



Dixon and Coles Negative Binomial Model for Football

By

Panagiotis Lazarou

A THESIS

Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece

July 2025







Dixon και Coles Μοντέλο Αρνητικής Διωνυμικής Κατανομής για το Ποδόσφαιρο

Παναγιώτης Λαζάρου

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Διπλώματος Μεταπτυχιακών Σπουδών στη Στατιστική

Αθήνα

Ιούλιος 2025





ACKNOWLEDGEMENTS

I would like to thank my parents for their constant and unconditional support. They have stood beside me in whatever I wished to pursue and that is what they did throughout the whole duration of the MSc Program. I would also like to express my sincere appreciation and gratitude to my supervisor Prof. Dimitris Karlis for his guidance. His great academic work has been truly inspiring. Most importantly I would like to thank him for his patience and complete understanding of my attempt to combine my job along with the completion of this thesis. I would also like to mention that I feel thankful to all the Professors of the MSc Program that I cooperated with.

Finally, I would like to thank my dear friend Evie. Her presence has given me the power to keep going.





ABSTRACT

Panagiotis Lazarou

Dixon and Coles Negative Binomial Model for Football

July 2025

This thesis explores extensions of the widely used Dixon and Coles model for football match outcomes. The original Dixon and Coles formulation relies on Poisson marginals and restricts dependence adjustments to a limited set of scorelines. This work addresses those limitations by introducing the use of the Sarmanov family of distributions. This flexible framework allows for marginals other than Poisson, in this case negative binomial, and provides the ability to shift probability across a greater set of scorelines with the use of various mixing functions (called q -functions). The proposed model is theoretically presented, estimated, and applied to real-world data from major European football leagues, including the Premier League, La Liga, Bundesliga, and Eredivisie. Comparative analysis demonstrates that negative binomial models perform similarly with Poisson models and in some leagues, they even outperform them. Finally, the negative binomial models show promising predictive performance, being able to predict the results of teams in the final ten matchdays relatively accurately. This work aims to provide one more tool to statisticians, analysts, fans and generally everyone involved with football who would like to analyze, explore and predict the results of the worlds' most beloved sport.





ΠΕΡΙΛΗΨΗ

Παναγιώτης Λαζάρου

Dixon και Coles Μοντέλο Αρνητικής Διωνυμικής Κατανομής για το Ποδόσφαιρο

Ιούλιος 2025

Αυτή η διατριβή εξετάζει επεκτάσεις του ευρέως χρησιμοποιούμενου μοντέλου των Dixon και Coles για τα αποτελέσματα ποδοσφαιρικών αγώνων. Η αρχική διατύπωση των Dixon και Coles βασίζεται σε περιθώριες κατανομές Πουασόν και περιορίζει τις τροποποιήσεις της δομής εξάρτησης σε ένα περιορισμένο σύνολο σκορ. Η παρούσα εργασία αντιμετωπίζει αυτούς τους περιορισμούς μέσω της χρήσης της οικογένειας κατανομών Sarmanov. Αυτό το ευέλικτο πλαίσιο επιτρέπει τη χρήση περιθωρίων κατανομών διαφορετικών της Πουασόν, σε αυτή την περίπτωση αρνητικής διωνυμικής, και παρέχει τη δυνατότητα μεταφοράς πιθανότητας σε μεγαλύτερο σύνολο σκορ με τη χρήση διάφορων συναρτήσεων ανάμειξης (γνωστών ως συναρτήσεις g). Το προτεινόμενο μοντέλο παρουσιάζεται θεωρητικά, εκτιμάται και εφαρμόζεται σε πραγματικά δεδομένα από σημαντικά ευρωπαϊκά ποδοσφαιρικά πρωταθλήματα, όπως η Premier League, η La Liga, η Bundesliga και η Eredivisie. Η συγκριτική ανάλυση δείχνει ότι τα μοντέλα με αρνητική διωνυμική κατανομή αποδίδουν παρόμοια με τα μοντέλα Πουασόν και σε ορισμένα πρωταθλήματα τα ξεπερνούν. Τέλος, τα μοντέλα με αρνητική διωνυμική κατανομή παρουσιάζουν ελπιδοφόρες προβλεπτικές επιδόσεις, επιτυγχάνοντας σχετικά ακριβείς προβλέψεις για τα αποτελέσματα των ομάδων στις δέκα τελευταίες αγωνιστικές. Η εργασία αυτή φιλοδοξεί να προσφέρει ένα ακόμη εργαλείο σε στατιστικούς, αναλυτές, φιλάθλους και γενικά σε όλους όσους ασχολούνται με το ποδόσφαιρο και επιθυμούν να αναλύσουν, να εξερευνήσουν και να προβλέψουν τα αποτελέσματα του πιο αγαπημένου αθλήματος στον κόσμο.

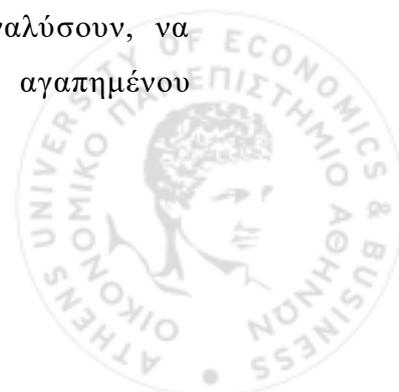




TABLE OF CONTENTS

1. Introduction	1
2. Literature Review.....	5
2.1 Basic Concepts	5
2.2 Football Results Modelling Through The Years	6
3. Negative Binomial Marginals Model	23
4. Application	29
4.1 Data	29
4.2 Estimation	33
4.3 Model Comparison	36
4.4 Prediction	41
5. Conclusions – Further Research	47
References	50





LIST OF FIGURES

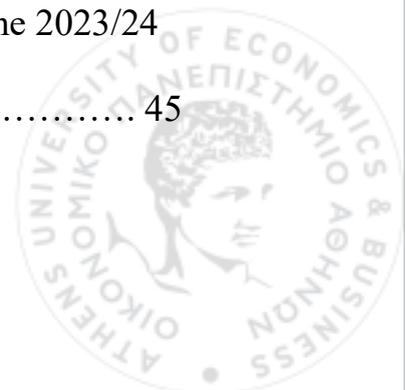
Figure 1 – Probability distributions for different q -functions.	26
Figure 2 – Correlation for different values of ω and different q -functions.	27
Figure 3 – Empirical correlation values.	29
Figure 4 – Mean and variance of goals scored by each team.	33
Figure 5 – Prediction under the q_1 – neg. binom. model.	42
Figure 6 – Prediction under the q_2 – neg. binom. model.	43
Figure 7 – Prediction under the q_4 – neg. binom. model.	44





LIST OF TABLES

Table 1 – Ratios for Premier League.	30
Table 2 – Ratios for La Liga.	31
Table 3 – Ratios for Eredivisie.	31
Table 4 – Ratios for Bundesliga.	31
Table 5 – Contingency table of Eredivisie data.	32
Table 6 – Attack and defence parameters of 2023/24 Eredivisie teams.	35
Table 7 – Rest model parameter values.	35
Table 8 – Parameter ω values for different q-functions and different data sets.	36
Table 9 – Details of the fitted models for the Eredivisie data.	38
Table 10 – Details of the fitted models for the Bundesliga data.	39
Table 11 – Details of the fitted models for the La Liga data.	39
Table 12 – Details of the fitted models for the Premier League data. ...	40
Table 13 - MAE and RMSE values for the fitted models for the 2023/24 Bundesliga data.	45





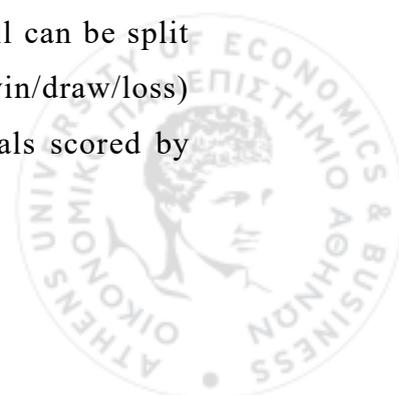
1.Introduction

Football has been labeled as the “king of sports”. Its’ ever-increasing popularity is the main reason behind the title that captivates the worldwide frenzy that fills the fans with emotion and leads to billions of euros being invested every year in different aspects of the game, like astonishing transfer fees for players, huge sponsorships for teams and players from the wealthiest companies, television rights in almost every country in the world etc. However, the world’s most beloved sport is, in its’ core, a simple game that can be played by anyone. No cultural, economic or demographic barrier can prevent people from playing or watching football.

Although football’s apparent simplicity it is a game that is really hard to predict. The media often refer to it as a game of “chaos”. Being a low scoring contest means that its result is vulnerable to small changes. One moment is proved sometimes the turning point where a player becomes the hero that scores a last-minute winning goal or misses the chance for just centimeters. An early red card, player injuries and substitutions are just some examples from the big list of decisive events. These small margins mean that football can only be treated with a probabilistic way of thinking.

Statisticians and mathematicians have developed over the years some methods that try to explain football’s uncertainty and have achieved reasonable accuracy. However, there is always room for improvement and the effort for the best possible result is almost never ending. Especially in the last few years when betting became a notable part of football’s ecosystem, generating huge amounts of revenue, the need for well-informed predictions is growing. Bet players are looking to beat bookmakers with the “best” statistical model that they can build their strategy around and betting companies are always trying to maximize their margin of profit.

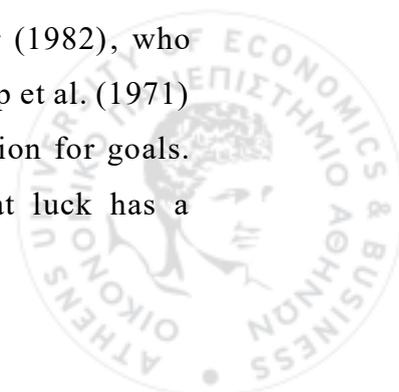
In academic literature, different approaches have been studied. According to McHale and Scarf (2007) the models about football can be split into direct and indirect ones. The first analyze match results (win/draw/loss) while indirect models assign probabilities to the number of goals scored by each team, thus occur the probabilities of the outcome.



Goals are the deciding factor of football matches. Modelling them is valuable for understanding and predicting the outcomes of matches, which has applications in sports analytics, betting markets, team strategy, and fan engagement. Accurate goal models help quantify team performance, assess player contributions, and evaluate tactical decisions. In competitive environments like professional leagues, these models aid in predicting league standings, optimizing player transfers, and assessing matchups against specific opponents. For betting markets, precise predictions improve odds-setting and risk management. Moreover, goal modelling supports academic research into the statistical properties of football matches, providing insights into scoring patterns, team dependencies, and home advantage effects. As the sport grows globally, modelling goals offers a pathway to better analytics and understanding of the game's dynamics.

When modeling the number of goals in football, two key considerations emerge. First, it is necessary to choose an appropriate marginal distribution for the number of goals, with the Poisson distribution being a common standard that is used in many instances throughout the academic literature (e.g. Maher (1982), Lee (1997), Dixon and Coles (1997), Karlis and Ntzoufras (2003)). Recently, researchers also suggested negative binomial marginals (McHale and Scarf (2007) and McHale and Scarf (2011)) and even Weibull count distribution marginals (Boshnakov et al. (2017)). Second, as the two opposing teams interact during the game, capturing the correlation between the goals of the two teams is crucial. This has been expressed with various methods, ranging from independence (no correlation) as in Maher (1982) and Lee (1997), correlation through a modification term (Dixon and Coles (1997)), through the model's formulation (Karlis and Ntzoufras (2003)), through copulas (McHale and Scarf (2007) and McHale and Scarf (2011)) and recently with flexible mixing functions that introduce dependence between the marginals (Michels, Ötting and Karlis (2023)).

The first thorough attempt to model goals was by Maher (1982), who drew inspiration from previous works of Moroney (1956) and Reep et al. (1971) that mainly examined the use of the negative binomial distribution for goals. Actually, the authors of the second publication concluded that luck has a

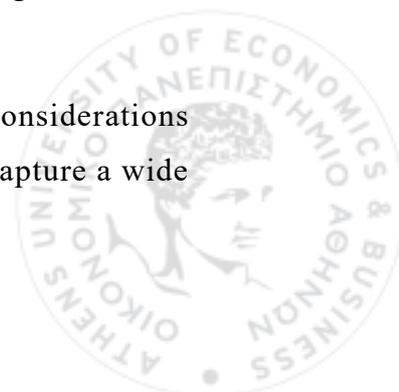


significant impact on football results and on results of other ball games. However, Hill (1974) proved that experts could predict final league rankings with relative accuracy. Thus, Maher modelled match scores using an independent double Poisson model that considered the attacking and defensive abilities of teams. In this way it is possible to incorporate the knowledge for a team's quality in the predictions for a particular match.

Dixon and Coles (1997) suggested that there is some dependence between the goals of the two competing teams. The authors found that the independence assumption is sensible for the data that they considered except for scores 0-0, 1-0, 0-1 and 1-1. In particular, they note that scores 0-0 and 1-1 are overestimated by the independence model while 1-0 and 0-1 are underestimated. Thus, they proposed a modification term that moves probabilities between those scores. This model has gained general acceptance and is widely used.

Karlis and Ntzoufras (2003) used the bivariate Poisson model and modified it to inflate probabilities of draws. This improved the fit on the otherwise underestimated count of draws. However, McHale and Scarf (2007) pointed out that both previously mentioned models allow only non-negative dependence. In an effort to model shots, which exhibit negative dependence in contrast to goals, the authors proposed the use of copula functions. These functions allow for negative dependence. Furthermore, McHale and Scarf (2011) noted that negative dependence is sensible for domestic leagues where teams of similar abilities face each other, while for international matches where teams with large differences in abilities play more often against one another the goals seem negatively correlated. In addition, the bivariate Poisson does not capture the overdispersion of goals data, so they proposed the use of a bivariate model that allows the use of marginals other than Poisson. To achieve this, they proposed again the use of copula functions, which are flexible enough to capture negative dependence and allow for negative binomial marginals which capture better the overdispersion characteristics.

Considering all the previously mentioned facts two main considerations emerge when modelling goals data: the model should be able to capture a wide



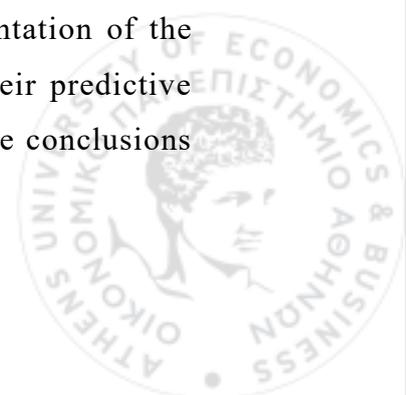
range of dependence between home and away goals, both positive and negative and also account for the overdispersion that some datasets may present.

Despite the wide use of the Dixon and Coles model it has two important limitations. First, one cannot shift probabilities to scores other than 0-0, 1-1, 1-0 and 0-1. Second, the use of a marginal distribution other than Poisson is not possible. To counter this, Michels, Ötting and Karlis (2023) provided an extension of the Dixon and Coles model. In their paper that focuses on women's football scores they use the properties of the Sarmanov family of distributions (Sarmanov (1966)) to provide flexibility on probability shifting (meaning that probabilities can be moved beyond 0-0, 1-1, 1-0 and 0-1). Additionally, these properties allow the use of marginal distributions other than Poisson and can capture a wider range of correlation. With the use of appropriate mixing functions (called q -functions) they examined model formulations that allowed to move score probabilities in different ways and to use Poisson and negative binomial marginal distributions.

This thesis aims to examine the models that were presented by Michels, Ötting and Karlis. These models provide the necessary flexibility to use any marginal distribution for goals and the ability to capture a wider range of dependence. This happens while moving probability among scorelines which is an important aspect that helps the fit to the actual data.

Throughout this thesis, these models are going to be presented in detail. Their properties are going to be derived and explained. Finally, the models will be applied to actual data from important European football leagues.

The structure of the thesis is the following: Section 2.1 provides some basic notions that are required for better understanding of what is presented in the rest of the thesis. Section 2.2 includes a review of models developed for goals in academic literature. Section 3 presents the theoretical framework of the models and their important characteristics and properties. Section 4 includes the application of the models to real data, with presentation of the data, model estimates, model comparisons and evaluation of their predictive capabilities. Finally, Section 5 provides a discussion where some conclusions are drawn and some thoughts about future work are provided.



2. Literature Review

2.1 Basic Concepts

Football modelling relies on a variety of statistical distributions to represent the underlying patterns of events occurring within a match. These distributions form the foundation for analyzing key aspects of the game, such as goal counts and team dynamics. By capturing the probabilistic nature of football events, these models help researchers and analysts understand complex phenomena like scoring likelihoods, and match outcomes. This section introduces the basic distributions commonly used in football modelling.

The Poisson distribution is a very popular discrete distribution that expresses the probability of an event occurring on a given unit of time, space, etc. These events happen with a known constant rate $\lambda > 0$ and are independent of time since the last event. The probability mass function is defined as:

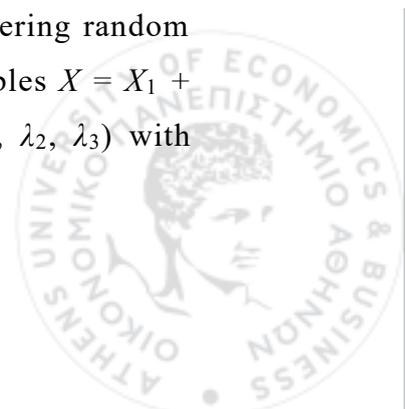
$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}, k = 0, 1, 2, \dots$$

The above equation provides the probability of observing k events within the given interval. Parameter λ represents the rate, the average number of events that occur in this interval. It's a measure of the distribution's central tendency and dispersion. This distribution has the equi-dispersion property. The variance is equal to the mean:

$$E(X) = \text{Var}(X) = \lambda.$$

Poisson distribution is also characterized by lack of memory, meaning that the probability of a future event is not affected by how much time has already elapsed since the last event.

