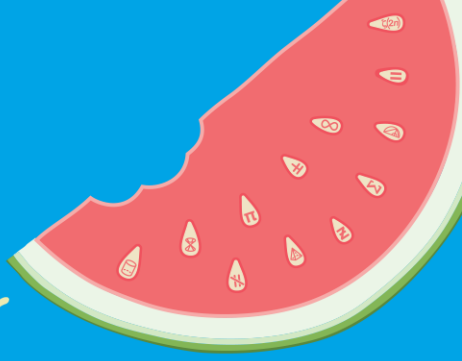
AMSI VACATIONRESEARCH SCHOLARSHIPS 2021–22

Get a taste for Research this Summer



Bayesian Estimation of Stationary Time Series Models with Exogenous Input

Mark Youssef

Supervised by **Dr Matias Quiroz**

University of Technology Sydney

AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Contents

Abstract

This project explores Bayesian estimation of stationary time series models for large data sets via Markov chain Monte Carlo methods. Processes with exogenous input are commonplace in real-world applications, although accommodation of these inputs can limit the effectiveness of model estimation. As such, an alternative approximate likelihood based on asymptotic independence of observations in the frequency domain was utilised for faster computation. The exogenous input in the model was accommodated via dynamic regression models with errors modelled as a time series process. The models considered were auto-regressive models with exogenous input, moving average models with exogenous input, and auto-regressive moving average models with exogenous input. The methods used ensured accurate estimation of the model parameters, while the choice of model allowed a simple interpretation of the effect of the exogenous input. The algorithm constructed in this project provides a strong framework that can be scaled to accommodate for more complex models found in real-world applications, as well as highly sophisticated subsets of Markov chain Monte Carlo methods.

1 Introduction

Bayesian methods have recently gained widespread use in Statistics and Data Science due to their natural probabilistic interpretations and ability to estimate complex models, as well as developing predictions that account for the uncertainty involved. Bayesian statistics express uncertainty in the parameters of a statistical model using the language of probability. The inferential object in Bayesian statistics is a posterior probability distribution of the parameters, obtained via Bayes' theorem, which combines both a priori information about the parameters, encoded as a probability distribution, and a probabilistic model for the data generating mechanism, known as the 'likelihood', which describes the plausibility of the observed data given a set of parameters. The posterior probability distribution corresponds to a known distribution in only a few toy model cases, and practitioners often need to resort to state-of-the-art Markov chain Monte Carlo (MCMC) simulation methods; in particular, the Metropolis-Hastings algorithm.

This project explores Bayesian estimation of stationary time series models for large data sets via Markov chain Monte Carlo methods. Accounting for the dependence of the observations through time typically results in an expensive likelihood for the model, especially for large data sets. A prominent solution utilised in this project is to transform the time series to the frequency domain, where an alternative approximate likelihood based on asymptotic independence of observations (in the frequency domain) can be formulated. The advantage of the so-called 'Whittle likelihood' (Whittle, 1953) is that it is computationally much faster than the time domain likelihood and thus enables inference for large data sets.

Practitioners are often interested in the effect of an exogenous variable on a system of endogenous variables. Such effects can be estimated using, for example, an ARMAX model for uni-variate time series, or a VARMAX model in the multivariate case. However, these models provide an unintuitive interpretation of the effect of the exogenous variable. To this end, a dynamic regression model was used as shown by Hyndman (2010) and used by Carter and Kohn (1997).

This project demonstrates the effectiveness of these methods when estimating auto-regressive models with exogenous input, moving average models with exogenous input, and auto-regressive moving average models with exogenous input.

1.1 Statement of Authorship

Dr Quiroz formulated the project idea and outline, as well as provided guidance and supervision. The algorithm was written in part by Mark Youssef and Dr Quiroz. Analysis and interpretation of results was performed by Youssef.

2 Example Application: Saginaw River Water Velocity

An exogenous variable is described as having an effect on a system, while not being affected by other parameter's within the system. Systems with exogenous variables are abundant in real-world practices. As such, the methods explored in this project could find use in a range of industries and data sets.

One example is the water velocity of the Saginaw River in Michigan measured at Bay City. The exogenous input in this case would be the water velocity measured further upstream at Saginaw. Because the river flows from Saginaw to Bay city, the velocity at the former will likely have a strong effect on the velocity at the latter. Figure 1 shows the water velocity at Saginaw, and figure 2 shows the water velocity at Bay City. It is visually apparent from both figures that there is a strong correlation between them. The methods explored in this project can be used to describe this effect numerically.

