空间广义线性混合效应模型及其应用

Spatial Generalized Linear Mixed Models with Its Applications

指导老师：李再兴　　答辩学生：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

绪论

- 文献综述分三段
  - 引出估计 SGLMM 模型参数的问题，实现其参数估计的算法的研究现状
  - SGLMM 模型的应用现状
  - 估计 SGLMM 模型参数在实践中遇到的瓶颈
- 论文结构
  - 论文后续章节的组织
  - 陈述自己的创新点或补充文献中的内容

空间广义线性混合效应模型是什么？

空间广义线性混合效应模型

$$g(\mu_i) = d(x_i)^\top \beta + S(x_i) + Z_i \tag{1}$$

- 观测数据向量：$d^\top(x_i)$ 表示 p 个协变量在第 i 个位置 x_i 的观察值。
- 回归参数向量：β
- 假定 $\mathcal{S} = \{S(x) : x \in \mathbb{R}^2\}$ 是均值为 0，方差为 σ^2，平稳且各向同性的空间高斯过程，$\rho(x, x') = \mathrm{Corr}\{S(x), S(x')\} \equiv \rho(\|x, x'\|)$，$\|\cdot\|$ 表示距离。
- 非空间随机效应：$Z_i \overset{i.i.d}{\sim} \mathcal{N}(0, \tau^2)$，$\tau^2 = \mathrm{Var}(Y_i|S(x_i)), \forall i = 1, 2, \ldots, N$，$N$ 是采样点的数目
- 联系函数：

与一般的广义线性混合混合效应模型区别在哪？

$S(x_i)$ 是与空间位置 x_i 相关的随机效应。

空间效应 $S(x_i)$ 给参数估计带来的困难在哪？

$$g(\mu_i) = d(x_i)^\top \beta + S(x_i) + Z_i \tag{2}$$

随机效应的协方差矩阵结构复杂，卷入的参数有 $\sigma^2, \tau^2, \phi, \kappa$

$$\text{Cov}(T_i(x), T_i(x)) = \sigma^2 + \tau^2, \text{Cov}(T_i(x), T_j(x)) = \sigma^2 \rho(u_{ij})$$

$$\rho(u) = \{2^{\kappa-1}\Gamma(\kappa)\}^{-1}(u/\phi)^\kappa \mathcal{K}_\kappa(u/\phi), u > 0 \tag{3}$$

$\mathcal{K}_\kappa(u)$ 是修正的第二类贝塞尔函数

$$I_{-\kappa}(u) = \sum_{m=0}^{\infty} \frac{1}{m!\Gamma(m+\kappa+1)} \left(\frac{u}{2}\right)^{2m+\kappa}$$
$$\mathcal{K}_\kappa(u) = \frac{\pi}{2} \frac{I_{-\kappa}(u) - I_\kappa(u)}{\sin(\kappa\pi)} \tag{4}$$

估计 SGLMM 模型参数的困难

设研究区域 $D \subseteq \mathbb{R}^2$，对于第 i 次观测，s_i 表示区域 D 内的位置，$y(s_i)$ 表示响应变量，$\mathbf{x}(s_i), i = 1, \ldots, n$ 是一个 p 维的固定效应，定义如下的 SGLMM 模型：

$$\mathrm{E}[y(s_i)|u(s_i)] = g^{-1}[\mathbf{x}(s_i)^\top \beta + \mathbf{u}(s_i)], \quad i = 1, \ldots, n$$

其中 $g(\cdot)$ 是实值可微的逆联系函数，β 是 p 维的回归参数向量，代表 SGLMM 模型的固定效应。随机过程 $\{U(\mathbf{s}) : \mathbf{s} \in D\}$ 是平稳的空间高斯过程，其均值为 $\mathbf{0}$，自协方差函数 $\mathrm{Cov}(U(\mathbf{s}), U(\mathbf{s}')) = C(\mathbf{s} - \mathbf{s}'; \theta)$，$\theta$ 是其中的参数向量。
$\mathbf{u} = (u(s_1), u(s_2), \ldots, u(s_n))^\top$ 是平稳空间高斯过程 $U(\cdot)$ 的一个实例。给定 $\mathbf{u}$ 的情况下，观察值 $\mathbf{y} = (y(s_1), y(s_2), \ldots, y(s_n))^\top$ 是相互独立的。