The Bivariate Poisson is an extension of the Poisson distribution. It models the joint occurrence of two correlated count-based random variables, so it incorporates dependence between these two variables. Considering random variables X_1, X_2, X_3 with parameters $\lambda_1, \lambda_2, \lambda_3 \geq 0$ then the variables $X = X_1 + X_2$ and $Y = X_2 + X_3$ follow jointly a bivariate Poisson $BP(\lambda_1, \lambda_2, \lambda_3)$ with probability function:



$$P(X = x, Y = y) = e^{-(\lambda_1 + \lambda_2 + \lambda_3)} \frac{\lambda_1^x \lambda_2^y}{x! y!} \sum_{k=0}^{\min(x,y)} \binom{x}{k} \binom{y}{k} k! \left(\frac{\lambda_3}{\lambda_1 \lambda_2}\right)^k, \quad x, y = 0, 1, 2, \dots$$

Marginally each variable follows a Poisson distribution with means $E(X) = \lambda_1 + \lambda_2$ and $E(Y) = \lambda_2 + \lambda_3$ with λ_3 being the covariance of X and Y . Thus, if $\lambda_3 = 0$ the variables are independent.

The negative binomial distribution is a discrete probability distribution that models the number of trials required to achieve a fixed number of successes in a sequence of independent Bernoulli trials, where each trial has the same probability of success. It is often used to model count data that exhibit overdispersion, meaning that the variance exceeds the mean, which is not accommodated by the Poisson distribution. The parameterization of the probability mass function we consider in this thesis is:

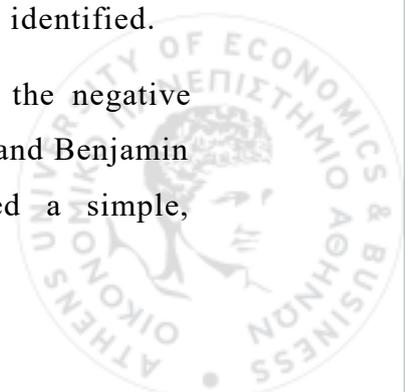
$$P(X = x) = \binom{x + \varphi - 1}{x} \left(\frac{\mu}{\mu + \varphi}\right)^x \left(\frac{\varphi}{\mu + \varphi}\right)^\varphi, \quad x = 0, 1, 2, \dots$$

Here $\mu > 0$ is the mean of the distribution and $\varphi > 0$ the dispersion parameter. The variance is $Var(X) = \mu + \frac{\mu^2}{\varphi}$. For large values of parameter φ the variance is close to μ , thus the distribution behaves like a Poisson.

2.2 Football Results Modelling Through The Years

Maher (1982) in his paper “Modelling association football scores” was the first to model the number of goals in football matches. Reep and Benjamin (1968) and Reep, Pollard and Benjamin (1971) by modelling passes using Poisson and negative binomial distributions concluded that “chance dominates the game”. However, Hill (1974) with a simple comparisons test showed that experts could predict quite accurately the final league standings. In an effort to answer to this apparent conflict of findings on the roles of luck and skill, Maher produced a model that could consider team-specific abilities which are measured by their performances in past matches. In this way, probabilities could be assigned to each outcome, and the better teams could be identified.

Studies on football data had shown until this point that the negative binomial distribution provided a better fit to football data (Reep and Benjamin (1968) and Reep et al. (1971)). However, Maher proposed a simple,



independent double Poisson model that includes parameters which “measure” the inherent qualities of teams on attacking and defending. It is defined in the following way:

$$X_{ij} | \lambda_1 \sim \text{Poisson}(\lambda_1),$$

$$Y_{ij} | \lambda_2 \sim \text{Poisson}(\lambda_2),$$

$$\lambda_1 = \alpha_i \beta_j,$$

$$\lambda_2 = \gamma_i \delta_j.$$

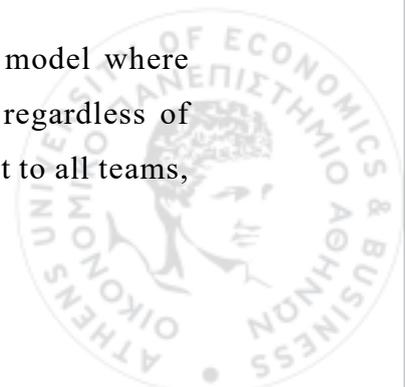
Here X_{ij} represents the goals scored by the home team and Y_{ij} the goals of the away team. These follow two independent Poisson distributions with means λ_1 and λ_2 respectively. The attacking ability of team i when playing at home is represented by α_i and the defensive ability of team j playing away by β_j . The defensive ability of the home team i is represented by γ_i and the attacking ability of the away team by δ_j . Attacking and defensive abilities represent how much better does a team compared to the average attacking or defensive level of the teams in the league. Larger positive attacking parameter values indicate that a team is performing better than average and larger (in absolute value) negative defensive parameter values suggest that a team defends better than average. In a league of n teams there will be $4n$ parameters and $n(n-1)$ observations (number of total matches). If all the α 's are multiplied by a factor k and all the β 's are divided by k then all the $\alpha_i\beta_j$ products are unchanged and thus a unique set of parameters can be produced with the constraints: $\sum_i \alpha_i = \sum_i \beta_i$ and $\sum_i \gamma_i = \sum_i \delta_i$.

The Poisson probability mass function for home goals is:

$$P(X_{ij} = x_i) = \frac{e^{-\alpha_i\beta_j}(\alpha_i\beta_j)^{x_i}}{x_i!}, x_i = 0, 1, 2, \dots$$

With the help of an iterative method one can extract the MLEs of α 's and β 's. In a similar way the γ 's and δ 's can be determined.

By comparing likelihoods Maher concluded to a simpler model where only the α 's and β 's are necessary to describe a team's ability regardless of playing at home or away. The home effect applies with equal effect to all teams,



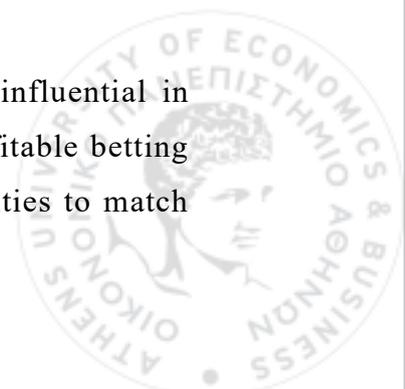
with the attacking ability diminishing by a constant factor when playing away. The same applies to defending.

Furthermore, Maher showed that although this model provides an overall accurate fit to the data it has some systematic deviations. By comparing observed and expected frequencies of goals scored he noted that the model underestimates the number of one and two goals being scored and overestimates the number of times that zero or more than four are scored. Notably, considering the whole dataset of twelve seasons that he examined, the χ^2 value is so inflated that it would lead to the rejection of the model. This is a sign towards the presence of some dependence.

Maher examined the distribution of the differences between team's scores $Z_{ij} = X_{ij} - Y_{ij}$. He understands that as the two opposing teams interact during the game there is a need for some sort of dependence between the goals that they eventually score, and this model aims to account for that. This Bivariate Poisson distribution has Poisson marginals with means $\mu_{ij} = \alpha_i \beta_j$, $\lambda_{ij} = k^2 \alpha_j \beta_i$ and correlation ρ . This model can be thought of as $X_{ij} = U_{ij} + W_{ij}$ and $Y_{ij} = V_{ij} + W_{ij}$ where U_{ij} , W_{ij} and V_{ij} are independent Poisson distributions with means $(\mu_{ij} - \eta_{ij})$, $(\lambda_{ij} - \eta_{ij})$ and η_{ij} , where $\eta_{ij} = \rho \sqrt{\mu_{ij} \lambda_{ij}}$ is the covariance between X_{ij} and Y_{ij} . By trying a range of values for ρ , Maher concluded that the most appropriate is 0.2. This model seemed to provide a better fit to the data and was able to estimate the differences in scores more accurately than the independent Poisson model.

Lee (1997) proceeded to an application of the independent Poisson model. He used it to calculate the attacking and defensive parameters of teams in the 95/96 English Premier League season. He also used it to assign probabilities to outcomes of particular matches and finally to do simulations of the league, all in an effort to determine if the final league positions were a true depiction of teams' quality or if luck had a role. The model was able to quite accurately predict the actual results.

The paper by Dixon and Coles (1997) has been highly influential in football modelling literature. The authors aimed to create a profitable betting strategy by building a model which accurately assigns probabilities to match



outcomes and then by comparing those probabilities to the odds provided by the bookmakers one is able to place bets on results that have high probability of occurrence but are underestimated by the bookmakers.

Dixon and Coles examined the assumption that home and away goals are independent through the use of the ratio $f(i, j) / f_H(i)f_A(j)$ with f being the observed joint probability function and f_H, f_A the marginal empirical probability functions under the independence assumption. They found that low scoring draws (0-0 and 1-1) are more common in actual data than under independence and the scores 1-0 and 0-1 are less frequent. To cope with these observations the authors provided the modifications of the independence model presented below.

Similar to the model of Maher (1982), Dixon and Coles include α and β which are the parameters that measure attacking and defensive abilities and introduce parameter γ which represents the home effect (advantage) which seems to be noteworthy in the game of football.

Home goals X_{ij} and away goals Y_{ij} are independent Poisson variables:

$$X_{ij} \sim \text{Poisson}(\alpha_i \beta_j \gamma),$$

$$Y_{ij} \sim \text{Poisson}(\alpha_j \beta_i).$$

Thus, they proposed a model with Poisson marginals with means λ and μ (for home and away team respectively) and an adjustment factor τ :

$$P(X_{ij} = x, Y_{ij} = y) = \tau_{\lambda, \mu}(x, y) \frac{\lambda^x e^{-\lambda}}{x!} \frac{\mu^y e^{-\mu}}{y!}, \quad x, y = 0, 1, 2, \dots$$

where

$$\lambda = \alpha_i \beta_j \gamma,$$

$$\mu = \alpha_j \beta_i.$$

and

$$\tau_{\lambda, \mu}(x, y) = \begin{cases} 1 - \lambda \mu \rho & \text{if } x = y = 0, \\ 1 + \lambda \rho & \text{if } x = 0, y = 1, \\ 1 + \mu \rho & \text{if } x = 1, y = 0, \\ 1 - \rho & \text{if } x = y = 1, \\ 1 & \text{otherwise.} \end{cases}$$



The adjustment factor provides the ability to move probability between the scores 0-0, 1-1, 1-0 and 0-1, departing from independence.

Dependence is captured by parameter ρ . Obviously for $\rho = 0$ we have independence but for $\rho \neq 0$ the independence assumption is perturbed for low scoring games to better fit the observed data. Parameter ρ also satisfies that:

$$\max(-1/\lambda, -1/\mu) \leq \rho \leq \min(1/\lambda\mu, 1).$$

Thus, it can be either positive or negative but is always bounded. For a positive ρ value the probability of scores 0-0 and 1-1 will be deflated and the probability of scores 0-1 and 1-0 will be increased. For a negative ρ value the converse is true. The probabilities of scores beyond these four combinations remain unchanged.

For n number of teams there is a need to estimate n number of attacking parameters, n number of defensive parameters, home effect γ and dependence parameter ρ . To prevent overparameterization they imposed the constraint:

$$n^{-1} \sum_{i=1}^n a_i = 1.$$

The parameters are obtained with numerical maximization of the likelihood function, with $k = 1, \dots, N$ the match index:

$$L(\alpha_i, \beta_i, \rho, \gamma; i=1, \dots, n) = \prod_{k=1}^N \tau_{\lambda_k, \mu_k}(x_k, y_k) \exp(-\lambda_k) \lambda_k^{x_k} \exp(-\mu_k) \mu_k^{y_k},$$

where

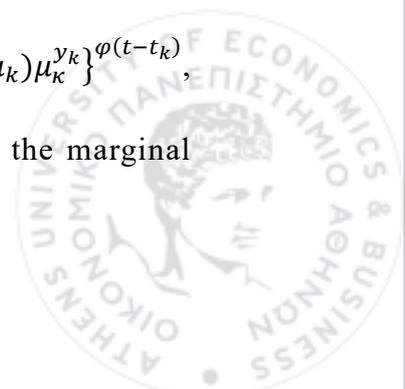
$$\lambda_k = a_{i(k)} \beta_{j(k)} \gamma,$$

$$\mu_k = a_{j(k)} \beta_{i(k)}.$$

The authors also further modified the model so that parameters are not static but affected more by recent performances of teams. This is achieved with the introduction of a weighting function which gives more value to recent information. So, the “pseudolikelihood” for each time point t is:

$$L_t(\alpha_i, \beta_i, \rho, \gamma; i=1, \dots, n) = \prod_{k \in A_t} \{ \tau_{\lambda_k, \mu_k}(x_k, y_k) \exp(-\lambda_k) \lambda_k^{x_k} \exp(-\mu_k) \mu_k^{y_k} \}^{\varphi(t-t_k)},$$

where t_k is the time of match k , $A_t = \{k: t_k < t\}$, λ_k and μ_k are the marginal means as above and φ is a non-increasing function of time.



The weighting function that the authors chose is $\varphi(t) = \exp(-\xi t)$, thus all previous results are downweighted exponentially according to $\xi > 0$. For $\xi=0$ the model is the static one. For larger values of ξ more weight is given to the most recent matches. To determine the value of ξ the authors define

$$S(\xi) = \sum_{k=1}^N (\delta_k^H \log p_k^H + \delta_k^A \log p_k^A + \delta_k^D \log p_k^D),$$

where δ_k^H , δ_k^A , δ_k^D are binary indexes for results corresponding to home win, away win and draw respectively and p_k^H , p_k^A , p_k^D are the maximum likelihood estimates of home win, away win and draw respectively. Through maximization of this function, they determine the value of ξ .

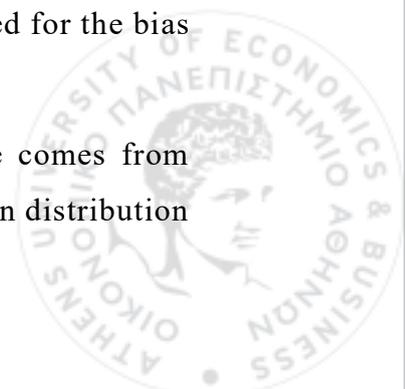
In this way, parameter ξ was chosen in a way to optimize the prediction of outcomes and they found it equal to 0.0065. Parameters were obtained by maximizing the above likelihood with $\xi=0.0065$. These parameters are dynamic meaning that they display some variation over time.

Furthermore, an equation that a betting strategy can be built upon was presented. It provided a way of comparing bookmakers' odds with the model-estimated probabilities and if the latter were accurate enough then there was positive expected return. They define expected gain as:

$$E(G) = p_k / b_k - 1.$$

Here p_k denotes the probability of a result provided by the model and b_k denotes the bookmaker's odds. Apparently, if the estimates of the model are more accurate than the odds then expected gain is positive. The betting strategy is determined by a value of r , for which it stands that $p_k / b_k > r$. Specifically, r is a predetermined level, larger than 1, which controls the bets that will be placed. For example, if $r = 1.2$ the difference between model probabilities and bookmaker odds should be high enough for the ratio to meet the above requirement, thus fewer bets will be placed. By examining the return that would have occurred with actual data if this strategy was adopted, the authors claimed that expected gain was positive for any $r > 1.1$ and even accounted for the bias of bookmakers' odds.

The next addition to football scores modelling literature comes from Karlis and Ntzoufras (2003). The authors use the Bivariate Poisson distribution



to model the differences of goals similarly to what Maher (1982) proposed but also add inflation terms on the diagonal to alter the probabilities of draws (in a score probability table the diagonal represents the draws). In this way they aim to introduce dependence between goals and also improve the fit on the count of draws which is a problem reported by Maher (1982). X (home goals) and Y (away goals) are modelled with three independent univariate Poisson random variables Z_1, Z_2, Z_3 for which it stands that $X = Z_1 + Z_3$ and $Y = Z_2 + Z_3$. Parameters λ_1, λ_2 and λ_3 are the rates of Z_1, Z_2 and Z_3 respectively.

The model is specified as follows:

$$(X_i, Y_i) \sim BP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i}),$$

where

$$\log(\lambda_{1i}) = \mu + \text{home} + \text{att}_{h_i} + \text{def}_{g_i},$$

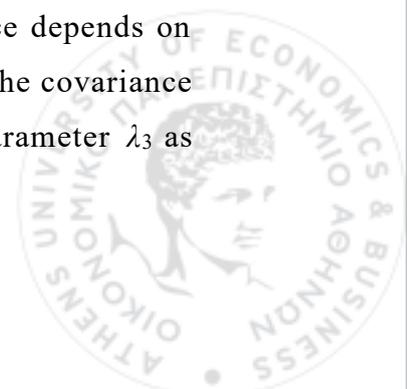
$$\log(\lambda_{2i}) = \mu + \text{att}_{g_i} + \text{def}_{h_i},$$

$$\log(\lambda_{3i}) = \beta^{\text{con}} + \gamma_1 \beta_{h_i}^{\text{home}} + \gamma_2 \beta_{g_i}^{\text{away}}.$$

The joint probability function of (X, Y) is given by:

$$P(X = x, Y = y) = e^{-(\lambda_1 + \lambda_2 + \lambda_3)} \frac{\lambda_1^x \lambda_2^y}{x! y!} \sum_{k=0}^{\min(x,y)} \binom{x}{k} \binom{y}{k} k! \left(\frac{\lambda_3}{\lambda_1 \lambda_2}\right)^k, \quad x, y = 0, 1, 2, \dots$$

Each random variable follows marginally a Poisson distribution with $E(X) = \lambda_1 + \lambda_3$ and $E(Y) = \lambda_2 + \lambda_3$. Obviously, as λ_1, λ_2 and λ_3 are Poisson rates they can only be non-negative. For λ_3 it stands that $\lambda_3 = \text{cov}(X, Y)$. Thus, this model allows only non-negative dependence as the covariance cannot be negative. For $\lambda_3 = 0$ the two variables are independent. Parameters *att* and *def* are interpreted like α and β in the previous models, μ is a constant parameter which specifies λ_1 and λ_2 when the two teams are of the same strength on a neutral field and *home* is the home advantage. Parameter β^{con} is also a constant while $\beta_{h_i}^{\text{home}}$ and $\beta_{g_i}^{\text{away}}$ depend on the home and away team respectively with γ_1 and γ_2 being dummy variables that indicate where the covariance depends on (e.g. if both are 0 the covariance is constant, if $\gamma_1 = 1$ and $\gamma_2 = 0$ the covariance depends on the home team only etc.). The authors interpret parameter λ_3 as



random effects that incorporate game conditions. To achieve identifiability, they propose either sum-to-zero or corner constraints.

