Fundação Getulio Vargas

Problem Set 2 : Estimating Dynamic US Yield Curve

Diogo D. Sánchez[1,†]

[1] *Sao Paulo School of Economics, FGV*

This manuscript was compile on April 20, 2025

Abstract

Este trabalho busca estimar modelo clássico de curva de juros *Nelson-Siegel* por meios bayesianos durante o período de 2002 até 2021 de forma dinâmica.

Keywords: *Monte Carlo, Markov Chain ,Bayes , Macroeconomia, Yield Curve, US*

Contents

1 Identify the model parameters and propose a prior distribution for them. 2

2 Propose a MCMC sampler for the model parameters. 4

2.1 Forward Filtering Backward Sampling - FFBS . . . 4
2.2 Normal sampler - For μ . . . 5
2.3 Normal-Wishart Φ and Q sampler - BVAR . . . 5
2.4 Posterior Sampling for H - Inversa Gamma . . . 5
2.5 Posterior Sampling for λ - Adaptative Random -Walk Metropolis Hasting . . . 6
2.6 Gibbs Sample with Metropolis . . . 7

3 Implement the sampler and show its diagnostics 8

4 Plot the posterior density of the elements of the autoregressive matrix A 14

5 Plot a predictive curve with respective predictive intervals for maturities from 3 to 360 months 15

6 Conclusão 15

Notations

Let T denote the number of time periods, and for each $t = 1, \dots, T$, let there be $N = 20$ cross-sectional observations $y_t^{(i)}$, $i = 1, \dots, N$, each associated with a fixed maturity τ_i.
We define the following model components:

- $\lambda^{-1} \in \mathbb{R}_+$: decay parameter of the Nelson-Siegel exponential components.
- $\beta_t \in \mathbb{R}^3$: vector of latent factors at time t:

$$\beta_t = \begin{bmatrix} \beta_{t,1} \\ \beta_{t,2} \\ \beta_{t,3} \end{bmatrix} = \begin{bmatrix} \text{Level } (L_t) \\ \text{Slope } (S_t) \\ \text{Curvature } (C_t) \end{bmatrix}$$

- $\Phi \in \mathbb{R}^{3\times3}$: transition matrix governing the autoregressive dynamics of the latent factors.
- $\mu \in \mathbb{R}^3$: fixed mean vector representing the unconditional average of the latent process.
- $Q \in \mathbb{R}^{3\times3}$: covariance matrix of the state innovation process.
- $H = \text{diag}(\sigma_1^2, \dots, \sigma_N^2) \in \mathbb{R}^{N\times N}$: diagonal covariance matrix of measurement errors, with one variance per maturity.

State equation:

$$\beta_t = \mu + \Phi \cdot \beta_{t-1} + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, Q) \tag{1}$$

Observation equation:

$$y_t = \Lambda(\lambda) \cdot \beta_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, H) \tag{2}$$

where $\Lambda(\lambda) \in \mathbb{R}^{N\times3}$ is the loading matrix whose i-th row is given by:

$$\Lambda_i(\lambda) = \begin{bmatrix} 1 & \dfrac{1 - e^{-\lambda^{-1}\tau_i}}{\lambda^{-1}\tau_i} & \dfrac{1 - e^{-\lambda^{-1}\tau_i}}{\lambda^{-1}\tau_i} - e^{-\lambda^{-1}\tau_i} \end{bmatrix}$$

All parameters and their prior distributions will be formally specified in section 1.

1. Identify the model parameters and propose a prior distribution for them.

The model has a total of $\beta_{(3T)} + \phi_{(9)} + \mu_{(3)} + Q_{(6)} + H_{(20)} = 3015 + 38 = 3053$ parameters.

Diebold and Li (2006) [3] fixed the value of $1/\lambda$ at 0.0609, which implies $\lambda = 16.42036$. Yu and Zivot (2011) [5] adopted $1/\lambda = 0.077$, leading to $\lambda = 12.98$. Taking the average of these two reference values yields a central tendency around $\lambda = 14.7036$, Using half the difference between them as a proxy for dispersion results in a standard deviation of approximately 1.75.

While a log-normal prior could match these moments directly, we choose to specify a Gamma distribution for computational convenience. We set:

$$\lambda \sim \text{Gamma}(\alpha = 59,\ \beta = 0.25)$$

This implies a mean of $\mathbb{E}[\lambda] = \alpha\beta = 14.75$ and a standard deviation of approximately 1.9

The prior for hyperaparameter $\mathbb{E}[\mu]$ is constructed from the mean estimation observed in Guidolin (2019) [6] for the period between 2001 and 2002. The prior for $\text{cov}(\mu)$ is constructed from observed parameters volatility between 1983 and 2002.

$$\mu \sim \mathcal{N}\left(\begin{bmatrix}6\\0\\0\end{bmatrix}; \begin{bmatrix}3^2 & 0 & 0\\0 & 2^2 & 0\\0 & 0 & 5^2\end{bmatrix}\right)$$

We begin by defining the transition matrix Φ, which governs the autoregressive dynamics of the latent factors L_t, S_t, C_t in the state-space representation of the dynamic Nelson-Siegel model:

$$\Phi = \begin{bmatrix}\phi_{LL} & \phi_{LS} & \phi_{LC}\\ \phi_{SL} & \phi_{SS} & \phi_{SC}\\ \phi_{CL} & \phi_{CS} & \phi_{CC}\end{bmatrix}$$

Empirical studies (e.g., Guidolin, 2019; Diebold et al., 2006) consistently report strong persistence in the level, slope, and curvature components of the yield curve. This empirical regularity implies that the current value of each factor is largely explained by its own past, with autoregressive coefficients typically close to unity. To encode this belief, we specify the prior mean of Φ as:

$$M_0 = I_3$$

That is, each latent factor is expected a priori to follow a persistent AR(1) process centered around identity.

Regarding the prior covariance, we adopt a structure that allows for a priori correlation between coefficients within the same row of Φ — that is, within the same latent factor equation — but assumes independence across rows. This is motivated by the fact that each row of Φ governs the dynamics of a distinct latent variable (L_t, S_t, or C_t), and it is not reasonable to expect that an increase in persistence of one component (e.g., level) should be compensated by parameter changes in another (e.g., slope).

To formalize this, we set:

$$V_0 = 2 \cdot I_3$$

The scaling factor 2 is chosen to match the empirical magnitude of the standard errors of the autoregressive coefficients as found in frequentist estimations (see Guidolin, 2019). Under this specification, the prior distribution for Φ given Q is:

$$\Phi \mid Q \sim \mathcal{MN}(M_0, V_0, Q)$$

which implies:

$$\text{vec}(\Phi) \mid Q \sim \mathcal{N}(\text{vec}(I_3),\ Q \otimes 2I_3)$$

This Kronecker product structure leads to the following block matrix for the prior covariance of $\text{vec}(\Phi)$:

$$Q \otimes V_0 = \begin{bmatrix}
\overset{\text{Var}(\phi_{LL})}{\boxed{2Q_{11}}} & 0 & 0 & \overset{\text{Cov}(\phi_{LL},\phi_{LS})}{\boxed{2Q_{12}}} & 0 & 0 & 2Q_{13} & 0 & 0\\
0 & 2Q_{11} & 0 & 0 & 2Q_{12} & 0 & 0 & 2Q_{13} & 0\\
0 & 0 & 2Q_{11} & 0 & 0 & 2Q_{12} & 0 & 0 & 2Q_{13}\\
2Q_{21} & 0 & 0 & 2Q_{22} & 0 & 0 & 2Q_{23} & 0 & 0\\
0 & 2Q_{21} & 0 & 0 & 2Q_{22} & 0 & 0 & 2Q_{23} & 0\\
0 & 0 & 2Q_{21} & 0 & 0 & 2Q_{22} & 0 & 0 & 2Q_{23}\\
2Q_{31} & 0 & 0 & 2Q_{32} & 0 & 0 & 2Q_{33} & 0 & 0\\
0 & 2Q_{31} & 0 & 0 & 2Q_{32} & 0 & 0 & 2Q_{33} & 0\\
0 & 0 & 2Q_{31} & 0 & 0 & 2Q_{32} & 0 & 0 & 2Q_{33}
\end{bmatrix}$$

This structure ensures that coefficients from the same row of Φ (e.g., ϕ_{LL} and ϕ_{LS}) are allowed to be a priori correlated, while coefficients from different rows (e.g., ϕ_{LL} and ϕ_{SL}) remain a priori uncorrelated. Independence across rows is therefore built directly into the prior via the Kronecker structure $Q \otimes V_0$, preserving both interpretability and parsimony in the dynamics of the latent yield curve factors.

The prior for Q is a covariance matrix with the following structure:

$$Q^{-1} \sim \mathcal{W}\left(\begin{bmatrix}0.063 & -0.04 & 0.032\\ -0.04 & 0.107 & -0.046\\ 0.032 & -0.046 & 0.436\end{bmatrix}^{-1},\ \nu_0 = 18\right)$$

The prior is defined with a Wishart distribution, and the precision parameter ν_0 is set to a very low value, indicating very little prior precision on the covariance matrix, 1 df for year (1983 -> 2001).

