Published in final edited form as:

J Comput Graph Stat. 2015 September 16; 24(3): 733–755. doi:10.1080/10618600.2014.948178.

copCAR: A Flexible Regression Model for Areal Data

John Hughes

Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455

John Hughes: hughesj@umn.edu

Abstract

Non-Gaussian spatial data are common in many fields. When fitting regressions for such data, one needs to account for spatial dependence to ensure reliable inference for the regression coefficients. The two most commonly used regression models for spatially aggregated data are the automodel and the areal generalized linear mixed model (GLMM). These models induce spatial dependence in different ways but share the smoothing approach, which is intuitive but problematic. This article develops a new regression model for areal data. The new model is called copCAR because it is copula-based and employs the areal GLMM's conditional autoregression (CAR). copCAR overcomes many of the drawbacks of the automodel and the areal GLMM. Specifically, copCAR (1) is flexible and intuitive, (2) permits positive spatial dependence for all types of data, (3) permits efficient computation, and (4) provides reliable spatial regression inference and information about dependence strength. An implementation is provided by R package copCAR, which is available from the Comprehensive R Archive Network, and supplementary materials are available online.

Keywords

Composite likelihood; Copula; Distributional transform; Generalized linear model; Markov random field; Non-Gaussian data; Spatial confounding; Spatial regression

1. INTRODUCTION

Spatially aggregated, or areal, data are common in many fields, including forestry, marketing, image analysis, ecology, epidemiology, geography, and dentistry. Researchers in these fields are often interested in scientific explanation rather than prediction, in which case reliable, i.e., accurate and precise, spatial regression inference is important.

The two most often used models for areal regression are the automodel (Besag 1974) and the areal generalized linear mixed model (GLMM) (Besag, York and Mollié 1991), both of which have long histories and enjoy widespread popularity. But these models pose numerous analytical, statistical, and computational challenges. We develop a new areal model that addresses many of these challenges. The new model is called copCAR because it

SUPPLEMENTARY MATERIALS

R code: This file contains the R code used to apply copCAR to the Slovenia data. (copCAR.R)

Slovenia data: This file contains the Slovenia data. (Slovenia.R)

is copula-based and because it borrows the conditional autoregression (CAR) from the areal GLMM.

In Section 2 we review the automodel and the areal GLMM and describe their drawbacks. In Section 3 we develop the copCAR model and discuss its advantages and limitations. In Section 4 we recommend three approaches to frequentist inference for copCAR. In Section 5 we describe how to efficiently fit copCAR to data. In Section 6 we present the results of a simulation study designed to reveal the behavior of the copCAR estimators for realistic sample sizes. In Section 7 we apply copCAR and the areal GLMM to Slovenian stomach cancer data (Zadnik and Reich 2006; Reich, Hodges and Zadnik 2006).

2. AREAL MODELS

In this section we provide the motivation for copCAR by reviewing the automodel and the areal GLMM and describing the difficulties they pose. Although the automodel and the areal GLMM can be distinguished by their dependence components (one is direct, the other hierarchical), the two models share the smoothing approach, which is to say they induce spatial dependence by adding a spatial term to the linear predictor of the classical GLM. (Although this article uses GLM terminology, these ideas also apply to similar models for which the response does not follow an exponential family distribution, e.g., beta regression models.) Although smoothing models have proved quite successful, smoothing can be problematic. copCAR, by contrast, exploits the modularity of copula-based modeling, which allows us to model the marginal distributions and dependence structure separately before joining them by way of the probability integral transform (Sklar 1959). In our view, this is the simplest, most natural, and most flexible way to construct a multivariate distribution. And this approach allows copCAR to overcome many of the challenges faced by the smoothing models.

2.1 The Automodel

The automodel is a Markov random field (MRF) model (Kindermann and Snell 1980). That is, the model is specified in terms of full conditional distributions that correspond to a valid joint distribution if certain conditions are satisfied. MRF models were developed by physicists in the 1920s—in connection with the study of ferromagnetism (Ising 1925)—but did not make their way into the statistics literature until the late Julian Besag formulated the automodel in a seminal 1974 paper.

The automodel can be specified as follows. Let $G = (V, E)$ be the underlying graph, where $V = \{1, 2, \dots, n\}$ are the vertices and $E \subset V \times V$ are the edges of G . Each vertex of G corresponds to a region over which measurements have been aggregated—e.g., pixel, political district, province. And each edge of G represents the spatial adjacency of two such regions. We assume that G is undirected and free of loops and parallel edges.

Associate with the i th vertex the random variable Z_i , so that $Z = (Z_1, \dots, Z_n)'$ is the random field of interest. Then the transformed conditional means are

$$g\{\mathbb{E}(Z_i|\boldsymbol{\theta}, Z_{-i})\} = \mathbf{x}_i' \boldsymbol{\beta} + \eta \sum_{j:(i,j) \in E} (Z_j - \mu_j), \quad (1)$$

where g is a link function, $\boldsymbol{\theta} = (\boldsymbol{\beta}', \eta)'$ is the complete vector of parameters, $Z_{-i} = Z \setminus \{Z_i\}$ is the field with the i th observation excluded, $\mathbf{x}_i$ is a p -vector of spatial predictors for the i th areal unit, $\boldsymbol{\beta}$ is a p -vector of regression coefficients, η is a spatial dependence parameter, and μ_j is the independence expectation of Z_j , i.e., $\mu_j = \mathbb{E}(Z_j | \boldsymbol{\beta}, \eta = 0) = g^{-1}(\mathbf{x}_j' \boldsymbol{\beta})$. Note that $\eta > 0$ implies positive dependence, $\eta = 0$ implies spatial independence, and $\eta < 0$ implies spatial repulsion. In the sequel we take η to be non-negative since repulsion is not often observed in the phenomena to which these models are typically applied.

The second term in (1) is called the autocovariate. It is by means of this “device” that the automodel induces dependence. To see this, note that η measures the reactivity of Z_i to its neighbors, conditional on the large-scale structure represented by the neighbors’ independence expectations $\{\mu_j : (i, j) \in E\}$. As η increases, Z_i becomes more dependent on its neighbors and so less dependent on $\mathbf{x}_i' \boldsymbol{\beta}$. Eventually the dependence becomes so strong that the large-scale structure is overwhelmed by small-scale interactions, making recovery of $\boldsymbol{\beta}$ impossible.

Although the automodel has much to recommend it for certain types of data, the model presents various theoretical and computational difficulties in general. For example,

- the auto-Poisson and autonegative binomial models permit only negative spatial dependence (Kaiser and Cressie 1997; Møller 1999), which we do not expect to find in the phenomena to which these models are usually applied;
- each new spatial data type requires new analytical work (Kaiser and Cressie 2000);
- depending on the approach to inference, a change in the clique number of G may require re-derivation of the objective function (Kaiser and Cressie 2000);
- it may not be feasible to construct an automodel for a given type of data, e.g., incidence-prevalence data (Hoef and Jansen 2007); and
- due to intractable normalizing functions, maximum likelihood and rigorous Bayesian inference may be infeasible for larger datasets because expensive sampling is required (Geyer 1994; Møller, Pettitt, Reeves and Berthelsen 2006; Hughes, Haran and Caragea 2011).

We turn now to the areal GLMM, which addresses some of the automodel’s challenges but faces formidable difficulties of its own.

2.2 The Areal GLMM

The areal GLMM also induces dependence through smoothing, but instead of modeling dependence directly, as the automodel does, the mixed model is hierarchical with an autonormal second stage. The model was introduced by Besag et al. in 1991, and it has since been the most commonly used areal model, owing to its flexible specification, the

availability of the WinBUGS software application (Lunn, Thomas, Best and Spiegelhalter 2000), and the difficulties posed by the automodel.

The first stage of the areal GLMM is

$$g\{\mathbb{E}(Z_i|\beta, S_i)\} = x_i'\beta + S_i. \quad (2)$$

Distortion of the large-scale structure is accomplished by way of spatially dependent random effects $\mathbf{S} = (S_1, \dots, S_n)'$, which are assumed to have come from a zero-mean Gaussian Markov random field (GMRF) (Rue and Held 2005). That is, $\mathbf{S} \sim \mathcal{N}\{\mathbf{0}, (\tau\mathbf{Q})^{-1}\}$, where $\tau > 0$ is a smoothing parameter and $\mathbf{Q}$ is a precision matrix. Two forms for $\mathbf{Q}$ are commonly used: the proper CAR and the improper (or intrinsic) CAR (Besag and Kooperberg 1995), where CAR stands for conditional autoregression. The proper CAR can be specified as follows.

