Electronic Journal of Statistics
Vol. 6 (2012) 1517–1549
ISSN: 1935-7524
DOI: 10.1214/12-EJS721

Gaussian copula marginal regression

Guido Masarotto

Department of Statistical Sciences
Università di Padova
Via Cesare Battisti, 241
I-35121 Padova, Italy
e-mail: guido.masarotto@unipd.it

and

Cristiano Varin

Department of Environmental Sciences, Informatics and Statistics
Università Ca' Foscari Venezia
San Giobbe Cannaregio, 873
I-30121 Venice, Italy
e-mail: sammy@unive.it

Abstract: This paper identifies and develops the class of Gaussian copula models for marginal regression analysis of non-normal dependent observations. The class provides a natural extension of traditional linear regression models with normal correlated errors. Any kind of continuous, discrete and categorical responses is allowed. Dependence is conveniently modelled in terms of multivariate normal errors. Inference is performed through a likelihood approach. While the likelihood function is available in closed-form for continuous responses, in the non-continuous setting numerical approximations are used. Residual analysis and a specification test are suggested for validating the adequacy of the assumed multivariate model. Methodology is implemented in a R package called gcmr. Illustrations include simulations and real data applications regarding time series, cross-design data, longitudinal studies, survival analysis and spatial regression.

Keywords and phrases: Discrete time series, Gaussian copula, generalized estimating equations, likelihood Inference, longitudinal data, marginal regression, multivariate probit, spatial data, survival data.

Received February 2012.

Contents

1. Introduction

Marginal regression models for non-normal correlated responses are typically fitted by the popular generalized estimating equations approach of Liang and Zeger [34]. Despite several theoretical and practical advantages, likelihood analysis of non-normal marginal regression models is much less widespread, see Diggle et al. [13]. The main reason is the difficult identification of general classes of multivariate distributions for categorical and discrete responses. Nevertheless, likelihood analysis of marginal models has been considered by many authors, see Molenberghs and Verbeke [39] for some references.

Gaussian copulas [47] provide a flexible general framework for modelling dependent responses of any type. Gaussian copulas combine the simplicity of interpretation in marginal modelling with the flexibility in the specification of the dependence structure. Despite this, Gaussian copula regression had still a limited use since for non-continuous dependent responses the likelihood function requires the approximation of high-dimensional integrals. The intents of this paper are to identify the class of Gaussian copula regression models and to show that methods developed for multivariate probit regression can be usefully adapted to the Gaussian copula models in a way to overcome numerical difficulties of the likelihood inference.

The identified class of models stems from and generalizes previous work on Gaussian copula regression models [47, 43] and multivariate probit models [7]. In particular, we show that with a proper parameterization of the correlation matrix of the Gaussian copula, the ideas of Song [47] can be applied also to the analysis of time series and spatially correlated observations. We suggest model fitting through maximum likelihood. In the continuous case, the likelihood function is available in closed-form. Otherwise, in the discrete and the categorical cases numerical approximations of the likelihood are needed. We propose to approximate the likelihood by means of an adaptation of the Geweke-Hajivassiliou-Keane importance sampling algorithm [30]. Another contribute of this manuscript is to interpret the normal scores of the Gaussian copula as errors and thus develop residuals analysis for model checking. We also suggest to validate the inferential conclusions on the marginal parameters through a Hausman-type specification test [22].

The methodology discussed in this paper has been implemented in the `R` [44] package `gcmr` – Gaussian copula marginal regression – available from the Comprehensive R Archive Network at url `http://cran.r-project.org/web/packages/gcmr`.

2. Framework

Let $\mathbf{Y} = (\mathbf{Y_1}, \dots, \mathbf{Y_n})^{\mathrm{T}}$ be a vector of continuous, discrete or categorical dependent random variables and let $\boldsymbol{y} = (y_1, \dots, y_n)^{\mathrm{T}}$ be the corresponding realizations. Dependence may arise in several forms, as for example repeated measurements on the same subject, observations collected sequentially in time, or georeferenced data. We consider situations where the primary scientific objective is evaluating how the distribution of Y_i varies according to changes in a vector of p covariates $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})^{\mathrm{T}}$. Dependence is regarded as a secondary, but significant, aspect.

Denote by $p_i(y_i; \boldsymbol{\lambda}) = p(y_i | \boldsymbol{x}_i; \boldsymbol{\lambda})$ the density function of Y_i given $\boldsymbol{x}_i$, so that covariates are allowed to affect not only the mean of Y_i but the entire univariate marginal distribution. Within this framework density $p_i(y_i; \boldsymbol{\lambda})$ identifies the regression model. Without further assumptions about the nature of the dependence among the responses, inference on the marginal parameters $\boldsymbol{\lambda}$ can be conducted with the pseudolikelihood constructed under working independence assumptions,

$$\mathscr{L}_{\mathrm{ind}}(\boldsymbol{\lambda}; \boldsymbol{y}) = \prod_{i=1}^{n} p_i(y_i; \boldsymbol{\lambda}). \tag{1}$$

If the marginals $p_i(y_i; \boldsymbol{\lambda})$ are correctly specified, then the maximum independence likelihood estimator $\hat{\boldsymbol{\lambda}}_{\mathrm{ind}} = \mathrm{argmax}_{\boldsymbol{\lambda}} \mathscr{L}_{\mathrm{ind}}(\boldsymbol{\lambda}; \boldsymbol{y})$ consistently estimates $\boldsymbol{\lambda}$ with no specification of the joint distribution of $\mathbf{Y}$ given the model matrix $\mathbf{X} = (\mathbf{x_1}, \dots, \mathbf{x_n})^{\mathrm{T}}$. Despite this important robustness property, there are various causes of concern with the above estimator. First, it may suffer from considerable loss of efficiency when dependence is appreciable. Second, standard likelihood theory does not apply and corrected standard errors should be based on sandwich-type formulas [57], whose computation can be difficult when the response vector $\mathbf{Y}$ does not factorize into independent subvectors. Finally, predictions that do not take into account dependence may be of poor quality.

For these reasons, complementing the regression model $p_i(y_i; \boldsymbol{\lambda})$ with a dependence structure is important for attaining more precise inferential conclusions and predictions. The ideal model would contain all the possible joint distributions of $\mathbf{Y}$ with univariate marginals $p_i(y_i; \boldsymbol{\lambda})$, $i = 1, \ldots, n$. However, this broader semiparametric model appears too general to be of practical use. Hence, in the following we identify and develop a narrower parametric model which is flexible enough for many applications.

3. Gaussian copula marginal regression models

In very general terms, a regression model is expressed as

$$\mathrm{Y}_i = g(\boldsymbol{x}_i, \epsilon_i; \boldsymbol{\lambda}), \quad i = 1, \ldots, n, \tag{2}$$

where $g(\cdot)$ is a suitable function of the regressors $\boldsymbol{x}_i$ and of an unobserved stochastic variable ϵ_i, commonly denoted as the *error* term. It is assumed that the regression model (2) is known up to a vector of parameters $\boldsymbol{\lambda}$. Among the possible specifications for the function $g(\cdot)$, a useful choice is

$$\mathrm{Y}_i = \mathrm{F}_i^{-1}\left\{\Phi(\epsilon_i); \boldsymbol{\lambda}\right\}, \quad i = 1, \ldots, n, \tag{3}$$

where ϵ_i is a standard normal variable and $\mathrm{F}_i(\cdot; \boldsymbol{\lambda}) = \mathrm{F}(\cdot|\boldsymbol{x}_i; \boldsymbol{\lambda})$ and $\Phi(\cdot)$ are the cumulative distribution functions of Y_i given $\boldsymbol{x}_i$ and of a standard normal variate, respectively. By the integral transformation theorem, the regression model (3) ensures the desired marginal distribution for the response Y_i and specifies ϵ_i in the familiar terms of a normal error. Specification (3) includes all possible parametric regression models for continuous and noncontinuous responses. For example, the Gaussian linear regression model $\mathrm{Y}_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \sigma\epsilon_i$ corresponds to set $\mathrm{F}_i(\mathrm{Y}_i; \boldsymbol{\lambda}) = \Phi\{(\mathrm{Y}_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})/\sigma\}$ in (3), with $\boldsymbol{\lambda} = (\boldsymbol{\beta}^{\mathrm{T}}, \sigma)^{\mathrm{T}}$, while the Poisson log-linear model is obtained by setting

$$\mathrm{F}_i(\mathrm{Y}_i; \boldsymbol{\lambda}) = \sum_{j=0}^{\mathrm{Y}_i} \frac{e^{-\mu_i}\mu_i^j}{j!}$$

where $\mu_i = \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, with $\boldsymbol{\lambda} \equiv \boldsymbol{\beta}$.

For subsequent developments, it is important to notice that the mapping between the response Y_i and the error term ϵ_i is one-to-one only in the continuous case, otherwise the mapping is one-to-many and hence the relationship (3) between Y_i and ϵ_i cannot be inverted.

This manuscript deals with regression analysis in presence of dependence. Model specification is then completed by assuming that the vector of errors $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^{\mathrm{T}}$ is multivariate normal,

$$\boldsymbol{\epsilon} \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Omega}), \tag{4}$$

where $\boldsymbol{\Omega}$ is a correlation matrix. The special case of independent observations corresponds to $\boldsymbol{\Omega} = \mathbf{I}_n$, the $n \times n$ identity matrix. Model identifiability requires that $\boldsymbol{\epsilon}$ has zero mean vector and unit variances because univariate characteristics are modelled separately in the marginals $p_i(y_i; \boldsymbol{\lambda})$.

The model specification conveniently separates the marginal component (3) from the dependence component (4), the latter being described in terms of a multivariate normal process. Various forms of dependence in the data can be modelled by suitably parametrizing the correlation matrix $\boldsymbol{\Omega}$ as a function of a vector parameter $\boldsymbol{\tau}$, see Section 3.1 for some common examples. The whole parameter vector is denoted by $\boldsymbol{\theta} = (\boldsymbol{\lambda}^{\mathrm{T}}, \boldsymbol{\tau}^{\mathrm{T}})^{\mathrm{T}}$. Thereafter, $\boldsymbol{\lambda}$ is termed the vector of marginal parameters and $\boldsymbol{\tau}$ the vector of dependence parameters.

The model (3)-(4) offers a natural interpretation in terms of the copula theory [28] where the *normal scores* ϵ_i are seen as nonlinear transformations of the variables Y_i and not as error terms. In this manuscript, we prefer the errors interpretation for the ϵ_i because we think this provides a clear connection to other regression approaches and because this interpretation facilitates the presentation of the residuals analysis described later in Section 7.1.

The approach investigated here differs from much of the existing literature about copula modelling where *viceversa* marginals are treated as nuisance components and interest lies on the dependence parameters of the copula. For an example of the latter use of copulas see the work on semiparametric Gaussian copula modelling by Hoff [23] and the references therein.

As stated in Section 2, ideal inferences on $\boldsymbol{\lambda}$ should be based on the semiparametric model of all the possible copulas with marginals $p_i(y_i; \boldsymbol{\lambda})$. The model discussed here restricts to a particular copula, namely the Gaussian copula [47]. This choice appears advantageous because it naturally inherits several well-known properties of the multivariate normal distribution, see *e.g.* Nikoloulopoulos et al. [41]. A limit of the normality assumption is that the full multivariate dependence structure is induced by the bivariate dependencies. Other copulas might be considered as well but at the cost of lessened interpretability. Further, simulation studies as that reported in Section 8 suggest some amount of robustness of the Gaussian copula to local misspecification of the dependence structure. In the rest of the paper, the class of models identified by the pair of equations (3)-(4) is termed Gaussian copula marginal regression (GCMR) models.

Song [47] introduced Gaussian copula generalized linear models for longitudinal data analysis. See also Song et al. [50] for an extension to multivariate longitudinal responses, also of mixed type. Pitt et al. [43] develop efficient Bayesian analysis of multivariate regression models using an unstructured correlation matrix for the copula. These models are examples of GCMR. Alternatively, GCMR may be seen as multivariate probit regression with marginals $\mathrm{F}_i(y_i; \boldsymbol{\lambda})$. The multivariate probit model with logistic marginals considered by Le Cessie and Van Houwelingen [33] for longitudinal binary data and the correlated probit model for joint modelling of clustered observations of mixed-type by Gueorguieva and Agresti [20] may also be interpreted as special cases of the model class discussed in this paper. Recent advances on the joint copula analysis of mixed dependent data can be found in a series of papers by A.R. de Leon and colleagues [11, 12, 58].