The prior for H is a Gamma distribution with the following form:

$$H \sim \mathcal{IG}\left(\begin{bmatrix}18.0\\18.0\\18.0\\18.0\\18.0\\18.0\\18.0\\18.0\\18.0\\18.0\\18.0\\18.0\\18.0\\18.0\\18.0\\18.0\\18.0\\18.0\\18.0\end{bmatrix}; 17 \times \begin{bmatrix}0.07004198\\0.06969731\\0.07000068\\0.0708754\\0.07223758\\0.07400236\\0.07608884\\0.07842325\\0.08358494\\0.08907398\\0.09997526\\0.10980066\\0.11801708\\0.12447086\\0.12919557\\0.13232266\\0.13403728\\0.1292663\\0.12042292\\0.12031612\end{bmatrix}\right)$$

The choice of an inverse-gamma distribution is natural for modeling variances under the assumption of normally distributed residuals, as it serves as the conjugate prior in this setting. The shape parameter $\alpha = 18$ of the Inverse-Gamma distribution was selected based on the assumption of assigning one degree of freedom per year of data in the historical sample (1983–2001). The scale parameters β_i correspond to the squared residuals that would be obtained under a random walk specification for each maturity. This is a reasonable choice, as discussed in Guidolin (2019) [6], where it is shown that autoregressive models yield only marginal gains in forecast performance. More substantial improvements are observed when factor structures are incorporated, as highlighted in Table 8 of his article.

The priors and their corresponding distributions were carefully chosen because the problem involves over 3000 parameters to be optimized. Given the complexity and the large number of parameters, it is crucial to have well-defined priors to ensure better convergence speed for the problem. By selecting informative priors, we aim to guide the optimization process more effectively, reducing the chances of the model diverge.

2. Propose a MCMC sampler for the model parameters.

The sampling strategy for estimating the model parameters is divided into Five components, which are discussed in detail in the following subsections. The overall approach combines elements of the Metropolis, Normal-Inverse Wishart distribution, Normal, Random-Walk, Forward Filtering Backward Sampling (FFBS), and the Gamma distribution.

2.1. Forward Filtering Backward Sampling - FFBS

Given the following objects obtained from the Kalman filter and smoother (as shown in Lecture 6, page 27 [8]) from Carter (1994) [1]:

Let $\beta_{i|i}$, $S_{i|i}$, $\beta_{i|T}$, and $S_{i|T}$ denote, respectively, the filtered means, filtered covariances, smoothed means, and smoothed covariances obtained externally via a Kalman filter and smoother (e.g., using `kf.filter(Y)` and `kf.smooth(Y)`).

Initialize $[\tilde{\beta}_1, \dots, \tilde{\beta}_T]$ as empty
$\tilde{\beta}_T \sim \mathcal{N}(\beta_{T|T}, S_{T|T})$ # sample last state from smoother
For $i = T-1, T-2, \dots, 1$:

$S_{i+1|i} = \Phi S_{i|i} \Phi^\top + Q$ # predictive variance of β_{i+1}
$\beta_{i+1|i} = \mu + \Phi(\beta_{i|i} - \mu)$ # predictive mean of β_{i+1}
$m_i = \beta_{i|i} + S_{i|i} \Phi^\top S_{i+1|i}^{-1} (\tilde{\beta}_{i+1} - \beta_{i+1|i})$ # conditional mean
$v_i = S_{i|i} - S_{i|i} \Phi^\top S_{i+1|i}^{-1} \Phi S_{i|i}$ # conditional variance
$\tilde{\beta}_i \sim \mathcal{N}(m_i, v_i)$ # sample β_i given $\tilde{\beta}_{i+1}$

The resulting trajectory $\tilde{\beta}_{1:T}$ is a full draw from the posterior distribution $p(\beta_{1:T} \mid Y_{1:T})$, conditional on the parameters Φ, Q, H, μ and $\Lambda(\lambda)$.

Below, to better illustrate the concept, we show the distribution of simulations given an arbitrary set of initial parameters:

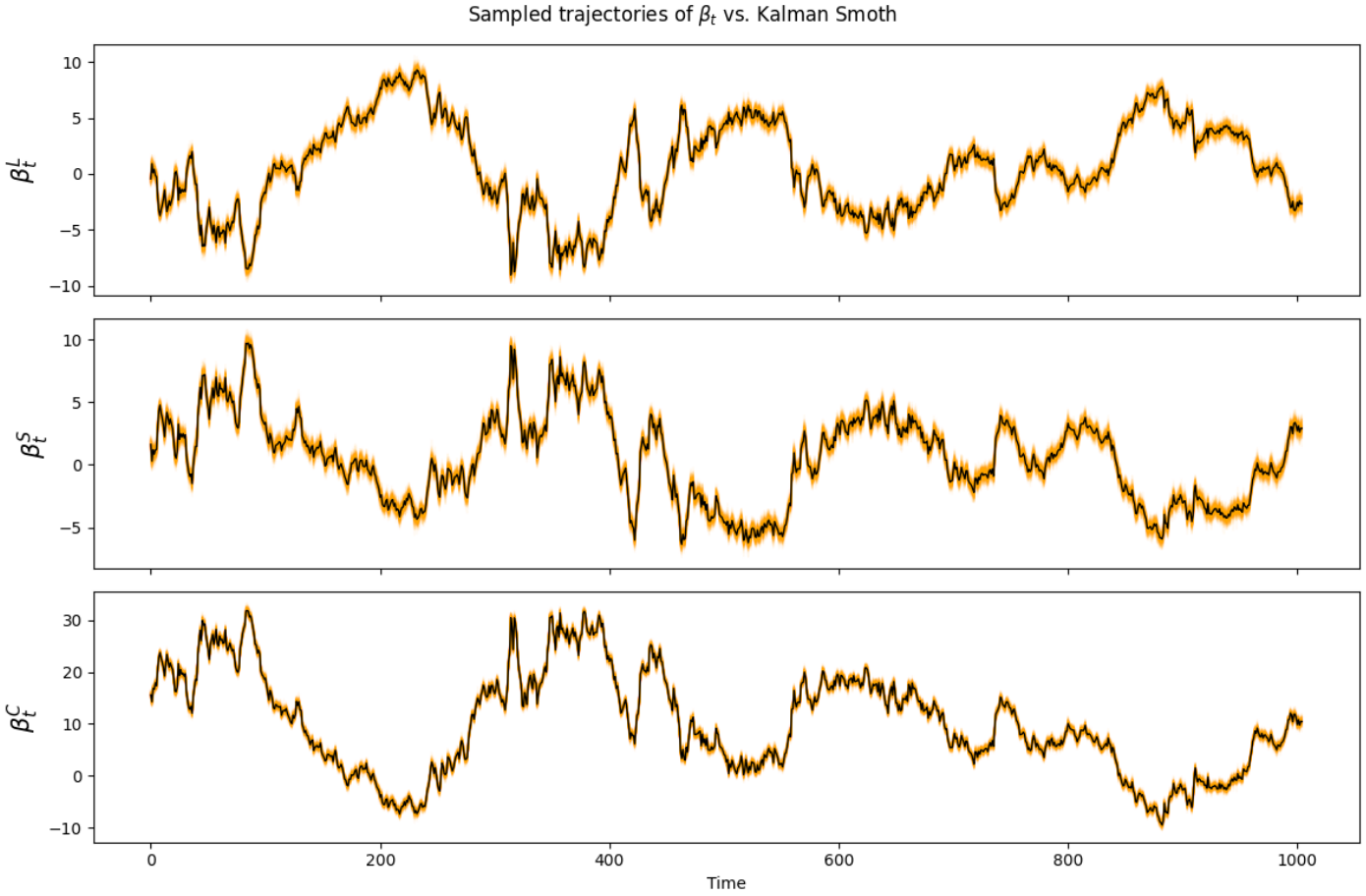

Figure 1. Sample Beta distribution example

2.2. Normal sampler - For μ

From the transition equation for the state vector in the Bayesian Dynamic Nelson-Siegel model, this equations are from Hoof Chapter 3 [4]:

$$\beta_t = \mu + \Phi\beta_{t-1} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, Q)$$

Define the transformed variable:

$$\Delta_t := \beta_t - \Phi\beta_{t-1}$$

Then, we can write:

$$\Delta_t = \mu + \eta_t \quad \Rightarrow \quad \Delta_t \sim \mathcal{N}(\mu, Q)$$

Assuming a conjugate prior:

$$\mu \sim \mathcal{N}(m_0, V_0)$$

The posterior distribution for μ conditional on the full sequence $\{\beta_t\}_{t=1}^T$, as well as Φ and Q, is:

$$V_n = \left(V_0^{-1} + (T-1)Q^{-1}\right)^{-1} \quad \text{and} \quad m_n = V_n\left(V_0^{-1}m_0 + Q^{-1}\sum_{t=2}^{T}\Delta_t\right)$$