边际似然函数含有空间随机效应带来的高维积分

给定 $u_i = u(s_i), i = 1, \ldots, n$ 的条件下，$y_i = y(s_i)$ 的条件概率密度函数是

$$f(y_i|u_i;\beta) = \exp[a(\mu_i)y_i - b(\mu_i)]c(y_i)$$

其中 $\mu_i = \mathrm{E}(y_i|u_i)$，$a(\cdot), b(\cdot)$ 和 $c(\cdot)$ 是特定的函数。SGLMM 模型的边际似然函数

$$L(\psi;\mathbf{y}) = \int \prod_{i=1}^{n} f(y_i|u_i;\beta)\phi_n(\mathbf{u};0,\Sigma_\theta)\mathrm{d}\mathbf{u} \tag{5}$$

记号 $\psi = (\beta, \theta)$ 表示 SGLMM 模型的全部参数，$\phi_n(\cdot;0,\Sigma_\theta)$ 表示 n 元正态密度函数，其均值为 $\mathbf{0}$，协方差矩阵为 $\Sigma_\theta = (c_{ij}) = (C(s_i - s_j;\theta)), i, j = 1, \ldots, n$。

困难

边际似然函数几乎总是卷入一个难以处理的积分，积分的维数等于观测点的个数。

估计 SGLMM 模型参数的算法主要有哪些？

- 基于似然函数的估计算法
  - 拉普拉斯近似算法 (Bonat and Ribeiro Jr., 2016)
  - 蒙特卡罗极大似然算法 (Diggle and Giorgi, 2016)
- 基于马尔科夫链蒙特卡罗的估计算法
  - R 实现的 Langevin-Hastings 算法 (Giorgi and Diggle, 2017)
  - 新提出：Stan 实现的汉密尔顿蒙特卡罗算法（简称 Stan-HMC）

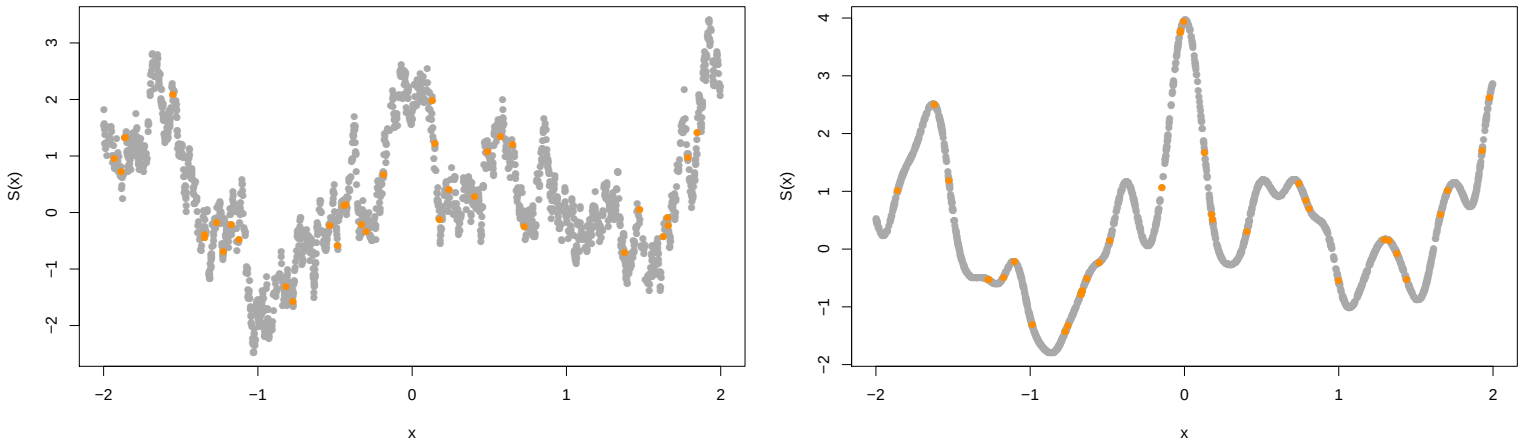
创新点在哪？怎么做的？效果怎么样？

1. 借助 Stan 实现汉密尔顿蒙特卡罗算法（简称 HMC）去估计 SGLMM 模型的参数，在响应变量服从二项分布和泊松分布的两组模拟实验中，与基于 R 包 geoRglm 实现的 Langevin-Hastings 算法相比，发现 HMC 算法在保持相似结果下能大大减少迭代次数，还不需要对算法进行调参；
2. 在真实数据分析中研究了基于似然函数的参数估计算法，发现这类算法容易陷入局部极值，因此，在小麦数据的分析中借助样本变差图选择初值，在核污染数据的分析中利用剖面似然轮廓来确定合适的初值。

