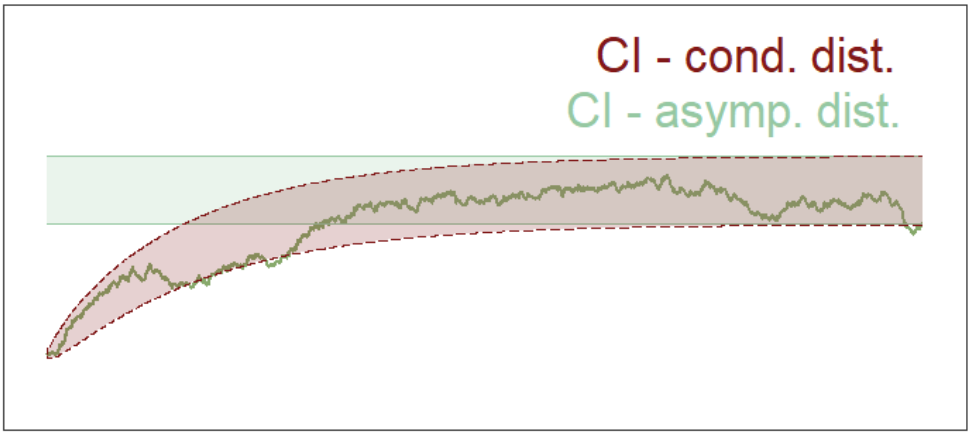
UNIVERSITY OF COPENHAGEN
DEPARTMENT OF MATHEMATICAL SCIENCES



Thesis Preparation Project

Youssef Raad
KU-id: zfw568

Regime-Switching in the CIR Model

Exploring an Autoregressive Hidden Markov Extension to the CIR Model

Date: 13-06-2025

Supervisor: Rolf Poulsen

Abstract

This thesis preparation project investigates extensions of one-factor short rate models using an Autoregressive Hidden Markov Model (ARHMM) framework. Specifically, we consider the Cox-Ingersoll-Ross (CIR) model. A ARHMM extension enables the model to account for structural shifts in interest rate behaviour across altering economic conditions. We outline the mathematical formulation of the CIR process, properties and limitations, particularly its numerical instability. A dissection of the CIR model is given to examine other estimation paths. Using U.S. 3-month Treasury bill data from 1984–2022, the study examines the in-sample performance of the CIR-ARHMM and discusses computational challenges of fitting. The main objectives are: 1) Examining the CIR process. 2) Checking for CIR-ARHMM robustness. Results and forecasts will mostly be saved for the thesis. The results indicate that, while the regime-switching extension enhances model flexibility, the CIR-ARHMM is sensitive to near-zero interest rate levels and suffers from numerical instability. The exact numerical instability is not measurable and the `R` built in solutions for the modified Bessel function of the first kind are not ripe for the task. Consequently, unless these issues can be mitigated, the CIR model may be unsuitable for further development within this framework. The findings support the use of simpler models like the Vašíček-ARHMM in the forthcoming thesis and highlight the need for robust approximation techniques.

Acknowledgements

I want to thank Rolf Poulsen, Professor of Mathematical Finance, University of Copenhagen, for his supervision during the preparation project. A special thanks to Théo Michelot, Assistant Professor of Statistics, Dalhousie University, for his help in applying the autoregressive hidden Markov model as a extension.

Note

Proofs/Derivations will be given in A.2 for most Theorems, Propositions and Lemmas. Some derivations will be given in the main text if they are meaningful for the further analysis of the paper and of short length.

ةزيزعلا يمأ ىلإ

CONTENTS

1 List of Symbols, Notation & Abbreviations

Symbol/Notation	Description
$\mathbb{P}$	Historical probability measure
$\mathbb{Q}$	Equivalent martingale measure
$\mathbb{Q}(\lambda)$	Some equivalent martingale measure specified by a fixed $\lambda \in \mathbb{R}$
$W_t^{\mathbb{Q}}$	Brownian motion under a measure (here the measure is exemplified with $\mathbb{Q}$)
Ω	Sample space
$\mathcal{F}$	Event space
$\{\mathcal{F}\}_{t\geq 0}$	Filtration
$(\Omega, \mathcal{F}, \{\mathcal{F}\}_{t\geq 0}, \mathbb{P})$	Filtered probability space
S_t	State occupied by Markov chain at time-t
p_i	(Conditional) density function in state i
$\boldsymbol{P}(r)$	Diagonal matrix with ith diagonal element p_i
I_N	$N \times N$-dimensional diagonal matrix with i element 1 (identity matrix)
r_t	Random variable at time-t
α_t	Forward probability, i.e. $\mathbb{P}(r_1, \ldots, r_t, S_t = i)$
$\boldsymbol{\alpha}_t$	(Row) vector of forward probabilities
$\boldsymbol{\Gamma}$	Transition probability matrix of a Markov chain
γ_{ij}	(i,j)'th element in $\boldsymbol{\Gamma}$; probability of transitioning from state i to state j in a Markov chain
$\boldsymbol{\delta}$	Stationary distribution of a Markov chain
$\mathbf{1}_N$	N-dimensional vector of 1's
$\mathbf{0}_N$	N-dimensional row vector of 0's
$\mathbf{1}_{N\times N}$	$N \times N$-dimensional matrix filled with 1's
T	Number of observations
N	Number of states
Δ	Time-increment between some observations r_t and r_s, $t \neq s$
$\mathcal{S}$	State space
$\mathcal{J}_\nu(x)$	Bessel function of the first kind of order ν evaluated at x
$\mathcal{I}_\nu(x)$	Modified Bessel function of the first kind of order ν evaluated at x
$\mathcal{I}(x)$	Fisher information matrix evaluated at x
H	Hessian matrix
$\text{SE}(\cdot)$	Standard Error of some estimator $\cdot$
$\mathcal{L}_T$	Likelihood function of T observations
ℓ_T	Log-likelihood function of T observations
$\xrightarrow{\mathcal{D}}$	Convergence in distribution
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$\mathcal{U}[a, b]$	Uniform distribution function over the range a to b
χ_k^2	χ-squared distribution with k degrees of freedom
$\chi_k'^2(\lambda)$	Non-central χ-squared distribution with k degrees of freedom and non-centrality parameter λ
$\Gamma(z)$	Gamma function evaluated at z
$\mathbb{1}_{\{A\}}$	Indicator function of some set A

Abbreviation	Description
CIR	Cox-Ingersoll-Ross
SDE	Stochastic differential equation
BESQ_x^δ	Bessel Square Root Process of dimension δ starting at x
HMM	Hidden Markov model
AR(p)	Autoregressive model of order p
ARHMM	Autoregressive hidden Markov model of order 1
t.p.m	Transition probability matrix
EMM	Equivalent martingale measure
AIC	Akaike information criterion
BIC	Bayesian information criterion
a.s.	Almost surely

2 Introduction

The ability to accurately model interest rate dynamics is crucial for both financial practitioners and researchers. One of such models, the Cox-Ingersoll-Ross (CIR) model, has long been a cornerstone for modelling such dynamics due to its ability to capture the mean-reverting behaviour of interest rates and non-negative rates under certain conditions. However, traditional models like the CIR model often struggle to fully account for abrupt shifts in market conditions or structural changes in the economy, which are key features of real-world financial systems. For example, one could easily imagine that the volatility and mean reverting parameters should vary across economic regimes. This limitation can lead to models that are overly simplistic or fail to adequately represent the complexities observed in actual interest rate behaviour.

In this paper, we propose an enhancement to the standard CIR model by integrating a regime-switching framework through the use of an Autoregressive Hidden Markov Model (ARHMM). The aim is to introduce multiple regimes within the CIR model, each with its own set of parameters, thereby capturing the shifts in market conditions that can significantly affect interest rate movements. By allowing the model to switch between different regimes, we enhance the flexibility and robustness of the CIR model, enabling it to better reflect the complex, dynamic nature of interest rate behaviour in financial markets.

While the underlying mathematical framework is inherently complex, this paper strives to bridge the gap between economic theory and quantitative modelling by firstly examining the possibility of using the extended CIR model in practice. By leveraging economic intuition, we aim to provide a deeper understanding of how different regimes might arise in financial markets, how to interpret regime-switching behaviour, and how to determine the optimal number of regimes a priori.

The methods, code and general framework that we develop in this paper can be extended to any process where a marginal or conditional density function is given for observations in the model. As such, it is not limited to interest rate modelling but could be extended to any model where an analytical density function is available.

3 Project Scope of Exploration

The paper serves as a "proof of concept" for the actual thesis after the summer. The focus will be on the application of the regime-switching autoregressive hidden Markov model framework when appended to one-factor short rate models. From grunt work done as preparation for the preparation project, it has become quite evident that the Vašíček model [43], where the solution to the stochastic differential equation is a Gaussian process, is well behaved when extended through the means of an autoregressive hidden Markov model. However, the CIR model [12], where the solution to the stochastic differential equation is a scaled non-central χ^2 process, is suspected to cause extreme numerical instability, under- and overflow errors when data isn't necessarily *well-behaved*. Indeed, the grunt work did show that the model could be fitted under nicely behaving parameter sub sets and rates larger than we usually observe, in a simulation study. As space is sparse it is excluded but code is provided for the keen reader. However, as rates usually concentrate around the origin, the non-centrality parameter will usually become extremely small, the degrees of freedom either extremely large or small. The modified Bessel function of the first kind in the non-central χ^2 distribution will at some points also have negative order which means a suboptimal approximation will be used[1]. Furthermore, the CIR model does yield problems when rates are non-positive.
As this thesis aims to explore the possibility of including the CIR model to the thesis analysis, whilst also outlining the mathematical framework for the model extension by the ARHMM for reusability, we will mark problem areas related to robustness for the CIR-ARHMM in the paper with a "$\triangledown$" for the discussion section of the paper.

Proceeding, modelling can be thought of as a two part problem, or rather, a two part process:

I. Understanding: How well does the model capture historical patterns during in-sample testing?

II. Forecasting: How well does the model perform on out-of-sample data?

By the nature of a **preparation** project, we limit ourself to point **I.**, which arguably makes the actual upcoming thesis a more interesting research paper. Furthermore, it makes us able to test if the CIR-ARHMM is suitable for such analysis of rates. We will not comment in depth on the actual results in aN interpretable sense, as this is not the focus.

To summarize, we know that the Vašíček model is stable, that the CIR model can work under extremely nice simulated circumstances but that the transition density in the CIR model is most likely going to yield extreme numerical instability as rates are small and parameters are suspected to be small as well. As such, we want to examine the performance of CIR-ARHMM in-sample whilst outlining the methodology for the actual thesis. We explore properties, discretizations

[1] Comments on this phenomena will be made later in the paper.

and possible different methods for parameter estimation under the CIR model, such as using the stationary distribution. The focus will be to dissect the CIR model properties and weaknesses as the square-root process is deceivingly difficult to assess both in theory and pratice.

If the CIR-ARHMM is concluded to be rendered useless by numerical instability the thesis will be on the Vašíček model extended by a ARHMM and a more exotic autoregressive continuous state-space model for both in- and out-of-sample testing.

4 Data

The preparation paper examines the modelling of short rates. In a short rate model, the stochastic state variable represents the instantaneous spot rate. The short rate is the continuously compounded interest rate at which an entity can borrow money for an infinitesimally small period starting at time-t. This begs the question: *"where is such a quantity observed?"*. The answer is simply that we have to find a proxy that is approximately instantaneous. Following the conclusion of [11], we use 3-month Treasury bill (3M T-bills), 3M rates, DGS3MO or more precisely: *Market Yield on U.S. Treasury Securities at 3-Month Maturity, Quoted on an Investment Basis* to test the performance of our models in a practical and an empirical setting. The data under analysis is a time series of the market yield on U.S. Treasury securities with a 3-month constant maturity, spanning 9638 business days from March 1, 1984, to July 15, 2022 shown in Figure 1. A *business day* is defined as having 5 observations per week, 21 per month, and 252 per year[2]. The data used can be accessed on our Github, formatted and cleaned, or through The Federal Reserve Bank of St. Louis [19] for the readers own data-suffering.

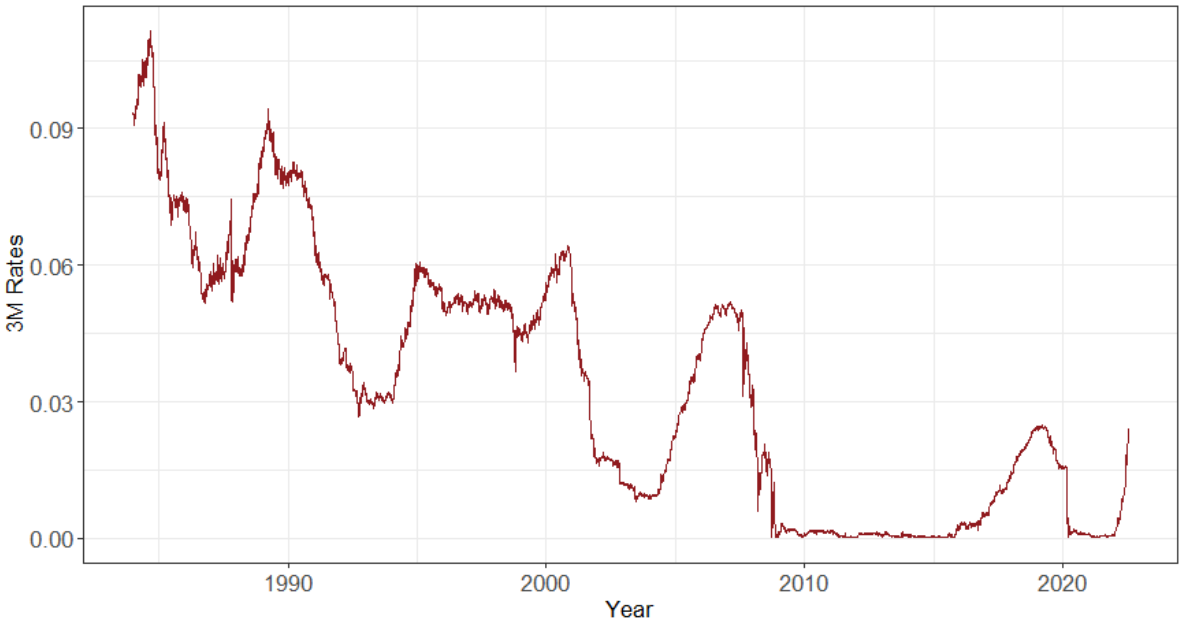


Figure 1: *DGS3MO Time Plot from 1984 to 2022.*

Usually, when working with HMM's, as this thesis shall, the data is that of animals. As nature is quite chaotic, rough and out of our control, tagging devices capturing various and numerous measurement data such as; angular momentum, speed, height, depth, stomach temperature changes etc. are highly subject to missing data and measurement errors because of the violent and turbulent environments governed by the physics of nature. Equipment might get damaged, be subject to changes in water densities making measurement errors differ or simply fall off. This

[2]Hence the funniest joke in finance: There is 252 days in a year.

means missing data should often be extrapolated or treated very carefully as it alters the likelihood computation significantly.

This is fortunately not an issue with financial data in general and specifically in our case. Indeed, observe Table 1 for descriptive statistics of our data set. It is quite evident that no missing values are present in our data, which is to be expected in a multi-trillion dollar financial sector. The largest value was observed in the year 1984 as countries, including the US, were trying to control the inflation caused by the economic recession [32]. The lowest rate is observed at the end of 2008, coinciding with the height of the global financial crisis. In response to the severe economic downturn, the Federal Reserve implemented aggressive monetary policy measures, including slashing the federal funds rate to near-zero levels. This unprecedented move aimed to inject liquidity into the financial system, restore investor confidence, and stimulate borrowing and spending. As a result, short-term yields plummeted to historic lows, reflecting both the accommodative policy stance and heightened risk aversion in financial markets. The near-zero interest rate environment persisted for several years, marking a fundamental shift in monetary policy strategy during times of systemic distress. However, the mean is approximately 0.034 with 75% of the observations being less than or equal to 0.054. In opposition to some Nordic and Asian countries, the US issued 3-month T-bills do not include negative rates, ever. Furthermore, zero-rates are rarely observed.

Metric	**Value**
Missing values	0
Minimum Value	0.00000
Number of Zeros	21
1st Quartile (Q1)	0.00310
Median	0.03120
Mean	0.03351
3rd Quartile (Q3)	0.05400
Maximum Value	0.11140

Table 1: *DGS3MO descriptive statistics.*

Considering the handful of missing data, we see that the count is a mere 20. However, we should address these cases for accuracy. The CIR model, by construction, can not handle negative rates when doing estimation as the transition probability is that of a non-central χ^2 distribution. This is because the degrees of freedom and non-centrality parameter is strictly positive. As there are only 20 counts of missing data, we choose to exclude them from the data set[a]. We will discuss this further as it is apparent from Figure 1 that 0 and near 0 rates are rapidly increasing. We choose to not use any later data as this will most likely be used in the actual thesis.

[a]The recommended method to dealing with missing data if data is missing (completely) at random (MACR/MAR) when using a HMM is to compute the *ignorable likelihood* by replacing $\boldsymbol{P}(r_i)$ with I when observation r_i is missing [45, p. 40] which is a reasonable basis for parameter estimation [35] (see section on Likelihood Estimation under HMM's).

The 3M T-bill data is evidently highly correlated by visualization in Figure 2 with 100 lags using the `ACF`-function in `R`:

$$\text{ACF}_k = \frac{\sum_{t=k+1}^{n}(r_t - \bar{r})(r_{t-k} - \bar{r})}{\sum_{t=1}^{n}(r_t - \bar{r})^2},$$

where r_t is the value of the time series at time t, $\bar{r}$ is the mean of the time series, n is the length of the time series, k is the lag at which the correlation is calculated, ACF_k is the empirical autocorrelation at lag k.

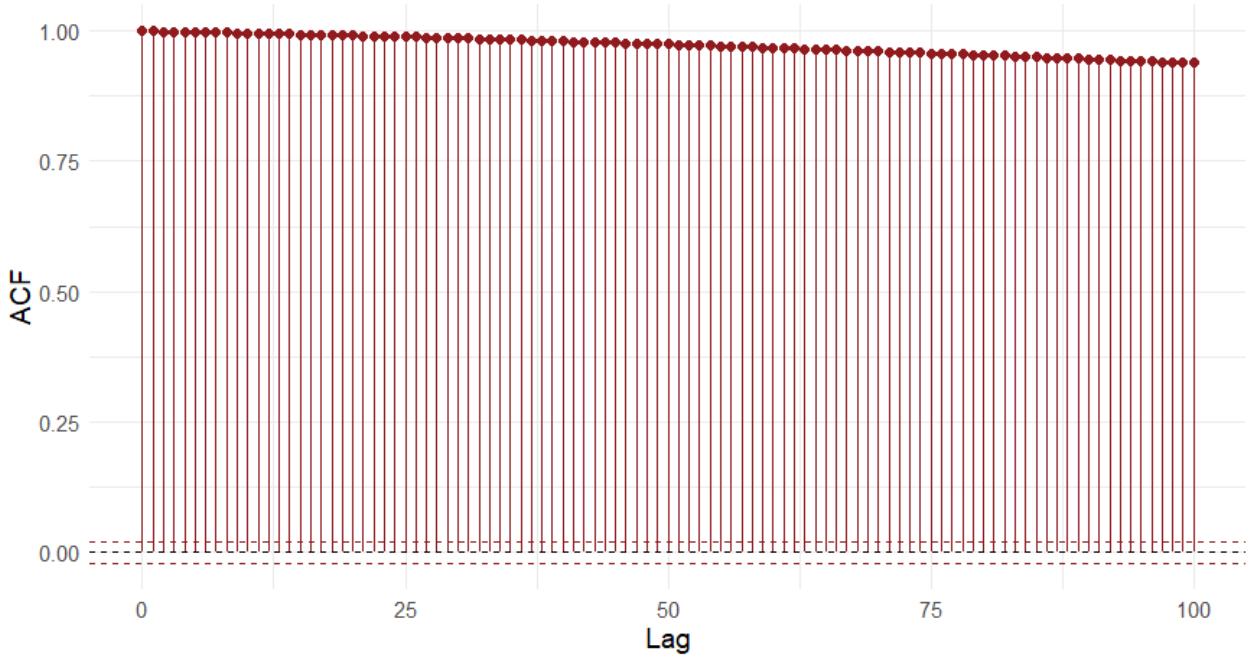


Figure 2: *DGS3MO empirical autocorrelation of 100 lags.*

The serial dependency will be addressed when we discuss what model might be adequate to handle the over-abundance of correlation.

5 Theory & Methodology

5.1 Short Rate Models: Cox-Ingersoll-Ross

5.1.1 BESQ and Cox-Ingersoll-Ross Processes

Throughout this section, let $(\Omega, \mathcal{F}, \{\mathcal{F}\}_{t\geq 1}, \mathbb{P})$ be a filtered probability space. To establish a general framework for the CIR model we first define BESQ_x^δ processes as follows.

Definition 5.1. *[27, Def. 6.1.2.1] For every $\delta \geq 0$ and $x \geq 0$, the unique strong solution to the equation*

$$\rho_t = x + \delta t + 2\int_0^t \sqrt{\rho_s}\, dW_s, \quad \rho_t \geq 0,$$

is called a squared Bessel process with dimension δ, starting at x and is denoted by $BESQ_x^\delta$.

We shall then show that a special kind of BESQ^δ process arises under some specific parameter assumptions.

Consider first the SDE

$$dr_t = \kappa(\theta - r_t)dt + \sigma\sqrt{r_t}dW_t, \quad t > 0,\ r_0 = x \in \mathbb{R}, \tag{1}$$

where W_t is a standard Brownian motion, κ the speed of mean reversion, θ the long term mean, and $\sigma > 0$ the volatility. In order to find the solution of Equation 1, consider the following theorem.

Theorem 5.1. *[27, Thm. 1.5.5.1] Consider the SDE*

$$X_t = X_0 + \int_0^t b(s, X_s)\, ds + \int_0^t \sigma(s, X_s)\, dW_s. \tag{2}$$

Suppose $\varphi : (0,\infty) \to (0,\infty)$ is a Borel function such that

$$\int_0^\infty \frac{1}{\varphi(a)} da = \infty.$$

Under any of the following conditions:

(i) The Borel function b is bounded, the function σ does not depend on the time variable and satisfies

$$|\sigma(x) - \sigma(y)|^2 \leq \varphi(|x-y|), \quad \text{and} \quad |\sigma| \geq \varepsilon > 0.$$

(ii) $|\sigma(s,x) - \sigma(s,y)|^2 \leq \varphi(|x-y|)$ and b is Lipschitz continuous.

(iii) The function σ does not depend on the time variable and satisfies

$$|\sigma(x) - \sigma(y)|^2 \leq |f(x) - f(y)|,$$

where f is a bounded increasing function, $\sigma \geq \varepsilon > 0$, and b is bounded.

Then Equation 2 admits a unique solution which is strong, and the solution X is a Markov process.

To prove that Equation 1 has a solution, consider first the altered form where we enforce positivity of r_t in the square root term by considering the absolute value, $|r_t|$:

$$dr_t = \kappa(\theta - r_t)dt + \sigma\sqrt{|r_t|}dW_t, \quad t > 0,\ r_0 = x_0 \in \mathbb{R}, \tag{3}$$

Take $\varphi(x) = cx$. It follows that

$$\begin{aligned}
\left|\sigma\sqrt{|x|} - \sigma\sqrt{|y|}\right|^2 &= \sigma^2 \left|\sqrt{|x|} - \sqrt{|y|}\right|^2 \\
&= \sigma^2 \left|\left(\sqrt{|x|} - \sqrt{|y|}\right)^2\right| \\
&= \sigma^2 \left|\left(\sqrt{|x|} - \sqrt{|y|}\right)\left(\sqrt{|x|} - \sqrt{|y|}\right)\right| \\
&\overset{\dagger}{\leq} \sigma^2 \left||x| - |y|\right| \\
&\overset{\dagger\dagger}{\leq} \sigma^2 |x - y|,
\end{aligned}$$

where † follows by $\left|\sqrt{x} - \sqrt{y}\right| = \sqrt{|x - y|}$ and †† by the reverse triangle inequality. As condition (i) in Theorem 5.1 is fulfilled, it follows that Equation 2 admits a strong solution which is also a Markov process.

Lastly, consider the Comparison Theorem.

Theorem 5.2. *[27, Thm. 1.5.5.9 (Comparison Theorem)] Let*

$$dX_i(t) = b_i(t, X_i(t))dt + \sigma(t, X_i(t))dW_t, \quad i = 1, 2,$$

where $b_i, i = 1, 2$ are bounded Borel functions and at least one of them is Lipschitz and σ satisfies (ii) or (iii) of Theorem 5.1. Suppose also that $X_1(0) \geq X_2(0)$ and $b_1(x) \geq b_2(x)$. Then $X_1(t) \geq X_2(t)$, $\forall t$, a.s.

Observe that for $r_0 = 0$ and $\theta = 0$, the solution of Equation 2 is $r_t = 0$ for all t. Then, by Theorem 5.2, it follows that assuming $\kappa\theta > 0$ implies that $r_t \geq 0$ whenever $r_0 > 0$. In this case, we simply omit the absolute value in Equation 3 and consider the positive solution of

$$dr_t = \kappa(\theta - r_t)dt + \sigma\sqrt{r_t}dW_t, \quad t > 0,\ r_0 = x \in \mathbb{R}.$$

This solution is called a Cox-Ingersoll-Ross (CIR) [12] process or a (Feller) square-root process [20]. It is of special interest not only in interest rate modelling but also for asset dynamics in i.e. the Heston model [25] where the variance follows such a square-root process in the bivariate system.

