The shaded area shows the standard errors based on a binomial distribution of the

percentage of correctly predicted athletes' over the total number of athletes. Decreasing coverage at higher years means that our model is not capturing enough uncertainty when projecting that far out. One potential cause of this is the selection bias that occurs as athletes retire. Stronger athletes tend to have longer careers, so we are not accounting for all career types when projecting far into the future. We see evidence of this in Figure 2.1 (left), where the majority of athletes have careers for less than 4 years. The drop in coverage shows that we are not capturing the proper variation for projections up to 4 years.

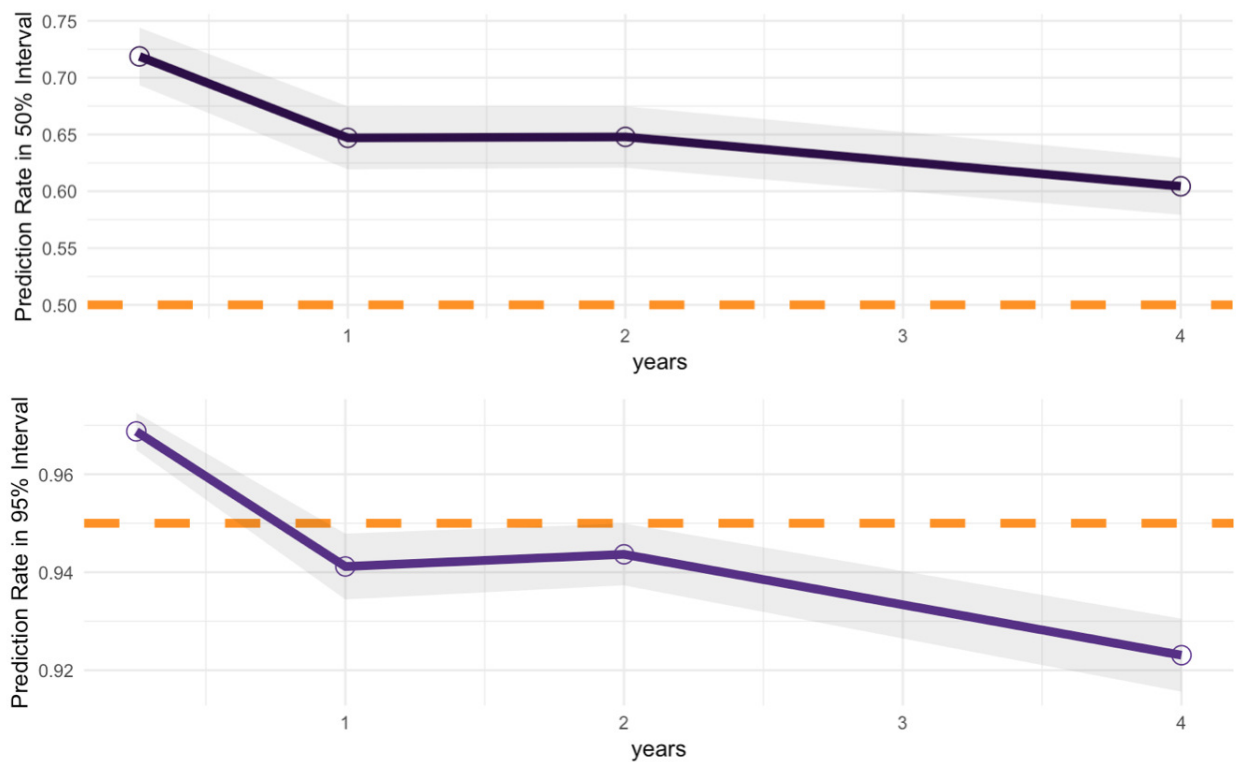


Figure 2.7: Coverage of the predictions projected out a quarter, one, two and four years. The dotted line shows the expected coverage for the 95% (top) and 50% (bottom) intervals. The shaded region shows standard errors based on a binomial distribution.

## 2.5 Clustering Career Trajectories

In Section 2.4, we compare sports by estimating the mean growth curve. Here, we perform an exploratory analysis to understand how athletes compare to one another. A natural way to visualize what creates a successful athlete is by comparing athletes against one another. For example, scouts may predict the success of a young athlete based on the style or physique of a successful athlete with the hope that they perform similarly in the future. An exploratory way of finding similarities is to cluster career trajectories. [Bartolucci and Murphy \(2015\)](#) fits a finite mixture model on runners in a 24-hour long endurance race with the intent of identifying strategies amongst clusters of runners based on their performance. Such an analysis can be useful in finding athletes with a potentially similar style or technique or help understand what makes athletes more successful during a certain career stage. To enable the exploration of the answers to these questions, we cluster the estimated career trajectories of the athletes. The goal of clustering is to create subgroups of the data, whereby the trajectories in each group have similarities. Instead of clustering by the predicted rating, we cluster the estimated growth curve parameters. This technique is more efficient than directly comparing the ratings since we have a fixed number of parameters within the growth curve. In contrast, the ratings can be of varying dimensions between athletes as some have had longer careers than others. Also, clustering based on estimated growth curve parameters can describe the athlete's entire career rather than just the observed years.

Clustering raw measurements is problematic for several reasons, including high dimensionality of the trajectories and trajectories of different observations covering different periods. A common approach is to use methods that reduce dimensionality by using decomposition methods to break the observations down into a simple basis. Conventional dimension reduction techniques include techniques such as principal components analysis (PCA), singular value decomposition (SVD), or dynamic factor analysis (DFA) ([Jolliffe, 2002](#); [Zuur et al., 2003](#)). Dimension reduction or decomposition techniques can produce better clustering results by smoothing away noise, but also suffer

from a loss of information and interpretability. Many of these decomposition techniques, such as PCA and SVD, are agnostic of time and do not perform well with missing data. Clustering raw data also does not make sense in the context of entire career trajectories because we are also interested in making out-of-sample predictions.

A better approach is to cluster functional representations of time-series data instead. Many techniques can be used, including two-stage, nonparametric, and model-based methods (Jacques and Preda, 2014). In particular, we focus on two-stage methods which first reduce the dimensionality of the data by providing a functional basis, and then use a non-parametric clustering approach to cluster the observations using this simplified basis. Abraham et al. (2003) first fits B-splines to the longitudinal data set and uses k-means clustering to the corresponding parameters. In an application to sports, Miller and Bornn (2017) cluster segments of the National Basketball Association athletes movement during possession by mapping them to a set of characteristic actions. These “templates” are represented by Bezier curves, a functional basis that maps time to a two-dimensional position on the court to determine the shape of the movement of the athlete. The athlete’s movement is then clustered using the functional form of the templates using the model parameters, thus constructing a vocabulary of types of possession.

Instead of clustering against point estimates of the parameters we want to utilize the entire distribution of these parameters to retain uncertainty of the estimates. The following definition of the Wasserstein distance is a good metric for measuring the distance between two empirical distributions which in our case are samples from the posterior. For simplicity, we use the distance between the marginal coefficients, although we recognize we ignore correlation. If we define  $B_{ib}$  as the empirical distribution of the posterior of  $\beta_{ib}$ , for athlete  $i$  and  $B_{jb}$  to be the equally sized empirical distribution of  $\beta_{jb}$  for athlete  $j$  then the distance between the two distributions is defined as a function of the ordered posterior draws,

$$W_2(B_{ib}, B_{jb}) = \left( \sum_{k=1}^n [\beta_{ib}^{(k)} - \beta_{jb}^{(k)}]^2 \right)^{1/2}, \quad (2.29)$$

Cluster Number	Athletes
1	13
2	163
3	3
4	205
5	6
6	13

Table 2.1: The distribution of athlete’s across clusters for a cut in the hierarchical clustering to create six clusters.

where  $\beta^{(k)}$  indicates the  $k$ -th ordered draw. The total distance between two athletes is defined as

$$d_{ij} = \sum_{p=1}^P W_2(B_{ip}, B_{jp}), \quad (2.30)$$

where  $P$  is the order selected through model selection. The  $\beta_{ib}$  are centered by  $\eta_p$  and standardized by  $\sigma_{\beta_b}$ . The intercept terms  $\beta_0, \alpha$  are ignored since we are more interested in the shape of the trajectories in this clustering exercise rather than the value of the rating.

We perform agglomerative hierarchical clustering using the distance matrix created via the Wasserstein distance (Ward, 1963). Hierarchical clustering methods are advantageous because they enable us to perform clustering before determining the number of partitions. Not having to fix the number of clusters is beneficial when we have no prior information on how many partitions exist. Hierarchical clustering also provides a network of the similarities between each of the athletes via a dendrogram (Figure 2.9 shows a partial dendrogram for a single cluster) which can be helpful when comparing two specific players to one another or to other groups of athletes (e.g., Ruta et al., 2019). Ward’s method is used to determine which clusters to combine. This method uses both distance and variance within a cluster to determine cluster merging; thus, it is shown to create more compact clustering in noisy circumstances. We use the average silhouette method to determine the number of clusters to be six (Rousseeuw, 1987). Table 2.1 records the number of athletes within each cluster. Details about the clustering methods and finding the optimal number of clusters can be found in Appendix B.



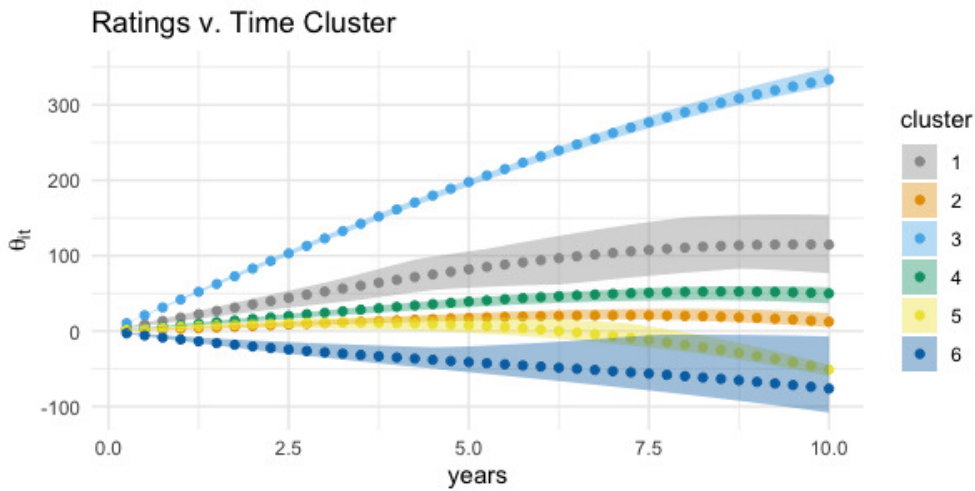


Figure 2.8: Mean and IQR of the estimated career trajectory for athletes within each of the six clusters.

Figure 2.8 shows the mean and interquartile range (IQR) of the estimates of the career trajectories within each cluster at each period taken from the fitted growth curves from all men’s slalom athletes. Immediately, we can see the separation between groups of athletes by simply clustering on three parameters. The clustering method appears to divide the trajectories into athletes with linearly increasing ratings (Cluster 1), athletes with decreasing ratings (Cluster 6), and athletes whose rating is projected to remain the same (Clusters 2,4). Cluster 1 contains trajectories that are first increasing, but then stop increasing and flatten out. Cluster 5 contains trajectories that decrease at a quicker rate as time goes on.

Figure 2.9 shows the section of the dendrogram and corresponding projected career trajectories for the athletes in Cluster 5. The trajectories feature a similar shape, with an accelerating decline in performance toward the end of the athlete’s career. We can use the dendrogram to find athletes who are more alike by following the branches from the top (larger distances) to the bottom (smaller distances). For example, Herbert and Pragner are the last branch in the dendrogram and have roughly the same trajectory shape with different intercept values. Kostelic was the final athlete to be included in the cluster and has a more downward-curved trajectory than the remaining athletes

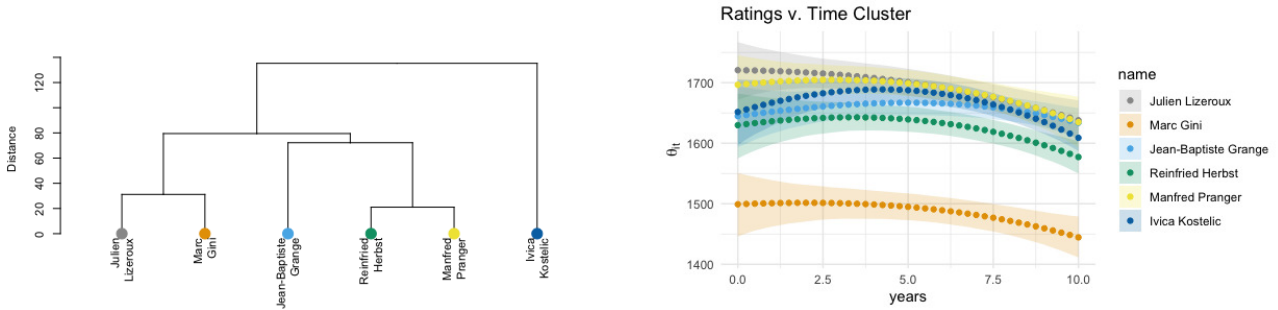


Figure 2.9: The section of the dendrogram representing the path of the clustering and the distance between each of athletes within Cluster 5 (right). The Estimated trajectories and 95 % CI for the members of Cluster 5 (right).

in Cluster 5.

## 2.6 Conclusions

The technique developed in this chapter can be useful for a wide variety of applications in sports analytics, including the comparison of performance across athletes and even across sports. This Chapter only addresses applications of the growth curve post-processing step to Elo, Glicko, and GH ratings, but the method can be used on any type of relevant rating. We used no rating-inherent or sport-inherent assumptions when building the growth-curve model. Multi-phase inference is advantageous because it can be applied as a post-processing step to any rating system. Multi-phase inference gives tremendous flexibility in analyzing the sport and application of interest and alleviating the need to create and fit complicated, time-varying rating systems.

We fit our model of ratings using an MCMC approach to obtain posterior samples. We validate the method by estimating the coverage of predictions made by projecting out the growth curve to times of varying lengths. Although the estimated coverage is the expected value for estimates up to two years, the main downfall of this implementation is that the coverage drops at four years. This drop is due to an underestimate of the variation in ratings at long-term projections. Calibrating

the model to account for this variance would increase the quality of predictions.

We present two applications of this growth curve. The first is an exploratory extension by clustering the trajectories across different athletes. Clustering athlete trajectories based on their growth curves is easier than clustering on raw ratings and can give us information about an athlete's entire career trajectory rather than only from observed periods. We also use the mean curve as a method to quickly compare trajectories across sports. Merely looking at plots of the mean curve can reinforce qualitative comparisons based on what we know about the sport, such as the physicality of the sport, the typical age of athletes, and other similar characteristics.

This page is intentionally left blank.

## Part II

# Defining Regions that Contain Complex Astronomical Structures

This page is intentionally left blank.

## Chapter 3

# Defining Regions that Contain Complex Astronomical Structures

### 3.1 Introduction

The universe contains an incredible variety and number of objects, some documented and some yet to be discovered. Understanding the morphology of these objects is an important step for astronomers to not only better explain the respective sources of emission, but also comprehend the underlying physics of the universe. Frequently these astronomical objects are diffuse and irregularly shaped, and images of these objects can be very noisy due to low photon counts. The morphology of these irregular objects cannot easily be modeled, and defining the boundaries of these objects becomes a difficult task. There are several sources of difficulty when drawing boundaries around distant astronomical objects. One is deconvolving the image from the point spread function (PSF), or measurable telescope error. Deconvolution is difficult in low-count, low-resolution images because the PSF is on the same spatial scale as the features of interest. Therefore, it is challenging to resolve detailed features. Another difficulty is that other, brighter objects in the image can overwhelm

the object of interest and obscure features of the extended source. Furthermore, diffuse sources do not have clearly visible edges where the emission ends, and typical image segmentation techniques that rely on edge detection are insufficient.

The motivation behind the development of the technique in this chapter is a series of X-ray images taken by the *Chandra* telescope. X-ray images are sparse compared to their counterparts with longer wavelengths, making analyzing the images more challenging. In the sources of interest, the structure we want to analyze has a small spatial resolution compared to the PSF. Two types of sources are the focus of this chapter. In both cases, our goal is to define a region of interest (ROI), where the sources are significantly different from the background. The first is a set of extragalactic jets, high-velocity mass ejections from a black hole, at a high redshift ( $z > 2.1$ ). It is challenging to make morphological observations about the jets at large redshifts since few high energy photons reach the detector; therefore, the features are drowned out by the stronger adjacent galaxy. Understanding the morphology of jets is essential in understanding the underlying physics that creates them. The second task is to find the boundary of the ROI in several images of the supernova SN 1987 A taken over time. Being able to track the changes in the boundary of structure in the supernova over time can help astronomers understand the development of the supernova at high energy frequencies. They also can use the segmented image to more objectively compare the morphology to other frequencies such as optical.