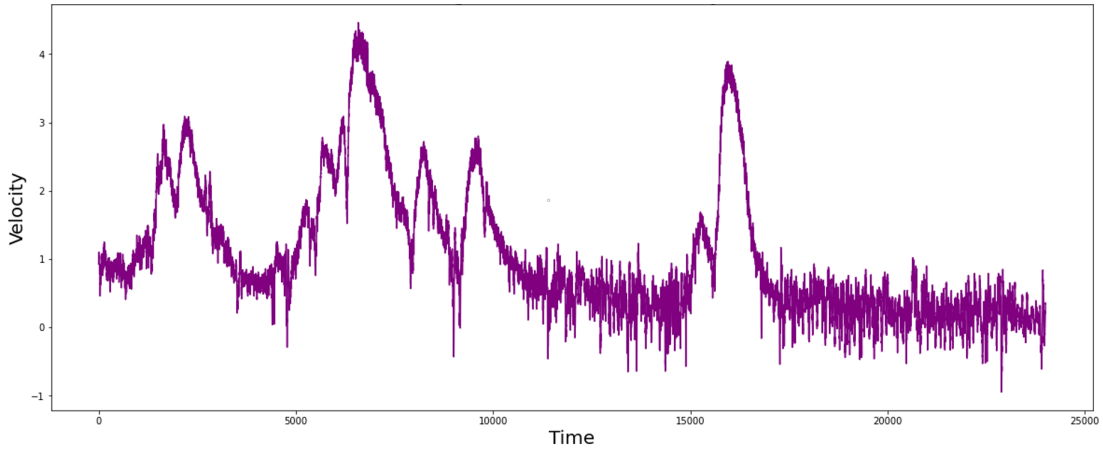


Figure 1: Saginaw Water Velocity

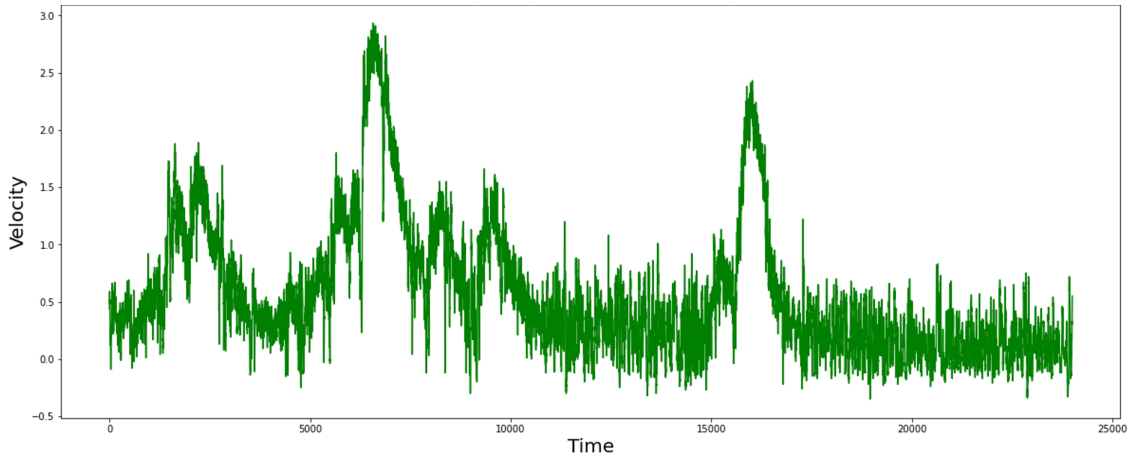


Figure 2: Bay City Water Velocity

3 Time Series Models

To define the dynamic regression models used in the project, the subsequent time series models enveloped within them must first be defined.

3.1 Auto-Regressive Moving Average Process

An auto-regressive time series is such that the value of the time series at time t is dependant on the value of the time series at previous time periods. Let $y_1 \dots y_n$ denote a zero-mean time series measured at n time points. An auto-regressive process with p lag terms $AR(p)$ is defined as

$$\begin{aligned} y_t &= \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + z_t \\ z_t &\sim N(0,1) \end{aligned}$$

A moving average process is such that the value of the process at time t is dependant on the value of the error term z at previous time periods. A moving average process with q lag terms $MA(q)$ is defined as

$$\begin{aligned} y_t &= z_t - \theta_1 z_{t-1} - \cdots - \theta_q z_{t-q} \\ z_t &\sim N(0,1) \end{aligned}$$

These models can be combined to form an auto-regressive moving average process $ARMA(p,q)$

$$\begin{aligned} y_t &= \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} - \theta_1 z_{t-1} - \cdots - \theta_q z_{t-q} + z_t \\ z_t &\sim N(0,1) \end{aligned}$$

3.2 Dynamic Regression Model

Let x_t denote the exogenous input and let β be its co-efficient. The dynamic regression model used in this project is defined below, along with the 3 different ways that η_t was defined

$$\begin{aligned} y_t &= \beta x_t + \eta_t \\ \eta_t &\sim AR(p) \\ \eta_t &\sim MA(q) \\ \eta_t &\sim ARMA(p,q) \end{aligned} \tag{1}$$

When using this model, we have a clear and intuitive interpretation of β. It represents the change in y given a one unit increase in x. It was due to this simple interpretation that the model was chosen for this project.

