**文章编号**:1006-3080(2017)04-0559-04 **DOI**:10.14135/j.cnki.1006-3080.2017.04.016

# 一种二元响应变量模型的分布式贝叶斯估计方法

吴 磊, 钱夕元
(华东理工大学理学院,上海 200237)

**摘要**:在海量数据背景下,传统的基于单个计算节点的算法很难满足分析要求。考察了一种分布式贝叶斯估计方法,通过在每台机器上单独运行蒙特卡洛抽样并做加权平均可以有效地解决算法效率问题。将该方法应用于基于广义极值模型的二元响应变量回归分析,并探讨其实用性。模拟研究表明分布式算法比传统方法更有效。

**关键词**:海量数据; 分布式贝叶斯方法; 极值模型
**中图分类号**:TP301.6 **文献标志码**:A

# A Distributed Bayesian Regression Method for Binary Response Massive Data

*WU Lei*, *QIAN Xi-yuan*
(*School of Science*, *East China University of Science and Technology*, *Shanghai* 200237, *China*)

**Abstract**: In the background of massive data, it is difficult to meet the analysis requirements for traditional one-node based algorithm. This paper considers a distributed Bayesian estimation method to solve the GEV based general linear regression model by running a separate Monte Carlo algorithm on each machine. The method is applied to regression analysis of binary response variables based on generalized extreme value model. The results show that the proposed distributed Bayesian regression algorithm is much faster than the traditional algorithm in the simulated data sets studying.

**Key words**: massive data; distributed Bayesian regression; GEV model

,

,

, , 。

0-1 。 ,

, 0-1 。 , Wang [2]

, (GEV)

, , GEV

, , , , ,

[1]。 ,

**收稿日期**:2016-10-31
**基金项目**: (“863”) (2015AA20107); “ ”(140304)
**作者简介**: (1992-), , , , 。
**通信联系人**: ,E-mail:xyqian@ecust.edu.cn

,

, 。

( ) 。

(MCMC) ,

。

,

,

。 ,

,

,

,

,

。 MCMC

,

,

[3] 。

,

, ,

。

## 1

(Be)

$\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_n)^{\mathrm{T}}$ $n$

0-1 $,\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)^{\mathrm{T}}$

,

$\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{ik})$, $\boldsymbol{\beta}_c=(\beta_1,\beta_2,\cdots,\beta_k)^{\mathrm{T}}$ $k$ 。 ,

$Y_i$ , $Y_i=1$

$p_i$, $Y_i=0$ $1-p_i$。 ,

$$p_i=Pr(Y_i=1\mid\boldsymbol{\beta}_c,\boldsymbol{x}_i)=\Psi(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_c)$$

$\Psi(\cdot)$ 0 1 ,

, $\Psi(\cdot)$ , $\Psi^{-1}(\cdot)$

。

logit

$\Psi^{-1}(p_i)=\lg\{p_i/(1-p_i)\}$, probit $\Psi^{-1}(p_i)=\Phi^{-1}(p_i)$,$\Phi^{-1}$ ,

cloglog $\Psi^{-1}(p_i)=-\lg\{-\lg(p_i)\}$ 。

,

。

Wang [2]

,

, :

$$G(x)=\exp\left[-\left\{1+\xi\frac{(x-\mu)}{\sigma}\right\}_+^{-1/\xi}\right]$$

$\xi$ ,

。 $\xi$ , ,

,

[4] 。 MCMC

, Metropolis-Hastings

(M-H) :

Step 0 $\boldsymbol{\theta}_0$;

Step 1

$q(\boldsymbol{\theta}^{(s-1)},\boldsymbol{\theta}^*)$ $\boldsymbol{\theta}^*$;

Step 2

$\alpha(\boldsymbol{\theta}^{(s-1)},\boldsymbol{\theta}^*)$;

Step 3 $\alpha(\boldsymbol{\theta}^{(s-1)},\boldsymbol{\theta}^*)$ $\boldsymbol{\theta}^{(s)}=\boldsymbol{\theta}^*$,

$1-\alpha(\boldsymbol{\theta}^{(s-1)},\boldsymbol{\theta}^*)$ $\boldsymbol{\theta}^{(s)}=\boldsymbol{\theta}^{(s-1)}$;

Step 4 Step 1,Step 2 Step 3 $S$ ;

Step 5 $\boldsymbol{\theta}_0,\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_S$ $\boldsymbol{\theta}$

