Chapter 1

Introduction

Joint modelling is a methodology which can be suitable for a wide range of applications. Wherever one might have two or more dependent variables, the question arises on how to account for dependencies between these variables. This chapter will introduce a few examples of situations, where such an approach could be deemed useful. The historical development for single examples and a stepwise improvement from rather simple models to more complicated ones will be shown, incorporating different forms of interdependency.

- *The sales numbers of competing products in a given time frame and location.* The marginal distributions can, depending on the situation, be assumed to be Poisson or Negative Binomial, both representing count data. It is obvious that some sort of dependency might exist, as consumers are deciding on only a single one of the products. Different car brands could be an example for this. Covariates might be each car brand's characteristics but also external conditions, such as buying power in a given country.

- *The number of points / goals scored by teams or single players in a given match in sports.* As above, depending on the ranges (so primarily on the sport), count data distributions, such as Poisson or Negative Binomial, can be used for marginal distributions. The necessity to account for dependency between the scores / goals is rather easy to justify, as scored points during the game heavily influence the further course of the match. Covariates might be specific team or player strengths and characteristics, influencing both the team/player itself as well as the opponent and external factors like location or weather.

- *The results of a blood panel.* One might be interested in the occurrence of different substances or cells in a blood sample. The density could be modelled with suitable marginal continuous distributions. If the number of cells per unit of blood is countable, count data approaches as mentioned above are applicable again. One could imagine to observe the number of observed red and white blood cells in a given sample. Potential covariates might cover information about the patient as well as other results from the blood sample. Even in situations where no specific dependency was expected, it might be useful to try to account for it. The resulting model can (and should) be analysed to decide if the incorporated dependency is needed in an explorative manner.

The easiest approach in the settings mentioned above (and numerous more) is the assumption of independence. In this case, the two or more marginal distributions can be

modelled independently from one another. This work is focussed on the usage of copulas to model and assess dependency structures in such bivariate outcomes.

1.1 Guideline Through the Thesis and Description of Originality and Contributed Work

This thesis is based on the idea of applying copula regression to model the outcome of football matches. As with everything in life, issues occurred that needed solving. Solving these issues and presenting the corresponding results is the common thread and backbone of this work.

The following parts of Chapter 1 contain a brief historical summary of the development for modelling football results. Afterwards, a short introduction to the notation of generalised regression settings as well as to copulas is given. The rest of the thesis is structured as follows:

Chapter 2 - A Penalisation Approach for Competitive Settings

Introduces a penalty to account for competitive settings, in which marginal covariates with identical interpretations should be forced to obtain similar or equal effects. This was directly motivated by the application to FIFA World Cup data. Imagine a covariate influencing both margins, such as weather or spectator count. Any difference in those influences between the first and second margin (first and second named teams) can only be artefacts, as no underlying reason for being first or second named exists, apart from FIFA tournament schedules. So it may be favourable to force the corresponding regression coefficients to be identical. The same can be applied for two different covariates that share an interpretation, such as each team's market value. Why would the first team's market value have a different influence on the first team's performance compared to the second team's market value on the second team's performance? A suitable penalisation scheme was proposed, implemented and tested in a simulation study. The methodology and simulation study were published in van der Wurp et al. (2020). An experiment in real time on Bundesliga data was carried out and previously unpublished results are included in Chapter 2.

This work relies heavily on the `GJRM` framework provided by Marra and Radice (2019b), who also cooperated in van der Wurp et al. (2020), mainly to describe `GJRM`'s main methodology. The research idea of introducing regularisation into the framework originates from Andreas Groll and Thomas Kneib, who co-authored the manuscript and improved it with valuable proof reading. The author of this thesis contributed the exact elaboration and formulation of the proposed penalisation scheme. He performed all implementations and created both the simulations and applications.

Chapter 3 - An Approach to Variable Selection via LASSO-Type Penalisation

Introduces a more general penalisation framework to obtain sparsity. While the penalty from Chapter 2 is able to reduce the complexity of a model, no true variable selection takes place. To incorporate this, a LASSO approximation was derived from existing approximations, implemented and, again, tested on simulated data. The methodology and simulation study were published in van der Wurp and Groll (2023a).