4 Spectral Information

4.1 Spectral Density

To obtain the spectral density, we must first derive the auto-co-variance function. The auto-covariance function of a zero-mean stationary process $y_t \in \mathbb{R}$ is defined as:

$$\gamma_\theta(\tau) = \mathbb{E}[y_t y_{t-\tau}]$$

$$\theta = \text{Model Parameters},\ \tau = 0, 1, \ldots$$

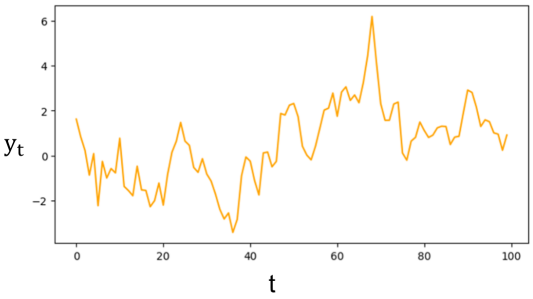
The spectral density is then the Fourier transform of the auto-covariance function, shown as

$$f_\theta(\omega) = \sum_{\tau=-\infty}^{\infty} \gamma_\theta(\tau)\exp(-i\omega\tau)$$

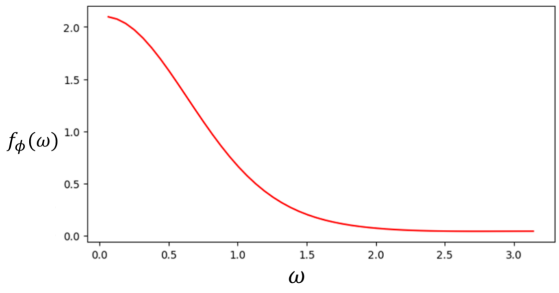
$$\omega \in (-\pi, \pi]$$

Shown in figure 3 is an AR(1) process with a positive coefficient alongside it's corresponding spectral density. Notice in 3a that the process is slow-moving. Consequently, 3b shows a spectral density composed of mostly low frequencies.

Figure 4, however, shows an AR(1) process with a negative co-efficient. Notice the process in 4a moves quickly and erratically. As such, this process has a spectral density composed of mostly high frequencies, shown in 4b.

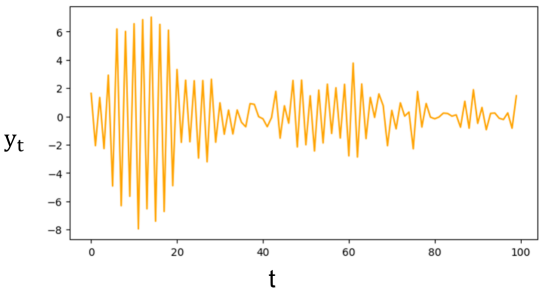


(a) AR(1) Process

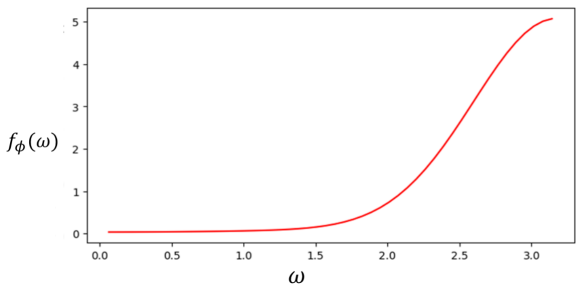


(b) Spectral Density

Figure 3: AR(1) Process, $\phi = 0.9$



(a) AR(1) Process



(b) Spectral Density

Figure 4: AR(1) Process, $\phi = -0.9$

4.2 Periodogram

To calculate the likelihood function given a set of parameters using spectral methods, the periodogram of the data is constructed. First, the discrete Fourier transformation (DFT) of the time series y_t is taken, shown in equation 2. Note that this can be computed in $O(n\ \log(n))$ with the fast Fourier transformation.

$$\begin{aligned} J(\omega_k) &\equiv \frac{1}{2\pi}\sum_{t=1}^{n} y_t\ \exp(-i\omega_k t) \\ \omega_k \in \Omega &= \{2\pi k n^{-1} \text{for } k = -\lceil \frac{n}{2} \rceil + 1, \ldots, \lfloor \frac{n}{2} \rfloor\} \\ n &= \text{ number of observations} \end{aligned} \tag{2}$$

The Periodogram is then the subsequent transformation of the DFT given in equation 3, where $I(\omega_k)$ represents data observation k in the frequency domain.

$$I(\omega_k) = n^{-1}|J(\omega_k)|^2 \tag{3}$$

5 Whittle Log-Likelihood

The Whittle Log-Likelihood (Whittle (1953)) is the likelihood function used to allow for fast computation. The key result is that as $n \to \infty$, the periodogram observations are independent and exponentially distributed with mean equal to the spectral density evaluated at their respective frequencies. This feature is demonstrated in figure 5, where 3 arbitrary frequencies of the periodogram have their exponential distribution plotted on the 'density' axis.