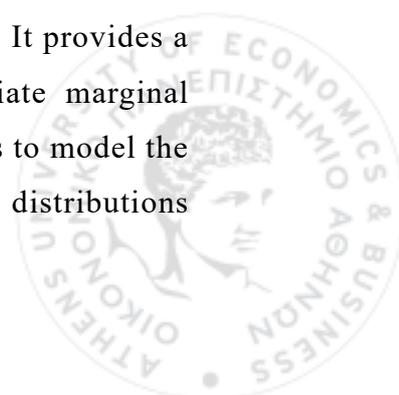
Karlis and Ntzoufras enhanced the Bivariate Poisson model by inflating the probabilities of draws. In literature, there were cases where the model could not approach accurately the observed number of draws which was usually underestimated. The general form of the joint probability function $P_D(x, y)$ of a diagonally inflated Bivariate Poisson model is:

$$P_D(x, y) = \begin{cases} (1 - p) BP(x, y | \lambda_1, \lambda_2, \lambda_3) & \text{if } x \neq y, \\ (1 - p) BP(x, y | \lambda_1, \lambda_2, \lambda_3) + p D(x, \theta) & \text{if } x = y. \end{cases}$$

Here, $D(x, \theta)$ is a discrete distribution like Poisson, Geometric or Bernoulli with parameter vector θ and p is the proportion of diagonal inflation. Now marginal distributions are no longer pure Poisson but become mixtures with one Poisson component, improving flexibility in fitting real-world data. Even without correlation ($\lambda_3 = 0$), inflation introduces a form of dependence between scores. These properties can combat both the overdispersion and correlation problems. The authors fit these models with the use of EM algorithm.

Fitting Bivariate Poisson models with different λ_3 formulations and diagonally inflated Bivariate Poisson models in 1991-92 Serie A data, the authors identified as best-fitting the Bivariate Poisson model with an extra parameter for score 1-1, which was underestimated in simpler models. Inflated models demonstrated better performance according to AIC, BIC, and Likelihood Ratio Tests compared to standard Bivariate Poisson models. However, fitting these models in 2000-2001 Champions League data showed that the diagonally inflated models did not improve the fit when the number of draws is not underestimated.

Presenting the next papers that appeared in football modelling literature requires to briefly explain copula functions. A copula is a statistical tool used to describe and model the dependence between random variables. It provides a way to construct multivariate distributions by linking univariate marginal distributions to a joint multivariate distribution. Copulas allow us to model the joint distribution of variables by combining the marginal distributions



(individual behaviors) with a dependence structure (copula). According to Sklar (1959) a copula is a multivariate cumulative distribution function where each marginal distribution is uniform on the interval $[0, 1]$. Any multivariate joint distribution can be written as a copula applied to the marginal distributions. Mathematically, for a bivariate distribution $F(x, y)$ with marginals $F_x(x)$ and $F_y(y)$, the copula is a distribution function $C: [0,1]^2 \rightarrow [0,1]$ that satisfies:

$$F(x, y) = C\{F_x(x), F_y(y)\}, (x, y) \in \mathbb{R}^2.$$

When F_x and F_y are continuous, the copula is uniquely determined on the unit square. When X and Y are discrete random variables taking values on some Ω then the copula is unique provided $(x, y) \in \Omega$. A valid bivariate distribution for X and Y arises whenever C , F_x and F_y are chosen from parametric families, discrete or continuous.

McHale and Scarf introduced the use of copulas to football data modelling. They use copula-based methods for generating bivariate distributions that can model both positive and negative dependence structures. In their first paper McHale and Scarf (2007) they describe the construction of Archimedean copulas (as in Nelsen (2006)), and employ two of them, Frank's:

$$C(u, v) = -k^{-1} \log \left\{ 1 - \frac{(1 - e^{-ku})(1 - e^{-kv})}{1 - e^{-k}} \right\}, k \in \mathbb{R} \setminus \{0\},$$

and Kimeldorf-Sampson's:

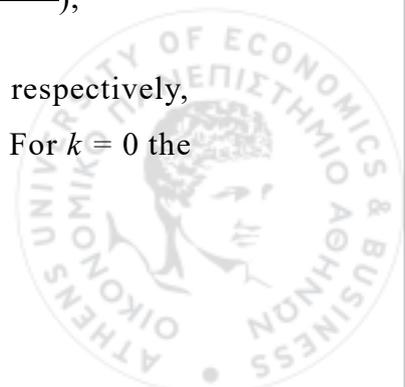
$$C(u, v) = \max \{(u^{-k} + v^{-k} - 1)^{-\frac{1}{k}}, 0\}, k \in \mathbb{R},$$

where $u, v \in [0, 1]$ are the cumulative distribution function values of the two random variables and parameter k controls the dependence.

They provide as an example a bivariate Poisson distribution with Frank's copula:

$$F_{XY}(x, y) = -\frac{1}{k} \log \left(1 - \frac{\{1 - \exp(-k \sum_{i=1}^x \frac{e^{-\mu_1} \mu_1^i}{i!})\} \{1 - \exp(-k \sum_{j=1}^y \frac{e^{-\mu_2} \mu_2^j}{j!})\}}{(1 - e^{-k})} \right),$$

where $x, y=0, 1, \dots$, are the number of home and away goals respectively, marginal means $\mu_1, \mu_2 > 0$, and dependence parameter $k \in (-\infty, \infty)$. For $k = 0$ the marginal distributions are independent.



By replacing u and v with the appropriate marginal functions, bivariate Geometric and negative Binomial distributions can be obtained. However, the authors noted that bivariate Geometric distributions provide a poor fit to the data.

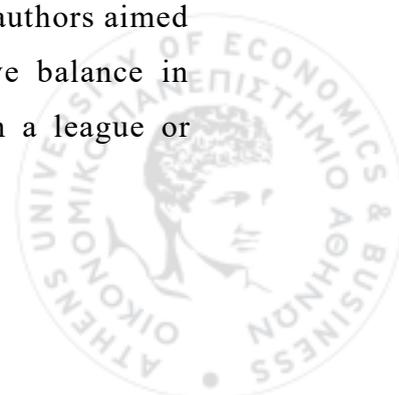
Thus, they employ these copulas with both Poisson and negative binomial marginals to model shots-for and shots-against in an effort to explore which in-game actions affect the number of goals. Goals according to the authors display only slight positive or no correlation at all. In contrast, shots show negative dependence, a feature that cannot be handled through the until then bivariate Poisson models. The copulas can capture this dependence and the negative binomial marginals can capture over or under dispersion in the data.

The likelihood function for a bivariate discrete distribution with marginals $F_x(x)$ and $F_y(y)$ and copula $C(u, v; k)$ for parameters (θ_1, θ_2, k) is:

$$L\{(\theta_1, \theta_2, k), (x_i, y_i)\} = C\{F_x(x_i), F_y(y_i)\} \\ - C\{F_x(x_i - 1), F_y(y_i)\} - C\{F_x(x_i), F_y(y_i - 1)\} + C\{F_x(x_i - 1), F_y(y_i - 1)\}.$$

They fit the models to data from the English Premier League for the period August 2003 to March 2006. Among the tested models, Frank's Copula with negative binomial Marginals provided the best fit. The authors also included covariates in the model and found that passes and crosses positively impact a team's shot production, supporting the notion that intricate play strategies ("the beautiful game") are effective. Tackles impact negatively on shots, possibly indicating defensive domination when tackle numbers are high.

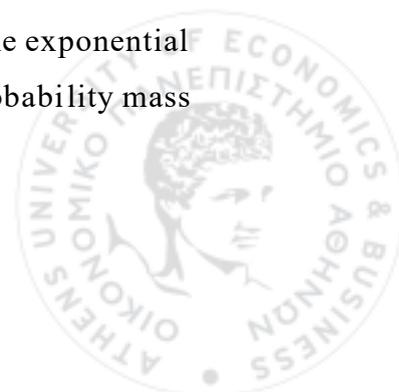
The flexibility of copulas to allow for negative dependence was also helpful for McHale and Scarf (2011) to model international matches scores. In national teams' matches it is more often for significantly better teams to play against weaker opposition, compared to domestic leagues where teams are split into divisions and thus the differences in ability are smaller. The authors aimed to associate dependence with competitive balance. Competitive balance in football (or any sport) refers to the degree to which teams in a league or competition have a relatively equal chance of success.



The goals in international matches showed negative correlation and the reason for this might be the uncompetitive nature of these matches. Exploring the ratios (as Dixon and Coles (1997)) $f(i, j) / f_H(i)f_A(j)$ where f is the observed joint probability function and f_H, f_A the marginal empirical probability functions under the independence assumption, the authors discovered two interesting features. There were more ($>3, 0$) outcomes and fewer ($>3, >0$) outcomes than expected under independence which indicated negative dependence. There were also more 1-1 and 2-2 draws which indicated positive dependence. These two features possibly explain why they observed a “net” weak negative correlation.

The authors explained this dependence structure with the presence of competitive balance, measured through the FIFA rank difference between teams. Matches with a small rank difference (closely matched teams) exhibit low dependence, while matches with a larger rank difference (unequal teams) exhibit stronger negative dependence. To model this they used Frank’s copula (which allows for both positive and negative dependence) with negative binomial marginals (which account for overdispersion). This model performed better compared to the one with Frank’s copula and Poisson marginals. They also allowed for the dependence parameter of the copula to be a linear function of rank difference, which improved the fit. This model showed that as the rank difference was increased (meaning that competitive balance was decreased), the dependence became more negative. Thus, a framework was set for measuring competitive balance which could be beneficial for forecasts or betting.

Copulas are used in literature with marginals other than Poisson or negative binomial. Boshnakov, Kharrat and McHale (2017) proposed the Weibull distribution. The model is based on a Weibull renewal process (McShane et al. (2008)) which assumes that goals occur by some time t when the inter-arrival times are independent and identically distributed Weibull random variables. This is done with a Taylor series expansion of the exponential in the Weibull density. It is called Weibull count model and the probability mass function is:



$$P(X(t) = x) = \sum_{j=x}^{\infty} \frac{(-1)^{x+j} (\lambda t^c)^j \alpha_j^x}{\Gamma(cj+1)}, \quad x = 0, 1, 2, \dots$$

where

$$\alpha_j^0 = \frac{\Gamma(cj+1)}{\Gamma(j+1)}, \quad j = 0, 1, 2, \dots,$$

and

$$\alpha_j^{x+1} = \sum_{m=x}^{j-1} a_m^x \Gamma(cj - cm + 1) / \Gamma(j - m + 1), \quad x = 0, 1, 2, \dots, j = x+1, x+2, \dots$$

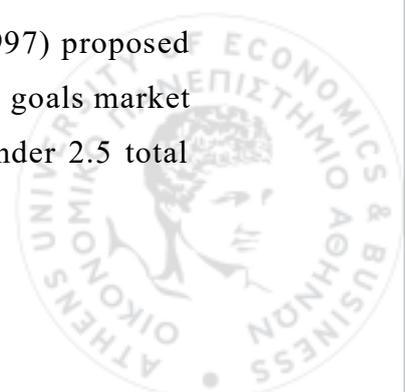
Parameter λ is a rate parameter and c a shape parameter. The observation unit here is the match which we consider as 1 time unit. So, λ is the scoring rate per match.

The hazard $h(t)$ associated to the count process can vary over time when using the Weibull distribution to model inter-arrival times. In survival analysis the Weibull hazard function describes the instantaneous rate of failure at a given time, assuming the subject has survived up to that time. Here, The Weibull hazard function allows the scoring intensity to change over the course of a match, which is more realistic for modeling football than the constant-rate Poisson model. Mathematically it is given by:

$$h(t) = \lambda c t^{c-1}.$$

For $c = 1$ the hazard is constant, implying that goals are equally likely to happen at any moment. This is the memoryless/time-homogeneous Poisson process. For $c < 1$, the hazard decreases over time, this suggests that the longer a team goes without scoring, the less likely it is to score soon. Finally, for $c > 1$, The hazard increases over time, this implies that the longer it takes for a goal to occur, the more likely it becomes.

This means that this distribution can handle both overdispersion and underdispersion in the data. To capture the dependence structure, they used Frank's copula, like McHale and Scarf (2007). The parameters for team abilities are built to be time-varying in the way that Dixon and Coles (1997) proposed with the ζ parameter but modified it to include the over-under 2.5 goals market (in betting it is common to bet whether there will be over or under 2.5 total goals scored in a match):



$$T(\zeta) = \sum_{k=1}^N (\delta_k^H \log p_k^H + \delta_k^A \log p_k^A + \delta_k^D \log p_k^D + \gamma_k^{02.5} \log p_k^{02.5} + \gamma_k^{U2.5} \log p_k^{U2.5}),$$

where $\gamma_k^{02.5} = 1$ if there are more than 2.5 goals in match k , $\gamma_k^{U2.5} = 1$ if there are less than 2.5 goals and $p_k^{02.5}$, $p_k^{U2.5}$ are the maximum likelihood estimates of over or under 2.5 in match k . This function is maximized at $\zeta = 0.002$ and this is the value that they use for their calculations.

The model proposed (Frank's copula - Weibull marginals) performed better in terms of log-likelihood and AIC compared to independent Weibull, copula – Poisson and independent Poisson models when fitted to data from the English Premier League for seasons from 2006/07 to 2015/16.

The paper by Michels, Ötting, and Karlis (2023) makes a significant contribution to the modeling of football scores by extending the foundational Dixon and Coles (1997) model. The Dixon and Coles model, which originally employs Poisson marginals, focuses on adjusting probabilities for specific score pairs (e.g., 0-0, 1-0, 0-1, and 1-1) to account for dependencies in football scorelines. The authors demonstrated that this model is, in fact, a special case of the Sarmanov family of distributions, a versatile framework for modeling bivariate data. However, the Dixon and Coles model does not allow for marginals other than Poisson and cannot modify probabilities of scores other than the four scores mentioned above. With the use of Sarmanov family one can extend the model to accommodate marginals other than Poisson, like negative binomial and also shift probability to scores beyond the four pairs.

The Sarmanov family was introduced by Sarmanov (1966) and further studied by Ting Lee (1996). Given that $P_1(x_1)$ and $P_2(x_2)$ are marginal probability mass functions for random variables X_1 and X_2 (home and away goals respectively, instead of X and Y) and $q_1(x_1)$, $q_2(x_2)$ are bounded non-constant functions, then a joint probability mass function can be defined by:

$$h(x_1, x_2) = P_1(x_1) P_2(x_2) [1 + \omega q_1(x_1)q_2(x_2)].$$

Functions $q_i(x_i)$ must satisfy that:

$$\sum_{x_i=-\infty}^{\infty} q_i(x_i)P_i(x_i) = 0.$$



Here the term $\omega q_1(x_1)q_2(x_2)$ introduces dependence between the variables and $\omega \in \mathbb{R}$ determines the strength and direction of the dependence between X_1 and X_2 with the condition that $1 + \omega q_1(x_1)q_2(x_2) \geq 0$. From this inequality it is relatively straightforward to derive the bounds for ω . For $\omega = 0$ we have the case of two independent marginal distributions.

According to Ting Lee (1996) the correlation between X_1 and X_2 is:

$$\rho = \frac{\omega u_1 u_2}{\sigma_1 \sigma_2},$$

where $u_i = E[X_i q_i(X_i)]$ and σ_i is the standard deviation of X_i for $i = 1, 2$.

Thus, by choosing different q -functions, the Sarmanov family can be used to produce bivariate distributions and model a wide variety of dependence structures, ranging from weak to strong, and positive to negative correlations.

It is plausible for someone to assume that the Sarmanov family is the same thing as copulas. However, there is one key difference. While both the Sarmanov family and copulas model dependencies, the first are constructive: one builds the joint distribution directly from the marginals. Copulas are transform-based: the marginals are linked via a copula function.

The authors suggest that the Dixon and Coles (1997) model is a member of the Sarmanov family. Specifically, it can be produced by setting $\omega = -\tilde{\omega}$ and selecting the following q -function:

$$q_{dc}(x_i) = \begin{cases} -\lambda_i & \text{if } x_i = 0 \\ 1 & \text{if } x_i = 1 \\ 0 & \text{if } x_i = 2, 3, \dots \end{cases}, \text{ for } i = 1, 2.$$

where λ_i is the mean of X_i . This q -function plugged in the expression $1 + \omega q_1(x_1)q_2(x_2)$ leads to the modification term proposed by Dixon and Coles. With Poisson marginals the condition $\sum q_i(x_i)P_i(x_i) = 0$ is satisfied.

In detail, for the Poisson distribution it stands that $P(X_i = 0) = e^{-\lambda}$ and $P(X_i = 1) = e^{-\lambda}\lambda$. Thus:

$$\begin{aligned} \sum q_i(x_i)P_i(x_i) &= q_{dc}(0)P(0) + q_{dc}(1)P(1) + 0 \cdot P(2) + \dots = \\ &= -\lambda e^{-\lambda} + e^{-\lambda}\lambda = 0. \end{aligned}$$



The flexibility of q -functions can be exploited to create different bivariate distributions with Poisson marginals that can shift probabilities in different ways and introduce correlation among different scores. The authors provide relative suggestions. These functions satisfy the condition $\sum q_i(x_i)P_i(x_i) = 0$ and one can prove it as above.

