Bayesian Subset ARMA Model Selection for Sparse Models

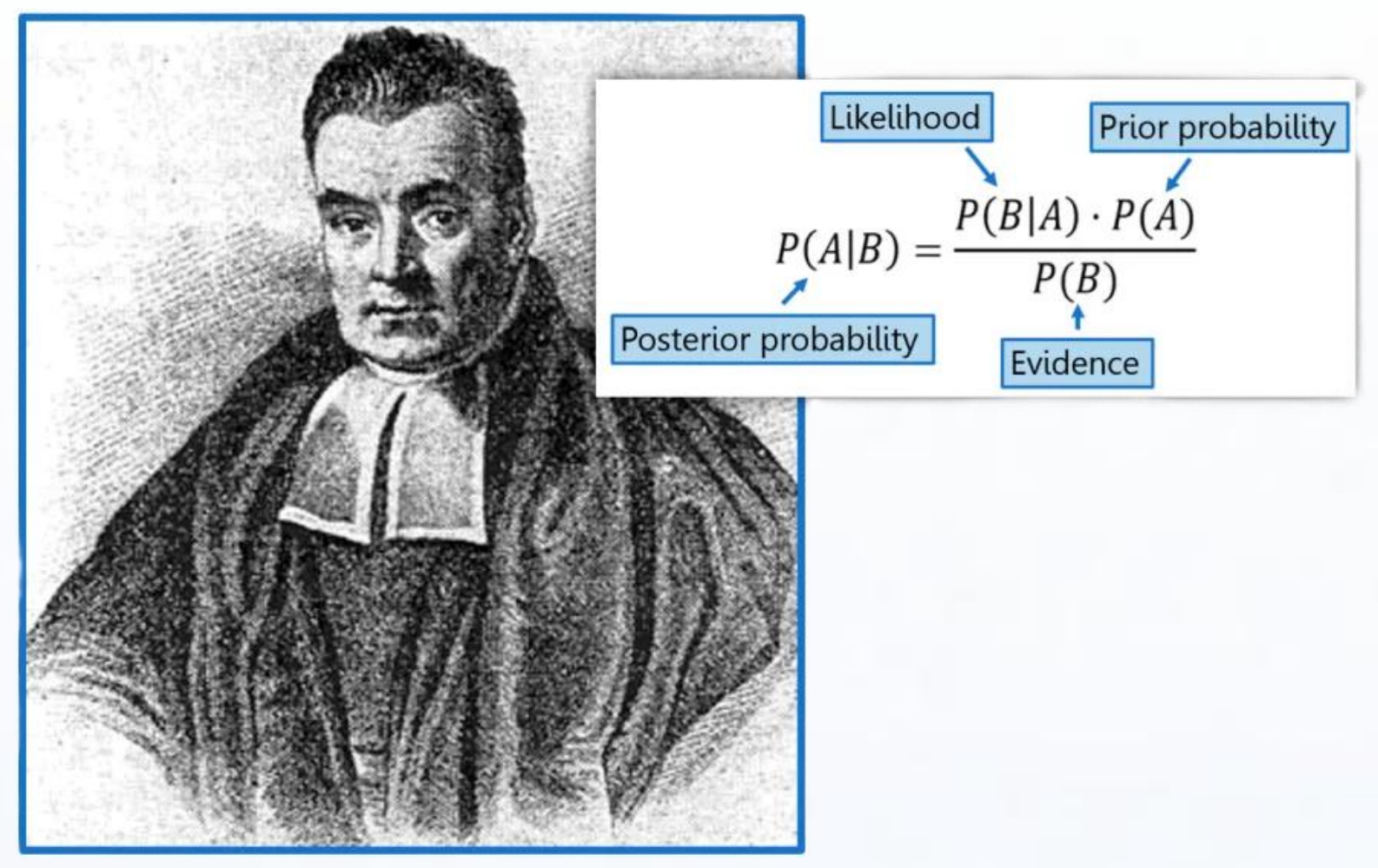
University of Crete

Applied & Computational Mathematics

Student: Jorgo Xhixho

Supervisor: Yiannis Kamarianakis

October 2025



Introduction

This work investigates **sparse ARMA** models for time-series data—settings where only a few lags matter. Motivated by a paper **"Subset ARMA selection via the adaptive Lasso"** for model selection and shrinkage, we design a **Bayesian** framework with two complementary approaches:

1. global–local shrinkage via Horseshoe/Horseshoe+ priors combined with predictive posterior projection, and
2. an INLA-based formulation with predictive posterior projections.