Hence:

$$\mu \mid \beta, \Phi, Q \sim \mathcal{N}(m_n, V_n)$$

$m_0 = \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix}$ `# Prior mean vector`
$V_0 = \begin{bmatrix} 9 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 25 \end{bmatrix}$ `# Prior covariance matrix (i.e., variances:` $3^2, 2^2, 5^2$`)`
$\Delta_t = \beta_t - \Phi\beta_{t-1}$ `# Compute residuals for` $t = 2, \dots, T$
$\bar{\Delta} = \sum_{t=2}^{T}\Delta_t$ `# Sum of residuals`
$V_n = \left(V_0^{-1} + (T-1)Q^{-1}\right)^{-1}$ `# Posterior covariance`
$m_n = V_n\left(V_0^{-1}m_0 + Q^{-1}\bar{\Delta}\right)$ `# Posterior mean`
$\mu^{(i)} \sim \mathcal{N}(m_n, V_n)$ `# Sample from posterior`

2.3. Normal-Wishart Φ and Q sampler - BVAR

To sample Φ and Q, we adopt a Bayesian VAR (BVAR) framework conditional on the previously sampled values of β and μ. This allows us to rely on the conjugate Normal-Inverse Wishart structure, as presented in Chapter 6 of Koop (2003) [2] and on slide 20 of Lecture 5 [7]. The joint sampling is performed according to the algorithm below.

$\tilde{\beta}_t = \beta_t - \mu$ # normalization
$X = [\tilde{\beta}_1^\top, \dots, \tilde{\beta}_{T-1}^\top]^\top$ # matrix X
$Y = [\tilde{\beta}_2^\top, \dots, \tilde{\beta}_T^\top]^\top$ # matrix Y
$V_n^{-1} = V_0^{-1} + X^\top X$ # posterior precision
$V_n = (V_0^{-1} + X^\top X)^{-1}$ # posterior cov
$A_n = V_n\left(V_0^{-1}\Phi_0 + X^\top Y\right)$ # mean posterior Φ
$S_n = S_0 + Y^\top X(X^\top X)^{-1}X^\top Y + A_0^\top V_0^{-1}A_0 + A_n^\top V_n^{-1}A_n$ # scale posterior
$\nu_n = \nu_0 + T - 1$ # degree of feedrom
$Q^{(i)} \sim \mathcal{IW}(\nu_n, S_n)$ # Sample Q
repeat
  $\Phi^{(i)} \sim \mathcal{MN}(A_n,\ Q^{(i)},\ V_n)$ # Sample Φ
until max $|\lambda_j(\Phi^{(i)})| < 1$ # stationary
  or iterations = 10

2.4. Posterior Sampling for H - Inversa Gamma

Given the observation equation in the Bayesian Dynamic Nelson-Siegel model:

$$y_t = \Lambda_t\beta_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, H)$$

we assume that the covariance matrix H is diagonal and that each variance element $h_j = H_{jj}$ follows an independent inverse-gamma prior:

$$h_j \sim \text{Inverse-Gamma}(\alpha, \beta_j)$$

where α is the shape parameter (common to all maturities) and β_j reflects the scale associated with the j-th maturity, calibrated from a benchmark model such as a univariate random walk. Conditioning on observed data $Y = [y_1, \dots, y_T]$, sampled states β_t, and known factor loadings Λ_t, the posterior distribution for each h_j is independent and given by:

$$h_j \mid Y, \beta, \Lambda \sim \text{Inverse-Gamma}\left(\alpha + \frac{T}{2},\ \beta_j + \frac{1}{2}\sum_{t=1}^{T}\left(y_{tj} - \Lambda_t^{(j)}\beta_t\right)^2\right)$$

where $\Lambda_t^{(j)}$ denotes the j-th row of the factor loading matrix at time t, and β_t is the latent state vector at time t. This formulation ensures that each variance component is informed by the squared residuals at its respective maturity over the entire sample period.

The following algorithm outlines the sampling procedure for the full diagonal matrix H:

$e_t = y_t - \Lambda_t \beta_t$ for $t = 1, \dots, T$ #Compute residuals
$E = [e_1^\top, \dots, e_T^\top]^\top \in \mathbb{R}^{T \times 20}$ # Stack all residuals
For each maturity $j = 1, \dots, 20$:

$e^{(j)} = E_{:,j}$ # extract residuals
Compute posterior parameters:
$\tilde{\alpha}_j = \alpha + \frac{T}{2}$
$\tilde{\beta}_j = \beta_j + \frac{1}{2} \sum_{t=1}^{T} (e_{tj})^2$
Sample: $h_j \sim$ Inverse-Gamma$(\tilde{\alpha}_j, \tilde{\beta}_j)$

Form the diagonal matrix: $H = \text{diag}(h_1, \dots, h_{20})$

2.5. Posterior Sampling for λ - Adaptative Random -Walk Metropolis Hasting

The log-likelihood and log-prior are used in this Metropolis-Hastings sampler to improve numerical stability. In practice, libraries such as `numpy` or `scipy` in Python do not handle very small floating-point numbers well, especially when computing likelihoods over many observations. Working in log scale prevents underflow and allows for more accurate and stable computation of acceptance probabilities.

The proposal scale κ is updated adaptively every $\Xi = 500$ iterations using the rule:

$$\kappa_{\text{new}} = \kappa_{\text{old}} \cdot r_t + \frac{2 \cdot \kappa_{\text{old}}}{3}$$

where r_t is the average acceptance rate observed over the last 50 steps. This rule is based on the assumption that the expected acceptance rate $\mathbb{E}[\alpha]$ is a monotonic function of κ. Under this assumption, the update function is guaranteed to converge to the target acceptance rate of $1/3$. This behavior was previously analyzed and validated in the solution to Problem Set 1, where this same adaptive strategy was applied successfully.

Furthermore, based on the insights from the earlier problem set, we know that a faster adaptation of κ can help accelerate convergence to the stationary distribution. For this reason, we adopt a relatively small update window ($\Xi = 500$), which allows the algorithm to respond more quickly to suboptimal acceptance behavior and adjust the proposal variance accordingly.

Initialization
Set initial value $\lambda_0 > 0$, initial scale $\kappa_0 > 0$, and acceptance counter $\alpha_i = 0$
Set prior: $\lambda \sim$ Gamma$(\alpha = 59,\ \beta = 0.25)$
For each iteration $t = 1, \dots, N$:

1. Proposal step
Propose: $\lambda^\star \sim |\mathcal{N}(\lambda_{t-1}, \kappa_{t-1}^2)|$

2. Compute log-likelihood
For each $t = 1, \dots, T$, compute $\Lambda_t(\lambda)$ based on τ
$e_t = y_t - \Lambda(\lambda)\beta_t$ # Compute residuals
$\log \mathcal{L}(\lambda) = -\frac{1}{2} \sum_{t=1}^{T} e_t^\top H^{-1} e_t$ # Compute log likelihood

3. Compute log-prior
$\log p(\lambda) = (\alpha - 1) \log \lambda - \beta\lambda$

4. Compute log acceptance ratio
$\log \alpha = [\log \mathcal{L}(\lambda^\star) + \log p(\lambda^\star)] - [\log \mathcal{L}(\lambda_{t-1}) + \log p(\lambda_{t-1})]$

5. Accept or reject
Sample $u \sim \mathcal{U}(0, 1)$
If $\log u < \log \alpha$:
    $\lambda_t = \lambda^\star$
Else:
    $\lambda_t = \lambda_{t-1}$

6. Adaptive tuning of κ every Ξ iterations
suppose $\Xi = 500$
If $mod(t, \Xi) = 0$:
    $r_t = \frac{1}{\Xi} \sum_{i=t+1-\Xi}^{t} \alpha_i$ # Compute recent acceptance rate
    $\kappa_{\text{new}} = \kappa_{\text{old}} \cdot r_t + \frac{2 \cdot \kappa_{\text{old}}}{3}$ # Update proposal scale

return $\lambda_t, \alpha_t, \kappa_t$ # Return parameters

2.6. Gibbs Sample with Metropolis

:

Once all conditional samplers have been defined, we proceed with the Gibbs sampling procedure. The initial values for the parameters are chosen based on Ordinary Least Squares (OLS) estimations:

Initial $\beta = \beta^{\text{OLS}}$
Initial $\Phi = \Phi^{\text{OLS}}$
Initial $\mu = \mu^{\text{OLS}}$
Initial Q computed from residuals of OLS β and Φ
Initial $H = H^{\text{observed}}$
Initial $\lambda = 124.75$

These choices are motivated by the high dimensionality of the model. Since Bayesian estimation should not diverge significantly from the frequentist point estimates (especially in small sample settings), we opt to initialize the chain near a region of likely convergence.

During the burn-in phase, we iteratively sample each parameter and store the outputs in a structured matrix containing:

$$[\text{iteration}],\ \theta,\ \alpha,\ \kappa$$

We monitor the stability of the adaptive proposal scale κ using stoping rule for alpha.