模拟实验 I: 模拟平稳空间高斯过程 $S(x)$

自协方差函数：

$$\mathrm{Cov}(S(x_i), S(x_j)) = \sigma^2 \exp\left\{ -\left(\frac{|x_i - x_j|}{\phi}\right)^{\kappa}\right\}, 0 < \kappa \leq 2 \tag{6}$$



(a) 平稳空间高斯过程 $S(x)$ 的协方差函数是指数型，均值向量为 $\mathbf{0}$，协方差参数 $\sigma^2 = 1$，$\phi = 0.15$，$\kappa = 1$

(b) 平稳空间高斯过程 $S(x)$ 的协方差函数是幂二次指数型，均值向量为 $\mathbf{0}$，协方差参数 $\sigma^2 = 1$，$\phi = 0.15$，$\kappa = 2$

图 1: 模拟平稳空间高斯过程 $S(x)$

模拟实验 II: 模拟 SGLMM 模型（二项型）

$$g(\mu_i) = \log\{\frac{p(x_i)}{1-p(x_i)}\} = \alpha + S(x_i) \tag{7}$$

- 固定效应参数 $\alpha = 0$，
- 协方差参数 $\boldsymbol{\theta} = (\sigma^2, \phi) = (0.5, 0.2)$，
- 采样点数目 $N = 36, 64, 81$，每个采样点抽取的样本数为 4，

模拟分两步走：

- 第一步：模拟平稳空间高斯过程 $S(x)$，在单位区域 $[0,1] \times [0,1]$ 划分为 8×8 的网格，格点选为采样位置，用 geoR 包提供的 `grf` 函数产生协方差参数为 $\boldsymbol{\theta} = (\sigma^2, \phi) = (0.5, 0.2)$ 的平稳空间高斯过程
- 第二步：根据 $p(x_i) = \exp[\alpha + S(x_i)]/\{1 + \exp[\alpha + S(x_i)]\}$，获得每个格点处二项分布的概率值，由 `rbinom` 函数产生服从二项分布的观察值 Y_i

