Gaussian copula regression using R

Guido Masarotto
University of Padua

Cristiano Varin
Ca' Foscari University, Venice

Framework

Gaussian copula marginal regression models

- different types of responses: continuous, discrete, categorical
- several forms of possible dependence:
  - time series
  - longitudinal/panel studies
  - spatial data
  - ...
- likelihood analysis
- alternative to generalized estimating equations
- R package `gcmr`

Model specification

1) response Y_i related to covariates vector $\boldsymbol{x}_i$ and error ϵ_i by

$$Y_i = F^{-1}\left\{\Phi(\epsilon_i)|\boldsymbol{x}_i;\boldsymbol{\lambda}\right\}, \quad i = 1,\ldots,n,$$

where

- $F(y_i|\boldsymbol{x}_i;\boldsymbol{\lambda})$ is c.d.f. of Y_i given $\boldsymbol{x}_i$
- $\Phi(z)$ is c.d.f. of $\mathsf{N}(0,1)$

2) multivariate normal errors

$$\boldsymbol{\epsilon} = (\epsilon_1,\ldots,\epsilon_n)^{\mathsf{T}} \sim \mathsf{MVN}(\mathbf{0},\boldsymbol{\Omega}),$$

correlation matrix $\boldsymbol{\Omega}$ parametrized so to reflect various types of dependence (time series, clustered data, spatial data, etc.)

Gaussian copula marginal regression

- copulas conveniently separate marginals (regression model) from the dependence component
- possible marginals: GLMs, skew-normal models, Weibull regression, beta regression, zero-inflated models, ...
- accommodate various forms of dependence through suitable choice of the copula correlation matrix:
  - longitudinal data: ARMA(p,q), exchangeable, unstructured
  - time series: ARMA(p,q)
  - spatial data: Matérn correlation matrix
- Gaussian copula attractive because inherits several well-known properties of multivariate normal
- special case: multivariate probit regression

Likelihood analysis

- continuous case: likelihood (nicely) in closed-form
- noncontinuous case: likelihood is awkward n-dimensional normal integral
  - hard to handle if n is not (very) small
  - generalization of the multivariate probit likelihood
  - many methods for fitting multivariate probit
  - gold standard (?): Geweke-Hajivassiliou-Keane simulator
  - efficient and relatively easy to program
  - implemented in gcmr

Package gcmr

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

gcmr: Gaussian Copula Marginal Regression

likelihood inference in Gaussian copula marginal regression models

Version:	0.1
Priority:	optional
Depends:	R (≥ 2.10), nlme, sandwich, sn, geoR, sp
Published:	2011-08-10
Author:	Guido Masarotto and Cristiano Varin
Maintainer:	Guido Masarotto <guido.masarotto@unipd.it> and Cristiano Varin <sammy at unive.it>
License:	GPL (≥ 2)
CRAN checks:	gcmr results

Downloads:

Package source:	gcmr_0.1.tar.gz
MacOS X binary:	gcmr_0.1.tgz
Windows binary:	gcmr_0.1.zip
Reference manual:	gcmr.pdf

Package gcmr

```
> library(gcmr)
```

```
> args(gcmr)
```

```
function(formula, data, subset, offset, contrasts = NULL,
marginal, cormat, start, fixed, options = gcmr.options())
```

nonstandard arguments:

- marginal marginal model
- cormat Gaussian copula correlation matrix
- start optional starting values
- fixed parameters fixed to a priori values
- options various options, such as fix the pseudorandom seed, number of Monte Carlo replications, choice of the optimizator (default is `nlminb`), ...

Package gcmr: model specification

marginal class marginal models

`gs.marg(link=linear)`	Gaussian
`bn.marg(link=logit)`	binomial
`ps.marg(link=log)`	Poisson
`nb.marg(link=log)`	negative binomial
`sn.marg(link=linear)`	skew-normal Azzalini (2005)

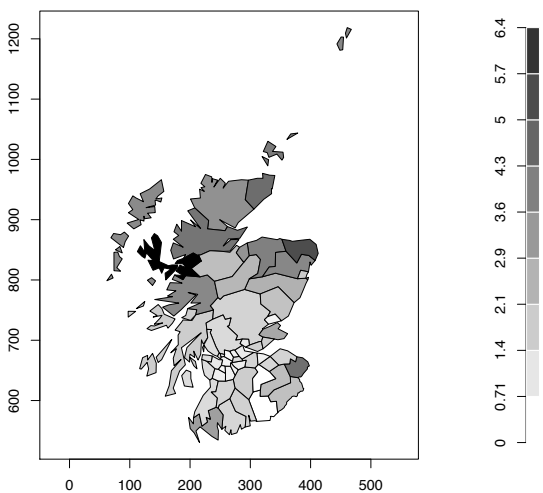
cormat class correlation matrices

`ind.cormat()`	working independence
`arma.cormat(p,q)`	ARMA(p,q)
`cluster.cormat(id, type)`	cluster/longitudinal
	type={AR(1), MA(1), exch, unstr}
`matern.cormat(D, k)`	Matérn spatial correlation
	D distance, k smoothness parameter