They define the following q -functions:

$$q_{pois}^{(1)}(x_i) = \begin{cases} -\lambda_i^2 & \text{if } x_i = 0 \\ \lambda_i & \text{if } x_i = 1 \\ 0 & \text{if } x_i = 2, 3, \dots \end{cases}, \quad q_{pois}^{(2)}(x_i) = \begin{cases} -\lambda_i^2 & \text{if } x_i = 0 \\ -\lambda_i & \text{if } x_i = 1 \\ 4 & \text{if } x_i = 2 \\ 0 & \text{if } x_i = 3, 4, \dots \end{cases}, \quad \text{for } i = 1, 2.$$

Function $q_{pois}^{(1)}$ allows to move probability with a quadratic term across the scores 0-0, 0-1, 1-0 and 1-1. Function $q_{pois}^{(2)}$ moves probability across all the scores from 0-0 to 2-2 (nine in total). This is the first time that such a “modification” to the original Dixon and Coles form has been presented in the literature.

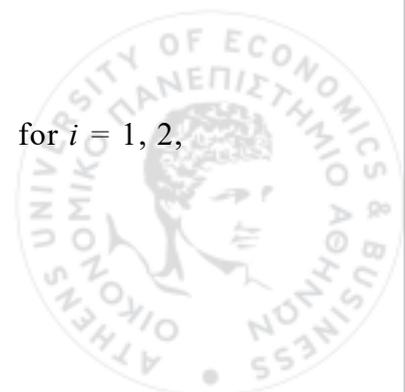
To generalize, the authors provided a q -function that can induce correlation among $(s + 1)^2$ scores. It is defined as:

$$q^{(s)}(x_i) = \begin{cases} -x_i! \lambda^{s-x_i} & \text{if } x_i = 0, 1, \dots, s-1 \\ ss! & \text{if } x_i = s \\ 0 & \text{if } x_i = s+1, \dots \end{cases}, \quad \text{for } i = 1, 2.$$

It is reported in the literature that the negative binomial distribution can handle better the overdispersion in football scores data. The Sarmanov family provides the ability to create bivariate distributions with any marginals given the appropriate q -functions. Thus, the authors use this property to introduce negative binomial marginals.

First, they present some new q functions that can be used generally for any discrete distribution marginals. Given that P_{x_i} is the probability of x_i and μ_i the expected value then:

$$q_{1P}(x_i) = \begin{cases} -\frac{P_1}{P_0} & \text{if } x_i = 0 \\ 1 & \text{if } x_i = 1 \\ 0 & \text{if } x_i = 2, 3, \dots \end{cases}, \quad q_{2P}(x_i) = \begin{cases} \mu_i & \text{if } x_i = 0 \\ -\mu_i \frac{P_0}{P_1} & \text{if } x_i = 1 \\ 0 & \text{if } x_i = 2, 3, \dots \end{cases}, \quad \text{for } i = 1, 2,$$



can be used to create bivariate distributions with any discrete distribution. These functions satisfy the required conditions ($\sum q_i(x_i)P_i(x_i) = 0$ and the bounds for parameter ω).

The authors propose q -functions for negative binomial marginals that occur from the above general forms:

$$q_{nb}^{(1)}(x_i) = \begin{cases} -\varphi_i \left(\frac{\mu_i}{\varphi_i + \mu_i}\right) & \text{if } x_i = 0 \\ 1 & \text{if } x_i = 1, \text{ for } i = 1, 2, \\ 0 & \text{if } x_i = 2, 3, \dots \end{cases}$$

with μ_i the mean and $\mu_i + \frac{\mu_i^2}{\varphi_i}$ the variance of the marginals. This q -function moves probability across 0-0, 1-1, 1-0 and 0-1 like the Dixon and Coles model.

For the marginal distributions it stands that $P(X_i = 0) = \left(\frac{\varphi_i}{\varphi_i + \mu_i}\right)^{\varphi_i}$ and $P(X_i = 1) = \varphi_i \left(\frac{\mu_i}{\varphi_i + \mu_i}\right) \left(\frac{\varphi_i}{\varphi_i + \mu_i}\right)^{\varphi_i}$. Thus, it is straightforward to show that $\sum q(x_i)P(x_i) = 0$ for $i = 1, 2$. This condition ensures that a proper bivariate distribution is provided. Specifically, for $q = q_{nb}^{(1)}$:

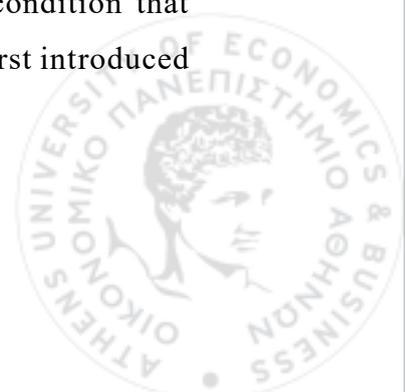
$$\begin{aligned} \sum q(x_i)P(x_i) &= q(x_1)P(x_1) + q(x_2)P(x_2) = \\ &-\varphi_i \left(\frac{\mu_i}{\varphi_i + \mu_i}\right) \left(\frac{\varphi_i}{\varphi_i + \mu_i}\right)^{\varphi_i} + 1 \cdot \varphi_i \left(\frac{\mu_i}{\varphi_i + \mu_i}\right) \left(\frac{\varphi_i}{\varphi_i + \mu_i}\right)^{\varphi_i} = 0. \end{aligned}$$

They showed that it is possible to extend to models that shift probabilities in different ways and to different scores with other q -functions:

$$q_{nb}^{(2)}(x_i) = \begin{cases} -\mu_i^2 & \text{if } x_i = 0 \\ \mu_i \frac{\varphi_i + \mu_i}{\varphi_i} & \text{if } x_i = 1 \\ 0 & \text{if } x_i = 2, 3, \dots \end{cases}, \quad q_{nb}^{(3)}(x_i) = \begin{cases} -\mu_i^2 & \text{if } x_i = 0 \\ -\mu_i \frac{\varphi_i + \mu_i}{\varphi_i} & \text{if } x_i = 1 \\ 4\mu_i \frac{\varphi_i}{\varphi_i + \mu_i} & \text{if } x_i = 2 \\ 0 & \text{if } x_i = 3, 4, \dots \end{cases}$$

for $i = 1, 2$.

More q -functions can be produced as required with the condition that $\sum q_i(x_i)P_i(x_i) = 0$. For example, the following q -function that is first introduced in this thesis can be used with negative binomial marginals:



$$q_{nb}^{(4)}(x_i) = \begin{cases} -\mu_i^2 & \text{if } x_i = 0 \\ -\mu_i \frac{\varphi_i + \mu_i}{\varphi_i} & \text{if } x_i = 1 \\ 4 \frac{(\mu_i + \varphi_i)^2}{\varphi_i^2 + \varphi_i} & \text{if } x_i = 2 \\ 0 & \text{if } x_i = 3, 4, \dots \end{cases}, \text{ for } i = 1, 2.$$

In detail, we show that:

$$\begin{aligned} \sum q_i(x_i)P_i(x_i) &= -\mu_i^2 \left(\frac{\varphi_i}{\varphi_i + \mu_i}\right)^{\varphi_i} - \mu_i \frac{\varphi_i + \mu_i}{\varphi_i} \varphi_i \left(\frac{\mu_i}{\varphi_i + \mu_i}\right) \left(\frac{\varphi_i}{\varphi_i + \mu_i}\right)^{\varphi_i} \\ &+ 4 \frac{(\mu_i + \varphi_i)^2}{\varphi_i^2 + \varphi_i} \frac{\varphi_i(\varphi_i + 1)}{2} \left(\frac{\mu_i}{\mu_i + \varphi_i}\right)^2 \left(\frac{\varphi_i}{\mu_i + \varphi_i}\right)^{\varphi_i} = \\ &- \mu_i^2 \left(\frac{\varphi_i}{\varphi_i + \mu_i}\right)^{\varphi_i} - \mu_i^2 \left(\frac{\varphi_i}{\varphi_i + \mu_i}\right)^{\varphi_i} + 2(\mu_i + \varphi_i)^2 \frac{\mu_i^2}{(\mu_i + \varphi_i)^2} \left(\frac{\varphi_i}{\varphi_i + \mu_i}\right)^{\varphi_i} = \\ &- 2\mu_i^2 \left(\frac{\varphi_i}{\varphi_i + \mu_i}\right)^{\varphi_i} + 2\mu_i^2 \left(\frac{\varphi_i}{\varphi_i + \mu_i}\right)^{\varphi_i} = 0. \end{aligned}$$

Thus, $q_{nb}^{(4)}$ is a plausible q -function that can be used with negative binomial marginals.

It is also possible to “mix” marginals of different distributions. The authors provide as an example the use of one Poisson and one negative binomial marginal with $q_{dc}(x_1)$ and $q_{nb}^{(1)}$ respectively. This provides greater flexibility that can be useful when modelling real world data.

As a result, with the Sarmanov family there is a method to construct many flexible and powerful bivariate distributions. It accommodates various dependence structures and marginal distributions. One can also benefit from its interpretability: the dependence parameter ω and q -functions have clear roles in defining the correlation and interaction between the variables. Finally, the construction is straightforward, and the resulting joint distribution retains the simplicity of the marginals.



3. Negative Binomial Marginals Model

This thesis aims to thoroughly examine a model built within the Sarmanov family framework along with negative binomial marginals. This model is particularly useful for football match outcomes because it offers a different way to model them, not considering only Poisson marginals as the “traditional” Dixon and Coles model. Additionally, by incorporating dependence via the Sarmanov family, the model allows for interactions between home and away goals, meaning it can reflect patterns such as high-scoring or low-scoring matches occurring more or less often than expected under independence.

In this section, we introduce the model that forms the foundation of our analysis. We begin by outlining its structure and the underlying assumptions that support its formulation. Following this, we present the mathematical framework, highlighting how the model captures the relationships within the data.

The basic assumption of the model is that home and away goals follow a negative binomial distribution. These are the marginal distributions of the model. Specifically, for variables X_1 and X_2 which symbolize the home and away goals respectively it stands that:

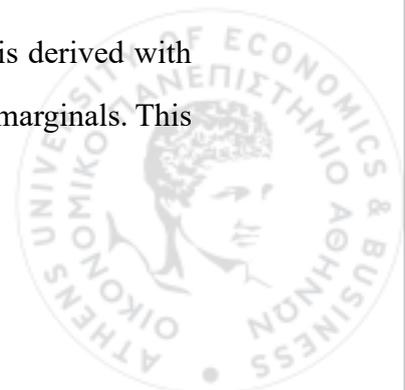
$$X_1 \sim \text{Negative Binomial}(\varphi_1, \mu_1),$$

$$X_2 \sim \text{Negative Binomial}(\varphi_2, \mu_2),$$

with $\mu_1 = \exp(a_{home} + \beta_{away} + home)$ and $\mu_2 = \exp(a_{away} + \beta_{home})$ the respective means.

Parameter a represents a team’s attacking strength and β the defensive strength with indexes *home* and *away* differentiating if a team is playing at home or away. Parameter *home* is the home advantage. The parameters φ_1 and φ_2 represent the dispersion in the negative binomial distribution for home and away team goals, respectively. These parameters control how much the variance of the goal distribution differs from a standard Poisson model, which assumes that the variance is equal to the mean.

The model is created through the Sarmanov family concept. It is derived with the use of previously mentioned q -function, $q_{nb}^{(1)}$ and negative binomial marginals. This leads to the following joint probability mass function:



$$\begin{aligned}
P(X_1 = x_1, X_2 = x_2) &= P(X_1 = x_1) P(X_2 = x_2) (1 + \omega q_{nb}^{(1)}(x_1) q_{nb}^{(1)}(x_2)) = \\
&\binom{x_1 + \varphi_1 - 1}{x_1} \left(\frac{\mu_1}{\varphi_1 + \mu_1}\right)^{x_1} \left(\frac{\varphi_1}{\varphi_1 + \mu_1}\right)^{\varphi_1} \times \\
&\binom{x_2 + \varphi_2 - 1}{x_2} \left(\frac{\mu_2}{\varphi_2 + \mu_2}\right)^{x_2} \left(\frac{\varphi_2}{\varphi_2 + \mu_2}\right)^{\varphi_2} \times \\
&[1 + \omega q_{nb}^{(1)}(x_1) q_{nb}^{(1)}(x_2)],
\end{aligned}$$

where:

$$q_{nb}^{(1)}(x_i) = \begin{cases} -\varphi_i \left(\frac{\mu_i}{\varphi_i + \mu_i}\right) & \text{if } x_i = 0 \\ 1 & \text{if } x_i = 1, \text{ for } i = 1, 2, \\ 0 & \text{if } x_i = 2, 3, \dots \end{cases}$$

with μ_i the mean and $\mu_i + \frac{\mu_i^2}{\varphi_i}$ the variance of the negative binomial distribution. Thus, probability is shifted between scores 0-0, 1-0, 0-1 and 1-1 with the magnitude of shifting depending on parameter ω .

Parameter $\omega \in \mathbb{R}$ controls the dependence between goals. From the Sarmanov family condition $1 + \omega q_1(x_1)q_2(x_2) \geq 0$ we can derive the range of ω . Substituting in the above inequality the corresponding branches of the q -function for the respective values of home and away goals (0-0, 0-1, 1-0, 1-1) we derive four inequalities which determine the lower and upper bounds:

- For $x_1 = 0, x_2 = 0$:

$$1 + \omega q_1(x_1)q_2(x_2) \geq 0 \Rightarrow$$

$$1 + \omega \left[-\varphi_1 \left(\frac{\mu_1}{\varphi_1 + \mu_1}\right) \left(-\varphi_2 \left(\frac{\mu_2}{\varphi_2 + \mu_2}\right)\right)\right] \geq 0 \Rightarrow \omega \geq -\frac{(\varphi_1 + \mu_1)(\varphi_2 + \mu_2)}{\varphi_1 \mu_1 \varphi_2 \mu_2}.$$

- For $x_1 = 0, x_2 = 1$:

$$1 + \omega \left[-\varphi_1 \left(\frac{\mu_1}{\varphi_1 + \mu_1}\right) \cdot 1\right] \geq 0 \Rightarrow 1 - \omega \varphi_1 \left(\frac{\mu_1}{\varphi_1 + \mu_1}\right) \geq 0 \Rightarrow \omega \leq \frac{\varphi_1 + \mu_1}{\varphi_1 \mu_1}.$$

- For $x_1 = 1, x_2 = 0$:

$$1 + \omega \left[1 \cdot \left(-\varphi_2 \left(\frac{\mu_2}{\varphi_2 + \mu_2}\right)\right)\right] \geq 0 \Rightarrow \omega \leq \frac{\varphi_2 + \mu_2}{\varphi_2 \mu_2}.$$

- For $x_1 = 1, x_2 = 1$:

$$1 + \omega (1 \cdot 1) \geq 0 \Rightarrow \omega \geq -1.$$

The bounds for ω can be summarized in the following inequality:



$$\max \left(-1, -\frac{(\varphi_1 + \mu_1)(\varphi_2 + \mu_2)}{\varphi_1 \mu_1 \varphi_2 \mu_2} \right) \leq \omega \leq \min \left(\frac{\varphi_1 + \mu_1}{\varphi_1 \mu_1}, \frac{\varphi_2 + \mu_2}{\varphi_2 \mu_2} \right).$$

Thus, omega can be either positive or negative. Naturally, the case $\omega = 0$ corresponds to the independence case.

Different q -functions alter the probabilities of scores in different ways. To understand these differences, we can compare q -functions and how they distribute probabilities across different numbers of home and away goals. Additionally, different q -functions imply different levels of correlation. The comparison below is between the independence case (negative binomial marginals without an interaction term) and the following q -functions:

- $q_{nb}^{(1)}(x_i) = \begin{cases} -\varphi_i \left(\frac{\mu_i}{\varphi_i + \mu_i} \right) & \text{if } x_i = 0 \\ 1 & \text{if } x_i = 1 \\ 0 & \text{if } x_i = 2, 3, \dots \end{cases}$
- $q_{nb}^{(2)}(x_i) = \begin{cases} -\mu_i^2 & \text{if } x_i = 0 \\ \mu_i \frac{\varphi_i + \mu_i}{\varphi_i} & \text{if } x_i = 1 \\ 0 & \text{if } x_i = 2, 3, \dots \end{cases}$
- $q_{nb}^{(4)}(x_i) = \begin{cases} -\mu_i^2 & \text{if } x_i = 0 \\ -\mu_i \frac{\varphi_i + \mu_i}{\varphi_i} & \text{if } x_i = 1 \\ 4 \frac{(\mu_i + \varphi_i)^2}{\varphi_i^2 + \varphi_i} & \text{if } x_i = 2 \\ 0 & \text{if } x_i = 3, 4, \dots \end{cases}$

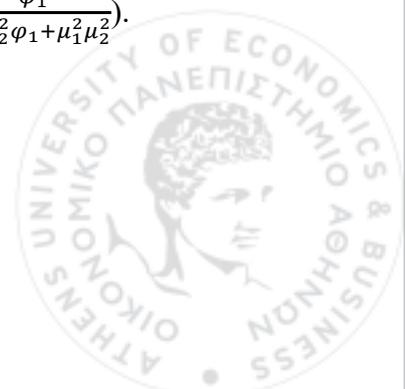
for $i = 1, 2$.

