Generalised joint regression for count data: a penalty extension for competitive settings

Hendrik van der Wurp[1] · Andreas Groll[1] · Thomas Kneib[2] · Giampiero Marra[3] · Rosalba Radice[4]

Abstract

We propose a versatile joint regression framework for count responses. The method is implemented in the R add-on package GJRM and allows for modelling linear and non-linear dependence through the use of several copulae. Moreover, the parameters of the marginal distributions of the count responses and of the copula can be specified as flexible functions of covariates. Motivated by competitive settings, we also discuss an extension which forces the regression coefficients of the marginal (linear) predictors to be equal via a suitable penalisation. Model fitting is based on a trust region algorithm which estimates simultaneously all the parameters of the joint models. We investigate the proposal's empirical performance in two simulation studies, the first one designed for arbitrary count data, the other one reflecting competitive settings. Finally, the method is applied to football data, showing its benefits compared to the standard approach with regard to predictive performance.

Keywords Count data regression · Joint modelling · Competitive Settings · Regularisation · Football

1 Introduction

There are many data situations where bivariate (or even multivariate) counts are the end point of interest and a priori assuming independence between such variables may be questionable. In particular, this is relevant in competitive settings. For example, in many team sports such as football, handball or ice hockey, one usually jointly observes the number of goals (or, more generally, the number of points as, for instance, in basket ball or Baseball) of both competing teams. These are certainly associated as the final scores are the outcome of many single game situations where the players of both teams are involved in. A second example could be the number of sales of competing products in a local store. Exemplarily, consider two competing luxury brands of cars. Each bivariate observation corresponds to the sales in a single region or store, which certainly depend on the cars'

characteristics (selling price, equipment, extras, etc.) and are correlated.

The use of copulae in this context and its historical development is portrayed in the example of modelling football scores. In this context, Poisson distribution approaches 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. Dixon and Coles (1997) were among the first to investigate dependency between scores of competing teams. They expanded the independent Poisson approach by an additional dependence parameter to adjust for certain under- and over-represented match results.

In this article, we present a flexible generalised joint regression framework for count responses. The linear or non-linear dependence between the outcomes is modelled via means of copulae. Moreover, all the parameters of the joint distribution can be specified as flexible functions of covariates. To account for competitive settings, we also provide an extension of the method which enforces the linear regression coefficients of the marginal predictors to be equal by introducing a penalty on the pairwise differences of the relevant effects. This is indeed particularly useful for modelling team sports data where the predictors of both competing teams are usually based on the same set of covariates whose effects are often assumed to be equal (e.g., Groll et al. 2018). The

✉ Hendrik van der Wurp
  vanderwurp@statistik.tu-dortmund.de

1 TU Dortmund University, Dortmund, Germany

2 Georg-August University Göttingen, Göttingen, Germany

3 University College London, London, UK

4 City University of London, London, UK

proposed method is incorporated in the R package GJRM (Generalized Joint Regression Modelling, Marra and Radice 2019).

A first approach for explicitly accounting for dependencies in football using the bivariate Poisson distribution was proposed by Karlis and Ntzoufras (2003). McHale and Scarf (2007) employed instead copula models with Poisson margins to model shots-for and shots-against. Copula models have already been considered in the context of count responses (see, for example, Nikoloulopoulos and Karlis (2010) and Trivedi and Zimmer (2017) and references therein). Here, we elected to extend the modelling and computational framework of Marra and Radice (2017) to the case of discrete margins because it allows for several types of covariate effects, for a rich variety of copula functions, and for any type of quadratic penalty which was a crucial aspect in our case study.

We investigate the proposed method's empirical performance in two simulation studies, the first one which does not assume equal regression coefficients and the other one which does, hence reflecting competitive settings. Finally, the method is applied to FIFA World Cup football data which shows that assuming equal coefficients yields a superior predictive performance when compared to the approach that does not impose such equality. Bookmakers are used as a natural benchmark for our model, and profits derived from adopting a given betting strategy are calculated.

The rest of the manuscript is structured as follows. The methodological background of the joint regression framework for count responses and the penalty extension specifically designed for competitive settings are introduced in Sect. 2. In Sect. 3, we present two simulation studies which analyse the proposed method's predictive performance in different data settings. The data employed in our application, covering all matches of the five preceding FIFA Football World Cups 2002 – 2018, are described in Sect. 4.1. Using these data, we then compare the predictive performance of the copula models whose parameters of the marginal distributions are assumed to be equal and not (Sect. 4.2). Assuming equal regression coefficients yields a superior performance, as elaborated in Sect. 4.3. We conclude in Sect. 5.

2 Methodology

This section provides the most salient details of the proposed generalised joint regression modelling framework for count data. In particular, to account for competitive settings, we will focus on the methodological aspects that are relevant to the model specification adopted in Sect. 4. Note that we have considered single-parameter copulae and marginal distributions with up to two parameters since they were deemed appropriate for our empirical application. However,

the computational framework can be conceptually easily extended to copulae and marginal distributions with more parameters.

2.1 Model structure and estimation approach

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 the paper that bivariate count data with integer realisations $\mathbf{y}_i = (y_{i1}, y_{i2})^T$, $i = 1, \ldots, n$, 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

$$P(Y_1 \leq y_1, \ Y_2 \leq y_2) = C_\theta \left(P(Y_1 \leq y_1), \ P(Y_2 \leq y_2) \right)$$
$$= C_\theta(F_1(y_1), \ F_2(y_2)),$$

where $F_1(\cdot)$ and $F_2(\cdot)$ are the marginal cdfs of Y_1 and Y_2 taking values in $(0, 1)$, $C_\theta : (0, 1)^2 \rightarrow (0, 1)$ is a two-place 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 copulae 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

$$c_\theta(F_1(y_1), \ F_2(y_2)) = C_\theta(F_1(y_1), \ F_2(y_2))$$
$$- C_\theta(F_1(y_1 - 1), \ F_2(y_2))$$
$$- C_\theta(F_1(y_1), \ F_2(y_2 - 1))$$
$$+ C_\theta(F_1(y_1 - 1), \ F_2(y_2 - 1)). \quad (1)$$

For the outcome variables Y_1 and Y_2, we have considered (and implemented in GJRM) four main discrete distribu-

tions, 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 Sect. 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 x_1, x_2 and x_θ, 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

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

where $\beta^{(1)}$, $\beta^{(2)}$ and $\beta^{(\theta)}$ 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 $x^{(1)}$, $x^{(2)}$ and $x^{(\theta)}$ are subsets of a complete set of covariates 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 in (2) 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 paper (see the next section), in this work we focus on linear effects as presented in (2).

The model's log-likelihood for the k-dimensional vector $\beta^T = \left((\beta^{(1)})^T, (\beta^{(2)})^T, (\beta^{(\theta)})^T \right)$ is

$$
\ell(\beta) = \sum_{i=1}^n \log \{ c_\theta \left(F_1(y_{i1}), F_2(y_{i2}) \right) \},
\tag{3}
$$

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 (3) 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 $\ell(\beta)$ with respect to β. To this end, we extended the estimation approach of the R package GJRM (Marra and Radice 2019) 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 copulae and discrete marginal distributions not included in the package. To facilitate the computational developments, when evaluating (1), 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 jth 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 any quadratic penalty of the form $\frac{1}{2}\beta^T S\beta$, where S is a penalty matrix. The next section discusses a penalty specification which is particularly useful for competitive settings.

