Metodología para la predicción con modelos SARIMA Bayesianos

Daniel Dala
Departamento de Estadística
Universidad Nacional Autónoma de Honduras
e-mail: daniel.dala@unah.hn

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE CUADROS

Metodología para la predicción con modelos SARIMA Bayesianos

***Resumen*—El resumen no solo hace referencia al trabajo reportado, también sintetiza el trabajo documentado en aproximadamente 200 palabras. Establece el propósito, reporta la información obtenida, provee conclusiones, y recomendaciones. En esencia, resume los puntos principales del estudio de forma adecuada y precisa. Es importante referirse a los resultados principales obtenidos en tiempo pasado al describir el trabajo realizado.**

I. Introducción

En la introducción se describe de forma concisa pero con un poco más de detalle el trabajo realizado.

En uno de los párrafos de esta sección se hace especial énfasis en la contribución realizada como parte del trabajo reportado, considerando como punto de partida el problema central que motivo el proyecto de investigación y las soluciones obtenidas por el autor del documento como resultado del trabajo de investigación realizado.

Se presentan además descripciones breves del resto de las secciones del documento.

II. Definición del Problema

Sea $\{y_1, y_2, \ldots, y_n\}$ una serie de tiempo cualquiera, generalmente se busca predecir los valores futuros y_{n+m} con $m = 1, 2, ...h$ que son desconocidos, para esto existen diferentes metodologías como modelos de espacio y estado [2], Prophet [3], redes neuronales [4], splines [5], procesos Gaussianos [6] entre otros. Algunos de los modelos más populares son los modelos SARIMA [7], [8] por su fácil interpretación y alta capacidad predictiva, pero su implementación con datos reales es compleja debido a que seleccionar el orden del modelo es una tarea complicada. Box y Jenkins (1970) [1] propusieron una metodología para una adecuada estimación de dichos modelos, la cual se basa en cinco etapas iterativas: *revisar los datos, seleccionar modelo, estimación de parámetros, diagnóstico, selección de modelo, y predicción*. Dicha metodología se ilustra en la Figura 1.

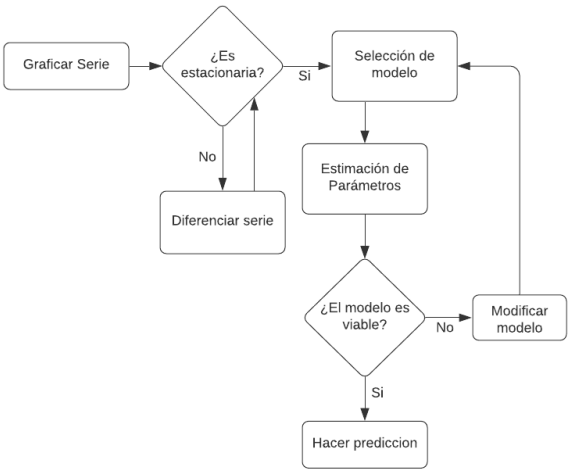

Figura 1. Metodología de Box-Jenkins (1970). El diagrama de flujo presenta el procedimiento a utilizar para un análisis de datos adecuado en un enfoque frequentista, para una mayor descripción de dicha metodología revisar [1].

Existen muchos esquemas para el proceso de inferencia, y en los últimos años la inferencia Bayesiana se ha vuelto una alternativa muy utilizada para el análisis de datos con muchas aplicaciones en economía, física, química, psicología, entre otras. Su creciente popularidad se debe a su capacidad de incorporar información externa al modelo mediante una distribución a priori, y actualizar las creencias mediante el Teorema de Bayes. Este enfoque de inferencia en la práctica es muy complicado, por lo cual en los últimos años se han aproximado los resultados mediante los métodos de Markov Chain Monte Carlo [9]. Estos métodos consisten en generar una cadena de Markov cuya distribución estacionaria es la distribución a posteriori del modelo, existen muchos procedimientos para implementar estos métodos, uno de los más comunes es el Monte-Carlo Hamiltoneano que por su flexible implementación en el lenguaje Stan ha sido de utilidad en múltiples aplicaciones [10].

