Statistics and Application 统计学与应用, 2016, 5(4), 350-358
Published Online December 2016 in Hans. http://www.hanspub.org/journal/sa
http://dx.doi.org/10.12677/sa.2016.54037

Hans汉斯

Bayes AR(p) Analysis of Container Throughput Based on the Gibbs Sampling

Shanwei Zhu

School of Economics & Management, Shanghai Maritime University, Shanghai
Email: 1377490996@qq.com

Received: Nov. 26th, 2016; accepted: Dec. 9th, 2016; published: Dec. 15th, 2016

Abstract

AR(p) model is widely used for time series forecasts; however, it's difficult for traditional static model to deal with emergencies, which lead to estimation bias. In view of the influences of emergencies for model estimation, we carry out Bayesian analysis of the model by aid of the Gibbs sampling. According to likelihood function's statistical structure of the time series samples, the prior distribution is obtained. After getting the posterior empirical distribution of parameters, the specific sampling strategy is proposed. Under the minimum mean square error estimation criterion, the simulation experiments show that the estimates are close to the true value. The analysis for the data of Shanghai port's container throughput from 1982 to 2015 indicates that by aid of the Bayesian analysis, the estimation bias from emergencies can be overcome so that the model prediction is more accurate.

Keywords

AR(p) Model, Bayes Analysis, Gibbs Sampling, Throughout Capacity

基于Gibbs抽样的集装箱吞吐量 Bayes AR(p)分析

朱善维

上海海事大学经济管理学院，上海
Email: 1377490996@qq.com

文章引用: 朱善维. 基于 Gibbs 抽样的集装箱吞吐量 Bayes AR(p)分析[J]. 统计学与应用, 2016, 5(4): 350-358.
http://dx.doi.org/10.12677/sa.2016.54037

收稿日期：2016年11月26日；录用日期：2016年12月9日；发布日期：2016年12月15日

摘　要

AR(p)模型广泛应用于时序预测，然而传统静态模型难以处理突发事件以致模型估计偏差。鉴于突发事件对模型估计的影响，采用Gibbs抽样方法对模型进行Bayes分析，根据时序样本似然函数的统计结构构造出模型各参数的先验分布。在导出模型参数后验条件分布后给出具体抽样策略。在最小均方误差估计准则下对中小样本的模拟显示，参数估计值与真值接近。对上海港1982~2015年集装箱吞吐量数据的分析表明：借助Bayes分析，可以克服由于突发事件导致的模型估计偏差，使模型预测更加准确。

关键词

AR模型，Bayes分析，Gibbs抽样，吞吐量

1. 引言

世界经济的发展与港口业休戚相关，世界各主要港口间的竞争如今正逐步转向以集装箱吞吐量为核心的港口综合能力的竞争。可见，对港口集装箱吞吐量的预测不可或缺，其在制定港口发展方向，经营策略，投资规模，泊位选址上都发挥着重要作用。

对于集装箱吞吐量的预测，田歆等[1]基于 TEI@I 方法，提出香港集装箱吞吐量预测研究框架，利用季节 ARIMA 及 VAR 等计量模型，不规则时间的量化方法以及径向基神经网络技术，建立了集装箱吞吐量的综合集成预测模型。杨金花等[2]基于灰色预测模型 GM(1,1)预测了上海港样本期外 3 年的集装箱吞吐量。余思勤等[3]采用带外生变量的非线性自回归神经网络模型对上海港集装箱吞吐量进行预测，发现训练后的网络误差较小，预测效果较好。已有预测模型多是传统静态模型，面对诸如 2009 年美国金融危机，全球贸易不景气，港口集装箱吞吐量锐减之类的突发事件就会导致模型预测偏差。鉴于此，本文提出 Bayes AR(p)模型，运用 Bayes 方法对时间序列进行预测分析，此不仅利用了模型信息和样本信息，也结合了模型中总体分布的未知参数信息。这样就能处理传统静态模型难以克服的缺陷，易于适应外部变化。

运用 Bayes 方法对时间序列模型进行分析时会遇到对高维概率分布积分的问题。Gibbs 抽样是解决高维积分的迭代 Monte Carlo 方法，解决了复杂表达式难以高维积分的问题，应用十分广泛。本文利用 Bayes 方法对 AR(p)模型进行分析，得到模型参数后验条件分布后，给出 Gibbs 抽样的具体策略。通过 Gibbs 抽样得到一系列模拟值构成 Markov 链，且链的平稳分布收敛于待估参数的后验条件分布，即可以将模拟值看作后验分布的独立样本对参数估计值进行推断。

