DOI:10.19789/j.1004-9398.2017.03.004

广义极值分布的参数估计及实例分析

樊利利 王艳永
(吕梁学院数学系,山西 吕梁 033000)

摘 要

对广义极值(GEV)分布的参数进行了ML、GML和BAYBETA估计. 利用R统计软件,通过用Markov Chain Monte Carlo(MCMC)方法与Metroplis-Hastings算法产生服从参数后验分布的模拟样本,进而对GEV分布的参数进行估计,并对上述几种估计的偏差和均方误差进行了模拟. 对澳大利亚南部的Pirie港海平面年最大海平面高度建立GEV模型,并对T年一遇的最高水位进行了估计. 实例研究表明,用BAY BETA方法估计的最高水位稍高于用极大似然估计得到的结果.

关键词: 广义极值分布,参数估计,GML估计,BAY BETA估计.
中图分类号: O211.4

1 引 言

在人们的认知范围内,极值事件较少出现,然而一旦发生影响重大. 极值统计理论的产生和发展,为这类随机事件的统计分析提供了理论依据. 例如在工程设计中,需要对一些带有严重破坏性的自然灾害的强度进行充分估计. 建造桥梁时,为了防止洪水冲塌桥梁这类事故的发生,设计时需要考虑保用期内该河流可能爆发的最高水位. 在解决实际问题的过程中,既要出于安全的考虑,又不能盲目加大安全系数,这就需要用极值统计的理论,对历史数据进行统计推断,从而得到更加符合实际的极值数据的统计分布. 传统的对极值数据进行分析的方法是极值类型定理,其理论研究见Fisher和Tippett[1]. 最早,他们给出的关于最大次序统计量标准化的广义极值分布极限理论,奠定了经典极值理论的基础,但这些定理局限于独立同分布的最大值样本数据. 随后Gnedenko[2]证明了标准化后的极值存在极限分布:Gumbel分布、Frechet分布、Weibull分布,分别称为第一型、第二型和第三型极值分布. Jekinson[3]将这三种极值分布合为一个,提出了广义极值分布(Generalized Extreme Value Distribution).

利用历史数据进行统计推断,关键是对分布的参数进行估计. 进而,极值估计是对极端变异数据进行分析,建立统计模型. 这样的数据一般较少,如果能够知道待估参数的某些信息,再综合样本信息,无疑会得到比较精确的判断. 鉴于极值数据的缺乏,需要考虑历史经验,而Bayes统计推断的出发点正是把分布中的参数看成随机变量,由于利用了样本信息和先验信息,可以对参数的估计进行补充和修正. 将Bayes估计与极值理论相结合,研究极值数据分布的极值分位数,将同时兼顾两方面的信息,根据历史经验得到的先验信息和已知的样本信息,得到极值数据的后验分布,这样计算出的极值分位数将更加切合实际.

文章主要研究了GEV分布的参数极大似然估计问题,通过对GEV分布的参数进行估计并通过对模型外推,进而间接对其高分位数进行估计. 最后利用R软件通过用Markov Chain Monte Carlo(MCMC)方法与Metroplis-Hastings算法产生长度为3 000的Markov链. Markov链的样本数据即为参数后验分布的模拟样本,基于模拟样本进而对GEV分布的参数进行估计,对上述几种估计的偏差和均方根误差进行了模拟,结果表明对GEV分布的高分位数,BAY BETA方法的估计效果更好. 在此基础上对澳大利亚南部的Pirie港海平面年最大海平面高度建立GEV模型,并对T年一遇的最高水位进行了估计. 实例研究表明,用BAY BETA方法估计的最高水位

收稿日期:2016-10-13

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

稍高于用极大似然估计得到的结果.

2 广义极值分布

设随机变量 X 服从广义极值分布,则其累积分布函数为

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

其中 $1+\xi\left(\frac{x-\mu}{\sigma}\right)>0$, $-\infty<\mu<\infty$ 是位置参数, $\sigma>0$ 是尺度参数和 $-\infty<\xi<\infty$ 是形状参数.

