and it would work. The problem with this is that the objects we construct have a formula that has, literally `QDPI.Y[, n]` as the dependent variable in the formula. If we want to do anything with the objects afterwards, such as prune them, update them, &c, we need to re-establish what `n` is in this particular case. The original object `n` was the loop variable and that is long gone. This is not difficult, of course, but it is an extra detail we need to carry along that we don't need. Essentially the formula part of the object we generate would not be self-contained and this can cause problems.

The strategy we have adopted has kept all the variables together in one data frame and explicitly encoded the correct response variable by name into the formula of each object as we go. At the end each fitted `rpart` object may be manipulated in the usual way witout this complication involving the now defunct loop variable.

Bibliography

W. N. Venables and B. D. Ripley. *S Programming*. Springer-Verlag, New York, 2000. 24

Bill Venables
CSIRO Marine Labs, Cleveland, Qld, Australia
Bill.Venables@cmis.csiro.au

geoRglm: A Package for Generalised Linear Spatial Models

by Ole F. Christensen and Paulo J. Ribeiro Jr

geoRglm is a package for inference in generalised linear spatial models using Markov chain Monte Carlo (MCMC) methods. It has been developed at the Department of Mathematical Sciences, Aalborg University, Denmark and the Department of Mathematics and Statistics, Lancaster University, UK. A web site with further information can be found at http://www.maths.lancs.ac.uk/~christen/geoRglm. **geoRglm** is an extension to the **geoR** package (Ribeiro, Jr. and Diggle, 2001). Maximum compatibility between the two packages has been intended and **geoRglm** also uses several of **geoR**'s internal functions.

Generalised linear spatial models

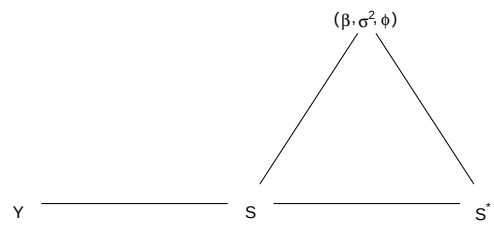
The classical geostatistical model assumes Gaussianity, which may be an unrealistic assumption for some data sets. The *generalised linear spatial model* (GLSM) as presented in Diggle et al. (1998), Zhang (2002) and Christensen and Waagepetersen (2002) provides a natural extension to deal with response variables for which a standard distribution other than the Gaussian more accurately describes the sampling mechanism involved.

The GLSM is a generalised linear mixed model in which the random effects are derived from a spatial process $S(\cdot)$. This leads to the following model specification.

Let $S(\cdot) = \{S(x) : x \in A\}$ be a Gaussian stochastic process with $\mathrm{E}[S(x)] = d(x)^{\mathrm{T}}\beta$, $\mathrm{Var}\{S(x)\} = \sigma^2$ and correlation function $\mathrm{Corr}\{S(x), S(x')\} = \rho(u;\phi)$ where $u = \|x - x'\|$ and ϕ is a parameter. Assume that the responses $Y_1, \ldots, Y_n$ observed at locations $x_1, \ldots, x_n$ in the sampling design, are conditionally independent given $S(\cdot)$, with conditional expectations $\mu_1, \ldots, \mu_n$, where $h(\mu_i) = S(x_i), i = 1, \ldots, n$, for a known link function $h(\cdot)$.

We write $S = (S(x_1), \ldots, S(x_n))^{\mathrm{T}}$ for the unobserved values of the underlying process at $x_1, \ldots, x_n$, and S^* for the values of $S(\cdot)$ at all other locations of interest, typically a fine grid of locations covering the study region.

The conditional independence structure of the GLSM is then indicated by the following graph.



The likelihood for a model of this kind is in general not expressible in closed form, but only as a high-dimensional integral

$$L(\beta, \sigma^2, \phi) = \int \prod_{i=1}^{n} f(y_i; h^{-1}(s_i)) p(s; \beta, \sigma^2, \phi) ds,$$

where $f(y;\mu)$ denotes the density of the error distribution parameterised by the mean μ, and $p(s;\beta,\sigma^2,\phi)$ is the multivariate Gaussian density

for the vector S. The integral above is also the normalising constant in the conditional distribution $[S|y, \beta, \sigma^2, \phi]$,

$$p(s \mid y, \beta, \sigma^2, \phi) \propto \prod_{i=1}^{n} f(y_i; h^{-1}(s_i)) p(s; \beta, \sigma^2, \phi).$$