2. 模型建立

对于随机变量 Y_t，其满足的 AR(p)模型为

$$Y_t = \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \cdots + \theta_p Y_{t-p} + \varepsilon_t, t = 1,2,\cdots,T, \tag{1}$$

其中，误差项 $\varepsilon_t \sim N\left(0,\sigma^2\right)$，且 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_T$ 相互独立；$\theta_1,\theta_2,\cdots,\theta_p$ 为模型的自回归系数；p 为模型的阶数；记 $\Phi\left(B\right)=\boldsymbol{\theta}^{\tau}\boldsymbol{B}$，其中，$\boldsymbol{\theta}=\left(\theta_1,\theta_2,\cdots,\theta_p\right)^{\tau}$，$\boldsymbol{B}=\left(B,B^2,\cdots,B^p\right)^{\tau}$，$B$ 为延迟算子。则(1)式可记为 $Y_t=\Phi\left(B\right)Y_t+\varepsilon_t$。若 AR(p)模型是平稳的，则必多项式 $\Phi\left(B\right)=0$ 的根在单位圆之内。本文假定所分析的

AR(p)模型均是平稳过程。对于未处理的时间序列数据，通过单位根检验来判定是否平稳，对非平稳的时间序列可通过差分化为平稳时间序列。

若已知$\{Y_{t-1},Y_{t-2},\cdots,Y_{t-p}\}$对于给定时间序列$\{Y_t,t=1,2,\cdots,T\}$，则

$$Y_t\left|Y_{t-1},Y_{t-2},\cdots,Y_{t-p}\right.\sim N\left(\Phi(B)Y_t,\sigma^2\right),$$

记Y_t的条件概率密度为$f_{Y_t|Y_{t-1},Y_{t-2},\cdots,Y_{t-p}}\left(y_t\left|y_{t-1},y_{t-2},\cdots,y_{t-p}\right.\right)$。对初始值$y_{-1},y_{-2},\cdots,y_{-p}$，样本$\Theta=\{y_1,y_2,\cdots,y_T\}$的条件似然函数为

$$l=\prod_{t=1}^{T}f_{Y_t|Y_{t-1},Y_{t-2},\cdots,Y_{t-p}}\left(y_t\left|y_{t-1},y_{t-2},\cdots,y_{t-p}\right.\right)=(2\pi)^{-\frac{T}{2}}\left(\sigma^2\right)^{-\frac{T}{2}}\exp\left\{-\frac{\sum_{t=1}^{T}\left(y_t-\boldsymbol{\theta}^{\tau}\boldsymbol{B}y_t\right)^2}{2\sigma^2}\right\},\tag{2}$$

其对数似然函数为

$$\log(l)=-\frac{T}{2}\log(2\pi)-\frac{T}{2}\log\left(\sigma^2\right)-\frac{\sum_{t=1}^{T}\left(y_t-\boldsymbol{\theta}^{\tau}\boldsymbol{B}y_t\right)^2}{2\sigma^2},$$

在进行 Bayes 分析时，由于参数的先验信息比较难确定，而不当先验信息会对估计结果产生错误影响。根据似然函数(2)式的结构，自回归系数$\boldsymbol{\theta}$的先验分布选为无信息先验，记$\pi(\boldsymbol{\theta})\propto\text{const}$，误差项$\varepsilon_t$的方差$\sigma^2$的先验分布选为倒 Gamma 分布(参见文[4])。其为σ^2的共轭先验分布。记

$$\pi\left(\sigma^2\right)\propto\sigma^{-\alpha-1}\mathrm{e}^{-\beta/\sigma^2}\sim IG\left(\frac{\alpha-1}{2},\beta\right)$$

其中α,β是先验分布参数。根据 Bayes 公式可知，参数$\boldsymbol{\theta}$和σ^2的后验条件分布为

$$\pi\left(\boldsymbol{\theta},\sigma^2\left|\Theta\right.\right)\propto\left(\sigma^2\right)^{-\frac{T+\alpha+1}{2}}\exp\left\{-\frac{\sum_{t=1}^{T}\left(y_t-\boldsymbol{\theta}^{\tau}\boldsymbol{B}y_t\right)^2+2\beta}{2\sigma^2}\right\}\tag{3}$$

