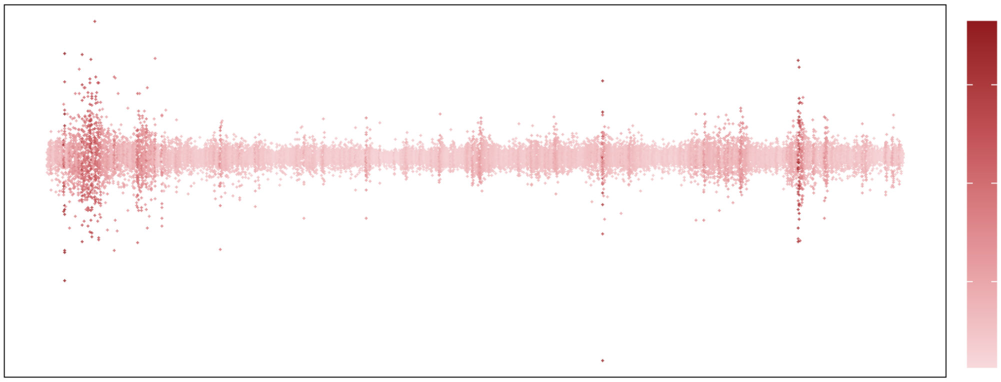
UNIVERSITY OF COPENHAGEN

DEPARTMENT OF MATHEMATICAL SCIENCES



Master's Thesis in Mathematical Finance

Youssef Raad

KU-id: zfw568

The Black-Scholes Asset Pricing Model

A Markov-Switching Extension

Date: 22-12-2025

Supervisor: Rolf Poulsen

Institute:	Institute for Mathematical Science
Department	Department of Mathematical Sciences
Author(s):	Youssef Mahmoud Raad
Title and subtitle:	The Black-Scholes Asset Pricing Model: A Markov-switching Extension
Description:	Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum porttitor rhoncus dui, vehicula pellentesque odio convallis id. Vestibulum mollis nulla ut sem luctus dictum. Sed scelerisque feugiat arcu, quis faucibus lorem viverra id. Aenean dignissim metus eu ligula sagittis, at commodo magna suscipit.
Advisor:	Rolf Poulsen
Date:	22-12-2025

Acknowledgements

A special thanks to Théo Michelot, Assistant Professor of Statistics at Dalhousie University, for his help throughout the process answering, at time, extremely trivial question.

I would like to thank Rolf Poulsen, Professor of Mathematical Finance at the University of Copenhagen, for his supervision during the preparation project and this thesis. Thank you for the opportunity and for the chance to collaborate as a teaching assistant in the course *Continuous-Time Finance II*.

Above all, and without comparison, my deepest thanks go to my mother. I would not have accomplished anything without her steady belief, patient encouragement, and constant support. She has been the light at the end of every bleak day, my reason to continue, and any success I have rests on her sacrifices and her power of will. Nothing I do will ever fully repay that debt—and I would not have it any other way. I love you.

Note

We only proof results which are not well known across both fields of statistics and mathematical finance and/or give rise to meaningful matters of learning. Furthermore, we will list definitions if they are rarely stuppled upon or subject to many varying forms of definitions. Results and definitions that we use to prove main results will be listed in Appendix A.2.

ةزيزعلا يمأ ىلإ

Abstract

CONTENTS

List of Symbols, Notation & Abbreviations

Symbol/Notation	**Description**
$\mathbb{N}$	Set of all positive integers
$\mathbb{R}$	Set of all real numbers
$\mathbb{P}$	Historical probability measure
$\mathbb{Q}$	Equivalent Martingale measure
$W_t^{\mathbb{P}}$	Brownian motion under a measure (here the measure is exemplified with $\mathbb{P}$)
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
C_t	State occupied by Markov chain at time-t
p_i	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)
S_t	Random variable at time-t (asset price)
X_t	Random variable at time-t (log-return)
$\mathbf{c}^{(-t)}$	$(c_1, \ldots, c_{t-1}, c_{t+1}, \ldots, c_T)$ (and similarly for $\mathbf{X}$, $\mathbf{S}$ and $\mathbf{V}$)
$\mathbf{C}^{(t)}$	$(C_1, C_2, \ldots, C_t)$ (and similarly for $\mathbf{X}$, $\mathbf{S}$ and $\mathbf{V}$)
$\mathbf{C}_t^T$	$(C_t, C_{t+1}, \ldots, C_T)$ (and similarly for $\mathbf{X}$, $\mathbf{S}$ and $\mathbf{V}$)
α_t	Forward probability, i.e. $\mathbb{P}(\mathbf{X}^{(t)} = \mathbf{x}^{(t)}, C_t = i)$
$\boldsymbol{\alpha}_t$	(Row) vector of forward probabilities
β_t	Backward probability, i.e. $\mathbb{P}(\mathbf{X}_{t+1}^T = \mathbf{x}_{t+1}^T \mid C_t = i)$
$\boldsymbol{\alpha}_t$	(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
$\boldsymbol{e}_i$	$(0, \ldots, 0, 1, 0, \ldots, 0)$ i.e. a (row) vector of dimension T with a 1 in the t'th entry
T	Number of observations
N	Number of states
$\boldsymbol{\phi}$	Normalized vector of forward probabilities
$\boldsymbol{\psi}$	Predicted state probabilities
μ	Drift of the Black-Scholes model
σ	Volatility of the Black-Scholes model
ρ	Autoregressive parameter
σ_ε^2	variance of the innovations ε of the AR(1) process
Δ	Time-increment between some observations
$\mathcal{C}$	State space
H	Hessian matrix
$\mathrm{SE}(\cdot)$	Standard Error of some estimator $\cdot$
$\mathcal{L}_T$	Likelihood function of T observations
ℓ_T	Log-likelihood function of T observations
$\xrightarrow{\mathcal{P}}$	Convergence in probability
$\xrightarrow{\mathcal{D}}$	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
$\mathbb{1}_{\{A\}}$	Indicator function of some set A

Abbreviation	**Description**
BS	Black-Scholes
BSM	Black-Scholes model
HMM	Hidden Markov model (sometimes we use the combination BS-HMM)
SSM	State space model (sometimes we use the combination BS-SSM)
AR(1)	Autoregressive process of order 1
t.p.m	Transition probability matrix
EMM	Equivalent Martingale measure
nlm	Non-linear maximization
AIC	Akaike information criterion
BIC	Bayesian information criterion
SDE	Stochastic differential equation
a.s.	Almost surely
CI	Confidence interval
LLN	Law of large numbers
CLT	Central limit theorem
DAG	Directed acylical graph

Introduction

The Black-Scholes model (BSM) of asset prices has long been a cornerstone of financial theory and practice. It models the asset price S_t as a geometric Brownian motion with constant expected rate of return μ and volatility σ, an assumption that leads to elegant analytical solutions for option pricing. Furthermore, the asset prices are log-normally distributed. This simplicity, however, comes at the cost of realism. In practice, asset returns exhibit time-varying volatility and occasional jumps or regime shifts that the log-normal Black–Scholes (BS) framework cannot capture. A proposed model to circumvent such jumps is Merton's Jump-Diffusion Model [33]. This model superimposes a jump component on a diffusion component of the asset price process. Formally, identifying jumps is challenging because large discrete returns can arise either from rare discontinuities or extreme diffusive shocks, making the two statistically indistinguishable in finite samples. Moreover, high-frequency data introduce microstructure noise and volatility clustering effects, which can mimic jumps and bias inference even in sophisticated econometric tests ([1], [4]). Indeed, [12] shows that jumps in financial asset prices are often erroneously identified and, in reality, are rare events that account for only a very small share of total price variation. Empirical returns often have heavier tails and more abrupt changes than the BSM assumes, indicating that the constant-σ assumption is too restrictive [41]. Indeed, it is often found that a single set of fixed parameters is inadequate to describe market dynamics across all periods. As market conditions evolve, the static BSM tends to lose predictive accuracy unless its parameters are frequently recalibrated [18]. This need for continual "tuning" of μ and σ underscores a key limitation of the classical model. That is, it is not well suited for forecasting in environments where volatility and other characteristics change over time.

To address the BSM's limitations in forecasting, this thesis considers two extensions that relax the assumption of constant parameters. The first is a BSM with a Hidden Markov Model (BS-HMM) for its parameters. In this regime-switching extension, the asset price still follows a diffusion as in BS, but its drift and/or volatility can switch between a finite set of states (regimes). These regime changes are governed by a hidden Markov chain. For instance, the market might alternate between "low-volatility" and "high-volatility" regimes, each with its own σ, and possibly different μ, with probabilistic transitions between regimes. Such an HMM-based approach can capture phenomena like bull and bear market regimes or sudden volatility shifts that the classical model would miss. By allowing discrete shifts in parameters, the BS-HMM can dynamically adapt to structural changes in the data. We hypothesize that a BS-HMM will improve forecast accuracy by accounting for the possibility of regime shifts (e.g. an upcoming turbulent period) that a single-regime model would underestimate.

The second extension is a BSM with a Continuous State-Space Model (BS-SSM) for the parameters. In this approach, the drift and volatility are treated as latent continuous-time state variables that evolve according to their own stochastic process, specifically, an autoregressive process of order 1. The observable data, the prices, are linked to these hidden states, forming a state-space model. Essentially, this is akin to a SVM. Volatility is no longer fixed, but changes over time in a continuous manner, and we infer its path from the observed prices. Such state-space formulations are very flexible, allowing σ_t and μ_t to vary at each time step and capturing gradual shifts or cyclical patterns in volatility that a BS-HMM might not fully reflect. Conceptually, the BS-SSM treats the BS equation as having time-dependent parameters μ_t, σ_t governed by an underlying dynamical system. This continuous adaptation may better capture the nuanced evolution of market risk over time. Like the BS-HMM, the BS-SSM relaxes the constant parameter assumption, but it does so in a way that allows for more gradual, continuous changes rather than abrupt switches. We expect that by tracking a latent volatility factor, the BS-SSM can forecast future price distributions more accurately during periods of slowly changing market conditions or volatility trends.

The overarching research question addressed in this thesis is: Do these extended BSMs deliver superior out-of-sample forecasting performance compared to the BSM? In other words, we will test whether incorporating either discrete regime shifts or continuous stochastic parameter evolution leads to more accurate predictions of future asset prices than the traditional model with fixed μ and σ. This question is of both academic interest and practical importance. If the extended models can demonstrably improve forecast accuracy, it would suggest a pathway to better risk management and derivative pricing by acknowledging and modeling the non-stationarity in asset dynamics. Conversely, if the extensions do not improve forecasts, that finding is also informative as it would imply that the added complexity is not worth the trouble for prediction, and the basic BS perhaps with frequent recalibration remains surprisingly hard to beat.

To answer this question, we conduct an empirical comparison using historical data on the S&P 500 index. The methodology involves estimating each model (the classical BS, the BS-HMM, and the BS-SSM) on a training sample and then evaluating their predictive performance on a hold-out sample. The models' parameters are fitted via maximum likelihood estimation, ensuring that each model is calibrated to the historical dynamics of the index. Once calibrated, each model generates out-of-sample forecasts of the S&P 500's price (or rather, log-return distribution) for the evaluation period. The key focus is on out-of-sample forecasting performance. By evaluating the predictions on data not used for estimation, we ensure a fair test of whether the additional flexibility of the HMM or state-space formulations translates into better predictive power. In short, we wish to extend the BSM by state-space models to examine if regime-switching does solve known issues of the BSM.

1 Data (I of II)

The Standard and Poor's 500 (S&P 500) will be used throughout for analysis. The S&P 500 is a free-float–adjusted, value-weighted index of large-capitalization U.S. equities maintained by S&P Dow Jones Indices. By construction it spans major sectors of the U.S. economy and concentrates on firms with substantial public float, liquidity, and operating history, yielding a diversified portfolio in which aggregate—rather than idiosyncratic—risk dominates variation. Its methodology and eligibility criteria (e.g., float adjustment, sector classification, profitability and size thresholds, and scheduled reconstitutions) are transparent and stable over time, producing a well-documented data-generating mechanism that supports reproducible empirical work. We will work with the price version of the index, using daily data from Yahoo Finance (`^GSPC`). This series excludes cash dividends. S&P also publishes total return (TR) and net total return (NTR) variants (`^SP500TR`, `^SP500NTR`) that reinvest dividends gross or net of withholding taxes, respectively (see [50, 49, 47, 56]).

For asset-price analysis, the index offers several practical advantages. It captures a large share of total U.S. equity market capitalization and trading volume, which enhances the signal-to-noise ratio of return observations and reduces the impact of microstructure frictions at daily horizons. The series is long, clean, and consistently adjusted for corporate actions, enabling inference across multiple macroeconomic episodes without survivorship bias tied to current constituents.

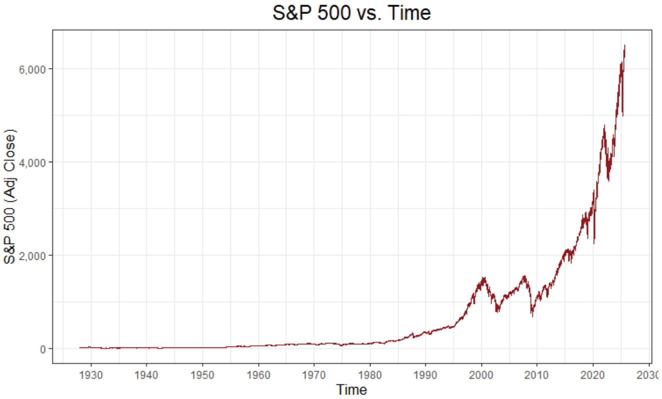


Figure 1: *S&P 500 index time series.*

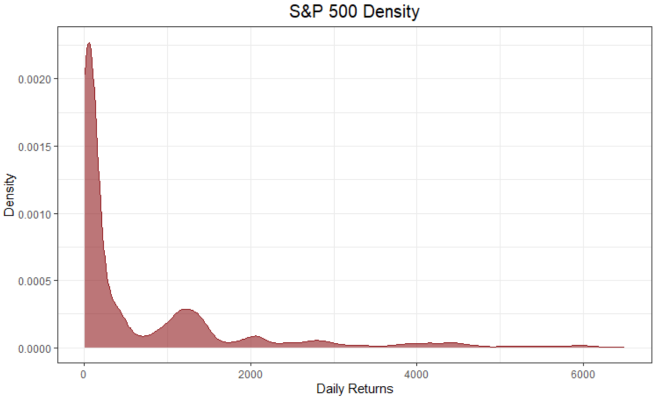


Figure 2: *Kernel density of the S&P 500 index.*

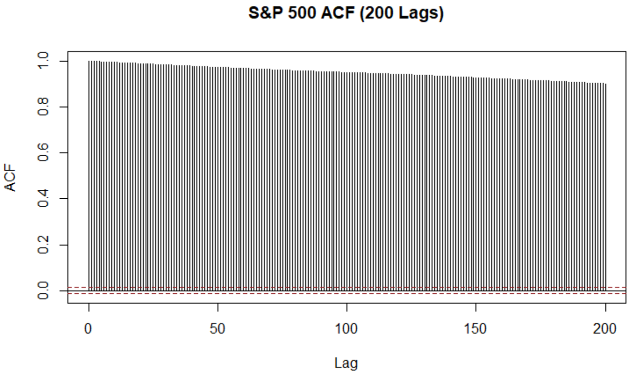


Figure 3: *Autocorrelation of closing prices (200 lags).*

Return dynamics display the canonical equity stylized facts—near-zero unconditional mean at short horizons, volatility clustering, heavy tails relative to Gaussian benchmarks, leverage effects, and episodic regime shifts—making the index a natural benchmark for assessing models of systematic risk. The first observation falls on 1927-12-30 and the last on the end of the trading day 2025-09-05 as analysis has begun at this time.

The time-series is shown in Figure 1. The density in Figure 2 places most mass near the origin: first quartile = 24.86, median = 103.21, and third quartile = 1076.22. The smallest observation is 4.40 and the largest is 6502.80. The shape appears multi-modal (roughly six modes, with the first three most prominent), which we keep in mind when discussing state-number selection. Finally, Figure 3 shows strong persistence in levels over 200 lags, as expected for nonstationary prices.

Dividend Treatment and Construction of a Daily Dividend–Yield Series In this thesis, we adopt the Black–Scholes specification with constant parameters (μ, σ, q). With a continuous dividend yield q, the ex-dividend price index S_t satisfies

$$\frac{dS_t}{S_t} = (\mu - q)dt + \sigma dW_t^{\mathbb{P}}, \qquad d\ln S_t = \big(\mu - q - \tfrac{1}{2}\sigma^2\big)dt + \sigma dW_t^{\mathbb{P}},$$

so estimation on the price index identifies the capital-gains drift $\mu - q$ rather than the total-return drift μ. By contrast, the S&P total-return (`^TR`) index reinvests ordinary cash dividends on the ex-date according to S&P's index mathematics; the net-total-return (NTR) variant additionally accounts for withholding taxes [48, 51].

Post-1988 (Daily): TR–PR Differencing Used to Estimate a Constant q Let P_t denote the S&P 500 price index level (e.g., `^GSPC`) and T_t the corresponding total-return level (e.g., `^TR`). Over one trading day, for $t = 2, 3, \ldots, T$

$$r_t^{PR} = \frac{P_t}{P_{t-1}} - 1, \qquad r_t^{TR} = \frac{T_t}{T_{t-1}} - 1, \qquad 1 + r_t^{TR} = (1 + r_t^{PR})(1 + r_t^{div}).$$

Hence the dividend simple return is $r_t^{div} = \frac{1+r_t^{TR}}{1+r_t^{PR}} - 1$, and the log (continuous) dividend yield is

$$q_t^{(\log)} = \ln\frac{T_t}{T_{t-1}} - \ln\frac{P_t}{P_{t-1}}.$$

We use $\left\{q_t^{(\log)}\right\}$ solely to estimate a single constant dividend yield by averaging and annualizing:

$$\widehat{q} \; := \; 252 \cdot \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} q_t^{(\log)},$$

where $\mathcal{T}$ is the chosen sample window (e.g., the full estimation period or a arbitrary (possibly non-consecutive) subperiod of dates). Public TR series are broadly available from the late 1980s onward [57, 42].

Pre-TR Period (Monthly): Backfill from Shiller Before daily TR series are readily available, we employ Shiller's long-run monthly S&P price and dividend data [45, 44]. Let P_m be the month-end price index and D_m^{TTM} the trailing-twelve-month dividend total reported for month m; the implied monthly flow is $d_m := D_m^{TTM}/12$. The monthly log dividend yield is

$$q_m^{(\log)} = \ln\left(1 + d_m/P_m\right).$$

To obtain a daily series that preserves each month's total, distribute $q_m^{(\log)}$ evenly across the N_m business days in month m:

$$q_t^{(\log)} = \frac{q_m^{(\log)}}{N_m} \quad (t \in m),$$

so that $\sum_{t\in m} q_t^{(\log)} = q_m^{(\log)}$ by additivity of log returns. This daily backfill is used only to compute the constant estimator $\widehat{q}$ over samples that include pre-1988 months. As a cross-check, S&P's Dividend Points indices track cumulative dividends in index points and reset annually [46].

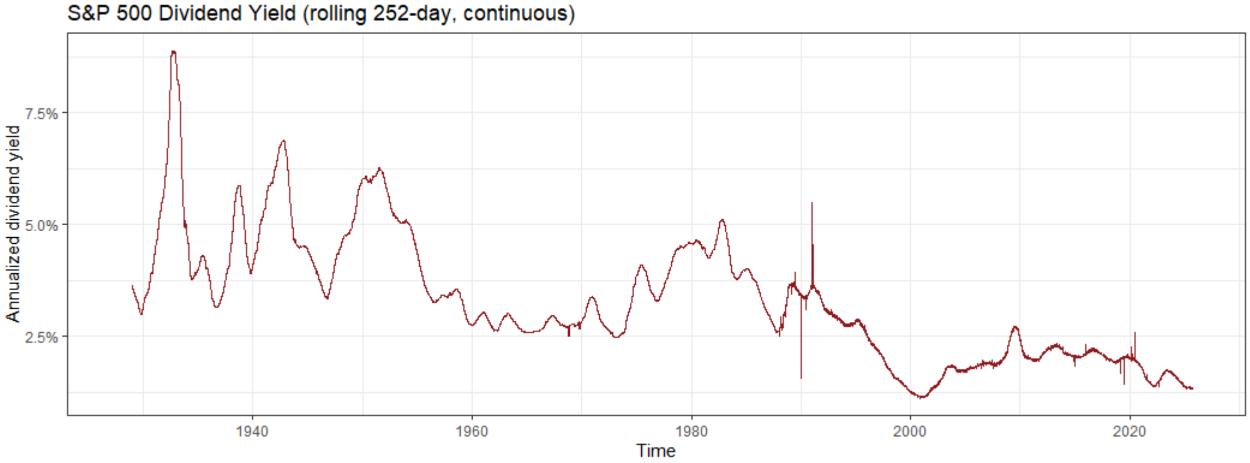


Figure 4: *Annualized dividend yield vs. time for the S&P 500 estimated using Shiller's data and difference of the price- and total return index.*

We visualize the estimated dividend yields in Figure 4. Over the full sample (1927–2025), the estimated average continuous annualized dividend yield is approximately 3.3% with a standard error of approximately $10^{-4}\times 9.2$[1]. The maximum is 8.9% on 2 November 1932, while the minimum is 1.1% on 27 November 2000. Dividends has been systematically lower over the years which could be a direct product of more liquid markets.

For the extensions of the BSM, we return to the issue of dividends after the Theory & Methodology section, namely in Section 3 of the thesis.

[1] How the standard error is calculated will be shown in Section 3 as we need to define concepts beforehand.

2 Theory & Methodology

2.1 The Black–Scholes Model

Model under $\mathbb{P}$ The Black and Scholes model [7] (or Black, Scholes and Merton model [32]) assumes that there is a riskless asset with interest rate r such that the bank/money account B has time-varying dynamics

$$dB_t = rB_t dt,$$

and that the dynamics of the price of the underlying asset under the historical probability measure $\mathbb{P}$ are

$$dS_t = (\mu - q)S_t dt + \sigma S_t dW_t^{\mathbb{P}}, \quad S_0 = s_0 \tag{1}$$

It is assumed in the Black-Scholes model that r, μ and σ, where $q \geq 0$ denotes a constant dividend yield, and that, for valuation purposes, that μ is an $\mathcal{F}$-adapted process. The solution of Equation 1 can easily be found by using the transformation $Z_t = \log(S_t)$, where we assume that a soluation exists and that S_t is a (strictly positive) solution . Itô's formula [6, Thm. 4.19] gives

$$\begin{aligned} dZ_t &= \frac{1}{S_t} dS_t + \frac{1}{2}\left(-\frac{1}{S_t^2}\right)(dS_t)^2 \\ &= \frac{1}{S_t}\left((\mu - q)S_t dt + \sigma S_t dW_t^{\mathbb{P}}\right) + \frac{1}{2}\left(-\frac{1}{S_t^2}\right)\sigma^2 S_t^2 dt \\ &= \left((\mu - q)dt + \sigma dW_t^{\mathbb{P}}\right) - \frac{1}{2}\sigma^2 dt. \end{aligned}$$

This leaves us with the equation

$$dZ_t = \left(\mu - q - \frac{\sigma^2}{2}\right) dt + \sigma dW_t^{\mathbb{P}}, \quad Z_0 = \log s_0. \tag{2}$$

Note that no occurences of the r.v. Z_t is on the RHS of Equation 2. By implication, integrating yields

$$Z_t = \log(s_0) + \left(\mu - q - \frac{\sigma^2}{2}\right) t + \sigma W_t^{\mathbb{P}},$$

which means we obtain the solution to Equation 1 by reversing the log-transformation

$$S_t = s_0 \exp\left((\mu - q)t + \sigma W_t^{\mathbb{P}} - \frac{\sigma^2}{2} t\right). \tag{3}$$

From this point forward, we shall only consider a fixed time horizon $[0, \infty]$.

Model under $\mathbb{Q}$ As we want to price options in the BSM, we derive the $\mathbb{Q}$-dynamics of Equation 1, i.e. the dynamics under the equivalent martingale measure, $\mathbb{Q}$.

Define the dividend-adjusted price $\bar{S}_t := e^{qt}S_t$ (i.e. ex-dividend price times the cum-dividend factor). By Itô's formula [6, Thm. 4.19],

$$d\bar{S}_t = e^{qt}dS_t + qe^{qt}S_t dt = (\mu\bar{S}_t)dt + \sigma\bar{S}_t dW_t^{\mathbb{P}}.$$

Thus $\bar{S}$ behaves like a non-dividend stock with drift μ under $\mathbb{P}$. Let the (constant) market price of risk be $\theta := \frac{\mu-r}{\sigma}$. Define the density process

$$\Lambda_t := \exp\left(-\theta W_t^{\mathbb{P}} - \tfrac{1}{2}\theta^2 t\right).$$

Set $d\mathbb{Q}|_{\mathcal{F}_t} := \Lambda_t d\mathbb{P}|_{\mathcal{F}_t}$. By Girsanov's theorem [6, Thm. 12.3],

$$W_t^{\mathbb{Q}} := W_t^{\mathbb{P}} + \theta t,$$

is a $\mathbb{Q}$–Brownian motion, and the $\bar{S}$ dynamics become

$$d\bar{S}_t = \mu\bar{S}_t dt + \sigma\bar{S}_t dW_t^{\mathbb{P}} = (\mu - \sigma\theta)\bar{S}_t dt + \sigma\bar{S}_t dW_t^{\mathbb{Q}} = r\bar{S}_t dt + \sigma\bar{S}_t dW_t^{\mathbb{Q}}.$$

Since $\bar{S}_t = e^{qt}S_t$, the $\mathbb{Q}$–dynamics of S are

$$dS_t = (r-q)S_t dt + \sigma S_t dW_t^{\mathbb{Q}}, \qquad S_0 = s_0.$$

Equivalently, the discounted cum-dividend price is a $\mathbb{Q}$–martingale:

$$\tilde{S}_t := B_t^{-1}\bar{S}_t = e^{(q-r)t}S_t, \qquad d\tilde{S}_t = \sigma\tilde{S}_t dW_t^{\mathbb{Q}}.$$

The explicit $\mathbb{Q}$–solution is

$$S_t = s_0 \exp\Big(\big(r - q - \tfrac{1}{2}\sigma^2\big)t + \sigma W_t^{\mathbb{Q}}\Big).$$

As the $\mathbb{Q}$-dynamics are now readily available, we proceed to pricing a European time-t call option under the BSM.