3.1. Dependence models

The dependence structure in GCMR is modelled through the specification of an appropriate correlation matrix $\mathbf{\Omega}$ of the errors vector $\boldsymbol{\epsilon}$. Although any correlation matrix $\mathbf{\Omega}$ is allowed, we identify here some particular model types that seem likely to have wide application. All the models described below are implemented in our package `gcmr`.

3.1.1. Longitudinal and clustered data

Suppose observations are grouped in m clusters of dimensions n_r, $r = 1, \dots, m$, with $\sum_{r=1}^{m} n_r = n$. This is the case of longitudinal or panel data where m subjects are observed on n_r occasions each. Under the standard assumption of independence between different subjects or groups, appropriate correlation matrices for the errors are block-diagonal,

$$\mathbf{\Omega} = \begin{pmatrix} \mathbf{\Omega}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{\Omega}_m \end{pmatrix},$$

where $\mathbf{\Omega}_r$ is a $n_r \times n_r$ correlation matrix. Similarly to the method of generalized estimating equations [34], we can identify some useful correlation structures for a generic block $\mathbf{\Omega}_r$. Consider indices i and j denoting two observations belonging to the same cluster r, *i.e.* $(n_1 + \cdots + n_{r-1}) + 1 \leq i < j \leq (n_1 + \cdots + n_r)$. Then, possible correlation structures are

1. *exchangeable*, $\text{corr}(\epsilon_i, \epsilon_j) = \tau$ for each choice of indices i and j;
2. *autoregressive* of order one, $\text{corr}(\epsilon_i, \epsilon_j) = \tau^{|i-j|}$;
3. *moving average* of order q, $\text{corr}(\epsilon_i, \epsilon_j) = \tau_{|i-j|}$ for $|i - j| \leq q$;
4. *unstructured*, corresponding to a correlation matrix without any restriction.

Song [47, 48] studies models of this type for the special case of marginals given by generalized linear models and calls them vector generalized linear models. However, Gaussian copulas can be used to join any type of marginals not only those belonging to the exponential family. As an example of this, censored clustered responses are later analysed with a Weibull regression model in Section 9.3.

The work by P. Song focuses on longitudinal data with small size clusters. One of the intents of this paper is to show that Gaussian copula regression models can be also used for the analysis of larger dimensional processes observed in time series and spatial statistics.

3.1.2. Time series

Marginal regression models with stationary time series errors for equi-spaced observations may be specified by assuming that $\mathbf{\Omega}$ is the correlation matrix induced by an autoregressive and moving average process of orders p and q. An illustration regarding serially correlated counts is discussed in Section 9.1.

3.1.3. Spatial data

Regression analysis with spatial and spatio-temporal dependent responses may be modelled by assuming that errors are generated from a stationary Gaussian random field [10]. A flexible choice for $\mathbf{\Omega}$ in the spatial case is the Matérn isotropic correlation function

$$\mathrm{corr}(\epsilon_i, \epsilon_j) = \frac{1}{2^{\tau_2 - 1}\Gamma(\tau_2)} \left(\frac{\|\mathbf{s}_i - \mathbf{s}_j\|_2}{\tau_1} \right)^{\tau_2} \mathrm{B}_{\tau_2} \left(\frac{\|\mathbf{s}_i - \mathbf{s}_j\|_2}{\tau_1} \right), \tag{5}$$

where $\mathbf{s_i}$ are the coordinates of the ith observation and B_{τ_2} is the modified Bessel function of order τ_2. The two parameters τ_1 and τ_2 both need to be strictly positive. For more details on these and other spatial correlation functions see, for example, Cressie [10] and Diggle and Ribeiro [14]. An illustration of spatial regression of counts is provided in Section 9.4.

4. Model properties

4.1. Distributional forms

GCMR models differ from other marginal models in the form of bivariate and higher order dimensional joint distributions. In the continuous case, the mapping (3) between ϵ_i and Y_i is one-to-one, so that marginal distributions of the responses are readily obtained by standard transformation rules from the distribution of the corresponding errors. For example, in the bivariate case we have

$$p_{ij}(y_i, y_j; \boldsymbol{\theta}) = p_i(y_i; \boldsymbol{\lambda}) p_j(y_j; \boldsymbol{\lambda}) q(\epsilon_i, \epsilon_j; \boldsymbol{\theta}), \tag{6}$$

where

$$q(\epsilon_i, \epsilon_j; \boldsymbol{\theta}) = \frac{p(\epsilon_i, \epsilon_j; \boldsymbol{\theta})}{p(\epsilon_i; \boldsymbol{\lambda}) p(\epsilon_j; \boldsymbol{\lambda})}$$

is the density of the bivariate Gaussian copula. Given the model assumptions, $p(\epsilon_i; \boldsymbol{\lambda})$ is a univariate standard normal density, while $p(\epsilon_i, \epsilon_j; \boldsymbol{\theta})$ is a bivariate normal density with zero means, unit variances and correlation given by the element at position (i, j) in matrix $\mathbf{\Omega}$.

In the categorical and discrete cases, mapping (3) is many-to-one. It follows that the joint marginal distributions are expressed by multivariate normal integrals. For example, the bivariate marginal distribution is the two-dimensional integral

$$p_{ij}(y_i, y_j; \boldsymbol{\theta}) = \int_{\mathscr{D}_i(y_i;\boldsymbol{\lambda})} \int_{\mathscr{D}_j(y_j;\boldsymbol{\lambda})} p(\epsilon_i, \epsilon_j; \boldsymbol{\theta}) \, \mathrm{d}\epsilon_i \mathrm{d}\epsilon_j, \tag{7}$$

whose domain is the Cartesian product of intervals

$$\mathscr{D}_i(y_i; \boldsymbol{\lambda}) = \left[\Phi^{-1}\{\mathrm{F}_i(y_i^-; \boldsymbol{\lambda})\}, \Phi^{-1}\{\mathrm{F}_i(y_i; \boldsymbol{\lambda})\} \right],$$

where $\mathrm{F}_i(y_i^-;\boldsymbol{\lambda})$ is the left-hand limit of $\mathrm{F}_i(\cdot;\boldsymbol{\lambda})$ at y_i. If the support of Y_i is contained in $\mathbb{N}$, as for binomial and Poisson marginals, then $y_i^- = y_i - 1$.

4.2. Dependence properties

In the special case of linear regression models with normally distributed errors, the correlation between pairs of responses, given the corresponding covariates, coincides with the correlation of the corresponding normal errors $\mathrm{corr}(\epsilon_i, \epsilon_j)$. Otherwise, the correlation between Y_i and Y_j is a nonlinear function of the correlation of ϵ_i and ϵ_j. This nonlinear function can be computed by a variety of numerical integration methods. Recently, Kugiumtzis and Bora-Senta [32] suggested to approximate these correlations making use of piece-wise linear approximations.

In the copula theory, alternative measures of association based on ranks are often used. The two most popular rank measures are the Kendall τ and the Spearman ρ. A recent illustration of the use of the Kendall τ in longitudinal regression analysis is given by Parzen et al. [42]. Song [48, §6.3.3] supplies a detailed discussion of the relative merits of the various association measures in Gaussian copula models. In particular, it is shown that Spearman ρ is very close to $\mathrm{corr}(\epsilon_i, \epsilon_j)$ and this is positively correlated with the Kendal τ index.

Interpretation of the dependence structured inherited by the Gaussian copula requires some care because the correlation of the responses is attenuated with respect to the correlation of the errors, as shown by Klaassen and Wellner [31]. The attenuation is often considerable, in particular in the noncontinuous case where the margins restrict the range of possible association between the responses. A special case that has received much attention is when the margins identify a probit model and $\mathrm{corr}(\epsilon_i, \epsilon_j)$ corresponds to the *tetrachoric correlation* [21]. However, the restricted range of dependence is a common problem for any multivariate analysis of discrete variables. See also Genest and Nešlehová [17] for a detailed discussion about the correct interpretation of dependence measures in copula models for count data.

Although closed-form computation for bivariate, and higher order, moments is not available, some key aspects of the dependence structure of the model (3)-(4) are easily derived.

Property 1. *If errors ϵ_i and ϵ_j are uncorrelated, then the corresponding pair of responses Y_i and Y_j are independent given the covariates $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.*

This coherency quality is obvious from bivariate distribution expressions in the continuous (6) and in the non-continuous (7) cases. See also Bodnar et al. [4] where this result is discussed together with other useful properties of Gaussian copula models. To appreciate Propriety 1, consider the special case of regression with stationary time series errors. The property states that if errors follow a moving average process of order q, then responses that are more than q units apart are independent.

Property 2. *The direction of the association between any pair of responses Y_i and Y_j given the covariates $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ coincides with the sign of the correlation between the corresponding pair of errors ϵ_i and ϵ_j.*

This is a direct consequence of mapping (3) being non-decreasing. This property ensures that the correlation structure of the normal errors does not modify the direction of the dependence between the responses given the covariates. As illustrated in the examples in Section 9, this simple result is very useful for interpretation of the fitted model.

Property 3. *If the error vector $\boldsymbol{\epsilon}$ follows a Markovian process of order p and all the marginals $p_i(y_i;\boldsymbol{\lambda})$ are continuous, also the response vector $\mathbf{Y}$ given the model matrix $\mathbf{X}$ then follows a Markovian process of order p.*

This property is easily verified in the special case of time-series or longitudinal data. In fact, the conditional density of a continuous response Y_i given its predecessors is

$$p_i(y_i|y_{i-1},\dots,y_1;\boldsymbol{\theta}) = \frac{p_i(y_i;\boldsymbol{\lambda})}{p(\epsilon_i;\boldsymbol{\lambda})} p(\epsilon_i|\epsilon_{i-1},\dots,\epsilon_1;\boldsymbol{\theta}).$$

If the errors are Markovian of order p, the above conditional density reduces to the limited memory conditional density, indeed

$$\begin{aligned} p_i(y_i|y_{i-1},\dots,y_1;\boldsymbol{\theta}) &= \frac{p_i(y_i;\boldsymbol{\lambda})}{p(\epsilon_i;\boldsymbol{\lambda})} p(\epsilon_i|\epsilon_{i-1},\dots,\epsilon_{i-p+1};\boldsymbol{\theta}) \\ &= p_i(y_i|y_{i-1},\dots,y_{i-p+1};\boldsymbol{\theta}) \end{aligned}$$

Similarly, but with notational complications, it can be shown that if the errors $\boldsymbol{\epsilon}$ are realizations from a Gaussian Markovian random field, then the multivariate continuous responses $\mathbf{Y}$ are realizations from a Markovian random field. For time series, if Property 1 states a kind of parallelism between moving average errors and responses, then Property 3 extends the parallelism also to autoregressive processes but only for continuous responses.

5. Maximum likelihood inference

Suggested fitting of GCMR models is through the method of maximum likelihood. A clear advantage of this approach is that standard tools for hypothesis testing and model selection, such as likelihood ratio statistics and information criteria, can be used. Along the following lines, details of likelihood computations are discussed. For this purpose, it is convenient to treat the continuous case and the discrete or categorical case separately. We start from the former because it is simpler and is propaedeutic to the latter.

The one-to-one relationship between responses $\mathbf{Y}$ and errors $\boldsymbol{\epsilon}$ in the continuous case yields the likelihood for $\boldsymbol{\theta}$ as

$$\mathscr{L}(\boldsymbol{\theta};\boldsymbol{y}) = \mathscr{L}_{\mathrm{ind}}(\boldsymbol{\lambda};\boldsymbol{y}) q(\boldsymbol{\epsilon};\boldsymbol{\theta}), \tag{8}$$

where the copula density

$$q(\boldsymbol{\epsilon};\boldsymbol{\theta})=\frac{p(\epsilon_1,\dots,\epsilon_n;\boldsymbol{\theta})}{p(\epsilon_1;\boldsymbol{\lambda})\times\cdots\times p(\epsilon_n;\boldsymbol{\lambda})}$$

can be interpreted as the likelihood ratio between the assumed multivariate normal model for the errors and that under the independence hypothesis. Hence the likelihood $\mathscr{L}(\boldsymbol{\theta};\boldsymbol{y})$ is obtained by sharpening the independence likelihood $\mathscr{L}_{\text{ind}}(\boldsymbol{\lambda};\boldsymbol{y})$ through a measure $q(\boldsymbol{\epsilon};\boldsymbol{\theta})$ of the evidence for dependence among the errors.