The commonly known *Feller condition* to ensure positivity of the solution can now be derived. Firstly, one has to recognize that the CIR process in Equation 1 can be found by a space-time change applied to r_t.

Proposition 5.1. *The CIR process in Equation 1 is a BESQ$^\delta$ process transformed by the following space-time changes:*

$$r_t = e^{-\kappa t}\rho\left(\frac{\sigma^2}{4\kappa}\left(e^{kt}-1\right)\right),$$

where $\{\rho(s), s \geq 0\}$ is a BESQ$^\delta$ process, with dimension $\delta = \frac{4\kappa\theta}{\sigma^2}$.

Proposition 5.2. *Let ρ be a δ-dimensional squared Bessel process. For $\delta = 0$, the point 0 is absorbing (the process remains at 0 as soon as it reaches it). For $0 < \delta < 2$, the BESQ$^\delta$ is reflected instantaneously.*

From Proposition 5.2, it follows that for $2\kappa\theta \geq \sigma^2$, a CIR process starting from a positive initial point x stays positive. For $0 \leq 2\kappa\theta < \sigma^2$, a CIR process starting from a positive initial point hits 0 with probability $p \in (0,1)$ in the case $\kappa < 0$ and a.s. if $\kappa \geq 0$. In the case $0 < 2\kappa\theta$, the boundary 0 is instantaneously reflecting, whereas in the case $2\kappa\theta < 0$, the process r starting from a positive initial point reaches 0 a.s..

5.1.2 Cox-Ingersoll-Ross as a Short Rate Model

Short Rate Dynamics: From $\mathbb{P}$ to $\mathbb{Q}(\lambda)$ The CIR process as a short interest rate model gained traction since the seminal paper [12] of Cox, Ingersoll and Ross. In this model, they assume that the rate, follows a Feller square-root process under the historical probability measure $\mathbb{P}$ given by

$$dr_t = \tilde{\kappa}(\tilde{\theta} - r_t)dt + \sigma\sqrt{r_t}dW_t^{\mathbb{P}}.$$

$\tilde{\kappa}(\tilde{\theta} - r_t)$ is a mean-reverting drift pulling the interest rate towards its long term mean $\tilde{\theta}$ at the speed $\tilde{\kappa}$. However, in a risk-adjusted economy where a agent would want to be compensated for the partaking of risk. As such, the dynamics by Girsanov's Theorem [6, Thm. 12.3] are given as

$$dr_t = (\tilde{\kappa}(\tilde{\theta} - r_t) - \lambda r_t)dt + \sigma\sqrt{r_t}dW_t^{\mathbb{Q}(\lambda)}, \quad W_t^{\mathbb{Q}(\lambda)} = W_t^{\mathbb{P}} + \frac{\lambda}{\sigma}\int_0^t \sqrt{r_t}ds,\ t \geq 0,$$

for $\lambda \in \mathbb{R}$. Note that [12] (inspired by [7]) assumes that the risk is proportional to the interest rate and thus the form λr_t. The dynamics under the EMM associated with $\lambda \in \mathbb{R}$, $\mathbb{Q}(\lambda)$, are then

$$dr_t = \underbrace{\kappa(\theta - r_t)}_{\text{drift}} dt + \underbrace{\sigma\sqrt{r_t}}_{\text{diffusion}} dW_t^{\mathbb{Q}(\lambda)}, \tag{4}$$

where $\kappa = \tilde{\kappa} + \lambda$, $\theta = \tilde{\kappa}\left(\tilde{\theta}/\tilde{\kappa}\right)$.

Distributional Properties Proceeding, results of the $\mathbb{Q}(\lambda)$-dynamics of the SDE in Equation 4 will be given. Firstly, the first two conditional moments of the r.v. r_t.

Theorem 5.3. *Let r be a CIR process, the solution of*

$$dr_t = \kappa(\theta - r_t)dt + \sigma\sqrt{r_t}dW_t^{\mathbb{Q}(\lambda)}.$$

The conditional expectation and the conditional variance of the r.v. r_t are given by, for $t > s$,

$$\begin{aligned}
\mathbb{E}[r_t \mid \mathcal{F}_s] &= r_s e^{-\kappa(t-s)} + \theta\left(1 - e^{-\kappa(t-s)}\right), \\
\mathbb{V}[r_t \mid \mathcal{F}_s] &= r_s \frac{\sigma^2\left(e^{-\kappa(t-s)} - e^{-2\kappa(t-s)}\right)}{\kappa} + \frac{\theta\sigma^2\left(1 - e^{-\kappa(t-s)}\right)^2}{2\kappa}.
\end{aligned}$$

Note: *if $\kappa > 0$ it is easily seen that $\mathbb{E}[r_t] \to \theta$ as $t \to \infty$, and hence why we say the CIR process enjoys the property of mean reversion.*

Arguably the most important result for the CIR model is now given, the transition density.

Proposition 5.3. *Let $t > s$. Denote by ${}^{\kappa}\mathbb{Q}^{\kappa\theta,\sigma}$ the law of the CIR process, r, solution of*

$$dr_t = \kappa(\theta - r_t)dt + \sigma\sqrt{r_t}dW_t^{\mathbb{Q}(\lambda)}.$$

The transition density ${}^{\kappa}\mathbb{Q}^{\kappa\theta,\sigma}(r_{t+s} \in dy \mid r_s = x) = f_t(x,y)dy$ is given by

$$f_t(x,y) = \frac{e^{\kappa t}}{2\eta}\left(\frac{ye^{\kappa t}}{x}\right)^{\nu/2} \exp\left(-\frac{x + ye^{\kappa t}}{2\eta}\right) \mathcal{I}_\nu\left(\frac{1}{\eta}\sqrt{xye^{\kappa t}}\right) \mathbb{1}_{\{y \geq 0\}},$$

where $\eta = \frac{\sigma^2}{4\kappa}\left(e^{\kappa t} - 1\right)$ and $\nu = \frac{2\kappa\theta}{\sigma^2} - 1$▽.

Note, that the density in Proposition 5.3 is that of a non-central χ^2 $\chi^2_\delta(\alpha)$ with $\delta = 2(\nu + 1)$ degrees of freedom, and non-centrality parameter $\alpha = x/\eta(t)$ such that we obtain

$${}^{\kappa}\mathbb{Q}_x^{\kappa\theta,\sigma}(r_t < y) = \chi'^2_{\frac{4\kappa\theta}{\sigma^2}}\left(\frac{x}{\eta(t)}; \frac{ye^{\kappa t}}{\eta(t)}\right),$$

where the function $\chi^2_\delta(\alpha;\cdot)$ is the cumulative distribution function associated with said density, i.e.

$$\begin{aligned} f(x;\delta,\alpha) &= 2^{-\delta/2}\exp\left(-\frac{1}{2}(\alpha+x)\right)x^{\delta/2-1}\sum_{n=0}^{\infty}\left(\frac{\alpha}{4}\right)^n\frac{1}{n!\Gamma(n+\delta/2)}x^n\mathbb{1}_{\{x>0\}} \\ &= \frac{e^{-\alpha/2}}{2\alpha^{\nu/2}}e^{-x/2}x^{\nu/2}\mathcal{I}_\nu(\sqrt{\alpha x})\mathbb{1}_{\{x>0\}}, \end{aligned}$$

where

$$\mathcal{I}_\nu(x) = \sum_{m=0}^{\infty}\frac{1}{m!\Gamma(m+\nu+1)}\left(\frac{x}{2}\right)^{2m+\nu}, \tag{5}$$

is a modified Bessel function of the first kind of order ν evaluated at x and $\Gamma(z)$ is the Gamma-function evaluated at z. The modified Bessel function can be written in terms of the (non-modified) Bessel function of the first kind

$$\mathcal{J}_\nu(x) = \sum_{m=0}^{\infty}\frac{(-1)^m}{m!\Gamma(m+\nu+1)}\left(\frac{x}{2}\right)^{2m+\nu},$$

as

$$\mathcal{I}_\nu(x) = i^{-\nu}\mathcal{J}_\nu(ix).$$

We use the built in `besselI`-function in base `R` [37] ▽. We drop the notation of $\mathbb{Q}(\lambda)$ and simply use $\mathbb{Q} := \mathbb{Q}(\lambda)$ for the rest of the paper.

To derive the stationary distribution r_t^* for the CIR model, we must (or rather can) use the Fokker-Planck equation. The Fokker–Planck equation for the probability density $f(r,t)$ of r_t is

$$\frac{\partial f}{\partial t} + \frac{\partial}{\partial r}\left[\kappa(\theta-r)f\right] = \frac{1}{2}\sigma^2\frac{\partial^2}{\partial r^2}(rf).$$

Our interest is in the particular case when $\partial f/\partial t \to 0$, which leads to the simplified equation

$$\kappa(\theta-r)f = \frac{1}{2}\sigma^2\left(f + r\frac{df}{dr}\right) \Rightarrow \frac{\alpha-1}{r} - \beta = \frac{d}{dr}\log f,$$

by defining $\alpha = \frac{2\kappa\theta}{\sigma^2}$ and $\beta = \frac{2\kappa}{\sigma^2}$ and rearranging. Integrating shows us that

$$\begin{aligned}
\frac{\alpha - 1}{r} - \beta &= \frac{d}{dr} \log f \\
&\iff \\
f &= C \exp\left(\int \left(\frac{\alpha - 1}{r} - \beta \right) dr \right) \\
&= C \exp\left(\int \frac{\alpha - 1}{r} dr - \int \beta \, dr \right) \\
&= C \exp\left((\alpha - 1) \log r - \beta r \right) \\
&= C r^{\alpha - 1} e^{-\beta r} \\
&\Rightarrow \\
f &\propto r^{\alpha - 1} e^{-\beta r}.
\end{aligned}$$

Over the range $f \in (0, \infty]$, this density describes a Γ-distribution. Therefore, the stationary distribution of the CIR model is a Γ-distribution. The probability density function of the r.v. $r_t^* \sim \Gamma\left(\frac{2\kappa\theta}{\sigma^2}, \frac{2\kappa}{\sigma^2}\right)$, is therefore

$$f(r_t^*; \kappa, \theta, \sigma) = \frac{\beta^\alpha}{\Gamma(\alpha)} (r_t^*)^{\alpha - 1} e^{-\beta r_t^*}, \quad r_t^* > 0.$$

The distribution of r_t^* does obviously not depend on the time-t. r_t is thus asymptotically stationary in the sense that it resembles the stationary process r_t^* for large t. Continuing we have the two moments using the known formulas for the Γ-distribution

$$\begin{aligned}
\mathbb{E}[r_t^*] &= \frac{\alpha}{\beta} = \frac{\frac{2\kappa\theta}{\sigma^2}}{\frac{2\kappa}{\sigma^2}} = \theta, \\
\mathbb{V}[r_t^*] &= \frac{\alpha}{\beta^2} = \frac{\frac{2\kappa\theta}{\sigma^2}}{\left(\frac{2\kappa}{\sigma^2}\right)^2} = \frac{\frac{2\kappa\theta}{\sigma^2}}{\frac{4\kappa^2}{\sigma^4}} = \frac{2\kappa\theta\sigma^2}{4\kappa^2} = \frac{\theta\sigma^2}{2\kappa}.
\end{aligned}$$

The asymptotic short rate r_t^* in the CIR model thus follows a Γ-distribution with shape parameter $\alpha = \frac{2\kappa\theta}{\sigma^2}$ and rate parameter $\beta = \frac{2\kappa}{\sigma^2}$. The distribution is depicted in Figure 3 by the associated 95%-confidence interval bands over time on a simulation of a rate using the QE-scheme developed by [2].

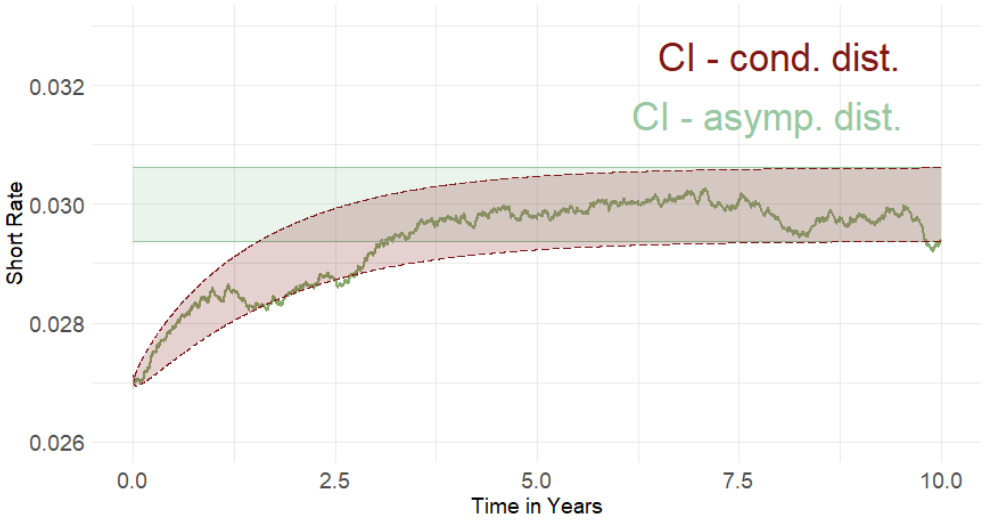


Figure 3: *One realization of the short rate over 10 years using the QE scheme from [2]. The actual and stationary distribution are depicted by their 95% confidence intervals.*

One could justify using the stationary distribution for maximum-likelihood estimation. However, that method is fundamentally flawed if one wants to estimate the CIR parameters. Consider the statistical model

$$\mathcal{P}_{\text{CIR}} = \left\{ f_{\boldsymbol{\rho}}(r^*) = \frac{\left(\frac{2\kappa}{\sigma^2}\right)^{\frac{2\kappa\theta}{\sigma^2}}}{\Gamma\left(\frac{2\kappa\theta}{\sigma^2}\right)} (r^*)^{\frac{2\kappa\theta}{\sigma^2}-1} e^{-\frac{2\kappa}{\sigma^2}r^*} \;\middle|\; \boldsymbol{\rho} = (\kappa,\ \theta,\ \sigma)^\top : \kappa, \theta, \sigma > 0 \right\}. \tag{6}$$

Consider two parameter sets:

$$\boldsymbol{\rho}_1 = (1,\ 2,\ 1)^\top, \quad \boldsymbol{\rho}_2 = (2,\ 1,\ \sqrt{2})^\top.$$

For these sets, as the Γ-distribution is exclusively parametrized through its parameters $(\alpha,\ \beta)^\top$, we calculate:

$$\alpha_1 = \frac{2\kappa_1\theta_1}{\sigma_1^2} = \frac{2 \cdot 1 \cdot 2}{1^2} = 4, \quad \beta_1 = \frac{2\kappa_1}{\sigma_1^2} = \frac{2 \cdot 1}{1^2} = 2,$$

$$\alpha_2 = \frac{2\kappa_2\theta_2}{\sigma_2^2} = \frac{2 \cdot 2 \cdot 1}{(\sqrt{2})^2} = 4, \quad \beta_2 = \frac{2\kappa_2}{\sigma_2^2} = \frac{2 \cdot 2}{(\sqrt{2})^2} = 2.$$

Thus, $f_{\boldsymbol{\rho}_1} = f_{\boldsymbol{\rho}_2}$, even though $\boldsymbol{\rho}_1 \neq \boldsymbol{\rho}_2$. This shows that $\boldsymbol{\rho} = (\kappa,\ \theta,\ \sigma)^\top$, or rather $(\kappa,\ \sigma)^\top$, is unidentifiable even though θ is identifiable. On the contrary, consider the statistical model parametrized through the Γ-distribution parameters $\alpha = \frac{2\kappa\theta}{\sigma^2}$ and $\beta = \frac{2\kappa}{\sigma^2}$

$$\mathcal{P}_\Gamma = \left\{ f_{\boldsymbol{\rho}}(r^*) = \frac{\beta^\alpha}{\Gamma(\alpha)} (r^*)^{\alpha-1} e^{-\beta^*} \;\middle|\; \boldsymbol{\rho} = (\alpha,\ \beta)^\top : \alpha, \beta > 0 \right\}. \tag{7}$$

If $f_{\boldsymbol{\rho}_1} = f_{\boldsymbol{\rho}_2}$ for $\boldsymbol{\rho}_1 = (\alpha_1,\ \beta_1)^\top$ and $\boldsymbol{\rho}_2 = (\alpha_2,\ \beta_2)^\top$, then

$$\frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)}(r^*)^{\alpha_1-1}e^{-\beta_1 r^*} = \frac{\beta_2^{\alpha_2}}{\Gamma(\alpha_2)}(r^*)^{\alpha_2-1}e^{-\beta_2 r^*}.$$

This equality cannot hold for all $r^* > 0$ unless $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$. Therefore, the mapping $(\alpha,\ \beta)^\top \mapsto \mathcal{P}$ is one-to-one and $(\alpha,\ \beta)^\top$ is therefore by [21, Def. 14.1] identifiable. We will discuss the implications and applications of which statistical model to use later.

OLS Estimation & Discretization Suitable initialization of the maximization/minimization algorithm to ensure global maximization of the log-likelihood in the estimation is crucial. As such, it is beneficial to find suitable estimates to initialize our models. Among many picks, we choose OLS.

Let $(\Omega, \mathcal{F}, \{\mathcal{F}\}_{t\geq 0}, \mathbb{P})$ be a filtered probability space. Assume the r.v. X_t is driven by the SDE

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dW_t^{\mathbb{Q}}, \tag{8}$$

where $W_t^{\mathbb{Q}}$ is a Wiener process under $\mathbb{Q}$. Equally-spaced time increments are used for notational convenience, allowing us to write $t_i - t_{i-1} := \Delta$. Integrating Equation 8 from t to $t + \Delta$ yields

$$X_{t+\Delta} = X_t + \int_t^{t+\Delta} \mu(X_u, u)du + \int_t^{t+\Delta} \sigma(X_u, u)dW_u^{\mathbb{Q}}. \tag{9}$$

At time-t, $\hat{X}_t$ is known. We aim to obtain the incremented, $\hat{X}_{t+\Delta}$. Euler scheme approximates the integrals using the left end-point rule, such that the deterministic integral of Equation 9 is approximated as the product of the integrand at time-t and the integration range Δ

$$\int_t^{t+\Delta} \mu(X_t, u)du \approx \mu(X_t, t)\int_t^{t+\Delta} du = \mu(X_t, t)\Delta.$$

Left end-points is a natural candidate as at time-t the value of $\mu(X_t, t)$ is known. Now, let $Z^{\mathbb{Q}} \sim \mathcal{N}(0, 1)$. The stochastic integral is approximated as

$$\int_t^{t+\Delta} \sigma(X_u, u)dW_u^{\mathbb{Q}} \approx \sigma(X_t, u)\int_t^{t+\Delta} dW_u^{\mathbb{Q}} = \sigma(X_t, u)(W_{t+\Delta}^{\mathbb{Q}} - W_t^{\mathbb{Q}}) = \sigma(X_t, u)\sqrt{\Delta}Z^{\mathbb{Q}},$$

because $W_{t+\Delta}^{\mathbb{Q}} - W_t^{\mathbb{Q}}$ and $\sqrt{\Delta}Z^{\mathbb{Q}}$ are identically distributed [6, Def. 4.3]. Assembling the results yields the general form of the Euler discretization scheme:

$$\hat{X}_{t+\Delta} = \hat{X}_t + \mu(X_t, t)\Delta + \sigma(X_t, t)\sqrt{\Delta}Z^{\mathbb{Q}}. \tag{10}$$

Applying Euler discretization to dr_t in Equation 4 by substituting the diffusion and drift of dr_t

into Equation 10 yields the final discretization Euler scheme of the CIR dynamics

$$\hat{r}_{t+\Delta} = \hat{r}_t + \kappa(\theta - \hat{r}_t)\Delta + \sigma\sqrt{\hat{r}_t\Delta}Z^{\mathbb{Q}} \iff \hat{r}_{t+\Delta} - \hat{r}_t = \kappa(\theta - \hat{r}_t)\Delta + \sigma\sqrt{\hat{r}_t\Delta}Z^{\mathbb{Q}}.$$

In practice we will observe negative values of the rate even when Proposition 5.1 holds [2, p. 2]. When this occurs we use *full truncation* yielding the complete scheme

$$\hat{r}_{t+\Delta} - \hat{r}_t = \kappa(\theta - \hat{r}_t^+)\Delta + \sigma\sqrt{\hat{r}_t^+\Delta}Z^{\mathbb{Q}}, \tag{11}$$

where $r_t^+ \equiv \max\{0, r_t\}$. Other scheme functions for discretized forms

$$\begin{aligned}\hat{r}_{t+\Delta} &= f_1(\hat{r}_t) + \kappa(\theta - f_2(\hat{r}_t))\Delta + \sigma\sqrt{f_3(\hat{r}_t)\Delta}Z^{\mathbb{Q}},\\ r_{t+\Delta} &= f_3(\hat{r}_{t+\Delta}),\end{aligned}$$

can be seen in Table 2.

Scheme	**Paper**	$f_1(r)$	$f_2(r)$	$f_3(r)$
Absorption	N/A	r^+	r^+	r^+
Reflection	[16], [15], [5]	$\|r\|$	$\|r\|$	$\|r\|$
Higham and Mao	[26]	r	r	$\|r\|$
Partial Truncation	[13]	r	r	r^+
Full Truncation	[36]	r	r^+	r^+

Table 2: *Euler scheme negative variance handling methods.*

Proceeding, using the Euler discretized full truncation form of Equation 4, namely, Equation 11, yields

$$\begin{aligned} r_{t+\Delta} - r_t &= \kappa(\theta - r_t)\Delta + \sigma\sqrt{r_t}\Delta Z^{\mathbb{Q}}\\ &\iff \frac{r_{t+\Delta} - r_t}{\sqrt{r_t^+}} = \kappa(\theta - r_t)\frac{\Delta}{\sqrt{r_t^+}} + \sigma\Delta Z^{\mathbb{Q}}\\ &= \kappa\theta\frac{\Delta}{\sqrt{r_t^+}} - \kappa\sqrt{r_t^+}\Delta + \sigma\Delta Z^{\mathbb{Q}}. \end{aligned} \tag{12}$$

Define now quantities:

$$\begin{aligned} y_i &= \frac{r_{t+\Delta} - r_t}{\sqrt{r^+}}, \quad \beta_1 = \kappa\theta, \qquad \beta_2 = -\kappa,\\ z_{1i} &= \frac{\Delta}{\sqrt{r_t^+}}, \qquad z_{2i} = \sqrt{r_t^+}\Delta, \quad \varepsilon_i = \sigma\Delta Z^{\mathbb{Q}}. \end{aligned} \tag{13}$$

Equation 12 using Equation 13 can then be written by the regression

$$y_i = \beta_1 z_{1i} + \beta_2 z_{2i} + \varepsilon_i \iff \boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \end{bmatrix}, \quad \boldsymbol{A} = \begin{bmatrix} \frac{\Delta}{\sqrt{r_1^+}} & \sqrt{r_1^+\Delta} \\ \frac{\Delta}{\sqrt{r_2^+}} & \sqrt{r_2^+\Delta} \\ \vdots & \vdots \\ \frac{\Delta}{\sqrt{r_{n-1}^+}} & \sqrt{r_{n-1}^+\Delta} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \quad \boldsymbol{\epsilon} = \sigma\sqrt{\Delta} \begin{bmatrix} \mathcal{N}_1(0,1) \\ \mathcal{N}_2(0,1) \\ \vdots \\ \mathcal{N}_{n-1}(0,1) \end{bmatrix}, \tag{14}$$

with $\mathcal{N}_i(0,1)$ i.i.d for $i = 1, \ldots, n-1$. The OLS estimator of $\boldsymbol{\beta}$ is then

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\| \boldsymbol{Y} - \boldsymbol{A}\boldsymbol{\beta} \right\|^2,$$

where $\|\cdot\|$ denotes the Euclidean norm. The solution is well known as

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \left(\boldsymbol{A}^\top \boldsymbol{A}\right)^{-1} \boldsymbol{A}^\top \boldsymbol{Y}.$$

It then follows from Equation 13 that κ and θ can be estimated as

$$\hat{\kappa} = -\hat{\beta}_2, \quad \hat{\theta} = \frac{\hat{\beta}_1}{\hat{\kappa}}.$$

Following [29, p. 371] and Equation 14, the estimator for σ is found as the standard deviation of residuals

$$\hat{\sigma}^2\Delta = \frac{\left\| \boldsymbol{Y} - \boldsymbol{A}\boldsymbol{\beta} \right\|^2}{n} \iff \hat{\sigma} = \sqrt{\frac{\left\| \boldsymbol{Y} - \boldsymbol{A}\boldsymbol{\beta} \right\|^2}{\Delta n}}.$$

Using OLS by the above described equations with $\Delta = 1/252$, yields the results seen in Equation 15

$$\hat{\kappa} = 0.2068172, \quad \hat{\theta} = 0.0247371, \quad \hat{\sigma} = 0.07450132. \tag{15}$$

Residuals We simply use the Euler discretization method given in Equation 11 of the CIR model dynamics to assess the model fit.

Limitations of the CIR Model The CIR model ensures non-negativity, which was advantageous for modelling interest rates in the pre-crisis period but may pose challenges in the post-crisis environment. However, the model's flexibility is limited, as it often fails to accurately reproduce the variety of zero-coupon yield curve shapes observed in financial markets. Additionally, being a single-factor model with a constant volatility parameter, the CIR framework does not account for market phenomena such as jumps.