Resting on the `GJRM` framework and the novel penalty from Chapter 2, Andreas Groll suggested to include true variable selection with a LASSO-type penalty approach. He proposed the approximation of the LASSO, based on the work of Oelker and Tutz (2017) and had contact to the authors thereof. The author of this thesis formalised and implemented the penalisation scheme in the underlying setting. He performed all simulations and applications.

Chapter 4 - An Application to FIFA World Cups

The penalties introduced in Chapters 2 and 3 are extensively evaluated on the FIFA World Cup dataset. Measures for predictive quality are defined and calculated for models without penalisation, with each penalty on its own and with both of them combined. This chapter is a combined and rewritten version with recalculated results from both van der Wurp et al. (2020) and van der Wurp and Groll (2023a).

Chapter 5 - Case Study: Football and Machine Learning

In this chapter, we compare classical univariate regression approaches with copula models explicitly accounting for the dependency structure as well as with modern machine learning techniques in the context of modelling and predicting football results in the major European leagues. Particularly, we want to present an extensive data set compiled from publicly available sources containing data and match results from the first men's football divisions from England, France, Germany, Italy, Spain (often referred to as the "big five"), the Netherlands and Turkey. We introduce several modelling approaches to predict upcoming matches and compare their predictive strengths. The gathered data set is presented in detail and made publicly available to motivate further work and modelling ideas.

The case study, published in van der Wurp and Groll (2023b), was mainly performed by the author of this thesis. Andreas Groll took part in proof reading, was consulted occasionally, and helped with revisions during the publication process.

As both methodology-based chapters (Chapter 2 and Chapter 3) should be comprehensible on their own, some redundant notational introductions and definitions (especially for quality of prediction measures) occur. The case study in Chapter 5 can be read completely on its own as the published version is very close to this chapter, with its own introduction and conclusion. All applications and simulations can be reproduced. Code and data was made available via GitHub, `https://github.com/H-vanderWurp/GJRM-mods`.

Historical Development for Modelling Football Results

This section is mainly taken from van der Wurp et al. (2020) and focusses on sports and especially football data. Due to the rather small and low range of scored goals during a match, the Poisson distribution is often deemed sufficient.

Poisson distributions to model football results are well established and have been widely used, see e.g. Lee (1997) or Dyte and Clarke (2000), who modelled the number of the teams' goals with independent Poisson distributions. Maher (1982) and Dixon and Coles (1997) were among the first to investigate dependency between scores of competing teams. They included an additional dependence parameter into the independent Poisson approach to adjust for certain under- and overrepresented match results, as their first results suggested an underestimation of matches ending with a draw.

Regularisation was introduced to the (mostly independent) Poisson approach plentifully, see e. g. Groll and Abedieh (2013) or Groll et al. (2015), who used LASSO penalisation (originally proposed by Tibshirani, 1996) in the context of football data.

Regarding dependencies, Karlis and Ntzoufras (2003) used a bivariate Poisson distribution approach which is explicitly accounting for dependencies. A lot of variations of this underlying idea have been presented. For example, Groll et al. (2018) created a re-parametrisation to tweak the model as they saw fit.

Due to its computational intense nature, the use of copulas in this context is rather new. Nikoloulopoulos and Karlis (2010) and Trivedi and Zimmer (2017) used copulas to account for dependency when modelling bivariate count data.

Parallel to this, other machine learning approaches such as random forests (e.g. Schauberger and Groll, 2018 and Groll et al., 2019, methodology from Breiman, 2001) have become popular alternatives to regression approaches and will regularly be used as benchmark models throughout this work.

1.2 Notation and Basics

For the introduction of regression modelling we will stick close to Fahrmeir et al. (2013) and the Chapters 2 and 3 thereof. Formulae and basic properties of regression are taken from there.