More difficult is the case of discrete or categorical-valued responses. In this case, likelihood evaluation requires the computation of the n-dimensional rectangular integral

$$\mathscr{L}(\boldsymbol{\theta};\boldsymbol{y})=\int_{\mathscr{D}_1(y_1;\boldsymbol{\lambda})}\cdots\int_{\mathscr{D}_n(y_n;\boldsymbol{\lambda})}p(\epsilon_1,\dots,\epsilon_n;\boldsymbol{\theta})\,\mathrm{d}\epsilon_1\dots\mathrm{d}\epsilon_n. \qquad (9)$$

It is convenient to re-express the above integral by considering the change of variable from $\mathscr{D}_1(y_1;\boldsymbol{\lambda})\times\cdots\times\mathscr{D}_n(y_n;\boldsymbol{\lambda})$ to $(0,1)^n$ with componentwise inverse

$$\epsilon_i(u_i)=\Phi^{-1}\left\{\mathrm{F}_i(y_i;\boldsymbol{\lambda})-u_i\,p_i(y_i;\boldsymbol{\lambda})\right\},\quad i=1,\dots,n. \qquad (10)$$

Then, likelihood (9) assumes the form

$$\mathscr{L}(\boldsymbol{\theta};\boldsymbol{y})=\mathscr{L}_{\text{ind}}(\boldsymbol{\lambda};\boldsymbol{y})\int_{[0,1]}\cdots\int_{[0,1]}q\{\epsilon_1(u_1),\dots,\epsilon_n(u_n);\boldsymbol{\theta}\}\,\mathrm{d}u_1\dots\mathrm{d}u_n, \qquad (11)$$

whose interpretation is much the same as in the continuous case (8) except for the adjustment term which is an average over likelihood ratios of type $q(\boldsymbol{\epsilon};\boldsymbol{\theta})$ but computed at the randomized errors $\epsilon_i(u_i)$ given by expression (10).

5.1. Likelihood computation

In the non-continuous case the likelihood is expressed in terms of the Gaussian probability integral (9). Whenever this integral factorizes in low-dimensional terms as in longitudinal studies with few observations per subject, we suggest to use precise deterministic approximations as the method of Joe [27] or other recent numerical methods by Miwa et al. [38] and Craig [9]. Joe's and Miwa's algorithms are both publicly available through R packages `mprobit`, authored by Joe, Chou and Zhang, and `mvtnorm` [24], respectively.

For larger dimensions, occurring with time series, spatial data, longer longitudinal studies, the computational cost of deterministic approximations is too large for practical use. Hence, randomized methods need to be considered. An example is the randomized quasi-Monte Carlo method of Genz and Bretz [18] and also included in the above mentioned `R` package `mvtnorm`. This numerical method is of general purpose and may not be efficient for the particular models considered in this paper. For this reason, in the next subsection we describe another Monte Carlo approximation of the likelihood for non-continuous responses, tailored to the GCMR model class.

5.2. Importance sampling

Expression (11) suggests the following simple Monte Carlo approximation of the likelihood:

1. for $k = 1, \dots, \mathrm{K}$,
   (a) generate n independent uniform (0,1) variates, $u_1^{(k)}, \dots, u_n^{(k)}$;
   (b) compute the randomized errors $\epsilon_i^{(k)} = \epsilon(u_i^{(k)})$ from equation (10);
   (c) compute the Gaussian copula density

$$q^{(k)}(\boldsymbol{\epsilon}; \boldsymbol{\theta}) = \frac{p(\epsilon_1^{(k)}, \dots, \epsilon_n^{(k)}; \boldsymbol{\theta})}{p(\epsilon_1^{(k)}; \boldsymbol{\lambda}) \times \cdots \times p(\epsilon_n^{(k)}; \boldsymbol{\lambda})};$$

2. approximate the likelihood by

$$\tilde{\mathscr{L}}(\boldsymbol{\theta}; \boldsymbol{y}) = \mathscr{L}_{\text{ind}}(\boldsymbol{\lambda}; \boldsymbol{y}) \frac{1}{\mathrm{K}} \sum_{k=1}^{\mathrm{K}} q^{(k)}(\boldsymbol{\epsilon}; \boldsymbol{\theta}). \tag{12}$$

Unfortunately, this approximation turns out to be quite inefficient. Indeed, consider the importance sampling approximation of the likelihood,

$$\hat{\mathscr{L}}^{\text{IS}}(\boldsymbol{\theta}; \boldsymbol{y}) = \frac{1}{\mathrm{K}} \sum_{k=1}^{\mathrm{K}} \frac{p(\boldsymbol{y}, \boldsymbol{\epsilon}^{(k)}; \boldsymbol{\theta})}{p^{\text{IS}}(\boldsymbol{\epsilon}^{(k)} | \boldsymbol{y}; \boldsymbol{\theta})}, \tag{13}$$

where $\boldsymbol{\epsilon}^{(k)}$ is a vector of randomized errors drawn from the importance sampling distribution $p^{\text{IS}}(\boldsymbol{\epsilon}|\boldsymbol{y}; \boldsymbol{\theta})$. It follows that approximation (12) corresponds to an importance density constructed under the working assumption that the errors ϵ_i are independent given the observations y_i,

$$p^{\text{IS}}(\boldsymbol{\epsilon}|\boldsymbol{y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{p(\epsilon_i; \boldsymbol{\lambda})}{p_i(y_i; \boldsymbol{\lambda})}.$$

Although valid, this importance density may be very inefficient because of the strong independence assumption. Obviously, the ideal importance density would be the exact conditional density $p(\boldsymbol{\epsilon}|\boldsymbol{y}; \boldsymbol{\theta})$: with this choice, each term in sum (13) is exactly equal to $p(\boldsymbol{y}; \boldsymbol{\theta})$ and thus a single draw ($\mathrm{K} = 1$) would be sufficient to restore the exact likelihood. Unfortunately, using this ideal distribution is unfeasible because the importance sampling weights depend on $p(\boldsymbol{y}; \boldsymbol{\theta})$.

By noticing that a draw from the ideal $p(\boldsymbol{\epsilon}|\boldsymbol{y}; \boldsymbol{\theta})$ could be obtained by sampling sequentially from $p_i(\epsilon_i | y_i, \dots, y_1; \boldsymbol{\theta})$, $i = 1, \dots, n$, we replace this unmanageable importance density with sequential sampling from density

$$p_i(\epsilon_i | y_i, \epsilon_{i-1}, \dots, \epsilon_1; \boldsymbol{\theta}), \quad i = 1, \dots, n. \tag{14}$$

For the special case of multivariate probit regression, the above importance sampling density corresponds to the popular Geweke-Hajivassiliou-Keane simulator (GHK), see for example Keane [30].

The extension of the GHK simulator to deal with GCMR is immediate. Under the model assumptions, $p(\epsilon_i|\epsilon_{i-1},\dots,\epsilon_1;\boldsymbol{\theta})$ is a normal density with mean $m_i = \mathrm{E}(\epsilon_i|\epsilon_{i-1},\dots,\epsilon_1)$ and variance $v_i^2 = \mathrm{var}(\epsilon_i|\epsilon_{i-1},\dots,\epsilon_1)$. Thus, (14) is a truncated normal density over the interval $\mathscr{D}_i(y_i;\boldsymbol{\lambda})$ and a draw from it is obtained by setting

$$\epsilon_i(u_i) = m_i + v_i\Phi^{-1}\left\{(1-u_i)a_i + u_ib_i\right\}, \quad i = 1,\dots,n, \tag{15}$$

where $u_1,\dots,u_n$ are n independent draws from a uniform (0,1) random variable and

$$a_i = \Phi\left[\frac{\Phi^{-1}\{\mathrm{F}_i(y_i^-;\boldsymbol{\lambda})\} - m_i}{v_i}\right],\ b_i = \Phi\left[\frac{\Phi^{-1}\{\mathrm{F}_i(y_i;\boldsymbol{\lambda})\} - m_i}{v_i}\right], \quad i = 1,\dots,n.$$

Hence, the resulting sequential sampler algorithm for approximating the likelihood is

1. for $k = 1,\dots,\mathrm{K}$,
    (a) generate n independent uniform (0,1) variates, $u_1^{(k)},\dots,u_n^{(k)}$;
    (b) compute the randomized errors $\epsilon_i^{(k)} = \epsilon(u_i^{(k)})$ from the (15);
2. approximate the likelihood by

$$\hat{\mathscr{L}}^{\mathrm{IS}}(\boldsymbol{\theta};\boldsymbol{y}) = \frac{1}{\mathrm{K}}\sum_{k=1}^{\mathrm{K}}\left\{\prod_{i=1}^{n}\frac{p(\epsilon_i^{(k)}|\epsilon_{i-1}^{(k)},\dots,\epsilon_1^{(k)};\boldsymbol{\theta})}{p_i(\epsilon_i^{(k)}|y_i,\epsilon_{i-1}^{(k)},\dots,\epsilon_1^{(k)};\boldsymbol{\theta})}\right\}.$$

A few comments on numerical aspects are in order. Quantities m_i and v_i^2 can be efficiently computed by the Cholesky factorization of $\boldsymbol{\Omega}$ which, in any case, is an integral component of the likelihood computation. Substantial computational saving is achieved by exploiting the fact that the error correlation matrix $\boldsymbol{\Omega}$ is the same for all the simulated error vectors $\boldsymbol{\epsilon}^{(k)}$, $k = 1,\dots\mathrm{K}$.

Other importance densities could be considered, for example following the ideas of Durbin and Koopman [16]. However, the increase in computational cost may not be justified in terms of numerical precision given the simplicity and the good results obtained using the suggested sampler. See the simulation study in Section 6.

In general, the total complexity for one likelihood approximation with the discussed importance sampling, as well as with any other likelihood approximation for GCMR models, is of order $\mathcal{O}(n^3)$, because of the necessary inversion of the correlation matrix $\boldsymbol{\Omega}$. However, for specific problems the computational cost is much lower. For example, in time series models with $\boldsymbol{\epsilon}$ following an autoregressive moving average process the Cholesky factorization can be efficiently implemented via the Kalman filter and only $\mathcal{O}(n)$ computations are needed. If n is large, the general computational cost $\mathcal{O}(n^3)$ is impractical. Some possible remedies are discussed in the final section within the directions for future research.

Maximum likelihood estimates are better computed from the log-likelihood. Importance sampling is however designed for giving unbiased estimates of the likelihood, while $\log \hat{\mathscr{L}}^{\mathrm{IS}}(\boldsymbol{\theta}; \boldsymbol{y})$ is a slightly biased estimator of the log-likelihood $\ell(\boldsymbol{\theta}; \boldsymbol{y})$. In our package gcmr, an approximately unbiased estimator of the log-likelihood is obtained with the correction given by Durbin and Koopman [16, §2.3].

The Monte Carlo size needed for correct inferential conclusions is typically a function of the degree of discreteness, the most difficult case being binary responses. As a general suggestion, it is advisable to repeat the analysis to check the adequacy of the Monte Carlo size, for example starting with a small Monte Carlo size and then increasing it until differences in the parameter estimates become practically insignificant. See Section 9.5 for some guidelines about the choice of the Monte Carlo size.