Let d_i be the degree of vertex i , let $\mathbf{D} = \text{diag}(d_1, \dots, d_n)$, let $\rho \in [0, 1)$, and let $\mathbf{A} = [1\{(i, j) \in E\}]$ be the adjacency matrix of G , where $1\{\bullet\}$ is the indicator function. Then the precision matrix for the proper CAR is $\mathbf{Q} = \mathbf{D} - \rho\mathbf{A}$. This marginal definition corresponds to the conditional distributions

$$S_i|\rho, \tau, \mathbf{S}_{-i} \sim \mathcal{N}\left(\frac{\rho}{d_i} \sum_{j: (i,j) \in E} S_j, \frac{1}{\tau d_i}\right),$$

which reveals that ρ is analogous to η in (1).

The restriction of ρ to $[0, 1)$ is not necessary to ensure that $\mathbf{Q}$ is nonsingular. It is sufficient to take $\rho \in (1/\min \lambda, 1/\max \lambda)$, where $\lambda = (\lambda_1, \dots, \lambda_n)'$ are the eigenvalues of $\mathbf{\Omega} = \mathbf{D}^{-1}\mathbf{A}$, the row-normalized adjacency matrix (Banerjee, Carlin and Gelfand 2004, p. 80). In short, the restriction ensures that $\mathbf{Q}$ is nonsingular and that the model has a sensible interpretation for the desired sphere of application (Gelfand and Vounatsou 2003).

The areal GLMM has at least the following three unappealing aspects.

- (1) It can be difficult to deal with $\mathbf{S}$. In a Bayesian framework, it is well known that a univariate Metropolis–Hastings algorithm for sampling from the posterior distribution of $\mathbf{S}$ leads to a slow mixing Markov chain because the coordinates of $\mathbf{S}$ exhibit strong *a posteriori* dependence. This has led to a number of approaches for updating $\mathbf{S}$ in a block(s). Constructing proposals for these updates is challenging, and the better mixing comes at the cost of increased running time per iteration—see, for instance, Knorr-Held and Rue (2002), Haran, Hodges and Carlin (2003), and Haran and Tierney (2012).
- (2) The spatial random effects are collinear with the fixed effects predictors. This spatial confounding causes variance inflation that may make important predictors appear insignificant. This was discovered by Clayton, Bernardinelli and Montomoli (1993) and proved (for the intrinsic CAR) by Reich et al.

(2006). The argument given by Reich et al. (2006) carries over essentially unchanged to the proper CAR model and to the generalized linear geostatistical model (Diggle, Tawn and Moyeed 1998).

- (3) The areal GLMM is generally unable to recover ρ (Banerjee et al. 2004). In a Bayesian setting, this problem is customarily handled by assigning ρ an informative prior distribution that encourages ρ 's posterior to be concentrated near 1 (Banerjee et al. 2004; Gelfand and Vounatsou 2003). This forces smoothing, which can improve prediction, but does not result in reliable inference for ρ .

The last two of these liabilities are the most troubling because they can lead to erroneous regression inference. Reich et al. (2006) and Hughes and Haran (2013) proposed reparameterized versions of the areal GLMM that alleviate the spatial confounding and also result in a faster mixing Markov chain. Hughes and Haran (2013) demonstrated the effectiveness of these reparameterizations for the intrinsic CAR, but it is not clear that the techniques can be carried over to the proper CAR in a way that permits efficient computation. By inducing dependence using a copula instead of an augmented linear predictor, copCAR eliminates problems (1) and (2) and is able to recover ρ .

3. THE COPCAR MODEL FOR AREAL DATA

During the last decade or so, copula regression models have been applied in several fields, most notably economics, finance, and insurance (Kolev and Paiva 2009). Only recently were copula models developed for spatial applications, and the copula approach has thus far been limited to geostatistical data (Madsen 2009; Kazianka and Pilz 2010). To our knowledge, copCAR is the first copula model for areal data.

The chief advantage of copula modeling is modularity. The dependence structure and the marginal distributions can be modeled separately (Sklar 1959) and then joined by way of the probability integral transform. This approach allows copCAR to overcome the problems associated with the smoothing models described in the previous section. (See Nelsen (2006) for an introduction to copulas.)

3.1 The Gaussian Copula

The focus of this article is the Gaussian copula, which can be constructed as follows. Let $\Phi_{\mathbf{R}}$ be the cdf of a multinormal random variable with mean $\mathbf{0}$ and correlation matrix $\mathbf{R}(\xi)$ (where ξ are dependence parameters), and let Φ be the standard normal cdf. Then the Gaussian copula is

$$C_{\mathbf{R}}(\mathbf{u}) = \Phi_{\mathbf{R}}\{\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_n)\}. \quad (3)$$

Note that $C_{\mathbf{R}}$ is defined using a correlation matrix rather than a covariance matrix—the variances, being properties of the marginal distributions, are irrelevant to the dependence structure and can be omitted.

3.2 The copCAR Specification

copCAR employs what we call the CAR copula, a Gaussian copula (or other suitable copula) based on the proper CAR described above. Recall that the proper CAR has precision matrix $\tau\mathbf{Q}$, where $\mathbf{Q} = \mathbf{D} - \rho\mathbf{A}$. Since a copula is scale free, we do not need τ , but omitting τ does not leave us with an inverse correlation matrix because the variances

$\sigma^2 = (\sigma_1^2, \dots, \sigma_n^2)' = \text{vecdiag}(\mathbf{Q}^{-1})$ are not equal to 1. We could rescale $\mathbf{Q}$ so that its inverse is a correlation matrix, i.e., we could construct a Gaussian copula using $\Sigma^{1/2}\mathbf{Q}\Sigma^{1/2}$, where $\Sigma = \text{diag}(\sigma^2)$. In fact, rescaling is necessary in the general case lest the model be unidentifiable with respect to the variances. For copCAR, however, rescaling is unnecessary because the variances σ^2 are not free parameters. The variances are entirely determined by $\mathbf{Q}$'s only dependence parameter, ρ , which is not a scale parameter. (Assunção and Krainski (2009) showed that ρ is best thought of as a range parameter.) Since using $\mathbf{Q}$ itself leads to an identifiable model, rescaling would slow computation but accomplish nothing. Thus copCAR employs the CAR correlation structure indirectly, by using $\mathbf{Q}$ along with the variances σ^2 . This leads to the CAR copula:

$$C_{\mathbf{Q}^{-1}}(\mathbf{u}) = \Phi_{\mathbf{Q}^{-1}}\{\Phi_{\sigma_1}^{-1}(u_1), \dots, \Phi_{\sigma_n}^{-1}(u_n)\}, \quad (4)$$

where Φ_{σ_i} denotes the distribution function of the normal distribution with mean 0 and variance σ_i^2 .

The model specification can be completed by pairing the CAR copula with a set of suitable marginal distributions for the outcomes. The copula and the marginals are linked by way of the probability integral transform. Specifically, if $\mathbf{Z} = (Z_1, \dots, Z_n)'$ are the observations, and $F_1, \dots, F_n$ are the desired marginal distribution functions, we have $Z_i = F_i^{-1}(U_i)$, where $\mathbf{U} = (U_1, \dots, U_n)'$ is a realization of the copula. In a spatial setting, Bernoulli or Poisson

marginals, with expectations $\{1 + \exp(-\mathbf{x}_i'\beta)\}^{-1}$ and $\exp(\mathbf{x}_i'\beta)$, respectively, are common, but there are many other options for the marginal specifications, e.g., extreme-value distributions, beta regression models, zero-inflated models, heavy-tailed distributions, etc.

3.3 Advantages and Limitations of copCAR

copCAR addresses the problems associated with the above mentioned smoothing models. Unlike the auto-model, copCAR is flexible—a new type of data or change in G does not necessitate new analytical work—and permits positive spatial dependence for any type of data. Unlike the areal GLMM, copCAR is able to recover ρ , as we will show in Section 6; is free of computationally burdensome random effects; and cannot exhibit spatial confounding. Additionally, copCAR is a marginal regression model, which is to say that estimates of regression coefficients have the same interpretation as in an independence model. Practitioners may find this less confusing than interpreting estimates conditional on random effects or neighboring observations.

Spatial confounding deserves further discussion. Let $\mathbf{P}$ be the orthogonal projection onto $\mathcal{C}(\mathbf{X})$, where $\mathbf{X}$ is the design matrix. Now, spectrally decompose the operators $\mathbf{P}$ and $\mathbf{I} - \mathbf{P}$

to acquire orthogonal bases $\mathbf{K}$ and $\mathbf{L}$ for $\mathcal{C}(\mathbf{X})$ and $\mathcal{C}(\mathbf{X})^\perp$, respectively. These bases allow us to rewrite (2) as

$$g\{\mathbb{E}(Z_i|\beta, S_i)\} = \mathbf{x}_i'\beta + S_i = \mathbf{x}_i'\beta + \mathbf{k}_i'\gamma + \mathbf{l}_i'\delta,$$

where γ and δ are random coefficients (Reich et al. 2006). This form exposes the source of the areal GLMM's spatial confounding: $\mathbf{K}$ and $\mathbf{X}$ have the same column space. This cannot happen for copCAR because copCAR uses the “bare” linear predictor $\mathbf{x}_i'\beta$.