We benchmark these against Auto-ARIMA and ADAM.

The aim is to compare the models and find which is more efficient finding the Data generating Model.

ARMA Model

$$\mathrm{Y_t} = \sum_{i=1}^{p} \alpha_i \mathrm{Y}_{t-i} + \sum_{j=1}^{q} \beta_j \varepsilon_{t-j} + \varepsilon_t,$$

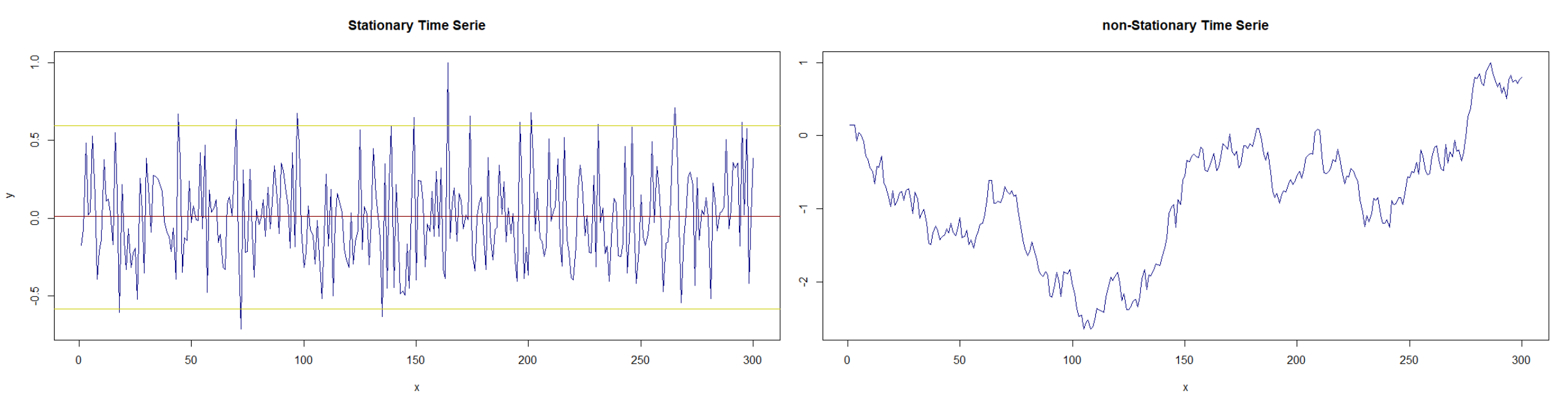
$$\varepsilon_t \sim \text{i.i.d.}\left(0, \sigma^2\right)$$

with $\alpha_i = 0$ or $\beta_j = 0$ for many i, j ⇒ sparse structure.

Combination of Autoregressive (AR) & Moving Average (MA) Models with (p,q) their finite lag orders.

Required Properties:

1. **Weak Stationarity: constant mean & variance.**
2. **Invertibility: unique representation of the process.**



Bayesian Framework

Bayesian methods offer a different approach to statistical analysis, parameters are random variables with prior distributions, updated by observed data to yield posterior distributions. This approach incorporates prior knowledge and quantifies uncertainty in parameter estimates.

Frequentist approaches, by comparison, view parameters as fixed unknowns and base inference entirely on the data, typically offering point estimates without a complete uncertainty distribution.

Bayesian Framework:

$$p(\theta|y) \propto p(y|\theta)\, p(\theta), \quad \theta = \left(\mu, \alpha, \beta, \sigma^2\right)$$

posterior likelihood prior

Model & Likelihood :

For observations $y = (y_1, \ldots, y_n)$,

$$y \sim N_n\left(\mu\, \mathbf{1},\, \Sigma\left(\alpha, \beta, \sigma^2\right)\right)$$

Where $\Sigma = \sigma^2 L L^{\top}$ and L is the lower-triangular matrix built from $MA(\infty)$ coefficients:

$$\frac{1 + \beta_1 B + \cdots + \beta_q B^q}{1 - \alpha_1 B - \cdots - \alpha_p B^p} = \sum_{j=0}^{\infty} \psi_j B^j$$

and B is a backshift operator : $B y_t = y_{t-1}$, $B^k y_t = y_{t-k}$.

$$p(\,y \mid \mu, \alpha, \beta, \sigma^2\,) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(y - \mu \mathbf{1})^{\top} \Sigma^{-1} (y - \mu \mathbf{1})\right)$$

Priors:

- $\mu \sim N(\mu_0, \sigma_0^2)$
- $\alpha \sim N_p(0, \Sigma_\alpha)$
- $\beta \sim N_q\left(0, \Sigma_\beta\right)$
- $\sigma^2 \sim Inv\text{–}Gamma(a, b)$

Posterior:

$$p(\,\mu, \alpha, \beta, \sigma^2 \mid y\,) \propto p(\,y \mid \mu, \alpha, \beta, \sigma^2\,), p(\mu)\, p(\alpha)\, p(\beta)\, p(\sigma^2)$$

In practice the innovations $\{\varepsilon_t\}$ are unobserved. Under invertibility, MA(q) can be estimated from an AR(∞) representation. We use a two-step Bayesian scheme, approximate $\widehat{\varepsilon_t}$ from data and then sample form the Priors.

I: Choose a high AR order $m = \max(\lceil 10 \log_{10}(n) \rceil, \max(p, q) + 1)$ and fit $y_t = \beta_0 + \sum_{i=1}^{m} \beta_i y_{t-i} + \eta_t,\ \eta_t \sim N\left(0, \sigma_\eta^2\right)$ using Markov Chain Monte Carlo method (Gibbs sampling).

II: Run MCMC for $y_t = \mu + \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{j=1}^{q} \theta_j\, \widehat{\varepsilon_{t-j}} + \nu_t, \quad \nu_t \sim N\left(0, \sigma_\nu^2\right)$
or : $Y = Xb + \nu$ with $b = [\alpha; \theta]$ and$X = \left[\, y_{t-1:t-p}\ \ \hat{\varepsilon}_{t-1:t-q} \right]$.
Priors: $b \sim N_{p+q}(0, \Sigma_b)$, $\sigma_\nu^2 \sim Inv - Gamma(\alpha_\nu, \beta_\nu)$

Global-Local shrinkage

To introduce sparsity in the ARMA framework, we use modern Bayesian global–local shrinkage priors: the **Horseshoe** & **Horseshoe+**. These priors apply strong shrinkage to small or irrelevant coefficients while allowing important ones to remain large. The Horseshoe prior achieves this through a hierarchical Half-Cauchy structure, while Horseshoe+ extends it with an additional layer of shrinkage, improving performance in highly sparse settings. To apply shrinkage, we have chosen different priors in the second step of the Framework.

Horseshoe Priors:

Formally, for each APMA coefficient (for simplicity we are using β) β_ι the Horseshoe prior is defined as:

1. Hierarchically (per coefficient):

$$\sigma^2 \sim Inv - Gamma(\alpha, \beta),\ \lambda_i \sim C^{+}(0,1),\ \tau \sim C^{+}(0,1)$$

$$\beta_i \mid \lambda_i, \tau, \sigma^2 \sim N\left(0,\ \lambda_i^2 \tau^2 \sigma^2\right),$$

2. Shrinkage factor:

$$\kappa_i \sim \mathrm{B}\,(\,0.5, 0.5), \qquad \kappa_i = \frac{1}{1 + \lambda_i^2 \tau^2},$$

This favors values near to 0 or 1 ⇒ keep large effects, crush noise.

The Horseshoe prior possesses some desirable theoretical properties for sparse signal recovery.

- **Optimal Posterior Concentration**: The posterior concentrates at the true sparse signal at a near-optimal rate, even in high-dimensional settings.
- **Oracle Properties**: The Horseshoe prior achieves similar performance to oracle methods that know the true non-zero coefficients, as shown through its shrinkage factor κ_i.
- **Robustness**: The heavy-tailed nature of the prior ensures that large coefficients are not over-shrunk, critical for capturing significant AR or MA lags in our models.

Horseshoe + Priors:

The Horseshoe+ prior extends the Horseshoe prior to address ultra-sparse settings, where the number of non-zero coefficients is extremely small, as often encountered in ARMA models with large orders (p,q), by introducing an additional layer of local shrinkage to enhances sparsity while maintaining robustness for large coefficients.