模拟实验 II: 模拟 SGLMM 模型（二项型）续

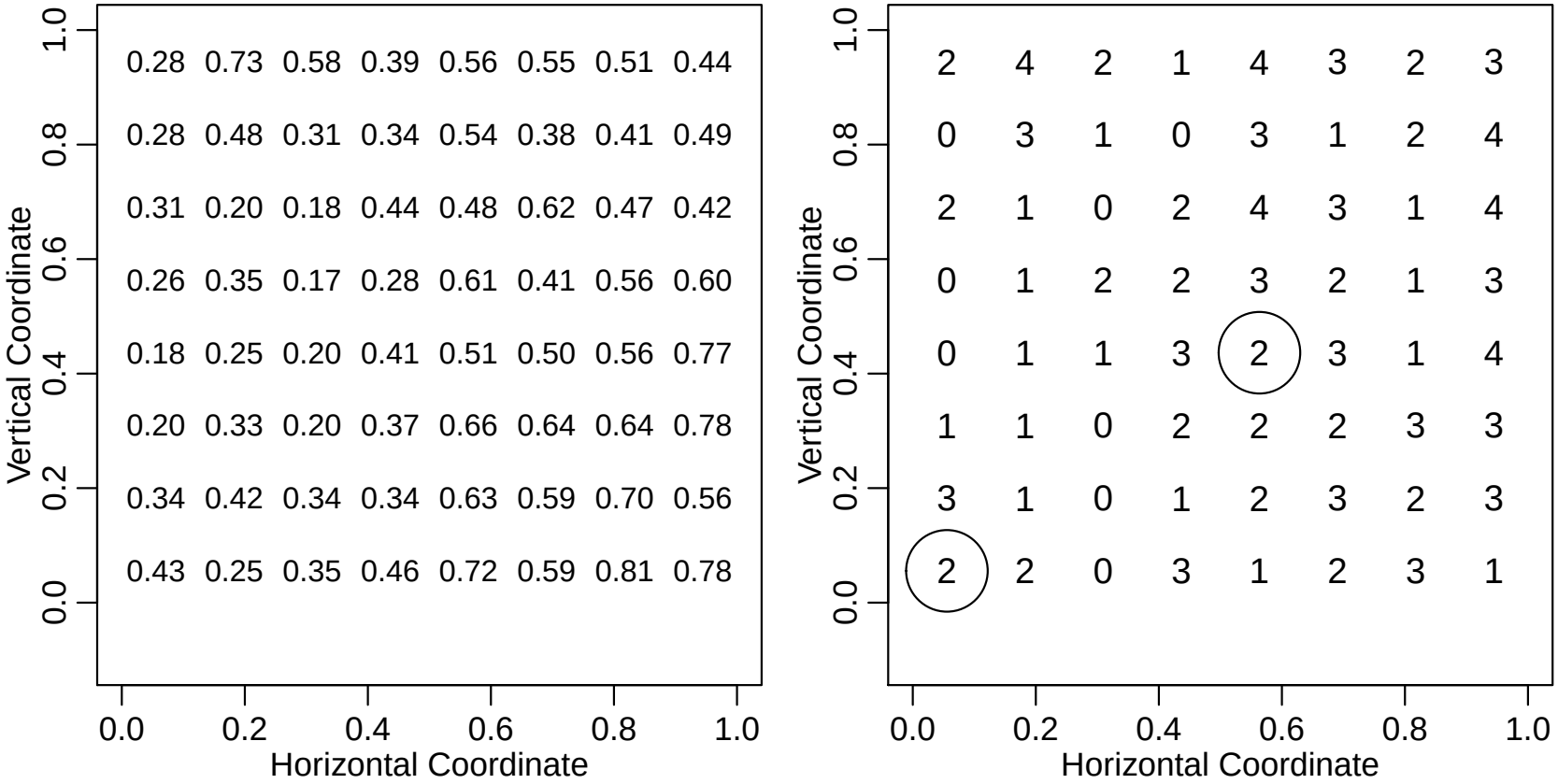


图 2: 左图表示二维规则平面上的平稳空间高斯过程，格点是采样点的位置，其上的数字是 $p(x)$ 的值，已经四舍五入保留两位小数，右图表示观察值 Y 随空间位置的变化，格点上的值即为观察值 Y，图中的两个圈分别是第 1 个 (左下) 和第 29 个 (右上) 采样点

模拟实验 II: 模拟 SGLMM 模型 (二项型) 续 : Langevin-Hastings 算法与 HMC 算法比较

表 5.1: 在模型(5.4)的设置下，Langevin-Hastings 算法与 HMC 算法的数值模拟比较

	true(init)	mean	var	2.5%	25%	50%	75%	97.5%	N	time(s)
α	0.0(0.387)	-0.354	0.079	-0.938	-0.524	-0.361	-0.173	0.215	36	600.12
ϕ	0.2(0.205)	0.121	0.006	0.005	0.055	0.110	0.180	0.285		
σ^2	0.5(1.121)	0.683	0.147	0.215	0.408	0.596	0.850	1.667		
α	0.0(0.157)	0.003	0.089	-0.596	-0.169	0.013	0.179	0.609	64	729.19
ϕ	0.2(0.110)	0.194	0.004	0.070	0.145	0.195	0.250	0.295		
σ^2	0.5(0.494)	0.656	0.096	0.254	0.449	0.592	0.781	1.453		
β	0.0(-0.006)	-0.155	0.044	-0.565	-0.284	-0.156	-0.03	0.273	81	844.56
ϕ	0.2(0.185)	0.116	0.006	0.005	0.055	0.105	0.17	0.280		
σ^2	0.5(0.403)	0.468	0.057	0.180	0.311	0.414	0.56	1.129		
α	0.0(-0.813)	-0.230	0.209	-1.127	-0.521	-0.214	0.056	0.653	36	6.65
ϕ	0.2(1.692)	1.103	0.364	0.459	0.721	0.936	1.284	2.669		
σ^2	0.5(0.144)	0.474	0.187	0.105	0.216	0.333	0.573	1.572		
α	0.0(0.155)	0.046	0.251	-0.947	-0.269	0.049	0.356	1.069	64	27.70
ϕ	0.2(1.766)	1.042	0.246	0.471	0.708	0.921	1.247	2.324		
σ^2	0.5(0.808)	0.647	0.228	0.170	0.338	0.524	0.779	1.958		
α	0.0(-0.369)	-0.082	0.170	-0.893	-0.321	-0.078	0.174	0.742	81	45.69
ϕ	0.2(0.911)	1.110	0.331	0.453	0.721	0.986	1.330	2.506		