Although copCAR is an appealing alternative to the automodel and the areal GLMM, it does have some limitations qua spatial model. copCAR obviously cannot capture non-Gaussian dependence, e.g., tail dependence or asymmetric dependence. But asymmetry could easily be accommodated by incorporating $\mathbf{Q}$ into a non-central χ^2 copula (Bárdossy and Li 2008). And it would be straightforward to accommodate tail dependence by using $\mathbf{Q}$ to construct an elliptical copula such as a t copula (Demarta and McNeil 2005), for example. Another attractive choice is the skew- t copula, which can accommodate both asymmetric and tail dependence (Smith, Gan and Kohn 2012). Neither the t nor the skew- t copula would permit spatial independence, however, which some practitioners may find unacceptable, or at least unappealing.

A second potential drawback has to do with identifiability. When the marginal distributions are discrete, the joint distribution is uniquely defined only on the support of the marginals, and the dependence between a pair of random variables depends on the marginal distributions as well as on the copula. Genest and Neslehova (2007) described the implications of this and warned that, for discrete data, “modeling and interpreting dependence through copulas is subject to caution.” But Genest and Neslehova (2007) go on to say that copula parameters can still be interpreted as dependence parameters, and estimation of copula parameters is often possible using fully parametric likelihood methods. It is precisely such methods that we recommend in the next section, and the simulation results in Section 6 suggest that copCAR's parameters can be recovered using those methods.

4. FREQUENTIST INFERENCE FOR COPCAR

In this section we recommend three approaches—continuous extension (CE), composite marginal likelihood (CML), and distributional transform (DT)—to frequentist inference for copCAR with discrete margins. We begin by describing maximum likelihood inference for copCAR with continuous margins, since the approximate likelihoods for the CE and DT approaches are reminiscent of the true likelihood for continuous data.

4.1 Continuous Marginals

Maximum likelihood inference is straightforward for Gaussian copula models with continuous marginal distributions. To construct the relevant likelihood for copCAR, we begin with the density for the CAR copula, which can be obtained by applying the chain rule to (4):

$$c_{\mathbf{Q}^{-1}}(\mathbf{u}) \propto |\mathbf{Q}|^{1/2} |\Sigma|^{1/2} \exp \left\{ -\frac{1}{2} \mathbf{y}' (\mathbf{Q} - \Sigma^{-1}) \mathbf{y} \right\},$$

where $\mathbf{y} = (y_1, \dots, y_n)' = (\Phi_{\sigma_1}^{-1}(u_1), \dots, \Phi_{\sigma_n}^{-1}(u_n))'$ and $\Sigma = \text{diag}(\sigma^2)$. Now, given observations z_i with continuous marginal cdfs F_i and corresponding pdfs f_i , the likelihood of the parameters θ given the data z is the product of the copula density and the marginal densities:

$$L(\theta|z) \propto c_{\mathbf{Q}^{-1}}\{F_1(z_1), \dots, F_n(z_n)\} \prod_{i=1}^n f_i(z_i).$$

Notice that the argument to the copula density is the vector of probability integral transformed observations, which are standard uniform random variables, as required. This implies the copCAR log likelihood

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

where $y_i = \Phi_{\sigma_i}^{-1}\{F_i(z_i)\}$. This log likelihood can be optimized to arrive at the maximum likelihood estimate $\hat{\theta}$ of θ .

4.2 Discrete Marginals

When the marginal distributions are discrete, the likelihood with respect to counting measure is

$$L(\theta|z) = \sum_{j_1=0}^1 \dots \sum_{j_n=0}^1 (-1)^k C_{\mathbf{Q}^{-1}}(u_{1j_1}, \dots, u_{nj_n}), \quad (5)$$

where $k = \sum_{i=1}^n j_i$, $u_{i0} = F_i(z_i)$, and $u_{i1} = \lim_{z \nearrow z_i} F_i(z) = F_i(z_i^-) = F_i(z_i - 1)$. The last equality holds when the marginals have integer support.

Unless n is quite small, computation of (5) is infeasible because the multinomial cdf is unstable in high dimensions and because the sum contains 2^n terms. To put this last fact in perspective, consider that evaluating the true likelihood for a dataset not much larger than the cancer dataset analyzed in Section 7 would require approximately the same number of operations as there are atoms in the known universe.

The following subsections describe three computationally feasible approaches to inference for copCAR with discrete margins. The first approach is a form of Monte Carlo maximum likelihood. We will see that this approach does not scale as well as the others, but, because it is exact (up to Monte Carlo error), the approach can serve as a standard against which to judge the other two approaches, each of which entails a misspecification.

The Continuous Extension—The continuous extension approach to maximum likelihood inference for Gaussian copula models with discrete marginals was developed by Madsen (2009). The approach gets its name from a technique whereby a discrete random variable is transformed to a continuous one by introducing an auxiliary random variable supported on the unit interval (Denuit and Lambert 2005).

To see how this can be accomplished, first suppose that $Z \sim F$ is a discrete random variable, and let f be the mass function corresponding to F . Let W be a continuous random variable supported on the unit interval, and suppose that W has distribution function G , density function g , and is independent of Z . Then the continuation of Z is the continuous random variable $Z^* = Z + (W - 1)$. Denuit and Lambert (2005) showed that Z^* has distribution function $F^*(z) = F([z]) + G(z - [z])f([z] + 1)$ and density function $f^*(z) = g(z - [z])f([z] + 1)$, where $[\bullet]$ returns the integer part of its argument. If we take W to be standard uniform, $Z^* = Z - W$ and the distribution and density functions simplify to $F^*(z) = F([z]) + (z - [z])f([z] + 1)$ and $f^*(z) = f([z] + 1)$, respectively.

For copCAR we continue $\mathbf{Z} = (Z_1, \dots, Z_n)'$ using n independent standard uniforms $\mathbf{W} = (W_1, \dots, W_n)'$ and form the expected likelihood

$$L(\boldsymbol{\theta}|z) \propto \mathbb{E}_{\mathbf{W}} \left[|\mathbf{Q}|^{1/2} \sum |\cdot|^{1/2} \exp \left\{ -\frac{1}{2} \mathbf{y}^*{}' (\mathbf{Q} - \sum^{-1}) \mathbf{y}^* \right\} \prod_{i=1}^n f_i(z_i) \right],$$

where $\mathbf{y}^* = (\Phi_{\sigma_1}^{-1}\{F_1^*(z_1^*)\}, \dots, \Phi_{\sigma_n}^{-1}\{F_n^*(z_n^*)\})'$. Using a result proved by Madsen (2009), one can show that this expectation is equal to the true likelihood given in (5).

We approximate the expectation using a sample-based approach. Let m be a positive integer, and simulate a vector of independent standard uniforms $\mathbf{W}_j = (W_{j,1}, \dots, W_{j,n})'$ for $j = 1, 2, \dots, m$. Then approximate the expected likelihood using

$$L_{\text{CE}}(\boldsymbol{\theta}|z) = \frac{1}{m} \sum_{j=1}^m |\mathbf{Q}|^{1/2} \sum |\cdot|^{1/2} \exp \left\{ -\frac{1}{2} \mathbf{y}_j^*{}' (\mathbf{Q} - \sum^{-1}) \mathbf{y}_j^* \right\} \prod_{i=1}^n f_i(z_i), \quad (6)$$

where $\mathbf{y}_{j,i}^* = \Phi_{\sigma_i}^{-1}\{F_i^*(z_i - w_{j,i})\}$. This approximate likelihood can then be optimized to arrive at an approximate maximum likelihood estimate $\hat{\boldsymbol{\theta}}_{\text{CE}}$ of $\boldsymbol{\theta}$.

The Distributional Transform—When the marginal distributions have fairly smooth cdfs, we recommend that the likelihood be approximated using the distributional transform. This approach was proposed by Kazianka and Pilz (2010) for fitting Gaussian copula geostatistical models.

It is well known that if $Z \sim F$ is continuous, $F(Z)$ has a standard uniform distribution. But if Z is discrete, $F(Z)$ tends to be stochastically larger, and $F(Z^-)$ tends to be stochastically smaller, than a standard uniform random variable. This can be remedied by stochastically “smoothing” F at its jumps, a technique that goes at least as far back as Ferguson (1967), who used it in connection with hypothesis tests. More recently, the DT has been applied to

stochastic ordering (Rüschendorf 1981), conditional value at risk (Burgert and Rüschendorf 2006), and the extension of limit theorems for the empirical copula process to general distributions (Rüschendorf 2009), for example.