One approach to defining object morphology is for an expert astronomer to define boundaries by hand. This task can be onerous if there are many sources and can contain errors caused by human subjectivity. The definition of the boundaries can be influential in future analyses. For example, in [McKeough et al. \(2016\)](#) the detection method used was occasionally sensitive to the boundary shape and size. Another approach involves using scientific models to describe the morphology of these objects. However, data sparsity makes evaluating these models tricky, and often it is these scientific models we would like to understand. We prefer a data-based, automated method to alleviate these downfalls.



Data based approaches for source detection and boundary algorithms are widely used in astronomy. However, no method currently exists that can accommodate hurdles that come with sparse, low-resolution, X-ray images. Wavelet-based approaches, including **wavdetect**, are useful in detecting point sources in a noisy environment including X-ray images (Starck et al., 2002; Freeman et al., 2002). Vikhlinin et al. (1998) uses a matched filter technique to detect sources in X-ray images. BASCS is a Bayesian technique developed by Jones et al. (2015) that separates two overlapping point sources. These approaches have only proven useful for detecting and reconstructing point sources, which does not help in cases with extended emission. Other attempts of modeling more complicated structure have been developed in high signal to noise settings. However, these approaches do not perform well with the sparse, low-signal, X-ray images. Extensive reconstruction and analysis have been done on maps from the Cosmic Microwave Background (CMB) (Bobin et al., 2016). Scientists have used adaptive binning to smooth images of galaxy clusters to help map physical parameters (Sanders and Fabian, 2001; Sanders, 2006). Picquetot et al. (2019) developed a segmentation method for extended sources, but it requires high photon counts as well as spectral homogeneity, an assumption not guaranteed in all observations. A method that works with complex extended sources is adaptive kernel smoothing, which creates a smooth representation of the input data. However, it remains how it can apply to scientific purposes (Ebeling et al., 2006). Spatial field reconstruction using a Gaussian Markov random field to model extended sources has performed well in the survey data of galaxies and could be extended to X-ray images (González-Gaitán et al., 2019). The technique **vtpdetect** is widely implemented, but is limited by computation cost and the need for global thresholding (Ebeling and Wiedenmann, 1993). Methods similar to seeded region growing has been adapted to capture irregular shapes, but only in high signal images (Bertin and Arnouts, 1996). Typical machine learning techniques, such as morphological snakes (Marquez-Neila et al., 2014) or seeded region growing (Adams and Bischof, 1994), tend to have trouble segmenting sparse images. Such techniques rely on large amounts of data and do not innately have the ability to include uncertainty in the estimates. All of the above techniques have weaknesses in addressing boundaries of low count, diffuse sources and therefore a specialized approach is needed.

To address the implications of using X-ray images, we develop a Bayesian method to model the ROI so we can utilize prior information to stabilize or fill in sparse regions of the image and account for the high uncertainty in our observations due to the low number of observed photon counts. We impose the Ising distribution a priori to help segment an image into the ROI and background and to induce cohesiveness between pixel assignments. The Ising distribution and the generalized version, the Potts model, were initially invented to model spin states in ferromagnetic materials (Ising, 1925; Potts, 1952). Since then has been widely used in image segmentation in many fields, including medical images and sonar images of Earth (Bentrem, 2010; Mignotte et al., 2000). We expand upon these applications of the Ising distribution by implementing image segmentation in the Bayesian setting on images with a low signal to noise ratio.

In Section 3.2, we formalize and describe a three-step image segmentation technique built to outline the ROI of complex astronomical sources. Section 2.3 highlights the computational details in fitting the model. Sections 3.4 and 3.5 contain validation of the method through several simulated scenarios and details of the application of this method to the extragalactic jet, supernova.

## **3.2 Model & Inference**

The procedure to obtain the optimal boundary for the ROI takes three steps. The first step uses a Bayesian reconstruction algorithm to infer the multi-scale structure of the extended source of interest to elicit details in the morphology at a higher resolution. The second step uses a novel Gibbs sampler to draw from the distribution of pixel assignments, dividing the pixels into the ROI and the background. The final step uses the distribution of the pixel assignments to determine an optimal boundary around the ROI.

### 3.2.1 Step 1: Image Reconstruction Using LIRA

The first step is to reconstruct the image so that we can resolve detailed morphological structure within an extended source. Due to the nature of X-ray imaging, the number of photon counts we observe is low, causing the information in the pixels in the images to be sparse, making it challenging to study detailed structures within the source. Reconstruction methods allow us to infer the true underlying distribution of the intensity of the source.

To accomplish this we use the tool Low-counts Image Reconstruction and Analysis (LIRA, [Esch et al., 2004](#); [Connors and van Dyk, 2007](#))<sup>1</sup>. LIRA is a fully Bayesian, Markov Chain Monte Carlo (MCMC) algorithm designed to simultaneously model the structure of an observed image at multiple scales. This multi-scale structure captures the residual emission in excess of a given baseline model. LIRA is a useful tool in eliciting the details of extended sources within this multi-scale component. LIRA can also be used in the presence of a PSF and is especially useful when the structure of interest is on the same scale or smaller than that of the PSF. LIRA was previously used in the detection of source components of extragalactic jets ([McKeough et al., 2016](#); [Stein et al., 2015](#)). LIRA was also used to recover the detailed structure of the supernova remnant SN 1987 A across multiple dates ([Kashyap et al., 2017](#)).

Bayesian techniques are useful for dealing with X-ray images since the process of collecting photons is inherently probabilistic. MCMC algorithms allow us to explore the full distribution of the parameters as well as obtain uncertainty measurements on our estimates of interest. Furthermore, a fully Bayesian reconstruction algorithm simplifies parameter inference by removing ambiguity about stopping rules, as well as providing estimates for all parameters simultaneously. Here, we are interested in inferring the true intensity  $\Lambda$  of our source, given an observed image containing pixel-wise photon counts  $Y = \{y_i, i = 1, \dots, n\}$ . Using Bayes rule, we can equate the distribution of

---

<sup>1</sup>LIRA is an open-source **R** package and is found on GitHub: <https://github.com/astrostat/LIRA>.

the underlying intensity given the observed image to other known distributions:

$$p(\Lambda|Y) \propto p(Y|\Lambda)p(\Lambda). \quad (3.1)$$

We denote the photon count across the entire image as  $N_y = \sum_i^n y_i$ . We can model the spatial distribution of these counts in terms of two components: the multi-scale component in which we are interested in performing inference on  $\Lambda = \{\lambda_i, i = 1, \dots, n\}$  and the known baseline component,  $\Lambda_b = \{\lambda_{bi}, i = 1, \dots, n\}$ . We model the observed photon counts as a Poisson distribution with the mean as a linear combination of the two photon count processes:

$$y_i|\Lambda \sim \text{Poisson}\left(\sum_{j=1}^n P_{ij}A_j(\lambda_j + \lambda_{bj})\right). \quad (3.2)$$

The PSF,  $P_{ij}$ , is the probability of a photon originating in  $j$  is detected in pixel  $i$ . This quantity needs to be estimated dependent on observational equipment but is included in this model as a known parameter. The exposure map  $A_j$  is the value of the efficiency of photons detected at pixel  $j$ . The total photon count across the image remains fixed by enforcing  $\sum_{j=1}^n \lambda_j + \lambda_{bj} = N_y$ . When using LIRA, we must consider square images of  $N = 2^d \times 2^d$  pixels. Ideally, we have images of dimension  $64 \times 64$  or  $128 \times 128$  pixels ( $d = 6, 7$ ) because any fewer we may have trouble uncovering interesting structure and any more we run into computational limits since the size of the image increases exponentially. The prior distribution on  $\Lambda$ , as described in [Esch et al. \(2004\)](#), imposes structure on the image by assigning prior distributions to partitions or groups of partitions within the image. The goal of this prior distribution is to achieve flexible smoothing at multiple scales, as well as to ensure stability in the fit.

The output of LIRA is a sequence of draws from the posterior distribution in which each draw is a single image with pixel-wise values of the multi-scale counts,  $\tilde{\Lambda} : \{\tilde{\lambda}_i \in \mathbb{R}, \tilde{\lambda}_i \geq 0\}$ . An individual draw of LIRA tends to be noisy given the low signal environment; therefore, it is not a good representation of the entire distribution. The best way to view the results of LIRA is in aggregate

rather than a single draw from the posterior. Figure 3.1 (b) shows an example of the average multi-scale counts as a result of applying LIRA to an extragalactic jet. Using LIRA, we elicit details of our source of interest at smaller resolutions than we would without it and in the absence of features that would hinder our inference on the source’s boundary. In the next step, we use the individual LIRA draws to assign regions of the image to the ROI and the background.

### 3.2.2 Step 2: Distribution of Pixel Assignments

This step uses the draws from LIRA to segment the image into the ROI and background. For simplicity, we assume that the image contains a single source. Therefore, the image can be entirely represented as a set of two-pixel assignments  $Z : z_i \in \{-1, +1\}$  where pixels are divided into the ROI ( $z_i = +1$ ) and the background ( $z_i = -1$ ). We label pixel assignments independently on each draw from LIRA to propagate uncertainty between the steps. We are interested in making inference about the spin state of the image  $Z$  by drawing from the posterior distribution,

$$p(Z|\Lambda, \tau_0, \tau_1, \sigma_0^2, \sigma_1^2, \beta) \propto f(\Lambda|Z, \tau_0, \tau_1, \sigma_0^2, \sigma_1^2)p(Z|\beta)\pi(\tau_0, \tau_1, \sigma_0^2, \sigma_1^2, \beta). \quad (3.3)$$

We define our likelihood to be distributed as a square-root-normal,

$$\sqrt{\tilde{\lambda}_i|z_i, \tau_0, \tau_1, \sigma_0^2, \sigma_1^2} \sim \text{Normal}(\tau_0, \sigma_0^2)\mathbb{I}_{z_i=-1} + \text{Normal}(\tau_1, \sigma_1^2)\mathbb{I}_{z_i=+1}, \quad (3.4)$$

where  $\tau_0$  is the mean intensity of the background and  $\tau_1$  is the mean intensity of the ROI. We estimate variances  $(\sigma_0^2, \sigma_1^2)$  separately since we expect background pixels to have less variability and be clustered close to zero while pixels in the ROI to have more variability in intensity. We set a distribution to the square root of  $\tilde{\lambda}_i$  rather than the log due to a commonly occurring underflow error in LIRA which produces zero values. The anomaly is described in Section 3.3.

We incorporate the Ising distribution as a prior on the pixel assignments to impose cohesion

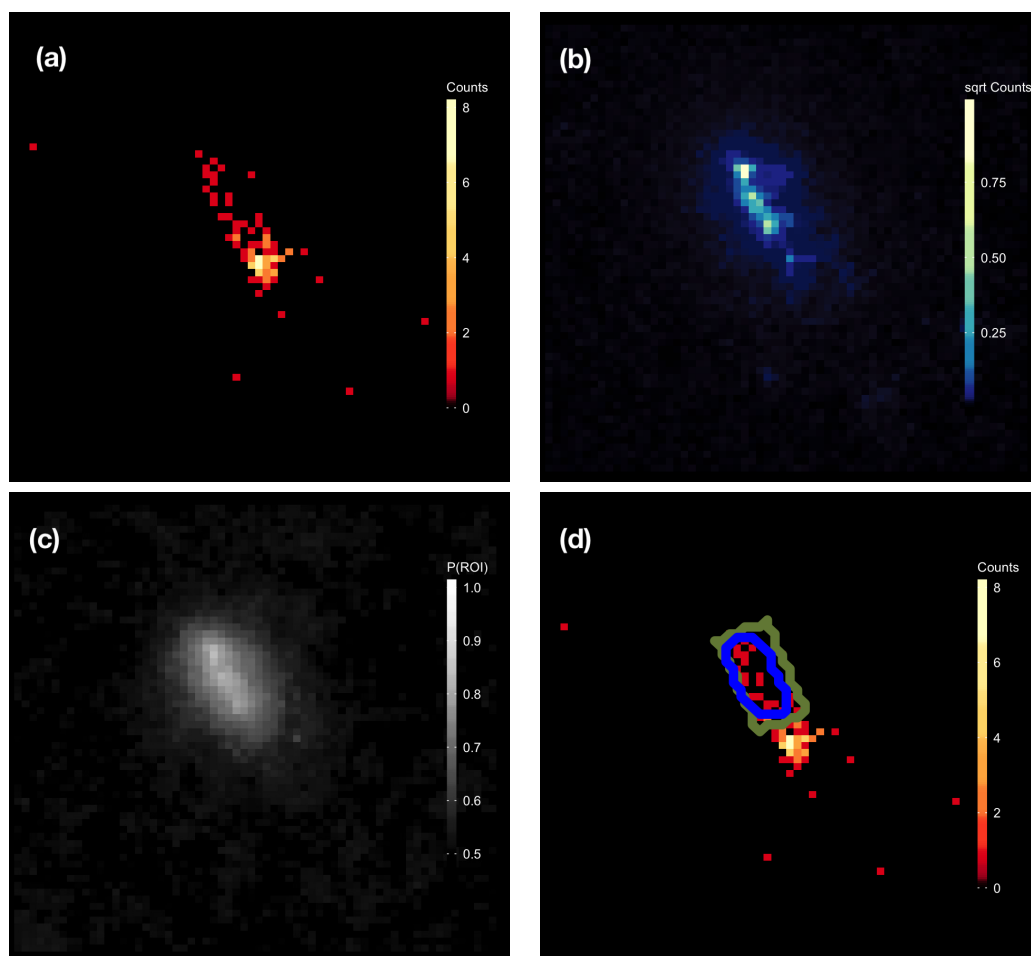


Figure 3.1: The Chandra observatory image of Obs ID 7873 (a). The average multi-scale counts of the jet from LIRA, absent of the galaxy (b). The 2D probability map aggregating across the distribution of pixel assignments based on the multi-scale counts (c). The optimal boundary overlaid on the original X-ray image (d).

between similarly labeled pixels. The Ising PDF is,

$$p(Z|\beta) = \frac{\exp(\beta \sum_{d(i,j)=1} z_j z_i)}{\tilde{Z}(\beta)}, \quad (3.5)$$

where  $\tilde{Z}(\beta)$ , the partition function, is the sum of all possible “energy” states. The distance metric  $d(i, j)$  is the distance between pixels  $i$  and  $j$  such that  $d(i, j) = 1$  means they are adjacent pixels. The parameter  $\beta > 0$  is proportional to the *inverse* temperature meaning high values of  $\beta$  result in “freezing” conditions of our pixel assignments, and more cohesion between the pixel assignments. Low values of  $\beta$  correspond to a more noisy pixel assignment array. The hyper-prior on the temperature parameter the Ising distribution  $\beta$  is,

$$\beta \sim \text{Gamma}(a_\beta, b_\beta).$$

Since  $\tilde{\lambda}_i|z_i, \tau_0, \tau_1, \sigma_0^2, \sigma_1^2$  follows a square-root-normal, we will set the mean and variance parameters to the following conjugate priors,

$$\tau_0|\sigma_0^2 \sim \text{Normal}(\mu_0, \sigma_0^2) \quad (3.6)$$

$$\tau_1|\sigma_1^2 \sim \text{Normal}(\mu_1, \sigma_1^2) \quad (3.7)$$

$$\sigma_0^2 \sim \text{Inv} - \chi^2(\nu_0, \omega_0^2) \quad (3.8)$$

$$\sigma_1^2 \sim \text{Inv} - \chi^2(\nu_1, \omega_1^2). \quad (3.9)$$

Sampling from the posterior defined in Equation 3.3 is not trivial. We build a novel Gibbs sampler that can be used to obtain samples from the posterior (Geman and Geman, 1984). The conditional distributions are iteratively drawn using the following steps:

1. Draw the inverse temperature parameter from  $p(\beta|Z)$ :

Using Bayes rule, we know that:

$$p(\beta|Z) \propto p(Z|\beta)\pi(\beta),$$

where we have already assumed  $Z|\beta$  follows an Ising distribution. The partition function, the density of states for a periodic 2D Ising lattice, is exactly calculated using Mathematica code by [Beale \(1996\)](#). Although our images are not periodic (edges do not wrap) in large scales the 2D lattice is periodic asymptotically, so it is an acceptable approximation. In fact, no new assumptions are made because LIRA already requires the periodicity of the image.