$$\text{As } n \to \infty,\ I(\omega_k) \sim^{ind} \text{Exp}(f(\omega_k)$$

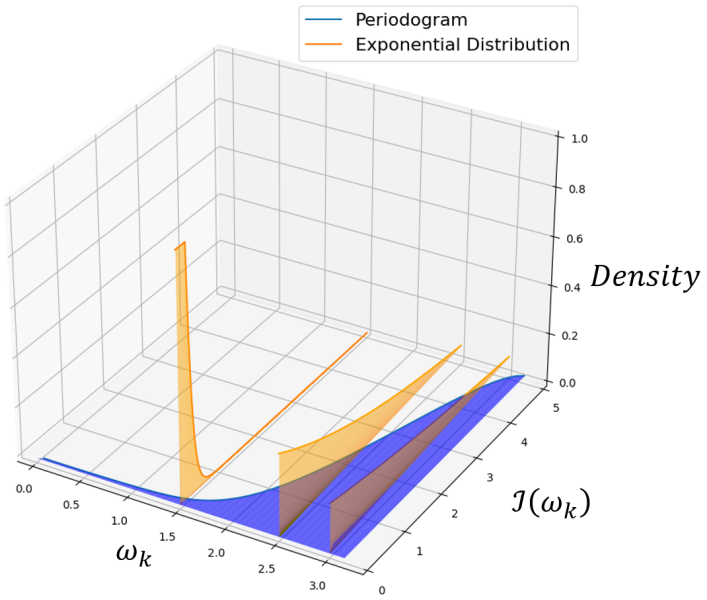


Figure 5: Example: Distribution of periodogram observations

The Whittle likelihood function of $I(\omega_k)|\theta$ is therefore

$$\prod_{\omega_k \in \Omega} p(I(\omega_k)|\theta) = \prod_{\omega_k \in \Omega} f_\theta(\omega_k)^{-1} \exp(-\frac{I(\omega_k)}{f_\theta(\omega_k)})$$

The log of the likelihood function is then taken, converting the function into a sum due to asymptotic independence, allowing for fast computation. The Whittle log-likelihood $l_W(\theta)$ is therefore

$$l_W(\theta) \equiv -\sum_{\omega_k \in \Omega} (\log(f_\theta(\omega_k)) + \frac{I(\omega_k)}{f_\theta(\omega_k)})$$

5.1 Re-arranging the Equation

When computing the periodogram and spectral density of the data based on the model in equation 1, the equation is first re-arranged. Rather than computing with respect to y_t, it is done with respect to $y_t - \beta x_t$. As shown in equation 4, this ensures the spectral information is equal to that of an $ARMA(p, q)$ process. Also note in equation 4 the dependence of the periodogram on the value of β. This means that the periodogram will need to be recomputed at each new value of β proposed.

$$\begin{aligned} y_t - \beta x_t &= \eta_t \\ \eta_t &\sim ARMA(p, q) \end{aligned} \tag{4}$$

6 Priors and Stationarity

For the spectral density and corresponding likelihood function to be accurate, η from equation 4 must have forced stationarity. Values of ϕ and θ are proposed in the partial auto correlation space, and then reparamaterised to the ordinary paramatrisation for computation of the posterior distribution.

Using this method, the process is certain to be stationary if all values of ϕ and θ are between -1 and 1 (Barndorff-Nielsen and Schou (1973)). A uniform prior between these values will therfore enforce stationarity.

σ^2, representing the variance of ϵ, must be a positive value. As such, values of σ^2 are proposed in the log scale to ensure that after performing the inverse exponential transformation, they will always be positive. A standard normal prior is therefore appropriate. We also use a standard normal prior for β, as alternate values of the variance of $p(\beta)$ were found to have an insignificant effect on estimation.

7 Markov Chain Monte Carlo

The algorithm used is an example of a Markov Chain Monte Carlo (MCMC) algorithm, a popular and powerful class of algorithms used for Bayesian inference.

7.1 Monte Carlo Methods

A Monte Carlo method can be described as one that uses repeated random sampling to obtain numerical results. For example, Monte Carlo methods can be used to obtain the Expected Value of a function. As shown in equation 5, by summing N draws of $h(\theta)$, then dividing by N, by the law of large numbers, this converges almost surely to the $\mathbb{E}[h(\theta)]$. The relevance to the project is that repeated random sampling is used to approximate the posterior distribution.

$$\frac{1}{N}\sum_{i=1}^{N} h(\theta^{(i)}) \xrightarrow{\text{a.s}} \mathbb{E}[h(\theta)] \tag{5}$$

7.2 Markovian Property

Let the collection of random variables $\{\theta^{(t)}\}_{t\geq 0}$ be a process indexed a period t. The process is 'Markovian' if the following condition is met.

$$\Pr(\theta^{(t)} = \phi^{(t)} | \theta^{(t-1)} = \phi^{(t-1)}, \dots \theta^{(1)} = \phi^{(1)}) = \Pr(\theta^{(t)} = \phi^{(t)} | \theta^{(t-1)} = \phi^{(t-1)})$$

where $\phi^{(t)}$ denotes the state of the process at period t

An intuitive interpretation of this condition is that the value of the process at period t only depends on the value of the process at period $t-1$. A sequence generated by a Markov process is called a Markov chain.

7.3 Metropolis MCMC Algorithm

The type of MCMC algorithm employed in this project is the Metropolis MCMC algorithm. The algorithm in general works as follows.

Let θ_c denote the current state of a Markov chain

- Choose arbitrary start point $\theta_c = \theta^{(0)}$

For N = chosen number of iterations, repeat:

- Propose draw $\theta_p = q(\theta|\theta_c)$, where q is the proposal distribution
- Compute acceptance probability α as

$\alpha(\theta_c, \theta_p) = \min(1, \frac{\pi(\theta_p)}{\pi(\theta_c)})$, $\quad \pi(\theta) = p(\theta|y)$

- Sample $u \sim \text{Uniform}(0, 1)$
- If $u < \alpha(\theta_c, \theta_p) \rightarrow \theta^{(i)} = \theta_p$, else, $\theta^{(i)} = \theta_c$
- Set $\theta_c = \theta^{(i)}$

End For

- Discard an appropriate proportion of initial draws

What this algorithm means intuitively is that if the proposed position is more likely than the current position, the draw is always accepted. If it is less likely, the draw is accepted with probability α. After the chain has reached N iterations, an appropriate proportion of initial draws are discarded, or 'burned'. This is to allow the chain to converge to sampling from the true posterior distribution, as this will take time depending on the efficiency of the chain and the accuracy of the starting values. The decision of an appropriate number of samples to burn can be made by inspecting a plot of the cumulative mean of the likelihood function. When this appears to stabilise to a constant value, previous draws should be burned and the remainder will be samples from the true posterior distribution.