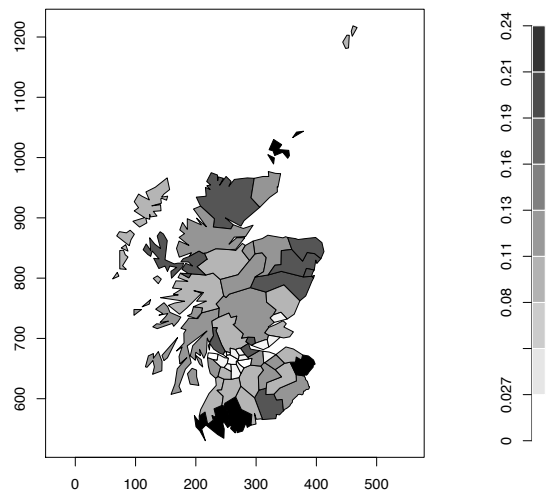
user can specify more marginals and correlation matrices

Scottish lip cancer

- incidence of male lip cancer in Scotland during 1975-1980
- response: observed cases Y_i in each county of Scotland ($n = 56$), also available expected cases E_i
- question: excess of cases associated with proportion of population employed in agriculture, fishing or forestry (AFF)?
- need model for spatially correlated counts



standardized morbidity ratio Y/ E



AFF

Scottish lip cancer

- standard non-spatial analysis of these data:
  - observed cases Y_i negative binomial with mean

$$\mu_i = E_i \exp(\beta_0 + \beta_1 \text{AFF} + \beta_2 \text{latitude})$$

  - scale parameter κ
- complement this independence model with spatial variability:
  - Gaussian copula with Matérn correlation matrix

Scottish lip cancer

```
> library(xtable); library(gcmr)
> data(scotland)
> D.scot <- spDists(cbind(scotland$longitude,
+     scotland$latitude), longlat = TRUE)
> m <- gcmr(observed ~ offset(log(expected)) +
+     AFF + I(latitude/100), data = scotland,
+     marginal = nb.marg(link=log),
+     cormat = matern.cormat(D.scot, k=0.5),
+     options = list(seed = 71271))
> xtable(cbind(coef = coef(m), se = se(m)))
```

	coef	se
(Intercept)	-20.80	4.58
AFF	4.30	1.43
I(latitude/100)	36.74	8.06
kappa	0.17	0.06
range	14.36	6.19

gcmr: methods

available methods:

- profile likelihood

  ```
  profile(fitted, which, low, up, npoints = 10,
          display = TRUE, alpha = 0.05, ...)
  ```

- quantile residuals

  ```
  residuals(object, type=c("conditional","marginal"),
            method=c("random","mid"),...)
  ```