When the CIR model was originally developed, the possibility of negative interest rates was not a consideration. In fact, it was desirable for a model to generate strictly positive rates. However, as observed in the post-financial crisis era, negative rates are a real and recurring phenomenon. Anecdotally, some banks welcomed this shift, as it meant they did not need to replace or modify their extended Vasicek models, which naturally allow for negative rates.

The practical limitations are numerous. The presence of the square root term, the non-central χ^2-distribution involving modified Bessel functions of the first kind, and exponential terms near zero all contribute to significant challenges in estimation. While many of these issues can be addressed, numerical instability is ultimately unavoidable. This is simply not a model for the faint-hearted in practice.

5.1.3 Likelihood Estimation

The joint density of the realization $\{r_t\}_{t=0}^{T}$ with the initial value r_0 fixed and known, by the Markov property of the CIR process, for consecutive observations of the short rate, the joint density factorizes as

$$\begin{aligned}
\mathcal{L}_T(\boldsymbol{\zeta}) &= f(r_T, r_{T-1}, \ldots, r_1 \mid r_0; \boldsymbol{\zeta}) \\
&= \prod_{t=1}^{T} f(r_t \mid r_{t-1}; \boldsymbol{\zeta}) \\
&= \prod_{t=1}^{T} \frac{\exp(\kappa\Delta)}{2\eta} \left(\frac{r_t \exp(\kappa\Delta)}{r_{t-1}}\right)^{\nu/2} \exp\left(-\frac{r_{t-1} + r_t \exp(\kappa\Delta)}{2\eta}\right) \mathcal{I}_\nu\left(\frac{1}{\eta}\sqrt{r_{t-1} r_t \exp(\kappa\Delta)}\right) \\
&\overset{\dagger}{\Longleftrightarrow} \\
\mathcal{L}_T(\boldsymbol{\zeta}) &= \prod_{t=1}^{T} c \cdot \exp(-u_{t-1} - v_t) \left(\frac{v_t}{u_{t-1}}\right)^{q/2} \mathcal{I}_q\left(2\sqrt{u_{t-1} v_t}\right),
\end{aligned}$$

where † is a extremely convenient rewriting[3] and adhering to the notation used throughout the literature (for example originally derived in [12]) with

$$\begin{aligned}
c &= \frac{2\kappa}{\sigma^2\left(1 - \exp\left(-\kappa\Delta\right)\right)}, \\
u_{t-1} &= c r_{t-1} \exp\left(-\kappa\Delta\right), \\
v_t &= c r_t, \\
q &= \frac{2\kappa\theta}{\sigma^2} - 1,
\end{aligned}$$

[3]The density is exactly the same. Indeed, $r_t \mid r_{t-1} \sim \frac{1}{2c}\chi'_{2q+2}(2u)$. The formulae and notation are those originally given in the seminal paper of Cox, Ingersoll and Ross [12, pp. 391–392] building on results taken directly from [20]. We show how the density is derived in A.2 as a derivation is impossible to find in the literature.

and where $\Delta = 1/252$ in our implementation for the simple CIR model and $\boldsymbol{\zeta} = (\kappa,\ \theta,\ \sigma)^\top$. By convention and to align units, it is usually chosen that the time-step is such that $\Delta = 1/252$ to align with number of business days per year, approximately, as rates are annualized. However, $\Delta = 1/252$ causes the exponent to become extremely small as κ initial values will be numerically small. However, from observing the density in Proposition 5.3, notice that we can easily rescale the parameters after optimization. Indeed, we simply have▽:

i. $\kappa_{\Delta=1/252} = \kappa_{\Delta=1} \cdot 252.$

ii. $\theta_{\Delta=1/252} = \theta_{\Delta=1}.$

iii. $\sigma_{\Delta=1/252} = \sigma_{\Delta=1} \cdot \sqrt{252}.$

As logarithms are strictly increasing functions, maximizing the likelihood is equivalent to maximizing the log-likelihood

$$\begin{aligned}
\ell_T(\boldsymbol{\zeta}) &= \log(\mathcal{L}_T(\boldsymbol{\zeta})) \\
&= \log\left(\prod_{t=1}^{T} f(r_i \mid r_{i-1}; \boldsymbol{\zeta})\right) \\
&= \sum_{i=1}^{T} \log f(r_i \mid r_{i-1}; \boldsymbol{\zeta}) \\
&= \sum_{i=1}^{T} \log\left(c \cdot \exp(-u_{t-1} - v_t)\left(\frac{v_{t-1}}{u_{t-1}}\right)^{q/2} \mathcal{I}_q\left(2\sqrt{u_{t-1}v_t}\right)\right).
\end{aligned} \tag{16}$$

The maximum-likelihood estimator $\boldsymbol{\zeta}^*$ of the parameter vector $\boldsymbol{\zeta}$ is found by maximizing the log-likelihood function in Equation 16 over its parameter space

$$\boldsymbol{\zeta}^* := (\kappa^*,\ \theta^*,\ \sigma^*)^\top = \underset{\boldsymbol{\zeta}}{\operatorname{argmax}}\, \ell_T(\boldsymbol{\zeta}),$$

or equivalently,

$$\boldsymbol{\zeta}^* := (\kappa^*,\ \theta^*\ ,\ \sigma^*)^\top = \underset{\boldsymbol{\zeta}}{\operatorname{argmin}}\, -\ell_T(\boldsymbol{\zeta}).$$

To maximize the log-likelihood function in Equation 16, evaluation of the the modified Bessel function of the first kind is required. This is a difficult task, especially when parameter values are extremely small. Direct usage of the modified Bessel function of the first kind will often yield highly imprecise results by its rapid divergence, which optimization routines struggle with [31, p. 4]. To (somewhat) counteract divergence complications, a exponentially scaled Bessel function $\mathcal{I}_q^*(x) = \mathcal{I}_q(x)\exp(-x)$ which dampens the function will be used[4]. Substituting into Equation 16 and adjusting for the exponential term yields the altered log-likelihood function

$$\ell_T(\boldsymbol{\zeta}) = T\log c + \sum_{i=1}^{T} -u_{t-1} - v_t + \frac{q}{2}\log\left(\frac{v_t}{u_{t-1}}\right) + \log\mathcal{I}_q^*\left(2\sqrt{u_{t-1}v_t}\right) + \underbrace{2\sqrt{u_{t-1}v_t}}_{\text{correction}}.$$

From results presented in Theorem 5.3, explicit formulae for the first two (conditional) moments of the CIR process are known. One, and many, have argued that a fitting approximation for numerically unstable non-central χ^2 distribution is a Gaussian moment-matched distribution. However, from [28, p. 450], this approximation is only fitting asymptotically as the non-centrality parameter approaches ∞. Indeed, the CIR process tends to have a strong affinity for the area around the origin [2, p. 4].

Unconstrained Optimization From the definition of the CIR process, κ, θ and σ have to be strictly positive. To obtain a unconstrained optimization problem, we re-parametrize to working parameters by defining

$$\kappa = \exp(\kappa^\star), \quad \theta = \exp(\theta^\star), \quad \sigma = \exp(\sigma^\star),$$

where we optimize over $\kappa^\star$, $\theta^\star$ and $\sigma^\star$. As the exponential function is strictly positive, this will assure $\kappa, \theta, \sigma > 0$. After the unconstrained estimation, simply take logarithms to achieve the constrained estimate by monotonicity of the functions.

Standard Error Estimation Let $\boldsymbol{\zeta} = (\kappa,\ \theta,\ \sigma)^\top$ denote the parameter vector of the CIR model. To ensure the positivity of each parameter, estimation is conducted over the transformed parameter vector $\boldsymbol{\eta} = (\log\kappa, \log\theta, \log\sigma)^\top$. The maximum-likelihood estimator, $\boldsymbol{\eta}^*$, is obtained by minimizing the negative log-likelihood function. Under standard regularity conditions, the MLE satisfies the asymptotic normality property, yielding

$$\sqrt{T}(\boldsymbol{\eta}^* - \boldsymbol{\eta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \mathcal{I}^{-1}(\boldsymbol{\eta}_0)\right),$$

[4]This is not an ideal solution and research into the area is limited. Ideally, a switching regime that uses asymptotic forms that can be found in [1, p. 390] of the modified Bessel function of the first kind, as it will constantly swap between extremely small and large inputs, would possibly have been beneficial. We will discuss this and its implication for the thesis.

where $\mathcal{I}(\boldsymbol{\eta}_0)$ is the Fisher information matrix evaluated at the true parameter value $\boldsymbol{\eta}_0$. In practice, $\mathcal{I}(\boldsymbol{\eta}_0)$ is approximated by the observed information matrix, which is taken to be the negative of the Hessian matrix H of the log-likelihood function evaluated at $\boldsymbol{\eta}^*$. The asymptotic covariance matrix of $\boldsymbol{\eta}^*$ is then estimated by H^{-1}, and the standard errors are given by

$$\mathrm{SE}(\eta_i^*) = \sqrt{[H^{-1}]_{ii}}, \quad i = 1, 2, 3.$$

To recover the standard errors for the original parameter vector $\boldsymbol{\zeta}$, we apply the delta method. Since $\zeta_i = \exp(\eta_i)$, it follows that

$$\mathrm{SE}(\zeta_i^*) = \exp(\eta_i^*) \cdot \mathrm{SE}(\eta_i^*), \quad i = 1, 2, 3.$$

This yields consistent estimates for the standard errors of κ, θ, and σ on their original scale.

5.2 Hidden Markov Models

A N-dimensional (or N-state) hidden Markov model (HMM) assumes that the distribution of the observed response variable r_t depends exclusively on a hidden state $S_t \in \mathcal{S}$, where $\mathcal{S} = \{S_t : t = 1, 2, \ldots, T\}$ is modelled by a discrete time N-state Markov chain, meaning, S_t satisfies the Markov property

$$\mathbb{P}(S_t = j \mid S_{t-1} = i, \ldots, S_1 = k) = \mathbb{P}(S_t = j \mid S_{t-1} = i),$$

where $S_t \in \mathcal{S}$ is the state at time $t = 1, 2, 3, \ldots, T$ and $\mathcal{S}$ is the *state space*. We will assume time-homogeneity of the Markov chain throughout this paper. The assumption of time-homogeneity of the Markov chain gives rise to the *state transition probabilities* in the $N \times N$ *transition probability matrix* (t.p.m.) Γ as

$$\boldsymbol{\Gamma} = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1N} \\ \vdots & \ddots & \vdots \\ \gamma_{N1} & \cdots & \gamma_{NN} \end{bmatrix}, \quad \gamma_{ij} = \mathbb{P}(S_{t+1} = j \mid S_t = i) \in [0, 1], \quad \sum_{j \in \mathcal{S}} \gamma_{ij} = 1, \tag{17}$$

for all $i, j = 1, \ldots, N$, where γ_{ij} denotes the probability of transitioning from state i at time-t to state j at time-$t+1$, where the assumption of time-homogeneity is seen in action by the fact that the transition probabilities do not depend on the time index. The unconditional probabilities of the state process refer to the probability of the process being in state i at time-t—unconditional

of all previous states of the process. These are summarized in the row vector of probabilities

$$\boldsymbol{\delta}^{(t)} = \underbrace{\begin{bmatrix} \mathbb{P}(S_t = 1) & \cdots & \mathbb{P}(S_t = N) \end{bmatrix}}_{1 \times N}, \tag{18}$$

where the number of probabilities equals the number of states of the Markov chain. We let $\boldsymbol{\delta}^{(1)}$ denote the *initial distribution* of the Markov chain, which provides the probabilities of the process being in the different states at time-1. This allows for a convenient and surprisingly useful result.

Theorem 5.4. *Let $\boldsymbol{\delta}^{(t)}$ be defined as in Equation 18. All future distributions of the Markov chain can then be found by*

$$\boldsymbol{\delta}^{(t+1)} = \boldsymbol{\delta}^{(t)}\boldsymbol{\Gamma} = \boldsymbol{\delta}^{(1)}\boldsymbol{\Gamma}^{(t)}.$$

We now turn our attention to the *stationary distribution*. A Markov chain with a t.p.m. $\boldsymbol{\Gamma}$ is said to have stationary distribution $\boldsymbol{\delta}$, a row vector with non-negative elements, if

$$\boldsymbol{\delta}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}, \quad \boldsymbol{\delta}\mathbf{1}^\top = 1, \tag{19}$$

where $\mathbf{1}$ is a vector with entries 1. The first of the requirements in Equation 19 expresses the stationarity, i.e. moving forward in time is independent of the t.p.m., $\boldsymbol{\Gamma}$. The second is the requirement that $\boldsymbol{\delta}$ is indeed a probability distribution. Consequently, a Markov chain started from its stationary distribution will continue to have that distribution at all subsequent time points and we shall refer to such a process as a stationary Markov chain [45, p. 17]. Intuitively, a stationary distribution reflects the long-term proportion of time the model spends in each state.

To find the stationary distribution, one can obtain a explicit expression by solving the following system of equations [45, p. 18]

$$\left(\boldsymbol{I}_N - \boldsymbol{\Gamma} + \mathbf{1}_{N\times N}\right)^\top \boldsymbol{\delta}^\top = \mathbf{1}_N^\top, \tag{20}$$

where $\boldsymbol{I}_N$ is a N-dimensional identity matrix, $\mathbf{1}_{N\times N}$ is a $N \times N$-dimensional matrix filled with 1's and $\mathbf{1}_N$ is a vector filled with 1's.

When the transition probabilities are time-varying (i.e. functions of covariates), the stationary distribution does not exist [38, p. 14]. However, for fixed covariate values, a single transition probability matrix can be determined, allowing for the computation of a stationary distribution. Throughout we will assume stationarity of the Markov chain. This is adequate as the considered data has a extremely long run time with almost 10000 observations. Furthermore, computational cost is currently extremely daunting which is alleviated by assuming stationarity as it will be evident in Proposition 5.4. We use Equation 20 throughout the code after fitting to find the stationary distribution to ease computational drag.

The univariate observed response variable is a state-dependent process, $\{r_t : t = 1, 2, 3, \ldots, T\}$, the short rate process. The short rate process is a noisy observation process in the sense that it is assumed to be produced by a underlaying state process, $\{S_t : t = 1, 2, 3, \ldots, T\}$. The distribution of r_t is conditionally independent of previous observations and all states except the current hidden state:

$$f(r_t \mid r_{t-1}, \ldots, r_0, S_t, S_{t-1}, \ldots, S_1) = f(r_t \mid S_t), \quad t = 1, 2, 3, \ldots, T,$$

where f denotes a probability density function. Note that we do not say that $r_t \mid r_s$, $t > s$ are unconditionally independent. The structure of a regular hidden Markov model can be seen in Figure 5.2.

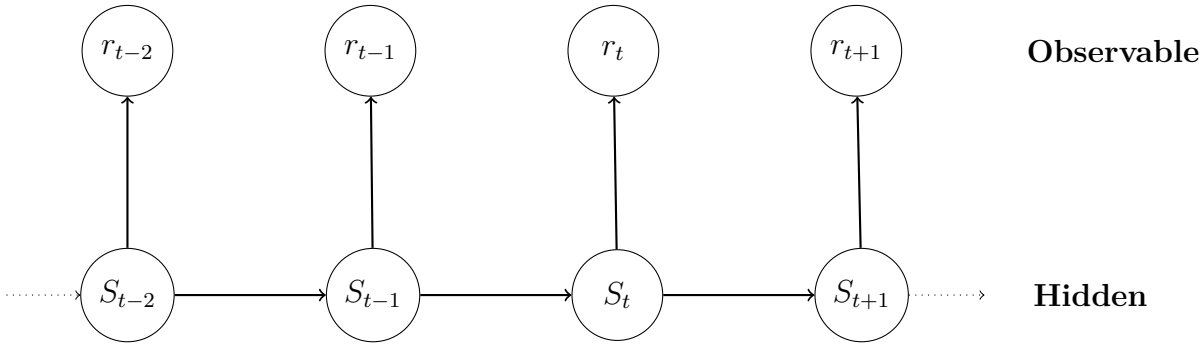


Figure 4: *A hidden Markov Model.*

The Markov chain induces dependence in the state-dependent process, meaning, the observations are independent of each other within states. Extensions of the regular hidden Markov model will be discussed and used throughout the paper. We will allow for some serial dependency.

Simulating a State Process To see why one would think a HMM would be beneficial in the modeling of interest rate consider the following:
The Euler discretization scheme with a incorporated N-state sequence for the CIR process is given as

$$\hat{r}_{t+\Delta} - \hat{r}_t = \kappa_i(\theta_i - \hat{r}_t^+)\Delta + \sigma_i\sqrt{\hat{r}_t^+\Delta}Z^{\mathbb{Q}}, \quad i \in \{1, 2, \ldots, N\}, \tag{21}$$

with $Z^{\mathbb{Q}} \sim \mathcal{N}(0, 1)$. With parameters values seen in Table 3 we simulate a extended CIR process via a 5-state Markov chain.

One can intuitively see in Figure 5, that the CIR model extended via a state sequence does capture the chaotic nature of the short rate seen in Figure 1 better than the standard CIR model seen in Figure 3, relatively speaking[5].

[5]Code to simulate and fit the extended CIR 5-state hidden Markov model is in A.1 for the extra keen reader, as the simulation is extremely slow.

Parameter	Value
$\boldsymbol{\kappa}$	$\begin{bmatrix} 0.04000 \\ 0.02000 \\ 0.02500 \\ 0.00100 \\ 0.00500 \end{bmatrix}$
$\boldsymbol{\theta}$	$\begin{bmatrix} 0.00100 \\ 0.01000 \\ 0.04000 \\ 0.05000 \\ 0.09000 \end{bmatrix}$
$\boldsymbol{\sigma}$	$\begin{bmatrix} 0.00400 \\ 0.00300 \\ 0.00100 \\ 0.00300 \\ 0.00400 \end{bmatrix}$
$\boldsymbol{\delta}^{(1)}$	$\begin{bmatrix} 0.20000 \\ 0.20000 \\ 0.20000 \\ 0.20000 \\ 0.20000 \end{bmatrix}$
$\boldsymbol{\Gamma}$	$\begin{bmatrix} 0.99379 & 0.00103 & 0.00103 & 0.00103 & 0.00103 \\ 0.00103 & 0.99379 & 0.00103 & 0.00103 & 0.00103 \\ 0.00103 & 0.00103 & 0.99379 & 0.00103 & 0.00103 \\ 0.00103 & 0.00103 & 0.00103 & 0.99379 & 0.00103 \\ 0.00103 & 0.00103 & 0.00103 & 0.00103 & 0.99379 \end{bmatrix}$

Table 3: *CIR 5-state Markov chain simulation parameters.*

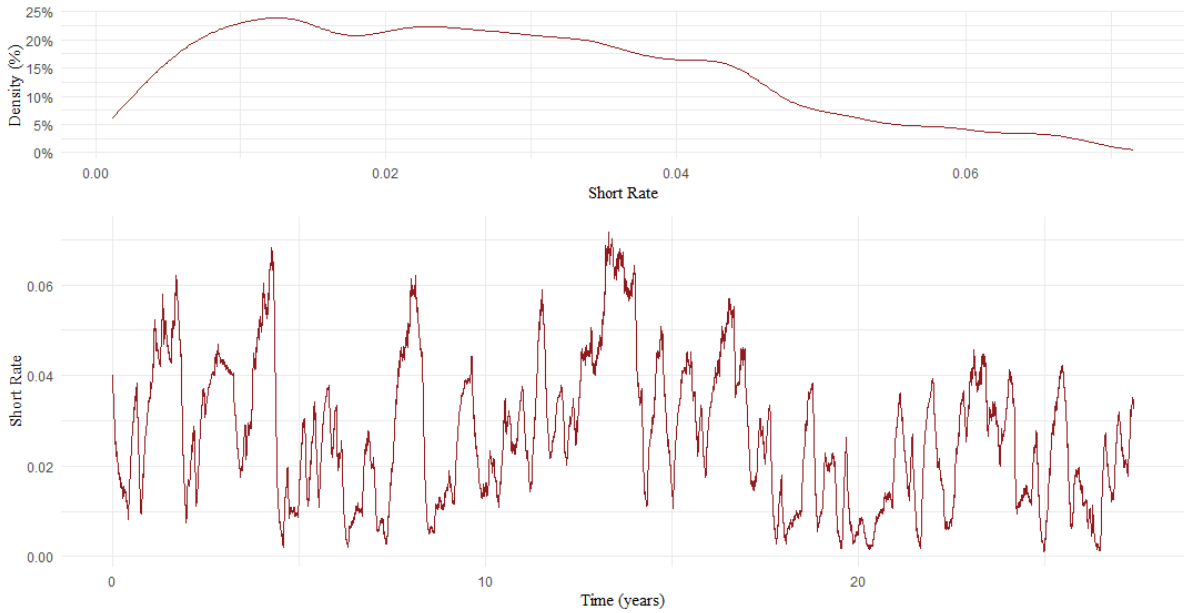


Figure 5: *A realization of the interest rate simulated under the CIR model extended via a 5-state Markov chain using the Euler Scheme with parameters* $\Delta = 1$ *and* $n = 10000$.

5.2.1 State-Dependent Distributions

The state-dependent distributions are the probability density functions of r_t given some state i at time-t:

$$f_i(r_t) := f(r_t \mid S_t), \quad i = 1, \ldots, N.$$

If the state process is stationary, the unconditional distribution of r_t can be given by

$$f(r_t) \overset{\dagger}{=} \sum_{i \in \mathcal{S}} f(r_t, S_t = i) \sum_{i \in \mathcal{S}} f(r_t \mid S_t = i) f(S_t = i) = \sum_{i \in \mathcal{S}} \delta_i^{(t)} f_i(r_t) \overset{\dagger\dagger}{=} \sum_{i \in \mathcal{S}} \delta_i f_i(r_t), \tag{22}$$

where † follows from the law of total probability and †† by stationarity.

As we don't have a explicit expression for the CIR solution marginal distribution, only the transition probability, we shall define a class of HMM's, namely, autoregressive hidden Markov models. Note that the CIR process is among one of the very few cases of diffusion models, where a closed-form expression for the transition probability is known. Furthermore, as we have explored, the data is highly correlated. As such, a natural choice is using a autoregressive model as one might imagine that a regular HMM will not account for the empirical autocorrelation.

Parameter Count As all the parameters are now defined for the HMM we proceed to count the number of parameters to be estimated. The state process is characterized by $\boldsymbol{\delta}$ and $\boldsymbol{\Gamma}$. The latter has $N \times (N-1) = N^2 - N$ free parameters according to the row sum constraint (last equality of Equation 17). For a stationary Markov chain, we need not estimate the initial distribution as this simply equals the stationary distribution, which otherwise would yield N extra parameters to be estimated in each model (see Equation 18). As previously stated, we simply use Equation 20 to find the stationary distribution estimates after estimations of the transition probabilities. By the assumption of conditional independence, the state-dependent process is characterized by the state-dependent distributions which require $3N$ parameters (κ, θ and σ) to be estimated for our N-state HMM. In total we have to estimate

$$\#\text{Parameters}_3 = N^2 - N + 3N = N^2 + 2N.$$

If, however, we wanted to model 2 of the 3 parameters κ, θ and/or σ as state-dependent, then we would have a total of

$$\#\text{Parameters}_2 = N^2 + 2N - (N-1) = N^2 + N + 1,$$

parameters to be estimated. Lastly modelling 1 of the 3 parameters κ, θ and/or σ as state-dependent yields

$$\#\text{Parameters}_1 = N^2 + 2N - (N-1) - (N-1) = N^2 + 2,$$

parameters to be estimated. For example, assume that we have one state-dependent variable. This yields $2^2 + 2 = 6$ parameters to be estimated for the simplest case, $N = 2$, but a whopping $5^5 + 2 \cdot 5 = 27$ parameters to be estimated for the more complicated case, $N = 5$. See Figure 6 for a visualization.

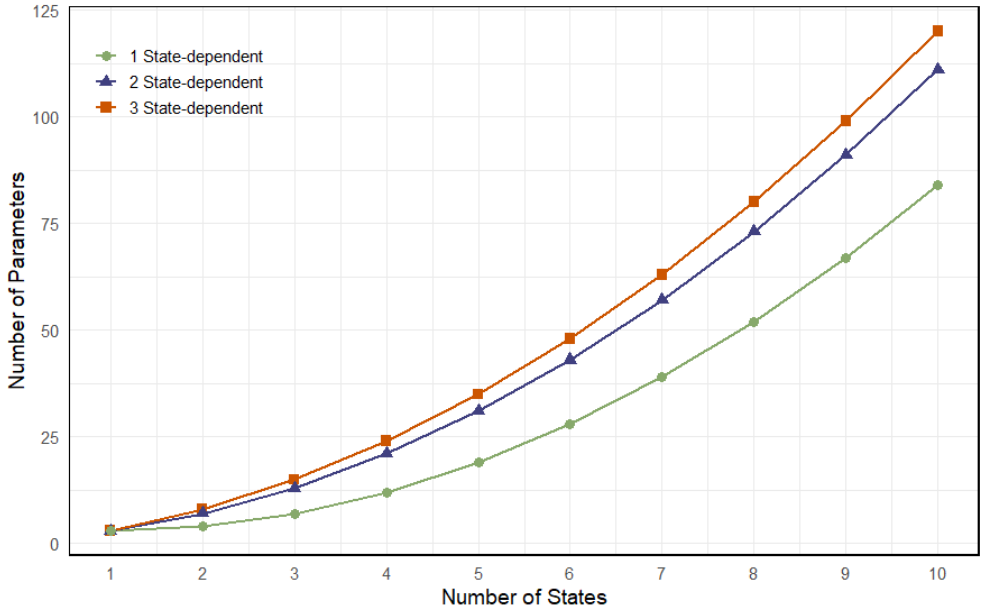


Figure 6: *Number of states visualized for some number of states, N, and some number of state-dependent parameters, κ, θ and/or σ.*

5.2.2 Likelihood Estimation

The likelihood of a hidden Markov model has a convenient recursive form which is seen in the next result. Throughout, let the vector of parameters for estimation be denoted by $\boldsymbol{\zeta} = (\boldsymbol{\Gamma}, \boldsymbol{\kappa}, \boldsymbol{\theta}, \boldsymbol{\sigma})^\top$.

