copCAR: An R Package for Copula-based Inference for Discrete Areal Data

Emily Goren

Master of Science Plan B

May 30, 2014

Areal Data

Spatial data are geographically referenced:

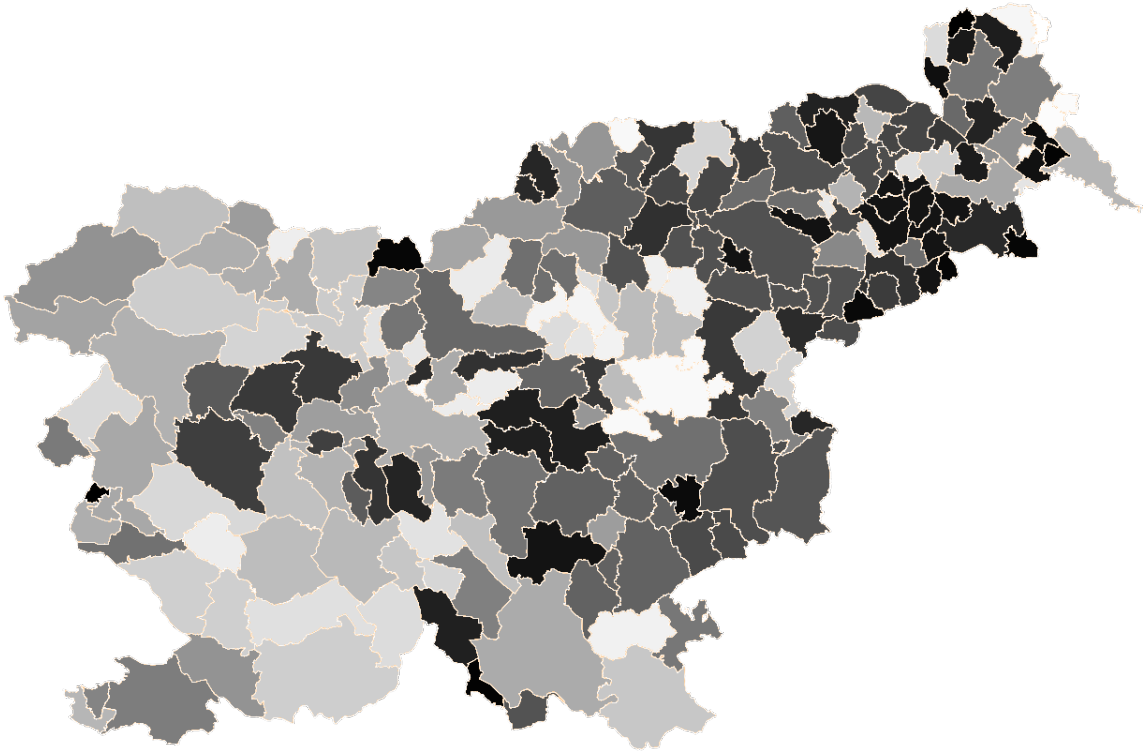
$$\{z(s) : s \in D \subset \mathbb{R}^d\},$$

where z is an observation at a spatial location s on a d-dimensional index set D.

Areal (or lattice) data occur when D is discrete and fixed, such as

- pixels in an image,
- geopolitical units.

An Example: Slovenia Stomach Cancer Data 1995–2001



Areal units are the municipalities of Slovenia.

$\text{SIR}_i = O_i/E_i$ for $i = 1, \ldots, 194$.

Dependence

Spatial data are almost always *dependent*.

Many familiar statistical methods assume observations are independent, such the generalized linear model (GLM), which includes ANOVA and the t-test as special cases.

Failure to account for dependence can result in erroneous inference due to:

(1) inaccurate estimates of regression coefficients, and/or
(2) underestimating the variability of the regression coefficients.

