van der Wurp, Hendrik; Groll, Andreas

Article — Published Version

Introducing LASSO-type penalisation to generalised joint regression modelling for count data

AStA Advances in Statistical Analysis

Provided in Cooperation with:
Springer Nature

Suggested Citation: van der Wurp, Hendrik; Groll, Andreas (2021) : Introducing LASSO-type penalisation to generalised joint regression modelling for count data, AStA Advances in Statistical Analysis, ISSN 1863-818X, Springer, Berlin, Heidelberg, Vol. 107, Iss. 1-2, pp. 127-151, https://doi.org/10.1007/s10182-021-00425-5

AStA Advances in Statistical Analysis (2023) 107:127–151
https://doi.org/10.1007/s10182-021-00425-5

ORIGINAL PAPER

Introducing LASSO-type penalisation to generalised joint regression modelling for count data

Hendrik van der Wurp[1] · **Andreas Groll**[1]

Received: 8 January 2021 / Accepted: 6 October 2021 / Published online: 12 November 2021

Abstract
In this work, we propose an extension of the versatile joint regression framework for bivariate count responses of the `R` package `GJRM` by Marra and Radice (R package version 0.2-3, 2020) by incorporating an (adaptive) LASSO-type penalty. The underlying estimation algorithm is based on a quadratic approximation of the penalty. The method enables variable selection and the corresponding estimates guarantee shrinkage and sparsity. Hence, this approach is particularly useful in high-dimensional count response settings. The proposal's empirical performance is investigated in a simulation study and an application on FIFA World Cup football data.

Keywords Count data regression · FIFA world cups · Football penalisation · Joint modelling · Regularisation

1 Introduction

Various scenarios with bivariate count data can be thought of, where the two outcomes are expected to depend on one another. In particular, the dependency between two outcomes often can be of a competitive nature. Jointly observed numbers of goals (or more generally, points) in a given football match; or sales numbers of two competing products like car brands in a given sales branch; or the observed count for red and white blood cells in a blood sample are examples where some form of dependency can be expected. Therefore, the inclusion of copula structures may be useful when statistical models are applied in such settings.

The historical development of copula models in this context (and especially in sports settings) is diverse. In general, bivariate Poisson modelling approaches are well established and started without any form of dependency. For example, in the case of modelling football scores, independent Poisson distributions were used e.g. by Lee (1997), Karlis and Ntzoufras (2000), Dyte and Clarke (2000),

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

[1] Department of Statistics, TU Dortmund University, Dortmund, Germany

Groll et al. (2015) or Ley et al. (2019). However, in recent years many different approaches have been proposed to include dependency. For example, in the context of football, Dixon and Coles (1997) were among the first to explicitly account for dependency between scores of competing teams by expanding the independent Poisson approach by an additional dependence parameter to adjust for certain under and overrepresented match results. An even more flexible approach to this is the usage of copulae, as a lot of different copula families can be applied to consider a wide range of different dependency structures. In football, a first specific copula, namely the bivariate Poisson distribution, was employed by Karlis and Ntzoufras (2003). Moreover, McHale and Scarf (2007) proposed alternative copula models with Poisson margins to model shots-for and shots-against. In general, for the use of copula models for count responses, see e.g. Nikoloulopoulos and Karlis (2010), Trivedi and Zimmer (2017) and the references therein.

The Generalised Joint Regression Modelling (GJRM) infrastructure by Marra and Radice (2020), implemented in the `GJRM` add-on package for `R`, is a powerful tool for joint regression modelling. But to this point, no classical penalisation approaches for shrinkage of regression coefficient estimates and variable selection are available. In this work, we extend the existing GJRM infrastructure by allowing for a lasso-penalty (Tibshirani 1996; Friedman et al. 2010) and by this means add tools for variable selection and sparsity.

The LASSO technique has already been successfully applied in the context of football. For example, in Groll and Abedieh (2013) a penalised generalised linear mixed model has been used for modelling and prediction of European championship match data, and in Groll et al. (2015), a similar LASSO model has been applied on FIFA World Cup data. An L_1-penalised approach for Bradley-Terry-type models has also been proposed by Schauberger et al. (2017) on data for the German Bundesliga. Here, we will build upon these ideas and introduce L_1-type penalisation to generalised joint regression modelling for count data, also with a specific emphasis on football applications.

The remainder of the manuscript is structured as follows. In Sect. 2, we give an overview about the general model specifications and introduce the LASSO-penalty and how it can be embedded into the framework. We investigate the penalised model's performance in both a simulation study (Sect. 3) and the real life situation of FIFA World Cup football data (Sect. 4), before we conclude in Sect. 5.

2 Methodology

In this section, we give a brief overview of the basic methodological framework into which the proposed penalty approach is embedded. It is essentially a more compact version of Sect. 2.1 in van der Wurp et al. (2020). Note that in the following all concepts are illustrated for the two-dimensional response case, but can principally easily be extended to higher dimensions.

2.1 Model structure and estimation approach

Given a bivariate response of count data $\boldsymbol{y}_i = (y_{i1}, y_{i2})^T$, $i = 1, \dots, n$, where some form of dependency is assumed and shall be taken into account (e.g. scores of two competing teams in sports like football or sales numbers of two competing products), the corresponding joint cumulative distribution function (cdf) $F(\cdot, \cdot)$ of the two underlying discrete outcome variables $Y_1, Y_2 \in \mathbb{N}_0$ is given by

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

with $F_1(\cdot)$ and $F_2(\cdot)$ denoting the marginal cdfs of Y_1 and Y_2, respectively, realising values in $(0,\ 1)$ and $C_\theta : [0,1]^2 \to [0,1]$ is a two-place copula function. $C_\theta(\cdot, \cdot)$ does not depend on the marginal distributions but rather on its parameter θ that can be scalar- or vector-valued and controls the dependency strength, depending on the copula class chosen. A list of common copula classes implemented in `GJRM` can be found in Table 1 of Marra and Radice (2019). The most important properties of copula classes as well as details on the most well-known copula families can be found in standard copula literature such as Nelsen (2006).

The joint probability mass function (pmf) $c_\theta(\cdot, \cdot)$ for a chosen copula class $C_\theta(\cdot, \cdot)$ and discrete, integer-valued responses only exists on the two-dimensional integer grid and is expressed as

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

Although several distributions suitable for count data are implemented in `GJRM`, here we focus on and use the notation of Poisson distributed marginals as it is deemed adequate for our application. The marginal distribution parameters are therefore λ_1 and λ_2 and are modelled by a set of covariate vectors, denoted by $\boldsymbol{x}_1, \boldsymbol{x}_2$ of length p_1 and p_2, respectively. Covariates influencing the copula parameter θ, which for simplicity and notational convenience in the following is taken to be scalar, are collected in the vector $\boldsymbol{x}_\theta$ of length p_θ. In general, $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_\theta$ can be of different lengths and may contain partly or completely the same set of covariates. The corresponding regression equations are

$$\log(\lambda_1) = \eta_1 \quad = \beta_0^{(1)} + x_1^{(1)}\beta_1^{(1)} + \cdots + x_{p_1}^{(1)}\beta_{p_1}^{(1)} = (\boldsymbol{x}^{(1)})^T\boldsymbol{\beta}^{(1)}\,, \tag{1}$$

$$\log(\lambda_2) = \eta_2 \quad = \beta_0^{(2)} + x_1^{(2)}\beta_1^{(2)} + \cdots + x_{p_2}^{(2)}\beta_{p_2}^{(2)} = (\boldsymbol{x}^{(2)})^T\boldsymbol{\beta}^{(2)}\,, \tag{2}$$

$$g(\theta) = \eta_\theta \quad = \beta_0^{(\theta)} + x_{1\theta}^{(\theta)}\beta_{1\theta}^{(\theta)} + \cdots + x_{p_\theta}^{(\theta)}\beta_{p_\theta}^{(\theta)} = (\boldsymbol{x}^{(\theta)})^T\boldsymbol{\beta}^{(\theta)}, \tag{3}$$

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 marginal regressions (1) and (2) stem from the usual GLM-approach in Poisson regression, while $g(\theta)$ from (3) denotes a link function that is

suitable for the chosen copula class $C_\theta(\cdot,\cdot)$ (see Marra and Radice 2019). The vectors $\boldsymbol{x}^{(1)}$, $\boldsymbol{x}^{(2)}$ $\boldsymbol{x}^{(\theta)}$ are subsets (not necessarily disjoint) of a complete set of covariates $\boldsymbol{x}$ of length d, with $p_1 + p_2 + p_\theta = p \geq d$. It should be noted that the linear predictors from equations (1)–(3) are a substantial simplification of the possibilities allowed for in the GJRM framework.