Proposition 5.4. *[45, Prop. 1] Let $\{S_t\}_{t=1}^T$ be a homogeneous, finite-state Markov chain on $\{1, 2, \ldots, N\}$, with transition probability $\boldsymbol{\Gamma} = (\gamma_{ij})_{i,j=1}^N$. Consider a observation process $\{r_t\}_{t=1}^T$. The joint likelihood of observing $\{r_t\}_{t=1}^T$ is then given by*

$$\mathcal{L}_T(\boldsymbol{\zeta}) = \boldsymbol{\delta}^{(1)} \boldsymbol{P}(r_1) \boldsymbol{\Gamma} \boldsymbol{P}(r_2) \boldsymbol{\Gamma} \boldsymbol{P}(r_3) \cdots \boldsymbol{\Gamma} \boldsymbol{P}(r_T) \mathbf{1}^\top,$$

where $\boldsymbol{\delta}^{(1)}$ is the initial distribution, $\boldsymbol{P}(r)$ is the diagonal matrix with the state-dependent distribution $f_1(r)$, $f_2(r)$, $\ldots$, $f_N(r)$ given in Equation 22 as elements and $\boldsymbol{\Gamma}$ is the t.p.m.. If $\boldsymbol{\delta}^{(1)}$ is the

stationary distribution $\boldsymbol{\delta}$ of the Markov chain, then in addition

$$\mathcal{L}_T(\boldsymbol{\zeta}) = \boldsymbol{\delta}\boldsymbol{\Gamma}\boldsymbol{P}(r_1)\boldsymbol{\Gamma}\boldsymbol{P}(r_2)\boldsymbol{\Gamma}\boldsymbol{P}(r_3)\cdots\boldsymbol{\Gamma}\boldsymbol{P}(r_T)\mathbf{1}^\top.$$

The recursive nature of the likelihood in Proposition 5.4 enables computationally efficient evaluation through numerical optimization. The likelihood is maximized using direct numerical methods, leveraging the forward algorithm (which we define in a second). This approach utilizes the forward probabilities, defined for $t = 1, 2, \ldots, T$ and $j \in \mathcal{S}$ as follows:

$$\alpha_t(j) = f(r_1, \ldots, r_T, S_t = j), \quad \boldsymbol{\alpha}_t = \big[\alpha_t(1) \ \ldots \ \alpha_t(N)\big]. \tag{23}$$

In other words, the forward probabilities contain information on the likelihood of the observations up to and including time-t. Consequently, Equation 23 allows us to write the likelihood from Proposition 5.4 as

$$\mathcal{L}_T(\boldsymbol{\zeta}) = f(r_1, r_2, \ldots, r_T) \overset{\dagger}{=} \sum_{j\in\mathcal{S}} f(r_1, r_2, \ldots, r_T, S_t = j) = \sum_{j\in\mathcal{S}} \alpha_t(j),$$

where † follows from the law of total probability. The probability of the Markov chain occupying state-$j \in \mathcal{S}$ at different times-t, is its proportion of the forward probability at time-t for state j:

$$f(S_t = j \mid r_1, \ldots, r_t) = \frac{f(S_t = j, r_1, \ldots, r_t)}{f(r_1, \ldots, r_t)} = \frac{\alpha_t(j)}{\sum_{i=1}^N \alpha_t(i)}.$$

We can then state the (row) vector of forward probabilities for $t = 1, 2, \ldots, T$ as

$$\boldsymbol{\alpha}_t = \boldsymbol{\delta}\boldsymbol{P}(r_1)\boldsymbol{\Gamma}\boldsymbol{P}(r_2)\cdots\boldsymbol{\Gamma}\boldsymbol{P}(r_t) = \boldsymbol{\delta}\boldsymbol{P}(r_1)\prod_{s=2}^{t}\boldsymbol{\Gamma}\boldsymbol{P}(r_s),$$

with the convention that an empty product is the identity matrix [45, p. 38]. Assembling, Proposition 5.4 states that $\mathcal{L}_T(\boldsymbol{\zeta}) = \boldsymbol{\alpha}_T\mathbf{1}^\top$ and for $t \geq 2$ we defined $\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\boldsymbol{P}(r_t)$. This allows us to define the *forward algorithm*:

$$\begin{aligned}
&\boldsymbol{\alpha}_1 = \boldsymbol{\delta}\boldsymbol{P}(r_1);\\
&\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\boldsymbol{P}(r_t), \quad \text{for } t = 2, 3, \ldots, T;\\
&\mathcal{L}_T = \boldsymbol{\alpha}_T\mathbf{1}^\top.
\end{aligned}$$

Note, for a N-state HMM, $\boldsymbol{\delta}$ has N elements, $\boldsymbol{P}(r_t)$ has N elements (all in the diagonal) and $\boldsymbol{\Gamma}$ $N \times N$ elements. For the forwards algorithm, this implies that $\boldsymbol{\alpha}_t$ is a sum of N products consisting of a previous iteration, $\boldsymbol{\alpha}_{t-1}$, a transition probability γ_{ij} and a state-dependent probability $f_i(r_t)$, $i \in \mathcal{S}$. Hence, for each $t \in \{1, 2, \ldots, T\}$, there are N elements to be computed of $\boldsymbol{\alpha}_t$. Finally,

this implies that the number of operations to calculate the likelihood of T observations is of order TN^2.

Scaling the Likelihood Let $\mathcal{L}_t$ denote the likelihood of the observations up to time t under a fixed parameter specification $\boldsymbol{\zeta}$ of a HMM. Then, under suitable regularity conditions, there exists a constant $h \in \mathbb{R}$ such that (see [33])

$$\lim_{t\to\infty} \frac{1}{t} \log \mathcal{L}_t(\boldsymbol{\zeta}) = h, \quad \text{a.s..}$$

In particular,

- if $h < 0$, the likelihood $\mathcal{L}_t(\boldsymbol{\zeta})$ converges to 0 exponentially fast as $t \to \infty$;
- if $h > 0$, the likelihood $\mathcal{L}_t(\boldsymbol{\zeta})$ diverges to ∞ exponentially fast as $t \to \infty$.

I.e. the likelihood approaches either 0 or ∞ a.s., exponentially fast. This is highly problematic as our model is already susceptible to numerical overflow complications.
As such, observe firstly from Proposition 5.4, that the HMM likelihood is a product of matrices and not scalars. Consequently, it is not possible to circumvent numerical underflow by computing the logarithm of the likelihood as the sum of logarithms of its factors. Therefore, we adapt the method used by [45, p. 48] (although heavily inspired by [17, p. 78]): For $t = 1, \dots, T$ define the standardised vector of forward probabilities at time-t as:

$$\boldsymbol{\phi}_t = \frac{\boldsymbol{\alpha}_t}{\boldsymbol{\alpha}_t \mathbf{1}^\top} = \frac{\boldsymbol{\alpha}_t}{\sum_{j\in\mathcal{S}} \alpha_t(j)}, \quad \boldsymbol{\phi}_t = [\phi_t(1) \dots \phi_t(N)], \quad \sum_{j\in\mathcal{S}} \phi_t(j) = 1.$$

This is yields the normalized forward probabilities, which are far less susceptible to numerical underflow:
For $t = 1$:

$$\boldsymbol{\phi}_1 = \frac{\boldsymbol{\alpha}_1}{\boldsymbol{\alpha}_1 \mathbf{1}^\top} = \frac{\boldsymbol{\delta}_0 \boldsymbol{P}(r_1)}{\boldsymbol{\delta}_0 \boldsymbol{P}(r_1) \mathbf{1}^\top}.$$

For $t = 2, \dots, T$:

$$\boldsymbol{\phi}_t = \frac{\boldsymbol{\alpha}_t}{\boldsymbol{\alpha}_t \mathbf{1}^\top} = \frac{\boldsymbol{\alpha}_{t-1} \boldsymbol{\Gamma} \boldsymbol{P}(r_t)}{\boldsymbol{\alpha}_{t-1} \boldsymbol{\Gamma} \boldsymbol{P}(r_t) \mathbf{1}^\top} = \frac{\boldsymbol{\alpha}_{t-1} \boldsymbol{\Gamma} \boldsymbol{P}(r_t) / (\boldsymbol{\alpha}_{t-1} \mathbf{1}^\top)}{\boldsymbol{\alpha}_{t-1} \boldsymbol{\Gamma} \boldsymbol{P}(r_t) \mathbf{1}^\top / (\boldsymbol{\alpha}_{t-1} \mathbf{1}^\top)} = \frac{\boldsymbol{\phi}_{t-1} \boldsymbol{\Gamma} \boldsymbol{P}(r_t)}{\boldsymbol{\phi}_{t-1} \boldsymbol{\Gamma} \boldsymbol{P}(r_t) \mathbf{1}^\top}.$$

I.e. scalar multiplication as opposed to matrix multiplication. To see why this is actually the case, we derive the likelihood, $\mathcal{L}_T(\boldsymbol{\zeta})$, in terms of $\boldsymbol{\phi}$ instead of $\boldsymbol{\alpha}$.

Firstly, using $\boldsymbol{\alpha}_0 = \boldsymbol{\delta}$, note that

$$\mathcal{L}_T(\boldsymbol{\zeta}) = \boldsymbol{\alpha}_T \mathbf{1}^\top = \frac{\boldsymbol{\alpha}_1 \mathbf{1}^\top}{\boldsymbol{\alpha}_0 \mathbf{1}^\top} \frac{\boldsymbol{\alpha}_2 \mathbf{1}^\top}{\boldsymbol{\alpha}_1 \mathbf{1}^\top} \cdots \frac{\boldsymbol{\alpha}_T \mathbf{1}^\top}{\boldsymbol{\alpha}_{T-1} \mathbf{1}^\top} = \prod_{t=1}^{T} \frac{\boldsymbol{\alpha}_t \mathbf{1}^\top}{\boldsymbol{\alpha}_{t-1} \mathbf{1}^\top}, \tag{24}$$

where $\frac{\boldsymbol{\alpha}_t \mathbf{1}^\top}{\boldsymbol{\alpha}_{t-1} \mathbf{1}^\top} \in \mathbb{R}$. This allows us to find the log-likelihood function using Equation 24

$$\begin{aligned}
\ell_T(\boldsymbol{\zeta}) &= \log \mathcal{L}_T(\boldsymbol{\zeta}) \\
&= \log \prod_{t=1}^{T} \frac{\boldsymbol{\alpha}_t \mathbf{1}^\top}{\boldsymbol{\alpha}_{t-1} \mathbf{1}^\top} \\
&= \sum_{t=1}^{T} \log \left(\frac{\boldsymbol{\alpha}_t \mathbf{1}^\top}{\boldsymbol{\alpha}_{t-1} \mathbf{1}^\top} \right) \\
&= \log \left(\frac{\boldsymbol{\alpha}_1 \mathbf{1}^\top}{\boldsymbol{\alpha}_0 \mathbf{1}^\top} \right) + \sum_{t=2}^{T} \log \left(\frac{\boldsymbol{\alpha}_t \mathbf{1}^\top}{\boldsymbol{\alpha}_{t-1} \mathbf{1}^\top} \right) \\
&= \log \left(\frac{\boldsymbol{\delta} \boldsymbol{P}(r_1) \mathbf{1}^\top}{\boldsymbol{\delta} \mathbf{1}^\top} \right) + \sum_{t=2}^{T} \log \left(\frac{\boldsymbol{\alpha}_{t-1}}{\boldsymbol{\alpha}_{t-1} \mathbf{1}^\top} \boldsymbol{\Gamma} \boldsymbol{P}(r_t) \mathbf{1}^\top \right) \\
&= \log \left(\boldsymbol{\delta} \boldsymbol{P}(r_1) \mathbf{1}^\top \right) + \sum_{t=2}^{T} \log \left(\boldsymbol{\phi}_{t-1} \boldsymbol{\Gamma} \boldsymbol{P}(r_t) \mathbf{1}^\top \right),
\end{aligned}$$

which is exactly stating that the log-likelihood is a sum of logarithmic-values.

Furthermore, we implement working parameters to address constraints of positivity for the CIR model parameters and row sums equaling one in the t.p.m..
For the rest of this paragraph, denote by $\hat{\cdot}$ the estimator of some parameter $\cdot$.
Assume, for example, $N = 3$. Firstly, set the working parameters $\eta_i = \log \rho_i$ for some parameter ρ_i. After we have maximized the likelihood with respect to the unconstrained parameters, the constrained parameter estimates can be obtained by transforming back: $\hat{\rho}_i = e^{\hat{\eta}_i}$. Next, start by defining the matrix with entries $\tau_{ij} \in \mathbb{R}$

$$\mathbf{T} = \begin{pmatrix} - & \tau_{12} & \tau_{13} \\ \tau_{21} & - & \tau_{23} \\ \tau_{31} & \tau_{32} & - \end{pmatrix},$$

and $g : \mathbb{R} \to \mathbb{R}^+$ (strictly increasing) function e^x. Define

$$\nu_{ij} = \begin{cases} g(\tau_{ij}) & \text{for } i \neq j \\ 1 & \text{for } i = j \end{cases},$$

and

$$\gamma_{ij} = \frac{\nu_{ij}}{\sum_{k=1}^{N} \nu_{ik}}, \quad i, j = 1, 2, \ldots, N,$$

and $\mathbf{\Gamma} = (\gamma_{ij})_{i,j=1}^{N}$. We perform the calculation of the likelihood-maximizing parameters in two steps:

I. Maximize $\mathcal{L}_T$ with respect to the working parameters $\mathbf{T} = \{\tau_{ij}\}$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_N)^\top$ which are all unconstrained by construction.

II. Transform the estimates of the working parameters to estimates of the natural parameters:

$$\widehat{\mathbf{T}} \to \widehat{\mathbf{\Gamma}}, \quad \widehat{\boldsymbol{\eta}} \to \widehat{\boldsymbol{\rho}}.$$

Consider $\mathbf{\Gamma}$ for the case $g(x) = \exp(x)$ and general N. Here we have

$$\gamma_{ij} = \frac{\exp(\tau_{ij})}{1 + \sum_{k \neq i} \exp(\tau_{ik})}, \quad i \neq j,$$

and the diagonal elements of $\mathbf{\Gamma}$ follow from the row sums of 1. The transformation in the opposite direction is

$$\tau_{ij} = \log\left(\frac{\gamma_{ij}}{1 - \sum_{k \neq i} \gamma_{ik}}\right) = \log(\gamma_{ij}/\gamma_{ii}), \quad \text{i} \neq j.$$

5.2.3 Number of States

HMM's are prone to overfitting [24, p. 2]. That is, HMM's are not well suited for order estimation as small variations in the data are known to cause such models to overestimate the number of groups as well as the frequency of transitions when the number of states is unknown. This begs the question; How does one adequately choose the number of states in a (AR)HMM?
In HMM's, the number of states must be specified a priori to the analysis rather than estimated during model fitting by the above-mentioned reason. However, this decision can be challenging, as standard model selection criteria like AIC and BIC often favour a large number of states, which can reduce interpretability[6]. In particular, AIC tends to select more states as increased model flexibility allows for better data fitting, though this can come at the expense of generalizability and interpretability. In other words, we might misspecify some variation in the data as an extra false state-i, as it gives a better model fitting, even though it might just be a (large) variation within a true state-j. [41] and [38, p. 4] highlighted this issue, recommending that the choice of

[6]This will become evident in Section 5.3.1.

states should be guided by domain expertise and model validation rather than relying solely on selection criteria. As such, our information criteria, AIC and BIC, will also be utilized for model selection, but not exclusively. We describe the information criteria in Section 5.3.1.

A proposed solution to the a priori number of state selection, is the layman method of counting modes in the distribution of the data. However, this can be severely problematic. For example, assume the known number of states is two. If the means are approximately equal but the variance differ it is virtually impossible by visual examination to determine that the number of states is one or some larger integer.

Following [41] and [38], we determine the number of states based on domain expertise only. However, we note that the application of HMM's to financial contexts—and especially to interest rates—remains extremely limited. In fact, our research has found no literature on the topic. Consequently, we must develop our own arguments to justify the chosen number of states based on domain expertise.

In the context of macroeconomic analysis, particularly for modelling interest rate regimes using the CIR model, selecting between 2 and 5 states seems appropriate. A two-state model typically distinguishes between stable and volatile interest rate environments, while additional states can capture more nuanced dynamics such as shifts in monetary policy, periods of financial crisis, or prolonged episodes of low interest rates.

The number of states in an HMM framework affects the estimation of key CIR parameters—speed of mean reversion, volatility, and the long-term mean. The speed of mean reversion κ determines how quickly interest rates revert to their long-term mean θ. If too many states are selected, κ may be underestimated, as the model attributes rate fluctuations to state transitions rather than reversion dynamics. Conversely, an insufficient number of states may lead to an overestimation of κ, failing to account for distinct macroeconomic phases. Similarly, volatility σ can be misrepresented if states do not properly capture varying levels of uncertainty in financial markets. The effect of different levels of the long-term mean θ is quite obvious as to why it might be of extreme importance to correctly capture the number of states.

Consider now a varying number of states, $N \in \{1, 2, 3, 4, 5\}$. $N = 1$ is trivial as it is our simple non-extended CIR model. $N = 2$ would, as stated, simply be a dichotomic approach; the economy is in two states, expansion and recession. $N = 3$ can be viewed as our classical economical states of business cycles; recession, expansion and recovery. $N = 5$ would have 2 more states than the $N = 3$ case which could be interpreted as some middle-ground between some (possibly extreme) expansion and recession business cycles but not quite recovery. It could potentially capture variation in the data, but at some possibility of overfitting because of exactly capturing that variation. For example, during a *non-extreme* recession, interest rates are cut moderately by central banks, which leads to a slight decrease or steady in the mean reversion speed, κ. The long-term mean rate, θ, slightly decreases, and volatility σ increases mildly due to some uncertainty, but overall, the economic stability ensures that interest rates return to normal

levels after a period of low rates.
In contrast, during a *extreme* recession, central banks lower rates aggressively, causing a significant decrease in the mean reversion speed, κ, because the rates stay low for a prolonged period. The long-term mean rate, θ, falls significantly, reflecting a low-growth, low-inflation environment, and volatility, σ, increases sharply due to heightened uncertainty and financial instability, which causes greater fluctuations in interest rates. Thus, in our analysis, we constrain the number of states to this range, ensuring a balance between model fit, interpretability, and macroeconomic relevance. $N = 4$ is difficult to justify in terms of domain expertise but it would be natural to include it as we are already considering number of states on both sides of the integer.

5.2.4 Autoregressive Hidden Markov Models

As demonstrated in Section 4, the observations exhibit strong autocorrelation—likely stronger than what the state process alone can induce. To capture this dependence, we let each observation at time-t depend on the previous observation at time-$t-1$. Figure 7 illustrates this autoregressive extension: the arrows indicate that dependence occurs both among hidden states, S, and between consecutive observations, r. This structure is known as an Autoregressive Hidden Markov Model (ARHMM) of order one[7].

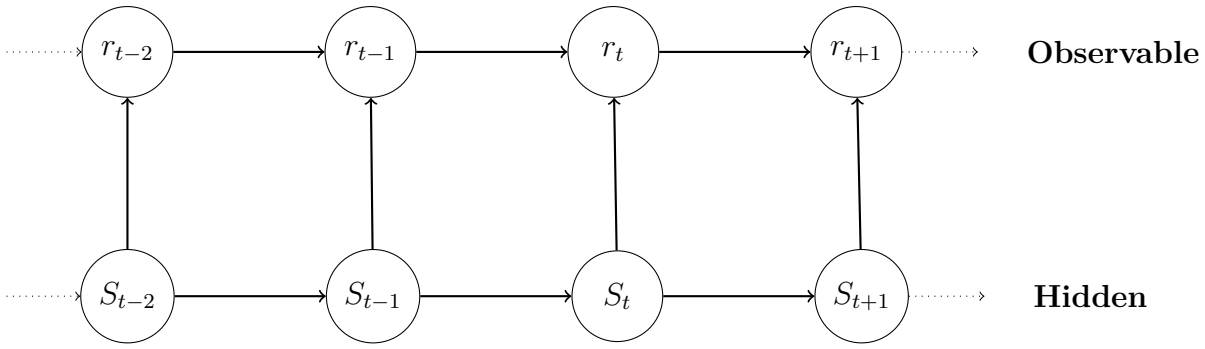


Figure 7: *A Markov-switching AR(1) model. Note that the observation at time t is not conditionally independent of $t-1$ but is of $t-2$, $t-3$ and so forth.*

A natural question now would be: How does the autoregressive nature change the likelihood expression derived in Proposition 5.4? The answer is surprisingly pleasant.
By the law of total probability, the likelihood is computed as

$$\mathcal{L}_T = f(r_1, r_2, \ldots, r_T) = \sum_{j \in \mathcal{S}} f(r_1, r_2, \ldots, r_T, S_T = j),$$

where f denotes a probability density function. The forward variable calculations can now be broken down into two cases:

[7]Or hidden Markov AR(1) model or Markov-switching autoregression model (kært barn har mange navne).

I. For $t = 1$:

$$\alpha_1(j) = f(r_1, S_1 = j) = f(r_1 \mid S_1 = j)\mathbb{P}(S_1 = j) = f_j(r_1)\delta_j,$$

where $f_j(r) = f(r_t = r \mid S_t = j)$ and δ is the initial distribution of the Markov chain. For convenience, we rewrite to matrix notation:

$$\boldsymbol{\alpha}_1 = \boldsymbol{\delta}\boldsymbol{P}(r_1).$$

Computing $\boldsymbol{P}(r_1)$ in the AR(1) case requires a previously unobserved value, say r_0. For the purpose of parameter estimation in the AR(1) case, one can conveniently circumvent this complication by conditioning on the first observation, meaning, we replace $\boldsymbol{P}(r_1)$ by $\boldsymbol{P}(r_0, r_1)$.

II. For $t > 1$: Using the law of total probability and subsequently taking advantage of the conditional independence assumptions yields

$$\begin{aligned}
\alpha_t(j) &= \sum_{i \in \mathcal{S}} f(r_1, \ldots, r_T, S_t = j, S_{t-1} = i) \\
&\overset{\dagger}{=} \sum_{i \in \mathcal{S}} f(r_t \mid r_{t-1}, S_t = j) f(r_{t-1}, \ldots, r_1, S_t = j, S_{t-1} = i) \\
&= \sum_{i \in \mathcal{S}} f_j(r_t \mid r_{t-1}) \mathbb{P}(S_t = j \mid S_{t-1} = i) f(r_{t-1}, \ldots, r_1, S_{t-1} = i) \\
&= \sum_{i \in \mathcal{S}} f_j(r_t \mid r_{t-1}) \gamma_{ij} \alpha_{t-1}(i).
\end{aligned}$$

The main difference with the regular HMM derivations is in †, where $f_j(r_t \mid r_{t-1}) = f(r_t \mid r_{t-1}, S_t = j)$ would simplify further to $f_j(r_t) = f(r_t \mid S_t = j)$. However, the rest of the derivation are unchanged. In matrix form, this gives us:

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\boldsymbol{P}(r_{t-1}, r_t),$$

where $\boldsymbol{P}(r_{t-1}, r_t)$ is the diagonal matrix with j-th diagonal element equal to $f_j(r_t \mid r_{t-1})$. We can use this relationship to find $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_T$, which yields

$$\boldsymbol{\alpha}_T = \boldsymbol{\delta}\boldsymbol{P}(r_0, r_1)\boldsymbol{\Gamma}\boldsymbol{P}(r_1, r_2) \cdots \boldsymbol{\Gamma}\boldsymbol{P}(r_{T-1}, r_T),$$

and so

$$\mathcal{L}_T(\boldsymbol{\zeta}) = \boldsymbol{\delta}\boldsymbol{P}(r_0, r_1)\boldsymbol{\Gamma}\boldsymbol{P}(r_1, r_2) \cdots \boldsymbol{\Gamma}\boldsymbol{P}(r_{T-1}, r_T)\mathbf{1}^\top. \tag{25}$$

Equation 25 states that the only substantive difference lies in the use of $\boldsymbol{P}(r_{t-1}, r_t)$ instead of

$\boldsymbol{P}(r_t)$, as in the standard non-autoregressive HMM. This substitution effectively completes our model, as all previously derived results can be extended to the ARHMM setting by replacing the diagonal elements of the matrix $\boldsymbol{P}(\cdot)$ with the appropriate conditional probability density functions. Nonetheless, as previously noted, we must still provide an explicit specification of the state-dependent distributions.

