AMSI SUMMER RESEARCH SCHOLARSHIPS 2025–26

Get a taste for Research this Summer

Frequency-Domain Model Selection for Large-Scale Time Series via Bayesian Variable Selection

Jamie Hill

Supervised by Dr. Matias Quiroz & Dr. Aishwarya Bhaskaran

University of Technology Sydney & University of New South Wales

Jane Street®

AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Contents

Abstract

This project develops an efficient algorithm for simultaneous variable selection and regression coefficient estimation for large-scale time series data. Bayesian inference for large-scale time series is computationally expensive. To reduce this cost, data are transformed from the time domain into the frequency domain using the Fourier transform and the Whittle likelihood (Whittle, 1953) is employed in place of the exact time-domain likelihood. This approach is computationally more efficient for large-scale time series analysis as the frequency-domain representation exploits conjugate symmetry, eliminating the redundant negative-frequency components and making the frequency-domain representation after elimination smaller in size than its time-domain counterpart. The Whittle likelihood is a more efficient method for likelihood evaluation than the exact time-domain likelihood through asymptotic independence reducing the number of operations necessary. Decreasing run-time considerably.

1 Introduction

This report will give a brief overview of time series analysis, the roadblocks that can be encountered for large-scale time series and the solutions.

The estimation of time series models is well established in the literature. The issue is that accounting for temporal dependence is typically a computationally expensive endeavour in the time domain when working with larger datasets. Considering the computational cost in the time domain, transforming the dataset into another domain is a possible solution. Therefore, this report will explore the analysis of such time series models in the frequency domain. The exact likelihood evaluation is defined in the time domain, therefore the algorithm will utilise the Whittle likelihood (Whittle, 1953) instead. The Whittle likelihood is an approximation of the regular Gaussian likelihood and a much more efficient evaluation method allowing valuable inference for large datasets while reducing computational cost.

A common approach for model selection is significance testing. This method contains several key flaws like non-parsimony and inference degradation in high-dimensional spaces. These flaws are mitigated by alternative methods. Information criteria (Akaike, 1974; Schwarz, 1978) methods provide formal criteria for comparing competing models, directly addressing the parsimony problem. The main flaw that exists for this approach is that it requires multiple runs to determine which models are the most plausible or have the most predictive power, making this method inefficient in moderately large dimensions.

This report will explain how using Bayesian variable selection, specifically the Metropolis–Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) is a much more computationally efficient method. It jointly proposes possible models ($\boldsymbol{\mathcal{I}}$) along with their corresponding estimates ($\boldsymbol{\beta}$). This allows us to build the posterior distribution of the data and draw valuable inference in a single run rather than multiple for previously stated methods. The report will explicitly focus on AR(p), MA(q), and ARMA(p,q) models as the formation of the Whittle likelihood (Whittle, 1953) requires the spectral density which requires a model-specific formula. This does not preclude extensions beyond ARMA(p,q) models, however, for clarity and simplicity, this report

does not venture further than ARMA(p,q) models.

This report will further explore the basics of time series data, a useful category of data for fields such as economics, biology, finance etc. Secondly, it will move on to time series data structures (models) and how inference is drawn from such models. Furthermore, we will provide a brief description of Bayesian variable selection and its consequent sampling methods, (Direct sampling, Gibbs sampling, Metropolis–Hastings sampling) (O'Hara and Sillanpää, 2009). After this, the frequency domain will be introduced along with the Whittle likelihood. Finally, an explanation of the algorithm and results will be produced.

2 Time Series Data

The defining feature of a time series dataset is that it is ordered in time, the order itself being important for analysis. It must be noted that the day, month, year or any time sequencing is known to exist in the time domain. The analysis of time series data often relies on the assumption of covariance stationarity due to many estimation and evaluation methods requiring such an assumption. Covariance stationarity refers to the property of a time series dataset that has a time-invariant mean, variance, and autocovariance. There are several methods to impose the assumption onto a process, although this report will not discuss such methods.

3 Time Series Models

There are a multitude of model structures that can represent time series data and can be exploited for prediction and inference. To maintain simplicity and relevance, this report will only cover the three models that the algorithm can support. These structures include autoregressive (AR) models, moving average (MA) models, and autoregressive moving average (ARMA) models.