Defining the composed vector of coefficients $\boldsymbol{\beta}^T := \left((\boldsymbol{\beta}^{(1)})^T, (\boldsymbol{\beta}^{(2)})^T, (\boldsymbol{\beta}^{(\theta)})^T\right)$, the log-likelihood of the copula regression model is given by

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log\left\{c_\theta\left(F_1(y_{i1}), F_2(y_{i2})\right)\right\}, \tag{4}$$

where for $i = 1, \ldots, n$ and $j = 1, 2$, we have

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

The log-likelihood is maximised in GJRM by using a trust region algorithm from the trust package by Geyer (2015). The corresponding required first and second order derivatives are included in the GJRM framework and were provided by Marra and Radice (2020).

Moreover, the GJRM infrastructure allows for the implementation of any quadratic penalty on (all or parts of) the regression coefficients $\boldsymbol{\beta}$ of the form $\frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{S}\boldsymbol{\beta}$, where $\boldsymbol{S}$ is a penalty matrix. This particularly includes the usage of splines to allow for nonlinear predictors. In this case, the smooth covariate effect is found in m corresponding entries of $\boldsymbol{\beta}$ and $\boldsymbol{S}$, respectively, where m depends on the number of spline basis functions. van der Wurp et al. (2020) tweaked $\frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{S}\boldsymbol{\beta}$ to incorporate a penalty on the differences between the marginal coefficient vectors $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ to force them to be equal, a feature that is specifically relevant in competitive settings such as sports applications. The next section discusses the specific construction of a suitable penalty matrix $\boldsymbol{S}$ to incorporate LASSO-type penalties.

2.2 Implementation of (adaptive) LASSO-type penalisation

The model's aforementioned log-likelihood $\ell(\boldsymbol{\beta})$ can be penalised in different ways. A first step in this direction could be ridge regression (Hoerl and Kennard 1970) penalising the squared regression coefficients. However, for better interpretability and stability with respect to multicollinearity issues, we prefer a method that can perform variable selection. Hence, we aim to incorporate a shrinkage penalty via a standard LASSO approach. In general, the penalised likelihood can be written as

$$\ell_p(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \nu \sum_l |\beta_l|, \tag{5}$$

with ν denoting a generic penalty strength parameter. Note that the sum in the penalty term in (5) typically does not contain any of the intercepts $\beta_0^{(j)}, j = 1, 2$, and $\beta_0^{(\theta)}$, as these usually shall not be penalised.

To incorporate this into the GJRM framework, which only allows for quadratic penalty structures, we adapt the approach and notation presented by Oelker and Tutz (2017), where a generic penalty term $|\xi|$ is quadratically approximated via $\sqrt{\xi^2 + c}$. Their notation leads to separate penalty matrices for each term $l \in \{1, \ldots, p\}$ that shall be penalised, namely

$$\mathbf{A}_l = \frac{1}{\sqrt{\left(\mathbf{a}_l^T \boldsymbol{\beta}^{[k]}\right)^2 + c}} \cdot \mathbf{a}_l \mathbf{a}_l^T, \tag{6}$$

where $\mathbf{a}_l^T \boldsymbol{\beta}^{[k]}$ with $\mathbf{a}_l \in \mathbb{R}^p$ can depict any linear transformation on the coefficient vector $\boldsymbol{\beta}$ in the k-th iteration step of the underlying fitting procedure. The arbitrary small value of $c > 0$ ($c = 10^{-9}$ in our application and simulation) guarantees a denominator greater than zero. For the specific choice of the standard LASSO, the vectors $\mathbf{a}_l$ are chosen such that $\mathbf{a}_0^T \boldsymbol{\beta} = \beta_0$, $\mathbf{a}_1^T \boldsymbol{\beta} = \beta_1$ and so forth. Hence, the matrix in (6) is a matrix of zeros, which only contains a single non-zero value $1/\sqrt{(\beta_l^{[k]})^2 + c}$ on the corresponding diagonal element.

The matrices $\mathbf{A}_l$ for each variable that is supposed to be penalised are finally simply added up. This can also be written via

$$\boldsymbol{S} = \nu \cdot \mathbf{W} = \nu \cdot \sum_{l=1}^{p} \mathbf{A}_l, \tag{7}$$

where $\mathbf{W}$ is a suitable chosen weight matrix such as

$$\mathbf{W} = diag\left[0, \frac{1}{\sqrt{\left(\beta_1^{[k]}\right)^2 + c}}, \ldots, \frac{1}{\sqrt{\left(\beta_p^{[k]}\right)^2 + c}}\right] \circ \begin{pmatrix} 0 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \ldots & 1 \end{pmatrix}, \tag{8}$$

with $diag[\ldots]$ denoting a diagonal matrix and "$\circ$" the Hadamard matrix product. Note that intercepts are not penalised in this setting, which is e.g. reflected by the value of 0 in the first diagonal element. The strength of the LASSO penalty is controlled by the penalty parameter ν, which needs to be tuned. Note that covariates are standardised to a standard deviation of 1, which is necessary for a balanced and comparable penalisation.

This weighting scheme leads to a penalisation of

$$\begin{aligned} \ell_p(\boldsymbol{\beta}) &= \ell(\boldsymbol{\beta}) - \frac{1}{2} \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta} \\ &= \ell(\boldsymbol{\beta}) - \frac{1}{2} \nu \cdot \sum_{i=1}^{p} \frac{\beta_i^2}{\sqrt{\left(\hat{\beta}_i^{[k]}\right)^2 + c}}, \end{aligned} \tag{9}$$

where the fraction is the approximation of $\beta_i^2 \,/\, |\hat{\beta}_i^{[k]}|$. Although this is not reducable, it could be argued that this is closely related to the absolute value $|\beta_i|$, which is the pursued penalisation from (5).

The proposed LASSO approximation by Oelker and Tutz (2017) yields "almost identical" (see Chapter 3.1 therein) results to the "actual" LASSO. As this approach is straightforward within the infrastructure of GJRM and generally much easier to implement than the actual LASSO (particularly in more complex settings), we deem it to be preferable in the present case. Typically, quadratically approximated LASSO is also faster with respect to computational complexity compared to actual LASSO, which involves numerical techniques as for example coordinate descent.

2.2.1 Group lasso for categorical predictors

For the case of categorical covariates, we follow the group lasso approach from Meier et al. (2008), leading to a group of coefficients being penalised and shrunk towards zero as a single entity (see also Oelker and Tutz 2017, and Groll et al. 2019). Assume that a set of j coefficients $\beta_{i+1}, \ldots, \beta_{i+j}$ correspond to e.g. the respective columns in the design matrix of a dummy-encoded factor variable.

With suitably chosen $\mathbf{a}_l$ and $\mathbf{A}_l$, the corresponding entries in the weight matrix $\mathbf{W}$ in (7) and (8) are then replaced by

$$\mathbf{W} = diag\left[\ldots, \frac{1}{\sqrt{\left(\beta_i^{[k]}\right)^2 + c}}, v_i, \ldots, v_i, \frac{1}{\sqrt{\left(\beta_{i+j+1}^{[k]}\right)^2 + c}}, \ldots\right] \circ \begin{pmatrix} 0 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \ldots & 1 \end{pmatrix}, \tag{10}$$

with v_i denoting the group weight that corresponds to all coefficients belonging to the same group. It is given by

$$v_i = \frac{1}{\sqrt{\sum_{r=i+1}^{i+j} \left(\beta_r^{[k]}\right)^2 + c}} \cdot \sqrt{j}, \tag{11}$$

where j refers to the size of the group $\beta_{i+1}, \ldots, \beta_{i+j}$ (i.e. for a dummy-encoded factor variable to the number of levels minus one), and, hence, to the group's complexity.

In this case, for the covariates' standardisation process we follow the Meier et al. (2008) standardisation technique for groups of categorical predictors. The corresponding block matrices $\mathbf{X}_g$ from the design matrix $\mathbf{X}$ are orthonormalised using a QR-decomposition. Finally, note that the coefficient estimates obtained by the LASSO fitting routine are transformed back in order to correspond to the original scale.