State-Dependent Distributions As the regular HMM state-dependent distribution framework is established, we proceed to the ARHMM state-dependent distributions. The ARHMM a natural choice as the transition density is available from Proposition 5.3. We allow the distribution parameters to admit a state-dependent speed of mean reversion, κ, long-term mean, θ, and volatility parameter, σ. The time-t conditional state-dependent distributions for $i \in \mathcal{S}$ and $t \in \{1, 2, \ldots, T\}$ for the CIR model extended through an ARHMM is

$$f_i(r_t \mid r_{t-1}; \boldsymbol{\zeta}) = c_i \cdot \exp(-u_{t-1,i} - v_{t,i}) \left(\frac{v_{t,i}}{u_{t-1,i}}\right)^{q_i/2} \mathcal{I}^*_{q_i}\left(2\sqrt{u_{t-1,i} v_{t,i}}\right) \cdot \underbrace{\exp\left(2\sqrt{u_{t-1,i} v_{t,i}}\right)}_{\text{correction}},$$

where

$$\begin{aligned} c_i &= \frac{2\kappa_i}{\sigma_i^2 \left(1 - e \exp\left(-\kappa_i \Delta\right)\right)}, \\ u_{t-1,i} &= c r_{t-1} \exp\left(-\kappa_i \Delta\right), \\ v_{t,i} &= c r_t, \\ q_i &= \frac{2\kappa_i \theta_i}{\sigma_i^2} - 1, \end{aligned}$$

and $\mathcal{I}^*_{q_i}(2\sqrt{u_{t-1,i} v_{t,i}})$ is the exponentially damped modified Bessel function of the first kind of order q_i. Note the double indexation on u and v but only singular indexation on c and q. The index i is for the set of states, i.e. the state-space $\mathcal{S}$ and t is for the set of observation, $t \in \{1, 2, \ldots, T\}$. This completes the model.

Inherent Modelling Error In continuous time, the likelihood of r_t given r_{t-1} depends not only on the state at time t, but also on how the underlying state S_{t-1} evolves throughout the entire interval $[t-1, t]$. However, in many practical implementations of regime-switching models, we make the simplifying assumption that the hidden (regime) state remains constant over each observation interval. This assumption, while common, constitutes an inherent approximation error.

The reasonableness of this approximation is generally justified under the following conditions:

I. The time-steps between observations are small.

II. Regime switches are infrequent.

Let us address both complications:

I. Small Time-Steps: In our context, the dataset described in Section 4 is observed on a daily business frequency, implying that the intervals between observations are relatively short. In the regime-switching and hidden Markov model literature, it is well-established that the error introduced by assuming a constant regime state over small intervals is negligible when the sampling frequency is high [23, 4]. That is, as the time-steps between observations decrease, the likelihood that an unobserved state-switch occurs within a single interval also decreases, mitigating the impact of this approximation on inference and estimation. This property underpins the widespread use of discrete-time regime-switching models for financial and macroeconomic data sampled at high frequency, which our data is.

II. Infrequent Regime Switches: The second justification relies on the empirical observation that regime switches—such as those corresponding to business cycles or structural shifts in economic variables—are typically rare events, spanning months or even years rather than days [23, 14, 3]. In such environments, the probability of multiple regime transitions occurring between two consecutive daily observations is extremely low, which makes the constant-regime approximation reasonable [22]. Additionally, estimated transition probabilities in Markov-switching models applied to macroeconomic and financial time series are typically small, indicating high persistence within each regime.

It is important to recognize, however, that this modelling error becomes more pronounced if the above conditions do not hold: i.e., if data is observed at a low frequency or if regime switches are rapid and frequent. In such cases, ignoring possible intra-interval transitions can bias inference and may underestimate the uncertainty in the latent state path. Alternative modelling approaches—such as continuous-time Markov-switching models or state-space models that explicitly account for state changes within observation intervals—have been developed to address this limitation [18, 10]. These models, while more complex, provide a more accurate representation of the latent process dynamics when the aforementioned assumptions are violated. However, the computational difficulties faced are much greater as we have a large array of states.

We keep these considerations in mind when interpreting our results and when evaluating alternative modelling approaches for the final thesis—such as continuous state-space models that can better accommodate intra-interval regime dynamics.

Numerical Optimization and Local Minima As with most statistical models encountered in practice, there is no closed-form solution for the maximum-likelihood estimate of an (AR)HMM. Therefore, parameter estimation must rely on direct numerical maximization of the likelihood function in Equation 25, or equivalently, minimization of the negative likelihood. In our work, we employ the Nelder-Mead algorithm [39], a derivative-free simplex method for function minimization.

While the likelihood is, in principle, differentiable and amenable to gradient-based methods,

practical challenges arise. The high dimensionality of the parameter space causes the Hessian matrix to grow quadratically with the number of parameters, making Newton-type methods prohibitively memory-intensive. For instance, the CIR model's parameter space (see [31, p. 7-8] for PRIBOR 3M time series applications) is notoriously flat, resulting in near-zero gradients and nearly singular Hessians—issues that are exacerbated by the typically small numerical values of the parameters, as also indicated by our OLS estimates in Equation 15. See Figure A.3.1 for log-likelihood surfaces for varying fixed optimal parameters. Here, we used our 3M T-bill data but fixed one of the parameters to plot the 3D log-likelihood space by lettering the two non-fixed values vary on a grid of values around the OLS from Equation 15.

Given these characteristics, gradient- and Hessian-based methods become unstable or inefficient, prompting the use of Nelder-Mead, which relies solely on function evaluations. The primary drawback, however, is its slow convergence. As an example, fitting the most computationally intensive model in our study—the 5-state autoregressive hidden Markov extension of the CIR model with state-dependent κ, θ, and σ—required approximately 42 hours, even with well-chosen initial values. Gradient- and Hessian-based methods simply do not have enough memory to allocate in `R` (or even by trial in `C`), as the vectors that has to be allocated are astronomically large.

To mitigate the risk of convergence to local minima, we initialize the optimizer from a range of starting values, following the recommendations of [45, p. 53]▽. Initial parameters are selected in the vicinity of the OLS estimates. The final parameter estimates are chosen as those yielding the highest log-likelihood (equivalently, the smallest negative log-likelihood). The optimization terminates when successive iterations fail to improve the objective function by more than $\varepsilon \times (|\mathcal{L}_T| + \varepsilon)$ in a single step, where ε is the relative tolerance. This scale-invariant stopping criterion is especially beneficial when parameters are numerically small, which the existing literature and our preliminary analysis, hints to be the case. $\varepsilon = 10^{-8}$ will be used throughout optimization of any kind.

5.3 Model Selection Criteria & Assessment

5.3.1 Information Criteria: AIC & BIC

Two of the most popular approaches to model selection for HMM's will be used: The Akaike Information Criterion (AIC) and The Bayesian Information Criterion (BIC). These are supplementary methods to those discussed previously.

Assume that $r_1, \ldots, r_T$ were generated by the data generating process, f, and that one is interested in determining which model to choose among two different approximating families $\{g_1 \in \mathcal{G}_1\}$ and $\{g_2 \in \mathcal{G}_2\}$ under some criteria of being "the best". We thus need some operator to determine the lack of fit between the true data generating model and the fitted models, $\Delta(f, \hat{g}_1)$ and $\Delta(f, \hat{g}_2)$. A immediate issue that arises is the lack of knowledge of f. As such, we can not determine from this discrepancy which model to select. However, we can use model selection criteria, $\widehat{\mathbb{E}}_f\big[\Delta(f, \hat{g}_1)\big]$ and $\widehat{\mathbb{E}}_f\big[\Delta(f, \hat{g}_2)\big]$. These quantities bases selection on estimators of the expected discrepancies. The model selection criterion simplifies to the Akaike information criterion [45, p. 98] which, in (dangerously) short, arises of the Kullback–Leibler discrepancy and conditions listed in [34, Appendix A]:

$$\text{AIC} = \underbrace{-2 \log \mathcal{L}_T}_{\text{measure of fit}} + \underbrace{2p}_{\text{penalty}}, \tag{26}$$

where $\mathcal{L}$ is the log-likelihood of the fitted model and p denotes the number of parameters of the model (see Section 5.2.3 for number of parameter determination). It is immediately clear that increasing the number of parameters, by increasing the number of states or state-dependent parameters, will penalize the AIC. To compare model performances in terms of AIC, we follow [8, pp. 270–272] to some degree; Let Δi denote the difference in AIC between the best model (i.e. smallest AIC) and the one of comparison. The rule of thumb then states that we can assess the relative merits of models by:

- $\Delta i \leq 2 \Rightarrow$ Substantial support (evidence).
- $4 \leq \Delta i \leq 7 \Rightarrow$ Considerably less support (evidence).
- $\Delta i > 10 \Rightarrow$ Essentially no support (evidence).

Note, that [9] relaxed the rule of thumb and thus $2 \leq \Delta i \leq 7$ have some support and should seldom be disregarded. However, this is not sufficient for model assessment as discussed in Section 5.2.3.

The Bayesian philosophy to model selection differs slightly to the AIC approach. The Bayesian philosophy is to select the family which is estimated to be most likely to be true. In true Bayesian fashion, in the first step before considering observations at hand, one specifies the prior probabilities, that f stems from the approximating families $\mathcal{G}_1, \mathcal{G}_2$, namely, $\mathbb{P}(f \in \mathcal{G}_1)$ and $\mathbb{P}(f \in \mathcal{G}_2)$.

Secondly, one computes and compares the posterier probabilities that f belongs to the approximating families given the observations, namely, $\mathbb{P}\big(f \in \mathcal{G}_1 \mid r_1, \ldots, r_T\big)$ and $\mathbb{P}\big(f \in \mathcal{G}_2 \mid r_1, \ldots, r_T\big)$. Again, in (dangerously) short, under conditions seen in [44], the Bayesian information criterion arises [45, p. 98]:

$$\text{BIC} = \underbrace{-2 \log \mathcal{L}_T}_{\text{measure of fit}} + \underbrace{p \log T}_{\text{penalty}}, \tag{27}$$

where $\mathcal{L}_T$ and p are as for the AIC and T is the number of observations, which is obviously not present whatsoever for the AIC. Compared to the AIC, the penalty term of the BIC has more weight for $T > e^2$, which holds in most practical applications. Thus, the BIC does, in general, favour models with fewer parameters than the AIC.

Summarizing, in both cases, the best model in the family is the one that minimizes these information criteria. Clearly, AIC does not depend directly on the sample size, T. Moreover, AIC presents the danger that it might overfit, whereas BIC presents the danger that it might underfit, simply in virtue of how each criterion penalize free parameters (see the under-braced penalty-terms in Equation 26 and Equation 27).

5.3.2 Pseudo-Residuals

Each r_t has a conditional distribution that depends on both the latent state $S_t = j$ and the previous observation r_{t-1}; that is,

$$r_t \mid (S_t = j, r_{t-1}) \sim \frac{1}{2c_j} \chi'^2_{\nu_j}(\lambda_j), \quad t = 2, \ldots, T,\ j \in \mathcal{S}.$$

As such, assessing outliers or model fit is non-trivial, since the conditional distribution of each r_t changes over time and depends on the hidden state sequence. A commonly used approach in (AR)HMM settings is to transform the observations to a common scale using pseudo-residuals $\{z_t\}_{t=2}^T$, constructed via the probability integral transform [45, pp. 101–106]:

I. Transform r_t to $u_t = F_t(r_t \mid r_{t-1}) \sim \mathcal{U}[0, 1]$, where F_t is the conditional CDF of r_t given r_{t-1}.

II. Transform u_t to $z_t = \Phi^{-1}(u_t) \sim \mathcal{N}(0, 1)$, where Φ is the standard normal CDF.

III. If the model is correctly specified, then the pseudo-residuals

$$z_t = \Phi^{-1}\left(F_t(r_t \mid r_{t-1})\right),$$

should be approximately independent and standard normally distributed. These can be evaluated using histograms and Q–Q plots.

To compute the forecast pseudo-residuals $z_t = \Phi^{-1}(F_t(r_t))$, we evaluate the one-step-ahead forecast distribution of r_t given the observed history up to time $t-1$. In the ARHMM setting, this forecast distribution is a mixture of state-dependent conditional distributions:

$$F_t(r_t) = \sum_{j \in \mathcal{S}} \xi_{t,j} \cdot F_j(r_t \mid r_{t-1}),$$

where $F_j(r_t \mid r_{t-1})$ denotes the conditional CDF of r_t under state j, and $\xi_{t,j} = \mathbb{P}(S_t = j \mid r_1, \ldots, r_{t-1})$ are the one-step-ahead predicted state probabilities. These are obtained by normalizing the forward probabilities at time $t-1$ and propagating them forward using the transition matrix $\mathbf{\Gamma}$:

$$\xi_t = \boldsymbol{\phi}_{t-1}\mathbf{\Gamma},$$

where $\boldsymbol{\phi}_{t-1}$ is the normalized forward probability vector at time $t-1$. The state-dependent conditional distributions $F_j(r_t \mid r_{t-1})$ are governed by the transition law of the CIR process:

$$r_t \mid (S_t = j, r_{t-1}) \sim \frac{1}{2c_j}\chi'^2_{\nu_j}(\lambda_j),$$

with degrees of freedom $\nu_j = 2q_j + 2$, non-centrality parameter $\lambda_j = 2c_j r_{t-1} e^{-\kappa_j \Delta t}$, and scaling constant $c_j = \frac{2\kappa_j}{(1-e^{-\kappa_j \Delta t})\sigma_j^2}$. The pseudo-residuals are then given by

$$z_t = \Phi^{-1}\left(\sum_{j \in \mathcal{S}} \xi_{t,j} \cdot F_j(r_t \mid r_{t-1})\right), \quad t = 2, \ldots, T.$$

If the model is correctly specified, the pseudo-residuals $\{z_t\}$ should be approximately independent and standard normally distributed.

For the first residual z_2, the one-step-ahead state probabilities $\xi_{2,j}$ cannot be computed from previous forward probabilities, since there is no observation before r_1. Instead, we approximate them using the stationary distribution $\boldsymbol{\delta}$, which solves $\delta^\top \mathbf{\Gamma} = \delta^\top$ and represents the long-run state probabilities of the Markov chain. Thus, the first residual is computed as:

$$z_2 = \Phi^{-1}\left(\sum_{j \in \mathcal{S}} \delta_j \cdot F_j(r_2 \mid r_1)\right).$$

Since the initial observation r_1 has no prior observation to condition on, pseudo-residuals are typically computed only for $t = 2, \ldots, T$.

Pseudo-Residuals: CIR-ARHMM As stated an enormous amounts of times, the transition density in the CIR model is that of a scaled non-central χ^2 distribution. As such, the distribution

function, $F_t(r_t \mid r_{t-1})$, will be a scaled version of said distribution. Specifically, the (conditional) cumulative distribution function of the non-central χ^2 distribution is represented by

$$F_t(r_t \mid r_{t-1}) = 1 - Q_{\frac{2\kappa\theta}{\sigma^2}}\left(\sqrt{\frac{4\kappa e^{-\kappa(\Delta)} r_{t-1}}{\sigma^2(1-e^{-\kappa(\Delta)})}}, \sqrt{\frac{2\sigma^2(1-e^{-\kappa(\Delta)})r_t}{4\kappa}}\right),$$

where $Q_m(a,b)$ is the Marcum Q-function of order m, defined as

$$Q_m(a,b) = \frac{1}{a^{m-1}} \int_b^\infty x^m \exp\left(-\frac{x^2+a^2}{2}\right) \mathcal{I}_{m-1}(ax)dx,$$

and $\mathcal{I}_{m-1}$ is the modified Bessel function of the first kind of order $m-1$. We rely on the implementation in `R` for the CDF and its approximation of the Marcum Q-function ▽.

6 Empirical Data Application

6.1 Model Selection & Assessments

The models were of much different computational strain. Indeed:

- Fitting the simple CIR model with two different sets of initial values around the OLS estimates took $\sim$ 15 minutes. They both converged to the same estimated parameters. The numerically differentiated Hessian matrix was also recovered. It was feasible to fit the models under $\Delta = 1/252$.
- Fitting the CIR-ARHMM with two different sets of initial values around the OLS estimates took $\sim$ 26 days. They both converged to the same estimated parameters. The numerically differentiated Hessian matrix was not recovered as computational efforts were already extremely high. The models were not feasible to be fitted under $\Delta = 1/252$ without using abnormally large values as initial values. As such, we used $\Delta = 1\triangledown$.

CIR Model: Assessment via Residuals We shall shortly touch upon the assessment of the simple CIR model. Residuals were calculated based on Equation 11 and can be assessed in Figure 8.

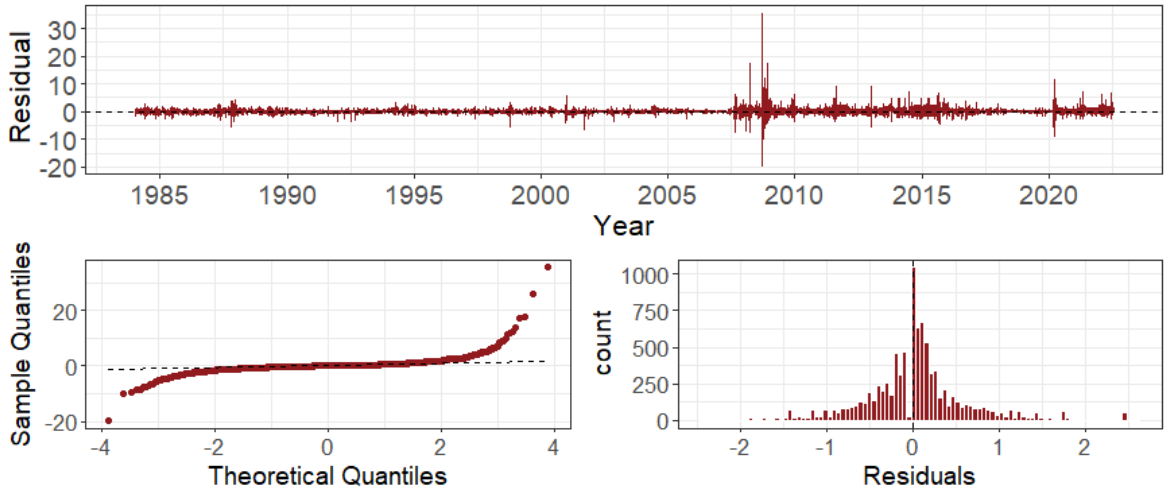


Figure 8: *Lower left: Theoretical quantiles vs. the sample quantiles. Lower right: Histogram of residuals from 1th to 99th quantile excluding extreme outliers. Upper: Time series plot of residuals over the sample period. Outliers are accessible for inspection in this plot.*

The CIR model does perform extremely well when we do not consider periods of extreme financial turmoil. When considering the 1st to 99th quantile, we see that all, except a few, fall within the range of -2 to 2. However, considering the time series plot of the residuals, we clearly see that the CIR model struggles to catch modern economic tendencies as the financial crisis (2007/2008)

hits and onwards including the COVID-19 pandemic (2019/2020). Furthermore, the Q-Q plot does show extremely heavy tails when we do not exclude the outliers as we did in the histogram for visibility. As such, we can conclude that the CIR model does produce reliable results for the pre-financial crisis era (pre-2007) but does fail to capture modern economic tendencies (2007 and onwards) with residuals ranging from almost -20 to 30.

CIR-ARHMM: Selection via AIC & BIC Now consider the information criteria described in section Section 5.3.1. The values for our 28 CIR-ARHMM's can be found in Table 4 where the models were fitted under $\Delta = 1$.

Discrete/Finite State Space CIR-ARHMM			
Model	p	**AIC**	**BIC**
2-state ARHMM (θ)	6	-119434.2	-119391.2
3-state ARHMM (θ)	11	-119571.6	-119492.7
4-state ARHMM (θ)	18	-119638.4	-119509.3
5-state ARHMM (θ)	27	-119837.4	-119643.8
2-state ARHMM (κ)	6	-119173.2	-119130.2
3-state ARHMM (κ)	11	-120581.0	-120502.1
4-state ARHMM (κ)	18	-120884.2	-120755.1
5-state ARHMM (κ)	27	-120750.5	-120556.8
2-state ARHMM (σ)	6	-126722.9	-126679.9
3-state ARHMM (σ)	11	-127878.4	-127799.5
4-state ARHMM (σ)	18	-128185.2	-128056.1
5-state ARHMM (σ)	27	**-128430.3★**	**-128236.7★**
2-state ARHMM (θ, σ)	7	-126793.1	-126742.9
3-state ARHMM (θ, σ)	13	-127944.4	-127851.2
4-state ARHMM (θ, σ)	21	-128305.9	-128155.3
5-state ARHMM (θ, σ)	31	**-128423.7■**	**-128201.4■**
2-state ARHMM (κ, σ)	7	-126798.2	-126748.0
3-state ARHMM (κ, σ)	13	-127677.0	-127583.8
4-state ARHMM (κ, σ)	21	-128126.8	-127976.2
5-state ARHMM (κ, σ)	31	-128098.9	-127876.6
2-state ARHMM (θ, κ)	7	-119364.6	-119314.4
3-state ARHMM (θ, κ)	13	-120281.7	-120188.4
4-state ARHMM (θ, κ)	21	-121141.1	-120990.5
5-state ARHMM (θ, κ)	31	-121210.9	-120988.6
2-state ARHMM (κ, θ, σ)	8	-126774.3	-126716.9
3-state ARHMM (κ, θ, σ)	16	-127637.1	-127529.5
4-state ARHMM (κ, θ, σ)	24	-127421.8	-127249.7
5-state ARHMM (κ, θ, σ)	35	**-128343.4▲**	**-128092.4▲**

Table 4: *AIC and BIC for all fitted ARHMM's. p is the number of parameters estimated. The parenthesis denotes which parameters are state-dependent of κ, θ and/or σ. **Red** ★ indicates the best model, **blue** ■ the second best, and **green** ▲ the third best according to AIC & BIC.*

We immediately notice that the best performing model in terms of both information criteria, is 5-state ARHMM where σ is state-dependent. However, the 5-state ARHMM where θ and σ are both state-dependent and 5-state ARHMM where θ, κ and σ are all state-dependent will be examined as their AIC & BIC values are *next in line*. Furthermore, as stated, the main objective of this project is to check for robustness and feasibility. As such, we consider the top three performing models. As will be evident shortly, some models suffer from numerical instability even with a good performance in terms of AIC & BIC.

CIR-ARHMM: Assessment via Pseudo-Residuals As described in Section 5.3.2, if the model is adequate, the pseudo-residuals should be approximately normally distributed. Consequently, we resort to Q-Q plots and histogram plots of the pseudo-residuals. We supplement these plots with time series and ACF plots for the models. These can be seen in Figure A.3.2, Figure A.3.3, Figure A.3.4 and Figure A.3.5. The pseudo-residuals are the culmination of most "$\triangledown$"'s given throughout the paper as most numerical instability complications are present in the calculation of pseudo-residuals. We will discuss exactly what these implies/implied in Section 7.

Immediately, note that the residual autocorrelation is considerably reduced, although most lags are still significant. Firstly, we consider the most appropriate model in respect to AIC & BIC, namely, 5-state ARHMM where σ is state-dependent. Considering the missing value counter in Figure A.3.2, we see that the model returns about $\sim 33\%$ of the residuals as missing. The missing pseudo-residuals are caused by numerical instability.
As model assessment is rendered useless for the model, we conclude that the 5-state ARHMM where σ is state-dependent, is not a feasible model to use for analysis.

Next in line is the 5-state ARHMM where θ and σ are state-dependent. This model did also suffer by numerical underflow/overflow. In fact, every model did but a large minority of models were rendered useless. However, for this model, the 5-state ARHMM where θ and σ are state-dependent, the Marcum Q-function did convergence when calculating the pseudo-residuals. Consequently, the model is fit for assessment as pseudo-residuals are available. Assessing the lower-tail quantiles of the data, the data is more extreme than what the theoretical distribution predicts. Indeed, if we observe the time series plot Figure A.3.3, we clearly see that the model pseudo-residuals are extremely large throughout the financial crisis (2007/2008) and post-financial crisis era (2008 and onwards) where near-0-rates were heavily in use. The model mismatch results in a large minority of pseudo-residuals being ~ -15 to ~ -5.
As this is the case, we conclude that the 5-state ARHMM where θ and σ are state-dependent, is a somewhat feasible but not an adequate model to use for analysis.

Next in line is the 5-state ARHMM where θ, σ, and κ are state-dependent. This model did suffer by numerical underflow/overflow as well, but the Marcum Q-function did converge when calculating the pseudo-residuals. As such, only one missing pseudo-residual was present. The pseudo-residuals for this model are distributed in a more Gaussian bell shape. The Q-Q plot

Figure A.3.2 does however indicates that the pseudo-residuals deviate from the theoretical normal distribution. Particularly, both tails suggest departure from normality but not by a large margin as with the other models. Indeed, observe the histogram displaying the distribution of residuals in Figure A.3.4. A large majority (95.00%) of the residuals fall within the range $[-2, 2]$. If we extended the interval to $[-3, 3]$ almost all residuals (99.91%) fall within the range. Examining the time series plot, we see that most of these outliers are around the start of the financial crisis but with very little outliers of size ~ 4.4.
As this is the case, we conclude that the 5-state ARHMM where θ, σ, and κ are state-dependent is feasible for analysis. However, we will discuss why we still think the CIR-ARHMM is not ripe for usage yet in practice.

6.2 Model Presentations

This section will be kept short, as presentation and forecasting are more relevant for the actual thesis.