Areal Models

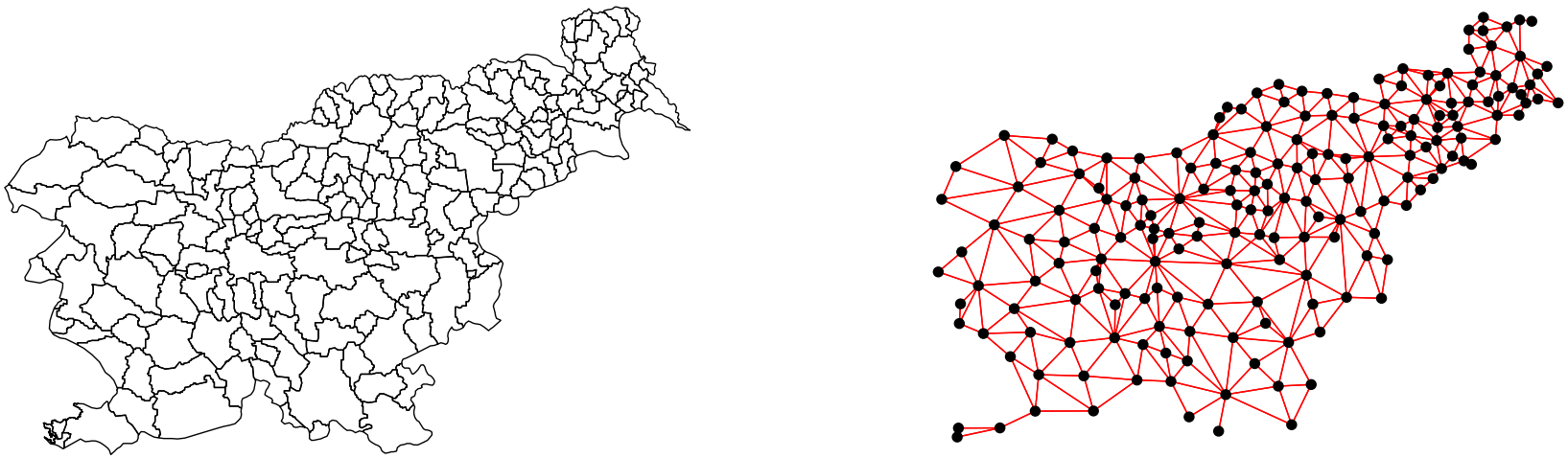
Appropriate models for spatial data incorporate spatial dependence.

For areal data, the adjacency structure between the areal units is often considered rather than any notion of distance.

Adjacency structure is captured by an underlying graph $G = (V, E)$, where

- $V = \{1, 2, \ldots, n\}$ are the vertices, and
- $E \subset V \times V$ are the edges.

Underlying Graph for Slovenia Stomach Cancer Data



Municipalities of Slovenia (left) and underlying graph (right).

Types of Areal Models

Common regression models for areal data:

- Automodel (Besag, 1974)
  - induces dependence directly via the autocovariate
- Spatial generalized linear mixed model (SGLMM) (Besag et al., 1991)
  - induces dependence hierarchically via a field of spatial random effects

Both models have statistical and computational drawbacks, so Hughes (2014) proposed the copCAR model, which employs a copula to induce dependence and uses the SGLMM's conditional autoregression (CAR).

Copulas

A copula C is a multivariate cdf with standard uniform marginals (Nelson, 2006).

$C : [0,1]^n \to [0,1]$.

Copulas are well suited for modeling dependence:

- captures the dependence structure between random variables,
- allows the dependence structure and marginals to be modeled separately.

Sklar's Theorem

Given random variables $X_1, \ldots, X_n$ with n-dimensional joint cdf

$$F(x_1, \ldots, x_n) = \Pr(X_1 \leq x_1, \ldots, X_n \leq x_n),$$

and marginal cdf's

$$F_i(x_i) = \Pr(X_i \leq x_i), \quad i = 1, \ldots, n,$$

by Sklar's Theorem (Sklar, 1959) there exists a copula C such that

$$F(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n)).$$

Gaussian Copula

By the probability integral transform, the random variable $U_i = F_i(X_i)$ is standard uniform, so that

$$C(u_1, \ldots, u_n) = F(F_1^{-1}(u_1), \ldots, F_n^{-1}(u_n)).$$

The Gaussian copula is specified as:

$$C_{\mathbf{R}}(u_1, \ldots, u_n) = \Phi_{\mathbf{R}}(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n)),$$

where

- $\Phi_{\mathbf{R}}$ is the cdf of a multinormal random variable with mean $\mathbf{0}$ and correlation matrix $\mathbf{R}$, and
- Φ is the cdf of a univariate standard normal random variable.