2.2 A penalty approach for competitive settings

The model adopted in Sect. 4 is based on $F(y_1, y_2|\boldsymbol{\vartheta}) = C\left(F_1(y_1|\lambda_1), F_2(y_2|\lambda_2); \theta \right)$, where $Y_1 \sim \text{Poi}(\lambda_1)$, $Y_2 \sim \text{Poi}(\lambda_2)$ and $\boldsymbol{\vartheta} = (\lambda_1, \lambda_2, \theta)^T$. Each Poisson marginal models the counts of competitor $j \in \{1, 2\}$, e.g. goals of football team j, and is characterised by parameter λ_j. The expected number of goals for team j in a match i is given by $\lambda_{ij} = \exp\left(\beta_0^{(j)} + x_{i1}\beta_1^{(j)} + \ldots + x_{ip}\beta_p^{(j)} \right)$ with $i = 1, \ldots, n$, $j = 1, 2$. Although including covariate information into θ is possible and allowed for by GJRM, for simplicity, the copula parameter θ is specified as function of an intercept $\beta_0^{(\theta)}$ only. This way, we achieve explicit comparability of dependence strengths in terms of Kendall's τ among different copula functions.

In contrast to the setting of the equations in (2), in competitive settings it is often sensible to consider the same set of covariates for both competitors (i.e., $p_1 = p_2 =: p$). Also, one needs to impose the same covariate effects across the predictors of the marginal distributions. Specifically, assum-

ing covariates that are ordered such that $x_{ir}^{(1)}$ and $x_{ir}^{(2)}$, $r = 1, \ldots, p$, correspond to the same regressors (e.g., the average age of football team 1 and 2, respectively or the price tag of car 1 and 2, respectively), we would like to achieve $\beta_r^{(1)} = \beta_r^{(2)}$ for all $r \in 0, \ldots, p$. Without this restriction, being first- or second-named competitor could affect the estimation of $\beta_k^{(j)}$ and thus make the interpretation of the coefficients questionable, as stressed in Groll et al. (2018).

Equal (or virtually equal) coefficients for both margins can be achieved using a properly defined penalty. To this end, we propose to use the following penalised version of log-likelihood (3), i.e.

$$\ell_p(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \frac{1}{2}\xi \sum_{j=0}^{p} w_j \left(\beta_j^{(1)} - \beta_j^{(2)}\right)^2, \quad (4)$$

where the ridge-type penalty acts on the differences of the pair of intercepts as well as the pairs of coefficients corresponding to the same covariates, with suitably chosen weights w_j and penalty parameter ξ. This penalty can be easily incorporated into GJRM via a suitably designed penalty matrix S, which in this case is equal to

$$S = \xi \cdot W \circ \begin{pmatrix} 1 & 0 & \cdots & 0 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 & -1 & \cdots & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & -1 & 0 \\ -1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & \cdots & 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

(5)

where 'o' denotes the Hadamard matrix product. Moreover, ξ is a tuning parameter controlling the strength of the penalty, and W a weight matrix of the form

$$W = \begin{pmatrix} \boldsymbol{w}^T & \boldsymbol{w}^T & 0 \\ \vdots & \vdots & \vdots \\ \boldsymbol{w}^T & \boldsymbol{w}^T & 0 \\ 0 & \cdots & 0 \end{pmatrix},$$

which is of the same dimension as the other matrix from the Hadamard product in (5). The vectors of weights $\boldsymbol{w} = (w_0, w_1, \ldots, w_p)^T$ depend on the current fit $\hat{\boldsymbol{\beta}}^{[k]}$, which is obtained at iteration k of the algorithm. In order to shrink all paired differences jointly to zero, we use $w_j = \left| \hat{\beta}_j^{(1)} - \hat{\beta}_j^{(2)} \right|$ (here suppressing the iteration index for notational convenience). If the penalty parameter ξ is chosen sufficiently large, we obtain (virtually) equal parameter estimates. In our case study ξ was set to 10^9.

2.3 Prediction

After fitting a model with equal or unequal coefficients, we can calculate probabilities for each possible pair of outcomes. We will sketch this modus operandi for our football application, but it could be easily generalised to different competitive 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 in Sect. 2.1. That is,

$$\begin{aligned} P((Y_1, Y_2) = (y_1, y_2)) &= c(F_1(y_1), F_2(y_2); \widehat{\theta}) \\ &= C(F_1(y_1), F_2(y_2); \widehat{\theta}) - C(F_1(y_1 - 1), F_2(y_2); \widehat{\theta}) \\ &\quad - C(F_1(y_1), F_2(y_2 - 1); \widehat{\theta}) \\ &\quad + C(F_1(y_1 - 1), F_2(y_2 - 1); \widehat{\theta}), \end{aligned} \quad (6)$$

where $F_1(\cdot)$ and $F_2(\cdot)$ are the Poisson cdfs with corresponding parameters $\widehat{\lambda}_{i1}$ and $\widehat{\lambda}_{i2}$.

In football it is typically also of high interest to obtain estimates of the three coarser match results *win*, *draw* and *defeat* (from the perspective of the first-named team); these are collected in the categorical (ordinal) outcomes $\tilde{y}_i \in \{1, 2, 3\}$. For each of these categories, the probabilities of all relevant precise match results (y_1, y_2) are simply added up. Because the calculation of the pmf for all possible results may quickly become computationally infeasible, for practical purposes, in the football application of Sect. 4.2, we only consider $y_1, y_2 \leq 20$. The results $\widehat{\pi}_{i1}, \widehat{\pi}_{i2}$ and $\widehat{\pi}_{i3}$ are estimates for the true probabilities

$$P(win) = P(\tilde{Y}_i = 1) = \pi_{i1},$$
$$P(draw) = P(\tilde{Y}_i = 2) = \pi_{i2},$$
$$P(defeat) = P(\tilde{Y}_i = 3) = \pi_{i3},$$

which can then be used, e.g., for comparison with bookmakers' odds.

"Appendix A" provides a short operational description of the software.

3 Simulation study

In this section, we present two sets of simulations for pairs of (marginally) Poisson-distributed outcomes. The first one shows that the proposed estimation approach is able to (a) select the correct copula and (b) deliver estimates that are close to the true coefficients $\boldsymbol{\beta}$. The second set shows that when the true underlying coefficients are equal across margins, a penalisation, as presented in Sect. 2.2, considerably improves the goodness-of-fit.