As seen before the range of acceptable values of parameter ω occurs from the condition $1 + \omega q_1(x_1)q_2(x_2) \geq 0$. This ensures that there will be no negative probabilities for any possible scoreline. We showed that for $q_{nb}^{(1)}$ it stands that:

$$\max \left(-1, -\frac{(\varphi_1 + \mu_1)(\varphi_2 + \mu_2)}{\varphi_1 \mu_1 \varphi_2 \mu_2} \right) \leq \omega \leq \min \left(\frac{\varphi_1 + \mu_1}{\varphi_1 \mu_1}, \frac{\varphi_2 + \mu_2}{\varphi_2 \mu_2} \right).$$

In the same way we find the limits for the other two q -functions. Specifically, the range of ω for $q_{nb}^{(2)}$ and $q_{nb}^{(4)}$ respectively takes the following forms:

1. $\max \left(-\frac{1}{\mu_1^2 \mu_2^2}, -\frac{\varphi_1 \varphi_2}{\mu_1 \mu_2 (\varphi_1 + \mu_1)(\varphi_2 + \mu_2)} \right) \leq \omega \leq \min \left(\frac{\varphi_2}{\mu_1^2 \mu_2 \varphi_2 + \mu_1^2 \mu_2^2}, \frac{\varphi_1}{\mu_1 \mu_2^2 \varphi_1 + \mu_1^2 \mu_2^2} \right).$



$$2. \max \left(-\frac{1}{\mu_1^2 \mu_2^2}, -\frac{\varphi_2}{\mu_1^2 \mu_2 \varphi_2 + \mu_1^2 \mu_2^2}, -\frac{\varphi_1}{\mu_1 \mu_2^2 \varphi_1 + \mu_1^2 \mu_2^2}, -\frac{\varphi_1 \varphi_2}{\mu_1 \mu_2 (\varphi_1 + \mu_1) (\varphi_2 + \mu_2)}, \right. \\ \left. -\frac{(\varphi_1^2 + \varphi_1)(\varphi_2^2 + \varphi_2)}{16(\mu_1 + \varphi_1)^2 (\mu_2 + \varphi_2)^2} \right) \leq \omega \leq \min \left(\frac{\varphi_2^2 + \varphi_2}{4\mu_1^2 (\mu_2 + \varphi_2)^2}, \frac{\varphi_1 \varphi_2^2 + \varphi_1 \varphi_2}{4\mu_1 (\varphi_1 + \mu_1) (\mu_2 + \varphi_2)^2}, \frac{\varphi_1^2 + \varphi_1}{4\mu_2^2 (\mu_1 + \varphi_1)^2}, \right. \\ \left. \frac{\varphi_1^2 \varphi_2 + \varphi_1 \varphi_2}{4\mu_2 (\varphi_2 + \mu_2) (\mu_1 + \varphi_1)^2} \right).$$

Obviously, the ranges of ω for $q_{nb}^{(2)}$ and $q_{nb}^{(4)}$ are more restrictive than $q_{nb}^{(1)}$ thus we expect generally smaller acceptable values.

Figure 1 provides a visual representation that can help understand the way that different q -functions work in altering probabilities. In each cell there is the probability of a particular score.

For all panels we consider negative binomial marginals with means 1.8 and 1.4 respectively. Dispersion parameters φ_1 and φ_2 are set to 5. Dependence parameter ω is set to 0.15 for $q_{nb}^{(1)}$ and $q_{nb}^{(2)}$, while for $q_{nb}^{(4)}$ it is set to 0.05 as this is the largest acceptable value.

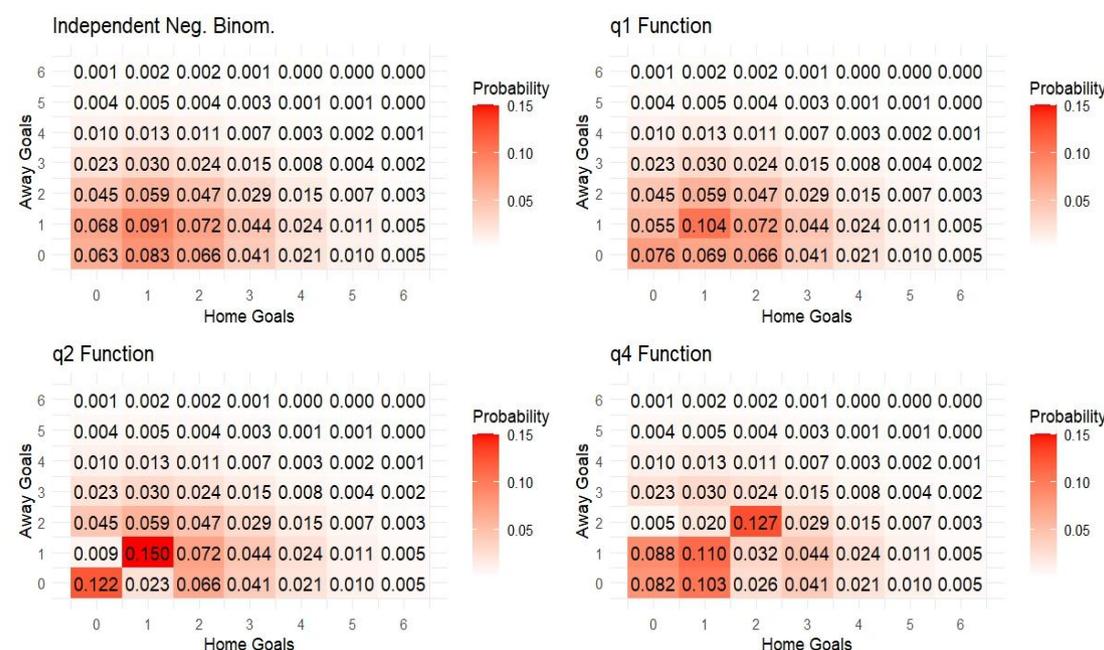
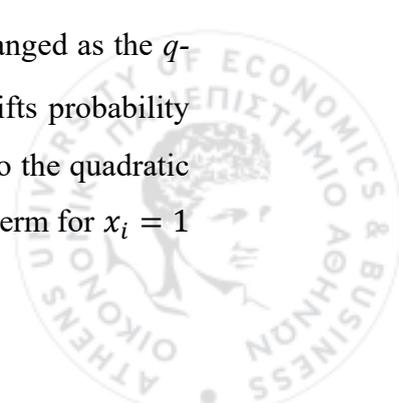


Figure 1 – Probability distributions for different q -functions.

We see that $q_{nb}^{(1)}$ compared to independence, moves probability from scores 1-0 and 0-1 to 0-0 and 1-1. Probabilities beyond these scores remain unchanged as the q -function takes the value of 0 for $x_i = 2, 3, \dots$. Function $q_{nb}^{(2)}$ again shifts probability from 1-0 and 0-1 to 0-0 and 1-1 but more strongly. This happens due to the quadratic term for $x_i = 0$ and the term for $x_i = 1$ which is larger than 1 (1 is the term for $x_i = 1$



of $q_{nb}^{(1)}$). Again, probabilities beyond these four pairs remain unchanged. Finally, $q_{nb}^{(4)}$ moves probability from pairs 0-2, 1-2, 2-0 and 2-1 to 0-0, 0-1, 1-0, 1-1 and 2-2. Thus, in this case, probability is shifted across nine pairs compared to the four of the two previous q -functions.

Furthermore, we examine the correlation that each q -function can achieve for different values of dependence parameter ω . Correlation is computed with the formula $Cov(X,Y) / \sqrt{Var(X)Var(Y)}$ where $Cov(X,Y) = E(XY) - E(X)E(Y)$, $Var(X) = E(X^2) - E(X)^2$ and $Var(Y) = E(Y^2) - E(Y)^2$. Figure 2 is a line plot in which each point of the red line is a value of correlation (y-axis) that can be achieved for the respective value of parameter ω (x-axis). For $\mu_1 = 1.8$, $\mu_2 = 1.4$ and $\varphi_1 = \varphi_2 = 5$ we can define the range of acceptable values of parameter ω based on the beforementioned limits. Specifically: for function $q_{nb}^{(1)}$ the lowest acceptable value of ω is -0.69 and the highest is 0.75, for $q_{nb}^{(2)}$ the range is from -0.15 to 0.15 and for $q_{nb}^{(4)}$ the range is significantly limited, from -0.029 to 0.055. Since only those values of ω are acceptable then only for these values we can get correlation values (points on the red line).

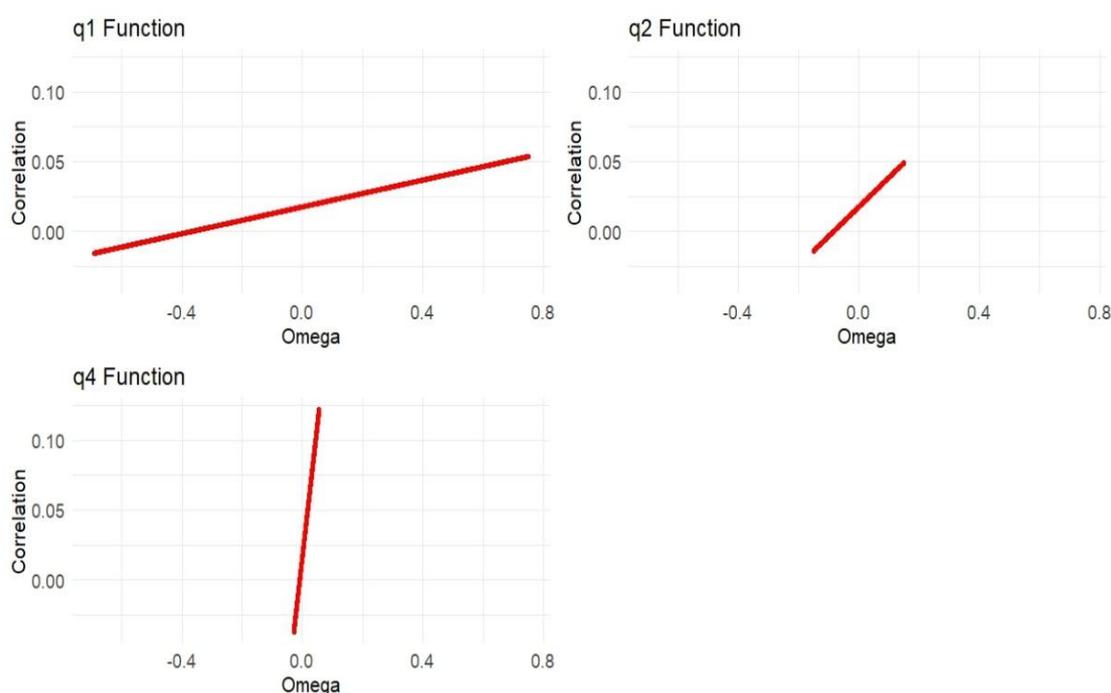


Figure 2 – Correlation for different values of ω and different q -functions.

From Figure 2 we see that with $q_{nb}^{(1)}$ and $q_{nb}^{(2)}$ we can get correlation values from around -0.015 up to around 0.05, while with $q_{nb}^{(4)}$ correlation can range significantly

more, from -0.03 up to 0.12. Function $q_{nb}^{(4)}$ shifts probability among nine scores (0-0 to 2-2) compared to $q_{nb}^{(1)}$ and $q_{nb}^{(2)}$ which shift probability among four scores (0-0 to 1-1) and thus is able to induce higher dependence between home and away goals. However, its limited ω can be restrictive in real world scenarios.



4. Application

4.1 Data

For the purposes of this thesis, we have obtained data from major European football leagues. Specifically, we fitted models on data of the men's English Premier League, Spanish La Liga, German Bundesliga and Dutch Eredivisie for the last three seasons (2021/22, 2022/23 and 2023/24). The data were obtained from Football-Data.co.uk (<https://www.football-data.co.uk/data.php>). These data include the four variables of interest: the home team, the away team, number of home team goals and away team goals.

We explore the dependence structure that lies between home and away goals. Figure 3 presents the empirical correlation values of goals observed in each league in different seasons. Each dot represents one season. In general, we have negative correlation values for all leagues and seasons except for two seasons of La Liga where correlation appears slightly positive. Negative correlation suggests that in matches where the home team scores more goals, the away team tends to score fewer goals meaning that we expect more one-sided games: dominance by one team often suppresses the other team's scoring. In contrast, positive correlation values in La Liga suggest that matches with high home goals also tend to have high away goals (more open games).

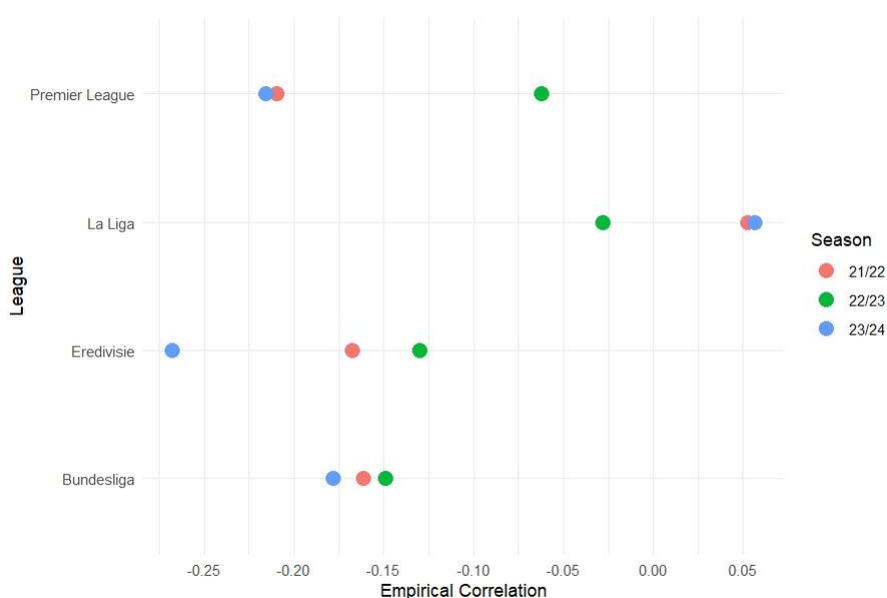


Figure 3 – Empirical correlation values.



Other than correlation, it is common in football modelling literature to check the ratio of the joint empirical probability function and the empirical probability function under independence for each score which is approximated as the product of the marginal empirical probabilities of X (home goals) and Y (away goals). Specifically, after getting the joint contingency table of home and away goals, the ratio can be considered in the following form:

$$\frac{f(i,j)}{f_H(i)f_A(j)}$$

where $f(i,j)$ represents the frequency of both variables having specific values (e.g., how often a game ended 2-1 divided by the total number of matches), $f_H(i)$ represents the frequency for a specific row value of the contingency table (e.g., all matches where the home team scored 2 goals divided by the total number of matches) and $f_A(j)$ represents the frequency for a specific column value of the contingency table (e.g., all matches where the away team scored 1 goal divided again by the total matches). Tables 1 to 4 present the estimates of this ratio for different leagues across the three seasons considered (2021/22, 2022/23, 2023/24). Standard errors are obtained via bootstrapping and are given in parentheses.

A ratio greater than 1 indicates that the combination occurs more frequently than expected under independence. Conversely, a ratio less than 1 indicates that the combination occurs less frequently than expected. In the Premier League data of Table 1, the overrepresentation of scores 3-0, 3-1, 5-0, 0-2, 0-3, 0-4 and 0-5 along with the underrepresentation of 0-0 might be an indication of negative dependence between home and away goals, possibly explaining the negative empirical correlation values of Figure 3 above. In simpler terms, there are more one-sided games where the dominant team scores many goals and the weaker one cannot respond, than games of evenly matched teams that could end goalless.

		Premier League					
H \ A	0	1	2	3	4	5	
0	0.79 (0.082)	0.89 (0.081)	1.17 (0.115)	1.18 (0.177)	1.35 (0.300)	2.19 (0.613)	
1	1.02 (0.067)	1.00 (0.062)	0.94 (0.083)	0.92 (0.129)	1.30 (0.217)	1.29 (0.425)	
2	0.97 (0.079)	0.98 (0.074)	1.15 (0.105)	1.01 (0.164)	0.87 (0.240)	0.55 (0.375)	
3	1.12 (0.122)	1.18 (0.108)	0.88 (0.139)	0.83 (0.213)	0.28 (0.193)	0.00 (0.000)	
4	1.04 (0.176)	1.09 (0.160)	0.75 (0.209)	1.29 (0.382)	0.95 (0.532)	0.00 (0.000)	
5	1.71 (0.323)	1.07 (0.269)	0.51 (0.277)	0.35 (0.354)	0.00 (0.000)	0.00 (0.000)	

Table 1 – Ratios for Premier League.

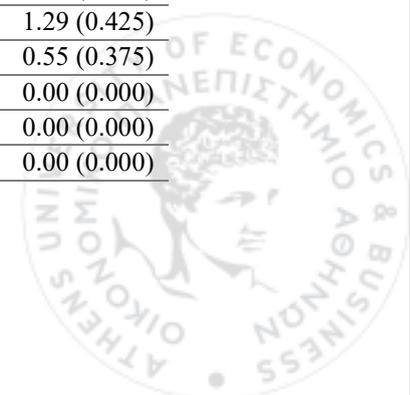


Table 2 presents the situation in Spain. There is some overrepresentation of 4-0 and 0-3, a sign towards negative dependence but also an overrepresentation of 3-3, a sign towards positive dependence. Other than those there are no significant deviations from independence. These facts might explain why there are two positive and one negative empirical correlation values with all three being close to 0.

		La Liga					
H \ A	0	1	2	3	4	5	
0	1.03 (0.073)	0.93 (0.072)	1.05 (0.109)	1.19 (0.194)	0.68 (0.280)	0.00 (0.000)	
1	1.07 (0.056)	1.05 (0.057)	0.93 (0.080)	0.54 (0.128)	1.27 (0.271)	0.73 (0.725)	
2	0.86 (0.075)	1.07 (0.079)	1.01 (0.113)	1.18 (0.215)	1.17 (0.361)	1.10 (1.059)	
3	0.97 (0.116)	0.91 (0.113)	1.11 (0.167)	1.18 (0.322)	0.88 (0.496)	4.38 (2.583)	
4	1.15 (0.182)	0.73 (0.164)	1.00 (0.266)	1.77 (0.625)	0.69 (0.702)	0.00 (0.000)	
5	0.61 (0.278)	1.36 (0.342)	0.79 (0.447)	2.20 (1.184)	0.00 (0.000)	0.00 (0.000)	

Table 2 – Ratios for La Liga.

In the ratios of Eredivisie data in Table 3 there are more 4-0, 0-3 and 0-4 results than expected under independence. Again, this could be a sign of negative dependence which also corresponds to the negative correlation values for the three Eredivisie seasons.