copCAR model: Dependence Structure

copCAR uses the CAR copula:

$$C_{\mathbf{Q}^{-1}}(u_1, \ldots, u_n) = \Phi_{\mathbf{Q}^{-1}}(\Phi_{\sigma_1}^{-1}(u_1), \ldots, \Phi_{\sigma_n}^{-1}(u_n)),$$

where $\mathbf{Q} = \mathbf{D} - \rho\mathbf{A}$ and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_n^2)' = \text{diag}(\mathbf{Q}^{-1})$.

$\mathbf{Q}$ is a precision matrix from the SGLMM's proper CAR, where

- $\mathbf{D} = \text{diag}(d_1, \ldots, d_n)$ where d_i is the degree of the ith vertex,
- $\rho \in [0, 1)$ is a spatial dependence parameter,
- $\mathbf{A}$ is the adjacency matrix of the graph G.

copCAR model: Margins

For observations $\mathbf{Z} = (Z_1, \ldots, Z_n)'$, let $F_1, \ldots, F_n$ be the desired marginal cdf's such that $Z_i = F_i^{-1}(U_i)$.

Expected value of the ith margin is

$$\mu_i = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}),$$

where

- g is a link function,
- $\mathbf{x}_i$ is a p-vector of covariates for the ith areal unit,
- $\boldsymbol{\beta}$ is a p-vector of regression coefficients.

$\mathbf{U} = (U_1, \ldots, U_n)'$ is a realization of the CAR copula.

Frequentist Inference for copCAR

For discrete marginals with integer support, the likelihood is intractable (contains a sum of 2^n terms).

Approaches to frequentist inference for $\boldsymbol{\theta} = (\boldsymbol{\beta}', \rho)'$ are:

(1) continuous extension (CE),

(2) distributional transform (DT),

(3) composite marginal likelihood (CML).

Continuous Extension

Developed by Madsen (2009) for Gaussian copula geostatistical models.

Exact up to Monte Carlo error.

Transform discrete observations into continuous ones by convolution with independent standard uniform random variables.

Uses a sampling-based approach of size m.

Continuous Extension

(1) Simulate independent standard uniform vectors:

$$\mathbf{W}_j = (W_{j,1}, \ldots, W_{j,n})' \text{ for } j = 1, \ldots, m.$$

(2) Use continued observations:

$$\mathbf{y}_j^* = (\Phi_{\sigma_1}^{-1}\{F_1^*(z_1 - w_{j,1})\}, \ldots, \Phi_{\sigma_n}^{-1}\{F_n^*(z_n - w_{j,n})\})'.$$

(3) Compute likelihood:

$$L_{\mathrm{CE}}(\boldsymbol{\theta} \mid \mathbf{z}) = \frac{1}{m}\sum_{j=1}^{m} |\mathbf{Q}|^{1/2}|\boldsymbol{\Sigma}|^{1/2} \exp\left\{-\frac{1}{2}\mathbf{y}_j^{*\prime}(\mathbf{Q} - \boldsymbol{\Sigma}^{-1})\mathbf{y}_j^*\right\} \prod_{i=1}^{n} f_i(z_i),$$

where $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\sigma}^2)$.