Once a stable $\kappa^\star$ is found, we fix it and execute the full Gibbs sampler in parallel using 10 cores. Each core independently performs full Gibbs iterations (now with fixed $\kappa = \kappa^\star$), allowing us to collect multiple posterior trajectories for inference and uncertainty quantification.

The algoritm:

THETA matrix ← empty # Stores all sampled parameters
kappa matrix ← empty # Stores all kappa values
alpha matrix ← empty # Stores all acceptance rates
$\Xi = 500$ # Window size for adaptation
for $t \in 1 : 10\,000$): # Main Gibbs-MH loop

$\beta_i = \text{sampler\_beta}(Y, \Phi_i, \Lambda_i, Q_i, H_i, \mu_i;\ \text{prior} = (\beta_0, Q_0))$ # FFBS
$\mu_i = \text{sampler\_mu}(\beta_i, \Phi_i, Q_i;\ \text{prior} = (\mu_0, Q_0))$ # Normal conjugate
$\Phi_i, Q_i = \text{sampler\_PhiQ}(\beta_i, \mu_i;\ \text{prior} = (\Phi_0, Q_0))$ # Normal-Wishart
$H_i = \text{sampler\_H}(Y, \beta_i, \Lambda_i;\ \text{prior} = (\alpha_\Gamma, \beta_\Gamma))$ # Inverse Gamma prior
$\lambda_i, \alpha_i, \kappa_i = \text{sampler\_lambda}(Y, \lambda_{i-1}, \beta_i, H_i;\ \text{prior} = (\alpha_\gamma, \beta_\gamma),\ \text{aux} = (t, \Xi),\ \text{history} = (\text{alpha}, \text{kappa}),\ \text{signal} = \text{False})$

THETA matrix ← append row # Save current parameter draw
alpha matrix ← append row # Save current acceptance rate
kappa matrix ← append row

if len(alpha) > 10000: # Start checking convergence after burn-in

MA_alpha = rolling_mean(alpha, Ξ)
max_alpha = max(alpha[$-5000 :$])
min_alpha = min(alpha[$-5000 :$])
if max_alpha < 0.4 and min_alpha > 0.3: # Acceptance rate is stable
break # Stop adaptation

$\beta,\ \mu,\ \Phi,\ H,\ \lambda = \arg\max_{\theta \in \text{THETA}[-5000:]} \hat{\text{post}}\text{erior}(\theta)$ # Select MAP after adaptation
for CPU **in** computer:

START Gibbs Sampler with signal = True, κ = mean(kappa[-500 // Ξ :]) # Run final chains with fixed kappa

That runs in my PC with 22 threads in parallel

3. Implement the sampler and show its diagnostics

Enquanto ajustava as funções no *Python*, observei que o parâmetro λ convergia para a casa dos 120, enquanto κ convergia para aproximadamente 0,45. Antes mesmo de rodar a versão final (*prime*) do meu algoritmo MCMC, eu já tinha uma boa ideia dos valores iniciais a serem atribuídos a esses parâmetros, o que me ajudou a poupar tempo de computação — especialmente considerando que, neste caso, temos cerca de 3.000 parâmetros, e o tempo de convergência sempre foi uma preocupação central.

Executei 10.000 iterações com boas aproximações iniciais para os parâmetros λ e κ. Os demais palpites iniciais foram obtidos a partir de estimativas por Mínimos Quadrados Ordinários (OLS), conforme mencionado anteriormente. É importante frisar que não utilizei o OLS para definir as prioris, e sim como aproximação inicial próxima da posteriori, com o intuito de acelerar a convergência.

Figure 2. Convergência dos parâmetros nas primeiras 10000 iterações, lembrando que ao longo da manutenção dos códigos os palpites inicias foram sendo ajustados

Visualmente, percebi que a cadeia já havia convergido nas últimas 5.000 iterações.

Salvei a média de $\hat{\theta}$ utilizando esse intervalo final de 5.000 iterações, incluindo o parâmetro κ (que é argumento da função de λ).

Com esses bons palpites iniciais de θ_0 e κ, rodei 10 cadeias MCMC paralelas. Cada cadeia foi executada com um máximo de 100.000 iterações, sendo os resultados parciais armazenados a cada mil iterações, com a memória sendo liberada após cada salvamento. Essa estratégia foi adotada por três motivos:

1. Meu computador possui pouca memória RAM, o que poderia levar o processo a travar.
2. O armazenamento frequente garante que os resultados estejam preservados mesmo em caso de falha.
3. Essa estrutura me permite continuar outras tarefas sem aguardar a conclusão total do processo — basta carregar os resultados parciais em outro terminal para não perder o dia.

Foi verificado se o valor de máxima a posteriori (MAP), calculado após as 10.000 iterações, estava próximo dos valores iniciais utilizados. De modo geral, os resultados confirmaram essa proximidade, com exceção da matriz Q. No entanto, observou-se que essa matriz convergiu rapidamente, estabilizando-se em menos de 1.000 iterações. Para os demais parâmetros, cujos valores iniciais foram obtidos via estimativas por Mínimos Quadrados Ordinários (OLS), todos permaneceram dentro de um intervalo de um desvio padrão em relação ao MAP observado.

É natural que as estimativas obtidas por Mínimos Quadrados Ordinários (OLS) diferenciem-se do valor de máxima a posteriori (MAP), por dois motivos principais. Primeiro, o MAP incorpora explicitamente a informação a priori, o que não ocorre no OLS. Segundo, as estimativas por OLS foram obtidas condicionalmente ao vetor β_{ols}, ou seja, os demais parâmetros foram otimizados a partir dessa condição. Já o MAP resulta de uma solução conjunta, que pondera simultaneamente a adequação da matriz Φ e a minimização da matriz H, normalizando os resíduos para cada variância esperada associada a cada H_τ.

Nas próximas páginas é apresentada as distribuições dos parâmetros estimados.