1. Hierarchically (per coefficient):

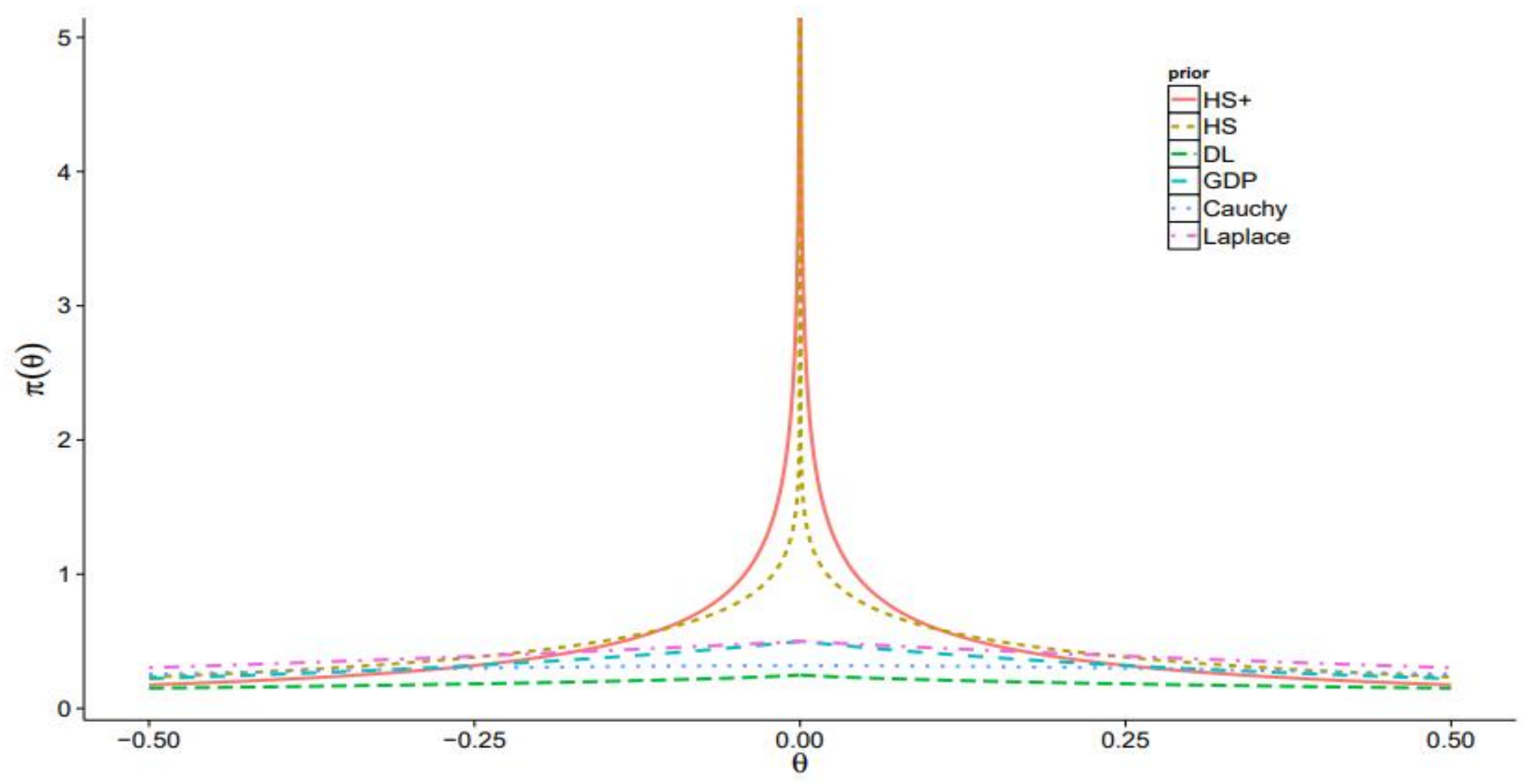
$$\beta_i \mid \lambda_{1,i}, \lambda_{2,i}, \tau, \sigma^2 \sim N\left(0, {\lambda_{1,i}}^2 {\lambda_{2,i}}^2, \tau^2, \sigma^2\right),$$

$$\tau \sim C^{+}(0,1),\ \lambda_{2,i} \sim C^{+}(0,1),\ \lambda_{1,i} \sim C^{+}(0,1),$$

$$\sigma^2 \sim Inv - Gamma(\alpha, \beta)$$

2. Shrinkage factor:

$$\kappa_i = \frac{1}{1 + {\lambda_{1,i}}^2 {\lambda_{2,i}}^2 \tau^2}$$



Model Selection

HS/HS+ priors yield nearly sparse posterior draws but do not set unwanted ARMA coefficients exactly to zero; we enforce true sparsity via a final model selection step using Kullback-Leibler divergence between posterior predictive distributions—favoring parsimonious submodels with minimal KL to the full model.

Kullback-Leibler divergence estimate

$$V_{\perp} \subset V_F = \{1, \ldots, p + q\}$$

$$D(V_F | V_{\perp}) = \frac{1}{S} \sum_{s=1}^{S} \log \frac{\sigma^{(s)}(V_{\perp})}{\sigma^{(s)}(V_F)}$$

Where S denotes the total number of MCMC draws used to approximate the posterior predictive distribution, $V_{\perp}$ is a submodel & V_F the full model.

