POLITECNICO DI TORINO

Master Degree
in Mathematical Engineering

Master degree thesis

Analyzing skiing performance: a Bayesian approach to account for athlete abilities and varying race conditions

Supervisor
prof. Mauro Gasparini
Supervisor sign

. . . . . .
. . . . . .

Candidate
Giulia Monchietto
Candidate sign

. . . . . .

Academic Year 2023-2024

t

Acknowledgements

Contents

Summary

This master thesis develops a statistical model to analyze athlete performance in alpine skiing competitions, while accounting for variability introduced by race-specific factors such as weather, track characteristics, snow conditions, and race length. In multi-competitor sports, rank ordering has commonly been employed for performance evaluation, with models like the Luce-Plackett and the Rank Ordered Logit model by Allison and Christakis (1994) facilitating more consistent comparisons of athlete performances across different races. In this thesis, the Luce-Plackett ranking model has been first implemented, highlighting challenges in the interpretation of latent variables. To address this issue, a more interpretable hierarchical Bayesian model has been developed, which quantifies skiers ability, by estimating the percentage deviation of each athlete's time, from the average competition time. The latter model incorporates both individual athlete abilities and race-specific factors, using race times as the primary data source. The abilities derived from both models were compared, resulting in similar rankings. Nevertheless, the Bayesian model offers a key advantage: it provides a directly interpretable measure of ability, and allows for model updates as new race data are introduced, without the need for a full model re-computation.

Knowledge is power.

[F. Bacone]

Chapter 1

Introduction

1.1 Background and motivation

The landscape of sports has been significantly transformed by the advent of innovative technologies exploiting statistical modeling, machine learning, and data management techniques. This transformation has given rise to the field of sports analytics, which involves the application of the latter advanced techniques to analyze the vast and complex datasets generated in sports contexts. At its simplest, sports analytics translates sport-related datasets into meaningful insights that can drive decision-making and performance improvements. While sports such as basketball or motor sports have already seen substantial advancements through analytics, alpine skiing is still in the early stages of integrating these scientific methods into performance analysis.

Alpine skiing is a sport defined by a multifaceted nature. Performance is influenced by a myriad of factors, including athlete abilities, weather conditions, and snow conditions. Different weather conditions such as wind, temperature, and visibility can strongly alter the race outcomes. Slopes and snow conditions can vary significantly too, influencing the techniques and abilities required from the athletes, as well as the speed and average timing of the competitions. The complexity is further heightened by track variations, not only between different specialties, like among giant slalom, slalom, super G and downhill, but also within each specialty. Thus, each race presents a unique set of track conditions, which is also influenced by the fact that every race has its own course setter. The course design variability results in differences in the number of gates present, the angle between one gate and the successive one, and the distance between them.

Given the significant variability inherent in alpine skiing, traditional methods for assessing athlete abilities, commonly employed in sports like athletics or swimming, prove inadequate. These sports, characterized by standardized environments and consistent conditions, allow for direct comparisons of performance across time and events. In athletics, for example, track dimensions, surface conditions, and even environmental factors such as wind resistance are highly regulated, permitting a relatively straightforward evaluation of an athlete's performance through time trials and distance measurements. In contrast, the dynamic and unpredictable nature of alpine skiing, coupled with the challenge of course

variability, renders such methods unsuitable, necessitating the development of adaptive approaches that can account for these complexities and identify new performance metrics.

1.2 Relevance and contribution

Measuring athletes' abilities is of paramount importance in every sport, including alpine skiing, as it provides insights into their performance levels at any given period. Athletes' abilities are not static, they evolve over time due to factors such as training, experience and physical growth. By capturing a snapshot of their abilities at different points throughout their evolution, coaches and sports scientists can track the progression of individual athletes and identify trends in performance, enabling a more nuanced understanding of each skier's development trajectory.

This ability to measure and monitor athletes' progression offers multiple benefits. First, it provides a concrete basis for adjusting training regimens. With objective data, coaches can identify areas of strength and weakness, allowing them to tailor training programs more effectively to address individual needs. For example, if an athlete consistently underperforms in certain disciplines, such as special slalom, the training focus can shift to improve that specific skill set. Conversely, strengths can be reinforced and refined to maximize competitive advantage. In order to do that, ability can be measured considering each type of race results alone, and computing an ability measure that is specialty-specific. On the other hand, an individual performance measure, considering all the competitions, across a group of athletes facilitates the creation of training subgroups based on equivalent skill levels. This can significantly enhance the effectiveness of training sessions. Grouping athletes with similar abilities together allows for more targeted coaching, where exercises and drills can be designed to challenge and improve skills at an appropriate level for the group. This minimizes the risk of top athletes not being challenged enough or lower-level athletes feeling overwhelmed, thus promoting balanced development within the group. Moreover, such an approach can also be strategically adapted to spread top performers across different training subgroups. By placing higher-performing athletes among groups with varying abilities, less experienced skiers can benefit from the presence of more skilled peers, leading to improved performance through peer-driven motivation and learning. In the long term, systematic measurement of athletes' abilities also supports talent identification, helping to recognize promising skiers early in their careers.

From a statistical perspective, this thesis seeks to advance the development of models, particularly Bayesian and ranking models, within the context of alpine skiing competitions. The objective is to build models that possess the following key characteristics:

- **Easy interpretability:** parameters should provide clear and meaningful quantification of abilities, allowing coaches, athletes, and analysts to easily understand the measures obtained.

- **Scalability:** the pool of athletes observed should be of arbitrary size. The number of athletes involved in a single race usually ranges from 10 up to a limit number of 150 athletes.

- **Accuracy:** they should capture the true skill levels of athletes as closely as possible. By accounting for different competition formats, conditions, and disciplines, the models should represent performance fairly and robustly, enabling more precise comparisons across varying race scenarios.

1.3 Problem Statement

In the context of alpine skiing, the study consider a pool of 12 young female athletes, aged 10 to 11, which took part in a competitive racing circuit throughout the winter season, going from December 2023 to April 2024.

The competition format is diverse: in some races, athletes from both age groups compete together, while in others, the 10-year-olds and 11-year-olds race separately. The athletes also participate in three distinct disciplines: giant slalom, special slalom, and flipper. Giant slalom involves gates that are more widely spaced, requiring broader turns. Special slalom demands quick and sharp turns due to its tighter gate placements. Flipper presents a mix of long stretches between gates followed by sequences of closer gates, adding another layer of complexity, and requiring more agility.

For each race, completion times have been recorded, and rankings are based on these times. However, because of the varying conditions—different race formats, age divisions, and external conditions—it is difficult to compare finish times across competitions. This variability complicates the evaluation of athletes' overall performance. Therefore, there is a need to develop a method that accurately quantifies each athlete's ability, identifies those who are particularly promising, and distinguishes between those with average or below-average performance across all these different race types. In this statistical analysis, we are interested in obtaining an overall performance measure, across all alpine skiing specialties.

1.4 Methodology

To address the challenges posed by the problem, the first step of this thesis has been to review the state of the art. This review provided foundational insights into the models previously used in similar contexts. Building on these examples, the initial approach focused on analyzing competition rankings by applying an order statistics model, Plackett-Luce model. This method enabled the estimation of abstract measures of athlete ability, that can be ordered to obtain a ranking of the top-performing athletes. However, this approach has a limitation: the ability measure was tied to latent variables, which lacked a direct and interpretable relationship with the actual race times, making it difficult to infer a meaning in practical terms. Additionally, since current model estimation methods works well with at maximum 15/20 parameters, scalability problems arise when the pool of athletes analyzed increases.

To improve interpretability, race times have been successively considered. However, analyzing directly track times introduces an additional complexity, as race times are influenced not only by the athletes' abilities but also by varying race conditions.

To address these complexities, an hierarchical Bayesian model was developed. This model introduced more interpretable parameters, separating the athlete's ability from

competition-specific factors. By doing so, it provided a clearer representation of both the intrinsic performance level of each athlete and the external conditions influencing their race times. The hierarchical structure allowed the model to handle these layers of complexity, and the possibility to incorporate informative priors offered the flexibility to include external data, such as the mean and standard deviation of each race completion times, further enhancing the model's robustness.

1.5 Structure of the thesis

The structure of the thesis is organized as follows. In Chapter 2, a literature review of the state of the art is conducted, providing context and identifying gaps in the existing research on statistical models in alpine skiing performance analysis. Chapters 3 and 4 focus on the theoretical foundations of the models used in the study. Chapter 3 outlines the framework of order statistics models, such as the Plackett-Luce model and its extension, the Rank Ordered Logit model, while Chapter 4 introduces the Bayesian approach. In Chapter 5, the methodology employed in the study is further detailed, describing how the models were applied and adapted to address the specific challenges posed by alpine skiing competitions. Finally, the results of the model implementations are presented.

Chapter 2

Literature review

The literature on sports analytics has evolved significantly over the past few decades, driven by the need to develop sophisticated models for assessing competitors' strengths and predicting outcomes in various sports. Traditional approaches have predominantly focused on ranking data, which provides a basis for comparing competitors in head-to-head competitions.

2.1 Rank Ordering Models

Early models, such as the *Bradley-Terry* and *Thurstone-Mosteller*, which were introduced in the mid-20th century, were pioneering in this regard. These models represent competitors' strengths as parameters within probability frameworks that predict the outcomes of one-to-one contests based on ranking data alone. However, these models were largely confined to paired comparisons and did not account for scenarios where multiple competitors participate simultaneously, such as in alpine skiing.

For sports where multiple competitors participate in a single event, the outcome has often been best determined by rank ordering in literature. This approach allowed for a more consistent comparison of an athlete's performance across different races, as scores or race times can be subject to greater variability, induced by external factors too. Additionally, collecting more specific data, other than the final rank, can be difficult in some cases. Thus, traditional rank orderings models, such as the *Plackett-Luce* model and its extensions, as the *Exploded Logit model*, by [1], have provided a framework for estimating competitor abilities based on observed rankings. The latter models assume that each competitor's unobservable performance follows a specific distribution (e.g., Gumbel, Normal, or Gamma), and inferences are drawn by modeling the probability of observed rank orderings across multiple competitions.

2.2 Extreme value distributions and Exploded Logit model

Extreme value distributions, traditionally employed to model maxima and minima, also offer significant utility in ranking models. Specifically, when the Gumbel distribution is used to model latent variables, it allows for the derivation of a closed-form expression for ranking probabilities, thereby enhancing the model's computational tractability and interpretability. Given that their approach was based solely on the availability of ranking data, latent variables were employed to represent the unobserved, underlying performance levels of athletes, which are inferred from the competition rankings.

In more detail, for each athlete i in a competition, there is an associated latent performance variable Y_i, which depends on the athlete's ability θ_i. Mathematically, this relationship can be expressed as:

$$Y_i \sim F(y \mid \theta_i).$$

Commonly, $F(y \mid \theta_i)$ is assumed to be a Gumbel distribution, which is appropriate for modeling maximum values, such as in this case, the highest performance in a competition.

The cumulative distribution function of the Gumbel distribution is given by:

$$F(y \mid \theta_i) = \exp\left(-\exp\left(-(y - \theta_i)\right)\right)$$

This distribution has been chosen because it leads to a convenient form for expressing rankings probabilities. The model constructed on this assumption is commonly refered to *Exploded logit model*.

The exploded logit model, also known as the rank-ordered logit model, has been used to model the probability of a particular ranking of athletes in a competition. Suppose to have a set of athletes competing, and their performances ranked, such that athlete 1 is ranked first, athlete 2 is ranked second, and so on. The probability of this ranking can be expressed as a product of conditional probabilities:

$$P(Y_1 > Y_2 > \cdots > Y_n \mid \theta_1, \theta_2, \ldots, \theta_n) = \prod_{i=1}^{n-1} \frac{\exp(\theta_i)}{\sum_{j=i}^{n} \exp(\theta_j)}$$

This probability function $P(Y_1 > Y_2 > \cdots > Y_n \mid \theta_1, \theta_2, \ldots, \theta_n)$ can be employed as likelihood, to estimate the abilities θ_i of each athlete based on observed ranking data from multiple competitions, in a frequentist approach, or used as a distribution function in a bayesian approach. The latter ranking probability distribution has been employed in different sports, in horse races as illustrated by [2] and in alpine skiing too, as in [3].

2.3 Bayesian Approaches in dynamic sports environments

In sports where athlete performance is expected to change over time, static models are insufficient. To address this limitation, [3] applied a Bayesian approach in the context of

women's world alpine skiing competitions. More in detail, their approach models competitors' performances using independent extreme value distributions, with each competitor's ability evolving over time as a Gaussian random walk.

Considering multiple time istants, each athlete's ability θ_i is thus allowed to change. This time-varying ability has been modeled by Glickman and Hennessy, as a latent variable following a the dynamic process of a Gaussian random walk:

$$\theta_{it+1} = \theta_{it} + \delta_{it+1}$$

where $\delta_{it+1} \sim \mathcal{N}(0, \tau^2)$ is a random variable representing changes in the athlete's ability between competitions, while the variance τ^2 controls the extent of ability variation over time. The paper utilizes Markov Chain Monte Carlo (MCMC) simulation to perform inference on the posterior distribution of the ability parameters θ_{it}. Given the observed rankings, the posterior distribution of the abilities is updated iteratively, allowing for the estimation of the competitors' abilities.

2.4 Handling ties and missing data in Bayesian Models

The Bayesian framework also provides a natural way to handle ties and missing data. For ties, the posterior distribution is computed by averaging the probability over all possible permutations of the tied competitors. For missing data, the Bayesian approach treats the missing values as latent variables, integrating over their possible values during the MCMC sampling process.

Mathematically, let Y represent the complete data, which consists of observed data Y_{obs} and missing data Y_{mis}. We aim to infer the parameters θ given the observed data. The joint distribution of the parameters and the complete data can be expressed as:

$$p(\theta, Y_{\text{obs}}, Y_{\text{mis}}) = p(\theta)\, p(Y_{\text{obs}}, Y_{\text{mis}} \mid \theta)$$

where $p(\theta)$ is the prior distribution of the parameters, and $p(Y_{\text{obs}}, Y_{\text{mis}} \mid \theta)$ is the likelihood of the complete data given the parameters.

The posterior distribution of the parameters given the observed data is obtained by marginalizing over the missing data:

$$p(\theta \mid Y_{\text{obs}}) = \int p(\theta, Y_{\text{mis}} \mid Y_{\text{obs}})\, dY_{\text{mis}}$$

Using the relationship $p(\theta, Y_{\text{mis}} \mid Y_{\text{obs}}) = p(\theta \mid Y_{\text{obs}})\, p(Y_{\text{mis}} \mid \theta, Y_{\text{obs}})$, this can be rewritten as:

$$p(\theta \mid Y_{\text{obs}}) = \int p(\theta \mid Y_{\text{obs}})\, p(Y_{\text{mis}} \mid \theta, Y_{\text{obs}})\, dY_{\text{mis}}$$

This integral accounts for the uncertainty about the missing data by averaging over all possible values Y_{mis} could take, given θ and Y_{obs}.

2.5 Direct use of race times in performance modeling

Race times have been used in sports literature, often in cases where large datasets were available and coupled with the use of machine learning algorithms. In [4], the authors applied different machine learning approaches to build a speed-skater performance prediction model. The study exploited competition results data directly, with other 71 features, to obtain knowledge in order to design appropriate training programs for the athletes. Diverse ML approaches have been tested: logistic regression, random forest, K-nearest neighbor, naive Bayes, neural networks and support vector machines, with the latter showing to be more viable to establish performance prediction models. In order to select the optimal features, and decrease computational complexity, lasso regression methods have been used. In further detail, *lasso regression*, short for Least Absolute Shrinkage and Selection Operator, is a type of linear regression that incorporates a regularization technique to enhance prediction accuracy and interpretability. The key feature of lasso regression is that it minimizes the sum of the absolute values of the model parameters (coefficients), effectively shrinking some coefficients to zero. This leads to a sparse model where only the most important predictors are retained. The formula for lasso regression can be expressed as:

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

where y_i is the observed response, x_{ij} represents the predictors, β_0 is the intercept, β_j are the coefficients for the predictors, λ is the regularization parameter that controls the strength of the penalty on the coefficients.

Another interesting direct use of race times can be found in [5]. The paper makes a significant contribution by leveraging race times, to develop a model that accurately reflects the performance of athletes in both Para and able-bodied cross-country skiing categories. The model employs as variables the race time differences for each skier, compared to the winner, in particular the *average percent difference*. All statistical analyses were performed using linear mixed modeling, testing for interactions among the different categories in which athletes can be identified.

Chapter 3

Data

3.1 Data description

To accurately evaluate athlete abilities, two primary datasets were utilized, each offering detailed insights into individual performances and overall competition statistics.

The data were collected during the 2023/2024 winter season, with competition results sourced from regional FISI (Federazione Italiana Sport Invernali) events in Piedmont, as well as FIE (Federazione Italiana Escursionismo) races held in Bardonecchia. The analysis focuses on female athletes, aged 10 to 11.

1. **Athlete performance data:** This dataset contains detailed results of 135 single performances of 12 athletes across 22 different ski competitions. Each entry includes:

   - Position: the rank of the athlete in a specific competition.
   - Athlete: the name of the athlete.
   - Time: the time taken by the athlete to complete the race, measured in seconds (with a precision of 0.01s).
   - Competition: the name of the competition in which the performance was recorded.

   We can observe that data are incomplete. Indeed, not all the 12 athletes have taken part in all 22 competitions. Competitions in which, just 2 or 3 of the athletes in the case study were present, have been included in the dataset too.

2. **Competition Statistics:** This dataset summarizes key statistics from 21 different ski competitions. Each record includes:

   - Competition: the name of the competition.
   - Mean time: the average race time across all participants in that competition (not just the 12 considered in the previous dataset).
   - Standard deviation: the variability in race times for each competition, indicating the spread of times around the mean.

3. **Athletes' Fitness Tests:** this dataset contains data from fitness tests conducted on 12 athletes. It includes a variety of physical measurements and test results that provide insights into the athletes' physical capabilities. Data have been collected on October 2023, thus before the start of the race season. Each entry includes:

    - Surname and name: identification of the athlete.
    - Age: the age of the athlete, given as a year of birth.
    - Weight: the weight of the athlete in kilograms.
    - Height: the height of the athlete in centimeters.
    - HCG: a specific body measurement (height of the center of gravity).
    - Foot size: the size of the athlete's right and left foot, respectively.
    - Flexibility: the flexibility test measures how far an athlete's hands can extend beyond the tips of their feet while sitting with legs straight. The result is recorded in centimeters, indicating the distance the hands surpass the feet.
    - Sprint times (mt30 and mt60): times recorded for 30-meter and 60-meter sprints, respectively, in seconds.
    - Parallel bars: the number of repetitions performed on parallel bars.
    - Harre test: time taken to complete the Harre test, a measure of agility, recorded in seconds. The Harre test is performed on a cross-shaped course. Starting at one end, the athlete performs a forward roll, then turns 90 degrees around a central cone, jumps over and passes under a 55 cm high obstacle, returns to the central cone, repeats the process with a second and a third identical obstacles, and finally sprints to the finish line.
    - Long jump test: the long jump test measures explosive strength and involves performing a standing long jump from a stationary position with feet together. The distance of the jump is measured in centimeters.
    - Backward roll plus jump: the backward roll plus jump test starts with the athlete standing, performing a half backward roll until their toes touch the ground. From this position, the athlete unfolds, stands up, and immediately executes a vertical jump with feet together. The test is scored with a "yes" or "no" based on whether the athlete can perform the sequence of movements continuously.
    - Ratio SLF-Height: measurements related to the athlete's long jump-to-height ratio.

