空间广义线性混合模型及其应用

黄湘云

目录

1 绪论

1.1 研究意义

1.2 文献综述

P. J. Diggle, Tawn, and Moyeed (1998) 提出基于模型的地统计学框架，将高斯空间随机过程和（广义）线性混合模型结合应用到空间流行病数据分析中，通过贝叶斯推断方法进行参数估计和预测。

P. Diggle et al. (2002) 提出空间广义线性混合模型分析冈比亚儿童疟疾的数据，在贝叶斯框架下，通过 Metropolis-Hastings 算法实现 MCMC（马尔科夫链蒙特卡罗）方法进行参数估计和模型预测。

数据如下：

从冈比亚的 5 个区域，65 个村庄，共采集 2035 个 5 岁以下儿童的血液样本，并记录儿童的年龄，村庄的位置坐标，血液样本中是否含有疟疾寄生虫，儿童是否睡在蚊帐中，是否使用杀虫剂对蚊帐杀虫，村庄附近的绿色植物的绿色度，村庄是否有健康中心。

变量如下：

1. (x, y) 村庄的坐标
2. pos 血样中是否出现疟疾（1 有 0 否）
3. age 儿童的年龄（按天计算）
4. netuse 儿童是否睡在蚊帐中（1 是 0 否）
5. treated 蚊帐是否消毒（1 是 0 否）

6. green 村庄附近的绿色植物的绿色度
7. phc 村庄是否有健康中心（1 是 0 否）

模型如下：

$$\log\{p_{ij}/(1-p_{ij})\} = \alpha + \beta' z_{ij} + U_i + S(x_i)$$

1. the effects of child level covariates(age and bed net use) 儿童的年龄（以天计）和蚊帐的使用情况（是否使用蚊帐和蚊帐是否杀虫），包含在 Z_{ij} 中；
2. village level covariates (the primary health care and greenness of surrounding vegetation) 村庄是否有初级卫生保健和周围植被绿色度，包含在 Z_{ij} 中；
3. separate components for residual spatial 平稳高斯过程 $S(x)$ 表示剩余的空间成分；
4. non-spatial extrabinomial variation 村庄水平上的非空间的额外二项变异 U_i。

说明：

1. $\mathcal{S} = \{S(x) : x \in \mathbb{R}^2\}$ 是均值为 0，方差 σ^2，相关函数 $\rho(x,x') = \mathrm{Corr}\{S(x), S(x')\}$ 的高斯过程；
2. 假定过程 $\mathcal{S}$ 是平稳且各向同性，则 $\mathrm{Corr}\{S(x), S(x')\} = \rho(\|x,x'\|)$，$\|\cdot\|$ 表示欧氏距离；
3. Z_{ij} 是第 i 个村庄的第 j 个儿童的观测值；
4. $S(x_i)$ 是与空间相关的随机效应；
5. U_i 相互独立且服从 $N(0,\tau^2)$ 的随机变量（效应）。

Mat*é*rn 参数族的选择问题：$\rho(u) \triangleq \rho(\|x,x'\|)$

1. 一般假设 $\rho(u)$ 单调不增，尺度参数 ϕ 控制 $\rho(u)$ 递减到 0 的速率，因此 $\rho(u) = \rho_0(u/\phi)$
2. 一个经验模型是
$$\rho_0(u) = \frac{1}{2^{\delta-1}\Gamma(\delta)} u^{\delta} \kappa_{\delta}(u)$$
其中 $\kappa_\delta(\cdot)$ 是阶数为 δ 的第二类修正的贝塞尔函数，$\delta > 0$ 是平滑参数，具有这样的相关函数，过程 $S(x)$ 是 $\lceil \delta \rceil - 1$ 次均方可微。
3. Mat*é*rn 参数族包含指数族，即当 $\delta = 0.5$ 时，$\rho_0(u) = \exp(-u)$，$S(x)$ 均方连续但不可微，当 $\delta \to \infty$ 时，$\rho_0(u) = \exp(-u^2)$，$S(x)$ 无限次均方可微。要从数据中估计 δ，为了节省计算，又不失一般性，取离散的 δ 先验，如 $\delta = 0.5, \delta = 1.5, \delta = 2.5$，分别对应 $S(x)$ 均方连续、一次可微和二次可微。
4. P. J. Diggle, Tawn, and Moyeed (1998) 使用幂指数族 $\rho_0(u) = \exp(-u^{\delta}), 0 < \delta \le 2$，其与 Mat*é*rn 参数族形状相似，且当 $0 < \delta < 2$ 时，$S(x)$ 均方连续但不可微，当 $\delta = 2$ 时，$S(x)$ 无限次均方可微。

模型中 U_i 与 $S(x_i)$ 项的可识别问题：令 $T_i = U_i + S(x_i)$，向量 T 是协方差为矩阵 $\nu^2 I + \sigma^2 R$ 的多元高斯分布，其中 $R_{ij} = \rho(u_{ij}, \phi)$，$u_{ij}$ 是 x_i 与 x_j 之间的距离，从而随机过程 $T(x)$ 的相关函

数在原点不连续。只要指定参数，使得 $\rho(u)$ 在原点连续，则参数 ν^2, σ^2, ϕ 就都是可识别的，显然这依赖于抽样的位置 x_i。

贝叶斯统计分析方法中，参数 $\boldsymbol{\theta} = \alpha, \beta, \nu^2, \sigma^2, \phi$ 先验分布的选择问题：为了使用 MCMC 算法实现贝叶斯推断，需要先指定参数 $\boldsymbol{\theta}$ 的先验分布，对 α, β，选择独立的不适当（真实的先验谁也不知道，也没有理论结果）的均匀先验。对于参数 ν^2, σ^2, ϕ，选取如下模糊先验：$f(\nu^2) \propto 1/\nu^2; f(\sigma^2) \propto 1/\sigma^2; f(\phi) \propto 1/\phi^2$，其中 ν^2 和 σ^2 为杰弗里斯先验（Jeffreys priors）[1]，这些模糊先验的选择是出于实用的考虑，如果由这些模糊先验导出的后验不合适，则 MCMC 算法将会不收敛，通过选取不同的初始值，而没有出现算法不收敛，则这样的先验是被允许的。