, $\boldsymbol{\theta}$ 。

$$\alpha(\boldsymbol{\theta}^{(s-1)},\boldsymbol{\theta}^*)=\min\left[\frac{p(\boldsymbol{\theta}=\boldsymbol{\theta}^*\mid\boldsymbol{y})q(\boldsymbol{\theta}^*;\boldsymbol{\theta}=\boldsymbol{\theta}^{(s-1)})}{p(\boldsymbol{\theta}=\boldsymbol{\theta}^{(s-1)}\mid\boldsymbol{y})q(\boldsymbol{\theta}^{(s-1)};\boldsymbol{\theta}=\boldsymbol{\theta}^*)},1\right]$$

, $p(\boldsymbol{\theta}=\boldsymbol{\theta}^*\mid y)$ $\boldsymbol{\theta}=\boldsymbol{\theta}^*$

, $q(\boldsymbol{\theta}^*;\boldsymbol{\theta}=\boldsymbol{\theta}^{(s-1)})$ $\boldsymbol{\theta}$

$\boldsymbol{\theta}=\boldsymbol{\theta}^{(s-1)}$ 。

## 2 (DBe)

### 2.1 模型参数估计

,

。

。

,

,

,

,

,

,

:

$\boldsymbol{y}$ ( $n$), $\boldsymbol{y}_s$ $s$

( $n_s$), $\boldsymbol{\theta}$ 。

, :

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \prod_{s=1}^{S} p(\boldsymbol{y}_s \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})^{1/S} \quad (1)$$

(1) ,

$S$ ,

。 ,

,

。 , :

$$p(\boldsymbol{y}_s \mid \boldsymbol{\theta}) = \prod_{i=1}^{n_s} y_i F_{\mathrm{GEV}}(\beta_0 + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_c; \xi) + (1 - y_i) \cdot [1 - F_{\mathrm{GEV}}(\beta_0 + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}_c; \xi)]$$

,

,

Metropolis-Hastings ,

,

[5]。

$s$ $G$

Metropolis-Hastings ,

$S$

:

$\boldsymbol{\theta}_g = \left(\sum_{s=1}^{S} \boldsymbol{W}_s\right)^{-1} \sum_{s=1}^{S} \boldsymbol{W}_s \boldsymbol{\theta}_{sg}$, $g = 1, 2, \cdots, G$

, $\boldsymbol{W}_s$ ,

$\boldsymbol{\theta}$ 。DBe

:

Step 1 $\boldsymbol{y}$ $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_s$;

Step 2 $p(\boldsymbol{\theta})^{1/S}$, $S$

M-H ( ) $\boldsymbol{\theta}_{sg} \sim p(\boldsymbol{\theta} \mid \boldsymbol{y}_s)$, $g = 1, \cdots, G$;

Step 3 $\boldsymbol{\theta}_{sg}$, $s = 1, \cdots, S; g = 1, \cdots, G$ ,

$\boldsymbol{\theta}_g$, $g = 1, \cdots, G$;

Step 4 $\alpha$,

$(1-\alpha)\,G$ ;

## 2.2 其他非参数估计策略

,

。 ( 1),

(Kernel density estimate)

$p(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y}_s)$, $p(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y}) \propto \prod_{s=1}^{S} p(\hat{\boldsymbol{y}}_s \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})^{1/S}$。

, $\boldsymbol{\theta}$ ,

[3]。

# 3

$n = 100\,000$, 3 ,

0、 1 ,

$\beta_0 = 1, \beta_1 = 2, \beta_2 = 3, \beta_3 = 4$,

$\xi = 2$, 50 。

, $S = 12, G = 1\,250, \alpha = 0.8$。

: Hadoop2. x: 1

Master 3 Slave , CPU

Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00 GHz * 8 * (1[M]+3[S]) = 32core, 32 (1[M]+3[S]) = 128 GB, JVM -Xmx768 M,