3.2 Summary statistics

The dataset encompasses a variety of statistics across different ski competitions, with particular attention to the mean and standard deviation of race times. The attached graph (Figure 3.1) illustrates the variability in race statistics across the 22 different competitions,

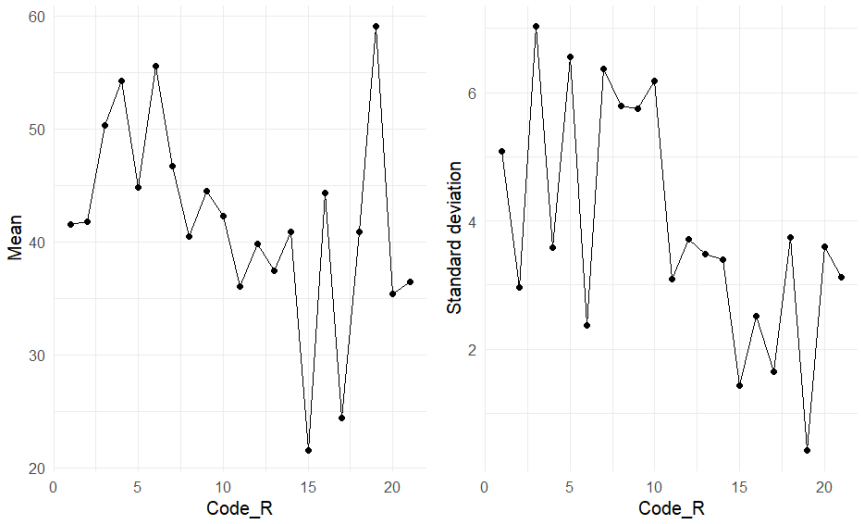

Figure 3.1: Mean and standard deviation of race times measured in seconds across 22 ski competitions.

showing mean and standard deviation for each competition, computed across all athletes race times.

In particular, the left panel of the graph shows the mean race times for each competition, revealing fluctuations between 40 and 60 seconds. These variations suggest that the conditions of each race, which could include factors such as weather, track difficulty, and competitive pressure, significantly impact overall performance. Certain competitions exhibit particularly high mean times, which may be indicative of more challenging conditions or a longer race track.

The right panel of the graph presents the standard deviation of race times within each competition, which ranges from about 2 to 7 seconds. A higher standard deviation indicates greater variability among competitors, suggesting that there was a wider disparity in the abilities of the participants or that conditions underlined more the different athletes' abilities. Usually, more challenging conditions lead to more time disparities.

The substantial variability in both the mean and standard deviation across the competitions underscores the necessity of considering these factors when analyzing athlete performance.

Regarding athlete participation in the competitions analyzed, Figure 3.2 illustrates the number of races each athlete attended. Notably, none of the athletes participated in all 22 competitions. Athletes 12 and 9 had the highest level of engagement, each competing in 14 events, while athlete 3 had the lowest participation, taking part in only 6 races.

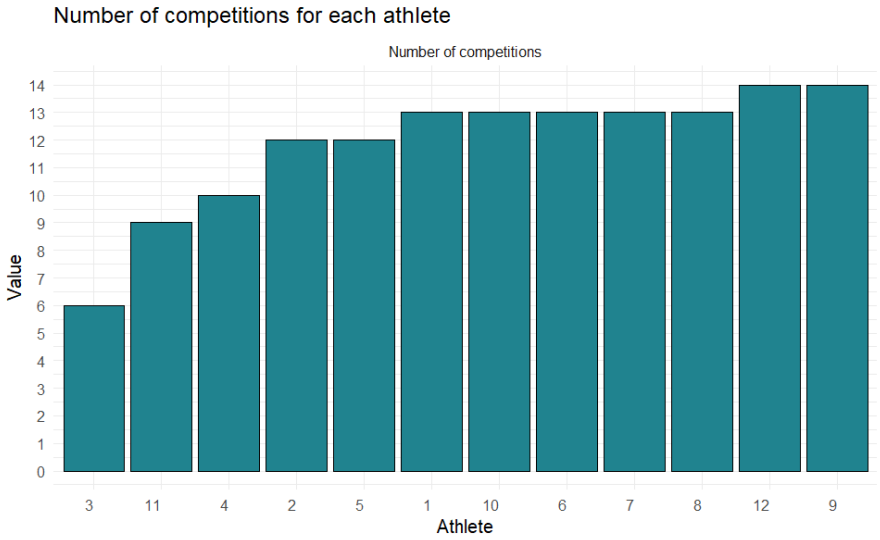

Figure 3.2: Plot illustrating the total number of competitions each athlete has competed in.

Chapter 4

Theoretical Framework of Luce-Plackett models

4.1 Rankings and orderings

Consider a set of objects $\mathcal{O} = \{O_1, O_2, ..., O_N\}$, where each object is uniquely identified by an integer i. Let the set of indices for these N objects be denoted by $\mathcal{I} = \{1, 2, ..., N\}$. A complete ranking assigns a specific order of preference to the items, identifying the best one, the second best, and so on, down to the worst. The two most used representations of a ranking are the rank vector and the order vector:

- The *rank vector* lists the ranks given to the objects, where '1' represents the best object and 'N' the worst. It presumes the objects have been previously indicized. For example, suppose four friends want to do a running competition. In this case $\mathcal{O} =$ {Anna, Barbara, Claudio, Davide} and index 1 is assigned to Anna, 2 to Barbara and so on. They compete, recording the completion times, and the final rank vector is given by (4,2,3,1). Thus, Davide has arrived first at the finish line, Barbara was the second one, Claudio third and the last one to complete the race was Anna.
- The *order vector* lists the object themselves, in order from the best to the worst. In the example above, the order would have been (Davide, Barbara, Claudio, Anna).

Both rankings and orderings are *permutations*. In this context, we will consistently identify objects using integers and define the set of all possible rankings as the set of permutations of the N integers, expressed as:

$$\mathcal{P}_N \equiv \{\textit{Permutations of the ranks 1,2,...,N}\},$$

We will work with a sample of M rank vectors, represented as:

$$r^1, r^2, ..., r^M \in \mathcal{P}_M$$

While the objects being ranked can represent a wide variety of entities—such as words, people, items, or ideas—throughout this thesis, we will focus on ranking athletes.

4.2 Order Statistics Model

Order statistics models were first introduced by Thurstone in his work [6]. These models are based on the premise that each object possesses an inherent quality or perceived value, which is not entirely deterministic but can instead be modeled as a random variable. Assuming the presence of different judges, that have to determine the rankings of N objects, the randomness reflects the variability in judgments or perceptions across different evaluators. If the object quality we want to measure is quantitative, different judges can measure it in different ways, with different levels of precision. On the other hand, if the quality under analysis is qualitative, evaluators may perceive it differently based on subjective criteria. In cases where the "judges" represent different points in time at which the quality is measured, the variability in outcomes can be influenced by the differing conditions at each time. For example, recalling the four friends from the previous section, suppose they decide to repeat the running competition every week, over the course of some months. The results may differ from week to week due to factors such as fatigue or changes in race conditions. In this scenario, the competitions themselves serve as the judges, while the objects being ranked are Anna, Barbara, Claudio, and Davide, each of whom has a performance measure, modeled as a random variable that produces different outcomes at each race instance. In Thurstone's model, the random variable associated with each object represents its latent utility, and the ranking of these objects is determined by the ranking of the r.v. outcomes.

4.2.1 Mathematical formulation of order statistics models

Consider a set of items having indices in $\mathcal{I} \subset \mathbb{N}_+$, and let's introduce a new set of indices $\mathcal{J} = \{1, ..., M\}$, which represents the M judges. Each judge j, assigns a ranking by sorting N random utilities, denoted as $y_{1j}, y_{2j}, ..., y_{Nj}$, where each utility corresponds to one object $i \in \mathcal{I}$. These utilities are stored in the vector $\mathbf{y}_j \in \mathbf{R}^N$, and we assume the utility vectors for different judges to be independent. Within this framework, the probability of observing a specific ranking $R_j \in \mathcal{P}_N$ for judge j is expressed as:

$$P(R_j) = P(y_{[1]j} > y_{[2]j} > \cdots > y_{[N]j}), \tag{4.1}$$

where the subscript $[i]$ refers to the index of the object placed in the i-th position, while $([1]_j, [2]_j, ..., [N]_j)$ denotes the rank vector of R_j. It is important to note that the rank vector probability (4.1) is invariant under any strictly increasing transformation of the latent variables $y_{[i]j}$, as long as the relative ordering between the objects is preserved.

To make the statistical model more tractable, we assume that the latent variables follow a particular family of distributions. Specifically, y_{ij} are assumed to be independent and expressed as:

$$y_{ij} = \alpha_{ij} + \epsilon_{ij}$$

where α_{ij} is the expected utility of object i given by judge j and ϵ_{ij} is the i.i.d. random component. Two main distributions for ϵ_{ij} are commonly employed in literature, leading to distinct ranking models:

- **Thurstone model**: the main idea is to assume ϵ_{ij} to be distributed as a standard normal r.v., whose cumulative distribution is given by:

$$F(\epsilon) = P(\epsilon \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$$

- **Plackett-Luce model**: In the model developed, ϵ_{ij} is assumed to be distributed following an extreme value distribution (Gumbel distribution). The Plackett-Luce model has gained significant popularity in applications due to the fact that assuming a Gumbel distribution for the latent variables yields a closed-form expression for the ranking probability, as we will discuss in the next section.

4.2.2 Plackett-Luce Model

The Plackett-Luce model is a probabilistic choice model widely used in decision theory and discrete choice analysis. The model has been developed, in first place by Luce in [7], as a top-choice model, subsequently it has been extended to accomodate rankings by Plackett in [8]. In this context, the objects of choice, referred to as alternatives, can range from products to sport competitors, or political candidates in elections. As outlined in [9], Plackett-Luce model can also be viewed as an extension of the conditional logit model, developed by McFadden in [10]. Both McFadden and Plackett-Luce models foundation lies in the assumption that the latent variable associated with each alternative follows an Extreme Value distribution.

Gumbel distribution

In top-choice models, the Gumbel distribution is frequently used to represent random variables associated with each alternative. In decision-making contexts, an extreme value distribution, such as the Gumbel distribution, is often a good fit because it is designed to model the minima or maxima of a set of random variables. This is particularly relevant when the goal is to select the best or worst option based on underlying quality measures of different objects. The Gumbel distribution, in particular, has full support over the real line, $S = (-\infty, +\infty)$. Its cumulative distribution function is given by:

$$F(x) = \exp\left(-\exp\left(-\frac{x-u}{\beta}\right)\right)$$

where:

- u is the location parameter, indicating the mode of the distribution.
- $\beta > 0$ is the scale parameter, controlling the spread of the distribution.

A different re-parametrization is often beneficial for simplifying computations and obtaining a certain interpretation of parameters. This alternative formulation is achieved by changing parameters variables, as described in [11], where:

$$\begin{cases} \alpha = \frac{1}{\beta} \\ \pi = \exp(\frac{u}{\beta}) \end{cases} \tag{4.2}$$

Exploiting this re-parametrization, the following Gumbel CDF is obtained:

$$\begin{aligned} F(x) &= \exp\left(-\exp\left(-\frac{x-u}{\beta}\right)\right) = \\ &= \exp\left(-e^{\alpha u} \cdot e^{-\alpha x}\right) = \\ &= \exp\left(-\pi \cdot e^{-\alpha x}\right) \end{aligned}$$

Top-choice probability

In order to develop the Plackett-Luce ranking model, the first essential step is to derive the formula for the top-choice ranking probability. We begin by considering the simplest case, involving the computation of the probability for a pairwise comparison between two alternatives. Let y_1 and y_2 be the two latent variables, relative to two alternatives, the probability of y_2 being greater than y_2, i.e., $P(y_2 > y_1)$, can be computed as follows:

$$P(y_2 > y_1) = \int_{-\infty}^{+\infty} p(y_2)(\int_{-\infty}^{y_2} p(y_1)\, dy_1\,)dy_2$$

where $p(y)$ denotes the y probability density function. Substituting the Gumbel CDF computed in y_2, we obtain:

$$P(y_2 > y_1) = \int_{-\infty}^{+\infty} \exp\left(-\pi_1 e^{-\alpha_1 y_2}\right) \cdot p(y_2)\, dy_2$$

Since both y_1 and y_2 are Gumbel-distributed with the common scale parameter α (i.e., $\alpha_1 = \alpha_2 = \alpha$), the density of y_2 can be written as:

$$p(y_2) = \frac{d}{dy_2} F(y_2) = \alpha \exp\left(\alpha(u_2 - y_2) - \exp\left(\alpha(u_2 - y_2)\right)\right)$$

Thus, substituting $p(y_2)$, and expressing the parameters in function of π_1 and π_2, as defined in the re-parametrization (4.2), the following expression is obtained:

$$P(y_2 > y_1) = \int_{-\infty}^{+\infty} \alpha\pi_2 \exp\left(-\alpha y_2 - (\pi_2 + \pi_1)e^{-\alpha y_2}\right)\, dy_2$$

Next, we transform the expression dividing and multiplying for the factor $(\pi_2 + \pi_1)$, leading to:

$$P(y_2 > y_1) = \int_{-\infty}^{+\infty} \alpha \cdot \frac{\pi_2}{\pi_2 + \pi_1} \cdot (\pi_2 + \pi_1) \exp\left(-\alpha y_2\right) \exp\left(-(\pi_2 + \pi_1)e^{-\alpha y_2}\right)\, dy_2$$

Lastly, since probability measure integrates to one, we obtain the following simplification:

$$P(y_2 > y_1) = \frac{\pi_2}{\pi_2 + \pi_1} \tag{4.3}$$

With similar computations to those used for pairwise comparisons, it is possible to derive the formula for the top-choice probability, when more than two alternatives are present, $P(y_1 > y_2, ..., y_N)$:

$$P(y_1 > y_2, ..., y_N) = \frac{\pi_1}{\pi_1 + \pi_2 + ... + \pi_N} \tag{4.4}$$

Plackett-Luce model formulation

Building upon the reasoning used to compute top-choice probabilities, we now extend this framework to model the ranking of N items. Specifically, the probability of a specific ranking is given by:

$$P(Y_1 > Y_2 > \cdots > Y_N \mid \pi_1, \pi_2, \ldots, \pi_N) = \prod_{i=1}^{N-1} \frac{\pi_i}{\sum_{j=i}^{N} \pi_j}$$

This formula represents the product of conditional probabilities, where each term describes the likelihood that an item is ranked ahead of the remaining items based on their respective utilities.

Breaking this down further:

- **First rank**: The probability that item 1 is ranked first among all N items is given by:

$$P(\text{item 1 is first}) = \frac{\pi_1}{\sum_{j=1}^{N} \pi_j}$$

- **Second rank**: Given that item 1 is ranked first, the probability that item 2 is ranked second among the remaining $N - 1$ items is:

$$P(\text{item 2 is second} \mid \text{item 1 is first}) = \frac{\pi_2}{\sum_{j=2}^{N} \pi_j}$$

- **Subsequent ranks**: This process continues for each subsequent rank, where the probability of an item being ranked i-th is conditioned on the prior $i - 1$ rankings.

Thus, the overall probability of observing the ranking R is the product of these conditional probabilities:

$$P(R) = \prod_{i=1}^{N-1} \frac{\pi_i}{\sum_{j=i}^{N} \pi_j}$$

This expression represents the closed-form solution for the Luce model, where the ranking probability is determined by the relative utility values π_i of the items.

Parameters interpretation

Let us explore the meaning of the Gumbel distribution parameters in more detail. Since we have assumed that α is fixed, where $\alpha = \frac{1}{\beta}$, this corresponds to fixing the scale of the distribution. The variability between the latent variables lies in the location parameter, represented by π or u in the original parametrization. To clarify this, we can rewrite the formula for pairwise comparison (4.3), and by introducing the Euler-Mascheroni constant $\gamma = 0.5772$, we get the following:

$$\begin{aligned} P(y_2 > y_1) &= \frac{\pi_2}{\pi_2 + \pi_1} = \\ &= \frac{e^{\alpha u_2}}{e^{\alpha u_1} + e^{\alpha u_2}} = \\ &= \frac{e^{\alpha u_2} e^{\gamma}}{e^{\gamma}\left(e^{\alpha u_1} + e^{\alpha u_2}\right)} = \\ &= \frac{e^{\alpha(u_2 + \beta\gamma)}}{e^{\alpha(u_1+\beta\gamma)} + e^{\alpha(u_2+\beta\gamma)}} \end{aligned}$$

Here, we can observe that $u_i + \beta\gamma$ represents the expected value of the Gumbel distribution. This demonstrates that the larger the location parameter of one variable compared to the other, the higher the probability that it will be greater than the other parameter. This relationship highlights the direct influence of the location parameter on the comparison outcome in the Gumbel-based model.

Independence of Irrelevant Alternatives Assumption

The *Independence of Irrelevant Alternatives (IIA)* assumption is a core principle in discrete choice theory, particularly in models such as the Plackett-Luce model. It states that the relative probability of preference between two alternatives should remain unchanged when other irrelevant alternatives are added or removed from the choice set. In other words, the ratio of probabilities of sorting two items preferences is independent of the presence of other items in the set.

To illustrate this, suppose to have to decide if apples (A) or bananas (B) are more tasty. The presence or absence of coffee (C) as an irrelevant alternative should not affect the comparison between apples and bananas. Thus, if a person prefers apples over bananas, the introduction of coffee should not change their relative preference.

Mathematically, the IIA assumption can be expressed as:

$$\frac{P(A \mid \{A, B, C\})}{P(B \mid \{A, B, C\})} = \frac{P(A \mid \{A, B\})}{P(B \mid \{A, B\})}.$$

In the Plackett-Luce model, where the probability of selecting an item is proportional to the exponential of its utility, this principle holds. Suppose the utilities for apples, bananas, and coffee are u_A, u_B, and u_C, respectively. The probabilities of selecting apples or bananas from the set $\{A, B, C\}$ are:

$$P(A \mid \{A, B, C\}) = \frac{e^{\alpha u_A}}{e^{\alpha u_A} + e^{\alpha u_B} + e^{\alpha u_C}},$$

$$P(B \mid \{A, B, C\}) = \frac{e^{\alpha u_B}}{e^{\alpha u_A} + e^{\alpha u_B} + e^{\alpha u_C}}.$$

The ratio of these probabilities is:

$$\frac{P(A \mid \{A, B, C\})}{P(B \mid \{A, B, C\})} = \frac{e^{\alpha u_A}}{e^{\alpha u_B}}.$$

Notice that this ratio is independent of u_C, the utility of the irrelevant alternative coffee. This demonstrates the IIA assumption, which ensures that irrelevant alternatives do not influence the relative probabilities of the relevant choices.

4.2.3 Rank ordered logit Model

The Rank Ordered Logit model (ROL) (often referred to as Exploded Logit Model) is an extension of the Plackett-Luce model previously introduced, used for analyzing sets of ranked items. According to [1], this model builds upon the assumption of a random utility framework, as Plackett-Luce model. In this scenario, each item $i \in \mathcal{I}$, being ranked by $j \in \mathcal{J}$, has an associated utility that consists of both a systematic and a random component. The systematic component captures the predictable part of the utility, while the random component represents unobserved factors.

$$y_{ij} = \alpha_{ij} + \epsilon_{ij}$$

The ROL model provides an extension to the Plackett-Luce framework, by allowing the incorporation of both *item-specific* and *judge-specific* attributes. This enhancement makes it possible to analyze how the characteristics of the items being ranked and the individuals doing the ranking, influence the final rankings.