These draws are implemented via Metropolis Hastings (e.g., [Gelman et al., 2013](#)) with  $p(\beta|z)$  as a target distribution and  $J(\beta^*|\beta)$  as a proposal distribution such that  $E[\beta^*|\beta] = \beta$  and  $\text{Var}[\beta^*|\beta] = \rho$ :

$$\beta^*|\beta \sim \text{Gamma}\left(\frac{\beta^2}{\rho^2}, \frac{\beta}{\rho^2}\right).$$

We recommend placing a weakly informative prior on  $\beta$  for a small value so that the chain does not get trapped within its critical temperature. Once at its critical temperature, the image ‘freezes,’ and the  $\beta$  continues to increase to infinity, giving nonsensical results. We recommend taking several iterations of the Metropolis-Hastings algorithm before updating the value of  $\beta$ .

2. Draw the mean and variance parameters from  $p(\tau_0, \tau_1, \sigma_0^2, \sigma_1^2|\tilde{\Lambda}, Z)$ :

Using Bayes rule we define the distribution of the likelihood parameters,

$$p(\tau_0, \tau_1, \sigma_0^2, \sigma_1^2|\tilde{\Lambda}, Z) \propto \prod_{i=1}^n f(\tilde{\lambda}_i|z_i, \tau_0, \tau_1, \sigma_0^2, \sigma_1^2)\pi(\tau_0, \tau_1, \sigma_0^2, \sigma_1^2).$$

Conditional on the pixel assignment  $Z$ , the distributions for  $\tau_0, \sigma_0^2$  and  $\tau_1, \sigma_1^2$  are independent and therefore can be drawn directly by first drawing the  $\sigma_1^2, \sigma_0^2$ , then using the values to draw  $\tau_0, \tau_1$ . For simplicity let’s say  $\mu_0 = \mu_1 = \mu$ ,  $\nu_0 = \nu_1 = \nu$  and  $\omega_0 = \omega_1 = \omega$ . The process of drawing



the parameters is,

$$\begin{aligned}
 \tau_1 | \sigma_1^2 &\sim \text{Normal}(\mu_{n_1}, \sigma_1^2 / (n_1 + 1)) \\
 \sigma_1^2 &\sim \text{Inverse} - \chi^2(\nu_{n_1}, \omega_{n_1}^2) \\
 \tau_0 | \sigma_0^2 &\sim \text{Normal}(\mu_{n_0}, \sigma_0^2 / (n_0 + 1)) \\
 \sigma_0^2 &\sim \text{Inverse} - \chi^2(\nu_{n_0}, \omega_{n_0}^2) .
 \end{aligned}$$

The definition for the parameters is,

$$\begin{aligned}
 \mu_{n_1} &= \frac{1}{1 + n_1} \mu + \frac{n_1}{1 + n_1} \bar{\lambda}_{n_1} \\
 \nu_{n_1} &= \nu + n_1 \\
 \nu_{n_1} \omega_{n_1} &= \nu \omega^2 + (n_1 - 1) s_{n_1}^2 + \frac{n_1}{1 + n_1} (\bar{\lambda}_{n_1} - \mu)^2 \\
 \bar{\lambda}_{n_1} &= \frac{1}{n_1} \sum_{i \in \{z_i=1\}} \lambda_i \\
 s_{n_1} &= \frac{1}{n_1 - 1} \sum_{i \in \{z_i=1\}} (\lambda_i - \bar{\lambda}_{n_1})^2 \\
 n_1 &= \sum_{i=1}^n \mathbb{I}_{z_i=1} ,
 \end{aligned}$$

for when  $z_i = +1$  and similarly defined when  $z_i = -1$ . The number of pixels assigned to the ROI  $n_1 = \sum_i \mathbb{I}_{z_i=+1}$  and  $n_1 + n_0 = N$ . As a precaution to avoid label switching, after each iteration if  $\tau_1$  is less than  $\tau_0$ , then the  $\tau_0, \tau_1$  and  $\sigma_0^2, \sigma_1^2$  values are swapped for the final update.

3. Draw from the posterior distribution on pixel assignments  $p(Z | \tilde{\Lambda}, \tau_0, \tau_1, \sigma_0^2, \sigma_1^2, \beta)$ :

We draw the pixel assignments using a modified version of the Swendsen-Wang method, here on referred to as the SW method (MacKay, 2003; Swendsen and Wang, 1987). The SW method takes a pixel assignment array  $Z$  and induces a larger parameter space that contains the original  $N$  pixel assignments and  $M$  additional bond variables. The set of additional bond

variables are denoted by  $D : d_i \in \{0, 1\}$  of which  $d_i = 1$  means ‘connected’ and  $d_i = 0$  means ‘disconnected’. We can define a joint distribution  $g$  that couples pixel assignments to bonds,

$$p(Z, D | \tilde{\Lambda}, \tau_0, \tau_1, \sigma_0^2, \sigma_1^2, \beta) \propto \prod_{m=1}^M g_m(z^{(m)}, d_m | \beta) \prod_i f(\tilde{\lambda}_i | z_i, \tau_0, \tau_1, \sigma_0^2, \sigma_1^2).$$

The notation  $z^{(m)}$  refers to the set of all pixel values  $z_i$  that are connected or disconnected by the bond  $d_m$ . The distribution  $g$  must meet the following conditions:

- The marginal distribution of  $Z$ , is equivalent to our original likelihood,

$$\sum_d p(Z, d | \tilde{\Lambda}, \tau_0, \tau_1, \sigma_0^2, \sigma_1^2, \beta) = p(Z | \tilde{\Lambda}, \tau_0, \tau_1, \sigma_0^2, \sigma_1^2, \beta).$$

- The conditional distributions of  $Z$  and  $D$ ,  $p(Z | D, \beta, -)$  and  $p(D | Z, \beta, -)$  are easy to sample from.

The proposed model for  $g_m(z^{(m)}, d_m)$  is,

$$g_m(z^{(m)}, d_m) = \begin{cases} & d_m = 0 & & d_m = 1 & \\ & z_i = -1 & z_i = +1 & z_i = -1 & z_i = +1 \\ z_j = -1 & e^{-\beta} & e^{-\beta} & e^{\beta} - e^{-\beta} & 0 \\ z_j = +1 & e^{-\beta} & e^{-\beta} & 0 & e^{\beta} - e^{-\beta} \end{cases}.$$

The distribution of  $Z, D$  is unchanged if we re-scale it by any constant factor. By defining  $p = 1 - e^{-2\beta}$  we can re-write the model as,

$$\tilde{g}_m(z^{(m)}, d_m) = \begin{cases} & d_m = 0 & & d_m = 1 & \\ & z_i = -1 & z_i = +1 & z_i = -1 & z_i = +1 \\ z_j = -1 & 1 - p & 1 - p & p & 0 \\ z_j = +1 & 1 - p & 1 - p & 0 & p \end{cases}.$$

Explicitly, the joint model for  $Z$  and  $D$  is defined as,

$$\begin{aligned} p(Z, D|\beta) &\propto \prod_{m=1}^M g_m(z^{(m)}, d_m|\beta) \prod_{i=1}^N f(\tilde{\lambda}_i|z_i, b, \tau_0, \tau_1, \sigma_0^2, \sigma_1^2) \\ &\propto \prod_{m=1}^M \tilde{g}_m(z^{(m)}, d_m|\beta) \prod_{i=1}^N f(\tilde{\lambda}_i|z_i, b, \tau_0, \tau_1, \sigma_0^2, \sigma_1^2). \end{aligned}$$

We iterate over the following steps to sample from  $D|Z, \beta$  and sequentially  $Z|D, \beta$ :

- (a) Sample bond assignments from  $p(D|Z, \beta)$ . If two pixel assignments surrounding a bond are identical ( $z_{m_1} = z_{m_2}$ ) set the bond  $d_m$  equal to one with probability  $p$  and set it to zero otherwise.
- (b) Sample pixel assignments from  $p(Z|D, \beta)$ . The bonds connect the pixel assignments into a number of clusters. A cluster  $C$  is defined by a collection of pixel assignments that are connected by tangential bonds equal to one. All pixel assignments in a cluster must adopt the same state as each other. For each cluster, the pixel assignments are +1 with a probability of  $p_+$  and -1 with a probability of  $p_- = 1 - p_+$  where,

$$\frac{p_+}{p_-} = \frac{\prod_{i \in C} f(\tilde{\lambda}_i|z_i = +1, \tau_1, \sigma_1^2)}{\prod_{i \in C} f(\tilde{\lambda}_i|z_i = -1, \tau_0, \sigma_0^2)}.$$

In the event of a pixel being in a cluster on its own we assign it to +1 with a probability of,

$$p_+ = \frac{f(\tilde{\lambda}_i|z_i = +1, \tau_1, \sigma_1^2)}{f(\tilde{\lambda}_i|z_i = +1, \tau_1, \sigma_1^2) + f(\tilde{\lambda}_i|z_i = -1, \tau_0, \sigma_0^2)}.$$

We recommend a burn-in period before accepting a draw for the pixel assignment matrix  $Z$ .

Using the Gibbs sampler, we were able to draw from the distribution of pixel assignments  $Z$ . Similar to the LIRA iterations, the information in aggregate across all draws of the pixel assignments is more valuable since each draw individually contains noise. In the next step, we use the posterior to find the most likely candidate for the boundary of the ROI.

### 3.2.3 Step 3: Optimal Boundary

To obtain our final boundary estimate, we optimize over the distribution of pixel assignments given the observation  $Y$  using a maximum a posterior (MAP) estimation,

$$P(Z|Y) = \int P(Z, \vec{\theta}, \Lambda|Y) d\vec{\theta} d\Lambda \quad (3.10)$$

$$= \int P(Z|\vec{\theta}, \Lambda) P(\vec{\theta}|\Lambda, Y) P(\Lambda|Y) d\vec{\theta} d\Lambda \quad (3.11)$$

$$= \int P(Z|\vec{\theta}, \Lambda) P(\vec{\theta}|\Lambda) P(\Lambda|Y) d\vec{\theta} d\Lambda, \quad (3.12)$$

where  $\vec{\theta}$  represents the nuisance parameters  $(\beta, \sigma_0^2, \sigma_1^2, \tau_0, \tau_1)$ . We assume  $\vec{\theta}$  is independent of  $Y$  when given  $\Lambda$ .

Ideally, we would approximate this by,

$$\hat{P}(Z|Y) = \frac{1}{\Omega} \sum_{k=1}^{\Omega} P(Z|\vec{\theta}^{(k)}, \tilde{\Lambda}^{(k)}), \quad (3.13)$$

where  $\Omega$  is the total number of iterations from steps 1 and 2. Evaluating this estimate is difficult since we cannot perform operations on the log probability since we are summing across the raw probability terms. The evaluated probabilities are too small, given our computational limits due to the overwhelming possible of pixel assignment arrangements. However, we are not necessarily concerned with the value of  $\hat{P}(Z|Y)$ , but rather we would like to find which  $Z$  gives us the maximum. That is, we just need to show that  $\hat{P}(Z_1|Y) > \hat{P}(Z_2|Y)$ , or equivalently,

$$\frac{\hat{P}(Z_1|Y)}{\hat{P}(Z_2|Y)} > 1, \quad (3.14)$$

to claim that  $Z_1$  brings us closer to the global maximum.

We can proceed with this technique by first writing the ratio in terms of values we can solve

without computational issues,

$$\begin{aligned}
 \frac{\hat{P}(Z_1|Y)}{\hat{P}(Z_2|Y)} &= \frac{\sum_{k=1}^{\Omega} \exp(\log P_k(Z_1|\vec{\theta}^{(k)}, \tilde{\lambda}^{(k)}))}{\sum_{k=1}^{\Omega} \exp(\log P_k(Z_2|\vec{\theta}^{(k)}, \tilde{\lambda}^{(k)}))} \\
 &= \frac{\sum_{k=1}^{\Omega} \exp(\log P_k(Z_1|\vec{\theta}^{(k)}, \tilde{\lambda}^{(k)}) - \log P_k(Z_2|\vec{\theta}^{(k)}, \tilde{\lambda}^{(k)})) \exp(\log P_k(Z_2|\vec{\theta}^{(k)}, \tilde{\lambda}^{(k)}) - l_{\max})}{\sum_{k=1}^{\Omega} \exp(\log P_k(Z_2|\vec{\theta}^{(k)}, \tilde{\lambda}^{(k)}) - l_{\max})} \\
 &= \sum_{k=1}^{\Omega} w_k \exp\left(\log \frac{P_k(Z_1|\vec{\theta}^{(k)}, \tilde{\lambda}^{(k)})}{P_k(Z_2|\vec{\theta}^{(k)}, \tilde{\lambda}^{(k)})}\right). \tag{3.15}
 \end{aligned}$$

where  $w_k = \exp(\log P_k(Z_2|\vec{\theta}^{(k)}, \tilde{\lambda}^{(k)}) - l_{\max}) / [\sum_{k=1}^{\Omega} \exp(\log P_k(Z_2|\vec{\theta}^{(k)}, \tilde{\lambda}^{(k)}) - l_{\max})]$  is defined as an additive weight and  $P_k(Z) = P(Z|\tilde{\lambda}^{(k)}, \vec{\theta}^{(k)})$  and  $l_{\max} = \max[\log P_k(Z_2)]$  is the maximum log-likelihood of denominator term. Given a set of pixel assignments  $Z$  we can find the global maximum by comparing the probability of each new  $P(Z_k|Y)$  in a ratio with the current maximum. Using this method we can find the global maximum using any set of pixel assignments even if we do not expect the corresponding probabilities to be monotonically increasing.

To find the global maxima we need to calculate the ratio for all  $Z$ , but since our images can be up to  $128 \times 128$  pixels, the number of possible pixel assignments are too numerous ( $\gg 1e300$ ). To narrow down our search, we find the pixel assignments that are most likely the optimal boundary. To build this set of best possible candidates, we first construct a statistic, we will call the neighbourhood statistic, that would be representative of pixel  $i$  belonging to a *set* of pixels within the ROI,

$$\Phi_i = \frac{\sum_{j \in d(i,j)=1} \zeta_i \zeta_j}{\sum_{j \in d(i,j)=1} 1}, \tag{3.16}$$

where  $\zeta_i = \{0, 1\}$  is a one to one mapping from  $z_i = \{-1, +1\}$ . The neighbourhood statistic can be thought of as the fraction of neighboring pixels that assigned to the ROI given the pixel itself is assigned to the ROI ( $z_i = +1$ ). If the pixel value is  $z_i = -1$  then  $\Phi_i = 0$ .

To create the set of candidates, we first take the average  $\Phi_i$  across all draws from the posterior denoted as  $\bar{\Phi}_i = \sum_i^{\Omega} \Phi_i$ . To determine the pixels within the ROI we rank the value at each pixel from largest to smallest  $\bar{\Phi}^{(1)}, \bar{\Phi}^{(2)}, \dots, \bar{\Phi}^{(n)}$ . We set the pixel with the highest neighbourhood statistic

$\bar{\Phi}^{(1)}$  to  $z_i = +1$  and the remainder to  $z_{j+i} = -1$ . For the next image, we set the pixel assignments for the highest and second highest neighborhood statistics to +1 and the remainder to -1. We repeat the process, including the pixel with the next highest average neighbourhood statistic until all pixels are +1. We also choose to include the posterior draws from the novel Gibbs sampler in step 2 since the draws likely come from exploring a more dense space of our posterior, and thus are suitable candidates for being the best-fit ROI. We then can optimize the posterior across this much smaller and carefully designed set of pixel assignment arrays that are within our computational limits.

### 3.3 MCMC Implementation

We suggest the following implementation for the Gibbs sampler described in Section 3.2. For step 1, we keep 1000 iterations of LIRA for inference from each simulation after a burn-in of 2000 iterations to ensure convergence. We apply step 2 to every posterior sample in step 1. To do so, we first run the Gibbs sampler to convergence on a single LIRA draw, for 500 iterations. We then take the final draw from this initial run and use it as a starting value for the Gibbs sampler independently run on all the posterior draws from LIRA. After a burn-in of 50 iterations, we sample a pixel assignment  $Z$  for each LIRA sample, ending up with 1000 pixel assignment draws from the posterior. We do not use more than a single draw per LIRA sample because there is not much variation between pixel assignments after convergence, and doing so lowers the computational cost. When drawing  $\beta$  within the Gibbs sampler, we suggest a jump standard deviation of  $\rho = 0.01$ , and we take a draw after a burn-in of 20 iterations. For drawing the intensity  $\tau_0, \tau_1$ , and intensity variances  $\sigma_0^2, \sigma_1^2$ , we can sample directly using the conjugate prior. For sampling the spin states, we iterate the SW algorithm and take a single sample after a burn-in of 50 iterations. We use this same sampling strategy for all simulations and applications in Sections 3.4 and 3.5. The process proved robust to the current length of burn in times. To save computational resources we found these burn in times

to be sufficiently long enough to have the same results of longer burn in periods. A more formal analysis of these burn in times could be seen as future work.

