空间广义线性混合效应模型及其在流行病预测中的应用

Spatial Generalized Linear Mixed Models and its Applications to Prevlence Mapping

黄湘云

2018-01-13

摘要

空间广义线性混合效应模型的应用背景介绍

空间广义线性混合效应模型及其应用

空间广义线性混合效应模型及其应用

目录

1 论文综述

线性模型到广义线性模型线性混合效应模型到广义线性混合效应模型再到空间广义线性混合效应模型

求解混合效应模型的各种方法

从 kriging 克里金插值到高斯过程

简单介绍提出本文的大纲思路

简单回顾广义线性模型和混合效应模型

定位统计计算

广义线性模型的应用广泛广义线性模型的算法实现进展综述

介绍空间数据统计包含地统计、离散空间变差、空间点过程

引出广义线性模型在 geostatistical data analysis 的应用及算法实现

在 S 语言中空间数据分析和建模 Modern Applied Statistics with S[36]

检验环境和基因效应在空间相关性中的存在性[29] 流行现象的时空分析[23]

近年来涉及空间数据分析和建模的书籍也越来越多，

用于空间数据分析的分层模型 Hierarchical Modeling and Analysis for Spatial Data[1]

基于 R-INLA 软件的空间和时空贝叶斯模型 Spatial and Spatio-temporal Bayesian Models with R-INLA[20]

特别地，基于地统计数据的有[33]

R 语言空间数据可视化方面呈现越来越流行的趋势，从早些年的 lattice[31] 到如今的 ggplot2[37]，操作空间数据的 sp 对象也发展为 sf 对象，同时整合了不少第三方软件和服务如基于 Google Earth 的空间可视化[17]，基于 Google Maps 的交互空间可视化[18]

一元和多元时空模型 spBayes 包[12]

2 模型介绍

2.1 线性模型

线性模型的一般形式为

$$Y = X'\beta + \epsilon, \mathrm{E}(\epsilon) = 0, \mathrm{Cov}(\epsilon) = \sigma^2 I \tag{2.1}$$

其中，$Y = (y_1, y_2, \cdots, y_n)'$ 是 n 维列向量，代表对响应变量 Y 的 n 次重复观测；$\beta = (\beta_0, \beta_1, \cdots, \beta_{p-1})'$ 是 p 维列向量，代表模型自变量 X 的系数，β_0 是截距项；$X' = (1'_{(1\times n)}, X'_{(1)}, X'_{(2)}, \cdots, X'_{(n)}), 1'_{(1\times n)}$ 是全 1 的 n 维列向量，而 $X'_{(i)} = (x_{1i}, x_{2i}, \cdots, x_{ni})'$ 代表对第 i 个自变量的 n 次观测；$\epsilon = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)'$ 是 n 维列向量，代表模型的随机误差，$\mathrm{E}(\epsilon_i \epsilon_j) = 0, i \neq j$。求解线性模型 (2.1) 的 R 函数是 `lm`，近年来，高维乃至超高维稀疏线性模型成为热门的研究方向，相关的 R 包也越来越多，比较著名的有 `glmnet`[34] 和 `SIS`[30]。

2.2 广义线性模型

广义线性模型的一般形式

$$Y \sim \text{指数族} \quad \eta = X'\beta \quad \mathrm{E}(Y) = g^{-1}(\eta) \tag{2.2}$$

简写之为

$$g(\mu) = X'\beta$$

其中 $\mu \equiv \mathrm{E}(Y)$,g 代表联系函数,特别地,当 $Y \sim N(\mu, \sigma^2)$ 时,$g(x) = x$;当 $Y \sim \mathrm{Binomial}(n, p)$ 时，$g(x) = \ln(\frac{x}{1-x})$；当 $Y \sim \mathrm{Possion}(\lambda)$ 时，$g(x) = \ln(x)$；此处不一一列举。[21] 模型(2.2)最早由 Nelder 和 Wedderburn[24] 提出，它弥补了模型(2.1) 的两个重要缺点：一是因变量只能取连续值的情况，二是期望与自变量只能用线性关系联系[40]。求解广义线性模型 (2.2) 的 R 函数是 `glm`，参数估计的办法一般是拟似然法。

