

基于Bayes推断的ARIMAX模型的建模分析

孙鹏越, 汤禧俊

华南农业大学数学与信息学院, 广东 广州

收稿日期: 2025年3月23日; 录用日期: 2025年4月14日; 发布日期: 2025年4月25日

摘要

ARIMAX模型在多个领域有着重要应用。但是针对ARIMAX模型的参数估计均是经典统计方法, 利用贝叶斯估计是一个值得探究的问题。考虑到共轭先验分布的性质特点, 对于ARIMAX模型中的系数的先验分布为正态分布, 噪声项的方差先验分布为逆伽马分布假定, 本文给出了参数的后验分布, 并使用Gibbs采样的方式, 给出各个参数的一个估计。模拟试验的结果表明, 本文的估计方法具有很好的功效, 借助参数的迭代图表明文中使用的方法具有稳健性。使用创业板数据与上证数据做出实证分析, 并发现文中给出的方法不仅具有很好的解释性, 同时能够提取出完整的数据信息。

关键词

ARIMAX模型, 贝叶斯, Gibbs采样, 平稳

Modeling Analysis of ARIMAX Model Based on Bayes Inference

Pengyue Sun, Xijun Tang

College of Mathematics and Informatics, South China Agricultural University, Guangzhou Guangdong

Received: Mar. 23rd, 2025; accepted: Apr. 14th, 2025; published: Apr. 25th, 2025

Abstract

The ARIMAX model has important applications in various fields. However, parameter estimation for the ARIMAX model has traditionally relied on classical statistical methods, and exploring Bayesian estimation presents a worthwhile direction. Considering the properties of conjugate prior distributions, this study assumes a normal prior distribution for the coefficients in the ARIMAX model and an inverse gamma prior for the variance of the noise term. The posterior distributions of the parameters are derived, and estimates for each parameter are obtained using Gibbs sampling. Simulation results demonstrate the effectiveness of the proposed estimation method. Iterative plots of the parameters

indicate the robustness of the method. An empirical analysis using data from the ChiNext Index and the Shanghai Composite Index further reveals that the proposed method not only offers strong interpretability but also effectively captures comprehensive information from the data.

Keywords

ARIMAX Model, Bayes, Gibbs Sampler, Stationary

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

Box 和 Tiao (1975) [1] 针对非平稳时间序列数据, 首先将 ARMA 模型(Autoregressive Moving Average, ARMA)结合差分运算, 提出求和移动平均自回归模型(Autoregressive Integrated Moving Average, ARIMA), 但是很多数据不仅仅受自身的影响, 还要受到其他多个变量的影响。为此 Engle 和 Granger (1987) [2] 在 ARIMA 的基础上, 引入外生变量 $\{x_{1t}\}, \{x_{2t}\}, \dots, \{x_{kt}\}$, 提出 ARIMAX 模型(Autoregressive Integrated Moving Average with Exogenous Variables, ARIMAX)。研究人员发现, 使用 ARIMAX 模型在经济学与金融领域往往有良好的预测以及拟合效果, 例如 Peter 和 Silvia (2012) [3] 使用人均 GDP 作为响应变量, 失业率作为外生变量, 发现 ARIMAX 的解释性要优于 ARIMA 模型; Sharma 和 Parkhar (2021) [4] 借助 ARIMAX 模型对股票市场预测进行敏感度分析。但是在使用 ARIMAX 模型应用之前, 各项系数的确定是一个值得探究的问题。传统的估计手段是 Engle 和 Granger (1987) [2] 使用协整回归的方式进行检验, 不同于传统的方式, 本文参考夏强和刘金山(2011) [5] 对 TAR 模型门限变量使用贝叶斯的思想进行检验的手段, 对 ARIMAX 模型的系数进行估计。假定先验分布已知, 对于计算后验概率的方式, 本文采取马尔可夫蒙特卡洛模拟(MCMC)方法中的 Gibbs 采样的方式计算。