Let $W \sim \mathcal{U}(0, 1)$, and suppose that $Z \sim F$ and is independent of W . Then the distributional transform $G(W, Z) = (1 - W)F(Z^-) + WF(Z)$ follows a standard uniform distribution and $F^{-1}\{G(W, Z)\}$ follows the same distribution as Z . See Rüschendorf (2009) for a proof.

Turning back to the problem at hand, the DT-based approximate likelihood for copCAR models with discrete marginals can be developed as follows. For each $i \in \{1, \dots, n\}$, let

$$G_i(W_i, Z_i) = (1 - W_i)F_i(Z_i^-) + W_iF_i(Z_i),$$

where the W_i are standard uniform random variables and are independent of one another and of the Z_i . Now put

$$u_i = \mathbb{E}\{G_i(W_i, Z_i) | Z_i = z_i\} = \{F_i(z_i^-) + F_i(z_i)\}/2 = (u_{i0} + u_{i1})/2. \quad (7)$$

Then the approximate likelihood for copCAR is

$$L_{\text{DT}}(\boldsymbol{\theta} | \mathbf{z}) = c_{\mathbf{Q}^{-1}}(u_1, \dots, u_n) \prod_{i=1}^n f_i(z_i),$$

which implies the approximate log likelihood

$$\ell_{\text{DT}}(\boldsymbol{\theta} | \mathbf{z}) = \frac{1}{2} \log |\mathbf{Q}| + \frac{1}{2} \log \left| \sum \right| - \frac{1}{2} \mathbf{y}' (\mathbf{Q} - \sum^{-1}) \mathbf{y} + \sum_{i=1}^n \log f_i(z_i), \quad (8)$$

where $y_i = \Phi_{\sigma_i}^{-1}(u_i)$. Optimization of (8) yields $\hat{\boldsymbol{\theta}}_{\text{DT}}$.

Kazianka and Pilz (2010) noted that the DT-based approximation performs well when the cdfs are not too rough. A subsequent investigation by Kazianka (2013) revealed that the marginal variance(s) is a suitable indicator of the level of discretization, and hence the performance of the DT-based approximation. Specifically, a smaller variance(s) leads to worse performance. Our simulation studies (Section 6) show that the approach performs well for copCAR when the marginals are Poisson with small means, and so we should expect the approach to perform well for copCAR analyses of similar types of data, e.g., negative binomial data. Other simulation studies, the results of which will not be detailed here, confirmed this conjecture. Specifically, the DT-based approximation performs well for copCAR with negative binomial marginals, but performance is unacceptable when the marginals are Bernoulli.

Composite Marginal Likelihood—When the marginal cdfs are rather rough (e.g., Bernoulli), we recommend a composite marginal likelihood approach (Varin 2008), where the objective function is a product of pairwise likelihoods:

$$L_{\text{CML}}(\boldsymbol{\theta}|z) = \prod_{\substack{i,j \in \{1, \dots, n\} \\ i \neq j}} \sum_{j_1=0}^1 \sum_{j_2=0}^1 (-1)^k C_{ij}(u_{ij_1}, u_{jj_2}),$$

where C_{ij} denotes the bivariate Gaussian copula with covariance matrix

$$\mathbf{V}^{ij} = \begin{pmatrix} \sigma_i^2 & (\mathbf{Q}^{-1})_{ij} \\ (\mathbf{Q}^{-1})_{ij} & \sigma_j^2 \end{pmatrix}.$$

This implies the log composite likelihood

$$\ell_{\text{CML}}(\boldsymbol{\theta}|z) = \sum_{\substack{i \in \{1, \dots, n-1\} \\ j \in \{i+1, \dots, n\}}} \log \left\{ \sum_{j_1=0}^1 \sum_{j_2=0}^1 (-1)^k \Phi_{\mathbf{V}^{ij}}(y_{ij_1}, y_{jj_2}) \right\}, \quad (9)$$

where $y_{\bullet 0} = \Phi_{\sigma_{\bullet}}^{-1}\{F_{\bullet}(z_{\bullet})\}$ and $y_{\bullet 1} = \Phi_{\sigma_{\bullet}}^{-1}\{F_{\bullet}(z_{\bullet} - 1)\}$. Optimization of (9) yields $\hat{\boldsymbol{\theta}}_{\text{CML}}$. (For more information about composite likelihood methods, see, e.g., the seminal paper by Lindsay (1988), or the more recent overview by Varin, Reid and Firth (2011).)

Computing Confidence Intervals—A preliminary simulation study revealed that $\hat{\beta}_{\text{CE}}$, $\hat{\beta}_{\text{CML}}$, and $\hat{\beta}_{\text{DT}}$ tend to be approximately normally distributed, whereas $\hat{\rho}_{\text{CE}}$, $\hat{\rho}_{\text{CML}}$, and $\hat{\rho}_{\text{DT}}$ tend to be left skewed when ρ is close to 1. This implies that asymptotic inference for ρ tends to result in poor coverage rates. This can be avoided by using a parametric bootstrap, but a parametric bootstrap is rather burdensome computationally, especially for the continuous extension.

Luckily, a simple reparameterization yields an approximately normally distributed estimator because the objective function for the reparameterized model is approximately quadratic with constant Hessian (Geyer 2013). This is true for all three procedures, at least for the realistic scenarios considered in this article. Specifically, for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \Phi^{-1}(\rho))'$, we have

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

where $\mathcal{J}_{\bullet}$ is the appropriate Fisher information matrix and $\mathcal{J}_{\bullet}$ is the variance of the score:

$$\mathcal{J}_{\bullet}(\boldsymbol{\theta}) = \mathbb{V} \nabla \ell_{\bullet}(\boldsymbol{\theta} | \mathbf{Z}).$$

Note that the asymptotic covariance matrices for the CML and DT estimators are Godambe information matrices (Godambe 1960) because ℓ_{CML} and ℓ_{DT} are misspecified.

A sensible estimator of $\mathcal{J}_\bullet$ is the observed Fisher information matrix $\hat{\mathcal{J}}_\bullet(\boldsymbol{\theta}) = \mathcal{H}_\bullet(\hat{\boldsymbol{\theta}}_\bullet)$, where $\mathcal{H}_\bullet$ denotes the Hessian matrix $[\nabla^2 \ell_\bullet(\boldsymbol{\theta} | \mathbf{Z})]$. This matrix can be approximated efficiently as a side effect of optimization. Estimation of $\mathcal{J}_\bullet$ is more challenging. One appealingly general approach is window subsampling (Heagerty and Lumley 2000), but the quality of the resulting estimate can be sensitive to the manner in which internal replicates are defined. We recommend that $\mathcal{J}_\bullet$ be estimated using a parametric bootstrap instead, i.e, our estimator of $\mathcal{J}_\bullet$ is

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

where b is the bootstrap sample size and the $\mathbf{Z}^{(k)}$ are datasets simulated from copCAR at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_\bullet$. This approach performs well, as our simulation results show, and is considerably more efficient computationally than a “full” bootstrap (it is much faster to approximate the score than to optimize the objective function). The procedure can be made even more efficient through parallelization. We recommend the snow package (Tierney, Rossini and Li 2009) for R (Ihaka and Gentleman 1996), which makes embarrassingly parallel computation embarrassingly easy to do.

5. COMPUTING ISSUES

In this section we will show how to do reliable and efficient computation for copCAR. The recipe for efficiency is threefold: for CE and DT estimation, (1) exploit the sparsity of $\mathbf{Q}$ to compute $|\mathbf{Q}|$ quickly, and (2) employ matrix identities that allow the CAR variances to be approximated quickly and accurately; and (3) use G ’s adjacency structure to speed up maximum CML estimation. The use of these techniques makes copCAR suitable for the analyses of larger areal datasets. Additionally, we describe how to ensure the accuracy and numerical stability of the continuous extension.

5.1 Computing $|\mathbf{Q}|$ Efficiently

Numerical methods for sparse matrices can be used to compute $|\mathbf{Q}|$ quickly. Let $\mathbf{C}$ be the lower Cholesky triangle of $\mathbf{Q}$, so that $\mathbf{Q} = \mathbf{C}\mathbf{C}'$. Then $|\mathbf{Q}| = |\mathbf{C}||\mathbf{C}'| = |\mathbf{C}|^2$, which implies that

$$\frac{1}{2} \log |\mathbf{Q}| = \sum_{i=1}^n \log \mathbf{C}_{ii}. \quad (10)$$

The righthand side of (10) can be computed efficiently because $\mathbf{C}$ can be computed efficiently after $\mathbf{Q}$ has been reordered to reduce its bandwidth. Also note that $\mathbf{C}$ needs to be computed just once—the structure of $\mathbf{C}$ depends on the sparsity structure of $\mathbf{Q}$ and not on ρ , and so $\mathbf{C}$ can simply be updated to reflect a change in ρ (Ng and Peyton 1993).