2.2.2 Adaptive weights

Additionally, following e.g. Zou and Hastie (2006), multiplicative adaptive weights w_l can be incorporated, which are based on the (inverse of) the unregularised maximum likelihood estimates, i.e. $w_l = 1/|\mathbf{a}_l^T \hat{\boldsymbol{\beta}}_{\text{ML}}|$ (see also Oelker and Tutz 2017). For ordinary lasso, this results in the weight

$$\mathbf{W} = diag\left[0, \frac{1}{|\hat{\beta}_{1,\text{ML}}|\sqrt{\left(\beta_1^{[k]}\right)^2 + c}}, \ldots, \frac{1}{|\hat{\beta}_{p,\text{ML}}|\sqrt{\left(\beta_p^{[k]}\right)^2 + c}}\right] \circ \begin{pmatrix} 0 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \ldots & 1 \end{pmatrix},$$

corresponding to equation (8).

In the case of group lasso with adaptive weights, the structure from equation (10) is modified by expanding the weights correspondingly. The group weights v_i are then calculated via

$$v_i = \frac{1}{||\boldsymbol{\beta}_{c,\text{ML}}||_2 \cdot \sqrt{\sum_{r=i+1}^{i+j} \left(\beta_r^{[k]}\right)^2 + c}} \cdot \sqrt{j}$$

with $\boldsymbol{\beta}_{c,\text{ML}}$ denoting the corresponding subvector $(\beta_{i+1,\text{ML}}, \ldots, \beta_{i+j,\text{ML}})^T$ of $\boldsymbol{\beta}_{\text{ML}}$ and are hence identical for all coefficients corresponding to the same group.

2.3 Optimising the penalty parameter ν

Next, different approaches to optimise the penalty parameter ν are presented. Note that there exists a variety of alternatives, though no obvious best choice exists.

First, the AIC from Akaike (1973) has been shown to be a viable option for copula models (see, e.g. Marra and Radice 2017 or van der Wurp et al. 2020). Second, the BIC (Schwarz 1978) can be suitable when a stronger penalisation of the number of coefficients estimated is preferable and, hence, if sparser models are desirable. Third and last, we implemented a K-fold cross validation (CV) approach (with $K = 10$ in both the simulation and application chapters), which uses the unpenalised predictive external log-likelihood $\ell(\hat{\boldsymbol{\beta}}|\boldsymbol{y}_{test}, \boldsymbol{X}_{test})$ (exLLH) as a goodness-of-fit measure on the unseen test fold, where $\boldsymbol{y}_{test}$ denotes the response and $\boldsymbol{X}_{test}$ the covariate design matrix from the left out fold and $\hat{\boldsymbol{\beta}}$ was estimated on all other folds used for training.

The resulting fit of each of those approaches is observed together with the corresponding optimal value of ν_{opt}. In the following simulation study, those models and each approach's tuning for ν are compared with regard to their performance, which is measured in two dimensions, the first being the SSE on the coefficients and the second being the true-positive and true-negative ratios of coefficients, indicating how many features and noise variables were identified as such (see Sect. 3).

To optimise ν, a suitable grid of values needs to be chosen. A possible pragmatic strategy is to start with an arbitrary small value of ν and increase the value stepwise until all penalised coefficients fall below a chosen threshold ε_{lasso} and are set to zero (see next section for more details). A suitable grid from 0 to the found value of ν is created, putting more emphasis (by using smaller steps) on the lower end.

2.4 Threshold ε_{lasso} for approximative LASSO

Note that due to the quadratic approximation of the penalty in equations (6) and (11), coefficient estimates cannot be set exactly to zero. Instead, coefficients that should be zero differ from zero in some late decimals. For this reason, one usually uses rounded coefficients, with the consequence that estimates very close to zero are simply set to zero, and the corresponding covariates are excluded from the model (see, e.g. Table 1 in Schauberger and Tutz 2017, and Footnote 18 in Hambuckers et al. 2018). In both our simulations (Sect. 3) and the application (Sect. 4), a threshold of $\varepsilon_{lasso} = 0.01$ turned out to be a good choice for these specific scenarios. As the approximative LASSO is using standardised covariates, the parameter of ε_{lasso} itself is quite easy to interpret. Our choice of 0.01 implies a change of 0.01 in the linear predictor if the corresponding covariate is changed by one standard deviation. Although, principally, the ε_{lasso} could also be tuned, reasonable choices corresponding to the aforementioned interpretation can be made in the context of the individual application. Our choice of 0.01 is both easy to interpret and yields satisfying results in both the application and the corresponding simulation (which was motivated by the application).

However, as this parameter can also be seen as kind of a "second layer tuning parameter", in future research one could think about more sophisticated choices, which lead to good results independent of the specific application at hand, such as e.g. choices based on certain quantiles of the unregularised maximum-likelihood estimates (in the spirit of the adaptive LASSO). This, however, is beyond the scope of the present work.

3 Simulation study

In this section, we present a thorough simulation study to show the usefulness of our proposed penalisation approach. A selection of different setups, copula classes and dependency strengths will be investigated.

3.1 Setting

We create covariates $x_1, \ldots, x_8$ sampled from correlated Gaussian distributions (see Table 1), which are chosen to have an influence on one of the marginal parameters each. Additionally, we create noise variables $z_1, \ldots, z_{30}$ of which each will be used in only one marginal regression approach and additional noise variables $z_{31}, \ldots, z_{35}$ that will be assigned to both margins simultaneously. This is done to simulate settings where some variables can be assumed to have an influence on both margins at the same time and some do not. The groups (x_1, x_2, x_3) and

Table 1 Sampling distributions for covariates

Covariate	Distribution	Covariate	Distribution
$(x_1, x_2, x_3)^T$	$\mathcal{N}((-1,-1,-1.5)^T, \Sigma)$	x_8	$\mathcal{N}(-1, 0.5^2)$
$(x_4, x_5, x_6)^T$	$\mathcal{N}((-0.25, 1, -1)^T, \Sigma)$	$z_1, \ldots, z_{30}$	$\mathcal{N}(5, 2^2)$
x_7	$\mathcal{N}(2, 0.5^2)$	$z_{31}, \ldots, z_{35}$	$\mathcal{N}(5, 2^2)$

Covariance matrix Σ yields correlations of $r_{x_1x_2} = 0.8, r_{x_1x_3} = 0.5$ and $r_{x_2x_3} = 0.3$ (and identical for x_4, x_5, x_6), respectively

(x_4, x_5, x_6) are sampled with high levels of correlation to include the setting of highly collinear covariates. One group is only used in the first marginal regression formula while the second is margin-overlapping. Another setting without multicollinearity was investigated as well. This yielded virtually the same results and was therefore excluded.

The marginal Poisson coefficients λ_{i1} and λ_{i2} were then specified via

$$\lambda_{i1} = \exp\left(\beta_0^{(1)} + x_1\beta_1^{(1)} + x_2\beta_2^{(1)} + x_3\beta_3^{(1)} + x_4\beta_4^{(1)}\right), \tag{12}$$

$$\lambda_{i2} = \exp\left(\beta_0^{(2)} + x_5\beta_1^{(2)} + x_6\beta_2^{(2)} + x_7\beta_3^{(2)} + x_8\beta_4^{(2)}\right), \tag{13}$$

while the copula parameter θ is not depending on covariates and is hence specified by an intercept $\beta_0^{(\theta)}$ only together with a suitably chosen link function $g(\cdot)$. Each pair of outcomes (y_{i1}, y_{i2}) is then sampled from a given copula with marginal Poisson parameters from equations (12) and (13). The selection of copula classes consists of the Archimedean copulas N (Gaussian), C0 (Clayton), F (Frank), G (Gumbel), C90 (Clayton rotated by 90 degrees), and J (Joe).

To depict settings with different dependency strengths, a grid of values for Kendall's τ is chosen as $(-0.75, -0.5, -0.25, -0.1, 0.1, 0.25, 0.5, 0.75)$. Respective copula parameters θ (note that all copulas mentioned above depend on a scalar parameter) are derived from these τ values with given conversions (for more details in this regard, see, e.g. van der Wurp et al., 2019).