El mayor obstáculo al momento de realizar una análisis de datos adecuado en un enfoque Bayesiano, es que los procedimientos de estimación, diagnostico, y selección utilizados en *Box y Jenkins (1970)* no son validos en este nuevo enfoque. *Gelman, Vehtari et. al. (2020)* [11] proponen una extensa y robusta metodología denominada "*Bayesian workflow*", que presenta diferentes herramientas para un análisis de datos adecuado. Esta metodología se basa en la propuesta por *Box y Jenkins (1970)*, y se generaliza para cualquier tipo de modelamiento que involucre un enfoque de inferencia probabilístico.

Los dos principales problemas del método de *Gelman, Vehtari et. al. (2020)* al ser aplicados en el análisis de series temporales son su compleja estructura, y que algunas herramientas no son adecuadas para datos con supuestos de dependencia, por lo tanto, en este estudio presentamos una simplificación del *Bayesian Workflow* con ligeras variaciones en las herramientas presentadas para el uso adecuado en series temporales.

Por último, una vez establecido el método simplificado, se realizarán tres diferentes pruebas en tres conjuntos de datos que miden el IPC en Honduras de 1980 al 2018, la tasa de cambio de divisas entre Alemania y Reino Unido de 1984 a 1991 y la afluencia de turistas en Australia de 1995 al 2015, cada uno de estos conjuntos se encuentran en el paquete

bayesforecast [12], que se especializa en el análisis Bayesiano de series temporales. Finalmente con los resultados de dichas pruebas se demostrará la funcionalidad del nuevo método propuesto.

III. Preliminares y Notación

Para los objetivos de este estudio un proceso estocástico es una colección arbitraria de variables aleatorias $\{Y_1, Y_2, \ldots\}$, y una serie de tiempo es simplemente una realización o muestra finita $\{y_1, y_2, \ldots, y_n\}$ del proceso. Una propiedad importante a considerar en series temporales es la estacionaridad, diremos que un proceso $\{y_i\}$ con $i \in \mathbb{Z}$ es estacionario fuerte si:

$$F_X(x_{t_1}, x_{t_2}, ..., x_{t_n}) = F_X(x_{t_{1+\tau}}, x_{t_{2+\tau}}, ..., x_{t_{n+\tau}})$$

para $t \in \mathbb{Z}$ con $n \in \mathbb{N}$ y cualquier $\tau \in \mathbb{Z}$, esto es, si para cualquier colección finita del proceso su distribución conjunta se mantiene constante en el tiempo. Por otro lado diremos que el proceso $\{x_i\}$ tiene la propiedad de estacionaridad débil si:

$$\mu(t) = \mu \ \ , \ \ \sigma^2(t) = \sigma^2$$

esto es, que el proceso tiene una media y varianza constante a través del tiempo.

Diremos que una serie $\{y_t\}$ presenta estacionalidad si muestra alguna tendencia u oscilaciones periódicas constantes sobre la media del proceso, en este sentido para estabilizar la media de la serie y reducir la estacionalidad se aplica una transformación en los datos llamada diferenciación denotada por el operador diferencia:

$$\nabla^d y_t = y_t - y_{t-1}$$

Diremos que una serie es diferenciada si al aplicar el operador diferencia se vuelve estacionaria, de manera similar la serie es de diferencias estacionales si al aplicar la diferenciación estacional:

$$\nabla_s^D y_t = y_t - y_{t-s}$$

donde s es el periodo estacional, se vuelve estacionaria. También diremos que una serie de tiempo es un ruido blanco si es i.i.d. y normalmente distribuida con media cero y varianza constante.

III-A. Modelos ARIMA no estacionales

En el análisis de series de tiempo si combinamos los modelos autorregresivos, modelos de medias móviles y la diferenciación obtenemos los modelos ARIMA (Autoregressive Integrated Moving Average) no estacionales denotado por:

$$\text{ARIMA}(p, d, q)$$

en donde p, d y q son los órdenes del modelo autorregresivo, diferenciación y modelo de medias móviles respectivamente, de manera explícita se expresa como:

$$\nabla^d y_t = c + \sum_{i=1}^{p} \phi_i \nabla^d y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t$$

donde ε_t es ruido blanco.