6. Simulations under model conditions

We have carried out several simulation studies to investigate the reliability of the importance sampling approximation. Here we only show the results for a time series log-linear regression model with a continuous time-varying covariate and two seasonal terms. Denote by $\mathrm{F}_i(y_i; \lambda)$ the cumulative distribution function of a Poisson random variable of mean $\mu_i = \mathrm{E}(\mathrm{Y}_i|\mathbf{x_i})$. Data are generated from the marginal regression model

$$\begin{aligned} \mathrm{Y}_i &= \mathrm{F}_i^{-1}\left\{\Phi(\epsilon_i); \lambda\right\}, \\ \log(\mu_i) &= \beta_0 + \beta_1 \sin\left(\frac{2\pi i}{12}\right) + \beta_2 \cos\left(\frac{2\pi i}{12}\right) + \beta_3 x_i, \\ x_i &= 0.6 x_{i-1} - 0.4 x_{i-2} + \zeta_i, \quad \zeta_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1), \end{aligned} \tag{16}$$

with the correlation matrix of the errors $\boldsymbol{\Omega}$ corresponding to that of an autoregressive process of order one with first-order autocorrelation equals to 0.8.

For values of the marginal parameters $\beta_0 = 1$, $\beta_1 = \beta_2 = \beta_3 = 0.2$, we generated 300 time series of different lengths. Then, for each simulated series, five different estimates of the parameters were independently computed. In all cases, the log-likelihood was approximated by using $\mathrm{K} = 300$ sequences of uniform pseudo-random numbers $u_1^{(k)}, \dots, u_n^{(k)}$, $k = 1, \dots, \mathrm{K}$. Columns 2 and 5 of Table 1 display the averages of the $300 \times 5 = 1,500$ estimates for time series of lengths $n = 100$ and $n = 500$, respectively. Clearly, the averages of the estimates are close to the true values. It is more interesting to evaluate the impact of the Monte Carlo errors. For this purpose, we compute the standard ANOVA decomposition of the total sum of squares into between and within parts, using the 300 time series as grouping factor. Columns 3-4 and 6-7 of Table 1 show the between and within sums of squares measuring the variability due to the true estimation errors and to Monte Carlo errors, respectively. With the chosen Monte Carlo sample size, the variability due to the simulation is much smaller than that due to estimation errors. Similar results have been obtained for other parameter values, other dependence structures and other discrete marginal distributions.

Table 1

Averages of 1,500 estimates of the parameter of model (16) with AR(1) errors and time series of lengths $n \in \{100, 500\}$ using the importance sampling approximation with a Monte Carlo sample size of $K = 300$. Also reported are the sums of squares pertaining to estimation errors (SS_{est}) and to Monte Carlo errors (SS_{MC})

		$n = 100$			$n = 500$		
	model	average	$\mathrm{SS}_{\mathrm{est}}$	$\mathrm{SS}_{\mathrm{MC}}$	average	$\mathrm{SS}_{\mathrm{est}}$	$\mathrm{SS}_{\mathrm{MC}}$
β_0	1.0	0.955	52.277	0.458	0.995	8.553	0.238
β_1	0.2	0.211	18.102	0.144	0.195	3.065	0.081
β_2	0.2	0.206	15.672	0.101	0.203	3.377	0.074
β_3	0.2	0.204	2.090	0.011	0.199	0.435	0.007
AR(1)	0.8	0.806	2.054	0.046	0.797	0.365	0.013

7. Model checking

Validation of the model assumptions is a crucial aspect in any regression analysis, particularly in our multivariate setting. Model checking by residual analysis and by a Hausman [22] type specification test is discussed below.

7.1. Residuals

In the continuous case, the Rosenblatt's transformations [45]

$$\mathrm{M}_i = \mathrm{F}_i(\mathrm{Y}_i|y_{i-1}, \dots, y_1; \boldsymbol{\theta})$$

are uniformly and independently distributed in the unit interval. Hence, model adequacy can be checked through residuals

$$r_i = \Phi^{-1}\{\mathrm{F}_i(y_i|y_{i-1}, \dots, y_1; \hat{\boldsymbol{\theta}})\},$$

which are realizations of n uncorrelated standard normal variables if the model is correctly specified. These type of Cox and Snell residuals [8] are termed (normalized) quantile residuals in Dunn and Smyth [15] and normal pseudo-residuals in Zucchini and MacDonald [63]. Normality of the quantile residuals can be inspected by normal probability plots and tests. Assessment of lack of correlation can involve graphical tools as autocorrelation plots for time series and longitudinal studies and variograms for spatial data.

In the non-continuous case, we define residuals r_i to be any arbitrary value belonging to the interval $[\Phi^{-1}(m_i^-), \Phi^{-1}(m_i)]$, where the lower extreme is defined as $m_i^- = \mathrm{F}_i(y_i^-|y_{i-1}, \dots, y_1; \hat{\boldsymbol{\theta}})$, the left-hand limit of m_i at y_i. Accordingly, we base model diagnostic on the so-called randomized quantile residuals $r_i^{\mathrm{rnd}}(u_i) = \Phi^{-1}\{m_i^- + u_i(m_i - m_i^-)\}$ where u_i is a draw from a $(0, 1)$ uniform variate [15]. Under model conditions, $r_i^{\mathrm{rnd}}(u_i)$ are realizations of uncorrelated standard normal variables and thus they can be used as ordinary residuals for checking model assumptions. Since residuals $r_i^{\mathrm{rnd}}(u_i)$ depend on the uniform variates u_i, it is appropriate to inspect several sets of residuals before taking a decision about the model.

Alternatively, it is possible to avoid randomization by considering the mid interval quantile residuals $r_i^{\mathrm{mid}} = \Phi^{-1}\{(m_i^- + m_i)/2\}$, as suggested by Zucchini and MacDonald [63] in the context of hidden Markov models. These residuals are, however, neither normally distributed nor uncorrelated. Zucchini and MacDonald [63] also suggest diagnostics based on the interval quantile residuals $r_i^{\mathrm{int}} = [\Phi^{-1}(m_i^-), \Phi^{-1}(m_i)]$ in order to preserve the data discreteness. However, interval quantile residuals are problematic when observations are the minimal or the maximal possible values, since in these cases r_i^{int} are left or right open (infinite) intervals.

7.2. Hausman-type specification test

In most cases there is scientific interest in the marginal parameters $\boldsymbol{\lambda}$ or a subset of regression parameters $\boldsymbol{\beta}$. The assumed marginals can be checked through a variety of well-developed graphical and numeric methods, thus correct specification of the independence likelihood is generally accomplished. This leads to the dilemma of basing inference on $\boldsymbol{\lambda}$ on the safe but inefficient independence likelihood or considering the complete model but with the risk of copula misspecification. In other terms, we are interested in assessing the null hypothesis

$$\mathrm{H}_0 : \text{the assumed multivariate model is correctly specified,}$$

against the alternative

$$\mathrm{H}_1 : \text{marginals are correct but the assumed multivariate normal distribution}$$
$$\text{for the errors is } \textit{not} \text{ (wrong copula).}$$

The independence likelihood estimator $\hat{\boldsymbol{\lambda}}_{\mathrm{ind}}$ is consistent under both null and alternative hypotheses, while the maximum likelihood estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\lambda}}^{\mathrm{T}}, \hat{\boldsymbol{\tau}}^{\mathrm{T}})^{\mathrm{T}}$ is consistent and efficient under the null hypothesis but inconsistent under the alternative. This suggests validating the correct model specification by the Hausman [22]-type statistic

$$\mathrm{h}(\mathbf{Y}) = (\hat{\boldsymbol{\lambda}}_{\mathrm{ind}} - \hat{\boldsymbol{\lambda}})^{\mathrm{T}} \mathbf{D}^{\mathbf{-1}} (\hat{\boldsymbol{\lambda}}_{\mathrm{ind}} - \hat{\boldsymbol{\lambda}}),$$

with the variance $\mathbf{D} = \mathrm{var}(\hat{\boldsymbol{\lambda}}_{\mathrm{ind}} - \hat{\boldsymbol{\lambda}})$ computed under the null hypothesis, where $\mathrm{h}(\mathbf{Y})$ is distributed as a chi-squared random variable with $\dim(\boldsymbol{\lambda})$ degrees of freedom.

Remark. The above framework differs from the usual Hausman test, which is focused on the complete parameter $\boldsymbol{\theta}$ and not on its subset $\boldsymbol{\lambda}$. Consequently, the Hausman orthogonality result does not apply, that is $\mathrm{cov}(\hat{\boldsymbol{\lambda}}_{\mathrm{ind}} - \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\lambda}}) \neq \mathbf{0}$. Hence, $\mathbf{D}$ does not simplify into the difference of variances $\mathrm{var}(\hat{\boldsymbol{\lambda}}_{\mathrm{ind}}) - \mathrm{var}(\hat{\boldsymbol{\lambda}})$.

Under the null hypothesis, the vector of estimates $(\hat{\boldsymbol{\lambda}}_{\mathrm{ind}}^{\mathrm{T}}, \hat{\boldsymbol{\theta}}^{\mathrm{T}})^{\mathrm{T}}$ has asymptotic variance matrix

$$\mathbf{V} = \begin{pmatrix} \mathbf{H}_1^{-1}\mathbf{J}_1\mathbf{H}_1^{-1} & \mathbf{H}_1^{-1}\mathbf{J}_{12}\mathbf{H}_2^{-1} \\ \mathbf{H}_2^{-1}\mathbf{J}_{21}\mathbf{H}_1^{-1} & \mathbf{H}_2^{-1} \end{pmatrix},$$

where $\mathbf{H}_1 = \mathrm{E}\{-\nabla^2 \ell_{\mathrm{ind}}(\boldsymbol{\lambda}; \mathbf{Y})\}$, $\mathbf{J}_1 = \mathrm{var}\{\nabla \ell_{\mathrm{ind}}(\boldsymbol{\lambda}; \mathbf{Y})\}$, $\mathbf{H}_2 = \mathrm{E}\{-\nabla^2 \ell(\boldsymbol{\theta}; \mathbf{Y})\}$, $\mathbf{J}_{12} = \mathrm{cov}\{\nabla \ell_{\mathrm{ind}}(\boldsymbol{\lambda}; \mathbf{Y}), \nabla \ell(\boldsymbol{\theta}; \mathbf{Y})\}$ and $\mathbf{J}_{21} = \mathbf{J}_{12}^{\mathrm{T}}$. Hence, $\mathbf{D} = \mathbf{C}^{\mathrm{T}}\mathbf{V}\mathbf{C}$ for a contrast matrix $\mathbf{C} = (\mathbf{I}^{\mathrm{T}}, -\mathbf{I}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})$, where the identity blocks $\mathbf{I}$ have the dimension of $\boldsymbol{\lambda}$. Generally, the components of matrix $\mathbf{V}$ are unavailable in closed-form. They can be estimated by Monte Carlo simulation from the assumed model. Alternatively, one can consider a more accurate, but computationally costly, direct estimation of $\mathbf{D}$ via parametric bootstrap. In the examples discussed in Section 9, the test statistic is computed via parametric bootstrap.

The Hausman-type test can be used for checking the correctness of inference about the global marginal parameter $\boldsymbol{\lambda}$ as well as for checking subsets of $\boldsymbol{\lambda}$, such as single regressors β_i, or combinations of regressors, in conformity with the scientific focus.

8. Robustness to the misspecification of the error distribution

Despite the residual analysis described in the previous section, a thorough validation of the multivariate normal assumption for the errors $\boldsymbol{\epsilon}$ remains a difficult task. Hence, it is natural to ask whether correct inferences for $\boldsymbol{\lambda}$ can be obtained under *local* misspecification of the multivariate distribution for the errors. A general answer to this question is also difficult, however simulation studies such those reported below give a preliminary positive indication for robustness against local misspecification of the error distribution.

We illustrate two different simulation studies. In the first study, we investigate the robustness to the misspecification of the ARMA correlation of the errors and to the presence of heavy tails in their marginal distribution. In the second study, we consider the effect of non-linearity in the interdependence among the errors and skewness in their marginal distribution.

8.1. Effect of heavy tails and unspecified moving average correlation

We consider the same simulation setting of Section 6 with marginal structure as in formulas (16), but with a different error model, namely:

1. the errors follow an ARMA(1,1) normal model;
2. the errors follow an AR(1) model but are distributed as Student t random variables with ν degrees of freedom;
3. the errors follow an ARMA(1,1) model and are distributed as Student t random variables with ν degrees of freedom.