We use the SSE in coefficients such as

$$\text{SSE} = \sum_{r=0}^{4}\left[\left(\beta_r^{(1)} - \hat{\beta}_r^{(1)}\right)^2 + \left(\beta_r^{(2)} - \hat{\beta}_r^{(2)}\right)^2\right] + \sum_{r=5}^{24}\left[\left(\hat{\beta}_r^{(1)}\right)^2 + \left(\hat{\beta}_r^{(2)}\right)^2\right]$$

with true coefficients

$$\boldsymbol{\beta}^{(1)} = (\beta_0^{(1)}, \ldots, \beta_4^{(1)}, 0, \ldots, 0)^T = (0.55,\ 0.1,\ 0.15, 0.1,\ -0.1,\ 0, \ldots, 0)^T,$$
$$\boldsymbol{\beta}^{(2)} = (\beta_0^{(2)}, \ldots, \beta_4^{(2)}, 0, \ldots, 0)^T = (0.75,\ -0.2,\ 0.10,\ -0.20,\ -0.25,\ 0, \ldots, 0)^T.$$

Note again that due to the quadratic approximation of the penalty terms, coefficients are never estimated to exact zero by the fitting routine (compare Sect. 2.4), even for very large values of the penalty parameter ν. Hence, for all simulations a suitable

threshold for the coefficients to be set to zero (after standardising the covariates) had to be chosen. Here, the choice $\epsilon_{lasso} = 0.01$ turned out to be adequate. Stricter or less strict thresholds may be discussed and investigated. We also derive the true positive rate (TPR) and true negative rate (TNR) on the coefficients to investigate if our LASSO model is able to correctly identify features and noise. They are calculated via

$$\text{TPR} = \frac{1}{10} \sum_{r=0}^{4} \mathbb{1}_{\left(\beta_r^{(1)} \neq 0\right)} + \mathbb{1}_{\left(\beta_r^{(2)} \neq 0\right)} \in [0, 1], \tag{14}$$

$$\text{TNR} = \frac{1}{40} \sum_{r=5}^{24} \mathbb{1}_{\left(\beta_r^{(1)} = 0\right)} + \mathbb{1}_{\left(\beta_r^{(2)} = 0\right)} \in [0, 1]. \tag{15}$$

Both TPR and TNR from (14) and (15) are to be maximised, while of course the SSE is to be minimised.

As copula classes C0, G, and J cannot depict negative correlation structures in terms of Kendall's τ, while C90 cannot depict positive ones, and N and F can do both, this approach overall creates $8 + 8 + 4 + 4 + 4 + 4 = 32$ different settings regarding copula class and it's parameter θ corresponding to 8 or 4, respectively, different values of Kendall's τ. A sample size of $n = 250$ was chosen and each setting repeated 100 times.

3.2 Results

The results have to be analysed in multiple dimensions. First, the raw performance, measured by the SSE, is displayed in Fig. 1 for the Gaussian (N) copula class (both used to generate data and to fit the model). Principally, the

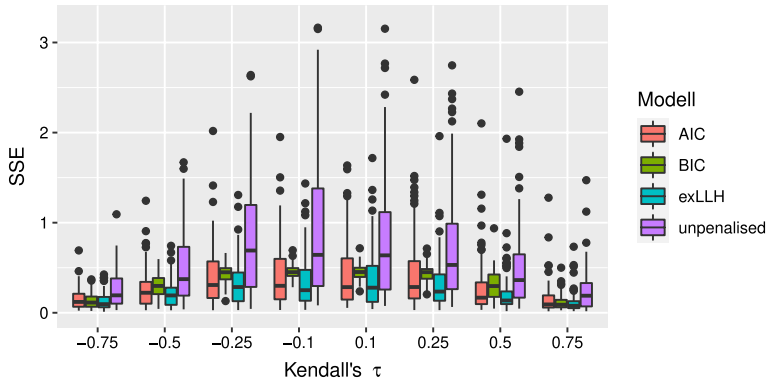

Fig. 1 Simulation results of the standard GJRM approach and LASSO-penalised versions of it for Gaussian (N) copula models in terms of SSE on the estimated and true coefficients for 100 settings of each correlation strength τ

LASSO-penalty approach clearly improves the estimations compared to the unpenalised one no matter which strategy is chosen for tuning the penalty strength. The differences between using AIC, BIC or a cross validation approach for the optimisation of the penalty parameter ν are rather small, though the results from AIC and CV show more variability. In this setting, the BIC can compete with the other two approaches. However, the simulation setup with 10 features and 40 noise variables should be kept in mind here – a higher ratio of features to noises might not favour the stronger penalisation that clearly.

Second, the feature selection ability, i.e. the ability to correctly identify features and noise, is depicted by the true positive and true negative rates. The TPR is the ability to identify features. Figure 2 illustrates the TPR exemplarily for the case of a Gaussian copula. The unpenalised ML-estimation, unsurprisingly, achieves a perfect score of 1 ($\hat{=}$ 100%). The results reflect the SSE-results from above, as again AIC and CV approaches are very similar in performance.

The same applies for the TNR (see Fig. 3, again exemplarily for a Gaussian copula): the stronger penalisation resulting from the BIC is able to detect more noise variables than all other approaches.

In summary, using the AIC or a CV approach leads to roughly the same performance in all mentioned dimensions. Taking the high computational expenditures for the CV into account, it is deemed redundant, and the AIC is preferable. However, comparing the AIC- and BIC-results reveals some substantial differences. As expected, a stronger penalisation will lead to a higher TNR and a lower TPR. In terms of the SSE on estimated and true coefficients, the BIC approach seems to be slightly better. But minimising the SSE might not always be the main goal. In particular, if prediction of new observations is the main objective, a more sparse model might be preferable. Moreover, a sparser model is typically easier to interpret and, hence, might be preferred by practitioners, as long as the loss in performance is relatively small.

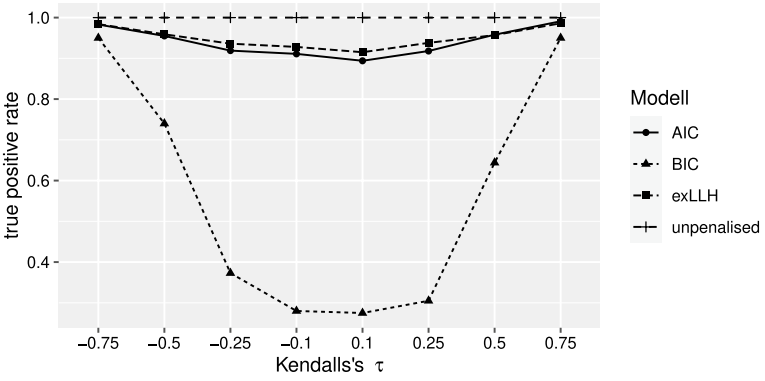

Fig. 2 Simulation results of the standard GJRM approach and LASSO-penalised versions of it for Gaussian (N) copula models in terms of true positive rate, depicting the mean ratio of correctly identified model features for 100 settings of each correlation strength τ

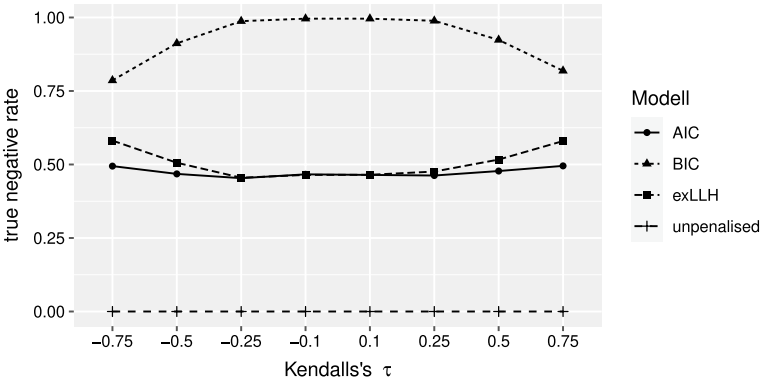

Fig. 3 Simulation results for of the standard GJRM approach and LASSO-penalised versions of it for Gaussian (N) copula models in terms of true negative rate, depicting the mean ratio of correctly identified noise variables for 100 settings of each correlation strength τ

Additionally, a structure depending on the copula parameter θ (and the correlation in terms of Kendall's τ), was observed. Interestingly, a stronger copula structure lessens the difference between results for AIC and BIC for both TPR and TNR (Figs. 2 and 3). This behaviour is also visible for the unpenalised ML approach (Fig. 1). In settings with stronger dependency, the resulting SSE in the estimates is lower compared to weaker dependency settings. As of now, this has no implications on how to optimise the penalisation parameter or our penalisation methodology itself.

Although Figs. 1, 2, 3 focus on the Gaussian (N) Copula class as an example, the results were virtually identical for copula classes Clayton (C0), Frank (F), Clayton rotated (C90), Gumbel (G0) and Joe (J), confirming the previous findings (see Fig. 7 for SSE, Fig. 8 for true positive rate, and Fig. 9 for true negative rate, all in appendix).