CIR Model The simple CIR model results under $\Delta = 1/252$ will be considered shortly. The maximum-likelihood estimated parameters can be assessed in Equation 28

$$\kappa^* = 0.1345\ (0.0643), \quad \theta^* = 0.01998\ (0.0094), \quad \sigma^* = 0.07103\ (0.0005), \tag{28}$$

where we remind that the OLS estimators from Equation 15 were

$$\hat{\kappa} = 0.2068172, \quad \hat{\theta} = 0.0247371, \quad \hat{\sigma} = 0.07450132,$$

which were also found under $\Delta = 1/252$.
The most noticeable difference is that our maximum-likelihood estimated value, $\hat{\kappa}$, is approximately 35% lower than that of the OLS and the long run mean is 19.2% lower and the volatility almost identical but still larger. This is explained by the Euler discretization scheme that is used for OLS estimation. It is a smoothing that linearizes the process. In contrast, the maximum-likelihood estimators use the exact transition density rather than an approximation. This corresponds exactly to results found in [30].

CIR HMM The maximum-likelihood estimated parameter values can be see in Equation 29 for the CIR-ARHMM

$$\boldsymbol{\delta}^* = \begin{bmatrix} 0.1798 \\ 0.1719 \\ 0.2000 \\ 0.2273 \\ 0.2210 \end{bmatrix} \quad \boldsymbol{\Gamma}^* = \begin{bmatrix} 0.9687 & 0.0068 & 0.0068 & 0.0068 & 0.0108 \\ 0.0068 & 0.9726 & 0.0068 & 0.0068 & 0.0068 \\ 0.0068 & 0.0068 & 0.9726 & 0.0068 & 0.0068 \\ 0.0069 & 0.0026 & 0.0069 & 0.9767 & 0.0069 \\ 0.0068 & 0.0068 & 0.0068 & 0.0068 & 0.9726 \end{bmatrix},$$

$$\boldsymbol{\kappa}^* = \begin{bmatrix} 0.0057 \\ 0.0028 \\ 0.0001 \\ 0.0001 \\ 0.0010 \end{bmatrix}, \quad \boldsymbol{\theta}^* = \begin{bmatrix} 0.0027 \\ 0.1283 \\ 0.0680 \\ 0.0258 \\ 0.0131 \end{bmatrix}, \quad \boldsymbol{\sigma}^* = \begin{bmatrix} 0.0064 \\ 0.0286 \\ 0.0021 \\ 0.0016 \\ 0.0032 \end{bmatrix}. \tag{29}$$

Firstly, we consider $\boldsymbol{\delta}^*$ which is the stationary distribution. As stated, it reflects the long-term proportion of time the model spends in each state. The proportion of time spend in each state is seen to be close at around $\sim 17\%$ to $\sim 23\%$. The least time spend is $\sim 17\%$ (state 2) and most time $\sim 23\%$ (state 4).

The diagonal probabilities in $\boldsymbol{\Gamma}^*$, representing γ_{ij} for $i = j$ (i.e. the probability of remaining in the state at the subsequent time step), are all of some large probability around $\sim 97\%$. This is to be expected as the regime-switching should, in the domain of economical cycles, not happen frequently. The transition probabilities γ_{ij} with $i \neq j$ of the chain jumping from a state $i \to j$ are around $\sim 0.26\%$ to $\sim 1.1\%$. The smallest probability is $\gamma_{42} = 0.0026$ and largest $\gamma_{15} = 0.0108$.

As transitioning probabilities between states has been established, we move onto interpreting the effect of these state transitions combined with our estimated CIR model parameters. We will use Low, Medium and High as relative labels to describe differences between parameters:

State 1 High κ_1, Low θ_1, Medium σ_1:

This state represents a low mean level with moderate persistence and volatility. It may correspond to post-shock stabilization or a consolidation phase, where rates or volatility have dropped but remain reactive. Could be associated with cautious optimism or early recovery conditions. The high speed of mean reversion suggests exactly that this is possibly a recovery state, as the economy is actively trying to stabilize to some moderate interaste rate level at a high pace.

State 2 High κ_2, High θ_2, High σ_2:

This regime is characterized by strong mean reversion (κ_2) toward a high long-term level (θ_2), alongside elevated volatility (σ_2). Such a configuration is indicative of turbulent macro-financial conditions, typically associated with periods of monetary policy tightening, infla-

tionary shocks, or systemic stress. The high reversion speed implies that interest rates spike sharply but are pulled quickly back toward a higher average level — consistent with central banks reacting forcefully to anchor expectations. The elevated volatility suggests markets are uncertain about the economic outlook, pricing in both aggressive policy moves and downside risks. Overall, this state reflects a high-pressure environment where rates adjust rapidly and unpredictably in response to shifting fundamentals or policy signals.

State 3 Low κ_3, Medium θ_3, Low σ_3:

A calm and stable regime defined by weak mean reversion, a moderate long-term rate level, and subdued volatility. This configuration is typical of mature expansions or periods of policy neutrality, where macroeconomic conditions are steady and expectations well anchored. Markets exhibit low sensitivity to shocks, and interest rates drift gently around a stable midpoint without strong directional pressures.

State 4 Low κ_4, Low θ_4, Low σ_4:

A stagnant regime marked by persistently low interest rates, minimal volatility, and weak reversion dynamics. This pattern aligns with liquidity trap scenarios, zero lower bound (ZLB) environments, or periods of prolonged economic slack. Monetary policy is constrained, and market participants show little expectation of rate normalization. The system exhibits inertia — neither strong recovery signals nor significant downside momentum.

State 5 Medium κ_5, Low θ_5, Low-to-Medium σ_5:

An early-stage recovery or transitional regime. Rates remain low on average, but moderate mean reversion suggests some responsiveness to changing conditions. Volatility is slightly elevated, pointing to cautious re-engagement by markets or policymakers. This regime may reflect tentative normalization efforts, risk reappraisal, or the initial phase of exit from accommodative policy stances — a state of latent dynamism under a still-muted surface.

With these state-interpretations in mind, we now revisit the estimated transition matrix in Equation 29, focusing on the most and least likely regime shifts, as estimated by $\boldsymbol{\delta}^*$. The relatively high transition probability from State 1 (a post-shock or corrective regime) to State 5 (an early-stage recovery or policy support regime) suggests a plausible macro-financial narrative/interpretation: after a sharp dislocation or stress event, the system tends to move toward gradual stabilization, supported by accommodative policies and cautious optimism. This path reflects typical monetary policy cycles, where authorities respond to shocks with easing measures that gradually restore investor confidence.

In contrast, the transition probability from State 4 (a stagnant, liquidity-trap regime) to State 2 (a high-volatility, tightening or stress regime) is negligible, which is consistent with the economic intuition. Such a direct jump would imply a highly implausible scenario where rates leap from near-zero with minimal volatility to a regime of aggressive mean reversion and elevated

uncertainty, without passing through any intermediate phase of reawakening or normalization. The asymmetry between these transitions supports the CIR-ARHMM's structural realism: economic regimes do not shift chaotically but evolve through discernible stages, each with distinct dynamics and memory.

Concludingly, the transition structure embedded in $\mathbf{\Gamma}^*$ reinforces the interpretability of the estimated regimes, aligning well with stylized facts from macroeconomic cycles and market behavior.

7 Discussion

The modified Bessel function of the first kind was extremely difficult to assess. Warnings about imprecision and accuracy loss were given multiple times during optimization of the CIR-ARHMM. However, [42] states that the current algorithms for the modified Bessel function of the first kind will give warnings about accuracy loss for large arguments. In some cases, these warnings are exaggerated, and the precision is quote on quote "perfect". For large q [order], say in the order of millions, the current algorithms are rarely useful. Optimization did yield large orders, but never remotely close to the order of millions. However, some near 0 and negative orders did arise, which is the case when $2\kappa\theta < \sigma^2$, which yields another complication. The base `R` function `besselI` will in this case use another formula given in Equation 5, namely the one from [1, formulae 9.1.2]. This function is not optimal as the modified Bessel function of the first kind is not symmetric in the order whereas the formula from [1, formulae 9.1.2] is. The degree of lack of optimality from this complication was two part. 1) How many times did negative orders arise? 2) How does the indcues error change our optimization? These questions were almost impossible to assess with the current `R` packages available.
The reality is, that implementing a switching scheme for Bessel function approximations, testing its accuracy in a justified manner against that of `R` (which is fundamentally already a switching scheme) is a thesis of its own.

We ran into many difficulties related to using $\Delta = 1/252$. First off, using the OLS estimates, and values around the OLS estimates, as initial values for the CIR-ARHMM, yielded initial values that could not initialize the optimizer as it would instantly suffer from instability. In short, the values were too small. For the optimizer to start, we would have to use unreasonably large initial values. Using such values does not align with our theory and domain expertise. One could then say that "Why do we not resort to scaling the CIR-ARHMM estimated parameters post-optimizing?"–the reason is simple; the estimators do not simply scale as given in section Section 5.1.3 when using the CIR-ARHMM. Indeed, when we apply the ARHMM extension, the likelihood is computed using the forward probabilities. The forward probabilities are defined partly of the matrix $\boldsymbol{P}$ which has diagonal elements given by the (conditional) state-dependent densities, $f_i(r_t \mid r_{t-1})$, for $i \in \mathcal{S}$ and $t \in \{1, 2, \ldots, T\}$. As such, when using the forward algorithm, the scaling changes constantly when iterating recursively through all observations. This implies that we can not use the scalings from Section 5.1.3 to scale parameters to a $\Delta = 1/252$. Furthermore, the transition probabilities will also be affected by Δ. The estimated parameter values for the CIR-ARHMM are thus extremely hard to assess and interpret by the fact that they were estimated under $\Delta = 1$.
We have tried to figure out theoretically what the scaling would be after fitting but without luck.

Continuing the issues regarding numerical instability, we turn our attention to the model assessment by pseudo-residuals. The pseudo-residuals would often not be computed as the conditional

cumulative distribution function related to the (scaled) non-central χ^2 distribution would fail to convergence when evaluating the Marcum Q-function. Indeed, we observed degrees of freedom and non-centrality parameters of orders $\sim 10^{13}$ (some smaller, some larger) which resulted in numerical instability and thus no convergence of the Marcum Q-function. Interestingly, the numerical instability arose exactly at the start of the financial crisis (2007/2008) which can be examined by the time series plot Figure A.3.3. This is of no surprise, as this is where the most extreme financial environment was observed. We tried to circumvent the issue of divergence by scaling the forward probabilities using the exact same trick described for the likelihood scaling. It did alleviate the issue, as convergence before scaling was impossible, but the final results speak for themselves; issues still arose. On a positive note, the models seem to fit increasingly better when increasing the number of state-dependent parameters and number of states. This is of course expected.

However, the failure of convergence is a major issue in terms of robustness. If the CIR-ARHMM should ever be applied in financial forecasting (or even in-sample testing), we would need some tools to assess the model fit. If the CIR-ARHMM fails to have any evaluation criterium in terms of fit applied, here pseudo-residuals, it is simply not robust enough. Our best performing model, the 5-state CIR-ARHMM with σ state-dependent did not qualify for assessment as pseudo-residuals were missing at large. The next model in line, the 5-state CIR-ARHMM with θ and σ state-dependent was simply just bad with a extremely heavy lower tail. A model that did fit quite nicely and was assessable was the 5 state CIR-ARHMM with κ, θ and σ state-dependent. However, it failed the rule of thumb described in Section 5.3.1 miserably.

Consider again the probability density function for $f(r_t \mid r_{t-1})$

$$f(r_t \mid r_{t-1}) = c \cdot \exp(-u_{t-1} - v_t) \left(\frac{v_t}{u_{t-1}}\right)^{q/2} \mathcal{I}_q\left(2\sqrt{u_{t-1} v_t}\right),$$

with

$$\begin{aligned} c &= \frac{2\kappa}{\sigma^2\left(1 - \exp\left(-\kappa\Delta\right)\right)}, \\ u_{t-1} &= c r_{t-1} \exp\left(-\kappa\Delta\right), \\ v_t &= c r_t, \\ q &= \frac{2\kappa\theta}{\sigma^2} - 1. \end{aligned}$$

At $r_{t-1} = 0$, it follows that $u_{t-1} = 0$ which in turns implies

$$f(r_t \mid r_{t-1}) = c \cdot \exp(-0 - v_t) \underbrace{\left(\frac{v_t}{0}\right)^{q/2}}_{\text{undefined}} \mathcal{I}_q\left(2\sqrt{0 \cdot v_t}\right).$$

Furthermore, negative rates would yield the square root in the dynamics undefined. Zero- and

negative-rates were not a real phenomenon to consider in prior economic environments. Indeed, [27, p. 120] even states as a point against the Vašíček model that "This [negative rates] is one of the reasons why this process is no longer used to model interest rates.". zero-rates were not a large issue for our data as only 20 rates were 0 and none negative. As such, we simply rmeoved these. However, negative- and near-zero rates have become much more prominent in later economic trends. As such, the CIR-ARHMM does become extremely unstable and, at times, non-feasible. This is a large limitation of the CIR model. Furthermore, this complication is exaggerated for the CIR-ARHMM as it suffers from numerically instability more frequently.
If the thesis were to consider data from Denmark, which is likely, where negative rates have been extremely popular, the CIR model and CIR-ARHMM would simply not capture the data well.
As stated in footnote *a*, missing data in the CIR-ARHMM can be treated by replaying $\boldsymbol{P}$ by the identity matrix to achieve the ignorable likelihood [45, p. 40]. One could assume negative or zero rates were missing data and replace such entries with the identity matrix which would cause the model to fit outside its support. However, this is not an elegant solution, as it is constructionally wrong.

Our results for the CIR-ARHMM did seem to overcome some of the inherent modelling errors as pseudo-residuals were adequate for the 5-state CIR-ARHMM where κ, θ and σ were assumed to be state-dependent.
We conjectured that the models would suffer less from inherent modelling error if the regime switching did not happen often. Evidently, the diagonal elements of $\boldsymbol{\Gamma}^*$, γ_{ij} with $i = j$, were all about $\sim 97\%$ which indicates a less than frequent regime-switching. However, we again refer to the fact that the models were fitted under $\Delta = 1$ which does make the interpretation of the transition probabilities difficult and awkward.

The majority of these before-mentioned problems could be alleviated if we were to consider the derived stationary distribution, r_t^*, instead of the conditional distribution $r_t \mid r_{t-1}$. We remind that the former is Γ-distributed. The Γ distribution is well behaved and seldom suffers from instability. However, if one were to use the statistical model $\mathcal{P}_{\text{CIR}}$ parametrized through $\boldsymbol{\rho} = (\kappa,\, \theta,\, \sigma)^\top$ (Equation 6) for estimation, we simply can not do maximum-likelihood estimation. The statistical model $\mathcal{P}_{\text{CIR}}$ was shown to be unidentifiable and thus would result in an inconsistent estimator by construction (see [40, Thm. 2.5]). Consequently, the unidentifiability would result in our estimators only being reliable up to some scalar, meaning, we would have an infinite number of optimal solutions. However, the statistical model $\mathcal{P}_\Gamma$ parametrized through $\boldsymbol{\rho} = (\alpha,\, \beta)^\top$ (Equation 7) is identifiable. Usually, applications of (AR)HMM's are for distributional parameter estimations as these applications are not extensions overlaid onto a established model, such as the CIR model. If we were to continue with $\mathcal{P}_\Gamma$, we would simply estimate Γ-distribution parameters. Although this would be of much gain in terms of performing forecasting, model fitting and simulation, it would

not provide much economic intuition or interpretation. It is easily seen that

$$\mathbb{E}[r_t^*] = \frac{\alpha}{\beta} = \frac{\frac{2\kappa\theta}{\sigma^2}}{\frac{2\kappa}{\sigma^2}} = \theta,$$
$$\mathbb{V}[r_t^*] = \frac{\alpha}{\beta^2} = \frac{\frac{2\kappa\theta}{\sigma^2}}{\left(\frac{2\kappa}{\sigma^2}\right)^2} = \frac{\frac{2\kappa\theta}{\sigma^2}}{\frac{4\kappa^2}{\sigma^4}} = \frac{2\kappa\theta\sigma^2}{4\kappa^2} = \frac{\theta\sigma^2}{2\kappa},$$

i.e. we would always be able to identify θ but we would never be able to distinguish between the sizes of κ and σ in a absolute sense, only relatively. As we want economic interpretation to be the main focus of the actual thesis, using $\mathcal{P}_\Gamma$ does not seem elegant. Although, it would alleviate every single computational aspect, and as such, we should definetly consider it.

8 Conclusion

The Market Yield on U.S. Treasury Securities at 3-Month Maturity, Quoted on an Investment Basis, was used as the proxy for the short rate in this study. The dataset was well-behaved, exhibiting no negative rates, and only 20 zero-rate observations out of approximately 9,600. As these 20 rates were missing, they were simply omitted from the analysis. While this is not ideal, a proposed solution was outlined in footnote *a*, highlighting a limitation of the CIR model.

First, the classical one-factor Cox-Ingersoll-Ross (CIR) short-rate model was fitted. The CIR model performed adequately in the pre-financial crisis period (before 2007). However, during and after the financial crisis (2007/2008), the CIR model failed to capture the observed dynamics, as indicated by residuals of far greater magnitude than acceptable. The model's performance improved during 2017–2020, when near-zero rates became obsolete and rates reverted toward average levels. Computationally, the CIR model fit the full dataset efficiently, taking approximately 15 minutes without numerical instability. The convergence was the same for two different sets of parameters around the OLS.

Next, the extended CIR model was considered, using an autoregressive hidden Markov model (ARHMM) framework. This extension was natural, as the CIR process admits an analytical transition density and empirical autocorrelation between observations was high. Although the CIR-ARHMM introduces inherent modeling errors—such as assuming discrete-time regime transitions—the CIR-ARHMM achieved promising results in some configurations. The best-performing model by AIC and BIC (5-state CIR-ARHMM with state-dependent σ) could not be fully evaluated for fit due to instability: roughly 33% of pseudo-residuals were missing, stemming from non-convergence in the Marcum Q-function approximation for the non-central χ^2 conditional CDF.

The next-best model by information criteria, the 5-state CIR-ARHMM with state-dependent θ and σ, avoided missing values thanks to convergence in the conditional CDF. Nevertheless, residual analysis revealed inadequacy in terms of model fit. The third-ranked model, with state-dependent θ, κ, and σ, provided an excellent fit: no missing residuals, only a handful of outliers, and interpretable regimes. Even with fewer states, the model maintained good behaviour, though residuals remained left-skewed. State interpretations were clear, and regime characteristics and transition probabilities aligned with plausible economic narratives. The frequent transition from State 1 (post-shock stabilization) to State 5 (early recovery/normalization) mirrors macro-financial progression from short-term stability to cautious risk-taking. Conversely, the extremely low probability of jumping from State 4 (stagnation/liquidity trap) directly to State 2 (high volatility/stress) emphasizes the model's economic realism: such abrupt transitions rarely occur without intermediate phases.

However, all CIR-ARHMM's were fitted under $\Delta = 1$, complicating interpretation. Unlike the simple CIR model, parameters do not scale directly between $\Delta = 1$ and $\Delta = 1/252$, and transition probabilities are expressed in a different unit, further complicating interpretation.

In summary, the autoregressive hidden Markov model approach enhances accuracy in modelling interest rates but comes with significant drawbacks: long computation times and a lack of robustness, both largely attributable to difficulties in evaluating the CIR transition density. As robustness is of primary concern, the CIR model extended with an ARHMM should not be used further until the stability issues surrounding the modified Bessel function of the first kind are addressed in `R`. A proposed solution was using the stationary distribution parametrized through the Γ-distribution parameters.

Bibliography

REFERENCES

[1] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Vol. 55. Courier Corporation, 1965.

[2] Leif BG Andersen. "Efficient simulation of the Heston stochastic volatility model." In: *Available at SSRN 946405* (2007).

[3] Andrew Ang and Geert Bekaert. "Regime switches in interest rates." In: *Journal of Business & Economic Statistics* 20.2 (2002), pp. 163–182.

[4] Christian Bauer and Alain Bernard. "Regime switching in interest rates: A model for the term structure of interest rates." In: *Mathematical Finance* 13.2 (2003), pp. 241–264.

[5] Abdel Berkaoui, Mireille Bossy, and Awa Diop. "Euler scheme for SDEs with non-Lipschitz diffusion coefficient: strong convergence." In: *ESAIM: Probability and Statistics* 12 (2008), pp. 1–11.

[6] Tomas Björk. *Arbitrage theory in continuous time*. 4th ed. Oxford university press, 2020.

[7] Douglas T Breeden. "An intertemporal asset pricing model with stochastic consumption and investment opportunities." In: *Journal of financial Economics* 7.3 (1979), pp. 265–296.

[8] Kenneth P Burnham and David R Anderson. "Multimodel inference: understanding AIC and BIC in model selection." In: *Sociological methods & research* 33.2 (2004), pp. 261–304.

[9] Kenneth P Burnham, David R Anderson, and Kathryn P Huyvaert. "AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons." In: *Behavioral ecology and sociobiology* 65 (2011), pp. 23–35.

[10] Young Shin Chang, Hyoung-Goo Kim, and Seong-Jong Kim. "Regime-switching models: A survey and new directions." In: *Mathematics* 6.3 (2018), p. 34.

[11] David A Chapman, John B Long Jr, and Neil D Pearson. "Using proxies for the short rate: when are three months like an instant?" In: *The Review of Financial Studies* 12.4 (1999), pp. 763–806.

[12] John C. Cox, Jonathan E. Ingersoll, and Stephen A. Ross. "A theory of the term structure of interest rates." In: *Econometrica* 53.2 (1985), pp. 385–407.

[13] Griselda Deelstra, Freddy Delbaen, et al. "Convergence of discretized stochastic (interest rate) processes with stochastic drift term." In: *Applied stochastic models and data analysis* 14.1 (1998), pp. 77–84.

[14] Francis X. Diebold and Joon-Haeng Lee. "Regime switching with time-varying transition probabilities." In: *Nonstationary Time Series Analysis and Cointegration* (1994), pp. 283–302.

[15] Awa Diop. "An efficient discretisation scheme for one-dimensional SDEs with a diffusion coefficient function of the form $|x|^{\alpha},\ \alpha \in [1/2, 1]$." In: (2004).

[16] Awa Diop. "Sur la discrétisation et le comportement à petit bruit d'EDS unidimensionnelles dont les coefficients sont à dérivées singulières." PhD thesis. Nice, 2003.

[17] Sean R Eddy. *Biological sequence analysis Probabilistic models of proteins and nucleic acids*. 1998.

[18] Christina Erlwein and Ser-Huang Poon. "Continuous-time regime-switching models for interest rates." In: *Journal of Financial Econometrics* 8.1 (2010), pp. 105–130.

[19] Federal Reserve Bank of St. Louis (FRED). *Market Yield on U.S. Treasury Securities at 3-Month Constant Maturity, Quoted on an Investment Basis (DGS3MO)*. `https://fred.stlouisfed.org/graph/?g=oc55`. Accessed: 2024-09-22. 2024.

[20] William Feller. "Two singular diffusion problems." In: *Annals of mathematics* 54.1 (1951), pp. 173–182.

[21] William H. Greene. *Econometric Analysis*. English. 7th, international. Includes bibliographic references (pp. 1155-1200) and index. Previous edition: 2008. Harlow, England: Pearson, 2012, p. 1228. ISBN: 9780273753568.

[22] Massimo Guidolin. "Markov switching models in empirical finance." In: *Advances in Econometrics* 23 (2007), pp. 1–86.

[23] James D. Hamilton. "A new approach to the economic analysis of nonstationary time series and the business cycle." In: *Econometrica* 57.2 (1989), pp. 357–384.

[24] Zoé van Havre et al. "Overfitting hidden Markov models with an unknown number of states." In: *arXiv preprint arXiv:1602.02466* (2016).

[25] Steven L Heston. "A closed-form solution for options with stochastic volatility with applications to bond and currency options." In: *The review of financial studies* 6.2 (1993), pp. 327–343.

[26] Desmond J Higham and Xuerong Mao. "Convergence of Monte Carlo simulations involving the mean-reverting square root process." In: *Journal of Computational Finance* 8.3 (2005), pp. 35–61.

[27] Monique Jeanblanc, Marc Yor, and Marc Chesney. *Mathematical methods for financial markets*. Springer Science & Business Media, 2009.

[28] Norman L Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. *Continuous univariate distributions, volume 2*. Vol. 289. John wiley & sons, 1995.

[29] Richard Arnold Johnson, Dean W Wichern, et al. "Applied multivariate statistical analysis." In: (2002).

[30] Mathieu Kessler. "Estimation of an ergodic diffusion from discrete observations." In: *Scandinavian Journal of Statistics* 24.2 (1997), pp. 211–229.

[31] Kamil Kladívko. "Maximum likelihood estimation of the Cox-Ingersoll-Ross process: the Matlab implementation." In: *Technical Computing Prague* 7.8 (2007), pp. 1–8.

[32] M Ayhan Kose, Naotaka Sugawara, and Marco E Terrones. "Global recessions." In: (2020).

[33] Brian G Leroux and Martin L Puterman. "Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models." In: *Biometrics* (1992), pp. 545–558.

[34] H Linhart and W Zucchini. "Model Selection, Wiley." In: *New York* (1986).

[35] Roderick Little. "Selection and pattern-mixture models." In: *Longitudinal data analysis*. Chapman and Hall/CRC, 2008, pp. 423–446.

[36] Roger Lord, Remmert Koekkoek, and Dick Van Dijk. "A comparison of biased simulation schemes for stochastic volatility models." In: *Quantitative Finance* 10.2 (2010), pp. 177–194.

[37] M Maechler. *Bessel: Computations and Approximations for Bessel Functions*. 2009. URL: `https://cran.r-project.org/web/packages/Bessel/index.html`.

