DOI:10.13546/j.cnki.tjyjc.2017.22.019

贝叶斯后验的快速计算方法

闵素芹[1],李　群[2]

(1.中国传媒大学 理学院,北京 100024;2.中国社会科学院 数量经济与技术经济研究所,北京 100732)

摘　要:贝叶斯推断的主要障碍是后验分布积分计算,INLA方法通过对边缘后验密度的精确近似克服了计算时间的问题,而且得到的计算结果与MCMC的精度几乎无异。文章首先对INLA方法的计算进行介绍,然后通过综述INLA与MCMC的比较研究阐述了INLA方法应用的优越性,并给出一个运用INLA进行参数估计的应用实例。

关键词:贝叶斯推断;后验边缘;INLA;MCMC

中图分类号:O212.8　　**文献标识码:**A　　**文章编号:**1002-6487(2017)22-0084-03

0　引言

贝叶斯推断需要计算后验分布的积分,而后验分布往往是复杂的、高维的、非标准形式的分布,积分难以计算。MCMC方法是常用的一种行之有效的贝叶斯计算方法,该方法渐进精确,但是存在运行速度相对较慢的问题,特别是在大数据背景下很难适应新的模型的要求。

INLA是由Rue等(2009)[1]针对结构可加回归模型提出的一种近似贝叶斯推断的快速方法,它基于近似数值积分计算贝叶斯后验边缘。该方法的估计精度与MCMC方法类似,但在计算时间上远远快于MCMC方法,对于INLA需要几秒到几分钟的问题,MCMC通常需要几分钟到几小时。而且R-INLA程序包目前已比较成熟,应用方便。

INLA方法提出后由于其运算速度上突出的优越性得到了广泛关注,应用领域包括广义线性模型、平滑样条模型、半参数回归、空间和时空模型、地统计模型等。例如,Martino等(2010)[2]运用INLA对金融时间序列的随机波动模型(SV模型)进行推断,基于1217天SP500指数日收盘数据和1292天微软收盘价数据进行实际数据分析。Fong等(2010)[3]运用INLA方法对广义线性混合模型(GLMM)的贝叶斯推断进行了研究。Yu和Rue(2011)[4]运用INLA方法对分位数τ=0.25,0.50和0.75分别进行参数估计,针对慕尼黑房屋租赁数据,因变量为每平米的租金,模型包括13个自变量的线性效应、面积和建筑年份的非线性效应、以及房子位置的空间效应。Natário等(2014)[5]利用INLA方法运用时空分层模型对葡萄牙森林火灾数据进行分析。Gómez-Rubio(2015)[6]利用INLA方法对空间计量模型参数进行推断,分析波士顿房产数据。

近几年,INLA方法在应用领域突显出优越的快速计算能力。本文对其计算方法、优点进行阐述,并给出一个应用实例,旨在为INLA方法的进一步研究和应用提供借鉴。

1　INLA方法

INLA算法由Rue等(2009)[1]提出,该算法应用于一类应用广泛的结构可加回归模型的贝叶斯推断。模型的因变量假定服从指数分布族,均值μ_i通过一个连接函数$g(\cdot)$连接到一个结构可加的预测变量η_i,即:$g(\mu_i)=\eta_i$。

$$\eta_i=\alpha+\sum_{j=1}^{n_f}f^{(j)}(u_{ji})+\sum_{k=1}^{n_\beta}\beta_k z_{ki}+\varepsilon_i \quad (1)$$

其中,$\{f^{(j)}(\cdot)\}$是自变量u的函数,可选用不同的形式,如:自回归模型、季节效应、空间效应等,$\{\beta_k\}$代表自变量z的线性效应,ε_i是非结构项。

结构可加回归模型是一类应用非常广泛的模型,包括广义线性模型、平滑样条模型、状态空间模型、半参数回归、空间和时空模型、对数高斯Cox过程和地统计模型等。

假定模型中的α,$\{f^{(j)}\}$,$\{\beta_k\}$,$\{\varepsilon_i\}$具有高斯先验。令向量x表示模型中η_i,α,$\{f^{(j)}\}$,$\{\beta_k\}$这n个变量,$\pi(\cdot|\cdot)$表示参数的条件密度,假定$\pi(x|\theta_1)$是零均值、精度矩阵$Q(\theta_1)$的高斯密度,n_d个观测变量$y=\{y_i:i\in I\}$的分布用$\pi(y|x,\theta_2)$表示,假定$y=\{y_i:i\in I\}$在给定x,θ_2时条件独立,令$\theta=(\theta_1^T,\theta_2^T)$,$\theta$是m维向量。