Proposition 2.1. *Fix constants $r \in \mathbb{R}$, $q \in \mathbb{R}$ and $\sigma > 0$, and let $(W_t^{\mathbb{Q}})_{t\geq 0}$ be a standard Brownian motion on a filtered probability space satisfying the usual conditions. Under the risk–neutral measure $\mathbb{Q}$, the (cum-dividend) stock price S follows*

$$dS_t = (r-q)S_t dt + \sigma S_t dW_t^{\mathbb{Q}}, \qquad S_t = s > 0.$$

For $0 \le t < T$, set $\tau := T - t$. Then the arbitrage–free time-t price $F(t,s)$ of a European call with strike $K > 0$ and maturity T is

$$F(t,s) = s\, e^{-q\tau}\, \Phi(d_1) - K\, e^{-r\tau}\, \Phi(d_2),$$

where

$$d_1 = \frac{\ln(s/K) + (r - q + \frac{1}{2}\sigma^2)\tau}{\sigma\sqrt{\tau}}, \qquad d_2 = d_1 - \sigma\sqrt{\tau}.$$

Proof. Set $\tau := T - t$. Solving the SDE under $\mathbb{Q}$ gives the (risk–neutral) GBM solution

$$S_T = s\exp\Big((r - q - \tfrac{1}{2}\sigma^2)\tau + \sigma\big(W_T^{\mathbb{Q}} - W_t^{\mathbb{Q}}\big)\Big),$$

where $S_t = s$ is the current underlying price. Write

$$Z := \big(r - q - \tfrac{1}{2}\sigma^2\big)\tau + \sigma\sqrt{\tau}\, Y, \qquad Y := \frac{W_T^{\mathbb{Q}} - W_t^{\mathbb{Q}}}{\sqrt{\tau}},$$

so that, conditional on $\mathcal{F}_t$ (or equivalently on $S_t = s$),

$$Y \sim \mathcal{N}(0,1) \text{ under } \mathbb{Q}, \qquad Z \sim N\big(m, v^2\big)\,,$$

with

$$m := (r - q - \tfrac{1}{2}\sigma^2)\tau, \qquad v^2 := \sigma^2\tau, \qquad \text{and} \qquad S_T = se^Z.$$

Let $a := \ln(K/s)$. Using risk–neutral valuation and conditioning on $S_t = s$,

$$\begin{aligned} F(t,s) &= e^{-r\tau}\, \mathbb{E}^{\mathbb{Q}}\big[(S_T - K)^+ \mid \mathcal{F}_t\big] \\ &= e^{-r\tau}\Big(s\, \mathbb{E}^{\mathbb{Q}}\big[e^Z \mathbb{1}_{\{Z \ge a\}} \mid \mathcal{F}_t\big] - K\, \mathbb{Q}(Z \ge a)\Big). \end{aligned}$$

We evaluate the two terms by a standard change of variables and completing the square. First, with $Y = (Z - m)/v$,

$$\begin{aligned} \mathbb{E}^{\mathbb{Q}}\big[e^Z \mathbb{1}_{\{Z \ge a\}} \mid \mathcal{F}_t\big] &= \int_a^\infty e^z\, \frac{1}{v\sqrt{2\pi}} \exp\left(-\frac{(z-m)^2}{2v^2}\right)\, dz \\ &= e^{m + \frac{1}{2}v^2} \int_{\frac{a-m}{v}}^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-v)^2}{2}\right)\, dy \\ &= e^{m + \frac{1}{2}v^2}\, \Phi\left(\frac{m + v^2 - a}{v}\right). \end{aligned}$$

Second,

$$\mathbb{Q}(Z \ge a) = \int_{\frac{a-m}{v}}^\infty \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy = \Phi\left(\frac{m-a}{v}\right).$$

Putting these together,

$$F(t,s) = e^{-r\tau}\Big(s\,e^{m+\frac{1}{2}v^2}\,\Phi\big(\tfrac{m+v^2-a}{v}\big) - K\,\Phi\big(\tfrac{m-a}{v}\big)\Big).$$

Now substitute $m = (r-q-\frac{1}{2}\sigma^2)\tau$, $v = \sigma\sqrt{\tau}$, $a = \ln(K/s)$. Noting that

$$e^{-r\tau}\,s\,e^{m+\frac{1}{2}v^2} = e^{-r\tau}\,s\,e^{(r-q)\tau} = s\,e^{-q\tau},$$

and

$$\frac{m+v^2-a}{v} = \frac{\ln(s/K)+(r-q+\frac{1}{2}\sigma^2)\tau}{\sigma\sqrt{\tau}} = d_1, \qquad \frac{m-a}{v} = \frac{\ln(s/K)+(r-q-\frac{1}{2}\sigma^2)\tau}{\sigma\sqrt{\tau}} = d_2,$$

we obtain the Black–Scholes call price with continuous dividend yield q:

$$F(t,s) = s\,e^{-q\tau}\,\Phi(d_1) - K\,e^{-r\tau}\,\Phi(d_2),$$

as claimed. □

Distributional Properties By normality of the Wiener process increments [6, Def. 4.1], zero-mean property [6, Prop. 4.5] and Itô isometry [6, Prop. 4.5] the distribution of Z_t (and equivalentily S_t) is

$$Z_t \sim \mathcal{N}\left(\log(s_0) + \left(\mu - q - \frac{\sigma^2}{2}\right)t, \sigma^2 t\right)$$
$$\iff$$
$$S_t \sim \text{Lognormal}\left(\log(s_0) + \left(\mu - q - \frac{\sigma^2}{2}\right)t, \sigma^2 t\right)$$

Now, define the continuously compounded log-return over the time horizon $[t, t-1]$ as

$$X_t = \log\left(\frac{S_t}{S_{t-1}}\right) \tag{4}$$

Using our solution Equation 3 for Equation 1 and by normality of the Wiener process increments [6, Def. 4.1] yields

$$X_t = \left(\mu - q - \frac{\sigma^2}{2}\right)t + \sigma\left(W_t^{\mathbb{P}} - W_{t-1}^{\mathbb{P}}\right) \sim \mathcal{N}\left(\left(\mu - q - \frac{\sigma^2}{2}\right)\Delta, \sigma^2\Delta\right), \tag{5}$$

such that the corresponding probability density function is

$$f_{X_t}(x_t) = \frac{1}{\sqrt{2\pi\sigma\Delta}} \exp\left(-\frac{\left(x_t - \left(\mu - q - \frac{\sigma^2}{2}\right)\Delta\right)^2}{2\sigma^2\Delta}\right), \quad x_t \in \mathbb{R} \tag{6}$$

where $\Delta := t_i - t_{i-1}$ is used to denote the time-increment. Note that if we have T observations of the asset price we will have $T-1$ observations of the log-returns.

2.1.1 Likelihood Formulation & Parameter Estimation

Let $\{X_t\}_{t=1}^T$ denote T observed log-returns with fixed step size $\Delta > 0$. From Equation 5, under $\mathbb{P}$ the returns are i.i.d.

$$X_t \sim \mathcal{N}\Big(\big(\mu - q - \tfrac{1}{2}\sigma^2\big)\Delta,\ \sigma^2\Delta\Big),$$

where μ and $\sigma > 0$ are constant drift and volatility parameters of the BSM.

Firstly note that, a parametric model $\{\mathcal{P}_\upsilon : \upsilon \in \mathcal{H} \subset \mathbb{R}^3\}$ is identifiable iff $\mathcal{P}_\upsilon = \mathcal{P}_{\upsilon'} \Rightarrow \upsilon = \upsilon'$. Suppose there exists a (non-injective) map $\tau : \mathcal{H} \to \Theta \subset \mathbb{R}^2$ and a two–parameter family $\{\mathcal{Q}_\iota : \iota \in \Theta\}$ such that $\mathcal{P}_\upsilon = \mathcal{Q}_{\tau(\upsilon)}$. Then for any $\upsilon \neq \upsilon'$ with $\tau(\upsilon) = \tau(\upsilon')$ we have $\mathcal{P}_\upsilon = \mathcal{P}_{\upsilon'}$, so $\upsilon \mapsto \mathcal{P}_\upsilon$ is not injective and the 3-parameter parametrization is not identifiable. Under standard regularity (differentiable τ, interior point, nonsingular $\mathcal{I}_\iota$), the Fisher information satisfies

$$\mathcal{I}_\upsilon(\upsilon) = D\tau(\upsilon)^\top \mathcal{I}_\iota(\tau(\upsilon))\, D\tau(\upsilon), \quad \Rightarrow \quad \operatorname{rank}\mathcal{I}_\upsilon \leq \operatorname{rank} D\tau(\upsilon) \leq 2 < 3,$$

confirming lack of local identifiability.

As such, for MLE we estimate the *capital-gains* drift μ_{cap} and retrieve the *total-return* drift via $\mu_{\text{total}} = \mu_{\text{cap}} + \hat{q}$, where $\hat{q}$ is the estimated continuous dividend yield (see Section 1). Henceforth we work explicitly with $\mu := \mu_{\text{cap}}$ unless otherwise specified

For a realization $\mathbf{x}^{(T)} = (x_1, \ldots, x_T)$ of log-returns with step Δ, the likelihood and log-likelihood are

$$\mathcal{L}_T(\mu, \sigma) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\,\sigma^2\Delta}} \exp\left\{-\frac{\left[x_t - (\mu - \frac{1}{2}\sigma^2)\Delta\right]^2}{2\,\sigma^2\Delta}\right\}$$

$$\Longleftrightarrow$$

$$\ell_T(\mu, \sigma) = -\frac{T}{2}\log(2\pi\sigma^2\Delta) \; - \; \frac{1}{2\sigma^2\Delta}\sum_{t=1}^{T}\Big[x_t - (\mu - \tfrac{1}{2}\sigma^2)\Delta\Big]^2. \tag{7}$$

2.1.2 Simulation

To simulate the standard Black-Scholes model (and the extended models), we need to develop a discretization scheme to simulate the continuous time process defined in Equation 1. To simulate paths for (S_t) at discrete times $\mathscr{T} = \{t_i\}_{i=1}^T$, we generate random samples of $(S_{t+\Delta})$ given (S_t) for any increment Δ. Repeatedly appending increments constructs the complete path $(S_t)_{t\in\mathscr{T}}$. We derive the Euler discretization scheme, often attributed to the work of [25].

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

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

where $W_t^{\mathbb{P}}$ is a Wiener process under $\mathbb{P}$. 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 the incremented distance $t + \Delta$ yields

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

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 $Z^{\mathbb{P}} \sim \mathcal{N}(0, 1)$. The stochastic integral is approximated as

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

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

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

Applying Euler discretization to dS_t in equation Equation 1 by substituting the diffusion and drift of dr_t into Equation 10 yields the final discretization Euler scheme of the Black-Scholes' model dynamics

$$\hat{S}_{t+\Delta} = \hat{S}_t + \mu\hat{S}_t\Delta + \sigma\hat{S}_t\sqrt{\Delta}Z^{\mathbb{P}}. \tag{11}$$

The discretization scheme does induce a discretization error to the continuous time process which is highly dependent on parameter subsets and discretization grid roughness as dictated by Δ. For a thorough examination and comments see the well-known work of [25, 8, 29].

Simulating We simulate the BSM with parameters $\mu = 0.05$, $\sigma = 0.15$, $S_0 = 100$ and $n = 25000$ over one realization of the stock price with daily observations, implying $\Delta = 1/252$ corresponding to approximately $25000/252 \approx 99$ years. Furthermore, we use the log-return X_t formulation in the implementation as given in Equation 4. The simulated stock price path is seen in Figure 5 where the used method of discretization was that developed in Section 2.1.2. The estimated values along side the true values are seen in Table 1. Parameter estimates were found using the `nlm`-function (Non-Linear Minimization) [14] in `R` for consistency, although closed-form solutions are available. `nlm` is extremely popular for HMMs (see [53], [35], [30], [38], [58]) and provide Hessians along side extremely fast function evaluations via a Newtonian-style algorithm, as compared to i.e. the Nelder–Mead technique which is a heuristic search method. As is quite evident, noting the Euler discretization scheme error, the estimates are accurate.

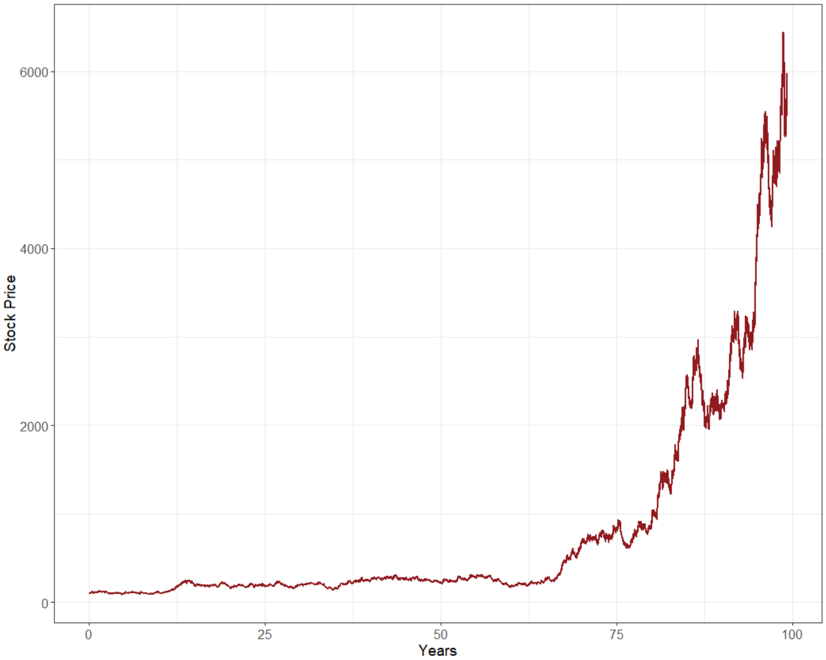


Figure 5: *A simulated stock price path S_t in the BSM.*

Parameter	True Value	Estimated Value	Relative Error (%)
μ	0.05000	0.05224	4.47
σ	0.15000	0.15002	0.0107

Table 1: *True vs. maximum likelihood estimated BSM parameters, μ and σ. Maximum likelihood estimates were found using direct numerical maximization using the `nlm`-function.*

2.2 Hidden Markov Models

2.2.1 Independent Mixture Models

A basic but useful method to dealing with overdispersed data is with that of a mixture model. Mixture models capture unobserved population heterogeneity: the overall population comprises latent groups, each with its own distribution for the observed variable.

A independent mixture distribution consists of $N \in \mathbb{N} \setminus \{1\}$ component distributions and a mixing distribution which choses between the different components. Let $\delta_1, \ldots, \delta_N$ denote the probabilities of the N different components and $f_1, \ldots, f_N$ denote their probability density functions. Let X denote the continuous r.v. that has mixture distribution and C the discrete r.v that performs the mixing, i.e.

$$\mathbb{P}(C=k)=\begin{cases}\delta_1, & k=1,\\ \vdots & \vdots\\ \delta_N, & k=N,\\ 0, & \text{otherwise,}\end{cases} \qquad \text{with} \;\; \delta_k \geq 0,\; \sum_{k=1}^{N}\delta_k = 1.$$

The density function of X is then

$$f_X(x)=\sum_{i=1}^{N} f_{X|C}(x \mid i)\mathbb{P}(C=i)=\sum_{i=1}^{N}\delta_i f_{X,i}(x).$$

Figure 6 illustrates the generative story of an independent finite mixture with two components. For each observation x_j, a selector $C_j \in \{1,2\}$ is first drawn with $P(C_j = i) = \delta_i$. In each row, the filled dot in the left panel marks the realized C_j (component 1 or 2). The middle panel shows the component densities f_1 and f_2 (stylized), a local x–axis, and a short vertical tick at the location of the realized x_j. The right panel lists the observation labels. Across rows the selectors C_j are i.i.d. with mixing weights $(\delta_1, \delta_2) = (0.68, 0.32)$; conditional on C_j, the x_j are drawn from the corresponding component distribution.

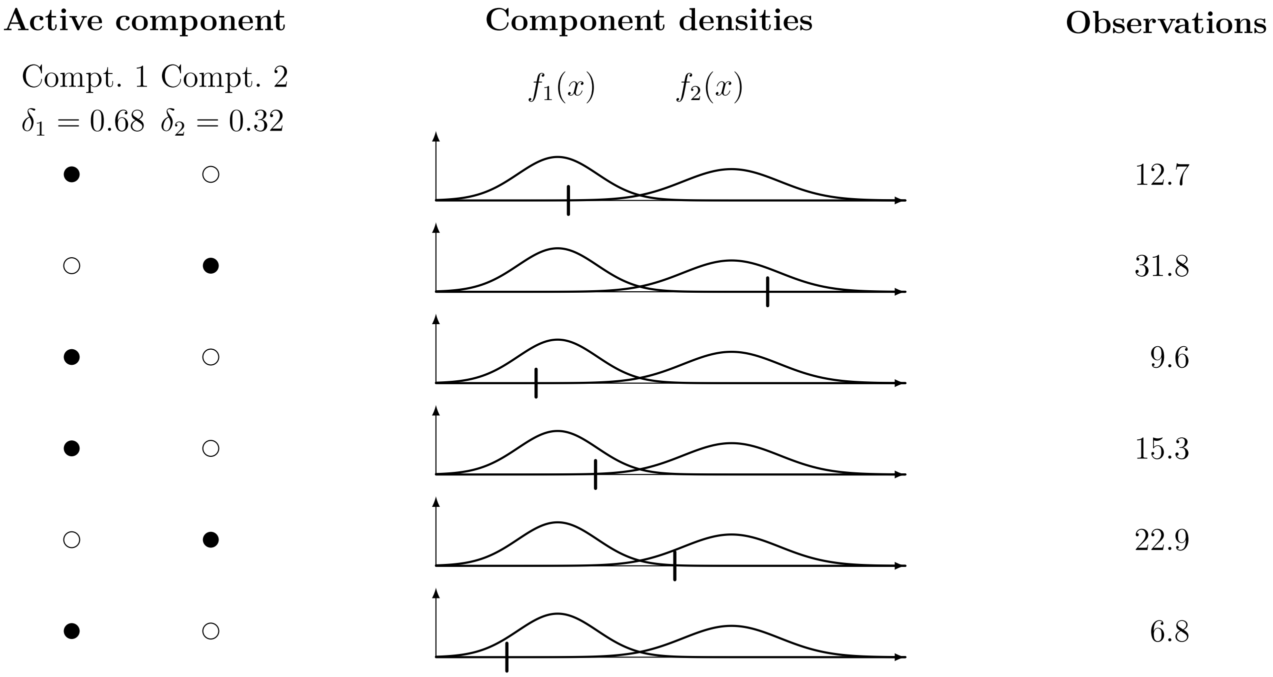


Figure 6: *A independent mixture model with 2 components, $\delta_1 = 0.68$ with corresponding density $f_1(x)$ and $\delta_2 = 0.32$ with corresponding density $f_2(x)$. The black veritcal line on the first axis represents the observation. A filled dot under a component denotes if it is active.*

Parameter Estimation Let $\boldsymbol{\zeta_1}, \ldots, \boldsymbol{\zeta_N}$ be the parameter vectors of the N-component distributions, $\delta_1, \ldots, \delta_N$ the mixing parameters where $\delta_k \geq 0$, $\sum_{k=1}^{N} \delta_k = 1$ and $x_1, \ldots, x_T$ be the T observations. The likelihood of a mixture model with N components is then given by

$$\mathcal{L}_T(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_N, \delta_1, \ldots, \delta_N) = \prod_{j=1}^{T} \sum_{i=1}^{N} \delta_i f_{X_j,i}(x_j). \tag{12}$$

Thus, if the components are specified only by a single parameter, $2m-1$ independent parameters have to be estimated by the component sum constraint.

Unbounded Likelihood in Mixtures The beforementioned theory relating to the independent mixture distribution could easily be expanded to a discrete case by use probability masses instead of densities. However, one key difference arises; it can happen that in the vicinity of certain parameter subsets, the likelihood is unbounded. In a Gaussian mixture, the likelihood can be made arbitrarily large by assigning a component mean to coincide with an observed data point and letting that component's variance approach zero. In the case of a unbounded likelihood it is often argued in the literature that the MLEs do not exist [43, p. 4630].

In this case, one can somewhat circumvent the complication via a discrete likelihood approximation of Equation 12

$$\mathcal{L}_T^{\text{discrete}}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_N, \delta_1, \ldots, \delta_N) = \prod_{j=1}^{T} \sum_{i=1}^{N} \delta_i \int_{a_j}^{b_j} f_{X_j,i}(x_j),$$

where the interval (a_j, b_j) consists of those values which would be recorded as x_j, if observed. For a set of r.v.'s $X_1, X_2, \ldots, X_T$, the discrete likelihood is of the form $\mathbb{P}\left(a_t < X_t < b_t\right)$ for all t. An alternative remedy is to enforce a positive lower bound on the component variances and then seek the best local maximum under this constraint.

2.2.2 Markov Chains & Hidden Markov Models

Let $\{C_t\}_{t\in\mathbb{N}}$ be a sequence of discrete r.v.'s. $\{C_t\}_{t\in\mathbb{N}}$ is said to be a discrete-time Markov chain if, for all $t \in \mathbb{N}$, it satisfies the Markov property

$$\mathbb{P}(C_{t+1} \mid C_t, C_{t-1}, \ldots, C_1) = \mathbb{P}(C_{t+1} \mid C_t).$$

In other words, condition on the history up to and including time-t only depends on time-t. We will when convinent use the notation $\mathbf{C}^{(t)} = (C_1, C_2, \ldots, C_t)$ such that

$$\mathbb{P}(C_t = j \mid C_{t-1} = i, \ldots, C_1 = k) = \mathbb{P}(C_t = j \mid C_{t-1} = i),$$

where $C_t \in \mathcal{C}$ is the state at time $t = 1, 2, 3, \ldots, T$ and $\mathcal{C}$ is the state space. The Markov property is a first relaxation of the assumption of independence and can be seen visualized in Figure 7.

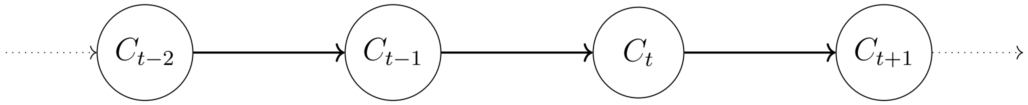


Figure 7: *A (first-order) Markov chain for the sequence of discrete r.v.'s* $\{C_t\}_{t\in\mathbb{N}}$.

A N-dimensional (or N-state) hidden Markov model (HMM) $\{X_t\}_{t\in\mathbb{N}}$ is a dependent mixture model that assumes that the distribution of the observed response variable X_t depends exclusively on a hidden state $C_t \in \mathcal{C}$, where $\mathcal{C} = \{C_t : t = 1, 2, \ldots, T\}$ is modelled by a discrete time N-state Markov chain, meaning, C_t satisfies the Markov .. Summerized, the model is

$$\mathbb{P}\left(C_t \mid \mathbf{C}^{(t-1)}\right) = \mathbb{P}\left(C_t \mid C_{t-1}\right), \quad t = 2, 3, \ldots,$$
$$\mathbb{P}\left(X_t \mid \mathbf{X}^{(t-1)}, \mathbf{C}^{(t)}\right) = \mathbb{P}\left(X_t \mid C_t\right), \quad t \in \mathbb{N}.$$

The model thus consists of a unobserved/hidden parameter process $\{C_t\}_{t\in\mathbb{N}}$ satisfying the Markov property and a state-dependent process $\{X_t\}_{t\in\mathbb{N}}$ in which the distribution of X_t depends exclusively on the time-t state, C_t.

Specifically for this thesis, the observed response variable is a state-dependent process, $\{X_t\}_{t\in\mathbb{N}}$, the log returns. The process is a noisy observation process in the sense that it is assumed to be produced by a underlaying unobserved hidden state process, $\{C_t\}_{t\in\mathbb{N}}$. The distribution of X_t, which is Gaussian, is conditionally independent of previous observations and all states except the current hidden state $i \in \mathcal{C}$:

$$f_{X_t|\mathbf{X}^{(t-1)},\mathbf{C}^{(t)}}(x_t \mid \mathbf{x}^{(t-1)}, \mathbf{c}^{(t)}) = f_{X|C_t}(x_t \mid i), \quad t = 1, 2, 3, \ldots, T,$$

cwhere f denotes a probability density function. Note that we do not say that $S_t \mid S_s$, $t > s$ are unconditionally independent. The structure of a regular hidden Markov model can be seen in Figure 8, where the conditional independence can be intuitively understood.

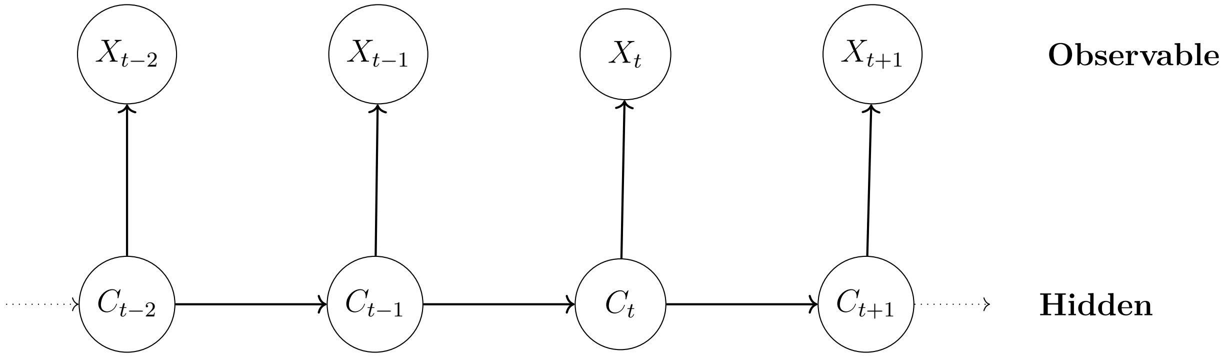


Figure 8: *A hidden Markov Model of order 1.*

The Markov chain induces dependence in the state-dependent process, meaning, the observations are independent of each other within states.

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

$$\mathbf{\Gamma} = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1N} \\ \vdots & \ddots & \vdots \\ \gamma_{N1} & \cdots & \gamma_{NN} \end{bmatrix}, \quad \gamma_{ij} = \mathbb{P}(C_{t+1} = j \mid C_t = i) \in [0,1], \quad \sum_{j \in \mathcal{C}} \gamma_{ij} = 1, \tag{13}$$

for all $i, j \in \mathcal{C}$, 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. We define two key concepts for a Markov chain.

Definition 2.1. *A Markov chain $\{C_t\}_{t=0,1,\dots}$ with state space $\mathcal{C}$ is said to be irreducible if*

$$\forall i, j \in \mathcal{C}, \exists t < \infty : \mathbb{P}\left(C_{n+t} = j \mid C_n = i\right) > 0$$

Equivalently, every state can be reached from every other state with positive probability in some finite number of steps. In this case, we say that all states communicate.

The concept of aperiodicity can be seen intuitively in Figure 9.