3.1 Autoregressive Models

The autoregressive structure or process of order p is denoted AR(p) and is simple in essence. It can be understood directly from its name which in plain English means that the process' past entries directly have some effect on the future entries e.g., for an autoregressive process of order 1 (AR(1)), where the time points are sequenced in days, yesterday has some effect on today, but the day before yesterday has no effect on today. The process is formulated below:

$$y_t = c \ + \ \phi_1 y_{t-1} \ + \ \phi_2 y_{t-2} \ + \cdots + \ \phi_p y_{t-p} \ + \ \varepsilon_t, \quad \varepsilon_t \ \sim \mathcal{N}(\mu, \sigma^2).$$

A verbal illustration of autoregressive processes can assist in building intuition. The data's most recent observation will be influenced by the preceding entry. Say t is today, then t (today) has some dependence on $t-1$ (yesterday). For an AR(2) process, t can be affected by both $t-1$ (yesterday) and $t-2$ (the day before yesterday), although it is only necessary for $t-2$ (the day before yesterday) to have an effect on t for it to be classed as an AR(2) process.

3.2 Moving Average Models

A moving average process of order q is denoted by MA(q) and has a very similar structure to the autoregressive process. The distinction between the two is that the most recent observation is not driven by the preceding value, but the preceding error term.

$$y_t = c + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q}, \quad \varepsilon_t \sim \mathcal{N}(\mu, \sigma^2),$$

where ε_t is a regular normally distributed error term generally understood as a Gaussian white noise process.

$$\varepsilon_t \sim \mathcal{N}(\mu, \sigma^2).$$

For intuition the moving average process can be understood as follows: yesterday's error term influences today's final value. E.g., Say the stock price yesterday was \$10. Then a random shock occurs and it drops to \$5, the halving of the stock price can cause investors to be worried about investments, and further have an effect on the next day as economic agents then pull their investments from the market.

3.3 Autoregressive Moving Average Models

An autoregressive moving average (ARMA) model combines both previously mentioned processes. Its formulation goes as follows:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q}, \varepsilon_t \sim \mathcal{N}(\mu, \sigma^2).$$

Essentially, for an ARMA(1,1) process we would see that both the preceding value and error term have some effect on today's value. This is a much more flexible and realistic process rather than the isolation of the components. With the understanding of both the AR and MA notation, it is clear to see that the combination is written as an ARMA(p,q) process. The mixing of these two processes also begins to increase the number of parameters for the process and increases potential model complexity for the regular significance testing method, generally leading to the use of more sophisticated methods like information criteria (Akaike, 1974; Schwarz, 1978) or Bayesian Variable Selection (O'Hara and Sillanpää, 2009).

4 Bayesian Variable Selection

Bayesian Variable Selection requires both additional notation and understanding of the spike-and-slab formulation which will be introduced here and used throughout the remainder of the report.

- $\boldsymbol{\mathcal{I}}$: The inclusion vector. This is a binary vector where $\mathcal{I}_j = 1$ means predictor j is included (IN) in the model and $\mathcal{I}_j = 0$ means predictor j will be excluded (OUT) from the model.
- $\boldsymbol{\beta}$: The beta vector. This vector works alongside the $\boldsymbol{\mathcal{I}}$ vector. when $\mathcal{I}_j = 1$, β_j will be the estimate or effect on the dependent variable y_t. When $\mathcal{I}_j = 0$, the prior distribution of β_j is a point mass at 0.

Assuming σ^2 known, the vector form linear regression model with variable selection can be written as follows. This spike-and-slab formulation is presented in the standard linear regression form for clarity and can be extended to the frequency domain by replacing the exact Gaussian time-domain likelihood with the approximate Whittle likelihood

$$y_t = \boldsymbol{x}_t^\top \boldsymbol{\beta} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n),$$

$$\boldsymbol{\beta} \mid \boldsymbol{\mathcal{I}}, \sigma^2 \sim \mathcal{N}(\mathbf{0}_{\boldsymbol{\mathcal{I}}} \Sigma_{0,\boldsymbol{\mathcal{I}}}), \quad \Sigma_{\mathbf{0},\boldsymbol{\mathcal{I}}} = \sigma^2 \Omega_{\mathbf{0},\boldsymbol{\mathcal{I}}}^{-1}$$

$$\boldsymbol{\beta} \mid \boldsymbol{\mathcal{I}}^c \sim \delta_{\mathbf{0}_{\boldsymbol{\mathcal{I}}^c}}$$

$$\mathcal{I}_j \overset{\text{ind}}{\sim} \text{Bernoulli}(\pi_j), \quad \pi_j = \Pr(\mathcal{I}_j = 1).$$

Let $\mathcal{I} \subset \{1, \ldots, p\}$ denote the set of included predictors and let $\mathcal{I}^c$ denote its complement. $\delta_{\mathbf{0}_{\boldsymbol{\mathcal{I}}^c}}$ denotes a generalised point density at zero.

