空间广义线性混合效应模型及其应用

Spatial Generalized Linear Mixed Models with Its Applications

导师：李再兴　学生：黄湘云

计算数学与统计系
中国矿业大学（北京）理学院

2015 级硕士学位论文答辩

研究背景

计算方法:

- MCMC: Peter J. Diggle et al. (1998) from Lancaster University
- MCML: Charles J. Geyer (1994) and Hao Zhang (2002) from Purdue University
- INLA: Håvard Rue et al. (2009) from Columbia University
- FRK: Noel A. C. Noel and Gardar (2008) from University of Wollongong

软件:

- Andrew Gelman (BUGS[1]) from Columbia University
- Bob Carpenter et al. (2017) (**Stan**) from Columbia University
- Robert Gentleman and Ross Ihaka (R) from University of Auckland

[1] **B**ayesian inference **U**sing **G**ibbs **S**ampling, WinBUGS/OpenBUGS/JAGS

我做了什么

- 综述和实现了空间广义线性混合效应模型的三类参数估计方法，分别是低秩近似、蒙特卡罗最大似然和近似贝叶斯（创新）
- 在同一模拟数据集上，比较了三类算法的优劣（创新）
- 在同样的准确度下，基于新的计算框架 Stan 实现了贝叶斯马尔科夫链蒙特卡罗算法（创新）
- 指出三类算法实现的关键技巧和使用场景（创新）

有什么意义

- 指导算法的选择和应用：

1. 空间流行疾病预测（如冈比亚和喀麦隆）

模型

高斯空间过程

$\mathcal{S} = \{S(w), w \in \mathbb{R}^2\}$ 是一个随机过程，满足：
任意给定一组位置 $w_1, w_2, \ldots, w_n, w_i \in \mathbb{R}^2$，对应的联合分布
$\mathcal{S} = \{S(w_1), S(w_2), \ldots, S(w_n)\}$ 是多元高斯分布，由均值 $\mathrm{E}[S(w)]$ 和协方差 $\mathrm{Cov}\{S(w_i), S(w_j)\}$ 完全确定

空间广义线性混合效应模型

$$\begin{aligned} g(\mu) &= d(x)'\beta + S(x) \\ \mu &= E[Y|S(X)] \\ Y &\sim \text{指数族} \end{aligned} \tag{1}$$

$S(x)$ 是平稳高斯过程，$d(x)$ 样本点的观测变量，详见论文公式 2.4

数值模拟：响应变量 Y 服从正态分布

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + S(w) + \epsilon \quad (2)$$

- 响应变量 Y 服从正态分布，样本量为 $N = 50$，$\beta_0 = 1.2, \beta_1 = 1, \beta_2 = 0.8, X_1 \sim N(0,1), X_2 \sim N(0,4)$，残差 $\epsilon \sim N(0, \tau^2), \tau = 1$
- $S(w)$ 服从 N 元高斯分布 $N(\mu_S, G)$，$w = (d_1, d_2) \in \mathbb{R}^2$，$d_1 \sim N(0,1), d_2 \sim N(0,1), \mu_S = \mathbf{0}_{N\times 1}$，$G_{(ij)} = \mathrm{Cov}(S(w_i), S(w_j)) = \sigma^2 * \rho(u_{ij})$
- $S(w)$ 的相关函数 $\rho(u_{ij}) = \exp(-u_{ij}/\phi), u_{ij} \equiv \|w_i - w_j\|_2$，$\sigma^2 = 1, \phi = 25$
- 困难：$\dim\{S(w)\} =$ 样本量N，意味着空间随机效应自带高维特点
- 目标：估计参数 $\beta_0, \beta_1, \beta_2, \sigma^2, \tau^2, \phi$

模拟结果

表 1: 正态分布情形下的数值模拟比较

估计	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\phi}$	$\hat{\sigma}^2$	$\hat{\tau}^2$	CPU (s)
真实值	1.20	1.00	0.80	25.00	1.00	1.00	-
Stan	2.75	1.17	0.74	29.58	0.36	0.69	117.17
ML/REML	2.75	1.18	0.74	10.13	0.97	0.00	3.68