本文剩余章节安排如下: 第二章介绍 ARIMAX 模型与各个参数的后验分布, 第三章为数值模拟, 包含数值分析情形与实证数据分析, 第四章是对文章的总结与概括。

2. 模型及估计方法

2.1. ARIMAX 模型介绍

假设响应序列 $\{y_t\}$ 和输入序列(自变量序列) $\{x_{1t}\}, \{x_{2t}\}, \dots, \{x_{kt}\}$ 平稳, 且响应序列和输入序列之间具有相关关系。响应序列 $\{y_t\}$ 和输入序列 $\{x_{1t}\}, \{x_{2t}\}, \dots, \{x_{kt}\}$ 最终建立的线性模型为:

$$y_t = \beta_0 + \sum_{i=1}^k \frac{\Psi_i(B)}{\Phi_i(B)} B^{r_i} x_{it} + \frac{\Psi(B)}{\Phi(B)} a_t \quad (1)$$

其中 $t \in \{1, 2, \dots, n\}$ 为时间参数, B 代表一阶滞后算子, r_i 表示第 i 个变量的延迟阶数, a_t 表示零均值白噪声序列, 以及 $\Phi_i(B)$ 表示第 i 个自变量序列 $\{x_{it}\}$ 的 p_i 阶自回归系数多项式:

$$\Phi_i(B) = 1 - \phi_{i1}B - \phi_{i2}B^2 - \dots - \phi_{ip_i}B^{p_i}, i = 1, \dots, k$$

且 $\Psi_i(B)$ 是第 i 个自变量序列 $\{x_{it}\}$ 的 m_i 阶移动平均系数多项式:

$$\Psi_i(B) = \psi_{i0} - \psi_{i1}B - \dots - \psi_{im_i}B^{m_i}, i = 1, \dots, k$$

经过一系列变换, 模型(1)可以转变为

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \sum_{i=1}^k \Theta_i(B) B^{l_i} x_{it} + \varepsilon_t \quad (2)$$

此时, 式中 $\Theta_i(B) = \theta_{i0} + \theta_{i1}B + \cdots + \theta_{iq_i}B^{q_i}$ 为第 i 个自变量 $\{x_{it}\}$ 的 q_i 阶移动平均系数多项式, q_i 为第 i 个自变量 x_{it} 对变量 y_t 的有效作用时长, ε_t 服从标准正态分布。现在模型中, 未知的参数为 $\{y_t\}$ 的自回归系数, 记 $\Phi = (\phi_1, \cdots, \phi_p)'$; 以及自变量序列的移动平均系数, 记为 $\Theta_i = (\theta_{i0}, \cdots, \theta_{iq_i})', i = 1, \cdots, k$ 。

2.2. ARIMAX 模型参数的贝叶斯推断

为了对参数进行贝叶斯推断, 首先要有对参数的先验分布的认识, 这样才能得到参数的条件后验分布(Gelfand 和 Smith (1990) [6]), 考虑到共轭先验分布的性质特点[7], 假定参数向量 Φ 服从多元正态分布 $N_p(\mu, \Sigma)$; Θ_i 之间互相独立, 并且服从多元正态分布 $N_{q_i}(\alpha_i, \Xi_i)$, 噪声项的方差 σ^2 服从逆伽马分布 $LG(\beta, \gamma)$ 。

为书写方便起见, 将模型(2)中的每个观测按行摆放, 且不妨设 t 足够大, 记

$$Y_0 = \begin{pmatrix} y_{t-1} & \cdots & y_{t-p} \\ \vdots & \ddots & \vdots \\ y_{n-1} & \cdots & y_{n-p} \end{pmatrix}, X_i = \begin{pmatrix} x_{i,t-l_i} & \cdots & x_{i,t-l_i-q_i} \\ \vdots & \ddots & \vdots \\ x_{i,n-l_i} & \cdots & x_{i,n-l_i-q_i} \end{pmatrix}, i = 1, \cdots, k$$

