UNIVERSITÀ CATTOLICA DEL SACRO CUORE

Campus of Milan

Faculty of Economics, Mathematical, Physical, and Natural Sciences.

Master program in Data Analytics for Business

UNIVERSITÀ
CATTOLICA
del Sacro Cuore

A Bayesian hierarchical approach for modeling the outcome of football matches

Graduand: Francesco Esposito

Student ID: 5108223

Supervisor:

Dott. Francesco Denti

Academic Year 2022−2023

Contents

Chapter 1

Introduction

For a considerable period of time, data analysis has consolidated its position as a discipline of fundamental importance, showing increasing development and finding application in various contexts, including sports, particularly football.

Football, as a sport characterised by highly variable and random events, presents a considerable diversity in the frequency and patterns of such events. This variability makes opinion-based interpretation susceptible to bias, undermining the objectivity and objectivity of evaluations.

The use of data analysis allows for generalisable, objectively based insights and opinions. This paves the way for the development of coherent visions of the game, supported by solidly

grounded future projections.

In recent years, several approaches have been proposed to estimate the key elements influencing the outcome of a game and to predict the score of specific matches. One crucial aspect concerns the distribution associated with the number of goals scored in a single match by the two contending teams. Although Binomial or Negative Binomial distributions were initially adopted in the late 1970s [Pollard, 1977], the Poisson distribution has been generally recognised as an appropriate model for such quantities. Often, the simplifying assumption of independence between the goals scored by the home team and the away team is adopted. For example, Maher [1982] implemented a model with two independent Poisson variables based on the attacking strength of one team and the defensive weakness of the opposing team. Over time, studies have shown relatively modest empirical levels of correlation between the two quantities. As a result, more advanced models were proposed, such as that of Dixon and Coles [1997], which introduced a correction factor to the independent Poisson model to improve predictions. Karlis and Ntzoufras [2003] proposed the adoption of a Bivariate

Poisson distribution with a more complex formulation, including a parameter for the covariance between the goals scored by the two contending teams. Within the Bayesian framework, it was shown that the use of a Bivariate Poisson model is avoidable. By assuming two conditionally independent Poisson variables for the number of goals scored, the observation on correlation will be handled through the stratification that characterises a hierarchical model.

Historically, all models have assigned significant importance to the effects of certain teams in order to allow for the evaluation of offensive and defensive forces, commonly used to model the scoring rate of a given team.

This dissertation is organized as follows.

In Chapter 2, we will introduce the theoretical aspects of the Bayesian statistic, computational strategies for Bayesian inference and delve into the hierarchical structure, highlighting the distinctive features of both approaches: a model with Beta distributions as prior and a model with Gaussian Processes.

Subsequently, in Chapter 3 we will describe the results obtained from the analysis and detail the structures of the two

models, highlighting the differences and key elements that distinguish one from the other.

Finally, in Chapter 4 we will analyze the results obtained from the two models, contextualizing them within the reality, emphasizing the advantages and disadvantages through a thorough comparison.

The discussion and a potential approach that gives continuity to the two seen in this paper, can be found in Chapter 5.

Chapter 2

Some methodological preliminaries

2.1 Frequentist and Bayesian Statistics compared

Frequentist statistics, a classical approach in statistical analysis, is founded on the idea that probabilities are tied to the frequency of random events in repeated trials. For instance, in the case of rolling a fair die, the probability of each number is interpreted as the frequency with which that number appears in a sufficiently large number of rolls. More technically, this approach employs the maximum likelihood estimation (MLE), a method whose

principle is evident in its name: it fits a model that maximizes the probability of having observed the actual data. Essentially, MLE produces a point estimate representing the most probable value of the parameter, given the observed data. The primary goal in this paradigm is to quantify the uncertainty through the provision of point estimates and confidence intervals.

On the other hand, Bayesian statistics offers a more flexible and dynamic approach to statistical analysis. In this context, parameters are treated as random variables, and subjective information is incorporated through the use of prior probability distributions. It begins with a prior belief about the probability of an event, encapsulating the *initial belief* which is the initial probability of an event or hypothesis before any evidence is observed. This belief is then continuously updated in light of new data or evidence, reflecting the *likelihood* of the observed data, which refers to the probability of observing the data given the hypothesis. As Bayesian statistics progresses, it refines and updates this initial belief, resulting in the *subsequent belief*, the updated probability of the hypothesis given the observed data. This iterative process captures the essence of Bayesian infer-

ence, providing a robust mathematical tool for incorporating both prior beliefs and recent evidence, ultimately generating new posterior beliefs.

2.2 Bayes Theorem

The foundation of Bayesian statistics lies in the fundamental assumption that the reference parameter in the distribution is also treated as a random variable, as mentioned in the previous paragraph. Consequently, this parameter follows a specific distribution defined as *a priori* in Bayesian statistics. The parameter, denoted as θ, can adopt various prior distributions, the choice of which depends on the available information. The goal is to assign θ a distribution that assigns higher probabilities to values considered more plausible for θ.

Considering the observed data $y = (y_1, ..., y_n)$ and the density function (or probability) $f(y|\theta)$, we define the likelihood function as:

$$L(\theta) = f(y_1, ..., y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

Subjecting θ to a specific prior distribution, indicated as $\theta \sim \pi(\theta)$, Bayes' theorem can be formulated as:

$$\pi(\theta|y) = \frac{f(y|\theta) \cdot \pi(\theta)}{f(y)} = \frac{f(y|\theta) \cdot \pi(\theta)}{\int_\Theta f(y|\theta) \cdot \pi(\theta)\, d\theta}$$

Subsequently, the posterior distribution can be approximated by considering only the numerator term since the denominator can be treated as a constant, not depending on the parameter of interest. The posterior distribution can be redefined proportional to the product of the data likelihood and the prior distribution:

$$\pi(\theta|y) \propto f(y|\theta) \cdot \pi(\theta)$$

In some cases, the integral in the denominator is omitted, especially when the posterior distribution $\pi(\theta|y)$ follows the same distribution as the prior $\pi(\theta)$. This case is known as the use of conjugate distributions for parameters, significantly simplifying the computation of the posterior. We will discuss about it in a later section.

Let us now look in detail at the main players in Bayesian statistics.

2.2.1 Likelihood and Prior distributions

Likelihood and prior distribution play pivotal roles in Bayesian statistics. Considering y as the observed data, it is assumed to be generated from a random variable Y with density $p(y;\theta)$, where θ represents an unknown parameter. The probability function $p(y;\theta)$ is considered known, except for the parameter θ. As mentioned earlier, in the Bayesian context, the parameter is treated as a random variable. This parameter follows a distribution with a density called the prior distribution, denoted as $\pi(\theta)$, summarizing prior knowledge about θ. The prior distribution $\pi(\theta)$ usually depends on other parameters, known as hyperparameters, which can be known or unknown and are also treated as random variables.

To apply Bayes' theorem in inference, the initial step involves defining the statistical model.

Prior information about the parameter is provided by its prior distribution, while the knowledge derived from observations is represented by the likelihood $L(\theta;y)$, defined as:

$$L(\theta;y) = p(y;\theta)$$

Assuming y is fixed, in the case where the sample consists of n independent observations, the likelihood takes the form:

$$L(\theta; y) = \prod_{i=1}^{n} p(y_i; \theta)$$

The choice of prior information plays a crucial role in inference, as it can influence results for the same y. The same experimental result could lead to substantially different inferential conclusions if the prior information introduced into the model were even slightly different. Therefore, the choice of priors can significantly influence inferential conclusions, emphasizing the importance of accurate selection to obtain robust estimates. An crucial role in these cases is played by the conjugate priors and non-informative priors, that allow the definition of the model while optimizing computational aspects.

Conjugate and Non-informative prior distributions

To overcome the complexity of the computetional challenges, one can turn to prior distributions specifically designed to ensure that the posterior distribution maintains the same functional form. These distributions are know as conjugate priors.

The term *conjugate* in Bayesian statistics refers to a special relationship between the prior distribution and the posterior distribution. Specifically, in the use of conjugate priors, the distribution family remains unchanged, but the parameters are updated based on new information obtained from observed data.

To better understand, let's consider a certain family of prior distributions, represented as $\pi(\theta|a, b, c, \ldots)$, where $a, b, c, \ldots$ are the initial parameters of the distribution. The posterior will still take the form $\pi(\theta|a', b', c', \ldots)$, where $a', b', c', \ldots$ are the updated parameters based on the observed data.

This parameter update reflects the incorporation of new information from the data and allows for maintaining consistency in the functional form of the distribution, thereby simplifying calculations in the context of Bayesian inference.

In certain contexts, the preference about the choice of the prior is to adopt *non-informative* prior distributions to minimize the impact of subjectivity in Bayesian inference. These distributions are designed to be *neutral* and have a limited effect on the final results. Commonly used distributions in this context include uniform distributions or improper distributions,

which provide little relevant prior information. The use of non-informative prior distributions proves particularly advantageous when aiming to mitigate the influence of subjective opinions in the inferential process, allowing the data to play a decisive role in estimates and predictions.

2.2.2 Inferential approaches

The Bayesian approach, in contrast to classical inference, simplifies the distinction between various inferential procedures, as many of them directly stem from synthesizing the posterior distribution.

Let us begin with the point estimation of the parameter, which relies on the use of position indices derived from the posterior distribution. A commonly used index is the posterior expected value $E(\theta|y)$, obtained by integrating θ multiplied by the posterior density $\pi(\theta|y)$.

Expressing the uncertainty associated with inferential conclusions is crucial in any statistical analysis. A common practice is defining a credibility region at level $1-\alpha$, representing a set of values Θ_α such that $P(\theta \in \Theta_\alpha|y) = 1-\alpha$. This re-

gion can be determined through the quantiles of the posterior distribution $(\theta_{\alpha/2}, \theta_{1-\alpha/2})$, where $P(\theta < \theta_\alpha|y) = \alpha$. An alternative, especially in the presence of asymmetric distributions, is using Highest Posterior Density (HPD) intervals based on the posterior density.