相应的概率密度函数为

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

注意到形状参数 ξ 决定了分布的尾部形状,当 $\xi>0$ 时,分布的尾部较长;当 $\xi=0$ 时,分布的尾部呈指数状;当 $\xi<0$ 时,分布具有有限的上端点.

$\xi=0$ 是 GEV 分布的特殊情况,这时的分布称作 Gumbel 分布且此时 $x\in R$,同时也注意到 $\xi>0$ 和 $\xi<0$ 时 GEV 分布分别趋向于 Frechet 分布和 Weibull 分布.

3 MCMC 方法

Monte Carlo 方法亦称为随机模拟方法,是一种具有独特风格的数值计算方法,它是通过随机变量函数的概率模拟、统计试验或随机抽样求解数学物理、工程技术、生产管理等各种不同问题的近似数值解的概率统计方法.

MCMC 方法的基本思想是:通过一个平稳分布为 $\pi(x)$ 的 Markov 链来得到 $\pi(x)$ 的随机样本,如果模拟方法是精确的,那么就可用模拟样本估计后验分布.并且,还可以用模拟样本的均值估计后验均值,用模拟样本的直方图估计后验分布的密度等等.

4 GML 估计法

Martins and Stedinger[4] 指出,对于 GEV 分布形状参数 ξ 为负值时,对其进行高分位估计会造成较大的偏差.因而提出了一种新的估计方法,即 Generalized maximum likelihood estimation (GML). GML 是基于半贝叶斯分布的 Beta 分布,具体形式如下[5]:

$$\pi(\xi) = \frac{(0.5+\xi)^{u-1}(0.5-\xi)^{v-1}}{B(u,v)},$$

$$-0.5\leqslant\xi\leqslant0.5, u=6, v=9.$$

其中

$$B(u,v) = \frac{\Gamma(u)\Gamma(v)}{\Gamma(u+v)}.$$

ξ 的期望和方差分别为

$$E(\xi) = -0.1, \mathrm{Var}(\xi) = 0.015.$$

GML 估计的具体做法与极大似然估计做法类似,给定样本 $x=\{x_1,x_2,\cdots,x_n\}$,其联合密度函数为

$$p(x_1,x_2,\cdots x_n) = \prod_{i=1}^{n}\frac{1}{\sigma}\exp\cdot\left\{-\left(1+\xi\frac{x_i-\mu}{\sigma}\right)^{-\frac{1}{\xi}}\right\}\left(1+\xi\frac{x_i-\mu}{\sigma}\right)^{-\left(1+\frac{1}{\xi}\right)},$$

ML 估计的似然函数用 L 来表示

$$L(\mu,\sigma,\xi) = \frac{1}{\sigma^n}\exp\cdot\left\{\sum_{i=1}^{n}-\left(1+\xi\frac{x_i-\mu}{\sigma}\right)^{-\frac{1}{\xi}}\right\}\prod_{i=1}^{n}\left(1+\xi\frac{x_i-\mu}{\sigma}\right)^{-\left(1+\frac{1}{\xi}\right)},$$

GML 估计的似然函数用 GL 来表示

$$\mathrm{GL}(\mu,\sigma,\xi) = L(\mu,\sigma,\xi)\pi(\xi),$$

其对数似然函数为

$$\ln\mathrm{GL}(\mu,\sigma,\xi) = -n\ln\sigma-\sum_{i=1}^{n}\left\{\left(-\frac{1}{\xi}-1\right)\ln(y_i)-(y_i)^{-\frac{1}{\xi}}\right\}+\ln\{\pi(\xi)\}$$

其中 $y_i = 1+\xi\frac{x_i-\mu}{\sigma}$.

然后将其分别对 μ,σ,ξ 求偏导,分别令 $\frac{\partial\ln L}{\partial\mu}=0$, $\frac{\partial\ln L}{\partial\sigma}=0$, $\frac{\partial\ln L}{\partial\xi}=0$,得到含三个参数的三个方程,解此方程组便可得到 $\hat{\mu},\hat{\sigma},\hat{\xi}$.