In this context, the systematic component of the utility function, α_{ij}, can be expanded to include explanatory variables. The general utility function can be written as:

$$\alpha_{ij} = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_j'\boldsymbol{\gamma}_i + \mathbf{w}_{ij}'\boldsymbol{\delta},$$

where $\mathbf{x}_j$ is the vector representing the attributes of item i, $\mathbf{z}_j$ denotes the characteristics of respondent j, and $\mathbf{w}_{ij}$ captures any interactions between the attributes of item i and respondent j. The vectors $\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}$ are the corresponding coefficients that quantify the effects of these attributes on the utility. Specifically, item covariates effect $\boldsymbol{\beta}$ is assumed to be equal for all judges. On the other hand, $\boldsymbol{\gamma}_i$, that quantifies judge characteristic effects, has an effect that can differ across items. Suppose, the items are runners, and the judges are running competition, the fact that a competition is a long distance one, e.g. a marathon, can have a positive effect for long distance runners, while having a remarkable negative effect for a 100m speed-runner.

Incorporating these attributes into the model allows for an analysis of how both the items and the individuals characteristics influence the rankings outcome. For instance, item-specific attributes could include characteristics such as price, quality, or other measurable features, while respondent-specific attributes might include demographic factors

like age or income. The interaction terms provide an additional layer of detail by allowing the effects of item attributes to vary according to judge-specific characteristics.

Consider again the example where the "items" being ranked are runners, and the rankings are conducted by different races, which act as judges. Each runner i is characterized by height (h), weight (w), and number of weekly training sessions (t). Meanwhile, each race j is characterized by its length (len) and total elevation gain (elev). The utility function for runner i in race j, can be modeled as:

$$\mu_{ij} = \beta_1 \cdot \mathrm{h}_i + \beta_2 \cdot \mathrm{w}_i + \beta_3 \cdot \mathrm{t}_i + \gamma_{1_i} \cdot \mathrm{len}_j + \gamma_{2_i} \cdot \mathrm{elev}_j + \delta_1 \cdot (\mathrm{len}_j \times \mathrm{w}_i).$$

In this example, the parameters $\beta_1, \beta_2, \beta_3$ describe how the runner-specific attributes (height, weight, number of trainings) influence the probability of a higher ranking. Similarly, γ_{1_i} and γ_{2_i} reflect the impact of race-specific characteristics (length and elevation) on the utility of the runner. The interaction term δ_1 allows for the possibility that the effect of a runner's weight depends on the length of the race, capturing more nuanced relationships between the attributes.

Chapter 5

Theoretical Framework of Bayesian Models

5.1 Probability foundations and Bayes' rule

5.1.1 Probability density

Given a sample space Ω (the set of possible outcomes) and a measurable space E, we define a random variable X as a measurable function $X : \Omega \to E$.
When E is a subset of $\mathbb{R}^n$ (as in this case, since from now on we will consider $E = \mathbb{R}_+$), the probability distribution of an absolutely continuous variable X can be described by a probability density function (PDF). The *density function* f_X associated with the random variable X is defined as a function $f_X : E \to [0, \infty)$ that satisfies the following condition:

$$P(X \in A) = \int_A f_X(x)\, dx, \tag{5.1}$$

for any measurable subset $A \subseteq E$. Here, $P(X \in A)$ represents the probability that the random variable X takes a value within the set A. The function $f_X(x)$ is non-negative and integrates to 1 over the entire space E:

$$\int_E f_X(x)\, dx = 1 \tag{5.2}$$

In our sport applications, as we will later see, the r.v. X will represent either ski racing times or various characteristics of athletes and races.

5.1.2 Conditional densities

Conditional density is a fundamental concept since it allows the computation of an event probability, given that certain outcomes are known to have occurred. Using the previously defined concepts of random variables and density functions, we introduce *Bayes' rule* for computing conditional densities.

Given two random variables X and Y defined on the same sample space Ω, the conditional probability density function of X, given that $Y = y$, is denoted by $f_{X|Y}(x|y)$ and is defined as:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \tag{5.3}$$

where:

- $f_{X,Y}(x,y)$ is the joint probability density function of X and Y.
- $f_Y(y)$ is the marginal probability density function of Y, that can also be defined as:

$$f_Y(y) = \int_E f_{X,Y}(x,y)\,dx. \tag{5.4}$$

5.1.3 Bayes' rule derivation

Similarly to 5.3, the conditional probability density function of Y given $X = x$ is:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \tag{5.5}$$

Thus, the joint probability density function $f_{X,Y}(x,y)$ can also be expressed as:

$$f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y) = f_{Y|X}(y|x)f_X(x) \tag{5.6}$$

Using this relationship, we can express the conditional probability density function $f_{X|Y}(x|y)$ in terms of the conditional probability density function $f_{Y|X}(y|x)$ and the marginal densities $f_X(x)$ and $f_Y(y)$, as follows:

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)} \tag{5.7}$$

This equation is known as **Bayes Rule**.
Often, the un-normalized density is used, which is:

$$f_{X|Y}(x|y) \propto f_{Y|X}(y|x)f_X(x) \tag{5.8}$$

Interpretation of Bayes' Rule

Suppose some data are available and the task is to gain insights about their distribution. Firstly, one has to assume the family of distributions to which data belongs. This choice of distribution family is typically characterized by certain parameters, that are not directly observable and must be inferred. Bayes' Rule, from which Bayesian inference originates, can be interpreted as a way to update the prior belief about the distribution of a parameter X (unobservable quantity), based on new evidence provided by data Y (observable data). In the context of Bayesian inference:

- $f_X(x)$ represents the **prior distribution**, which reflects the initial belief about the parameter X before observing the data Y.

- $f_{Y|X}(y|x)$ is the **likelihood**, representing the probability of observing the data Y given that the parameter X takes a specific value x.

- $f_{X|Y}(x|y)$ is the **posterior distribution**, which reflects the updated belief about the parameter X after observing the data Y. It represents the probability that the underlying distribution has parameter X, given that data Y have been generated.

- $f_Y(y)$ is the **marginal likelihood**, which serves as a normalizing constant ensuring that the posterior distribution integrates to one. This interpretation stems from expression (5.4).

5.2 Bayesian modeling

Bayesian modeling poses its foundations on Bayes' rule. As [12] explains, Bayesian modeling has two fundamental ideas:

1. Bayesian inference is a re-allocation of credibility across all the existent different possibilities, adjusting belief according to observed data.

2. The possibilities over which we allocate credibility, are parameter values within meaningful mathematical models.

5.2.1 Bayesian update

To illustrate the first concept, consider the example of modeling a runner's performance in 10km races. Initially, if we have no prior knowledge about the runner's ability, we might allocate credibility evenly across a wide range of possible race times, from the fastest possible times, such as values in the neighborhood of the world record, to much slower times. This represents a non-informative prior distribution (such as the uniform distribution), where all possible times are considered equally plausible and we do not have an idea yet.

Once the runner completes his first race, and we observe that his time is 50 minutes, our belief, or credibility, about his true average race time needs to be updated. Specifically, we increase the credibility for those times around 50 minutes, while drastically reducing the credibility for extreme values, such as the possibility that the runner could break the world record of 26:24 minutes. This process of updating our beliefs based on new data leads to what is known as the *posterior distribution*, that represents Bayesian modeling's outcome.

The posterior distribution represents the re-allocated distribution of credibility, where the sum of credibility over all possible times still equals one, ensuring it remains a valid probability distribution. Posterior distribution reflects our updated understanding of the runner's performance, shifting our focus from the broad, uninformed prior to a more concentrated belief around the observed time, thus providing a more accurate model of the athlete's abilities.

Mathematically, the posterior is expressed using Bayes' rule:

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)}, \tag{5.9}$$

Now, suppose we observe additional data, denoted Y'. We can update our beliefs again, from $f_{X|Y}(x|y)$, representing the new prior, to the new posterior distribution $f_{X|Y',Y}(x|y', y)$.

Example: Bayesian update

We now proceed to graphically illustrate how Bayesian updating works by sequentially incorporating two observations into the model. Assume a runner race time is a random variable T belonging to the family of gaussian distribution. Thus, f_T is characterized by two parameters, the mean μ and the variance σ^2. Consider for simplicity σ^2 to be known in advance.

$$T \sim \mathcal{N}(\mu, \sigma^2). \tag{5.10}$$

The goal is to get to know the true value of μ. In order to do that, we start with a uniform prior distribution, which reflects our initial state of lack of knowledge regarding the runner's true race time distribution. The prior assumes that any time between 20 and 80 minutes is equally likely.

$$\mu \sim \mathcal{U}(20, 80).$$

Upon observing the first race time of 50 minutes ($T_1 = 50$), we update our prior beliefs to obtain the posterior distribution $f(\mu|T_1)$. The posterior is derived by combining the uniform prior with the likelihood function, as previously explained. In this example the likelihood $f(T_1|\mu)$, is chosen to be (5.10).

$$f(\mu|T_1) = \frac{f(T_1|\mu) \cdot f(\mu)}{\int f(T_1|\mu) \cdot f(\mu) \cdot d\mu}$$

As we can see from the first plot of 5.1, this update shifts our belief, concentrating the probability density around the observed value, i.e. $T_1 = 50$, while reducing the credibility of extreme times, such as those near the world record.

Next, we incorporate a second observation $T_2 = 47$ minutes. The posterior from the first update $f(\mu|T_1)$, now serves as the prior for this new update. Again, we combine this prior with the new likelihood function, which is centered at the newly observed time of 47 minutes.

This second update $f(\mu|T_1, T_2)$, further refines our beliefs, shifting the posterior distribution slightly towards the value of 47 minutes and narrowing the range of plausible race times, as we can clearly see from the second plot in 5.1.

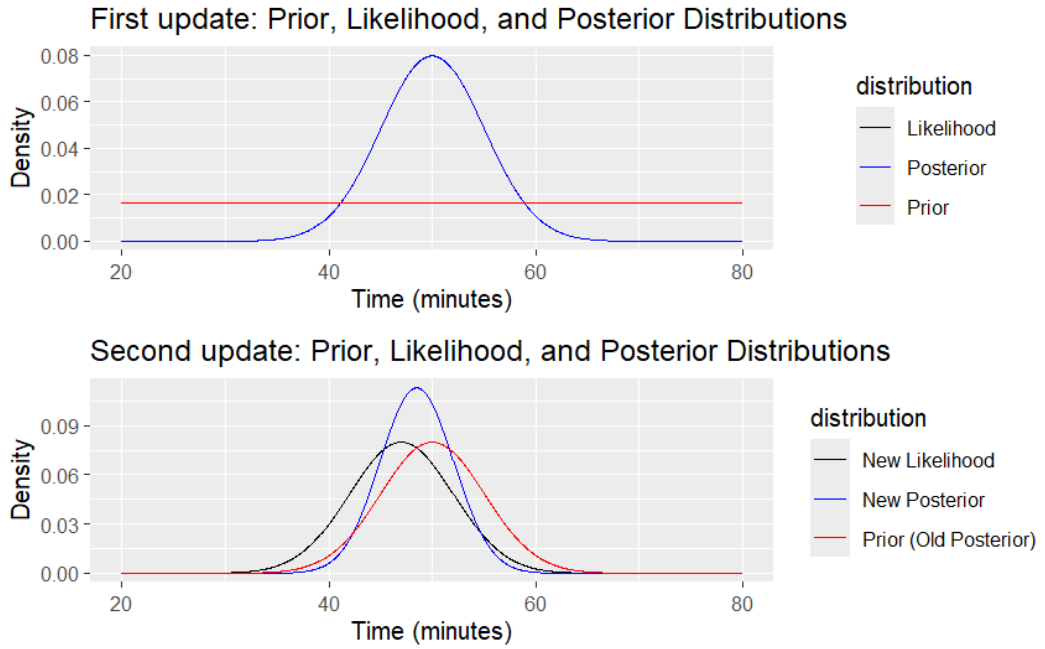

Figure 5.1: Plot of prior, likelihood, and posterior distributions after the two updates.

5.2.2 Data order invariance

In Bayesian inference, the property of *data order invariance* ensures that the final posterior distribution is unaffected by the sequence in which data points are incorporated, provided that data are *conditionally independent*.

Mathematically, the property can be stated as follows. Assume the likelihood of Y, $f_{Y|X}$ is conditionally independent of other data's likelihood $f_{Y'|X}$, meaning that:

$$f_{Y,Y'|X}(y, y'|x) = f_{Y|X}(y|x) \cdot f_{Y'|X}(y'|x) \tag{5.11}$$

Under this assumption, the order of updating has no effect on the final posterior distribution. We can demonstrate this proposition, as follows:

$$f_{X|Y',Y}(x|y', y) = \frac{f_{Y',Y|X}(y', y|x) f_X(x)}{\int_x f_{Y',Y|X}(y', y|x^*) f_X(x^*) dx^*}$$

Using the independence assumption:

$$f_{X|Y',Y}(x|y', y) = \frac{f_{Y'|X}(y'|x) f_{Y|X}(y|x) f_X(x)}{\int_x f_{Y'|X}(y'|x^*) f_{Y|X}(y|x^*) f_X(x^*) dx^*}$$

Since multiplication is commutative, this can be rewritten as:

$$f_{X|Y,Y'}(x|y, y') = \frac{f_{Y|X}(y|x) f_{Y'|X}(y'|x) f_X(x)}{\int_x f_{Y|X}(y|x^*) f_{Y'|X}(y'|x^*) f_X(x^*) dx^*}$$

In conclusion, we find that:

$$f_{X|Y',Y}(x|y',y) = f_{X|Y,Y'}(x|y,y'). \tag{5.12}$$

Data order invariance implies that, in the context of the earlier example with a runner's competition times, if we assume that the times are independent of each other, the order in which the results of different competitions are entered into the model does not influence the final posterior distribution of the runner's performance.

However, it is important to note that this assumption of independence may not always hold true in practice. For instance, an athlete's performance can improve or deteriorate over time due to factors such as training, fatigue, or changes in strategy, which introduce a dependency between successive competition times. Nevertheless, when competitions in a sufficiently short time-lapse are considered, these effects are often minimal, and the events can be reasonably approximated as independent.

5.2.3 Model interpretability

For what concerns the second key idea 2, a crucial step in bayesian modeling is defining the set of possibilities over which credibility is allocated. This is practically done by choosing the appropriate family of distributions that best represent the data and the context of the problem. For example, if we are modeling race times, a normal distribution might be an appropriate choice if we believe that the times are symmetrically distributed around a mean.

On the other hand, if the observed data do not seem to be well described by the chosen distribution, we can consider expanding the set of possibilities by selecting a different family of distributions or modifying the model. This check is typically performed during the posterior predictive check, which will be explained in greater detail later. Thus, two crucial concepts stand out:

1. The mathematical distributions chosen for the model must be comprehensible and equipped with meaningful parameters that can be interpreted within the context of the problem.

2. The chosen distribution should closely reflect the possible data.

5.2.4 Bayesian data analysis

The process of Bayesian data analysis can be summarized in the following steps:

1. Define an adequate likelihood for all observable quantities in the problem, depending on parameters (unobservable quantities), ensuring the model is meaningful and aligned with the underlying scientific problem.

2. Choose a prior that reflects initial beliefs or assumptions about the parameters.

3. Update the model by conditioning on observed data, yielding a posterior distribution.

4. Conduct a posterior predictive check to compare model predictions with observed data.

5.3 Markov Chain Monte Carlo approaches (MCMC)

5.3.1 Conjugate priors

In Bayesian statistics, determining the posterior distribution directly from Bayes' Rule often requires computing the marginal likelihood. For continuous parameters, this process typically involves solving an integral, as

$$f_Y(y) = \int_E f_{X,Y}(x, y)\, dx, \tag{5.13}$$

that can be analytically intractable. Historically, this challenge has been addressed by utilizing a specific class of prior distributions known as *conjugate priors*. Conjugate priors are selected so that the posterior distribution remains within the same family as the prior distribution. This choice allows for the update process to involve only the modification of the parameters within the already known distribution family, using specific formulas. As a result, the need for computationally intensive calculations is avoided.

Example: conjugate priors

To make an example, consider the case of a binomial likelihood function, which is often used to model the number of successes Y in a fixed number of trials, N, with p representing the probability of success.

$$Y|p \sim \text{Binomial}(N, p),$$

If we assume a prior distribution for p that is a Beta distribution:

$$p \sim \text{Beta}(\alpha, \beta),$$

the posterior distribution after observing the data will also be a Beta distribution. Specifically, if we observe $Y = k$ successes out of N trials, the posterior distribution becomes:

$$p|Y = k \sim \text{Beta}(\alpha + k, \beta + n - k).$$

This result simplifies the process of updating our beliefs based on new data and avoids the need for complex numerical integration.

Limitations of conjugate priors

While conjugate priors facilitate analytical solutions, they are not always available or suitable, particularly for more complex models. In such cases, alternative methods are necessary to approximate the posterior distribution. One common approach is the *variational approximation*, where the complex posterior distribution is approximated by a simpler, more tractable distribution. This approximation is chosen by minimizing the difference, often measured by Kullback-Leibler divergence, between the true posterior and the approximating distribution. However, we will not delve into the details of this method since it is out of the purpose of this master thesis.

5.3.2 Numerical integration via grid-based methods

When dealing with Bayesian inference for models with a small parameter space, such as one or two parameters, numerical methods can be employed to approximate the intractable integrals involved in determining the posterior distribution. A straightforward and effective technique in such cases is the grid-based method, which involves discretizing the parameter space into a grid of equally spaced points and computing the integral by summing over these points.

To illustrate this approach, let's revisit the example of estimating the 10 km running time distribution of a runner. Suppose we model the runner's competition times Y using a normal distribution, with known variance σ^2,

$$Y|\mu \sim \mathcal{N}(\mu, \sigma^2)$$

where the parameter μ represents the mean running time. Our goal is to estimate the posterior distribution for μ, given observed competition times $Y_1, ..., Y_N$.

In the grid-based method, we first define a range of plausible values for μ, say from $\mu_{\min}$ to $\mu_{\max}$, and divide this range into a finite number of points, choosing appropriate step size, creating a grid. Each point on this grid represents a possible value of μ.

Next, we calculate the likelihood of the observed data for each value of μ on the grid. Assuming a normal likelihood function, the likelihood for a specific μ_i, given observed data $y = (y_1, y_2, \ldots, y_N)$ can be expressed as:

$$f_{Y|\mu}(y|\mu_i) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_j - \mu_i)^2}{2\sigma^2}\right),$$

assuming $\{Y_j\}_{j=1}^{N}$ pairwise independent.

The posterior for each μ_i is then proportional to the product of the prior $f_\mu(\mu_i)$ and the likelihood $f_{Y|\mu}(y|\mu_i)$:

$$f_{\mu|Y}(\mu_i|y) \propto f_{Y|\mu}(y|\mu_i) \cdot f_\mu(\mu_i).$$

To approximate the integral for the marginal likelihood, which is necessary to normalize the posterior, we sum the posterior values over all grid points:

$$\int_\mu f_{Y|\mu}(y|\mu) f_\mu(\mu) d\mu \approx \sum_i f_{Y|\mu}(y|\mu_i) \cdot f_\mu(\mu_i) \cdot \Delta\mu,$$

where $\Delta\mu$ represents the spacing between grid points.

By summing these products across the grid, we obtain an approximation for the marginal likelihood, as shown in figure 5.2.

This grid-based approach, though simple, is effective when the parameter space is low-dimensional.