In the context of multivariate regression frameworks we are to model an outcome variable y depending on covariate variables $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ which leads to the goal of estimating the expected value $E(y|\boldsymbol{x})$ conditioned on given covariates. The regression model can be written as

$$y = g(x_1, \ldots, x_p) + \varepsilon,$$

with ε denoting the random noise. The classical linear regression model is using

$$g(x_1, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \tag{1.1}$$

with purely linear and additive covariates' effects. For assumptions and detailed properties see appropriate literature, e.g. Fahrmeir et al. (2013), Chapter 3. The (ordinary) least squares estimator $\hat{\boldsymbol{\beta}} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is given in closed form by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}, \quad \text{with design matrix} \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix},$$

and $\boldsymbol{y} = (y_1, \ldots, y_n)^T$. The (ordinary) least squares solution is viable whenever $g(\boldsymbol{x}) = \boldsymbol{x}$ given assumptions.

One core assumption usually is Gaussian errors, i.e. $\varepsilon \sim N(0, \sigma^2)$, which will not hold in settings with very distinct distributions. As this work focusses on count data and sports applications, the Poisson distribution is an intuitive example:

We call a random variable X to be Poisson-distributed, written $X \sim \text{Poi}(\lambda)$, if the underlying probability mass function is

$$f(x) = \frac{\lambda^x}{x!} \exp(-\lambda),$$

with its parameter $\lambda > 0$. The support of X is $\{0, 1, \ldots\} = \mathbb{N}_0$. Further properties are $E(X) = \lambda$ and $\text{Var}(X) = \lambda$.

As the value set of $g(\cdot)$ from Equation (1.1) is the full real scale $\mathbb{R}$, which is incompatible with modelling $E(y) = \lambda > 0$, some transformation needs to be done. The exp-function is sufficient, so we obtain

$$\begin{aligned} \lambda = \exp(\eta) &= \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p) \\ &= \exp(\beta_0) \cdot \exp(\beta_1 x_1) \cdot \ldots \cdot \exp(\beta_p x_p) \\ \Leftrightarrow \qquad \log(\lambda) = \eta &= \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \end{aligned}$$

with η denoting the shortened linear predictor. In general, other functions instead of $\log(\cdot)$ can be used on the left hand side. They are called link functions and are chosen depending on the underlying distribution. These approaches are called *generalised linear models* and are usually abbreviated via GLM. The coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ can be interpreted as a linear influence for the standard (Gaussian) regression, i.e. adding the value of β_k to the prediction of $\hat{y}$ for each unit increase in x_k. In the case of a Poisson regression the effects are interpreted multiplicatively, as $\exp(\beta_k x_k) = (\exp(\beta_k))^{x_k}$. So with each increase in x_k of one unit the response $\hat{\lambda}$ is multiplied by a factor of $\exp(\beta_k)$.

1.3 Motivation for Copulas

As discussed before, the dependency structure within bivariate football match outcomes is widely discussed in the literature and a multitude of different approaches have been presented. To get some sense of what these bivariate structures could look like, see Table 1.1. A home advantage is easy to see, as almost all results (x, y) with $x > y$ (upper triangle above **bold** diagonal) were occurring more often than their respective counterparts (y, x) (lower triangle below **bold** diagonal). The underlying structure can be highlighted in color, see Figure 1.1. The data can be found in the `EUfootball` `R`-package (van der Wurp, 2022). But as the grid for football results is rather coarse, we take a look at another sport, namely basketball and the NBA (National Basketball Association) in Figure 1.2 with data from `kaggle.com` (source Lauga, 2021). A positive dependence (in terms of correlation) is easy to see. And while copulas are not needed to grasp a simple correlation, they are immensely flexible and can depict a huge selection of different dependency structures. Figure 1.3 is an example to show the possibilities in terms of different structures that can be obtained with copulas. Even within a single class, here Frank, a lot of different shapes can be depicted. We assume that a "best fitting" copula class can be found for every situation, e.g. every type or sports. Even different occasions within a given sport, say FIFA World Cups vs. club football in the German Bundesliga vs. amateur or youth football matches could be modelled using different copula classes.