For both simulations, we draw covariates $x_1, \ldots, x_6$ from independent uniform distributions on the interval $[0, 1]$, i.e. $x_r \sim \mathcal{U}[0, 1]$, for $n = 250$ observations and $r = 1, \ldots, 6$. For each observation, the Poisson parameters λ_{ij} ($i = 1, \ldots, 250$, $j = 1, 2$) are determined with given coefficient vectors $\boldsymbol{\beta}^{(1)} = (\beta_0^{(1)}, \beta_1^{(1)}, \beta_2^{(1)}, \beta_3^{(1)})^T$ and $\boldsymbol{\beta}^{(2)} = (\beta_0^{(2)}, \beta_1^{(2)}, \beta_2^{(2)}, \beta_3^{(2)})^T$ via

$$\lambda_{i1} = \exp\left(\beta_0^{(1)} + x_{i1}\beta_1^{(1)} + x_{i2}\beta_2^{(1)} + x_{i3}\beta_3^{(1)}\right), \quad (7)$$

$$\lambda_{i2} = \exp\left(\beta_0^{(2)} + x_{i4}\beta_1^{(2)} + x_{i5}\beta_2^{(2)} + x_{i6}\beta_3^{(2)}\right). \quad (8)$$

Each pair of outcomes (y_{i1}, y_{i2}) is sampled from a set of different copulae with marginal Poisson parameters λ_{i1} and λ_{i2}. For each copula, the respective parameter θ is determined by specifying fixed values for Kendall's $\tau \in \{-0.8, -0.6, -0.4, -0.2, 0.1, 0.3, \ldots, 0.9\}$ (although exact τ values are to be taken with caution due to identifiability issues on the copula class in our discrete setting, see Sect. 2.1). This is done to account for fundamentally different dependency scenarios. For *Gumbel* and *Clayton* this can be done analytically via direct transformation of the respective inverse functions, i.e.

$$\theta_{\text{Gumbel0}} = \frac{1}{1 - \tau}, \qquad \tau > 0,$$

$$\theta_{\text{Clayton0}} = 2\frac{\tau}{1 - \tau}, \qquad \tau > 0,$$

$$\theta_{\text{Clayton90}} = -2\frac{-\tau}{1 + \tau}, \qquad \tau < 0,$$

while for *Frank*, *Joe* and *Gaussian* this needs to be done numerically solving the following equations:

$$\tau = 1 - \frac{4}{\theta_{\text{Frank}}}(1 - d_1(\theta_{\text{Frank}})), \quad (9)$$

$$\tau = 1 - 4\sum_{k=1}^{\infty} \frac{1}{k(\theta_{\text{Joe0}}k + 2)(\theta_{\text{Joe0}}(k - 1) + 2)}, \quad (10)$$

$$\tau = \frac{2}{\pi}\arcsin(\theta_{\text{Gaussian}}). \quad (11)$$

Term d_1 denotes the first Debye function. Formulae (9) and (10) are from Hofert et al. (2012) and formula (11) from

Lindskog et al. (2003). To each pair of outcomes (Y_{i1}, Y_{i2}) a bivariate distribution function is assigned, from which we sample a single (bivariate) realisation.

Goodness-of-fit is measured by calculating the mean squared errors for the regression coefficients $\boldsymbol{\beta}$ via

$$\text{MSE} = \frac{1}{8}\left(\sum_{r=0}^{3}\left(\beta_r^{(1)} - \hat{\beta}_r^{(1)}\right)^2 + \left(\beta_r^{(2)} - \hat{\beta}_r^{(2)}\right)^2\right). \quad (12)$$

3.1 Classical count data set up

This section shows that the proposed estimation framework with unequal coefficients is able to detect the true copula and estimate the parameters reliably. We define two sets of coefficients, namely $\boldsymbol{\beta}^{(1)} = (0.5, 0.2, -0.2, 0)^T$ and $\boldsymbol{\beta}^{(2)} = (0.2, -0.3, 0.1, 0.5)^T$, determining the linear predictors in equation (7) and (8), respectively, and a chosen copula from the pre-defined set {C0, C90, F, G0, independence, J0, N}, containing seven different copulae (here C0 denotes the classical Clayton, C90 its 90 degree rotated version, F stands for Frank, G0 for Gumbel, J0 for Joe and N for Gaussian). Similarly to our case study, we fix the sample size to $n = 250$ and, as stated in Sect. 3, covariates $x_{i1}, \ldots, x_{i6}$ are generated from a uniform $\mathcal{U}[0, 1]$ distribution.

Each copula from the aforementioned list is combined with suitable values for Kendall's τ from the set $\{-0.8, -0.6, \ldots, 0.7, 0.9\}$ and each setting is repeated 100 times. For positive τ, five copulae (N, F, G0, C0, J0) are used, which leads to $100 \times 5 \times 5 = 2500$ samples with $n = 250$ bivariate observations each. The three copulae N, F and C90, which can depict negative correlation, are also combined with the four negative τ values, hence leading to $100 \times 3 \times 4 = 1200$ data sets. The penalisation approach to force equal coefficients in both marginal distributions from Sect. 2.2 is not yet applied here. The use of other copulae and respective rotations, implemented in GJRM, led to similar conclusions.

Figure 1 displays boxplots of the resulting MSEs of the regression coefficients from equation (12) for different true copulae and a selection of fitted ones in a scenario of weak positive correlation ($\tau = 0.1$). Due to the weak correlation, the resulting copula structures are similar to an independence approach and, hence, no visible differences in goodness-of-fit occur. When simulating from a setup with a considerably stronger dependence structure, e.g., $\tau = 0.7$, the results show substantial differences regarding the selection of the copula function (see Fig. 2). Stronger association implies a better ability to select the appropriate copula; increasing τ emphasises each copula's individual shape.

The results for C90 are numerically identical to those of the independence copula for most samples. Due to the copula's inability to account for positive correlation ($\tau > 0$), the fitting process results in $\hat{\theta} \approx 0$, which practically implies an
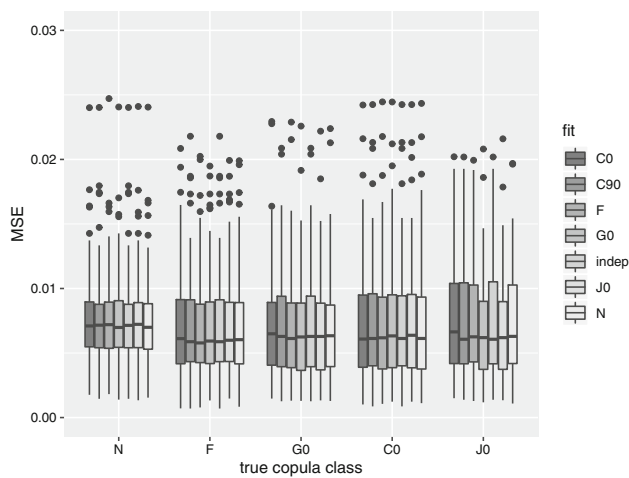
Fig. 1 Results for MSE of the regression coefficients for different true copulae with each copula parameter θ derived from $\tau = 0.1$
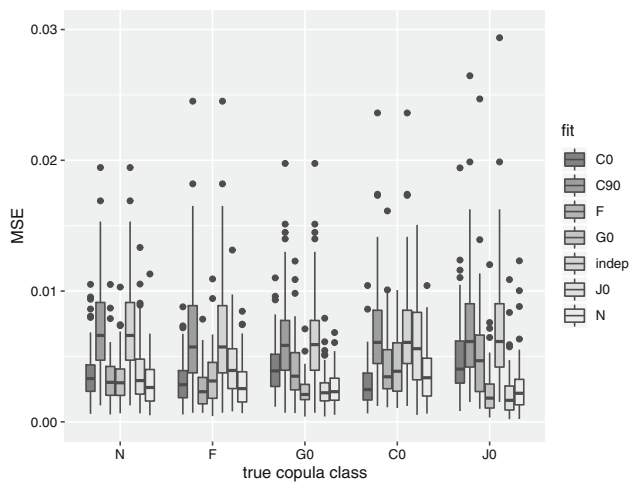

Fig. 2 Results for MSE of the regression coefficients for different true copulae with each copula parameter θ derived from $\tau = 0.7$

Fig. 3 Results for MSE of the regression coefficients for different true copulae with each copula parameter θ derived from $\tau = -0.6$

Table 1 Absolute number of times each copula is chosen by AIC as compared to the true function, for $\tau = 0.1$ (correct in bold)

		fitted class						
		indep	N	F	G0	J0	C0	C90
true class	N	20	**35**	12	13	5	15	0
	F	12	13	**31**	7	4	32	1
	G0	10	10	9	**23**	43	5	0
	J0	3	4	2	21	**70**	0	0
	C0	20	9	8	6	1	**55**	1

independence copula. This occurs in both directions; if data are sampled from copulae with $\tau < 0$ (see Fig. 3) the copulae C0, G0 and J0 will lead to fits reflecting (approximately) the independence copula as they can only account for positive association.