In each case, estimates are obtained from the misspecified model with normal AR(1). Table 2 displays $1,000$ simulated estimates for the covariate parameter β_3 for (i) various choices of the degrees of freedom ν, (ii) with or without a moving average component and (iii) different sample sizes. We choose to show results only for the time-varying covariate parameter for space limitations: results for the other marginal parameters are similar. Table 2 reports (i) the averages of the estimates, (ii) their standard deviations, (iii) the averages of standard errors

Table 2

Averages, standard deviations (sim. s.e.), averages of estimated standard errors (est. s.e.) and 95% *confidence intervals coverages for* 1,000 *estimates of parameter* β_3 *in model (16) with errors that follow an ARMA(1,1) model and are distributed as Student t random variables with* ν *degrees of freedom. The estimates are computed either with a misspecified dependence model with normal AR(1) errors (*`gcmr`*) or under the working assumption of independence with robust standard errors (ind.). The table displays results for sample sizes* $n \in \{50, 100, 300\}$, *moving average parameter MA* $\in \{0, 0.2\}$ *and degrees of freedom* $\nu \in \{1, 10, +\infty\}$

	MA	0	0.2	0	0.2	0	0.2
	ν	$+\infty$	$+\infty$	10	10	1	1
				$n = 50$			
`gcmr`	average	0.202	0.201	0.203	0.202	0.203	0.203
	sim. s.e.	0.052	0.048	0.055	0.051	0.064	0.063
	est. s.e.	0.053	0.049	0.055	0.051	0.067	0.064
	coverage	0.949	0.954	0.956	0.958	0.964	0.964
ind.	average	0.205	0.204	0.204	0.205	0.204	0.204
	sim. s.e	0.081	0.082	0.081	0.081	0.079	0.081
	est. s.e.	0.065	0.064	0.067	0.066	0.070	0.069
	coverage	0.863	0.847	0.867	0.859	0.896	0.887
				$n = 100$			
`gcmr`	average	0.203	0.203	0.204	0.203	0.203	0.203
	sim. s.e.	0.038	0.036	0.039	0.037	0.046	0.046
	est. s.e.	0.036	0.034	0.038	0.036	0.045	0.043
	coverage	0.941	0.950	0.945	0.945	0.957	0.944
ind.	average	0.204	0.204	0.204	0.205	0.203	0.203
	sim. s.e.	0.059	0.060	0.059	0.060	0.058	0.059
	est. s.e.	0.050	0.050	0.050	0.050	0.051	0.051
	coverage	0.899	0.888	0.898	0.897	0.920	0.915
				$n = 300$			
`gcmr`	average	0.204	0.203	0.206	0.205	0.206	0.206
	sim. s.e.	0.023	0.022	0.023	0.020	0.027	0.025
	est. s.e.	0.021	0.020	0.022	0.021	0.026	0.025
	coverage	0.930	0.910	0.940	0.970	0.940	0.940
ind.	average	0.202	0.202	0.203	0.202	0.204	0.203
	sim. s.e.	0.034	0.034	0.033	0.033	0.033	0.032
	est. s.e.	0.031	0.031	0.031	0.031	0.032	0.032
	coverage	0.920	0.900	0.950	0.950	0.940	0.940

computed from the Fisher observed information assuming the model is correct and (iv) the empirical coverages of 95% confidence intervals. For comparison, we also report the corresponding estimates using the independence likelihood with standard errors computed from the sandwich heteroskedasticity and autocorrelation consistent estimator for time-series data (Zeileis, 2006) to account for the serial dependence.

The first column of Table 2 corresponds to correct specification of the error distribution. Subsequent columns describe increasing levels of misspecification, the extreme case being errors that follow the ARMA(1,1) process with autoregressive parameter 0.8 and moving average parameter 0.2 and are distributed as Cauchy random variables ($\nu = 1$).

We start by describing the results of the first column, where the two estimation methods are both correctly specified. As expected, maximum independence likelihood estimates show loss of efficiency with respect to estimates based on the fully specified dependence model. It is however more interesting that standard errors of the maximum independence likelihood estimates are strongly biased with $n = 50$ and 100, leading to over-optimistic confidence intervals. This result illustrates the well-known very slow convergence of robust variance estimation using sandwich-type methods; see Kauermann and Carroll (2001) for a general discussion on this point.

The other five columns of Table 2 show what happens if the dependence model is misspecified. In this case, the maximum independence likelihood estimate $\hat{\boldsymbol{\lambda}}_{\text{ind}}$ remains correct, while the estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\lambda}}^{\mathrm{T}}, \hat{\boldsymbol{\tau}}^{\mathrm{T}})^{\mathrm{T}}$ derived from the incorrect dependence model is inconsistent for the complete parameter $\boldsymbol{\theta}$. However, our simulations show that the component $\hat{\boldsymbol{\lambda}}$ of $\hat{\boldsymbol{\theta}}$ relative to the marginal parameters remains essentially correct under local misspecification of the errors. Furthermore, standard errors wrongly computed by inverting the observed Fisher information matrix do not exhibit a practically relevant bias. In contrast, robust sandwich standard errors for the maximum independence likelihood grossly underestimate the uncertainty if the sample size is not large enough.

8.2. Effect of skewness and non-linear dependence among the errors

In the second simulation study, we again consider the same simulation setting of Section 6 with marginal structure as in formulas (16). In this case, the errors are generated from the following scaled threshold autoregressive model of order one TAR(1) [52]

$$\epsilon_i = \tau|\epsilon_{i-1}| + \sqrt{1-\tau^2}\,\eta_i, \quad \eta_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1). \tag{17}$$

This model yields two forms of misspecification with respect to the assumed scaled AR(1) model for the errors

$$\epsilon_i = \tau\epsilon_{i-1} + \sqrt{1-\tau^2}\,\eta_i \quad \eta_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1). \tag{18}$$

First, there is a non-linear dependence among the errors because of the absolute value of the past error in the recursive equation (17). Secondly, the marginal distribution of the errors is not standard normal anymore but it is the skew-normal distribution [3] with location parameter equal to zero, unit scale parameter and skewness parameter $\tau/\sqrt{1-\tau^2}$ [1],

$$\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{SN}\left(0, 1, \frac{\tau}{\sqrt{1-\tau^2}}\right).$$

Hence, as the autoregressive parameter τ raises in absolute value, both the non-linear dependence and the skewness of the errors increase together. Figure 1 illustrates the increasing difference between the stationary distribution of the scaled AR(1) process and the scaled TAR(1) process for values of $\tau \in \{0.3, 0.6, 0.9\}$.

Table 3 displays $1,000$ simulated estimates of the covariate coefficient β_3 for various values of the autoregressive parameter τ and different sample sizes. As

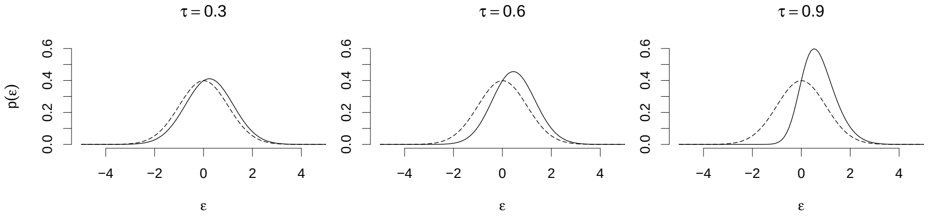

Fig 1. *Comparison between the marginal distribution of TAR(1) errors (solid line) and the marginal distribution of AR(1) errors (dotted line). The three plots correspond to increasing values of the autoregressive parameter* $\tau \in \{0.3, 0.6, 0.9\}$.

Table 3

Averages, standard deviations (sim. s.e.), averages of estimated standard errors (est. s.e.) and 95% *confidence intervals coverages for* 1,000 *estimates of parameter* β_3 *in model (16) with errors that follow the TAR(1) model (17). The estimates are computed either with a misspecified dependence model with normal AR(1) errors (*`gcmr`*) or under the working assumption of independence with robust standard errors (ind.). The table displays results for sample sizes* $n \in \{50, 100, 300\}$ *and autoregressive parameter* $AR \in \{0.3, 0.6, 0.9\}$

	AR	0.3	0.6	0.9
		$n = 50$		
`gcmr`	average	0.197	0.197	0.205
	sim. s.e.	0.079	0.079	0.062
	est. s.e.	0.073	0.076	0.059
	coverage	0.927	0.938	0.953
ind.	average	0.197	0.197	0.208
	sim s.e.	0.079	0.084	0.084
	est s.e.	0.070	0.070	0.065
	coverage	0.899	0.881	0.844
		$n = 100$		
`gcmr`	average	0.197	0.197	0.200
	sim s.e.	0.054	0.052	0.040
	est s.e.	0.051	0.053	0.040
	coverage	0.944	0.962	0.952
ind.	average	0.197	0.197	0.200
	sim s.e.	0.053	0.055	0.056
	est s.e.	0.050	0.051	0.050
	coverage	0.932	0.929	0.911
		$n = 300$		
`gcmr`	average	0.200	0.199	0.199
	sim s.e.	0.028	0.028	0.022
	est s.e.	0.029	0.030	0.023
	coverage	0.960	0.960	0.954
ind.	average	0.200	0.199	0.200
	sim s.e.	0.028	0.030	0.032
	est s.e.	0.029	0.030	0.031
	coverage	0.959	0.947	0.935

for the previous simulation study, standard errors for the maximum likelihood independence estimator are based on the sandwich heteroskedasticity and autocorrelation consistent estimator, while those of the misspecified Gaussian copula model are obtained from the Fisher observed information as if the model were instead correct.

Once more, the maximum likelihood estimates of the marginal parameters $\hat{\boldsymbol{\lambda}}$ based on the Gaussian copula model do not show significant bias despite the misspecification of the errors. For values of τ equal to 0.3 and 0.6, the misspecified maximum likelihood estimator $\hat{\boldsymbol{\lambda}}$ has essentially the same efficiency as the maximum independence likelihood estimator $\hat{\boldsymbol{\lambda}}_{\text{ind}}$. However, the standard errors of the latter estimator tend to underestimate the uncertainty, especially for moderately small sample sizes. Accordingly, the empirical coverage of the confidence intervals is sensibly smaller than the nominal value. By comparison, the coverage of the confidence intervals based on the misspecified Gaussian copula model are much closer to the nominal values. For values of τ that approach the limit of one, the maximum likelihood estimates based on the Gaussian copula model are remarkably more efficient than the maximum independence likelihood estimates.

In synthesis, the two simulation studies show how inferences from an incorrect dependence model can be more precise than inferences obtained from a correct model that avoids specification of the dependence structure. This form of robustness is a consequence of the ability of the nuisance parameters $\boldsymbol{\tau}$ to accommodate departures in the dependence structure in such a way as to keep inferences on marginal parameters $\boldsymbol{\lambda}$ broadly correct. Similar findings were obtained for other models.

9. Examples

In this section we illustrate the model flexibility through a variety of examples covering applications to time series, crossed design experiments, survival analysis and spatial data. All examples are based on well known data sets, in order best to facilitate comparisons with other possible models and fitting methods. Some comments about computational aspects are provided in Section 9.5.

9.1. Generalized linear models with time-series errors

The time series of monthly Polio incidences in the USA from 1970 to 1983 has been analyzed by several authors with different observation- and parameter-driven models since Zeger [59]. Among many others, some useful references may be found in Song [48, §12]. The scientific question is whether or not there is evidence of a decreasing trend of Polio infections in the observation period.

Following previous analyses of these data, we consider a log-linear model with covariates designed to capture possible linear trend and seasonality effects,

$$\log\{\mu_i\} = \beta_0 + \beta_1 t_i + \beta_2 \cos\left(\frac{2\pi t_i}{12}\right) + \beta_3 \sin\left(\frac{2\pi t_i}{12}\right) + \\ + \beta_4 \cos\left(\frac{2\pi t_i}{6}\right) + \beta_5 \sin\left(\frac{2\pi t_i}{6}\right),$$

where t_i is the time of the ith observation. Marginals are modelled through a negative binomial distribution with mean $\mathrm{E}(\mathrm{Y}_i|\boldsymbol{x}_i) = \mu_i$ and variance $\mathrm{var}(\mathrm{Y}_i|\boldsymbol{x}_i) =$

Table 4

Polio data. Estimates (est.) and standard errors (s.e.) for the marginal parameters from (a) the independence model and (b) from the dependence model with ARMA(2,1) errors. Also displayed are estimates, standard errors and z values for the differences of the estimates

	independence		ARMA(2,1)		difference		
	est.	s.e.	est.	s.e.	est.	s.e.	z value
β_0	0.21	0.10	0.21	0.12	0.00	0.01	−0.05
β_1 $(\times 10^3)$	−4.33	1.84	−4.31	2.30	−0.02	0.75	−0.03
β_2	−0.14	0.13	−0.12	0.15	−0.02	0.02	−0.81
β_3	−0.50	0.14	−0.50	0.16	0.00	0.02	−0.19
β_4	0.17	0.13	0.19	0.13	−0.02	0.02	−1.37
β_5	−0.42	0.13	−0.40	0.13	−0.02	0.02	−0.92
κ	0.57	0.16	0.57	0.17	0.00	0.03	−0.13

$\mu_i + \kappa\mu_i^2$. Serial correlation between polio counts is accommodated by assuming an ARMA(p,q) model for the errors.