Table 1.1: Amount of Bundesliga match outcomes of seasons 2010/2011 until 2019/2020, $n = 3060$. Diagonal in bold to highlight tied matches.

Goals of Home Team	Goals of Away Team									
	0	1	2	3	4	5	6	7	8
9	0	0	1	0	0	0	0	0	0
8	3	1	0	0	0	0	0	0	**0**
7	3	2	0	0	0	0	0	**0**	0
6	14	9	7	0	0	0	**0**	0	0
5	29	25	7	4	1	**0**	0	0	0
4	65	63	36	6	**6**	3	0	0	0
3	142	158	80	**43**	6	4	1	0	0
2	234	268	**154**	59	24	7	3	0	0
1	228	**342**	196	118	49	12	6	0	1
0	**190**	185	137	78	38	7	4	1	0

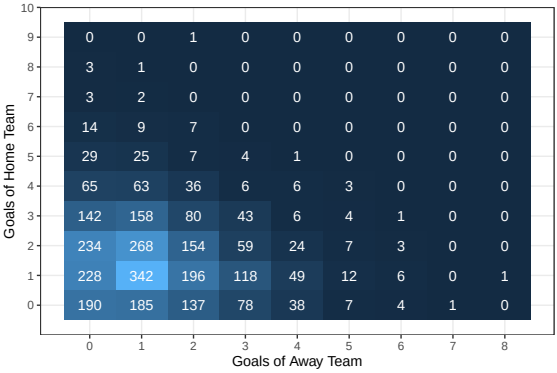

Figure 1.1: Amount of Bundesliga match outcomes of season 2010/2011 until 2019/2020, $n = 3060$. Colour brightness indicates more frequent outcomes.

1.4 Introduction to Copulas

In this section we give a brief notational overview about multivariate distributions via copulas. For more details we recommend Nelsen (2006) (especially Chapter 2 therein), from where properties and formulae were taken.

Assume two random variables Y_1 and Y_2 with their respective probability mass functions f_1, f_2 and their distribution functions F_1, F_2. We are interested in the joint distribution function $H(y_1, y_2)$, $H : S_1 \times S_2 \to [0, 1]$. S_1 and S_2 are denoting the supports of Y_1 and Y_2, respectively, which are $\mathbb{N}_0$ in the case of Poisson distributions.

The well known Sklar's theorem (originally by Sklar, 1959) states for every such H with given marginal distributions F_1 and F_2 a copula function C such as

$$H(y_1, y_2) = C(F_1(y_1), F_2(y_2))$$

exists. The copula itself is therefore a function $C : [0, 1]^2 \to [0, 1]$. Although this work will only use this bivariate setting, the methodology is generally not bound to the two-

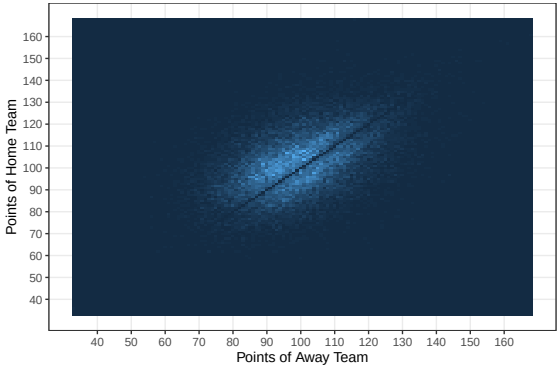

Figure 1.2: Amount of NBA match outcomes between 2004 and 2021, $n = 24\,195$. Colour brightness indicates more frequent outcomes.