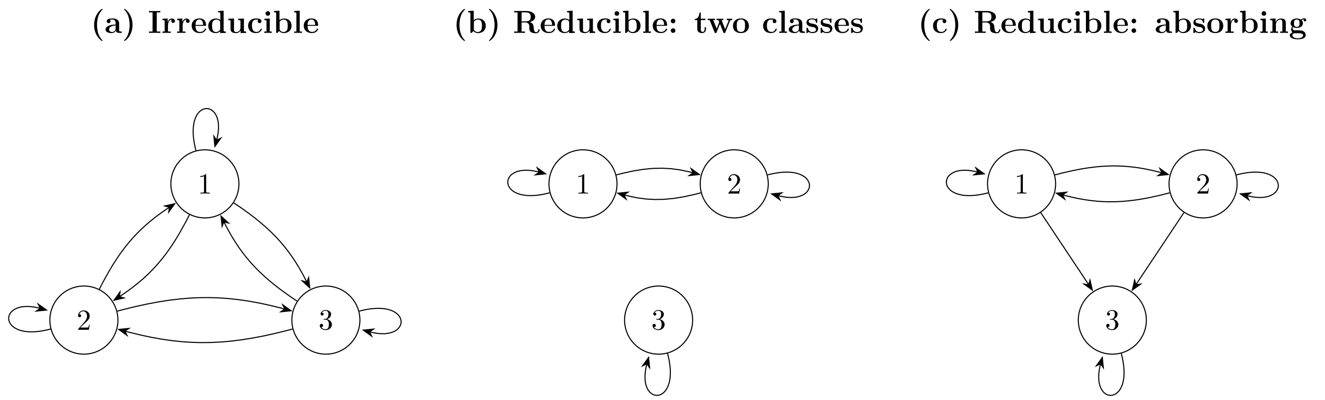

(a) Irreducible **(b) Reducible: two classes** **(c) Reducible: absorbing**

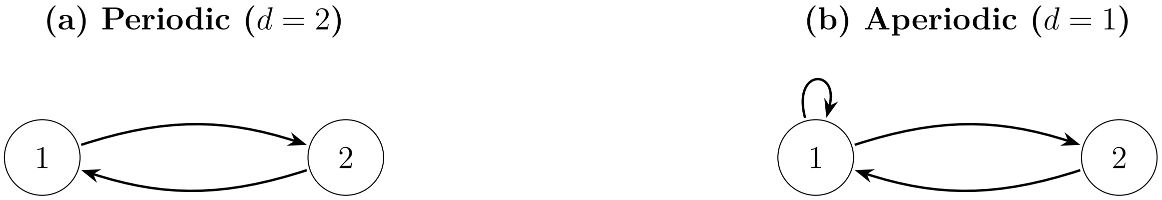
Figure 9: *Three finite Markov chains: (a) irreducible; (b) reducible with two communicating classes; (c) reducible with an absorbing class.*

Definition 2.2. *For a state $i \in \mathcal{C}$, define its period as*

$$d(i) := \gcd\{\, t \geq 1 : (\mathbf{\Gamma}^t)_{ii} > 0 \,\}.$$

A state i is said to be aperiodic if $d(i) = 1$. A Markov chain is called aperiodic if all its states are aperiodic.

The concept of aperiodic is visualized in Figure 10.


(a) Periodic $(d = 2)$

(b) Aperiodic $(d = 1)$

Figure 10: *Periodic vs aperiodic Markov chains. Adding a self-loop breaks periodicity.*

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

$$\boldsymbol{\delta}^{(t)} = \underbrace{\begin{bmatrix}\mathbb{P}(C_t = 1) & \cdots & \mathbb{P}(C_t = N)\end{bmatrix}}_{1\times N}, \tag{14}$$

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 2.1. *Let $\boldsymbol{\delta}^{(t)}$ be defined as in Equation 14. All future distributions of the Markov chain can then be found by*

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

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

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

Lastly, equality two and three follows from:

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

□

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

$$\boldsymbol{\delta}\boldsymbol{\Gamma} = \boldsymbol{\delta}, \quad \boldsymbol{\delta}\mathbf{1}^\top = 1, \tag{15}$$

where $\mathbf{1}_N$ is a N-dimensional vector with entries 1. The first of the requirements in Equation 15 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. To see why this holds, 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

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

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 [58, 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 [58, p. 18]

$$\left(\boldsymbol{I}_N - \boldsymbol{\Gamma} + \mathbf{1}_{N\times N}\right)^\top \boldsymbol{\delta}^\top = \mathbf{1}_N^\top, \tag{16}$$

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.

Proposition 2.2. *Let $\boldsymbol{\Gamma} \in \mathbb{R}^{N\times N}$ be row-stochastic, i.e. $\boldsymbol{\Gamma}\mathbf{1}_N = \mathbf{1}_N$, and define $J = \mathbf{1}_N\mathbf{1}_N^\top$. A row vector $\boldsymbol{\delta} \in \mathbb{R}^{1\times N}$ satisfies*

$$\boldsymbol{\delta\Gamma} = \boldsymbol{\delta}, \qquad \boldsymbol{\delta}\mathbf{1}_N = 1$$

if and only if

$$(\boldsymbol{I}_N - \boldsymbol{\Gamma} + J)^\top \boldsymbol{\delta}^\top = \mathbf{1}_N.$$

If $\boldsymbol{\Gamma}$ is irreducible, then $\boldsymbol{A} := (\boldsymbol{I}_N - \boldsymbol{\Gamma} + \boldsymbol{J})^\top$ is nonsingular and

$$\boldsymbol{\delta}^\top = \boldsymbol{A}^{-1}\mathbf{1}_N.$$

Proof. **(⇒)** If $\delta\boldsymbol{\Gamma} = \delta$, then

$$(\boldsymbol{I}_N - \boldsymbol{\Gamma}^\top)\delta^\top = 0.$$

Since $\delta\mathbf{1}_N = 1$, we have

$$\boldsymbol{J\delta}^\top = (\boldsymbol{\delta}\mathbf{1}_N)\mathbf{1}_N = \mathbf{1}_N.$$

Adding gives

$$(\boldsymbol{I}_N - \boldsymbol{\Gamma}^\top + \boldsymbol{j})\boldsymbol{\delta}^\top = (\boldsymbol{I}_N - \boldsymbol{\Gamma}^\top)\boldsymbol{\delta}^\top + \boldsymbol{J\delta}^\top = \mathbf{1}_N,$$

equivalently $(\boldsymbol{I}_N - \boldsymbol{\Gamma} + J)^\top\delta^\top = \mathbf{1}_N$.

(⇐) Assume $(\boldsymbol{I}_N - \boldsymbol{\Gamma}^\top + \boldsymbol{J})\boldsymbol{\delta}.^\top = \mathbf{1}_N$. Left-multiplying by $\mathbf{1}_N^\top$ and using $\mathbf{1}_N^\top\boldsymbol{\Gamma}^\top = \mathbf{1}_N^\top$ and $\mathbf{1}_N^\top\boldsymbol{J} = N\mathbf{1}_N^\top$ yields

$$N(\boldsymbol{\delta}\mathbf{1}_N) = \mathbf{1}_N^\top\mathbf{1}_N = N,$$

so $\boldsymbol{\delta}\mathbf{1}_N = 1$. Substituting this back gives

$$(\boldsymbol{I}_N - \boldsymbol{\Gamma}^\top)\boldsymbol{\delta}^\top = \mathbf{1}_N - \boldsymbol{J\delta}^\top = \mathbf{1}_N - \mathbf{1}_N = 0,$$

so $\boldsymbol{\Gamma}^\top\delta^\top = \delta^\top$, i.e. $\delta\boldsymbol{\Gamma} = \delta$. Suppose $(\boldsymbol{I}_N - \boldsymbol{\Gamma}^\top + J)x = 0$. Left-multiplying by $\mathbf{1}_N^\top$ gives

$$\mathbf{1}_N^\top(\boldsymbol{I}_N - \boldsymbol{\Gamma}^\top + \boldsymbol{J})x = N\,\mathbf{1}_N^\top x = 0,$$

so $\mathbf{1}_N^\top x = 0$. The equation then reduces to

$$(\boldsymbol{I}_N - \boldsymbol{\Gamma}^\top)x = 0 \quad \Longrightarrow \quad \boldsymbol{\Gamma}^\top x = x.$$

For irreducible $\mathbf{\Gamma}$, the eigenspace of eigenvalue 1 is one-dimensional and spanned by the strictly positive stationary vector. Any nonzero such eigenvector has strictly positive sum, contradicting $\mathbf{1}_N^\top x = 0$. Thus $x = 0$, so $\boldsymbol{I}_N - \mathbf{\Gamma}^\top + \boldsymbol{J}$ is invertible. Therefore

$$\boldsymbol{\delta}^\top = (\boldsymbol{I}_N - \mathbf{\Gamma}^\top + \boldsymbol{J})^{-1}\mathbf{1}_N.$$

□

When the transition probabilities are time-varying (i.e. functions of covariates), the stationary distribution does not exist [35, 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. Furthermore, computational cost is currently extremely daunting which is alleviated by assuming stationarity as it will be evident in Proposition 2.3. We use Equation 16 throughout the code after fitting to find the stationary distribution to ease computational drag.

2.2.3 State-Dependent Distributions

The state-dependent distributions are the probability density functions of X_t given some state $i \in \mathcal{C}$ at time-t given by[2]

$$f_{i,X_i}(x_t) := f_{X_t|C_t}(x_t \mid i).$$

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

$$f_{X_t}(x_t) \overset{\dagger}{=} \sum_{i\in\mathcal{C}} f_{X_t,i}(x_t, i) = \sum_{i\in\mathcal{C}} f_{X_t|C_t}(x_t \mid i) f(C_t = i) = \sum_{i\in\mathcal{C}} \delta_i^{(t)} f_{i,X_t}(x_t) \overset{\dagger\dagger}{=} \sum_{i\in\mathcal{C}} \delta_i f_{i,X_t}(x_t), \tag{17}$$

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

As the log-returns follow a normal distribution, we can directly specify the densities in Equation 17. Namely

$$f_{i,X_t}(x_t) = \frac{1}{\sqrt{2\pi\,\sigma_i^2\Delta}} \exp\left(-\frac{\left(x_t - (\mu_i - \frac{1}{2}\sigma_i^2)\Delta\right)^2}{2\,\sigma_i^2\Delta}\right), \quad x_t \in \mathbb{R}$$

Parameter Count As all the parameters are now defined for the BS-HMM we proceed to count the number of parameters to be estimated. The state process is characterized by $\boldsymbol{\delta}$ and $\mathbf{\Gamma}$. The

[2] The notation of the state-dependent density functions should not be confused with a joint density function. We will explicitly write i as a lower and first index when state-dependent densities are intended and not joint density functions.

latter has $N \times (N-1) = N^2 - N$ free parameters due to the row-sum constraint (last equality of Equation 13). For a stationary Markov chain, we need not estimate the initial distribution as this equals the stationary distribution, which would otherwise yield N additional parameters (see Equation 14). As previously stated, we simply use Equation 16 to obtain the stationary distribution after estimating the transition probabilities.

Under conditional independence, the state-dependent process is governed by the state-dependent distributions. In the Black–Scholes setting these are parameterized by (μ, σ), so if both are state-dependent we require $2N$ parameters. The $N \times N$ transition probawbilities. However, as the row-constraint states that the sum of transition probabilities in row i has to equal 1 this leaves us with $N \times (N-1)$ parameters for the T.P.M. The stationary distribution is estimated from eq. ?? which then requires no extra parameters to be estiamted. In total we therefore estimate

$$\#\text{Parameters}_2 = N^2 - N + 2N \;=\; N^2 + N.$$

If, instead, exactly one of μ or σ is state-dependent (the other being globally state-independent), we replace the $2N$ by $N+1$, yielding

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

Lastly, if *neither* μ nor σ is state-dependent (both globally state-independent), we only need to estimate one count of both μ and σ, yielding

$$\#\text{Parameters}_0 = N^2 - N + 2.$$

For a visualization of the relation between state-dependent parameters and number of states see Figure 11.

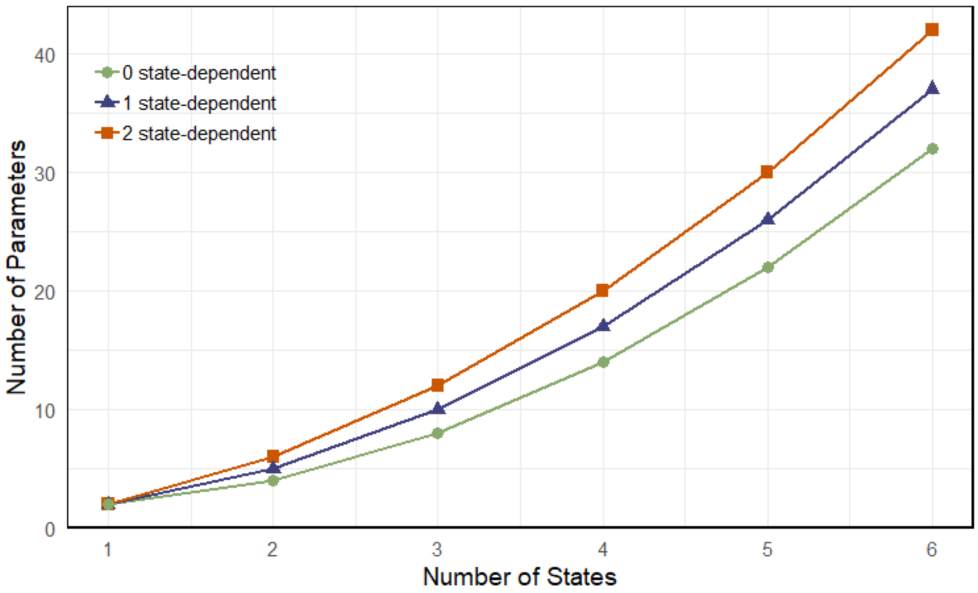


Figure 11: *Number of parameters as a function of the number of states N when 0, 1, or 2 of the BS parameters (μ, σ) are modeled as state-dependent.*

2.2.4 Likelihood Formulation & Parameter 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 $\boldsymbol{\zeta} = (\boldsymbol{\Gamma},\ \boldsymbol{\mu},\ \boldsymbol{\sigma})^\top$.

Proposition 2.3. *Let $\{C_t\}_{t=1}^T$ be a homogeneous, finite N-state Markov chain on $\mathcal{C}$, with transition probability $\boldsymbol{\Gamma} = (\gamma_{ij})_{i,j=1}^N$. Consider a observation process $\{X_t\}_{t=1}^T$. The joint likelihood of observing $\{X_t\}_{t=1}^T$ is then given by*

$$\mathcal{L}_T(\boldsymbol{\zeta}) = \boldsymbol{\delta}^{(1)} \boldsymbol{P}(x_1) \boldsymbol{\Gamma} \boldsymbol{P}(x_2) \boldsymbol{\Gamma} \boldsymbol{P}(x_3) \cdots \boldsymbol{\Gamma} \boldsymbol{P}(x_t) \mathbf{1}^\top,$$

where $\boldsymbol{\delta}^{(1)}$ is the initial distribution, $\boldsymbol{P}(x)$ is the diagonal matrix with the state-dependent distribution $f_{1,X}(x),\ f_{2,X}(x),\ \ldots,\ f_{N,X}(x)$ given in Equation 17 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

$$\mathcal{L}_T(\boldsymbol{\zeta}) = \boldsymbol{\delta} \boldsymbol{\Gamma} \boldsymbol{P}(x_1) \boldsymbol{\Gamma} \boldsymbol{P}(x_2) \boldsymbol{\Gamma} \boldsymbol{P}(x_3) \cdots \boldsymbol{\Gamma} \boldsymbol{P}(x_T) \mathbf{1}_N^\top.$$

Proof. Note that

$$\begin{aligned} \mathcal{L}_T(\boldsymbol{\zeta}) &= f_{\mathbf{X}^{(T)}}(\mathbf{x}^{(T)}) \\ &= \sum_{c_1,\ldots,c_T=1}^{N} f_{\mathbf{X}^{(T)} \mid \mathbf{C}^{(T)}}\left(\mathbf{x}^{(T)} \mid \mathbf{c}^{(T)}\right) \mathbb{P}\Big(\mathbf{C}^{(T)} = \mathbf{c}^{(T)}\Big) \end{aligned}$$

and by Lemma A.2.1

$$\begin{aligned}\mathcal{L}_T(\boldsymbol{\zeta}) &= f_{\mathbf{X}^{(T)}}(\mathbf{x}^{(T)}) \\ &= \mathbb{P}(C_1)\prod_{k=2}^{T}\mathbb{P}(C_t \mid C_{t-1})\prod_{k=1}^{T} f_{X_k|C_k}(x_k \,|\, C_k = c_k)\,.\end{aligned}$$

It then follows that

$$\begin{aligned}\mathcal{L}_T(\boldsymbol{\zeta}) &= \sum_{c_1,c_2,\ldots,c_T=1}^{N}(\delta_{c_1}\gamma_{c_1,c_2}\gamma_{c_2,c_3}\ldots\gamma_{c_{T-1},c_T})(f_{c_1,X_1}(x_1)f_{c_2,X_2}(x_2)\ldots f_{c_T,X_T}(x_T)) \\ &= \sum_{c_1,c_2,\ldots,c_T=1}^{N}\delta_{c_1}f_{c_1,X_1}(x_1)\gamma_{c_1,c_2}f_{c_2,X_2}(x_2)\gamma_{c_2,c_3}\ldots\gamma_{c_{T-1},c_T}f_{c_T,X_T}(x_T) \\ &= \boldsymbol{\delta}\boldsymbol{P}(x_1)\boldsymbol{\Gamma}\boldsymbol{P}(x_2)\ldots\boldsymbol{\Gamma}\boldsymbol{P}(x_T)\mathbf{1}_N^{\top}.\end{aligned}$$

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 [58, Ex. 7(b)] or Lemma A.2.5). If $\boldsymbol{\delta}$ is the stationary distribution of the Markov chain, we simply have

$$\boldsymbol{\delta}\boldsymbol{P}(x_1) = \boldsymbol{\delta}\boldsymbol{\Gamma}\boldsymbol{P}(x_1).$$

□

The recursive nature of the likelihood in Proposition 2.3 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).

Forward Probabilities The forward algorithm utilizes the forward probabilities which for $t = 1, 2, \ldots, T$ and $j \in \mathcal{C}$ are given as

$$\alpha_t(j) = f_{\mathbf{X}^{(t)},C_t}(\mathbf{x}^{(t)}, j) = \mathbb{P}(C_t = j)\, f_{\mathbf{X}^{(t)}|C_t=j}\big(\mathbf{x}^{(t)}\big), \quad \boldsymbol{\alpha}_t = \big[\alpha_t(1)\ \ldots\ \alpha_t(N)\big]. \tag{18}$$

In other words, the forward probabilities contain information on the likelihood of the observations up to and including time-t. Also note that from the defintition of $\boldsymbol{\alpha}_t$ that, for $t = 1, 2, \ldots, T-1$, $\boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t\boldsymbol{\Gamma}\boldsymbol{P}(x_{t+1})$ which can be written in scalar form as

$$\alpha_{t+1}(j) = \left(\sum_{i\in\mathcal{C}}\alpha_t(i)\gamma_{ij}\right) f_{j,X_{t+1}}(x_{t+1}). \tag{19}$$

We are now able to prove the following result to justify their description as probabilities by utilizing the recursive form in Equation 19 and Lemma A.2.1.

Proposition 2.4. *For $t = 1, 2, \ldots, T$ and $j \in \mathcal{C}$,*

$$\alpha_t(j) = f_{\mathbf{X}^{(t)}, C_t}\big(\mathbf{x}^{(t)}, j\big).$$

Proof. Firstly, since $\boldsymbol{\alpha}_1 = \boldsymbol{\delta}\, \mathbf{P}(x_1)$ it follows that for $t = 1$

$$\alpha_1(j) = \delta_j f_{j, X_1}(x_1) = \mathbb{P}(C_1 = j)\, f_{X_1 | C_1}(x_1 \mid j) = f_{X_1, C_1}(x_1, j).$$

For some $t \in \mathbb{N}$ we then show it holds for $t + 1$:

$$\begin{aligned}
\alpha_{t+1}(j) &\overset{\dagger}{=} \sum_{i \in \mathcal{C}} \alpha_t(i)\, \gamma_{ij} f_{j, X_{t+1}}(x_{t+1}) \\
&= \sum_{i \in \mathcal{C}} f_{\mathbf{X}^{(t)}, C_t}\big(\mathbf{x}^{(t)}, i\big)\ \mathbb{P}(C_{t+1} = j \mid C_t = i)\ f_{X_{t+1} | C_{t+1}}(x_{t+1} \mid j) \\
&= \sum_{i \in \mathcal{C}} f_{\mathbf{X}^{(t)}, C_t, C_{t+1}}\big(\mathbf{x}^{(t)}, i, j\big)\ f_{X_{t+1} | C_{t+1}}(x_{t+1} \mid j) \\
&= \sum_{i \in \mathcal{C}} f_{\mathbf{X}^{(t+1)}, C_t, C_{t+1}}\big(\mathbf{x}^{(t+1)}, i, j\big) \\
&\overset{\dagger\dagger}{=} f_{\mathbf{X}^{(t+1)}, C_{t+1}}\big(\mathbf{x}^{(t+1)}, j\big),
\end{aligned}$$

where † is the scalar forward recursion (matrix–vector form $\boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t\, \boldsymbol{\Gamma}\, \mathbf{P}(x_{t+1})$), and we used the HMM conditional independences $X_{t+1} \perp (\mathbf{X}^{(t)}, C_t) \mid C_{t+1}$ and $\mathbf{X}^{(t)} \perp C_{t+1} \mid C_t$. Finally, †† is marginalization over the discrete r.v. C_t: summing over i yields $f_{\mathbf{X}^{(t+1)}, C_{t+1}}(\mathbf{x}^{(t+1)}, j)$. □

Consequently, Equation 18 allows us to write the likelihood from Proposition 2.3 as

$$\mathcal{L}_t(\boldsymbol{\zeta}) = f_{\mathbf{X}^{(t)}}\left(\mathbf{x}^{(t)}\right) \overset{\dagger}{=} \sum_{j \in \mathcal{C}} f_{\mathbf{X}^{(t)}, C_t}\left(\mathbf{x}^{(t)}, j\right) = \sum_{j \in \mathcal{C}} \alpha_t(j).$$

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

$$f_{C_t | \mathbf{X}^{(t)}}(j \mid \mathbf{x}^{(t)}) = \frac{f_{C_t, \mathbf{X}^{(t)}}(j, \mathbf{x}^{(t)})}{f_{\mathbf{X}^{(t)}}(\mathbf{x}^{(t)})} = \frac{\alpha_t(j)}{\sum_{i \in \mathcal{C}} \alpha_t(i)}.$$

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

$$\boldsymbol{\alpha}_t = \boldsymbol{\delta} \boldsymbol{P}(x_1) \boldsymbol{\Gamma} \boldsymbol{P}(x_2) \cdots \boldsymbol{\Gamma} \boldsymbol{P}(x_t) = \boldsymbol{\delta} \boldsymbol{P}(x_1) \prod_{s=2}^{t} \boldsymbol{\Gamma} \boldsymbol{P}(x_s),$$

following the convention that an empty product is the identity matrix [58, p. 38]. Assembling, Proposition 2.3 states that $\mathcal{L}_T(\boldsymbol{\zeta}) = \boldsymbol{\alpha}_T \mathbf{1}_N^\top$ and for $t \geq 2$ we defined $\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1} \boldsymbol{\Gamma} \boldsymbol{P}(x_t)$. This allows us to define the *forward algorithm*:

$$\begin{aligned}
\boldsymbol{\alpha}_1 &= \boldsymbol{\delta}\boldsymbol{P}(x_1);\\
\boldsymbol{\alpha}_t &= \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\boldsymbol{P}(x_t), \quad \text{for } t = 2, 3, \ldots, T;\\
\mathcal{L}_T &= \boldsymbol{\alpha}_T\mathbf{1}_N^\top.
\end{aligned}$$

Note, for a N-state HMM, $\boldsymbol{\delta}$ has N elements, $\boldsymbol{P}(x)$ has N elements (all in the diagonal) and $\boldsymbol{\Gamma}$ has $N \times N$ elements. For the forward 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 $f_{i,X_t}(x_t)$, $i \in \mathcal{C}$. 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.

Backwards Probabilities Define

$$\boldsymbol{\beta}_t^\top = \boldsymbol{\Gamma}\boldsymbol{P}(x_{t+1})\boldsymbol{\Gamma}\boldsymbol{P}(x_{t+2})\cdots\boldsymbol{\Gamma}\boldsymbol{P}(x_T)\mathbf{1}_N^\top = \left(\prod_{s=t+1}^{T}\boldsymbol{\Gamma}\boldsymbol{P}(x_s)\right)\mathbf{1}_N^\top,$$

with the convention that an empty product is the identity matrix, The cast $t = T$ yields $\boldsymbol{\beta}_T = \mathbf{1}_N$. We thne show that $\beta_t(j)$, the jth component of $\boldsymbol{\beta}_t$, can be identified as the the conditional probability

$$f_{\mathbf{X}_{t+1}^T|C_t}\left(\mathbf{x}_{t+1}^T \mid i\right)$$

It then follows that for $t = 1, 2, \ldots, T$,

$$\alpha_t(j)\beta_t(j) = f_{\mathbf{X}^{(t)},C_t}(\mathbf{x}^{(t)}, j)$$

From the definition of $\boldsymbol{\beta}_t$ it follows that $\boldsymbol{\beta}_t^\top = \boldsymbol{\Gamma}\boldsymbol{P}(x_{t+1})\boldsymbol{\beta}_{t+1}^\top$ and hence the name *backwards* probabilities. We now identify the backward probabilities as probabilities by the proposition below.

Proposition 2.5. *Assume* $\mathbb{P}(C_t = i) > 0$. *For* $t = 1, 2, \ldots, T-1$ *and* $i \in \mathcal{C}$,

$$\beta_t(i) = f_{\mathbf{X}_{t+1}^T,C_t}\left(\mathbf{x}_{t+1}^T \mid i\right)$$

Proof. We proof the proposition by induction. For $t = T-1$

$$\beta_{T-1}(i) = \sum_j \mathbb{P}(C_T = j \mid C_{T-1} = i) f_{X_T,C_T}\left(x_T \mid j\right), \quad (\dagger)$$

since $\boldsymbol{\beta}_{T-1}^{\top} = \boldsymbol{\Gamma}\boldsymbol{P}(x_T)\mathbf{1}_N^{\top}$. Furtermore, by Lemma A.2.3

$$\begin{aligned}\mathbb{P}\left(C_T = j \mid C_{T-1=i}\right) f_{X_T|C_T}\left(x_T \mid j\right) &= \mathbb{P}(C_T = j \mid C_{T-1} = i) f_{X_t|C_{T-1},C_T}(x_T \mid i, j) \\ &= f_{X_T,C_{T-1},C_t}(x_T, i, j)/\mathbb{P}(C_{T-1} = i). \quad (\dagger\dagger)\end{aligned}$$

Substituting (††) into (†) gives

$$\begin{aligned}\beta_{T-1}(i) &= \frac{1}{\mathbb{P}(C_{T-1} = i)} \sum_j f_{X_T,C_{T-1},C_t}(x_T, i, j) \\ &= f_{X_T,C_{T-1}}(x_t, i)/\mathbb{P}(C_{T-1} = i) \\ &= f_{X_T|C_{T-1}}(x_t \mid i)\end{aligned}$$

as required.