2.3 混合效应模型

广义线性混合模型的一般形式

$$g(\mu) = X'\beta + Z'b \tag{2.3}$$

其中 Z' 是 q 维随机效应的 $n \times q$ 的向量值矩阵。混合效应模型包含线性混合效应模型、广义线性混合效应模型、非线性混合效应模型等，之所以称之为混合效应，是因为模型既包含固定效应 (fixed-effects) 又随机效应 (random effects)。如前所述的线性和广义线性模型中的自变量就是固定效应，而随机效应是那些不能直接观察到的潜变量。求解模型(2.3)的 R 包有 `nlme`[25]，`mgcv`[38] 和 `lme4`[2]，参数估计的方法一般有限制极大似然法。

2.4 空间广义线性混合效应模型

本文重点关注空间广义线性混合效应模型 (Spatial Generalized linear mixed-effects models, 简写为 SGLMM)[3;35;39]，它既是对模型(2.1)、(2.2)和(2.3)的延伸也是对空间数据分析的应用，属于空间统计下的地统计 (geostatistics) 分支，因此在有些参考文献[7;8;10] 中也称为广义线性地统计模型 (Generalized linear geostatistical models)，SGLMM 具有广泛的应用，如分析核污染浓度的空间分布[11]，预测热带流行病的分布[9]

3　算法描述

3.1　贝叶斯方法

coda: Convergence Diagnosis and output analysis for MCMC 用于对 MCMC 的收敛性诊断和输出分析[26] geoR: 用于空间数据分析和预测的贝叶斯方法[27] geoRglm: 空间广义线性混合效应模型[6] brms[4]

```r
library(coda)
```

MCMCvis— Tools to Visualize, Manipulate, and Summarize MCMC Output

Performs key functions for MCMC analysis using minimal code - visualizes, manipulates, and summarizes MCMC output. Functions support simple and straightforward subsetting of model parameters within the calls, and produce presentable and 'publication-ready' output. MCMC output may be derived from Bayesian model output fit with JAGS, Stan, or other MCMC samplers.

mgcv and JAGS

Stan 是一种概率编程语言[5]，可以替代 BUGS (**B**ayesian inference **U**sing **G**ibbs **S**ampling)[19] 作为 MCMC 的高效实现，可用于贝叶斯框架下，标准地统计模型的参数估计，Stan 提供多种语言的接口实现，方便起见，本文采用它提供的 R 语言接口 – rstan 包[5;13]。此外，还有[22]

lme4[2]

gstat[16]

glmmBUGS RStan 实现 MCMC

3.2　最大似然方法

最大似然[14] 将 MCML 和 MCMC 方法应用于空间广义线性混合效应模型的参数估计和预测，

3.3　拉普拉斯近似

方法[15]

4　数值模拟

`RandomFields` 是模拟多元随机场的 R包[32]，`geoR` 包[28] 的 `grf` 函数只适合模拟少量数据点

分泊松、二项两种情况、上述 3 种方法，在空间广义线性混合效应模型下的模拟情况（使用标准的地统计模型，无时间项和混合分布）

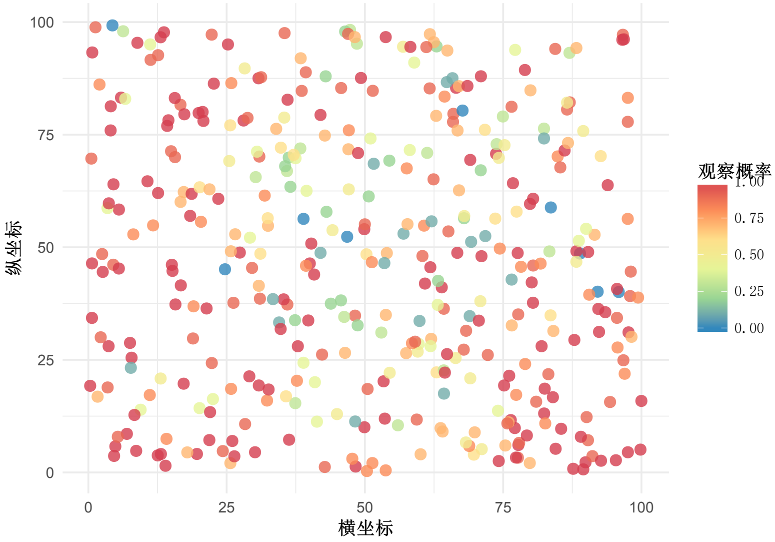


图 1: 二项分布

标准地统计流行抽样模型（Standard Geostatistical Prevalence Sampling Model）

$$\log[p(x_i)/\{1-p(x_i)\}] = d(x_i)'\beta + S(x_i) + Z_i \tag{4.1}$$

其中 $\boldsymbol{S} = \{S(x) : x \in \mathbb{R}^2\}$

```
## grf: simulation(s) on randomly chosen locations with  400  points
## grf: process with  1  covariance structure(s)
## grf: nugget effect is: tausq= 1
## grf: covariance model 1 is: powered.exponential(sigmasq=1, phi=25, kappa = 1)
## grf: decomposition algorithm used is:  cholesky
## grf: End of simulation procedure. Number of realizations: 1
```