8 Algorithm Implementation

The Metropolis MCMC algorithm was combined with the spectral methods described above to construct the following algorithm that was implemented in the project. The variance of the proposal distribution is scaled by $\tilde{c} = 2.38 \cdot (\text{Number of paramaters})^{-1/2}$, as this is shown to obtain an optimal acceptance rate of approximately 2.38 (Roberts and Gilks (1997)).

- The posterior distribution is first optimised to obtain the *Maximum a posteriori*(MAP), the mode of the posterior distribution
- The posterior covariance is calculated as the inverse of the negative Hessian matrix, evaluated at the MAP

  For $N = 10,000$:
- θ_p vector of parameters is proposed in the reparamatrised space. The proposal distribution used is $q \sim N(MAP\ ,\ \tilde{c}\ (\text{Posterior covariance})\ )$
- The parameters are transformed to the ordinary parametrisation to compute the periodogram and spectral density of $y_t - \beta x_t$ given θ_p and given $\theta^{(c)}$
- The Whittle log-likelihood of θ_p and $\theta^{(c)}$ is computed
- The prior probabilities are computed in the log scale and summed to their respective likelihoods
- Acceptance probability α is computed and draw is accepted or rejected as per Metropolis MCMC

  End For
- The cumulative mean of the likelihood is inspected over the N iterations and an appropriate number of samples are burned
- The accepted proposals are transformed to the ordinary paramaterisation to obtain samples from the posterior distribution

9 Results

To test the algorithm, data was randomly generated for 4 dynamic regression models with differing distributions of the errors and differing number of parameters. The number of MCMC iterations N was set to $10,000$ for each test. The number of data points n generated for each model was 14,401. The exogenous input $x_1, \dots x_n$ was generated as indpendant draws from a $N(0,1)$ distribution, otherwise known as a white-noise process. The 4 models tested were

- $y_t = \beta x_t + \eta_t \quad , \qquad \eta_t \sim AR(2)$
- $y_t = \beta x_t + \eta_t \quad , \qquad \eta_t \sim MA(2)$
- $y_t = \beta x_t + \eta_t \quad , \qquad \eta_t \sim ARMA(1,1)$
- $y_t = \beta x_t + \eta_t \quad , \qquad \eta_t \sim ARMA(3,1)$