$$p(\boldsymbol{\beta}_{\mathcal{I}^c} \mid \boldsymbol{\mathcal{I}}) = \delta_{\mathbf{0}_{\mathcal{I}^c}}(\boldsymbol{\beta}_{\mathcal{I}^c}) = \begin{cases} \infty, & \boldsymbol{\beta}_{\mathcal{I}^c} = \mathbf{0}_{\mathcal{I}^c} \\ 0, & \boldsymbol{\beta}_{\mathcal{I}^c} \neq \mathbf{0}_{\mathcal{I}^c}, \end{cases}$$

with the property

$$\int \delta_{0_{\mathcal{I}^c}}(\boldsymbol{\beta}_{\mathcal{I}^c})\, d\boldsymbol{\beta}_{\mathcal{I}^c} = 1,$$

and the marginal prior distribution of $\boldsymbol{\beta}$ being

$$p(\boldsymbol{\beta}) = \sum_{\boldsymbol{\mathcal{I}}} p(\boldsymbol{\beta} \mid \boldsymbol{\mathcal{I}}) p(\boldsymbol{\mathcal{I}}).$$

$$p(\boldsymbol{\beta} \mid \boldsymbol{\mathcal{I}}) = p(\boldsymbol{\beta}_{\mathcal{I}}) \; + \delta_{0_{\mathcal{I}^c}}(\boldsymbol{\beta}_{\mathcal{I}^c}).$$

4.1 Direct Sampling

Direct sampling is a brute force sampling method which directly samples all possible outcomes of $\boldsymbol{\mathcal{I}}$ (as discussed previously, a binary vector which decides whether parameters are included in the model or excluded from the model) to determine the most probable inclusion set. This is done via sampling from the marginal posterior distribution of the inclusion vector:

$$p(\boldsymbol{\mathcal{I}} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{\mathcal{I}})\, p(\boldsymbol{\mathcal{I}}),$$

$$\text{where} \qquad p(\boldsymbol{y} \mid \boldsymbol{\mathcal{I}}) = \int p(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{\mathcal{I}})\, p(\boldsymbol{\beta} \mid \boldsymbol{\mathcal{I}})\, d\boldsymbol{\beta}$$

The algorithm is an exact method for finding the most probable model, although it becomes intractable for moderately large dimensions. For p parameters there will be 2^p possible models. The number of possible models to sample when $p = 3$ would be 8 models, and not a problem for the algorithm. When dealing with higher dimensions, say $p = 20$, there will then be $1,048,576$ models to be analysed, a computationally expensive ordeal, therefore suggesting a more sophisticated algorithm for posterior exploration.

4.2 Gibbs Sampling

The Gibbs sampler is a smarter approach when encountering larger dimensions. That being said, it is not necessarily a *fast* algorithm, but is considerably more efficient than direct enumeration. The Gibbs sampler rather than brute forcing the models of $\boldsymbol{\mathcal{I}}$, updates the components conditionally. Iteratively sampling from the full conditional distribution, allowing efficient exploration of the model space $p(\boldsymbol{\mathcal{I}} \mid y)$. The full conditional posterior can be seen below:

$$p(\mathcal{I}_j \mid \boldsymbol{\mathcal{I}}, \boldsymbol{y}, \sigma^2) \propto p(\boldsymbol{y} \mid \boldsymbol{\mathcal{I}}, \sigma^2) p(\boldsymbol{\mathcal{I}}).$$

4.3 Metropolis–Hastings

Thus far, only sampling the $\boldsymbol{\mathcal{I}}$ vector has been discussed. The Metropolis–Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) jointly samples both the $\boldsymbol{\beta}$ and $\boldsymbol{\mathcal{I}}$ vectors. This allowing for exploration of the posterior distributions of both $\boldsymbol{\beta}$ and $\boldsymbol{\mathcal{I}}$.

The Metropolis–Hastings algorithm draws samples from the posterior distribution:

$$p(\boldsymbol{\beta}, \boldsymbol{\mathcal{I}} \mid \boldsymbol{y}),$$

through the joint proposal distribution:

$$q(\boldsymbol{\beta}' \mid \boldsymbol{\mathcal{I}}') q(\boldsymbol{\mathcal{I}}' \mid \boldsymbol{\mathcal{I}}^{(m-1)}),$$

It is clear that the proposal distribution is broken into two parts. To start, we will discuss the $\boldsymbol{\mathcal{I}}$ proposal.

4.3.1 The $\boldsymbol{\mathcal{I}}$ Proposal

$$q(\boldsymbol{\mathcal{I}}' \mid \boldsymbol{\mathcal{I}}^{(m-1)}).$$

This proposal distribution proposes an independent Bernoulli flip from 0 to 1, or from 1 to 0 with some predetermined probability.