Apart from the direction of the association being an important factor, we found that the proposed approach is able to select the true copula in terms of the regression coefficients' MSEs. Moreover, some copulae lead to results of similar quality. For example, Fig. 2 shows that when the data were sampled from a G0 and J0 structure, both copulae deliver satisfying results. In fact, given our setting, these copulae are rather similar in terms of implied dependence structure. Note also that employing the incorrect copula might still lead to good results as long as the general association 'direction', be it a positive or negative value of Kendall's τ, can be accounted for. To this end, it is often useful to fit a selection of copulae
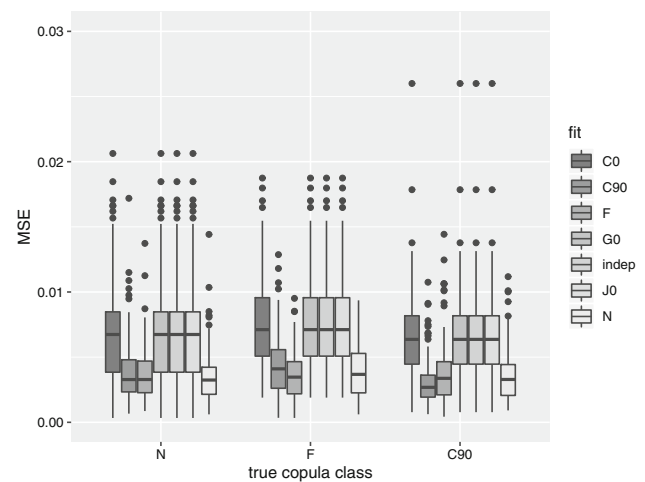
capturing different types of dependence and then choose a model based on cross-validation or information criteria. Previous research by Fang et al. (2014) and Marra and Radice (2017) suggests that the Akaike's information criterion (AIC; Akaike, 1973) is able to detect the true underlying copula function. In the following, we investigate the AIC's ability to identify the correct copula structure.

For a weak dependence structure ($\tau = 0.1$) AIC delivers mixed results (see Table 1). Due to our setting with a small sample size and a rather small range for the response values (the specified covariate distributions and coefficient values yield maximal values for λ_j of about $\exp(1.1) \approx 3$ for the Poisson marginals), some of the copula functions will lead to very similar bivariate structures. This is supported by the results displayed in Fig. 2, where the G0 and J0 yield similar results in terms of MSE. Nevertheless, the true copula is mostly the one that is more frequently selected (see bold numbers on the diagonal). When increasing the strength of dependence to $\tau = 0.7$, Table 2 shows that copula selection improves, although on a somewhat questionable level. Overall, in 334 out of 500 data sets the AIC was able to detect the true underlying copula. Again, the considered small sample

Table 2 Absolute number of times each copula is chosen by AIC as compared to the true function, for $\tau = 0.7$ (maximum in bold)

		fitted class						
		indep	N	F	G0	J0	C0	C90
true class	N	0	**62**	0	27	10	1	0
	F	0	2	**59**	0	0	39	0
	G0	0	24	0	**47**	29	0	0
	J0	0	6	0	15	**79**	0	0
	C0	0	0	13	0	0	**87**	0

size and a small range for the response values play a role here.

3.2 Count data in competitive settings

If both coefficient vectors $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ are equal or expected to be, our approach from Sect. 2.2 can be used. This is particularly relevant for competitive settings (e.g. sports data). As compared to the previous simulation set up, coefficients are now chosen as $\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(2)} = (0.25, 0.2, -0.35, 0)^T$, hence leading to

$$\lambda_{i1} = \exp(0.25 + 0.2x_{i1} - 0.35x_{i2} + 0x_{i3}),$$
$$\lambda_{i2} = \exp(0.25 + 0.2x_{i4} - 0.35x_{i5} + 0x_{i6}).$$

This setup depicts the same influence of the corresponding covariates in both marginal distributions. Note that choosing $\beta_3^{(j)} = 0$ creates a noise variable. This time, the model is always fitted using the true underlying copula structure only, but we now distinguish between the unpenalised estimation approach and the penalised approach proposed in Sect. 2.2. The results from 100 simulation runs per copula are visualised in Fig. 4. They clearly show the superior performance of the penalised approach as compared to the unpenalised one. In the latter, unequal coefficient estimates occur over the replicates.

4 FIFA football world cup application

We will now apply the proposed penalised approach to a real world data set containing FIFA Football World Cup matches and then compare the method's predictive performance to that of the unpenalised technique (Sect. 2.1).

4.1 Data

Model building was based on a data set constructed from the five past FIFA World Cups 2002 – 2018 with 64 matches each. The basic data set (without the World Cup 2018 data)
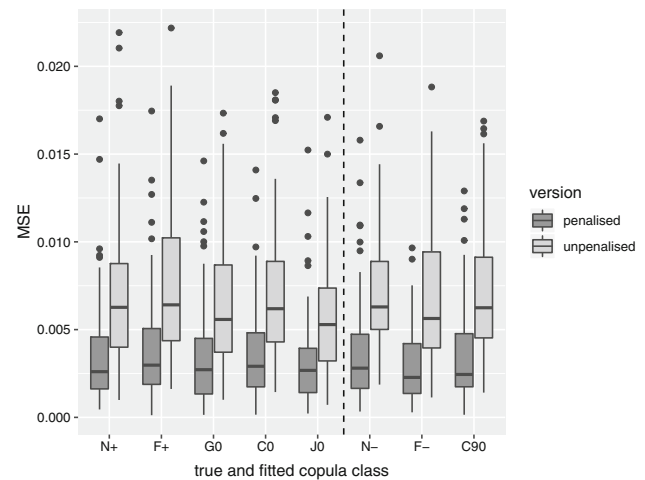

Fig. 4 Results for the MSE of the regression coefficients obtained using the penalised (left boxes) and unpenalised (right boxes) estimation approaches for a set of different copulae and associations; $\tau = 0.25$ for copulae N, F, G0, C0, J0 and $\tau = -0.25$ for copulae N, F, C90

was introduced and described in detail in Groll et al. (2015) and Schauberger and Groll (2018). It was then used in Groll et al. (2019) to make predictions for the World Cup 2018. We extended the data set by including two more variables, namely *Knockout* and *Titleholder*, which together with all the other covariates, are described below.