We expect the distribution of  $\tilde{\lambda}_i$  to be skewed and positive, but due to a non-correctable underflow issue in the LIRA algorithm when applied to real data, many of the  $\tilde{\lambda}_i$  are exactly 0. The distribution we choose for the likelihood is knowingly mis-specified. A better-specified model would be a piece-wise distribution that includes the point mass on zero. Such a distribution would be more complicated to sample from, thus we did not try this but view this as a possible avenue of future research. We conclude that the square-root-normal distribution matches the shape well enough and use it to simplify our model. We are cautious in believing intensity and variance estimates from the likelihood model in the presence of many zero-valued pixels in the LIRA output. Furthermore, since some values are zero, we choose the square-root-normal instead of the log-normal.

### **3.4 Validation**

Here we present two types of simulations to validate the multi-phase model. The first is a set of two dimensional Gaussians with varying widths on backgrounds of different noise levels. These simulations are to emulate a real but well-defined and simple source to observe how the multi-phase model performs. The second set of simulations is geometric shapes with hard edges in varying intensities, noise levels, and arrangements. The purpose of these simulations is testing the classification error of the multi-phase method in the presence of a “true” boundary. Although this is unrealistic in the case of extended sources, it is still essential to understand how the model performs in these ideal scenarios.

### 3.4.1 2D Gaussian Simulation

We create realizations of a 2D Gaussian with a varying amount of background noise and of varying size within a  $64 \times 64$  pixel image. We simulate the sources as Gaussians with three different variances (4, 8, 16 pixels). For each set of images with the same size Gaussian, we vary the background intensity (0.01, 0.1, 1 photons). The set of simulations use a peak intensity of 5 photons regardless of the background intensity and size. Poisson realizations are made from each of the 2D Gaussian and noise templates. The same realization of noise is used across the same noise intensity for consistency. The realizations are convolved with a 2-pixel standard deviation, 2D Gaussian PSF, to simulate telescope blurring in typical observations. The final images for the simulations are in Figure 3.2. Row (2) most closely resembles real astronomy scenarios.

We then proceed with the multi-phase method to obtain a boundary estimate around the 2D Gaussian in each image. First, we run LIRA independently on each of the nine simulations. In all cases, the same PSF used to convolve the images is used as input in LIRA for deconvolution. We use a flat baseline array of zero photons in each pixel as there are no extraneous sources we wish to remove from the images. Figure 3.3 shows the average multi-scale counts output from the LIRA iterations for each of the nine simulations. The aggregate of the average multi-scale counts emphasizes the features of the source in the absence of the noise in individual draws. However, in some cases, we are also picking up structure in the background.

The next step is to obtain posterior draws from the Gibbs sampler described in step 2 in Section 3.2.2. One result is a probability map that is created by averaging the  $\zeta_i$  for each pixel across the iterations to obtain the probability of pixel  $i$  is contained within the ROI. An astronomer can use these probabilities as a distribution describing the ROI across the entire image. Quantities such as expected values of luminosity or flux can easily be calculated from these probability maps. The probability map resulting from this step for the size varying simulations is shown in Figure 3.4.

Finally, we optimize over the posterior distribution of pixel assignments to get our final boundary



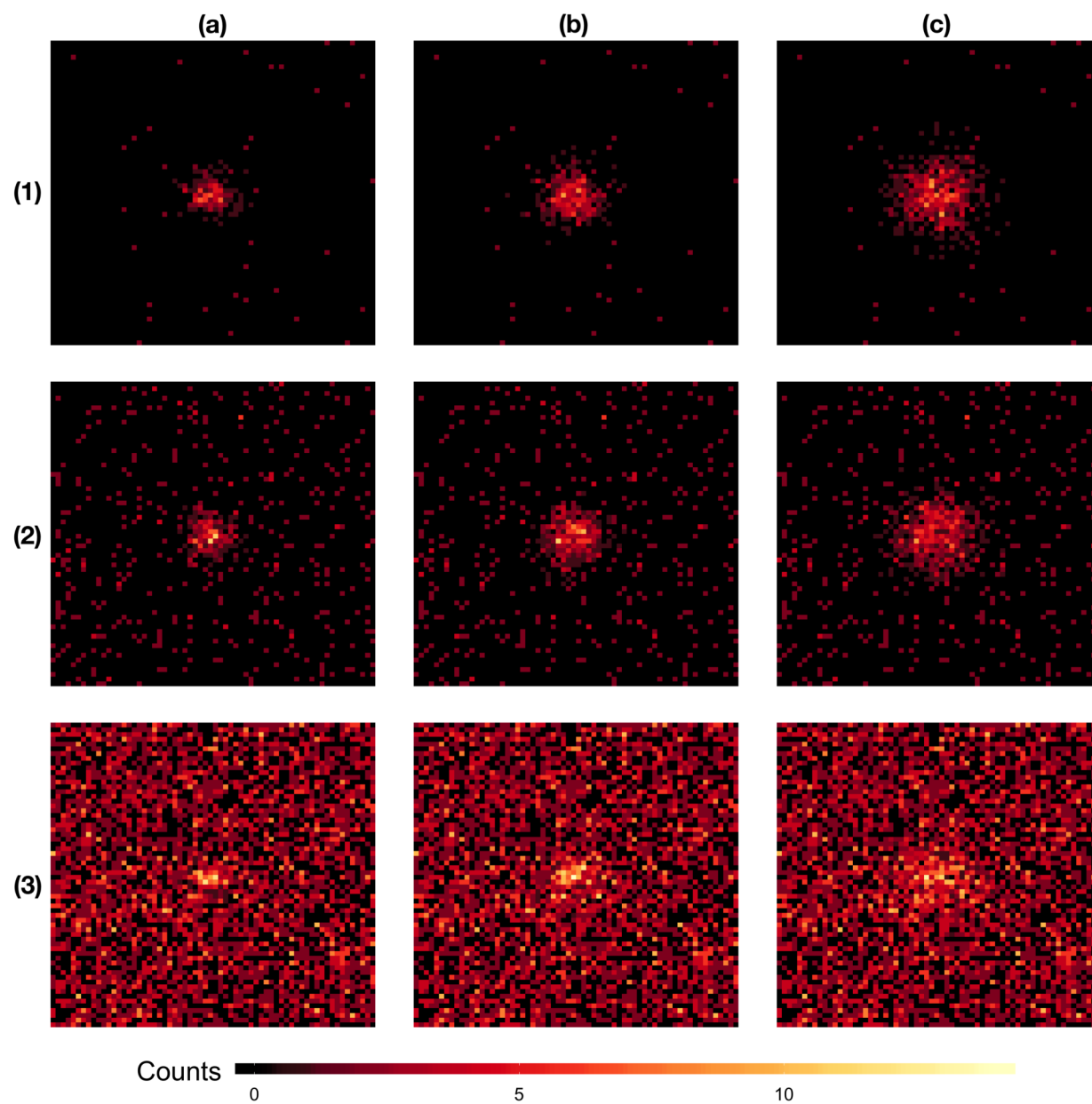


Figure 3.2: 2D Gaussian simulated images. Rows 1,2, and 3 contain noise intensity 0.01,0.1,1 photons respectively. Columns a,b, and c contain a 2D Gaussian with a peak intensity of 5 photons, with a variance of 4,8,16 pixels respectively.

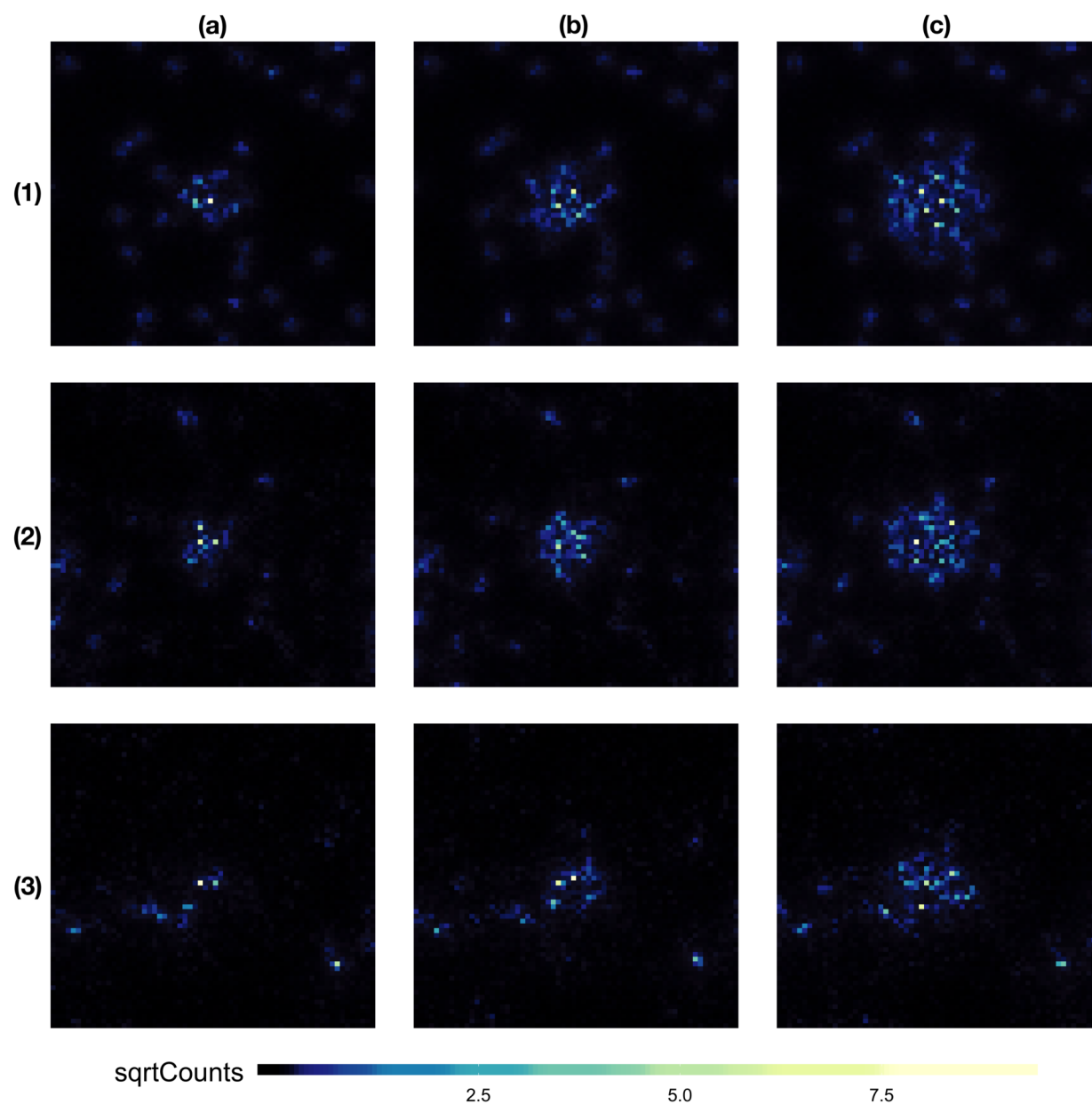


Figure 3.3: The average multi-scale counts calculated using the square root of the posterior draws from LIRA. Row and column arrangements are identical to that of Figure 3.2.

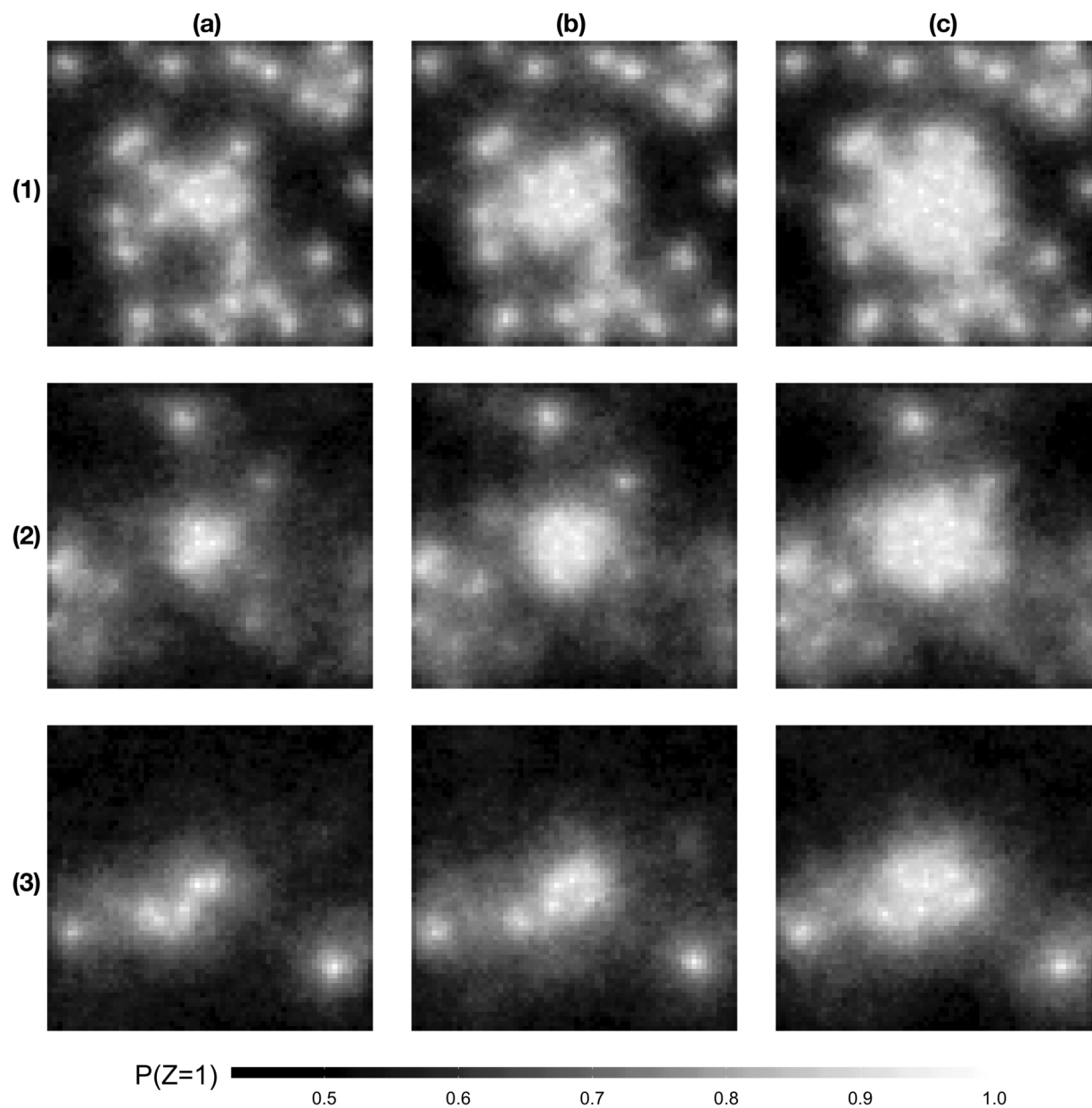


Figure 3.4: Probability map as an aggregation of draws from the posterior distribution of pixel assignments. Each pixel value is the probability of that pixel being within the ROI. Row and column arrangements are identical to that of Figure 3.2.

shown in Figure 3.4 (blue) with two and three standard deviations from the center of the Gaussian for comparison (green).

In rows (1) and (2), the boundary corresponds to being slightly bigger than the three  $\sigma$  contour in the 2D Gaussian. In the case with the most noise, row (3), the boundary corresponds more with the two  $\sigma$  contour. We note that the value of  $\beta$  tends to increase with the width of the 2D Gaussian. The posterior mode and 95% credible interval (CI) for the  $\beta$  is shown in Figure 3.6. The higher the  $\beta$ , the more cohesive we expect the pixel assignments to be. We expect cohesiveness to be favored when the source is larger, but this has a hidden consequence of also increasing the boundary around noisy regions. In Figure 3.5 as we move from columns (a) to (c) we see the ROI grow with the source. We also see the boundary around the noise grow from the first to the second column, before decreasing again once the signal is stronger. This effect correlates with the increasing estimate of  $\beta$ .

### Simulations with Edges

We perform a validation procedure by estimating the boundary for two types of simulated images with edges. The simulated images come from a “truth” that contains edges so that we know which pixels belong to the ROI. This type of image is not what we expect to see in complex astronomical structures. However, it gives us an idea of relative intensity to the background the source needs to be to see good performance of the boundary in “ideal” circumstances.