Christensen (2004) 将 MCML 方法应用到空间广义线性混合模型的参数估计和预测，通过对真实数据 Rongelap 的分析

$$\log \mu_i = \log t_i + \beta + S(x_i)$$

1. $Y_i|S(\cdot) \sim \mathrm{Poiss}(\mu_i)$
2. $S(\cdot)$ 是平稳高斯过程

Peter J. Diggle and Ribeiro Jr (2007) 将经典的广义线性模型 P. McCullagh (1989) 、混合模型 Breslow and Clayton (1993) 和贝叶斯推断方法统一到基于模型的地统计框架下，提出广义线性地统计模型。

Eidsvik, Martino, and Rue (2009) 近似贝叶斯推断方法在空间广义线性混合模型中的应用

Peter J. Diggle and Giorgi (2016) 将 MCML 方法应用于空间广义线性混合模型的参数估计和预测，分析喀麦隆及周边地区的 Loa loa 疟疾流行情况（Crainiceanu, Diggle, and Rowlingson (2008) 已研究），还考虑并解决了以下三类问题：

1. 组合随机调查数据和非随机调查数据（即潜在有偏的数据），以肯尼亚疟疾流行数据为例，组合了学校和社区的调查数据
2. 时空扩展，即将时间因素考虑进模型，以马拉维 2010 年 5 月至 2013 年 6 月的疟疾流行数据为例，

$$\begin{aligned}\log[p_j(x_i,t_i)/\{1-p_j(x_i,t_i)\}] =& \beta_0 + \beta_1 Z_{ij} + \beta_2 t_i + \beta_3 \sin(2\pi t_i/12) \\ &+ \beta_4 \cos(2\pi t_i/12) + \beta_5 \sin(2\pi t_i/6) \\ &+ \beta_6 \cos(2\pi t_i/6) + S(x_i) + U(t_i)\end{aligned}$$

3. 考虑 zero-inflation，即响应变量 Y_i 是混合二项分布

$$P(Y_i = y|S(x_i)) = \begin{cases} [1-\pi(x_i)] + \pi(x_i)\mathrm{Bin}(0; m_i, p(x_i)) & \text{if} \quad y = 0, \\ \pi(x_i)\mathrm{Bin}(y; m, p(x_i)) & \text{if} \quad y > 0. \end{cases}$$

[1] https://en.wikipedia.org/wiki/Jeffreys_prior

1.3 主要内容

在 MCML 算法、贝叶斯实现算法 MCMC-Metropolis-Hastings 或近似贝叶斯推断方法中寻求改进，并基于 Giorgi and Diggle (2016) 提出的 PrevMap 包、John Salvatier and Fonnesbeck (2016) 提出的 PyMC3 软件或 Rue, Martino, and Chopin (2009) 提出的 R-INLA 软件实现

参考文献

Breslow, N. E., and D. G. Clayton. 1993. "Approximate Inference in Generalized Linear Mixed Models." *Journal of the American Statistical Association* 88 (421): 9–25. doi:10.1080/01621459.1993.10594284.

Christensen, Ole F. 2004. "Monte Carlo Maximum Likelihood in Model-Based Geostatistics." *Journal of Computational and Graphical Statistics* 13 (3): 702–18. doi:10.1198/106186004X2525.

Crainiceanu, Ciprian M, Peter J Diggle, and Barry Rowlingson. 2008. "Bivariate Binomial Spatial Modeling of Loa Loa Prevalence in Tropical Africa." *Journal of the American Statistical Association* 103 (481): 21–37. doi:10.1198/016214507000001409.

Diggle, P. J., J. A. Tawn, and R. A. Moyeed. 1998. "Model-Based Geostatistics." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 47 (3). Blackwell Publishers Ltd.: 299–350. doi:10.1111/1467-9876.00113.

Diggle, Peter J., and Emanuele Giorgi. 2016. "Model-Based Geostatistics for Prevalence Mapping in Low-Resource Settings." *Journal of the American Statistical Association* 111 (515): 1096–1120. doi:10.1080/01621459.2015.1123158.

Diggle, Peter J., and Paulo J. Ribeiro Jr. 2007. *Model-Based Geostatistics*. Springer-Verlag New York.

Diggle, Peter, Rana Moyeed, Barry Rowlingson, and Madeleine Thomson. 2002. "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. doi:10.1111/1467-9876.00283.

Eidsvik, Jo, Sara Martino, and Havard Rue. 2009. "Approximate Bayesian Inference in Spatial Generalized Linear Mixed Models." *Scandinavian Journal of Statistics* 36 (1). Blackwell Publishing Ltd: 1–22. doi:10.1111/j.1467-9469.2008.00621.x.

Giorgi, Emanuele, and Peter J. Diggle. 2016. *PrevMap: Geostatistical Modelling of Spatially Referenced Prevalence Data*. https://CRAN.R-project.org/package=PrevMap.

John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. 2016. "Probabilistic Programming in Python Using Pymc3." *PeerJ Computer Science*. doi:10.7717/peerj-cs.55.

P. McCullagh, J. A. Nelder FRS. 1989. *Generalized Linear Models*. Monographs on Statistics and Applied Probability. Springer US.

Rue, Havard, Sara Martino, and Nicolas Chopin. 2009. "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71 (2): 319–92.