模拟实验 I

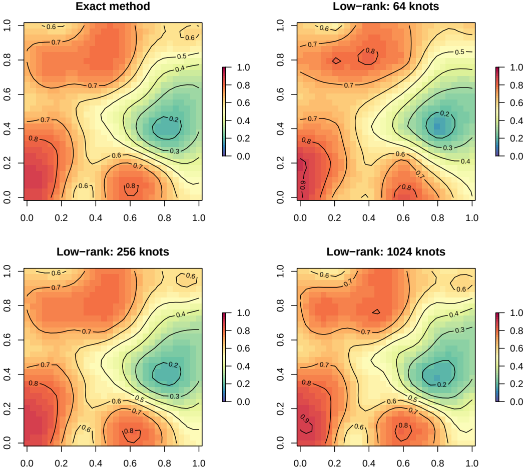

图 1: 精确方法 MCML 与近似方法 Low-Rank 比较

模拟实验 II

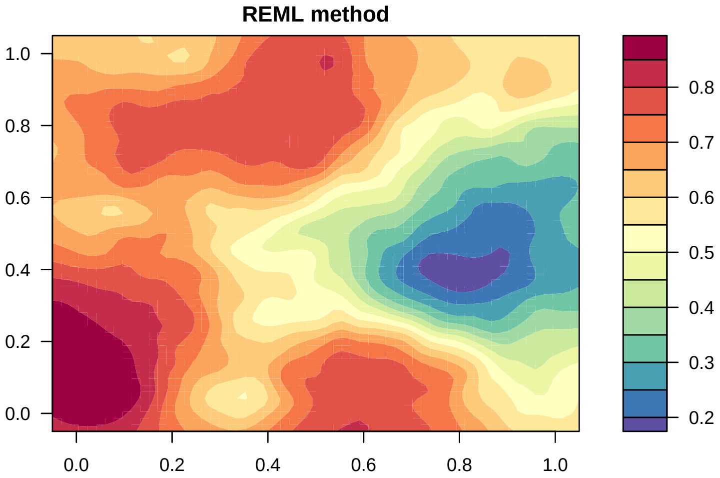

图 2: ML/REML 方法

模拟实验 III

表 2: 正态分布情形下的数值模拟比较

估计	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\phi}$	$\hat{\sigma}^2$	$\hat{\tau}^2$	CPU (s)
真实值	1.200	1.000	0.800	25.000	1.000	1.000	-
RSA	1.977	1.016	0.803	21.937	0.857	0.960	298.250
ML/PQL	1.966	1.007	0.796	28.172	1.365	0.516	464.420
ML/REML	1.958	1.007	0.796	38.114	1.159	0.970	634.720
ML/LR	1.935	1.008	0.796	44.317	3.916	0.264	326.780

RSA 不仅计算效率高，而且也比较准确，PQL 在空间效应的参数估计中效果不及 REML，LR 牺牲一些计算精度可以大大缩短运行时间。

案例 I

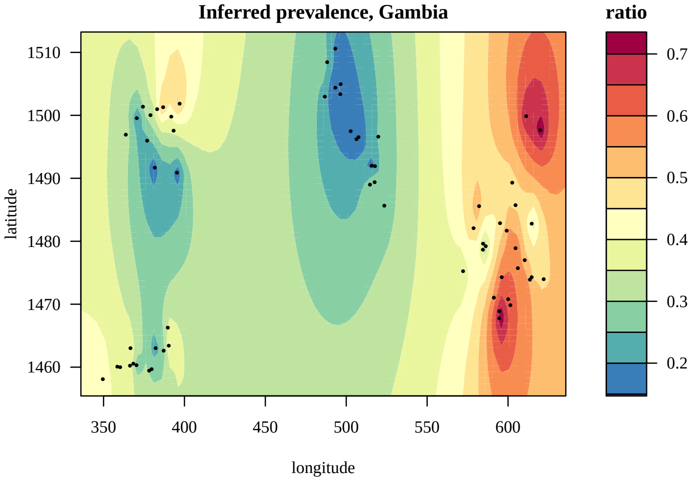

图 3: ML/REML 方法

案例 II

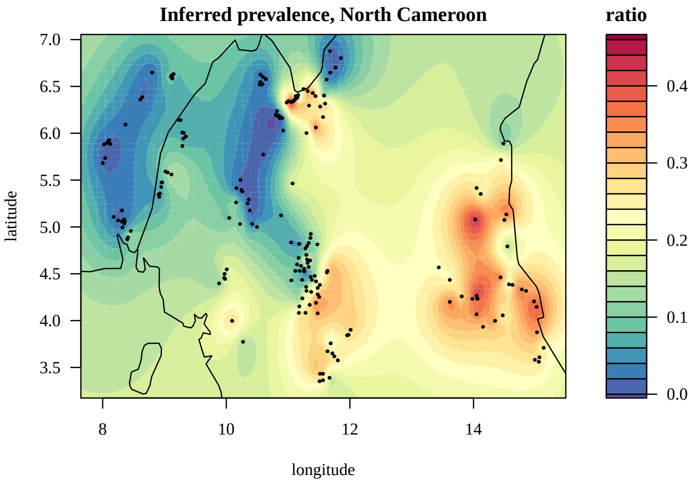

图 4: ML/REML 方法

参考文献 I

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society, Series C*, 47(3):299–350.

Geyer, C. J. (1994). On the convergence of monte carlo maximum likelihood calculations. *Journal of the Royal Statistical Society, Series B*, 56(1):261–274.

Noel, C. and Gardar, J. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226.

参考文献 II

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models using integrated nested laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, 71(2):319–392.

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