We use the `spam` package (Furrer and Sain 2010) for R to do sparse matrix computations. The `chol` and `update` functions of package `spam` perform the fast Cholesky decomposition and updating described in the preceding paragraph. (See Rue and Held (2005) for more information about the use of sparse matrix techniques in statistical computing.)

5.2 Approximating the CAR Variances

Since computing σ^2 requires inversion of $\mathbf{Q}$ (the relevant `spam` function is `chol2inv`), it would seem that using the CAR copula given in (4) must leave us unable to fully exploit the sparsity of $\mathbf{Q}$. This is not the case, however, because the variances in (4) can be replaced with approximations so that $\mathbf{Q}$ need not be inverted. Here we present an algorithm for computing the approximate variances quickly and with arbitrary accuracy. This scheme has a greater upfront cost than inversion of $\mathbf{Q}$, but the amortized cost of approximating σ^2 is much lower. And the difference between the approximate CAR copula and (4) is negligible except with respect to computational complexity.

First, choose a maximum value for ρ , say $\rho^{\max}$, and a tolerance $\epsilon > 0$. Then if k is chosen such that

$$k > \frac{\log(1 - \rho^{\max}) + \log \epsilon - \log c(G)}{\log \rho^{\max}} - 1,$$

where

$$c(G) = \max_i \left\{ \frac{(\Omega^2)_{ii}}{d_i} \right\},$$

the approximate variances $\tilde{\sigma}^2 = (\tilde{\sigma}_1^2, \dots, \tilde{\sigma}_n^2)'$ will satisfy $(\sigma_i^2 - \tilde{\sigma}_i^2) < \epsilon$ for all i and for every $\rho \in [0, \rho^{\max}]$. See the appendix for a proof of this claim and an efficient algorithm for finding the smallest possible value of k .

For a given value of ρ , the approximate variances are $\tilde{\sigma}^2 = \mathbf{D}^{-1} \mathbf{M} \boldsymbol{\rho}$, where $\mathbf{M}_{\cdot 1} = \mathbf{1}$; $\mathbf{M}_{\cdot j} = \text{vecdiag}(\Omega^{j-1})$ for $j = 2, \dots, k+1$; and $\boldsymbol{\rho} = (1, \rho, \rho^2, \dots, \rho^k)'$. The diagonal of Ω^{j-1} can be computed quickly using the eigendecomposition of Ω . Let $\Omega = \Gamma \Lambda \Gamma^{-1}$, where $\Lambda = \text{diag}(\boldsymbol{\lambda})$ and Γ has as its columns the eigenvectors of Ω . Then

$$\text{vecdiag}(\Omega^{j-1}) = \{\Gamma \circ (\Gamma^{-1})'\} \boldsymbol{\lambda}^{j-1} \quad (11)$$

(Horn and Johnson 1994, p. 305), where $\circ$ denotes the Hadamard product and $\boldsymbol{\lambda}^{j-1} = (\lambda_1^{j-1}, \dots, \lambda_n^{j-1})'$. This simple yet somewhat obscure identity is crucial for making computation of $\tilde{\sigma}^2$ much more efficient than computation of σ^2 . Naive computation of the desired vectors is rather expensive, whereas the use of (11) allows $\mathbf{M}$ to be computed efficiently, even for large datasets. For example, with $\rho^{\max} = 0.999$, $\epsilon = 0.01$, and G the 50×50 square lattice, k is equal to 4,100. Under these circumstances, a pure R implementation of (11) can build $\mathbf{M}$ in approximately eleven minutes on a 2.8 GHz quad-core Intel Xeon processor. A C++ implementation based on the Armadillo library (Sanderson 2010) built $\mathbf{M}$ in just eight minutes. (Armadillo code can be called from R using the `Rcpp` and `RcppArmadillo` packages (Eddelbuettel and Francois 2011).)

In our simulation studies, and for the analysis of the cancer data, we used $\rho^{\max} = 0.999$ and $\epsilon = 0.01$. Not once did $\hat{\rho}_{\bullet}$ equal $\rho^{\max}$ for any of the thousands of datasets analyzed. This is because the estimator $\hat{\rho}_{\bullet}$ becomes very precise as ρ approaches 1. A practitioner who wishes to be more conservative might use, say, 0.9999 instead, but this would be unnecessary in most cases and would increase the computational burden.

The value of 0.01 for ϵ is already conservative because the proof given in the appendix does not provide a tight upper bound for the discrepancy between the true and approximate variances. For the scenarios described in this article, using $\epsilon = 0.01$ leads to approximate variances that are numerically indistinguishable from the true variances.

A thoughtful referee suggested that, in some scenarios, it might be more efficient to simply invert $\mathbf{Q}$ by inverting $\mathbf{C}$. This is true for small datasets, but inversion does not scale well. For the graph used in our simulation study, i.e., the square lattice with n vertices, matrix inversion has a running time in $O(n^2)$. Our approach, which exploits the special structure of the CAR precision matrix (see the proof for details), has a running time in $O(n)$. This advantage can reduce the running time considerably when n is large (greater than 1,000, say).

5.3 Fast Maximum CML Estimation

The form of ℓ_{CML} given in (9) requires four evaluations of the bivariate normal cdf for each of the $n(n-1)/2$ pairs of observations. This computation is rather expensive even for fairly small samples.

In a spatial setting we can expect a pair of nearby observations to carry more information about dependence than a pair of more distant observations. Others have found, in a variety of contexts, that retaining the contributions to the CML made by more distant pairs of observations decreases not only the computational efficiency of the procedure but also the statistical efficiency of the estimator (Varin and Vidoni 2009; Apanasovich, Ruppert, Lupton, Popovic, Turner, Chapkin and Carroll 2008). We investigated this by simulation and found that including second-order, not to mention third-order, neighbors in the copCAR CML does yield a less efficient estimator. Hence, we allow only pairs of adjacent observations to contribute to the copCAR CML. This means replacing (9) with

$$\ell_{\text{CML}}(\boldsymbol{\theta}|z) = \sum_{\substack{i,j:(i,j) \in E \\ i < j}} \log \left\{ \sum_{j_1=0}^1 \sum_{j_2=0}^1 (-1)^k \Phi_{\mathbf{v}_{ij}}(y_{ij_1}, y_{ij_2}) \right\}. \quad (12)$$

If thoughtfully implemented, optimization of (12) is efficient enough to permit the analyses of larger areal datasets. For example, Brenton Kenkel's pbivnorm package (Kenkel 2012) for R offers a vectorized implementation of the bivariate normal cdf. The vectorization minimizes the number of calls to the underlying Fortran workhorse function. Through further specialization of Kenkel's R wrapper (e.g., removal of many error-checking statements that are unnecessary in our setting), we were able to speed up the evaluation of (12) by a factor of ten relative to a more naive implementation.

5.4 Stabilizing the Continuous Extension

Computation of L_{CE} is numerically stable for smaller sample sizes but becomes unstable as the sample size increases. This is because the quadratic form $-v_j = \mathbf{y}_j^{*'}(\mathbf{Q} - \sum^{-1})\mathbf{y}_j^*$ can be so large that computation of $\exp(v_j)$ results in underflow. If the sample size is large enough, computation of $\exp(v_j)$ may cause underflow for every j , in which case application of the log function then causes overflow. And even if this worst case scenario does not obtain, many of the exponentials may underflow, which means the procedure is not able to benefit fully from the current Monte Carlo sample size m .

This instability can be alleviated, using the so called log-sum-exp trick, as follows. Suppose our aim is to compute $\log(e^a + e^b)$, where $a, b < 0$ are potentially large in absolute value. Choose an appropriate c , and rewrite the expression as $\log\{(e^a + e^b)e^{-c}e^c\} = \log(e^{a-c} + e^{b-c}) + c$. Choosing $c = \max\{a, b\}$ sets one of the exponents to 0 and shifts the other toward 0, which reduces the likelihood of underflow. The resulting stabilized form of the log likelihood is

$$\ell_{CE}(\boldsymbol{\theta}|z) = \frac{1}{2}\log|\mathbf{Q}| + \frac{1}{2}\log\left|\sum\right| + \sum_{i=1}^n \log f_i(z_i) + \log \sum_{j=1}^m \exp\left\{\frac{1}{2}(u_j - s)\right\},$$

where $s = \max\{v_1, \dots, v_m\}$. The added computational cost is in $O(m)$ and so is dominated by the cost of evaluating the quadratic forms.