[38] Théo Michelot, Roland Langrock, and Toby Patterson. "moveHMM: An R package for the analysis of animal movement data." In: *Computer software* (2019).

[39] John A Nelder and Roger Mead. "A simplex method for function minimization." In: *The computer journal* 7.4 (1965), pp. 308–313.

[40] Whitney K. Newey and Daniel McFadden. "Large sample estimation and hypothesis testing." In: *Handbook of Econometrics*. Ed. by Robert Engle and Dan McFadden. Vol. 4. Elsevier Science, 1994. Chap. 36, pp. 2111–2245. ISBN: 978-0-444-88766-5.

[41] Jennifer Pohle et al. "Selecting the number of states in hidden Markov models: pragmatic solutions illustrated using animal movement." In: *Journal of Agricultural, Biological and Environmental Statistics* 22 (2017), pp. 270–293.

[42] R Core Team. *Bessel Functions: Bessel Functions of Integer and Fractional Order*. R documentation for Bessel functions. R Foundation for Statistical Computing. Vienna, Austria, 2025. URL: `https://stat.ethz.ch/R-manual/R-devel/library/base/html/Bessel.html`.

[43] Oldrich Vasicek. "An equilibrium characterization of the term structure." In: *Journal of financial economics* 5.2 (1977), pp. 177–188.

[44] Larry Wasserman. "Bayesian model selection and model averaging." In: *Journal of mathematical psychology* 44.1 (2000), pp. 92–107.

[45] Walter Zucchini and Iain L MacDonald. *Hidden Markov models for time series: an introduction using R*. Chapman and Hall/CRC, 2009.

Appendix

A.1 Code

Code used for this paper (and some for the keen reader) is available on this hyper link to a GitHub repository dedicated to this paper. The data frames are also found along the source code.

A.2 Derivations & Proofs

A.2.1 Proof of Proposition 5.1

Proof. Let $\rho(s)$ be a squared Bessel process of dimension $\delta \geq 0$, satisfying

$$d\rho_s = \delta\, ds + 2\sqrt{\rho_s}\, d\mathcal{B}_s,$$

where $\mathcal{B}_s$ is a standard Brownian motion. Define the time change

$$s(t) = \frac{\sigma^2}{4k}(e^{kt} - 1)$$

and set

$$r_t = e^{-kt}\rho(s(t)).$$

Applying Itô's formula to the composition yields

$$dr_t = -ke^{-kt}\rho(s(t))\, dt + e^{-kt}d\rho(s(t)).$$

Since

$$\frac{ds}{dt} = \frac{d}{dt}\left(\frac{\sigma^2}{4k}(e^{kt} - 1)\right) = \frac{\sigma^2}{4}e^{kt},$$

and using the chain rule for Itô processes, we obtain

$$d\rho(s(t)) = \delta \cdot \frac{\sigma^2}{4}e^{kt}dt + 2\sqrt{\rho(s(t))} \cdot \sqrt{\frac{\sigma^2}{4}e^{kt}}\, d\widetilde{W}_t,$$

where $\widetilde{W}_t$ is a Brownian motion with respect to time-t. Substituting into the expression for dr_t, we find

$$dr_t = -kr_t dt + \frac{\delta\sigma^2}{4}dt + \sigma\sqrt{r_t}d\widetilde{W}_t.$$

This is of the form

$$dr_t = k(\theta - r_t)dt + \sigma\sqrt{r_t}d\widetilde{W}_t$$

if and only if $\delta = \frac{4k\theta}{\sigma^2}$. Therefore, the process r_t defined by the given transformation satisfies the CIR stochastic differential equation and is thus obtained from a squared Bessel process via a space-time change. □

A.2.2 Proof of Proposition 5.2

Proof. **Case 1:** $\delta = 0$**.** The process satisfies the SDE $d\rho_t = 2\sqrt{\rho_t}\, dW_t$ with $\rho_0 \geq 0$. Since the diffusion term vanishes at $\rho_t = 0$, the process remains at 0 once it hits it. Formally, if $\tau := \inf\{t \geq 0 : \rho_t = 0\}$, then for all $t \geq \tau$,

$$\rho_t = \rho_\tau + \int_\tau^t 2\sqrt{\rho_s}\, dW_s = 0.$$

Thus, 0 is absorbing.

Case 2: $0 < \delta < 2$**.** Now ρ_t satisfies $d\rho_t = \delta\, dt + 2\sqrt{\rho_t}\, dW_t$. This is a continuous non-negative semimartingale. Applying the occupation time formula with $f(x) = \frac{1}{4x}\mathbf{1}_{x>0}$, we get:

$$\begin{aligned}\int_0^t \frac{1}{4\rho_s}\mathbf{1}_{\{\rho_s>0\}}\, d\langle\rho\rangle_s &= \int_0^t \mathbf{1}_{\{\rho_s>0\}}\, ds \\ &= \int_0^\infty \frac{1}{4a} L_t^a(\rho)\, da.\end{aligned}$$

For this integral to be finite, we must have $L_t^0(\rho) = 0$, since $\frac{1}{4a}$ is not integrable at $a = 0$ otherwise. From the general identity for squared Bessel processes,

$$L_t^0(\rho) = 2\delta \int_0^t \mathbf{1}_{\{\rho_s=0\}}\, ds,$$

so the vanishing of local time at 0 implies

$$\int_0^t \mathbf{1}_{\{\rho_s=0\}}\, ds = 0.$$

Therefore, the process spends zero Lebesgue measure time at 0, i.e., it is instantaneously reflected. □

A.2.3 Proof of Theorem 5.3

Proof. For $s \leq t$, one has by definition that

$$r_t = r_s + \kappa \int_s^t (\theta - r_u)du + \sigma \int_s^t \sqrt{r_u} dW_u^{\mathbb{Q}(\lambda)}$$

for $\lambda \in \mathbb{R}$. Itô's formula [6, Prop. 4.12] then gives

$$\begin{aligned} r_t^2 &= r_s^2 + 2\kappa \int_s^t (\theta - r_u) r_u du + 2\sigma \int_s^t (r_u)^{3/2} dW_u^{\mathbb{Q}(\lambda)} + \sigma^2 \int_s^t r_u du \\ &= r_s^2 + (2\kappa\theta + \sigma^2) \int_s^t r_u du - 2\kappa \int_s^t r_u^2 du + 2\sigma \int_s^t (r_u)^{3/2} dW_u^{\mathbb{Q}(\lambda)} \end{aligned}$$

By [27, Prop. 6.3.1.1], r admits moments of any order. Therefore, the expectation of r_t is given as

$$\mathbb{E}[r_t] =^{\kappa} \mathbb{Q}_x^{\kappa\theta,\sigma}(r_t) = r_0 + \kappa \left(\theta t - \int_0^t \mathbb{E}[r_u] du \right).$$

We now introduce a helper function $\psi(t) = \mathbb{E}[r_t]$. The integral equation

$$\psi(t) = r_0 + \kappa \left(\theta t - \int_0^t \psi(t) du \right)$$

can then be written in different form as $\psi'(t) = \kappa(\theta - \psi(t))$ where $\psi(0) = r_0$. Hence we achieve

$$\mathbb{E}[r_t] = \theta + (r_0 - \theta) e^{-\kappa t}.$$

Similarly, we have

$$\mathbb{E}[r_t^2] = r_0^2 + (2\kappa\theta + \sigma^2) \int_0^t \mathbb{E}[r_u] du - 2\kappa \int_0^t \mathbb{E}[r_u^2] du,$$

and setting $\psi(t) = \mathbb{E}[r_t^2]$ leads to $\psi'(t) = (2\kappa\theta + \sigma^2)\psi(t) - 2\kappa\psi(t)$ and hence

$$\mathbb{V}[r_t] = \frac{\sigma^2}{\kappa} \left(1 - e^{-\kappa t}\right) \left(r_0 e^{-\kappa t} + \frac{\theta}{2} (1 - e^{-\kappa t} \right).$$

Then, by the Markovian character of r and by Theorem 5.2, the conditional expectation can also be computed as

$$\begin{aligned} \mathbb{E}[r_t \mid \mathcal{F}_s] &= \theta + (r_s - \theta) e^{-\kappa(t-s)} = r_s e^{-\kappa(t-s)} + \theta \left(1 - e^{-\kappa(t-s)}\right), \\ \mathbb{V}[r_t \mid \mathcal{F}_s] &= r_s \frac{\sigma^2 \left(e^{-\kappa(t-s)} - e^{-2\kappa(t-s)}\right)}{\kappa} + \frac{\theta\sigma^2 \left(1 - e^{-\kappa(t-s)}\right)^2}{2\kappa}. \end{aligned}$$

□

A.2.4 Proof of Proposition 5.3

Proof. Firstly, note that we have shown that squared Bessel processes are Markov processes and their transition densities are known. Expectation under $\mathbb{Q}_x^\delta$ will be denoted by $\mathbb{Q}_x^\delta[\cdot]$ and ρ will denote a square Bessel process under the $\mathbb{Q}^\delta$-law. By [27, Prop. 6.2.1.1], the Laplace transformation of ρ_t satisfies the relation

$$\mathbb{Q}_x^\delta[\exp(-\lambda\rho)] = \mathbb{Q}_x^1[\exp(-\lambda\rho_t)]\left[\mathbb{Q}_0^1[\exp(-\lambda\rho_t)]\right]^{\delta-1},$$

and since, under $\mathbb{Q}_x^1$, the random variable ρ_t is the square of a Gaussian variable, we can apply a Laplace transformation for the Square of Gaussian Law [27, Exer. 1.1.12.3]

$$\mathbb{Q}_x^1\left[\exp(-\lambda\rho_t)\right] = \frac{1}{\sqrt{1+2\lambda t}}\exp\left(-\frac{\lambda x}{1+2\lambda t}\right)$$

Investing this Laplace transformation yields the transition density $q_t^{(v)}$ of some BESQ$^{(\nu)}$ for $\nu > -1$ as

$$q_t^{(\nu)}(x,y) = \frac{1}{2t}(y/x)^{\nu/2}\exp\left(-\frac{x+y}{2t}\right)\mathcal{I}_\nu\left(\frac{\sqrt{xy}}{t}\right).$$

So, from the relation $r_t = e^{-\kappa t}\rho_{\eta(t)}$, where ρ is a BESQ$^{(\nu)}$, we obtain

$${}^\kappa\mathbb{Q}^{\kappa\theta,\sigma}(r_{t+s}\in dy \mid r_s = x) = e^{\kappa t}q_{c(t)}^{(\nu)}(x, ye^{\kappa t})dy.$$

Now, denote by $(r_t(x); t \geq 0)$ the CIR process with initial value $r_0 = x$, the random variable $Y = r_t(x)e^{\kappa t}[\eta(t)]^{-1}$ has density

$$\begin{aligned}\mathbb{P}(Y_t \in dy)/dy &= \eta(t)e^{-\kappa t}f_t(x, y\eta(t)e^{-\kappa t})\mathbb{1}_{\{y>0\}}\\ &= \frac{e^{-\alpha/2}}{2\alpha^{\nu/2}}e^{-y/2}y^{\nu/2}\mathcal{I}_\nu(\sqrt{y\alpha})\mathbb{1}_{\{y\geq 0\}},\end{aligned}$$

where $\alpha = x/\eta(t)$ □

A.2.5 Derivation of the CIR Solution Distribution and More

By definition, the sum of the squares of n standard normal variables follows a χ^2 distribution with n degrees of freedom. If the individual normal variables have non-zero means (but still standard deviation 1), then the sum of their squares follows a non-central χ^2 distribution with non-centrality parameter equal to the sum of the squares of the means. Let X_i be normal random variables. We posit:

$$r_t = X_{1,t}^2 + X_{2,t}^2 + \cdots + X_{n,t}^2,$$

where each $X_{i,t}$ follows the dynamics (similar to Vašíček but with mean-reversion to zero):

$$dX_{i,t} = -\frac{1}{2}\kappa X_{i,t}dt + \frac{1}{2}\sigma dW_t^{\mathbb{Q}}$$

Multiply both sides by the integrating factor and integrate from s to t, $t > s$:

$$\begin{aligned}
dX_{i,t} + \frac{1}{2}\kappa X_{i,t}\,dt &= \frac{1}{2}\sigma\,dW_t^{\mathbb{Q}} \\
&\iff \\
\exp(\kappa t/2)dX_{i,t} + \frac{1}{2}\exp(\kappa t/2)\kappa X_{i,t}\,dt &= \frac{1}{2}\exp(\kappa t/2)\sigma\,dW_t^{\mathbb{Q}} \\
&\iff \\
d\left(\exp(\kappa t/2)X_{i,t}\right) &= \frac{1}{2}\exp(\kappa t/2)\sigma\,dW_t^{\mathbb{Q}} \\
&\iff \\
\int_s^t d\left(\exp(\kappa u/2)X_{i,u}\right) &= \frac{1}{2}\sigma\int_s^t \exp(\kappa u/2)\,dW_u^{\mathbb{Q}} \\
&\iff \\
\exp(\kappa t/2)X_{i,t} - \exp(\kappa s/2)X_{i,s} &= \frac{1}{2}\sigma\int_s^t \exp(\kappa u/2)\,dW_u^{\mathbb{Q}} \\
&\iff \\
X_{i,t} &= \exp(-\kappa(t-s)/2)X_{i,s} + \frac{1}{2}\sigma\int_s^t \exp(-\kappa(t-u)/2)\,dW_u^{\mathbb{Q}}
\end{aligned}$$

The conditional mean and variance are found by [6, Prop. 4.5] and [6, Lem. 4.18]:

$$\begin{aligned}
\mathbb{E}[X_{i,t} \mid X_{i,s}] &= \exp(-\kappa(t-s)/2)X_{i,s} \\
\mathbb{V}[X_{i,t} \mid X_{i,s}] &= \frac{\sigma^2}{4}\int_s^t \exp(-\kappa(t-u))du = \frac{\sigma^2}{4\kappa}(1-\exp(-\kappa(t-s)))
\end{aligned}$$

To obtain unit variance, define the standardized variable:

$$X'_{i,t} = \frac{X_{i,t}}{\sqrt{\frac{\sigma^2}{4\kappa}(1-\exp(-\kappa(t-s)))}}.$$

Then

$$\mathbb{E}[X'_{i,t} \mid X_{i,s}] = \frac{\exp(-\kappa(t-s)/2)X_{i,s}}{\sqrt{\frac{\sigma^2}{4\kappa}(1-\exp(-\kappa(t-s)))}}, \quad \mathbb{V}[X'_{i,t}] = 1.$$

such that

$$r_t = \sum_{i=1}^n X_{i,t}^2 = \frac{\sigma^2}{4\kappa}(1-\exp(-\kappa(t-s)))\sum_{i=1}^n X_{i,t}'^2.$$

So r_t is a scalar times a Non-central χ^2 variable with non-centrality parameter

$$\begin{aligned}
\lambda &= \sum_{i=1}^{n} \mathbb{E}[X'_{i,t}]^2 \\
&= \sum_{i=1}^{n} \left(\frac{\exp(-\kappa(t-s)/2) X_{i,s}}{\sqrt{\frac{\sigma^2}{4\kappa}(1-\exp(-\kappa(t-s)))}} \right)^2 \\
&= \frac{\exp(-\kappa(t-s))}{\frac{\sigma^2}{4\kappa}(1-\exp(-\kappa(t-s)))} \sum_{i=1}^{n} X_{i,s}^2 \\
&= \frac{4\kappa \exp(-\kappa(t-s))}{\sigma^2(1-\exp(-\kappa(t-s)))} r_s
\end{aligned}$$

Now we verify the degrees of freedom. Using Itô's formula [6, Prop. 4.12]:

$$\begin{aligned}
d\left(X_{i,t}^2\right) &= 2X_{i,t} dX_{i,t} + \frac{1}{2} \cdot 2 \cdot (dX_{i,t})^2 \\
&= 2X_{i,t} \left(-\frac{1}{2}\kappa X_{i,t} dt + \frac{1}{2}\sigma dW_t^{\mathbb{Q}} \right) + \frac{\sigma^2}{4} dt \\
&= \left(-\kappa X_{i,t}^2 + \frac{\sigma^2}{4} \right) dt + \sigma X_{i,t} dW_t^{\mathbb{Q}}.
\end{aligned}$$

Thus

$$\begin{aligned}
dr_t &= dX_{1,t}^2 + dX_{2,t}^2 + dX_{3,t}^2 + \cdots + dX_{n,t}^2 \\
&= \sum_{i=1}^{n} dX_{i,t}^2 \\
&= \sum_{i=1}^{n} \left(-\kappa X_{i,t}^2 + \frac{\sigma^2}{4} \right) dt + \sum_{i=1}^{n} \sigma X_{i,t} dW_t^{\mathbb{Q}} \\
&= \left(-\kappa \sum_{i=1}^{n} X_{i,t}^2 + n\frac{\sigma^2}{4} \right) dt + \sigma \sum_{i=1}^{n} X_{i,t} dW_t^{\mathbb{Q}} \\
&= \left(-\kappa r_t + n\frac{\sigma^2}{4} \right) dt + \sigma\sqrt{r_t} \sum_{i=1}^{n} X_{i,t}\sqrt{r_t} dW_t^{\mathbb{Q}} \\
&= \left(-\kappa r_t + n\frac{\sigma^2}{4} \right) dt + \sigma\sqrt{r_t} d\tilde{W}_t^{\mathbb{Q}},
\end{aligned}$$

where $d\tilde{W}_t^{\mathbb{Q}}$ is a Brownian motion[8]. So, by comparing dynamics to the CIR dynamics we see that the degrees of freedom n is

$$dr_t = \kappa(\theta - r_t)dt + \sigma\sqrt{r_t}dW_t^{\mathbb{Q}} \Rightarrow \frac{n\sigma^2}{4} = \kappa\theta \Rightarrow \quad n = \frac{4\kappa\theta}{\sigma^2}$$

However, [12] formula for the density of short rate is presented in a slightly different form, which we can easily transform our equations into by making the same substitutions that [**cox**] chose to make. Define

$$\begin{aligned} c &= \frac{2\kappa}{\sigma^2\left(1-\exp(-\kappa(\Delta))\right)} \\ u &= c\exp(-\kappa(\Delta))r_s \\ v &= cr_t \\ q &= \frac{2\kappa\theta}{\sigma^2} - 1 \\ x &\sim \chi_n'^2(\lambda) \\ f_X(x;n,\lambda) &= \frac{1}{2}\exp\left(-\frac{x+\lambda}{2}\right)\left(\frac{x}{\lambda}\right)^{\frac{n}{2}-1}\mathcal{I}_{\frac{n}{2}-1}\left(\sqrt{\lambda x}\right) \end{aligned}$$

Making the substitutions, we get:

$$\begin{aligned} n &= \frac{4\kappa\theta}{\sigma^2} = 2\left(\frac{2\kappa\theta}{\sigma^2} - 1\right) + 2 = 2q + 2 \\ \lambda &= \frac{4\kappa\exp(-\kappa(\Delta))}{\sigma^2\left(1-\exp(-\kappa(\Delta))\right)r_s}r_s \\ r_t &= \frac{\sigma^2}{4\kappa}\left(1-\exp(-\kappa(t-s))\right)x = \frac{1}{2c}x \quad \Rightarrow \quad r_t \sim \frac{1}{2c}\chi_n'^2(\lambda) \end{aligned}$$

Finally, we derive the formula for the density of r_t by transforming the density of x:

$$\begin{aligned} f_X(x;n,\lambda) &= \frac{1}{2}\exp\left(-\frac{x+\lambda}{2}\right)\left(\frac{x}{\lambda}\right)^{\frac{n-2}{4}}\mathcal{I}_{\frac{n-2}{4}}\left(\sqrt{\lambda x}\right) \\ f_r(r;n,\lambda) &= f_X\left(x^{-1}(r);n,\lambda\right)\left|\frac{dx}{dr}\right|, \end{aligned}$$

where

$$\begin{aligned} r = \frac{1}{2c}x \Rightarrow x = 2cr \Rightarrow \frac{dx}{dr} = 2c \\ f_r(r;n,\lambda) = 2cf_X(2cr;n,\lambda) \end{aligned}$$

[8]Note that $\sum_{i=1}^n X_{i,t}dW_t^{\mathbb{Q}} = \sqrt{r_t}\cdot\sum_{i=1}^n \frac{X_{i,t}}{\sqrt{r_t}}dW_t^{\mathbb{Q}} = \sqrt{r_t}d\tilde{W}_t^{\mathbb{Q}}$.

Substitute for f_X:

$$
\begin{aligned}
f_r(r; n, \lambda) &= 2c \cdot \frac{1}{2} \exp\left(-\frac{x+\lambda}{2}\right) \left(\frac{x}{\lambda}\right)^{\frac{n-2}{4}} \mathcal{I}_{\frac{n-2}{2}}\left(\sqrt{\lambda x}\right) \\
&\iff \\
f_r(r; n, \lambda) &= 2c \cdot \frac{1}{2} \exp(-(v+u)) \left(\frac{2v}{2u}\right)^{\frac{2q+2-2}{4}} \mathcal{I}_{\frac{2q+2-2}{2}}\left(\sqrt{2u2v}\right) \\
&\iff \\
f_r(r; n, \lambda) &= c \exp(-(v+u)) \left(\frac{v}{u}\right)^{\frac{q}{2}} \mathcal{I}_q\left(2\sqrt{uv}\right).
\end{aligned}
$$

A.2.6 Proof of Theorem 5.4

Proof. The first equality simply follows from the law of total probability:

$$
\begin{aligned}
\delta_i^{(t+1)} &= \mathbb{P}(S_{t+1} = i) \\
&= \sum_{j \in \mathcal{S}} \underbrace{\mathbb{P}(S_t = j)}_{\delta_i^{(t)}} \underbrace{\mathbb{P}(S_{t+1} = j \mid S_t = i)}_{\gamma_{ij}} \\
&\Rightarrow \\
\boldsymbol{\delta}^{(t+1)} &= \begin{bmatrix} \delta_1^{(t+1)} & \cdots & \delta_N^{(t+1)} \end{bmatrix} \\
&= \begin{bmatrix} \delta_1^{(t)} & \cdots & \delta_N^{(t)} \end{bmatrix} \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1N} \\ \vdots & \ddots & \vdots \\ \gamma_{N1} & \cdots & \gamma_{NN} \end{bmatrix} \\
&= \boldsymbol{\delta}^{(t)} \boldsymbol{\Gamma}.
\end{aligned}
$$

Lastly, equality two and three follows from:

$$
\boldsymbol{\delta}^{(t+1)} = \boldsymbol{\delta}^{(t)} \boldsymbol{\Gamma} = \boldsymbol{\delta}^{(t-1)} \boldsymbol{\Gamma}\boldsymbol{\Gamma} = \boldsymbol{\delta}^{(t-2)} \boldsymbol{\Gamma}\boldsymbol{\Gamma}\boldsymbol{\Gamma} = \ldots = \boldsymbol{\delta}^{(1)} \boldsymbol{\Gamma}^{t-1}.
$$

□

A.2.7 Calculating the Stationary Distribution (Equation 20)

Firstly, note that

$$
\boldsymbol{\delta}\boldsymbol{\Gamma} = \boldsymbol{\delta} \iff \boldsymbol{\delta} - \boldsymbol{\delta}\boldsymbol{\Gamma} = \mathbf{0}_N \iff \boldsymbol{\delta}(\boldsymbol{I}_N - \boldsymbol{\Gamma}) = \mathbf{0}_N,
$$

where $\mathbf{0}_N$ is an N-dimensional row vector of zeros. Now, note that

$$\begin{aligned}
\sum_i \delta_i = 1 &\iff \begin{bmatrix}\delta_1 & \cdots & \delta_N\end{bmatrix}\begin{bmatrix}1\\ \vdots \\ 1\end{bmatrix} = 1 \\
&\iff \begin{bmatrix}\delta_1 & \cdots & \delta_N\end{bmatrix}\begin{bmatrix}1 & \cdots & 1\\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1\end{bmatrix} = \begin{bmatrix}1 & \cdots & 1\end{bmatrix} \\
&\iff \boldsymbol{\delta}\mathbf{1}_{N\times N} = \mathbf{1}_N.
\end{aligned}$$

Adding the two equations, factoring out $\boldsymbol{\delta}$ and transposing, then yields the desired result

$$\begin{aligned}
\boldsymbol{\delta}(\boldsymbol{I}_N - \boldsymbol{\Gamma} + \mathbf{1}_{N\times N}) = \mathbf{1}_N &\iff (\boldsymbol{I}_N - \boldsymbol{\Gamma} + \mathbf{1}_{N\times N})^\top \boldsymbol{\delta}^\top = \mathbf{1}_N^\top \\
&\iff \left(\boldsymbol{I}_N - \boldsymbol{\Gamma} + \begin{bmatrix}1 & \cdots & 1\\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1\end{bmatrix}\right)^\top \begin{bmatrix}\delta_1\\ \vdots \\ \delta_N\end{bmatrix} = \begin{bmatrix}1\\ \vdots \\ 1\end{bmatrix}
\end{aligned}$$