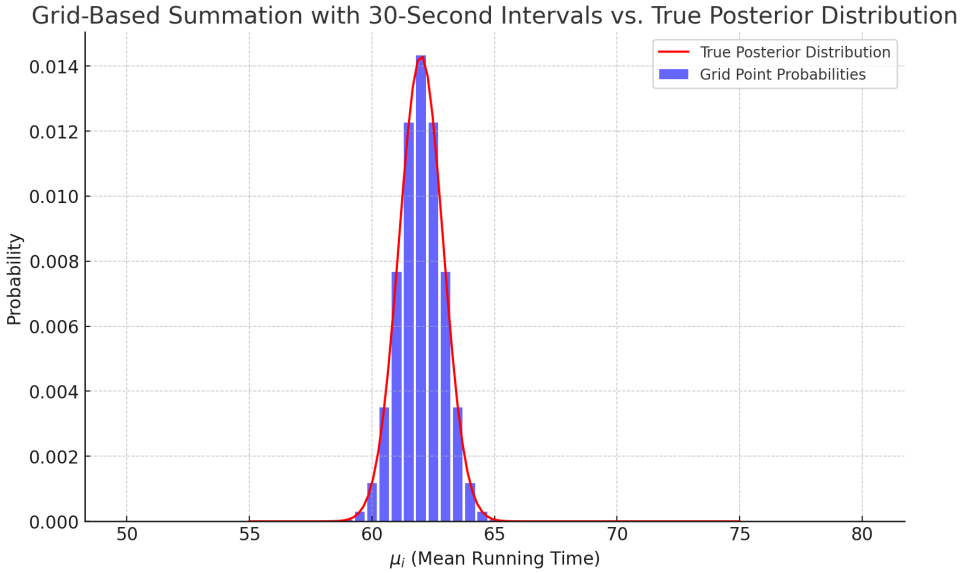

Figure 5.2: Grid-based summation with 30-Second Intervals vs. true posterior distribution (in red).

5.3.3 MCMC methods

An introductory metaphor

Imagine a company that wants to understand customer preferences for a new product line, they are planning to launch. To gain insights, the company decides to conduct a survey among a sample of potential customers. Instead of asking every potential customer, which would be time-consuming and expensive, they collect responses from a large but manageable sample, say 10,000 customers. By analyzing the responses from this sample, the company can estimate the distribution of preferences across the entire customer base. For example, they might find that 60% of the sampled customers prefer the blue colored product, while the 40% prefer the red one. This sample proportion provides an estimate of the true proportion of all customers who prefer that color. The larger the sample size, usually, the more confident the company can be that the sample accurately reflects the true customer preferences. This is because a larger sample usually reduce the likelihood of sampling errors and provide a more accurate representation of the population. This approach is similar to the Markov Chain Monte Carlo (MCMC) method in that it relies on using a large, representative sample to approximate the underlying distribution of interest.

General framework

When neither conjugate priors nor variational or grid approximations provide adequate solutions, an alternative approach involves randomly sampling a large number of representative combinations of parameter values from the posterior distribution. This method is commonly referred to as *Markov Chain Monte Carlo (MCMC)*. By definition, a Markov

chain Monte Carlo method for the simulation of a distribution f, is any method producing an ergodic Markov chain, whose stationary distribution is f.

Let's take a Bayesian inference problem as an example. Suppose the posterior distribution is complex and cannot be computed analytically, thus, MCMC methods allow to simulate a large number of samples, that collectively approximate this distribution. These samples can then be used to estimate quantities of interest, such as means or credible intervals, without the need for an explicit analytical solution.

Through techniques such as the Metropolis-Hastings algorithm or Gibbs sampling, MCMC facilitates Bayesian inference across a wide range of applications, providing flexibility where conjugate priors and variational methods fall short. We will delve further into MCMC methods in the upcoming section.

Before detailing the procedures of Metropolis-Hastings algorithm and Gibbs sampling, some basic concepts about Markov chains and their definition will be presented.

5.3.4 Markov chains

Given a state space S, a *stochastic process* is a collection of random variables $\{X_t\}_{t\in T}$, where T is an index set, where each t typically represents a time instant, and each X_t takes values in the state space S. Mathematically, a stochastic process is defined as a function

$$X : T \times \Omega \to S,$$

where Ω is a sample space within a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For each $t \in T$, X_t is a random variable defined on Ω, meaning $X_t : \Omega \to S$. The distribution of these random variables evolves over time, according to certain probabilistic rules.

A *Markov chain* is a specific type of stochastic process that satisfies the *Markov property*. This property implies that the probability of transitioning to any particular state is dependent solely on the current state and the time elapsed. For the sake of simplicity, we discuss first Markov chains with a discrete state space S. Formally, the Markov property of order T can then be expressed as follows:

$$\mathbb{P}(X_{t+1} = s_{t+1} \mid X_t = s_t, X_{t-1} = s_{t-1}, \ldots, X_0 = s_0) = \mathbb{P}(X_{t+1} = s_{t+1} \mid X_t = s_t, ..., X_{t-T+1})$$

for all $t \in T$ and for all states $s_0, s_1, \ldots, s_t, s_{t+1} \in S$. This equation states that the conditional probability of moving to the next state s_{t+1} depends only on the current state s_t and the $T - t + 1$ states before, and not on any prior states $s_{t-T}, s_{t-2}, \ldots, s_0$.

For a Markov chain, the transition probability at time t, thus the probability to go in a certain state j at the successive step $t + 1$, given that the chain current state is j, is defined by

$$P_{ij}^t = \mathbb{P}(X_{t+1} = j \mid X_t = i)$$

for all $i, j \in S$. The collection of all transition probabilities P_{ij} forms the *transition matrix* P.

A Markov chain is called *time homogeneous* if

$$\mathbb{P}(X_{t+1} = j \mid X_t = i) = p_{ij}, \ \forall t$$

. In words, the transition probabilities are not changing as a function of time.

Ergodicity

A Markov chain $\{X_t\}$ is called *ergodic* if the limit

$$\pi(j) = \lim_{t\to\infty} \mathbb{P}_i\{X_t = j\}$$

exists for every state j and does not depend on the initial state i. The vector π is called the *stationary distribution*.

In other words, the probability $\pi(j)$ of being in state j after a long time is independent of the initial state i.

Consider the following important implication. If a Markov chain is ergodic, then

$$\pi(j) \equiv \lim_{t\to\infty}(P^t)_{ij} = \lim_{t\to\infty}(P^{t+1})_{ij} = \lim_{t\to\infty}(P^t P)_{ij} = \lim_{t\to\infty}\sum_{s\in S}(P^t)_{is}P_{sj} = \sum_{s\in S}\pi(s)P_{sj}.$$

Where the second equality above holds because, if the limit exists, the distinction between t and $t+1$ does not matter. So we can write this as

$$\pi^\top = \pi^\top P,$$

where π is a column vector. Hence the name *stationary probability* for π. It is a distribution that does not change over time.

The ergodicity property of Markov chains will be crucial when discussing Markov Chain Monte Carlo (MCMC) methods. Indeed, ergodicity ensures that the Markov chain has a unique stationary distribution π, towards which the chain converges, that is independent of the initial state. This property will be essential for the convergence of MCMC methods.

5.3.5 Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm, as outlined in [13] and [14], is a fundamental method within the family of Markov Chain Monte Carlo (MCMC) algorithms. It is designed to generate samples from a target probability distribution $\pi(x)$, when direct sampling is challenging, often due to the complexity or intractability of the distribution. Indeed, in Bayesian modeling, when often normalizing constants involve intractable integrals, direct methods are not applicable because they would require the full specification of the distribution.

Overview of the Algorithm

The core idea of the Metropolis-Hastings algorithm is to construct a Markov chain, having as stationary distribution, the desired target distribution $\pi(x)$. This can be achieved through an iterative process that generates a sequence of samples, $\{X^{(t)}\}_{t=1}^{T}$, such that the distribution of $X^{(t)}$ converges to $\pi(x)$ as t increases up to infinity. In other words, for large values of t, the probability of visiting state x in the chain, should be approximately equal to the probability of $X = x$, where X is a r.v. distributed according to the target distribution.

Detailed Algorithm Steps

Given the current state $X^{(t)} = x^{(t)}$ at iteration t, the algorithm proceeds as follows:

1. **Proposal Step**: the algorithm generates a new candidate state Y by sampling from the distribution:
$$q(y|x^{(t)}),$$
called *proposal distribution*. Y is the sampled value, representing a specific proposed state that the algorithm will evaluate, and potentially move to, in the next iteration. The proposal distribution $q(y|x)$ is a critical component of the Metropolis-Hastings algorithm, as it determines how new candidate states (proposals) are generated from the current state of the Markov chain.

   The choice of the proposal distribution is based on several key considerations: simplicity and computational efficiency, adaptation to the target distribution and problem-specific requirements. For example, if the target distribution is known to have certain symmetries or constraints (e.g., positivity or boundedness), proposals should be chosen accordingly. Thus, for a symmetrical target, a normal distribution could be a good fit. Common choices of proposals include:

   - *Gaussian distribution*: for continuous target distributions, a common choice is a normal distribution centered at the current state x, $q(y|x) \sim \mathcal{N}(x, \sigma^2)$, where σ^2 is a variance parameter that controls the spread of the proposed samples.
   - *Uniform distribution*: another simple choice is a uniform distribution within a specified interval around the current state x, such as $q(y|x) \sim \text{Uniform}(x - \alpha, x + \alpha)$, where α is a parameter controlling the width of the interval.
   - *Laplace distribution*: in some cases, a Laplace (double exponential) distribution is used to allow for heavier-tailed proposals.

2. **Acceptance Probability**: Compute the acceptance probability, $\rho(x^{(t)}, y)$, which is defined as:
$$\rho(x^{(t)}, y) = \min\left\{1, \frac{\pi(y)q(x^{(t)}|y)}{\pi(x^{(t)})q(y|x^{(t)})}\right\}. \tag{5.14}$$
   The ratio expressed in (5.14), represents the relative probability of the proposed state y, compared to the one of the current state $x^{(t)}$. Thus, the proposed jump is accepted with certainty if the posterior distribution is higher at the proposed position than at

the current position. While the proposed jump is probabilistically accepted, if the posterior is lower at the proposed position than at the current position. Should the proposed move be rejected, the algorithm retains the current position for the next iteration, effectively counting it again in the sampling process.

Note that target distribution is known up to a normalizing constant. Since the latter constant cancels out in expression (5.14), the evaluation of the resulting ratio is feasible.

3. **Transition Step**:

   - Draw u_t from a uniform distribution on $[0, 1]$.
   - If $u_t \leq \rho(x^{(t)}, y)$, then accept the candidate, setting $X^{(t+1)} = Y$. Otherwise, reject the candidate and retain the current state, setting $X^{(t+1)} = x^{(t)}$.

Key Properties

- **Ergodicity**: The Markov chain produced by this algorithm is ergodic, meaning that as $t \to \infty$, the distribution of $X^{(t)}$ approaches the target distribution $\pi(x)$ regardless of the initial state $X^{(1)}$. As a consequence, $X^{(1)}$ can be arbitrarily chosen and the sequence in which the sampled points appeared, and the trajectory are irrelevant.
- **Stationarity**: If the chain reaches stationarity, the distribution of $X^{(t)}$ is exactly $\pi(x)$.

Metropolis-Hastings limitation

One significant limitation of the Metropolis algorithm is its sensitivity to the choice of proposal distribution. For the algorithm to perform efficiently, the proposal distribution must be well-aligned with the posterior distribution. If the proposal distribution is either too narrow or too broad, the algorithm may result in a high rejection rate for proposed jumps, causing the Markov chain to stagnate in a localized region of the parameter space. This inefficiency reduces the effective sample size, meaning that the number of independent samples generated is substantially lower than the total number of iterations.

5.3.6 Gibbs sampling method

Gibbs sampling is a specialized form of Markov chain Monte Carlo (MCMC) that, like the Metropolis algorithm, involves a random walk through the parameter space. The key difference lies in how Gibbs sampling progresses from one point to the next in this space. The algorithm begins at an arbitrary point, and the next step depends solely on the current position, without considering any previous steps, thus satisfying Markov property.

In Gibbs sampling, rather than proposing a new position for all parameters at once (as in Metropolis-Hastings), the algorithm updates one parameter at a time. Typically, parameters are cycled through in a fixed order, and then repeating, rather than selecting parameters at random. This cycling ensures that all parameters are regularly updated.

Fixed update order is particularly important in models with many parameters, where randomly selecting parameters could result in some parameters being infrequently updated, if iterations are not enough.

In particular, at the beginning of the proposal step, a particular parameter θ_i is selected. Then, Gibbs sampling draws a new value for that parameter directly from its conditional probability distribution, given the fixed current values of all other parameters and the observed data, denoted as

$$p(\theta_i \mid \{\theta_j\}_{j \neq i}, S).$$

Successively, this newly sampled value for θ_i is combined with the current values of the other parameters to form the new position in the parameter space. The algorithm then repeats this process for each parameter in turn.

An important feature of Gibbs sampling is that the proposal distribution for each parameter is exactly the conditional posterior distribution of that parameter, given the current values of the other parameters. This alignment between the proposal distribution and the posterior distribution ensures that the proposed move is always accepted. Unlike the Metropolis-Hastings algorithm, where proposed moves can be rejected if they do not improve the posterior probability, Gibbs sampling inherently avoids this issue, leading to a more efficient exploration of the parameter space. This characteristic makes Gibbs sampling particularly advantageous when the complete joint posterior, $p(\{\theta_i\}_{i=1}^n \mid S)$, cannot be analytically determined and cannot be directly sampled, but all the conditional distributions, $p(\theta_i \mid \{\theta_j\}_{j \neq i}, S)$, can be determined and directly sampled.

Moroever, as we can derive from the two methods explanations, effective size of the Gibbs sample is larger than the effective size of the Metropolis sample for the same length of chain.

Illustrative example: Gibbs sampling in practice.

Let's consider a bivariate random variable $\mathbf{x} = (x_1, x_2)$, which is distributed according to a normal distribution with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and a 2×2 covariance matrix $\boldsymbol{\Sigma}$ given by:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{pmatrix}$$

Sampling directly from this bivariate normal distribution can be challenging. However, by using Gibbs sampling, we can simplify the problem by utilizing the conditional distributions $f(x_1 \mid x_2)$ and $f(x_2 \mid x_1)$, which are both univariate normal distributions.

The conditional distribution $f(x_1 \mid x_2)$ is given by:

$$x_1 \mid x_2 \sim \text{Normal}\left(\mu_1 + \frac{\sigma_{1,2}}{\sigma_2^2}(x_2 - \mu_2),\ \sigma_1^2 - \frac{\sigma_{1,2}^2}{\sigma_2^2}\right)$$

Similarly, the conditional distribution $f(x_2 \mid x_1)$ is given by:

$$x_2 \mid x_1 \sim \text{Normal}\left(\mu_2 + \frac{\sigma_{1,2}}{\sigma_1^2}(x_1 - \mu_1),\ \sigma_2^2 - \frac{\sigma_{1,2}^2}{\sigma_1^2}\right)$$

By iteratively sampling from these conditional distributions, we can efficiently generate samples from the joint distribution of $\mathbf{x}$.

As we can see from 5.3, just either x_1 or x_2 are sampled at each iteration.

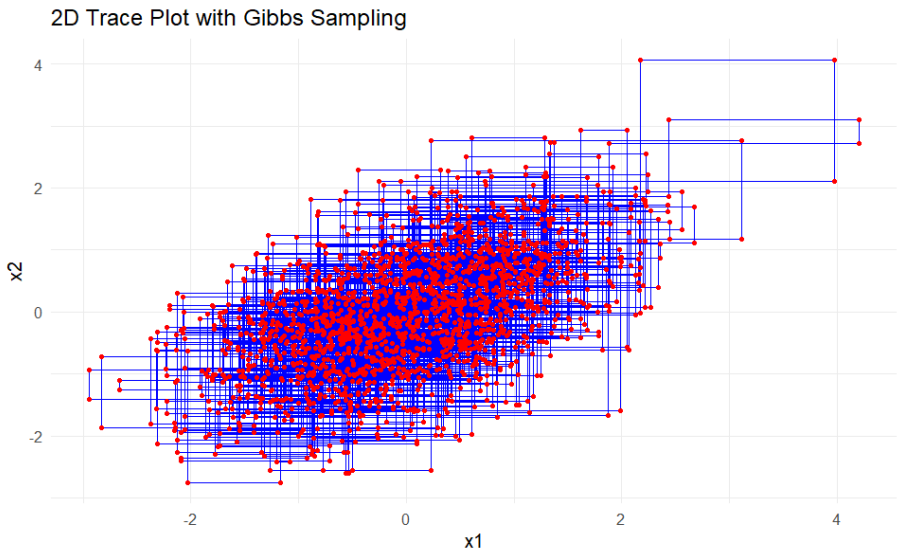

Figure 5.3: Bidimensional trace plot of the samples generated using Gibbs sampling algorithm.

MCMC methods assumptions

To effectively implement Markov Chain Monte Carlo methods, certain fundamental assumptions must be met to ensure the accuracy and efficiency of the sampling process. First and foremost, the prior distribution must be specified by a function that is also computationally tractable. This implies that the prior should be expressible in a form that allows for straightforward evaluation at any given point within the parameter space. In conjunction with the specification of the prior, the likelihood function, denoted as $f(Y|\theta)$, where Y represents the observed data and θ denotes the parameter of interest, must also be computable for any given value of Y and θ. Moreover, a minimal regularity condition on both the target and the proposal distributions should be imposed. Defined as S, the support of target distribution f, S should be connected and the support of f should be contained in the proposal q support. Indeed, assuming that there exists $A \in S$, such that

$$\int_A f(x)ds > 0, \forall x \in S$$

but

$$\int_A q(y|x)dy = 0, \forall x \in S$$

Then, in this case, f is not the limiting distribution of the Markov chain, since for $x_0 \notin A$, the chain $\{X_t\}_{t\in T}$ never visits A. On the other hand, if S is not connected, more than one

connected components are present. Thus, starting from a given connected component, states belonging to another connected component cannot be visited by the chain.

5.4 Hierarchical models

Bayesian hierarchical models extend the concept of Bayesian inference by introducing a structure in which parameters are organized into different levels, reflecting their dependencies. In these models, the parameters at one level are treated as random variables whose distributions are influenced by parameters at higher levels. This hierarchical structure allows for the sharing of information across different groups or levels, leading to more informed and robust inferences.

5.4.1 Example: Hierarchical Bayesian model for runners race times

We now extend the previous example by introducing a Bayesian hierarchical model, where runners are grouped into two categories: long-distance runners and short-distance runners. Each category has its own average race time, denoted as μ_L for long-distance runners and μ_S for short-distance runners. The race time for each individual runner, denoted as T_i, depends on the category mean μ_C where $C \in \{L, S\}$.

Hierarchical Model Structure

- *Category-Level Distribution*: Each category mean μ_C is assumed to be drawn from a normal distribution centered at a global mean μ_G, with known variance τ^2 for simplicity, reflecting the variability between different categories.

$$\mu_C \sim \mathcal{N}(\mu_G, \tau^2) \quad \text{for } C \in \{L, S\}$$

- *Runner-Level Distribution*: Each runner's race time T_i is drawn from a normal distribution centered at the mean of the category to which the runner belongs, with a known variance σ^2, reflecting the variability in times within the category.

$$T_i \sim \mathcal{N}(\mu_C, \sigma^2) \quad \text{for } C \in \{L, S\}$$

- *Hyperprior Distributions*: non-informative prior distributions on the global mean μ_G are placed, reflecting our initial uncertainty about these parameters.

$$\mu_G \sim \mathcal{U}(20, 80)$$

Formally, we define *hyper-parameters* as parameters that define the prior distributions of other parameters within a hierarchical Bayesian model. In the latter case, μ_G is an hyper-parameter. *Hyperpriors* are the prior distributions assigned to these hyperparameters. In the example, the uniform distribution followed by μ_G is an hyperprior.