5.5 Choosing the Monte Carlo Sample Size for the Continuous Extension

One way to choose m is to compute Monte Carlo standard errors (Flegal, Haran and Jones 2008). In the case of $\hat{\rho}_{CE}$, for example, observe that

$\mathbb{V}\hat{\rho}_{CE} = \mathbb{E}\{\mathbb{V}(\hat{\rho}_{CE}|\mathbf{W})\} + \mathbb{V}\{\mathbb{E}(\hat{\rho}_{CE}|\mathbf{W})\}$, where $\mathbf{W}$ denotes the collection of the $\mathbf{W}_j$. The second term in this sum is the Monte Carlo variance of $\hat{\rho}_{CE}$, and it can be estimated by computing $\hat{\rho}_{CE}^{(k)}$ for many different $\mathbf{W}^{(k)}$ ($k = 1, \dots, b$) and taking the sample standard deviation of the estimates:

$$\text{mcse}(\hat{\rho}_{CE}) = \left\{ \frac{1}{b-1} \sum_{k=1}^b \left(\hat{\rho}_{CE}^{(k)} - \hat{\mu} \right)^2 \right\}^{1/2},$$

where $\hat{\mu}$ is the sample mean $b^{-1} \sum_{k=1}^b \hat{\rho}_{CE}^{(k)}$ (Madsen and Fang 2011). If the Monte Carlo standard error of $\hat{\rho}_{CE}$ is small relative to the sample mean, i.e., if the estimated coefficient of variation $\text{cv}(\hat{\rho}_{CE}) = \text{mcse}(\hat{\rho}_{CE})/\hat{\mu}$ is sufficiently small, the current value of m is sufficiently large.

Another approach is to fit a sequence of models for values of m that range over several orders of magnitude. We used this method while preparing to conduct our simulation studies and found that $m = 100$ is generally sufficient when the dependence is weak to moderate in strength. When the dependence is strong, however, m should be 1,000 or even 10,000. Thus

the CE approach scales poorly relative to the CML and DT approaches, even though the latter two approaches require bootstrap estimation of $\mathcal{J}$. For example, we took G to be the 50×50 square lattice, and simulated from copCAR with Poisson marginals (see below for details). Analyzing such data using the CE approach with $m = 10,000$ took about 225 minutes, on average, while CML and DT analyses (with a bootstrap sample size of 500 and using five CPU cores in parallel) averaged approximately 75 minutes and 30 minutes, respectively.

6. SIMULATION STUDY

To determine the finite-sample behavior of the CE, CML, and DT estimators for copCAR, we conducted an extensive simulation study. For this study we simulated data on the 20×20 and 30×30 square lattices, where the coordinates of the vertices were restricted to the unit square. We chose to focus on these two lattice sizes for two reasons: (1) most areal datasets have a few hundred observations, and so the 20×20 lattice represents a realistic scenario, and (2) the 30×30 lattice is large enough by comparison to reveal large-sample behavior.

The marginal distributions were Poisson with means $\exp(x_i + y_i)$, where x_i and y_i are the coordinates of vertex i . Each coordinate was treated as a separate spatial predictor, and so $\beta = (\beta_1, \beta_2)' = (1, 1)'$. The resulting means range from 1 to approximately 7.4 and increase in the direction $(1, 1)'$.

We used three values for ρ : 0.8, 0.975, and 0.995. These correspond to moderate, strong, and very strong dependence, respectively. This is illustrated in Figure 1. The dots in Panel A mark the correlations between the first CAR variable and every other variable in the same row of the 20×20 lattice. The level plots in Panel B show simulated realizations of the CAR copula, one for each value of ρ .

For each lattice size and value of ρ we analyzed 1,000 simulated datasets, 6,000 datasets altogether. We used a BFGS quasi-Newton method (Byrd, Lu, Nocedal and Zhu 1995) to optimize the objective functions. For the CE and DT procedures we chose $\rho^{\max} = 0.999$ and constrained the estimate of ρ to $(0, \rho^{\max})$. For the CE approach we used $m = 100$ for $\rho = 0.8$, and $m = 1,000$ for the two larger values of ρ .

We obtained approximate 95% confidence intervals for the coordinates of θ using the approach described near the end of Section 4, with a bootstrap sample size of 500 for estimation of $\mathcal{J}_{\text{CML}}$ and $\mathcal{J}_{\text{DT}}$.

Tables 1 and 2 show mean estimates, mean standard errors, simulation standard deviations, and coverage rates for the various scenarios. We see that the three approaches perform well—and comparably, for the most part. As expected, $\hat{\theta}_{\text{CE}}$ is the most efficient of the three estimators, and $\hat{\theta}_{\text{CML}}$ is the least efficient. What is perhaps surprising is the excellent performance of the DT estimator: $\hat{\beta}_{\text{DT}}$ is nearly as efficient as $\hat{\beta}_{\text{CE}}$, and $\hat{\rho}_{\text{DT}}$ is more efficient than $\hat{\rho}_{\text{CE}}$.

And we see that copCAR, unlike the areal GLMM, is able to recover ρ . In fact, $\hat{\rho}_\bullet$ becomes considerably more precise as ρ approaches 1. This is a useful property—when ρ is close to 1, we can be more certain of that fact and so more certain that the data are strongly dependent, in which case regression inference is less likely to be reliable.

7. COPCAR APPLIED TO THE SLOVENIA DATA

The plot in Figure 2 shows stomach cancer incidence data for Slovenia for the period 1995–2001. These data were first analyzed by Zadnik and Reich (2006), who investigated the relationship between stomach cancer incidence and socioeconomic factors. Then Reich et al. (2006) reconsidered the data while exploring the spatial confounding of the areal GLMM. The plotted values are observed standardized incidence ratios (SIR), i.e., the ratio of the observed to the expected (O_i/E_i) number of cases, for 194 municipalities.

The covariate of interest is each municipality's socioeconomic status (SE) as determined by Slovenia's Institute of Macroeconomic Analysis and Development. Each status takes one of five ordered values. Visual inspection reveals a negative relationship between SIR and SE—western Slovenia tends to exhibit low SIR and high SE, while eastern Slovenia tends to exhibit high SIR and low SE.

To these data we fitted one nonspatial and two spatial Poisson regressions. For the areal GLMM with proper CAR, the model is

$$\log \mathbb{E}O_i = \log E_i + \alpha + \beta \text{ SE} + S_i,$$

where α is an intercept and β is the fixed effect for SE. (Note that SE was standardized prior to the analyses.) The corresponding expressions for the classical GLM and copCAR are the same except S_i does not appear.

Since the areal GLMM is usually applied in a Bayesian setting, we fitted the areal GLMM using MCMC for Bayesian inference. We chose uninformative prior distributions for $(\alpha, \beta)'$ and ρ : $\mathcal{N}(\mathbf{0}; 1,000,000 \mathbf{I})$ and $\mathcal{U}(0, 1)$, respectively. And the prior for τ (which is present only for the areal GLMM) was gamma with shape parameter 0.5 and scale parameter 2,000 (Kelsall and Wakefield 1999). This prior is appealing because it corresponds to the prior assumption that the fixed effects are sufficient to explain the data, and because it discourages artifactual spatial structure in the posterior. We used a Metropolis–Hastings random walk to draw a sample of size 100,000. The MCMC estimates stabilized after approximately 75,000 iterations. The effective sample size for β was just under 1,700, and the estimated coefficients of variation were smaller than 0.02.

For copCAR we used $m = 1,000$ (for CE), $\rho^{\max} = 0.999$ (for CE and DT), and a bootstrap sample size of 500 (for CML and DT).

The results of the analyses are shown in Table 3. copCAR found SE to be a significant predictor of SIR, while the areal GLMM analysis indicates no regression. The copCAR estimates of ρ were small, which suggests that SE accounts for nearly all of the spatial clustering in the response. The copCAR intervals are appreciably wider than the intervals

from the nonspatial fit, however, which implies that a spatial model is indicated for these data. The areal GLMM had to find a much larger degree of spatial dependence in order to explain the data. This is presumably due to confounding, which would explain the much wider intervals for α and β .

These results are comparable to those given by Reich et al. (2006). They applied both the traditional ICAR GLMM and their reparameterized ICAR GLMM, the RHZ model, to the Slovenia data. copCAR produced inference very similar to that produced by the RHZ model. And the proper CAR GLMM and the traditional ICAR GLMM, both of which are known to be spatially confounded, resulted in 95% HPD intervals for β that cover 0.

We have not provided measures of model complexity and fit because such measures would be misleading. We would expect the areal GLMM to provide a better fit than copCAR simply because the former has spatial random effects. But closeness of fit is beside the point. copCAR's purpose is regression, not smoothing. And the mixed model's facility for smoothing does not imply that the model also performs well with respect to regression. The results of Reich et al. (2006) and Hughes and Haran (2013) show that the mixed model tends to result in a close fit and erroneous regression inference. copCAR tends to provide a poorer fit, but regression inference is improved because copCAR cannot be spatially confounded.