4.3.2 The $\boldsymbol{\beta}$ Proposal

$$q(\boldsymbol{\beta}' \mid \boldsymbol{\mathcal{I}}').$$

The proposal for $\boldsymbol{\beta}$ is certainly much more complex than the $\boldsymbol{\mathcal{I}}$ proposal. This proposal draws samples from a multivariate normal distribution which is centered around the MAP, given all the parameters are included in the model. Thinking about this in simple terms, if we were to include every parameter we have in our data, it would be the most probable model that could exist. This is then solved through a maximisation problem. The covariance will then follow to be the inverse Hessian of the MAP multiplied by some constant c, where $c \geq 1$. For concreteness, the distribution goes as follows:

$$q(\boldsymbol{\beta}) \sim \mathcal{N}\left(\hat{\boldsymbol{\beta}}_{\mathcal{I}},\ (c\,\mathcal{H}_{\mathcal{I}})^{-1}\right).$$

Proposing values then brings us to the acceptance condition. The Metropolis–Hastings accepts or rejects values based on the equation:

$$\alpha_{\text{MH}} = \min(1, a), \qquad \text{where}$$

$$a = \frac{p(\boldsymbol{y} \mid \boldsymbol{\beta}', \boldsymbol{\mathcal{I}}')\, p(\boldsymbol{\beta}' \mid \boldsymbol{\mathcal{I}}')\, p(\boldsymbol{\mathcal{I}}')\, / q(\boldsymbol{\beta}^{(m-1)} \mid \boldsymbol{\mathcal{I}}^{(m-1)})\, q(\boldsymbol{\mathcal{I}}^{(m-1)} \mid \boldsymbol{\mathcal{I}}')}{p(\boldsymbol{y} \mid \boldsymbol{\beta}^{(m-1)}, \boldsymbol{\mathcal{I}}^{(m-1)})\, p(\boldsymbol{\beta}^{(m-1)} \mid \boldsymbol{\mathcal{I}}^{(m-1)})\, p(\boldsymbol{\mathcal{I}}^{(m-1)})\, / q(\boldsymbol{\beta}' \mid \boldsymbol{\mathcal{I}}')\, q(\boldsymbol{\mathcal{I}}' \mid \boldsymbol{\mathcal{I}}^{(m-1)})}.$$

The equation above is the Metropolis–Hastings ratio, including both the posterior density ratio, and the proposal density ratio. The algorithm always accepts models with Metropolis–Hastings ratios above one, and proposals with ratios lower than one are accepted with probability a. Initially, this seems counterintuitive, but the acceptance of such proposals allows the algorithm to generate samples from the posterior distribution via the Markov chain, enabling the user to then quantify model uncertainty, rather than performing optimisation for a single most probable model.

5 The Frequency Domain

For large-scale time series, all methods discussed become computationally inefficient, therefore, we introduce the frequency domain.

The frequency domain transforms the dataset into a sum of sinusoidal waves to which the time points become frequencies and the power becomes our value of interest. The frequency domains transform uses the Fourier transform which goes as follows:

$$J(\omega_k) \equiv \frac{1}{2\pi} \sum_{t=1}^{n} y_t \exp(-i\omega_k t),$$

where:

$$\omega_k \in \Omega = \left\{2\pi k n^{-1} \text{ for } k = -\left\lceil \frac{n}{2} \right\rceil + 1,\ \ldots,\ \left\lfloor \frac{n}{2} \right\rfloor \right\}.$$

It is important to note that the periodogram measures the power at each frequency and is the basis for further analysis and can be seen below.

$$I(\omega_k) \ = \ n^{-1} \left| J(\omega_k) \right|^2.$$

There are two key reasons as to why this transformation allows for more efficient Bayesian inference. Firstly, due to conjugate symmetry, the negative frequencies produced contain redundant information and therefore allow for the removal of half the Fourier coefficients without information loss. Secondly, likelihood calculations in the time domain for large-scale time series come with a heavy computational burden, but when substituting the exact likelihood evaluation for the Whittle likelihood (Whittle, 1953) the burden decreases significantly due to the asymptotic independence (Whittle, 1953; Gray, 2006) which will be discussed in the next section.

5.1 Conjugate Symmetry

One of the most important features which contribute to the frequency domain's usefulness is its conjugate symmetry. When a real valued time series is transformed via the Fourier transform, the following identity is used:

$$e^{i\theta} = \cos(\theta) + i\sin(\theta).$$

This allows us to decompose the data into a linear combination of sinusoidal components for different frequencies where the number of frequencies would be the size of the dataset T. Conjugate symmetry is then used as the negative-frequency Fourier coefficients are implied to be the complex conjugates of their positive value counterparts. This means that they offer no new information and can be discarded, changing the size of the dataset being analysed from T to $\frac{T}{2}$.