Posterior predictive distribution

The posterior distribution represents our updated knowledge of the parameter θ after observing data from an experiment. However, when we want to make predictions based on this knowledge, it is necessary to consider the predictive distribution for new observations, such as y^*.

The predictive distribution for y^* is obtained through the marginalization of the parameter θ in the likelihood calculated for the new observation, weighted by the posterior distribution of θ. Mathematically, this operation is expressed as:

$$p(y^*|y) = \int_\Theta p(y^*|\theta)\pi(\theta|y)\, d\theta$$

This process reflects how our uncertainty about the parameter estimate influences future predictions. Let us highlight some

key points:

1. *Conditional Likelihood*: $p(y^*|\theta)$ represents the probability of observing y^* given the parameter θ and the observed data y.

2. *Posterior Distribution of* θ: $\pi(\theta|y)$ is the posterior distribution of the parameter θ after observing the data. It represents our updated knowledge of the parameter.

3. *Marginalization*: The integration over θ ($\int_\Theta$) is a marginalization step. This takes into account all possible estimates of θ weighted by our posterior knowledge, allowing us to incorporate uncertainty into the prediction process.

Making predictions based on the predictive distribution is crucial as it provides a more realistic estimate of the variability in future data, considering our uncertainty about the parameter estimate.

2.3 Hierarchical Bayesian models

An extension of the Bayesian approach can be attributed to the Bayesian hierarchical models. In this context, a model is

structured across multiple levels, known as a hierarchical form, allowing for more robust parameter estimation. The integration of sub-models within the hierarchical model is achieved through the application of Bayes' theorem, considering all sources of present uncertainty. The outcome of this integration is the posterior distribution, often referred to as the updated probability estimate, as it incorporates new evidence on the initial parameter distribution.

In contrast to frequentist statistics, which may lead to seemingly incompatible conclusions, Bayesian statistics emphasizes the importance of relevant information in decision-making and belief updating and the role of the robustness of Bayesian hierarchical modeling becomes crucial due to their ability to manage uncertainty.

Bayesian hierarchical models rely on two fundamental concepts to derive the posterior distribution:

1. Hyperparameters: are parameters of the prior distributions;

2. Hyperpriors: representing the distributions of the hyperparameters.

The structure of a Bayesian hierarchical model involves observations (y_j) and parameters (θ_j) governing the data generation process for the j-th unit, where the units j range from 1 to n. It is assumed that the parameters are generated interchangeably from a common population, with a distribution governed by a hyperparameter (ϕ). The model has three phases:

1. *Phase I:*

$$y_j \mid \theta_j, \phi \sim f(y_j \mid \theta_j, \phi)$$

2. *Phase II:*

$$\theta_j \mid \phi \sim \pi(\theta_j \mid \phi)$$

3. *Phase III:*

$$\phi \sim \pi(\phi)$$

The probability in Phase I, $f(y_j \mid \theta_j, \phi)$, is based on the conditional distribution of y_j given θ_j and ϕ, and extends to all units j from 1 to n.

The joint prior distribution in Phase II can be decomposed as follows:

$$\pi(\theta_j, \phi) = \pi(\theta_j \mid \phi)\pi(\phi) \quad \text{for each } j = 1, \ldots, n$$

where $\pi(\theta_j \mid \phi)$ is the prior distribution for θ_j given ϕ, and $\pi(\phi)$ is the prior distribution for the hyperparameter ϕ.

The posterior distribution, proportional to the product of the likelihood, the prior distribution of θ_j given ϕ, and the prior distribution of ϕ, is given by:

$$\pi(\phi, \theta_j \mid y) \propto \prod_{j=1}^{n} f(y_j \mid \theta_j)\pi(\theta_j \mid \phi)\pi(\phi)$$

This expression follows from the Bayes' theorem. In summary, Bayesian hierarchical models allow for the incorporation of flexibility and information from multiple levels, enabling robust estimation of the parameters of interest.

2.4 Computational strategies for Bayesian inference

One significant limitation in this field is the inability to always obtain the posterior distribution analytically. This factor played a pivotal role in limiting the impact of Bayesian statistics in the scientific community during its early stages. However, with advancements in technology, engineering, and industry, powerful

computers have been developed, enabling extensive computational capabilities in a short timeframe. In many cases, particularly when non-conjugate prior distributions are employed, obtaining a closed-form posterior distribution is not feasible. To overcome this challenge, two distinct approaches can be pursued: deriving the posterior distribution through simulation or employing analytical approximation methods. In the following sections, we will delve into the first approach and its variants, which lead to greater computational efficiency, ultimately describing the algorithm we will use for the model in this thesis.

2.4.1 Monte Carlo Markov Chains

The most widely used methods to obtain an approximation of the posterior distribution rely on simulating random values from the same distribution. These methods are known as Monte Carlo methods, and are particularly prevalent. These methods enable the approximation of integrals such as:

$$Q_\theta = E(f(\theta)) = \int_\Theta f(\theta)\pi(\theta|y)\, d\theta$$

The estimation of this integral is given by the empirical mean:

$$\hat{Q}_\theta = \frac{1}{N} \sum_{n=0}^{N} f(\theta_n)$$

which converges in probability to the value Q_θ according to the weak law of large numbers.

It is important to note that the exact generation of values from the distribution of interest is not always possible or straightforward, for example when the inherent complexity of the distribution or the absence of a known analytical form of the normalizing constant may make it difficult or impossible to obtain a closed-form expression for the inverse cumulative distribution function. Even when an analytical form is known, numerical computations can be complex or computationally intensive.

In these cases the role of the Monte Carlo Markov Chain (MCMC) becomes crucial. MCMC is a subset of Monte Carlo methods, relying on the utilization of Markov chains to generate approximate samples from complex probability distributions, such as the posterior distribution in Bayesian inference.

Markov Chain Monte Carlo (MCMC) utilizes Markov chains, sequences of interdependent random variables. These chains

traverse a state space, generating a sequence of values where the transition probability depends solely on the current state, independent of how it was reached.

Ergodicity, a fundamental property for MCMC's Markov chains, implies that the chain can, over a sufficient number of steps, reach any possible state and maintain a stationary distribution.

Stationarity and invariance characterize the stationary distribution of a Markov chain, representing the limit distribution reached after an adequate number of steps. Invariance indicates that once achieved, the stationary distribution remains unchanged over time.

The MCMC algorithm, exemplified by the Metropolis-Hastings algorithm, involves proposing a new state in each step, accepting it with a certain probability determined by the target distribution and the transition probability of the Markov chain.

The process of generating values involves starting from an initial state, proposing new states based on a specific rule, accepting the new state with a certain probability, and repeating the process to generate a sequence of states.

Filtration, or thinning, is sometimes applied due to the corre-

lation between successive states in Markov chains. This involves considering only a fraction of the generated values to reduce correlation.

Convergence in MCMC occurs when the sequence of states reaches the stationary distribution, and the generated values provide a good approximation of the target distribution.

After outlining the functioning of MCMC, we can now explore the main algorithms based on this concept, differing in the approaches adopted in the sampling process.

2.4.2 Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm (1953-1970) is devised to generate Markov chains that converge to a limiting distribution corresponding to the one of interest. Considering the states of the chain, represented by points in the parametric space Θ, the objective is to formulate an algorithm that, starting from an initial state, explores other states with a certain probability of transitioning or remaining in the current state. In essence, given the state θ of the chain, the algorithm proposes a transition to the state θ^*, generated from a chosen density $q(\theta^*|\theta)$

(from which samples can be drawn), and accepts this transition with a probability $\alpha(\theta, \theta^*)$. The acceptance probability is calibrated to ensure that its limiting distribution corresponds to $\pi(\theta|y)$.

Exploiting the reversibility property of the chain (crucial for convergence), expressed as:

$$\pi(\theta|y)q(\theta^*|\theta)\alpha(\theta, \theta^*) = \pi(\theta^*|y)q(\theta|\theta^*)\alpha(\theta^*, \theta)$$

the acceptance probabilities can be defined as follows:

$$\alpha(\theta, \theta^*) = \min\left(1, \frac{\pi(\theta^*|y)q(\theta|\theta^*)}{q(\theta^*|\theta)\pi(\theta|y)}\right)$$

Iterating this algorithm for R steps allows the generation of values from the ergodic chain with distribution $\pi(\theta|y)$.

Since the chain starts from a predefined initial state, some *burn-in* iterations will be necessary before reaching convergence, during which the approach gradually converges to the true distribution. Once this phase is surpassed, the generated values are considered to come from the true limiting distribution, and the initial *burn-in* values are discarded.

2.4.3 Hamiltonian Monte Carlo

The Hamiltonian Monte Carlo (HMC) algorithm is an advanced Markov sampling technique that leverages the concept of kinetic and potential energy, inspired by the principles of Hamiltonian physics. The goal of HMC is to generate samples from the posterior distribution of a parameter of interest.

Consider a random variable θ with posterior distribution $\pi(\theta|y)$ that we want to sample. HMC introduces an auxiliary variable p called momentum and defines a new state space (θ, p). The total energy of the Hamiltonian is given by:

$$H(\theta, p) = U(\theta) + K(p)$$

where $U(\theta)$ is the potential energy associated with the posterior distribution $\pi(\theta|y)$, $K(p)$ is the kinetic energy associated with the auxiliary variable p, and H represents the total energy. The introduction of momentum p allows for a more efficient exploration of the phase space since the total energy is conserved during the Hamiltonian dynamics.

Let us see in details the algorithm:

1. *Initialization*: Start from an initial point (θ_0, p_0) in the

phase space, assigning random values to θ_0 and sampling p_0 from a momentum distribution.

2. *Hamiltonian Proposal*: Simulate the system's evolution through Hamiltonian dynamics using numerical integration. This yields a new proposal (θ', p').