To visualize the hierarchical structure, consider the following graph representing the dependency structure:

$$\begin{array}{ccc} & \mu_G & \\ \swarrow & & \searrow \\ \mu_L & & \mu_S \\ \downarrow & & \downarrow \\ T_{L,1}, T_{L,2}, \ldots & & T_{S,1}, T_{S,2}, \ldots \end{array}$$

In this graph:

- μ_G (level 1) influences both μ_L and μ_S (level 2),
- μ_L and μ_S each influences the observed times $T_{L,i}$ and $T_{S,j}$ (level 3) of the athletes in their respective categories.

This hierarchical Bayesian model not only refines our estimates of individual runner times but also accounts for the variability both *within* and *across* runner categories.

Bayesian Updating Process

We begin by observing the race time $T_{L,1} = 50$ minutes for a runner in the long-distance category. The likelihood of this observation, given the category mean μ_L, is:

$$T_{L,1}|\mu_L \sim \mathcal{N}(\mu_L, \sigma^2)$$

The prior for μ_L is the normal distribution centered at μ_G, which itself is still uncertain:

$$\mu_L \sim \mathcal{N}(\mu_G, \tau^2)$$

After observing $T_{L,1}$, we update our belief about the parameters (μ_G, μ_L) using Bayes' rule, combining the hyperprior $f(\mu_G)$ with the likelihood $f(T_{L,1}|\mu_L)$, as follows:

$$f(\mu_L, \mu_G|T_{L,1}) \propto f(T_{L,1}|\mu_L) \cdot f(\mu_L|\mu_G) \cdot f(\mu_G)$$

The process of observing multiple runners from both categories allows us to refine the global mean μ_G and the category-level means μ_L and μ_S. In particular, the posterior for μ_G is updated based on the observed data from both categories, as it indirectly influences μ_L and μ_S through their priors.

5.5 Comparison with the Frequentist Approach

The Bayesian approach differs from the frequentist approach, particularly in how uncertainty is treated and how new data is incorporated. In the frequentist paradigm, we typically obtain a point estimate of the parameter of interest, in the example above, the runner's average 10km time. This estimate is derived by maximizing the likelihood of the observed data, which is a procedure known as Maximum Likelihood Estimation (MLE).

Suppose we estimate the runner's race time after observing half of the race season. In the frequentist approach, if we wish to update our estimate at the end of the season, we would have to recompute the entire model by combining the new data with the old ones, and finding the parameter value that maximizes the likelihood across all the collected data. This approach does not naturally accommodate incremental updates; instead, it requires a complete reanalysis every time new data is added.

In contrast, the Bayesian approach offers a more flexible framework. After each race, we can update our model by incorporating the new time observation into the posterior distribution obtained from the previous data. This updating process allows us to refine our model continuously without needing to recompute everything from the beginning. With each new piece of data, the posterior distribution from the previous analysis becomes the prior distribution for the next analysis, making the Bayesian approach adaptive to new information.

Moreover, the Bayesian approach does not just yield a point estimate like the frequentist method. Instead, it provides a full probability distribution over the possible values of the parameter. This posterior distribution encapsulates all the information about the parameter, including its mean, variance, and credible intervals, thereby offering a more nuanced understanding of the uncertainty associated with the parameter.

Chapter 6

Methodology

In this study, we apply two distinct methodologies—a Bayesian hierarchical model and a ranking model—to analyze athlete performance in ski competitions. The Bayesian hierarchical model incorporates both individual athlete abilities and competition-specific factors, using track times as the primary data source. On the other hand, the ranking model focuses on rankings rather than track times, not taking into account of race-specific variabilities. The final objective of both models is to assess athletes ability, obtaining an ability measure that does not depend on factors of variability.

6.1 Plackett-Luce model approach

The first approach that we have undertaken, has been to focus solely on ranking data, employing the methods mentioned in literature ([3], [2]). As discussed in the introduction section, performance measures are subjected to great variability, while rankings, focusing only on comparison outcomes, enable to reduce the noise present in track times due to factors such as race conditions.

6.1.1 Model specification

In this light of this, we have developed a Plackett-Luce model based on alpine skiing competitions ranking results. In this context, the skiers represent the items to be ordered, while the skiing competitions serve as the judges that rank the athletes. Each skier is assigned an index $i \in \mathcal{I}$, where $|\mathcal{I}| = N$, and each competition is indexed by $j \in \mathcal{J}$, with $|\mathcal{J}| = M$. Skiing performances, for which precise track times are considered unobservable, are modeled using a Gumbel probability distribution. Every athlete is characterized by their own location parameter, while the scale parameter is assumed to be constant across all athletes.

6.1.2 Model parameter interpretation

The resulting model, as described in Section 4.2.2, provides estimates of skiers' abilities, which are captured by the "worth" parameter for each item. Specifically, the worth values

π_i are estimated, with:

$$\pi_i = e^{\frac{u_i}{\beta}}.$$

Here, β represents the scale parameter of the Gumbel distribution, while u_i denotes the location parameter. The location parameter can be interpreted as proportional indicator of each skier's ability: the higher the parameter value, the greater the athlete's perceived ability. In order to identify the model, the average ability among the skiers considered, has been chosen as a reference. It is important to clarify that, since we do not employ precise track time data, the Gumbel distribution obtained for each athlete does not correspond to the actual distribution of race times. Indeed, a higher ability parameter (i.e., a higher location parameter u_i) implies that the athlete's performance distribution is skewed toward higher values. However, since faster athletes have shorter race times, this distribution reflects an abstract measure of performance, that is inversely proportional to recorded completion times.

6.1.3 Implementation details

The model has been implemented in the R environment using the `PlackettLuce` package, which provides functions for preparing ranking data and fitting the Plackett-Luce model. This package has been particularly selected because it accommodates ties of any order and can handle partial rankings effectively.

Specifically, partial rankings refer to rankings in which some of the athletes considered in the study are absent from the ranking list. This can occur when an athlete either fails to finish the race or does not participate at all. In other words, partial rankings do not provide a complete ordering of all competitors, yet they still offer valuable information for the model.

To prepare the data, we used the `ranking` function to create a "ranking object" from a dataframe structured with the following three columns:

- ID: representing the indices of the rankings, with each index corresponding to a specific competition instance.
- Item: identifying the athlete by their respective index in the dataset.
- Rank: indicating the rank assigned to the athlete (referred to by the Item index) in the competition corresponding to the given ID.

Once the ranking object was created, it was passed to the `PlackettLuce` function, which fits the model and estimates the athletes' abilities.

6.1.4 Ties handle

In this section, we present how ties are handled in the `PlackettLuce` function, as described in [15]. A single ranking can be expressed as:

$$R = \{C_1, C_2, \ldots, C_J\}$$

where the items in set C_1 are ranked higher than those in C_2, and so forth. If any set C_j contains multiple items, these items are tied in the ranking. For a set of items S, we can define the function:

$$f(S) = \epsilon_{|S|} \left(\prod_{i \in S} \pi_i \right)^{\frac{1}{|S|}}$$

where $|S|$ represents the cardinality of the set, ϵ_n is a parameter related to the frequency of ties of order n (with $\epsilon_1 = 1$), and π_i is the parameter representing the worth of item i. A model is said to accept ties up to order D, if the maximum number of ties, thus the maximum number of elements in each set C_j, is given by D. Under the Plackett-Luce model with ties up to order D, the probability of observing ranking R is given by:

$$P(R) = \prod_{j=1}^{J} \frac{f(C_j)}{\prod_{k=1}^{\min(D_j, D)} \sum_{S \in \binom{A_j}{k}} f(S)}$$

where D_j denotes the cardinality of A_j, the set of remaining alternatives from which C_j is selected, and $\binom{A_j}{k}$ represents all possible combinations of k items from A_j. If no ties are present, meaning $D = 1$, the parameters simplify to the form used in the theoretical framework from previous chapters. For example, if C_1 contains only one item indexed by k (i.e., $C_1 = \{k\}$), the function returns the following:

$$f(C_1) = \epsilon_1 \left(\prod_{i \in C_1} \pi_i \right)^{\frac{1}{1}} = \pi_k.$$

For what concerns the $\epsilon_{|S|}$ parameter, [15] suggests to relate it to the probability that $|S|$ items, of equal worth, tie for first place, given that the first place is not a tie of higher order.

6.1.5 Partial rankings

In our study, the dataset contains *subset rankings*, where only a subset of athletes is fully ranked in each observation. This occurs when some athletes do not complete the race or are missing from certain competitions, resulting in partial rankings. To handle this, the function in Plackett-Luce package, adjusts the likelihood function to account for the fact that the set of alternatives in the denominator only includes the remaining items from the subset of ranked athletes.

6.2 Bayesian hierarchical model approach

Building on the approach outlined in [5], we quantify skier's ability by estimating the *percentage deviation of the athlete's time from the average competition time*. Importantly, we did not select the first-place finisher as the reference time as in [5]. This decision was motivated by the fact that each of the analyzed competitions involves a slightly different group of athletes. Using the first-place finisher as the reference could introduce bias, when top athletes are absent from the race. In contrast, the average competition time offers a more stable reference, being less affected by the absence of individual competitors. In this problem, the performance of athletes in ski competitions involves multiple levels of variability, which necessitates the use of an hierarchical model. In particular, a two-levels description has been identified:

1. **Athletes' ability**: at the most granular level, we are interested in modeling the performance of individual skiers. Each one of them has a latent, unobservable "ability" that determines their performance in a competition. This ability can be thought of as a personal characteristic, influenced by factors such as amount and quality of the training, experience, motivation, and overall physical fitness. However, the observed race time for an athlete is not a direct reflection of their ability, as it is also affected by external conditions of the specific competition in which they are taking part and the environment.

2. **Competition specific effects**: at a higher level, the competition itself introduces variability. Each race has different characteristics, such as course difficulty, weather conditions, altitude, and snow quality, all of which affect the performance of the athletes. In this study, it is assumed that all skiers in a given race compete under the same conditions, though this may not be entirely accurate in practice.

The key challenge of the problem lies in the fact that athlete performance and competition conditions are intertwined. An hierarchical model enables us to capture the interactions. In particular, it is assumed that, the race time observed, results from a combination of the athlete's underlying ability and the external conditions of the competition. Moreover, the hierarchical model structure reflects the nested nature of the data: athletes compete in multiple events, and each event conditions influence performances of multiple athletes.

6.2.1 Model specification

Let I denote the set of indices representing the athletes, where each athlete is identified by an index $i \in I$, and J representing the set of indices corresponding to the competitions, where each competition is identified by an index $j \in J$.
The core of the model is the assumption that the observed race time for athlete i in competition j, denoted by y_{ij}, follows a log-normal distribution, thus:

$$\log(y_{ij}) \sim \mathcal{N}(\theta_i + \beta_j, \sigma^2)$$

Where:

- θ_i represents the model parameter accounting for the ability of the athlete,
- β_j accounts for the average time for competition j,
- σ^2 represents the variance that captures the residual variability in race times not accounted for by athlete abilities or competition characteristics. Alternatively, it can be interpreted as an inverse measure of the athlete's consistency and performance stability across competitions.

The logarithmic transformation is appropriate for two key reasons: the *positivity* of race times and their typically *right-skewed* nature.

- Firstly, $y_{ij} > 0$ for each i and j. Indeed, an athlete cannot complete a race in negative time. The log-normal distribution is well-suited to model such strictly positive data because it ensures that all the values predicted by the model are positive, without employing domain constraints. This is achieved by modeling the logarithm of the race times, which can take any real value, but when exponentiated back to the original scale, the outcomes are positive.
- Secondly, race times in competitive settings frequently display right-skewed distributions, characterized by a concentration of faster times around lower values, which correspond to the most competitive athletes. In contrast, slower performances form a long tail extending towards higher values, spread over a wider range of times. This pattern is illustrated in Figure 6.1, which shows histograms from six different competitions.

6.2.2 Model parameters interpretation

Ability parameter: an inverse measure

The ability of an athlete is interpreted as the percentage variation in his race time, with respect to the average competition time. This means that an athlete's ability is quantified by how much faster or slower they are compared to the average participant. Specifically, it is an inverse measure, an athlete who completes the race in less time than the average has a lower ability coefficient, indicating higher performance, as they are outperforming the average competitor.

Mathematically, the ability of an athlete a_i is defined as:

$$a_i = \exp(\theta_i) - 1$$

In this formulation, θ_i is the model parameter associated with the athlete i.
The interpretation of this equation is as follows:

- when θ_i is **negative**, $\exp(\theta_i)$ is less than 1, leading to a_i being negative. This implies that the athlete's race time is expected to be shorter than the average competition time, indicating a superior performance. A more negative θ_i corresponds to a faster athlete, as it leads to a lower a_i value.

- When θ_i is **zero**, $\exp(\theta_i) = 1$, and hence $a_i = 0$. The latter implies that athlete's performance is equal to the average competition time. In this case, the athlete's ability is neutral, neither outperforming nor underperforming relative to the average.
- When θ_i is **positive**, $\exp(\theta_i)$ is greater than 1, resulting in a positive a_i. This indicates that the athlete usually takes longer than the average to complete the track, implying lower performance compared to the other competitors. The larger θ_i, the greater the deviation above the average time, and thus the lower the athlete's ability.

In the hierarchical model, θ_i is assumed to follow a normal distribution, reflecting our prior belief that athletes' abilities are normally distributed around a central tendency with some variability. Specifically, we define the prior for θ_i as follows:

$$\theta_i \sim \mathcal{N}(\mu_{\theta_i}, \sigma^2_{\theta_i})$$

Here, μ_{θ_i} is the prior mean of the log-ability for athlete i, which we assume follows a non-informative distribution:

$$\mu_{\theta_i} \sim \mathcal{N}(0, 1)$$

This hyperprior reflects our belief that, in the absence of observed data, each athlete is supposed to perform as an average competitor. A value $\theta_i = 0$ corresponds to $a_i = 0$, indicating that the athlete's race time is expected to be equal to the average competition time. A moderate variance of 1 has been chosen to ensure the model adheres more closely to the interpretation of a_i as a percentage.

The variance of the log-ability $\sigma^2_{\theta_i}$ is given a Gamma prior::

$$\sigma^2_{\theta_i} \sim \text{Gamma}(0.001, 0.001)$$

This weak prior allows to give more importance to the data, in order to determine the consistency of athletes across competitions. The Gamma distribution with hyperparameters 0.001 and 0.001 represents a minimally informative prior, providing little constraint on the possible values of $\sigma^2_{\theta_i}$.

Competition-specific parameters

For each race $j \in \mathcal{J}$, we define μ_j as the logarithm of the competition-specific parameter β_j, where μ_j directly represents the average track time in race j, taking into account all participating athletes. β_j, also follows a normal prior distribution, and its prior is expressed as:

$$\beta_j \sim \mathcal{N}(\mu_\beta, \sigma^2_\beta)$$

Where:

- μ_β is exactly computed from data as the observed log-transformation of the race times mean in competition j, providing an empirical estimate of the central tendency of that specific competition,

- σ_β^2 is computed as the observed log-variance of the race times in competition j, capturing the inherent variability in race times within that competition.

This informative prior ensures that each competition has its own baseline average race time, which can vary depending on factors such as weather, track conditions, or competition difficulty. By incorporating the observed mean and variance, the model can account from the beginning of the characteristics of each event. Since the analysis is limited to only 12 skiers out of 50 to 80 participants per competition, it is not practical to reliably estimate the mean and standard deviation from this smaller subset of data. Therefore, the use of an informative prior is advantageous.

Normal distribution mean interpretation

In the light of the more interpretable parameters a_i and μ_j, the mean $\theta_i + \beta_j$ of the normal distribution can be expressed as:

$$\theta_i + \beta_j = log(a_i + 1) + log(\mu_j) = (a_i + 1) \cdot \mu_j$$

The latter expression allows us to retrieve the interpretation of athlete's ability a_i as the percentage deviation from the average competition time b_j, consistent with the interpretation provided in the general framework section.

Residual Variability

The model also accounts for residual variability in race times, denoted by σ_y^2. This residual variability captures factors that are not explained by athlete ability or competition-specific effects, such as random noise, day-to-day performance fluctuations, or unmeasured environmental factors. We place a Gamma prior on σ_y^2:

$$\sigma_y^2 \sim \text{Gamma}(0.001, 0.001)$$

As with the prior for $\sigma_{\theta_i}^2$, this Gamma prior is weakly informative, allowing the data to largely determine the level of residual variability. This choice reflects the lack of knowledge about external and unpredictable factors.

6.2.3 Hierarchical model graphical representation

Here, the graphical representation of the hierarchical model is presented, offering a visualization of the model's structure. The diagram illustrates how the various components are interconnected, with arrows indicating dependencies. Specifically, for example, y_{ij} distribution depends on θ_i, β_j and σ_y^2, while θ_i distribution depends on both μ_{θ_i} and $\sigma_{\theta_i}^2$.

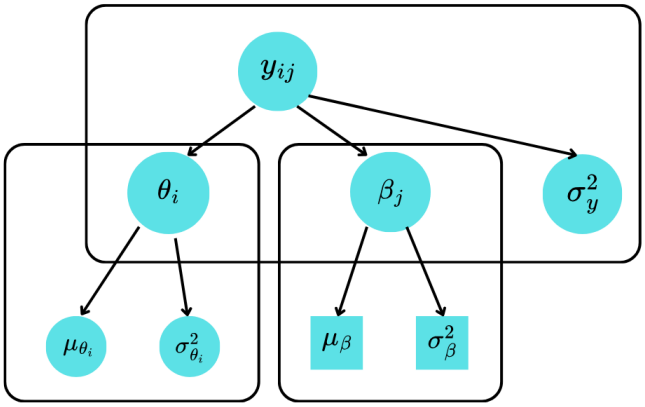

Figure 6.2: Graphical representation of the hierarchical model used to model athlete abilities across various alpine skiing competitions.

6.2.4 Athletes performance measure uncertainty quantification

Given the inherent uncertainty surrounding race conditions, which are not known in advance, the hierarchical Bayesian model does not provide direct estimates of the mean track times for upcoming competitions before they occur. Instead, the model accomplishes a descriptive purpose, by extracting the insights of athletes' abilities from the data. A possibly useful application of this model is the construction of an overall ability ranking of the athletes. In such a scenario, we can use credible intervals, or just the expected value of the posterior distribution of the ability parameters, to establish the skiers ordering.

Intervals of credibility

In more detail, we define a *credible interval* (or Highest Density Interval, HDI) for a parameter θ, as the range of values for θ, that have a posterior probability above a certain threshold, such that the total probability of these values is $1 - \alpha$, where α is typically chosen to be 5%. This interval gives us a measure of the uncertainty around the parameter estimate and the most credible range of values for θ.

