

Air Traffic Forecast Empirical Research Based on the MCMC Method

Jian-bo Wang¹, Chong-jun Fan¹ & Lei Bai¹

¹ Business School, University of Shanghai for Science and Technology, Shanghai 200093, China

Correspondence: Chong-jun Fan, Business School, University of Shanghai for Science and Technology, Shanghai 200093, China. Tel: 86-136-2197-4830. E-mail: wangjianbai@hotmail.com

Received: June 11, 2012 Accepted: July 2, 2012 Online Published: July 27, 2012

doi:10.5539/cis.v5n5p50

URL: <http://dx.doi.org/10.5539/cis.v5n5p50>

Abstract

Airport air traffic is one of the most important and hardest one among all the airport data forecasts. In this paper, the Markov Chain Monte Carlo (MCMC) method of applied theory of statistics has been introduced into the aviation sector, and the discussion on airport air traffic forecast has been conducted taking Shanghai airport as the application background. MCMC method is designed to solve the problem parameters expectations where some bayesian inference can not be directly calculated. This paper conducts the iterative computation using WinBUGS, EViews is used to estimate the initial iteration parameter of the regression model, and it is on this basis that the regression equation under the MCMC method is obtained.

Keywords: MCMC, airport, traffic data, multiple linear regression

1. Introduction

In Bayesian statistical learning, the estimation of overall parameter often needs high-dimensional probability distribution function to do the integral. Under normal circumstances, there is no clear expression of this integral, so it is very difficult to do it. In this case, the Markov Chain Monte Carlo (MCMC) method is a simple and effective calculation method.

The basic idea of MCMC method is to construct a Markov Chain, making its stationary distribution the posterior distribution of the parameters to be estimated, thus generating the samples of posterior distribution through this Markov Chain, and doing the Monte Carlo integration of the samples of Markov Chain when they reach stationary distribution. For the reason that the joint distribution of the parameters and variables in many time series models are complex and high dimensional, it is impossible to extract samples directly from the distribution. MCMC can generate samples from the joint density of parameters, without having to know the specific forms of expression of density. The most commonly used MCMC algorithm is Gibbs sampling, which is more accessible to gain the conditional distribution density through simulation, and it adopts iterative procedure to construct the Markov Chain. Under the simulation of large sample, it is expected to converge to the target distribution (Lin & Han, 2007; Stephen, 1998; Lin & Han, 2005).

WinBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009) is a kind of specialized statistical software based on MCMC method, in which all unknown parameters are regarded as random variables in order to find out the solution of this kind of probability model then. In this paper, basing on the regression model, pretreatment is done to the initial value of the MCMC model, and empirical study is done regarding the relationship between air traffic, GDP growth rate and population growth (David, 2001). The GDP means the money value of all the final goods and services produced in the domestic territory of a country in a year's time. In this paper, we conduct three parts, including MCMC method description, Empirical research on air traffic forecast and conclusions.

2. MCMC Method and Gibbs Sampling

The MCMC method is a general simulation method for sampling from posterior distributions and computing posterior quantities of interest. MCMC methods sample succeed from a target distribution, and each sample depends on the previous one, hence the notion of the Markov Chain.

A Markov Chain is a sequence of random variables, $\theta^1, \theta^2, \dots$, in which the random variable θ^t depends on all previous θ^s only through its immediate predecessor θ^{t-1} . The Markov Chain applied to sampling as a mechanism that traverses randomly through a target distribution without having any memory of where it has been. Where it moves next is entirely dependent on where it is now.

Monte Carlo, as in Monte Carlo integration, is mainly used to approximate an expectation by using the Markov Chain samples. In the simplest version

$$\int_S g(\theta)p(\theta)d\theta \cong \frac{1}{n} \sum_{t=1}^n g(\theta^t) \quad (1)$$

where $g(\cdot)$ is a function of interest and θ^t are samples from $p(\theta)$ on its support S . This approximates the expected value of $g(\theta)$.