- variance-covariance matrix of estimators

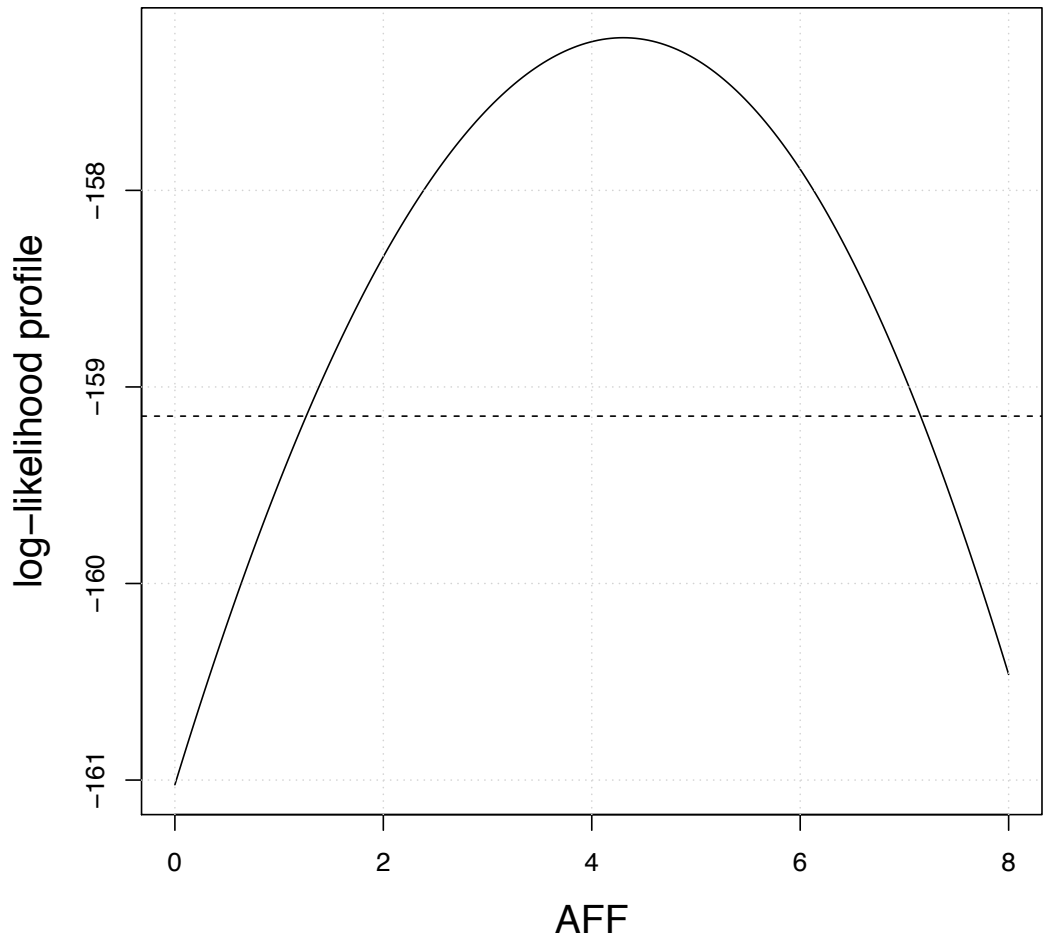
  ```
  vcov(object, type = c("hessian", "sandwich", "vscore",
                        "cluster", "hac"), ...)
  ```

  `se(object, type)` for standard errors

- and other common methods: `coefficients`, `loglik`, `estfun`, `bread`, ...,

Scottish lip cancer

```
> profile(m, which = 2, low = 0, up = 8, alpha=0.05)
```



dashed line is 0.95% confidence interval

Model checking

- continuous case:
  - model adequacy checked through quantile residuals

$$R_i = \Phi^{-1}\{F(Y_i|y_{i-1},\ldots,y_1;\hat{\boldsymbol{\theta}})\}$$

    if model correct $R_i \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$

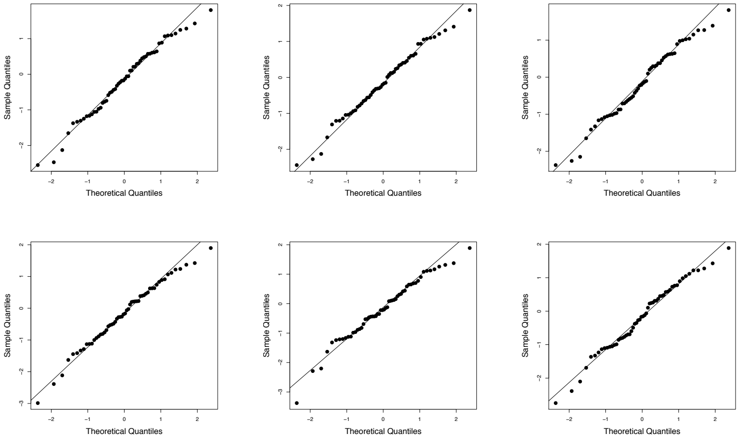
- noncontinuous case:
  - randomized quantile residuals Dunn and Smyth (1996)

$$R_i^{\text{rnd}}(U_i) = \Phi^{-1}\{M_i^- + U_i(M_i - M_i^-)\},$$

    where $U_i \overset{\text{iid}}{\sim} \mathsf{U}(0,1)$ and $M_i^- = F(Y_i^-|y_{i-1},\ldots,y_1;\hat{\boldsymbol{\theta}})$
    (M_i^- is left-hand limit at Y_i^-)
  - if model correct $R_i^{\text{rnd}} \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$ but
  - because of the randomness, it is appropriate to inspect several sets of residuals before taking a decision about the model