Integrated Nested Laplace Approximation

As an alternative methodology—novel in sparse ARMA selection, to our knowledge—we propose **INLA** for fast posterior inference in latent Gaussian models, combined with KL-based model selection via posterior predictive projection, achieving sparsity without MCMC convergence concerns.

- ➢ Deterministic approximations for marginal posteriors in Latent Gaussian Models (LGMs
- ➢ Exploits GMRF sparsity for fast linear algebra (sparse Cholesky).
- ➢ Orders-of-magnitude speedups vs. MCMC in typical LGM settings.

Model Type	Sample Size	MCMC Time (s)	INLA Time (s)
Simple Linear Regression	100	4.19	0.176
Simple Linear Regression	5000	381.573	2.787
Poisson GLM with IID Effect	100	30.394	0.383
Poisson GLM with IID Effect	100000	> 6 hours	166.819

INLA structure can be presented into three layers:

Laten Gaussian Models (LGMs):

- Likelihood: $\mathrm{y_i} \sim \mathrm{p}(\mathrm{y_i} |\, \eta_i \theta_2)$
- Latent field: $\mathrm{x} | \theta_1 \sim \mathrm{N}(0, \Sigma)$
- Hyperparameters: $\theta = [\theta_1, \theta_2 \ldots.]^{\top} \sim p(\theta)$

$$\mathrm{p}(\mathrm{x_i} | \mathrm{y}) = \int \mathrm{p}(\mathrm{x_i} |\, \theta, \mathrm{y}) \mathrm{p}(\theta | \mathrm{y}) \mathrm{d}\theta$$

Gaussian Markov Random Fields(GMRFs):

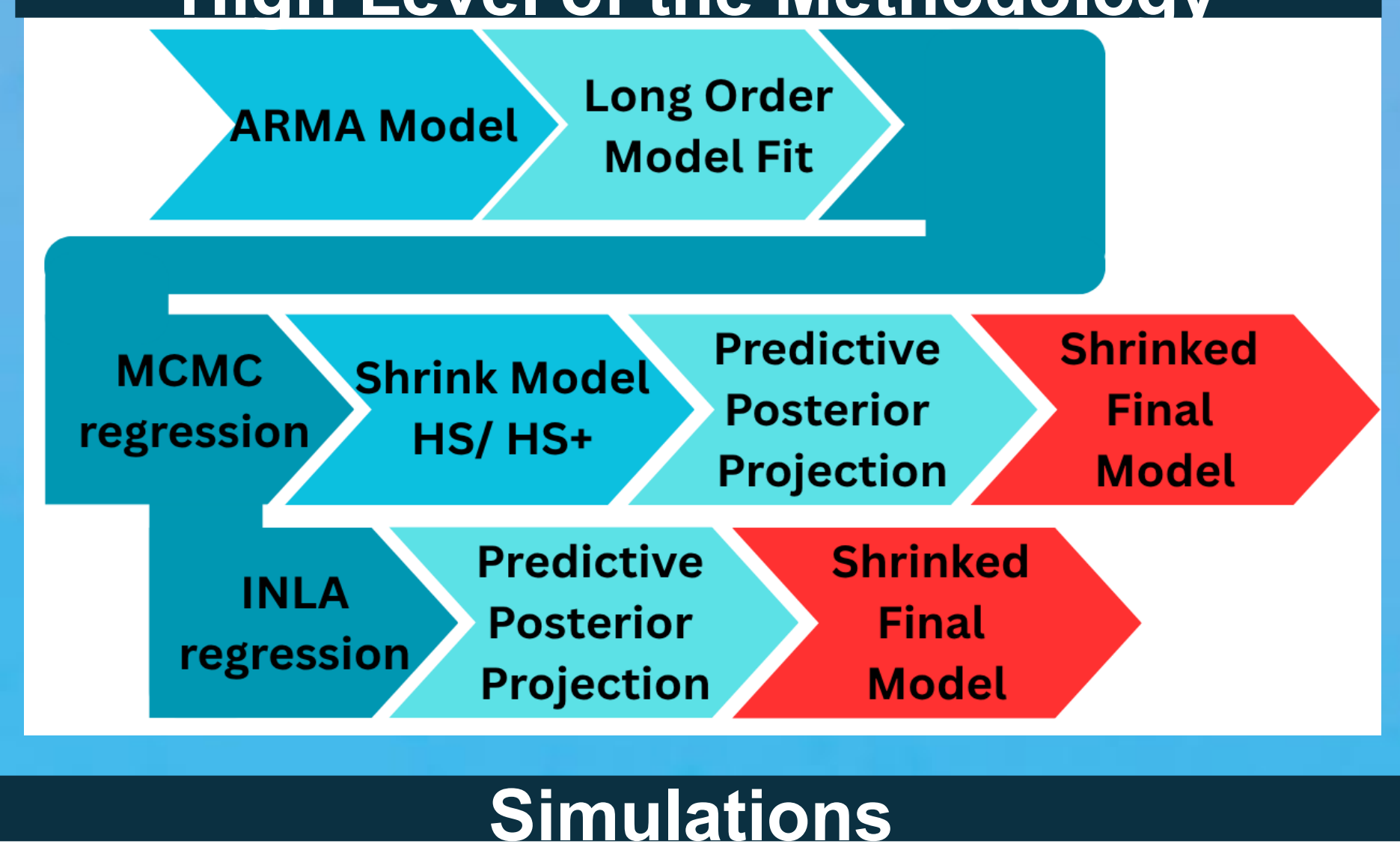
- Latent field has a sparse precession matrix $\mathrm{Q} = \Sigma^{-1}$
- Sparsity encode conditional independence $\mathrm{Q_{ij}} = 0 \iff \mathrm{x_i} \perp \mathrm{x_j}\ (\mathrm{i} \neq \mathrm{j})$