To demonstrate that validity at time $t+1$ implies validity at time t, we begin by observing that the recursive definition of β_t, in combination with the inductive hypothesis, gives

$$\beta_t(i) = \sum_j \gamma_{ij} f_{X_{t+1}|C_{t+1}}(x_{t+1}, j) f_{\mathbf{X}_{t+2}^{T}|C_{t+1}}(\mathbf{x}_{t+2}^{T} \mid j). \quad (\dagger\,\dagger\,\dagger)$$

However, Lemma A.2.2 and Lemma A.2.3 imply that

$$f_{X_{t+1}|C_{t+1}}(x_{t+1}, j) f_{\mathbf{X}_{t+2}^{T}}(\mathbf{x}_{t+2}^{T} \mid j) = f_{\mathbf{X}_{t+1}^{T}|C_t,C_{t+1}}(\mathbf{x}_{t+1}^{T} \mid i, j). \quad (\dagger\,\dagger\,\dagger\,\dagger)$$

Substitute from († † † †) into († † †) which yields

$$\begin{aligned}\beta_t(i) &= \sum_j \mathbb{P}(C_{t+1} = j \mid C_t = i) f_{\mathbf{X}_{t+1}^{T}|C_t,C_{t+1}}(\mathbf{x}_{t+1}^{T} \mid i, j) \\ &= \frac{1}{\mathbb{P}(C_t = 1)} \sum_j f_{\mathbf{X}_{t+1}^{T},C_t,C_{t+1}}(\mathbf{x}_{t+1}^{T}, i, j) \\ &= \frac{f_{\mathbf{X}_{t+1}^{T},C_t}\left(\mathbf{x}_{t+1}^{T}, i\right)}{\mathbb{P}(C_t = i)} \\ &= f_{\mathbf{X}_{t+1}^{T}|C_t}\left(\mathbf{x}_{t+1}^{T} \mid i\right)\end{aligned}$$

which is the required conditional probability. □

Scaling the Likelihood Let $\mathcal{L}_t(\boldsymbol{\zeta})$ 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 [27])

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

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 2.3, 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 [58, 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:

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

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

$$\boldsymbol{\phi}_1 = \frac{\boldsymbol{\alpha}_1}{\boldsymbol{\alpha}_1 \mathbf{1}_N^\top} = \frac{\boldsymbol{\delta}_0 \boldsymbol{P}(x_1)}{\boldsymbol{\delta}_0 \boldsymbol{P}(x_1) \mathbf{1}_N^\top}.$$

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

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

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

$$\mathcal{L}_T(\boldsymbol{\zeta}) = \boldsymbol{\alpha}_T \mathbf{1}^\top = \frac{\boldsymbol{\alpha}_1 \mathbf{1}^\top}{\boldsymbol{\alpha}_0 \mathbf{1}^\top} \frac{\boldsymbol{\alpha}_2 \mathbf{1}^\top}{\boldsymbol{\alpha}_1 \mathbf{1}^\top} \cdots \frac{\boldsymbol{\alpha}_T \mathbf{1}^\top}{\boldsymbol{\alpha}_{T-1} \mathbf{1}^\top} = \prod_{t=1}^{T} \frac{\boldsymbol{\alpha}_t \mathbf{1}^\top}{\boldsymbol{\alpha}_{t-1} \mathbf{1}^\top}, \tag{20}$$

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 20

$$\begin{aligned}
\ell_T(\boldsymbol{\zeta}) &= \log \mathcal{L}_T(\boldsymbol{\zeta}) \\
&= \log \prod_{t=1}^{T} \frac{\boldsymbol{\alpha}_t \mathbf{1}_N^\top}{\boldsymbol{\alpha}_{t-1}\mathbf{1}_N^\top} \\
&= \sum_{t=1}^{T} \log \left(\frac{\boldsymbol{\alpha}_t \mathbf{1}_N^\top}{\boldsymbol{\alpha}_{t-1}\mathbf{1}_N^\top} \right) \\
&= \log \left(\frac{\boldsymbol{\alpha}_1 \mathbf{1}_N^\top}{\boldsymbol{\alpha}_0 \mathbf{1}_N^\top} \right) + \sum_{t=2}^{T} \log \left(\frac{\boldsymbol{\alpha}_t \mathbf{1}_N^\top}{\boldsymbol{\alpha}_{t-1}\mathbf{1}_N^\top} \right) \\
&= \log \left(\frac{\boldsymbol{\delta} \boldsymbol{P}(x_1) \mathbf{1}_N^\top}{\boldsymbol{\delta} \mathbf{1}_N^\top} \right) + \sum_{t=2}^{T} \log \left(\frac{\boldsymbol{\alpha}_{t-1}}{\boldsymbol{\alpha}_{t-1}\mathbf{1}_N^\top} \boldsymbol{\Gamma} \boldsymbol{P}(x_t) \mathbf{1}_N^\top \right) \\
&= \log \left(\boldsymbol{\delta} \boldsymbol{P}(x_1) \mathbf{1}_N^\top \right) + \sum_{t=2}^{T} \log \left(\boldsymbol{\phi}_{t-1} \boldsymbol{\Gamma} \boldsymbol{P}(x_t) \mathbf{1}_N^\top \right),
\end{aligned}$$

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* $\eta_i = \log \lambda_i$ for some parameter λ_i. After we have maximized the likelihood with respect to the unconstrained parameters, the constrained parameter estimates can be obtained by transforming back: $\hat{\lambda}_i = e^{\hat{\eta}_i}$. 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

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

and

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

and $\boldsymbol{\Gamma} = (\gamma_{ij})_{i,j=1}^{N}$. 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 $\boldsymbol{\eta} = (\eta_1, \dots, \eta_N)^\top$ which are all unconstrained by construction.

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

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

Consider $\boldsymbol{\Gamma}$ 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 $\boldsymbol{\Gamma}$ follow from the row sums of 1. The transformation in the opposite direction is

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

2.2.5 Standard Errors & Confidence Intervals

Unfortunately, relatively little is known about the properties of the MLEs of HMMs [58, p. 56]. However, asymptotic results are avaliable but requires the estimation of the variance-covariance matrix of the estimators of the parameters. It is possible to estimate the standard errors from the Hession of the log-likelihood evaluated at the maximum, however, this causes issues when parameters are on the boundary of their parameter space. This phenomenon does unfortunately happen quite often [58, p. 56]. Another method is that of parametric bootstrapping but such methods are heavily relient on computational power. As is becoming quite evident, the thesis is already extremely relient on computational power and as such we restrict ourself to the former method.

Standard Errors via the Hessian Point eestimates of $\hat{\boldsymbol{\Theta}} = (\hat{\boldsymbol{\Gamma}}, \hat{\boldsymbol{\lambda}})$ are not difficult to compute. However, exact intervals are not avaliable. However, under certain regularity conditions, the MLEs of a HMM parameters are consistent, asymptotically normal and efficient [11, Chapter 12]. Thus, estimating the standard errors of the MLEs can then be used to find approximate confidence intervals by the property of asymptotic normality. It is fairly well known in the litterature, that for (independent) mixture models the sample size has to be very large and that mixtures with small weights or too many components (overfitting) can be of great pratical concern [31, p. 68], [21, p. 53].

In order to estimate the standardr errors of the MLEs of an HMM, we use the approximate Hessian of the minus log-likelihood at the minimum supplied by our `nlm`-optimizer in `R`. This Hessian is inverted to estimat the asymptotic variance-covariance matrix of the estimators of the parameters. However, as the parameters have been transformed, the Hessian is that of the working parameters η_i and not the original natural parameters ζ_i. In other words, we have the Hessian at the minimum of $-\ell$ with respect to the working parameters

$$\boldsymbol{H}_w = -\left(\frac{\partial^2 \ell}{\partial \eta_i \partial \eta_j}\right),$$

but we are interested in the Hessian with respect to the natural parameters

$$\boldsymbol{H}_n = -\left(\frac{\partial^2 \ell}{\partial \zeta_i \partial \zeta_j}\right).$$

From [36, p. 247], the following relation holds between $\boldsymbol{H}_w$ and $\boldsymbol{H}_n$ at the minimum (and for the rest of the paragraph, every matrix is evaluated at the minimum)

$$\boldsymbol{H}_w = \boldsymbol{M}\boldsymbol{H}_n\boldsymbol{M}^\top \quad \text{and} \quad \boldsymbol{H}_n^{-1} = \boldsymbol{M}\boldsymbol{H}_w\boldsymbol{M}^\top, \tag{21}$$

where $\boldsymbol{M}$ has entries $m_{ij} = \partial\zeta_j/\partial\eta_i$. Using Equation 21 and that $\boldsymbol{M}$ is readily avaliable, we deduce $\mathbf{H}_n^{-1}$ from $\boldsymbol{H}_w^{-1}$ and then use $\boldsymbol{H}_n^{-1}$ to find standard errors for the natural parameters, provided such parameters are not on the boundary of the parameter space.

2.2.6 Forecasting, Decoding and State Prediction

In this section, $\boldsymbol{\delta}$ denotes the initial distribution, but every result is identical if it were to be the stationary distribution.

Conditional Densities Using the HMM likelihood formulation discussed in Section ??, we obtain for $t = 2, 3, \ldots, T$ that

$$\begin{aligned} f_{X_t|\mathbf{X}^{(-t)}}(x_t \mid \mathbf{x}^{(-t)}) &= \frac{\boldsymbol{\delta}\boldsymbol{P}(x_1)\boldsymbol{B}_2\cdots\boldsymbol{B}_{t-1}\boldsymbol{\Gamma}\boldsymbol{P}(x_t)\boldsymbol{B}_{t+1}\cdots\boldsymbol{B}_T\mathbf{1}_N^\top}{\boldsymbol{\delta}\boldsymbol{P}(x_1)\boldsymbol{B}_2\cdots\boldsymbol{B}_{t-1}\boldsymbol{\Gamma}\boldsymbol{B}_{t+1}\cdots\boldsymbol{B}_T\mathbf{1}_N^\top} \\ &\propto \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\boldsymbol{P}(x_t)\boldsymbol{\beta}_t^\top, \end{aligned} \tag{22}$$

where $\boldsymbol{B}_t = \boldsymbol{\Gamma}\boldsymbol{P}(x_t)$, $\boldsymbol{\alpha}_t = \boldsymbol{\delta}\boldsymbol{P}(x_1)\boldsymbol{B}_2\cdots\boldsymbol{B}_t$, and $\boldsymbol{\beta}_t^\top = \boldsymbol{B}_{t+1}\cdots\boldsymbol{B}_T\mathbf{1}_N^\top$.

For $t = 1$, we similarly have

$$f_{X_1|\mathbf{X}^{(-1)}}(x_1 \mid \mathbf{x}^{(-1)}) = \frac{\boldsymbol{\delta}\boldsymbol{P}(x_1)\boldsymbol{B}_2 \cdots \boldsymbol{B}_T\mathbf{1}_N^\top}{\boldsymbol{\delta}\boldsymbol{I_N}\boldsymbol{B}_2 \cdots \boldsymbol{B}_T\mathbf{1}_N^\top} \\ \propto \boldsymbol{\delta}\boldsymbol{P}(x_1)\boldsymbol{\beta}_1^\top. \tag{23}$$

This ratio represents the conditional density of x_t, where the numerator corresponds to the joint likelihood with the observation at time t replaced by x_t, while the denominator is the full likelihood of the observed data, treating x_t as missing.

These conditional densities can be expressed as mixtures of the state-dependent densities. Since $\boldsymbol{P}(x) = \text{diag}(f_1(x), \ldots, f_N(x))$ is diagonal, both Equation 22 and Equation 23 yield

$$f_{X_t|\mathbf{X}^{(-t)}}(x_t \mid \mathbf{x}^{(-t)}) \propto \sum_{i \in \mathcal{C}} d_i(t) f_{i,X_t}(x_t),$$

where for Equation 22, $d_i(t)$ equals the product of the i'th entry of $\boldsymbol{\beta}_t$ and the i'th entry of $\boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}$, while for Equation 23, it is the product of the i'th entry of $\boldsymbol{\beta}_1$ and the i'th entry of $\boldsymbol{\delta}$. Normalizing these weights gives

$$f_{X_t|\mathbf{X}^{(-t)}}(x_t \mid \mathbf{x}^{(-t)}) = \sum_{i \in \mathcal{C}} w_i(t) f_{i,X_t}(x_t), \qquad w_i(t) = \frac{d_i(t)}{\sum_{j \in \mathcal{C}} d_j(t)}.$$

Here, $w_i(t)$ are mixing weights that depend on the model parameters and on the remaining observations $\mathbf{x}^{(-t)}$.

Forecast Density Forecast densities are a special case of conditional densities. Let $h \in \mathbb{Z}^+$ denote the forecast horizon. For continuous-valued observations, the h-step-ahead forecast density $f_{X_{T+h}|\mathbf{X}^{(T)}}(x_{T+h} \mid \mathbf{x}^{(T)})$ is obtained analogously to Equation 22:

$$\begin{aligned} f_{X_{T+h}|\mathbf{X}^{(T)}}(x_{T+h} \mid \mathbf{x}^{(T)}) &= \frac{f_{\mathbf{X}^{(T)},X_{T+h}}(\mathbf{x}^{(T)}, x_{T+h})}{f_{\mathbf{X}^{(T)}}(\mathbf{x}^{(T)})} \\ &= \frac{\boldsymbol{\delta}\boldsymbol{P}(x_1)\boldsymbol{B}_2 \cdots \boldsymbol{B}_T\boldsymbol{\Gamma}^h\boldsymbol{P}(x_{T+h})\mathbf{1}_N^\top}{\boldsymbol{\delta}\boldsymbol{P}(x_1)\boldsymbol{B}_2 \cdots \boldsymbol{B}_T\mathbf{1}_N^\top} \\ &= \frac{\boldsymbol{\alpha}_T\boldsymbol{\Gamma}^h\boldsymbol{P}(x_{T+h})\mathbf{1}_N^\top}{\boldsymbol{\alpha}_T\mathbf{1}_N^\top} \\ &= \boldsymbol{\phi}_T\boldsymbol{\Gamma}^h\boldsymbol{P}(x_{T+h})\mathbf{1}_N^\top, \qquad \boldsymbol{\phi}_T = \frac{\boldsymbol{\alpha}_T}{\boldsymbol{\alpha}_T\mathbf{1}_N^\top}. \end{aligned}$$

Thus, the forecast density is also a mixture of the N state-dependent densities:

$$f_{X_{T+h}|\mathbf{X}^{(T)}}(x_{T+h} \mid \mathbf{x}^{(T)}) = \sum_{i \in \mathcal{C}} \psi_i(h)\, f_{i,X_{T+h}}(x_{T+h}),$$

where $\psi_i(h)$ is the i'th entry of $\boldsymbol{\phi}_T\boldsymbol{\Gamma}^h$. Since the full forecast distribution is available, it is possible to construct both point and full interval forecasts. As the forecast horizon h increases, the predictive density converges to the marginal stationary density of the HMM, i.e.

$$\lim_{h\to\infty} f_{X_{T+h}|\mathbf{X}^{(T)}}(x_{T+h} \mid \mathbf{x}^{(T)}) = \lim_{h\to\infty} \boldsymbol{\phi}_T\boldsymbol{\Gamma}^h\boldsymbol{P}(x_{T+h})\mathbf{1}_N^\top = \boldsymbol{\delta}^*\boldsymbol{P}(x_{T+h})\mathbf{1}_N^\top,$$

where $\boldsymbol{\delta}^*$ is the stationary distribution of the Markov chain. The limit follows from the fact that for any nonnegative (row) vector $\boldsymbol{\vartheta}$ whose entries sum to 1, the vector $\boldsymbol{\vartheta}\boldsymbol{\Gamma}^h$ approaches $\boldsymbol{\delta}^*$ as $h \to \infty$, provided that the chain is irreducible and aperiodic [19, p. 394].

Decoding We turn to determining the states of the Markov chain that are most likely to have given rise to the observation sequence under the fitted model.

Consider again the vectors of forward $\boldsymbol{\alpha}_t$ and backward probabilities $\boldsymbol{\beta}_t$. For the derivation of the most likely state of the Markov chain at time $t \in \{1, 2, \dots, T\}$, we remind of the equation

$$\alpha_t(i)\beta_t(i) = f_{\mathbf{X}^{(t)},C_t}(\mathbf{x}^{(t)}, i)$$

We can then rewrite the conditional distribution of C_t given the observations, for $i \in \mathcal{C}$ as

$$\begin{aligned}\mathbb{P}\left(C_t = i \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}\right) &= \frac{f_{\mathbf{X}^{(T)},C_t}(\mathbf{x}^{(T)}, i)}{f_{\mathbf{X}^{(T)}}(\mathbf{x}^{(T)})} \\ &= \frac{\alpha_t(i)\beta_t(i)}{\mathcal{L}_T}\end{aligned}$$

For each $t \in \{1, \dots, T\}$, given the observations, the most probable state i_t^*, is defined as

$$i_t^* = \operatorname*{argmax}_{i=1,\dots,N} \mathbb{P}\left(C_t = i \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}\right). \tag{24}$$

In other words, this approach determines the most likely state seperately for each time-t by maximizing the conditional prbabolity. We referer to this as local decoding.

However, we are most interested in the most likely sequence of hidden states. As such, we are interested in the quantity

$$(i_1^*, \dots, i_T^*) = \operatorname*{argmax}_{(i_1,\dots,i_T)\in\mathcal{C}^T} \mathbb{P}\Big(\mathbf{C}^{(T)} = \mathbf{C}^{(T)} \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}\Big). \tag{25}$$

Finding the solution of Equation 25 over all possible state sequences involves N^T function evaluations which is not feasible. A feasible solution is the so called Viterbi algorithm ([54], [20]).

Define

$$\xi_{1i} = f_{X_1,C_1}(x_1, i) = \delta_i f_{i,X_1}(x_1),$$

and for $t = 2, 3, \ldots, T$,

$$\xi_{ti} = \max_{c_1,c_2,\ldots,c_{t-1}} f_{\mathbf{X}^{(t)},C_t,\mathbf{C}^{(t-1)}}\left(\mathbf{x}^{(t)}, i, \mathbf{c}^{(t-1)}\right)$$

This leads us to the recursion of ξ.

Proposition 2.6. *For $t = 2, 3, \ldots, T$ and $j \in \mathcal{C}$, it follows that*

$$\xi_{ij} = \left(\max_i(\xi_{t-1,i}\gamma_{ij}\right) f_{j,X_t}(x_t).$$

Proof. Fix $t \geq 2$ and $j \in \mathcal{C}$. For any $(c_1, \ldots, c_{t-1}) \in \mathcal{C}^{t-1}$, by the HMM conditional independences,

$$\begin{aligned} f_{\mathbf{X}^{(t)},C_t,\mathbf{C}^{(t-1)}}(\mathbf{x}^{(t)}, j, \mathbf{c}^{(t-1)}) &= f_{j,X_t}(x_t)\mathbb{P}(C_t = j \mid C_{t-1} = c_{t-1}) \\ &\quad \times f_{\mathbf{X}^{(t-1)},C_{t-1},\mathbf{C}^{(t-2)}}(\mathbf{x}^{(t-1)}, c_{t-1}, \mathbf{c}^{(t-2)}) \\ &= f_{j,X_t}(x_t)\gamma_{c_{t-1}j} f_{\mathbf{X}^{(t-1)},C_{t-1},\mathbf{C}^{(t-2)}}(\mathbf{x}^{(t-1)}, c_{t-1}, \mathbf{c}^{(t-2)}). \end{aligned}$$

Maximizing over $\mathbf{c}^{(t-1)}$ and extracting the factors that do not depend on the maximization variables gives

$$\xi_{tj} = f_{j,X_t}(x_t) \max_{c_{t-1}\in\mathcal{C}} \left\{ \gamma_{c_{t-1}j} \max_{c_1,\ldots,c_{t-2}\in\mathcal{C}} f_{\mathbf{X}^{(t-1)},C_{t-1},\mathbf{C}^{(t-2)}}\left(\mathbf{x}^{(t-1)}, c_{t-1}, \mathbf{c}^{(t-2)}\right) \right\}.$$

By the definition of $\xi_{t-1,i}$,

$$\max_{c_1,\ldots,c_{t-2}} f_{\mathbf{X}^{(t-1)},C_{t-1},\mathbf{C}^{(t-2)}}(\mathbf{x}^{(t-1)}, i, \mathbf{c}^{(t-2)}) = \xi_{t-1,i},$$

so

$$\xi_{tj} = f_{j,X_t}(x_t) \max_{i\in\mathcal{C}}(\gamma_{ij}\xi_{t-1,i}),$$

which is the desired recursion. The initialization $\xi_{1i} = \delta_i f_{i,X_1}(x_1)$ follows from $f_{X_1,C_1}(x_1, i) = f_{i,X_1}(x_1)\mathbb{P}(C_1 = i)$. □

The required maximizing sequence of states $\{i\}_{i=1}^{T}$ can then be determined recursively from

$$i_T = \underset{i=1,\ldots,N}{\operatorname{argmax}}\, \xi_{Ti},$$

and for $t = T-1, T-2, \dots, 1$ from

$$i_t = \underset{i=1,2,\dots,N}{\operatorname{argmax}} (\xi_{ti}\gamma_{i,i_{t+1}}).$$

Because the global-decoding objective is a product of probabilities, it's convenient to maximize its logarithm to avoid numerical underflow; the Viterbi recursions translate directly to the log domain. As an alternative, you can use likelihood-style scaling by normalizing each time-t row of the matrix $\{\xi_{ti}\}$ so that the entries sum to 1. The Viterbi algorithm applies to both stationary and non-stationary (time-inhomogeneous) Markov chains; the initial distribution need not be the stationary distribution.

State Prediction We turn our attention to finding conditional distributions of C_t when $t > T$, i.e. state prediction.

Proposition 2.7. *Given observations $x_1, \dots, x_T$, it follows that*

$$\mathcal{L}_T \mathbb{P}\left(C_t = i \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}\right) = \begin{cases} \boldsymbol{\alpha}_T \, \mathbf{\Gamma}^{t-T} \, \mathbf{e}_i^\top, & \textit{for } t > T \quad \textit{(state prediction)}, \\ \alpha_T(i), & \textit{for } t = T \quad \textit{(filtering)}, \\ \alpha_t(i)\, \beta_t(i), & \textit{for } 1 \le t < T \quad \textit{(smoothing)}. \end{cases}$$

The filtering and smoothing parts (for present or past states) are identical to the state probabilities and could be combined as $\beta_T(i) = 1$. The state prediction formula is therefore a generalization to $t > T$ and can be restated as

$$\mathbb{P}\left(C_{T+h} = i \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}\right) = \boldsymbol{\alpha}_T \boldsymbol{\Gamma}^h \boldsymbol{e}_i^\top / \mathcal{L}_T = \boldsymbol{\phi}_T \boldsymbol{\Gamma} \boldsymbol{e}_i^\top$$

Proof. Fix $h \ge 1$ and $i \in \{1, \dots, N\}$. By the law of total probability over the current state C_T,

$$\mathbb{P}\big(C_{T+h} = i \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}\big) = \sum_{j=1}^{N} \mathbb{P}\big(C_{T+h} = i \mid C_T = j,\ \mathbf{X}^{(T)} = \mathbf{x}^{(T)}\big)\, \mathbb{P}\big(C_T = j \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}\big).$$

By the (time-homogeneous) Markov property of $\{C_t\}$ and the HMM conditional-independence structure, the future of the chain depends on the past observations only through C_T, hence

$$\mathbb{P}\big(C_{T+h} = i \mid C_T = j,\ \mathbf{X}^{(T)} = \mathbf{x}^{(T)}\big) = \mathbb{P}\big(C_{T+h} = i \mid C_T = j\big) = (\mathbf{\Gamma}^h)_{ji}.$$

Therefore,

$$\begin{aligned}\mathbb{P}\big(C_{T+h} = i \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}\big) &= \sum_{j=1}^{N} (\boldsymbol{\Gamma}^{h})_{ji}\, \mathbb{P}\big(C_T = j \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}\big) \\ &= \boldsymbol{\phi}_T\, \boldsymbol{\Gamma}^{h}\, \mathbf{e}_i^{\top}.\end{aligned}$$

Using $\boldsymbol{\phi}_T = \boldsymbol{\alpha}_T / \mathcal{L}_T$ yields the equivalent form $\mathbb{P}(C_{T+h} = i \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \boldsymbol{\alpha}_T \boldsymbol{\Gamma}^{h} \mathbf{e}_i^{\top} / \mathcal{L}_T$.

For the special cases: when $h = 0$, $\boldsymbol{\Gamma}^{0} = \mathbf{I}$, so $\mathbb{P}(C_T = i \mid \mathbf{X}^{(T)}) = \phi_T(i) = \alpha_T(i)/\mathcal{L}_T$ (filtering). For $t < T$, the standard forward–backward identity $\mathbb{P}(C_t = i \mid \mathbf{X}^{(T)}) = \alpha_t(i)\beta_t(i)/\mathcal{L}_T$ holds, with $\beta_T(i) = 1$ by definition of the backward variables, giving the stated smoothing expression. □

Note that as $h \to \infty$, $\boldsymbol{\phi}_T \boldsymbol{\Gamma}^h \to \boldsymbol{\delta}$ and so $\mathbb{P}\left(C_{T+h} = i \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}\right) \to \delta_i$.

2.2.7 Number of States

HMMs are prone to overfitting [23, p. 2]. That is, HMMs 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 HMM?

In HMMs, 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[3]. 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. [39] and [35, p. 4] highlighted this issue, recommending that the choice of 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 2.4.1.

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 [39] and [35], we determine the number of states based on domain expertise only. However, we note that the application of HMMs to financial contexts—and especially to interest

[3] This will become evident in Section 2.4.1.

rates—remains extremely limited. Consequently, we must develop our own arguments to justify the chosen number of states based on domain expertise.

In the context of equity market modeling, selecting between 1 and 5 states provides a meaningful balance between model complexity, interpretability, and macro-financial relevance. A two-state model may capture broad bull and bear market regimes. A third state can correspond to recovery or neutral phases in market cycles. Higher state counts may reflect nuanced market phases, such as asset bubbles, mild corrections, or crashes.

The number of states affects the parameter estimation in a regime-switching Black-Scholes model. In particular, the volatility parameter σ becomes state-dependent. If too many states are included, temporary fluctuations in volatility may be mistaken for persistent regime shifts. Conversely, too few states may obscure significant differences in market regimes, such as between stable bull markets and volatile rallies.

We interpret (or rather hypothesize) the possible values of $N \in \{1, 2, 3, 4, 5\}$ as follows:

- $N = 1$: The trivial case—no regime-switching. The standard Black-Scholes model with constant drift and volatility.
- $N = 2$: A dichotomy of bull and bear markets, capturing broad upturn and downturn market regimes.
- $N = 3$: Extension to classical business cycle phases: expansion, recession, and recovery.
- $N = 4$: Potential to differentiate mild versus severe market states, e.g., modest bull markets vs. overheated bubbles, or shallow vs. deep downturns.
- $N = 5$: Allows capturing even finer distinctions, such as neutral/stagnant markets or extreme panic phases during financial crises.

For example, during a moderate downturn, the drift μ may slightly decline and volatility σ rise modestly, reflecting controlled risk aversion. In contrast, in an extreme crisis, μ becomes strongly negative and σ spikes due to panic-driven trading, credit contractions, and liquidity crises. Capturing both with the same state may understate risk in crisis scenarios or overstate it in moderate corrections.

In summary, based on economic reasoning and business cycle theory, we restrict our analysis to $N \in \{1, 2, 3, 4, 5\}$. This reflects plausible macroeconomic regimes and aims to avoid overfitting while maintaining explanatory power and interpretability in financial modeling.

2.2.8 Simulation

We simulate the BS-HMM with parameters $n = 25000$, and daily observations $\Delta = 1/252$ (approximately $25000/252 \approx 99$ years). The model parameters are seen in Table 2. To simulate from a N-state hidden Markov model, we extend the Euler discretized version of the BS SDE given in Equation 11 to include a state sequence from a simulated Markov chain, simply by using the `sample()`-function in base `R`. Combining the simulated Markov chain with the discretized BS SDE, we achieve the state-dependent Euler-discretized version of the BS SDE

$$\hat{S}_{t+\Delta} = \hat{S}_t + \mu_i \hat{S}_t \Delta + \sigma_i \hat{S}_t \sqrt{\Delta} Z^{\mathbb{P}}, \quad i \in \mathcal{C}$$

We limit ourself to the most daunting and computationally dragging case, which is the 5-state BS-HMM. The results are seen in Table 2. The simulated price path is shown in figure ??. We use `nlm` in `R` to maximize the likelihood given in Equation 20 for the 5-state BS-HMM with both μ and σ state-dependent.