To compute the credible interval for an athlete's ability θ_i, we start with the posterior distribution for $\hat{\theta}_i$, which is given by the following equation:

$$P(\hat{\theta}_i \mid D) = \int_{\mu_{\theta_i}} \int_{\sigma^2_{\theta_i}} P(\hat{\theta}_i \mid \mu_{\theta_i}, \sigma^2_{\theta_i}) P(\mu_{\theta_i} \mid D) P(\sigma^2_{\theta_i} \mid D) d\mu_{\theta_i} d\sigma^2_{\theta_i} \tag{6.1}$$

Here, $\hat{\theta}_i$ represents the predicted ability for skier i, D represents the observed data,

μ_{θ_i} and $\sigma^2_{\theta_i}$ are the mean and variance of the ability parameter's prior distribution. To compute the credible interval from this posterior distribution, we follow these steps:

1. **Generate Posterior Samples**: Obtain posterior samples for $\hat{\theta}_i$ from the posterior distribution using Markov Chain Monte Carlo sampling methods.
2. **Sort the Samples**: Sort the posterior samples of $\hat{\theta}_i$ in ascending order.
3. **Select the Central Interval**: For a 95% credible interval, discard the lowest and highest 2.5% of the samples. The remaining 95% of the posterior samples represent the credible interval. More formally, if $S_{\hat{\theta}_i}$ denotes the sorted array of posterior samples, the 95% credible interval is defined as:

$$[S_{\hat{\theta}_i}[0.025], S_{\hat{\theta}_i}[0.975]]$$

This interval provides a range where the skier's ability parameter θ_i is likely to fall with 95% probability, based on the observed data. This credible interval can then be used to rank athletes by comparing their expected abilities and the uncertainty associated with each estimate. Suppose the posterior distribution of a parameter θ, to be a Poisson distribution with parameter 0.02. In this scenario, the credible interval is about $(9, 26)$, as shown in the following plot:

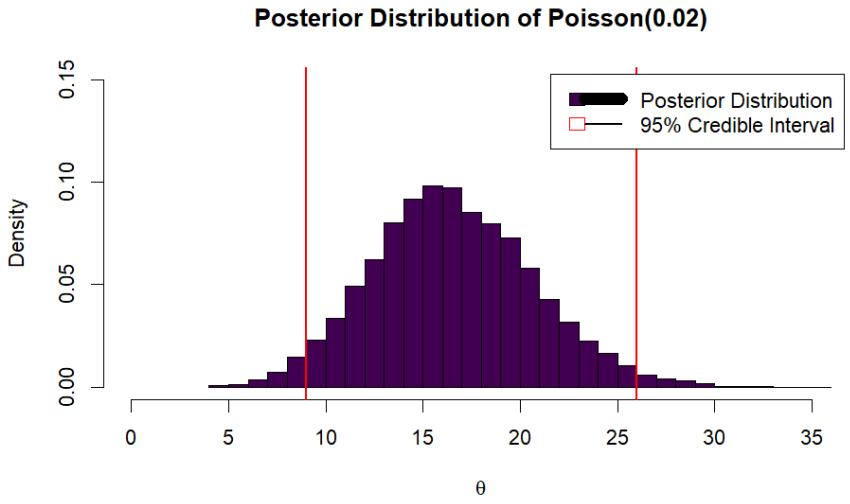

Figure 6.3: Credible interval representation of a Poisson(0.02) posterior distribution.

6.2.5 Model validation

In Bayesian analysis, the posterior distribution provides relative probabilities for the various parameter values under consideration. This means that the posterior tells us which parameters are less unlikely than others but does not necessarily indicate whether the

most probable parameters actually provide a good fit to the data. For example, consider a scenario where we are deciding if a runner is a professional athlete or a beginner. In both cases, we assume that their 5km times follow a normal distribution, with a mean of 15 minutes for a professional and 40 minutes for a beginner, and a fixed standard deviation of 3 minutes. Suppose we record the following times: $T_1 = 25$, $T_2 = 20$, $T_3 = 23$. According to the posterior distribution, the model indicating the runner is a professional athlete might be favored. However, the data clearly do not align well with the assumption that times are normally distributed around a mean of 15 minutes, as the observed times are significantly higher. To assess whether the most credible parameter values provide qualitatively a good fit to the data, we can use a *posterior predictive check*, as proposed in [12]. This involves simulating data using parameter values with high posterior credibility and comparing these simulated data to the actual observed data. If the simulated data resemble the real observations, then we can conclude that the parameter values are a good description of the data. On the other hand, if the simulated data pattern deviates significantly from the real data one, even the most probable parameter values may not be adequate for explaining the observed outcomes.

6.2.6 Implementation details

The hierarchical Bayesian model in 6.3 has been implemented using JAGS (Just Another Gibbs Sampler) within the R statistical computing environment. JAGS is a software package designed for the Bayesian analysis of complex hierarchical models. It provides a flexible and powerful framework for specifying and fitting Bayesian models using Markov Chain Monte Carlo methods, particularly Gibbs sampling.

R packages

In this implementation, the `r2jags` package has been employed to interface between R and JAGS. The `r2jags` package is a widely used tool that facilitates the execution of JAGS models from within the R environment. It provides functions to compile JAGS models, to handle the conversion of R data frames and matrices into formats that JAGS can process and to run the MCMC algorithms. The package also supports the retrieval and processing of the posterior samples, making it easier to conduct Bayesian inference and perform model diagnostics. In our implementation, `r2jags` has been utilized to fit the hierarchical Bayesian model by specifying the model structure and priors in JAGS and running the MCMC simulations to obtain posterior distributions. While for what concerns the posterior density plots and the credible intervals, they have been obtained using the R package `ggmcmc`.

.

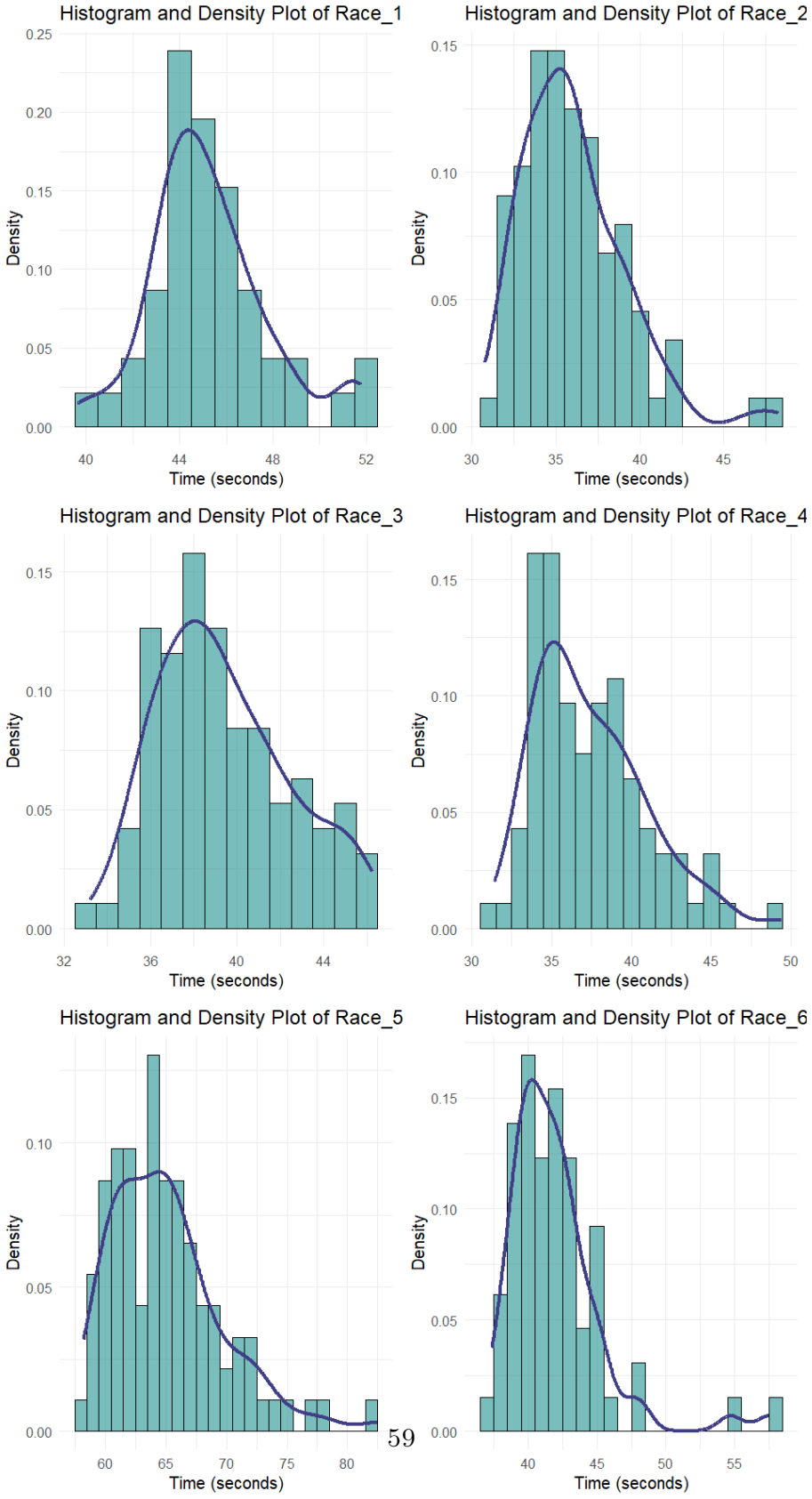

Figure 6.1: Histograms and density plots of 6 competitions among the races considered in the case study.

Chapter 7

Results and discussion

Building upon the methodology outlined in the previous chapter, this section provides a comprehensive overview and discussion of the results derived from the implementation of the previously defined models: the order statistics model Plackett-Luce and a Bayesian hierarchical approach. Both have been applied to analyze athlete performance in ski competitions, using rankings and competitions track times, respectively, as data.

7.1 Plackett-Luce model results

For what concerns the approach employing Plackett-Luce model, we have exploited the `PackettLuce` package. Given that the directed graph constructed from the ranking data is connected, we have not required the use of pseudo-ranks.

In this regard, ranked items can be represented as nodes in a directed graph, where directed edges between nodes denote the implied wins and losses between pairs of items. Specifically, an adjacency matrix can be defined to represent the graph, enabling us to verify the connectivity of the graph and ensuring that all nodes are reachable from one another. This property is crucial for the proper application of the Plackett-Luce model, as it relies on the assumption of a connected ranking structure, in order for the maximum likelihood estimate to be defined.

Therefore, the *Standard Maximum Likelihood* method was employed to fit the data, utilizing the default "Iterative Scaling" algorithm. The algorithm achieved convergence within 15 iterations. It is important to note that, given the case study involved 12 athletes, the procedure was computationally manageable, without the need of further exploring acceleration techniques.

The Z-tests for the estimated ability parameters assess the null hypothesis that the difference between each parameter and the mean worth value is zero, using the group's mean as a reference point. Table (7.1) presents the estimated coefficients, standard errors, and p-values obtained from the analysis. From these results, it is evident that skiers indexed as 3, 4, and 11 display abilities that significantly deviate from the average, underperforming compared to the group mean. In contrast, skiers indexed as 2, 5, 6, 8, and 9 show performance above the average. The standard error values provide further insight into the

consistency of each athlete's performance, with most skiers exhibiting a standard error of approximately 0.4. However, athletes 3, 4, and 11, who also underperformed, showed higher standard error values, indicating either greater inconsistency in their results or low partecipation in competitions.

Table 7.1: Table of skier estimated ability, standard error and p-value choosing as reference the mean ability value with Plackett-Luce model.

Skier	**Estimate**	**Std. Error**	**P-value**
1	0.68901	0.41570	0.0974
2	1.61279	0.44047	0.0003
3	-4.99757	1.18099	2.32e-05
4	-3.57481	0.78092	4.70e-06
5	1.85377	0.44042	2.56e-05
6	2.77352	0.43873	2.59e-10
7	-0.01022	0.40090	0.9797
8	1.65475	0.42593	0.0001
9	2.43365	0.42462	9.96e-09
10	-0.03666	0.42026	0.9305
11	-2.30488	0.63875	0.0003
12	-0.09335	0.41245	0.8209

As outlined in [16], standard errors vary depending on the chosen reference level. Quasi-standard errors (QSE) solve this issue by remaining constant regardless of which item is set as the reference. The latter result is achieved by QSE by introducing in the computation the covariance terms of all the other parameters with respect to the reference one. While a detailed proof of QSE's independence from the reference is beyond the scope of this master thesis, we will provide a formal definition of QSE. For further explanation, readers can refer to [16]. *Quasi-standard error* for item i, can be defined as follows:

$$\text{QSE}_i = \sqrt{\Sigma_{ii} - 2\frac{\Sigma_{ir}}{n} + \frac{\sum_{k=1}^{n} \Sigma_{rk}}{n^2}}$$

where:

- Σ_{ii} is the variance of the estimated parameter for item i,
- Σ_{ir} is the covariance between the parameter for item i and the reference item r,
- n is the number of items,
- Σ_{rk} is the covariance between the reference item and item k.

This ensures that even when a natural reference exists, as in this case, the mean group performance, the uncertainty surrounding the worth of that item can be assessed, allowing possible further comparisons where the same variables but different reference levels are used. Thus log-abilities with quasi-standard errors have been visually represented in Figure 7.1.

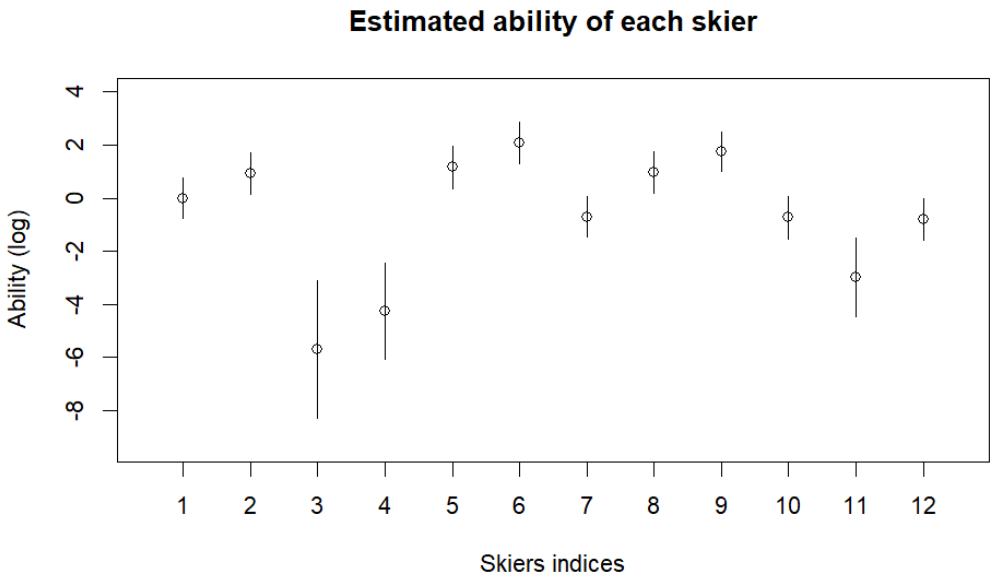

Figure 7.1: Skiers' abilities estimated log-values and quasi standard errors obtained applying Plackett-Luce model.

One aspect worth to be underlined of models relying on latent variables is poor interpretability. The ability coefficients obtained from the Plackett-Luce model rely on latent variables which quantify the underlying ability of each individual, but they are abstract constructs that do not correspond to any directly observable measures.

7.2 Hierarchical Bayesian model results

The following section presents the results obtained from the implementation of the hierarchical Bayesian model. The model was implemented using the JAGS framework, with three Markov chains initialized and run in parallel. Each chain was executed for 100,000 iterations, with an initial burn-in period of 2,000 iterations to ensure that the model reached a stable sampling phase. Convergence of the chains was assessed using the Rhat diagnostic coefficient, which indicated that all chains successfully converged within the allocated number of iterations.

As discussed in the methodology chapter, ability of an athlete has been interpreted as the percentage variation in his race time, with respect to the average competition time. Mean estimates, standard errors and the ability measure computed from the parameter mean estimate are displayed in Table (7.2). Specifically, we recall that in order to obtain the previous interpretation, ability measure has been computed as:

$$a_i = \exp(\theta_i) - 1.$$

Thus, for example, taking the skier indexed by 6, she is expected to perform about 13.5% better than the average competitor, given the pool of athletes considered.

As we can see from both numerical results and HDI displayed in Figure 7.2, SD assume quite high values, meaning that either insufficient data have been provided to obtain more accurate estimates, or athletes performance was inconsistent.

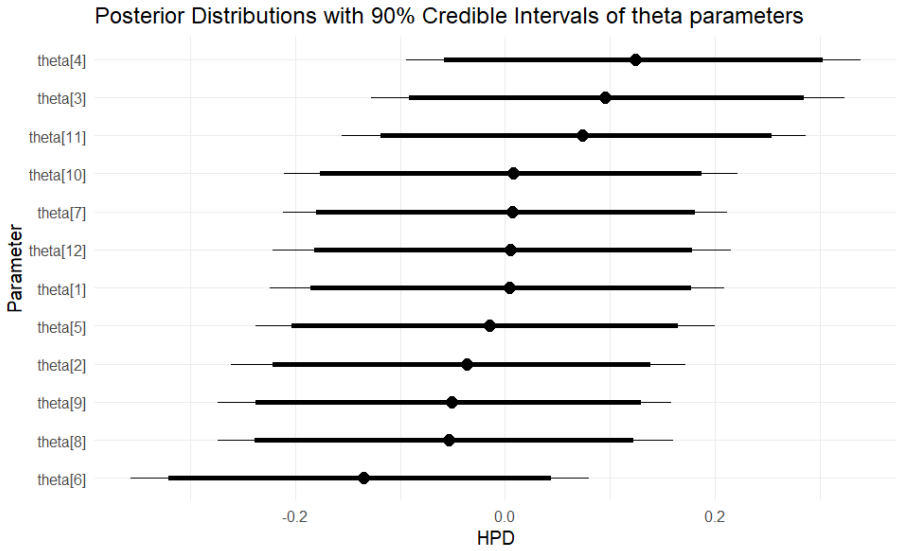

Figure 7.2: HDI for athlete abilities sorted in decreasing order of mean parameter value.

Parameter	**Mean**	**SD**	**Ability (%)**
theta[1]	0.001	0.109	0.001
theta[2]	-0.039	0.111	-0.038
theta[3]	0.097	0.115	0.102
theta[4]	0.125	0.111	0.133
theta[5]	-0.017	0.110	-0.017
theta[6]	-0.136	0.111	-0.135
theta[7]	0.004	0.110	0.007
theta[8]	-0.054	0.111	-0.054
theta[9]	-0.053	0.111	-0.053
theta[10]	0.007	0.110	0.008
theta[11]	0.071	0.113	0.074
theta[12]	0.002	0.110	0.005

Table 7.2: Summary of results applying the Bayesian hierarchical approach: Parameter, Mean, SD, and Ability (%)

7.2.1 Model validation

In order to validate the model two different validations have been carried out:

- **Rank comparison**: Rankings have been constructed using the estimated abilities from the hierarchical Bayesian model. The latter rankings have been compared against two benchmarks: the rankings derived from the Plackett-Luce model and the average rank for each athlete, computed directly from the observed data.

- **Posterior predictive check**: A posterior predictive check has been conducted to evaluate how well the model reproduces the observed data. Specifically, distributions of the athlete abilities and competition-specific parameters were simulated based on the posterior estimates of the model. These simulated distributions were then compared to the actual data to identify any discrepancies.

Rank comparison

Ranking estimated abilities obtained from both models (hierarchical Bayesian model and Plackett-Luce), enables to make a comparison of the orderings obtained, in order to check for consistency of results between the models. Rank orderings obtained are shown in Figure 7.3 and Table 7.3. The results show a strong overall consistency in the rankings, demonstrating that both models capture similar underlying patterns in athlete performance. Notably, the positioning of three athletes coincides exactly between the two models. However, for what concerns the remaining athletes, while the specific ranks differ between the models, these discrepancies can largely be attributed to the proximity of the ability estimated measures of those athletes. In both the hierarchical Bayesian model and the Plackett-Luce model, several athletes have very similar estimated ability values, resulting in minor differences in rank assignments.

Subsequently, the rankings obtained have been compared with the average rank of each skier, directly computed from data. As we can see from Table 7.3, the Plackett-Luce model outcome almost coincides with average rank orderings, except from minor differences in the positioning between the 7-th and the 10-th positions.