为了避免对参数的高维概率密度函数作积分的复杂问题，我们利用马尔科夫链蒙特卡罗(Markov Chain Monte Carlo, MCMC)算法。

MCMC 的基本思想是：对于给定样本集Φ，为得到参数向量$\boldsymbol{\beta}$的后验分布$\pi\left(\boldsymbol{\beta}\left|\Phi\right.\right)$的样本，建立一个平稳分布为$\pi\left(\boldsymbol{\beta}\left|\Theta\right.\right)$的 Markov 链$\left\{\boldsymbol{\beta}^{(0)},\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)},\cdots\right\}$，对于任意时刻$t$，下一状态$\boldsymbol{\beta}^{(t+1)}$来自分布$\pi\left(\boldsymbol{\beta}^{(t+1)}\left|\boldsymbol{\beta}^{(t)},\Theta\right.\right)$的抽样，即其只和当前时刻$t$的状态有关，与历史状态$\left\{\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)},\cdots,\boldsymbol{\beta}^{(t-1)}\right\}$无关。在这些迭代样本的基础上就可以对参数做统计推断，如下式计算后验均值

$$E\left[\boldsymbol{\beta}\left|\Theta\right.\right]\approx\frac{1}{M-m}\sum_{j=m+1}^{M}\boldsymbol{\beta}^{(j)}$$

其中，M为链的迭代状态数，m为链的截断状态数。在 MCMC 算法中，为构造状态转移概率使已知后验分布$\pi\left(\boldsymbol{\beta}\left|\Theta\right.\right)$为平稳分布，常采用的构造方法有 Gibbs 抽样法，Metropolis-Hastings 算法等。本文便是采用 Gibbs 抽样的 MCMC 算法。

考虑到联合后验分布(3)式的特点，如文[5]，在给定σ^2的条件下，可选取建议分布是p元正态分布的 Metropolis-Hastings 算法对自回归系数$\boldsymbol{\theta}$进行后验抽样。为确定建议分布的参数，参考[6]中定理给出如下引理。

引理：对于样本 Φ，模型的 p 维参数向量 $\boldsymbol{\theta}$ 的极大似然估计量为 $\hat{\boldsymbol{\theta}}$，$\hat{\boldsymbol{\theta}}$ 满足：$\sqrt{n}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)\xrightarrow[n\to\infty]{L} N_{\mathrm{p}}\left(0, I^{-1}\left(\boldsymbol{\theta}\middle|\Phi\right)\right)$，其中 $\boldsymbol{\theta}=\left(\theta_1,\theta_2,\cdots,\theta_p\right)$，$\hat{\boldsymbol{\theta}}=\left(\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_p\right)$，$I^{-1}\left(\boldsymbol{\theta}\middle|\Phi\right)$ 为 p 阶 Fisher 信息阵，有

$$I_{ij}\left(\boldsymbol{\theta}\middle|\Phi\right)=-E\left[\frac{\partial^2\log\left(l\left(\boldsymbol{\theta}\middle|\Phi\right)\right)}{\partial\theta_i\partial\theta_j}\right],\ i,j=1,2,\cdots,p,$$

其中，$l\left(\boldsymbol{\theta}\middle|\Phi\right)$ 为样本 Φ 下 $\boldsymbol{\theta}$ 的似然函数。则根据引理，Metropolis-Hastings 算法的建议分布的协差阵取为

$$\Sigma^{(t)}=\begin{bmatrix}-\dfrac{\partial^2\log\left(l\left(\theta^{(t)}\middle|\Theta\right)\right)}{\partial^2\theta_1} & -\dfrac{\partial^2\log\left(l\left(\theta^{(t)}\middle|\Theta\right)\right)}{\partial\theta_1\partial\theta_2} & \cdots & -\dfrac{\partial^2\log\left(l\left(\theta^{(t)}\middle|\Theta\right)\right)}{\partial\theta_1\partial\theta_p}\\ -\dfrac{\partial^2\log\left(l\left(\theta^{(t)}\middle|\Phi\right)\right)}{\partial\theta_2\partial\theta_1} & -\dfrac{\partial^2\log\left(l\left(\theta^{(t)}\middle|\Phi\right)\right)}{\partial^2\theta_2} & \cdots & -\dfrac{\partial^2\log\left(l\left(\theta^{(t)}\middle|\Phi\right)\right)}{\partial\theta_2\partial\theta_p}\\ \vdots & \vdots & \vdots & \vdots\\ -\dfrac{\partial^2\log\left(l\left(\theta^{(t)}\middle|\Phi\right)\right)}{\partial\theta_p\partial\theta_1} & -\dfrac{\partial^2\log\left(l\left(\theta^{(t)}\middle|\Phi\right)\right)}{\partial\theta_p\partial\theta_2} & \cdots & -\dfrac{\partial^2\log\left(l\left(\theta^{(t)}\middle|\Phi\right)\right)}{\partial^2\theta_p}\end{bmatrix}^{-1}$$