$$\lambda(x) = \frac{\exp(\mu + S(x))}{1 + \exp(\mu + S(x))}$$

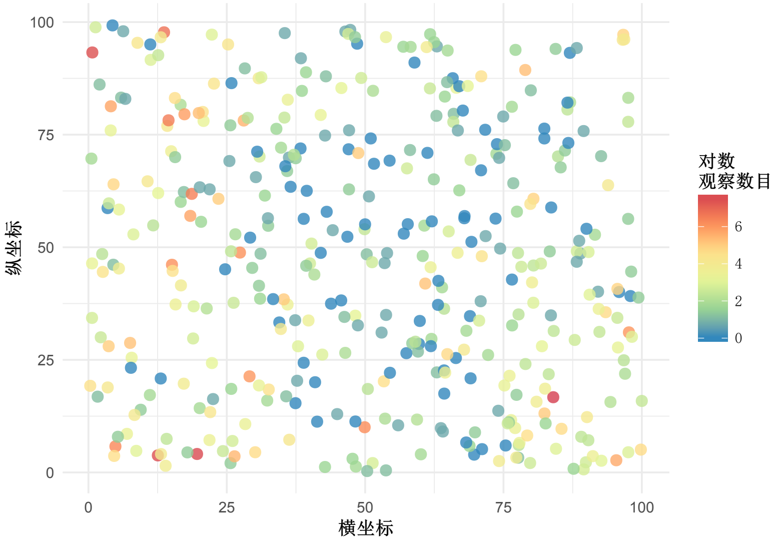
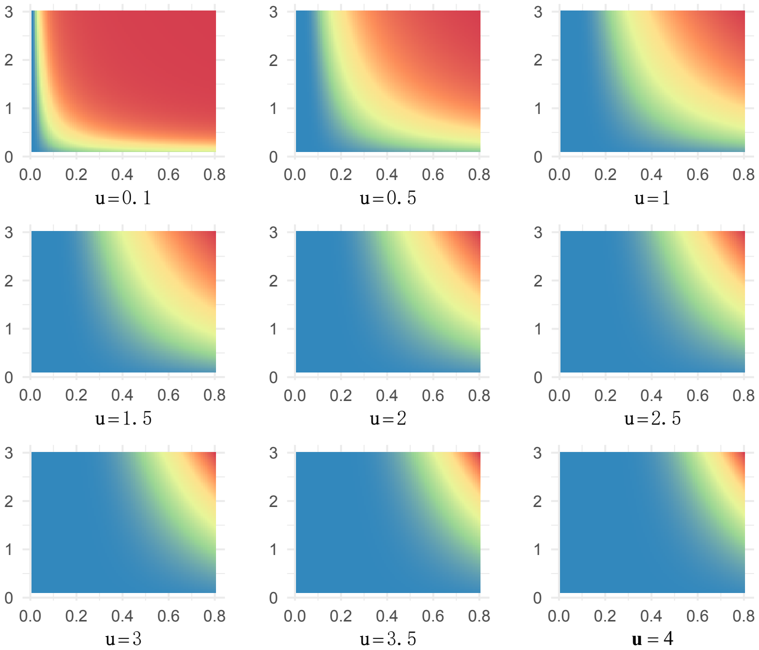


图 2: 泊松分布



从蓝到红，值由小变大

4.1 模拟二项分布数据集

此处模拟数据集 `data_sim` 来自 `PrevMap` 包，零均值高斯过程单元格上 30 × 30 参数 $\sigma^2 = 1, \phi = 0.15, \kappa = 2$，块金效应 (nugget effect) $\tau^2 = 0$，每个格点上重复实验 10 次，得到响应变量二项分布的概率值