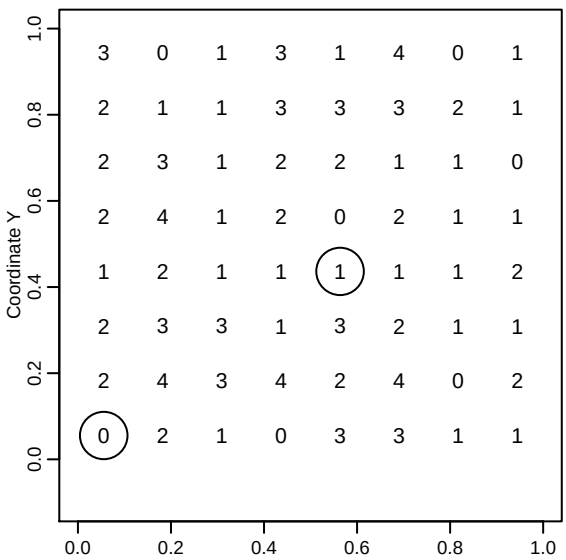
In practice, the high dimensionality of the integral prevents direct calculation of the predictive distribution $[S^* \mid y, \beta, \sigma^2, \phi]$. Markov chain Monte Carlo provides a solution to this. First by simulating a Markov chain we obtain a sample $s(1), \ldots, s(m)$ from $[S|y, \beta, \sigma^2, \phi]$, where each $s(j)$ is an n-dimensional vector. Second, by direct sampling from $[S^*|s(j), \beta, \sigma^2, \phi]$, $j = 1, \ldots, m$ we obtain a sample $s^*(1), \ldots, s^*(m)$ from $[S^*|y, \beta, \sigma^2, \phi]$. The MCMC algorithm uses Langevin-Hastings updates of S which are simultaneous updates based on gradient information.

In a Bayesian analysis priors must be assigned to the parameters in the model. For (β, σ^2) a conjugate prior exists such that these parameters can be integrated out analytically, whereas for ϕ one has to extend the MCMC algorithm above with updates of this parameter. We use a Metropolis random walk-type proposal for updating ϕ.

In its current version **geoRglm** implements the spatial Poisson model and the spatial binomial model.

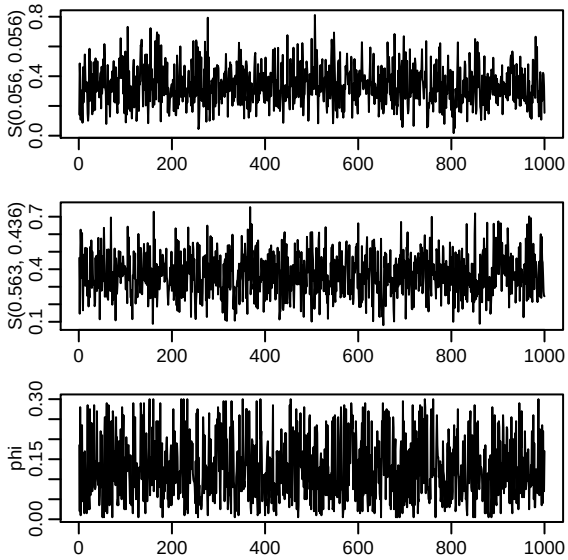
Package features

The following example gives a short demonstration of an analysis for a binomial spatial model with logistic link using the function `binom.krige.bayes`. We omit the specific commands here, but refer to the **geoRglm** homepage for further details. Consider the simulated data set shown below which consists of binomial data of size 4 at 64 locations.

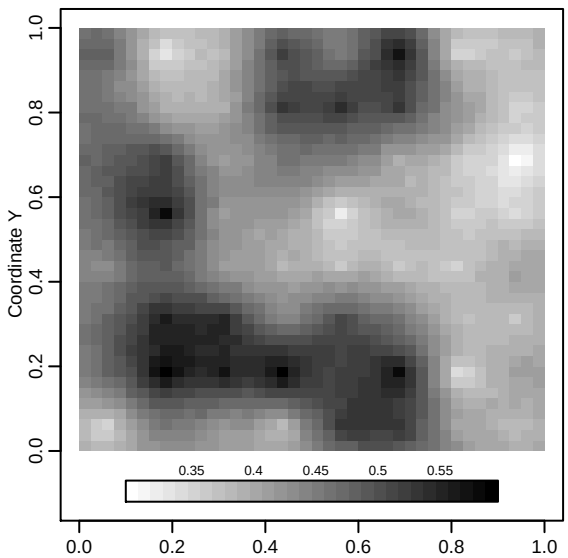


Priors for the parameters and options for the MCMC algorithm are set using the `prior.glm.control` and `mcmc.control` functions, respectively. As a rule of thumb the proposal variances must be tuned such that the acceptance rates for updating the random effects and the parameter ϕ are approximately 60% and 25%, respectively.