The first set of simulations we create is a sequence of four rectangular steps within a  $64 \times 64$  pixel image. In each case, the background intensity level is an average of 0.1 photon counts. We create images with four different maximum average intensities of 1, 3, 5, and 10 photon counts; therefore, the region of the highest strength comes from an expected strength of 10, 30, 50, and 100 times that of the background. The remaining levels take the form of stairs have intensity levels in sequence, linearly from the background to the maximum. The second set of simulations mirror the

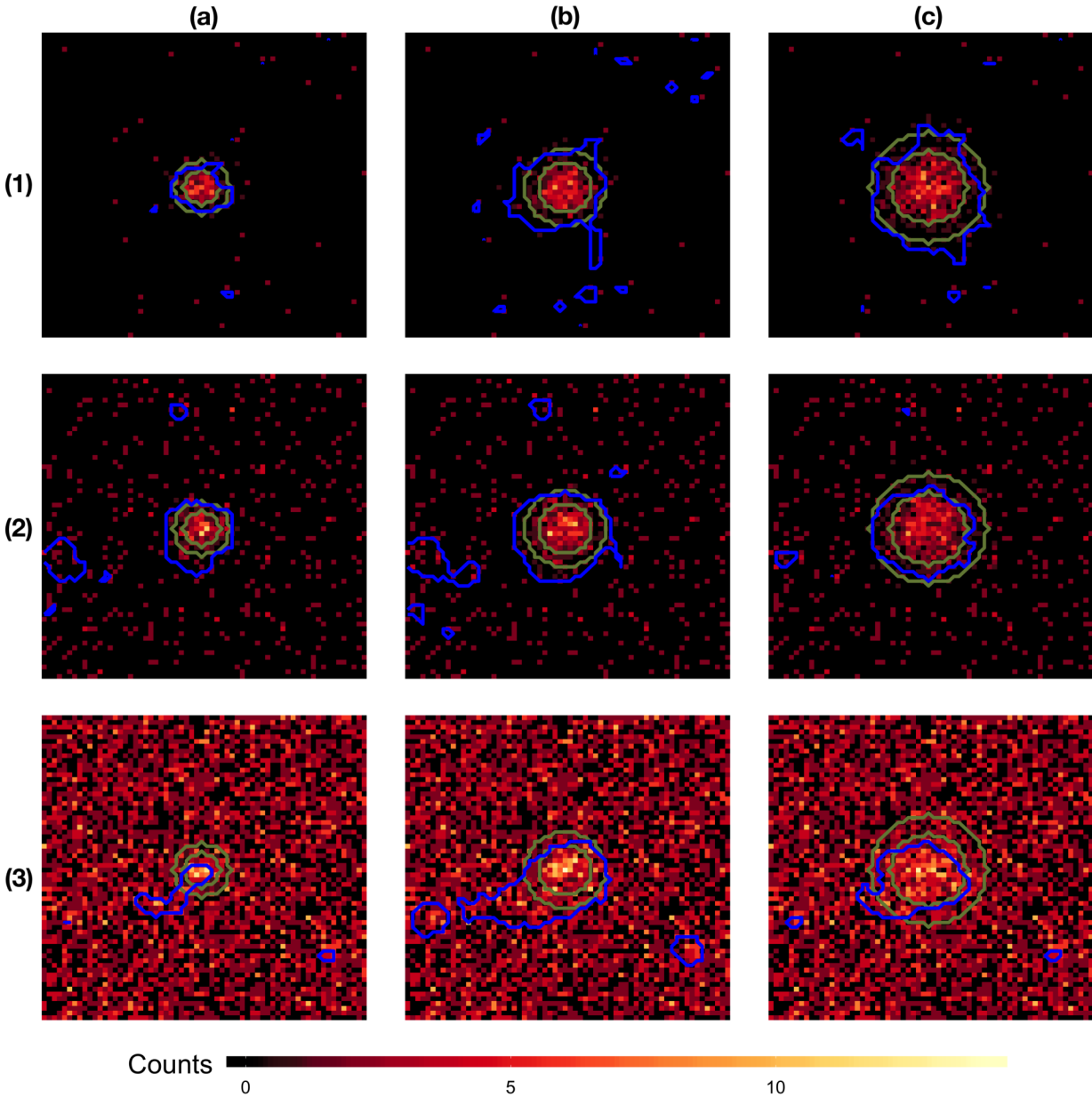


Figure 3.5: The optimal boundary (blue), compared to the one and two sigma levels of the 2D Gaussian (green), overlaid on the observed image. Row and column arrangements are identical to that of Figure 3.2.

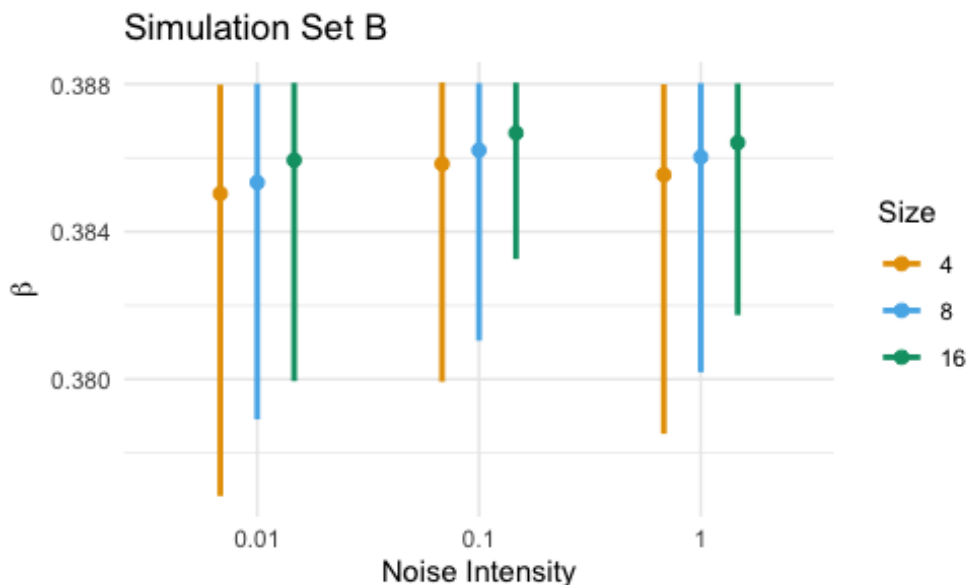


Figure 3.6: Beta scaling with respect to source size (pixels) and intensity (average photon counts). The 95% credible interval around  $\beta$ . The  $\beta$  increases with the size of the source, which is intuitive since the higher the  $\beta$ , the more cohesive our pixel assignments.

same intensity levels as the steps, but the levels are in a concentric square pattern with the highest strength in the center and lowest on the outside.

We take a Poisson realization of the truth image to simulate noise. For each pixel given strength  $\lambda_i$  we draw from the distribution  $\text{Poisson}(\lambda_i)$ . The Poisson realizations are in column (a) of Figure 3.7 and 3.8 for the stairs and concentric squares shapes respectively. The rows of the figures indicate the increasing strengths (1) corresponding to a maximum 10 times the background and (4) corresponding to a maximum 100 times the background. Next, we apply LIRA to the observed image using a flat baseline image of 0 counts and a PSF of a single pixel. The average multiscale counts are shown in column (b) in Figures 3.7 and 3.8. We then sample draws from the posterior of the pixel assignments. Column (c) in Figures 3.7 and 3.8 show the probability map, the percentage of times each pixel is assigned to the ROI. Finally, we use the posterior draws of the pixel assignments to determine the optimal boundary around the ROI by maximizing the posterior. The maximum boundary is shown overlaid on the average multi-scale counts in column (d) of Figure 3.7 and 3.8.

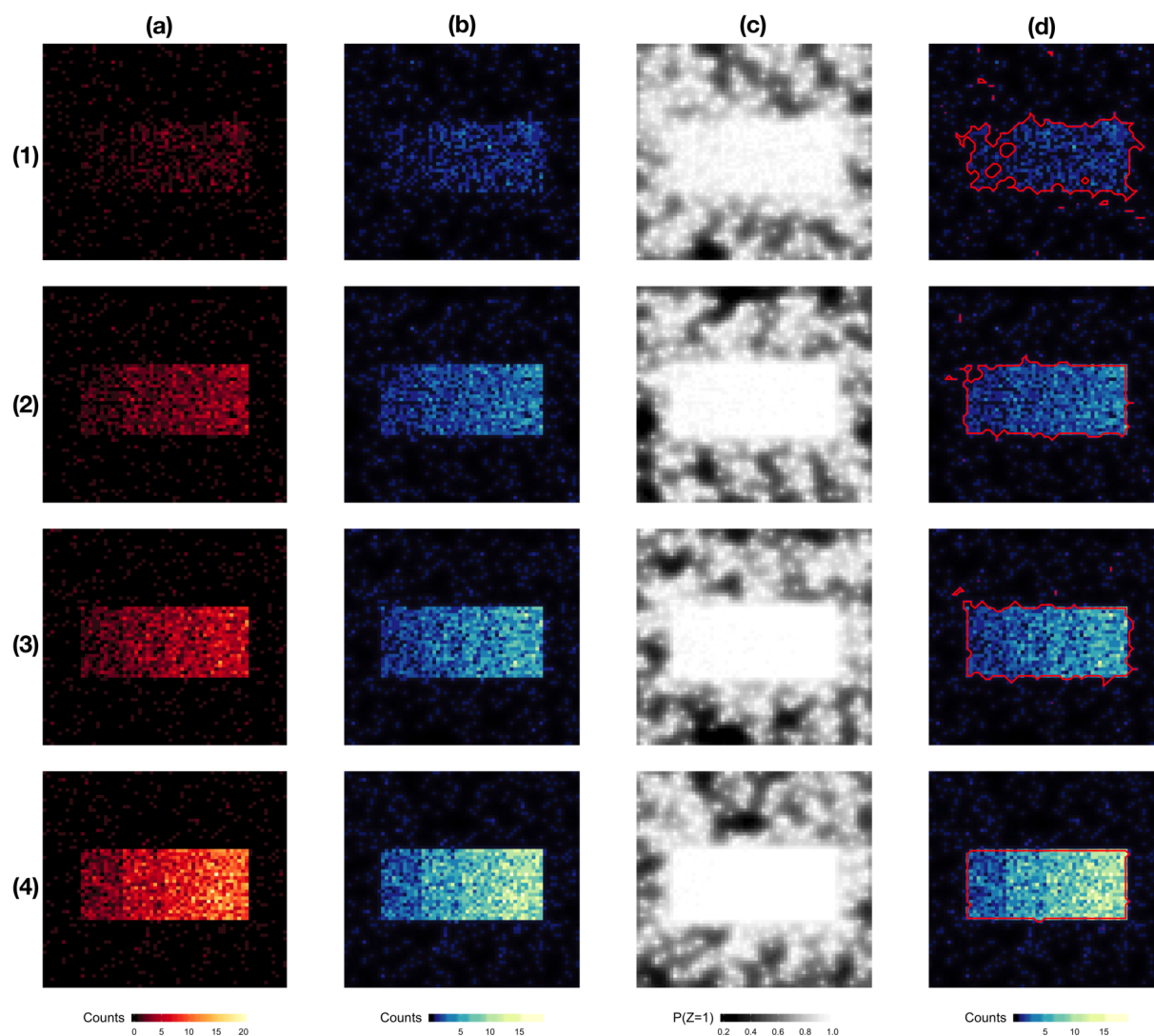


Figure 3.7: The observed image (a), average multi-scale counts (b), probability of a pixel being in the ROI (c), and the final boundary estimate (d) for the simulation in the shape of stairs. The maximum intensity for the simulations is 10, 30, 50, and 100 times that of the background in rows (1), (2), (3), and (4) respectively.

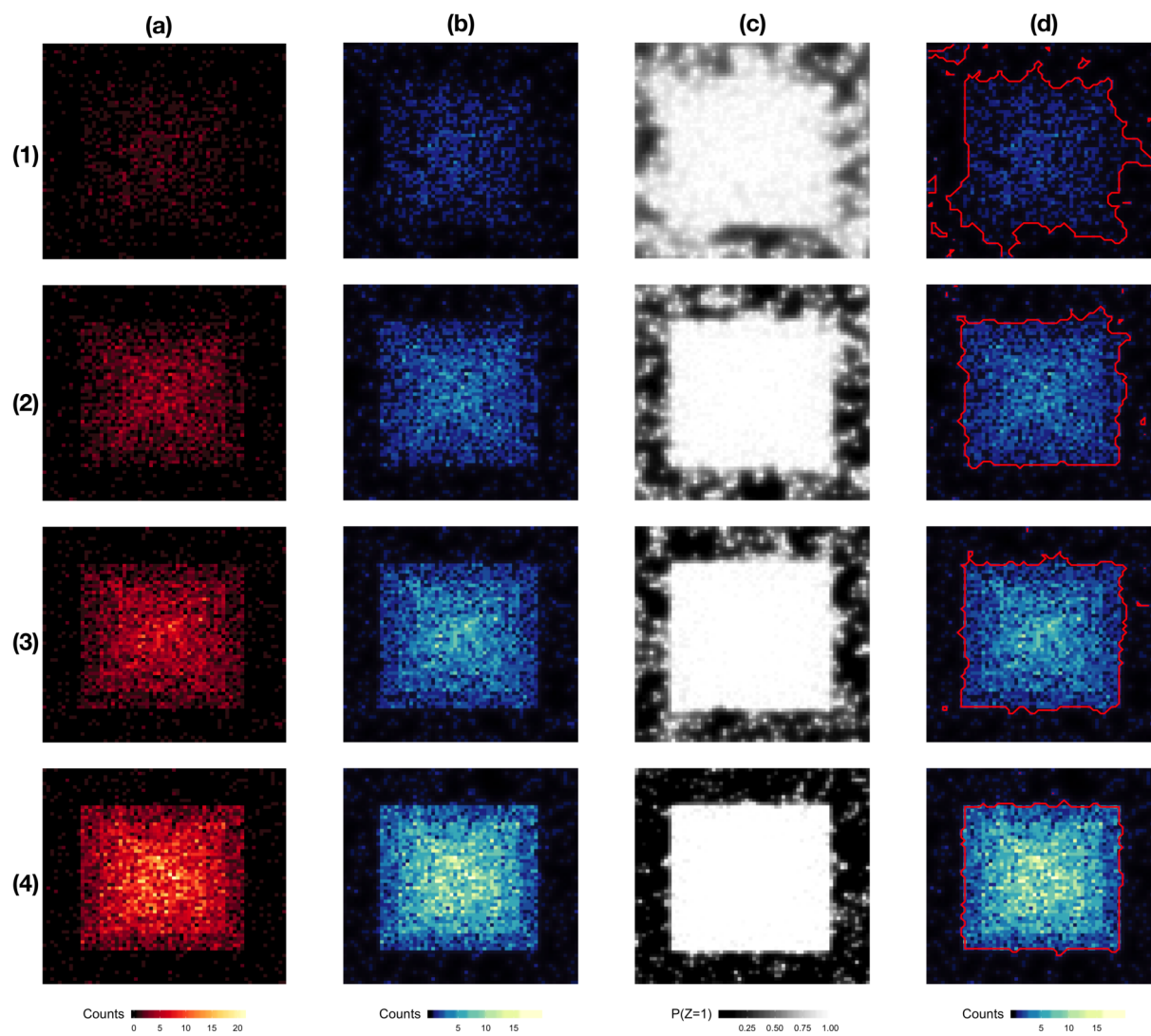


Figure 3.8: The same layout of images as with Figure 3.7, but with the concentric squares pattern.



It is visually apparent that we get a more accurate definition of the boundary as the magnitude of the source increases relative to the background. We quantify this by calculating the error of our boundary estimate in three ways: classification, type I, and type II error. Classification error is determined by the percentage of pixels we incorrectly labeled ROI or background. Type I error is the percentage of pixels incorrectly labeled as the ROI, and type II is the percentage of pixels incorrectly labeled as the background. Figure 3.9 shows the error for each of the four strengths and styles of simulation (steps:bottom, concentric squares:top). Overall the concentric squares had smaller errors than the stairs case, even approaching error rates of less than 1% when the maximum strength is 100 times that of the background, which is roughly representative of the SN 1987 A supernova. We do not expect our performance on SN 1987 A to have the same classification error because it is a diffuse source and has a more complicated PSF, but we expect a source of similar intensity relative to the background to have great performance in ideal circumstances.