主要目标是计算后验边缘密度$\pi(x_i|y)$,$\pi(\theta|y)$和$\pi(\theta_j|y)$。

边缘后验密度记为:

$$\pi(x_i|y)=\int\pi(x_i|\theta,y)\pi(\theta|y)d\theta$$

基金项目:北京高等学校"青年英才计划"项目(YETP0611);中国传媒大学优秀中青年教师培养工程(YXJS201330)

作者简介:闵素芹(1978—),女,山东青州人,博士,副教授,研究方向:分层模型、空间统计理论。

李　群(1961—),男,山东聊城人,教授,博士生导师,研究方向:计量经济模型。

$$\pi(\theta_j|y)=\int\pi(\theta|y)d\theta_{-j}$$

INLA方法构建一个嵌套近似：

$$\tilde{\pi}(x_i|y)=\int\tilde{\pi}(x_i|\theta,y)\tilde{\pi}(\theta|y)d\theta \quad (2)$$

$$\tilde{\pi}(\theta_j|y)=\int\tilde{\pi}(\theta|y)d\theta_{-j}$$

通过近似 $\pi(\theta|y)$ 和 $\pi(x_i|\theta,y)$，运用数值积分来计算 $\pi(x_i|y)$ 的近似。INLA方法近似边缘后验分为3步来完成：

（1）基于如下拉普拉斯近似计算 θ 的边缘后验：

$$\tilde{\pi}(\theta|y)\propto\left.\frac{\pi(x,\theta,y)}{\tilde{\pi}_G(x|\theta,y)}\right|_{x=x^*(\theta)} \quad (3)$$

其中，$\tilde{\pi}_G(x|\theta,y)$ 是 x 的完全条件高斯近似，$x^*(\theta)$ 是给定 θ 时 x 的众数。

（2）计算 $x_i|\theta,y$ 密度的拉普拉斯近似：

$$\tilde{\pi}(x_i|\theta,y)=N\{x_i;\mu_i(\theta),\sigma_i^2(\theta)\}$$

其中，$\mu(\theta)$ 是高斯近似的均值，$\sigma^2(\theta)$ 是相应的边缘方差向量。

（3）根据式（2）对 θ 进行数值积分，通过下式将前两步结合起来：

$$\tilde{\pi}(x_i|y)=\sum_k\tilde{\pi}(x_i|\theta_k,y)\tilde{\pi}(\theta_k|y)\Delta_k$$

INLA方法是一种集成的嵌套拉普拉斯近似，联合解析近似和数值积分得到近似后验边缘，可以用来计算后验均值、方差和分位数等。该方法实现了快速计算，代价是解析近似在计算后验概率时可能会有误差。另外，INLA计算的是边缘后验而非联合后验。但模拟研究均表明，在提高计算速度的同时，INLA方法可以保证较高的估计精度。

2 INLA与MCMC的比较

由于MCMC方法是目前贝叶斯推断中应用最为广泛的一种算法，INLA通过与其比较，可以直观的阐明INLA方法的优越性。

在对模型进行贝叶斯推断时需要计算后验分布的积分。MCMC方法的研究为推广贝叶斯推断方法的应用做出了贡献，使得之前不能实施计算的一些统计方法变得比较容易。求普通方法无法得到的后验分布密度常常运用高维积分运算，MCMC方法可以得到一条或几条收敛的马尔可夫链，该马尔可夫链的极限分布就是所需的后验分布。MCMC方法的基本思想是通过建立一个平稳分布为 $p(\theta|x^n)$ 的马尔可夫链来得到 $p(\theta|x^n)$ 的样本，基于这些样本就可以作各种统计推断[7]。

然而，结构可加回归模型运用MCMC方法时暴露出两个缺点，一是 x 中的各分量之间强相关，二是 θ 和 x 也相关，尤其当是n很大时。而INLA方法在收敛问题和运算速度上得到很大改善，该方法提出后，很多学者通过模拟数据或实际数据对其运行时间和估计精度与MCMC方法进行了比较研究。