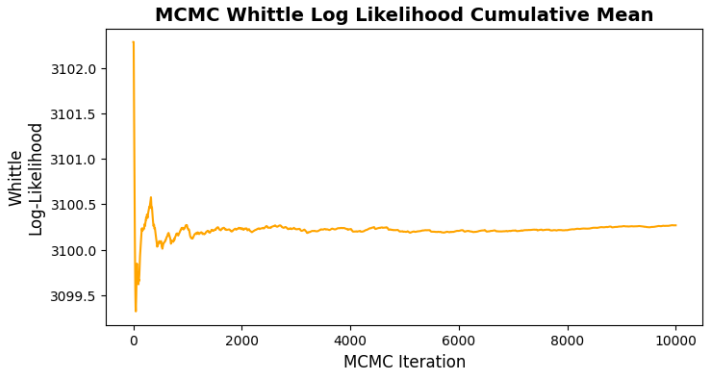


Figure 6: Cumulative mean for $\eta_t \sim AR(2)$

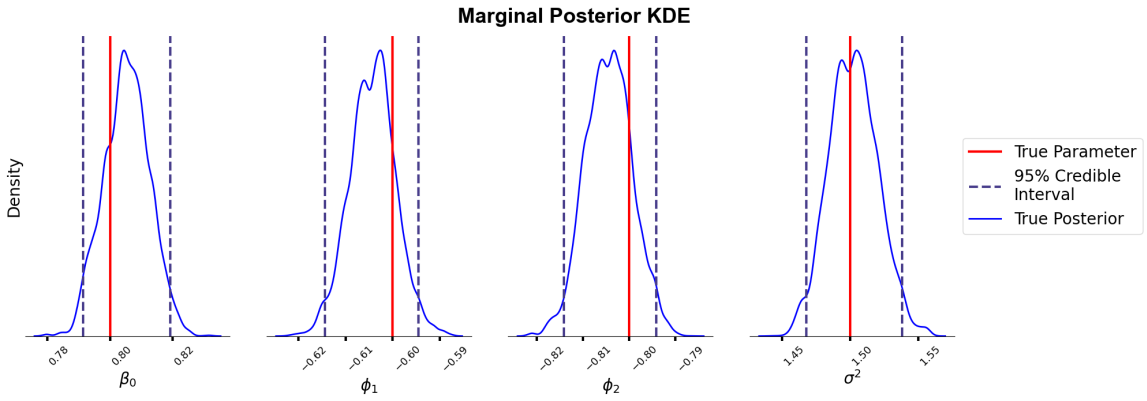


Figure 7: Marginal Posterior Distributions for $\eta_t \sim AR(2)$

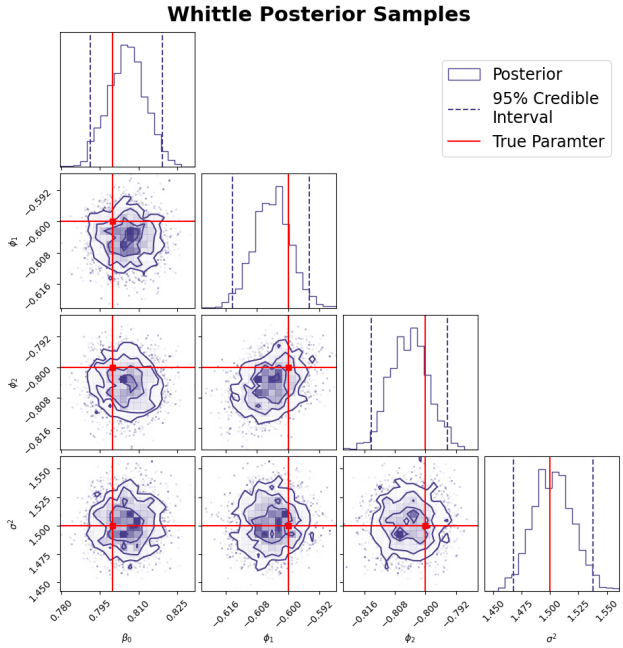


Figure 8: Bivariate Marginal Posterior Distributions for $\eta_t \sim AR(2)$

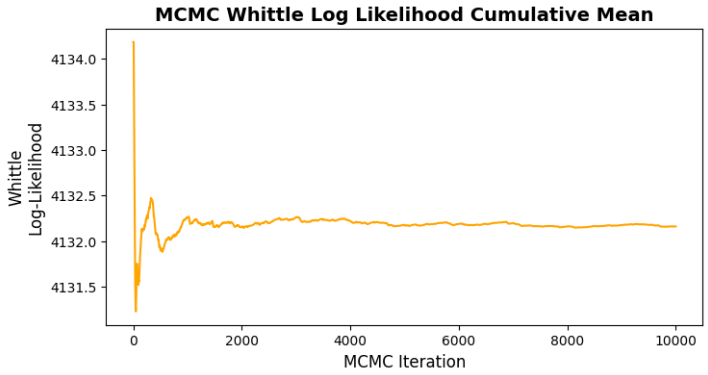


Figure 9: Cumulative mean for $\eta_t \sim MA(2)$

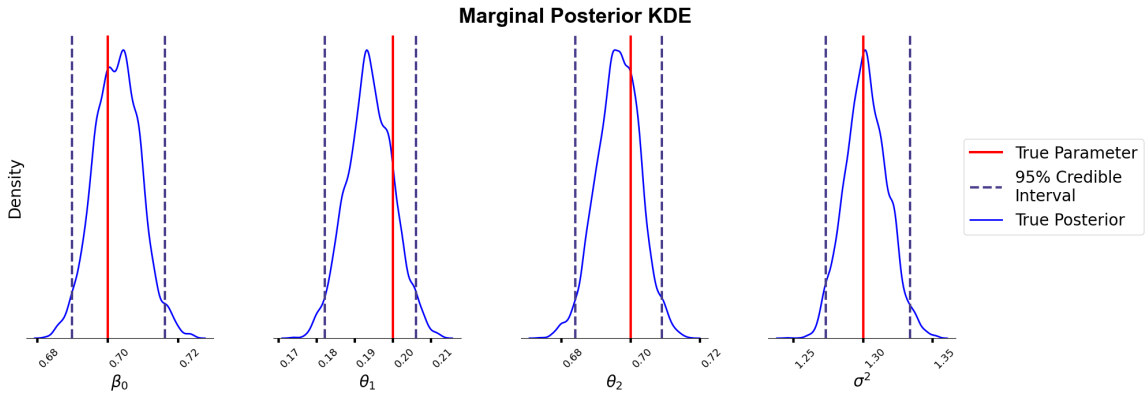