模拟实验 II: 模拟 SGLMM 模型（泊松型）

$$g(\mu_i) = \log[\lambda(x_i)] = \alpha + S(x_i) \tag{8}$$

- 模型参数真值为 $\alpha = 0.5, \phi = 0.2, \sigma^2 = 2.0, \kappa = 1.5$,
- 采样点数目分别为 $N = 36, 64, 100$

模拟分两步走：

- 第一步：首先产生服从平稳空间高斯过程 $S(x)$ 的随机数 $S(x_i), i = 1, \ldots, N$,
- 第二步：因为 $\lambda(x_i) = \exp(\alpha + S(x_i))$, 且响应变量 $Y_i \sim \mathrm{Poisson}(\lambda(x_i))$, 根据 R 内置函数 `rpois` 即可产生服从参数为 $\lambda(x_i)$ 的泊松分布的随机数。

模拟实验 II: 模拟 SGLMM 模型 (泊松型) 续 : Langevin-Hastings 算法与 HMC 算法比较

表 5.2: 在模型(5.6)的设置下，Langevin-Hastings 算法和 HMC 算法的比较

	true(init)	mean	var	2.5%	25%	50%	75%	97.5%	N	time(s)
α	0.5(1.201)	0.527	0.418	-0.759	0.189	0.514	0.855	1.864	36	642.66
ϕ	0.2(0.420)	0.401	0.052	0.100	0.240	0.360	0.520	0.960		
σ^2	2.0(1.038)	1.311	0.660	0.365	0.766	1.081	1.584	3.562		
α	0.5(1.211)	0.866	1.517	-1.610	0.059	0.870	1.666	3.159	64	883.76
ϕ	0.2(0.480)	0.682	0.073	0.300	0.480	0.640	0.820	1.380		
σ^2	2.0(2.232)	3.932	2.594	1.667	2.800	3.642	4.744	7.740		
α	0.5(0.189)	0.323	0.657	-1.449	-0.124	0.416	0.812	1.831	100	1223.28
ϕ	0.2(0.540)	0.617	0.085	0.220	0.400	0.560	0.785	1.320		
σ^2	2.0(1.395)	1.479	0.498	0.545	0.941	1.352	1.822	3.195		
α	0.5(0.335)	0.483	0.310	-0.608	0.094	0.488	0.851	1.613	36	11.25
ϕ	0.2(0.066)	0.631	0.036	0.362	0.501	0.602	0.722	1.090		
σ^2	2.0(0.347)	1.370	0.298	0.455	0.977	1.317	1.714	2.566		
α	0.5(1.021)	0.498	0.402	-0.775	0.082	0.534	0.917	1.798	64	113.04
ϕ	0.2(0.370)	0.385	0.003	0.285	0.343	0.380	0.422	0.509		
σ^2	2.0(2.610)	2.473	0.292	1.585	2.102	2.416	2.804	3.734		
α	0.5(0.613)	0.400	0.297	-0.723	0.062	0.415	0.767	1.412	100	272.58
ϕ	0.2(0.294)	0.299	0.005	0.181	0.243	0.289	0.343	0.465		

数据分析：以核污染数据分析为例

表 1: 拉普拉斯近似算法（简记 LAL）和蒙特卡罗极大似然算法（简记 MCL）估计泊松型 SGLMM 模型的参数

算法	$\hat{\beta}(\beta_0)$	$\hat{\sigma}^2(\sigma_0^2)$	$\hat{\phi}(\phi_0)$	$\hat{\tau}^2_{rel}(\tau^2_{rel_0})$	$\log L_m$
LAL	1.821(2.014)	0.264(0.231)	151.795(50)	0.133(0.1)	−1317.195
MCL	1.821(2.014)	0.265(0.231)	151.859(50)	0.132(0.1)	−8.8903
MCL	6.190(6.200)	2.401(2.400)	338.126(340)	2.053(2.074)	−5.8458