(a) *GDP per capita*. This is used as ratio of the GDP per capita for each respective country and the worldwide average GDP per capita (source: https://unstats.un.org/unsd/snaama/Index).

(b) *Population*. The population size of each country is used as ratio of the global population to take global population growth into account (source: https://data.worldbank.org/indicator/SP.POP.TOTL).

(c) *ODDSET probability*. The odds (taken from the German state betting agency ODDSET) are converted into winning probabilities. Therefore, the variable reflects the probabilities for each team to win the respective World Cup; these odds were calculated before the start of each tournament.

(d) *FIFA ranking*. The FIFA ranking provides a ranking system for all national teams measuring the performance of the team over the last four years (source: https://de.fifa.com/fifa-world-ranking/).

(e) *Host*. A dummy variable indicating if a national team is the hosting country.

(f) *Continent*. A dummy variable indicating if a national team is from the same continent as the host of the World Cup (including the host itself).

(g) *Confederation*. This categorical variable comprises the confederation of the respective team with (in principle) six possible values: Africa (CAF); Asia (AFC);

Table 3 Exemplary table showing the results of four matches (*a*) and a subset of the covariates of the involved teams (*b*). The matched data sets for each game are shown in (*c*)

(a) Table of results

FRA 🇫🇷	0:1	🇸🇳 SEN
URU 🇺🇾	1:2	🇩🇰 DEN
DEN 🇩🇰	1:1	🇸🇳 SEN
FRA 🇫🇷	0:0	🇺🇾 URU
⋮	⋮	⋮

Year	Team	Age	Rank	Oddset	...
(b) Table of covariates					
2002	France	28.3	1	0.149	...
2002	Senegal	24.3	42	0.006	...
2002	Uruguay	25.3	24	0.009	...
2002	Denmark	27.4	20	0.012	...
⋮	⋮	⋮	⋮	⋮	⋱

y_1	y_2	Team1	Team2	Year	Age1	Age2	...
(c) Matched data set							
0	1	FRA 🇫🇷	SEN 🇸🇳	2002	28.3	24.3	...
1	2	URU 🇺🇾	DEN 🇩🇰	2002	25.3	27.4	...
1	1	DEN 🇩🇰	SEN 🇸🇳	2002	27.4	24.3	...
0	0	FRA 🇫🇷	URU 🇺🇾	2002	28.3	25.3	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

Europe (UEFA); North, Central America and Caribbean (CONCACAF); Oceania (OFC); South America (CONMEBOL). The confederations OFC and AFC had to be merged because in the data set only one team (New Zealand, 2006) from OFC participated in one of the considered World Cups.

(h) *(Second) maximum number of teammates*. For each squad, both the maximum and second maximum number of teammates playing together in the same national club.

(i) *Average age*. The average age of each squad.

(j) *Number of Champions League (Europa League) players*. As a measurement of the success of the players at the club level, the number of players in the semi finals (taking place only a few weeks before the respective World Cup) of the UEFA Champions League (CL) and UEFA Europa League.

(k) *Number of players abroad*. For each squad, the number of players playing in clubs abroad (in the season previous to the respective World Cup).

(l) *Factors describing the team's coach:* For the coach of each national team, *age* and duration of his *tenure* are observed. Furthermore, a dummy variable is included, whether the coach has the same *nationality* as his team or not.

(m) *Knockout*. A dummy variable indicating if a match is a knockout one.

(n) *Titleholder*. A dummy variable indicating if a team is the current World Championship title holder.

There are, therefore, 18 variables which were collected separately for each World Cup and each participating team. In our data set each bivariate response (y_1, y_2), representing the number of goals each respective team scored in a certain match, is combined with the respective covariates of both teams. For both teams the same set of covariates is used. A shortened example of the overall data is given in Table 3. Our final data set (Table 3c) was created by matching the teams' covariates (Table 3b) with the match result data (Table 3a).

Next, we will fit the unpenalised and penalised versions of the proposed estimation method to the FIFA World Cup data from Table 3c, and use a cross-validation-type strategy to compare their performance.

4.2 Comparison of predictive performance

With prediction ability being our main objective, we have to validate all possible models with out-of-sample data. To do so, we fit the models (with different copulae) on four out of five World Cups and predict the marginal parame-

ters λ_1, λ_2 for the matches of the left-out tournament in a cross-validation-type strategy, cycling through all five tournaments. Using the estimate for θ of each copula, we can obtain probabilities for each possible match result (y_1, y_2) via equation (6) from Sect. 2.3.

Probabilities for the three-way results *win*, *draw* and *loss* are computed by aggregating all corresponding results; for example, for a draw we sum up the probabilities for the results $(0, 0), (1, 1), \ldots, (20, 20)$, cutting of at a reasonable maximum number of goals. We settled at 20 as cut-off, due to the maximal estimates of $\hat{\lambda}_j \approx 3$. The same strategy is applied for all the match results that lead to a win or a loss. Note that our estimated three-way probabilities practically always added up to one, which indicates that limiting the analysis to all results up to 20 goals was sufficient. The resulting *three-way probabilities* are denoted as $\hat{\pi}_{il}$ for the i-th match and $l = 1, 2, 3$ indicates *win*, *draw* and *loss*. Such three-way-outcomes could also have been modelled directly by using models for categorical responses. However, our approach to model the exact number of goals exploits more information, which is why ordinal/multinomial models were not considered.

For the estimation approaches considered, we employed several measures to compare their predictive performance. Similar to Schauberger and Groll (2018), we use the rank probability score (RPS), the multinomial likelihood and the classification rate as goodness-of-prediction measures, and average the results over all matches of all cross-validation cycles. In our setting, for a single game the RPS is defined via

$$\text{RPS}_i = \frac{1}{2} \sum_{r=1}^{2} \left(\sum_{l=1}^{r} \hat{\pi}_{il} - \delta_{il} \right)^2.$$

The true result is a dummy coding for *win*, *draw*, *loss* and denoted by Kronecker's delta δ_{li}. The mutinomial likelihood for a single match is defined as

$$\text{LLH}_i = \hat{\pi}_{i1}^{\delta_{i1}} \hat{\pi}_{i2}^{\delta_{i2}} \hat{\pi}_{i3}^{\delta_{i3}},$$

which is essentially the predicted probability $\hat{\pi}_{il}$ for the true outcome. Additionally, the classification rate is given as

$$\text{CR}_i = \mathbb{I}(\tilde{y}_i = \underset{l \in \{1,2,3\}}{\arg\max} (\hat{\pi}_{il})),$$

indicating whether match i was correctly classified. All measures are calculated for the unpenalised and penalised models and for each match prediction, and are then averaged over all $n = 320$ FIFA World Cup matches.

Beside focusing on three-way-outcomes, we can also analyse the performance in predicting the exact number of goals each team scored. Hence, we include the Euclidean distance

between observation and prediction. With the bivariate mean $\boldsymbol{\lambda}_i = (\lambda_{i1}, \lambda_{i2})^T$ of the bivariate distribution for a single match corresponding to a given copula model, we have

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left\| \mathbf{y}_i - \hat{\boldsymbol{\lambda}}_i \right\|_2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \sqrt{(y_{i1} - \hat{\lambda}_{i1})^2 + (y_{i2} - \hat{\lambda}_{i2})^2},$$

which is indeed calculated over n observed and predicted match results.