5 BAY BETA 估计法

BAY BETA 是基于全贝叶斯框架的一种方法,将 GEV 分布的形状参数 ξ、位置参数 μ 与尺度参数 σ 均作为随机变量,分别考虑其先验分布,具体形式如下:

$$\pi(\xi) = \frac{(0.5+\xi)^{u-1}(0.5-\xi)^{v-1}}{B(u,v)},$$

$$\pi(\mu) = \frac{1}{\sqrt{2\pi}\times10^2}\exp\left(-\frac{\mu^2}{2\times10^4}\right),$$

$$\pi(\xi) = \frac{1}{\sqrt{2\pi}\times10^2}\exp\left(-\frac{\Phi^2}{2\times10^4}\right),$$

$$\pi(\xi,\mu,\sigma) = \pi(\xi)\pi(\mu)\pi(\Phi),$$

其中 $-0.5\leqslant\xi\leqslant0.5$, $u=6.499$, $v=8.7927$, $\Phi=\log\sigma$.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

这里选取有较大方差的正态分布作为先验分布,使分布密度曲线比较平坦,以反映出先验信息的缺乏. BAY BETA 估计的似然函数用 BGL 来表示

$$\mathrm{BGL}(\mu,\sigma,\xi) = L(\mu,\sigma,\xi)\,\pi(\xi)\,\pi(\mu)\,\pi(\sigma),$$

其对数似然函数为

$$\ln\mathrm{BGL}(\mu,\sigma,\xi) = -n\ln\sigma - \sum_{i=1}^{n}\left\{\left(-\frac{1}{\xi}-1\right)\ln(y_i) - (y_i)^{-\frac{1}{\xi}}\right\} + \ln\{\pi(\xi)\} + \ln\{\pi(\mu)\} + \ln\{\pi(\sigma)\},$$

其中 $y_i = 1 + \xi\frac{x_i-\mu}{\sigma}$. 将其分别对 μ,σ,ξ 求偏导,令 $\frac{\partial \ln L}{\partial \mu}=0$, $\frac{\partial \ln L}{\partial \sigma}=0$, $\frac{\partial \ln L}{\partial \xi}=0$ 得到含三个参数的三个方程,解此方程组便可得到 $\hat{\mu},\hat{\sigma},\hat{\xi}$.

6　几种估计的模拟结果对比

前面我们用 MLE,GMLE 和 BAY BETA 三种不同的方法对 GEV 的参数进行了估计. 通过模拟比较这三种估计的偏差(Bias)以及均方根误差(RMSE),从而对它们估计的效果进行对比.

首先,我们是在假设形状参数 $-0.35 \leqslant \xi \leqslant 0.35$ 的情况下,对参数估计进行模拟. 通常我们通过计算参数估计的偏差和均方误差来评价估计的有效性. 这里我们假设位置参数 $\mu = 0$,尺度参数 $\sigma = 1$,位置参数取不同值 $\xi = -0.1$ 与 $\xi = 0.1$,由 MCMC 方法生成样本 $(X_1,\cdots,X_n)$,对样本容量 $n = 30, n = 100$ 分别用 GLME 和 BAY BETA 方法计算 Bias 和 RMSE. 结果如表 1,表 2.

表 1　对 GEV 分布分位数进行估计的 Bias 及 RMSE　$\xi = -0.1$

p	Bias			RMSE		
	ML	GML	BAYBETA	ML	GML	BAYBETA
Size = 30						
0.1	0.067	0.070	0.033	0.206	0.208	0.199
0.2	0.056	0.051	0.033	0.210	0.201	0.194
0.5	0.018	0.008	0.033	0.253	0.235	0.236
0.8	-0.042	-0.050	0.038	0.387	0.380	0.392
0.9	-0.071	-0.084	0.046	0.596	0.527	0.545
0.95	-0.075	-0.111	0.060	0.940	0.710	0.732
0.99	0.090	-0.137	0.123	2.469	1.315	1.334
0.998	0.740	-0.082	0.256	5.743	2.276	2.259
0.999	1.283	-0.019	0.346	8.086	2.839	2.790
Size = 100						
0.1	0.018	0.019	0.007	0.094	0.095	0.094
0.2	0.016	0.015	0.008	0.092	0.091	0.090
0.5	0.006	0.002	0.010	0.124	0.120	0.121
0.8	-0.016	-0.019	0.015	0.202	0.200	0.203
0.9	-0.032	-0.034	0.021	0.309	0.289	0.291
0.95	-0.043	-0.045	0.033	0.482	0.418	0.418
0.99	-0.028	-0.050	0.096	1.202	0.927	0.919
0.998	0.104	0.007	0.241	2.525	1.823	1.810
0.999	0.222	0.064	0.343	3.352	2.366	2.349