于是，回归系数 $\boldsymbol{\theta}=\left[\theta_1,\theta_2,\cdots,\theta_p\right]^{\tau}$ 的抽样方法的具体步骤见算法 1。

算法 1：步 1：抽取 $\boldsymbol{\theta}'\sim N_p\left(\boldsymbol{\theta}^{(t)},\Sigma^{(t)}\right)$；

步 2：抽取 $c\sim U\left(0,1\right)$；

步 3：若 $c\le\alpha\left(\boldsymbol{\theta}',\boldsymbol{\theta}^{(t)}\right)$，则 $\boldsymbol{\theta}^{(t+1)}=\boldsymbol{\theta}'$，否则 $\boldsymbol{\theta}^{(t+1)}=\boldsymbol{\theta}^{(t)}$，其中

$$\alpha\left(\boldsymbol{\theta}',\boldsymbol{\theta}^{(t)}\right)=\min\left\{1,\frac{\pi\left(\boldsymbol{\theta}'\middle|\Theta\right)}{\pi\left(\boldsymbol{\theta}^{(t)}\middle|\Theta\right)}\right\}.$$

对在参数 $\boldsymbol{\theta}$ 给定下，参数 σ^2 的后验分布核为

$$\pi\left(\sigma^2\middle|\boldsymbol{\theta},\Theta\right)\propto\left(\frac{\sum_{t=1}^{T}\left(y_t-\boldsymbol{\theta}^{\tau}\boldsymbol{B}y_t\right)^2+2\beta}{\sigma^2}\right)^{\frac{T+\alpha+3}{2}-1}\exp\left\{-\frac{\sum_{t=1}^{T}\left(y_t-\boldsymbol{\theta}^{\tau}\boldsymbol{B}y_t\right)^2+2\beta}{2\sigma^2}\right\}$$

也就是

$$\frac{\sum_{t=1}^{T}\left(y_t-\boldsymbol{\theta}^{\tau}\boldsymbol{B}y_t\right)^2+2\beta}{\sigma^2}\sim\chi^2_{T+\alpha+3}$$

其中，χ^2_f 表示自由度为 f 的 χ^2 分布。则参数 σ^2 的后验分布核是标准分布，可直接抽样。

于是 Gibbs 抽样方法的具体迭代步骤如下：给定迭代初始值 $\left\{\boldsymbol{\theta}^{(1)},\sigma^{2(1)}\right\}$，依次取 $j=2,3,\cdots,M$，那么第 j 轮迭代的实现步骤如下：

步 1：从分布 $\pi\left(\boldsymbol{\theta}^{(j)}\middle|\sigma^{2(j-1)},\Theta\right)$ 里产生 $\boldsymbol{\theta}^{(j)}$，

步 2：从分布 $\pi\left(\sigma^{2^{(j)}}\middle|\boldsymbol{\theta}^{(j)},\Theta\right)$ 里产生 $\sigma^{2^{(j)}}$。

3. 模拟实验

为了考察 Gibbs 抽样方法的效果，本节进行模拟实验，为不失一般性，取 $p=2$，则模拟的 AR(p)模型如下

$$Y_t = \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \varepsilon_t, t = 1,2,\cdots,T,$$

其中，$\varepsilon_t \sim N\left(0,\sigma^2\right)$，记 $\boldsymbol{\theta} = \left(\theta_1,\theta_2\right)$。令参数的真值 $\boldsymbol{\theta} = \left(0.7,-0.1\right)$，$\sigma^2 = 0.01$，超参数 $\left(\alpha,\beta\right) = \left(200,2\right)$，取 Y_t 的初始值 $Y_0 = 0, Y_{-1} = 0$。分别取时间跨度为 50, 100, 200 三种情况进行模拟比较，模拟迭代 10,000 次，并舍弃轨道的前 5000 次迭代，各参数的后验估计值和后验标准差如表 1，其中括号内为后验标准差。

由表 1 可见，当样本量较小时就已经接近真实值，且随着样本时间跨度的增加，参数的估计标准差逐渐缩小。

4. 实证分析

自 2010 年来，上海港集装箱吞吐量一直居世界第一，对其集装箱吞吐量的预测对我国经济的发展至关重要。本文选取 1982~2015 年的上海港集装箱吞吐量数据(如表 2)，数据来源于上海市统计年鉴[7]。