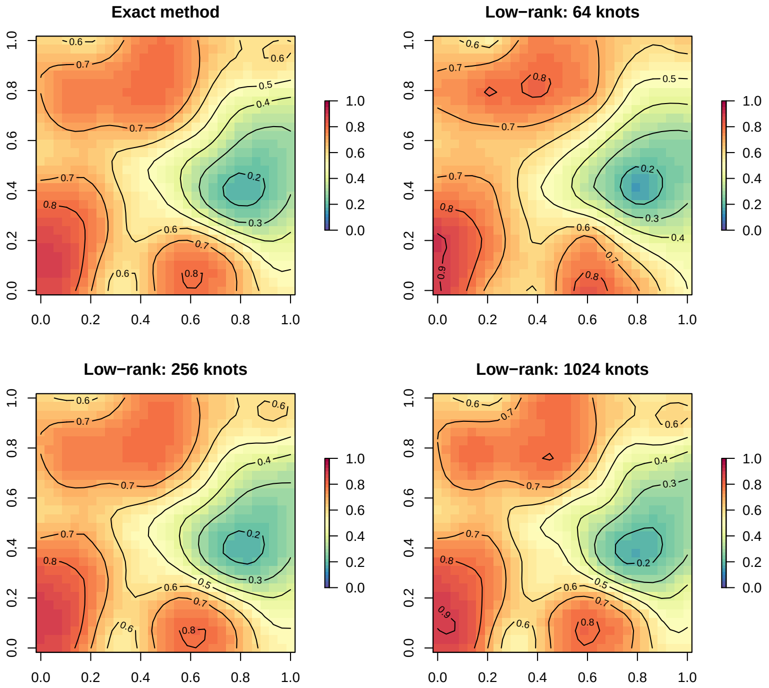


图 3: 数值模拟

参考文献

[1] Sudipto Banerjee, Bradley P. Carlin, and Alan E Gelfand. *Hierarchical Modeling and Analysis for Spatial Data, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, second edition, 2015. ISBN 978-1-4398-1918-0,1439819181.

[2] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

[3] Wagner Hugo Bonat and Paulo J. Ribeiro Jr. Practical likelihood analysis for spatial generalized linear mixed models. *Environmetrics*, 27(2):83–89, 2016.

[4] Paul-Christian Bürkner. brms: An R package for bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28, 2017. doi: 10.18637/jss.v080.i01.

[5] Bob Carpenter, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, 76(1):1–32, 2017. ISSN 1548-7660. doi: 10.18637/jss.v076.i01. URL https://www.jstatsoft.org/v076/i01.

[6] O.F. Christensen and P.J. Ribeiro Jr. geoRglm: a package for generalised linear spatial models. *R-NEWS*, 2(2):26–28, 2002. URL https://cran.r-project.org/doc/Rnews. ISSN 1609-3631.

[7] Ole F Christensen. Monte carlo maximum likelihood in model-based geostatistics. *Journal of Computational and Graphical Statistics*, 13(3):702–718, 2004.

[8] Peter Diggle, Rana Moyeed, Barry Rowlingson, and Madeleine Thomson. Childhood malaria in the gambia: a case-study in model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(4):493–506, 2002. ISSN 1467-9876. doi: 10.1111/1467-9876.00283. URL http://dx.doi.org/10.1111/1467-9876.00283.

[9] Peter J. Diggle and Emanuele Giorgi. Model-based geostatistics for prevalence mapping in low-resource settings. *Journal of the American Statistical Association*, 111(515):1096–1120, 2016. doi: 10.1080/01621459.2015.1123158.

[10] Peter J. Diggle and Paulo J. Ribeiro Jr. *Model-based Geostatistics*. Springer-Verlag, New York, 2007. ISBN 978-1-4419-2193-2. doi: 10.1007/978-0-387-48536-2.

[11] Peter J. Diggle, J. A. Tawn, and R. A. Moyeed. Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350, 1998. ISSN 1467-9876. doi: 10.1111/1467-9876.00113. URL http://dx.doi.org/10.1111/1467-9876.00113.

[12] Andrew O. Finley, Sudipto Banerjee, and Alan E.Gelfand. spBayes for large univariate and multivariate point-referenced spatio-temporal data models. *Journal of Statistical Software*, 63 (13):1–28, 2015. URL http://www.jstatsoft.org/v63/i13/.

[13] Andrew Gelman, Daniel Lee, Jiqiang Guo, and et al. Stan: A probabilistic programming language for bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, 40 (5):837–840, 2015.

[14] Emanuele Giorgi and Peter J. Diggle. PrevMap: An R package for prevalence mapping. *Journal of Statistical Software*, 78(8):1–29, 2017. doi: 10.18637/jss.v078.i08.

[15] Virgilio Gómez-Rubio and Håvard Rue. Markov chain monte carlo with the integrated nested laplace approximation. *ArXiv e-prints*, 2017.

[16] Benedikt Gräler, Edzer Pebesma, and Gerard Heuvelink. Spatio-temporal interpolation using gstat. *The R Journal*, 8:204–218, 2016.