As predictive ability is our main aim, we also investigate the betting outcome for the most recent FIFA World Cup 2018 as another measure of performance. Using the (average) betting odds of the 64 matches (obtained from `oddsportal.com`, which provides averaged three-way-odds from a selection of bookmaker companies) as well as the corresponding predicted probabilities from our respective models, different betting strategies can be applied (see, e.g. Groll et al., 2018). For every match i and each of the possible three outcomes $l \in \{1, 2, 3\}$, one can calculate the expected return as follows: $E[return_{il}] = \hat{\pi}_{il} \cdot odds_{il} - 1$. Then, one could choose the outcome with the highest expected return, but only place the bet if the expected maximum return is positive, i.e. if $\max_{l \in \{1,2,3\}} E[return_{il}] > \varepsilon$ holds, with $\varepsilon = 0$. Koopman and Lit (2015) used different values of the threshold $\varepsilon > 0$ and showed that this way the overall average return could be increased. While they used constant stake sizes (one arbitrary unit) for each bet, alternative betting strategies with varying stake sizes based on Kelly's criterion (Kelly 1956) can be applied; see for example Boshnakov et al. (2017). With this criterion the optimal stake for single bets can be determined in order to maximize the return by considering the size of the odds and the winning probability. We will use the simple strategy with constant betting stakes and $\tau = 0$ in our copula selection process and provide a more detailed look into results afterwards (Table 4).

4.3 Results

The assumption of Poisson-distributed margins can be checked using randomised normalised quantile residuals (see Marra and Radice, Marra and Radice 2017, and references therein). In this case, using for instance a Gaussian copula model with Poisson margins fitted to data from all World Cups, the choice of marginal distributions seems appropriate (see Figure 5).

The results for all the performance measures described in the previous section can be found in Table 5 in "Appendix A". For each copula, the predictive performance of the penalised

Table 4 Ranks according to selected measures for fits based on different copulae

Copula	RPS	LLH	CR	bets	MSE	Σ
F	6	3	14	5	10	38
FGM	8	6	14	2	8	38
N	9	9	10	1	12	41
J90	15	15	2	9	1	42
G270	14	14	2	9	3	42
C180	11	11	9	3	9	43
indep	13	13	1	9	7	43
C0	3	7	14	5	16	45
G90	16	16	2	9	2	45
J0	12	12	2	9	11	46
G180	7	5	12	5	17	46
J180	5	8	12	5	18	48
J270	17	17	2	9	4	49
G0	10	10	11	4	15	50
AMH	2	4	14	17	14	51
C270	18	18	2	9	5	52
PL	4	2	18	17	13	54
C90	19	19	2	9	6	55
T	1	1	19	17	19	57
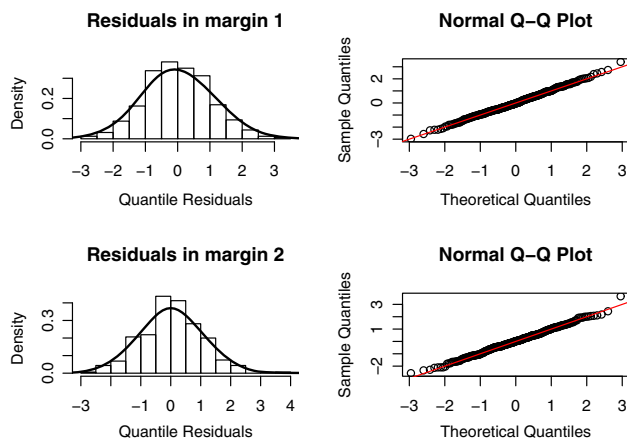
Fig. 5 Histograms and normal Q–Q plots of randomised normalized quantile residuals for the margins produced after fitting a Gaussian copula model with Poisson-distributed marginals to data from all World Cups

estimation approach is substantially better than that of the unpenalised method.

In our setup, the selected measures yield their bests values (highlighted in bold font) for different copulae, namely T (Student's T copula with 3 degrees of freedom), N (Gaussian copula), J90 (Joe copula rotated by 90 degrees), and even an independence model fitted outside of the GJRM framework. To aid comparability one could aggregate for each copula the corresponding ranks of all the considered measures (taking the different layouts of the selected measures into account). The respective results can be found in Table 4 where we see that F (Frank copula), followed by FGM (Farlie-Gumbel-Morgenstern copula) and N are the best for this specific scenario. Interestingly, these are symmetrical copulae.

Compared to the results of the models with no dependence structure (independence copula model obtained via a univariate Poisson regression approach on the single numbers of goals with three-way probabilities estimated via the Skellam distribution), we can see that the copula models improve the values for the chosen measures by a small margin or not at all. This was expected since a relatively weak dependence structure in the scores of international football matches was shown in previous work (see, e.g., Groll et al. 2018 who found that no additional dependence modelling was needed in a bivariate Poisson model when suitably structured predictors are employed). The estimated parameter $\hat{\theta} = 0.904$ and its corresponding value of Kendall's $\hat{\tau} = 0.100$ indicate a rather weak dependence structure for the F copula (however, keeping in mind the identifiability issues discussed in Sect. 2.1). The second and third ranked copula models with estimated values of $\hat{\theta} = 0.405$ leading to $\hat{\tau} = 0.09$ for FGM and $\hat{\theta} = 0.116$ leading to $\hat{\tau} = 0.0738$ for N (fitted on all World Cups) support the presence of a rather weak dependence structure. Table 6 in "Appendix B" shows the estimated regression coefficients for the F copula model.

Using the AIC for copula selection did not confirm the previous results: PL (Plackett copula) – which achieved the 17th place with respect to our five prediction measures (RPS, likelihood, classification, betting results, MSE) – provided the best fit. The F and FGM copulae, however, are ranked 3rd and 7th according to the AIC, and performed the best among our measures. It is important to stress that when using information criteria (but not only) the selection of the copula function is expected to improve as the sample grows large. Because predictive ability is our main goal, we rely more on the aforementioned measures over the AIC. The next paragraph shows that it can be advantageous to use the proposed copula models for betting strategies.