6 The Whittle Likelihood

The Whittle likelihood (Whittle, 1953) acts as an approximation of the exact Gaussian likelihood. It is arguably the most significant reason for why frequency domain analysis for large-scale time series is more efficient than remaining in the time domain. The Whittle likelihood's formulation goes as follows:

$$L_W(\theta) \propto \prod_{k=1}^{\lfloor n/2 \rfloor} \frac{1}{f_\theta(\omega_k)} \exp\left(-\frac{I(\omega_k)}{f_\theta(\omega_k)}\right),$$

$$\text{where:} \quad f_\theta(\omega_k) = \frac{\sigma^2}{2\pi} \frac{|1 + \sum_{j=1}^{q} \theta_j e^{-i\omega j}|^2}{|1 - \sum_{k=1}^{p} \phi_k e^{-i\omega k}|^2}, \ \omega \in (-\pi, \pi].$$

The Whittle likelihood itself is a general function, although the spectral density $f_\theta(\omega_k)$ is model specific. Therefore this project focused specifically on ARMA models so as to maintain the simplicity of the algorithm.

6.1 Computational Efficiency of the Whittle Likelihood

As $T \to \infty$, the ordinates of the periodogram become asymptotically independent. This property implies that the Whittle log-likelihood is a sum of frequencies even in the case that the data are dependent in the time domain (Salomone et al., 2020).

As well as this, the periodogram can be computed once using the fast Fourier transform (FFT) at cost $O(n \log n)$ since it does not depend on the $\boldsymbol{\beta}$ vector. This then allows each Whittle likelihood evaluation to require $O(n)$ number of operations in contrast to the exact time-domain likelihood which necessitates $O(n^3)$ number of operations.

7 Algorithm Development

The initial implementation in this report applied the Metropolis–Hastings algorithm in the time domain with a random walk $\boldsymbol{\beta}$ proposal. There were three features of the implementation that were then changed to adapt

it to the frequency domain; the likelihood function, the $\boldsymbol{\beta}$ proposal, and the parameterisation of the proposal coefficients.

7.1 The Likelihood Function

Due to Whittle's formulation and how accurate the approximation was when T is large, it was possible to substitute the Whittle likelihood for the exact likelihood. Therefore, the Metropolis–Hastings ratio is accepted and rejected as normal.

7.2 The $\boldsymbol{\beta}$ Proposal

It was necessary to change the $\boldsymbol{\beta}$ proposal as the random walk proposal can suffer from mismatched dimensions and get stuck proposing non-optimal values. Intuitively, this can be understood as follows, since the random walk was a multivariate normal distribution centered at zero with variance of σ^2, it would only step as far as σ^2 would allow it, when the independent Bernoulli flip occurred, it would be sent back to zero and result in poor mixing. Therefore, the previously explained $\boldsymbol{\beta}$ proposal was used. We centered a multivariate normal distribution at the *model a posteriori* given all parameters were in, and used its inverse Hessian as the covariance. This largely improved mixing, leading to better quantification of model uncertainty.

7.3 Coefficient Reparameterisation

In the original algorithm, the values were the direct coefficient parameterisation. The issue is that due to the assumption of stationarity bounding analysis, the models being proposed had to be stationary otherwise the evaluations would be invalid. Therefore, we used the Barndorff-Nielsen and Schou (Barndorff-Nielsen and Schou, 1973) reparameterisation. Rather than proposing actual parameter values, we proposed strictly PACF values, which were bounded between $[-1, 1]$, which we could reparameterise back to regular parameter values at a later time. This enforced stationarity and invertibility.