According to the corrected Akaike information criterion [25], the best fit is provided by GCMR with ARMA(2,1) errors. The first and second order autoregressive parameters are estimated as (standard errors in brackets) −0.53(0.21) and 0.31(0.09), while the moving average parameter as 0.71(0.22). The first two autocorrelations coefficients induced by the fitted ARMA(2,1) model are $\widehat{\text{corr}}(\epsilon_i, \epsilon_{i-1}) = 0.15$ and $\widehat{\text{corr}}(\epsilon_i, \epsilon_{i-2}) = 0.24$. By Property 2, it follows that, conditionally to the covariates, there is positive association between Y_i and Y_{i-1} and between Y_i and Y_{i-2}. Table 4 summarizes estimates for the marginal parameters from the working independence model and from the ARMA(2,1) dependence model.

Standard errors of the independence likelihood estimates are computed using the heteroskedasticity and autocorrelation consistent (HAC) sandwich estimator for time series of Andrews [2]. The estimates of the overdispersion parameter κ are significantly different from zero and thus negative binomial marginals are preferable to Poisson marginals. The point estimate of the trend parameter, that is the parameter of interest, is −4.33 under working independence. This value is very close to the estimate −4.31 obtained via the GCMR model with ARMA(2,1) errors. However, the latter model provides a standard error (s.e. 2.30) substantially larger than that derived under working independence assumptions using the HAC sandwich (s.e. 1.84). Correspondingly, the Wald test statistic for the one-sided hypothesis of negative trend ($\mathrm{H}_0 : \beta_1 = 0$ vs. $\mathrm{H}_1 : \beta_1 < 0$) has a p-value of 0.03 with ARMA(2,1) errors and a p-value of 0.009 under the working independence assumption. In both the cases, the conclusion is in favor of a positive trend, although at a very different level of confidence.

As regard model checking, the Hausman-type test is passed overall ($h = 3.49$, p-value 0.83) and also for all single marginal parameters as shown in last three columns of Table 4. Randomized quantile residuals computed several times sustain the distributional assumptions and suggest that no residual serial correlation is present in the data. Normal probability and autocorrelation plots for a set of residuals are displayed in Figure 2.

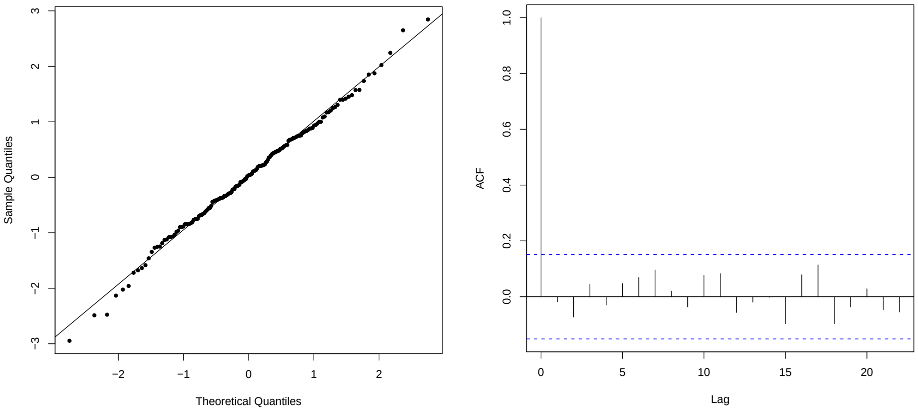

Fig 2. *Polio data. Normal probability (left panel) and autocorrelation (right panel) plots for a set of randomized residuals.*

9.2. Cross-correlated binary data

The Salamander mating data come from a study on barriers to interbreeding in two geographically isolated species of salamanders called Whiteside (W) and Rough-Butt (R). These much studied data have a rather complicated incomplete, balanced, crossed design. The data consist of three experiments, one conducted in Summer 1986 and two in the Fall of the same year. A total of 80 animals were used: the same 40 salamanders in Summer 1986 and in the first Fall experiment. A new set of 40 salamanders was used in the second Fall experiment. Each experiment involved 10 females and 10 males for each of the two populations. Each female was paired six times: three times with males of her own population and three times with males of the other population. The binary response records whether the mating event was successful or not. See McCullagh and Nelder [37] for more details.

These data have been analysed in a number of papers with crossed-random effects logistic or probit models. The random effects are designed to account for the correlation between the results of two matings involving the same female or the same male. The likelihood for this mixed-effects model is very difficult to manage because of the high-dimensional integrals involved. Typical solutions are based on simulation methods. Examples include Markov chain Monte Carlo [60], Monte Carlo expectation-maximization [5] and importance sampling [51] algorithms.

Here we consider a marginal regression analysis of the Salamander data. The proposed model is the marginal counterpart of model "B" of Zeger and Karim [60]. For the marginal part we assume the logistic model

$$\operatorname{logit}\{\mu_i\} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \beta_4 x_{4i},\ i = 1, \ldots, 360, \tag{19}$$

with x_{1i} indicating whether the female used in the ith mating belongs to the R race, x_{2i} whether the male belongs to the R race and x_{4i} whether the experiment took place in Fall ($x_{4i} = 1$) or in summer ($x_{4i} = 0$). The correlation matrix $\mathbf{\Omega}$ of the errors is partitioned in blocks of dimension 120×120,

$$\mathbf{\Omega} = \begin{pmatrix} \mathbf{\Omega}_{\mathrm{S}} & \mathbf{\Omega}_{\mathrm{S,F}} & \mathbf{0} \\ \mathbf{\Omega}_{\mathrm{S,F}} & \mathbf{\Omega}_{\mathrm{F}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{\Omega}_{\mathrm{F}} \end{pmatrix}.$$

Block $\mathbf{\Omega}_{\mathrm{S}}$ describes the cross-correlation structure in the Summer experiment. Apart from the diagonal, the correlation matrix $\mathbf{\Omega}_{\mathrm{S}}$ has zero entries everywhere but in the cells corresponding to two different matings done in summer involving the same female and/or the same male. Thus, there are three types of non-zero entries: the correlation between errors due to the use of the same female in two distinct experiments in Summer is measured by parameter τ_1, the correlation between errors due to the use of the same male by parameter τ_2, finally the sum $\tau_1 + \tau_2$ measures the correlation between errors when both the same female and male are used in Summer.

Similarly, block $\mathbf{\Omega}_{\mathrm{F}}$ accounts for cross-correlation between errors in the two Fall experiments and it is parameterized by τ_3 and τ_4, the first parameter related to female cross-correlation, while the second one to male cross-correlation. Finally, block $\mathbf{\Omega}_{\mathrm{S,F}}$ accounts for the association between the Summer and Fall experiments sharing the same animals and it is parameterized by parameters τ_5 and τ_6 for females and males, respectively.

Regressor coefficients are estimated as $\hat{\beta}_0 = 0.98$ (0.35), $\hat{\beta}_1 = -1.96$ (0.37), $\hat{\beta}_2 = -0.46$ (0.35), and $\hat{\beta}_3 = 2.51$ (0.40). The season dummy coefficient is not significant, $\hat{\beta}_4 = -0.39$ (0.30). The estimated correlation parameters show appreciable cross-correlation between the errors. In particular, the estimated parameters are $\hat{\tau}_1 = 0.29$ (0.14), $\hat{\tau}_2 = 0.07$ (0.09), $\hat{\tau}_3 = 0.18$ (0.07), $\hat{\tau}_4 = 0.31$ (0.09), $\hat{\tau}_5 = -0.04$ (0.09), and $\hat{\tau}_6 = 0.23$ (0.08). The Hausman test is largely passed both for all single regressors and overall (with a p-value of 0.96).

The main scientific focus is however on the marginal probabilities of matings. Denote by π_{RW} the probability that an R female successfully mates with a W male, and similarly denote by π_{RR}, π_{WR} and π_{WW} the remaining probabilities. The estimates of these quantities are easily obtained from the estimated $\boldsymbol{\beta}$ coefficients. For example the maximum likelihood estimate of π_{WR} is $\exp(\hat{\beta}_0 + \hat{\beta}_1)/\{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1)\} = 0.27$. All the other estimates are listed in Table 5 along with 95% confidence intervals. For comparison, we also report estimates and confidence intervals obtained by Zeger and Karim [60] with a random effects logistic model fitted by Gibbs sampling. Results from the two models are very similar, although the marginal-model confidence intervals are narrower in most cases. The clear conclusion is that the probability of a successful mating is much smaller when a W female is matched with an R male, than in all of the other three cases.

9.3. Matched time-to-event data

For an illustration of survival data analysis, we reanalyse the data of Mantel et al. [36] about times to tumor appearance in litter-matched rats. Three female

Table 5

Salamander data. Estimated mating probabilities and their 95% confidence intervals from the marginal model compared with the estimates provided by Zeger and Karim [60]

	Summer		Fall	
	est.	95%CI	est.	95%CI
π_{RR}	0.73	(0.60, 0.83)	0.64	(0.52, 0.74)
π_{RW}	0.63	(0.49, 0.75)	0.53	(0.41, 0.64)
π_{WR}	0.27	(0.17, 0.41)	0.20	(0.12, 0.31)
π_{WW}	0.74	(0.62, 0.84)	0.66	(0.54, 0.76)
	Zeger and Karim [60]			
	Summer		Fall	
	est.	95%CI	est.	95%CI
π_{RR}	0.73	(0.58, 0.84)	0.64	(0.51, 0.76)
π_{RW}	0.62	(0.46, 0.76)	0.52	(0.39, 0.65)
π_{WR}	0.24	(0.13, 0.41)	0.18	(0.10, 0.28)
π_{WW}	0.73	(0.58, 0.84)	0.64	(0.51, 0.76)

litter mates are observed for each of 50 litters. One of the three mates received a treatment, while the other two serve as controls. Responses consist of either the week of tumor occurrence or the week of death before the instance of any tumor. In any case, all rats were sacrificed after 104 weeks. Scientific interest is addressed to evaluating a possible association between treatment and time to tumor.

Let $\mathrm{Y}_i = \min\{\mathrm{T}_i, c_i\}$ denote the response with T_i being the time to tumor and c_i the censoring time, $i = 1, \dots, 150$. Marginally, we assume a Weibull regression model for the survival times T_i, that is $\mathrm{F}_i(t_i; \boldsymbol{\lambda}) = 1 - \exp\{-(t_i/\eta_i)^{\alpha}\}$, where α denotes the shape parameter, $\eta_i = \exp(\beta_0 + \beta_1 x_i)$ with x_i being the indicator for treatment and $\boldsymbol{\lambda} = (\alpha, \beta_0, \beta_1)^{\mathrm{T}}$. This model corresponds to both an accelerated life model and a proportional hazard model.