3. *Acceptance/Rejection*: Accept the proposal (θ', p') with a probability depending on the total energy. Acceptance occurs with a probability $\min\{1, \exp[-[H(\theta', p') - H(\theta, p)]]\}$.

4. *Update*: If the proposal is accepted, update the current state to (θ', p'); otherwise, retain the current state.

5. *Iteration*: Repeat steps 2-4 for a desired number of iterations.

Therefore, HMC employs principles of Hamiltonian mechanics to simulate an artificial physical system, enabling an efficient exploration of the phase space and improving the convergence of the algorithm compared to other Markov Chain Monte Carlo approaches.

NUTS algorithm and STAN application

In the following section, we will delve into the workings of the No-U-Turn Sampler (NUTS) algorithm, a sophisticated sampling technique that serves as the foundation for STAN [Team, 2024], which is the statistical language employed in this thesis, well-suited for the development and compilation of hierarchical Bayesian models.

The innovative approach of NUTS, grounded in the concept of *U-Turns*, enables efficient and adaptable Markov Chain Monte Carlo (MCMC) sampling, providing a robust framework for analyzing complex models in Bayesian contexts. The choice of STAN as the implementation language is driven by its ability to fully harness the capabilities of NUTS, facilitating the analysis of hierarchical and intricate data structures.

In a more detailed manner, the No-U-Turn Sampler (NUTS) is a Markov Chain Monte Carlo (MCMC) sampling algorithm developed to overcome some limitations of the Hamiltonian Monte Carlo (HMC), especially when dealing with high-dimensional parameter spaces.

Unlike HMC, which requires a predefined number of time

steps, NUTS dynamically determines the length of the sampling path. This is made possible by introducing the concept of *U-Turns*. When the sampling path starts to return on itself, indicated by a *U-Turn*, NUTS terminates the sampling. This approach provides dynamic control over the path length, allowing for larger jumps in regions of the parameter space with higher density. A *U-Turn* occurs when the sample begins to turn back in the direction from which it came. In the presence of many *U-shaped curves*, the algorithm detects that the path has reached a turning point and concludes the sampling. This feature makes NUTS particularly efficient in high-dimensional spaces, where a longer path might be ineffective. NUTS is designed to autonomously terminate sampling when it detects the start of a *U-Turn*, thus avoiding unnecessary cycles. This feature is crucial to ensure that the algorithm is efficient and adaptable to different posterior distribution geometries.

In conclusion, NUTS overcomes traditional limitations of HMC by introducing autonomous termination based on the concept of *U-Turns*. This makes it a highly adaptable and efficient algorithm, particularly useful in Bayesian modeling contexts with

complex and high-dimensional parameter spaces.

Chapter 3

Bayesian model for football data

The hierarchical Bayesian approach to football predictions provides a powerful methodological framework that adapts well to the complexity of football. The use of these models offers numerous advantages that align well with the complex and dynamic nature of this sport. This claim can be supported by the excellent results achieved in previous studies, both in modeling football outcomes [Egidi, 2018] and in analyzing the individual performance of players [Egidi and Jonah, 2018]. First and foremost, in the world of football, characterized by significant variability in the skills and performances of teams over time, the imple-

mentation of a hierarchical Bayesian model provides a robust framework to capture this diversity. The hierarchy allows for the consideration of both common factors shared by all teams, such as general gameplay dynamics, and the allowance for each team to deviate from this global average. In other words, the model recognizes the complexity of team performances, offering a more accurate and detailed insight into their capabilities and variations.

The availability of historical results spanning multiple seasons is a treasure trove of information that can be effectively leveraged by a hierarchical Bayesian model. This approach allows the use of past team performances as informative priors to initialize the model parameters.

These capabilities of hierarchical models will be exploited for two different approaches, as mentioned in the introduction, each with specific purposes.

The first approach focuses on enhancing the interpretability of attack and defense parameters for teams, providing a generalization across seasons and models. In contrast, the second approach utilizes Gaussian Processes to incorporate temporal

dependency, allowing for a more robust and adaptable modeling of team performance evolution over time.

Therefore, the inherently Bayesian nature of the model allows for addressing and quantifying uncertainty in parameter estimates. This is particularly significant in the context of football predictions, where numerous factors can influence the outcome of a match. The ability to measure and clearly communicate uncertainty in parameters provides a more realistic and transparent perspective. This information can be crucial for analysts, coaches, and football enthusiasts as it offers a more comprehensive understanding of predictions and potential unpredictable developments in the game.

3.1 The motivating dataset

In the following section, we will explore the specifics of the dataset we have employed. Our modeling efforts aim to capture and portray the attack and defense capabilities of teams. Consequently, the core dataset comprises historical match results, serving as the foundation upon which our model is built.

The data have been extracted from the website https://www.football-

Season	Home	Away	Home Goals	Away Goals
2013	Hellas Verona	AC Milan	2	1
2013	Sampdoria	Juventus	0	1
2013	Inter	Genoa CFC	2	0
...	...	...	...	...
2021	Sassuolo Calcio	AC Milan	0	3
2021	Salernitana 1919	Udinese Calcio	0	4
2021	AC Venezia	Cagliari Calcio	0	0

Table 3.1: Serie A matches from 2013 to 2021

data.co.uk/. The selection covers a time span of 9 seasons, from 2013 to 2021.

The dataset (as in Table 3.1) comprises 5 columns: *Season, Home, Away, HomeGoals, AwayGoals*. These columns contain all the information related to match results for the 9 considered seasons. It is important to note that, given each involved league consists of 20 teams, which play each other twice in a football season, there are a total of 380 matches in a single season. Having considered multiple seasons (9), this figure needs to be multiplied by 9, resulting in a total of 3420 records.

For a better understanding of the data, is essential to consider that, over multiple seasons, some teams may be relegated or promoted from the lower league, leading to their variable presence in different years and generating less meaningful or incomparable

descriptive data. For purely descriptive purposes, we will filter the original dataset, keeping only the matches played in seasons where at least one of the two involved teams participated in the Serie A league. The considered teams in this context are 13: *Sampdoria, Inter, Lazio Roma, Torino FC, Juventus, Atalanta, Genoa CFC, SSC Napoli, Sassuolo Calcio, AS Roma, AC Milan, Udinese Calcio, ACF Fiorentina.* Applying this filter, the number of matches in the dataset has been reduced to 3042 (338 for each season).

Considering the model's objective, it is crucial to understand the data and get an idea of their distribution, distinguishing between seasons, teams, and the performances of home and away teams. This last point is particularly important because the hierarchical model, which will be examined later, estimates the goal-scoring intensity for each team in a match.

For the away team, the scoring intensity (*Away score intensity*) is positively influenced by its attacking strength and negatively influenced by the defensive strength of the home team. In contrast, for the home team, the scoring intensity (*Home score intensity*) is positively influenced by its attacking strength and

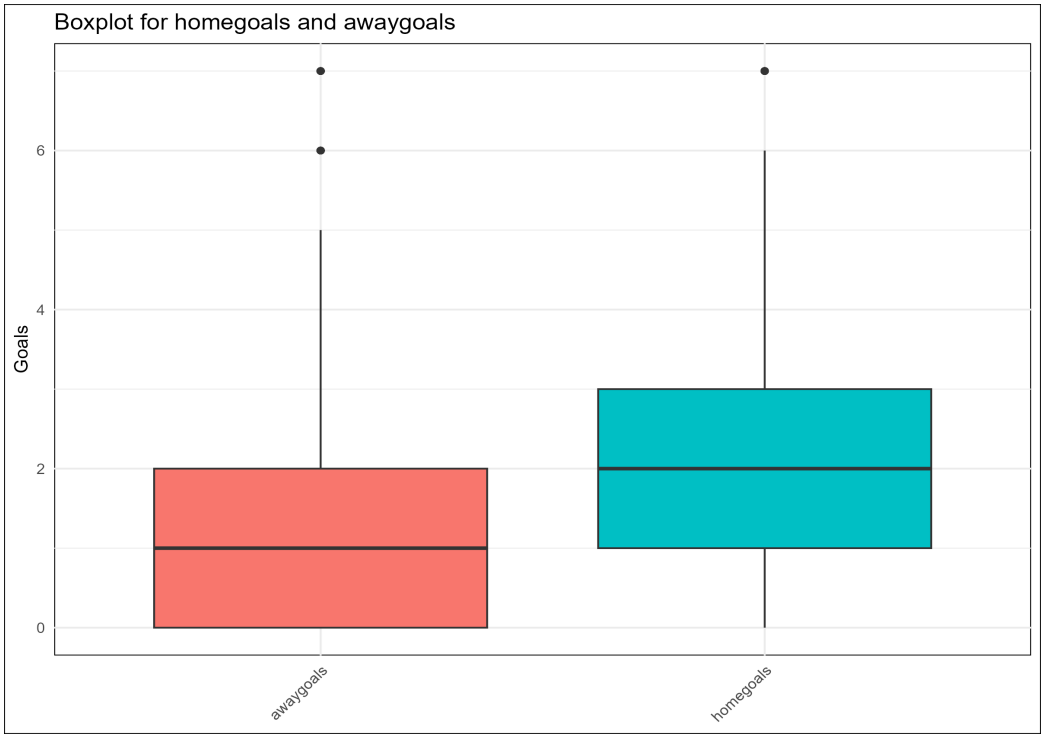


Figure 3.1: Box plot for Home and Away goal scored

negatively influenced by the defensive strength of the opposing team. Additionally, a positive factor defined as *home* in our formulas, contributes to the overall scoring intensity. This constant, recognized over the years as the *home advantage*, plays a role in enhancing the scoring intensity of the home team.

Thus it might be a good idea to try to describe our dataset and at the same time justify the choice of this parameter.