Note that we focused here only on simulation settings with $p < n$, as the type of football applications we are considering, i.e. single matches as observations with covariates on the team level, most typically are embedded in this framework. In other contexts, particularly in medical ones with e.g. gene information on the covariate side, p could potentially be much larger or maybe even exceed n (the latter is usually known as the "p larger than n case"). Principally, the proposed LASSO approach can also be applied to such settings, but we recommend to extend the simulations presented here in this direction then.

4 Application on FIFA world cup data

The proposed penalisation approach is now applied to a real world football data set containing FIFA World Cup matches from the tournaments in 2002, 2006, 2010, 2014 and 2018 with 64 matches each.

4.1 Data

The data set originates from Groll et al. (2015) and was also already used by Schauberger and Groll (2018), by Groll et al. (2019) and, finally, used and expanded by van der Wurp et al. (2020), from where essentially the following summary was taken.

(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); 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, World Cup 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) *Deviation from average age*. The absolute deviation of each squad's age from the average age of all teams, i.e. $|\text{age}_i - \overline{\text{age}}|$ (here: $\overline{\text{age}} = 27.171$).
(j) *Number of Champions League (Europa League) players*. As a measurement of the success of the players at 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.

Table 2 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
⋮	⋮	⋮

(b) Table of covariates

Year	Team	AgeDev	Rank	Oddset	...
2002	France	1.129	1	0.149	...
2002	Senegal	2.871	42	0.006	...
2002	Uruguay	1.871	24	0.009	...
2002	Denmark	0.229	20	0.012	...
⋮	⋮	⋮	⋮	⋮	⋱

(c) Matched data set

y_1	y_2	Team1	Team2	Year	AgeDev1	AgeDev2	...
0	1	FRA	SEN	2002	1.129	2.871	...
1	2	URU	DEN	2002	1.871	0.229	...
1	1	DEN	SEN	2002	0.229	2.871	...
0	0	FRA	URU	2002	1.129	1.871	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

These variables are available for each team and are used to model the number of goals scored by each team per individual match. Table 2 shows a shortened version of the full data set. The final data set (Table 2c) was created by matching the teams' covariates (Table 2b) with the match result data (Table 2a). Due to the FIFA tournament structures, the order in the matches (a team being first or second named, as World Cups do not have a home and away team for every match) are not completely random. To avoid influences correlating with the marginal intercepts, we removed these structures by performing a coin-toss for each match, deciding to flip the first and second named teams or to keep them as they are.

The model formulas in R-pseudo code are given by

$$\begin{aligned}\lambda_1 \sim\ & 1 + CL.players(1) + UEFA.players(1) + nation.coach(1) + age.coach(1) + \\ & tenure.coach(1) + legionaires(1) + max.teammates(1) + sec.max.teammates(1) + age.dev(1) + \\ & rank(1) + GDP(1) + host(1) + confederation(1) + continent(1) + odds(1) + population(1) + \\ & knockout + titleholder(1),\end{aligned}$$

$$\begin{aligned}\lambda_2 \sim\ & 1 + CL.players(2) + UEFA.players(2) + nation.coach1(2) + age.coach(2) + \\ & tenure.coach(2) + legionaires(2) + max.teammates(2) + sec.max.teammates(2) + age.dev(2) + \\ & rank(2) + GDP(2) + host(2) + confederation(2) + continent(2) + odds(2) + population(2) + \\ & knockout + titleholder(2),\end{aligned}$$

and

$$\theta \sim 1,$$

where the inclusion of 1 corresponds to the use of an intercept. All covariates are team-specific and denoted by (1) and (2) for the respective team, except for the variable Knockout. The copula parameter θ is only modelled via an intercept, meaning $\hat{\theta}$ will be a fixed value for all matches.

4.2 Comparison of predictive performance

This section presents six measures to investigate the predictive performance and is principally a shortened version of van der Wurp et al. (2020), as both the setting and the principal aim of the analysis were rather similar.

As prediction of future matches here is the main objective, we focus on out-of-sample data to validate the models. We therefore incorporated a cross-validation approach. Each time four out of five World Cups were used to fit the model (including optimisation of ν from Sect. 2.2). The left-out World Cup was then used to validate the resulting models in multiple dimensions.

We take multiple measures into account to observe the quality of predictions. A natural candidate to evaluate the predictive performance of the models is to derive the (predictive log-)likelihood (exLLH) from equation (4) on all observations of the respective left-out World Cup.

Closely related to this, we also calculate the Euclidean distance between observation (y_1, y_2) and prediction $\hat{\boldsymbol{\lambda}}_i = (\hat{\lambda}_{i1}, \hat{\lambda}_{i2})^T$. Note that $\boldsymbol{\lambda}_i = (\lambda_{i1}, \lambda_{i2})^T$ is the bivariate mean of the bivariate distribution for a single match i in a given copula model. The corresponding MSE can then be calculated via

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{y}_i - \hat{\boldsymbol{\lambda}}_i||_2^2 = \frac{1}{n}\sum_{i=1}^{n}(y_{i1} - \hat{\lambda}_{i1})^2 + (y_{i2} - \hat{\lambda}_{i2})^2$$

over a set of n matches.

Alternatively, when it comes to modelling football matches, the three match outcomes *win*, *draw* and *loss* are also of particular interest. Hence, we also provide several performance measures that focus on these categorical outcomes. First, we calculate the resulting *three-way probabilities* $\hat{\pi}_{il}$ for the i-th match. The index $l = 1, 2, 3$ indicates *win*, *draw* and *loss* from the first mentioned team's perspective. Alternatively, these outcomes principally could have been also modelled directly by using models for categorical responses. Instead, we model the outcome in terms of goals (y_1, y_2) by decision to include more information.

With $\hat{\pi}_{il}$ from above, the following measures can be applied: The rank probability score (RPS see, e.g. Schauberger and Groll 2018). For a given match i, let the dummy coding of the true result of *win*, *draw*, *loss* be denoted by Kronecker's delta δ_{il}. In the case of three possible outcomes, the RPS is defined via

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

Second, the multinomial likelihood (MLLH) is defined as

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

which is essentially the predicted probability $\hat{\pi}_{il}$ for the true outcome. Third, the classification rate (CR) indicates how many matches have been classified into the three-way outcomes correctly, assuming that the highest probability refers to a classification. It is calculated via

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

The last dimension of predictive performance is the result of betting strategies for the FIFA World Cup 2018. With given predicted probabilities from our models and corresponding betting odds from bookmakers (taken from oddsportal.com), the expected return per bet can be calculated such as $E[\text{return}_{il}] = \pi_{il} \cdot \hat{\text{odds}}_{il} - 1$. The simplest possible betting strategy, to bet only on outcomes with a positive expected return and only one outcome per match, can be expanded by adding a threshold ε. Only bets with an expected return of $> \varepsilon$ should be placed in this case, indicating a more or less careful (depending on the choice of ε) approach. Here, we simply used $\varepsilon = 0$.

4.3 Results

All calculations have been performed with adaptive weights and a Group LASSO approach for the factor covariates of the confederations. Before presenting results, we will give a short overview of the used benchmarks and models. The reference model, an independent Poisson approach, using two Poisson distributions with no dependence structure at all, was used as the most simple benchmark. Another benchmark model accounts for the copula structure and is the standard approach via `GJRM`. van der Wurp et al. (2020) improved these results by incorporating a novel penalty to penalise the differences between corresponding marginal coefficients, which can be assumed to be equal.

Due to the computational intensity of the CV approach, we only focus on the promising copula classes, namely Frank (F), which already performed well in prior research. First, the GJRM with LASSO penalty structure from Sect. 2.2 is applied and evaluated. Second, we combine this penalty with the one from van der Wurp et al. (2020) to allow for general sparsity and the assumption of equal marginal coefficients at the same time. For the LASSO penalty, multiple tuning approaches for the penalty parameter ν are used, namely AIC, BIC and the predictive log-likelihood (exLLH). Other CV-based tuning approaches (e.g. CV-measures on the three-way probabilities like RPS, multinomial likelihood and classification rate) could generally also easily be implemented.

But before going into detailed results, we will present a short overview about the model's properties in the regularisation context, showing LASSO path plots and how the presented optimisation approaches for ν from Sect. 2.3 are deciding on penalty strengths.