A.2.8 Proof of Proposition 5.4

´

Proof. We abuse notation a bit by simply noting that he random variables r and S take on values but we shorten the notation. First, note that

$$\mathcal{L}_T(\boldsymbol{\zeta}) = \mathbb{P}(r_1, r_2, \ldots, r_T) = \sum_{S_1, S_2, \ldots, S_T = 1}^{M} \mathbb{P}(r_1, r_2, \ldots, r_T, S_1, S_2, \ldots, S_T)$$

and by [45, Eq. 2.5]

$$\mathcal{L}_T(\boldsymbol{\zeta}) = \mathbb{P}(r_1, r_2, \ldots, r_T, S_1, S_2, \ldots, S_T) = \mathbb{P}(S_1)\prod_{k=2}^{T}\mathbb{P}(S_t \mid S_{t-1})\prod_{k=1}^{T}\mathbb{P}(r_k \mid S_k).$$

It then follows that

$$\begin{aligned}
\mathcal{L}_T(\boldsymbol{\zeta}) &= \sum_{S_1, S_2, \ldots, S_T}^{M} (\delta_{S_1}\gamma_{S_1,S_2}\gamma_{S_2,S_3}\ldots\gamma_{S_{T-1},S_T})(p_{S_1}(r_1)p_{S_2}(r_2)\ldots p_{S_T}(r_T)) \\
&= \sum_{S_1, S_2, \ldots, S_T}^{M} \delta_{S_1} p_{S_1}(r_1)\gamma_{S_1,S_2}p_{S_2}(r_2)\gamma_{S_2,S_3}\ldots\gamma_{S_{T-1},S_T}p_{S_T}(r_T) \\
&= \boldsymbol{\delta}\boldsymbol{P}(r_1)\boldsymbol{\Gamma}\boldsymbol{P}(r_2)\ldots\boldsymbol{\Gamma}\boldsymbol{P}(r_T)\mathbf{1}^\top.
\end{aligned}$$

The last equality exploits the fact that a multiple sum of terms having a certain simple multiplicative form can in general be written as a matrix product (see [45, Ex. 7(b)]). If $\boldsymbol{\delta}$ is the stationary distribution of the Markov chain, we simply have

$$\boldsymbol{\delta}\boldsymbol{P}(r_1) = \boldsymbol{\delta}\boldsymbol{\Gamma}\boldsymbol{P}(r_1).$$

□

A.2.9 Derivation of the Forward Algorithm

Firstly, for $j = 1, \ldots, N$

$$\alpha_1 = f(r_1, S_i = j) = f(r_1 \mid S_1 = j) f(S_1 = j) f_j(r_1) \delta_j^{(1)},$$

which we can write in a row vector from state 1 to N as

$$\boldsymbol{\alpha}_1 = \begin{bmatrix} \alpha_1(1) & \cdots & \alpha_1(N) \end{bmatrix} = \begin{bmatrix} \delta_1^{(1)} & \cdots & \delta_N^{(1)} \end{bmatrix} \begin{bmatrix} f_1(r_1) & 0 & \cdots & 0 \\ 0 & f_2(r_1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & f_N(r_1) \end{bmatrix} = \boldsymbol{\delta}\boldsymbol{P}(r_1)$$

Next, for the recursion part of the algorithm, for $i = 1, \ldots, N$, using the conditional probability, independence and Markov property of the Markov chain yields

$$\begin{aligned}
a_t(j) &= f(r_1, \ldots, r_t, S_j = j) \\
&= f(r_t \mid r_1, \ldots, r_{t-1}, S_t = j) f(r_1, \ldots, r_{t-1}, S_j = j) \\
&= f(r_t \mid S_t = j) \cdot \sum_{i \in \mathcal{S}} f(r_1, \ldots, r_{t-1}, S_{t-1} = i, S_t = j) \\
&= f_j(r_t) \cdot \sum_{i \in \mathcal{S}} f(S_t = j \mid r_1, \ldots, r_{t-1}, S_{t-1} = i) f(r_1, \ldots, r_{t-1}, S_{t-1} = i) \\
&= f_j(r_t) \cdot \sum_{i \in \mathcal{S}} f(S_t = j \mid S_{t-1} = i) \alpha_{t-1}(i) \\
&= f_j(r_t) \cdot \sum_{i \in \mathcal{S}} \gamma_{ij} \alpha_{t-1}(i) \\
f_j(r_t) &\cdot \sum_{i \in \mathcal{S}} \gamma_{ij} \alpha_{t-1}(i) \\
&= \begin{bmatrix} \alpha_{t-1}(1) & \ldots & \alpha_{t-1}(N) \end{bmatrix} \begin{bmatrix} \gamma_{ij} \\ \vdots \\ \gamma_{Nj} \end{bmatrix} f_j(r_t)
\end{aligned}$$

which implies that

$$\begin{aligned}\boldsymbol{\alpha}_t &= \begin{bmatrix}\alpha_t(1) & \cdots & \alpha_t(N)\end{bmatrix} \\ &= \begin{bmatrix}\alpha_{t-1}(1) & \dots & \alpha_{t-1}(N)\end{bmatrix} \begin{bmatrix}\gamma_{11} & \cdots & \gamma_{1N} \\ \vdots & \ddots & \vdots \\ \gamma_{N1} & \cdots & \gamma_{NN}\end{bmatrix} \begin{bmatrix}f_1(r_t) & 0 & \cdots & 0 \\ 0 & f_2(r_t) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & f_N(r_t)\end{bmatrix},\end{aligned}$$

which is exactly the algorithm.

A.2.10 Proof of Probability Integral Transformation for Pseudo-Residuals

Proof. For any $u \in [0, 1]$,

$$\begin{aligned}\mathbb{P}(u_t \leq u) &= \mathbb{P}\left(F_t(r_t \mid r_{t-1}) \leq u\right) \\ &= \mathbb{P}\left(r_t \leq F_t^{-1}(u \mid r_{t-1}) \mid r_{t-1}\right) \\ &= F_t\left(F_t^{-1}(u \mid r_{t-1}) \mid r_{t-1}\right) \\ &= u,\end{aligned}$$

so $u_t \sim \mathcal{U}[0, 1]$.
For any $z \in \mathbb{R}$,

$$\begin{aligned}\mathbb{P}(z_t \leq z) &= \mathbb{P}\left(\Phi^{-1}(u_t) \leq z\right) \\ &= \mathbb{P}\left(u_t \leq \Phi(z)\right) \\ &= \Phi(z),\end{aligned}$$

so $z_t \sim \mathcal{N}(0, 1)$. □

A.3 Figures

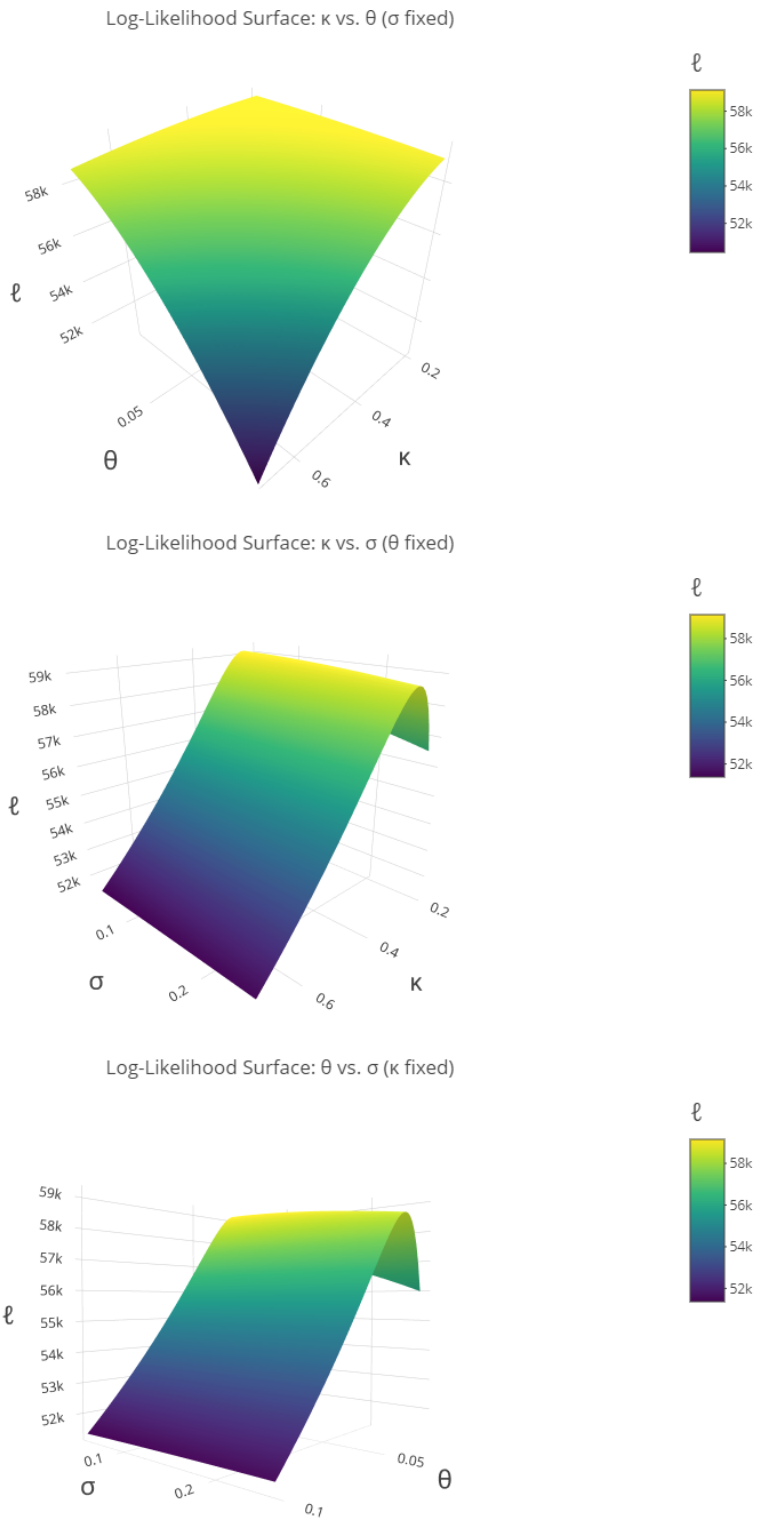


Figure A.3.1: *Log-likelihood surface of κ and θ for optimal $\sigma = 0.0745013$ in the first upper panel, κ and σ for optimal $\theta = 0.024737$ in the second middle panel and κ and σ for optimal $\kappa = 0.206817$ in the lower last panel.*

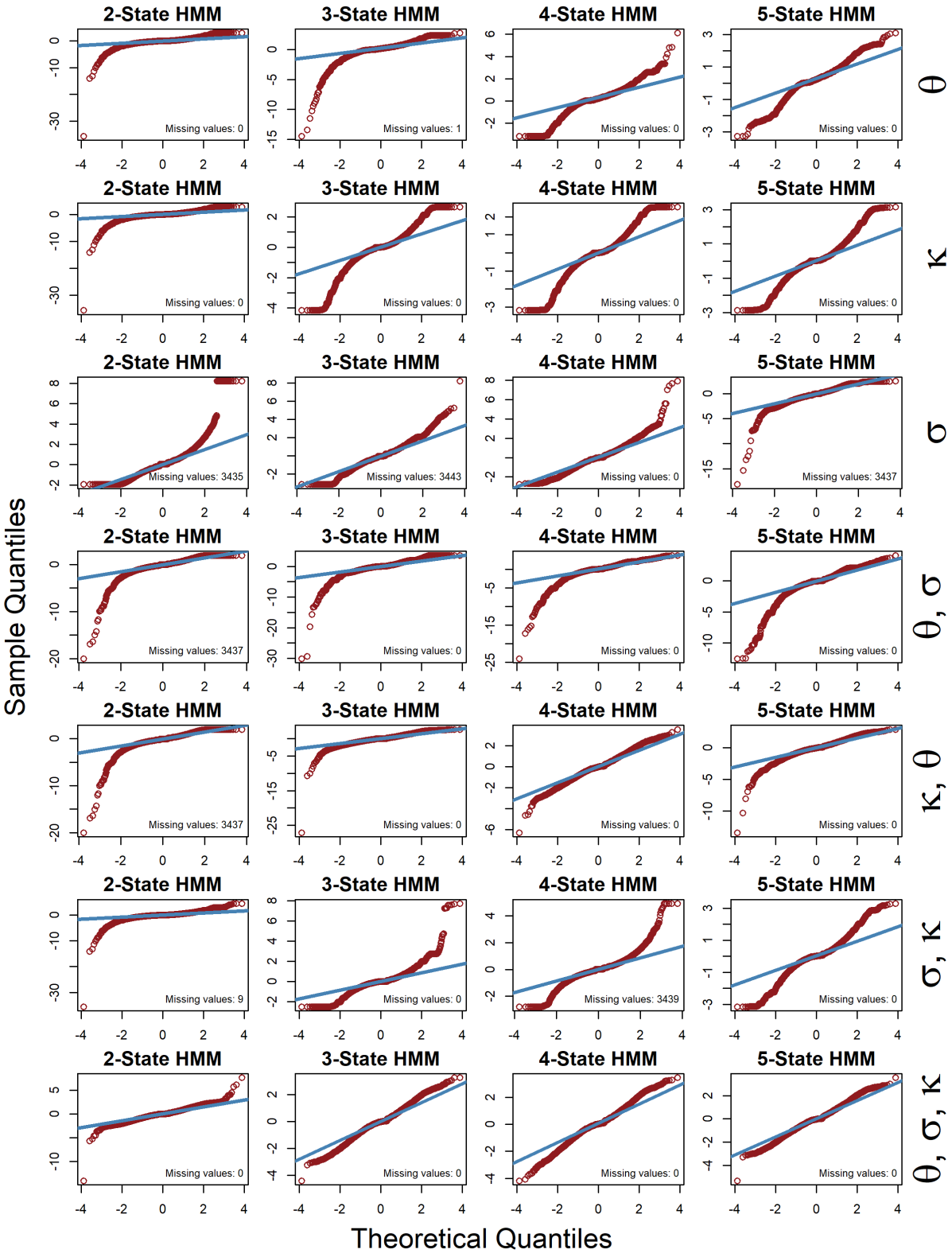


Figure A.3.2: *Pseudo-residuals Q-Q plot for the 28 fitted CIR-ARHMM's using* $\Delta = 1$.

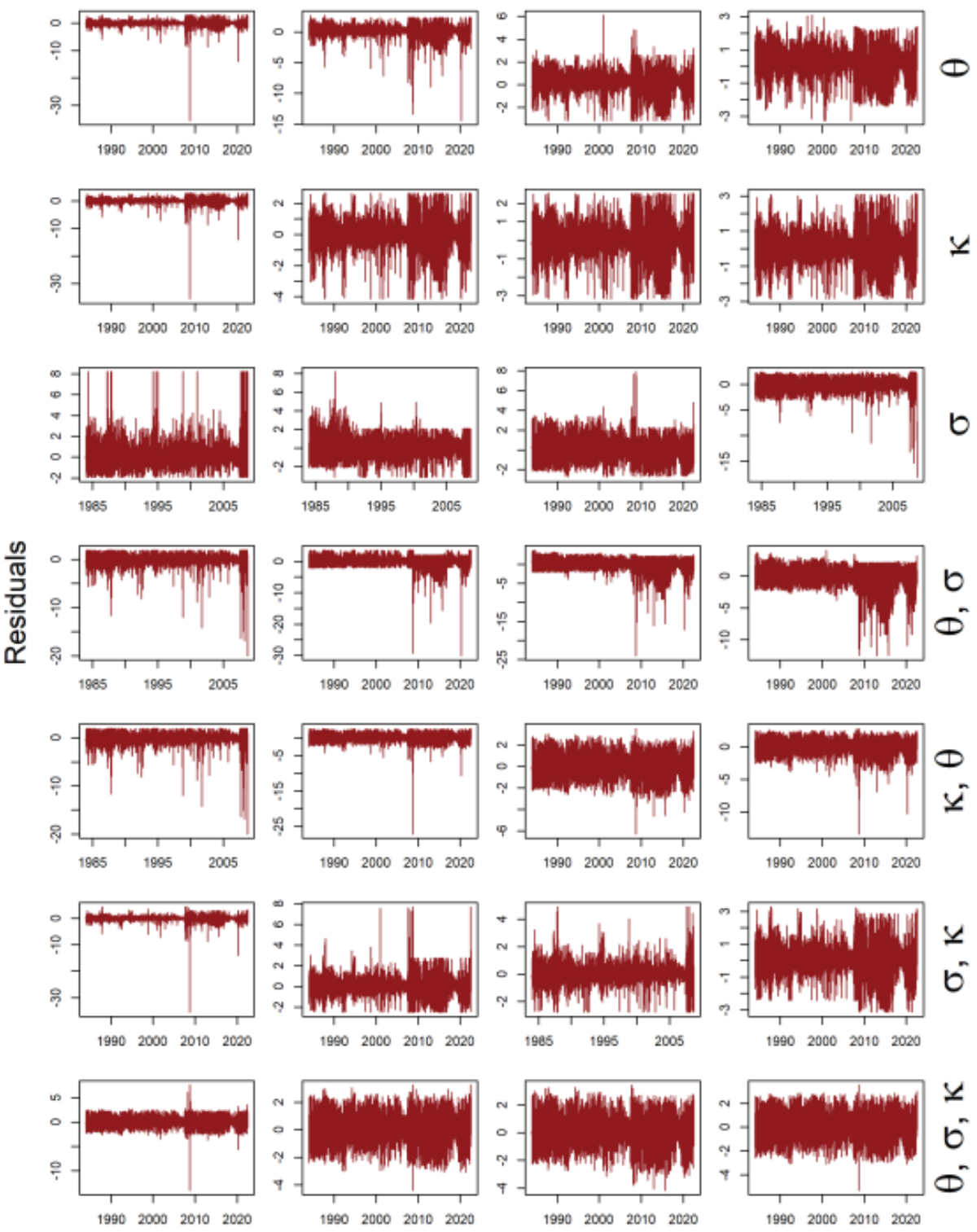


Figure A.3.3: *Time-series plot of pseudo-residuals for the 28 fitted CIR-ARHMM's using* $\Delta = 1$.

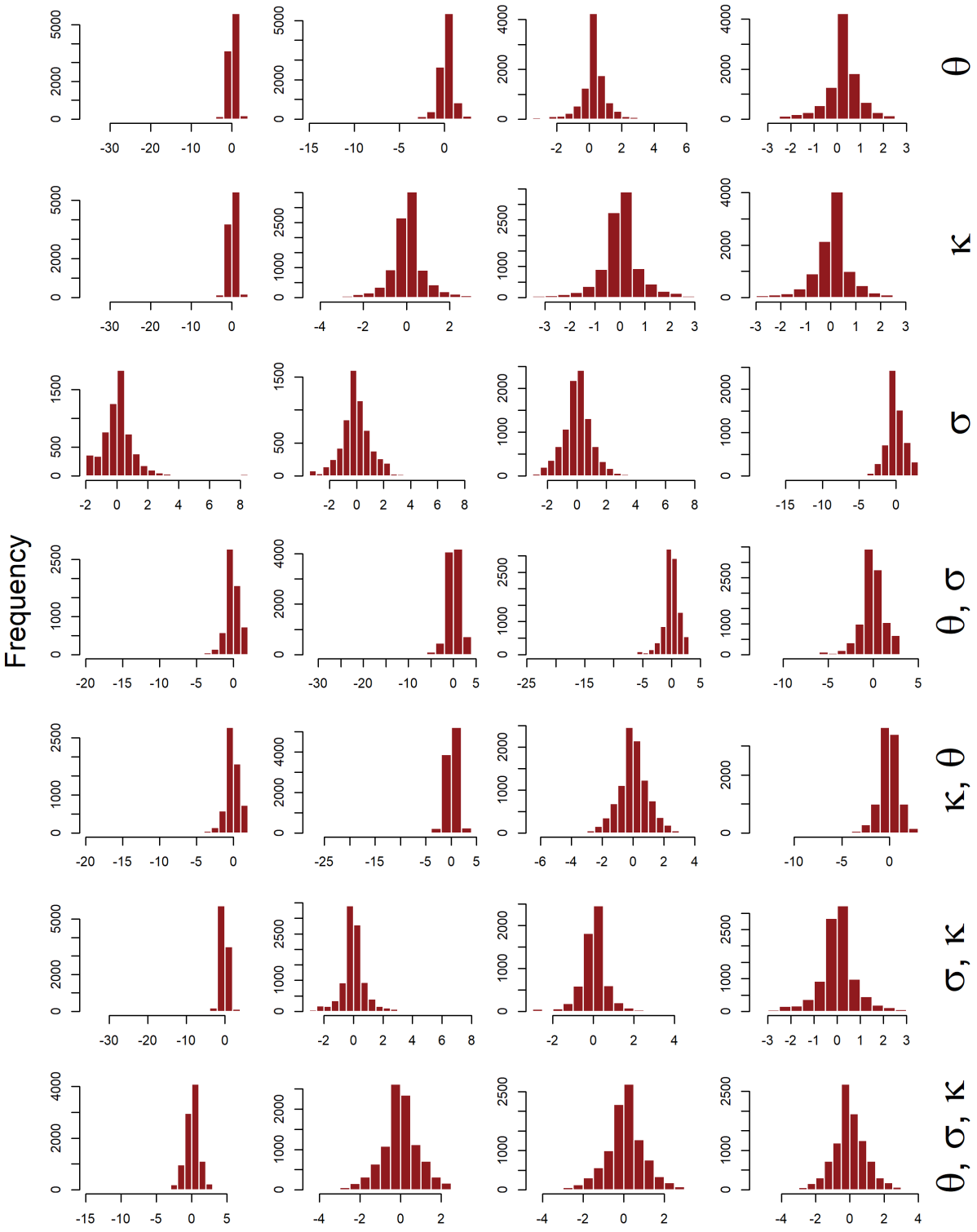


Figure A.3.4: *Histogram plot of pseudo-residuals for the 28 fitted CIR-ARHMM's using* $\Delta = 1$.

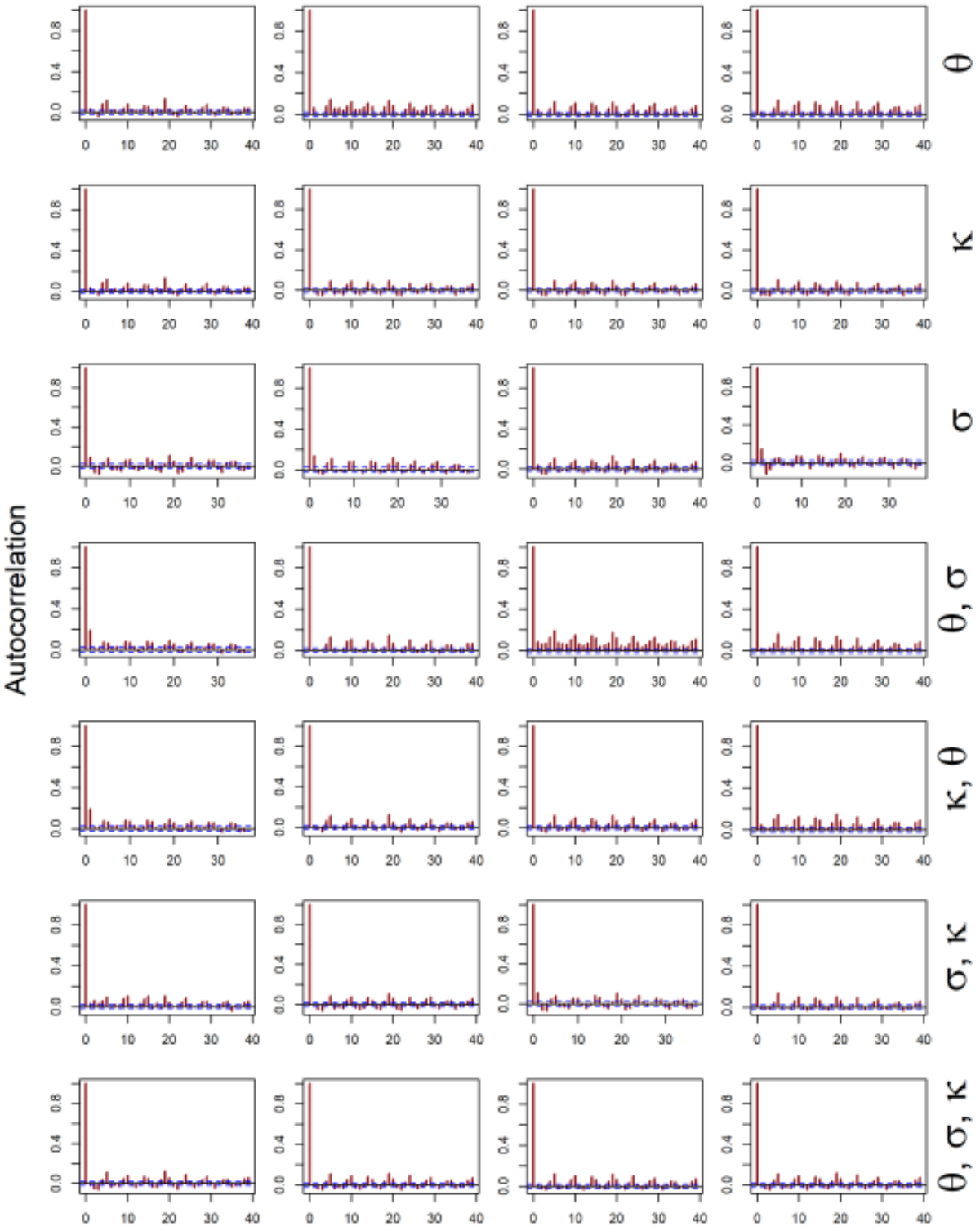


Figure A.3.5: *Autocorrelation plot of pseudo-residuals for the 28 fitted CIR-ARHMM's using* $\Delta = 1$.