Scottish lip cancer

```
> for (i in 1:6) {
+       qqnorm(residuals(m))
+       qqline(residuals(m))
+ }
```



Ohio children wheeze data

- package `geepack`: the dataset is a subset of the six-city study, a longitudinal study of the health effects of air pollution.
- 537 children, four annual measurements each
- variables:

`resp`	wheeze status (1=yes, 0=no)
`id`	children id
age	children' age
`smoke`	indicator of maternal smoking

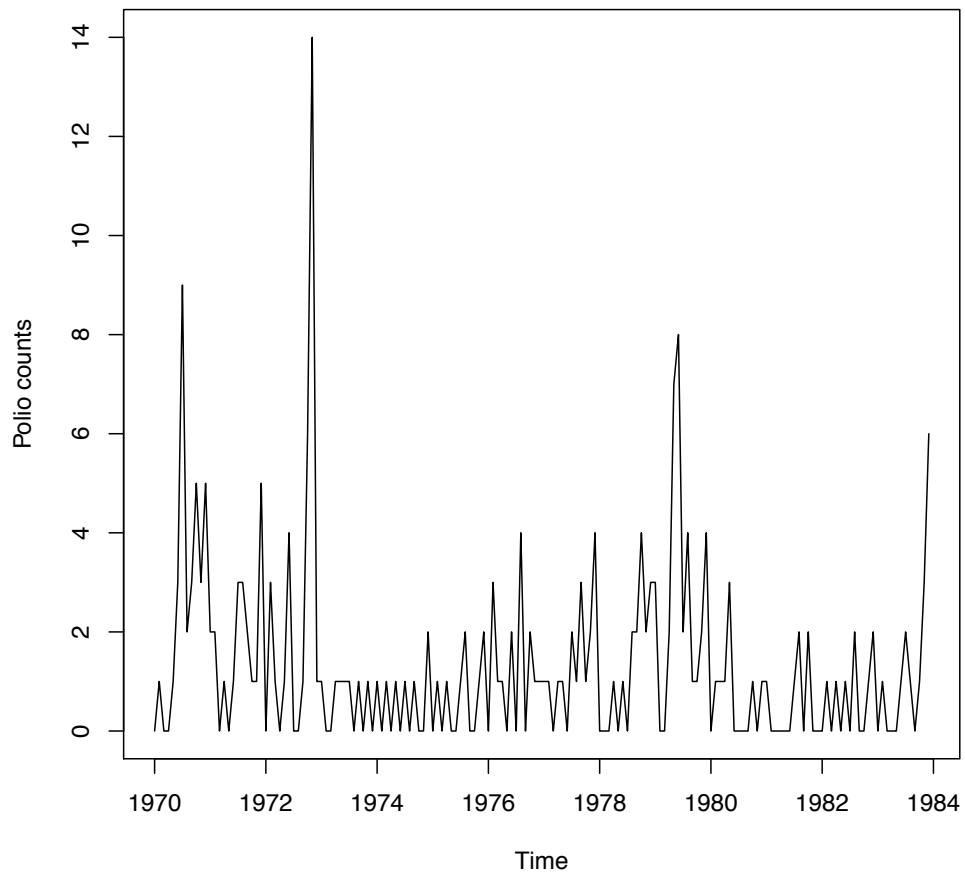
- standard analysis: logistic GEEs
- regressors: age, `smoke` and interaction

```
> library(geepack)
> data(ohio)
> m <- gcmr(cbind(resp, 1 - resp) ~
+     age + smoke + age:smoke,
+     data = ohio, marginal = bn.marg(link=logit),
+     cormat = cluster.cormat(ohio$id,  type = "ar1"),
+     options = list(seed = 123))
> xtable(cbind(coef = coef(m),
+     se.hessian = se(m, "hessian"),
+     se.vscore = se(m, "vscore"),
+     se.sandwich = se(m, "sandwich")))
```

	coef	se.hessian	se.vscore	se.sandwich
(Intercept)	-1.91	0.12	0.12	0.11
age	-0.15	0.07	0.07	0.06
smoke	0.29	0.18	0.18	0.18
age:smoke	0.08	0.10	0.11	0.10
ar1	0.67	0.04	0.03	0.04

Polio incidences in USA

- time series of monthly Polio incidences in USA in 1970-1983
- question: evidence of decreasing trend of Polio in 1970-1983?
- need a regression model for serially correlated counts
- correct for seasonality