The Gibbs sampler is a special case of the Metropolis sampler in which the proposal distributions exactly match the posterior conditional distributions and proposals are accepted 100% of the time. The main features of the Gibbs sampling, a Markov Chain sample, is extracted from a series of conditional distributions (George & Edward, 1992; Gelfand & Smith, 1990).

Suppose $\theta = (\theta_1, \dots, \theta_k)'$ is the parameter vector, $p(y|\theta)$ is the likelihood, and $\pi(\theta)$ is the prior distribution. The full posterior conditional distribution of $\pi(\theta_i | \theta_j, i \neq j, y)$ is proportional to the joint posterior density; that is,

$$\pi(\theta_i | \theta_j, i \neq j, y) \propto p(y|\theta)\pi(\theta) \quad (2)$$

For instance, the one-dimensional conditional distribution of θ_1 given $\theta_j = \theta_j^*, 2 \leq j \leq k$ is computed as the following:

$$\pi(\theta_1 | \theta_j = \theta_j^*, 2 \leq j \leq k, y) = p(y | (\theta = \theta_1, \theta_2^*, \dots, \theta_k^*))\pi(\theta = (\theta_1, \theta_2^*, \dots, \theta_k^*)) \quad (3)$$

The Gibbs sampler works as follows:

Set $t = 0$ and choose an arbitrary initial value of $\theta^{(0)} = \{\theta_1^{(0)}, \dots, \theta_k^{(0)}\}$.

Generate each component of θ as follows:

draw $\theta_1^{(t+1)}$ from $\pi(\theta_1 | \theta_2^{(t)}, \dots, \theta_k^{(t)}, y)$
draw $\theta_2^{(t+1)}$ from $\pi(\theta_2 | \theta_1^{(t+1)}, \theta_3^{(t)}, \dots, \theta_k^{(t)}, y)$
...
draw $\theta_k^{(t+1)}$ from $\pi(\theta_k | \theta_1^{(t+1)}, \dots, \theta_{k-1}^{(t+1)}, y)$

Set $t = t + 1$. If $t < T$ the number of desired samples, return to step 2.

When Gibbs sampling iterates enough number of steps, we can believe the sample is subject to the target distribution, or converge to the target distribution. The choice of the initial value will affect the convergence rate of MCMC.

3. Empirical Research on Air Traffic Forecast

It is rather important to choose proper factors ahead of modeling and estimation. The following parts include factors analysis and parameters estimation.

3.1 Influencing Factors Analysis

Air traffic forecast has always been a difficult and key point of the airport aviation forecasts. There are many aspects of factors that influence air traffic, in this paper, macro-economic factor and demographic impact factor are used to analyze the air traffic. For economic factor, GDP is an important standard to measure the comprehensive level of the economic development; population growth rate index is used as demographic factor. At the same time, because the data of GDP and population growth rate are rather consecutive, the reference value of the data is quite high. We conduct a linear regression of air traffic growth based on GDP growth rate and population growth rate. Therefrom a binary regression model is built as follows:

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 \quad (4)$$

Where $\alpha_0, \alpha_1, \alpha_2$ are parameters to be estimated, X_1 and X_2 are know vectors stand for GDP growth and population growth. Y stands for Airport traffic volume growth. This article uses Eviews simulation, according to data released by National Bureau of Statistic from 1990 to 2009, and then Equation (4) would have the form:

$$Y = 3.071 + 0.5453X_1 + 3.771X_2 \quad (5)$$

3.2 Parameters Estimation Based on the MCMC Method

WinBUGS is a special package based on MCMC method, which can easily conduct Gibbs sampling on many models and distributions. Before implementing Bayesian analysis, programmers need to set prior distribution and initial values of variables. The analysis processes include modeling, data loading, initial values and iterative analysis. In this case, the priors are defined as