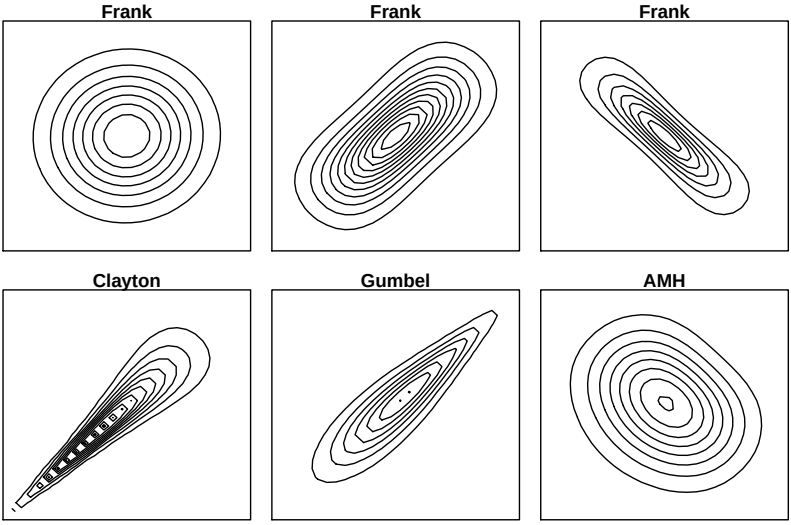

Figure 1.3: Contour plots for different copula classes. Arbitrarily chosen copula parameters to show the range of different dependency structures that can be depicted. Contour lines (inner to outer) show a decreasing density.

dimensional case. Extensive lists of copula functions, their densities and properties can be found in Nelsen (2006). The AMH (Ali-Mikhail-Haq) copula for example is the function

$$C_\theta(z_1, z_2) = \frac{z_1 z_2}{1 - \theta(1 - z_1)(1 - z_2)} \qquad \text{with} \qquad z_1, z_2 \in [0, 1].$$

The copula's parameter θ is limited to $[-1, 1)$ for this class, but has widely varying support

in different copula classes. Hence, θ itself is not comparable between different copulas. To allow for comparisons, we will use Kendall's τ as a measure for dependency strength. For the AMH copula, for example, it is calculated via (Kumar, 2010)

$$\tau = \frac{3\theta - 2}{3\theta} - \frac{2(1-\theta)^2 \log(1-\theta)}{3\theta^2} \in [-1, 1].$$

Kendall's τ is a measure on concordance and discordance, defined via

$$\tau = P(\text{concordance}) - P(\text{discordance}).$$

Two bivariate samples $(X_1, Y_1), (X_2, Y_2)$ are concordant, if $(X_1 - X_2)(Y_1 - Y_2) > 0$ holds, so if both factors are having the same sign. And they are discordant, if $(X_1 - X_2)(Y_1 - Y_2) < 0$ holds, so both parentheses are yielding different signs. Kendall's τ can be calculated both empirically (if samples are available) or theoretically. See Chapter 5 in Nelsen (2006) for more details. For most parts of this work we will refer to Kendall's τ as a general measure for dependency strength when comparing different copula classes, as their θ parameters are generally not comparable.

1.5 Underlying Methodology of the GJRM Framework

The following section will present the underlying methodology of joint modelling and how it is implemented in the `R` add-on package `GJRM` by Marra and Radice (2019b). The package's authors have presented their work and it's benefits in Marra and Radice (2017) and showed a set of different approaches applicable via their package in Marra and Radice (2019a). The section is mostly taken from van der Wurp et al. (2020).

For notational convenience, we drop the conditioning on parameters (of the marginal distributions and of the copula function) and observation index i. It is clear, however, from the context of this work that bivariate count data with integer realisations $\mathbf{y}_i = (y_{i1}, y_{i2})^T$, with $i = 1, \ldots, n$, for a sample of size n, are available (e.g. football scores or sales numbers) for modelling purposes and that covariate effects have to be accounted for.

We assume that the joint cumulative distribution function (cdf) $F(\cdot, \cdot)$ of two discrete outcome variables, $Y_1 \in \mathbb{N}_0$ and $Y_2 \in \mathbb{N}_0$, can be expressed as

$$\begin{aligned} P(Y_1 \leq y_1,\ Y_2 \leq y_2) &= C_\theta \left(P(Y_1 \leq y_1),\ P(Y_2 \leq y_2) \right) \\ &= C_\theta(F_1(y_1),\ F_2(y_2)), \end{aligned}$$