从整体上看,当 $\xi < 0$ 时,BAY BETA 估计效果相对较好,它在三种估计中有较小的 RMSE,但 GMLE 估计有较小的 Bias. 进一步,BAY BETA 估计在 $p > 0.95$ 时 Bias 为正,说明 BAY BETA 估计法对于水文极值模型是有保证的. 比如降雨量和洪水,Bias 为正意味着在将来我们不会低估极值分位数,即不会低估灾难.

7　BAY BETA 方法的实际应用

本节将 BAY BETA 估计方法应用于实际数据,得到高分位数的估计,讨论了所得结果的实际意义.

例 1: Pirie 港海平面年最大值图 7.1 是 1923—1987 年位于澳大利亚南部的 Pirie [6]. 港海平面年最大值.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表 2 对 GEV 分布分位数进行估计的 Bias 及 RMSE $\xi = 0.1$

p	Bias			RMSE		
	ML	GML	BAYBETA	ML	GML	BAYBETA
Size = 30						
0.1	0.034	0.046	0.005	0.220	0.220	0.218
0.2	0.044	0.000	−0.025	0.204	0.199	0.202
0.5	0.034	−0.052	−0.039	0.224	0.212	0.210
0.8	−0.022	−0.009	0.067	0.271	0.273	0.290
0.9	−0.068	0.088	0.212	0.350	0.350	0.414
0.95	−0.111	0.230	0.403	0.478	0.479	0.597
0.99	−0.176	0.722	1.020	0.942	1.002	1.242
0.998	−0.166	1.431	1.870	1.610	1.823	2.177
0.999	−0.132	1.801	2.306	1.967	2.265	2.668
Size = 100						
0.1	0.024	0.030	0.013	0.127	0.128	0.125
0.2	0.024	0.010	−0.002	0.121	0.118	0.117
0.5	0.020	−0.012	−0.011	0.130	0.122	0.122
0.8	0.011	0.008	0.037	0.152	0.151	0.157
0.9	0.006	0.050	0.102	0.189	0.198	0.220
0.95	0.003	0.111	0.188	0.254	0.277	0.320
0.99	0.012	0.315	0.461	0.496	0.575	0.674
0.998	0.045	0.598	0.826	0.824	1.005	1.169
0.999	0.068	0.742	1.008	0.985	1.225	1.421

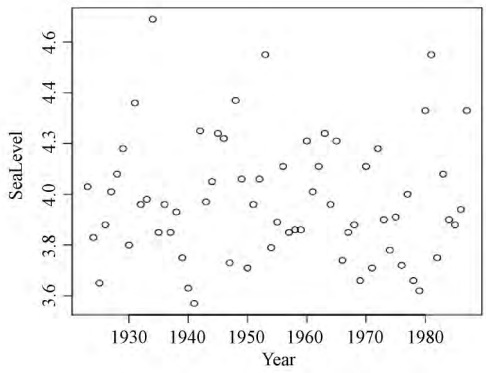


图 1 澳大利亚南部的 Pirie 港海平面年最大值

由于事先对该数据的分布无明确信息,将 GEV 分布的形状参数 ξ、位置参数 μ 与尺度参数 σ 均作为随机变量,分别考虑其先验分布,具体形式如下

$$\pi(\xi) = \frac{(0.5+\xi)^{u-1}(0.5-\xi)^{v-1}}{B(u,v)},$$

$$\pi(\mu) = \frac{1}{\sqrt{2\pi}\times 10^2}\exp\left(-\frac{\mu^2}{2\times 10^4}\right),$$