Regression equation:

$$muy[i] < -\alpha_0 + \alpha_1 X_1[i] + \alpha_2 X_2[i] \quad (6)$$

Prior distribution of the parameters:

$$\alpha_0 \sim N(0, 1.0E-6); \alpha_1 \sim N(0, 1.0E-6); \alpha_2 \sim N(0, 1.0E-6); \tau \sim Gamma(0.001, 0.001); Y_i \sim N(muy[i], \tau).$$

The initial parameter value:

$$\alpha_0 = 3.071; \alpha_1 = 0.5453; \alpha_2 = 3.771; \tau = 1.$$

In order to ensure the convergence of the parameters, we first conduct 5000 Gibbs pre-iterations, and then discard the initial iteration. We construct the model using a graphical interface Doodle (Figure 1). Meanwhile, we set up three Markov Chains in WinBUGS, and at the same time enter the initial values. This is because, for the convergence parameters, (and) three iterations results of the Markov Chain tend to be overlapping. It is conducive to the observation of the results.

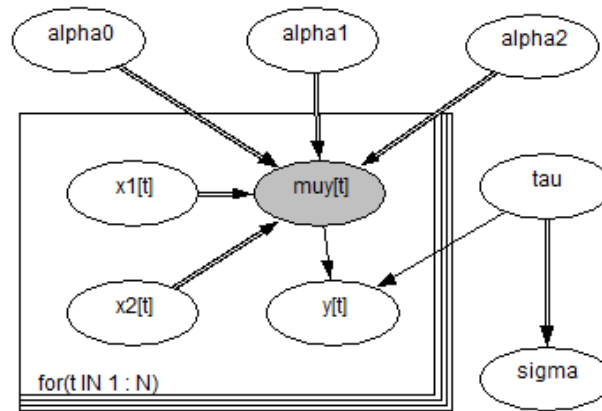


Figure 1. Bayesian modeling in WinBUGS

Iterations start from 5001 to 60000. The statistical results show as in Table 1, which produces summary statistics for the variable, pooling over the chains selected. In this case, we choose 2.5% and 97.5% as the confidence interval. The quantity reported in the MC error column gives an estimate of $\sigma / N^{1/2}$, the Monte Carlo standard error of the mean. During the 95% Confidence intervals, the average value of α_0 , α_1 , α_2 and τ are the interval (-14.13, 20.24), (-0.1383, 1.241), (-17.26, 24.69), (0.004114, 0.009146) respectively.

Table 1. Posterior estimate statistics for monitored parameters with 60000 iterations

node	mean	sd	MC error	2.50%	median	97.50%	start	sample
α_0	3.054	8.653	0.03505	-14.13	3.062	20.24	5001	60000
α_1	0.546	0.3488	0.001483	-0.1383	0.5455	1.241	5001	60000
α_2	3.756	10.56	0.04296	-17.26	3.755	24.69	5001	60000
τ	0.009543	0.00338	1.53E-05	0.004114	0.009146	0.01723	5001	60000

Before drawing our conclusion, we need to have a judgment on density estimation and convergence statistics of the parameters. The following plots a smoothed Posterior Kernel Density Estimate for the variable if it is continuous or a histogram if it is discrete (Figure 2).

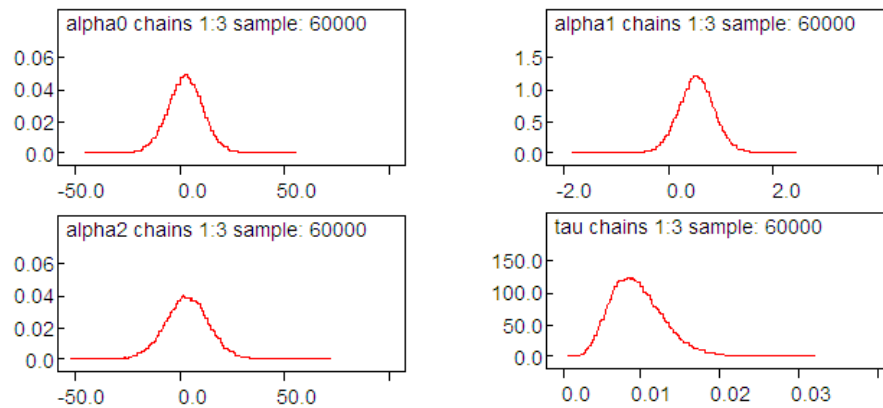


Figure 2. Posterior Kernel Density Estimation for monitored parameters with 60000 iterations