Initially, we observe a difference in the descriptive values of goals scored at home compared to those scored away (Fig-

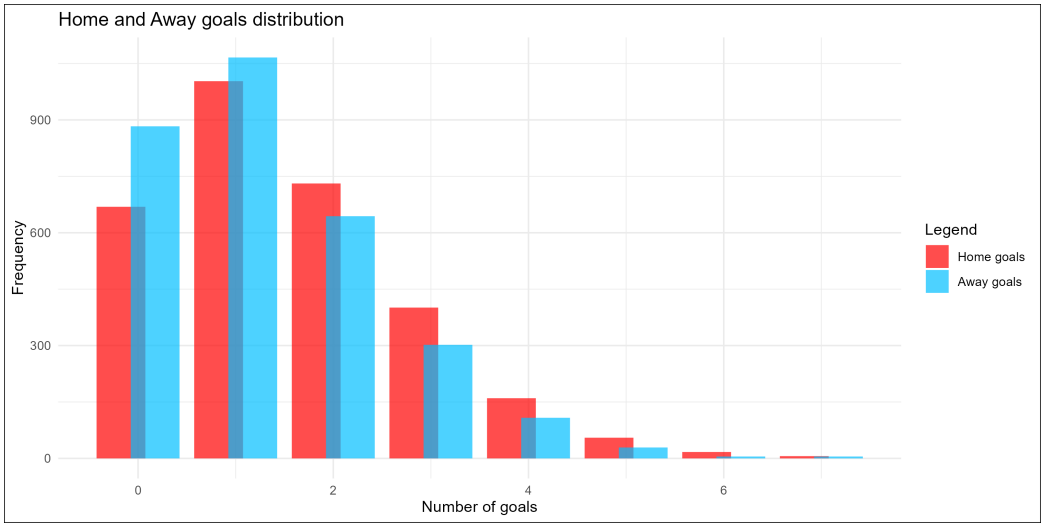


Figure 3.2: Bar plot for comparison of goals scored at home and goals scored away from home, grouped by team

ure 3.1). For the former, we find a *median* of 2, while for the latter, it is 1. This discrepancy, along with the two *means* (homegoals = 1.72, awaygoals = 1.44), indicates an initial slight distinction over the 9 seasons regarding the home advantage. However, relying solely on these data may not be sufficient to justify the home advantage.

Analyzing the distributions of goals scored at home and away by the considered teams, the graph in Figure 3.2 shows a constant gap between goals scored at home and those scored away. This gap is consistent, except for the bar plot corresponding to the value 0 (which indicates the frequency of no goals scored at

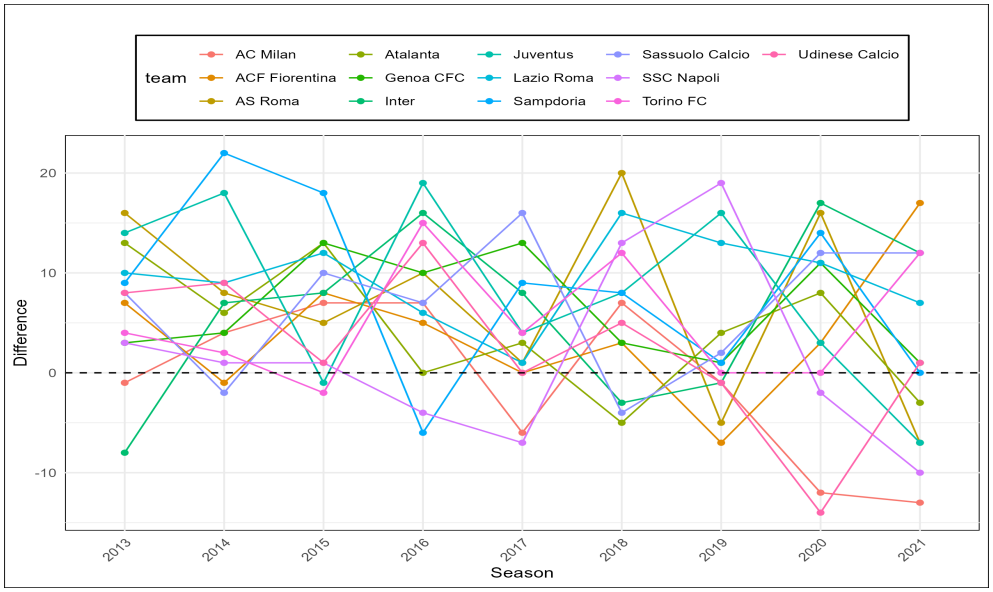


Figure 3.3: Time trend per season of the difference between goals scored at home and away, grouped by team

home), where it is higher when playing away, and the value 1 (a margin that will be compensated with bars representing higher values).

Continuing with the comparison of these two value distributions, an additional insight can be extracted from Figure 3.3. Considering the teams analyzed so far and aggregating the values per season, the graph illustrates the trend from 2013 to 2021 of the difference between goals scored at home and those scored away for each team in each season. The black dashed line indicates a difference of 0; therefore, observations below it represent

evidence against our home advantage thesis. However, in our favor, there is clear diversity in the density of observations above and below the 0 threshold. This suggests that over these years, goals scored at home by teams are generally higher than those scored away. More precisely, we have 26 observations below the threshold and 91 observations above.

Another analysis that could highlight the empirical influence of the home advantage is the comparison between the difference in goals scored and conceded when playing at home and the difference in goals scored and conceded when playing away (Figure 3.4).

The calculated difference signifies how representative the offensive performance is compared to the defensive performance for both home and away games. This difference easily translates into victories (if positive), draws (if zero), or defeats (if negative), with the black line indicating the sign of the difference. In the left graph, it is evident that over the 9 seasons considered, the thirteen teams consistently demonstrated significantly better offensive performance at home than defensive performance. The density of observations above the reference line is notably

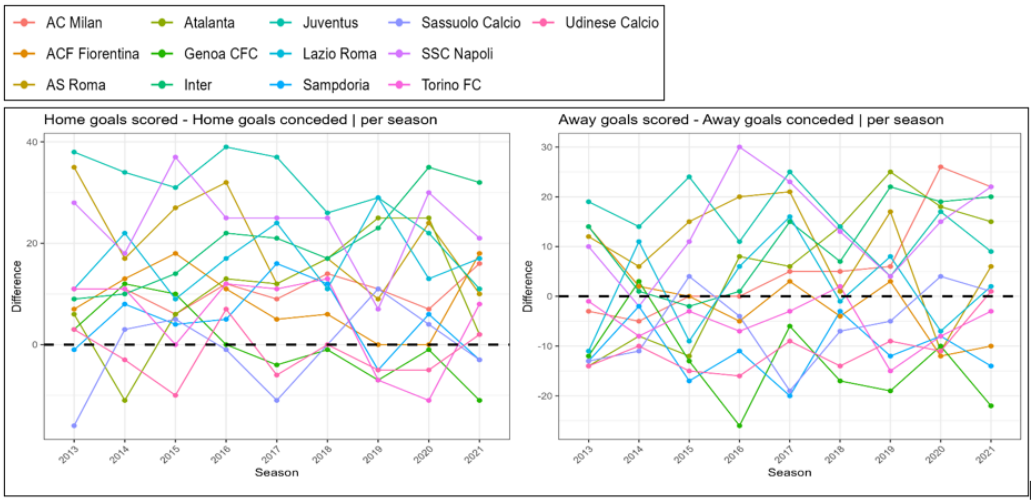


Figure 3.4: On left: difference between goal scored and conceded, grouped by home team, per season. On right: difference between goal scored and conceded, grouped by away team, per season.

higher (97) than below (20), indicating a substantial advantage when playing at home. Conversely, in the right graph, during the same period, the thirteen teams seem to respond almost indifferently to away matches. As highlighted by the graph, the densities appear to be fairly distributed relative to the reference line (62 above, 55 below). This suggests that playing away, in general, does not have a clear impact on the outcome.

3.2 Poisson likelihood specification

In the context of statistical modeling of goals scored in a soccer match, the Poisson distribution is often employed. This choice is not arbitrary but reflects the discrete and rare nature of the events we aim to represent. In this introduction, we will explore why the Poisson distribution is the appropriate choice for modeling the number of goals and analyze the advantages it offers in this specific context.

The Poisson distribution is a discrete probability distribution that models the number of rare events occurring in a fixed interval of time or space. In this specific context, it is used to model the number of goals scored by a team in a single soccer match.

The Poisson distribution is characterized by the probability mass function (PMF) given by the formula:

$$P(Y = k; \lambda) = \frac{e^{-\lambda} \cdot \lambda^k}{k!}$$

Where:

- Y is the discrete random variable representing the number of events (in our case, goals);

- k is the number of events we are counting (number of goals);
- λ is the parameter of the distribution (also known as rate), representing the average number of events in a fixed interval.

In the context of modeling goals scored by a team in a match, we can contextualize the Poisson-distributed formula as follows:

$$P(Y_{g,j} = y_{g,j}|\theta_{g,j}) = \frac{e^{-\theta_{g,j}} \cdot \theta_{g,j}^{y_{g,j}}}{y_{g,j}!}$$

Where:

- $y_{g,j}$ is the actual number of goals scored by team j in match g;
- $\theta_{g,j}$ is the parameter representing the scoring intensity of team j in match g (Poisson parameter).

Thus, the formula expresses the probability that team j scores exactly $y_{g,j}$ goals in a given match g, given the scoring intensity $\theta_{g,j}$.

More specifically, we assume two conditionally independent Poisson variables for the numbers of goals scored by the home team and away team.

We denote the number of goals scored by the j-th team ($j = h, a$) home team and the away team in the g-th match of the season ($g = 1, ..., G = 380$) as $y_{g,h}$ and $y_{g,a}$ respectively. The components of the observation vector $y = (y_{g,h}, y_{g,a})$ are modeled as independent Poisson distributions, i.e.,

$$y_{g,h}|\theta_{g,h} \sim \text{Poisson}(\theta_{g,h}), \qquad y_{g,a}|\theta_{g,a} \sim \text{Poisson}(\theta_{g,a}).$$

Once the distribution for the scores has been defined, it is now necessary to define those related to the parameters composing the scores equation. This aspect of the hierarchical model is among the most discussed in all the literature written on Bayesian hierarchical models for football outcomes. This is because assigning priors to the model variables involves specifying the shape of the probability distribution deemed most appropriate before observing the data. This specification can be based on prior knowledge, past experiences, or theoretical considerations, making it relatively subjective.