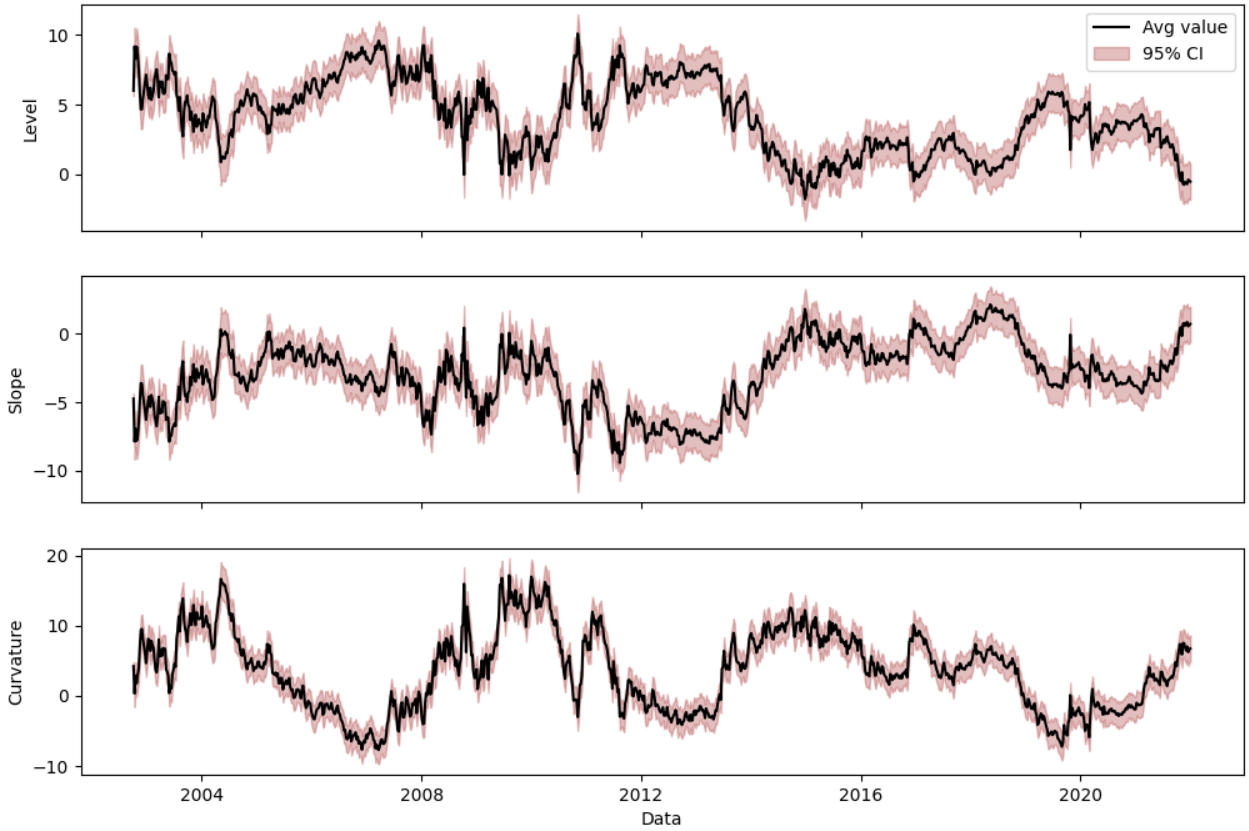

Figure 3. Betas a posteriori e seus intervalos de credibilidade

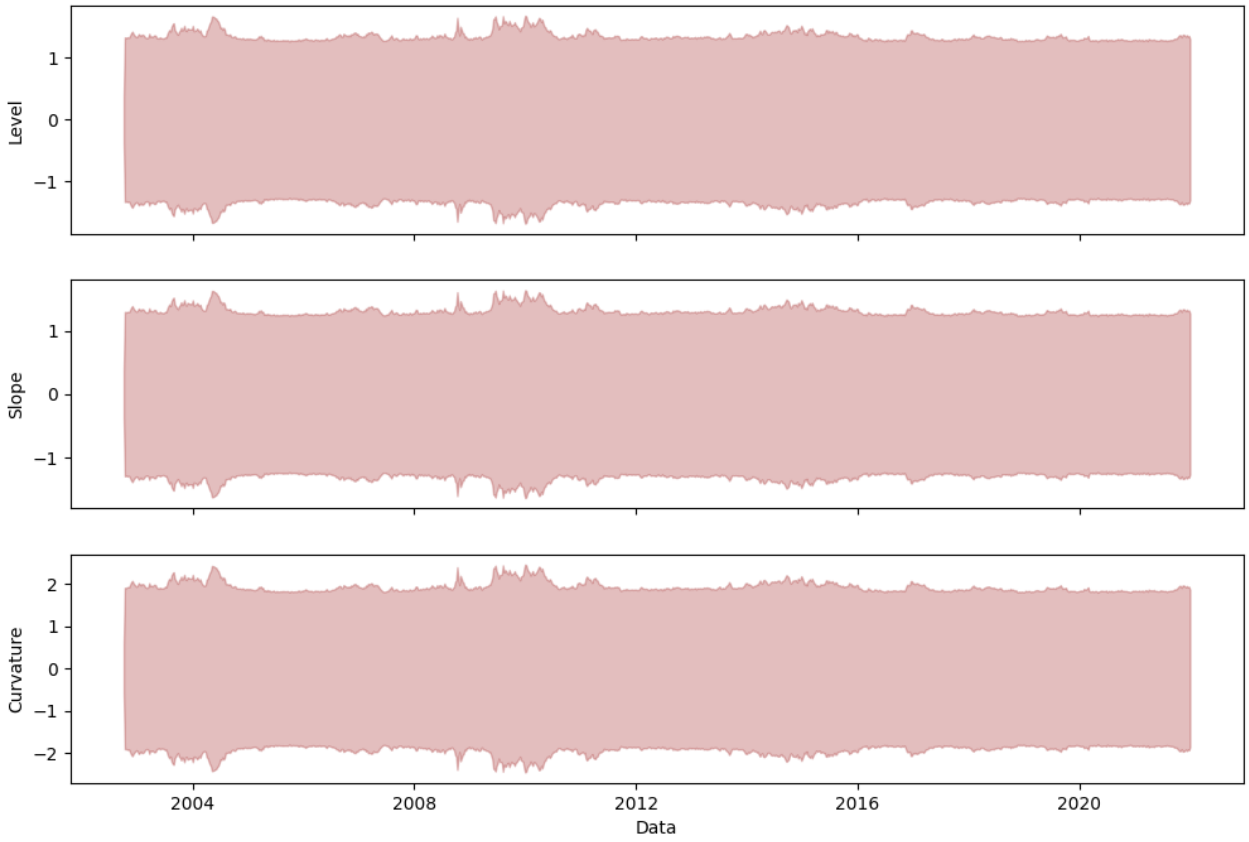

Figure 4. Tamanho do intervalo de confiança para cada beta a posteriori

Perceba, no gráfico abaixo, que os termos da diagonal da matriz Φ (A) apresentam maior magnitude, conforme previsto pela distribuição a priori adotada — isto é, uma matriz próxima da identidade, com os elementos fora da diagonal próximos de zero. Um ponto que chama atenção é que a curvatura é explicada predominantemente por sua própria defasagem, sugerindo forte autocorrelação estrutural nesse componente. Além disso, observa-se que o fator *Level* se mostra como o mais estável ao longo das iterações. Uma hipótese plausível é que este fator é mais associado aos vértices de menor maturidade, os quais tendem a ser mais sensíveis às condições atuais do mercado e, portanto os agentes possuem mais informação.