Betting
Fictional betting results can be calculated by predicting the World Cup 2018 outcomes from the F copula model fitted on World Cups 2002 – 2014. Figure 6 (top) depicts the average return percentage (i.e., the ratio between profit and investment) of two strategies for varying threshold sizes $\varepsilon \geq 0$. Note that with increasing values of ε the number of matches on which bets are placed decreases (see bottom of Fig. 6). For the FIFA World Cup 2018, solid positive returns can be achieved for values of $\varepsilon > 0.25$ with a simple betting strategy of constant stakes. Expanding this strategy with flexible stakes via Kelly's approach leads to positive returns for all ε. Overall, Kelly's strategy is clearly superior to placing con-
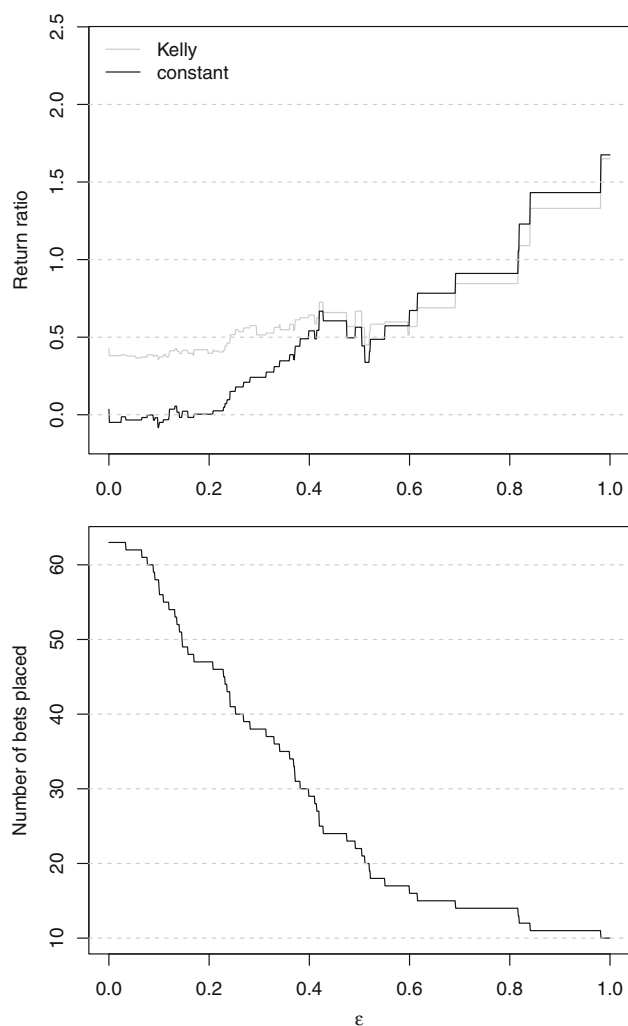
Fig. 6 Top: Return ratios for different betting strategies vs. threshold ε; Bottom: Number of bets placed vs. threshold ε

stant stakes indicated by higher (or equal) returns for smaller values of ε. An investment of 100 units of an arbitrary currency with a betting strategy of a fixed $\varepsilon = 0.4$ would yield a profit of about 50 units with constant stakes and 60 units with flexible stakes via Kelly's approach. Though remarkable at a first glance, these results have to be analysed cautiously. Due to the rather small sample size, the betting results very much depend on single match results and are probably very variable, especially for higher values of ε and therefore a smaller number of placed bets. Note, in fact, that the model is prone to extreme betting odds. For example, the bookmaker's odds for South Korea's victory against Germany were on average 19.52. Our model would recommend to bet on such outcomes - thus the betting results are likely to suffer from high variability. Despite these limitations, the results of the betting strategies are in favour of copula-structured models.

5 Conclusion

In this work, we have proposed a generalised joint regression framework for count responses. The method allows for linear and non-linear dependence structures through the use of different copulae, and for each parameter of the model to be specified as function of flexible covariate effects. We have also provided an extension which forces the regression coefficients of the marginal (linear) predictors to be equal via the use of a suitable penalisation. This is relevant for modelling competitive settings (e.g., sports data or competing product sales), because otherwise being first- or second-named competitor could affect the regression coefficients' estimates and their interpretation. The proposed method is available via the R add-on package GJRM.

We investigated the method's performance in two simulation studies, the first one designed for arbitrary count data, the other one reflecting competitive settings. In the first study, when the outcomes are weakly associated, copula structures are similar to an independence model and, hence, no tangible differences in goodness-of-fit are observed. With stronger dependence between the outcomes, results show substantial gains when using copula models. Generally, we found that the proposed method is able to select the true copula in terms of evaluating the regression coefficients' MSE. In the second simulation study, we assumed equal coefficients for the two marginal distributions. Under this scenario, the penalised method delivers an improved performance as compared to the unpenalised technique.

The method was applied to FIFA Football World Cup data; by using a cross-validation-type strategy based on several prediction measures, the penalised version of the method clearly outperformed the unpenalised approach. Moreover, the penalised approach performed better with regard to certain betting strategies.

Future research will address several extensions. Firstly, although not yet considered in our case study, we would like to extend the penalty discussed in this paper to the context of more complex predictor structures (allowing, for instance, for non-linear effects via P-splines). Moreover, we believe that the method's predictive performance can be further improved by penalising covariate effects via LASSO-type penalties (Tibshirani 1996; Friedman et al. 2010) or via boosting (e.g., Bühlmann and Hothorn 2007; Hothorn et al. 2010), a technique that stems from machine learning. These methods already proved to be effective in the context of predicting football matches (e.g., Groll et al. 2015, 2018).

source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Appendix

A Software

Copula models with discrete margins (but not only) can be fitted using the GJRM package by Marra and Radice (2019) for the R environment (R Core Team, 2019). The syntax is similar to that of established methods and packages for generalised linear and additive models. In the following, we provide an example. The model's formulae are provided, for instance, as

```
eq1 <- Goals ~ CL.players + UEFA.players +
    Nation.Coach + Age.Coach +
    Tenure.Coach + Legionaires + max.teammates +
    sec.max.teammates + age + Rank + GDP + host +
    confed + continent + odds + Population + Knockout +
    titleholder
eq2 <- Goals.oppo ~ CL.players.oppo +
    UEFA.players.oppo + Nation.Coach.oppo +
    Age.Coach.oppo + Tenure.Coach.oppo +
    Legionaires.oppo + max.teammates.oppo +
    sec.max.teammates.oppo +age.oppo + Rank.oppo +
    GDP.oppo + host.oppo + confed.oppo +
    continent.oppo + odds.oppo + Population.oppo +
    Knockout + titleholder.oppo
eq3 <- ~ 1
eqlist <- list(eq1, eq2, eq3)
```

where Goals, Goals.oppo are the discrete responses and the variables on the right hand side represent the covariates (regressors can either have or not the suffix oppo, depending on the margin considered). The same covariate can be used in more than one equation if desired (e.g., Knockout). The model is fitted by the call

```
fit <- gjrm(eqlist, data = WorldCup, BivD = "N",
    Model = "B", margins = c("PO", "PO"))
```

with BivD denotes the chosen copula (here, N for Gaussian), and the margins are Poisson distributed. Flexible covariate effects can be accounted for via splines by using for example s(covariate).

Convenience functions such as summary() and plot() are used in the same fashion as those for generalised linear and additive models. Residual diagnostics such as those displayed in Fig. 5 can be obtained via post.check().

More details, options, and extensive examples are given in the documentation of the GJRM package.

For this work the function gjrm() was adapted to allow for a new boolean argument linear.equal. For future work flexible penalisation utilities (including the proposed one in this paper) are planned to be implemented into the GJRM package.

B Tables

See Tables 5 and 6.