Parameter	True Values	Estimated Values	Relative Error (%)
μ	$\begin{bmatrix} 0.02000 \\ 0.04000 \\ 0.05000 \\ 0.08000 \\ 0.10000 \end{bmatrix}$	$\begin{bmatrix} 0.02168 \\ 0.04511 \\ 0.02251 \\ 0.08053 \\ 0.09927 \end{bmatrix}$	$\begin{bmatrix} 8.40 \\ 12.8 \\ 54.9 \\ 0.656 \\ 0.730 \end{bmatrix}$
σ	$\begin{bmatrix} 0.00500 \\ 0.01000 \\ 0.01500 \\ 0.02000 \\ 0.02500 \end{bmatrix}$	$\begin{bmatrix} 0.00510 \\ 0.01039 \\ 0.01419 \\ 0.01842 \\ 0.02424 \end{bmatrix}$	$\begin{bmatrix} 2.00 \\ 3.90 \\ 5.40 \\ 7.90 \\ 3.04 \end{bmatrix}$
δ	$\begin{bmatrix} 0.20000 \\ 0.20000 \\ 0.20000 \\ 0.20000 \\ 0.20000 \end{bmatrix}$	$\begin{bmatrix} 0.21346 \\ 0.20233 \\ 0.15377 \\ 0.17229 \\ 0.25815 \end{bmatrix}$	$\begin{bmatrix} 6.73 \\ 1.16 \\ 23.1 \\ 13.9 \\ 29.1 \end{bmatrix}$
Γ	$\begin{bmatrix} 0.93174 & 0.01707 & 0.01707 & 0.01707 & 0.01707 \\ 0.01707 & 0.93174 & 0.01707 & 0.01707 & 0.01707 \\ 0.01707 & 0.01707 & 0.93174 & 0.01707 & 0.01707 \\ 0.01707 & 0.01707 & 0.01707 & 0.93174 & 0.01707 \\ 0.01707 & 0.01707 & 0.01707 & 0.01707 & 0.93174 \end{bmatrix}$	$\begin{bmatrix} 0.92380 & 0.01792 & 0.01986 & 0.01009 & 0.02833 \\ 0.01530 & 0.87999 & 0.02855 & 0.04281 & 0.03336 \\ 0.02949 & 0.00005 & 0.84818 & 0.03968 & 0.05340 \\ 0.02303 & 0.02058 & 0.04967 & 0.87150 & 0.03522 \\ 0.01962 & 0.01225 & 0.03552 & 0.00732 & 0.92529 \end{bmatrix}$	$\begin{bmatrix} 0.850 & 4.97 & 16.3 & 40.9 & 66.0 \\ 10.3 & 5.56 & 67.3 & 151 & 95.4 \\ 72.8 & 99.7 & 8.98 & 132 & 213 \\ 35.0 & 20.5 & 191 & 6.47 & 106 \\ 14.9 & 28.2 & 108 & 57.1 & 0.690 \end{bmatrix}$

Table 2: *True, estimated, and relative error (%) for the Black–Scholes 5-state hidden Markov model parameters where μ and σ are modeled AS state-dependent.*

The issue with the maximum likelihood estimates in the 5-state BS-HMM with both μ and σ state-dependent is that the Euler discretization error is exaggerated by the nature of the Markov-switching. A transition from state-i to state-j will cause the convergence of the Euler discretized stock price path to be throttled. A change of state will inevitably reset the convergence and thus induce bias. As such, it is difficult and perhaps a thesis or Ph.d. of its own, to examine the convergence of discretizations in a Markov-switching model.

2.3 Continuous State-Space Models

The structure of sates in HMMs is often not well suited to problems at hand. As discussed in Section 2.2.7, the number of states is often not known a priori. As such, one has to rely on domain expertise to specify the number of states. In some instances, the choice of the number of states can be determined by model selection criteria and examination of residuals. Furthermore, the states can be intuitive and visited a *reasonably* number of times. However, in fact, most of the time the number of states remains difficult in pratice when the underlying data generating process is unknown. Furthermore, as the number of states can possibly be extremely large, the number of parameters rise extremely fast. If we for example have 10 states and 2 state-dependent variables and no state-independent variables, we need to estimate a whopping $10^2 + 10 = 110$ parameters for the 10-state HMM with 2 state-dependent parameters and no state-independent parameters. In these cases it can be advantageous to consider alternative models formulations where the state process is continuous-valued as opposed to discrete-valued and is relatively parsimonious in terms of the number of parameters for estimation.

2.3.1 Autoregressive Processes

For the S&P 500 data, it would be intuitive to assume that the rate of occurence is continuous-valued, as that would allow for gradual change over the years. As means of investment, methods and behavior is constantly changing. A simple model that would capture such changes could be formulated as

$$\begin{aligned} &C_t = \rho\, C_{t-1} + \varepsilon_t, \\ &X_t \sim \mathcal{N}\Big(\big(e^{C_t}\mu - \tfrac{1}{2}\left(e^{C_t}\sigma\right)^2\big)\,\Delta,\; \left(e^{C_t}\sigma\right)^2\Delta\Big)\,. \end{aligned} \tag{26}$$

for $t = 1, 2, ...$ and with the recursion initiated in $C_0 = C \in \mathbb{R}_+$[4]. The autoregressive parameter is $\rho \in \mathbb{R}$ and the innovations ε_t are independently and identically distributed with a normal distribution with mean zero and variance σ_ε^2[5]. In other words ε_t are i.i.d. $\mathcal{N}(0, \sigma_\varepsilon^2)$. This is called an autoregressive process of order 1. It follows that $\mathbb{E}[C_t \mid C_{t-1}] = \rho C_{t-1}$ while $\mathbb{V}[C_t \mid C_{t-1}] = \sigma_\epsilon^2$, and therefore the time-dependence, or dynamics, is modelled through the conditional mean of C_t given the past.

The dynamics of the AR(1) process is clearly seen to be dependent on the autoregressive parameter, ρ. The simple recurssion in Equation 26 can be written as

$$C_t = \rho^t C + \sum_{i=0}^{t-1} \rho^i \varepsilon_{t-i}.$$

In particular, C_t is normally distributed with time-varying parameters

$$\begin{aligned} \mathbb{E}[C_t] &= \rho^t C \\ \mathbb{V}[C_t] &= \left(1 + \rho^2 + \rho^4 + ... + \rho^{2(t-1)}\right)\sigma_\varepsilon^2 \end{aligned}$$

[4] Note that we index from 0 for the AR(1) theory rather than 1 as the HMM theory. This is to adhere to the general literature and is a simple index shift.

[5] We use the subscript ε in σ_ε to avoid confusion with the BSM parameter σ.

If $|\rho| < 1$, $\mathbb{E}[C_t] \to 0$ as $\rho^t \to 0$ for $t \to \infty$. Concludingly, for $|\rho| < 1$, as $t \to \infty$, C_t will resemble the so called linear-process,

$$C_t^* = \sum_{i=0}^{\infty} \rho^i \varepsilon_{t-i},$$

in terms of the sequence $\{\varepsilon_t\}_{t=\dots,-1,0,1,\dots}$ of i.i.d. $\mathcal{N}(0, \sigma_\varepsilon^2)$ variables. The process $\{C_t^*\}_{t=0,1,\dots}$ is then Gaussian distributed with

$$\mathbb{E}[C_t^*] = 0,$$
$$\mathbb{V}[C_t^*] \overset{\dagger}{=} \frac{\sigma_\varepsilon^2}{(1-\rho^2)},$$

as

$$\frac{1-\rho^{2t}}{1-\rho^2}\sigma_\varepsilon^2 \xrightarrow{\dagger\dagger} \frac{\sigma_\varepsilon^2}{(1-\rho^2)}, \quad t \to \infty,$$

where † and †† follows from Lemma A.2.6. The distribution of C_t^* clearly is clearly indepedent of time and is an example of a stationary process. Thus, if $|\rho| < 1$, C_t is asymptotically stationary in the sense that it resembles the stationary process C_t^* for $t \to \infty$[6].

Stationarity & Distributions Proceeding, we formalize the notion of stationarity and examine distributional properties of the AR(1) process by introducing the notion of a drift function.

Definition 2.3. *The process $\{C_t\}_{t=0,1,\dots}$ is said to be a staionary process if for all $t, h \geq 0$, the joint distribution of $(C_t, \dots, C_{t+h})$ does not depend on $t \geq 0$.*

By Definition 2.3, note that for a stationary process with well-defined second order moments, $\mathbb{E}[C_t]$ and $\mathbb{V}[C_t]$ are constant and that the covariance between C_t, C_{t+h}, i.e. $\text{Cov}[C_t, C_{t+h}]$ depends only on h and not t.

The definition of stationarity comments only on the joint distribution of the variables but nothing about dependence over time. Assume $\{C_t\}_{t\in\mathbb{Z}}$ is a stationary process that is dependent over time with finite second order moment $\mathbb{E}\left[|C_t|^2\right] < \infty$. A often used indicator to detect dependence is the auto-correlation. For a stationary process $C_t \in \mathbb{R}$, the autocovariance function is given by

$$v(h) = \text{Cov}[C_t, C_{t+h}],$$

[6] We will formalize the concepts in the next paragraph.

and autocorrelation function (ACF) defined by,

$$\mathrm{ACF}(h) = \mathrm{Corr}[C_t, C_{t+h}] = \frac{\mathrm{Cov}[C_t, C_{t+h}]}{\sqrt{\mathbb{V}[C_t]\mathbb{V}[C_{t+h}]}} \overset{\dagger}{=} \frac{v(h)}{v(0)},$$

where † holds by stationarity. The functions for various h describe the correlation and hence indicate dependence over time.

In general, "mixing" (i.e., asymptotic independence) captures that the dependence between C_t, C_{t+h} vanishes as $h \to \infty$. This idea is crucial for time series and replaces the concept of indepence. The idea is that a stationary process $\{C_t\}_{t=0,1,\dots}$ is said to be mixing[7] (or, ergodic) if for all t, h and sets A, B,

$$\mathbb{P}((C_0, \dots C_t) \in A, (C_h, \dots, C_{t+h}) \in B) \to \mathbb{P}((C_0, \dots, C_t) \in A)\mathbb{P}((C_0, \dots, c_t) \in B), \quad h \to \infty$$

The notion is intuitively, that events removed far in time from one another are independent. Importantly, they imply that various Laws of Large Numbers apply.

We now turn our attention to the *drift criterion* which establishes conditions under which LLNS and central limit theorems (CLTs) hold for time series. Let $\{C_t\}_{t=0,1,\dots}$ be a Markov chain that satisfies the drift criterion. The first implication of satisfying said criterion is that the initial value, C_0, can be assigned a distribtuion such that X_t is stationary. The second implication is finiteness of certain moments for the stationary version. Moreover, variations of LLN and CLT can be applied. Let $\{C_t\}_{t=0,1,\dots}$ denote a AR(1) process. Then, the distribtution of $C_t \mid (C_{t-1}, \dots, C_0)$, $t \geq 1$ depends only on C_{t-1}, meaning, $C_t \mid C_{t-1} \sim \mathcal{N}(C_{t-1}\rho, \sigma_\varepsilon^2)$. As can be seen, the conditional distribution is Gaussian, which has some attrative properties.

We now state 2 assumptions based on those of [52] and [34].

Assumption 2.2. *Assume that for $\{C_t\}_{t=1,0,\dots}$ with $C_t \in \mathbb{R}^p$ it holds that:*

(i) The conditional distribution of C_t given $(C_{t-1}, C_{t-2}, \dots, C_0)$ depends only on C_{t-1}, that is

$$C_t \mid C_{t-1}, C_{t-2}, \dots, C_0 \overset{d}{=} C_t \mid C_{t-1}.$$

(ii) The conditional distribution of C_t given C_{t-n}, for some $n \geq 1$, has a positive (n-step) conditional density $f(y \mid x) > 0$, which is continuous in both arguments.

Now, simply note that (i) in Assumption 2.2 implies that $\{C_t\}_{t=0,1,\dots,}$ is a Markov chain on $\mathbb{R}^p$, or sometimes called a *Markov chain on a general state space.*

Example: For the purpose of our analysis, consider the AR(1) process. As ε_t are i.i.d. $\mathcal{N}(0, \sigma_\varepsilon^2)$

[7] The litterature differs a lot on the notion on mixing; some even include different kind of mixing (α, β-mixing.

and independent of $(C_{t-1},\dots,C_0)$, C_t conditional on $(C_{t-1},\dots,C_0)$ has density

$$f_{C_t|C_{t-1}}(c_t \mid c_{t-1}) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(-\frac{(c_t-\rho c_{t-1})^2}{2\sigma_\varepsilon^2}\right),$$

which only depends on C_{t-1} and is well known to be positive and continuous in both arguments.

Next, define the so-called *drift function* that satisfies Assumption Assumption 2.2. A drift function for some time series C_t, is some function $\delta(C_t) \geq 1$ and which is not identically ∞. The role of a drift function is to measure the dynamical drift of X_t by studying the dynamics of the corresponding drift function $\delta(C_t)$. That is, we are interested in $\mathbb{E}[\delta(C_t) \mid C_{t-m}]$ for some $m \geq 1$.
Example: Consider again the AR(1) process with drift function $\delta(C_t) = 1 + C_t^2$ and $m = 1$. Then, using that C_{t-1} and ε_t are independent, we obtain

$$\begin{aligned}
\mathbb{E}[\delta(C_t) \mid C_{t-1}] &= \mathbb{E}[1 + (\rho C_{t-1} + \varepsilon_t)^2 \mid C_{t-1}] \\
&= 1 + \rho^2 \mathbb{E}[C_{t-1}^2 \mid C_{t-1}] + 2\rho C_{t-1}\mathbb{E}[\varepsilon_t \mid C_{t-1}] + \mathbb{E}\left[\varepsilon_t^2 \mid C_{t-1}\right] \\
&= 1 + \rho^2 C_{t-1}^2 + 2\rho C_{t-1}\mathbb{E}[\varepsilon_t] + \mathbb{E}\left[\varepsilon_t^2\right] \\
&= 1 + \sigma^2 + \rho^2 C_{t-1}^2 \\
&= \rho^2\delta(C_{t-1}) + c, \quad c = (1 - \rho^2 + \sigma_\varepsilon^2).
\end{aligned}$$

Thus we obtain a process that mimics a AR(1) process in $\delta(C_t)$, apart from some constant c. In other words,

$$\delta(C_t) = \rho^2\delta(C_{t-1}) + c + \eta_t,$$

with $\eta_t = (\delta(C_t) - \mathbb{E}[\delta(C_t) \mid C_{t-1}])$ such that $\mathbb{E}[\eta_t] = 0$. As such, if $\rho^2 < 1$, $\delta(C_t)$ resembles a stationary AR(1) process. This leads us to the final assumption.

Assumption 2.3. *Assume that $\{C_t\}_{t=0,1,\dots}$, with $C_t \in \mathbb{R}^p$, satisfies Assumption Assumption 2.2. With drift function δ, $\delta(C_t) \geq 1$, assume that there exist positive constants M, C, and φ with $\varphi < 1$, such that for some $m \geq 1$,*

(i) $\mathbb{E}\big[\delta(C_{t+m}) \,\big|\, C_t = C\big] \leq \varphi\delta(C)$, *for* $\|C\| > M$,

(ii) $\mathbb{E}\big[\delta(C_{t+m}) \,\big|\, C_t = C\big] \leq C < \infty$, *for* $\|C\| \leq M$.

Example: Consider again the AR(1). To obtain the desired properties for C_t by making restrictions on ρ, we apply the drift criterion with $\delta(C_t) = 1 + C_t^2$. Using our prievous example, we saw that:

$$\mathbb{E}[\delta(C_t) \mid C_{t-1} = C] = \rho^2\delta(C) + c, \quad c = (1 - \rho^2 + \sigma_\varepsilon^2).$$

Since

$$\lim_{|C|\to\infty} \frac{\mathbb{E}[\delta(C_t) \mid C_{t-1} = C]}{\delta(C)} = \rho^2,$$

we must require that $|\rho| < 1$ for the existence of positive constants M, φ with $\varphi < 1$, such that $\mathbb{E}[\delta(C_t) \mid C_{t-1} = C] \leq \varphi\delta(C)$ for $|C| > M$. By continuity of $\delta(C)$ and c, it automatically follows that $\mathbb{E}[\delta(C_t) \mid C_{t-1} = C] \leq C$ for some $C > 0$ for $|y| \leq M$. In other words, if $|\rho| < 1$, the AR(1) process satisfies the drift criterion in Assumption 2.3.

Theorem 2.4. *Assume that $\{C_t\}_{t\geq 1}$ satisfies Assumption 2.3 with drift function δ. Then C_1 can be given an initial distribution such that C_t initiated in X_1 is stationary. With C_t denoting the stationary version, we have $\mathbb{E}\big[\delta(C_t)\big] < \infty$. Moreover, C_t is mixing in the sense that, for any initial value C_1, the LLN Lemma 2.1 (seen below) applies.*

We turn our attention to distributional properties of the MLE estimators in the AR(1) process. By definition, the density of C_t, conditional on C_{t-1} is the Gaussian density with mean ρC_{t-1} and variance σ_ε^2. Moreover, the joint density of $\{C_t\}_{t=1}^T$ with the initial value $C_0 = C$ fixed, factorizes as follows

$$f(C_T, C_{T-1}, \ldots, C_1 \mid C_0) = \prod_{t=1}^{T} f(C_t \mid C_{t-1}) \tag{27}$$

Denote the likelihood function as $\mathcal{L}(\rho, \sigma^2)$. Then by the factorization in Equation 27 of the joint density of $C_1, \ldots, C_T$ given C_0, gives the log-likelihood function

$$\begin{aligned}\ell(\rho, \sigma^2) = \log \mathcal{L}(\rho, \sigma^2) &= \log\left(\prod_{t=2}^{T} f(C_t \mid C_{t-1})\right) \\ &= -\frac{T}{2}\underbrace{\log(2\pi)}_{*} - \frac{T}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=2}^{T}(C_t - \rho C_{t-1})^2.\end{aligned} \tag{28}$$

Note that the term $*$ does not contain any parameters and does not matter for parameter estimation. However, we include it for likelihood calculations as AIC and BIC will be reported.

Theorem 2.5. *The MLEs of ρ and σ_ε for the AR(1) model are given by*

$$\hat{\rho} = \frac{\frac{1}{T}\sum_{t=2}^{T} C_t C_{t-1}}{\frac{1}{T}\sum_{t=1}^{T} C_{t-1}^2}$$

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}(C_t - \hat{\rho} C_{t-1})^2$$

Ignoring a constant factor, the maximized likelihood function is given by

$$\mathcal{L}(\hat{\rho}, \hat{\sigma}^2) = (2\pi e\hat{\sigma}^2)^{-T/2}.$$

Proof. Differentiating Equation 28 with respect to the parameters (ρ, σ^2) gives us FOCs:

$$(1): \quad \sum_{t=2}^{T}(C_t - \rho C_{t-1})C_{t-1} = 0, \qquad (2): \quad \frac{1}{T}\sum_{t=2}^{T}(C_t - \rho C_{t-1})^2 = \sigma^2.$$

The first equality (1) instantly leads to the MLE of $\hat{\rho}$ by multiplying the parenthesis and isolating ρ. Substituting $\hat{\rho}$ into the second FOC (2) gives us that $\hat{\sigma}^2$ is exactly the residual sum of squares

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}(C_t - \hat{\rho}C_{t-1})^2.$$

Lastly, the second order derivatives evaulated at $(\hat{\rho}, \hat{\sigma}^2)$ equal

$$\begin{aligned}
\frac{\partial^2}{\partial \rho^2}\log\mathcal{L}\Big|_{(\hat{\rho},\hat{\sigma}^2)} &= -\frac{T}{\hat{\sigma}^2}\left(\sum_{t=1}^{T} C_{t-1}^2\right), \\
\frac{\partial^2}{\partial (\sigma^2)^2}\log\mathcal{L}\Big|_{(\hat{\rho},\hat{\sigma}^2)} &= -\frac{T}{2\hat{\sigma}^4}, \\
\frac{\partial^2}{\partial \sigma^2 \partial \rho}\log\mathcal{L}\Big|_{(\hat{\rho},\hat{\sigma}^2)} &= 0,
\end{aligned}$$

implying that $\mathcal{L}(\rho, \sigma^2)$ has maximum given by $(2\pi)^{-T/2}e^{-T/2}(\hat{\sigma}^2)^{-T/2} = (2\pi e\hat{\sigma}^2)^{-T/2}$. □

Next, we consider asymptotic properties of the MLEs. $_0$ will denote the values of the parameters under which the probabilistic arguments are made, or in a more intuitive sense, the true-values. However, firstly, one of the infamous Law of Large Numbers (LLN) [40, p. 17] and Central Limit Theorem (CLT) [40, p. 20].

Lemma 2.1. *Assume that with $C_t \in \mathbb{R}^p$, $\{C_t\}_{t=0}^T$ is a geometrically ergodic markov chain with statioanry solution $\{C_t^*\}$. Assume furthermore that the function $g : \mathbb{R}^{p(m+1)} \to \mathbb{R}$, $m \geq 0$, satisfies $\mathbb{E}\left[g(C_t^*, C_{t-1}^*, \ldots, C_{t-m}^*)\right] < \infty$, then as $T \to \infty$,*

$$\frac{1}{T}\sum_{t=1}^{T} g(C_t, C_{t-1}, \ldots, X_{t-m}) \xrightarrow{\mathcal{P}} \mathbb{E}[g(C_t^*, C_{t-1}^*, \ldots, C_{t-m}^*)].$$

Lemma 2.2. *For a given sequence $\{C_t\}_{t\geq 1}$, consider $C_t = f(C_t, C_{t-1}, \ldots, C_m)$, with f continuous,* *with* $\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] = 0$*, where* $\mathcal{F}_t = (C_t, \ldots, C_0)$*. If*

$$\mathbf{I}: \quad \frac{1}{T}\sum_{t=1} \mathbb{E}[Y_t^2 \mid \mathcal{F}_{t-1}] \xrightarrow{\mathcal{P}} \sigma^2 > 0$$

and either **II** *or* **II'** *hold,*

$$\mathbf{II}: \quad \frac{1}{T}\sum_{t=1} \mathbb{E}\left[Y_t^2 \mathbb{1}_{\{|Y_t| > \delta T^{1/2}\}}\right] \to 0,$$

$$\mathbf{II'}: \quad \frac{1}{T}\sum_{t=1} \mathbb{E}\left[Y_t^2 \mathbb{1}_{\{|Y_t| > \delta T^{1/2}\}} \mid \mathcal{F}_{t-1}\right] \xrightarrow{\mathcal{P}} 0,$$

for any $\delta > 0$*, then* $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Y_t \xrightarrow{\mathcal{P}} \mathcal{N}(0, \sigma^2)$*.*

Theorem 2.6. *For* $|\rho_0| < 1$*, the MLEs of the AR(1) model are consistent,* $\hat{\rho} \xrightarrow{\mathcal{P}} \rho_0$ *and* $\hat{\sigma} \xrightarrow{\mathcal{P}} \sigma_0$ *as* $T \to \infty$*. Moreover,*

$$\sqrt{T}(\hat{\rho} - \rho_0) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, 1 - \rho_0^2\right),$$

and a consistent estimator of the asymptotic variance is given by $\hat{\sigma}^2\left(\frac{1}{T}\sum_{t=1}^{T} C_{t-1}^2\right)$*, such that*

$$\sqrt{T}(\hat{\rho} - \rho_0)\sqrt{\frac{\frac{1}{T}\sum_{t=1}^{T} C_{t-1}^2}{\hat{\sigma}^2}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \quad T \to \infty.$$

Note: *The above distributional statements do not hold without the crucial condition* $|\rho_0| < 1$*.*

Proof. Since $|\rho_0| < 1$ we recall that C_t is geometrically ergodic and the LLN Lemma 2.1 applies. Consider $\hat{\rho}$ as given by,

$$\hat{\rho} = \left(\frac{1}{T}\sum_{t=1}^{T} C_t C_{t-1}\right)\left(\frac{1}{T}\sum_{t=1}^{T} C_{t-1}^2\right)^{-1}.$$

By the LLN applied to $C_t C_{t-1}$ and C_{t-1}^2, and as $\mathbb{E}[C_t^{*2}] < \infty$, it follows directly that

$$\hat{\rho} \xrightarrow{\mathcal{P}} \mathrm{Cov}\left[C_t^*, C_{t-1}^*\right] / \mathbb{V}\left[C_t^*\right] = \rho_0.$$

Furthermore,

$$\begin{aligned}\hat{\sigma}^2 &\xrightarrow{\mathcal{P}} \mathbb{V}\left[C_t^*\right] - \left(\mathrm{Cov}[C_t^*, C_{t-1}^*]\right)^2 / \mathbb{V}[C_t^*] \\ &= \frac{\sigma_0^2}{1 - \rho_0^2} - \rho_0^2 \frac{\sigma_0^2}{1 - \rho_0^2} = \sigma_0^2,\end{aligned}$$

as claimed and it follows that

$$\hat{\sigma}^2 \left(\frac{1}{T} \sum_{t=1}^{T} C_{t-1}^2 \right) \xrightarrow{\mathcal{P}} 1 - \rho_0^2.$$

For the asymptotic distribution, we have that

$$\sqrt{T}(\hat{\rho} - \rho_0) = \frac{\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \varepsilon_t C_{t-1}}{\frac{1}{T} \sum_{t=1}^{T} C_{t-1}^2},$$

where $\frac{1}{T} \sum_{t=1}^{T} C_{t-1}^2 \xrightarrow{\mathcal{P}} \frac{\sigma_0^2}{1-\rho_0^2}$ by the LLN Lemma 2.1. Define $Y_t \equiv \varepsilon_t C_{t-1}$. As $\varepsilon_t = C_t - \rho_0 C_{t-1}$, Y_t is a martingale difference sequence with respect to $\mathcal{F}_t$, and thus the CLT Lemma 2.2 can be applied. Observat that

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] = \sigma_{\varepsilon,0}^2 \left(\frac{1}{T} \sum_{t=1}^{T} C_{t-1}^2 \right) \xrightarrow{\mathcal{P}} \frac{\sigma_{\varepsilon,0}^4}{1 - \rho_0^2}.$$

Next let $A = \{|X| > \delta\sqrt{T}\}$ with $\delta > 0$, $T > 0$, and assume $\mathbb{E}[X^4] < \infty$. Then by Cauchy–Schwarz and Markov's inequality (see, e.g., [16, 5, 24])

$$\begin{aligned}
\mathbb{E}\big[X^2 \, \mathbb{1}_{\{A\}}\big] &\le \Big(\mathbb{E}[X^4]\Big)^{1/2} \Big(\mathbb{E}[\mathbb{1}_{\{A\}}]\Big)^{1/2} && \text{(by Cauchy–Schwarz)} \\
&= \Big(\mathbb{E}[X^4]\Big)^{1/2} \mathbb{P}(A)^{1/2} \\
&\le \Big(\mathbb{E}[X^4]\Big)^{1/2} \left(\frac{\mathbb{E}[X^4]}{\delta^4 T^2}\right)^{1/2} && \text{(by Markov on } X^4\text{)} \\
&= \frac{\mathbb{E}[X^4]}{\delta^2 T}.
\end{aligned}$$

Using the inequality yields

$$\begin{aligned}
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[Y_t^2 \mathbb{I}_{\{|Y_t| > \delta\sqrt{T}\}} \mid \mathcal{F}_{t-1}\right] &\le \frac{1}{T^2 \delta^2} \sum_{t=1}^{T} \mathbb{E}\left[Y_t^4 \mid \mathcal{F}_{t-1}\right] \\
&= \frac{1}{T\delta^2} \left(\frac{1}{T} \sum_{t=1}^{T} C_{t-1}^4 \right) \mathbb{E}\left[\varepsilon_t^4\right] \xrightarrow{\mathcal{P}} 0.
\end{aligned}$$

We conclude that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} Y_t = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \varepsilon_t C_{t-1} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{\sigma_{\varepsilon,0}^4}{1 - \rho_0^2}\right).$$

Lastly, substituting in gives us the final resul for $\hat{\rho}$. □

Unit roots The underlying assumption of the before-mentioned analyses relied on the fact that $|\rho| < 1$. As stated in cite ?? (under theorem II.2.2), often met in the analysis of stock prices, it will be the case that $\rho = 1$. However, we shall shortly argue on why $\rho = 1$, i.e. the unit root case and cointegration, is not examined further in this thesis. When $\rho = 1$, it follows that the AR(1) process with $C_0 = C$ fixed, ε_t i.i.d. $\mathcal{N}(0, \sigma_\varepsilon^2)$,

$$C_t = C_{t-1} + \varepsilon_t = \sum_{i=1}^{t} \varepsilon_i + C. \tag{29}$$

In other words, C_t is the sum of a random walk $\sum_{i=1}^{t} \varepsilon_i$ and the initial value C. When $\rho = 1$, x_t is not stationary, not even assymptotically. This can easily be seen by using the fact that ε_t i.i.d. $\mathcal{N}(0, \sigma_\varepsilon^2)$, to see that the variance of C_t in Equation 29 is given by

$$\mathbb{V}\left[C_t\right] = \mathbb{V}\left[\sum_{i=1}^{t} \varepsilon_i + C\right] = t\sigma_\varepsilon^2,$$

which is increasing in t. Furthermore, note that

$$\Delta C_t = C_t - C_{t-1} = \sum_{i=1}^{t} \varepsilon_i + C - \sum_{i=1}^{t-1} \varepsilon_i + C = \varepsilon_t,$$

implying that the differenced process is stationary. Unit root analysis provides a framwork to discriminate between the pair: The former is a non-stationary random walk case (the hypothesis of non-stationarity) and the latter a stationary case (the hypothesis of stationarity).