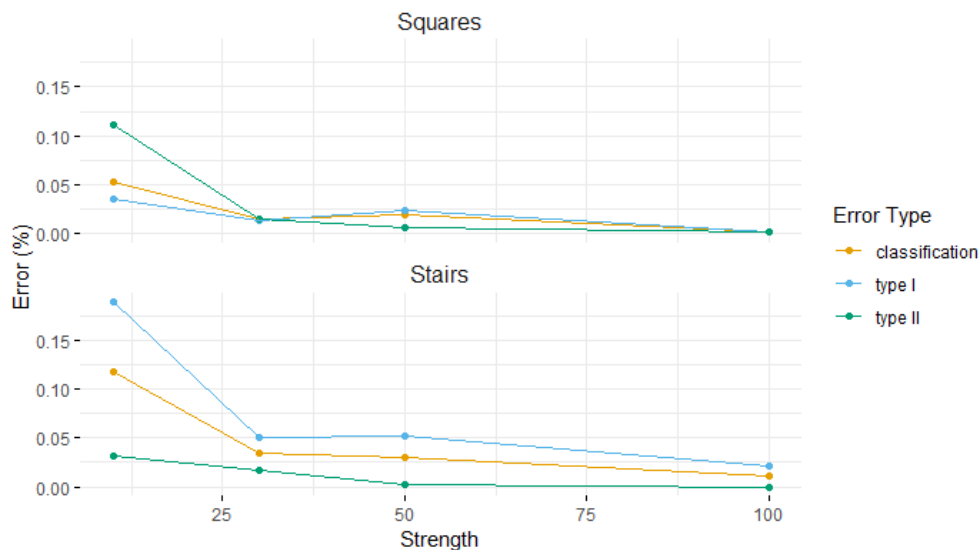


Figure 3.9: The classification, type I, and type II errors for the concentric squares (top) and stairs (bottom) across the four different maximum intensities.

## 3.5 Results

We revisit our motivating examples: the extragalactic jet Obs ID 7873 and supernova SN 87 A. The goal in both cases is to define the optimal boundary around the ROI to help us better understand the morphology of the astronomical source of interest.

### 3.5.1 Extragalactic Jets

Extragalactic jets are highly energetic ejections from supermassive black holes. The matter from jets can span enormous distances ( $> 100$  kpc), far beyond the galaxies' borders. Less than 100 jets have observations at X-ray wavelengths in which we can study their morphology (Massaro et al., 2011). Even fewer observations exist at high redshifts ( $z > 2$ ). In this section, we walk through the multi-phase process of the source 1418–064 (Obs ID 7873), which resides at a redshift of  $z = 3.689$ , taken on July 27, 2007. The jet was first observed in the radio wavelength, where the detail of the morphology is much clearer (Beasley et al., 2002; Ellison et al., 2001). Once targeted, X-ray observation was made with the *Chandra* X-ray observatory (AO8 program [PI: Cheung] from 2007 Jan - June). In McKeough et al. (2016) the authors apply the statistical method described in Stein et al. (2015) to detect the presence of high redshift jets in pre-defined regions. One complication with this method is that in some cases, the power of the detection was not robust to the size and shape of the boundary. We take a different approach by assuming that a single source in the image exists and using our multi-phase method to define a boundary around the ROI. We focus on this particular source because in McKeough et al. (2016), the astronomer originally defined a single boundary rather than multiple nodes as with some other jets, thus simplifying the procedure.

Figure 3.1 (a) shows the original X-ray image taken by *Chandra*. It is difficult to determine the morphology of the jet because only a handful of high energy photons were collected by the telescope, particularly compared to the galaxy in the center of the view. Figure 3.1 (b) shows the average

multi-scale output from LIRA. We implement the same LIRA procedure as described in [McKeough et al. \(2016\)](#). LIRA removes the implicating galaxy source and deconvolves the image from the PSF so that the details within the jet are much easier to make out. For each LIRA sample, we conduct the Gibbs sampler in step 2 to draw samples from the posterior on pixel assignments. The aggregate of the samples is shown in Figure 3.1 (c) in the form of a 2D probability map indicating the pixel-wise estimated probability that the pixel is contained in the ROI. We optimize the posterior to obtain the final boundary estimate. Figure 3.1 (d) shows this final boundary (blue) overlaid on the original X-ray image. We also display the astronomer-defined boundary from [McKeough et al. \(2016\)](#) (green). Our boundary roughly resembles that of the boundary in [McKeough et al. \(2016\)](#), but with a more form-fitting shape. In Obs ID 7873, both the astronomer and our multi-phase method define a single boundary around the object. However, it is typical in other sources to have multiple nodes and varying structure within a single jet, each potentially requiring a different boundary and an expansion of our method to account for multiple sources in an image.

### 3.5.2 Source 87A

SN 1987 A is the nearest observed supernova remnant discovered by astronomers and has become one of the most studied type II supernovas in history. A supernova is an explosion caused by the death of a star as it collapses, with the resulting wave of ejected material and interstellar medium forming a luminous shock called a supernova remnant (SNR). SN 1987 A, from here on referred to as 87A, is of particular interest to X-ray astronomers. Therefore, it has been observed several times throughout the lifetime of the *Chandra* telescope, starting in 2000 ([Burrows et al., 2000](#)). There have been several morphological changes since the first time it was observed with *Chandra*. [Kashyap et al. \(2017\)](#) reanalyzes archival data by reconstructing high-resolution images using LIRA. They reconstruct previously unresolved structures across four different dates 2000-Dec, 2007-Jul, 2011-Mar, and 2015-Sep. Figure 3.10 (a) shows the X-ray observation in 2000-Dec. Using the reconstructed images, we can compare morphological details that have changed across time for 87A

at scales of approx 1/4 arcsec. We revisit this analysis by estimating the boundary on a sequence of images over time to better understand the changing morphology of the supernova.

We estimate the boundary for each observation of 87A to see how the morphology of the SNR changes across time. For step 1, we use the same baseline and empirical PSF as described in [Kashyap et al. \(2017\)](#). Figure 3.10 shows the average multi-scale counts taken across the samples from LIRA. In step 2, we obtain draws from the posterior distribution of pixel assignments. Figure 3.10 (c) contains the pixel-wise probabilities of each pixel being labeled as the ROI. Finally, we optimize the posterior distribution to estimate the boundary around the ROI of the supernova remnant. The boundary estimate for the 2000-Dec observation is shown in Figure 3.10 (d).

We can get an idea of the uncertainty of our estimate by replicating the same procedure on the same image ten times. Figure 3.11 shows two different aggregations of these ten replicated boundaries overlaid on the original X-ray image. The image on the left shows all 10 overlapping ROI. Lighter tiles show where there is little overlap among the ROI, and darker tiles show where more ROI overlap. The figure on the right simplifies this image by drawing contours at the boundary of where at least 5 ROI overlap and all 10 ROI overlap. At the level of only 5 boundaries overlapping, we lose the separated regions that we see in an overlap of all 10 boundaries.

We run this method on the four observations presented in [Kashyap et al. \(2017\)](#). Unfortunately, the changes in the type of observation and PSF between the first two observations and the second two make the series of images challenging to compare <sup>2</sup>. After the first image, the bright spots in the center are no longer differential from the background. Ideally, we would be able to see the same ROI in each observation and be able to track its morphology over time; however, the significant amount of background structure in the multi-scale counts prohibits meaningful comparisons. An improved PSF will allow us to perform improved inference on the multi-scale counts, thus reducing uncertainty on the boundary, and giving more consistency across time.

---

<sup>2</sup>The first two are ACIS-S observations, and the second two are ACIS-S + HETGS observations

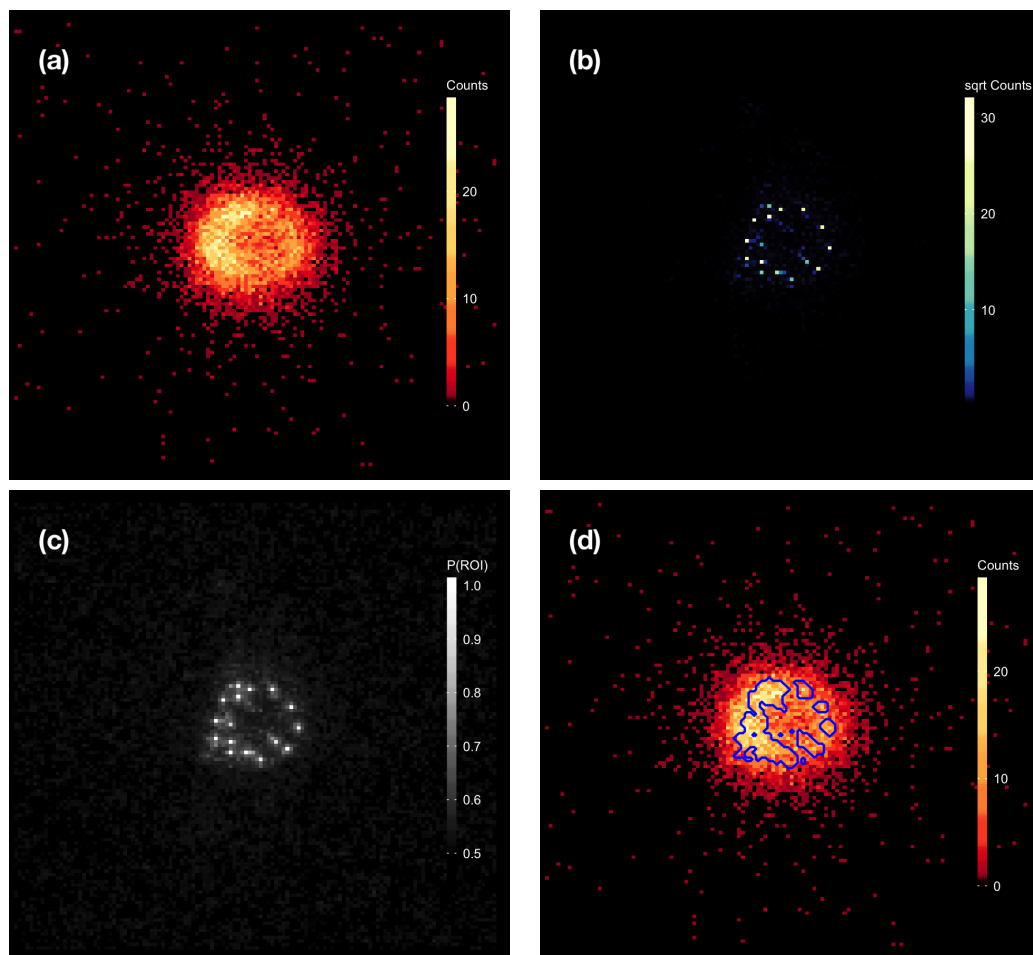


Figure 3.10: The broadband image of 87A from December 2000 (a). The average multi-scale counts from LIRA (b). The 2D probability map aggregated across the distribution of pixel assignments (c). The optimal boundary overlaid on the original X-ray image (d).

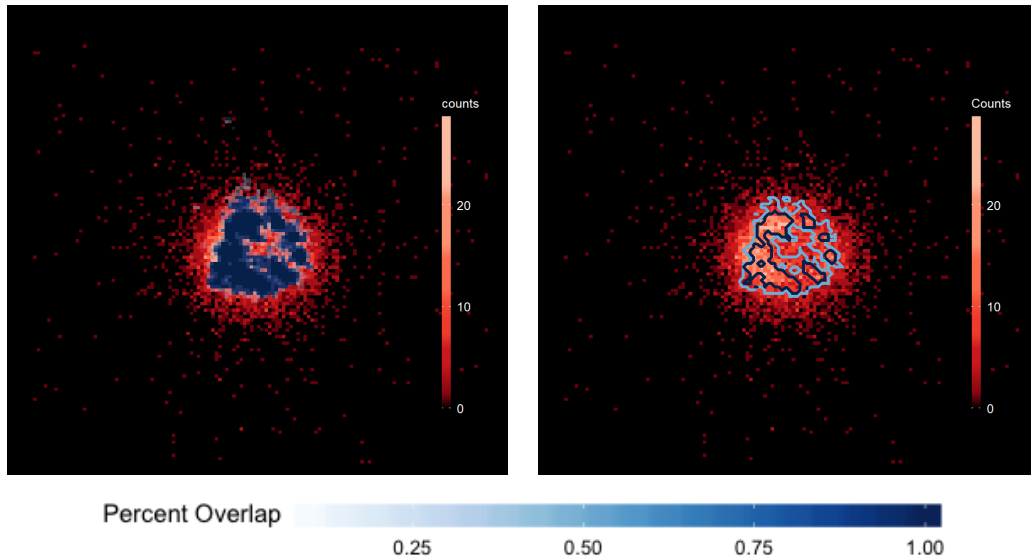


Figure 3.11: Aggregation of 10 repetitions of the estimated boundary overlaid on the original X-ray image. On the left we plot all ROI on top of one another with darker colors where most boundaries overlap, lighter colors are where fewer boundaries overlapped. On the Right we draw a boundary where 5 ROI and all 10 ROI overlap with the lighter and darker blue respectively.

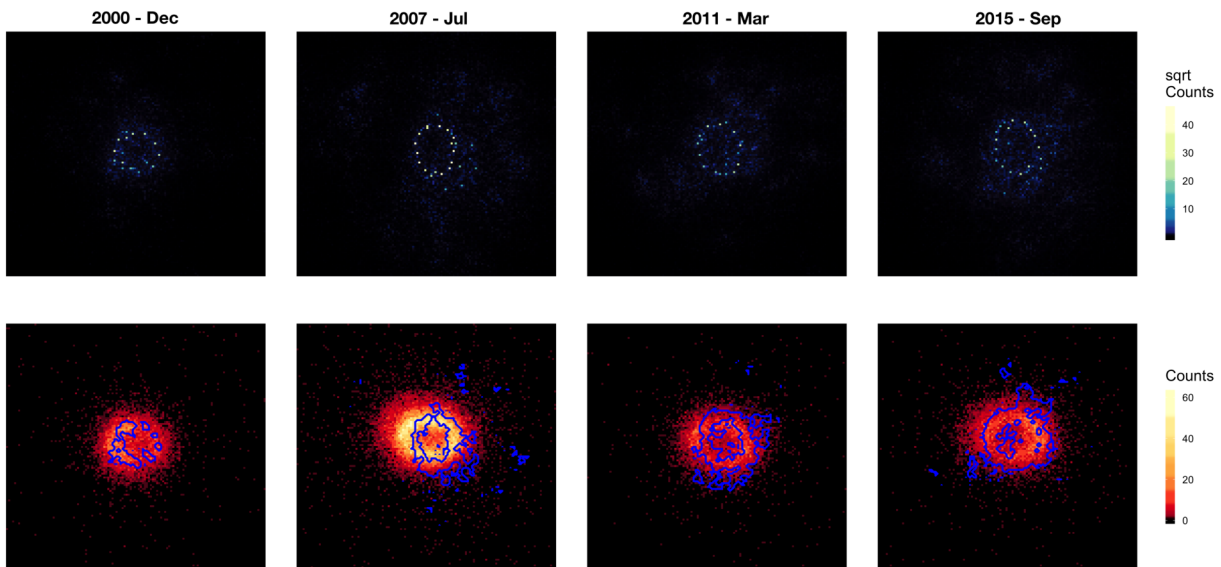


Figure 3.12: *Chandra* broad band images are shown in the bottom row overlaid with the optimal boundary determined by our method. On the top row we have the LIRA realizations of each of these images. From left to right we have 87A observations from 2000-Dec, 2007-Jul, 2011-Mar, and 2015-Sep.

## 3.6 Conclusion

Extensions to improve the boundary algorithm are generalizing the model to allow for more than one ROI and a method for estimating the uncertainty. The application to X-ray sources present a clear motivation for extending this model to allow for more than two ROI in a single image with varying intensities. In the jet example, the jets can be divided into different segments representing different nodes or other features. In the 87A example, particularly in the later observations, it seems that there are two sources of radiation, the bright spots, and a more extended fainter source. By separating these two from one another, we would get a better idea of the shape of the bright spots we wish to map over time. The MAP method we present here only produces a *point estimate* of the boundary. We try to alleviate this by running the algorithm many times and aggregating the results. Defining a model-based uncertainty on a boundary is difficult due to the computational restrictions of evaluating exact posterior probabilities. We suggest a more rigorous way of estimating uncertainty as a future extension to this work.

In our modeling decision, we could have chosen to update LIRA jointly model the pixel assignments and multi-scale counts, but instead, we choose to use a multi-phase model. This decision was for the simplicity of extending existing work as LIRA is already a frequently used algorithm. In Appendix C we explore the conditions in which our multi-phase model must meet in order to have equivalent inference with a fully joint model.

In summary, we construct a novel multi-phase method that defines a boundary around the ROI in images of complex, irregularly shaped, and diffuse astronomical sources. Estimating this boundary can be useful in understanding the morphology of these images in an objective approach. We demonstrated that this boundary algorithm was useful in delineating extragalactic jets as well as depicting how a SNR changes over time.