INLA:

1. Approximate p(θ|y) with Laplace around conditional mode.
2. Approximate p(x|θ, y) with Gaussian or Laplace approximation.
3. Numerical integration over θ

$$p(\,x_i \mid y\,) \approx \sum_{K=1}^{K} p(\,x_i \mid \theta_k, y\,) p(\,\theta_k \mid y\,) \Delta_k$$

High Level of the Methodology



ARMA Model → Long Order Model Fit

MCMC regression → Shrink Model HS/ HS+ → Predictive Posterior Projection → Shrinked Final Model

INLA regression → Predictive Posterior Projection → Shrinked Final Model

Simulations

We Compared Bayesian regularization with Horseshoe (HS/HS+) to Auto ARIMA, INLA, and ADAM on synthetic seasonal ARMA. The Monte Carlo Setup included 1000 replicates with 4 different sample sizes nϵ{120,240,360500}, using the same models from the paper **"Subset ARMA selection via the adaptive Lasso"**.

Data-generating models (where B is the backshift operator):

I: $(1 - 0.8B)(1 - 0.7B^6)\, y_t = \epsilon_t$

II: $(1 - 0.8B)(1 - 0.7B^6)\, y_t = (1 + 0.8B)(1 + 0.7B^6)\epsilon_t$

III: $y_t = (1 + 0.8B)\left(1 + 0.7B^6\right)\epsilon_t$

IV: $y_t = \left(1 - 0.6B - 0.8B^{12}\right)\epsilon_t$

Evaluation:

- **"ACC" Relative frequencies of pic all the significant variables.**
- **"TPR" Picking Correct model**
- **"FPR" False Positive rate**
- **"FNR" False Negative rate**

		Actual Coefficients	
Predicted Coefficients		$\beta_i = 0$	$\beta_i \neq 0$
	$\hat{\beta}_i = 0$	T.N.	F.N.
	$\hat{\beta}_i \neq 0$	F.P.	T.P.

Error/accuracy metrics.

$$\mathrm{Var}_j = \frac{1}{N-1} \sum_{i=1}^{N} (\hat{\theta}_{ij} - \bar{\theta}_j)^2, \quad \text{where } \bar{\theta}_j = \frac{1}{N} \sum_{i=1}^{N} \hat{\theta}_{ij}, \quad \mathrm{Bias}_j = \frac{1}{N} \sum_{i=1}^{N} \hat{\theta}_{ij} - \theta_j$$