$$\pi(\Phi) = \frac{1}{\sqrt{2\pi}\times 10^2}\exp\left(-\frac{\Phi^2}{2\times 10^4}\right),$$

$$\pi(\xi,\mu,\sigma) = \pi(\xi)\pi(\mu)\pi(\Phi),$$

其中 $-0.5 \leqslant \xi \leqslant 0.5$, $u = 6.499$, $v = 8.7927$, $\Phi = \log\sigma$. 建议分布的方差为 $(0.000\,4, 0.01, 0.01)$,参数 $\theta' = (\mu, \log\sigma, \xi)$ 的初始值为 $(\mu, \log\sigma, \xi) = (5, 0, 0.1)$. 根据 MCMC 方法,产生长度为 5 000 的 Markov 链,见图 2.

对于尺度参数 σ,令 $\sigma_i = \exp^{\Phi}$ 可得到尺度参数后验分布的一个模拟样本,进一步得到后验分布的模拟样本 (μ_i, σ_i, ξ_i). 根据 Markov 链的样本数据,对 GEV 分布的参数进行估计,分别利用 ML, GML, BAY BETA 法估计结果如下,

$$\hat{\mu}_1 = 3.8747, \hat{\sigma}_1 = 0.1980, \hat{\xi}_1 = 0.0501,$$

$$\hat{\mu}_2 = 3.8684, \hat{\sigma}_2 = 0.1944, \hat{\xi}_2 = 0.0097,$$

$$\hat{\mu}_3 = 3.8032, \hat{\sigma}_3 = 0.2119, \hat{\xi}_3 = 0.0020.$$

实际上,我们更关心极值分布的高分位数估计. 在水文工程中,就是重现水平. 重现水平就是在给定概率水平下的极值分布的分位数. 即由 GEV 发布可得

$$\hat{x}_p = \begin{cases} \hat{\mu} - \dfrac{\hat{\sigma}}{\hat{\xi}}\{1 - (-\log p)^{-\hat{\xi}}\}, & \hat{\xi} \neq 0, \\ \hat{\mu} - \hat{\sigma}\log\{(-\log p)\}, & \hat{\xi} = 0. \end{cases}$$

将向量 (μ_i, σ_i, ξ_i) 的每个模拟值代入上式,得到相应的 $1/(1-p)$ 年重现期的重现水平的后验分布的样本. 百年重现期重现水平模拟值样本均值估计结果如下:

$$\hat{x}_{0.99} = 4.78, \hat{x}_{0.99} = 4.71, \hat{x}_{0.99} = 4.84.$$

从以上三种方法得到的结果来看,可得以下结论: 第一,BAY BETA 方法和极大似然估计方法得到的 $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ 的估计值比较接近. 然而,BAY BETA