In any case, copCAR is capable of smoothing if smoothing is desired. This can be accomplished using a technique known as spatial filtering (Griffith 2003), the details of which are beyond the scope of this article.

8. CONCLUSION

This article developed copCAR, a flexible regression model for areal data. Because copCAR induces spatial dependence using a copula instead of an augmented linear predictor, copCAR overcomes many of the problems posed by the most commonly used areal models. Specifically, copCAR is flexible, permits positive spatial dependence for all types of data, is able to recover the dependence parameter ρ , cannot exhibit spatial confounding, and does not require conditional interpretation of estimated regression coefficients. Moreover, copCAR permits efficient computation for approximate maximum likelihood and maximum composite marginal likelihood inference. This makes copCAR suitable for the analyses of larger areal datasets.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The author is grateful to Jim Hodges, Brian Reich, and Vesna Zadnik for providing the Slovenia data; and to an associate editor and two referees, whose many helpful comments led to a much better article.

This work was partially supported by a grant from the National Cancer Institute, National Institutes of Health (R03 CA179555).

APPENDIX

Proposition 1

Let $\mathbf{Y} \sim \mathcal{N}\{\mathbf{0}, (\mathbf{D} - \rho\mathbf{A})^{-1}\}$. Then for any $\epsilon > 0$,

$$\mathbb{V}Y_i - \frac{1}{d_i} \{1 + \rho(\boldsymbol{\Omega})_{ii} + \rho^2(\boldsymbol{\Omega}^2)_{ii} + \dots + \rho^k(\boldsymbol{\Omega}^k)_{ii}\} < \epsilon$$

for all i so long as

$$k > \frac{\log(1 - \rho) + \log \epsilon - \log c(G)}{\log \rho} - 1,$$

where

$$c(G) = \max_i \left\{ \frac{(\boldsymbol{\Omega}^2)_{ii}}{d_i} \right\}.$$

Proof

Observe that

$$\mathbb{V}\mathbf{Y} = (\mathbf{D} - \rho\mathbf{A})^{-1} = \{\mathbf{D}(\mathbf{I} - \rho\boldsymbol{\Omega})\}^{-1} = (\mathbf{I} - \rho\boldsymbol{\Omega})^{-1}\mathbf{D}^{-1},$$

which implies that $\mathbb{V}Y_i = \frac{1}{d_i} \{(\mathbf{I} - \rho\boldsymbol{\Omega})^{-1}\}_{ii}$. Now, if $|\rho| < 1$, $(\mathbf{I} - \rho\boldsymbol{\Omega})^{-1}$ can be written as

$$(\mathbf{I} - \rho\boldsymbol{\Omega})^{-1} = \mathbf{I} + \rho\boldsymbol{\Omega} + \rho^2\boldsymbol{\Omega}^2 + \dots + \rho^k\boldsymbol{\Omega}^k + \dots$$

(Iosifescu 1980, p. 45), and so we see that

$$\mathbb{V}Y_i = \frac{1}{d_i} \{1 + \rho(\boldsymbol{\Omega})_{ii} + \rho^2(\boldsymbol{\Omega}^2)_{ii} + \dots + \rho^k(\boldsymbol{\Omega}^k)_{ii}\} + r_i,$$

Where $r_i = \frac{1}{d_i} \sum_{m=k+1}^{\infty} \rho^m (\boldsymbol{\Omega}^m)_{ii}$. Since $(\boldsymbol{\Omega}^m)_{ii} \leq (\boldsymbol{\Omega}^2)_{ii}$ when $m > 2$, we have

$$r_i \leq \frac{1}{d_i} \sum_{m=k+1}^{\infty} \rho^m (\boldsymbol{\Omega}^2)_{ii} = \frac{(\boldsymbol{\Omega}^2)_{ii}}{d_i} \sum_{m=k+1}^{\infty} \rho^m = \frac{(\boldsymbol{\Omega}^2)_{ii}}{d_i} \frac{\rho^{k+1}}{1 - \rho}.$$

Thus $r_i < \epsilon$ for all i if

$$\max_i \left\{ \frac{(\Omega^2)_{ii}}{d_i} \right\} \frac{\rho^{k+1}}{1-\rho} < \epsilon.$$

And a bit of algebra establishes the result.

Note that k can be decreased considerably by using a larger power of Ω in place of Ω^2 . An efficient way to find the smallest possible value for k is to first generate the sequence

$$c(G)_2, c(G)_4, \dots, c(G)_{2^j}, \dots = \max_i \left\{ \frac{(\Omega^2)_{ii}}{d_i} \right\}, \max_i \left\{ \frac{(\Omega^4)_{ii}}{d_i} \right\}, \dots, \max_i \left\{ \frac{(\Omega^{2^j})_{ii}}{d_i} \right\}, \dots$$

until the sequence converges, to $c(G)_*$, say. Then take

$$k > \frac{\log(1-\rho) + \log \epsilon - \log c(G)_*}{\log \rho} - 1.$$

References

- Apanasovich T, Ruppert D, Lupton J, Popovic N, Turner N, Chapkin R, Carroll R. Aberrant crypt foci and semiparametric modeling of correlated binary data. *Biometrics*. 2008; 64(2):490–500. [PubMed: 17725810]
- Assunção R, Krainski E. Neighborhood dependence in Bayesian spatial models. *Biometrical Journal*. 2009; 51(5):851–869. [PubMed: 19827056]
- Banerjee, S.; Carlin, B.; Gelfand, A. *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton: Chapman & Hall Ltd; 2004.
- Bárdossy A, Li J. Geostatistical interpolation using copulas. *Water Resour Res*. 2008; 44(7):W07412.
- Besag J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society Series B (Methodological)*. 1974; 36(2):192–236.
- Besag J, Kooperberg C. On conditional and intrinsic autoregression. *Biometrika*. 1995; 82(4):733–746.
- Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*. 1991; 43(1):1–20.
- Burgert C, Rüschendorf L. On the optimal risk allocation problem. *Statistics & Decisions*. 2006; 24(1/2006):153–171.
- Byrd R, Lu P, Nocedal J, Zhu C. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*. 1995; 16(5):1190–1208.
- Clayton D, Bernardinelli L, Montomoli C. Spatial correlation in ecological analysis. *International Journal of Epidemiology*. 1993; 22(6):1193–1202. [PubMed: 8144305]
- Demarta S, McNeil A. The t copula and related copulas. *International Statistical Review*. 2005; 73(1): 111–129.
- Denuit M, Lambert P. Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis*. 2005; 93(1):40–57.
- Diggle P, Tawn J, Moyeed R. Model-based geostatistics. *Applied Statistics*. 1998:299–350.
- Eddelbuettel D, Francois R. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*. 2011; 40(8):1–18.
- Ferguson, T. *Mathematical Statistics: A Decision Theoretic Approach*. New York: Academic Press; 1967.
- Flegal J, Haran M, Jones G. Markov chain Monte Carlo: Can we trust the third significant figure? *Statistical Science*. 2008; 23(2):250–260.