It is now apparent why the raw S&P 500 data was transformed to returns. However, note that we assumed that the state process is modelled through an autoregressive process of order 1 and not the asset prices.

A unit-root state behaves like a random walk, drifting without pullback. This destroys a stable baseline for "high/low" regimes and undermines interpretability. It also blurs identification (level shifts can be traded between the intercept and the state), and the state's uncertainty grows with sample length, so there is no steady-state signal-to-noise. If the state feeds the mean or (especially) the variance, implied moments can blow up over time. For stable inference and meaningful regimes, we therefore impose $|\rho| < 1$. Furthermore, as the state-process is unobserved/hidden/latent, we have no means of testing for the hypothesis of non-stationarity for the state-process but it is possible for the asset price process.

2.3.2 Likelihood Formulation & Parameter Estimation

We consider the basic SSM in which the state process is univartiate. Such a SSSM is characterized by two prcoesses (almost identical to that of the discrete-valued hidden Markov model):

1. A continuous-valued hidden Markov state process, $\{C_t\}_{t\in\mathbb{N}}$.
2. A observed process, $\{X_t\}_{t\in\mathbb{N}}$, whose realizations are assumed to be conditionally independent, given the states.

Formally, the assumptions are:

$$\begin{aligned} f_{C_t|\mathbf{C}^{(t-1)}}\left(c_t \mid \mathbf{c}^{(t-1)}\right) &= f(c_t \mid c_{t-1}), \quad t = 2, 3, \ldots, \\ f_{X_t|\mathbf{X}^{(t-1)},\mathbf{C}^{(t)}}\left(x_t \mid \mathbf{x}^{(t-1)}, \mathbf{c}^{(t)}\right) &= f(s_t \mid c_t), \quad t \in \mathbb{N}, \end{aligned}$$

where f is either a density or a probability. The only difference in models between that of a HMM and a SSM is that the Markov process $\{C_t\}$ is continuous-valued in the latter. However, As we will discretize the state space into a sufficiently large but finite number of states, we can evaluate an approximation of the likelihood of any given SSM by using the forward algorithm described in section ??, exactly like the discrete-valued HMM.

The discretization procedure is as follows: For some given SSM, we consider an essential range $[b_0, b_m]$ of possible values of C_t. This range is then subdivided into m subintervals $B_i = (b_{i-1}, b_i)$, $i = 1, \ldots, m$. These subintervals need not be on a equidistant grid, however, for simplicity and computational ease, we assume they are equidistant. As such, they are all of the length $h = (b_m - b_0)/m$. Denote b_i^* a representative point in B_i, for example as we shall choose, the midpoint . By making use of the SSM dependence strucutre and repeatedly apprixmating integrals $\int_a^b f(c)dc$ by simple expressions of the form $(b-a)f(c^*)$, the likelihood of the observations $x_1, \ldots, x_T$ can be approximated as follows

$$\begin{aligned} \mathcal{L}_T &= \underbrace{\int \ldots \int}_{T-\text{integrals}} f_{\mathbf{X}^{(T)},\mathbf{C}^{(T)}}(\mathbf{x}^{(T)}, \mathbf{c}^{(T)}) \mathrm{d}c_T \ldots \mathrm{d}c_1 \\ &\overset{\dagger}{=} \int \ldots \int f_{\mathbf{X}^{(T)}|\mathbf{C}^{(T)}}(\mathbf{x}^{(T)} \mid \mathbf{c}^{(T)}) f_{\mathbf{C}^{(T)}}(\mathbf{c}^{(T)}) \mathrm{d}c_T \ldots \mathrm{d}c_1 \\ &\overset{\dagger\dagger}{=} \int \ldots \int f_{C_1}(c_1) f_{X_1|C_1}(x_1 \mid c_1) \prod_{t=2}^{T} f_{C_t|C_{t-1}}(c_t \mid c_{t-1}) f_{X_t|C_t}(x_t \mid c_t) \mathrm{d}c_T \ldots \mathrm{d}c_1 \qquad (30) \\ &\overset{\dagger\dagger\dagger}{\approx} \int_{b_0}^{b_m} \ldots \int_{b_0}^{b_m} f_{C_1}(c_1) f_{X_1|C_1}(x_1 \mid c_1) \prod_{t=2}^{T} f_{C_t|C_{t-1}}(c_t \mid c_{t-1}) f_{X_t|C_t}(x_t \mid c_t) \mathrm{d}c_T \ldots \mathrm{d}c_1 \\ &\overset{\dagger\dagger\dagger\dagger}{\approx} h^T \sum_{i_1=1}^{m} \ldots \sum_{i_T=1}^{m} f_{B_{i_1}}(b_{i_1}^*) f_{X_1|B_{i_1}}(x_1 \mid b_{i_1}^*) \prod_{t=2}^{T} f_{B_{i_t}|B_{i_{t-1}}}(b_{it}^* \mid b_{it-1}^*) f_{X_t|B_{i_t}}(x_t \mid b_{it}^*). \end{aligned}$$

† follows from definition of joint probability, †† is simply rewriting into a product, † † † is splitting the integrals into the essential range and † † † † from the approximation, where the innermost

integral has been approximated as follows:

$$\int_{b_0}^{b_m} f_{C_T|C_{T-1}}(c_T \mid c_{T-1}) f_{X_T|C_T}(x_T \mid c_T) \mathrm{d}c_T \approx h \sum_{i_T}^{m} f_{B_{i_T}|C_{T-1}}(b_{i_T}^* \mid c_{T-1}) f_{X_T|B_{i_T}}(x_T \mid b_{i_T}^*).$$

The terms appearing in the approximation in Equation 30 are simple, however, the likelihood cannot be evaluated as is because of the extremely high number of summands (m^T). However, the likelihood with a discrete state space, as the approximation is, yields a convenient form which allows us to employ our previously developed technique of using the forward algorithm to evaluate the likelihood.

This is not the only way to discretize the likelihood and there is an infinite number of ways to approach the problem of discretization. We choose a simple midpoint quadrature as it is computationally efficient to implement.

Evaluation of the Approximate Likelihood As stated, the discretization of the state space into some large number of intervals m corresponds to an approximation of the SSM by an m-state HMM. However, it is now possible to specify the components of this approximating HMM with ease. First, Consider the initial distribution of the state process. To obtain the exact expressions given in the last line of Equation 30, we define the i'th component of the m-dimensional vector $\boldsymbol{\delta}$ to be $\delta_i = hf(b_i^*)$. then δ_i is the approximate probability of the state process fallin in the interval B_i at time 1 (as it is the initial distribution). For example, assume that the state process is in its stationary distribution at the time of the first observation. Then $f(b_i^*)$ is the density of the normal distribution evaluated at b_i^* with parameters

$$\mathbb{E}[C_t^*] = 0,$$
$$\mathbb{V}[C_t^*] = \frac{\sigma_\varepsilon^2}{(1-\rho^2)}.$$

In the exact same manner, define an $m \times m$ t.p.m $\boldsymbol{\Gamma} = (\gamma_{ij})_{i,j=1}^m$ by specifying $\gamma_{ij} = hf(b_j^* \mid b_i^*)$. The transition probabilities γ_{ij} are the approximate probability of the value of the state process falling into the intervals B_j at time t given that the process is in interval B_i at time $t-1$. For the Gaussian AR(1) state process, the values of γ_{ij} is h times the density of the normal distribution with mean ρb_i^* and variance σ_ε^2 evaluated at b_j^*. Lastly, we define the component $\boldsymbol{P}(x_t)$ to be the $m \times m$ diagonal matrix with ith entry corresponding ot $f(x_t \mid b_i^*)$. This corresponds to an approximation of the conditional density of x_t given that the state process takes some value in the interval B_i at time t.

Assembling the components just defined, we can rewrite the multiple-sum expression for the apprximate likelihood given in Equation 30 in the form of a matrix product

$$h^T \sum_{i_1=1}^{m} \cdots \sum_{i_T=1}^{m} f_{B_{i_1}}(b^*_{i_1}) f_{X_1|B_{i_1}}(x_1 \mid b^*_{i_1}) \prod_{t=2}^{T} f_{B_{i_t}|B_{i_{t-1}}}(b^*_{i_t} \mid b^*_{i_{t-1}}) f_{X_t|B_{i_t}}(x_t \mid b^*_{i_t})$$
$$= \boldsymbol{\delta} \boldsymbol{P}(x_1) \boldsymbol{\Gamma} \boldsymbol{P}(x_2) \boldsymbol{\Gamma} \boldsymbol{P}(x_3) \cdots \boldsymbol{\Gamma} \boldsymbol{P}(x_{T-1}) \boldsymbol{\Gamma} \boldsymbol{P}(x_T) \mathbf{1}_N^\top.$$

This matrix product, the computational effort required to evaluate the approximate likelihood is linear in the number of observations T and quadratic in the number of intervals used in the discretization m, exactly as the HMMs.

Estimation Issues and Assessment According to [58, p. 160], numerical maximization of the likelihood given in equation Equation 30 is feasible even when the observation count and m is fairly large, which is what would be required for a relatively close approximation to the likelihood. In general, it seems to be that values around $m = 50$ stabilize [58, 26]. Furthermore, as can easily be seen, the number of parameters does not depend on the magnitude of m. The entries of the approximate tpm $\boldsymbol{\Gamma}$, which we remind is a $m \times m$-matrix, depends only on state process parameters of the SSM. The range $[b_0, b_m]$ has to be chosen such that it is sufficiently large to cover the essential range of the state proces. However, if it is chosen too large, the do not maintain sufficient fineness of the grid. Assessing the fitting of the SSM by examining marginal distributions does indicate whether the chosen range is indeed large enough.

The other issues remain the same as for the N-state HMM as we are essentially fitting a m-state hidden Markov model; Local maxima, parameter constraints, and especially, numerical under- and overflow. We use the exact same techniques as described for the HMM to deal with the latter issues of numerical under- and overflow.

As for the HMM, we extract the numerically estimated Hessian of the log-likelihood for the estimated parameters using the base R function `nlm`. Furthermore, we can use the Viterbi algorithm for decoding, pseudo-residuals and forecasts, exactly as for the HMM.

Simulating We simulate the BS-SSM. That is, we simulate the BSM with an AR(1) latent state $C_t = \rho C_{t-1} + \varepsilon_t$ with ε i.i.d. $\mathcal{N}(0, \sigma_\varepsilon^2)$ that multiplicatively scales both drift and volatility in the log-return formulation X_t (cf. Equation 4). Parameters are $\mu = 0.05$, $\sigma = 0.15$, $\rho = 0.98$ to avoid stationarity, $\sigma_\varepsilon = 0.10$, with $S_0 = 100$, $n = 25000$, and daily observations $\Delta = 1/252$ (approximately $25000/252 \approx 99$ years). The simulated price path is shown in Figure 12. Estimated vs. true parameters are reported in Table 3. We used `nlm` in `R` to maximize the likelihood given in Equation 30 for the BS-SSM.

Noting the Euler schme discritization error, the relative estimation errors are small and do not raise any concerns. However, note that the error induced by the discretization might or might not reduce the estimation error, depending on which direction the induced discretization bias is.

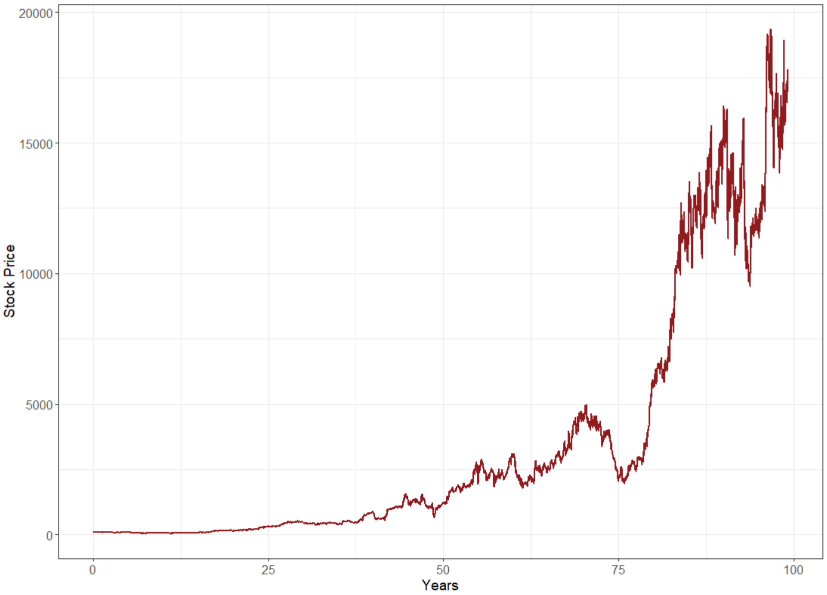


Figure 12: *A simulated stock price path S_t in the BS–SSM with AR(1) latent scale.*

Parameter	True Value	Estimated Value	Relative Error (%)
ρ	0.98000	0.98039	0.04
σ_ε	0.10000	0.09967	0.33
μ	0.05000	0.05062	1.24
σ	0.15000	0.14653	2.31

Table 3: *True vs. estimated parameters for the BS–SSM, including relative estimation errors.*

2.4 Model Selection Criteria & Assessment

2.4.1 Information Criteria: AIC & BIC

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

Assume that $x_1, \ldots, x_T$ were generated by the true 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, $\widehat{\mathbb{E}}_f\left[\Delta(f, \hat{g}_1)\right]$ and $\widehat{\mathbb{E}}_f\left[\Delta(f, \hat{g}_2)\right]$. These quantities bases selection on estimators of the expected discrepancies. The model selection criterion simplifies to the Akaike information criterion [58, p. 98] which, in (dangerously) short, arises of the Kullback–Leibler discrepancy and conditions listed in [28, Appendix A]:

$$\text{AIC} = \underbrace{-2\log \mathcal{L}_T}_{\text{measure of fit}} + \underbrace{2p}_{\text{penalty}}, \tag{31}$$

where $\mathcal{L}$ is the log-likelihood of the fitted model and p denotes the number of parameters of the model[8]. 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 [9, 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 [10] 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 2.2.7.

Another approach to model selection is the Bayesian philosopy. 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, $\mathbb{P}\left(f \in \mathcal{G}_1 \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}\right)$ and $\mathbb{P}\left(f \in \mathcal{G}_2 \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}\right)$. Again, in (dangerously) short, under conditions seen in [55], the Bayesian information criterion arises [58, p. 98]:

$$\text{BIC} = \underbrace{-2\log \mathcal{L}_T}_{\text{measure of fit}} + \underbrace{p\log T}_{\text{penalty}}, \tag{32}$$

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,

[8] see Section 2.2.7 for number of parameter determination.

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 31 and Equation 32).

2.4.2 Pseudo-Residuals

Even after selecting what seems to be the best model according to some chosen criterion, it is still necessary to determine whether the model actually provides a good fit to the data. This requires tools that can assess the overall adequacy of the model and help identify potential outliers. In classical settings such as normal-theory regression, residuals are a well-known and widely used method for checking model fit. In this section, we introduce quantities called pseudo-residuals (also referred to as quantile residuals), which extend this idea to more general models and serve a similar purpose in the context of HMMs. We present two types of pseudo-residuals, both of which rely on the ability to compute likelihoods efficiently, something that HMMs naturally allow. The theory presented is based on that of [15, pp. 236-244] which they note is a special case of Cox–Snell residuals [13].

Each X_t has a distribution that depends on some latent state in the state space. As such, assessing outliers or model fit is non-trivial, since the conditional distribution of each X_t changes over time and depends on the hidden state sequence. A commonly used approach in HMMs for assessment of model fit is to transform the observations to a common scale using pseudo-residuals $\{z_t\}_{t=1}^{T}$, constructed via the probability integral transform [58, pp. 101–106]:

I. Transform a observation x_t to $u_t = F_{X_t}(x_t) \sim \mathcal{U}[0,1]$, where F_{X_t} is the CDF of X_t.

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, $z_t = \Phi^{-1}\left(F_{X_t}(x_t)\right)$, should be approximately independent and standard normally distributed. These can be evaluated using histograms and Q–Q plots.

We show the properties in **I.** and **II.** to be the case.

Proposition 2.8. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X : \Omega \to \mathbb{R}$ be a real-valued r.v. with cumulative distribution function F_X. Suppose F_X is continuous and strictly increasing (hence invertible with inverse F_X^{-1}). Define*

$$U := F_X(X) \qquad \textit{and} \qquad Z := \Phi^{-1}(U),$$

where Φ is the standard normal distribution function. Then $U \sim \mathcal{U}[0,1]$ and $Z \sim \mathcal{N}(0,1)$.

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

$$\begin{aligned}\mathbb{P}\left(U \leq u\right) &= \mathbb{P}\left(F_X\left(X\right) \leq u\right)\\ &\overset{\dagger}{=} \mathbb{P}\left(X \leq F_X^{-1}\left(u\right)\right)\\ &= F_X\left(F_X^{-1}\left(u\right)\right)\\ &= u,\end{aligned}$$

where † follows since F_X is strictly increasing. Next, for any $z \in \mathbb{R}$,

$$\begin{aligned}F_Z\left(z\right) &= \mathbb{P}\left(Z \leq z\right)\\ &= \mathbb{P}\left(\Phi^{-1}\left(U\right) \leq z\right)\\ &= \mathbb{P}\left(U \leq \Phi\left(z\right)\right)\\ &= F_U\left(\Phi\left(z\right)\right)\\ &= \Phi\left(z\right).\end{aligned}$$

Therefore $Z \sim \mathcal{N}(0,1)$. □

Ordinary Pseudo-Residuals The first approach examines each observation individually, identifying those that appear unusually extreme relative to the model and the remaining data in the series, indicating that they may differ in nature or origin. In practice, this involves computing a pseudo-residual $\{z_t\}_{t=1}^{T}$ from the conditional distribution of $X_t \mid \mathbf{X}^{(-t)}$. For continuous observations the normal pseudo-residual is

$$z_t = \Phi^{-1}\left(\mathbb{P}\left(X_t \leq x_t \mid \mathbf{X}^{(-t)} = \mathbf{x}^{(-t)}\right)\right). \tag{33}$$

z_t is approximately a realization of a standard normal r.v., if the model is correctly specified. Using results from Section 2.2.6, specifically by integrating over Equation 22 and Equation 23, the ordinary pseudo-residuals can be calculated. However, we will use forecast pseudo-residuals for the assessment of models. This is because we aren't inherently interested in idiosyncratic outliers relative to the model. However, we are interested in the models predictive powers. As such, forecast pseudo-residuals are of advantageous use.

Forecast Pseudo-Residuals The second approach to detecting outliers focuses on identifying observations that appear unusually extreme when compared to what the model predicts based on all previous observations, rather than the entire data sequence. Here, the key quantity is the conditional distribution of X_t given $\mathbf{X}^{(t-1)}$. For continuous observations the pseudo-residual is

$$z_t = \Phi^{-1}\left(\underbrace{\mathbb{P}\left(X_t \leq x_t \mid \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)}\right)}_{(*)}\right). \tag{34}$$

To compute the forecast pseudo-residuals, we evaluate the one-step-ahead forecast distribution of X_t given the observed history up to time $t-1$ by examining $(*)$ in Equation 34. Note that

$$\begin{aligned}\mathbb{P}\left(X_t \leq x_t \mid \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)}\right) &= \sum_{j\in\mathcal{C}} \mathbb{P}\left(X_t \leq x_t, C_t = j \mid \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)}\right) \\ &\overset{\dagger}{=} \sum_{j\in\mathcal{C}} \underbrace{\mathbb{P}\left(X_t \leq x_t \mid \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)}, C_t = j\right)}_{:=F_{X_t,j}(x_t)} \underbrace{\mathbb{P}\left(C_t = j \mid \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)}\right)}_{:=\psi_t(j)} \\ &\overset{\dagger\dagger}{=} \sum_{j\in\mathcal{C}} \psi_t(j) F_{X_t,j}(x_t),\end{aligned}$$

where † follows from the Law of Total Probability and †† from the HMM conditional-independence assumption. In short, in the HMM setting, this forecast distribution is a mixture of state-dependent conditional distributions with $\psi_t(j) = \mathbb{P}\left(C_t = j \mid \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)}\right)$ the one-step-ahead predicted state probabilities and for $q, p \in \mathbb{R}$

$$F_{X_t,j}(x_t) = \Phi\left(\frac{x_t - m_j}{s_j}\right),$$

with $m_i := \left(\mu_i - \frac{1}{2}\sigma_i^2\right)\Delta$ and $s_i^2 = \sigma_i^2\Delta$. The predicted state probabilities are obtained by propagating them forward using the transition probabilities and the normalized vector of forward variables

$$\begin{aligned}\psi_t(j) &= \mathbb{P}\left(C_t = j \mid \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)}\right) \\ &= \sum_{i\in\mathcal{C}} \underbrace{\mathbb{P}\left(C_t = j \mid C_{t-1} = i\right)}_{:=\gamma_{i,j}} \underbrace{\mathbb{P}\left(C_{t-1} = i \mid \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)}\right)}_{:=\phi_{t-1}(j)} \\ &= \sum_{i\in\mathcal{C}} \gamma_{ij}\phi_{t-1}(i) \\ &\Rightarrow \\ \boldsymbol{\psi}_t &= \boldsymbol{\phi}_{t-1}\boldsymbol{\Gamma}.\end{aligned}$$

For numerical stability we use exponentiated normalized forward probability vectors. Specifically, in the BS-HMM, the state-dependent cumulative distributions $F_{X_t,j}(x_t)$ are governed by the

Gaussian distribution

$$X_t \mid \{C_t = i\} \sim \mathcal{N}\left(\left(\mu_i - \frac{1}{2}\sigma_i^2\right)\Delta, \sigma_i^2\Delta\right).$$

The pseudo-residuals are then given by

$$z_t = \Phi^{-1}\left(\sum_{j\in\mathcal{S}} \psi_t(j) \cdot F_{X_t,j}(x_t)\right), \quad t = 2, \dots, T.$$

For the first residual z_1, the one-step-ahead state probabilities $\psi_1(j)$ cannot be computed from previous forward probabilities, since there is no observation before $X_1 = x_1$. A convinient circumvention for this complication is to approximate them using the stationary distribution $\boldsymbol{\delta}$ which represents the long-run state probabilities of the Markov chain. Thus, the first residual is computed as

$$z_1 = \Phi^{-1}\left(\sum_{j\in\mathcal{S}} \delta_j \cdot F_{X_t,j}(x_t)\right).$$

If the model is correctly specified, the pseudo-residuals $\{z_t\}_{t=1}^T$ should be approximately independent and standard normally distributed.

3 Data (II of II)

Throughout the empirical analysis we work with logarithmic returns rather than price levels. Let S_t denote the closing level of the S&P 500 at trading day t. The one–day log return over some fixed time horizon $t = 1, 2, \dots, T$ is

$$X_t = \log S_t - \log S_{t-1} = \log\left(\frac{S_t}{S_{t-1}}\right) \qquad (35)$$

This transformation is standard in financial econometrics for several reasons.

First, price levels for broad equity indices are well known to be non–stationary and to exhibit unit–root behavior over long samples, while their first differences, i.e. returns, are much closer to weak stationarity. Stationarity (or approximate stationarity) is a prerequisite for a wide class of likelihood-based time–series methods (ARMA, GARCH, and state–space models). Working in returns mitigates spurious regression phenomena associated with trending levels and yields series with stable mean near zero and finite variance, albeit with conditional heteroskedasticity. However, we again note, we never state that we model the raw S&P 500 data as an autoregressive process but rather log-normal as we use the BSM.

Second, log returns aggregate additively across time. For any horizon $h \in \mathbb{N}$,

$$X_{t:t+h} = \log\left(\frac{S_{t+h}}{S_t}\right) = \sum_{j=1}^{h} X_{t+j}. \qquad (36)$$

Additivity greatly simplifies multi–horizon likelihoods, forecasting, and decomposition of long–horizon performance into daily contributions. By contrast, simple (arithmetic) returns compound multiplicatively. For high–frequency horizons where $|X_t| = (S_t - S_{t-1})/S_{t-1}$ is small, $\log(1 + X_t) \approx X_t$, so log and simple returns coincide to first order while preserving exact additivity in Equation 36.