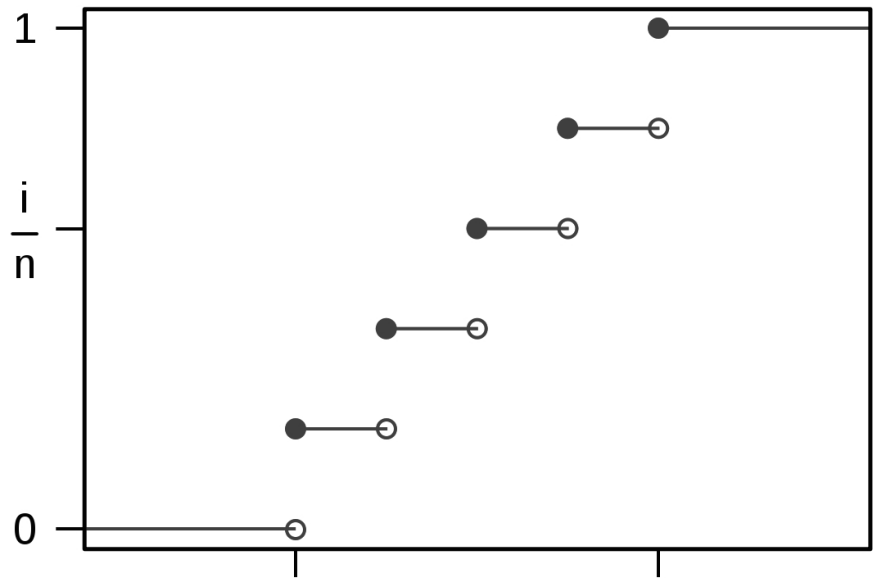
(4) Optimize to obtain MLE:

$$\hat{\boldsymbol{\theta}}_{\mathrm{CE}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\max}\, L_{\mathrm{CE}}(\boldsymbol{\theta} \mid \mathbf{z}).$$

Distributional Transform

Developed by Kazianka and Pilz (2010) for Gaussian copula geostatistical models.

Stochastically "smooths" jump discontinuities in the cdf of a discrete random variable.



http://en.wikipedia.org/wiki/Uniform_distribution_(discrete)

Distributional Transform

(1) Formulate means:

$$u_i = \frac{F_i(z_i^-) + F_i(z_i)}{2} \text{ for } i = 1, \ldots, n.$$

(2) Compute log likelihood:

$$\ell_{\text{DT}}(\boldsymbol{\theta} \,|\, \mathbf{z}) = \frac{1}{2} \log |\mathbf{Q}| + \frac{1}{2} \sum_{i=1}^{n} \log(\sigma_i^2) - \frac{1}{2} \mathbf{y}'(\mathbf{Q} - \boldsymbol{\Sigma}^{-1})\mathbf{y} + \sum_{i=1}^{n} \log f_i(z_i),$$

where $y_i = \Phi_{\sigma_i}^{-1}(u_i)$.

(3) Optimize to obtain MLE:

$$\hat{\boldsymbol{\theta}}_{\text{DT}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ell_{\text{DT}}(\boldsymbol{\theta} \mid \mathbf{z}).$$

Composite Marginal Likelihood

A composite marginal likelihood is the product of valid marginal likelihoods used in place of the ordinary likelihood.

To estimate dependence parameter(s), necessary to model blocks (i.e., pairs) of observations (Varin, 2008).

Composite Marginal Likelihood

(1) Compute log composite marginal likelihood:

$$\ell_{\text{CML}}(\boldsymbol{\theta} \mid \mathbf{z}) = \sum_{\substack{i \in \{1,\ldots,n-1\} \\ j \in \{i+1,\ldots,n\}}} \log \left\{ \sum_{j_1=0}^{1} \sum_{j_2=0}^{1} (-1)^k \Phi_{\mathbf{V}^{ij}}(y_{ij_1}, y_{jj_2}) \right\},$$

where

- $\mathbf{V}^{ij} = \begin{bmatrix} \sigma_i^2 & (\mathbf{Q}^{-1})_{ij} \\ (\mathbf{Q}^{-1})_{ij} & \sigma_j^2 \end{bmatrix}$.
- $y_{\bullet 0} = \Phi_{\sigma_\bullet}^{-1}\{F_\bullet(z_\bullet)\}$,
- $y_{\bullet 1} = \Phi_{\sigma_\bullet}^{-1}\{F_\bullet(z_\bullet - 1)\}$.

(2) Optimize to obtain MLE:

$$\hat{\boldsymbol{\theta}}_{\text{CML}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ell_{\text{CML}}(\boldsymbol{\theta} \mid \mathbf{z}).$$

Confidence Intervals

Use reparameterization $\boldsymbol{\theta} = (\boldsymbol{\beta}', \Phi^{-1}(\rho))'$.