This page is intentionally left blank.



# Afterword

In this dissertation, we apply a variety of statistical modeling techniques to the domains of sports analytics and astronomy. In Part I, we create two methods to model changes in athlete ability over time. First, we extend the Plackett-Luce model to incorporate a growth curve to model athlete ability over time for multi-competitor sports. The growth curve model is flexible enough to accommodate any career type, regardless of the sport. This new model proves to be successful in predicting the performance of athletes. We evaluate the model by estimating the correlation of our predictions and the actual competition results and find that even projecting out a couple of seasons, we still maintain a relatively high correlation. Second, we use the same growth curve and apply it as a non-linear mixed-effects regression model to estimated athlete rating over time across both head-to-head and multi-competitor sports. The growth curve model is useful for describing the time-varying nature of Elo, Glicko, and GH rating systems. We validate the model by creating predictive intervals and estimating the coverage of our predictions to show that coverage is as expected in the 95% CI for up to four years. We use the growth curve for not only prediction but also an exploratory analysis of the fitted career trajectories through clustering.

In Part II, we establish a method to define boundaries around complex astronomical objects in X-ray images. Typical boundary algorithms do not perform well due to the inherent sparsity in X-ray images. The first step is to use LIRA to reconstruct the X-ray image to elicit detailed structure within the ROI. The next step is to assign each pixel to the ROI or the background. We build a distribution for these pixel assignments through a novel model that imposes an Ising distribution

a priori on the pixel assignments. Finally, we optimize the posterior on the pixel assignments to estimate the boundary of the ROI. We apply this method on two different sources: a high-redshift, extragalactic jet, and a series of images over time of a supernova.

## Final Comments

- **Modeling Decisions**

The models presented in this dissertation are the product of an immense amount of exploration and decision making. Deciding the functional form of the growth curve was crucial in the success of the predictive models presented in Chapters 1 and 2. Through combing the literature, we discovered that the growth curves already being used were simple (e.g., quadratic or cubic) or were designed for specific purposes and not generalizable (e.g., Moudud et al., 2008). The first step was creating a more flexible, functional form for the growth curve. Candidate models consisted of a mixture of generalized gamma and logistic components to capture any possible career trajectory shape. However, a combination of these models suffered from identifiability issues, which made inference on the parameters difficult. We settled on a product of a polynomial with orthogonal components and exponential decay. Making the polynomial coefficients and the intercept vary per athlete gives the model flexibility in the shape between athletes while implementing this as random-effects allows for sharing information across the sport. Finally, this model is appealing because the decay rate can easily be compared across sports revealing trends such as how quickly an athlete's performance declines. The shape of the athletes' career trajectories can be compared against one another using the polynomial coefficients as done in the clustering exercise at the end of Chapter 2.

The major modeling decision in Chapter 3 was the decision to impose the Ising distribution as a prior on the pixel assignments. The spatial characteristics translate well to image segmentation because if a pixel is within the ROI, then it is much more likely the adjacent pixel is also within the ROI. Furthermore, since it is a widely known and commonly used distribution,

there exist simple and well-tested ways for sampling from the distribution, which we can take advantage of in our MCMC. Lastly, it can be extended to more than a single source through its generalized form, the Potts model. Although there exist examples of the Ising distribution used in image segmentation, it traditionally sued for modeling 2D magnetic fields. Applying the Ising distribution often gets initially disregarded as a misuse of the Ising distribution by physics experts. It is important to realize that the Ising distribution was made to model scientific phenomena, and that itself does not have hard science encoded into its form. As mentioned in Chapter 3, many examples exist of successfully using it in image segmentation.

- **Importance of Estimating & Propagating Uncertainty**

Being able to quantify the uncertainty in estimates is crucial when making claims based on statistical models. Therefore, it is essential to propagate uncertainty through the model, particularly in cases of multi-phase inference. When making predictions, estimating the uncertainty of the projection gives the reader a better understanding of future results than a point estimate. For example, stating that we predict the New England Patriots will win the season is less informative than saying they have a 60% of winning the season. In Part I, we incorporate uncertainty into all predictions using a 95% CI and use these probabilistic predictions when comparing athletes to one another. In Chapter 2, we encourage using the Glicko and GH rating methods as they update the estimate of the uncertainty (RD) along with updating the estimate of rating. Athletes who appear in more competitions have less uncertainty on their rating estimates. Quantifying uncertainty in boundary estimation is a more difficult task due to the model's spatial nature. We solve this problem through a bootstrap approach, replicating the process to flesh out many probable boundary arrangements. Approaching this problem formally through using the probability distribution of the pixel assignments is limited computationally by the enormous number of possible arrangements. Overall, defining an uncertainty around a boundary is an unsolved task and one that would be impactful in astronomy and other fields.

- **Multi-Phase Inference**

We compare the applications of the models in Chapter 1 (one-step approach) and in Chapter 2 (two-step approach) to multi-competitor sports. The two models are similar because they both use the Plackett-Luce model to infer the ratings and the growth curve model to infer the time-varying nature of the ratings. In both cases we are interested in finding the distribution of the growth curve parameters  $(\beta, \alpha, \omega)$  and the ability parameter  $(\theta)$ . In the one-step approach, we use time-varying parameters and event data to estimate both the rating and the growth curve parameters at the same time using the Plackett-Luce likelihood in Equation 1.2. In this case, we constrain the ratings under the assumptions of our growth curve, shown by the equality in Equation 1.3. In the two-step scenario, we first use event data to estimate player rating using the GH rating system described in Section 2.2.1. GH is not identical to the full Bayesian case described in the one-step approach, but rather provides an estimate of athlete rating over time and an estimate on the uncertainty by updating the rating after each period. In the second step, we impose the growth curve on the rating estimates as non-linear, mixed-effects, regression model through the relationship defined in Equation 2.16. In this case, the original estimate of the rating is unaffected by the growth curve relationship in the second step to describe the athletes' career trajectories. The main reason for using the two-step model is that the second step in the two-step model can be applied to a variety of rating schemes, thus increasing the generalizability of the application. The one-step model is limited to the multi-competitor use case since the growth curve is ingrained into the Plackett-Luce Likelihood.

In Chapter 3, we discuss a method that uses the output of the pre-existing LIRA algorithm  $\tilde{\Lambda}$  to build a distribution of pixel assignments  $Z$  for a particular observation  $Y$ . Rather than model both the parameters, multi-scale counts  $\Lambda$  and the pixel assignments  $Z$ , jointly, we choose to use a multi-phase method. One reason for using a multi-phase method is simplicity. LIRA was published 10 years ago, and since then, it has been used in a variety of settings (McKeough et al., 2016; Kashyap et al., 2017). Rather than modifying the entire

LIRA algorithm to estimate the pixel assignments and multi-scale counts jointly, it is much simpler to create an additional algorithm that uses LIRA output as a pre-processing step. Furthermore, LIRA is well-tested and well-formulated, whereas we know the model on pixel assignments is misspecified. We assign pixels of a diffuse source with no hard edges to a binary state, thus creating hard edges where there are none. Using LIRA as a pre-processing step for estimating the pixel assignments alleviate the effects that the misspecified model for pixel assignments that could bias our estimates on multi-scale counts.

Applying a model to observed data simplifies “real-life” mechanics. As statisticians, we strive to make this model as accurate and as assumption-free as possible to minimize error due to this simplification. However, if our data is pre-processed, then how we build our model reflects not the raw data, but the analyses put into creating the data we are given. As [Meng \(2014\)](#) remarks, these types of analyses are becoming more prevalent in a variety of applied settings since there is often a degree of separation between the experts in data collection and the experts in data analysis. We have a better understand and can collect more detailed data about the underlying system that analyses are complicated enough that modeling all components jointly is computationally difficult. [Meng](#) cites multi-phase inference as one of the open problems in the field of Statistics “that could win a Nobel Prize”. We barely scratch the surface of the depth and breadth of the impact this field of inference has on the science and social science communities.

This page is intentionally left blank.

# Appendix A

## Orthogonal Polynomials

Understanding the creation of the orthogonal polynomials is important because the process must be replicated to project new time points into the orthogonal space. We start with a sequence of discretized time points  $\mathbf{t} = \{0, 1, \dots, T\}$ . We want to create an orthogonal basis of  $\mathbf{t}, \mathbf{t}^2, \dots, \mathbf{t}^p$ . We will denote the orthonormal basis as  $\mathbf{t}_{-1}^*, \mathbf{t}_0^*, \mathbf{t}_1^*, \dots, \mathbf{t}_p^*$  and corresponding orthogonal basis as  $\mathbf{u}_{-1}, \mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_p$  where  $\mathbf{t}_i^* = \mathbf{u}_i / \|\mathbf{u}_i\|$ . This can be done by the following steps:

1. Center all elements of  $\mathbf{t}$  by the mean  $\bar{t}$ :

$$\{\mathbf{t}' = \mathbf{t} - \bar{t}; \bar{t} = \frac{1}{T+1} \sum_{i=0}^T i\}$$

2. Set basis vectors  $\mathbf{u}_{-1} = \mathbf{0}, \mathbf{u}_0 = \mathbf{1}, \mathbf{u}_1 = \mathbf{t}'$  which are trivially orthogonal.

3. Define  $\ell_i = \langle \mathbf{u}_i, \mathbf{u}_i \rangle$ ,  $\alpha_i = \frac{\langle \mathbf{u}_i, \mathbf{t}' \rangle}{\ell_i}$ . Note that the L2-norm is  $\|\mathbf{u}_i\| = \sqrt{\ell_i}$ .

4. For  $i = 2, \dots, p$  we can now solve for the orthogonal polynomial components using the Gram

Schmidt process on increasing orders of time, rewritten using a recursion relation.

$$\begin{aligned}\sqrt{\ell_i} \mathbf{t}_i^* &= \mathbf{u}_i = \mathbf{t}' \mathbf{u}_{i-1} - \sum_{j=0}^{i-1} \frac{\langle \mathbf{t}' \mathbf{u}_{i-1}, \mathbf{u}_j \rangle}{\ell_j} \mathbf{u}_j \\ &= (\mathbf{t}' - \alpha_{i-1}) \mathbf{u}_{i-1} - \frac{\ell_{i-1}}{\ell_{i-2}} \mathbf{u}_{i-2}\end{aligned}$$

Ignoring  $\mathbf{t}_{-1}^*$  and  $\mathbf{t}_0^*$ , we now have a set of orthogonal polynomials  $\mathbf{t}_1^* \dots \mathbf{t}_p^*$  up to time  $T$ . The model is fit using these orthogonal polynomials. In practice the orthogonal polynomial basis was created using the **poly** function in **R**, which also returns all  $\ell_i$  and  $\alpha_i$  coefficients.

In order to make projections to include future time periods  $\mathbf{s} = \{0, 1, \dots, T, T+1, T+2, \dots\}$  we must project these values into the same orthogonal space we use to fit the parameters in our model. To do so we keep the same coefficients,  $\ell_i$  and  $\alpha_i$ , and use them to project the time points  $\mathbf{s}$  into the same space using the equation in step 4s. Explicitly, if we created our orthogonal basis using  $\mathbf{t}_i^{*(0,1,\dots,T)}$ ;  $i = 1, \dots, p$ , but we want to extend it to  $\mathbf{t}_i^* = \mathbf{u}_i = \{\mathbf{t}_i^{*(0,1,\dots,T)}, \mathbf{t}_i^{*(T+1,T+2,\dots)}\}$  then we modify step 4 to be:

$$\sqrt{\ell'_i} \mathbf{t}_i^* = ((\mathbf{s} - \bar{t}) - \alpha'_{i-1}) \mathbf{u}_{i-1} - \frac{\ell'_{i-1}}{\ell'_{i-2}} \mathbf{u}_{i-2},$$

where  $\ell'_i$ ,  $\alpha'_i$ , and  $\bar{t}$  are calculated only using  $\mathbf{t}_i^{*(0,1,\dots,T)}$ . Note that the new polynomial vectors are not themselves orthonormal, but are projected into the same space we fit the coefficients.



## Appendix B

# Hierarchical Clustering

Agglomerative hierarchical clustering begins with all observations as separate clusters, and iteratively combines observations and clusters in the order of smallest to largest distance. We choose a type of linkage or method to merge clusters one at a time, in this case we choose the Ward's method. Hierarchical clustering will always merge the observations or clusters with the smallest distance as determined by Ward's method at each step. A way of visualizing this process is looking at the partial dendrogram in Figure 2.9, where the branches represent at what distance (y-axis) each cluster or observation was merged. Ward's method provides a decision metric based on the distance between two clusters and the noise within each cluster. This method tends to make compact clusters and perform well when clusters are noisy. The Ward's metric is between two clusters  $C_a$  and  $C_b$  is

$$\Delta(C_a, C_b) = \sum_{i \in C_a} \sum_{j \in C_b} d_{ij}, \quad (\text{B.1})$$

where  $d_{ij}$  is the distance between athletes  $i$  and  $j$ . The main downside to this type of clustering is that it is greedy: once a cluster is merged together it will never be separated.

Using the silhouette statistic is a good way to evaluate cluster fit. The silhouette statistic is calculated for each observation and measures how well the observation fits within it is labeled

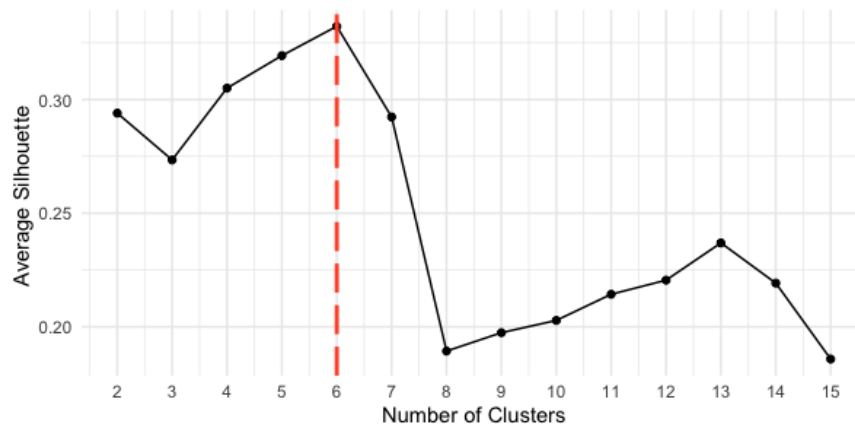


Figure B.1: Average silhouette values for number of clusters 2 through 15. The maximum average silhouette occurs with 6 well defined clusters, denoted by the red dotted line.

cluster, and how far away it is from the next closest cluster. Mathematically this concept is written as:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}, \quad (\text{B.2})$$

where  $a_i$  is the average distance between an observation and other observations in its assigned cluster and  $b_i$  is the average distance between an observation and other observations in the next closest cluster. If  $s_i$  is close to 1, then the observation is thought to be well assigned and if it is negative it is likely incorrectly assigned. The silhouette statistic can also be used in selecting the optimal number of clusters. If we define

$$\bar{s}_k = \frac{1}{n} \sum_i^n s_i$$

then the optimal number of clusters would be  $\max_k \bar{s}_k$ . Figure B.1 shows the  $s_i$  for 2 through 15 clusters. The optimal number of clusters is 6 which is where the maximum average silhouette statistic occurs.

## Appendix C

# Model Compatibility for the LIRA + Ising Model

We formulate the joint probabilities for  $\Lambda$  and  $Z$  in the joint setting and in the multi-phase setting to set conditions where the two methods may be equal. We can build a joint probability model that estimates the pixel assignments  $Z$  and multi-scale counts  $\Lambda$  simultaneously. We will indicate the joint model as  $\mathcal{J}$ :

$$P_{\mathcal{J}}(\Lambda, Z|Y) \propto f(Y|\Lambda, Z)\pi_{\mathcal{J}}(\Lambda, Z)$$

Instead, we approach this in two steps. The two steps of our LIRA-Ising model are explicitly written out here. For simplicity we ignore the extraneous parameters and only focus on  $\Lambda, Z$  and the observed image  $Y$ . We can write the probability model for the first step, denote as  $S_1$ :

$$P_{S_1}(\Lambda|Y) \propto f(Y|\Lambda)\pi_{S_1}(\Lambda)$$

Taking a single draw of the multi-scale counts  $\tilde{\Lambda}$  from  $S_1$ , we can write our probability model for

$S_2$

$$P_{S_2}(Z|\tilde{\Lambda}) \propto P_{S_2}(\tilde{\Lambda}|Z)\pi_{S_2}(Z)$$

The combination of the two gives us the complete multi-phase process outlined in Chapter 3 ( $S$ ):

$$P_S(\tilde{\Lambda}, Z|Y) = P_{S_1}(\Lambda|Y)P_{S_2}(Z|\tilde{\Lambda})$$

$$P_S(\tilde{\Lambda}, Z|Y) \propto f(Y|\tilde{\Lambda})\pi_{S_1}(\tilde{\Lambda})\frac{P_{S_2}(\tilde{\Lambda}|Z)\pi_{S_2}(Z)}{P_{S_2}(\tilde{\Lambda})}$$

Model  $S$  is equivalent to model  $\mathcal{J}$  under the following two conditions.