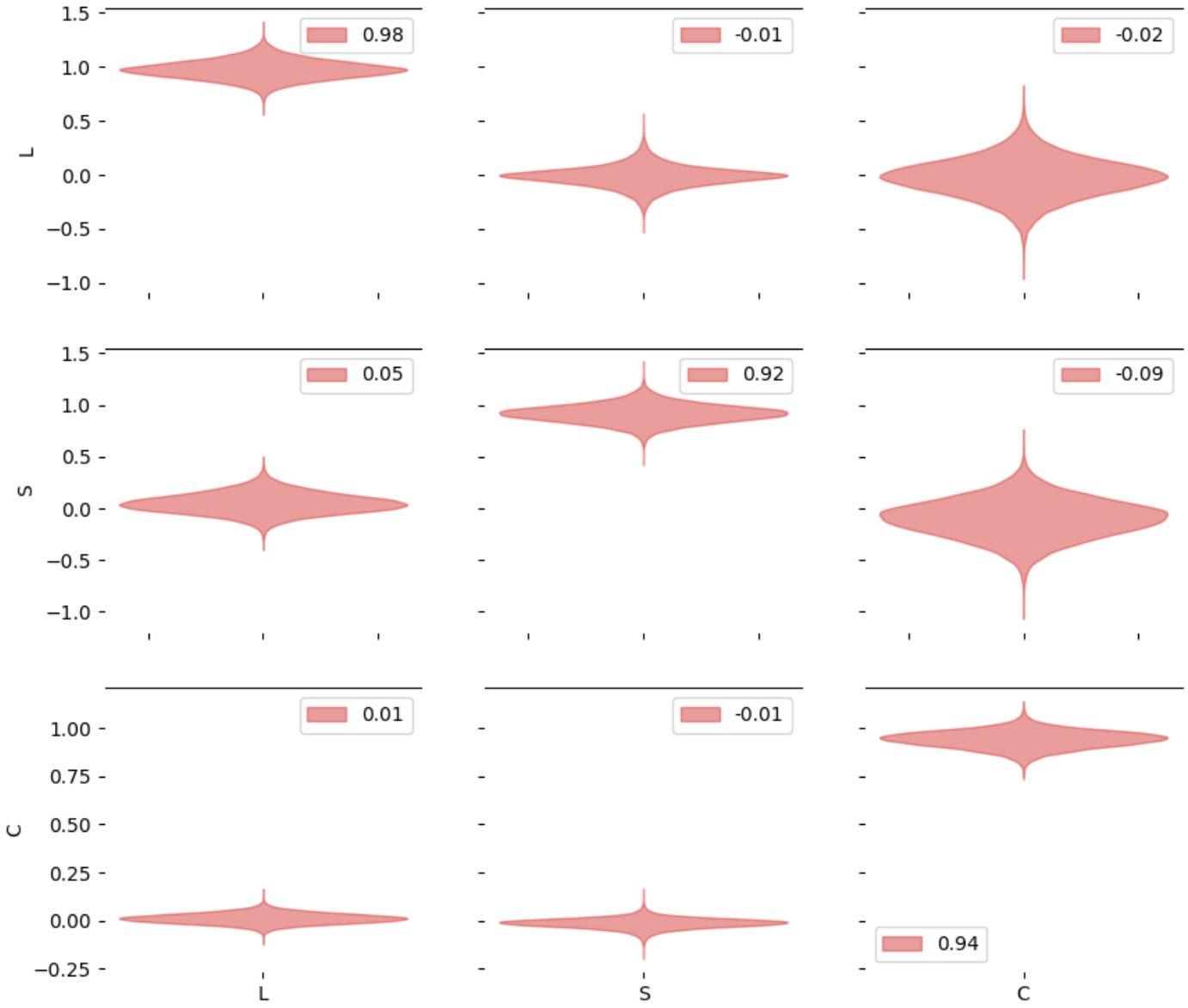

Figure 5. Violin plot para cada termo da matriz de transição A

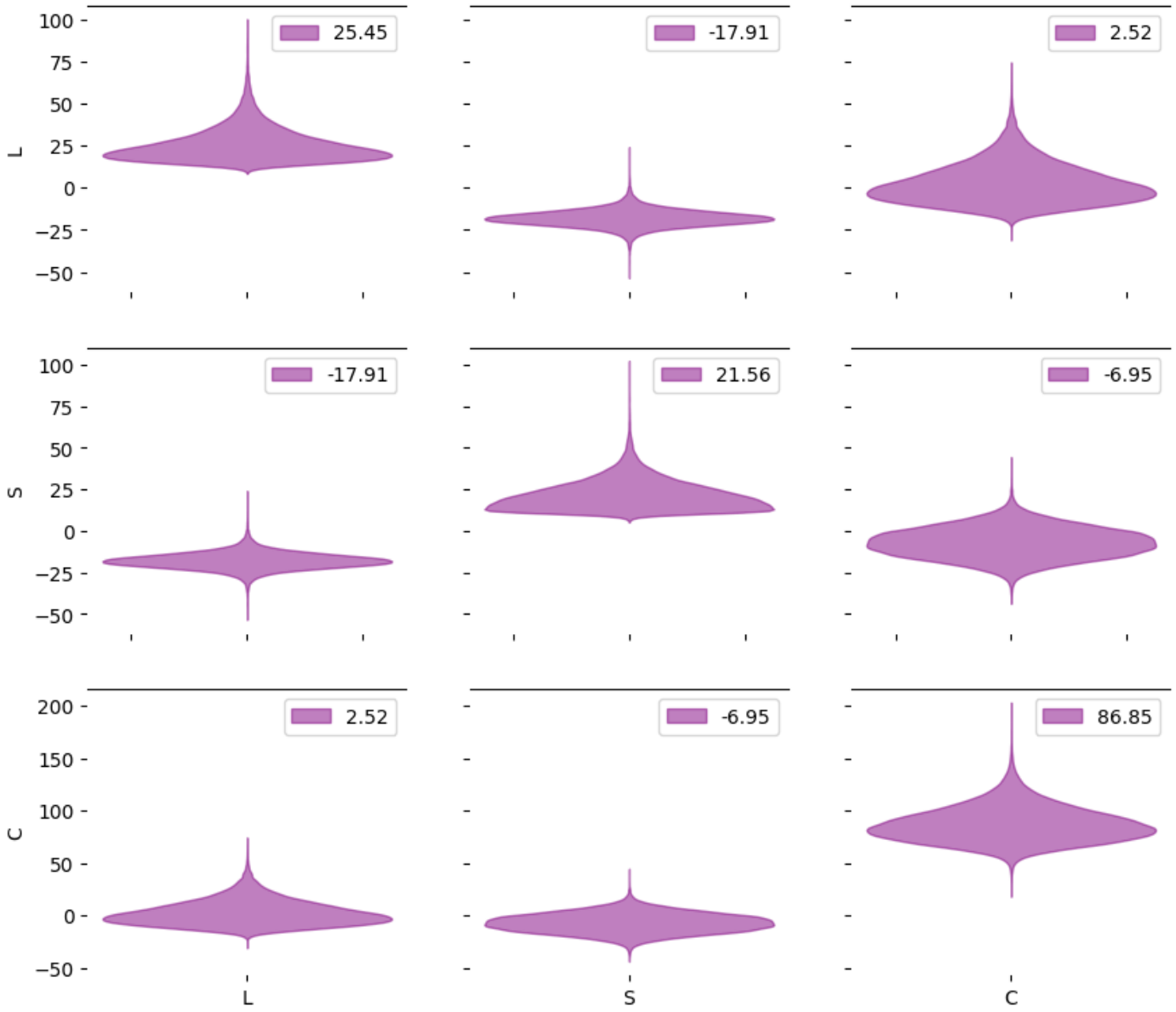

Figure 6. Violin plot para cada termo da matriz de transição A

À primeira vista, o gráfico acima pode sugerir a presença de valores anômalos ou fora do esperado. No entanto, é importante destacar que a matriz V_n, que acompanha esse gráfico, apresenta elementos com magnitudes extremamente pequenas, como ilustrado a seguir:

$$V_n = \begin{bmatrix} 2{,}89 \times 10^{-4} & 3{,}38 \times 10^{-4} & 7{,}55 \times 10^{-6} \\ 3{,}38 \times 10^{-4} & 4{,}65 \times 10^{-4} & 1{,}73 \times 10^{-5} \\ 7{,}55 \times 10^{-6} & 1{,}73 \times 10^{-5} & 2{,}43 \times 10^{-5} \end{bmatrix}$$

Dessa forma, ao considerar a multiplicação da matriz $Q \otimes V_n$, obtém-se valores mais razoáveis.

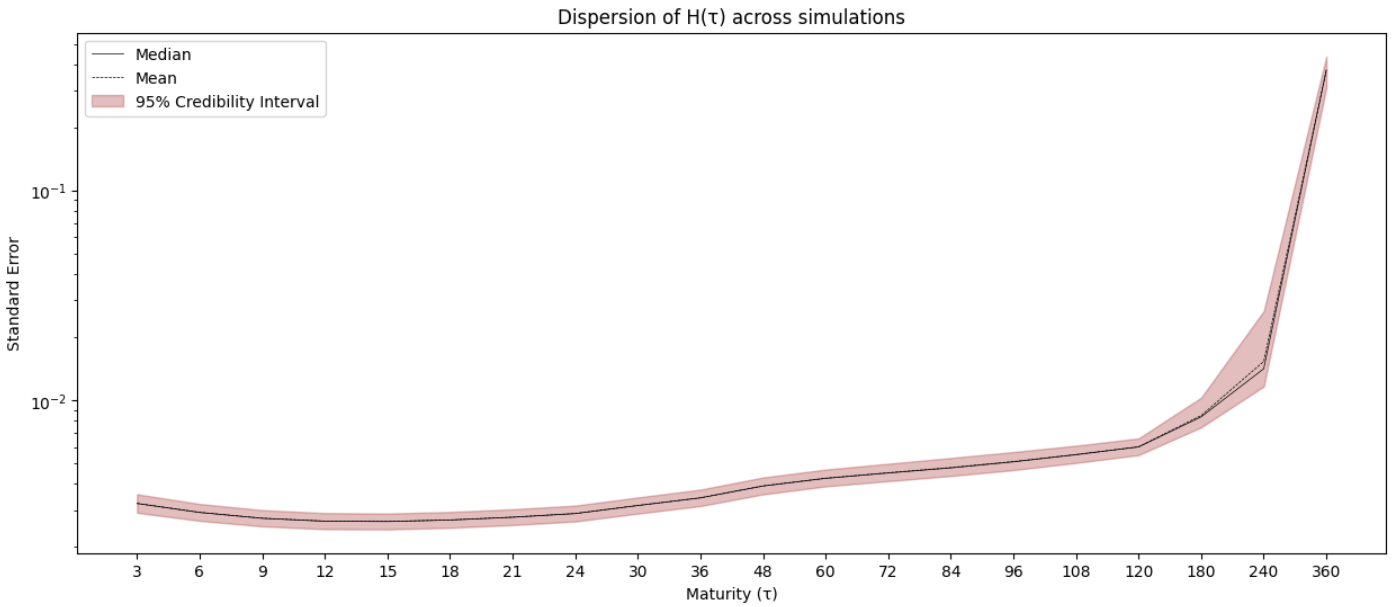

Figure 7. H posteriori para cada Tau

No gráfico acima é possível observar que o resíduo quadrático médio é maior para maturidades maiores, a ordinariedade de h era esperada pela priori, mas as ordens de grandeza são bem diferentes.

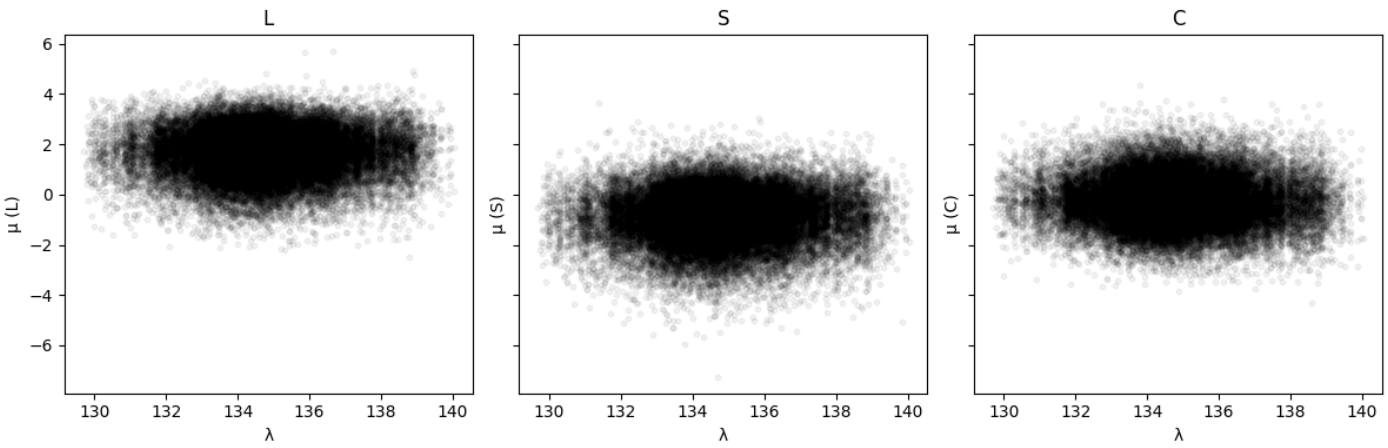

Figure 8. Relação entre μ e a variável λ posteriori

Da para observar que o lambda não tem efeito no μ, o que é esperado, principalmente pela ordem de grandeza do λ, a derivada de $\Lambda(\tau)$ para valores muito grandes de λ tende a ser constante, ou seja, quase uma linearização.