By Geyer (2013),

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{CE}} - \boldsymbol{\theta}) &\xrightarrow{\text{d}} \mathcal{N}\{\mathbf{0}, \mathcal{I}_{\text{CE}}^{-1}(\boldsymbol{\theta})\} \\
\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{DT}} - \boldsymbol{\theta}) &\xrightarrow{\text{d}} \mathcal{N}\{\mathbf{0}, \mathcal{G}_{\text{DT}}^{-1}(\boldsymbol{\theta})\} \\
\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{CML}} - \boldsymbol{\theta}) &\xrightarrow{\text{d}} \mathcal{N}\{\mathbf{0}, \mathcal{G}_{\text{CML}}^{-1}(\boldsymbol{\theta})\},
\end{aligned}$$

where

- $\mathcal{I}_\bullet$ is the appropriate Fisher information matrix,
- $\mathcal{G}_\bullet = \mathcal{I}_\bullet \mathcal{J}_\bullet^{-1} \mathcal{I}_\bullet$ is the Godambe information matrix,
- $\mathcal{J}_\bullet = \text{var}\{\nabla \ell_\bullet(\boldsymbol{\theta} \mid \mathbf{Z})\}$.

Confidence Intervals

Estimate $\mathcal{I}_\bullet$ by the observed Fisher information matrix:

$$\mathcal{H}_\bullet(\hat{\boldsymbol{\theta}}_\bullet) = \nabla^2 \ell_\bullet(\hat{\boldsymbol{\theta}}_\bullet \mid \mathbf{Z}).$$

Estimate $\mathcal{J}_\bullet$ using a parametric bootstrap:

$$\hat{\mathcal{J}}_\bullet(\boldsymbol{\theta}) = \frac{1}{b} \sum_{k=1}^{b} \nabla \nabla' \ell_\bullet(\hat{\boldsymbol{\theta}}_\bullet \mid \mathbf{Z}^{(k)}),$$

where

- b is the bootstrap sample size,
- $\mathbf{Z}^{(k)}$ are datasets simulated from the copCAR model at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_\bullet$.

copCAR Package for R

Main functionality:

(1) simulate Bernoulli or Poisson margins from a given copCAR model,

(2) fit the copCAR model to Poisson marginals (CE, DT, CML) or to Bernoulli marginals (CML).

Includes S3 methods `summary` and `vcov`.

Employs multiple methods to increase computational speed and reliability:

- log-sum-exp trick to prevent underflow in the CE,
- numerical methods for sparse matrices to compute $|\mathbf{Q}|$,
- approximation of the CAR variances $\boldsymbol{\sigma}^2$,
- uses only pairs of adjacent observations when evaluating ℓ_{CML},
- includes C++ implementation.