并记因变量 $Y = (y_t, \cdots, y_n)'$, 于是模型(2.2)可化为:

$$Y = Y_0\Phi + \sum_{i=1}^k X_i\Theta_i + \varepsilon$$

其中 $\varepsilon = (\varepsilon_t, \cdots, \varepsilon_n)'$, 且服从多元正态分布 $N_{n-t+1}(0, \sigma^2 I_{n-t+1})$ 。因此可以得到参数的联合先验分布:

$$\pi(\Phi, \Theta_i, \sigma^2) \propto \frac{1}{\sigma^n} \exp \left\{ -\frac{1}{2\sigma^2} \left[\left(Y - Y_0\Phi - \sum_{i=1}^k X_i\Theta_i \right)' \left(Y - Y_0\Phi - \sum_{i=1}^k X_i\Theta_i \right) \right] \right\}$$

于是可以得到参数 σ^2 条件后验分布为逆伽马分布:

$$P(\sigma^2 | \Phi, \Theta_i) \sim LG \left(\beta + \frac{n}{2}, \gamma + E'E \right) \quad (3)$$

其中

$$E = Y - Y_0\Phi - \sum_{i=1}^k X_i\Theta_i$$

参数 Φ 的后验分布为

$$P(\Phi | \sigma^2, \Theta_i) \sim \mathcal{N}_p(\mu^*, \Sigma^*) \quad (4)$$

其中

$$\mu^* = (Y_0\Sigma_0^{-1}Y_0 + \Sigma^{-1})^{-1} (Y_0'\Sigma_0^{-1}E + \mu), \Sigma^* = (Y_0\Sigma_0^{-1}Y_0 + \Sigma^{-1})^{-1}$$

参数 $\Theta_i, i = 1, \cdots, k$ 的后验分布为正态分布

$$P(\Theta_i | \sigma^2, \Phi) \sim N_{q_i}(\alpha_i^*, \Xi_i^*) \quad (5)$$

其中

$$\alpha_i^* = \left(X_i' \Sigma_0^{-1} X_i + \Xi_i^{-1} \right)^{-1} \left(X_i' \Sigma_0^{-1} E + \Xi_i^{-1} \alpha_i \right), \quad \Xi_i^* = \left(X_i' \Sigma_0^{-1} X_i + \Xi_i^{-1} \right)^{-1}$$

可以发现各个参数的后验分布都是标准的分布, 那么就可以使用 Gibbs 采样[8]的方法, 得到各个参数的估计量。

3. 数值分析

3.1. 试验模拟

将模型设置为:

$$y_t = 0.5y_{t-1} + 0.4y_{t-2} + 0.5x_{t-1} + 0.3x_{t-2} + \varepsilon_t$$

ε_t 服从标准正态分布。自变量序列通过下述方式迭代产生:

$$x_t = 0.5x_{t-1} + \rho_t$$

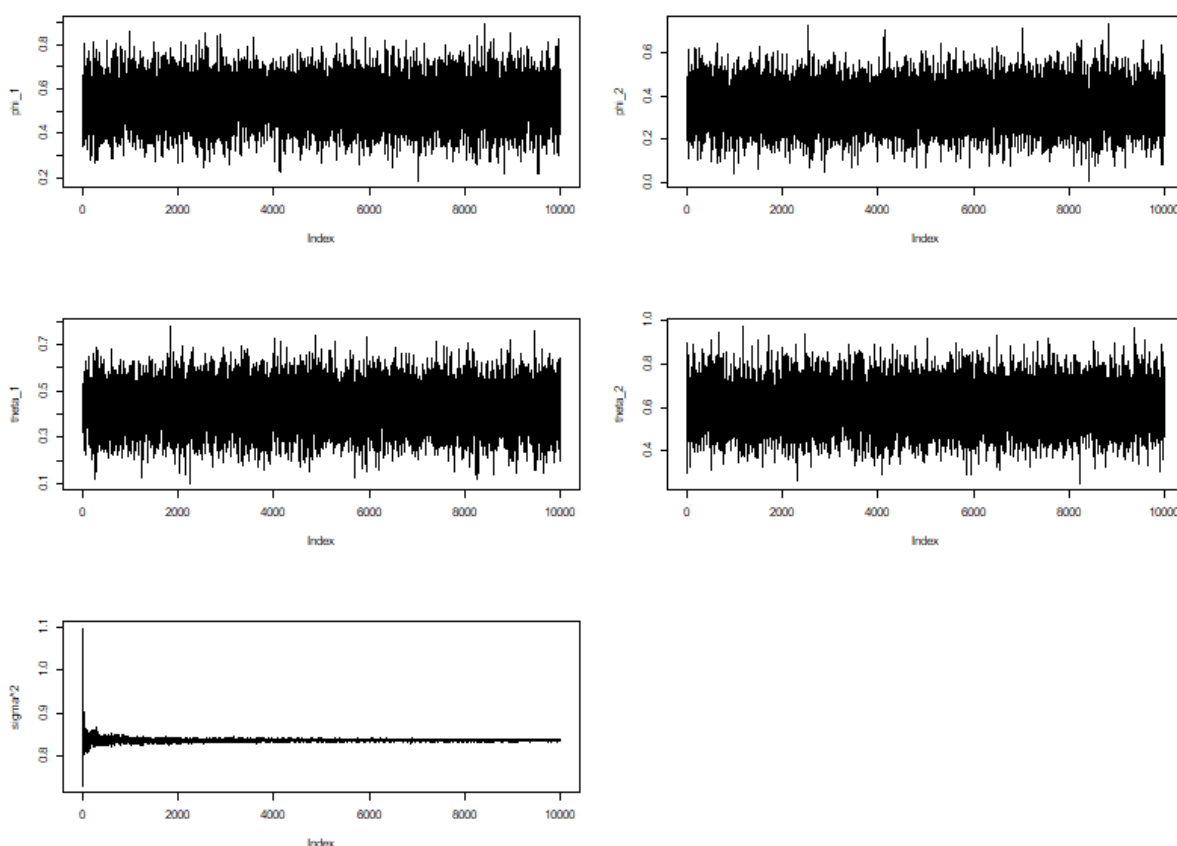


Figure 1. Diagram of parameter Φ, Θ, σ^2 iteration

图 1. 参数 Φ, Θ, σ^2 的迭代图

ρ_t 服从标准正态分布, 初始值 x_1 为一个(0,1)的均匀分布的随机数。根据模型(2), 可以观察到各个参数的真实值为 $\sigma^2 = 1, \Phi = (0.5, 0.4)', \Theta = (0.5, 0.3)'$ 。由于超参数对实验结果的选择影响不大, 见文献 Casella (2001) [9]与 Doss (2012) [10], 于是可以设置噪声项的方差 σ^2 的先验分布为 $LG(1.5, 1)$, 系数 Φ 与 Θ 均为标准多元正态分布。采用 Gibbs 采样的方法, 估计各个参数。但是使用 Gibbs 抽样时的一个问题是, 在什

么时候抽样可以停止, 即收敛, 为此我们设置迭代一万次, 并舍弃前五千次的数据, 使用后五千次的数据进行后验分析, 并绘制迹图。算法流程如下:

算法 1. 试验流程

(1) 生成 100 个数据, 记为 $y_1, y_2, \dots, y_{100}$, 将第 3 到第 100 个数据作为因变量向量 $Y = (y_3, y_4, \dots, y_{100})'$, 设置初始值

$$Y_0 = \begin{pmatrix} y_1, y_2, \dots, y_{98} \\ y_2, y_3, \dots, y_{99} \end{pmatrix}', \quad X_0 = \begin{pmatrix} x_1, x_2, \dots, x_{98} \\ x_2, x_3, \dots, x_{99} \end{pmatrix}';$$

- (2) 利用(3)抽取 σ^2 , 利用(4)抽取 Φ , 利用(5)抽取参数 Θ ;
- (3) 重复(1)与(2) 10000 次;
- (4) 剔除前 5000 次数据, 求后 5000 次的均值, 得到参数的估计 $\hat{\sigma}^2, \hat{\Phi}, \hat{\Theta}$;
- (5) 重复(1)至(4) 100 次, 对得到的数据求均值以及标准差。

表 1 中给出各个参数的估计值, 可以发现, 通过 Gibbs 取样, 在进行一百次抽样试验中, 可以把误差控制在 0.05 以内。图 1 为使用 Gibbs 采样时, 迭代的迹图, 在迭代 5000 次之后发现参数已经是收敛的, 因此认为估计是稳健的。