1.  $Y$  and  $Z$  are independent given  $\Lambda$ ,

$$f(Y|\Lambda) = f(Y|\Lambda, Z) .$$

2. The prior for the LIRA step is equivalent to the marginal prior of  $\Lambda$  in the joint model. That is,  $P_{S_2}(\Lambda|Z)\pi_{S_2}(Z) = \pi_{\mathcal{J}}(\Lambda, Z)$  which implies

$$\pi_{S_1} = \int \pi_{\mathcal{J}}(\Lambda, Z)dZ = \int P_{S_2}(\Lambda|Z)\pi_{S_2}(Z)dZ$$

It is safe to assume that the first condition is trivially satisfied. The second condition is not guaranteed to be true.

# Bibliography

- Abraham, C., P. A. Cornillon, E. Matzner-Løber, and N. Molinari (2003). Unsupervised Curve Clustering using B-splines. *Scandinavian Journal of Statistics* 30(3), 581–595.
- Adams, R. and L. Bischof (1994). Seeded Region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16(6), 641–647.
- Airoldi, E. M., C. Huttenhower, D. Gresham, C. Lu, A. A. Caudy, M. J. Dunham, J. R. Broach, D. Botstein, and O. G. Troyanskaya (2009). Predicting cellular growth from gene expression signatures (predicting cellular growth). *PLoS Computational Biology* 5(1), e1000257.
- Alqallaf, F. and P. Gustafson (2001). On cross-validation of Bayesian models. *Canadian Journal of Statistics* 29(2), 333–340.
- Baker, R. D. and I. G. McHale (2013). Forecasting exact scores in National Football League games. *International Journal of Forecasting* 29(1), 122–130.
- Bartolucci, F. and T. B. Murphy (2015). A finite mixture latent trajectory model for modeling ultrarunners’ behavior in a 24-hour race. *Journal of Quantitative Analysis in Sports* 11(4), 193–203.
- Béal, D., P. Brasseur, J. M. Brankart, Y. Ourmières, and J. Verron (2010). Characterization of mixing errors in a coupled physical biogeochemical model of the North Atlantic: Implications for nonlinear estimation using Gaussian anamorphosis. *Ocean Science* 6(1), 247–262.
- Beale (1996). Exact distribution of energies in the two-dimensional ising model. *Physical review letters* 76(1), 78.
- Beasley, A. J., D. Gordon, A. B. Peck, L. Petrov, D. S. MacMillan, E. B. Fomalont, and C. Ma (2002, jul). The VLBA calibrator survey—VCS1. *The Astrophysical Journal Supplement Series* 141(1), 13–21.
- Bell, A., J. Smith, C. E. Sabel, and K. Jones (2016). Formula for success: Multilevel modelling of Formula One Driver and Constructor performance, 1950-2014. *Journal of Quantitative Analysis in Sports* 12(2), 99–112.
- Bennett, J. and J. Wakefield (2001). Errors-in-variables in joint population pharmacokinetic/pharmacodynamic modeling. *Biometrics* 57(3), 803–812.

- Bentrem, F. W. (2010). A Q-Ising model application for linear-time image segmentation. *Central European Journal of Physics* 8(5), 689–698.
- Bertin, E. and S. Arnouts (1996). SExtractor: Software for source extraction. *Astronomy and Astrophysics Supplement Series* 117(2), 393–404.
- Blangiardo, M., A. Hansell, and S. Richardson (2011). A Bayesian model of time activity data to investigate health effect of air pollution in time series studies. *Atmospheric Environment* 45(2), 379–386.
- Blocker, A. W. and X.-L. Meng (2013). The potential and perils of preprocessing: Building new foundations. *Bernoulli* 19(4), 1176–1211.
- Bobin, J., F. Sureau, and J. L. Starck (2016). Cosmic microwave background reconstruction from WMAP and Planck PR2 data. *Astronomy and Astrophysics* 591, 1–12.
- Bornn, L., D. Cervone, A. Franks, and A. Miller (2017). Studying Basketball Through the Lens of Player Tracking Data. In *Handbook of Statistical Methods and Analyses in Sports*.
- Bradley, R. A. and M. E. Terry (1952). Biometrika Trust Rank Analysis of Incomplete Block Designs : I . The Method of Paired Comparisons Published by : Oxford University Press on behalf of Biometrika Trust Stable URL : <https://www.jstor.org/stable/2334029>. 39(3), 324–345.
- Bradlow, E. T. and P. S. Fader (2001). A Bayesian lifetime model for the hot 100 billboard songs. *Journal of the American Statistical Association* 96(454), 368–381.
- Brander, J. A., E. J. Egan, and L. Yeung (2014). Estimating the effects of age on NHL player performance. *Journal of Quantitative Analysis in Sports* 10(2), 241–259.
- Breslow, N. and J. Crowley (1974). A Large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship. *Annals of Statistics* 2(3), 437–453.
- Bürkner, P.-C., J. Gabry, and A. Vehtari (2019). Approximate leave-future-out cross-validation for Bayesian time series models.
- Burrows, D. N., E. Michael, U. Hwang, R. McCray, R. A. Chevalier, R. Petre, G. P. Garmire, S. S. Holt, and J. A. Nousek (2000, nov). The x-ray remnant of SN 1987a. *The Astrophysical Journal* 543(2), L149–L152.
- Caron, F. and Y. W. Teh (2012). Bayesian nonparametric models for ranked data. *Advances in Neural Information Processing Systems* 2, 1520–1528.
- Chambers, J. and T. Hastie (1992). *Statistical Models in S*. CRC Press.
- Connors, A. and D. A. van Dyk (2007). How to win with non-Gaussian data: Poisson goodness-of-fit. *Statistical Challenges in Modern Astronomy IV* 371, 101–117.
- Ebeling, H., D. A. White, and F. V. N. Rangarajan (2006). ASMOOTH: a simple and efficient algorithm for adaptive kernel smoothing of two-dimensional imaging data. *Monthly Notices of the Royal Astronomical Society* 368(1), 65–73.

- Ebeling, H. and G. Wiedenmann (1993). Detecting structure in two dimensions combining Voronoi tessellation and percolation. *Physical Review E* 47(1), 704–710.
- Ellison, S., Y. L., I. Hook, M. Pettini, J. Wall, and P. Shaver (2001). The CORALS survey I: New estimates of the number density and gas content of damped Lyman alpha systems free from dust bias. *Astronomy and Astrophysics* 379, 393–406.
- Elo, A. E. (1978). *The rating of chessplayers, past and present*. Batsford chess books. London: Batsford.
- Esch, D. N., A. Connors, M. Karovska, and D. A. van Dyk (2004). An Image Restoration Technique with Error Estimates. *The Astrophysical Journal* 610(2), 1213–1227.
- Freeman, P. E., V. Kashyap, R. Rosner, and D. Lamb (2002). A wavelet-based algorithm for the spatial analysis of poisson data. *The Astrophysical Journal Supplement Series* (138), 185–218.
- Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.
- Gelman, A. and D. B. Rubin (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science* 7(4), 457–511.
- Geman, S. and D. Geman (1984, Nov). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-6*(6), 721–741.
- Glickman, M. E. (1999, nov). Parameter estimation in large dynamic paired comparison experiments. *Applied Statistics* 48, 377–394.
- Glickman, M. E. and J. Hennessy (2015). A stochastic rank ordered logit model for rating multi-competitor games and sports. *Journal of Quantitative Analysis in Sports* 11(3), 131–144.
- Glynn, C. and S. T. Tokdar (2017). A switching dynamic generalized linear model to detect abnormal performances in Major League Baseball. In *MIT Sloan Sports Analytics Conference*, pp. 1–29.
- González-Gaitán, S., R. S. De Souza, A. Krone-Martins, E. Cameron, P. Coelho, L. Galbany, and E. E. Ishida (2019). Spatial field reconstruction with INLA: Application to IFU galaxy data. *Monthly Notices of the Royal Astronomical Society* 482(3), 3880–3891.
- Herbrich, R., T. Minka, and T. Graepal (2007). TrueSkill: A Bayesian Skill Rating System. In *Advances in Neural Information Processing Systems*.
- Hoffman, M. D. and A. Gelman (2011). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo.
- Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Physik* 31, 253–258.

- Jacob, P. E., L. M. Murray, C. C. Holmes, and C. P. Robert (2017). Better together? Statistical learning in models made of modules. pp. 1–31.
- Jacques, J. and C. Preda (2014). Functional data clustering: A survey. *Advances in Data Analysis and Classification* 8(3), 231–255.
- Jensen, S. T., B. B. McShane, and A. J. Wyner (2009). Hierarchical bayesian modeling of hitting performance in baseball. *Bayesian Analysis* 4(4), 631–652.
- Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed. ed.). Springer series in statistics. New York: Springer.
- Jones, D. E., V. L. Kashyap, and D. A. van Dyk (2015). Disentangling overlapping astronomical sources using spatial and spectral information. *The Astrophysical Journal* 808(2), 137.
- Kashyap, V. L., D. Van Dyk, K. McKeough, F. Primini, D. Jerius, A. Gowrishankar, and A. Siemiginowska (2017). X-raying the evolution of SN 1987A. *Proceedings of the International Astronomical Union* 12(S331), 284–289.
- Li, N. and M. Stephens (2003). Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics* 165(December), 2213–2233.
- Liu, F., M. J. Bayarriy, and J. O. Berger (2009). Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis* 4(1), 119–150.
- Luce, R. D. R. D. (1959). *Individual choice behavior; a theoretical analysis*. New York: Wiley.
- Lunn, D., J. Barrett, M. Sweeting, and S. Thompson (2013). Fully Bayesian hierarchical modelling in two stages, with application to meta-analysis. *Journal of the Royal Statistical Society. Series C: Applied Statistics* 62(4), 551–572.
- Luo, Y., K. Al-Harbi, Y. Luo, and K. Al-Harbi (2017). Performances of LOO and WAIC as IRT model selection methods. *Psychological Test and Assessment Modeling* 59(2), 183–205.
- MacKay, D. (2003). The Swendsen – Wang method The extended Ising model. *Information Theory, Inference, and Learning Algorithms* 1(4), 0–3.
- Malcata, R. M., W. G. Hopkins, and S. N. Pearson (2014). Tracking career performance of successful triathletes. *Medicine and Science in Sports and Exercise* 46(6), 1227–1234.
- Marquez-Neila, P., L. Baumela, and L. Alvarez (2014). A morphological approach to curvature-based evolution of curves and surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(1), 2–17.
- Massaro, F., D. E. Harris, and C. C. Cheung (2011). Large-scale extragalactic jets in the Chandra era. I. Data reduction and analysis. *Astrophysical Journal, Supplement Series* 197(2).



- McKeough, K., A. Siemiginowska, C. C. Cheung, L. Stawarz, V. L. Kashyap, N. Stein, V. Stampoulis, D. A. van Dyk, J. F. C. Wardle, N. P. Lee, D. E. Harris, D. A. Schwartz, D. Donato, L. Maraschi, and F. Tavecchio (2016, dec). Detecting relativistic x-ray jets in high -redshift quasars. *The Astrophysical Journal* 833(1), 123.
- Meng, X.-L. (2014). A trio of inference problems that could win you a Nobel Prize in statistics (if you help fund it). *Past, Present, and Future of Statistical Science*, 537–562.
- Mignotte, M., C. Collet, P. Pérez, and P. Bouthemy (2000). Sonar image segmentation using an unsupervised hierarchical MRF model. *IEEE Transactions on Image Processing* 9(7), 1216–1231.
- Miller, A. C. and L. Bornn (2017). Possession Sketches : Mapping NBA Strategies. *Proc. 11th Annual MIT Sloan Sports Analytics Conference*, 1–12.
- Mosteller, F. (1951). Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika* 16(1).
- Moudud, A., K. Carling, R. Chen, and Y. Liang (2008, 12). How to determine the progression of young skiers? *CHANCE* 21.
- Murphy, K. M. and R. H. Topel (1985). Estimation and Inference in Two-Step.
- Neal, R. M. (2012). MCMC using hamiltonian dynamics.
- Panik, M. J. (2014). *Growth curve modeling : theory and applications*. Hoboken, New Jersey: John Wiley Sons, Inc.
- Perrier, D. and M. Gibaldi (1973). Relationship between plasma or serum drug concentration and amount of drug in the body at steady state upon multiple dosing. *Journal of Pharmacokinetics and Biopharmaceutics* 1(1), 17–22.
- Picquenot, A., F. Acero, J. Bobin, P. Maggi, J. Ballet, and G. W. Pratt (2019). Novel method for component separation of extended sources in X-ray astronomy. *Astronomy and Astrophysics* 627.
- Piironen, J. and A. Vehtari (2017). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing* 27(3), 711–735.
- Pinheiro, J. C. and D. M. Bates (2000). *Mixed-effects models in S and S-PLUS*. New York, NY [u.a.]: Springer.
- Plackett, R. (1975). The Analysis of Permutations. *Applied Statistics* 24(2), 193–202.
- Plummer, M. (2014). Cuts in Bayesian graphical models. *Statistics and Computing* 25(1), 37–43.
- Potts, R. B. (1952). Some generalized order-disorder transformations. *Mathematical Proceedings of the Cambridge Philosophical Society* 48(1), 106–109.
- Rauch, H. E., F. Tung, and C. T. Striebel (1965). Maximum likelihood estimates of linear dynamic systems. *AIAA Journal* 3(8), 1445–1450.

- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20(C), 53–65.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley series in probability and mathematical statistics. Applied probability and statistics. New York ;: Wiley.
- Ruta, N., N. Sawada, K. McKeough, M. Behrisch, and J. Beyer (2019). SAX Navigator: Time Series Exploration through Hierarchical Clustering. *2019 IEEE Visualization Conference, VIS 2019* 1(c), 236–240.
- Sanders, J. S. (2006). Contour binning: A new technique for spatially resolved X-ray spectroscopy applied to Cassiopeia A. *Monthly Notices of the Royal Astronomical Society* 371(2), 829–842.
- Sanders, J. S. and A. C. Fabian (2001). Adaptive binning of X-ray galaxy cluster images. *Monthly Notices of the Royal Astronomical Society* 325(1), 178–186.
- Silver, N. (2019, Oct). How our raptor metric works. <https://fivethirtyeight.com/features/how-our-raptor-metric-works/>.
- Spearman, C. (1904). The proof and measurement of association between two things. *The American journal of psychology* 100(3-4), 441–471.
- Starck, J. L., E. Pantin, and F. Murtagh (2002). Deconvolution in astronomy: a review. *Publications of the Astronomical Society of the Pacific* 114(800), 1051–1069.
- Stein, N. M., D. A. Van Dyk, V. L. Kashyap, and A. Siemiginowska (2015). Detecting unspecified structure in low-count images. *Astrophysical Journal* 813(1), 66.
- Swendsen and Wang (1987). Nonuniversal critical dynamics in monte carlo simulations. *Physical review letters* 58(2), 86.
- Tu, X., X. Meng, and M. Pagano (1993). Survival differences and trends in patients with aids in the united-states. *Journal Of Acquired Immune Deficiency Syndromes And Human Retrovirology* 6(10), 1150–1156.
- Vehtari, A., A. Gelman, and J. Gabry (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and Estimating out-of-sample pointwise predictive accuracy using posterior simulations. (September), 1–29.
- Vikhlinin, A., B. R. McNamara, W. Forman, C. Jones, H. Quintana, and A. Hornstrup (1998). A Catalog of 203 Galaxy Clusters Serendipitously Detected in the ROSAT PSPC Pointed Observations. *The Astrophysical Journal* 502(2), 558–581.
- Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* 58(301), 236–244.
- Wishart, J. (1938). Growth-rate determinations in nutrition studies with the bacon pig, and their analysis. *Biometrika* 30(1/2), 16–28.

Xie, X. and X. L. Meng (2017). Dissecting multiple imputation from a multi-phase inference perspective: What happens when god's, imputer's and analyst's models are uncongenial? *Statistica Sinica* 27(4), 1485–1545.

Zuur, A. F., I. D. Tuck, and N. Bailey (2003). Dynamic factor analysis to estimate common trends in fisheries time series. *Canadian Journal of Fisheries and Aquatic Sciences* 60(5), 542–552.