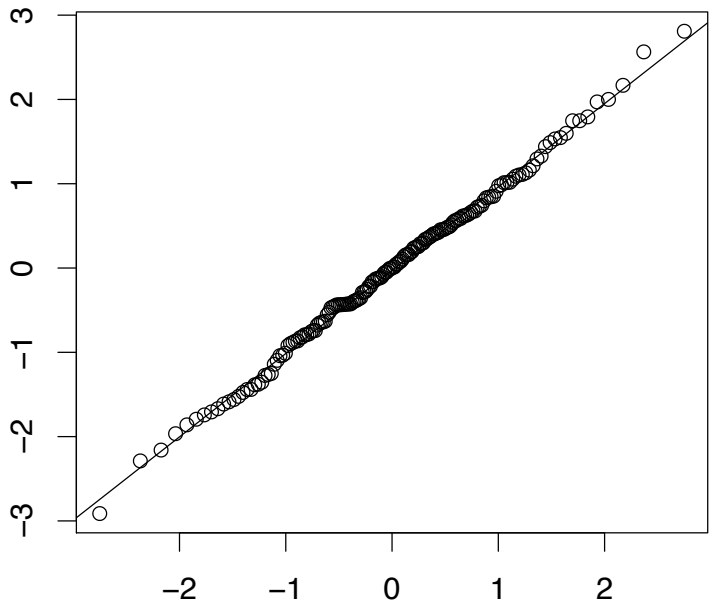


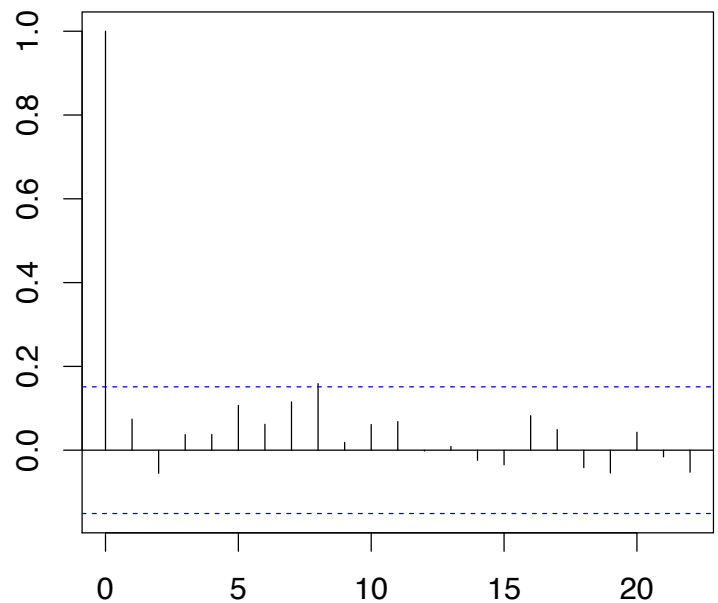
```
> data(polio)
> fit <- gcmr(y ~ ., data = polio,
+     marginal = nb.marg(link=log),
+     cormat = arma.cormat(2, 1),
+     options = list(seed = 71271))
> xtable(cbind(coef = coef(fit),
+     se.hessian = se(fit, "hessian"),
+     se.sandwich = se(fit, "sandwich"),
+     se.hac = se(fit, "hac")))
```

	coef	se.hessian	se.sandwich	se.hac
(Intercept)	0.21	0.12	0.12	0.13
trend	-4.29	2.30	2.32	2.56
$\cos(2\pi t/12)$	-0.12	0.15	0.15	0.15
$\sin(2\pi t/12)$	-0.50	0.16	0.16	0.19
$\cos(2\pi t/6)$	0.19	0.13	0.13	0.13
$\sin(2\pi t/6)$	-0.40	0.13	0.13	0.13
κ	0.57	0.17	0.17	0.17
ar1	-0.51	0.23	0.26	0.27
ar2	0.31	0.09	0.09	0.09
ma1	0.68	0.24	0.27	0.28

Polio time series

```
> qqnorm(residuals(fit))
> qqline(residuals(fit))
> acf(residuals(fit))
```





Conclusions

- Gaussian copula marginal regression: flexible framework for modelling dependence
- can be used to extend many regression models for independent data already available in other `R` packages
- future (!): more models, in particular zero-inflated responses, ordinal and multinomial responses
- future (?): pairwise likelihood for high-dimensional problems
- future (??): other methods for maximum simulated likelihood (MCMC?), MCEM
- some theory and computational details can be found in Masarotto, G. and Varin, C. (2011). Gaussian copula marginal regression (preprint available upon request)
- future theoretical work: robustness to copula misspecification