3.3 First approach: a hierarchical Bayesian model with Beta priors to enhance interpretability

The Beta distribution is a continuous probability distribution defined on the interval (0, 1). Its functional form is determined by two parameters, α and β. It is well-suited for representing the probability of success in a binary experiment, making it an intuitive choice for modeling football abilities, ultimately correlated with the probability of scoring or conceding goals. This approach introduces greater interpretability to the parameters, as they can be directly interpreted as a latent score confined to an interval.

Its probability density function is expressed as:

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

where $B(\alpha, \beta)$ is the Beta function, a normalizing constant ensuring that the integral of the probability density over the entire range is equal to 1.

Using a Beta distribution for the attack (att) and defense

(def) parameters implies that they take values between 0 and 1, providing a natural mechanism to model offensive and defensive abilities more restrictively than normal distributions with support on the entire real domain.

To model the rates $\theta_{g,h}$ and $\theta_{g,a}$ for the home team and the away team, respectively, we adopt a well-established log-linear random effect model. This approach, extensively discussed in statistical literature (see Karlis and Ntzoufras [2003]), formulates the logarithm of the rate parameters as follows:

$$\log(\theta_{g,h}) = home + att_{h[g]} - def_{a[g]}$$
$$\log(\theta_{g,a}) = att_{a[g]} - def_{h[g]}$$

$home \sim Beta(1, 1)$

The $home$, att and def parameters are then distributed as follows:

- $home \sim Beta(\alpha_{home}, \beta_{home})$
- $att_j \sim Beta(\alpha_{att}, \beta_{att})$
- $def_j \sim Beta(\alpha_{def}, \beta_{def})$

While discussing the interpretation of these parameters, for

the sake of conciseness, we suppress the index j in the parameter notation.

The *home* parameter represents the advantage for the home team (widely justified in the dataset introduction section), presumed to be the same across all teams and throughout the season. This parameter is assigned a prior distribution $home \sim Beta(1, 1)$, with both parameters equal to 1. When both parameters are 1, in practical terms, it represents a uniform distribution over [0, 1], where each point within this interval has an equal probability of being drawn. There is no preference or prior knowledge about the position of the probability within this interval.

Additionally, the score intensity depends on the combined attack and defense abilities of the teams, indicated by the *att* and *def* parameters, respectively. The nested indices $h[g]$ and $a[g]$ identify the home and away teams, respectively, in the g-th game of the season. In the expression $att, def \sim Beta(5, 5)$, we designate the shape parameters α and β as both being equal to 5. In this context, the distribution is centred at 0.5 and assumes a symmetrical bell shape. This distribution is characterized by

a relatively flat and symmetric profile centered around its mean of 0.5. This choice signifies a weakly informative, introducing a moderate level of uncertainty regarding the true value of the parameter. Consequently, this prior is not expected to exert a strong influence on the resulting posterior distribution.

About the Defense parameter

In scientific publications, it is common to observe that the effect of defense is included in the score intensity formula through addition. This approach is justified by assuming a normal distribution for the defense parameter, which can vary symmetrically around zero, including negative values. In this context, negative values of defense can be interpreted as a lower defensive ability of the team.

The term *att* continues to have a positive effect in the formulas. This reflects the concept that a team with a stronger attack will have a higher probability of scoring.

What changes in our model will therefore be the sign of the defense parameter, which entails logical changes of interpretation compared to traditional approaches.

The *def* component in the formulas now represents:

1. A penalty for the opponents' scoring ability if *def* tends to maximum values (close to 1). The following extreme cases will occur:

    - If both *home* and *att* tend to 1, $\log(\theta)$ will assume maximum values, tending to 1, so the score intensity $-\theta-$ will tend towards the mathematical constant e, which means 2,7 goals scored;
    - If both *home* and *att* tend to 0, $\log(\theta)$ will assume minimum values, tending to -1, so the score intensity $-\theta-$ will tend towards e^{-1}, which means approximately 0.35 goals scored;

2. An advantage for the opponents' scoring ability if *def* tends to minimum values (close to 0). The following extreme cases will occur:

    - If both *home* and *att* tend to 1, $\log(\theta)$ will assume maximum values, tending to 2, so the score intensity $-\theta-$ will tend towards e^2, which means approximately 7 goals scored;
    - If both *home* and *att* tend to 0, $\log(\theta)$ will assume min-

imum values, tending to 0, so the score intensity $-\theta-$ will tend towards e^0, which means 1 goal scored;

The above assessment is focused on extreme cases, but it does not exclude the assumption of intermediate results, where $\log(\theta)$ varies between -1 and 2, thus allowing the score intensity to take values between e^{-1} and e^2.

3.4 Second approach: Gaussian Processes for modeling time dependence

A crucial point in the following dissertation concerns the implementation of Gaussian Processes. As mentioned earlier, these processes define a temporal dependency among seasons by leveraging their structural characteristics.

We are discussing models that constitute a probabilistic framework for supervised machine learning, widely used for regression and classification tasks. It is a regression model based on processes that can make predictions incorporating prior knowledge (kernels) and provide uncertainty measures on predictions.

In the following section, we will define the fundamental fac-

tors of these models and elaborate on their structure [Wang, 2022].

The Multivariate Normal distribution

In the context of regression, the goal is to fit a function that best represents a set of observed data, enabling subsequent predictions on new data. However, when considering a set of observed data, it opens the door to an infinite number of possible functions that could fit those data points.

The advantage of using multivariate Gaussian distributions in Multivariate Gaussian Processes over univariate Gaussian distributions lies in their ability to capture correlations and dependencies among different random variables. In multivariate GPs, random variables are considered jointly, enabling the modeling of more complex relationships and capturing interactions across various dimensions of the data.

The Multivariate Gaussian distribution has a probability density function (PDF) defined as:

$$\phi_D(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

Here, D is the number of dimensions, x represents the variable, $\mu = E[X] \in \mathbb{R}^D$ is the mean vector, and $\Sigma = \text{cov}[x]$ is the $D \times D$ covariance matrix. The theory allows modeling complex relationships between variables, highlighting dependencies through Σ, a symmetric matrix representing covariances between variables.

The conditional probability $P(x_1|x_2)$ plays a crucial role in Multivariate Gaussian Processes during regression tasks. In the context of regression, the goal is to predict or model the probability distribution of a random variable given observations of other random variables.

The conditional distribution $P(x_1|x_2)$, in any context, is also Gaussian. Therefore, knowing the observation of x_2, one can obtain a predictive probability distribution for x_1 that takes into account the correlations between x_1 and x_2. In practical terms, this allows making predictions about the variable of interest (x_1) based on information provided by other variables (x_2), proving particularly useful in modeling complex relationships and multivariate dependencies.

Kernels

During the implementation of Gaussian Processes, a key role is covered by covariance functions, which reflect our ability to express prior knowledge about the shape of the function being modeled. In regression, the outputs of the function should be similar when two inputs are close to each other. One possible formulation of the equation is the dot product $A \cdot B = \|A\|\|B\| \cos\theta$, where θ is the angle between two input vectors. When two input vectors are similar, the output value of their dot product is high.

If a function is defined solely in terms of inner products in the input space and is positive definite, then the function $k(x, x_0)$ is a kernel function.

In particular, the kernel function we will use in our model is the Matérn covariance function with a smoothness parameter $v = 3/2$, known for its flexibility in modeling abrupt variations in data compared to other covariance functions like the squared exponential.

The Matérn covariance function with $v = 3/2$ is defined as:

$$k(x, x_0) = \sigma^2 \left(1 + \sqrt{3}\frac{|\mathbf{t}_1 - \mathbf{t}_2|}{\rho}\right) \exp\left(-\sqrt{3}\frac{|\mathbf{t}_1 - \mathbf{t}_2|}{\rho}\right)$$

where:

1. σ^2 represents the variance, indicating how much the data varies around its mean value.

2. r is the Euclidean distance between the two points x and x_0.

3. ρ represents he length scale parameter, which indicates the distance over which the correlation between observations significantly decreases.

This choice is often appropriate when modeling processes that may exhibit sudden changes or peaks in the data.

Adding covariance functions results in smoother lines that begin to resemble functions. It is natural to consider further increasing the dimension of Multivariate Normal distribution, where dimension refers to the number of its multiple variables. As the dimension increases, the region of interest will be filled with more points. When the dimension becomes infinite, there

will be a point to represent any possible input point.

3.4.1 Gaussian Processes

Before delving into the definition of Gaussian Processes, it's crucial to clarify the distinction between parametric and non-parametric models.

Parametric models assume that the data distribution can be represented by a finite set of parameters. In regression, these models seek to predict the function value $y = f(x)$ based on a specific x. For instance, in simple linear regression, parameters like θ_1 and θ_2 are identified to define the function $y = \theta_1 + \theta_2 x$. In cases where linear assumptions prove insufficient, more intricate models with additional parameters, such as $y = \theta_1 + \theta_2 x + \theta_3 x^2$, might be necessary. The training process involves utilizing a dataset with n observed points to establish the mapping of x to y through basis functions $f(x)$.

Non-parametric models, in contrast, involve an infinite number of parameters. These models dynamically adapt to increasing dataset sizes, offering enhanced flexibility. The relationship between the number of parameters and the dataset size classi-

fies a model as non-parametric, enabling better adaptability to intricate data patterns.

We can now precisely define Gaussian Processes as models that describe a probability distribution over possible functions fitting a set of points. This distribution encompasses all conceivable functions, enabling the calculation of means to represent the function and variances to indicate the confidence level of predictions. Let us delve into the details of how these processes operate.