Table 1. Bayes estimates of parameter σ^2, Φ, Θ under 100 simulations

表 1. 参数 σ^2, Φ, Θ 贝叶斯估计值在 100 次模拟下的估计值

参数	真值	实验结果		误差		标准差	
σ^2	1.0	1.0811		0.0811		0.1855	
Φ	0.5 0.4	0.4549	0.4242	-0.0451	-0.0309	0.0733	0.0718
Θ	0.5 0.3	0.4867	0.3492	-0.0133	0.0492	0.1048	0.1279

在表 2 中, 给出协整的方式估计的参数值, 通过对比可以发现使用 Bayes 的方式进行估计, 大部分情况下优于经典方法(Engle 和 Granger (1987) [2])使用协整的方式进行估计。

Table 2. Co-integration estimates of parameter σ^2, Φ, Θ co-integration estimates under 100 simulations

表 2. 参数 σ^2, Φ, Θ 协整估计值在 100 次模拟下的估计值

参数	真值	实验结果		误差		标准差	
σ^2	1.0	1.129		0.129		0.7975	
Φ	0.5 0.4	0.5769	0.3757	0.0769	-0.0343	0.2556	0.1613
Θ	0.5 0.3	0.4357	0.3419	-0.0643	0.0419	0.1851	0.1186

3.2. 实证分析

借助 ARIMAX 模型, 对金融数据进行分析建模的有很多, 例如 Badshah 等人(2023) [11], Zolfaghari 和 Gholami (2021) [12]. 本文选取自 2019 年 2 月 18 日至 2025 年 3 月 5 日的创业板指数与上证指数数据, 选取上证指数作为外生变量序列, 创业板指数作为响应变量序列。为使数据更加平稳, 对数据采取对数差分的手段。针对自回归阶数与移动平均阶数 p, q 的确定, 采用信息准则 AIC、BIC、AICc 最小的方式进行确定, 发现这是一个带有外生变量的 ARIMA(1,0,0)的模型。外生变量的滞后阶数采用逐步回归的方

式, 基于 AIC 最小的准则, 发现外生变量的滞后阶数为 0。在进行参数的估计时, 对先验分布设置为 σ^2 服从逆伽马分布 $LG(1.5,1)$, 其余两个参数均服从标准正态分布。得到的拟合结果为

$$Y_t = 0.0915Y_{t-1} + 1.3496X_t + \varepsilon_t$$

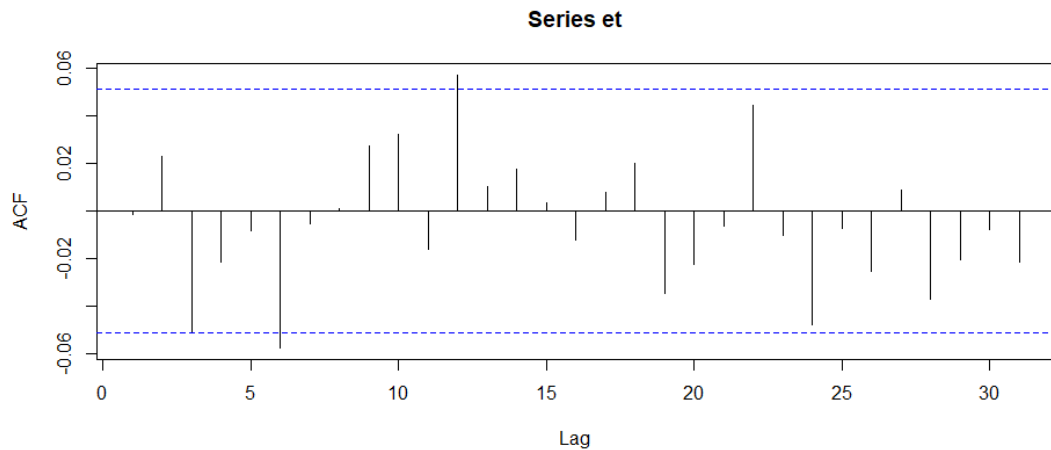


Figure 2. Diagram of autocorrelation in measure error
图 2. 测量误差的自相关图

图 2 给出模型拟合的残差的自相关函数图。表 3 给出针对实证数据模型拟合的误差输出, 从图 2 以及残差的自相关系数为 -0.0017 十分接近于 0 得知, 可以通过白噪声检验, 同时也表明 ARIMAX 模型对于创业板指数与上证指数的解释性很强, 同时将模型中的具有同响应变量具有相关信息的数据提取得十分充分。