While passing through the ν-grid, the estimation process is given a vector of zeroes as starting values when using the LASSO penalty only. However, when both penalties are active simultaneously, the previous estimates are given as starting values for the next ν-step. In both our application and simulations, this yielded best results.

4.3.1 LASSO results and properties

To create the exemplary plots and to show the influence of ν , we use the full dataset of all five World Cups from 2002 to 2018 and the Frank copula class.

The resulting path plots are depicted in Fig. 4. The penalty's quadratic properties lead to superlinear growth in coefficients β_i for decreasing ν. For better readability, it may be advised to show results on the index of ν instead. 'Wiggly' paths (paths that are not increasing monotonously from right to left) are the result of substantial multicollinearity and correlations and are expected in real life data. The tuning for ν was performed with AIC, BIC and the cross validated predictive log-likelihood. In all applications related to this dataset, tuning by BIC results in the highest penalty strength, usually yielding a very sparse model, often simply the intercept model (see also Fig. 4). Tuning by AIC results in the least penalised models, often rather close to unregularised ML estimation. Third, tuning by the cross validated predictive log-likelihood yields models in between, offering a compromise. As prediction strength is deemed important in this setting, we deem this to be the most sensible approach.

Figure 5 depicts the course of the predictive log-likelihood. The very common structure of lower values for a strong penalty (reading from the right end) that increase for a decreasing penalty strength is clearly visible. And at the left end (small ν values), the model enters the area of overfitting, resulting in a decreasing likelihood.

The AIC curve (depicted in Fig. 6) has roughly the expected structure as well. The uptick in increase in degrees of freedom (dfs; here estimated by the number of non-zero coefficients, see also top graph in Fig. 6) for the smallest values of ν manages to overcome the increase of likelihood, resulting in a slightly increased AIC close to the ML estimation. The AIC yields the most complex model. The BIC, on the other hand, prefers stronger penalised models. In all applications to this data set, optimising the BIC actually coincided with minimising the dfs, resulting in the mere intercept model. As discussed earlier, we deem the predictive log-likelihood to be the most sensible approach. This is corroborated by the results from Fig. 6.

4.3.2 Prediction results

The results for our cross-validation-like approach cycling through all five tournaments can be found in Table 3. Rows 1 to 3 are the updated results from van der Wurp et al. (2020), improving the model by first incorporating a (in this scenario rather weak) copula structure and then penalising the differences between marginal coefficients. The small differences between row 1 and 2 indicate that the Frank copula structure is only a small improvement, if at all, in this specific setting.

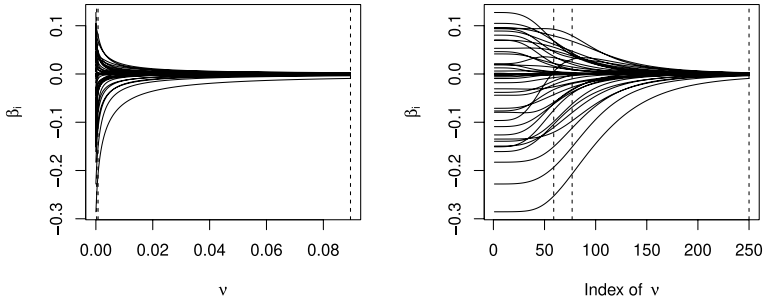

Fig. 4 LASSO path plots depending on the penalty strength ν (left) and (for better readability) by the index points of our ν-grid (right). The vertical marks are the chosen optimal penalty parameters by internal AIC, predictive log-likelihood and internal BIC, left to right in that order. Intercepts paths are not included, as they are not penalised

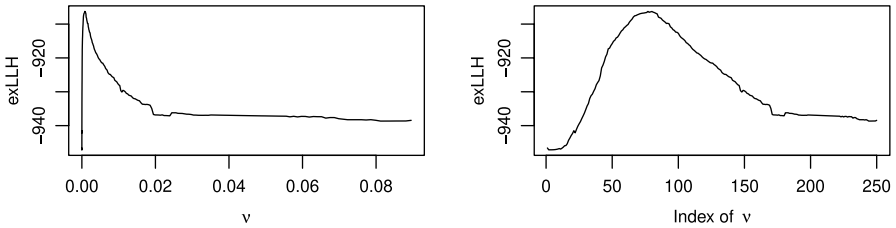

Fig. 5 Curve for predictive log-likelihood (exLLH) in cross validation, depending on ν (left) and the index points of our ν-grid (right), summed up across all observations and folds

By extending these approaches with a LASSO penalty, we were able to achieve further improvements regarding the predictive performance measures. Row 4 of Table 3 shows a strongly improved MSE and predictive log-likelihood compared to the unpenalised model from row 2. In this application, the CV approach seems superior compared to the AIC and BIC equivalents. We deem this is the most sensible tuning procedure if prediction strength is one of the main goals. Additionally, we focus on the results for MSE and predictive log-likelihood instead of the other measures, as they rely on the three-way probabilities and observations, which the model itself does not take into account.

It is notable that if coefficients are forced to be equal (row 3) this yields roughly the same improvements as the LASSO model from row 4. The combination of both approaches, however, which is easily obtained with our extensions to the `GJRM` infrastructure, yields no further improvements (row 7) but performs decent as well with regard to all measures. As the resulting model (row 7) might be more sparse compared to the LASSO-only version (row 4), due to the vectors of coefficients being equal between the margins, we deem both models to be potential winners with a high value regarding interpretability. By no means, this implies that the penalties

should be combined in general. It is easy to imagine settings where both penalties are suitable and where only one of them is.

As we use the application mostly for illustrative purposes, we abstain from interpreting the obtained estimates in detail. The model's coefficients can be found in Table 4 for the model with both penalties from row 7 and in Table 6 in the appendix for the LASSO-only model from row 4.

Due to substantial multicollinearities and correlations the coefficients need to be interpreted with caution. The missing coefficient for being the current titleholder (despite France's miserable performance in 2002, Italy's in 2010, Spain's in 2014 and of course Germany's in 2018) is noted as a tongue-in-cheek remark. Interestingly, in the LASSO-model (see Table 6 in the appendix), the titleholder's curse can be found, but only for the first named team. As we swapped the teams randomly, this is considered to be an artefact and a prime example for the usefulness of both penalties combined. Some other coefficients are quite intuitive, like the negative influence of a match being a knockout-game (bad teams already dropped out at this phase of the tournament) or the negative influence of the FIFA rank (high numerical values indicate bad ranked teams).

4.4 Stage-wise prediction of FIFA world cup 2018

To demonstrate possible prediction approaches, we predicted the FIFA World Cup of 2018 in each stage. The group stage of 48 matches was predicted using the four

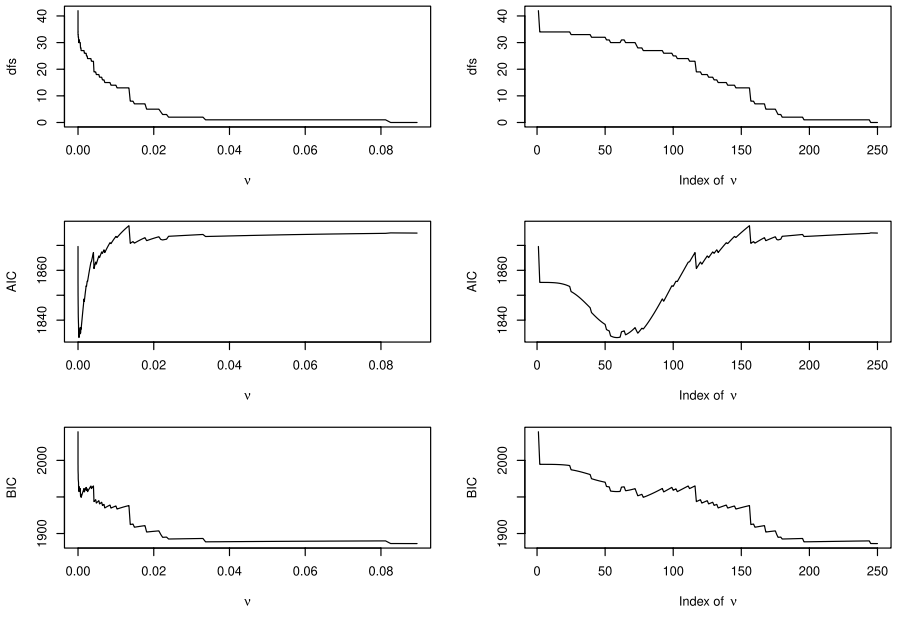

Fig. 6 Curves for degrees of freedom (top row), AIC (2nd row) and BIC (bottom row), depending on ν (left) and the index points of our ν-grid (right)