Paul M等（2010）[8]基于端粒酶数据，运用广义双随机效应meta分析模型，对INLA（运用R软件的INLA程序包）和MCMC（运用SAS软件的PROC NLMIXED）进行比较，INLA运行时间小于0.2秒，而MCMC通过10000次后验抽样（5000次预迭代），运行时间10分钟。通过进一步模拟研究比较两者的运行时间和失败次数（即：产生诸如优化数值计算不收敛所导致的不可靠结果）。每个meta分析选用25个研究，共模拟72种不同的情况，每种情况由1000次meta分析组成。所有分析均在Intel(R) Core(TM) 2 Duo T7200 processor 2.00 GHz进行，INLA方法平均花费只有3.8分钟（2.2 ~6.0分钟），而MCMC方法平均花费42.2分钟（23.1 ~74.7分钟）。对于不收敛率，INLA方法非常稳定，仅失败两次；而MCMC方法收敛问题与meta分析的设计相关，总体来看，不收敛率为1.6%（72000次分析中失败409次），并且72000次中有11037次的相关系数标准误差非常大。另外，两种方法得到的灵敏度（即真正率）和特指度（即真负率）的偏倚和MSE表现非常接近。综合来看，基于MCMC抽样的贝叶斯方法往往需要很长的运行时间，INLA方法快速，并且估计更稳健、具有更高的精度和更小的偏倚。

Yu和Rue（2011）[4]针对包含非线性项和随机效应项的可加混合分位数回归模型对INLA和MCMC进行比较，模拟生成了两个数据集，样本量均为400，重复200次，对每个数据集估计0.10,0.25,0.50,0.75,0.90分位数。MCMC方法基于40000次迭代（15000次预迭代）。采用平均绝对偏差（MADE）对估计进行评估，INLA和MCMC的模拟研究结果几乎无异。并且，当估计极端的分位数时MCMC逊于INLA。为了比较运行时间，模拟了样本量400, 800, 1500, 3000和6000的情况。INLA运行速度明显更快，各样本量下的计算时间分别为1.01秒、1.97秒、3.84秒、8.25秒、18.09秒；而MCMC方法相应为22.64秒、46.82秒、87.89秒、172.29秒、338.99秒，MCMC计算速度随着样本量的增加而严重变缓。其中，INLA在single-processor 2.13-GHz laptop上运行，运用R-INLA；MCMC在quad-core Intel Pentium 2.84 GHz CPUs上运行，运用Fortran程序。

Grilli L等（2015）[9]针对二元logit混合模型，通过模拟研究对INLA和MCMC（Gibbs抽样器）的贝叶斯推断进行比较。对不同的先验设定，INLA方法总是收敛的更快，计算迅速。当组数（层-2单位数）较大时（大于70）所有的先验设定均能得到较好的估计结果，先验的选择会影响INLA和MCMC的估计。模拟不同的组数和每组观测数时组方差估计的偏倚，最坏的情况（10组，每组10个观测）下，INLA的偏倚仅增加10%，但MCMC增加20.8%。回归系数点估计的偏倚，两种方法的结果类似。综合来看，INLA比MCMC更精确，计算时间上更优越，例如：组数100，每组观测数50时，INLA耗时11秒，而MCMC却需要534秒。

Schrödle等（2011）[10]基于牛病毒性腹泻数据的时空模型，对所有参数的后验分布MCMC方法得到的直方图与

INLA得到的近似曲线是一致的。MCMC抽样所得的DIC值与由INLA计算的DIC值非常接近。INLA计算时间为233.23秒，而MCMC（3030000次迭代，30000次预迭代）速度为每秒进行246次迭代，即：花费时间远远高于INLA。

Taylor（2012）[11]针对空间连续高斯过程，通过模拟研究比较INLA与MCMC。MCMC运用M-H算法，进行100000次迭代（10000次预迭代）。通过不同的参数设置模拟了18种情况，综合来看，INLA计算时间降低了30%~50%，两种方法的MSE无显著差异。

3 应用实例

下面通过一个简单的例子来说明如何运用INLA方法进行参数估计。以微博评论数影响因素研究数据为例，因变量为 y（评论数），微博的评论数是计数数据，数据非负、分布尖峰、严重偏斜、方差明显大于均值、过度离散，因此适用广义线性模型中的负二项回归。自变量为每条微博具有的特点，包括 x_1（时新性）、x_2（时间性）、x_3（重要性）、x_4（趣味性）、x_5（延续性）、x_6（显著性）、x_7（个体性）、x_8（正面性）。x_1 和 x_2 是多分类数据，转换为虚拟变量来处理；x_3-x_8 为二分类数据。