In this case, we are running three chains simultaneously. The trace and history plots show each chain in a different color. As the plots show (Figure 3), all the chains appear to be overlapping one another, therefore, we can be reasonably confident that convergence has been achieved.

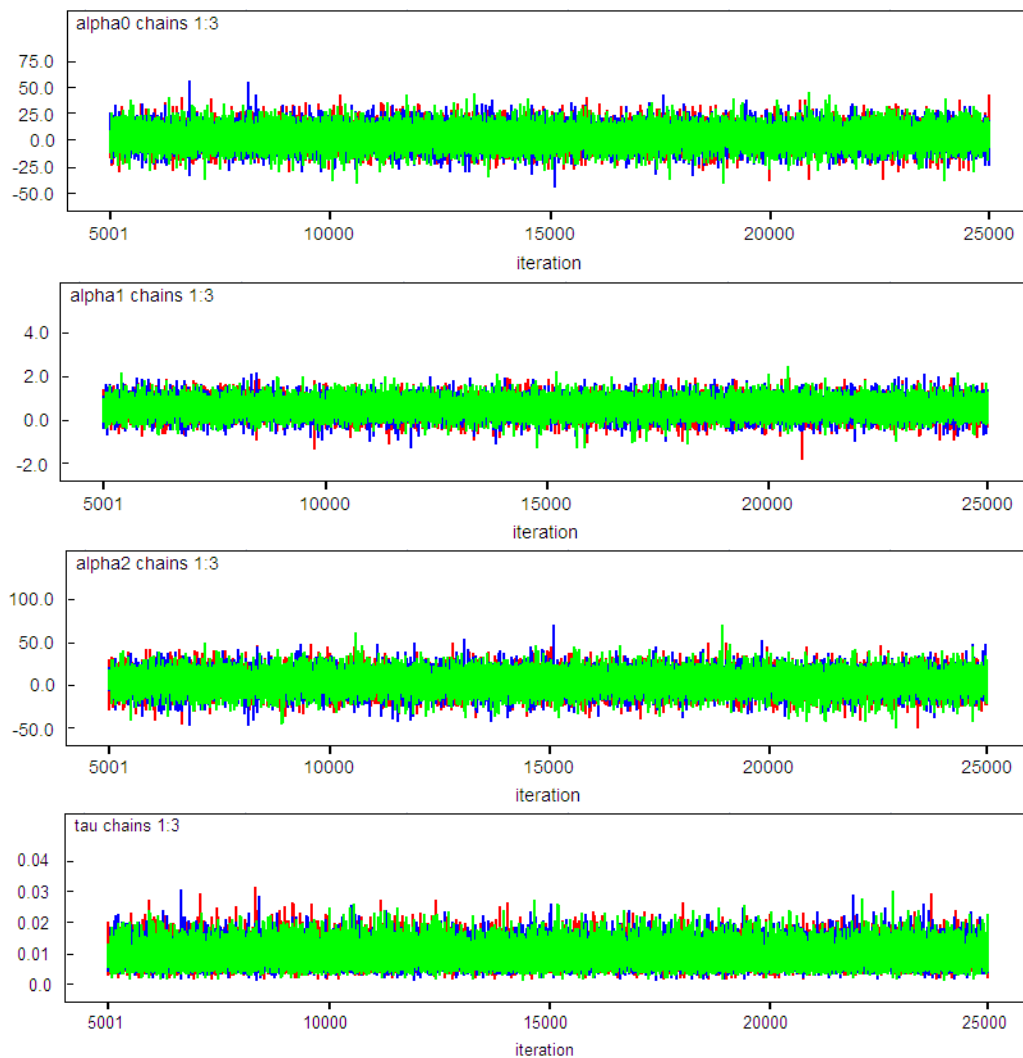


Figure 3. Trace plots for monitored parameters' three Chains iterations

The following plots (Figure 4) show the Gelman-Rubin convergence statistics. The width of the central 80% interval of the pooled runs is green, the average width of the 80% intervals within the individual runs is blue, and their ratio R (= pooled / within) is red. R would generally be expected to be greater than 1 if the starting values are suitably over-dispersed. The proper one should be concerned both with convergence of R to 1, and with convergence of both the pooled and within interval widths to stability.

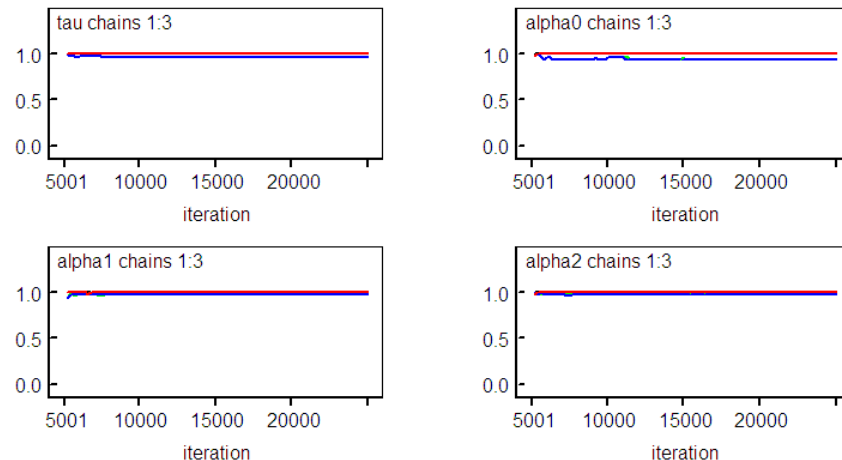


Figure 4. Parameters' convergence diagnostics

According to the data of Table 1, as well as Posterior Kernel Density Estimation, Trace Plots, Parameters' Convergence Diagnostics, we believe that parameters estimation based on MCMC method is valid. The linear regression model equation is shown as in Equation (7).

$$Y = 3.054 + 0.546X_1 + 3.756X_2 \quad (7)$$

4. Conclusions