Survival times for different rats are assumed independent if coming from different litters and associated if coming from the same litter. Accordingly, the correlation matrix of the errors is modelled by the Kronecker product $\boldsymbol{\Omega} = \mathbf{I}_{50} \otimes \boldsymbol{\Omega}_1$, where $\boldsymbol{\Omega}_1$ is a 3×3 exchangeable correlation matrix with equicorrelation parameter τ. The likelihood is the product of the contributions from the different litters. Each of these is a trivariate density which assumes a different form depending on the number of censored observations. For notational simplicity, consider the first triplet $p(y_1, y_2, y_3; \boldsymbol{\theta})$. Then, we have the following possibilities:

1. no censored observations,

$$p(y_1, y_2, y_3; \boldsymbol{\theta}) = p_1(t_1; \boldsymbol{\lambda}) p_2(t_2; \boldsymbol{\lambda}) p_3(t_3; \boldsymbol{\lambda}) q(\epsilon_1, \epsilon_2, \epsilon_3; \boldsymbol{\theta}),$$

   where $p_i(\cdot; \boldsymbol{\lambda})$ is the density function of the Weibull regression model;
2. one censored observation, say the third one,

$$\begin{aligned} p(y_1, y_2, y_3; \boldsymbol{\theta}) &= \int_{c_3}^{\infty} p(t_1, t_2, t_3; \boldsymbol{\theta}) \mathrm{d}t_3 \\ &= p_1(t_1; \boldsymbol{\lambda}) p_2(t_2; \boldsymbol{\lambda}) q_{12}(\epsilon_1, \epsilon_2; \boldsymbol{\theta}) \int_{\Phi^{-1}\{\mathrm{F}_3(c_3; \boldsymbol{\lambda})\}}^{\infty} p(\epsilon_3 | \epsilon_1, \epsilon_2; \boldsymbol{\theta}) \mathrm{d}\epsilon_3, \end{aligned}$$

where $p(\epsilon_3|\epsilon_1, \epsilon_2; \boldsymbol{\theta})$ is the density of a normal variable with mean $\tau/(1+\tau)(\epsilon_1+\epsilon_2)$ and variance $1-2\tau^2/(1+\tau)$;

3. two censored observations, say the last two,

$$\begin{aligned} p(y_1, y_2, y_3; \boldsymbol{\theta}) &= \int_{c_2}^{\infty}\int_{c_3}^{\infty} p(t_1, t_2, t_3; \boldsymbol{\theta})\mathrm{d}t_2\mathrm{d}t_3 \\ &= p_1(t_1; \boldsymbol{\lambda})\int_{\Phi^{-1}\{\mathrm{F}_2(c_2;\boldsymbol{\lambda})\}}^{\infty}\int_{\Phi^{-1}\{\mathrm{F}_3(c_3;\boldsymbol{\lambda})\}}^{\infty} p(\epsilon_2, \epsilon_3|\epsilon_1; \boldsymbol{\theta})\mathrm{d}\epsilon_2\mathrm{d}\epsilon_3, \end{aligned}$$

where $p(\epsilon_2, \epsilon_3|\epsilon_1; \boldsymbol{\theta})^{\mathrm{T}}$ is the density of a bivariate normal variable with mean vector $(\tau\epsilon_1, \tau\epsilon_1)$ and variance matrix $(1-\tau)\left(\begin{smallmatrix} 1+\tau & \tau \\ \tau & 1+\tau \end{smallmatrix}\right)$;

4. all censored observations

$$\begin{aligned} p(y_1, y_2, y_3; \boldsymbol{\theta}) &= \int_{c_1}^{\infty}\int_{c_2}^{\infty}\int_{c_3}^{\infty} p(t_1, t_2, t_3; \boldsymbol{\theta})\mathrm{d}t_2\mathrm{d}t_3 \\ &= \int_{\Phi^{-1}\{\mathrm{F}_1(c_1;\boldsymbol{\lambda})\}}^{\infty}\int_{\Phi^{-1}\{\mathrm{F}_2(c_2;\boldsymbol{\lambda})\}}^{\infty}\int_{\Phi^{-1}\{\mathrm{F}_3(c_3;\boldsymbol{\lambda})\}}^{\infty} p(\epsilon_1, \epsilon_2, \epsilon_3; \boldsymbol{\theta})\mathrm{d}\epsilon_1\mathrm{d}\epsilon_2\mathrm{d}\epsilon_3, \end{aligned}$$

where $p(\epsilon_1, \epsilon_2, \epsilon_3; \boldsymbol{\theta})$ is the density of a trivariate normal variable with zero mean and variance matrix $\boldsymbol{\Omega}_1$.

Hence, likelihood computations require approximation of rectangular normal probabilities of dimension two and three. For these low dimension integrals it is not necessary to use the importance sampling algorithm stated in Section 5.2. Instead, we use the precise deterministic approximation provided by the Miwa et al. [38] algorithm. The estimated marginal parameters are $\hat{\alpha} = -3.79$ (0.55), $\hat{\beta}_0 = 4.98$ (0.08) and $\hat{\beta}_1 = -0.24$ (0.09), the latter supporting a treatment effect. The equicorrelation parameter estimate $\hat{\tau} = 0.53$ (0.15) confirms by Property 2 the expected presence of positive association between survival times of rats from the same litter.

9.4. Spatial regression with count data

Data on incidence of male lip cancer in Scotland during years 1975-1980 have been analysed by many authors for illustrating varying disease mapping methods, see Waller and Gotway [55] and Wakefield [56]. Data consist of observed Y_i and expected number of cases e_i in each of the 56 counties of Scotland and they are available through the home page of Waller and Gotway's book at `http://www.sph.emory.edu/~lwaller/book/ch2/scotland.dat`. Interest lies in studying whether excess of cases can be associated with the proportion of the population employed in agriculture, fishing, or forestry (AFF), whose map is displayed in the right panel of Figure 3. Some authors refer to a possible spatial trend effect along the south-north direction, with a demographic interpretation because the most northern counties of Scotland are sparsely populated. The left panel of Figure 3 displays the standardized morbidity ratio (SMR) defined as the ratio of the observed to expected cases. This map appears to confirm the south-north trend.

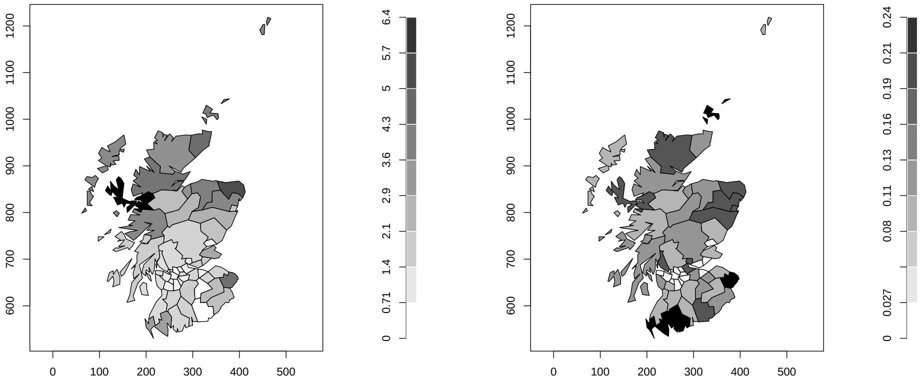

Fig 3. *Scotland lip cancer data: SMR (left panel) and AFF (right panel) maps.*

The standard non-spatial model for these data consists in assuming that the observed cases Y_i are distributed as a Poisson variable with mean $\mu_i = \phi_i e_i \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$ where x_{1i} is the AFF covariate and x_{2i} is the latitude coordinate (divided by 100). Quantity ϕ_i is an overdispersion parameter distributed as a Gamma variable with mean 1 and scale κ. In other words, a marginal negative binomial model for Y_i is assumed.

Residual spatial dependence is modelled by assuming that the errors ϵ_i are realizations of an underlying continuous zero mean Gaussian random field model. Under the assumption of uniform spatial distribution within each county, the correlation between the errors of county i and county j is given by the average

$$\mathrm{corr}(\epsilon_i, \epsilon_j) = \frac{1}{|\mathscr{A}_i||\mathscr{A}_j|} \int_{\mathscr{A}_i} \int_{\mathscr{A}_j} \rho(\|\mathbf{s}_i - \mathbf{s}_j\|_2; \boldsymbol{\tau}) \mathrm{d}\mathbf{s}_i \mathrm{d}\mathbf{s}_j, \tag{20}$$

where $\mathscr{A}_i$ denotes the ith county, $|\mathscr{A}_i|$ its area and $\rho(\cdot; \boldsymbol{\tau})$ is a spatial correlation function. Given its flexibility, we consider the Matérn spatial correlation function defined in equation (5).

We start the analysis approximating $\mathrm{corr}(\epsilon_i, \epsilon_j)$ by the spatial correlation function between the centroids of the two areas, $\mathrm{corr}(\epsilon_i, \epsilon_j) \approx \rho(\|\tilde{\mathbf{s}}_i - \tilde{\mathbf{s}}_j\|_2)$, with $\tilde{\mathbf{s}}_i$ denoting the centroid of the ith county. Diggle and Ribeiro [14] affirm that the shape parameter τ_2 of the Matérn correlation function is difficult to identify and suggest choosing its value from the discrete set $\{0.5, 1.5, 2.5\}$, which represents various degrees of mean-square differentiability of the underlying signal process. For these data, we found little difference by varying the shape parameter τ_2. In the following, we present estimates obtained from the model with $\tau_2 = 0.5$, corresponding to the so-called exponential correlation function $\mathrm{corr}(\epsilon_i, \epsilon_j) = \exp\left(-\|\tilde{\mathbf{s}}_i - \tilde{\mathbf{s}}_j\|_2/\tau_1\right)$. The estimated marginal parameters yield the mean response

$$\hat{\mathrm{E}}(\mathrm{Y}_i|x_{1i}, x_{2i}) = e_i \exp(-20.80_{(4.58)} + 4.31_{(1.43)} x_{1i} + 36.74_{(8.06)} x_{2i}),$$

and the non-spatial overdispersion parameter ϕ_i is distributed as a Gamma variable with mean of one and estimated scale parameter of $\hat{\kappa} = 0.17$ (0.06). Covariate AFF (x_1) is significantly positively associated with excess of tumor incidence. The estimated spatial trend coefficient (x_2) is also highly significant although this association should be interpreted with care because the most northern counties are very sparsely populated resulting in zero expected cases.

There is evidence of some local residual spatial correlation. The estimate of the correlation parameter of the exponential correlation function is 14.36 Km (6.19 Km). Hence, two counties more than $3 \times 14.36 = 43.08$ Km apart have errors with a correlation lower than 0.05. There are 157 distinct pairs of counties with centroids at distance lower than 43 Km among the observed $56!/(2!54!) = 1,540$ pairs. It follows by Property 1 that there is weak residual spatial dependence in the responses once accounting for the AFF covariates and the south-north trend.

A more precise analysis should consider the geography of Scotland and require the computation of the correlation between the errors generated by an underlying continuous process as in formula (20). However, this more elaborated analysis seems likely to be nonessential because of the weak local spatial dependence once the covariates AIFF and the spatial trend are included.

9.5. Computational details

The models in examples 9.1, 9.2 and 9.4 are fitted with the GHK algorithm. Each model is first fitted with K=100 Monte Carlo replications using the maximum independence likelihood estimates as starting values for the marginal parameters. Then, the model is re-fitted with a larger value of K=1,000 replications using the initial estimates as starting values. Our experience is that K=1,000 is a sufficient size to obtain statistically stable estimates with data set up to few hundreds of observations. The computational time needed to fit the various models depends on various factors, not only the number of observations. Other crucial factors are the number of dependence parameters, the degree of dependence and the level of discreteness of the responses. For example, with a MacBook Air notebook with processor 1.8 Ghz Intel Core i7 and 4 Gb of memory we need only 0.08 minutes to fit the spatial model to Scotland Lip Cancer data, which involve 56 observations, only one dependence parameter and the degree of dependence is very low. The computational time for fitting the ARMA(2,1) model to the Polio data is instead 1.2 minutes. This longer time is due to the larger sample size ($n = 168$), the presence of three dependence parameters and significant serial dependence. The analysis of Salamander data is more time consuming as model fitting needs 4.78 minutes. This time is explained not only by the size of 360 observations, but also because there are six dependence parameters and the responses are binary.

10. Concluding discussion

In this paper we have discussed the class of Gaussian copula models for marginal regression analysis of correlated non-normal data. The convenient model specification allows for simple interpretation of marginal parameters and great flexibility in specification of the dependence structure. Applications to time series, longitudinal studies, spatial data and survival analysis have been used to illustrate this flexibility. Residual analysis and the Hausman-type specification test can be used to validate the adequacy of the assumed multivariate model and simulation studies reported in Section 8 suggest that a certain level of local misspecification can be tolerated. Investigation of this apparent robustness is perhaps the most interesting direction for future research. Other possible future research directions include the following aspects.