Figure 10: Marginal Posterior Distributions for $\eta_t \sim MA(2)$

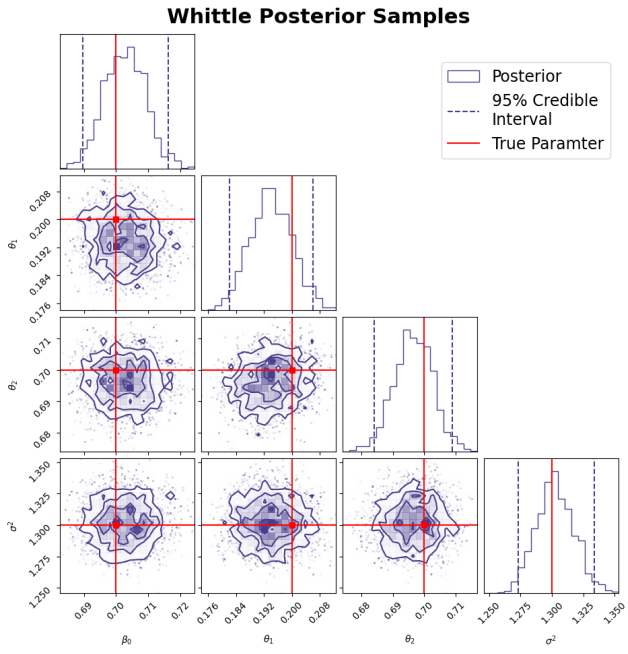


Figure 11: Bivariate Marginal Posterior Distributions for $\eta_t \sim MA(2)$

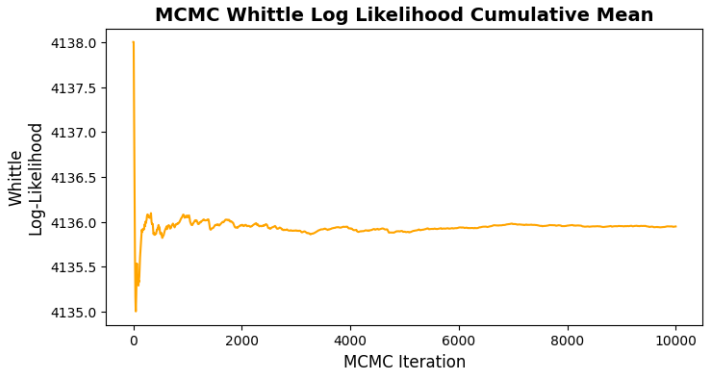


Figure 12: Cumulative mean for $\eta_t \sim ARMA(1,1)$

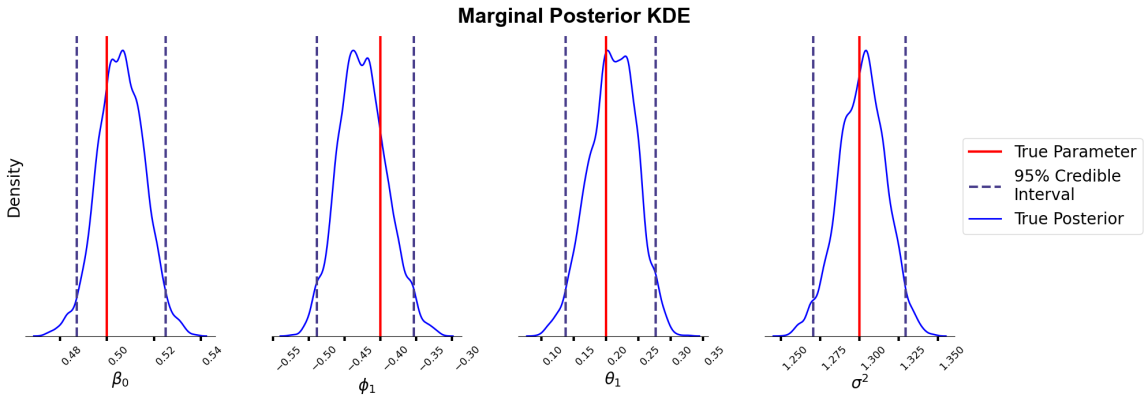


Figure 13: Marginal Posterior Distributions for $\eta_t \sim ARMA(1,1)$

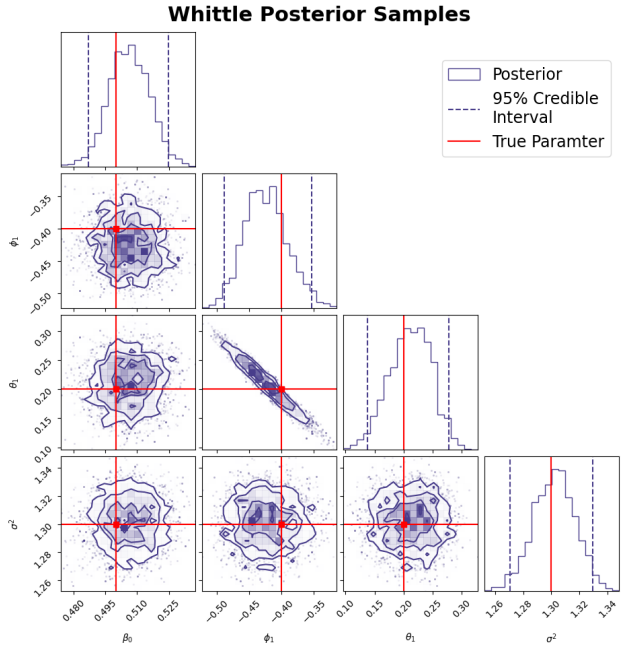


Figure 14: Bivariate Marginal Posterior Distributions for $\eta_t \sim ARMA(1,1)$