Table 3 Results for all six performance measures and a selection of models

	Cop	Penalty	Tuner	MSE	exLLH	RPS	MLLH	Cl. rate	Betting
1	–	–	–	2.9829	−187.0996	0.2012	0.4011	0.5250	0.0623
2	F	–	–	3.0154	−187.2698	0.2008	0.4086	0.4969	0.0959
3	F	Equal	–	2.8191	−183.7188	0.1990	0.4007	0.5219	0.0494
4	F	LASSO	CV	2.7783	−182.8722	0.2012	0.3885	0.5219	−0.1230
5	F	LASSO	AIC	2.8567	−184.1996	0.1996	0.3994	0.5125	−0.0609
6	F	LASSO	BIC	2.9302	−187.3526	0.2331	0.3401	0.3500	0.1298
7	F	Equal+LASSO	CV	2.7940	−183.7778	0.2038	0.3878	0.5188	0.1648

MSE and predictive log-likelihood (exLLH) correspond to the original response, while RPS, predictive multinomial log-likelihood (MLLH), classification rate and betting (gains relative to sum of stakes) correspond to measures on the three-way probabilities

previous tournaments and then added to the training data set to predict the round of 16, continuing this procedure throughout the tournament. The final match was grouped together with the match for the 3rd place. The results can be found in Table 5.

This corroborates the results from before. In a purely predictive setting, both penalty structures improve the quality of predictions in terms of MSE and predictive log-likelihood. The other measures are based on the three-way-outcome and are not directly taken into account in the fitting process of the model. Especially the betting results are to be taken cautiously, as these are extremely reliant on single events, e.g. South Korea's win against Germany with bookmaker's odds of 19.2. All penalties yield a visible improvement in prediction.

5 Conclusions

In this work, we proposed and incorporated LASSO penalties into the `GJRM` framework used to model bivariate count data responses with a copula structure. We also included adaptive weights into the estimation scheme and, additionally, implemented a group LASSO methodology for the handling of categorical predictors. The proposed approach adds previously missing regularisation in this context and is able to provide some control over sparsity.

We investigated the penalty's performance in a simulation study and showed its usage in a real life data set of FIFA World Cup matches. The simulation results are pretty clear: If some variables are expected to be noise, the LASSO penalty can be used to identify relevant model features and detect noise variables. It clearly outperforms the corresponding unpenalised models for all investigated copula classes. Both the AIC the BIC are reasonable choices for tuning the LASSO penalty strength. Additionally, we identified a cross validation approach, which is based on the predictive out-of-sample log-likelihood, as a sensible compromise between the very sparse BIC-results and the barely penalised AIC-results.

One aspect that turned out to be problematic in the present application was the presence of substantial multicollinearity among the covariates. Hence, in future work the

Table 4 Coefficients for the fitted model with a combination of both LASSO and equalisation penalty

	β		β
(Intercept)	0.830	GDP	0.024
CL.players	0.033	Host1	0.299
UEFA.players	0.039	ConfedCAF	0.089
Nation.coach1		ConfedCONCACAF	0.001
Age.coach	−0.007	ConfedCONMEBOL	0.319
Tenure.coach	−0.033	ConfedUEFA	0.176
Legionaires		Continent1	
Max.teammates		Odds	
Sec.max.teammates		Population	0.018
Age		KnockoutTRUE	−0.380
Rank	−0.007	Titleholder	

Missing values imply an estimate of zero

presented LASSO penalty could be generalised to allow for an elastic net approach (see Zou and Hastie 2005), combining LASSO and ridge penalisation. This is known to be particularly useful in situations with collinearity and correlation in the design matrix.

As this work was motivated by the application to football data, high dimensional cases with p very large or even larger than n were not investigated. However, principally, the proposed LASSO approach can also be used in such settings, which are common e.g. in gene expression data. Hence, in future research such settings could be further investigated.

Moreover, the presented modelling approach is so far restricted to linear or simple polynomial covariate effects only. Hence, in future research it is planned to combine the penalty approach proposed in van der Wurp et al. (2020) with the option to model smooth, nonlinear effects, e.g. via P-splines (see Eilers and Marx 1996, or Wood 2017). In order to maintain the merits of sparsity and variable selection, here the adaptation of existing boosting approaches designed for generalised additive models (see, e.g. Tutz and Binder 2006, Schmid and Hothorn 2008, or Groll and Tutz 2012) or generalised additive models for location, scale and shape (see Mayr et al. 2012; Hofner et al. 2014) seems promising.

Table 5 Results for all six performance measures and a selection of models for the stage-wise prediction of FIFA World Cup 2018

	Cop	Penalty	Tuner	MSE	exLLH	RPS	MLLH	cl. rate	betting
1	–	–	–	2.9577	−2.9528	0.2217	0.3851	0.5313	0.0431
2	F	–	–	3.0213	−2.9904	0.2242	0.3897	0.4219	−0.3661
3	F	Equal	–	2.7010	−2.8881	0.2154	0.3784	0.4531	0.0122
4	F	LASSO	CV	2.7314	−2.8841	0.2197	0.3808	0.5156	0.1127
5	F	LASSO	AIC	2.8380	−2.9161	0.2219	0.3829	0.4688	0.0779
6	F	LASSO	BIC	2.6566	−2.8558	0.2396	0.3440	0.3594	0.1298
7	F	Equal+LASSO	CV	2.6270	−2.8581	0.2091	0.3830	0.5000	0.0652

MSE and predictive log-likelihood (average per game, exLLH) correspond to the original response, while RPS, predictive multinomial log-likelihood (MLLH), classification rate and betting (gains relative to sum of stakes) correspond to measures on the three-way probabilities

Appendix

See Figs. 7, 8, 9 and Table 6

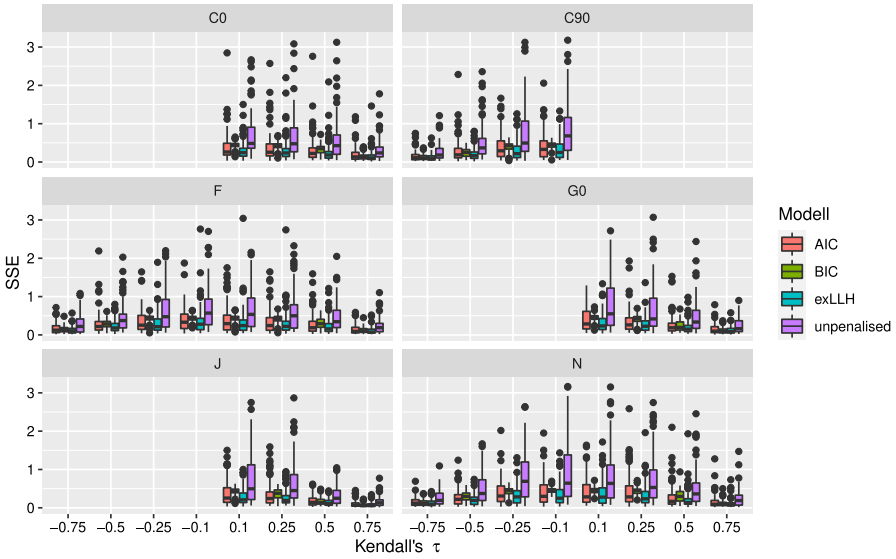

Fig. 7 Simulation results of the standard GJRM approach and LASSO-penalised versions of it for different copula models in terms of SSE on the estimated and true coefficients for 100 settings of each correlation strength τ