III-B. Modelos ARIMA estacionales

Los modelos ARIMA pueden modelar datos con estacionalidad esto se puede lograr agregando parámetros estacionales al modelo y lo denotaremos como:

$$\text{SARIMA } (p, d, q) \times (P, D, Q)$$

SARIMA (Multiplicative seasonal autoregressive integrated moving average) de manera explícita:

$$Z_t = c + \sum_{i=1}^{p} \phi_i Z_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \sum_{k=1}^{P} \Phi_k Z_{t-km} + \sum_{w=1}^{Q} \Theta_w \varepsilon_{t-m} + \varepsilon_t$$

$$Z_t = \nabla_m^D \nabla^d y_t,$$

donde ε_t es un ruido blanco, los parámetros (p, d, q) y (P, D, Q) representan el orden de la parte no estacional y la parte estacional del modelo respectivamente y m es el periodo de oscilación de la media en los datos.

III-C. Modelos SARIMA Bayesianos

En base a la definición previa de un modelo ARIMA estacional podemos definir un Modelo SARIMA Bayesiano como:

$$\text{Modelo} \sim SARIMA(p, d, q)_{\times}(P, D, Q)_m$$

$$\phi_i \sim priori_{\phi i}, \quad i = 1, ..., p$$

$$\theta_j \sim priori_{\theta j}, \quad j = 1, ..., q$$

$$\Phi_k \sim priori_{\Phi k}, \quad k = 1, ..., P$$

$$\Theta_w \sim priori_{\Theta w}, \quad w = 1, ..., Q$$

$$\mu_0 \sim \text{priori}_{\mu_0}$$

$$\sigma_0 \sim \text{priori}_{\sigma_0}$$

Referencias

[1] R. J. Hyndman, *Box-Jenkins modelling*. Hans Daellenbach and Robert Flood, 2002, ch. Informed Student Guide to Management Science. [Online]. Available: https://robjhyndman.com/papers/BoxJenkins.pdf

[2] J. Durbin and S. Koopman, *Time Series analysis by state space methods*. Oxford University press, 2012, no. Second Edition.

[3] S. Taylor and B. Letham, "Forecasting at scale," *PeerJ Preprints 5:e3190v2*, 2017. [Online]. Available: https://doi.org/10.7287/peerj.preprints.3190v2

[4] O. Bogdan and C. Stefan, Cristian, "Time series forecasting using neural networks," *CoRR*, vol. abs/1401.1333, 2014. [Online]. Available: http://arxiv.org/abs/1401.1333

[5] U. Lotrič and A. Dobnikar, "Using smoothing splines in time series prediction with neural networks," in *Artificial Neural Nets and Genetic Algorithms*. Vienna: Springer Vienna, 1999, pp. 121–126.

[6] S. Roberts, M. Osborne, M. Ebden, S. Reece, N. Gibson, and S. Aigrain, "Gaussian processes for time-series modelling," *Phil. Trans. R. Soc. A.3712011055020110550*, 2013. [Online]. Available: http://doi.org/10.1098/rsta.2011.0550

[7] R. J. Hyndman, *ARIMA processes*. Hans Daellenbach and Robert Flood, 2002, ch. Informed Student Guide to Management Science. [Online]. Available: https://robjhyndman.com/papers/ARIMA.pdf

[8] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*, 2nd ed. OTexts Melbourne, Australia, 2018, ch. 8.9. [Online]. Available: https://otexts.com/fpp2/seasonal-arima.html

[9] J. S. Speagle, “A conceptual introduction to markov chain monte carlo methods,” 2020. [Online]. Available: https://arxiv.org/abs/1909.12313
[10] T. S. D. Team. Stan. [Online]. Available: https://mc-stan.org/
[11] A. Gelman, A. Vehtari, D. Simpson, C. C. Margossian, B. Carpenter, Y. Yao, L. Kennedy, J. Gabry, P.-C. Bürkner, and M. Modrák, “Bayesian workflow,” 2020. [Online]. Available: https://arxiv.org/abs/2011.01808
[12] I. A. Alonzo and C. Cruz, “bayesforecast: Bayesian time series modeling with Stan,” *ARXIV Preprint*, 2021. [Online]. Available: https://CRAN.R-project.org/package=bayesforecast