以第 4 行为例，块金效应的估计值应为 $\hat{\tau}^2 = \hat{\sigma}^2 \times \hat{\tau}^2_{rel} = 4.929$

数据分析：以核污染数据分析为例（续）

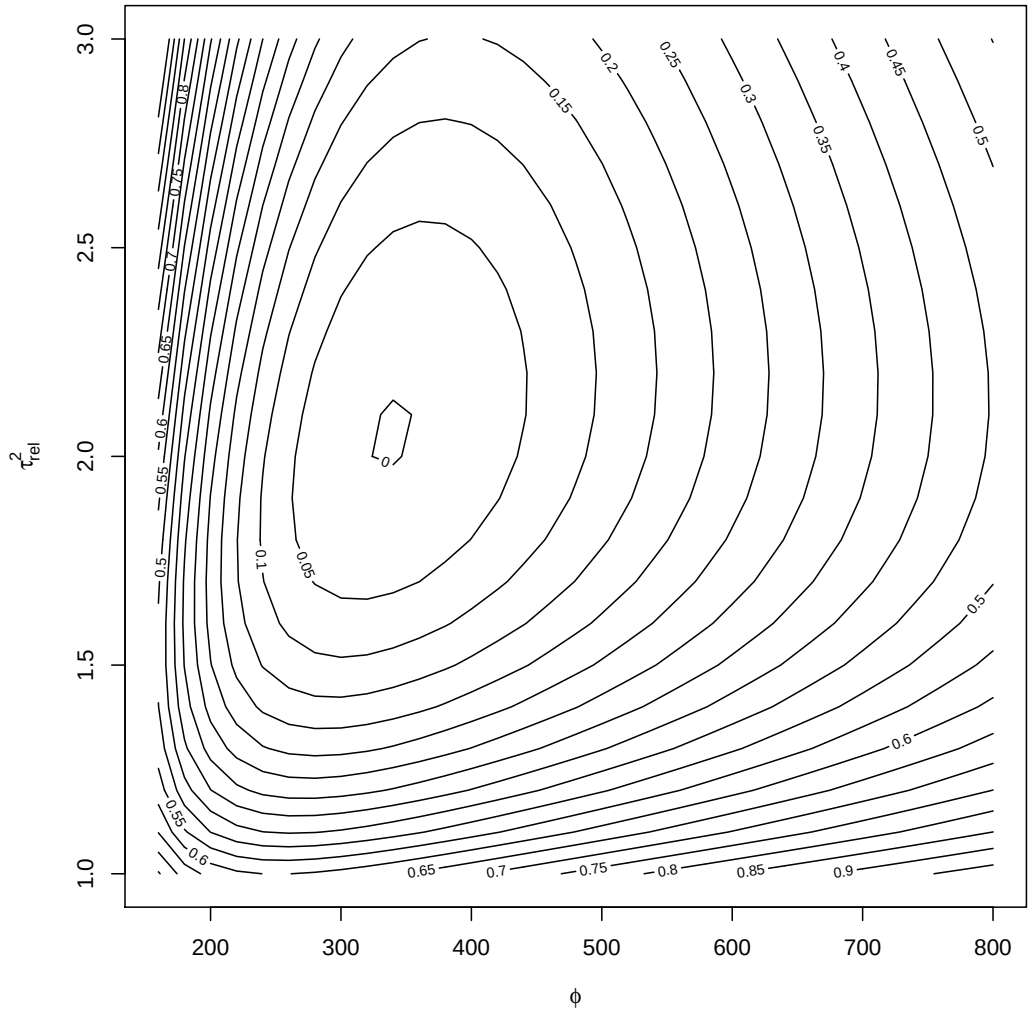


图 3: 泊松型 SGLMM 模型关于 ϕ 和相对块金效应 $\tau^2_{rel} = \tau^2/\sigma^2$ 的剖面似然函数轮廓

致谢

谢谢！

参考文献 I

Bonat, W. H. and Ribeiro Jr., P. J. (2016). Practical likelihood analysis for spatial generalized linear mixed models. *Environmetrics*, 27(2):83–89.

Diggle, P. J. and Giorgi, E. (2016). Model-based geostatistics for prevalence mapping in low-resource settings. *Journal of the American Statistical Association*, 111(515):1096–1120.

Giorgi, E. and Diggle, P. J. (2017). PrevMap: An R package for prevalence mapping. *Journal of Statistical Software*, 78(8):1–29.