绘出历年吞吐量的散点图如图 1。

由图可见，数据呈现指数上升的趋势，2009 年由于世界金融危机上海港集装箱吞吐量锐减，可视为突发事件。因此可对数据做对数处理，处理结果如图 2。

Table 1. The estimation of parameters posterior mean
表 1. 参数后验均值估计结果

N	真值	50	100	200
θ_1	0.7	0.6939(0.1719)	0.7011(0.1142)	0.7048(0.0923)
θ_2	−0.1	−0.098(0.1556)	−0.1056(0.1191)	−0.1089(0.1085)
σ^2	0.01	0.0160(0.0015)	0.010(0.0010)	0.0134(0.0011)

Table 2. Shanghai port's container throughput from 1982 to 2015
表 2. 1982~2015 年上海港集装箱吞吐量

年份	吞吐量 (万 TEU)	年份	吞吐量 (万 TEU)	年份	吞吐量 (万 TEU)
1982	6.6	1994	119.9	2006	2171.9
1983	8.0	1995	152.6	2007	2615.2
1984	11.5	1996	197.1	2008	2800.6
1985	20.2	1997	93.5	2009	2500.2
1986	20.4	1997	119.9	2010	2906.9
1987	22.4	1998	152.6	2011	3173.9
1988	31.3	2000	197.1	2012	3252.9
1989	35.4	2001	252.8	2013	3361.7
1990	45.6	2002	306.6	2014	3528.5
1991	57.7	2003	421.6	2015	3653.7
1992	73.1	2004	1455.4		
1993	93.5	2005	1808.4		

经对数处理后，吞吐量随时间的变化有明显的线性趋势。所以对变换后的数据进行时间序列分析。

利用 R 软件 tseries 包中的 adf.test()函数对变换后的吞吐量做单位根检验， p-value = 0.99，说明不能拒绝存在单位根的假设。对变换后的数据做两阶差分后对应的 p-value < 0.01，说明差分后的序列平稳，

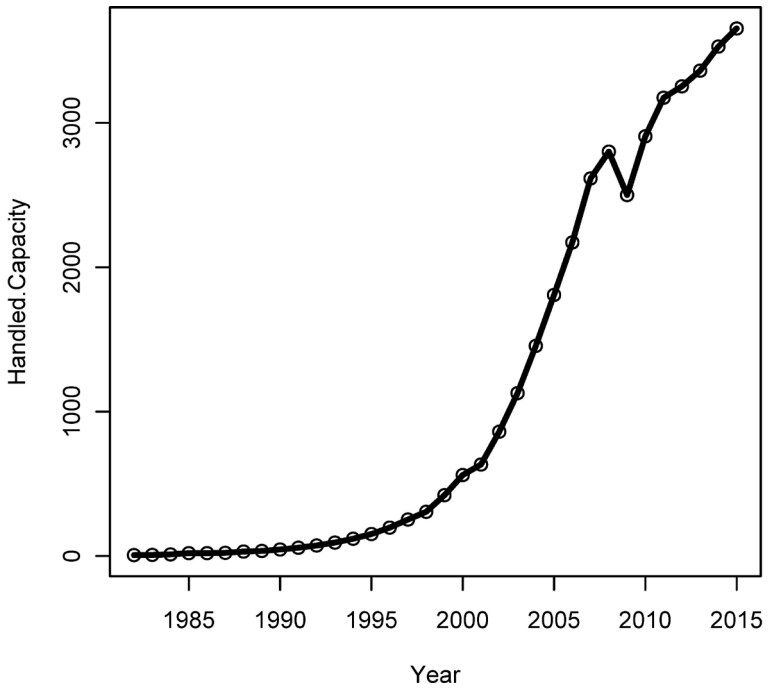


Figure 1. Shanghai port's container throughput from 1982 to 2015
图 1. 1982~2015 年集装箱吞吐量