8 Results & Findings

8.0.1 AR(1) Process Results

True $\mathcal{I}$	MH $\mathcal{I}$	True β	MH β
1	1	0.200	0.202
0	0	0.000	0.000

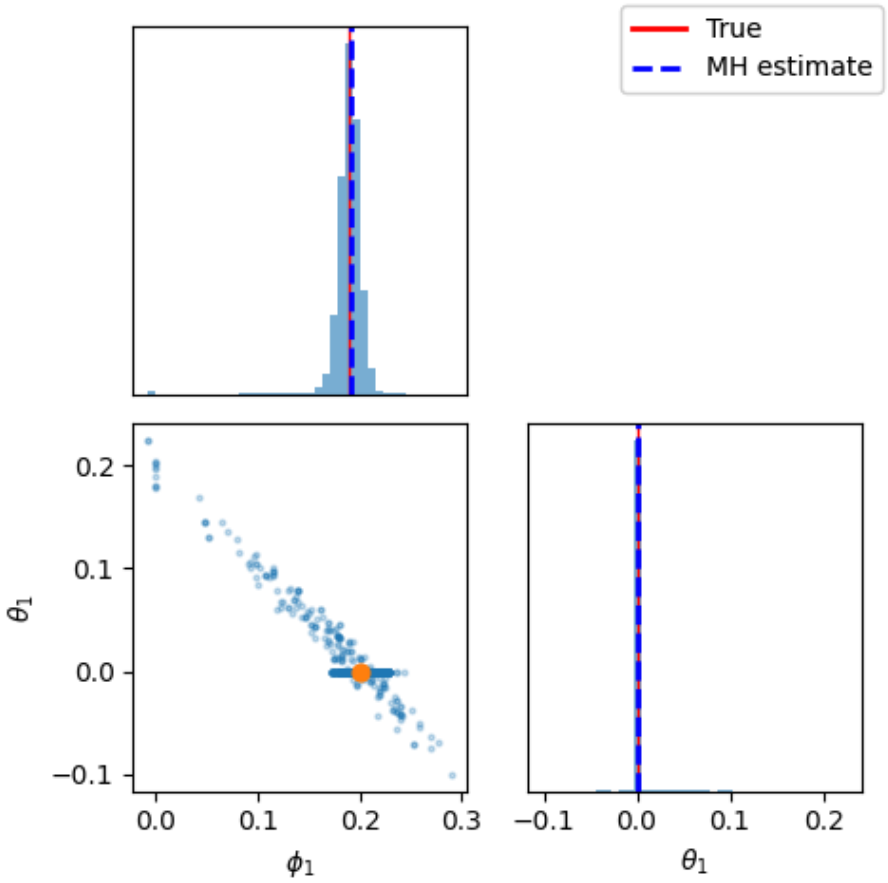

Figure 1: AR(1) Frequency-domain Metropolis–Hastings Posterior Distributions

8.0.2 MA(1) Process Results

True $\mathcal{I}$	MH $\mathcal{I}$	True β	MH β
0	0	0.000	0.000
1	1	-0.300	-0.298

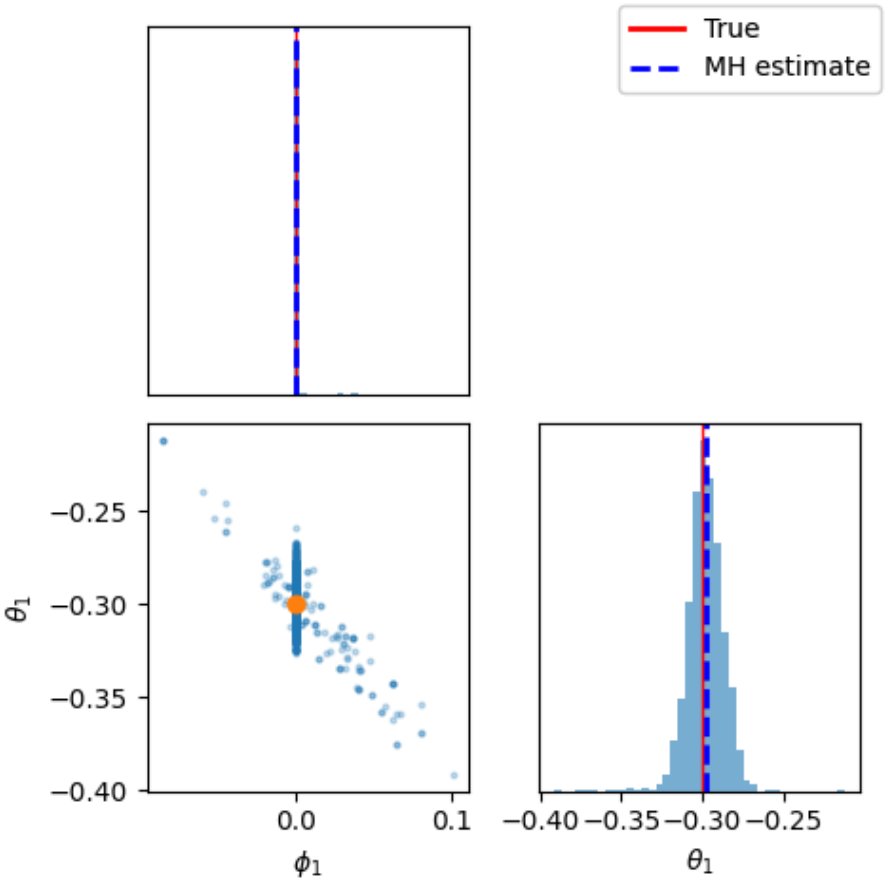

Figure 2: MA(1) Frequency-domain Metropolis–Hastings Posterior Distributions

8.0.3 ARMA(4,1) Process Results

True $\mathcal{I}$	MH $\mathcal{I}$	True β	MH β
1	1	0.200	0.208
1	1	-0.300	-0.302
0	0	0.000	0.000
1	1	0.500	0.488
1	1	-0.300	-0.310