Table 3. Errors of model
表 3. 模型误差

残差方差	AIC	AICc	BIC	平均误差	均方误差	ACF
0.0001	-9131.55	-9131.53	-9115.68	-0.0001	0.0107	-0.0017

4. 结论

从设定的模型出发, 基于贝叶斯推断, 文中给出一种新的针对 ARIMAX 模型参数估计的手段, 这个方法主要利用 Gibbs 采样计算后验概率分布, 从后验分布中抽取的均值就是参数的一个估计。在模拟试验中发现, 文中的方法的功效很好, 并且估计的方法具有稳健性, 使用这种方法在实证分析中发现对上证指数与创业板指数具有很高的解释性。

参考文献

- [1] Box, G.E.P. and Tiao, G.C. (1975) Intervention Analysis with Applications to Economic and Environmental Problems. *Journal of the American Statistical Association*, **70**, 70-79. <https://doi.org/10.1080/01621459.1975.10480264>
- [2] Engle, R.F. and Granger, C.W.J. (1987) Co-Integration and Error Correction: Representation, Estimation, and Testing. *Econometrica*, **55**, -276251. <https://doi.org/10.2307/1913236>
- [3] Peter, D. and Silvia, P.A. (2012) ARIMA vs. ARIMAX—Which Approach Is Better to Analyze and Forecast Macroeconomic Time Series. *Proceedings of 30th International Conference Mathematical Methods in Economics*, Karviná, 11-13 September 2012, 136-140.

- [4] Sharma, A., Tiwari, P., Gupta, A. and Garg, P. (2021) Use of LSTM and ARIMAX Algorithms to Analyze Impact of Sentiment Analysis in Stock Market Prediction. In: Hemanth, J., Bestak, R. and Chen, J.I.Z., Eds., *Intelligent Data Communication Technologies and Internet of Things*, Springer, 377-394. https://doi.org/10.1007/978-981-15-9509-7_32
- [5] 夏强, 刘金山. 基于贝叶斯推断的 TAR 模型的门限非线性检验[J]. 应用概率统计, 2011, 27(3): 276-282.
- [6] Gelfand, A.E. and Smith, A.F.M. (1990) Sampling-based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, **85**, 398-409. <https://doi.org/10.1080/01621459.1990.10476213>
- [7] 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.
- [8] Geman, S. and Geman, D. (1987) Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. In: Fischler, M.A. and Firschein, O., Eds., *Readings in Computer Vision*, Elsevier, 564-584. <https://doi.org/10.1016/b978-0-08-051581-6.50057-x>
- [9] Casella, G. (2001) Empirical Bayes Gibbs Sampling. *Biostatistics*, **2**, 485-500. <https://doi.org/10.1093/biostatistics/2.4.485>
- [10] Doss, H. (2012) Hyperparameter and Model Selection for Nonparametric Bayes Problems via Radon-Nikodym Derivatives. *Statistica Sinica*, **22**, 1-26. <https://doi.org/10.5705/ss.2009.259>
- [11] Waqar, B., Ibrahim, G., Cosmin, D. and Adriana, D. (2023) Identifying the Inter-Dynamics between Gold Prices of Turkey and Key Economic Indicators: An Application of Three Different Models. *Economic Computation and Economic Cybernetics Studies and Research*, **57**, 221-234. <https://doi.org/10.24818/18423264/57.3.23.13>
- [12] Zolfaghari, M. and Gholami, S. (2021) A Hybrid Approach of Adaptive Wavelet Transform, Long Short-Term Memory and ARIMA-GARCH Family Models for the Stock Index Prediction. *Expert Systems with Applications*, **182**, Article ID: 115149. <https://doi.org/10.1016/j.eswa.2021.115149>