The regression function modeled by a multivariate Gaussian is expressed as:

$$P(f|X) = N(f|\mu, K)$$

Here, $X = [x_1, \ldots, x_n]$, $f = [f(x_1), \ldots, f(x_n)]$, $\mu = [m(x_1), \ldots, m(x_n)]$, and $K_{ij} = k(x_i, x_j)$. In this context, X denotes the observed data points, m represents the mean function, and k is a positive definite kernel function. The covariance matrix K captures the relationships between different data points.

In the absence of observations, the mean function defaults to $m(X) = 0$, assuming that the data is often normalized to have a zero mean. The Gaussian processes model is a distribution

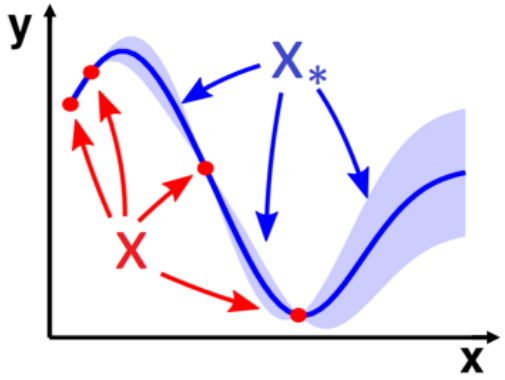


Figure 3.5: A illustrative process of conducting regressions by Gaussian processes. The red points are observed data, blue line represents the mean function estimated by the observed data points, and predictions will be made at new blue points [Wang, 2022].

over functions, and the shape or smoothness of these functions is defined by the kernel function K. If two points x_i and x_j are considered similar according to the kernel, the function outputs at these points, $f(x_i)$ and $f(x_j)$, are expected to be similar.

The process of performing regressions using Gaussian processes is illustrated in Figure 3.5. Given the observed data (depicted as red points) and a mean function f (depicted as a blue line) estimated from these observed data points, predictions are made at new points X_* denoted as $f(X_*)$.

The joint distribution of f and f_* is given by:

$$\begin{bmatrix} f \\ f_* \end{bmatrix} \sim N\left(\begin{bmatrix} m(X) \\ m(X_*) \end{bmatrix}, \begin{bmatrix} K & K_* \\ K_*^T & K_{**} \end{bmatrix}\right)$$

Where $K = K(X, X)$, $K_* = K(X, X_*)$, and $K_{**} = K(X_*, X_*)$. The mean $m(X), m(X_*) = 0$.

This is the equation for the joint probability distribution $P(f, f_*|X, X_*)$ over f and f_*, but regressions necessitate the conditional distribution $P(f_*|f, X, X_*)$ over f_* exclusively. The derivation from the joint distribution $P(f, f_*|X, X_*)$ to the conditional $P(f_*|f, X, X_*)$ is as follows.

In more realistic situations, we lack access to true function values but instead have noisy versions expressed as $y = f(x) + \epsilon$. Assuming the presence of additive independent and identically distributed (i.i.d.) Gaussian noise with variance σ_n^2, the prior on noisy observations becomes $\text{cov}(y) = K + \sigma_n^2 I$. The joint distribution of observed values and function values at new testing points becomes:

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K + \sigma_n^2 I & K_* \\ K_*^T & K_{**} \end{bmatrix}\right)$$

By deriving the conditional distribution, we obtain the pre-

dictive equations for Gaussian processes regression as:

$$\bar{f}_*|X, y, X_* \sim N(\bar{f}_*, \bar{K}_*)$$

In the variance function $\bar{K}_*$, it is noteworthy that the variance is independent of the observed output y but solely depends on the inputs X and X_*.

Gaussian Processes for football data

The implementation of Gaussian Processes in our case is used to introduce a temporal dependency to the model.

In the initial approach, by adopting a Beta distribution as a prior for attack and defense abilities, we achieved excellent interpretative results but limited practical use. This was because we lost the temporal information across seasons. Each played match was treated independently of when it occurred, resulting in an information loss that could be valuable for making predictions on unobserved seasons.

The goal is to introduce additional information to the attack and defense abilities provided by the model at time t, meaning *temporal dependency*.

This implies that the offensive and defensive abilities of individual teams provided by the model in the observed year are not simply based on a cumulative result from the dataset. Instead, each result will have a different weight based on the season to which it belongs. This allows us to highlight how the observed results in 2018, 2019, or 2020 will have much more impact on defining the teams' abilities in the last season considered in our dataset, i.e., 2021, compared to those observed in 2016, 2015, or earlier years.

3.4.2 Hierarchical model with Gaussian Process

In the context of Gaussian Processes, the model for scores changes as we introduce the indicator for seasons $t = 1...T$, which in our case ranges from 1 to 9. In detail:

$$y_{g,h}^{(t)}|\theta_{g,h}^{(t)} \sim Poisson(\theta_{g,h}^{(t)}), \qquad y_{g,a}^{(t)}|\theta_{g,a}^{(t)} \sim Poisson(\theta_{g,a}^{(t)}).$$

The distribution used to model the scores of individual teams is always a Poisson distribution.

The modeling of the logarithm of score intensities will un-

dergo modifications. In this case, depending on our objective, the attack and defense parameters will no longer depend solely on the team but also on the season in which the data is observed.

$$\log(\theta_{g,h}^{(t)}) = home + att_{h[g],t} - def_{a[g],t};$$
$$\log(\theta_{g,a}^{(t)}) = att_{a[g],t} - def_{h[g],t};$$

In the following model, the hierarchical structure is evidently more complex. The parameter home is defined with a normal distribution, i.e.:
$home \sim Normal(0, 5)$;

The priors assigned to attack and defense functions are defined as follows:

$$\boldsymbol{att}_j \sim \mathcal{GP}(\mathbf{0}, \boldsymbol{\Sigma}^{att});$$

$$\boldsymbol{def}_j \sim \mathcal{GP}(\mathbf{0}, \boldsymbol{\Sigma}^{def});$$

This suggests that, before observing the data, it is assumed that the attack and defense abilities of teams are normally distributed with a mean of zero and a variation defined by the associated covariance matrix, reflecting the relationship and vari-

ability among different components of team abilities.

In this model, compared to the first approach, it is necessary to impose a "zero sum" identifiability constraint for the parameters *att* and *def*. In practical terms, it is stipulated that the sum of all random effects for teams, both for offensive abilities (denoted as att) and defensive abilities (denoted as def), must be zero. This constraint helps avoid identifiability issues and ensures that the model has a unique solution.

$$\sum_{t=1}^{T} att_t = 0, \quad \sum_{t=1}^{T} def_t = 0, \quad for\ t = 1, \ldots, T.$$

The key factor in this model is the Matérn 3/2 covariance function, used to model the temporal dependency in the attack and defense abilities of teams. It is a common choice in non-parametric Bayesian models as it provides a balance between flexibility and smoothness. Analytically, the covariance functions for attack and defense abilities are defined as follows:

$$\mathbf{\Sigma}^{\text{att}}_{t_1,t_2} = \sigma^2_{\text{att}} \left(1 + \sqrt{3}\frac{|t_1 - t_2|}{\rho_{\text{att}}}\right) \exp\left(-\sqrt{3}\frac{|t_1 - t_2|}{\rho_{\text{att}}}\right);$$
$$\mathbf{\Sigma}^{\text{def}}_{t_1,t_2} = \sigma^2_{\text{def}} \left(1 + \sqrt{3}\frac{|t_1 - t_1|}{\rho_{\text{def}}}\right) \exp\left(-\sqrt{3}\frac{|t_1 - t_2|}{\rho_{\text{def}}}\right),$$

where $t_1, t_2 \in \{1, \ldots, T\}$. In defining the covariances between different seasons, the parameters of the kernel function, σ_q^2 and ρ_q, play a crucial role, where $q \in \{att, def\}$.

1. σ_q^2 represents the standard deviations of Gaussian functions in the process and can be interpreted as a measure of how much the attack and defense abilities vary around their mean values. In the context of the Matern 3/2 kernel function, this parameter influences the "smoothness" of the function, determining how quickly the covariance decreases with temporal distance. In the model, it is defined as $\sigma_q^2 \sim Gamma(2, 0.1)$.

2. ρ_q indicates the characteristic spatial or temporal scale of the process. Essentially, it describes the characteristic length over which the team abilities are correlated over time. A larger value of this parameter implies a greater correla-

tion between abilities of seasons temporally distant, while a smaller value suggests greater short-term variability. In the model, it is defined as $\rho_q \sim Gamma(2, 0.1)$.

The distribution of these parameters through Gamma priors reflects the uncertainty associated with our prior knowledge, allowing the model to learn coherently from available observations.

Chapter 4

Application and results

In the this section, the results obtained by the models are discussed. We showcase both models, allowing us to highlight the pros and cons of each approach, comparing them side by side.