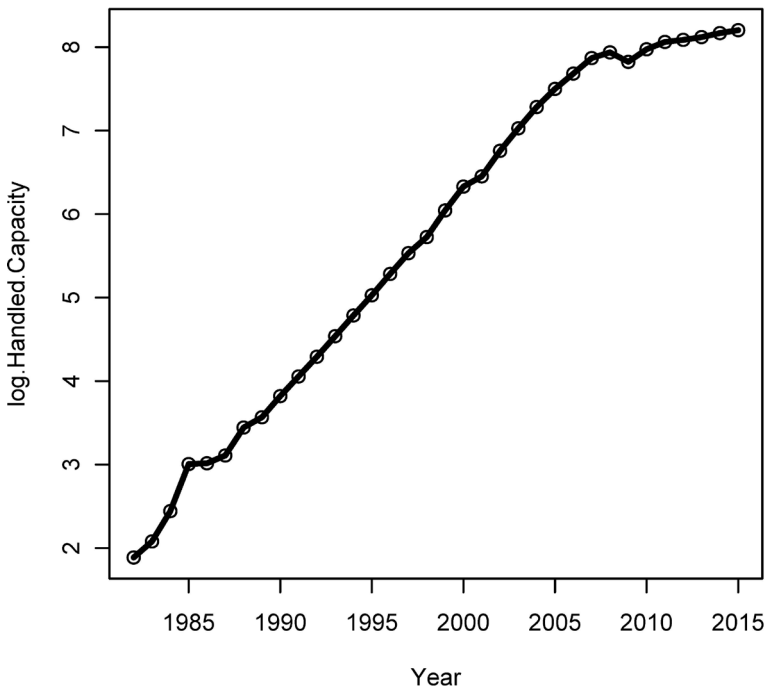


Figure 2. Shanghai port's logarithmic container throughput from 1982 to 2015
图 2. 1982~2015 年集装箱对数吞吐量

且序列的 ACF，PACF 图如下。

由图 3 可知，ACF 拖尾，PACF 两步截尾。可以采用 Bayes AR(2)模型 $\dot{Y}_t = \theta_1 \dot{Y}_{t-1} + \theta_2 \dot{Y}_{t-2} + \varepsilon_t, t = 3,4,\cdots T$