运用R软件INLA程序包中的inla函数进行估计。

```
library (INLA)
wbdata =read.table("wb.txt",header=T)
print(head(wbdata))
formula<-y~as.factor(x1) + as.factor(x2) + as.factor(x3) + as.factor(x4) + as.factor(x5) + as.factor(x6) + as.factor(x7) + as.factor(x8)
nm<-inla(formula, data=wbdata, family="nbinomial")
summary(nm)
```

模型的参数估计见表1所示。http://www.r-inla.org给出了R-INLA的相关说明。

4 结论

INLA方法是一种灵活易用的计算工具。在数据量越来越大的今天，计算开销是重点考虑的问题之一，算法的计算速度显得更加重要，INLA方法在保证计算精度的同时，具有突出的快速计算能力。由于贝叶斯方法在机器学习领域的重要应用，INLA方法除了对这一大类隐高斯模型进行贝叶斯推断，在机器学习领域也必然会展现其优势。

表1 负二项回归基于INLA方法的贝叶斯推断

	mean	sd	0.025quant	0.5quant	0.975quant	mode
Intercept	2.9821	0.1131	2.7633	2.981	3.2071	2.9787
as.factor(x_1)1	−0.487	0.2564	−0.9673	−0.4956	0.0416	−0.5132
as.factor(x_1)2	−0.1811	0.1085	−0.3953	−0.1807	0.0308	−0.1799
as.factor(x_1)3	0.0585	0.1388	−0.2118	0.0577	0.3329	0.0562
as.factor(x_1)4	0.3789	0.1285	0.1275	0.3786	0.6318	0.378
as.factor(x_2)1	0.464	0.0877	0.2918	0.464	0.636	0.464
as.factor(x_2)2	2.1568	1.2541	0.0905	2.0105	4.9795	1.679
as.factor(x_3)1	0.1829	0.1008	−0.0163	0.1833	0.3795	0.1842
as.factor(x_4)1	0.3565	0.1542	0.0614	0.3538	0.6671	0.3484
as.factor(x_5)1	−0.1704	0.0898	−0.3455	−0.1709	0.0071	−0.1718
as.factor(x_6)1	−0.9897	0.18	−1.3336	−0.993	−0.6265	−0.9997
as.factor(x_7)1	0.3063	0.1396	0.0355	0.3052	0.5834	0.3029
as.factor(x_8)1	−0.1905	0.0834	−0.3543	−0.1905	−0.0267	−0.1906

参考文献：

[1]Rue H, Martino S, Chopin N. Approximate Bayesian Inference for latent Gaussian Models Using Inte-Grated Nested Laplace Approximations[J]. Journal of the Royal Statistical Society, Series B, 2009,71(2).

[2]Martino S, et al. Estimating Stochastic Volatility Models Using Integrated Nested Laplace Approximations[J]. European Journal of Finance, 2011, 17(7).

[3]Fong Y, Rue H, Wakefield J. Bayesian Inference for Generalized Linear Mixed Models[J]. Biostatistics, 2010, 11(3).

[4]Yu R Y, Rue H. Bayesian Inference for Additive Mixed Quantile Regression Models[J]. Computational Statistics & Data Analysis, 2011, 55(1).

[5]Natário I, Oliveira M M, Susete M.Using INLA to Estimate a Highly Dimensional Spatial Model for Forest Fires in Portugal[M]. Berlin: Springer, 2014.

[6]Gómez-Rubio V, Bivand R, Rue H.A New Latent Class to Spatial Econometrics Models With Integrated Nested Laplace Approximations [J]. Procedia Environmental Sciences, 2015,(27).

[7]茆诗松，王静龙，濮晓龙.高等数理统计[M].北京：高等教育出版社，2000.

[8]Paul M, Riebler A, Bachmann L M, et al. Bayesian Bivariate Meta-Analysis of Diagnostic Test Studies Using Integrated Nested Laplace Approximations[J]. Statistics in Medicine, 2010, 29(12).

[9]Grilli L, Metelli S, Rampichini C. Bayesian Estimation With Integrated Nested Laplace Approximation for Binary Logit Mixed Models[J]. Journal of Statistical Computation and Simulation, 2015,85(13).

[10]Schrödle B, Held L, Riebler A, et al. Using Integrated Nested Laplace Approximations for the Evaluation Of Veterinary Surveillance Data From Switzerland: A Case-Study[J]. Applied Statistics, 2011, 60(2).

[11]Taylor B M, Diggle P J. INLA or MCMC? A Tutorial and Comparative Evaluation for Spatial Prediction in Log-Gaussian Cox Processes[J]. Journal of Statistical Computation & Simulation, 2012, 84(10).

（责任编辑/浩　天）