Rank	Model 1	Model 2	Avg Rank
1	6	6	6
2	9	8	9
3	5	9	5
4	8	2	8
5	2	5	2
6	1	1	1
7	7	12	10
8	10	7	12
9	12	10	7
10	11	11	11
11	4	3	4
12	3	4	3

Table 7.3: Comparison of Model 1: Plackett-Luce, Model 2: hierarchical Bayesian model, and average rank athlete orderings. In columns Model 1, Model 2 and Avg Rank are present the index of the athlete assigned to the position given by Rank column.

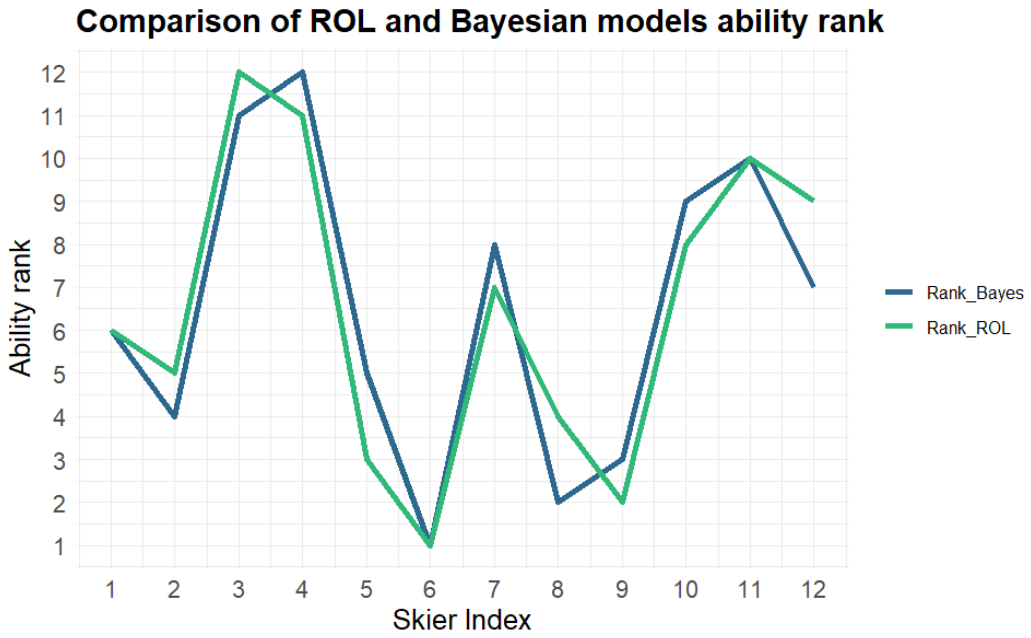

Figure 7.3: Comparison of abilities ranking obtained from the model fitting of a Bayesian hierarchical model and Plackett-Luce model.

Posterior predictive check

In this section, we validate the hierarchical Bayesian model by generating and analyzing simulated race times for each competition. Specifically, for each athlete i in competition

j, we simulate 100 race times from the distribution

$$\hat{y_{ij}} \sim \mathcal{N}(\theta_i + \beta_j, \tau)$$

where θ_i represents the estimated athlete's ability, β_j denotes the average race time estimate for that competition, and τ is the estimated precision parameter controlling the variability in race times. The idea behind posterior predictive check is to compare the model's simulated predictions from the posterior distribution with the actual race times recorded for each athlete in the corresponding competition. For each competition, we simulate 100 race times for every athlete. These simulated race times are then plotted alongside the actual observed race times. Importantly, if an athlete did not participate in a particular competition, the simulated data are represented, but real data to compare the results are missing.

As we can see from the subsequent plots, where real data is represented with an "X" symbol, visually the mean of the simulated data mostly coincide with real collected data. Exceptions exist, such as in Race 1 in Figure 7.4, where athlete 1 and 5 real performances constitute outliers of their posterior distribution. This discrepancy can be attributed to the fact that both athletes in that competition went off the track at some point during the competition, which caused them to stop and then resume the race from the point where they left off, thereby losing valuable time. The other exception is given by race 12 in Figure 7.9, where skier 6 outperformed all the performance expectations from the posterior distribution. This can be considered as a proper outlier, since all other predictions for the same athlete are coherent with the posterior distribution given by the model.


Simulated Data and True Race Times for race 1
Race Times (centiseconds)
6000
5000
4000
3000
1
2
3
4
5
6
7
8
9
10
11
12
Athlete
Simulated Data and True Race Times for race 2
Race Times (centiseconds)
6000
5000
4000
3000
1
2
3
4
5
6
7
8
9
10
11
12
Athlete

Figure 7.4


Simulated Data and True Race Times for race 3
Race Times (centiseconds)
8000
7000
6000
5000
4000
1
2
3
4
5
6
7
8
9
10
11
12
Athlete
Simulated Data and True Race Times for race 4
Race Times (centiseconds)
8000
7000
6000
5000
4000
1
2
3
4
5
6
7
8
9
10
11
12
Athlete

Figure 7.5

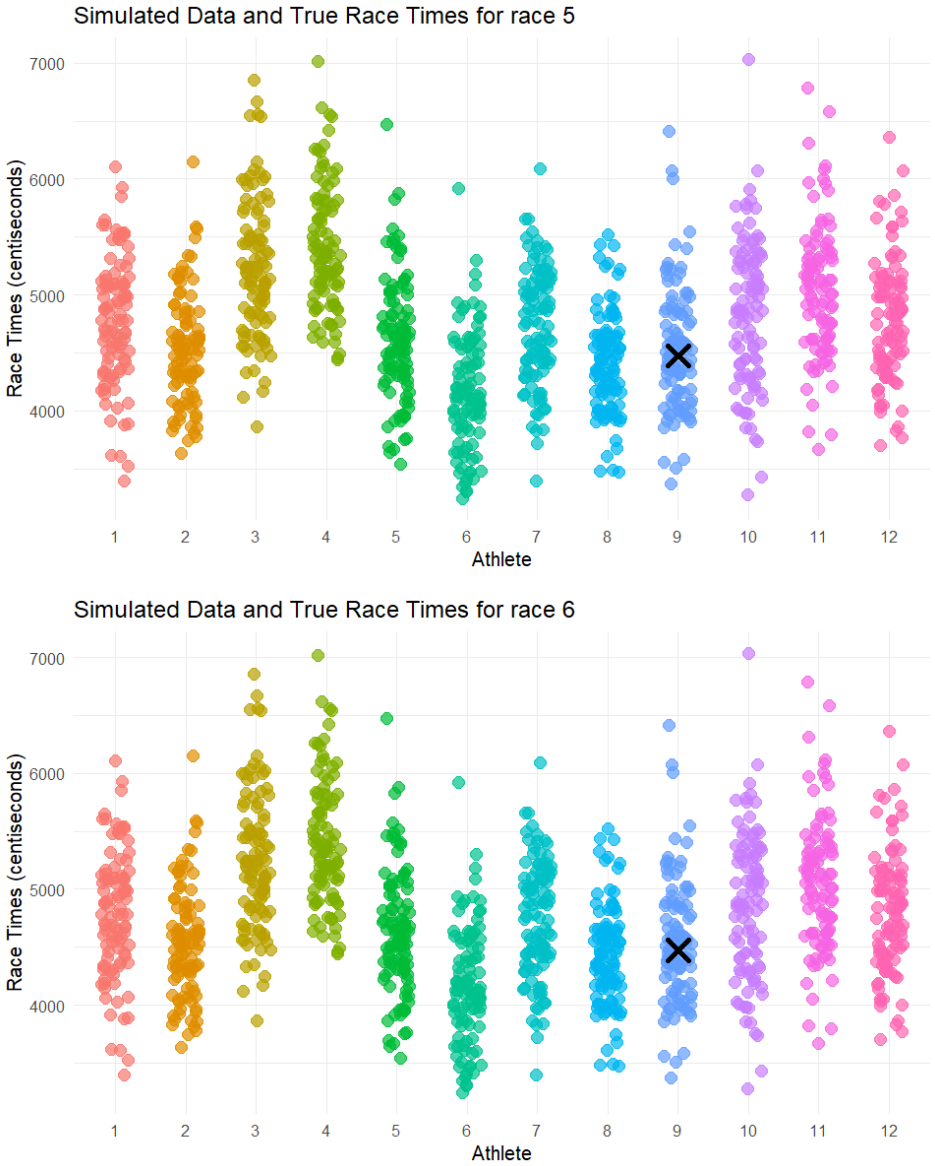
Simulated Data and True Race Times for race 5
Race Times (centiseconds)
7000
6000
5000
4000
1
2
3
4
5
6
7
8
9
10
11
12
Athlete
Simulated Data and True Race Times for race 6
Race Times (centiseconds)
7000
6000
5000
4000
1
2
3
4
5
6
7
8
9
10
11
12
Athlete

Figure 7.6

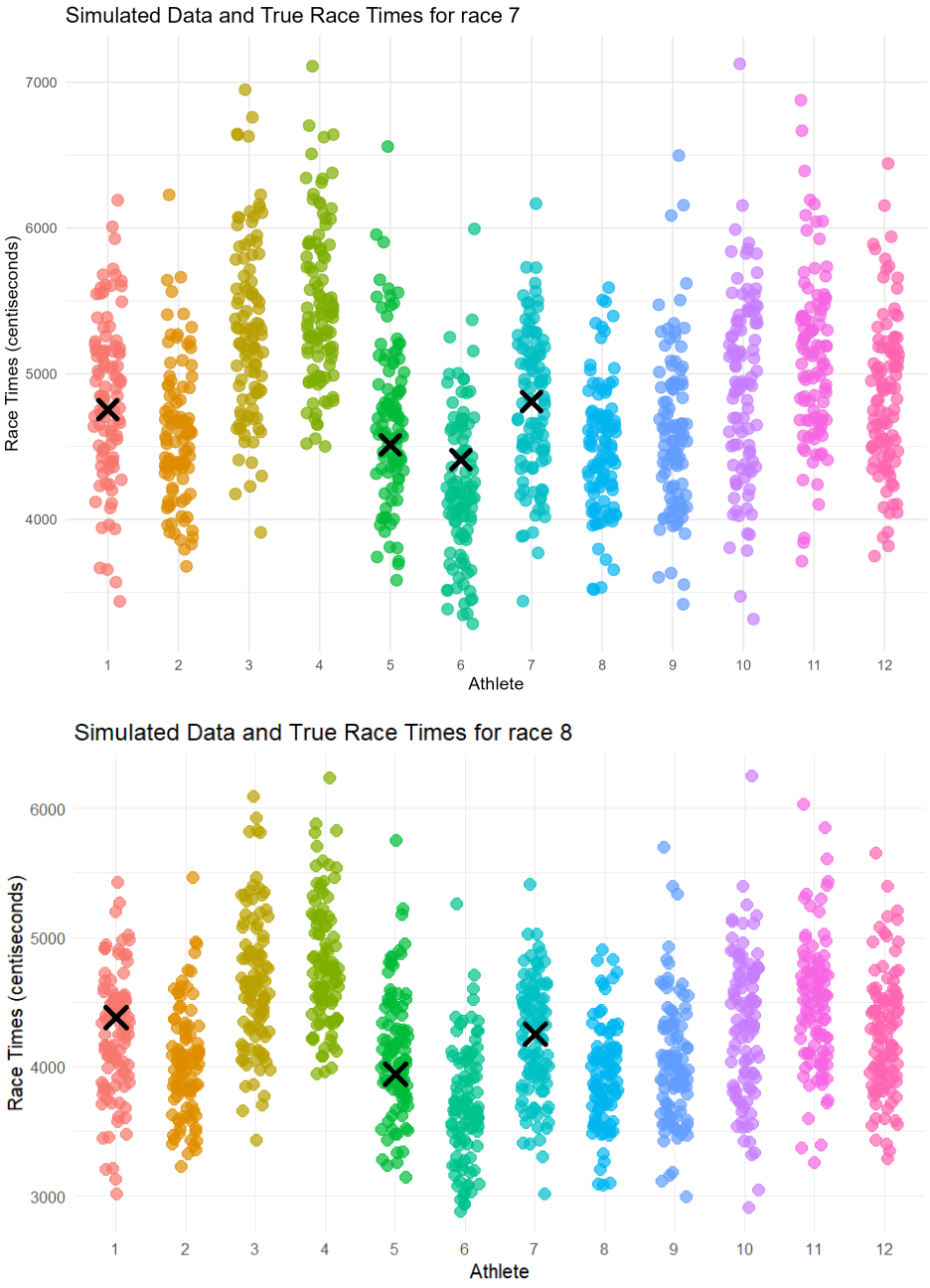
Simulated Data and True Race Times for race 7
Race Times (centiseconds)
7000
6000
5000
4000
1
2
3
4
5
6
7
8
9
10
11
12
Athlete
Simulated Data and True Race Times for race 8
Race Times (centiseconds)
6000
5000
4000
3000
1
2
3
4
5
6
7
8
9
10
11
12
Athlete

Figure 7.7

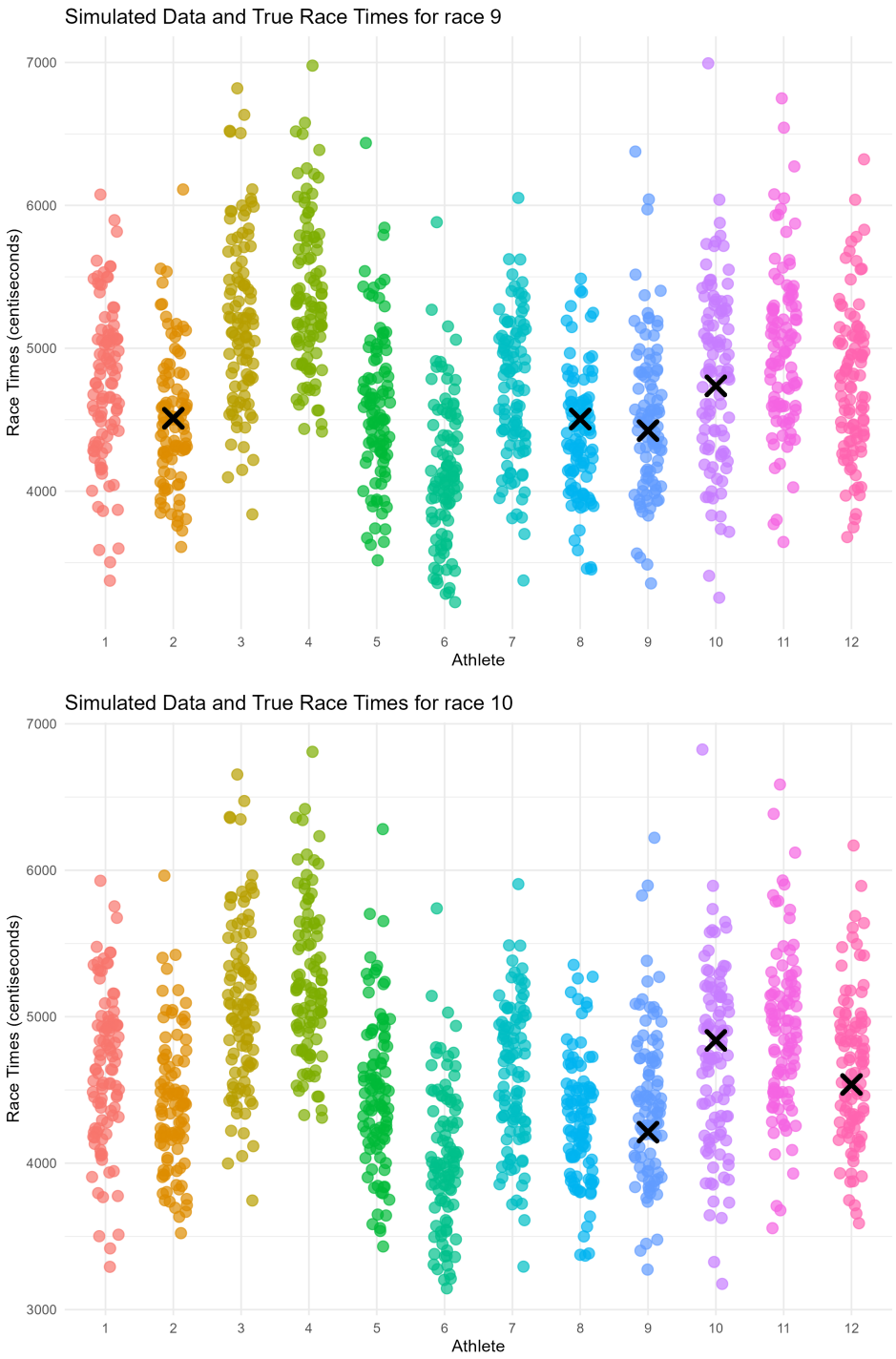
Simulated Data and True Race Times for race 9
Race Times (centiseconds)
7000
6000
5000
4000
1
2
3
4
5
6
7
8
9
10
11
12
Athlete
Simulated Data and True Race Times for race 10
Race Times (centiseconds)
7000
6000
5000
4000
3000
1
2
3
4
5
6
7
8
9
10
11
12
Athlete

Figure 7.8

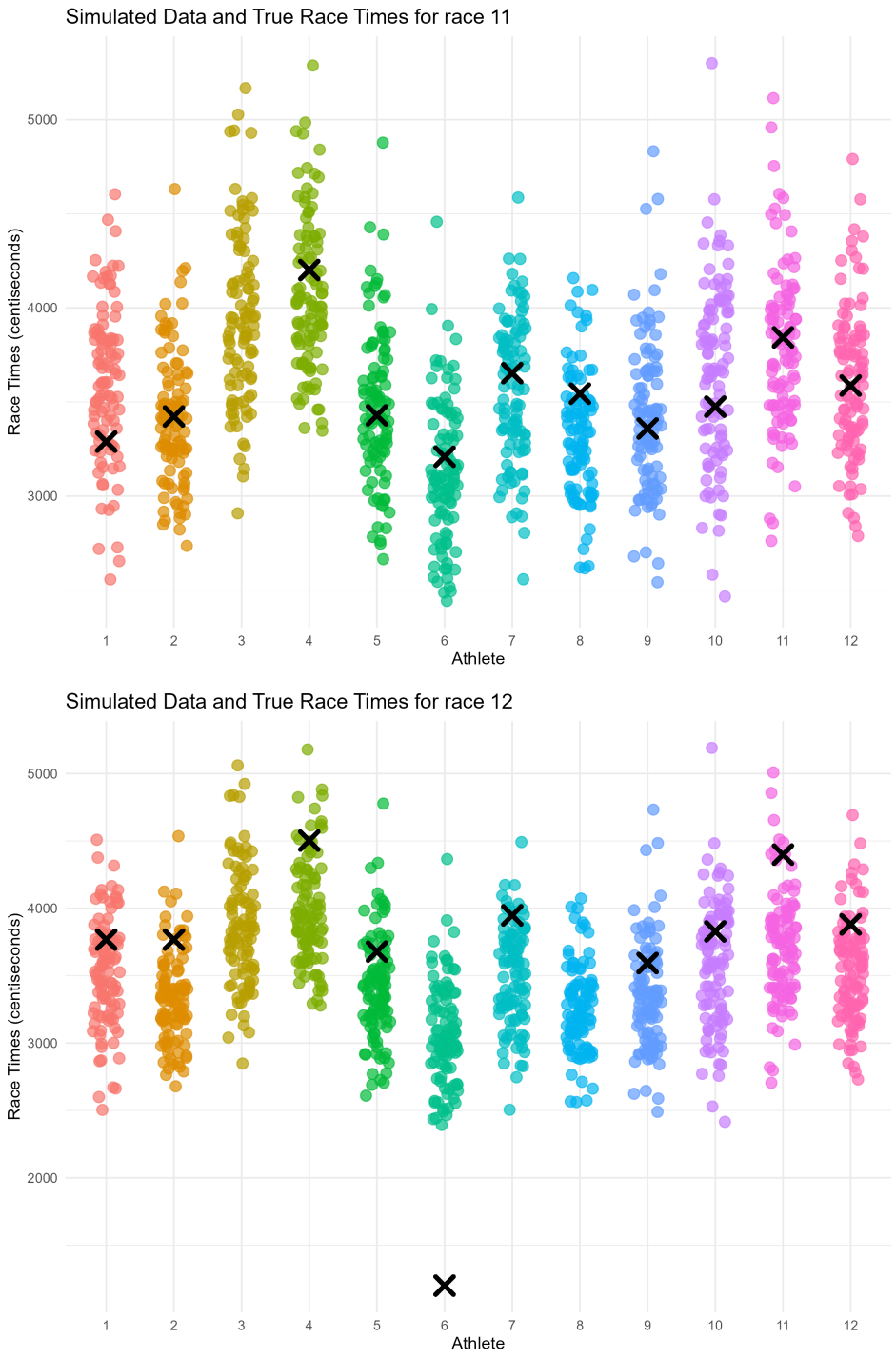
Simulated Data and True Race Times for race 11
Race Times (centiseconds)
5000
4000
3000
1
2
3
4
5
6
7
8
9
10
11
12
Athlete
Simulated Data and True Race Times for race 12
Race Times (centiseconds)
5000
4000
3000
2000
1
2
3
4
5
6
7
8
9
10
11
12
Athlete