Air traffic forecast has always been a problem troubling airport senior decision-makers to make scientific decisions, but the use of conventional forecasting methods often cannot get satisfactory results. In this paper, MCMC simulation methods and WinBUGS software applications based on Gibbs sampling are used to conduct the volatility analysis of the airport air traffic data. The analytical method proposed in this paper can provide new ideas and methods for the scientific decision-making of the airport senior decision-makers. We hope that, based on the fundamental prediction research of the MCMC method, airports can have a deeper level of study.

References

- David, P. M. S. (2001). Actuarial modeling with MCMC and BUGS. *North American Actuarial Journal*, 5(2), 96-125.
- Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *J. American Statistical Association*, 85, 398-409. <http://dx.doi.org/10.1080/01621459.1990.10476213>
- George, C., & Edward, I. G. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46, 167-174.
- Lin, J., & Han, Y. (2005). Exponential Regression Model Based on MCMC Method and Its Application. *Operations Research and Management Science*, 14(4), 95-100.
- Lin, J., & Han, Y. (2007). Gamma Process Priors Model Based on MCMC Simulation and Its Application in Reliability. *Journal of System Simulation*, 19(22), 5099-5102.
- Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 28, 3049-3067. <http://dx.doi.org/10.1002/sim.3680>
- Stephen, P. B. (1998). Series D (The Statistician): Chain Monte Carlo Method and Its Application. *Journal of the Royal Statistical Society*, 47(1), 69-100.