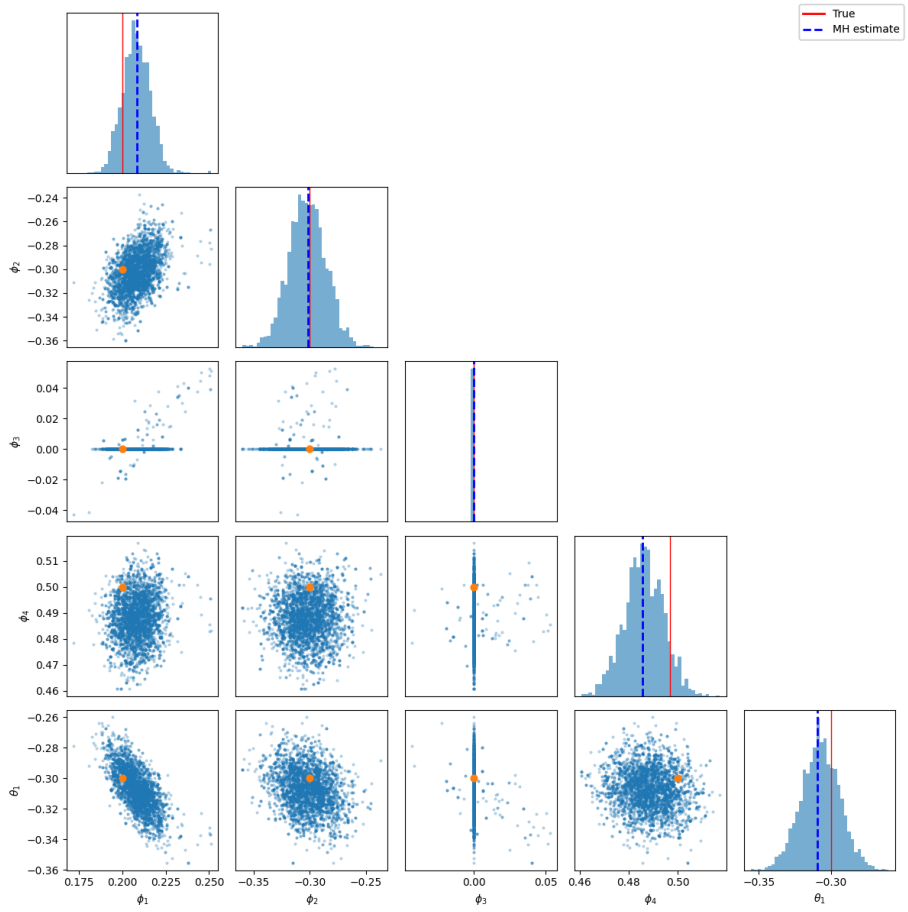

Figure 3: ARMA(4,1) Frequency-domain Metropolis–Hastings Posterior Distributions

8.0.4 ARMA(11,1) Process Results

True $\mathcal{I}$	MH $\mathcal{I}$	True β	MH β
1	1	0.200	0.193
1	1	-0.300	-0.306
0	0	0.000	0.000
1	1	0.500	0.490
1	1	0.200	0.201
1	1	-0.300	-0.284
0	0	0.000	0.000
1	1	-0.500	-0.497
1	1	0.200	0.177
0	0	0.000	0.000
1	1	0.200	0.198
1	1	0.100	-0.123