where $F_1(\cdot)$ and $F_2(\cdot)$ are the marginal cdfs of Y_1 and Y_2 taking values in $(0, 1)$, $C_\theta : (0, 1)^2 \to (0, 1)$ is a copula function which does not depend on the marginals, and θ denotes the copula parameter measuring the dependence between the two random variables. The adopted dependence structure relies on $C_\theta(\cdot, \cdot)$ and its parameter θ; the copulas implemented in `GJRM` are reported, for instance, in Table 1 of Marra and Radice (2019a). It should be pointed out that in a setting with discrete marginal distributions the copula function C_θ is not unique (see Schweizer and Sklar, 1983, Chapter 6; or Faugeras, 2017). However, as elaborated by several authors including Nikoloulopoulos and Karlis (2010) and Trivedi and Zimmer (2017), this is not an issue of practical concern in regression

settings. Potentially, another copula function C^* exists that can create the same probabilities on the grid implied by discrete marginal distributions. As the marginal distributions and their respective predictors are not influenced by this, we retain interpretability on corresponding estimated regression coefficients.

Following Trivedi and Zimmer (2017), the joint probability mass function (pmf) $c_\theta(\cdot,\cdot)$ for a given copula C_θ on the two-dimensional integer grid can be obtained as

$$\begin{aligned} c_\theta(F_1(y_1),\ F_2(y_2)) = C_\theta(&F_1(y_1),\ F_2(y_2)) \\ &- C_\theta(F_1(y_1-1),\ F_2(y_2)) \\ &- C_\theta(F_1(y_1),\ F_2(y_2-1)) \\ &+ C_\theta(F_1(y_1-1),\ F_2(y_2-1)). \end{aligned} \tag{1.2}$$

For the outcome variables Y_1 and Y_2, the authors of `GJRM` have considered (and implemented) four main discrete distributions, namely Poisson, negative binomial type I, negative binomial type II, and Poisson inverse Gaussian; these have been parametrised according to Rigby and Stasinopoulos (2005). In the following, we focus on Poisson marginals since they were found to be appropriate for modelling our count responses (see applications in Section 2.3 and Chapter 4).

Let now the parameters of the two marginal distributions as well as of the copula parameter θ be connected with sets of covariates of sizes p_1, p_2 and p_θ, respectively. Moreover, let the corresponding covariate vectors be denoted by $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_\theta$, including entries for intercepts and/or dummy variables for categorical variables. For two Poisson-distributed margins with rate parameters λ_1 and λ_2 and a copula function characterised by one parameter, we may have

$$\begin{aligned} \log(\lambda_1) = \eta_1 \quad &= \beta_0^{(1)} + x_1^{(1)}\beta_1^{(1)} + \ldots + x_{p_1}^{(1)}\beta_{p_1}^{(1)} \\ &= (\boldsymbol{x}^{(1)})^T\boldsymbol{\beta}^{(1)}, \\ \log(\lambda_2) = \eta_2 \quad &= \beta_0^{(2)} + x_1^{(2)}\beta_1^{(2)} + \ldots + x_{p_2}^{(2)}\beta_{p_2}^{(2)} \\ &= (\boldsymbol{x}^{(2)})^T\boldsymbol{\beta}^{(2)}, \\ g(\theta) = \eta_\theta \quad &= \beta_0^{(\theta)} + x_{1\theta}^{(\theta)}\beta_{1\theta}^{(\theta)} + \ldots + x_{p_\theta}^{(\theta)}\beta_{p_\theta}^{(\theta)} \\ &= (\boldsymbol{x}^{(\theta)})^T\boldsymbol{\beta}^{(\theta)}, \end{aligned} \tag{1.3}$$

where $\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}$ and $\boldsymbol{\beta}^{(\theta)}$ are p_1-, p_2- and p_θ-dimensional vectors of regression effects, respectively. The logarithmic link function guarantees positivity of the two Poisson parameters λ_1 and λ_2. Other distributions may require different link functions. The vectors $\boldsymbol{x}^{(1)}$, $\boldsymbol{x}^{(2)}$ and $\boldsymbol{x}^{(\theta)}$ are subsets of a complete set of covariates $\boldsymbol{x}$ of size d, with $p_1 + p_2 + p_\theta = k \geq d$. Finally, $g(\cdot)$ is a link function whose choice will depend on the employed copula (see Marra and Radice, 2019a).