Simulating Data

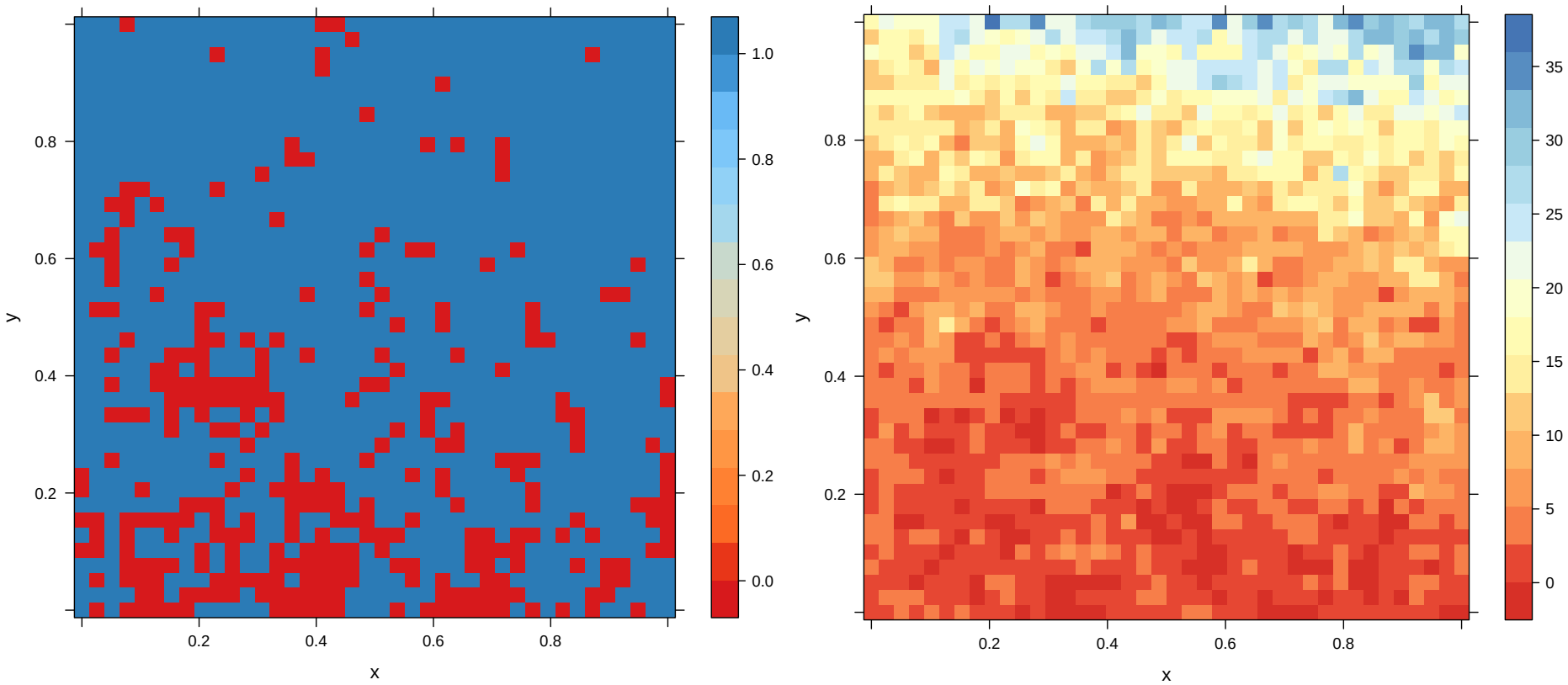
The function to draw data from a given copCAR model has signature

```
rcopCAR(rho, beta, X, A, family)
```

`rcopCAR` simulates data from the copCAR model with given

- spatial dependence parameter `rho`,
- regression coefficient vector `beta`,
- design matrix `X`,
- adjacency matrix `A`,
- marginal distribution and link function specified by `family`.

Example: Simulated Data with $\rho = 0.87$, $\boldsymbol{\beta} = (0.5, 3.0)'$



Bernoulli (left) and Poisson (right) simulated data using the 40×40 square lattice as the underlying graph.

Fitting the copCAR Model

The function to fit the copCAR model has signature

```
copCAR(formula, A, family, method = c("CML", "DT", "CE"),
       conf.int = c("none", "bootstrap", "asymptotic"),
       data, offset = NULL, control = list())
```

- `formula`, `family`, `data`, and `offset` are analogous to those in `glm`,
- `A` is the adjacency matrix for the underlying graph,
- `method` specifies the approach for inference,
- `conf.int` specifies how the confidence interval is computed,
- `control` allows the user to change default settings.

Fitting the copCAR Model: Control List

Arguments allowed in the control list (with default):

- confidence level (0.95),
- bootstrap sample size (500),
- m for the CE (1000),
- Monte Carlo standard error (FALSE) and sample size (100) for the CE,
- maximum value of ρ (0.999) and tolerance (0.001) for approximating the CAR variances,
- verbose (FALSE).

Fitting the copCAR Model: Output

`copCAR` returns an S3 object of class "`copCAR`", a list that includes:

- point estimate of the coefficients,
- confidence interval type,
- response vector, design matrix, and model frame used,
- number of model parameters,
- linear predictors and fitted values for the the margins,
- matched call and supplied formula,
- method used for inference,
- integer code and message returned by `optim` indicating convergence status,
- the terms object used,
- the data argument,
- the levels of the factors used in fitting (if relevant),
- the control list used,
- value of $-\ell_\bullet$ at its minimum.