Table 5 Results of selected measures for model fits based on different copulae obtained using the unpenalised and penalised approaches

Copula	RPS pen	RPS unp	likelihood pen	likelihood unp	class. rate pen	class. rate unp	betting pen	betting unp	MSE_Goals pen	MSE_Goals unp
N	0.196	0.210	0.403	0.395	0.522	0.506	**0.199**	−0.225	1.421	1.490
C0	0.196	0.210	0.404	0.396	0.512	0.484	0.035	−0.154	1.424	1.496
C90	0.198	0.211	0.398	0.390	0.528	0.509	−0.012	−0.240	1.418	1.486
C180	0.198	0.212	0.400	0.392	0.525	0.506	0.040	−0.240	1.421	1.490
C270	0.198	0.212	0.398	0.390	0.528	0.506	−0.012	−0.240	1.418	1.486
J0	0.198	0.211	0.400	0.392	0.528	0.500	−0.012	−0.240	1.421	1.490
J90	0.198	0.212	0.398	0.390	0.528	0.506	−0.012	−0.240	**1.418**	1.486
J180	0.196	0.210	0.404	0.396	0.516	0.478	0.035	−0.225	1.425	1.500
J270	0.198	0.211	0.398	0.390	0.528	0.509	−0.012	−0.240	1.418	1.486
G0	0.197	0.212	0.402	0.394	0.519	0.503	0.038	−0.278	1.422	1.492
G90	0.198	0.211	0.398	0.390	0.528	0.509	−0.012	−0.240	1.418	1.486
G180	0.196	0.210	0.404	0.396	0.516	0.484	0.035	−0.236	1.424	1.495
G270	0.198	0.211	0.398	0.390	0.528	0.509	−0.012	−0.240	1.418	1.486
F	0.196	0.210	0.405	0.396	0.512	0.494	0.035	−0.284	1.421	1.487
AMH	0.196	0.210	0.405	0.396	0.512	0.491	−0.049	−0.110	1.421	1.489
FGM	0.196	0.210	0.404	0.396	0.512	0.491	0.117	−0.223	1.420	1.486
T	**0.195**	0.212	**0.407**	0.398	0.506	0.469	−0.049	−0.219	1.430	1.500
PL	0.196	0.210	0.405	0.396	0.509	0.484	−0.049	−0.284	1.421	1.487
indep	0.198	0.211	0.398	0.390	**0.531**	0.509	−0.012	−0.240	1.419	1.486

Table 6 Estimated coefficients and standard errors for the F model fitted on all World Cups

Covariate	$\hat{\beta}$	$se(\hat{\beta})$	Covariate	$\hat{\beta}$	$se(\hat{\beta})$
Intercept	2.226	1.078	GDP	0.042	0.026
CL Players	0.038	0.025	Host	0.369	0.186
EL Players	0.048	0.027	conf. CAF	0.094	0.206
Nationality Coach	0.097	0.098	conf. CONCACAF	0.109	0.211
Age Coach	−0.008	0.005	conf. CONMEBOL	0.622	0.217
Tenure Coach	−0.043	0.022	conf. UEFA	0.395	0.184
Players Abroad	0.005	0.011	Continent	−0.023	0.088
Max. Teammates	−0.011	0.043	Odds	−0.593	1.521
2nd Max. Teammates	0.004	0.064	Population	0.069	0.042
Age	−0.054	0.038	Knockout	−0.443	0.096
Rank	−0.007	0.004	Titleholder	−0.274	0.262

References

Akaike, H.: Information theory and the extension of the maximum likelihood principle. Second International Symposium on Information Theory. pp. 267–281, Springer, New York (1973)

Boshnakov, G., Kharrat, T., McHale, I.G.: A bivariate Weibull count model for forecasting association football scores. Int. J. Forecast. (2017). https://doi.org/10.1016/j.ijforecast.2016.11.006

Bühlmann, P., Hothorn, T.: Boosting algorithms: regularization, prediction and model fitting. Stat. Sci. **22**, 477–505 (2007)

Dixon, M.J., Coles, S.G.: Modelling association football scores and inefficiencies in the football betting market. J. R. Stat. Soc. Ser. C (Appl. Stat.) **46**(2), 265–280 (1997)

Dyte, D., Clarke, S.R.: A ratings based Poisson model for World Cup soccer simulation. J. Oper. Res. Soc. **51**(8), 993–998 (2000)

Fang, Y., Madsen, L., Liu, L.: Comparison of two methods to check copula fitting. Int. J. Appl. Math. **44**(1), 53–61 (2014)

Faugeras, O.P.: Inference for copula modeling of discrete data: a cautionary tale and some facts. Depend. Model. **5**(1), 121–132 (2017)

Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. **33**(1), 1 (2010)

Geyer, C.J.: Trust: trust region optimization. https://CRAN.R-project.org/package=trust. R package version 0.1-7 (2015)

Groll, A., Kneib, T., Mayr, A., Schauberger, G.: On the dependency of soccer scores - a sparse bivariate Poisson model for the UEFA European football championship 2016. J. Quant. Anal. Sports **14**(2), 65–79 (2018)

Groll, A., Ley, C., Schauberger, G., Van Eetvelde, H.: A hybrid random forest to predict soccer matches in international tournaments. J. Quant. Anal. Sports. **15**(4), 271–287 (2019)

Groll, A., Schauberger, G., Tutz, G.: Prediction of major international soccer tournaments based on team-specific regularized Poisson regression: an application to the FIFA World Cup 2014. J. Quant. Anal. Sports **11**(2), 97–115 (2015)

Hofert, M., Kojadinovic, I., Maechler, M., Yan, J.: copula: Multivariate Dependence with Copulas. https://CRAN.R-project.org/package=copula. R package version 0.999-18 (2017)

Hofert, M., Mächler, M., McNeil, A.J.: Estimators for Archimedean copulas in high dimensions. arXiv preprint arXiv:1207.1708 (2012)

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., Hofner, B.: Model-based boosting 2.0. J. Mach. Learn. Res. **11**, 2109–2113 (2010)

Karlis, D., Ntzoufras, I.: Analysis of sports data by using bivariate poisson models. The Statistician **52**, 381–393 (2003)

Kelly, J.L.: A new interpretation of information rate. Bell Syst. Tech. J. **35**(4), 917–926 (1956). https://doi.org/10.1002/j.1538-7305.1956.tb03809.x

Koopman, S.J., Lit, R.: A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. J. R. Stat. Soc. Ser. A (Stat. Soc.) **178**(1), 167–186 (2015)

Lee, A.J.: Modeling scores in the Premier League: is Manchester United really the best? Chance **10**, 15–19 (1997)

Lindskog, F., Mcneil, A., Schmock, U.: Kendall's tau for elliptical distributions. In: Credit Risk, pp. 149–156. Springer, Berlin (2003)

Marra, G., Radice, R.: Bivariate copula additive models for location, scale and shape. Comput. Stat. Data Anal. **112**, 99–113 (2017)

Marra, G., Radice, R.: GJRM: generalised joint regression modelling. R package version 0.2 (2019)

Marra, G., Radice, R.: Copula link-based additivemodels for right-censored event time data. J. Am. Stat. Assoc. **115**(530), 886–895 (2020)

McHale, I., Scarf, P.: Modelling soccer matches using bivariate discrete distributions with general dependence structure. Stat. Neerl. **61**(4), 432–445 (2007)

Nikoloulopoulos, A.K., Karlis, D.: Regression in a copula model for bivariate count data. J. Appl. Stat. **37**, 1555–1568 (2010)

R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (2019)

Rigby, R.A., Stasinopoulos, D.M.: Generalized additive models for location, scale and shape. J. R. Stat. Soc. Ser. C **54**, 507–554 (2005)

Schauberger, G., Groll, A.: Predicting matches in international football tournaments with random forests. Stat. Model. **18**(5–6), 1–23 (2018)

Schweizer, B., Sklar, A.: Probabilistic Metric Spaces, North-Holland Series in Probability and Applied Mathematics (1983)

Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B **58**, 267–288 (1996)

Trivedi, P., Zimmer, D.: A note on identification of bivariate copulas for discrete count data. Econometrics **5**(1), 10 (2017)