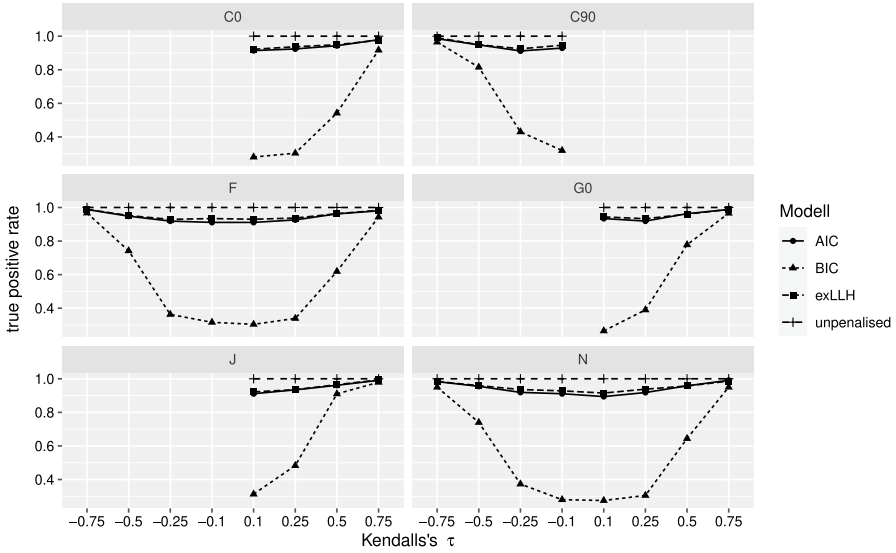

Fig. 8 Simulation results of the standard GJRM approach and LASSO-penalised versions of it for different copula models in terms of true positive rate, depicting the mean ratio of correctly identified model features for 100 settings of each correlation strength τ

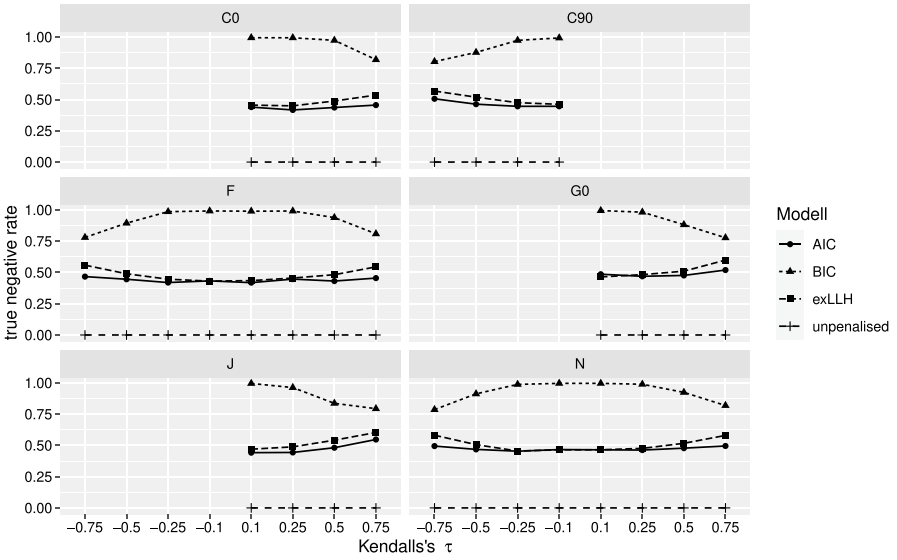

Fig. 9 Simulation results of the standard GJRM approach and LASSO-penalised versions of it for different copula models in terms of true negative rate, depicting the mean ratio of correctly identified noise variables for 100 settings of each correlation strength τ

Table 6 Coefficients for the fitted model with LASSO penalisation

	$\boldsymbol{\beta}^{(1)}$	$\boldsymbol{\beta}^{(2)}$		$\boldsymbol{\beta}^{(1)}$	$\boldsymbol{\beta}^{(2)}$
(Intercept)	1.338	0.261	GDP	0.007	0.019
CL.players	0.016	0.035	host1	0.194	0.188
UEFA.players	0.050		confedCAF	−0.164	0.033
Nation.Coach1	0.077		confedCONCACAF	−0.063	−0.015
Age.Coach	−0.013		confedCONMEBOL	0.102	0.011
Tenure.Coach	−0.041		confedUEFA	0.062	0.125
Legionaires			continent1		
Max.teammates	−0.020		odds	0.642	0.501
Sec.max.teammates			Population	0.007	0.038
Age	−0.026		KnockoutTRUE	−0.349	−0.144
Rank	−0.012		Titleholder	−0.765	

Missing values imply an estimation of zero

Acknowledgements The authors like to thank M. Oelker for the fruitful discussions regarding potential tuning of the threshold parameter ϵ_{lasso} of the approximative LASSO. Moreover, the authors are grateful for the comments of two anonymous referees and the associate editor, which helped to substantially improve this work.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

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

Dixon, M.J., Coles, S.G.: Modelling association football scores and inefficiencies in the football betting market. J. Roy. 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)

Eilers, P.H.C., Marx, B.D.: Flexible smoothing with B-splines and penalties. Stat. Sci. **11**, 89–121 (1996)

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

Geyer, CJ.: Trust: Trust Region Optimization. (2015). https://CRAN.R-project.org/package=trust, r package version 0.1-7

Groll, A., Abedieh, J.: Spain retains its title and sets a new record: generalized linear mixed models on European football championships. J. Quant. Anal. Sports **9**(1), 51–66 (2013)

Groll, A., Tutz, G.: Regularization for generalized additive mixed models by likelihood-based boosting. Methods Inf. Med. **51**(2), 168–177 (2012)

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)

Groll, A., Hambuckers, J., Kneib, T., Umlauf, N.: Lasso-type penalization in the framework of generalized additive models for location, scale and shape. Comput. Stat. Data Anal. **140**, 59–73 (2019)

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**, 271–287 (2019)

Hambuckers, J., Groll, A., Kneib, T.: Understanding the economic determinants of the severity of operational losses: A regularized generalized Pareto regression approach. J. Appl. Economet. **33**(6), 898–935 (2018)

Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics **12**, 55–67 (1970)

Hofner, B., Mayr, A., Schmid, M.: gamboostlss: An r package for model building and variable selection in the gamlss framework. (2014). arXiv preprint arXiv:14071774

Karlis, D., Ntzoufras, I.: On modelling soccer data. Student **3**(4), 229–244 (2000)

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

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

Ley, C., Van de Wiele, T., Van Eetvelde, H.: Ranking soccer teams on basis of their current strength: a comparison of maximum likelihood approaches. Stat. Model. **19**, 55–73 (2019)
Marra, G., Radice, R.: GJRM: generalised joint regression modelling. R package version 0.2-3 (2020)
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.: Copula link-based additive models for right-censored event time data. J. Am. Stat. Assoc. **115**, 886–985 (2019)
Mayr, A., Fenske, N., Hofner, B., Kneib, T., Schmid, M.: Generalized additive models for location, scale and shape for high-dimensional data: a flexible approach based on boosting. J. Roy. Stat. Soc. Ser. C Appl. Stat. **61**(3), 403–427 (2012)
McHale, I., Scarf, P.: Modelling soccer matches using bivariate discrete distributions with general dependence structure. Stat. Neerl. **61**(4), 432–445 (2007)
Meier, L., Van de Geer, S., Bühlmann, P.: The group lasso for logistic regression. J. Roy. Stat. Soc. B **70**, 53–71 (2008)
Nelsen, R.B.: An Introduction to Copulas. Springer, New York (2006)
Nikoloulopoulos, A.K., Karlis, D.: Regression in a copula model for bivariate count data. J. Appl. Stat. **37**, 1555–1568 (2010)
Oelker, M.R., Tutz, G.: A uniform framework for the combination of penalties in generalized structured models. Adv. Data Anal. Classif. **11**(1), 97–120 (2017)
Schauberger, G., Groll, A.: Predicting matches in international football tournaments with random forests. Stat. Model. **18**(5–6), 1–23 (2018)
Schauberger, G., Tutz, G.: Subject-specific modelling of paired comparison data: A lasso-type penalty approach. Stat. Model. **17**(3), 223–243 (2017)
Schauberger, G., Groll, A., Tutz, G.: Analysis of the importance of on-field covariates in the German Bundesliga. J. Appl. Stat. (2017). https://doi.org/10.1080/02664763.2017.1383370
Schmid, M., Hothorn, T.: Boosting additive models using component-wise P-splines. Comput. Stat. Data Anal. **53**, 298–311 (2008)
Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Roy. 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)
Tutz, G., Binder, H.: Generalized additive modelling with implicit variable selection by likelihood based boosting. Biometrics **62**, 961–971 (2006)
van der Wurp, H., Groll, A., Kneib, T., Marra, G., Radice, R.: Generalised joint regression for count data: a penalty extension for competitive settings. Stat. Comput. **30**(5), 1419–1432 (2020)
Wood, S.N.: Generalized Additive Models: An Introduction with R, 2nd edn. Chapman & Hall/CRC, London (2017)
Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. B **67**, 301–320 (2005)
Zou, H., Hastie, T.: The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. **101**, 1418–1429 (2006)