Fitting the copCAR Model: Output

If confidence intervals were requested:

- confidence intervals for $(\rho, \boldsymbol{\beta}')'$,
- confidence level,
- estimated inverse observed Fisher or Godambe information matrix for $(\Phi^{-1}(\rho), \boldsymbol{\beta}')'$,
- estimated standard errors for $(\rho, \boldsymbol{\beta}')'$,
- bootstrap sample size (if parametric bootstrap was used).

If the Monte Carlo standard error was requested for the CE:

- Monte Carlo standard error for ρ,
- Monte Carlo standard error sample size,
- estimated coefficient of variation for ρ.

Slovenia Stomach Cancer Data

Ordinary GLM for these data is

$$\log \mathrm{E}(O_i) = \log E_i + \beta_1 + \beta_2 x_i,$$

where

- O_i is the observed number of cases in the ith municipality,
- E_i is the expected number of cases in the ith municipality,
- x_i is the standardized socioeconomic status of the ith municipality.

The R code to fit the ordinary GLM is:

```
> fit.GLM = glm(Z ~ x + offset(log(O)), family = poisson)
```

The R code to fit the copCAR model using the CE approach is:

```
> fit.CE = copCAR(Z ~ x + offset(log(O)), A, family = poisson,
      method = "CE", conf.int = "asymptotic")
```

Slovenia Stomach Cancer Data

Applying the summary function to the fitted object gives:

```
> summary(fit.CE)
Call:

copCAR(formula = Z ~ x + offset(log(O)), A = A, family = poisson,
    method = "CE", conf.int = "asymptotic")

Control parameters:

m          1000.00
conf.level    0.95

Coefficients:

      Estimate   Lower    Upper mcse
rho     0.2895  0.1001  0.56800   NA
beta1   0.1532  0.1120  0.19430   NA
beta2  -0.1276 -0.1705 -0.08473   NA

Convergence:

[1] "Optimization converged at loglik = -565.3"
```

Slovenia Stomach Cancer Data Results

Method	β_1	β_2
Ordinary GLM	0.156 (0.120, 0.192)	-0.137 (-0.175, -0.098)
CE	0.153 (0.112, 0.194)	-0.128 (-0.171, -0.085)
DT	0.153 (0.112, 0.194)	-0.128 (-0.169, -0.086)
CML	0.170 (0.128, 0.211)	-0.148 (-0.192, -0.105)

Slovenia Stomach Cancer Data Results

Method	ρ	Time (seconds)
Ordinary GLM	NA	0.031
CE	0.290 (0.100, 0.568)	159.3
DT	0.283 (0.142, 0.469)	98.87
CML	0.279 (0.137, 0.469)	346.1

Future Directions

Incorporate option for parallel computation of the bootstrap and Monte Carlo standard error.

Add method to extract residuals from a fitted copCAR model. Masarotto and Varin (2012) define randomized quantile residuals for a Gaussian copula regression model that can be used as ordinary residuals for checking model fit.

Create a package vignette.

Allow specification of starting values for optimization in the control list.

References

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):92236.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):120.

Geyer, C. J. (2013). Asymptotics of Maximum Likelihood without the LLN or CLT or Sample Size Going to Infinity. Institute of Mathematical Statistics, Beachwood, OH.

Hughes, J. (2014). copcar: A flexible regression model for areal data. To appear in *Journal of Computational and Graphical Statistics*.

Kazianka, H. and Pilz, J. (2010). Copula-based geostatistical modeling of continuous and dis- crete data including covariates. *Stochastic Environmental Research and Risk Assessment*, 24(5):661673.

Madsen, L. (2009). Maximum likelihood estimation of regression parameters with spatially dependent discrete data. *Journal of Agricultural, Biological, and Environmental Statistics*, 14(4):375391.

Masarotto, G. and Varin, C. (2012). Gaussian copula marginal regression. *Electronic Journal of Statistics*, 6:15171549.

Nelson, R. B. (2006). An Introduction to Copulas. Springer, New York.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229231.

Varin, C. (2008). On composite marginal likelihoods. *AStA Advances in Statistical Analysis*, 92(1):128.