Figure 7.9

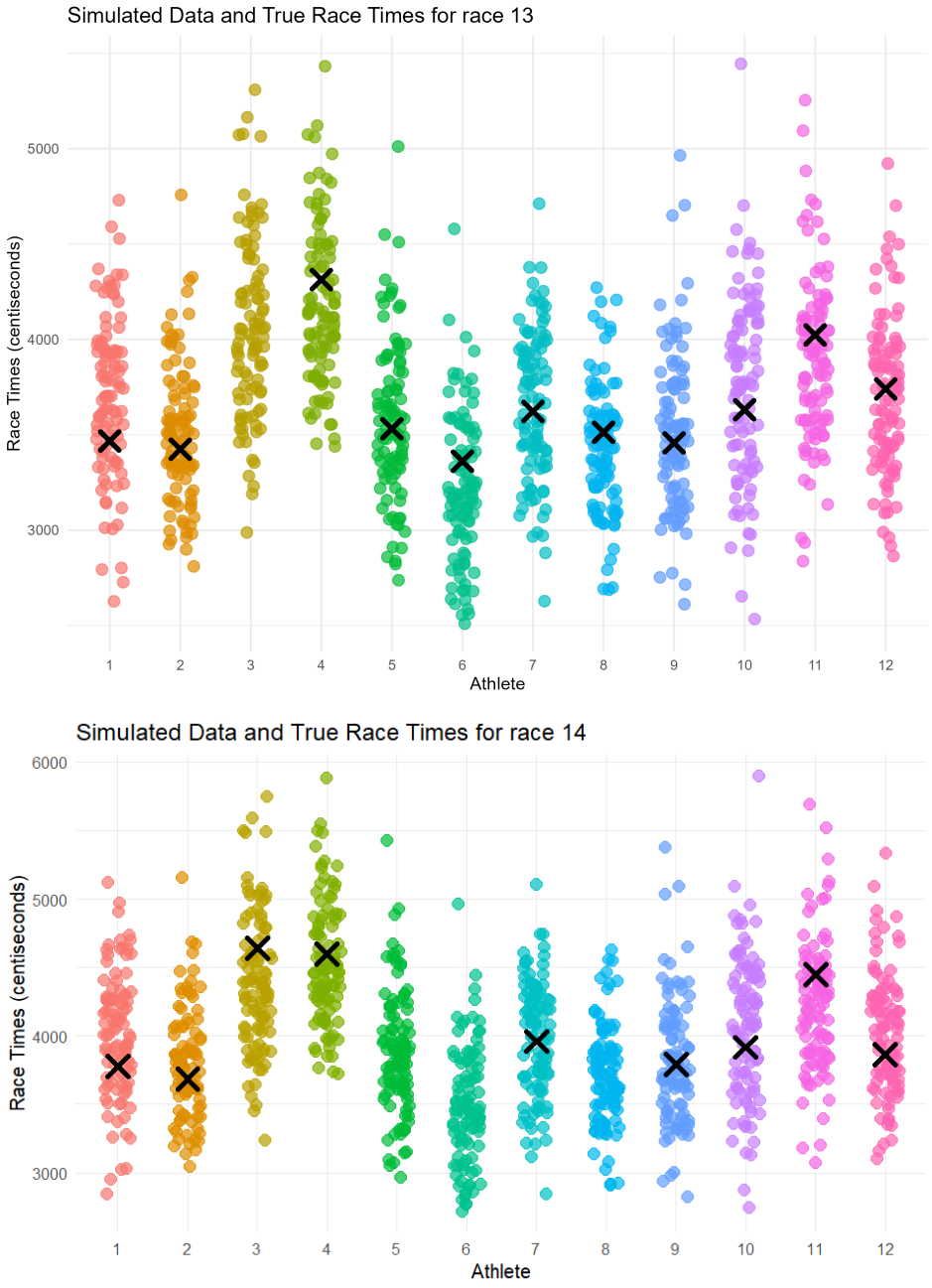
Simulated Data and True Race Times for race 13
Race Times (centiseconds)
5000
4000
3000
1
2
3
4
5
6
7
8
9
10
11
12
Athlete
Simulated Data and True Race Times for race 14
Race Times (centiseconds)
6000
5000
4000
3000
1
2
3
4
5
6
7
8
9
10
11
12
Athlete

Figure 7.10


Simulated Data and True Race Times for race 15
Race Times (centiseconds)
3000
2500
2000
1500
1
2
3
4
5
6
7
8
9
10
11
12
Athlete
Simulated Data and True Race Times for race 16
Race Times (centiseconds)
6000
5000
4000
3000
1
2
3
4
5
6
7
8
9
10
11
12
Athlete

Figure 7.11

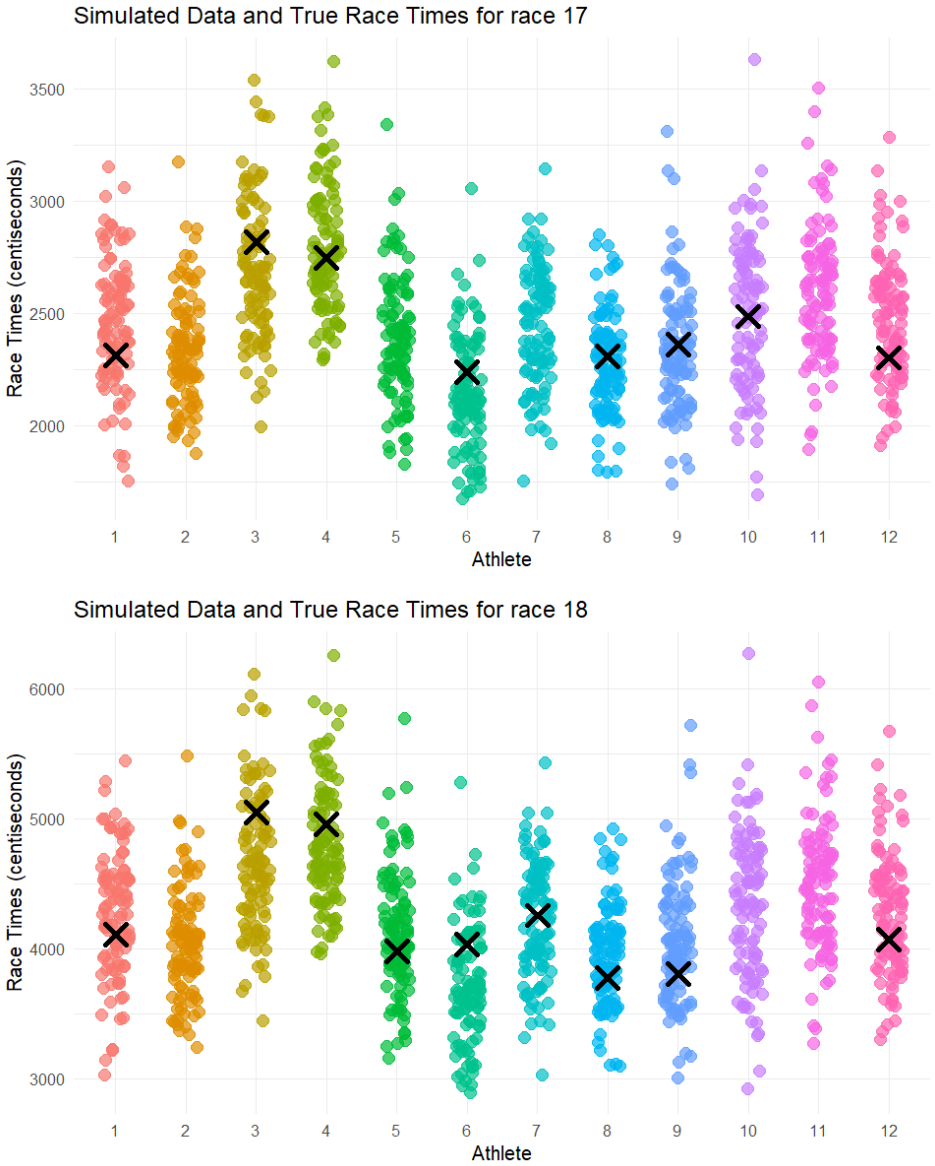
Simulated Data and True Race Times for race 17
Race Times (centiseconds)
3500
3000
2500
2000
1
2
3
4
5
6
7
8
9
10
11
12
Athlete
Simulated Data and True Race Times for race 18
Race Times (centiseconds)
6000
5000
4000
3000
1
2
3
4
5
6
7
8
9
10
11
12
Athlete

Figure 7.12

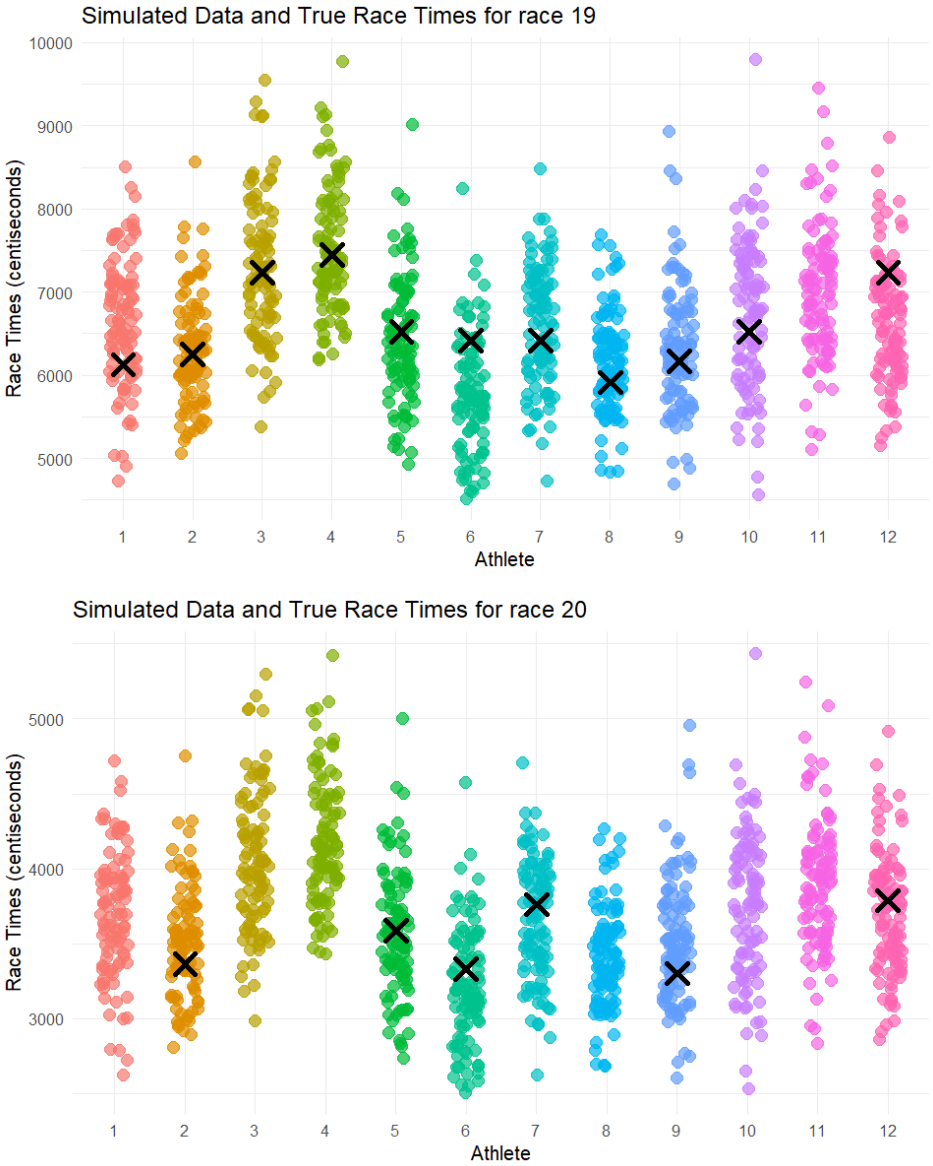
Simulated Data and True Race Times for race 19
Race Times (centiseconds)
10000
9000
8000
7000
6000
5000
1
2
3
4
5
6
7
8
9
10
11
12
Athlete
Simulated Data and True Race Times for race 20
Race Times (centiseconds)
5000
4000
3000
1
2
3
4
5
6
7
8
9
10
11
12
Athlete

Figure 7.13

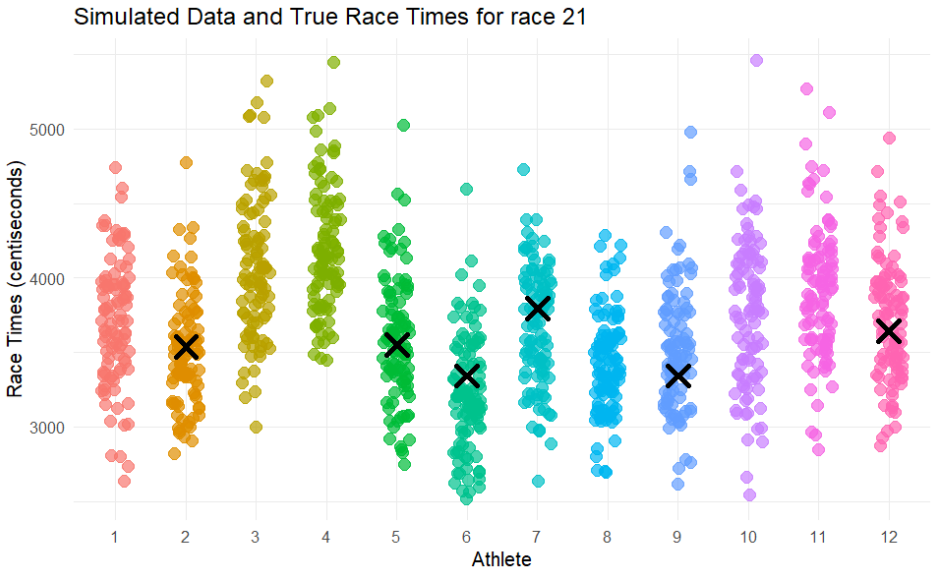

Figure 7.14: Posterior predictive check for each of the 21 analyzed alpine skiing competitions.

Chapter 8

Conclusions

The primary objective of this thesis was to develop statistical models for quantifying and comparing athlete abilities in alpine skiing competitions, accounting for the varying race conditions and disciplines. The goal was to develop an easily interpretable, scalable and accurate model.

The results revealed both advantages and limitations in using ranking models. These models fail to provide a clear relationship between the estimated parameters and actual race completion times, making interpretation challenging. Additionally, ranking models are intrinsically not updatable, requiring a complete re-computation whenever new data is introduced. Nevertheless, in situations where only ranking data are available and the analysis involve less than 15-20 athletes, ranking models proved to be a valid alternative for evaluating and comparing athlete abilities within the considered group.

In contrast, the hierarchical Bayesian model offered a more interpretable and flexible framework for assessing athlete abilities. This approach utilized informative priors, assuming prior knowledge of the overall mean and standard deviation for each race. Athlete-specific abilities were interpreted as the average percentage difference from the race's mean time, while competition-specific parameters captured the overall race time and variability.

However some limitations exist. It is only valid if the competitions being analyzed feature a relatively consistent set of athletes, ensuring the stability of the average completion times across races. For instance, if a below-average athlete competes in two different races—one with highly skilled competitors and another with significantly less experienced participants—their estimated ability would vary drastically. In the first race, their ability might be judged as below average, while in the second, they could appear far superior, leading to inconsistencies in ability measurement.

Despite these limitations, posterior predictive checks demonstrated that, under the assumptions discussed, the hierarchical Bayesian model effectively captured athlete performance and race-specific parameters. This suggests that the model is a reliable tool for assessing skier abilities, provided the data are carefully selected to ensure consistency across competitions.

Bibliography

[1] Paul D. Allison and Nicholas A. Christakis. Logit models for sets of ranked items. *Sociological Methodology*, 24:199–228, 1994.

[2] Mukhtar M. Ali. Probability models on horse-race outcomes. *Journal of Applied Statistics*, 25(2):221–229, 1998.

[3] Mark E. Glickman and Jonathan Philip Hennessy. A stochastic rank ordered logit model for rating multi-competitor games and sports. *Journal of Quantitative Analysis in Sports*, 11:131 – 144, 2015.

[4] Meng Liu, Yan Chen, Zhenxiang Guo, Kaixiang Zhou, Limingfei Zhou, Haoyang Liu, Dapeng Bao, and Junhong Zhou. Construction of women's all-around speed skating event performance prediction model and competition strategy analysis based on machine learning algorithms. *Frontiers in Psychology*, 13, 2022.

[5] Camilla H. Carlsen, Cecilia Severin, Øyvind Sandbakk, and Julia K. Baumgart. Comparison of race time-differences between and within para and able-bodied cross-country skiers. *Frontiers in Sports and Active Living*, 3, 2022.

[6] Louis Leon Thurstone. A law of comparative judgement. *Psychological Review*, 34:278–286, 1927.

[7] R. Duncan Luce. *Individual Choice Behavior: A Theoretical analysis*. Wiley, New York, NY, USA, 1959.

[8] R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):193–202, 1975.

[9] Mayer Alvo and Philip L.H. Yu. *Statistical Methods for Ranking Data*. Springer Publishing Company, Incorporated, 2014.

[10] Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka, editor, *Fontiers in Econometrics*, pages 105–142. Academic press, New York, 1974.

[11] Ian Hamilton, Nick Tawn, and David Firth. The many routes to the ubiquitous bradley-terry model, 2023.

[12] John Kruschke. *Doing Bayesian Data Analysis (Second Edition).* Academic Press, Boston, 2015.

[13] Christian P. Robert. *The Metropolis–Hastings Algorithm*, pages 1–15. John Wiley Sons, Ltd, 2015.

[14] C.P. Robert and G. Casella. *Monte Carlo statistical methods.* Springer Verlag, 2004.

[15] Heather Turner, Jacob van Etten, David Firth, and Ioannis Kosmidis. Modelling rankings in r: the plackettluce package. *Computational Statistics*, 35, 09 2020.

[16] David Firth. Overcoming the reference category problem in the presentation of statistical models. *Sociological Methodology*, 33(1):1–18, 2003.