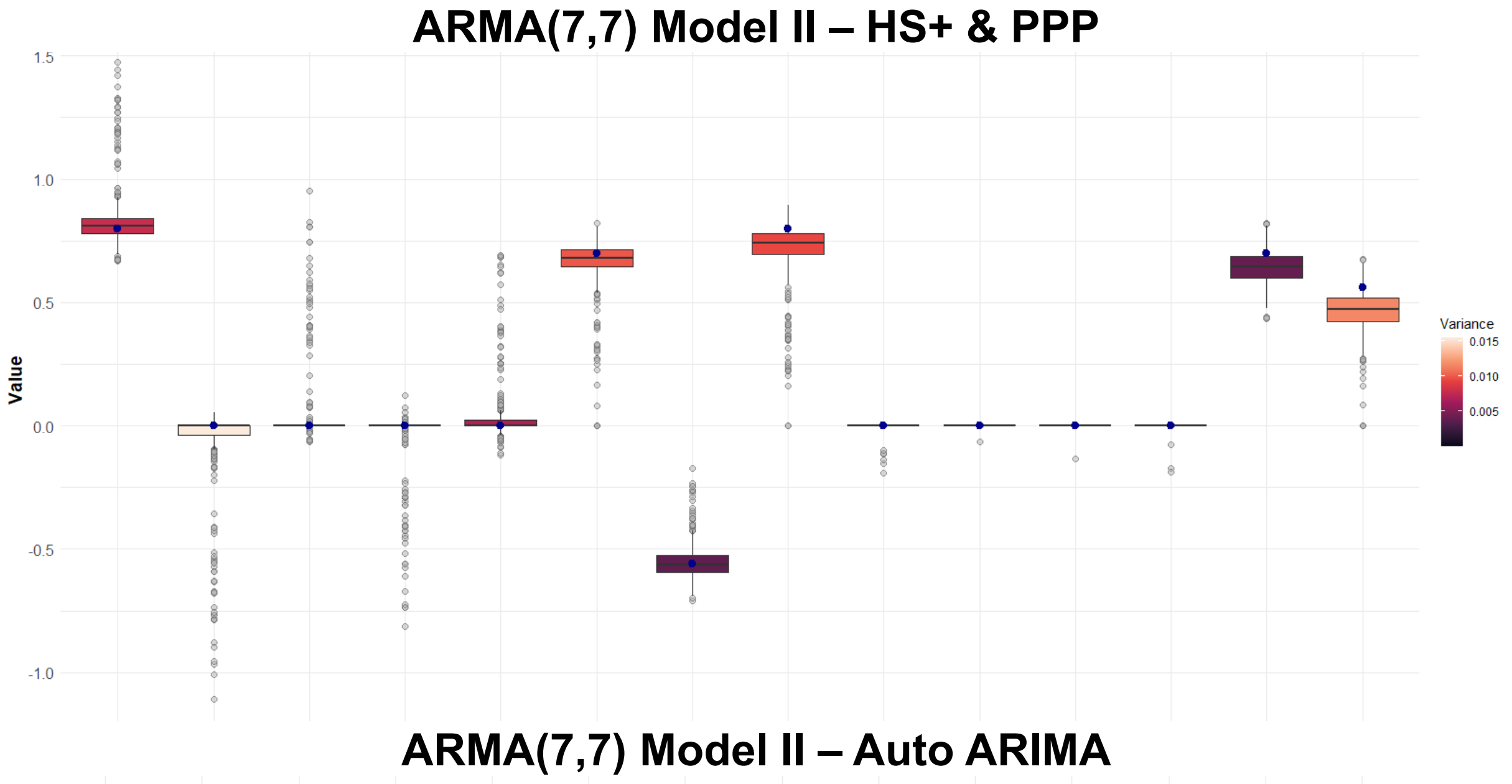
$$\mathrm{MSE}_j = \mathrm{Bias}_j^2 + \mathrm{Var}_j, \qquad \mathrm{MSE_{av}} = \sqrt{\sum_{j=1}^{p} \mathrm{MSE}_j}$$

$$\mathrm{ACC} = \frac{TP + TN}{TP + FP + TN + FN}, \qquad \mathrm{SSc} = \frac{\mathrm{MSE_{av}^{LS}}}{\mathrm{MSE_{av}^{PL}}}, \qquad \mathrm{NewACC} = \mathrm{SSc} \times \mathrm{ACC_{projection}}$$