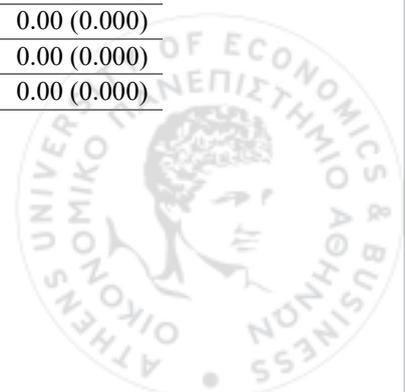
		Eredivisie					
H \ A	0	1	2	3	4	5	
0	0.92 (0.094)	0.92 (0.086)	0.81 (0.122)	1.34 (0.202)	1.35 (0.298)	2.66 (0.585)	
1	0.79 (0.076)	1.02 (0.071)	1.15 (0.106)	1.00 (0.148)	1.52 (0.245)	0.44 (0.311)	
2	1.03 (0.100)	0.91 (0.086)	1.23 (0.133)	0.99 (0.181)	0.69 (0.241)	0.90 (0.490)	
3	1.06 (0.131)	1.24 (0.121)	0.76 (0.151)	0.90 (0.228)	0.62 (0.308)	0.00 (0.000)	
4	1.64 (0.223)	0.86 (0.174)	0.83 (0.256)	0.70 (0.330)	0.00 (0.000)	0.00 (0.000)	
5	1.55 (0.291)	0.90 (0.239)	0.99 (0.337)	0.27 (0.275)	0.00 (0.000)	1.70 (1.755)	

Table 3 – Ratios for Eredivisie.

Finally, the ratios of the Bundesliga data suggest similar facts with the previous ones. Scores 3-0, 4-0, 0-3 and 0-4 appear more frequently than independence and also score 3-3 is less frequent. These point towards negative dependence.

		Bundesliga					
H \ A	0	1	2	3	4	5	
0	1.02 (0.119)	0.78 (0.094)	1.08 (0.141)	1.28 (0.232)	1.21 (0.401)	1.43 (0.764)	
1	0.72 (0.075)	1.19 (0.067)	0.87 (0.090)	1.06 (0.155)	1.48 (0.270)	1.36 (0.499)	
2	1.01 (0.093)	0.95 (0.076)	1.07 (0.108)	1.21 (0.183)	0.60 (0.254)	1.33 (0.571)	
3	1.15 (0.130)	1.00 (0.108)	0.99 (0.153)	0.77 (0.226)	1.04 (0.427)	0.00 (0.000)	
4	1.49 (0.225)	0.89 (0.160)	1.09 (0.241)	0.33 (0.226)	0.44 (0.447)	0.00 (0.000)	
5	1.17 (0.306)	0.96 (0.240)	1.43 (0.368)	0.32 (0.319)	0.00 (0.000)	0.00 (0.000)	

Table 4 – Ratios for Bundesliga.



However, it is important to note that Chi-squared tests do not reject independence between home and away goals for any league. The Chi-squared test for independence is a statistical test used to determine whether there is a significant association between two variables. The null hypothesis assumes that the two variables are independent and the alternative that they are associated. It requires the contingency table of goals. For the three seasons of Eredivisie this table is the following:

H \ A	0	1	2	3	4	5	6	7	8
0	55	65	33	29	14	9	1	0	1
1	63	96	62	29	21	2	2	1	0
2	60	63	49	21	7	3	0	0	0
3	39	54	19	12	4	0	0	0	0
4	26	16	9	4	0	0	0	0	0
5	16	11	7	1	0	1	0	0	0
6	4	5	0	0	0	0	0	0	0
7	0	2	1	0	0	0	0	0	0
9	1	0	0	0	0	0	0	0	0

Table 5 – Contingency table of Eredivisie data.

The test statistic of the Chi-squared test is:

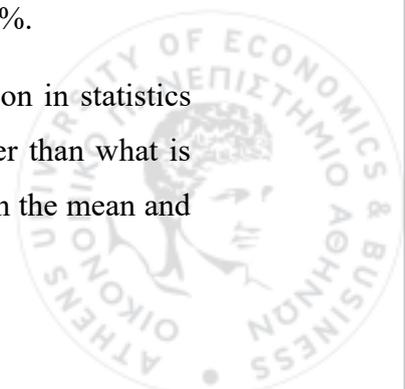
$$x^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i},$$

where O_i = observed frequency in category i and E_i = expected frequency in category i assuming independence, calculated for every cell as (row total) \times (column total) / grand total and k the number of categories.

The degrees of freedom are calculated as $df = (\text{number of rows} - 1) \times (\text{number of columns} - 1)$. Thus, for the corresponding degrees of freedom we find the critical value for which we reject the null hypothesis using a Chi-squared distribution table. The p-value is the probability of getting a Chi-squared value as extreme as (or more extreme than) the observed x^2 , assuming the null hypothesis is true. Formally: p-value = $P(x_{df}^2 \geq \text{observed } x^2)$. This is the right-tail probability of the Chi-squared distribution.

The p-values are 0.3217, 0.8438, 0.189 and 0.3589 for all three seasons of Premier League, La Liga, Eredivisie and Bundesliga respectively. Thus, we cannot reject the null hypothesis of independence for a significance level $\alpha = 5\%$.

Another point of the data regards overdispersion. Overdispersion in statistics refers to a situation where the observed variance in a dataset is greater than what is expected based on the chosen statistical model. In a Poisson distribution the mean and



the variance are assumed to be equal. Figure 4 presents the mean and the variance of goals scored by each team across the 2021/22, 2022/23 and 2023/24 seasons. Each blue point represents a team, where the x-axis corresponds to the mean numbers of goals that they scored (regardless of playing at home or away) and the y-axis to the variance of these goals. The diagonal dotted line represents mean and variance equivalence. For all four leagues there are several points that lie above the diagonal line. This may indicate that there is some overdispersion in the data, and possibly the negative binomial dist. would be more appropriate than the Poisson.

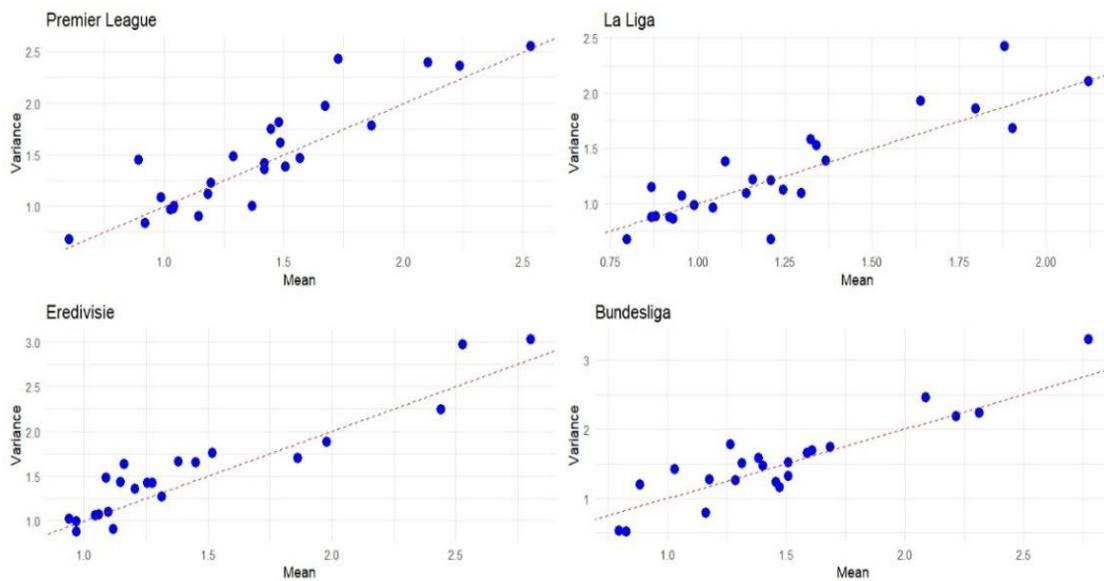


Figure 4 – Mean and variance of goals scored by each team.

4.2 Estimation

In this section, we fit the Sarmanov with negative binomial marginals model

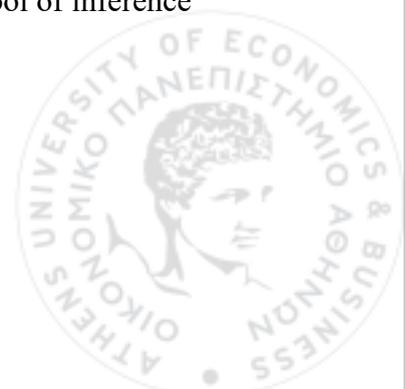
$P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1) P(X_2 = x_2) (1 + \omega q_{nb}^{(1)}(x_1) q_{nb}^{(1)}(x_2))$ where:

$$q_{nb}^{(1)}(x_i) = \begin{cases} -\varphi_i \left(\frac{\mu_i}{\varphi_i + \mu_i} \right) & \text{if } x_i = 0 \\ 1 & \text{if } x_i = 1, \text{ for } i = 1, 2, \\ 0 & \text{if } x_i = 2, 3, \dots \end{cases}$$

to data from the 2023/24 Eredivisie season.

First, we define the log-likelihood function which is our basic tool of inference through Maximum Likelihood Estimation in R:

$$\ell = \sum_{i=1}^N [\log P(X_{1i} | \varphi_1, \mu_{1i}) + \log P(X_{2i} | \varphi_2, \mu_{2i}) + \log(1 + \omega q_{1i}(X_{1i} | \varphi_1, \mu_{1i}) q_{2i}(X_{2i} | \varphi_2, \mu_{2i}))],$$



for $i = 1, \dots, N$ number of matches.

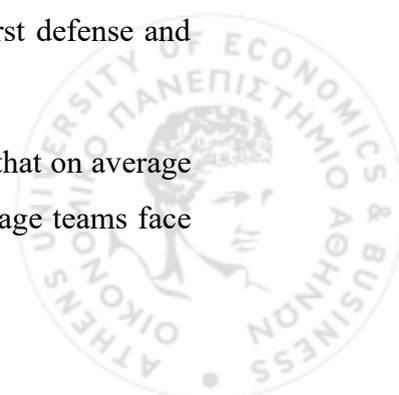
In the model for n teams there are n number of attacking parameters (a_1, \dots, a_n) and defence parameters (β_1, \dots, β_n), a home advantage parameter, a dependence parameter ω and dispersion parameters φ_1 and φ_2 . To prevent the model from being overparameterized the following constraint is imposed: $\sum_{i=1}^n a_i = 0$.

With the use of “optim” function in R we get the Maximum Likelihood Estimates for the parameters of the model. These include the attacking and defensive parameters, the home advantage parameter, dependence parameter ω and dispersion parameters φ_1 and φ_2 .

Attacking and defensive abilities represent how much better a team does compared to the average attacking or defensive level. Larger positive attacking parameter values indicate that a team is performing better than the average attacking level of the teams in the league and larger (in absolute value) negative defensive parameter values suggest that a team defends better than average meaning that they concede fewer goals in general. Table 6 and 7 present the values of these parameters with standard errors given in parentheses. The sum-to-zero constraint applied on the attacking parameters leads to calculating $n-1$ attacking parameters and the last attacking parameter is calculated as the negative sum of the $n-1$ parameters. Thus, the last parameter to be calculated in Table 6 does not have a standard error: in this case RKC Waalwijk.

Champions PSV Eindhoven which scored the most goals have the highest attacking parameters followed by second placed Feyenoord. Scoring 111 goals is reflected in their 0.732 parameter. Feyenoord scored 92 goals and achieved a 0.559 attacking parameter. Twente, Alkmaar and Ajax with 69, 70 and 74 goals respectively have attacking parameters that are significantly lower than those of the first two teams. PSV had also a remarkable defensive performance with the best defensive parameter value. They conceded just 21 goals and their defensive parameter of -0.598 is by far the best in the league. On the other hand, relegated Volendam and Vitesse naturally have bad attacking and defensive performance with the first having the worst defense and the latter having the worst attack output.

Home advantage parameter value of 0.186 in Table 7 indicates that on average the home team scores around 20% ($e^{0.186}$) more goals when two average teams face



each other. Parameter value of ω which is 0.185 indicates that there is some dependence between home and away goals.

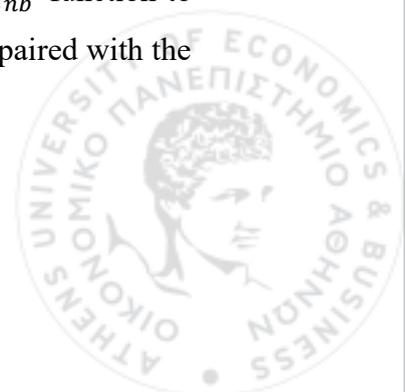
Team	Attack	Defence
Ajax	0.375 (0.115)	0.454 (0.133)
Almere City	-0.406 (0.167)	0.371 (0.135)
AZ Alkmaar	0.286 (0.0.117)	0.000 (0.164)
Excelsior	0.003 (0.137)	0.611 (0.122)
Feyenoord	0.559 (0.104)	-0.388 (0.198)
Fortuna Sittard	-0.292 (0.158)	0.334 (0.138)
Go Ahead Eagles	-0.090 (0.141)	0.139 (0.151)
Heerenveen	0.039 (0.134)	0.562 (0.125)
Heracles	-0.220 (0.150)	0.608 (0.122)
NEC Nijmegen	0.266 (0.119)	0.249 (0.144)
PSV Eindhoven	0.732 (0.095)	-0.598 (0.221)
RKC Waalwijk	-0.340 (-)	0.313 (0.138)
Sparta Rotterdam	-0.013 (0.136)	0.163 (0.149)
FC Twente	0.246 (0.118)	-0.097 (0.170)
FC Utrecht	-0.084 (0.138)	0.134 (0.150)
FC Volendam	-0.398 (0.112)	0.766 (0.112)
Vitesse	-0.527 (0.174)	0.601 (0.122)
PEC Zwolle	-0.134 (0.144)	0.498 (0.127)

Table 6 – Attack and defence parameters of 2023/24 Eredivisie teams.

Home advantage parameter	0.186 (0.063)
Dependence parameter ω	0.185 (0.093)
Dispersion parameter φ_1	39389.32 (765.78)
Dispersion parameter φ_2	30324.16 (230.54)

Table 7 – Rest model parameter values.

A reasonable question that might arise is why we choose the $q_{nb}^{(1)}$ function to obtain the parameter estimates and not some other function that can be paired with the negative binomial marginals like:



$$q_{nb}^{(2)}(x_i) = \begin{cases} -\mu_i^2 & \text{if } x_i = 0 \\ \mu_i \frac{\varphi_i + \mu_i}{\varphi_i} & \text{if } x_i = 1 \\ 0 & \text{if } x_i = 2, 3, \dots \end{cases} \quad \text{or} \quad q_{nb}^{(4)}(x_i) = \begin{cases} -\mu_i^2 & \text{if } x_i = 0 \\ -\mu_i \frac{\varphi_i + \mu_i}{\varphi_i} & \text{if } x_i = 1 \\ 4 \frac{(\mu_i + \varphi_i)^2}{\varphi_i^2 + \varphi_i} & \text{if } x_i = 2 \\ 0 & \text{if } x_i = 3, 4, \dots \end{cases},$$

for $i = 1, 2$.

These two move probability across scores in different ways. Function $q_{nb}^{(2)}$ introduces a quadratic term and $q_{nb}^{(4)}$ shifts probability across more scores than the four pairs (0-0, 0-1, 1-0 and 1-1). The answer to this question is that when using these q -functions the ω parameter that is derived from the optimization process is close to 0.

Specifically, in Table 8 it is clear that parameter ω is optimized to small values when using $q_{nb}^{(2)}$ or $q_{nb}^{(4)}$ regardless of the 2023/24 data of the four leagues used to calculate it. In a model with the form $h(x_1, x_2) = P_1(x_1) P_2(x_2) [1 + \omega q_1(x_1) q_2(x_2)]$, when ω is close to 0, the term $[1 + \omega q_1(x_1) q_2(x_2)]$ is close to 1, leading to something almost identical to the case of two independent negative binomial marginals.

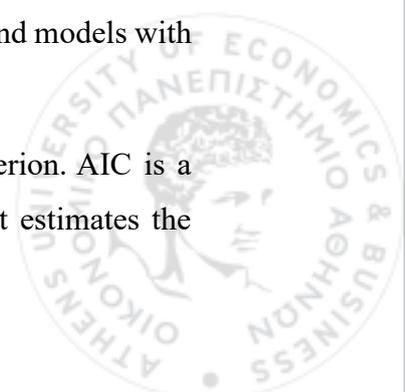
	Eredivisie	Bundesliga	La Liga	Premier League
$q_{nb}^{(1)}$	0.185 (0.093)	0.169 (0.074)	0.082 (0.071)	0.020 (0.076)
$q_{nb}^{(2)}$	0.047 (0.07)	0.051 (0.035)	0.025 (0.032)	-0.002 (0.033)
$q_{nb}^{(4)}$	0.008 (0.011)	0.011 (0.011)	0.015 (0.014)	0.002 (0.008)

Table 8 – Parameter ω values for different q -functions and different data sets.

4.3 Model comparison

In this section we fit the models to data of the 2023/24 season of the aforementioned leagues. Starting from independence with Poisson marginals, the model that Maher (1982) suggested, we then explore the fit of the model with Poisson marginals and q_{ac} which is the model proposed by Dixon and Coles (1997) and the model with Poisson marginals and $q_{pois}^{(1)}$ which involves a quadratic term. Furthermore, we compare these models with the independent negative binomial one and models with negative binomial marginals and the q -functions examined previously.

The main point of comparison is the Akaike Information Criterion. AIC is a metric developed by Japanese statistician Hirotugu Akaike in 1974. It estimates the



relative quality of a statistical model for a given set of data. It is used to compare multiple models, not to test a model in isolation. It helps compare different models and choose the one that best balances goodness of fit with model simplicity. Lower AIC values indicate better models.