[17] Tomislav Hengl, Pierre Roudier, Dylan Beaudette, and Edzer Pebesma. plotKML: Scientific visualization of spatio-temporal data. *Journal of Statistical Software*, 63(5):1–25, 2015. URL http://www.jstatsoft.org/v63/i05/.

[18] Milan Kilibarda and Branislav Bajat. plotgooglemaps: the r-based web-mapping tool for thematic spatial data. *GEOMATICA*, 66:37–49, 2012.

[19] David Lunn, David Spiegelhalter, Andrew Thomas, and Nicky Best. The bugs project: Evolution, critique and future directions. *Statistics in Medicine*, 28(25):3049–3067, 2009. ISSN 1097-0258.

[20] Michela Cameletti Marta Blangiardo. *Spatial and Spatio-temporal Bayesian Models with R-INLA*. Wiley, 2015. ISBN 1118326555, 9781118326558.

[21] Peter McCullagh and John Nelder. *Generalized Linear Models*. Chapman and Hall/CRC, Boca Raton, second edition, 1989. ISBN 0-412-31760-5.

[22] Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC, New York, 2015. URL http://xcelab.net/rm/statistical-rethinking/. ISBN 978-1482253443.

[23] Sebastian Meyer, Leonhard Held, and Michael Höhle. Spatio-temporal analysis of epidemic phenomena using the R package surveillance. *Journal of Statistical Software*, 77(11):1–55, 2017. doi: 10.18637/jss.v077.i11.

[24] John A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135(3):370–384, 1972.

[25] Jose Pinheiro, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2017. URL https://CRAN.R-project.org/package=nlme. R package version 3.1-131.

[26] Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. Coda: Convergence diagnosis and output analysis for mcmc. *R-News*, 6(1):7–11, 2006. URL https://www.r-project.org/doc/Rnews/Rnews_2006-1.pdf.

[27] Paulo J. Ribeiro Jr. and Peter J. Diggle. geoR: A package for geostatistical analysis. *R-News*, 1 (2):14–18, June 2001. URL https://cran.r-project.org/doc/Rnews/Rnews_2001-2.pdf.

[28] Paulo J. Ribeiro Jr. and Peter J. Diggle. *geoR: Analysis of Geostatistical Data*, 2016. URL https://CRAN.R-project.org/package=geoR. R package version 1.7-5.2.

[29] François Rousset and Jean-Baptiste Ferdy. Testing environmental and genetic effects in the presence of spatial autocorrelation. *Ecography*, 37(8):781–790, 2014. URL http://dx.doi.org/10.1111/ecog.00566.

[30] Diego Franco Saldana and Yang Feng. Sis: An r package for sure independence screening in ultrahigh dimensional statistical models. *Journal of Statistical Software*, page to appear, 2016.

[31] Deepayan Sarkar. *Lattice: Multivariate Data Visualization with R*. Springer-Verlag, New York, 2008. URL http://lmdvr.r-forge.r-project.org. ISBN 978-0-387-75968-5.

[32] Martin Schlather, Alexander Malinowski, Peter J. Menck, Marco Oesting, and Kirstin Strokorb. Analysis, simulation and prediction of multivariate random fields with package RandomFields. *Journal of Statistical Software*, 63(8):1–25, 2015. URL http://www.jstatsoft.org/v63/i08/.

[33] Daniela K Schlüter, Martial L Ndeffombah, Innocent Takougang, and et al. Using community-level prevalence of loa loa infection to predict the proportion of highly-infected individuals: Statistical modelling to support lymphatic filariasis and onchocerciasis elimination programs. *Plos Neglected Tropical Diseases*, 10(12), 2016.

[34] Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5): 1–13, 2011. URL http://www.jstatsoft.org/v39/i05/.

[35] Cristiano Varin, Gudmund Høst, and Øivind Skare. Pairwise likelihood inference in spatial generalized linear mixed models. *Computational Statistics & Data Analysis*, 49(4):1173–1191, 2005.

[36] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer-Verlag, New York, fourth edition, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0.

[37] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, second edition, 2016. ISBN 978-3-319-24277-4. URL http://ggplot2.org.

[38] Simon N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, second edition, 2017. ISBN 978-1498728331.

[39] Hao Zhang. On estimation and prediction for spatial generalized linear mixed models. *Biometrics*, 58(1):129–36, 2002.

[40] 陈希孺. 广义线性模型的拟似然法. 中国科学技术大学出版社, 合肥, 2011. ISBN 978-7-312-02284-5.