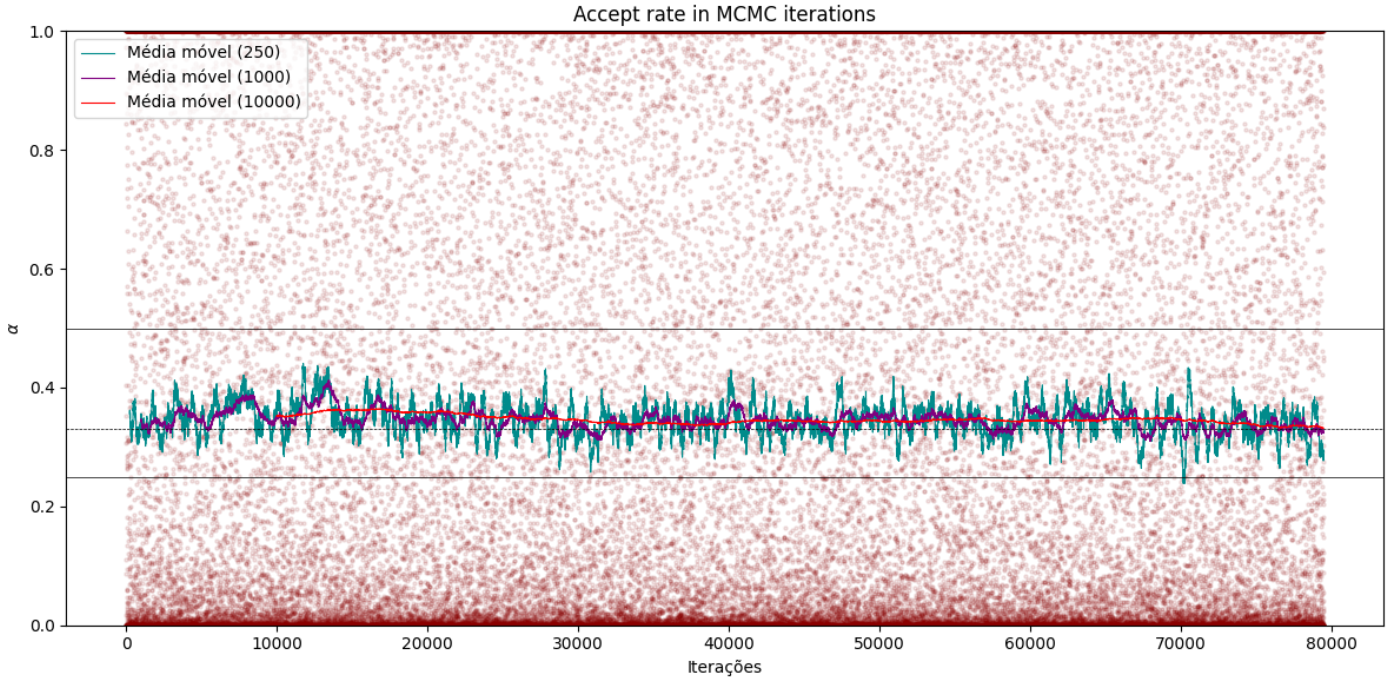

Figure 9. Taxa média de aceitação nas iterações

O gráfico acima mostra que as taxas de aceitação médias permanecem, em geral, dentro da faixa entre 0,25 e 0,50. Quando utilizamos janelas de médias móveis maiores, observa-se uma estabilização mais clara em torno do intervalo entre 0,30 e 0,40, o que está em conformidade com o nosso objetivo de manter a taxa de aceitação próxima de 0,33. Esse comportamento indica um bom ajuste da escala de proposta ao longo das iterações, contribuindo para uma exploração eficiente do espaço amostral.

4. Plot the posterior density of the elements of the autoregressive matrix A

$$cov(vec(\Phi^{\top})) = \begin{bmatrix} \overset{\text{Var}(\phi_{LL})}{\boxed{0.00785}} & -0.00747 & -0.00084 & \overset{\text{Cov}(\phi_{LL},\phi_{SL})}{\boxed{0.00763}} & -0.00647 & -0.00010 & 0.00066 & -0.00084 & -0.00044 \\ -0.00747 & 0.00987 & -0.00345 & -0.00699 & 0.00793 & -0.00280 & -0.00076 & 0.00138 & -0.00048 \\ -0.00084 & -0.00345 & 0.03379 & 0.00015 & -0.00281 & 0.02935 & -0.00105 & -0.00050 & 0.00358 \\ 0.00763 & -0.00699 & 0.00015 & 0.01018 & -0.00775 & 0.00058 & 0.00048 & -0.00072 & -0.00023 \\ -0.00647 & 0.00793 & -0.00281 & -0.00775 & 0.00828 & -0.00235 & -0.00064 & 0.00109 & -0.00053 \\ -0.00010 & -0.00280 & 0.02935 & 0.00058 & -0.00235 & 0.03541 & -0.00043 & -0.00093 & 0.00286 \\ 0.00066 & -0.00076 & -0.00105 & 0.00048 & -0.00064 & -0.00043 & 0.00072 & -0.00055 & -0.00016 \\ -0.00084 & 0.00138 & -0.00050 & -0.00072 & 0.00109 & -0.00093 & -0.00055 & 0.00076 & -0.00011 \\ -0.00044 & -0.00048 & 0.00358 & -0.00023 & -0.00053 & 0.00286 & -0.00016 & -0.00011 & 0.00208 \end{bmatrix}$$

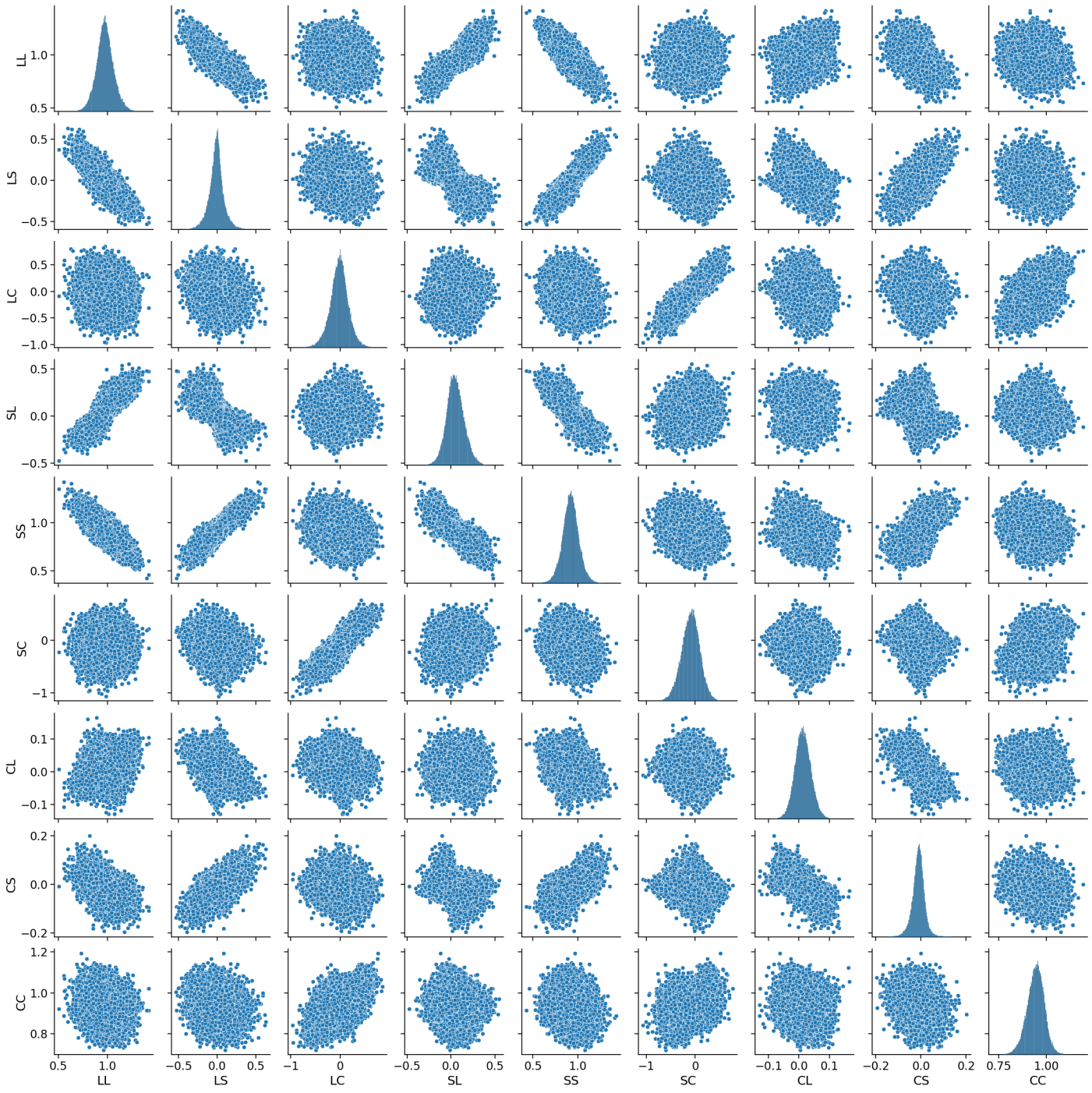

Figure 10. Grágico de distribuição conjunta da matriz A

Nota-se que as variáveis associadas à explicação do nível (*Level* – L) e da inclinação (*Slope* – S) apresentam boa correlação entre si, sugerindo uma estrutura dinâmica compartilhada. Por outro lado, os termos relacionados à curvatura (*Curvature* – C) demonstram baixa correlação tanto com os fatores L e S quanto entre si, indicando uma maior autonomia ou ruído no processo que governa sua evolução ao longo do tempo.