Output from the MCMC algorithm is presented below for the parameter ϕ and for the two random effects at locations marked with a circle in the figure above.



Predicted values of the probabilities $p(x) = \exp(S(x))/(1 + \exp(S(x)))$ at 1600 locations are plotted below using the function `image.kriging` from **geoR**.



Further details about this example and an introduction to the models can be found in Diggle et al. (2002) and in the files 'inst/doc/bookchap.pdf' and 'inst/doc/geoRglm.intro.pdf' distributed with the package.

Future developments

Work in progress with Gareth Roberts and Martin Sköld aims to improve convergence and mixing of the MCMC algorithm by using a more appropriate parameterisation.

Acknowledgements

Some of the C code in the package is based on code originally developed together with Rasmus Waagepetersen. We are grateful to Peter J. Diggle for giving encouragement and support to the development of the package. Ole acknowledges support from DINA, NERC and the EU TMR network. Paulo acknowledges CAPES/Brasil grant 1676/96-2.

Bibliography

O.F. Christensen and R.P. Waagepetersen. Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics*, 58:280–286, 2002. 26

P. J. Diggle, P. J. Ribeiro Jr, and O. F. Christensen. An introduction to model-based geostatistics. In M. B. Hansen and J. Møller, editors, *Spatial statistics and computational methods*. Springer Verlag, 2002. (to appear). 27

P. J. Diggle, J. A. Tawn, and R. A. Moyeed. Model-based geostatistics (with discussion). *Appl. Statist.*, 47:299–350, 1998. 26

Paulo J. Ribeiro, Jr. and Peter J. Diggle. geoR: A package for geostatistical analysis. *R News*, 1(2):14–18, 2001. 26

H. Zhang. On estimation and prediction for spatial generalised linear mixed models. *Biometrics*, 58:129–136, 2002. 26

Ole F. Christensen
Lancaster University, UK
o.christensen@lancaster.ac.uk

Paulo J. Ribeiro Jr
Universidade Federal do Paraná, Brasil
and *Lancaster University, UK*
Paulo.Ribeiro@est.ufpr.br

Querying PubMed

Web Services

by Robert Gentleman and Jeff Gentry

Introduction

While many view the world wide web (WWW) as an interactive environment primarily designed for interactive use, more and more sites are providing web services that can be accessed programmatically. In this article we describe some preliminary tools that have been added to the *annotate* package in the Bioconductor project www.bioconductor.org. These tools facilitate interaction with resources provided at the National Center for Biotechnology Information (NCBI) located at www.ncbi.nlm.nih.gov. These ideas represent only a very early exploration of a single site and we welcome any contributions to the project in the form of enhancements, new tools, or tools adapted to other sites providing web services.

We believe that web services will play a very important role in computational biology. In part this is because the data are complex and gain much of their relevance by association with other data sources. For example, knowing that there is a particularly high level of messenger RNA (mRNA) for some gene (or set of genes) does not provide us with much insight. However, associating these genes with the relevant scientific literature and finding common themes often does provide new insight into how these genes interact.

We can think of a cellular pathway as a set of genes that interact (through the proteins that they produce) to provide a particular function or protein. A second way of obtaining insight into the role of certain genes would be to examine the expression of mRNA for a set of genes in a particular pathway, or to take a set of genes and determine whether there is a particular pathway that contains (most of) these genes.

Both of these examples rely on associating experimental data with data that are available in databases or in textual form. These latter data sources are often large and are continually evolving. Thus, it does not seem practical nor prudent to keep local versions of them suitable for querying. Rather, we should rely on retrieving the data when it is wanted and on tools to process the data that are obtained from on-line sources.

It is important to note that most of the processes we are interested in can be carried out interactively. However, there are two main advantages to designing programmatic interfaces. First, interactive use introduces a rate limiting step. The analysis of genomic data needs to be high throughput. A second reason to prefer programmatic access is that it allows for the possibility of combining data from several sources, possibly filtered through online resources, to provide a new product. Another reason to prefer a programmatic approach is that it makes fewer mistakes and