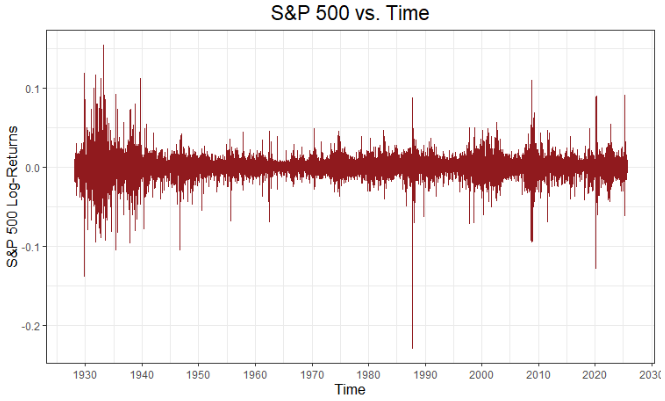


Figure 13: *S&P 500 index time series of log-returns.*

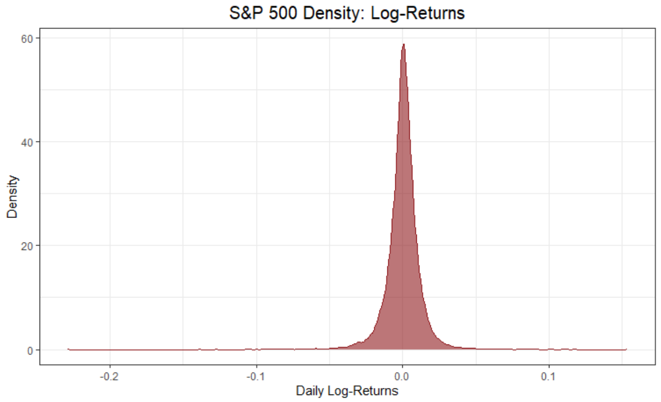


Figure 14: *Kernel density of S&P 500 log-returns.*

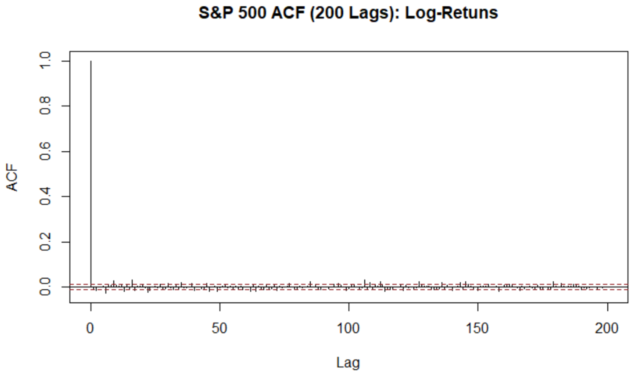


Figure 15: *Autocorrelation of S&P 500 log-returns (200 lags).*

Third, log returns are scale–free and invariant to rescaling of the numeraire. If prices are multiplied by any constant $c > 0$ (e.g., index rebasings), the difference of logs in Equation 35 is unchanged. This facilitates cross–asset and cross–period comparisons and stabilizes variance relative to the price level.

Finally, using log returns improves numerical stability in estimation. Likelihood functions built on price levels inherit the explosive scale and collinearity of S_t, whereas those in differences avoid ill–conditioning and reduce sensitivity to arbitrary index base values.

To examine the data after the transformation we examine the same descriptive statistics and plots. The time series plot can be seen in Figure 13. As claimed, the data are now scale-free and returns do not compound multiplicatively as with the non-transformed data.

Figure 14 shows the density of the log-returns. It is apparent that the data are unimodal, symmetric, and approximately normal with low variance and approximately zero mean. Indeed, the first quartile is -0.0045519, the median is 0.0004940, and the third quartile is 0.005458. The largest observation is 0.1536613 and the smallest is -0.2289973.

Including 200 lags (200 closing trading days), the S&P 500 log-returns ACF is shown in Figure 15. The autocorrelation is notably reduced by the simple transformation.

Because of the issues related to raw prices and the advantages of using the transformed data, we use log-returns for the remainder of the thesis, but we do report MLEs on both raw and transformed data for comparison.

Estimating a State-Wise Dividend Yield $\widehat{q}_i$ As detailed in Section 1, we construct a daily log-dividend series

$$q_t^{(\log)} \;=\; \log\frac{T_t}{T_{t-1}} \;-\; \log\frac{P_t}{P_{t-1}},$$

from total-return T_t and price P_t. As the most likely sequence of states are most likely not always consecutive, i.e. that all states-i are ordered before states-j, $i \neq j$ and $i, j \in \mathcal{C}$, we can't simply estimate a state-wise dividend yield from consecutive observations. As such, we resort to using the Viterbi algorithm to find the most probable state sequence $(i_1^*, \ldots, i_T^*)$. From this state sequence, we find the annualized, constant state-wise dividend yield by summing over estimated daily dividends in each (identical) state, and multiply by 252. In other words, let $\mathcal{T}_i := \{t \in \{2, \ldots, T\} : i_t^* = i\}$ and $N_i := |\mathcal{T}_i|$. Then

$$\widehat{q}_i = 252 \cdot \frac{1}{N_i} \sum_{t \in \mathcal{T}_i} q_t^{(\log)}.$$

For each state we report the pair governing price dynamics and the total-return counterpart:

$$\widehat{\mu}_{\text{cap},i},\; \widehat{\sigma}_i \quad \text{and} \quad \widehat{\mu}_{\text{total},i} = \widehat{\mu}_{\text{cap},i} + \widehat{q}_i\ .$$

Dividend Estimator Standard Errors The estimator $\widehat{q}$ is the annualised sample mean of the daily log dividend yields $\{q_t^{(\log)}\}_{t\in\mathcal{T}}$,

$$\bar{q}^{(\log)} \;:=\; \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} q_t^{(\log)}, \qquad \widehat{q} \;=\; 252\,\bar{q}^{(\log)}.$$

We model $\left\{q_t^{(\log)}\right\}$ as a strictly stationary and ergodic time series with finite $2+\delta$ moment for some $\delta > 0$ and absolutely summable autocovariances

$$\Psi_h \;:=\; \mathrm{Cov}\big(q_t^{(\log)},\, q_{t+h}^{(\log)}\big), \qquad \sum_{h=-\infty}^{\infty} |\Psi_h| < \infty.$$

Define the long–run variance

$$\Omega_q \;:=\; \sum_{h=-\infty}^{\infty} \Psi_h \;=\; \Psi_0 + 2\sum_{h=1}^{\infty} \Psi_h.$$

Under these conditions, a central limit theorem for stationary, weakly dependent time series (see, e.g., [2, 22]) implies

$$\sqrt{|\mathcal{T}|}\left(\bar{q}^{(\log)} - \mathbb{E}\left[q_t^{(\log)}\right]\right) \xrightarrow{d} \mathcal{N}\big(0, \Omega_q\big).$$

Since $\widehat{q} = 252\,\bar{q}^{(\log)}$, we obtain

$$\sqrt{|\mathcal{T}|}\,(\widehat{q} - q) \xrightarrow{d} \mathcal{N}\left(0,\, 252^2\,\Omega_q\right)$$

where the asymptotic variance is $\frac{252^2\,\Omega_q}{|\mathcal{T}|}$, and where $q := 252\,\mathbb{E}\left[q_t^{(\log)}\right]$ is the continuous annualised dividend yield.

In practice, the long–run variance Ω_q is unknown and is replaced by a heteroskedasticity– and autocorrelation–consistent (HAC) estimator of the so called Newey–West type [37, 3]. Let $\hat{\Psi}_h$ denote the sample autocovariance of $q_t^{(\log)}$ at lag h, and let $\{w_h\}_{h=0}^{H_{|\mathcal{T}|}}$ be kernel weights (e.g., Bartlett) with bandwidth $H_{|\mathcal{T}|} \to \infty$ and $H_{|\mathcal{T}|}/|\mathcal{T}| \to 0$. The HAC estimator of the long–run variance is

$$\widehat{\Omega}_q \;=\; \hat{\Psi}_0 + 2\sum_{h=1}^{H_{|\mathcal{T}|}} w_h\,\hat{\Psi}_h,$$

which is consistent for Ω_q under the same weak–dependence conditions [37, 3]. A consistent estimator of the asymptotic variance of $\widehat{q}$ is then $\frac{252^2}{|\mathcal{T}|}\,\widehat{\Omega}_q$ and $\widehat{\mathrm{SE}} = (\widehat{q})252\sqrt{\frac{\widehat{\Omega}_q}{|\mathcal{T}|}}$, and standard Wald confidence intervals for q follow.

4 Empirical Data Application

5.1 Model Selection & Assessment

In what follows, let p denote the number of estimated parameters of a given model.

The Black-Scholes Model The model criteria for the BSM is seen in Table 4 and the residuals in Figure A.3.1.

Black-Scholes Model (BSM)			
Model	p	**AIC**	**BIC**
BSM	2	-139353.758	-139337.662

Table 4: *AIC and BIC for the standard BSM.*

It is quite evident that the BSM yields extremely heavy tails. Heavy tails imply that large positive and negative returns occur with a much higher frequency than predicted by the Gaussian assumption in the BSM, so the model severely underestimates tail risk. Specifically, we notice from the time-series plot that the overly large residuals stem from:

- The Wall Street crash of 1929 and recession of 1937-1938.
- The early 1980s recession and Black Monday 1987.
- The Dot-com Bubble late 1990s & early 2000.
- The 2008 Financial Crisis.

The autocorrelation for the returns is mostly unchanged and the model does still largely over- and underestimate returns outside of economical crises. These findings suggest that the BSM does not capture extreme economical environments adequately.

Hidden Markov Models It is evident from Table 5 that the best performing model in terms of the model information criteria, AIC and BIC, was the 5-state BS-HMM where we allow for state-dependency of the variables σ and μ. Secondly is the 5-state BS-HMM where σ is allowed state-dependency but μ is state-independent. Thirdly, and last up for consideration, is the 4-state BS-HMM where we allow for state-dependency of the variables σ and μ. In figure Figure A.3.2 we examine the state-dependent density plots to examine for any unreasonable fittings that would point towards overfitting issues.

Black-Scholes Hidden Markov Model (BS-HMM)			
Model	$\boldsymbol{p}$	**AIC**	**BIC**
2-state BS-HMM (μ)	5	-139348.0	-139308.0
3-state BS-HMM (μ)	10	-139338.0	-139257.0
4-state BS-HMM (μ)	17	-139324.0	-139187.0
5-state BS-HMM (μ)	26	-139306.0	-139097.0
2-state BS-HMM (σ)	5	-150725.0	-150685.0
3-state BS-HMM (σ)	10	-152591.0	-152510.0
4-state BS-HMM (σ)	17	-153156.0	-153019.0
5-state BS-HMM (σ)	26	-153266.0	-153057.0
2-state BS-HMM (σ & μ)	6	-150746.0	-150698.0
3-state BS-HMM (σ & μ)	12	-152613.0	-152517.0
4-state BS-HMM (σ & μ)	20	-153199.0	-153038.0
5-state BS-HMM (σ & μ)	30	**-153342.0**	**-153101.0**

Table 5: *AIC and BIC for fitted HMMs by state-dependent parameter family. Here, p denotes the number of estimated parameters, and the model order (number of states) is indicated in the row label. The globally lowest (best) AIC and BIC are shown in bold.*

Firstly, we examine state-dependent densities seen in Figure A.3.2 for the best performing model in terms of information criteria, 5-state BS-HMM where we allow for state-dependency of the parameters σ and μ. We notice that states 2 and 3 are strikingly similar to state 2 for the 4-state BS-HMM where we allow for state-dependency of the parameters σ and μ, which could suggest a case of overfitting. Furthermore, by examining the proportion of time spend in each state, it is almost exactly the sum of states 2 and 3 in the 4-state model that gives the time spent in state 2 for the 5-state model However, it is advantageous for us to examine parameter estimates as well, which are seen in Table 6.

Parameter	Estimate (95% CI)	Std. Error
$\widehat{\mu}_{\text{cap},1}$	0.2181 (0.1833, 0.2529)	0.01777
$\widehat{\mu}_{\text{cap},2}$	0.6049 (0.4268, 0.7831)	0.09089
$\widehat{\mu}_{\text{cap},3}$	-1.387 (-1.764, -1.010)	0.1925
$\widehat{\mu}_{\text{cap},4}$	-0.06715 (-0.1861, 0.05178)	0.06068
$\widehat{\mu}_{\text{cap},5}$	-0.3986 (-0.9461, 0.1488)	0.2793
$\widehat{\sigma}_1$	0.07022 (0.06755, 0.07288)	0.001359
$\widehat{\sigma}_2$	0.1156 (0.1111, 0.1202)	0.002326
$\widehat{\sigma}_3$	0.1186 (0.1103, 0.1270)	0.004254
$\widehat{\sigma}_4$	0.2314 (0.2218, 0.2410)	0.004892
$\widehat{\sigma}_5$	0.5665 (0.5340, 0.5990)	0.01659

Table 6: *5-state BS-HMM state-dependent parameter estimates with 95% CIs based on the inverse Hessian of the minimized log-likelihood.*

As is now quite evident, the volatility parameter of states 2 and 3 for the 5-state BS-HMM with μ and σ state-dependent, are almost identical for the first 3 digits at $\sigma_2 \approx \sigma_3 \approx 0.11$. This estimate is reasonable but does raise concern by the similarity. We turn our attention to μ_2 and μ_3. Immediately we see that the estimates for μ differ by quite a large margin, which is not a concern. However, both estimates are unreasonable. Especially, $\mu_3 \approx -1.39$ emphasizes that the model for consideration provides unreasonable parameter estimates by overfitting to the data at hand. We therefore move our attention to the 4-state BS-HMM with μ and σ state-dependent. Parameter estimates are seen in Table 7.

Parameter	Estimate (95% CI)	Std. Error
$\widehat{\mu}_{\text{cap},1}$	0.2346 (0.1990, 0.2703)	0.01820
$\widehat{\mu}_{\text{cap},2}$	0.1062 (0.06214, 0.1502)	0.02246
$\widehat{\mu}_{\text{cap},3}$	-0.1031 (-0.2193, 0.01315)	0.05930
$\widehat{\mu}_{\text{cap},4}$	-0.3818 (-0.9260, 0.1623)	0.2776
$\widehat{\sigma}_1$	0.07068 (0.06795, 0.07342)	0.001394
$\widehat{\sigma}_2$	0.1270 (0.1229, 0.1311)	0.002096
$\widehat{\sigma}_3$	0.2301 (0.2205, 0.2396)	0.004870
$\widehat{\sigma}_4$	0.5651 (0.5329, 0.5974)	0.01646

Table 7: *4-state BS-HMM state-dependent parameter estimates with 95% CIs based on the inverse Hessian of the minimized log-likelihood.*

The parameter estimates are reasonable, with no immediate outliers other than some large volatility parameters that could be problematic. We will present these in the next section and why they allow for a reasonable market interpretation.

Continuous State-Space Models As described in Section 2.3.2, there is no exact result in choosing $b_{\max}$ and m but only general experimental knowledge. As such, we firstly calculate the negative log-likelihood using a range and combination of $b_{\max} \in \{0.5, 1, 2, 3, 4\}$ and $m \in \{20, 40, 70, 100, 200\}$. This is done to find the model yielding the smallest negative log-likelihood that also yields reasonable parameter estimates. The results are seen in Table 8. All models marked with a "$*$" yield near identical parameter estimates. Models marked with a "—" yield parameter estimates that are unreasonable and suggest the model is too coarse.

	m				
$b_{\max}$	20	40	70	100	200
0.5	—	—	-74572.223296	-74568.761073	-74566.365696
1	—	-76247.016297	-76243.344459	-76242.462436	-76241.832368
2	—	-76635.707635	-76635.587101	-76635.558157	-76635.537345
3	—	—	**-76637.766845***	-76637.766841*	-76637.766841*
4	—	—	—	-76637.766841*	-76637.766841*

Table 8: *BS-SSM negative log-likelihoods rounded to 6 decimals; the overall minimum is in bold. Dashes indicate runs that didn't converge or were too coarse. A $\star$ marks values with identical 10 first digits (i.e., $-76637.76684x$), indicating they are essentially tied with the best model in negative log-likelihood values.*

Models with values $(b_{\max}, m) \in \{(3, 70), (3, 100), (4, 100), (3, 200), (4, 200)\}$ are therefore adequate candidates.

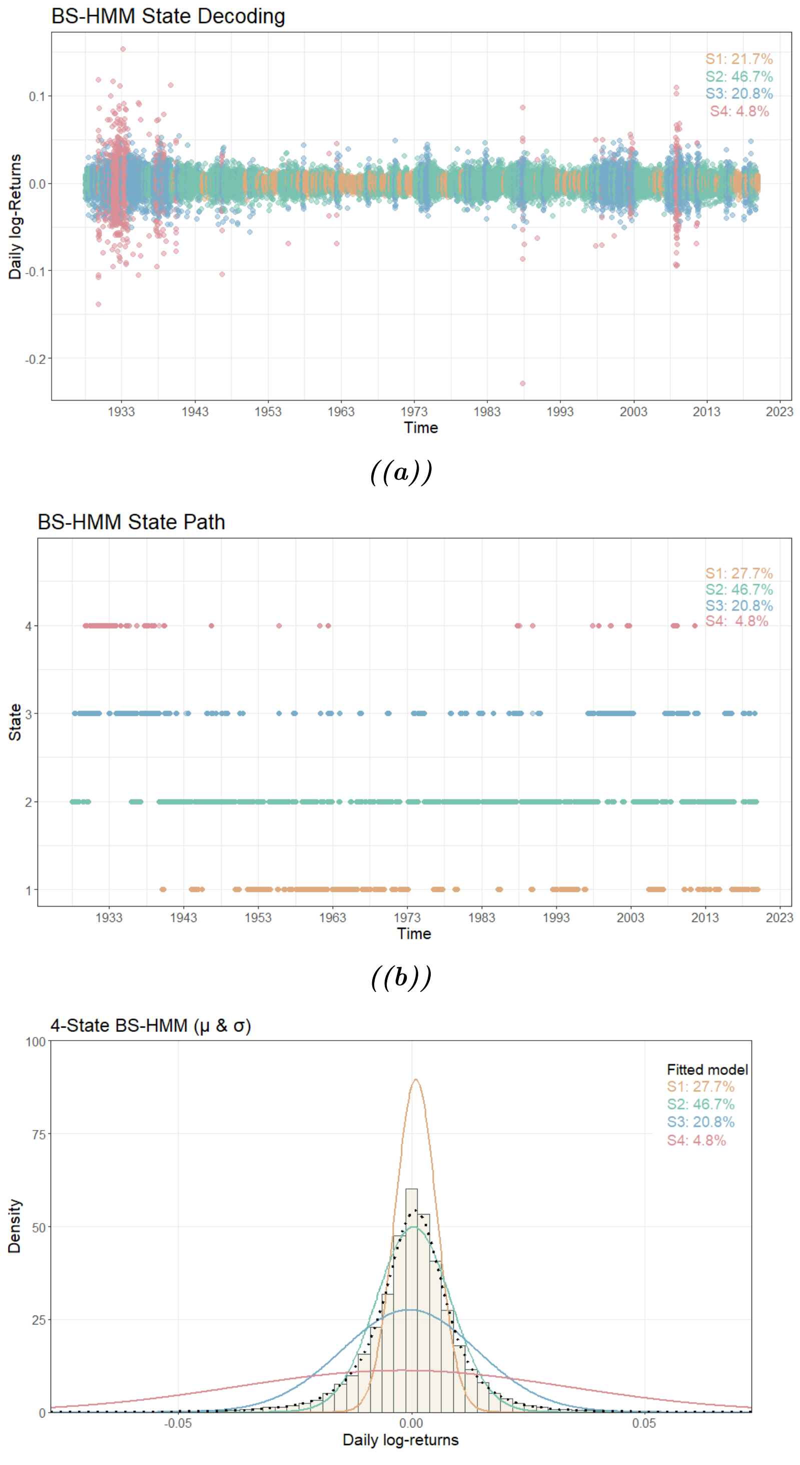
Proceeding to the model assessment we examine pseudo-residuals which can be seen in Figure A.3.7. The QQ-plot shows that the residuals are heavy-tailed and negatively skewed, with a particularly pronounced left tail relative to the assumed normal distribution. This implies that large negative residuals occur more frequently and are more extreme than the model predicts, so it systematically understates downside moves in returns and violates the normality assumption for the innovations. This is mostly similar to the BSM pseudo-residuals. Furthermore, the large residuals are present on the exact same days as the BSM suggesting a lack of fit on extreme changes in economical environments that deviates from the norm. However, a key difference is seen in the size of residuals. The BS-SSMs predicted returns are systematically too high, whereas they are largely symmetrical for the BSM. Pseudo-residuals of magnitude $|5| >$ are not a rare occurance. The large pseudo-residuals imply that one should continue with the model with extreme uncertainty and that conclusions drawn are most likely not adequate. We will discuss why the BS-SSM intuitively is not great for asset pricing in Section 5.

5.2 Model Presentation

Parameter	Estimate (95% CI)	Std. Error
$\widehat{\mu}_{\text{cap}}$	0.07454 (0.03599, 0.1131)	0.01967
$\widehat{q}$	0.03297 (0.03205, 0.03389)	0.0009171
$\widehat{\sigma}$	0.1883 (0.1866, 0.1900)	0.0008761

Table 9: *BSM parameter estimates with 95% CIs based on the inverse Hessian of the minimized log-likelihood.*

TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT
TEXT



((a))

((b))

((c))

Figure 16: *Black–Scholes HMM diagnostics. Panels: (a) decoded state probabilities, (b) most likely state path, and (c) fitted return density.*

Black-Scholes Model

Continuous State Space Model

- Viterbi
- Density plot

Parameter	Estimate (95% CI)		Std. Error
$\widehat{\rho}$	0.9842	(0.9810, 0.9873)	0.001614
$\widehat{\sigma}_\varepsilon$	0.09327	(0.08623, 0.1003)	0.003590
$\widehat{\mu}_{\text{cap}}$	0.1390	(0.1097, 0.1684)	0.01498
$\widehat{\sigma}$	0.1314	(0.1214, 0.1414)	0.005123

Table 10: *Parameter estimates with 95% confidence intervals in parentheses.*

5.3 Prediction

5 Discussion

SSM

- Makes sense it is bad as economic regimes and shocks often happen abruptly
- another continuous state process
- another discretization
- another interval other than simple quadrature
- could be great for gradual regime switches that are obviois (heat prices as seasonal effects happen gradually for example)

6 Conclussion

Bibliography

REFERENCES

[1] Yacine Aït-Sahalia and Jean Jacod. "Testing for Jumps in a Discretely Observed Process." In: *Annals of Statistics* 37.1 (2009), pp. 184–222.

[2] Theodore W. Anderson. *The Statistical Analysis of Time Series.* New York: John Wiley & Sons, 1971.

[3] Donald W. K. Andrews. "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation." In: *Econometrica* 59.3 (1991), pp. 817–858. DOI: 10.2307/2938229.

[4] Ole E. Barndorff-Nielsen and Neil Shephard. "Power and Bipower Variation with Stochastic Volatility and Jumps." In: *Journal of Financial Econometrics* 2.1 (2004), pp. 1–37.

[5] Patrick Billingsley. *Probability and Measure.* 3rd ed. Wiley, 1995.

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

[7] Fischer Black and Myron Scholes. "The pricing of options and corporate liabilities." In: *Journal of political economy* 81.3 (1973), pp. 637–654.

[8] Mark Broadie and Özgür Kaya. "Exact simulation of stochastic volatility and other affine jump diffusion processes." In: *Operations research* 54.2 (2006), pp. 217–231.

[9] 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.

[10] 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.

[11] Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in hidden Markov models.* Springer, 2005.

[12] Kim Christensen, Roel CA Oomen, and Mark Podolskij. "Fact or friction: Jumps at ultra high frequency." In: *Journal of Financial Economics* 114.3 (2014), pp. 576–599.

[13] D. R. Cox and E. J. Snell. "A General Definition of Residuals (with discussion)." In: *Journal of the Royal Statistical Society: Series B (Methodological)* 30 (1968), pp. 248–275.

[14] J. E. Dennis Jr. and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Englewood Cliffs, NJ: Prentice-Hall, 1983.

[15] Peter K. Dunn and Gordon K. Smyth. "Randomized Quantile Residuals." In: *Journal of Computational and Graphical Statistics* 5 (1996), pp. 236–244.

[16] Rick Durrett. *Probability: Theory and Examples.* 5th ed. Cambridge University Press, 2019.

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

[18] Lorella Fatone et al. “The Use of Statistical Tests to Calibrate the Black-Scholes Asset Dynamics Model Applied to Pricing Options with Uncertain Volatility.” In: *Journal of Probability and Statistics* 2012 (2012), Article ID 931609, 20 pages. DOI: 10.1155/2012/931609.

[19] William Feller. *An introduction to probability theory and its applications, Volume 2.* Vol. 2. John Wiley & Sons, 1991.

[20] Jr. Forney G. David. “The Viterbi Algorithm.” In: *Proceedings of the IEEE* 61.3 (1973).

[21] Sylvia Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models.* Springer Series in Statistics. New York: Springer, 2006.

[22] James D. Hamilton. *Time Series Analysis.* Princeton, NJ: Princeton University Press, 1994.

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

[24] Achim Klenke. *Probability Theory: A Comprehensive Course.* 3rd ed. Springer, 2020.

[25] Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations.* Vol. 23. Applications of Mathematics. New York, NY: Springer, 1992.

[26] Roland Langrock, Iain L MacDonald, and Walter Zucchini. “Some nonstandard stochastic volatility models and their estimation using structured hidden Markov models.” In: *Journal of Empirical Finance* 19.1 (2012), pp. 147–161.

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

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

[29] 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.

[30] Brett T. McClintock and Théo Michelot. “momentuHMM: R package for generalized hidden Markov models of animal movement.” In: *Methods in Ecology and Evolution* 9.6 (2018), pp. 1518–1530. DOI: 10.1111/2041-210X.12995. URL: https://doi.org/10.1111/2041-210X.12995.

[31] Geoffrey J McLachlan and David Peel. *Finite mixture models.* John Wiley & Sons, 2000.

[32] Robert C Merton. “An intertemporal capital asset pricing model.” In: *Econometrica: Journal of the Econometric Society* (1973), pp. 867–887.

[33] Robert C. Merton. “Option Pricing When Underlying Stock Returns Are Discontinuous.” In: *Journal of Financial Economics* 3.1-2 (1976), pp. 125–144. DOI: 10.1016/0304-405X(76)90022-2.

[34] Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability.* Springer Science & Business Media, 2012.

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

[36] John F Monahan. *Numerical methods of statistics.* Cambridge University Press, 2011.

[37] Whitney K. Newey and Kenneth D. West. “A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix.” In: *Econometrica* 55.3 (1987), pp. 703–708. DOI: 10.2307/1913610.

[38] Manh Cuong Ngô, Mads Peter Heide-Jørgensen, and Susanne Ditlevsen. “Understanding narwhal diving behaviour using Hidden Markov Models with dependent state distributions and long range dependence.” In: *PLoS computational biology* 15.3 (2019), e1006425.

[39] 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.

[40] Anders Rahbek and Rasmus Søndergaard Pedersen. *Lecture Notes for NMAK24011U Financial Econometric Time Series Analysis (FinMetrics).* Course lecture notes, Department of Mathematical Sciences. Aug. 2024.

[41] Cyrus A. Ramezani and Yong Zeng. “Maximum Likelihood Estimation of the Double Exponential Jump-Diffusion Process.” In: *Annals of Finance* 3.4 (2007), pp. 487–507.