5. Plot a predictive curve with respective predictive intervals for maturities from 3 to 360 months

No gráfico abaixo é possível observar a distribuição de trajetórias previstas para o ano de 2021, bem como os valores efetivamente realizados. Conforme esperado pelo modelo, as variações para prazos mais longos são mais acentuadas — efeito que é majoritariamente capturado pela matriz H.

Acredito que a escolha de uma prior condicional para A/Φ dada Q tenha contribuído para facilitar a formulação das equações, mas prejudicado a estimação de Q, que acabou sendo amostrado com valores extremamente elevados — aproximadamente 10 vezes superiores à covariância observada no residuo de $\hat{\beta}_{t|t-1} - \beta_t$ (i.e., Cov(η)).

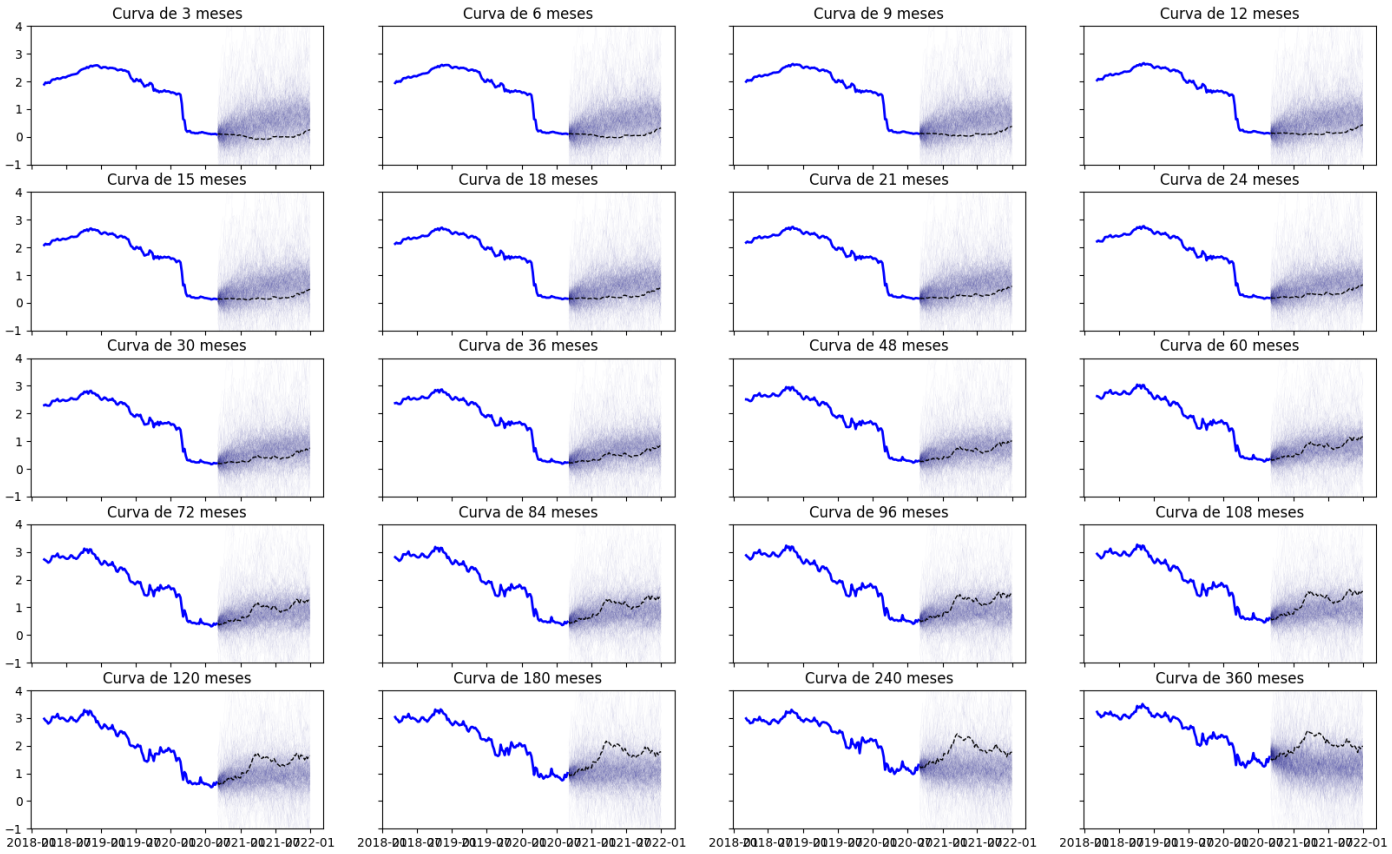

Figure 11. Curva de previsão do modelo para os vértices entre 3 e 360 meses.

6. Conclusão

Este trabalho permitiu compreender de forma mais profunda as consequências práticas das escolhas de prioris e a dinâmica dos algoritmos MCMC. Implementar manualmente o algoritmo foi fundamental para interpretar os resultados gerados, especialmente ao observar que a matriz Q foi superestimada. Esse problema decorre, provavelmente, da escolha de uma prior condicional para Φ em função de Q, o que, apesar de facilitar a implementação matemática, introduziu uma dominância excessiva de Q sobre a variância total do modelo.

Como consequência, o estimador de Q apresentou oscilações maiores que o desejado, enquanto os valores atribuídos à matriz H foram reduzidos. Isso ocorre porque, uma vez que a variância total do erro é limitada, um aumento em Q exige uma redução proporcional em H, criando um desequilíbrio indesejado na decomposição da incerteza.

Durante o processo de estimação, diversos erros foram cometidos e posteriormente ajustados. A escolha do parâmetro Ξ mostrou-se sensível: valores muito altos tornaram o processo ineficiente, enquanto valores muito baixos geraram oscilações desnecessárias em κ.

Por fim, ressalta-se que trabalhar com as probabilidades posteriores e de máxima verossimilhança no formato logarítmico pode ser a única abordagem computacionalmente viável, dada a complexidade e dimensionalidade do modelo.

É mais fácil assimilar a intuição das equações que definem o problema depois de estimar todas as distribuições posteriores delas e como elas se relacionam. Fora todo o desafio computacional que levou a necessidade de soluções relacionadas a Data Science para o desenvolver da tarefa

References

[1] C. Carter and R. Kohn, "On gibbs sampling for state space models", *Biometrika*, vol. 81, no. 3, pp. 541–553, 1994. DOI: 10.1093/biomet/81.3.541.

[2] G. Koop, *Bayesian Econometrics*. Chichester: John Wiley & Sons, 2003, ISBN: 9780470845677.

[3] F. X. Diebold and C. Li, "Forecasting the term structure of government bond yields", *Journal of Econometrics*, vol. 130, no. 2, pp. 337–364, 2006.

[4] P. D. Hoff, *A First Course in Bayesian Statistical Methods* (Springer Texts in Statistics). Springer, 2009, ISBN: 9780387922997. DOI: 10.1007/978-0-387-92407-6.

[5] W.-C. Yu and E. Zivot, "Forecasting the term structures of treasury and corporate yields using dynamic nelson-siegel models", *International Journal of Forecasting*, vol. 27, no. 2, pp. 579–591, 2011.

[6] M. Guidolin and M. Pedio, "Forecasting and trading monetary policy effects on the riskless yield curve with regime switching nelson-siegel models", *Journal of Economic Dynamics & Control*, vol. 107, p. 103 723, 2019. DOI: 10.1016/j.jedc.2019.103723. [Online]. Available: https://doi.org/10.1016/j.jedc.2019.103723.

[7] R. Duda, *Bayesian time series i*, https://eesp.fgv.br, Lecture slides, FGV/EESP – Bayesian Econometrics Series (Lecture 5), 2024.

[8] R. Duda, *Bayesian time series ii*, https://eesp.fgv.br, Lecture slides, FGV/EESP – Bayesian Econometrics Series (Lecture 6), 2024.