4.1 Home parameter

In the descriptive analysis of the dataset, we delved into the investigation ofthe consistency of defining a constant advantage affecting the score intensity for home-playing teams. This element is identified by the variable *home* in the model. In Figure 4.1, we display the boxplot of the posterior distribution of the home advantage estimated in both models. Contextualiz-

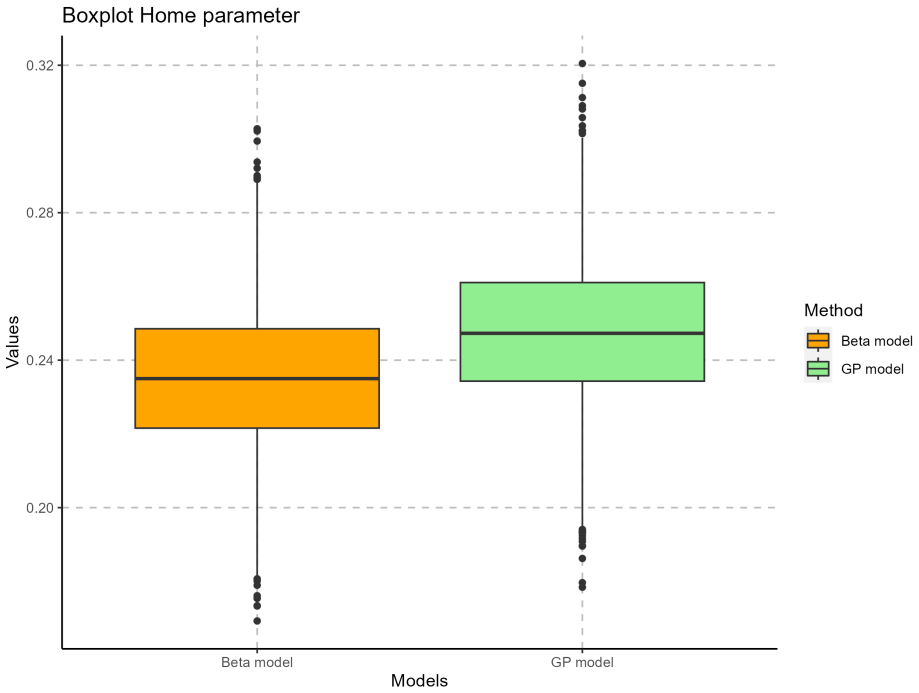


Figure 4.1: Boxplots of the posterior distributions of the home advantage for the Beta model and GPs model.

ing this aspect to reality and maintaining coherence with what emerged in the descriptive analysis, both models indicate that the advantage of playing at home translates into an initial positive margin of approximately 0.25 on the $\log(\theta)$, so θ is equal to $e^{0.25}$, which is approximately 1.28. In other words, we could interpret this advantage as a starting point with a margin of about one goal for the home team compared to the visiting teams.

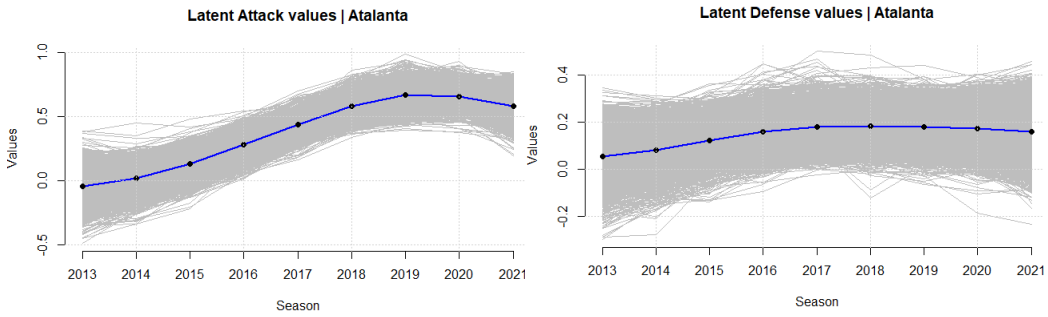
4.2 Attack and Defense functions

Both models enable the modeling of teams' offensive and defensive abilities, considering matches played during the period 2013-2021. The distinction between the two models lies in the handling of the time factor.

On the one hand, in the first approach (Beta model), each match contributes uniformly to the final values of teams' abilities, as previously mentioned. In other words, permuting the matches in the dataset would leave the final results unchanged compared to maintaining the chronological order of the dataset. The model will return a single value for the offensive ability and one for the defensive ability of each team, obtained from the mean of the 4000 iterations performed. Consequently, there are no temporal variations in the values, but rather a single retrospective observation, guided by the results of the nine seasons and the prior knowledge defined by the Beta distributions.

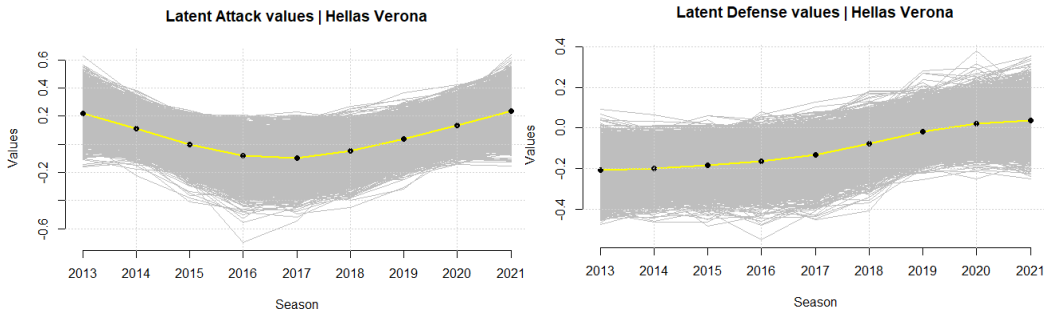
On the other hand, the second approach (Gaussian Process model) assigns a different weight to each match based on the belonging season, defined by the covariance function (Matérn). This temporal dependence allows for observing values that de-

velop along a time series, represented by the nine seasons. For each season, it is possible to observe a representative average of the defensive and offensive abilities of each team. These are calculated as the average of the values provided by the 4000 iterations performed by the model to determine the final parameters, represented by faint gray lines visible in the background in the graphs below.



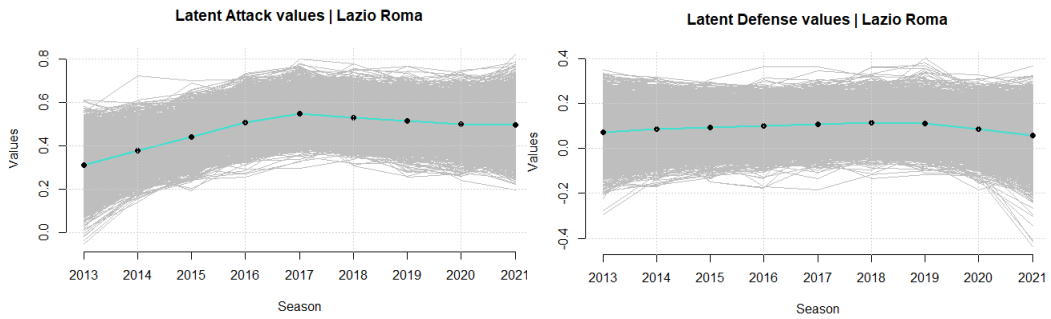
(a) Atalanta latent Attack values (b) Atalanta latent Defense values

Figure 4.2



(a) Hellas Verona latent Attack values (b) Hellas Verona latent Defense values

Figure 4.3



(a) Lazio latent Attack values (b) Lazio latent Defense values

Figure 4.4

To demonstrate the obtained results, we will focus on three clubs primarily: Atalanta, Hellas Verona, and Lazio.

In Figure 4.2, we present the evolution over time of Atalanta's attack and defense parameters. The continuity provided by Gaussian Processes makes evident the growth of the team's offensive abilities over the nine seasons, reflecting reality as it mirrors the team's development over time under the guidance of its coach, who has ensured continuity to a multi-year project. In particular, a significant improvement is highlighted in recent years around 2021, with the team equaling and surpassing several records related to goals scored in a single season, as well as in individual matches. Defensive performances also ensure growth over the years, certainly less pronounced but still reassuring.

In Figure 4.3, we report the evolution over time of Hellas Verona's attack and defense parameters. In the seasons after 2013, the team achieved good results but then was relegated twice in the 2015 and 2017 seasons. Surprisingly, from 2019 to 2021, it achieved exciting results, ranking among the top for goals scored and showing good defensive performance. This history is evident in both graphs provided by the model: offensive abilities have a "U" shape, highlighting how in the intermediate period, negative values of offensive performance are reached before rebounding, while defensive abilities tend to grow, improving defensive skills, although not reaching high absolute values.

Finally, in Figure 4.4, we present the evolution over time of Lazio's attack and defense parameters. In this case, the graphs seem relatively constant over the years, with offensive ability slightly increasing, assuming relatively high values from the beginning to the end. Stability is also evident in defensive performance, with a slight decline in recent years around 2021.

4.3 The real advantages of the two models

Now let us analyze the tangible advantages that the implemented models bring to the modeling of football outcomes with the Bayesian approach.

4.3.1 When time matters: Gaussian Process

As highlighted in the previous section, the implementation of Gaussian Processes enables us to have a temporal function describing the development of teams' defensive and offensive abilities across seasons.

This advantage, in addition to the previously discussed graphs, becomes evident through the comparison between Figure 4.5(a) and Figure 4.5(b).

In these graphs, we depict the posterior parameter values (att and def) obtained from both models (with Gaussian Processes and with Beta as the prior distribution) on a Cartesian axis, showcasing *only* the teams that participated in the Serie A championship in 2021, the most recent season available in our dataset.

Firstly, in both models, we apply a clustering algorithm known

as K-means to distinguish teams' performances into 4 classes: HIGH, MID-HIGH, MID-LOW, LOW. This algorithm enables the generation of team clusters by minimizing variance within each cluster and maximizing variance between them. A crucial role in this algorithm is played by the parameter K, specified by the user, defining the desired number of clusters in the output.

Furthermore, in the following graphs, in addition to estimating the parameters, indicators of standard deviation for each observation have been incorporated concerning the coordinates (i.e., the values of attack and defense for each team).

This classification, along with contextualizing the models' results to reality, allows for a better understanding of the influence of the time factor on the model.

As seen earlier, two significant cases highlighting a notable difference in the graphs are Atalanta, a team consistently improving offensively in recent years, and Hellas Verona, particularly intriguing for its oscillating performance over the years.

In the graphs, the region where Atalanta and Lazio concentrate their values is highlighted. In the Beta model, Lazio is considered more offensive, thanks to a cumulative better perfor-

mance over the nine seasons compared to Atalanta.

Conversely, in the model with Gaussian Processes, Atalanta's offensive performance is superior to Lazio's in the highlighted region. This is because, although there are fewer seasons in which Atalanta performed better than Lazio, these have a greater weight on the final result, being closer to the target season 2021.