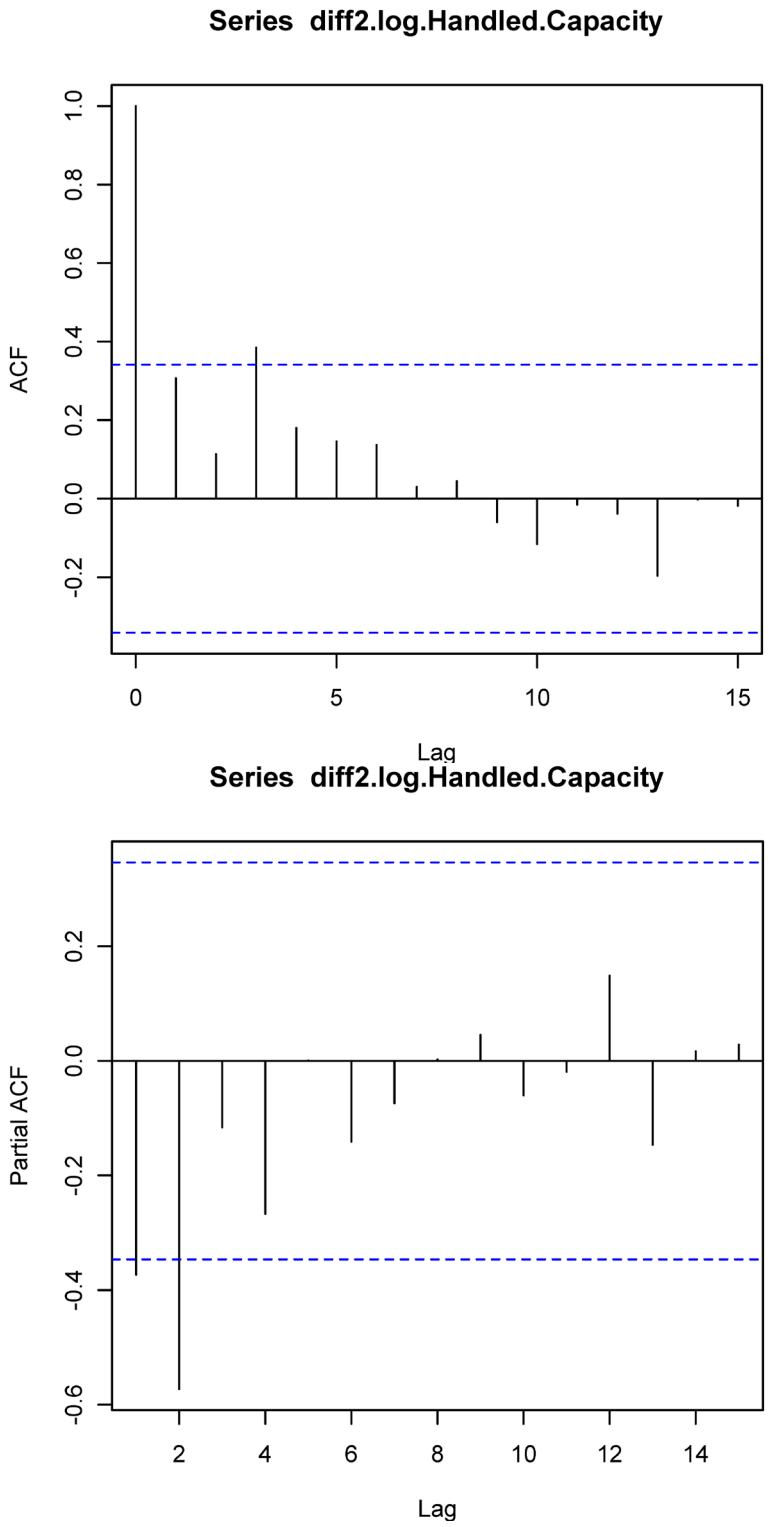


Figure 3. ACF (up), PACF (down) of logarithmic container throughput's second order difference
图 3. 对数吞吐量二阶差分的 ACF(上)、PACF(下)

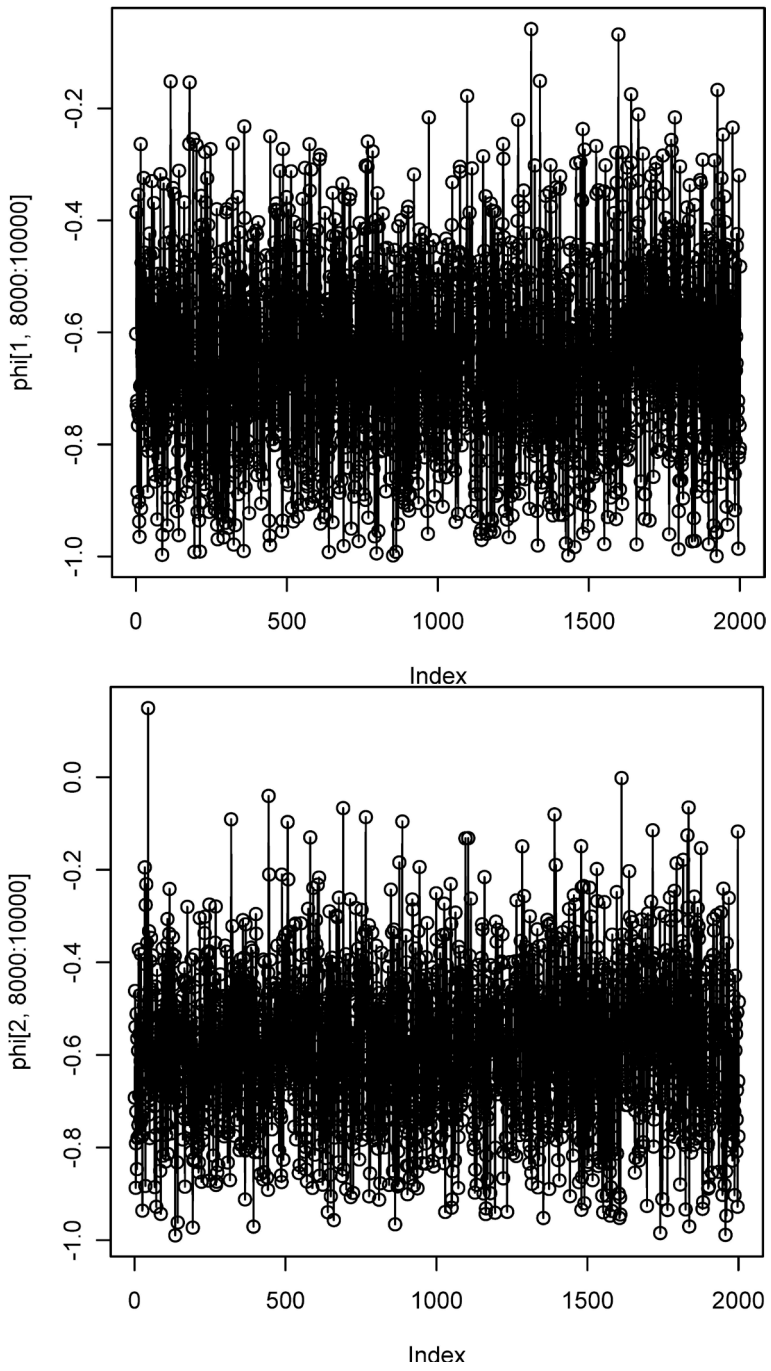
对数据进行拟合，利用上文的 Gibbs 抽样方法对模型参数进行估计，且先验参数与模拟实验部分一致，Gibbs 抽样模拟轨道长 10,000，绘出参数的后 2000 次迭代过程(见图 4)。