AIC is calculated using the formula:

$$\text{AIC} = 2k - 2 \ell(\hat{\theta}),$$

where: k is the number of parameters in the model and $\ell(\hat{\theta})$ is the maximized log-likelihood of the model which is obtained using command “optim” in R.

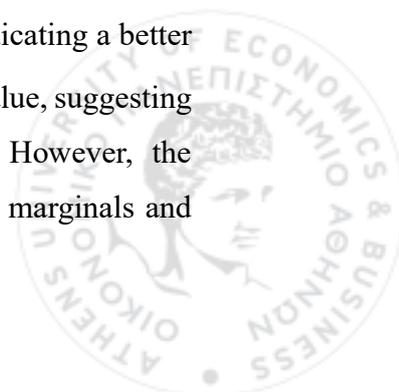
The log-likelihood represents how well the model fits the data. A higher log-likelihood means the model better explains the observed data. However, higher likelihoods often come from more complex models. Each parameter in a model increases its complexity, possibly though, leading to overfitting. This occurs when a model becomes too complex and captures noise rather than the underlying pattern.

Thus, AIC tries to avoid overfitting by balancing both goodness of fit and complexity. Term $2 \ell(\hat{\theta})$ rewards goodness of fit and the term $2k$ penalizes complexity. So, the model with lower AIC value is generally preferred.

Table 9 presents the number of parameters, the log-likelihood values and the AIC values of the models for the 2023/24 Eredivisie data. The lower AIC values for the Poisson models' group and the negative binomial models' group are underlined and the lowest value in general is in bold.

For the Eredivisie data with $n=18$ teams, the independent double Poisson model estimates 36 parameters: $n-1=17$ attacking parameters, 18 defense parameters and a home advantage parameter. The models with q_{dc} and $q_{pois}^{(1)}$ also have those parameters plus the dependence parameter ω . The double negative binomial model has 17 attacking parameters, 18 defence and two dispersion parameters φ_1 and φ_2 . The negative binomial models with q -functions have all these parameters and additionally an ω parameter.

The Poisson model with $q_{pois}^{(1)}$ has the highest log-likelihood indicating a better fit to the data than the alternatives. This model also has the lowest AIC value, suggesting it provides the best trade-off between model fit and complexity. However, the differences between this model and the model with negative binomial marginals and



$q_{nb}^{(2)}$ are significantly small, both for Log-likelihood and AIC, indicating that the Poisson model is not decisively better.

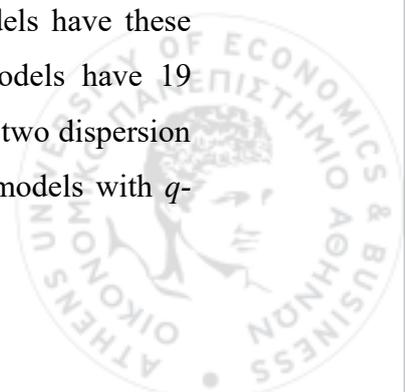
This minimal difference suggests that while Poisson with $q_{pois}^{(1)}$ is slightly better, negative binomial models could still be preferred if overdispersion is theoretically or empirically justified (a fact that is plausible according to Figure 4). A negative binomial model may provide more robust estimates in the presence of excess variability.

Eredivisie			
	No. of parameters	Log-likelihood	AIC
Independent double Poisson	36	-899.236	1870.473
Poisson with q_{dc}	37	-896.654	1867.309
Poisson with $q_{pois}^{(1)}$	37	-894.814	<u>1863.629</u>
Indep. double neg. binom.	38	-899.237	1874.475
Neg. binom. with $q_{nb}^{(1)}$	39	-896.656	1871.313
Neg. binom. with $q_{nb}^{(2)}$	39	-894.828	<u>1867.658</u>
Neg. binom. with $q_{nb}^{(4)}$	39	-897.885	1873.771

Table 9 – Details of the fitted models for the Eredivisie data.

For the Bundesliga data with $n=18$ teams, the models have the same number of parameters as for the Eredivisie data. Table 10 presents a similar situation as previously. Again, the best performing model is a Poisson one, the one with q_{dc} which creates the Dixon and Coles model. It has the highest log-likelihood and the lowest AIC. However, we see that once more there is a negative binomial marginals model, the one with $q_{nb}^{(1)}$, that performs almost equivalently good in terms of log-likelihood, while the difference in AIC is notable but not conclusive. In a case where there is overdispersion in the data the negative binomial marginals model might be preferable.

In the La Liga dataset we have $n=20$ teams. Thus, the double Poisson model estimates $n-1=19$ attacking parameters, 20 defence parameters and a home advantage parameter, summing up to a total of 40. The other two Poisson models have these parameters plus an ω parameter. The double negative binomial models have 19 attacking and 20 defence parameters, a home advantage parameter and two dispersion parameters (φ_1 and φ_2), a total of 42. Finally, the negative binomial models with q -



functions add a dependence parameter to the 42 of the independent one so these have 43 parameters. In contrast to Eredivisie and Bundesliga, where a Poisson model was performing slightly better, we observe in Table 11 that the 2023/24 La Liga dataset seems to be better modeled with a negative binomial model, the one with $q_{nb}^{(4)}$ which has the highest log-likelihood and the lowest AIC. However, the Poisson models present similar performance with this model in terms of AIC.

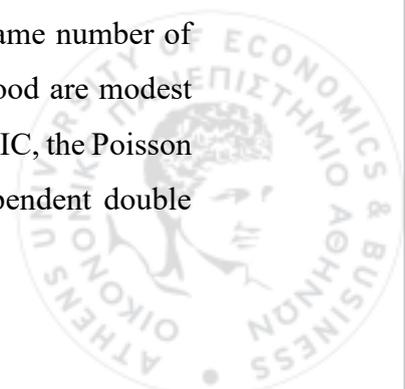
Bundesliga			
	No. of parameters	Log-likelihood	AIC
Independent double Poisson	36	-894.187	1860.374
Poisson with q_{dc}	37	-891.918	<u>1857.837</u>
Poisson with $q_{pois}^{(1)}$	37	-892.279	1858.560
Indep. double neg. binom.	38	-894.19	1864.38
Neg. binom. with $q_{nb}^{(1)}$	39	-891.921	<u>1861.843</u>
Neg. binom. with $q_{nb}^{(2)}$	39	-892.283	1862.566
Neg. binom. with $q_{nb}^{(4)}$	39	-892.824	1863.649

Table 10 – Details of the fitted models for the Bundesliga data.

La Liga			
	No. of parameters	Log-likelihood	AIC
Independent double Poisson	40	-1050.265	<u>2180.531</u>
Poisson with q_{dc}	41	-1049.602	2181.204
Poisson with $q_{pois}^{(1)}$	41	-1049.985	2181.969
Indep. double neg. binom.	42	-1050.266	2184.532
Neg. binom. with $q_{nb}^{(1)}$	43	-1049.603	2185.206
Neg. binom. with $q_{nb}^{(2)}$	43	-1049.995	2185.989
Neg. binom. with $q_{nb}^{(4)}$	43	-1046.895	<u>2179.79</u>

Table 11 – Details of the fitted models for the La Liga data.

Finally, the models for the Premier League data estimate the same number of parameters as the La Liga models. Here, the differences in log-likelihood are modest with differences only in the second and third decimal place. In terms of AIC, the Poisson models seem to outperform the negative binomial models. The independent double



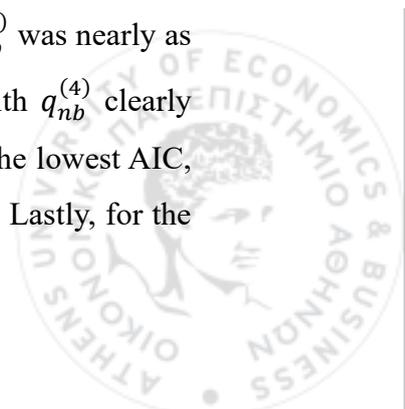
Poisson model is the best performing model among the Poisson ones, and the independent double negative binomial model is the best performing among negative binomial models.

Premier League			
	No. of parameters	Log-likelihood	AIC
Independent double Poisson	40	-1135.286	<u>2350.571</u>
Poisson with q_{dc}	41	-1135.251	2352.502
Poisson with $q_{pois}^{(1)}$	41	-1135.283	2352.566
Indep. double neg. binom.	42	-1135.287	<u>2354.574</u>
Neg. binom. with $q_{nb}^{(1)}$	43	-1135.254	2356.508
Neg. binom. with $q_{nb}^{(2)}$	43	-1135.284	2356.568
Neg. binom. with $q_{nb}^{(4)}$	43	-1135.226	2356.452

Table 12 – Details of the fitted models for the Premier League data.

This might be a sign that there is not strong dependence between home and away goals. In Table 8 we saw that the ω parameter is estimated at 0.02 when using negative binomial marginals and $q_{nb}^{(1)}$, at -0.002 when using $q_{nb}^{(2)}$ and at 0.002 when using $q_{nb}^{(4)}$. These values are relatively small for a dependence parameter, thus possibly explaining the better performance of the independence models.

Across the four major European football leagues analyzed for the 2023/24 season—Eredivisie, Bundesliga, La Liga, and Premier League—model performance varied slightly, with Poisson-based models generally performing better in Eredivisie and Bundesliga, while a negative binomial model stood out in La Liga. Specifically, the Poisson model with $q_{pois}^{(1)}$ achieved the best fit for the Eredivisie data, offering the lowest AIC and highest log-likelihood, though the negative binomial model with $q_{nb}^{(2)}$ performed nearly as well, suggesting the potential benefit of modeling overdispersion. In the Bundesliga, the Poisson model with q_{dc} (Dixon-Coles model) had the best AIC and log-likelihood, although again, a negative binomial model with $q_{nb}^{(1)}$ was nearly as competitive. Conversely, in La Liga, the negative binomial model with $q_{nb}^{(4)}$ clearly outperformed all others, showing both the highest log-likelihood and the lowest AIC, providing stronger indication of overdispersion in Spanish league data. Lastly, for the



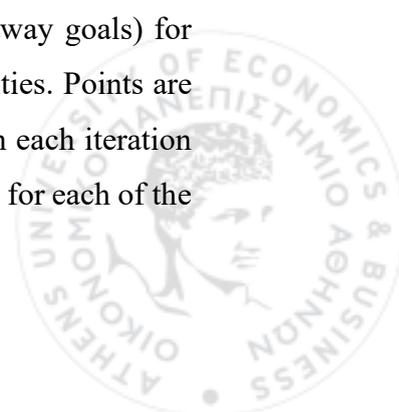
Premier League, differences in model fit were minimal, with the independent double Poisson model slightly outperforming others, including both dependent Poisson and negative binomial models. This suggests little evidence of goal dependence or overdispersion in the Premier League data, as also supported by the near-zero estimates of the dependence parameter ω . Overall, the choice between Poisson and negative binomial models appears to depend on the league-specific characteristics such as dispersion and goal dependence structure.

4.4 Prediction

In this section we examine the predictive capabilities of the models. Specifically, we split the data to two parts: most of them are used as a “train” set, meaning that this is going to be used for parameter estimation, and the second part will be a “test” set, which will be used for evaluation of the predictive performance of the respective models. The teams’ dynamics are calculated using the first matchdays and then are used to perform 1,000 Monte Carlo simulations of the final matchdays. Thus, we can examine the predictive ability of the models by comparing the simulation results with the actual results.

However, the form of the Sarmanov family model, which is given by $P_1(x_1)P_2(x_2) [1 + \omega q_1(x_1)q_2(x_2)]$ does not permit direct sampling from the joint distribution. As a result, Monte Carlo simulation methods are required to generate samples from this distribution.

First, we fit the model to the first twenty-four matchdays of the 2023/24 Bundesliga season, leaving the last ten to be predicted. The model, using the train set which is the first 216 matches of the dataset, estimates the attacking and defense parameters, the home advantage parameter, the dependence parameter ω and finally the dispersion parameters φ_1 and φ_2 . Then, these parameters are used to create a score probability matrix for each individual match in the test set, meaning that a probability is assigned to every possible score, which is a pair of numbers of home and away goals. To simulate the results of the ten final matchdays we use a rather simple but effective algorithm that randomly draws a result (a combination of home and away goals) for every match of the final ten matchdays based on the assigned probabilities. Points are assigned accordingly to teams (3-1-0). This happens for every match in each iteration out of the total 1,000. Thus, we end up with 1,000 simulated final points for each of the



18 teams. From these final points we calculate the mean number of points that each team got through the 1,000 iterations, as well as the 2,5% and 97,5% quantiles to get the 95% prediction intervals of each team.

In this study, prediction intervals were derived using empirical quantiles computed in R. The quantiles were obtained using the default interpolation method implemented in R's `quantile()` function, corresponding to Type 7 of the nine quantile definitions proposed by Hyndman and Fan (1996).

All sample quantiles $Q(p)$ are calculated as weighted averages of two consecutive ordered observations, $x_{(j)}$ and $x_{(j+1)}$:

$$Q(p) = (1 - \gamma)x_{(j)} + \gamma x_{(j+1)},$$

where p is the quantile probability for which it stands that $\frac{j-m}{n} \leq p < \frac{j-m+1}{n}$, n is the sample size, x_j is the j^{th} order statistic, the value of γ is a function of $j = \lfloor np + m \rfloor$ and $g = np + m - j$ and m is a constant that is defined by the quantile type.

For Type 7, used by default in R:

$$m = 1 - p, \text{ and } p_k = \frac{k-1}{n-1}.$$

This formula assigns a probability position p_k to each ordered value $x_{(k)}$.

Hence, the estimated quantile is obtained by linearly interpolating between the order statistics at the fractional index $h = 1 + (n - 1)p$, that is:

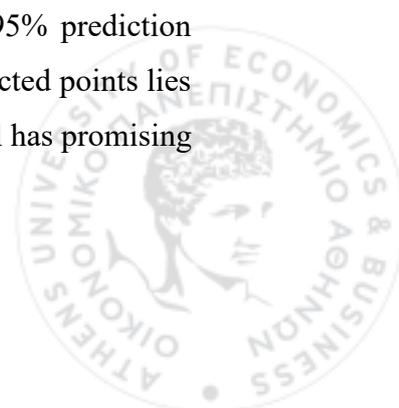
$$Q(p) = (1 - \gamma)x_{(\lfloor h \rfloor)} + \gamma x_{(\lfloor h \rfloor + 1)},$$

with $\gamma = h - \lfloor h \rfloor$.

Finally, we calculate the actual points that each team accumulated in the final ten matchdays in order to compare them with the predicted points.

Figure 5 presents the above when using the q_1 function along with negative binomial marginals. For each of the 18 teams in the Bundesliga season the blue dots represent the mean number of points that each team accumulated in the final ten matchdays through the 1,000 Monte Carlo simulations. The red triangles are the actual numbers of points of each team and finally the light-blue bars represent the 95% prediction intervals.

The actual points for all teams except Mainz lie within the 95% prediction intervals. Furthermore, for most of the teams the mean number of predicted points lies relatively close to the actual number of points, indicating that the model has promising predictive performance.



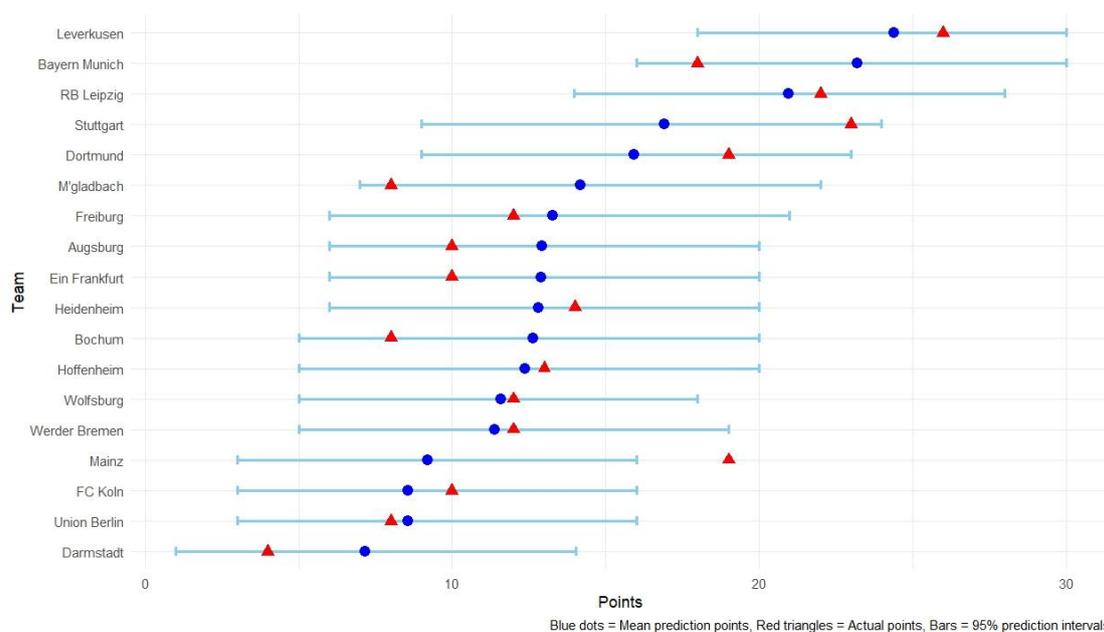


Figure 5 – Prediction under the q_1 – neg. binom. model.

There is only one “outlier” that stands out, Mainz. Their actual points lie outside the prediction intervals meaning that the model could not predict accurately their performance in the last ten matchdays.

To investigate why this happens we looked up the results of their matches across the season. In the first twenty-four matchdays which make up the “training” set, Mainz had a record of two wins, ten draws and twelve defeats scoring 19 goals and conceding 38. As a result, the model during the estimation phase assigns weak attacking and defensive parameters to Mainz, leading the prediction for their ten last matchdays to be worse than their actual performance: in these matches Mainz suffered just one defeat, a heavy 8-1 from Bayern Munich. They won five of these matches and drew four. They scored 20 goals and conceded just 13. It is obvious that the performance of Mainz drastically changed in their last ten matches. One possible reason for this change might lie in Mainz changing their manager not once but twice: They started the season with Bo Svensson until the ninth matchday, when Jan Siewert took charge until the twenty-first matchday and finally they ended the season with Bo Henriksen who led them to six wins out of their thirteen final matches. In football, changing manager often results in an improved performance of the team and this is not easy to predict using this statistical model which relies on goals scored and conceded.