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

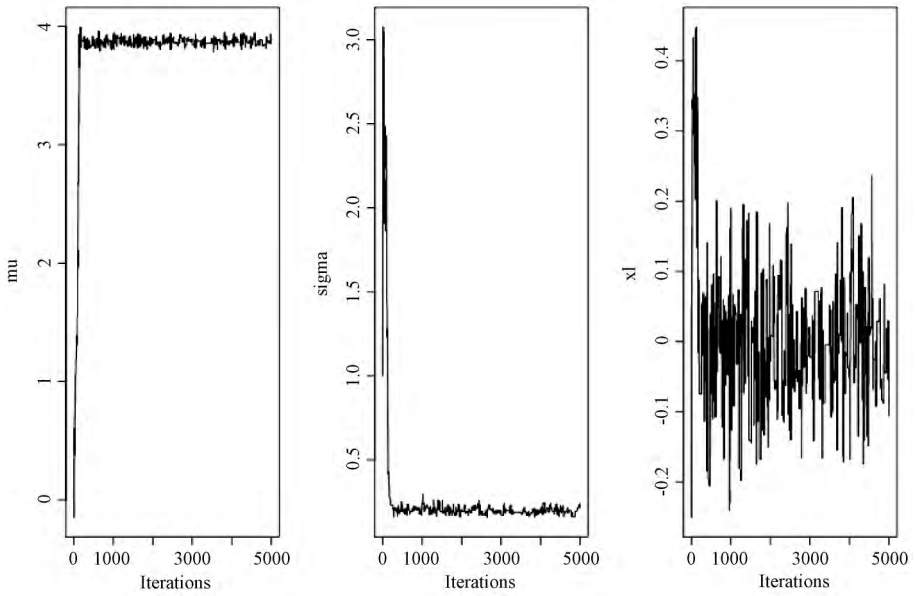


图 2　GEV 模型参数估计的 MCMC 模拟

方法估计的重新水平 x_p 略大于其他两种估计的结果,这可以理解为,正是由于 BAY BETA 方法中把参数 θ 看成了随机变量,增加了模型的不确定性的结果,这实际上更加符合实际情况; 第二,尽管 BAY BETA 方法和极大似然估计方法得到的结果比较接近,但是,应该说 BAY BETA 方法提供了一个更加全面的信息. 特别是对于极值来说,它总是对未来发生事件的概率分析. 因此,充分利用先验信息,并且增加模型的随机性是有必要的.

参 考 文 献

[1] Fisher R A, Tippett L H C. Limittimg forms of the frequency distribution of the largest or smallest member of a sample [J]. Procs Cambridge Philos Soc,1928,24: 180 – 190.

[2] Gnedenko B. Sur la distribution limite du terme d'une serie aleatoire [J]. Ann Math,1943,44: 423 – 453.

[3] Jenkinson A F. The frequency distribution of the annual maximum (or minimum) of meteorological elements [J]. The Quarterly Journal of the Royal Meteorological Society,1955,81: 158 – 197.

[4] Martins E S, Stedinger J R. Generalized maximum likelihood GEV quantile estimators for hydrologic data [J]. Water Resources Research, 2000,36(3): 737 – 744.

[5] Park J. A simulation-based hyperarameter selection for quantile estimation of the generalized extreme value distribution [J]. Math Comput Simul,2005, 70(4): 277 – 234.

[6] Hastings W. Monte Carlo samling methods using Markov C hains and their applications [J]. Biometria, 1970,57: 97 – 109.

[7] 汤银才. R 语言与统计分析 [M]. 北京: 高等教育出版社,2008.

[8] Eric P Smith. An introduction to statistical modeling of extreme values [J]. Technometrics, 2002,44 (4): 397 – 397.

[9] 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,2006.

[10] 史道济. 实用极值统计分析 [M]. 天津: 天津科学技术出版社,2006.

[11] Stuart Coles. 极值统计建模导论 [M]. 北京: 世界图书出版公司北京公司,2008.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

Parameter Estimation and Case Analysis of Based on Generalized Extreme Value Distribution

Fan Lili Wang Yanyong

(Department of Mathematics, Lüliang University, Lüliang Shanxi 033000)

Abstract

In this paper, the parameters of the generalized extreme value (GEV) distribution provide ML, GML and BAYBETA estimates. Using of R statistical software, analog samples of obedience posterior distribution of parameters are generated by Markov Chain Monte Carlo (MCMC) methods and Metroplis-Hastings algorithm, then GEV distribution parameters are estimated, and bias and root mean square error of the several estimation are simulated. Then we establish GEV model of the biggest sea level of Port Pirie in South Australia, and the maximum water level of T-year return is estimated. Case study shows that the highest level of using BAY BETA method is slightly higher estimation to that using maximum likelihood estimation results obtained.

Key words: generalized extreme value distribution, parameter estimation, GML estimation, BAY BETA estimation.

(上接第 5 页)

Singular Integral Equations of Convolution Type with Reflection and a Finite Number of Translations

Li Pingrun

(School of Mathematical Sciences, Qufu Normal University, Qufu Shandong 273165)

Abstract

By Fourier transform, we transform these equations into a system of equation or Riemann boundary value problems with both discontinuous coefficient and reflection. According to the methods of solution for classical boundary value problem of analytic function, the general solutions and conditions of solvability are obtained in class (0) respectively.

Key words: Riemann boundary value problems, singular integral equations, reflection, a finite number of translation shifts.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net