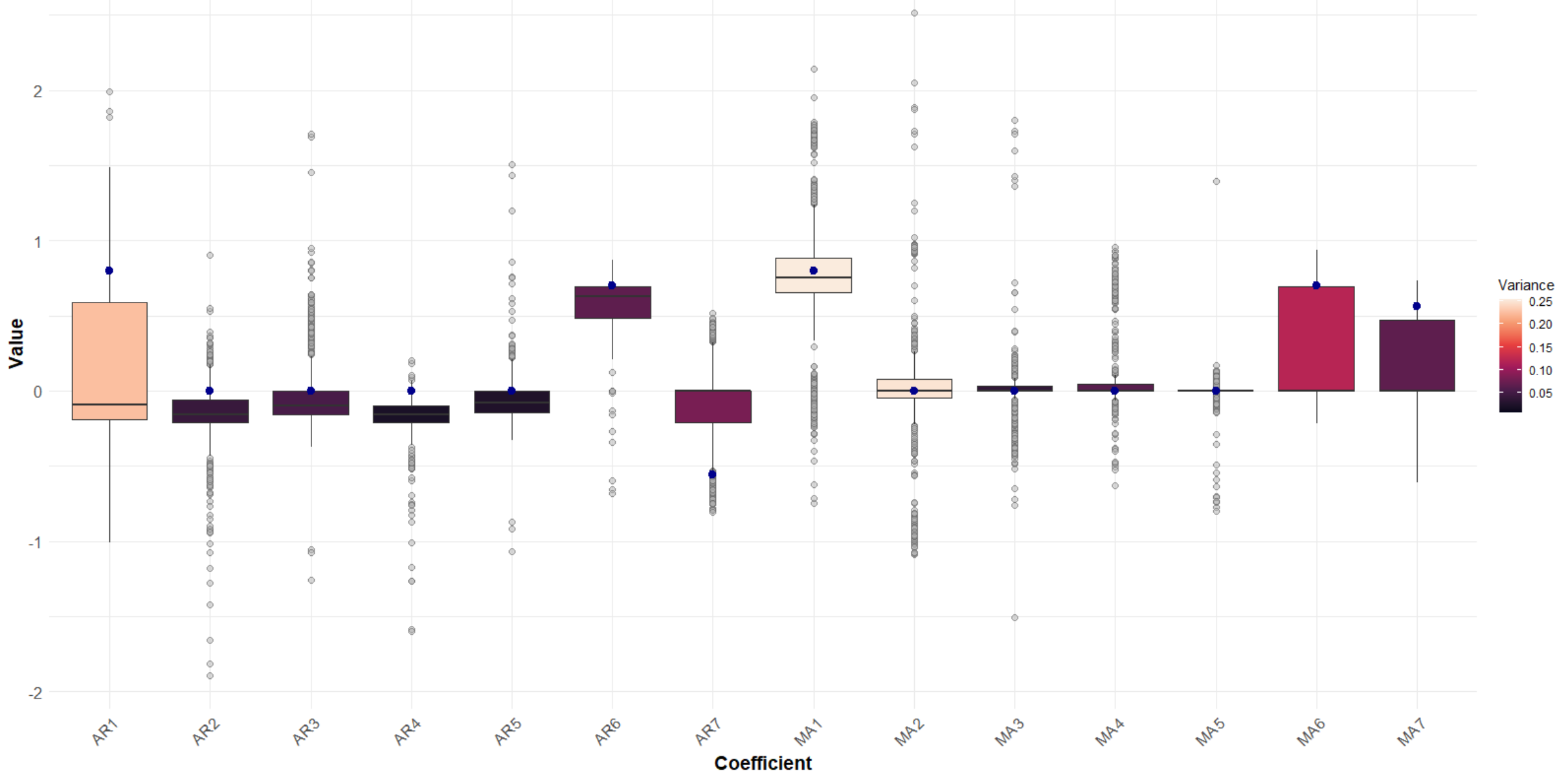
Results from ARMA(7,7) Model II

Methods	Metrics	120	240	360	500	Comp. Time
HS+ & P	MSE av	0.70	0.57	0.39	0.33	🟥
	ACC	0.84	0.88	0.90	0.90	
	TPR	0.93	0.97	0.99	0.99	
INLA	MSE av	1.78	1.14	0.72	0.56	🟨
	ACC	0.61	0.76	0.84	0.88	
	TPR	0.95	0.96	0.97	0.99	
Auto ARMA	MSE av	1.87	1.80	1.62	1.61	🟩
	ACC	0.51	0.45	0.44	0.44	
	TPR	0.40	0.54	0.63	0.67	
ADAM	MSE av	1.63	1.61	1.61	1.61	🟩
	ACC	0.55	0.55	0.55	0.56	
	TPR	0.17	0.17	0.17	0.17	
Adaptive Lasso	MSE av	NA	NA	NA	NA	🟩
	ACC	0.40	0.81	0.92	0.93	
	TPR	0.01	0.03	0.04	0.09	

ARMA(7,7) Model II – HS+ & PPP



ARMA(7,7) Model II – Auto ARIMA


Value
Variance
AR1
AR2
AR3
AR4
AR5
AR6
AR7
MA1
MA2
MA3
MA4
MA5
MA6
MA7
Coefficient