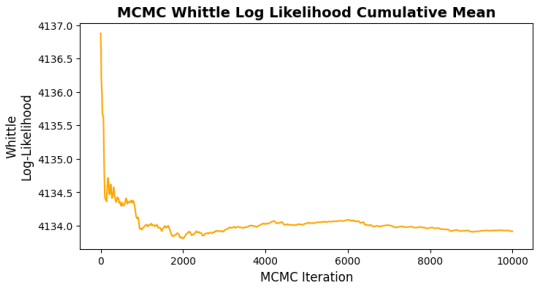


Figure 15: Cumulative mean for $\eta_t \sim ARMA(3,1)$

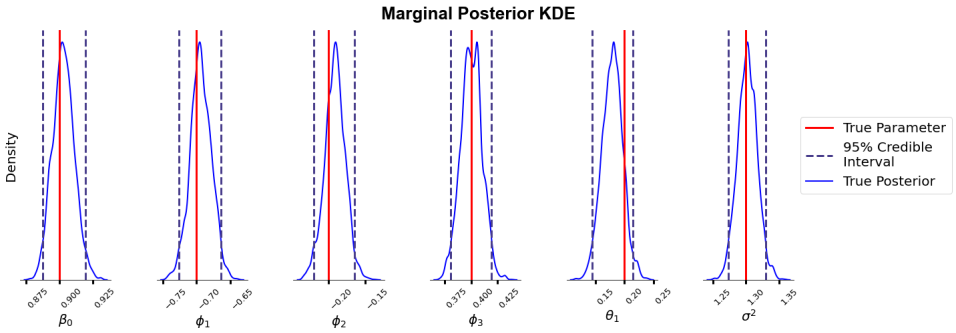


Figure 16: Marginal Posterior Distributions for $\eta_t \sim ARMA(3,1)$

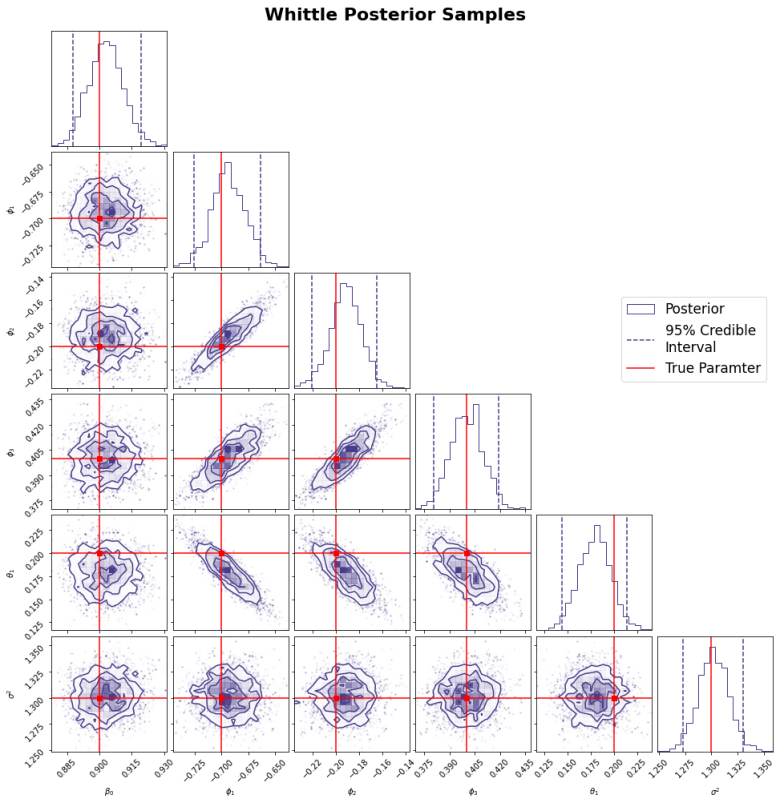


Figure 17: Bivariate Marginal Posterior Distributions for $\eta_t \sim ARMA(3,1)$

10 Conclusion and Future Research

As shown in figures 6 to 17, the parameters of the models were accurately estimated in all cases. The values of the true model parameters all fell well within the 95% credible intervals. While an approximation of the likelihood function was made to obtain independent observations, accurate results were still obtained. However, as shown in equation 4, the dependency of the periodogram on the value of β means it must be re-computed at every iteration, increasing the computational cost of the algorithm.

Future research in this area would be to employ state of the art 'subsampling MCMC' methods (Quiroz, Kohn, Villani, and Tran (2019)) to estimate the likelihood function using a sub-sample of the data. This would ensure faster computation of the periodogram, mitigating the issue of dependence on the model parameters.

When limiting the distribution of η_t to an $ARMA(p, q)$ process, the ability of the model to capture long-term dependencies in real world data is negatively effected. A future extension of this research would be to expand the range of processes that η_t could be modelled by. This would increase the applicability of the algorithm to many more data sets. Combining these models with sub-sampling methods would provide a faster and highly accurate algorithm to estimate parameters for a wider range of processes.

References

Barndorff-Nielsen, O. and G. Schou. 1973. "On the parametrization of autoregressive models by partial autocorrelation." *Journal of Multivariate Analysis* 3 (4):408–419.

Carter, Christopher K and Robert Kohn. 1997. "Semiparametric Bayesian inference for time series with mixed spectra." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59 (1):255–268.

Hyndman, Rob J. 2010. "The ARIMAX model muddle." URL `https://robjhyndman.com/hyndsight/arimax`.

Quiroz, Matias, Robert Kohn, Mattias Villani, and Minh-Ngoc Tran. 2019. "Speeding up MCMC by efficient data subsampling." *Journal of the American Statistical Association* 114 (526):831–843.

Roberts, Gelman A., G. O. and W. R. Gilks. 1997. "Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms." *The Annals of Applied Probability* 7:110–120.

Whittle, Peter. 1953. "Estimation and information in stationary time series." *Arkiv för matematik* 2 (5):423–434.