可见参数迭代平稳，并抛弃轨道的前 5000 个点，计算参数的后验均值与标准差见表 3。其中括号内数字代表参数后验标准差。

则模型的估计结果为

$$Z_t = -0.6465Z_{t-1} - 0.5856Z_{t-2}, t = 3,4,\cdots T$$

其中，$Z_t = \log(Y_t) - 2\log(Y_{t-1}) + \log(Y_{t-2})$，于是我们可得未来三年的集装箱吞吐量为 3782.107TEU，



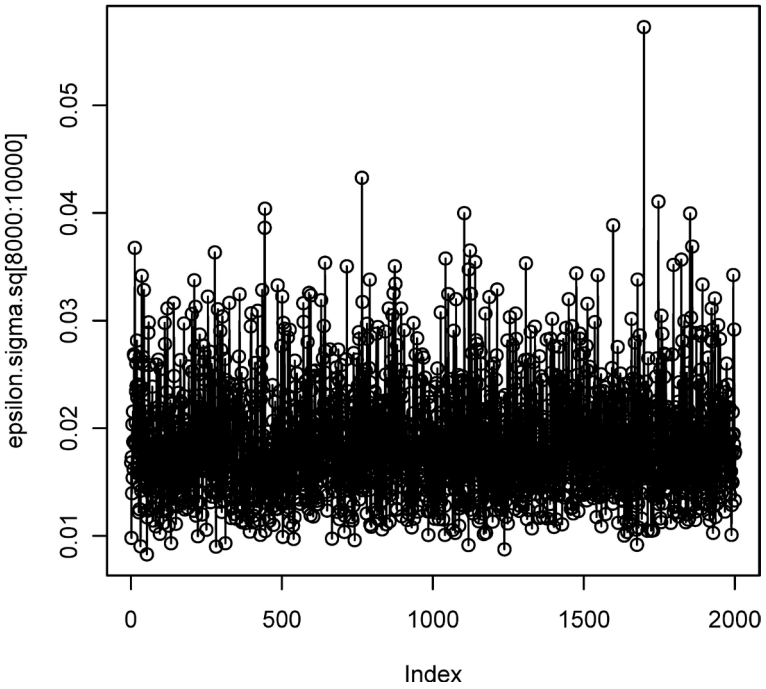


Figure 4. The last 2000 iterations procedure of the parameter
图 4. 参数后 2000 次迭代过程

Table 3. The estimation of parameters posterior mean
表 3. 参数后验均值结果

θ_1	θ_2	σ^2
−0.6532(0.166)	−0.5847(0.172)	0.0195(0.0053)

3947.072TEU，4098.362TEU。

由于贝叶斯参数估计方法充分利用了样本的信息和模型信息，估计方法更加灵活，对诸如金融危机之类的突发性事件有着较好的应变能力，也能克服传统估计方法中因为样本不足或者质量不佳导致结果误差较大的缺陷。用贝叶斯估计未知参数的方法得到的模型更适合预测，更能反映现实问题。

参考文献 (References)

[1] 田歆, 曹志刚, 骆家伟, 等. 基于 TEI@I 方法论的香港集装箱吞吐量预测方法[J]. 运筹与管理, 2009, 18(4): 82-89.

[2] 杨金花, 杨艺. 基于灰色模型的上海港集装箱吞吐量预测[J]. 上海海事大学学报, 2014, 35(2): 28-32.

[3] 余思勤, 范莹莹. 基于 NARX 神经网络的港口集装箱吞吐量预测[J]. 上海海事大学学报, 2015, 36(4): 1-5.

[4] Fernandez, C., Osiewalski, J. and Steel, M. (1997) On the Use of Panel Data in Stochastic Frontier Models with Improper Priors. *Journal of Econometrics*, **79**, 169-193. https://doi.org/10.1016/S0304-4076(97)88050-5

[5] Altaleb, A. and Chauveau, D. (2002) Bayesian Analysis of the Logit Model and Comparison of Two Metropolis-Hastings Strategies. *Computational Statistics and Data Analysis*, **39**, 137-152. https://doi.org/10.1016/S0167-9473(01)00055-X

[6] 茆诗松, 王静龙, 濮晓龙. 高等数理统计[M]. 第 2 版. 北京: 高等教育出版社, 2006: 120-121.

[7] 上海统计网[Z/OL]. http://www.stats-sh.gov.cn/