- Furrer R, Sain SR. spam: A sparse matrix R package with emphasis on MCMC methods for Gaussian Markov random fields. *Journal of Statistical Software*. 2010; 36(10):1–25.
- Gelfand A, Vounatsou P. Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*. 2003; 4(1):11. [PubMed: 12925327]
- Genest C, Neslehova J. A primer on copulas for count data. *Astin Bulletin*. 2007; 37(2):475.
- Geyer C. On the convergence of Monte Carlo maximum likelihood calculations. *Journal of the Royal Statistical Society Series B (Methodological)*. 1994:261–274.
- Geyer, CJ. Le Cam made simple: Asymptotics of maximum likelihood without the LLN or CLT or sample size going to infinity. In: Jones, GL.; Shen, X., editors. *Advances in Modern Statistical Theory and Applications: A Festschrift in honor of Morris L. Eaton*. Beachwood, Ohio, USA: Institute of Mathematical Statistics; 2013.
- Godambe V. An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*. 1960:1208–1211.
- Griffith, DA. *Spatial Autocorrelation and Spatial Filtering: Gaining Understanding Through Theory and Scientific Visualization*. Berlin: Springer; 2003.
- Haran M, Hodges J, Carlin B. Accelerating computation in Markov random field models for spatial data via structured MCMC. *Journal of Computational and Graphical Statistics*. 2003; 12(2):249–264.
- Haran M, Tierney L. On automating Markov chain Monte Carlo for a class of spatial models. 2012 arXiv preprint arXiv:1205.0499.
- Heagerty P, Lumley T. Window subsampling of estimating functions with application to regression models. *Journal of the American Statistical Association*. 2000; 95(449):197–211.
- Hoef JV, Jansen J. Space-time zero-inflated count models of harbor seals. *Environmetrics*. 2007; 18(7):697–712.
- Horn, RA.; Johnson, CR. *Topics in Matrix Analysis*. New York: Cambridge University Press; 1994.
- Hughes J, Haran M. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2013; 75(1):139–159.
- Hughes J, Haran M, Caragea PC. Autologistic models for binary data on a lattice. *Environmetrics*. 2011; 22(7):857–871.
- Ihaka R, Gentleman R. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*. 1996; 5:299–314.
- Iosifescu, M. *Finite Markov Processes and Their Applications*. New York: Wiley; 1980.
- Ising E. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*. 1925; 31(1):253–258.
- Kaiser M, Cressie N. Modeling Poisson variables with positive spatial dependence. *Statistics & Probability Letters*. 1997; 35(4):423–432.
- Kaiser M, Cressie N. The construction of multivariate distributions from Markov random fields. *Journal of Multivariate Analysis*. 2000; 73(2):199–220.
- Kazianka H. Approximate copula-based estimation and prediction of discrete spatial data. *Stochastic Environmental Research and Risk Assessment*. 2013; 27(8):2015–2026.
- Kazianka H, Pilz J. Copula-based geostatistical modeling of continuous and discrete data including covariates. *Stochastic Environmental Research and Risk Assessment*. 2010; 24(5):661–673.
- Kelsall, J.; Wakefield, J. Discussion of “Bayesian models for spatially correlated disease and exposure data”, by Best et al. In: Bernardo, J.; Berger, J.; Dawid, A.; Smith, A., editors. *Bayesian Statistics 6*. New York: Oxford University Press; 1999.
- Kenkel, B. pbivnorm: Vectorized Bivariate Normal CDF. 2012. R package version 0.5-1. URL:<http://CRAN.R-project.org/package=pbivnorm>
- Kindermann, R.; Snell, J. *Markov Random Fields and Their Applications*. Providence, RI: American Mathematical Society; 1980.
- Knorr-Held L, Rue H. On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics*. 2002; 29(4):597–614.

- Kolev N, Paiva D. Copula-based regression models: A survey. *Journal of statistical planning and inference*. 2009; 139(11):3847–3856.
- Lindsay B. Composite likelihood methods. *Contemporary Mathematics*. 1988; 80(1):221–239.
- Lunn D, Thomas A, Best N, Spiegelhalter D. WinBUGS – A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*. 2000; 10(4):325–337.
- Madsen L. Maximum likelihood estimation of regression parameters with spatially dependent discrete data. *JABES*. 2009; 14(4):375–391.
- Madsen L, Fang Y. Joint regression analysis for discrete longitudinal data. *Biometrics*. 2011; 67(3): 1171–1175. [PubMed: 21039391]
- Møller J. Perfect simulation of conditionally specified models. *Journal of the Royal Statistical Society, Series B, Methodological*. 1999; 61:251–264.
- Moller J, Pettitt A, Reeves R, Berthelsen K. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*. 2006; 93(2):451–458.
- Nelsen, RB. *An Introduction to Copulas*. New York: Springer; 2006.
- Ng EG, Peyton BW. Block sparse Cholesky algorithms on advanced uniprocessor computers. *SIAM Journal on Scientific Computing*. 1993; 14(5):1034–1056.
- Reich B, Hodges J, Zadnik V. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*. 2006; 62(4):1197–1206. [PubMed: 17156295]
- Rue, H.; Held, L. *Gaussian Markov Random Fields: Theory and Applications* Vol. 104 of *Monographs on Statistics and Applied Probability*. London: Chapman & Hall; 2005.
- Rüschendorf L. Stochastically ordered distributions and monotonicity of the OC-function of sequential probability ratio tests. *Statistics*. 1981; 12(3):327–338.
- Rüschendorf L. On the distributional transform, Sklar’s theorem, and the empirical copula process. *Journal of Statistical Planning and Inference*. 2009; 139(11):3921–3927.
- Sanderson, C. *Armadillo: An open source C++ linear algebra library for fast prototyping and computationally intensive experiments*. 2010. Technical report, NICTA
- Sklar A. Fonctions de répartition à n dimensions et leurs marges. *Publ Inst Statist Univ Paris*. 1959; 8(1):11.
- Smith MS, Gan Q, Kohn RJ. Modelling dependence using skew t copulas: Bayesian inference and applications. *Journal of Applied Econometrics*. 2012; 27(3):500–522.
- Tierney L, Rossini A, Li N. Snow : A parallel computing framework for the R system. *International Journal of Parallel Programming*. 2009; 37(1):78–90.
- Varin C. On composite marginal likelihoods. *AStA Advances in Statistical Analysis*. 2008; 92(1):1–28.
- Varin C, Reid N, Firth D. An overview of composite likelihood methods. *Statistica Sinica*. 2011; 21(1):5–42.
- Varin C, Vidoni P. Pairwise likelihood inference for general state space models. *Econometric Reviews*, 28. 2009; 1(3):170–185.
- Zadnik V, Reich B. Analysis of the relationship between socioeconomic factors and stomach cancer incidence in Slovenia. *Neoplasma*. 2006; 53(2):103. [PubMed: 16575465]

Figure 1.

Panel A shows the correlations between the first CAR variable of the simulation study and every other variable in the same row of the 20×20 lattice, for three values of the dependence parameter ρ . The circles mark the correlations; the curves merely guide the eye. The horizontal line allows us to see the effective range of the correlation function, i.e., the “distance” at which the correlation has dropped to 0.05. Panel B shows a simulated realization from the CAR copula for each value of ρ .

Figure 2.
The Slovenia stomach cancer data. A darker shade of gray indicates a larger SIR. The values range from 0 to approximately 4.1.

Table 1

The results of the simulation study for the 20×20 lattice and Poisson marginals.

Parameter	Mean Estimate			Mean SE			Simulation SD			Coverage		
	CE	CML	DT	CE	CML	DT	CE	CML	DT	CE	CML	DT
$\beta_1 = 1$	1.002	1.008	0.999	0.130	0.136	0.142	0.122	0.128	0.124	94%	96%	95%
$\rho = 0.8$	0.769	0.772	0.806	0.065	0.067	0.070	0.065	0.065	0.063	96%	96%	96%
$\beta_1 = 1$	0.981	0.969	0.991	0.236	0.267	0.261	0.252	0.287	0.253	94%	94%	95%
$\rho = 0.975$	0.971	0.968	0.977	0.013	0.016	0.011	0.013	0.012	0.008	94%	96%	95%
$\beta_1 = 1$	0.966	0.956	0.980	0.303	0.377	0.331	0.322	0.406	0.336	94%	95%	97%
$\rho = 0.995$	0.993	0.992	0.995	0.003	0.005	0.002	0.003	0.004	0.002	93%	97%	98%

Table 2

The results of the simulation study for the 30×30 lattice and Poisson marginals.

Parameter	Mean Estimate			Mean SE			Simulation SD			Coverage		
	CE	CML	DT	CE	CML	DT	CE	CML	DT	CE	CML	DT
$\beta_1 = 1$	1.004	1.004	1.001	0.089	0.095	0.090	0.085	0.088	0.087	96%	97%	97%
$\rho = 0.8$	0.779	0.791	0.821	0.042	0.042	0.049	0.042	0.044	0.041	94%	96%	96%
$\beta_1 = 1$	0.988	0.991	1.001	0.178	0.197	0.190	0.181	0.202	0.185	94%	94%	96%
$\rho = 0.975$	0.972	0.970	0.976	0.009	0.010	0.008	0.009	0.010	0.008	96%	93%	94%
$\beta_1 = 1$	0.985	1.005	1.016	0.244	0.311	0.285	0.251	0.326	0.283	95%	94%	96%
$\rho = 0.995$	0.995	0.993	0.995	0.002	0.003	0.002	0.002	0.002	0.002	95%	98%	98%

Table 3

The results of applying the classical GLM, copCAR, and the areal GLMM to the stomach cancer data.

Model	α	β	ρ	τ
Nonspatial	0.156 (0.120, 0.193)	-0.137 (-0.176, -0.098)	NA	NA
copCAR (CE)	0.153 (0.111, 0.195)	-0.128 (-0.171, -0.084)	0.289 (0.097, 0.574)	NA
copCAR (CML)	0.170 (0.124, 0.216)	-0.149 (-0.195, -0.101)	0.279 (0.139, 0.466)	NA
copCAR (DT)	0.153 (0.112, 0.194)	-0.128 (-0.174, -0.081)	0.283 (0.138, 0.475)	NA
areal GLMM	0.137 (0.003, 0.216)	-0.070 (-0.144, 0.022)	0.980 (0.494, 1.000)	4.54 (2.66, 8.49)