1. *High dimensional data.* Numerical difficulties may arise from the inversion of matrix $\mathbf{\Omega}$ with high-dimensional data and from the dimension of the integral needed to compute the likelihood in the non-continuous case.

   In these cases it is necessary to simplify either the correlation model for the errors or the estimation procedure. An example of model simplification for spatial applications is the approximation of errors following a Gaussian random field by a Gaussian Markov random field in a fine grid, as shown in Rue and Tjelmeland [46]. With this approximation, the computational cost for inversion of $\mathbf{\Omega}$ roughly reduces from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^{3/2})$ in two dimensions and $\mathcal{O}(n^2)$ in three dimensions.

   An example of simplified estimation procedure is the method of maximum composite marginal likelihood [35, 54]. This consists of maximizing a pseudolikelihood constructed from a product of marginal densities, typically low-dimensional. The basic example is the pairwise likelihood formed by bivariate densities

$$\mathscr{L}_{\text{pair}}(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} p_{ij}(y_i, y_j; \boldsymbol{\theta})^{w_{ij}},$$

   where w_{ij} are suitable weights and the bivariate densities are given in formulas (6) and (7). The pairwise likelihood does not require inversion of $\mathbf{\Omega}$ and has a computational cost of $\mathcal{O}(n^2)$ if all pairs are used. Furthermore, in the non-continuous case only bivariate integrals (7) need to be evaluated. The computational saving can be substantial and can be further improved by a suitable choice of the weights w_{ij} to remove less informative pairs, such as those formed by too-distant observations in spatial applications. Maximum pairwise likelihood estimators are consistent and asymptotically normal under appropriate regularity conditions, and in many applications they have competitive efficiency. Zhao and Joe [62] study the performance of pairwise likelihood estimators for Gaussian copula models and conclude that this method perform well.
2. *Elliptic distributions other than normal for the errors.* Although simulation results reported in Section 8 suggest that inference is relatively robust

to misspecification of the joint distribution for the errors, in some applications it may useful to consider more flexible elliptic distributions for handling heavy tails or skewness of the errors. For example, the assumed Gaussian copula is not appropriate for modelling extreme values events because of its well-known lack of tail dependence. In this case, it may be appropriate to assume instead a Student-t distribution for the errors [40].

3. *Maximization by parts.* Maximization by parts is a numerical iterative algorithm proposed by Song et al. [49] to optimize complex log-likelihoods that can be partitioned into a manageable "working log-likelihood" plus a more complex "remainder log-likelihood". See also Song [48]. The algorithm aims to enhance numerical stability relative to the direct numerical optimization of the likelihood function. Among other applications, maximization by parts has been proposed for fitting continuous Gaussian copula regression models.
4. *Maximum simulated likelihood via Markov chain Monte Carlo algorithms.* Recently, Jeliazkov and Lee [26] show that Markov chain Monte Carlo algorithms designed for marginal likelihood computation in Bayesian inference can be used for efficient maximum simulated likelihood analysis. The key ingredient of this proposal is the method of Chib [6] for marginal likelihood estimation from the output of Gibbs sampling algorithms. Translated to the context of this paper, one possibility described in Jeliazkov and Lee [26] is drawing from $p(\boldsymbol{\epsilon}|\boldsymbol{y};\boldsymbol{\theta})$ with the Gibbs sampling algorithm by Geweke [19] and then estimate the marginal likelihood with Chib's method with Rao-Blackwellization. Jeliazkov and Lee [26] discuss other variants of this idea and compare them against the GHK algorithm in the context of multivariate probit models.

Acknowledgements

The authors would like to thank R. Bellio, C. Bhat, M. Cattelan and A. Guolo for comments. A special thank goes to D. Firth for valuable discussion and continuous encouragement. The authors are also thankful to a Referee and an Associate Editor for comments which have improved the presentation of the paper.

References

[1] Anděl, J., Netuka, I. and Svara, K. (1984). On threshold autoregressive processes. *Kybernetika* **20**, 89–106. MR0747062

[2] Andrews, D.W.K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* **59**, 817–858. MR1106513

[3] Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* **12**, 171–178. MR0808153

[4] Bodnar, O. Bodnar, T., and Gupta, A.K. (2010). Estimation and inference for dependence in multivariate data. *Journal of Multivariate Analysis* **101**, 869–881. MR2584905

[5] Booth, J.G. and Hobert, J.P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B* **61**, 265–285.

[6] Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* **90**, 1313–1321. MR1379473

[7] Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* **85**, 347–361.

[8] Cox, D.R., and Snell, E.J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society, Series B* **30**, 248–275. MR0237052

[9] Craig, P. (2008). A new reconstruction of multivariate normal orthant probabilities. *Journal of the Royal Statistical Society, Series B* **70**, 227–243. MR2412640

[10] Cressie, N. (1993). *Statistics for Spatial Data*. Wiley, New York. MR1239641

[11] de Leon, A.R. and Wu, B. (2011). Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Statistics in Medicine* **30**, 175–185. MR2758273

[12] de Leon, A.R., Wu, B., and Withanage, N. (2012). Joint analysis of mixed discrete and continuous outcomes via copula models. Preprint `http://math.ucalgary.ca/~adeleon/Chapter11.pdf`.

[13] Diggle, P.J., Heagerty, P., Liang, K.-Y. and Zeger, S.L. (2002). *Analysis of longitudinal data*. Second edition. Oxford University Press, Oxford. MR2049007

[14] Diggle, P.J. and Ribeiro, P.J.J. (2007). *Model-based Geostatistics*. Springer, New York. MR2293378

[15] Dunn, P.K. and Smyth, G.K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics* **5**, 236-244.

[16] Durbin, J. and Koopman, S.J. (2001). *Time Series Analysis by State Space Methods*. Oxford University Press. MR1856951

[17] Genest, C. and Nešlehová, J. (2007). A primer on copulas for count data. *Astin Bulletin* **37**, 475–515. MR2422797

[18] Genz, A. and Bretz, F. (2002). Methods for the computation of multivariate t-probabilities. *Journal of Computational and Graphical Statistics* **11**, 950–971. MR1944269

[19] Geweke, J. (1991). Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints. In Proceedings of the 23rd Symposium in the Interface, Interface Foundation of North America, Fairfax.

[20] Gueorguieva, R.V. and Agresti, A. (2001). A correlated probit model for joint modelling of clustered binary and continuous responses. *Journal of the American Statistical Association* **96**, 1102–1112. MR1947258

[21] Harris, B. (1988). Tetrachoric correlation coefficient, *in* L. Kotz and N. Johnson (eds.) *Encyclopedia of Statistical Sciences* **9**, 223–225. Wiley.

[22] Hausman, J.A. (1978). Specification tests in econometrics. *Econometrica*, **46**, 1251–1271. MR0513692

[23] Hoff, P.D. (2007). Extending the rank likelihood for semiparametric copula estimation. *Annals of Applied Statistics* **1**, 265–283. MR2393851

[24] Hothorn, A., Bertz, F., and Genz, A. (2001). On multivariate T and Gaussian probabilities in R. *R News* **1**, 27–29.

[25] Hurvich, C.M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297–307. MR1016020

[26] Jeliazkov, I. and Lee, E.H. (2010). MCMC perspectives on simulated likelihood estimation. *Advances in Econometrics* **26**, 3–40.

[27] Joe, H. (1995). Approximation to multivariate normal rectangle probabilities based on conditional expectations. *Journal of the American Statistical Association* **90**, 957–964. MR1354012

[28] Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman and Hall. MR1462613

[29] Kauermann, G. and Carroll, R.J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association* **96**, 1387–1396. MR1946584

[30] Keane, M.P. (1994). A computationally practical simulation estimator for panel data. *Econometrica* **62**, 95–116.

[31] Klaassen, C.A. and Wellner, J.A. (1997). Efficient estimation in the bivariate normal copula model: normal margins are least favourable. *Bernoulli* **3**, 55–77. MR1466545

[32] Kugiumtzis, D. and Bora-Senta, E. (2010). Normal correlation coefficient of non-normal variables using piece-wise linear approximation. *Computational Statistics* **25**, 645–662. MR2738694

[33] Le Cessie, S. and Van Houwelingen, J.C. (1994). Logistic regression for correlated binary data. *Applied Statistics* **43**, 95–108.

[34] Liang, K.-L. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22. MR0836430

[35] Lindsay, B.G. (1988). Composite likelihood methods. *Contemporary Mathematics* **80**, 221–240. MR0999014

[36] Mantel, N., Bohidar, N.R. and Ciminera, J.L. (1977). Mantel-Haenszel analysis of litter-matched time-to-response data, with modifications to recovery of interlitter information. *Cancer Research* **37**, 3863–3868.

[37] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* Second edition. Chapman and Hall. MR0727836

[38] Miwa, T., Hayter, A.J. and Kuriky, S. (2003). The evaluation of general non-centred orthant probabilities. *Journal of the Royal Statistical Society, Series B* **65**, 223-234. MR1959823

[39] Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*, Springer. MR2171048

[40] Nikoloulopoulos, A.K., Joe, H. and Li, H. (2011). Weighted scores method for regression models with dependent data. *Biostatistics* **12**, 653–665.

[41] Nikoloulopoulos, A.K., Joe, H. and Chaganty, N.R. (2011). Extreme value properties of multivariate t copulas. *Extremes* **12**, 129–148. MR2515644

[42] Parzen, M., Ghosh, S., Lipsitz, S., Sinha, D., Fitzmaurice, G.M., Mallick, B.K., Ibrahim, J.G. (2011). A generalized linear mixed model for longitudinal binary data with a marginal logit link function. *Annals of Applied Statistics* **5**, 449–467. MR2810405

[43] Pitt, M., Chan, D. and Kohn, R. (2006). Efficient Bayesian inference for Gaussian copula regression models. *Biometrika* **93**, 537–554. MR2261441

[44] R Development Core Team (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org.

[45] Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics* **23**, 470–472. MR0049525

[46] Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian random fields to Gaussian fields. *Scandinavian Journal of Statistics* **29**, 31–50.

[47] Song, P.X.-K. (2000). Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics* **27**, 305–320. MR1777506

[48] Song, P.X-K. (2007). *Correlated Data Analysis: Modeling, Analytics and Applications*. Springer-Verlag. MR2377853

[49] Song, P.X.-K., Fan, Y. and Kalbfleisch, J.D. (2005). Maximization by parts in likelihood inference (with discussion). *Journal of the American Statistical Association* **100**, 1145–1167. MR2236431

[50] Song, P.X.-K., Li, M. and Yuan, Y. (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics* **65**, 60–68. MR2665846

[51] Sung, Y.J. and Geyer, C.J. (2007). Monte Carlo likelihood inference for missing data models. *The Annals of Statistics* **35**, 990–1011. MR2341695

[52] Tong, H. (1990). *Non-Linear Time Series: A Dynamical System Approach*. Oxford: Oxford University Press. MR1079320

[53] Train, K.E. (2003). *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press. MR2003007

[54] Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* **21**, 5–42. MR2796852

[55] Waller, L.A. and Gotway, C.A. (2004). *Applied Spatial Statistics for Public Health Data*. New York: John Wiley and Sons. MR2075123

[56] Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics* **8**, 158–183.

[57] White, H. (1994). *Estimation, Inference and Specification Analysis*. Cambridge University Press. MR1292251

[58] Wu, B. and de Leon, A.R. (2012). Flexible random effects copula models for clustered mixed bivariate outcomes in developmental toxicology. Preprint http://math.ucalgary.ca/~adeleon/re_copula_paper_may2012.pdf.

[59] Zeger, S.L. (1988). A regression model for time series of counts. *Biometrika* **75**, 822–835. MR0995107

[60] Zeger, S.L. and Karim, M.R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association* **86**, 79–86. MR1137101

[61] Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software* **16**, issue 9.

[62] Zhao, Y. and Joe, H. (2005). Composite likelihood estimation in multivariate data analysis. *The Canadian Journal of Statistics* **33**, 335–356. MR2193979

[63] Zucchini, W. and MacDonald, I.L. (2009). *Hidden Markov Models for Time Series*. Chapman & Hall/CRC. MR2523850