When using q_2 the situation remains similar. In Figure 6 all the predictions, except for Mainz, lie again within the 95% prediction interval. This suggests similar predictive capabilities of the q_1 and the q_2 models.

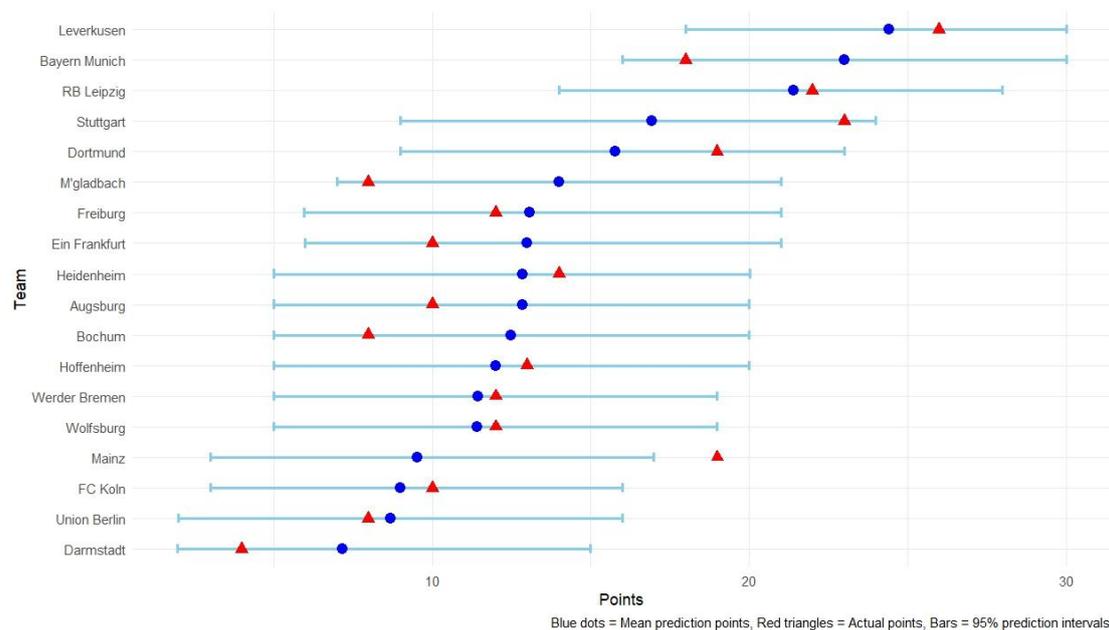


Figure 6 – Prediction under the q_2 – neg. binom. model.

Finally, in Figure 7 the prediction capabilities of the model with q_4 seem adequate. Indeed, the mean prediction of points for some teams like Hoffenheim, Wolfsburg and Werder Bremen is closer to their actual points than it is when using the previous q -functions.

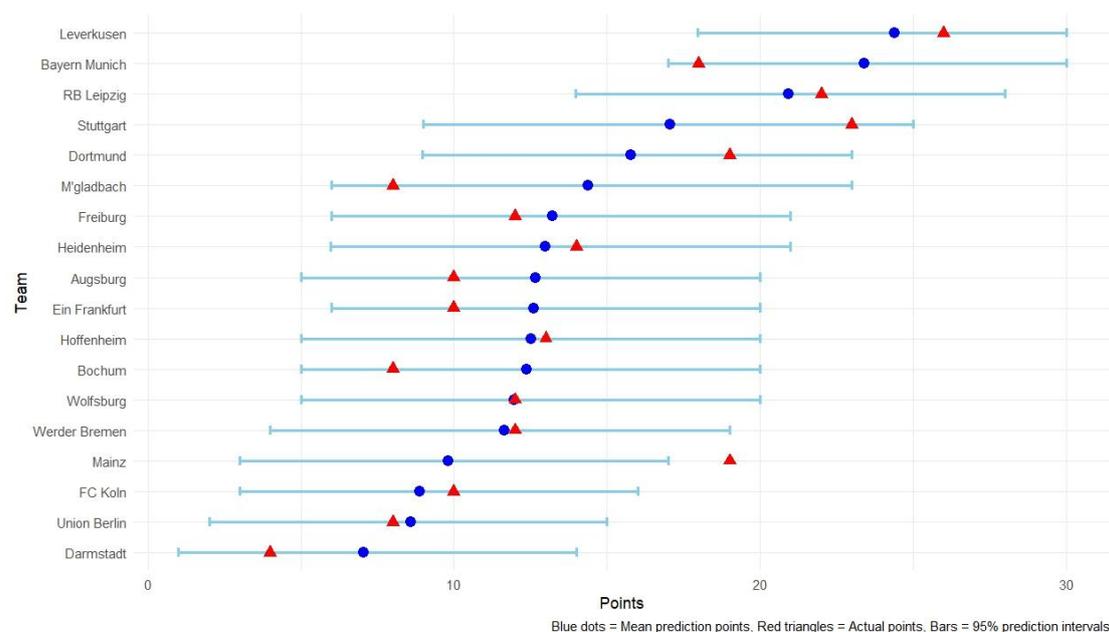
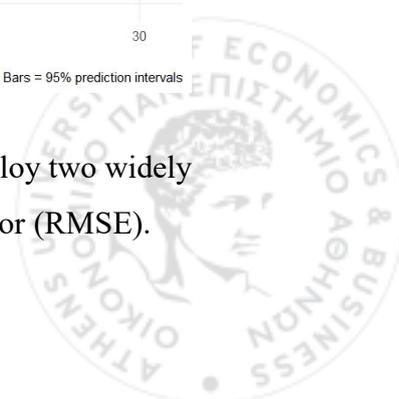


Figure 7 – Prediction under the q_4 – neg. binom. model.

In order to compare the predictions of the above models we employ two widely used metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).



MAE measures the average magnitude of errors in a set of predictions, without considering their direction (positive or negative). It's the average of the absolute differences between predicted values and actual values. It is given by the following formula:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|,$$

where n is the number of data points, \hat{y}_i are the predicted values and y_i the actual values. MAE provides straightforward interpretation of the average error as it is on the same scale as the original data. For example, if MAE is 5, the model's predictions are on average 5 units off from the actual values.

RMSE is the square root of the average squared differences between predicted and actual values. It penalizes large errors more heavily because of squaring. This means that this metric is more sensitive to outliers. It is given by the formula:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2},$$

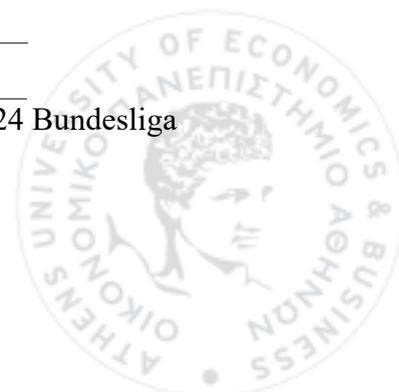
where n is the number of data points, \hat{y}_i are the predicted values and y_i the actual values. Again, RMSE is also on the same scale as the original data.

In Table 13 we can observe the MAE and RMSE of mean prediction points and actual points. The MAE values suggest that the models' predictions are on average around 2.9 points off the actual points of the teams. RMSE which penalizes more the larger errors suggest that the models' predictions are around 3.8 off the actual results. In football forecasting, where randomness in goals and points is high, that's quite decent performance.

In general, we see that the model with q_4 has the lowest MAE meaning that on average its predictions are slightly closer to the actual values. It also has the lowest RMSE meaning that it is handling outliers or large deviations a bit better. However, the differences between q_1 , q_2 and q_4 are small indicating that all three models are performing similarly with q_4 being slightly more accurate.

Model	MAE	RMSE
q_1 – neg. binom	2.901	3.799
q_2 – neg. binom	2.887	3.798
q_4 – neg. binom	2.863	3.764

Table 13 – MAE and RMSE values for the fitted models for the 2023/24 Bundesliga data.



In conclusion, the predictive evaluation of the Sarmanov family model across different q -functions demonstrates encouraging performance, with most teams' actual points falling within the 95% prediction intervals and the mean simulated outcomes closely matching the real results. The relatively low values of MAE and RMSE across all q -functions further confirm that the models are not only accurate on average but also manage larger deviations effectively, reinforcing their reliability. This suggests that the models are capable enough of capturing the teams' dynamics in the Bundesliga. The consistent outlier, Mainz, highlights a critical limitation of the model, its inability to anticipate off-field changes such as managerial shifts that can drastically alter a team's trajectory. Nonetheless, the fact that the predictions hold well across different q -functions reinforces the model's robustness and adaptability. While future enhancements might aim to incorporate external factors like managerial changes or player transfers, the current approach provides a solid foundation for probabilistic forecasting in football match outcomes.



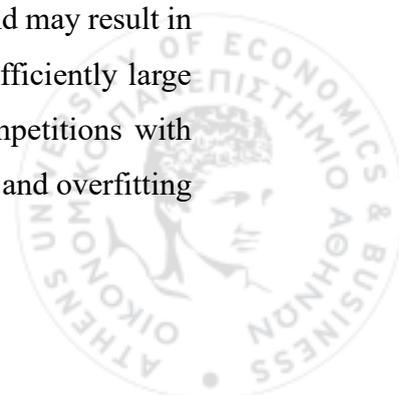
5. Conclusions – Further Research

The Dixon and Coles model has been widely used when modelling the number of goals of football matches. It is the first model that provided the ability to shift probability to some scores in order to better fit the actual data. Michels, Ötting, and Karlis proved that this model is in fact a special case of the Sarmanov family of distributions. Relying on the paper of these authors “Extending the Dixon and Coles model: an application to women’s football data” (2023) in which they extend the foundational Dixon and Coles model, in this thesis we examined thoroughly some Sarmanov family models with negative binomial marginals. This approach addresses two significant limitations of the original model: the inability to use marginals other than Poisson, which can be restrictive when there is overdispersion in goal counts and the restriction in shifting probabilities only among a limited set of scorelines.

Through theoretical derivation, empirical estimation, and practical application to data from major European football leagues, we demonstrated that the extended models offer greater flexibility. The use of various q -functions allowed for richer dependency structures between home and away team goals, possibly capturing complex interactions and reflecting real-world scoring behaviours more accurately.

Model comparisons however, showed that there is no clear “winner” among the models with Poisson and negative binomial marginals. All models presented relatively similar performance. In some cases, a Poisson model performed better in terms of AIC and log-likelihood and in some others a negative binomial one. Finally, the negative binomial models seem able to predict with relative accuracy the final ten matchdays of the 2023/24 Bundesliga season, indicating that potentially these models could be used in practise in fields like sports betting or by football analysts who assess team strengths and dynamics.

Despite these encouraging results and the demonstrated flexibility of the negative binomial Sarmanov models, certain limitations must be acknowledged. First, the model’s complexity is inherently higher than that of the traditional Dixon and Coles formulation. The inclusion of additional parameters, such as the dispersion parameters and the dependence parameter ω , increases the computational burden and may result in convergence issues during estimation. This type of models demand sufficiently large datasets for reliable parameter estimation. For smaller leagues or competitions with limited numbers of matches, parameter estimates may become unstable, and overfitting



may arise. Additionally, while the Sarmanov framework offers interpretability through its constructive nature, the intuition behind the chosen q -functions can become less transparent as their structure grows more elaborate.

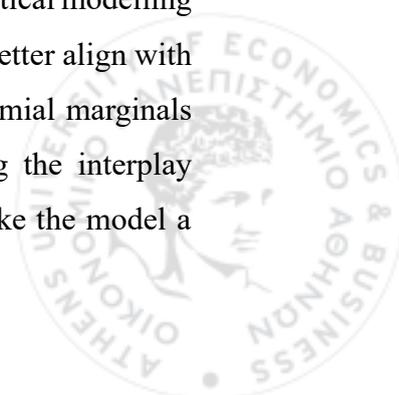
Furthermore, while the current model assumes static team parameters across a season, incorporating time-varying effects, as the dynamic extensions of the original Dixon and Coles model where more weight is given to the recent matches, could further improve predictive performance. Allowing team parameters to vary over time or adopting a Bayesian hierarchical framework could accommodate temporal evolution and provide a way to quantify parameter uncertainty. Second, integrating covariate information such as recent team form, player statistics, weather conditions, betting market odds etc. could enhance the model's explanatory power. To these covariates one could also add some more off-field parameters such as player transfers or managerial changes.

The empirical evaluation of this thesis was limited to some of the top European leagues. The performance of the proposed models in lower divisions, women's football, or international tournaments remains an open question and could exhibit different degrees of overdispersion or dependence.

Several avenues for future research emerge from this study. One natural extension involves the use of copula functions to model the dependence between home and away goals. Copulas allow for more general, possibly nonlinear dependence structures and can capture both symmetric and asymmetric tail behavior. Possibly a hybrid Sarmanov – Copula framework could combine the interpretability of the Sarmanov approach with the flexibility of copulas.

Finally, future work could systematically compare Sarmanov-based models with alternative approaches, such as copula-based models, bivariate Poisson formulations, or even modern machine learning algorithms, using common predictive evaluation metrics (e.g., log-likelihood or Brier score). Extending the application of the proposed model to other low-scoring sports, such as hockey or handball, would also be of practical and theoretical interest.

In summary, this thesis offers an addition in the literature of statistical modelling of football match outcomes by extending an established framework to better align with the characteristics of real-world data. The integration of negative binomial marginals within the Sarmanov family provides greater flexibility in capturing the interplay between home and away team performances. These enhancements make the model a

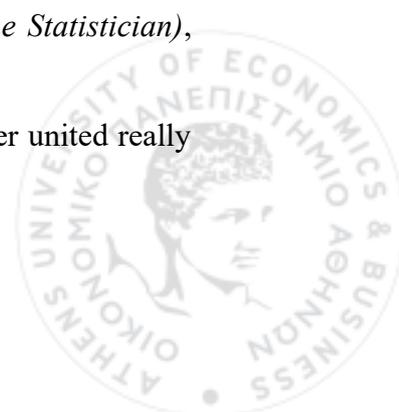


valuable addition to the toolkit of sports statisticians, analysts, and practitioners alike. As the demand for robust, interpretable, and data-driven insights in football continues to grow, whether for strategic planning, performance evaluation or informed betting decisions, models such as those developed here can play a key role in shaping the future of sports analytics.



References

- Baker, R. and Scarf, P. (2006).** Predicting the outcomes of annual sporting contests. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 55(2):225-239.
- Boshnakov, G., Kharrat, T., and McHale, I. G. (2017).** A bivariate Weibull count model for forecasting association football scores. *International Journal of Forecasting*, 33(2):458 – 466.
- Carmichael, F. and Thomas, D. (2005).** Home-field effect and team performance: evidence from english premiership football. *Journal of Sports Economics*, 6(3):264-281.
- Dixon, M. J. and Coles, S. G. (1997).** Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2):265 – 280.
- Famoye, F. (2010).** On the bivariate negative binomial regression model. *Journal of Applied Statistics*, 37(6):969-981.
- Goddard, J., and Asimakopoulos, I. (2004).** Forecasting football match results and the efficiency of fixed-odds betting. *Journal of Forecasting*, 23, 51– 66.
- Goddard, J. (2005).** Regression models for forecasting goals and match results in association football. *International Journal of Forecasting*, 21, 331-340.
- Groll, A., Kneib, T., Mayr, A., and Schauburger, G. (2018).** On the dependency of soccer scores-a sparse bivariate Poisson model for the UEFA European Football Championship 2016. *Journal of Quantitative Analysis in Sports*, 14(2):65-79.
- Hill, I. D. (1974).** Association Football and Statistical Inference. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 23(2), 203–208.
- Hyndman, R. J., & Fan, Y. (1996).** *Sample quantiles in statistical packages*. The American Statistician, 50(4), 361–365.
- Karlis, D. and Ntzoufras, I. (2003).** Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):381 – 393.
- Lee, A. J. (1997).** Modeling scores in the premier league: is manchester united really the best? *Chance*, 10(1):15 – 19.



- Maher, M. J. (1982).** Modelling association football scores. *Statistica Neerlandica*, 36(3):109 – 118.
- McHale, I. and Scarf, P. (2007).** Modelling soccer matches using bivariate discrete distributions with general dependence structure. *Statistica Neerlandica*, 61(4):432 – 445.
- McHale, I. and Scarf, P. (2011).** Modelling the dependence of goals scored by opposing teams in international soccer matches. *Statistical Modelling*, 11(3):219 – 236.
- Michels, R., Ötting, M., & Karlis, D. (2023).** Extending the Dixon and Coles model: an application to women’s football data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 74(1):167-186.
- Nelsen, R.B. (2006).** An Introduction to Copulas, *2nd Edition*. Springer, New York.
- Reep, C., and Benjamin, B. (1968).** Skill and Chance in Association Football. *Journal of the Royal Statistical Society. Series A (General)*, 131(4), 581–585.
- Reep, C., Pollard, R. and Benjamin, B. (1971).** Skill and Chance in Ball Games. *Journal of the Royal Statistical Society. Series A (General)*. 134, 623–629.
- Rue, H., and Salvesen, O. (2000).** Prediction and retrospective analysis of soccer matches in a league. *Statistician*, 49,399–418.
- Sarmanov, O. V. (1966).** Generalized normal correlation and two-dimensional Frechet classes. In *Doklady Akademii Nauk*, volume 168, pages 32-35. Russian Academy of Sciences.
- Sklar, A. (1973).** Random variables, joint distribution functions, and copulas. *Kybernetika*, 9(6):449 - 460.
- Ting Lee, M.-L. (1996).** Properties and applications of the Sarmanov family of bivariate distributions. *Communications in Statistics-Theory and Methods*, 25(6):1207-1222.