Container 1 GB[6]。

1

。 ,

, ,

3 。 ,

Hadoop Map ( 12 ) ,

。

, 。

, , M-H

0, 0.005 0.1

,50

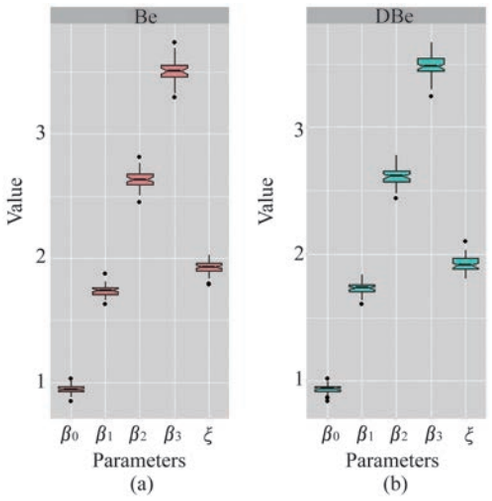
,

1 。

表 1

**Table 1** Comparative analysis of distributed Bayes vs. classical method

| MH Sampling | CPU Time/s | $\xi$ | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ |
|---|---|---|---|---|---|---|
| DBe | 46 | 1.924 6 (0.066 9) | 0.938 8 (0.036 0) | 1.731 8 (0.044 9) | 2.616 3 (0.069 5) | 3.488 3 (0.091 5) |
| Be | 124 | 1.925 7 (0.050 8) | 0.947 7 (0.036 4) | 1.741 8 (0.044 1) | 2.632 2 (0.068 2) | 3.510 1 (0.088 3) |




1

Fig. 1 Comparative analysis of distributed Bayes vs. classical method

4

,

,

(Metropolish-Hastings )

, ,

。

Gibbs

。

**参考文献：**


[1] CZADO C, SANTNER T J. The effect of link misspecification on binary regression inference [J]. Journal of Statistical Planning and Inference, 1992, 33(2): 213-231.

[2] WANG X, DEY D K. Generalized extreme value regression for binary response data: An application to B2B electronic payments system adoption [J]. The Annals of Applied Statistics, 2010, 4(4): 2000-2023.

[3] STEVEN S L, BLOCKER A W, BONASSI F V. Bayes and big data: The consensus Monte Carlo algorithm [J]. International Journal of Management Science and Engineering Management, 2016, 11(2): 78-88.

[4] GHOSH S K, MUKHOPADHYAY P, LU J C. Bayesian analysis of zero-inflated regression models [J]. Journal of Statistical Planning and Inference, 2006, 136(4): 1360-1375.

[5] CHIPMA H A, GEORGE E I, MCCULLOCH R E. BART: Bayesian additive regression trees [J]. The Annals of Applied Statistics, 2010, 4(1): 266-298.

[6] DEAN J, GHEMAWAT S. Mapreduce: Simplified data processing on large clusters [J]. Communications of the ACM, 2008, 51(1): 107-113.


**(上接第 532 页)**


[10] GANDOMI A H, YANG X S. Chaotic bat algorithm [J]. Journal of Computational Science, 2014, 5(2): 224-232.

[11] NIKNAM T, BAVAFA F, AZIZIPANAH-ABARGHOOEE R. New self-adaptive bat-inspired algorithm for unit comi--tment problem [J]. IET Science Measurement Technology, 2014, 8(6): 505-517.

[12] WANG G G, CHU H C E, MIRJALILI S. Three-dimensional path planning for UCAV using an improved bat algorithm [J]. Aerospace Science & Technology, 2016, 49: 231-238.

[13] AFRABANDPEY H, GHAFFARI M, MIRZAEI A, *et al*. A novel bat algorithm based on chaos for optimization tasks [C]//Iranian Conference on Intelligent Systems. USA: IEEE, 2014: 1-6.

[14] RAGHAVAN S, MARIMUTHU C, SARWESH P, *et al*. Bat algorithm for scheduling workflow applications in cloud [C]// International Conference on Electronic Design, Computer Networks & Automated Verification. USA: IEEE, 2015: 139-144.

[15] YıLMAZ S, KÜÇÜKSILLE E U. A new modification approach on bat algorithm for solving optimization problems [J]. Applied Soft Computing, 2014, 28(5): 259-275.

[16] YOUNG G F, SCARDOVI L, CAVAGNA A, *et al*. Starling flock networks manage uncertainty in consensus at low cost [J]. PloS Computational Biology, 2013, 9(1): e1002894.

[17] NETJINDA N, ACHALAKUL T, SIRINAOVAKUL B. Particle swarm optimization inspired by starling flock behavior [J]. Applied Soft Computing, 2015, 35(C): 411-422.

[18] XIE J, ZHOU Y, CHEN H. A novel bat algorithm based on differential operator and Lévy flights trajectory [J]. Computational Intelligence & Neuroscience, 2013(2013): 453-812.

[19] , . [J]. , 2015(2): 23-26.

[20] . PID MATLAB [M]. : , 2004.