We would like to stress that the Equations (1.3) represent a substantial simplification of the possibilities allowed for in the proposed modelling framework. In particular, our implementation allows to include non-linear functions of continuous covariates, smooth interactions between continuous and/or discrete variables and spatial effects, to name but a few. For this purpose, the penalised regression spline approach was adopted and the reader is referred to, e.g., Marra and Radice (2017) for some examples. Due to the specific type of penalisation employed in this work (see Chapters 2 and 3), here we focus on linear effects as presented in (1.3).

The model's log-likelihood for the k-dimensional vector

$$\boldsymbol{\beta}^T = \left((\boldsymbol{\beta}^{(1)})^T, (\boldsymbol{\beta}^{(2)})^T, (\boldsymbol{\beta}^{(\theta)})^T\right)$$

is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log\left\{c_\theta\left(F_1(y_{i1}), F_2(y_{i2})\right)\right\}, \tag{1.4}$$

where, for $j = 1, 2$,

$$F_j(y_{ij}) = \exp(-\exp(\eta_{ij})) \sum_{m=0}^{y_{ij}} \frac{\exp(\eta_{ij})^m}{m!}.$$

If spline terms appear in the model specification then (1.4) has to be augmented by a quadratic penalty term whose role would be to enforce specific properties on the respective functions, such as smoothness.

Simultaneous estimation of all the parameters is based on maximising $\ell(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. To this end, we extended the estimation approach of the `R` package `GJRM` (Marra and Radice, 2019b) to accommodate discrete margins. The fitting algorithm is based on iterative calls of a trust region algorithm, which requires first and second order analytical derivatives, which have been tediously derived and verified numerically. In `R`, the algorithm is realised in the `trust()` function from the `trust` package by Geyer (2015). The modularity of the implementation means that, in principle, it will be easy to extend our modelling framework to parametric copulas and discrete marginal distributions not included in the package. To facilitate the computational developments, when evaluating (1.2), we replaced $F_j(y_j - 1)$ with $F_j(y_j) - f_j(y_j)$ for $j = 1, 2$, where $f_j(\cdot)$ denotes the j^{th} marginal pmf. This is especially relevant for the case $y_j = 0$ where $F_j(-1)$ needs to be set to 0.

As hinted above, the `GJRM` infrastructure allows one to incorporate[1] any quadratic penalty of the form $\frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{S}\boldsymbol{\beta}$, where $\boldsymbol{S}$ is a penalty matrix. The next section discusses a specification of penalty which is particularly useful for competitive settings.

Prediction

After fitting a model, we can calculate probabilities for each possible pair of outcomes. We will sketch this modus operandi for the following football application, but it could be easily generalised to different data situations and marginal distributions. First, based on the two teams' estimated coefficients $\hat{\boldsymbol{\beta}}^{(j)}, j = 1, 2$, for an arbitrary match i, we estimate the marginal Poisson parameters λ_1 and λ_2 using

$$\widehat{\lambda}_{i1} = \exp(\widehat{\eta}_i^{(1)}) = \exp\left((\boldsymbol{x}_i^{(1)})^T\hat{\boldsymbol{\beta}}^{(1)}\right),$$

$$\widehat{\lambda}_{i2} = \exp(\widehat{\eta}_i^{(2)}) = \exp\left((\boldsymbol{x}_i^{(2)})^T\hat{\boldsymbol{\beta}}^{(2)}\right).$$

We then use the `copula` package (Hofert et al., 2017) to obtain the joint function for a specific chosen copula with Poisson margins and parameters $\widehat{\lambda}_{i1}, \widehat{\lambda}_{i2}$ and $\widehat{\theta}$. The probability for a specific match outcome (y_1, y_2) can be calculated using the joint pmf described above.

[1] On details of the implementation modifications see Appendix B.