Another relevant example is the situation of Hellas Verona. This trend is well represented in the two graphs: in the model without temporal dependence, Hellas Verona is placed in the cluster of least performing teams (in blue), while in the model with temporal dependence, it is considered the best among the mid-low performance cluster (in red). This happens because the weight of the results obtained during the 2019 and 2020 seasons is higher than the previous ones, more strongly influencing the estimation of parameters in the model with Gaussian processes.

Thus, the advantage of the time factor emerges clearly for practical purposes, as modeling football performances, like many other phenomena, is highly dependent on time.

4.3.2 When interpretation matters: Beta as prior distribution

Up until now, we have employed Figure 4.5 to underscore the significance of temporal dependence. In our initial approach, utilizing the Beta distribution in static model, our emphasis was on interpretability. As observed in the Beta model graph, the domain of skill values ranges from 0 to 1, which is a crucial aspect within the context of a Bayesian hierarchical model and should not be underestimated.

The capability to generate values within a specific range plays a pivotal role, allowing for a more precise quantification and interpretation of the obtained results. This becomes particularly significant when contrasting outcomes across different seasons or when dealing with Beta distribution parameters set a priori.

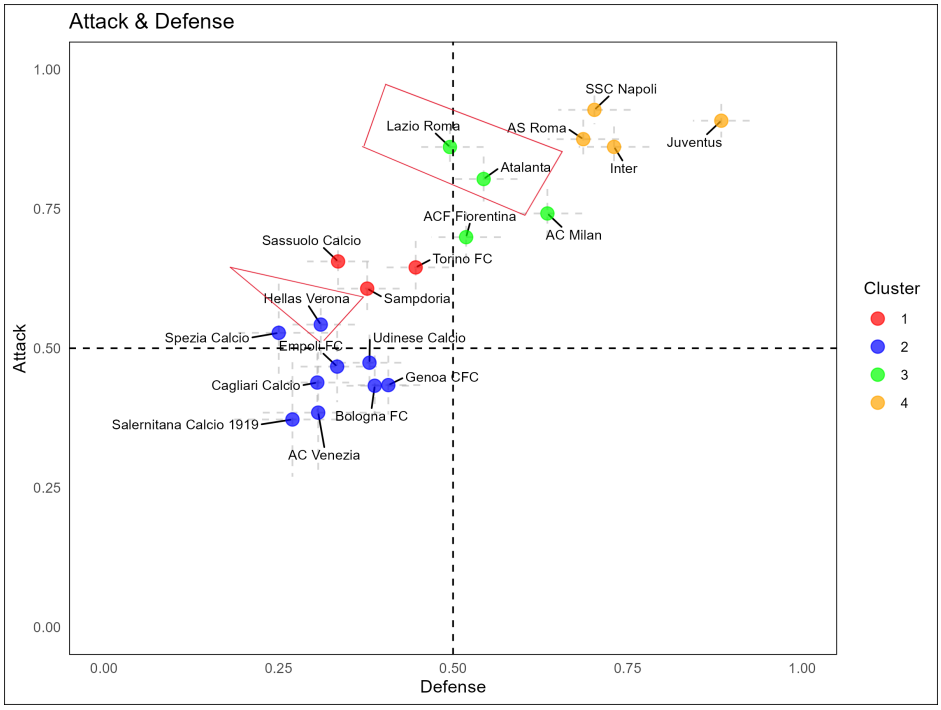
Additionally, the presentation of results in Table 4.1 further elucidates the advantages of defining a specific range of values between 0 and 1. This approach enables the observation of parameters within a single quadrant of the Cartesian graph. This, in turn, offers a notable advantage – the ability to calculate the distance from the origin, providing an immediate interpretation

1.	Juventus	1.2674727	11.	Sampdoria	0.7145239
2.	SSC Napoli	1.1636054	12.	Hellas Verona	0.6250260
3.	Inter	1.1293536	13.	Udinese Calcio	0.6076963
4.	AS Roma	1.1121952	14.	Genoa CFC	0.5950845
5.	Lazio Roma	0.9934138	15.	Spezia Calcio	0.5840315
6.	AC Milan	0.9765069	16.	Bologna FC	0.5810630
7.	Atalanta	0.9702849	17.	Empoli FC	0.5742004
8.	ACF Fiorentina	0.8706616	18.	Cagliari Calcio	0.5344266
9.	Torino FC	0.7844427	19.	AC Venezia	0.4918978
10.	Sassuolo Calcio	0.7364378	20.	Salernitana Calcio	0.4597847

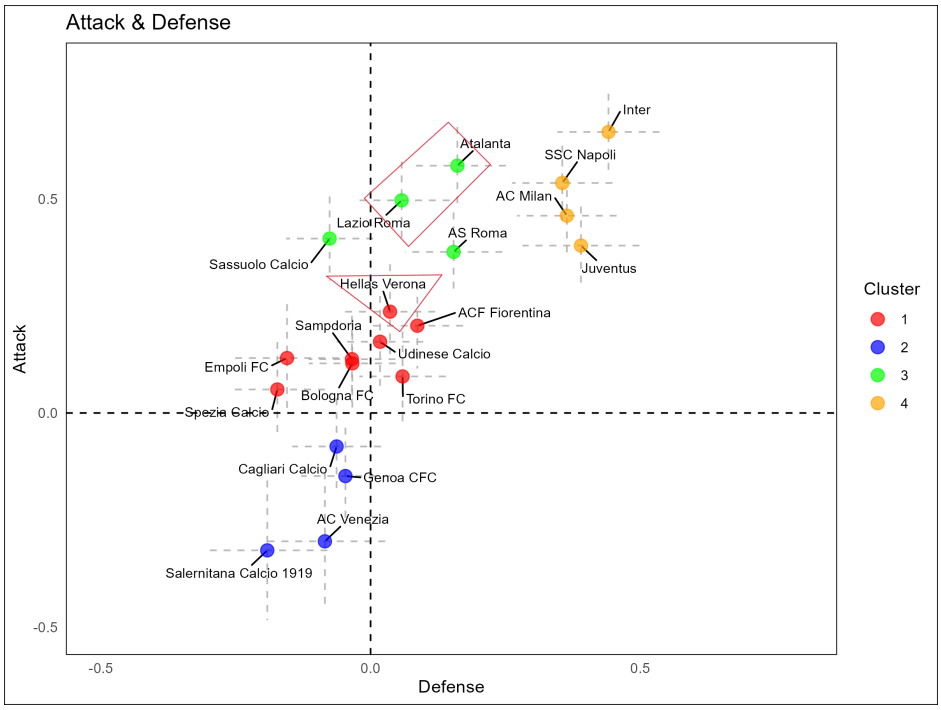
Table 4.1: Ranking provided by the computed distances from the origin, based on the values in Figure 4.5(a)

of the ranking, as illustrated in table.

Hence, the importance of the model *without* Gaussian Processes cannot be overstated, as its flexibility facilitates an in-depth study and empirical research to achieve optimal results in shorter time frames.



(a) Latent posterior abilities values of the teams participating in Serie A 2021, with a *Beta model*



(b) Latent posterior abilities values of the teams participating in Serie A 2021, with *Gaussian Process*.

Figure 4.5

Chapter 5

Discussion and further work

It is evident how in the context of football, there are numerous statistical-mathematical approaches to model, analyze, and attempt to predict results, injuries, and specific performances. This is a field in constant growth, where each analyst adopts a unique approach, structuring their own model in an original way.

The uniqueness of this work lies in the adoption of two distinct approaches, highlighting the pros and cons of each. In the first approach, a Bayesian hierarchical model was implemented using a Beta distribution for the parameters of the offensive and defensive abilities of the teams. The clear advantage of this model is its immediate interpretability and the possibility to

define a weighted ranking for both parameters (attack and defense) through a simple calculation of distances from the origin, thanks to the parameter constraints between 0 and 1 of the Beta distribution.

However, the drawback of this model lies in its neglect of the time factor. In this approach, each match has the same weight regardless of when it took place, leading to results that are not always accurate, especially in a context like football, where temporal dependence is crucial.

The weakness of the first approach becomes the strength of the second. In the second approach, a Gaussian Process was implemented in a Bayesian hierarchical model, introducing a temporal dependence on the seasons. Each match has a different weight on the target season (specifically, 2021), allowing for the weighting of the final results of the parameters based on temporal distance.

The drawback of this second model, compared to the first, is the lack of immediate interpretability. The distributions used are not defined in a limited domain, so there is not immediate clarity in understanding and interpreting the results.

A possible extension to combine the two models could be the adoption of a Probit model. This would represent an approach of merging Bayesian models, using Probit as a linking function between the latent parameters of the two models.

In conclusion, this study has highlighted the pros and cons of the adopted approaches. Both can be considered as solid foundations to improve the model, focusing on the quality of predicting team scores in future matches.

Bibliography

E. Pollard. A method for assessing changes in the abundance of butterflies. *Biological Conservation*, 12(2):115–134, 1977. ISSN 0006-3207. doi: https://doi.org/10.1016/0006-3207(77)90065-9. URL `https://www.sciencedirect.com/science/article/pii/0006320777900659`. 6

M. Maher. Modelling association football scores. *STATNED*, 1982. 6

Mark J. Dixon and Stuart G. Coles. Title of the article. *Journal Name*, 1997. 6

Dimitris Karlis and Ioannis Ntzoufras. Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 52(3):381–93, 2003. URL

http://www.jstor.org/stable/4128211. Accessed 23 Jan. 2024. 6, 47

Stan Development Team. Stan modeling language users guide and reference manual, 2.34. 2024. URL https://mc-stan.org. 29

Leonardo Egidi. Developments in bayesian hierarchical models and prior specification with application to analysis of soccer data. 2018. 33

Egidi and Jonah. Bayesian hierarchical models for predicting individual performance in soccer. *Journal of Quantitative Analysis in Sports*, 14(3):143–157, 2018. doi: 10.1515/jqas-2017-0066. 33

Jie Wang. An intuitive tutorial to gaussian processes regression. 2022. 52, 58