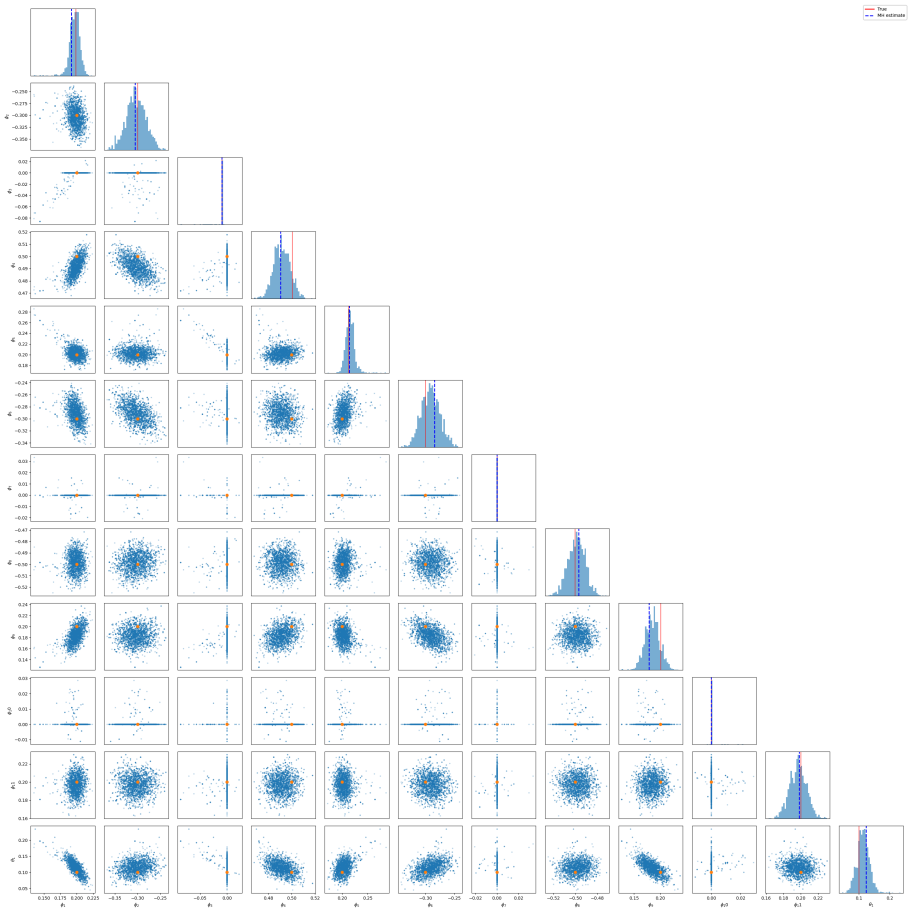

Figure 4: ARMA(11,1) Frequency-domain Metropolis–Hastings Posterior Distributions

8.1 Limitations

There were two main limitations found. Firstly, the autoregressive process estimation was limited to the order of 11. The accuracy of such estimates less than 11 proved to be high. With the correct tuning, inclusion vector estimates would be exact, or, if they were not, their PACF values would reflect their insignificance.

Secondly, the largest of the limitations was that the moving average process was largely inaccurate after an order of 1. This meant that estimating an MA(1) process would be highly accurate, but moving to larger orders would both affect the inclusion vector estimates and PACF value estimates. The reasoning for this has been left to further research.

Combining the two limitations, the highest order process the algorithm could reliably estimate was an ARMA(11,1) process with exact inclusion vector estimates after tuning the $\boldsymbol{\mathcal{I}}$ proposal through the independent Bernoulli flip's predetermined probability.

Finally, the optimisation used was limited to estimating only parameter values, therefore the algorithm was restricted to cases with known variance (σ^2).

9 Conclusion

This report developed a frequency-domain Metropolis–Hastings algorithm for Bayesian variable selection of ARMA(p,q) models. The algorithm has replaced the exact time-domain likelihood with the frequency-domain approximation method, the Whittle likelihood (Whittle, 1953) as it provides approximate likelihoods and is computationally more efficient. Simulations demonstrate accurate recovery of both the inclusion indicators and parameter estimates for ARMA(p,q) models up to the order of ARMA(11,1).

The limitation of the moving average process to an order of one and restriction of only endogenous parameters provides a clear pathway for future research.

10 Further Research

As stated previously, the restriction of the moving average component to order one has been left up to further research. Extending the algorithm beyond the ARMA(11,1) process will improve applicability to real economic data, which often exhibits higher-order moving average processes.

In addition, future work could extend the optimisation component beyond the *model a posteriori* to include $\log \sigma^2$. Estimation of $\log \sigma^2$ would allow for real data analysis rather than restricting it to simulated data.

Finally, future work could involve developing the algorithm further and allowing for simultaneous evaluation of alternative processes such as ARDL(p,q) along with the ARMA(p,q) specification. This would improve the flexibility of the algorithm substantially, as it would allow for exogenous variables.

References

Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

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

Robert M. Gray. Toeplitz and circulant matrices: A review. *Foundations and Trends in Communications and Information Theory*, 2(3):155–239, 2006.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1): 97–109, 1970.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

Robert B. O'Hara and Mikko J. Sillanpää. A review of bayesian variable selection methods: what, how and which. *Bayesian Analysis*, 4(1):85–117, 2009.

Robert Salomone, Matias Quiroz, Robert Kohn, Mattias Villani, and Minh-Ngoc Tran. Spectral subsampling MCMC for stationary time series. *International Conference on Machine Learning (ICML2020)*, 2020.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

Peter Whittle. The analysis of multiple stationary time series. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1953.