[42] Thomas Ruf, Johannes Schiele, and Maximilian Schuster. *Constructing a Historical S&P 500 Total Return Index.* Discusses public availability of TR data and reconstruction before 1988. 2023. URL: https://www.tidy-finance.org/blog/historical-sp-500-total-return/ (visited on 11/09/2025).

[43] F. W. Scholz. “Maximum likelihood estimation.” In: *Encyclopedia of Statistical Sciences.* Ed. by Samuel Kotz et al. 2nd. Hoboken, NJ: Wiley, 2006, pp. 4629–4639.

[44] Robert J. Shiller. *Data Appendix: U.S. Stock Market (notes and documentation).* Documentation of sources and splicing for price, dividends, and earnings. 2025. URL: https://www.econ.yale.edu/~shiller/data/chapt26.html (visited on 11/09/2025).

[45] Robert J. Shiller. *Online Data: U.S. Stock Market Prices, Dividends, Earnings, and CPI.* Monthly S&P price & dividend series starting in 1871. 2025. URL: https://www.econ.yale.edu/~shiller/data.htm (visited on 11/09/2025).

[46] S&P Dow Jones Indices. *FAQ: S&P 500 Dividend Points Index.* Explains the Dividend Points index and annual reset. 2023. URL: https://www.spglobal.com/spdji/en/documents/additional-material/faq-sp-500-dividend-points-index.pdf (visited on 11/09/2025).

[47] S&P Dow Jones Indices. *Index Mathematics Methodology.* URL: https://www.spglobal.com/spdji/en/methodology/article/index-mathematics-methodology/ (visited on 11/09/2025).

[48] S&P Dow Jones Indices. *Index Mathematics Methodology.* Defines price, total return, and net total return; dividends are reinvested on ex-date. 2024. URL: https://www.spglobal.com/spdji/zh/documents/methodologies/methodology-index-math.pdf (visited on 11/09/2025).

[49] S&P Dow Jones Indices. *Index Mathematics Methodology.* 2025. URL: https://www.spglobal.com/spdji/zh/documents/methodologies/methodology-index-math.pdf (visited on 11/09/2025).

[50] S&P Dow Jones Indices. *S&P U.S. Indices Methodology.* 2025. URL: https://www.spglobal.com/spdji/en/documents/methodologies/methodology-sp-us-indices.pdf (visited on 11/09/2025).

[51] S&P Dow Jones Indices. *S&P U.S. Indices Methodology.* 2025. URL: https://www.spglobal.com/spdji/en/documents/methodologies/methodology-sp-us-indices.pdf (visited on 11/09/2025).

[52] Dag Tjøstheim. “Non-linear time series and Markov chains.” In: *Advances in applied probability* 22.3 (1990), pp. 587–611.

[53] Ingmar Visser and Maarten Speekenbrink. “depmixS4: An R Package for Hidden Markov Models.” In: *Journal of Statistical Software* 36.7 (2010), pp. 1–21. DOI: 10.18637/jss.v036.i07. URL: https://www.jstatsoft.org/v36/i07/.

[54] Andrew J. Viterbi. “Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm.” In: *IEEE Transactions on Information Theory* 13.2 (1967).

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

[56] Yahoo Finance. *S&P 500 (ˆGSPC) — Quote and Historical Data.* Data source. 2025. URL: https://finance.yahoo.com/quote/%5EGSPC/ (visited on 11/09/2025).

[57] Yahoo Finance. *S&P 500 (Total Return) ŜP500TR.* Daily total return index series used alongside ĜSPC. 2025. URL: https://finance.yahoo.com/quote/%5ESP500TR/ (visited on 11/09/2025).

[58] 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.

A.2 Derivations & Proofs

The HMM and SSM We start by defining two key concepts to assist in most of the proofs: a directed acyclical graph and a parent of a r.v.[9].

Definition A.2.1. *Let $G = (V, E)$ be a directed graph, where V is a finite set of vertices (or nodes), and $E \subseteq V \times V$ is a set of directed edges, where an edge $(u, v) \in E$ indicates a directed link from u to v. The graph G is called a* directed acyclic graph (DAG) *if and only if it contains no directed cycles; that is, there do not exist distinct vertices $v_1, v_2, \ldots, v_k \in V$ such that $(v_i, v_{i+1}) \in E$ for all $i = 1, \ldots, k-1$, and $(v_k, v_1) \in E$. Equivalently, G is acyclic if there exists a topological ordering of the vertices $v_1, v_2, \ldots, v_n$ such that $(u, v) \in E \implies$ the index of u is less than that of v.*

Definition A.2.2. *Let $G = (V, E)$ be a directed acyclic graph (DAG), where $V = \{V_1, \ldots, V_n\}$ denotes the set of vertices (r.v.'s) and $E \subseteq V \times V$ denotes the set of directed edges. For a node $V_i \in V$, the* parent set *of V_i is defined as*

$$\mathrm{pa}(V_i) := \{\, V_j \in V : (V_j, V_i) \in E \,\}.$$

That is, V_j is said to be a parent *of V_i if and only if there exists a directed edge from V_j into V_i in the graph.*

The driving tool for any of the proofs is the following factorization for the joint distribution of the set of r.v.'s V_i $i \in \{1, \ldots, N\}$ in a directed acyclic graph

$$f_{\mathbf{V}^{(N)}}\left(\mathbf{v}^{(N)}\right) = \prod_{i=1}^{N} f_{V_i|\mathrm{pa}(V_i)}(v_i \mid \mathrm{pa}(v_i)), \tag{37}$$

where $\mathrm{pa}(V_i)$ denotes all the parents of V_i in the set $\{V_1, V_2, \ldots, V_N\}$. For example, consider our usual hidden Markov model setup such as that in Figure 8. The only parent of X_k is C_k and for $k = 2, 3, \ldots$ the only parent of C_k is C_{k-1} (obviously, C_1 has no parent). As an example, the joint distribution of $\mathbf{X}^{(t)}$ and $\mathbf{C}^{(t)}$ is therefore given by

$$f_{\mathbf{X}^{(t)},C_t}(\mathbf{x}^{(t)}, c) = \mathbb{P}(C_1) \prod_{k=2}^{t} \mathbb{P}(C_k \mid C_{k-1}) \prod_{k=1}^{t} f_{X_k,C_k}(x_k, c_k). \tag{38}$$

Lemma A.2.1. *For $t \in \mathbb{Z}^+$ and histories $\mathbf{X}^{(t)}$ and $\mathbf{C}^{(t)}$ we have that*

$$f_{\mathbf{X}^{(t+1)},C_t,C_{t+1}}(\mathbf{x}^{(t+1)}, c_t, c_{t+1}) = f_{\mathbf{X}^{(t)},C_t}(\mathbf{x}^{(t+1)}, c_t)\mathbb{P}\left(C_{t+1} \mid C_t\right) f_{X_{t+1},C_{t+1}}(x_{t+1}, c_{t+1})$$

[9] Proceeding in the appendix, we use the notation, that a draw of a r.v. (say C) is simply denoted by the lower caps of said r.v. (say c) and not necessarily the usual state values $i, j \in \mathcal{C}$.

Proof. By Equation 38 and the analogous expression for the expression $f_{\mathbf{X}^{(t+1)},\mathbf{C}^{(t+1)}}\left(\mathbf{x}^{(t+1)},\mathbf{c}^{(t+1)}\right)$ imply that

$$f_{\mathbf{X}^{(t+1)},\mathbf{C}^{(t+1)}}\left(\mathbf{x}^{(t+1)},\mathbf{c}^{(t+1)}\right) = \mathbb{P}(C_{t+1} \mid C_t) f_{\mathbf{X}^{(t)},\mathbf{C}^{(t)}}\left(\mathbf{x}^{(t)},\mathbf{c}^{(t)}\right) f_{X_{t+1}|C_{t+1}}(x_{t+1}, c_{t+1})$$

Summing over $\mathbf{C}^{(t-1)}$ yields the desired result. □

Lemma A.2.2. *For $t = 1, 2, \ldots, T-1$,*

$$f_{\mathbf{X}_{t+1}^T|C_{t+1}}(\mathbf{x}_{t+1}^T \mid c_{t+1}) = f_{X_{t+1}|C_{t+1}}(x_{t+1} \mid c_{t+1}) f_{\mathbf{X}_{t+2}}(\mathbf{x}_{t+2})$$

Proof. The result follows by

$$\begin{aligned} f_{\mathbf{X}_{t+1}^T,\mathbf{C}_{t+1}^T}(\mathbf{X}_{t+1}^T, \mathbf{C}_{t+1}^T) &= f_{X_{t+1}|C_{t+1}}(x_{t+1} \mid c_{t+1}) \left(\mathbb{P}\left(C_{t+1}\right) \prod_{k=t+2}^{T} \mathbb{P}\left(C_k \mid C_{k-1}\right) \prod_{k=t+2}^{T} f_{X_k|C_k}(x_k \mid c_k) \right) \\ &= f_{X_{t+1}|C_{t+1}}(x_{t+1} \mid c_{t+1}) f_{\mathbf{X}_{t+2}^T,\mathbf{C}_{t+1}^T}(\mathbf{x}_{t+2}^T, \mathbf{c}_{t+1}^T) \end{aligned}$$

and then summing over $\mathbf{C}_{t+2}^T$ and dividing by $\mathbb{P}\left(C_{t+1}\right)$. □

Lemma A.2.3. *For $t = 1, 2, \ldots, T-1$,*

$$f_{\mathbf{X}_{t+1}^T|C_{t+1}}(\mathbf{x}_{t+1}^T \mid c_{t+1}) = f_{\mathbf{X}_{t+1}^T|C_{t+1},C_t}(\mathbf{x}_{t+1}^T \mid c_t, c_{t+1}). \quad (\dagger)$$

Proof. Simply rewrite the RHS of (†) to

$$\frac{1}{\mathbb{P}(C_t, C_{t+1})} \sum_{\mathbf{C}_{t+2}^T} f_{\mathbf{X}_{t+1}^T,\mathbf{C}_t^T}(\mathbf{x}_{t+1}^T, \mathbf{c}_t^T),$$

which by Equation 37 reduces to

$$\sum_{\mathbf{C}_{t+2}^T} \prod_{k=t+2}^{T} \mathbb{P}(C_k \mid C_{k-1}) \prod_{k=t+1}^{T} f_{X_k|C_k}(x_k \mid c_k).$$

The LHS of (†) is

$$\frac{1}{\mathbb{P}(C_t)} \sum_{\mathbf{C}_{t+2}^T} f_{\mathbf{X}_{t+1}^T,\mathbf{C}_t^T}(\mathbf{x}_{t+1}^T, \mathbf{c}_t^T) = \sum_{\mathbf{C}_{t+2}^T} \prod_{k=t+2}^{T} \mathbb{P}(C_k \mid C_{k-1}) \prod_{k=t+1}^{T} f_{X_k|C_k}(x_k \mid c_k),$$

which show that both sides reduce to the same expression. □

Lemma A.2.4. *For $r = 1, \ldots, T$ and $i_r \in \{1, \ldots, m\}$, the vectors defined by*

$$\alpha_1(i_1) \equiv h(i_1), \qquad \alpha_{r+1}(i_{r+1}) \equiv \sum_{i_r=1}^{m} \alpha_r(i_r)\, f_{r+1}(i_r, i_{r+1})$$

satisfy

$$\alpha_r(i_r) = \sum_{i_1=1}^{m} \cdots \sum_{i_{r-1}=1}^{m} h(i_1) \prod_{t=2}^{r} f_t(i_{t-1}, i_t). \quad (\dagger)$$

Proof. We proceed by induction on r.

Base case ($r = 1$). The product over an empty index set equals 1, so

$$\alpha_1(i_1) = h(i_1) = \sum_{\text{empty}} h(i_1) \cdot 1,$$

which is (†) for $r = 1$.

Inductive step. Assume (†) holds for some $r \in \{1, \ldots, T-1\}$. Then

$$\begin{aligned}\alpha_{r+1}(i_{r+1}) &= \sum_{i_r=1}^{m} \alpha_r(i_r)\, f_{r+1}(i_r, i_{r+1}) \\ &= \sum_{i_r=1}^{m} \left[\sum_{i_1=1}^{m} \cdots \sum_{i_{r-1}=1}^{m} h(i_1) \prod_{t=2}^{r} f_t(i_{t-1}, i_t) \right] f_{r+1}(i_r, i_{r+1}) \\ &= \sum_{i_1=1}^{m} \cdots \sum_{i_r=1}^{m} h(i_1) \prod_{t=2}^{r+1} f_t(i_{t-1}, i_t),\end{aligned}$$

which is (†) with r replaced by $r+1$. By induction, the claim holds for all r, and in particular for $r = T$:

$$\alpha_T(i_T) = \sum_{i_1=1}^{m} \cdots \sum_{i_{T-1}=1}^{m} h(i_1) \prod_{t=2}^{T} f_t(i_{t-1}, i_t).$$

□

Lemma A.2.5. *Let $\mathbf{F}_t$ be the $m \times m$ matrix with (i, j) entry $f_t(i, j)$ and let $\mathbf{1}_N$ denote the $m \times 1$ vector of ones. Then*

$$\mathbb{S} = \sum_{i_1=1}^{m} \cdots \sum_{i_T=1}^{m} h(i_1) \prod_{t=2}^{T} f_t(i_{t-1}, i_t) = \sum_{i_T=1}^{m} \alpha_T(i_T) = \boldsymbol{\alpha}_T \mathbf{1}_N = \boldsymbol{\alpha}_1 \mathbf{F}_2 \mathbf{F}_3 \cdots \mathbf{F}_T \mathbf{1}_N.$$

Proof. The first equality is the definition of $\mathbb{S}$. The second follows from Lemma A.2.4 with $r = T$, which shows $\alpha_T(i_T)$ is the sum over all indices except i_T. Summing over i_T gives $\mathbb{S} = \sum_{i_T} \alpha_T(i_T) = \boldsymbol{\alpha}_T \mathbf{1}_N$. Finally, by construction $\boldsymbol{\alpha}_{r+1} = \boldsymbol{\alpha}_r \mathbf{F}_{r+1}$, hence $\boldsymbol{\alpha}_T = \boldsymbol{\alpha}_1 \mathbf{F}_2 \cdots \mathbf{F}_T$, yielding the displayed formula. □

The AR(1) process

Lemma A.2.6. *We prove a Lemma to help us rewrite the AR(1) process in a more convenient form under certain conditions. With $\rho \in \mathbb{R}$ and $\rho \neq 1$, then*

$$1 + \rho + \rho^2 + \ldots + \rho^n = \sum_{i=0}^{n} \rho^i = \left(1 - \rho^{n+1}\right)/(1 - \rho).$$

If moreover $|\rho| < 1$, $\rho^n \to 0$ as $n \to \infty$, and

$$\sum_{i=0}^{\infty} \rho^i = 1/(1 - \rho)$$

Proof. Let $S_n := \sum_{i=0}^{n} \rho^i$ with $\rho \in \mathbb{R}$ and $\rho \neq 1$. Then

$$(1 - \rho)S_n = \sum_{i=0}^{n} \rho^i - \sum_{i=0}^{n} \rho^{i+1} = \left(1 + \rho + \rho^2 + \ldots + \rho^n\right) - \left(\rho + \rho^2 + \ldots + \rho^{n+1}\right) = 1 - \rho^{n+1}.$$

Hence

$$S_n = \frac{1 - \rho^{n+1}}{1 - \rho},$$

which proves the finite–sum identity. If moreover $|\rho| < 1$, then $|\rho|^n \to 0$ as $n \to \infty$. Taking limits in the identity above yields

$$\sum_{i=0}^{\infty} \rho^i = \lim_{n\to\infty} S_n = \lim_{n\to\infty} \frac{1 - \rho^{n+1}}{1 - \rho} = \frac{1 - 0}{1 - \rho} = \frac{1}{1 - \rho}.$$

□

A.3 Figures

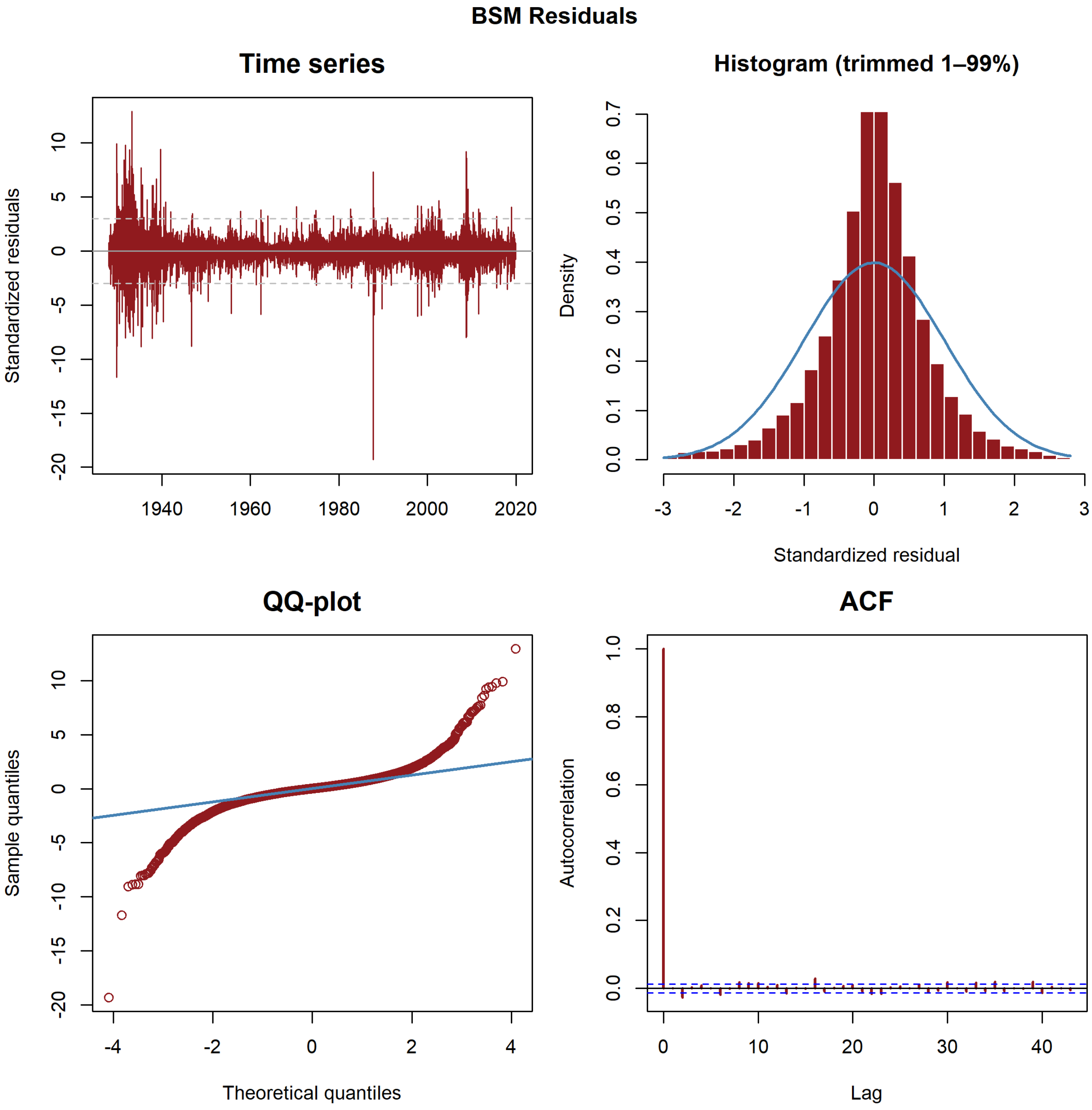


Figure A.3.1: *Standarized residuals for the BSM. The histogram is trimmed for the purpose of inspection.*

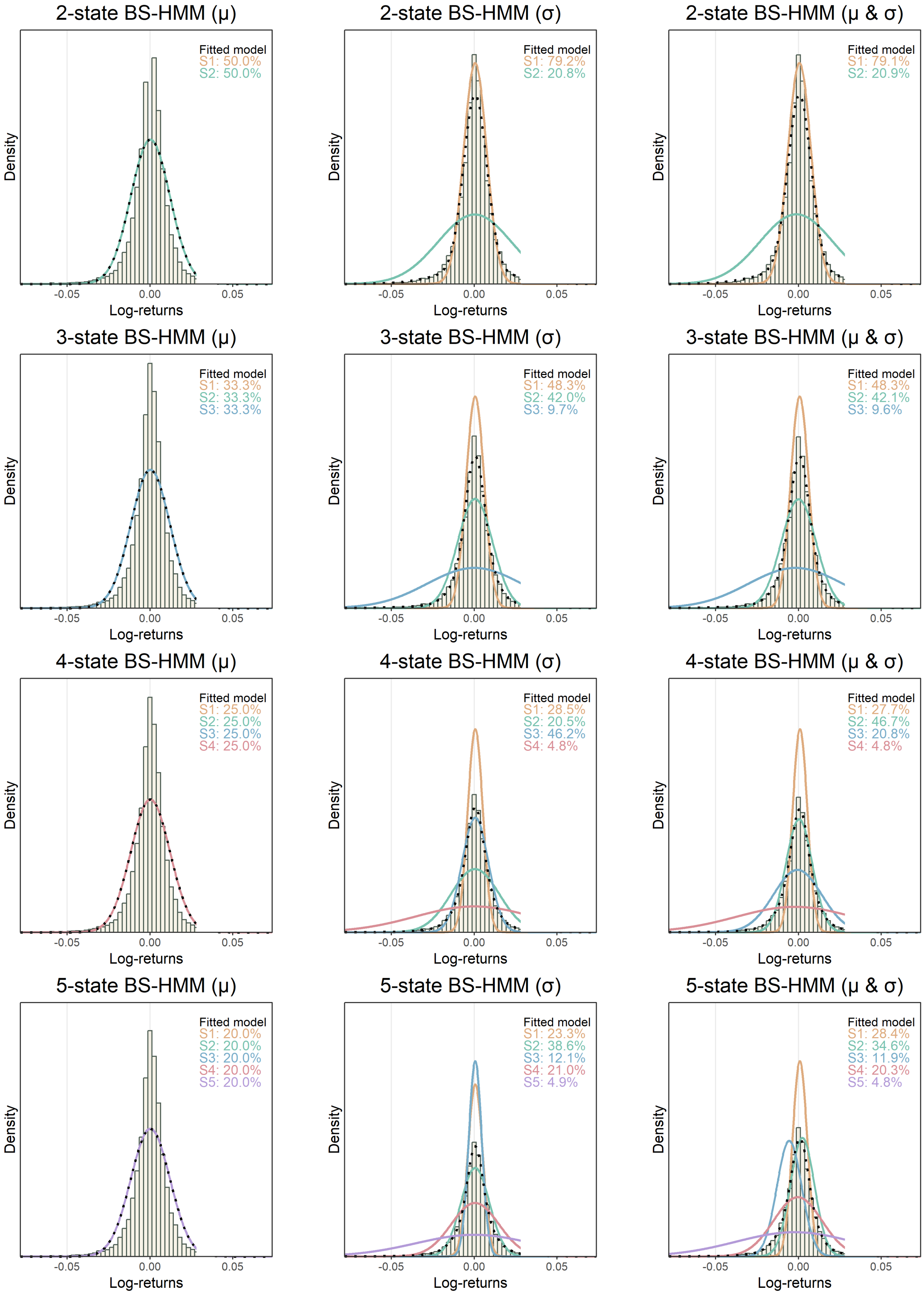


Figure A.3.2: *State-dependent density plots for the BS-HMMs*

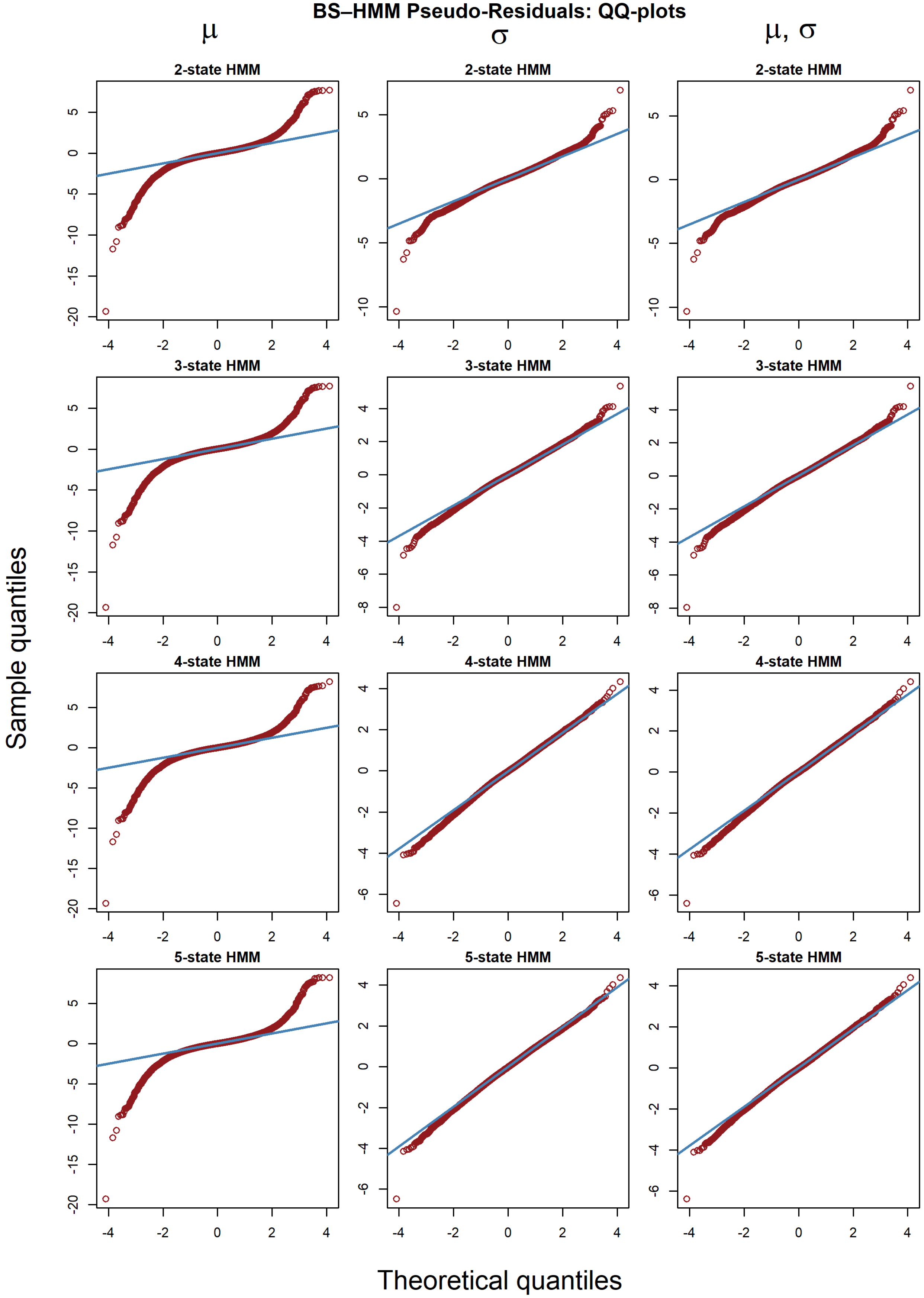


Figure A.3.3: *QQ-plot for the BS-HMMs.*

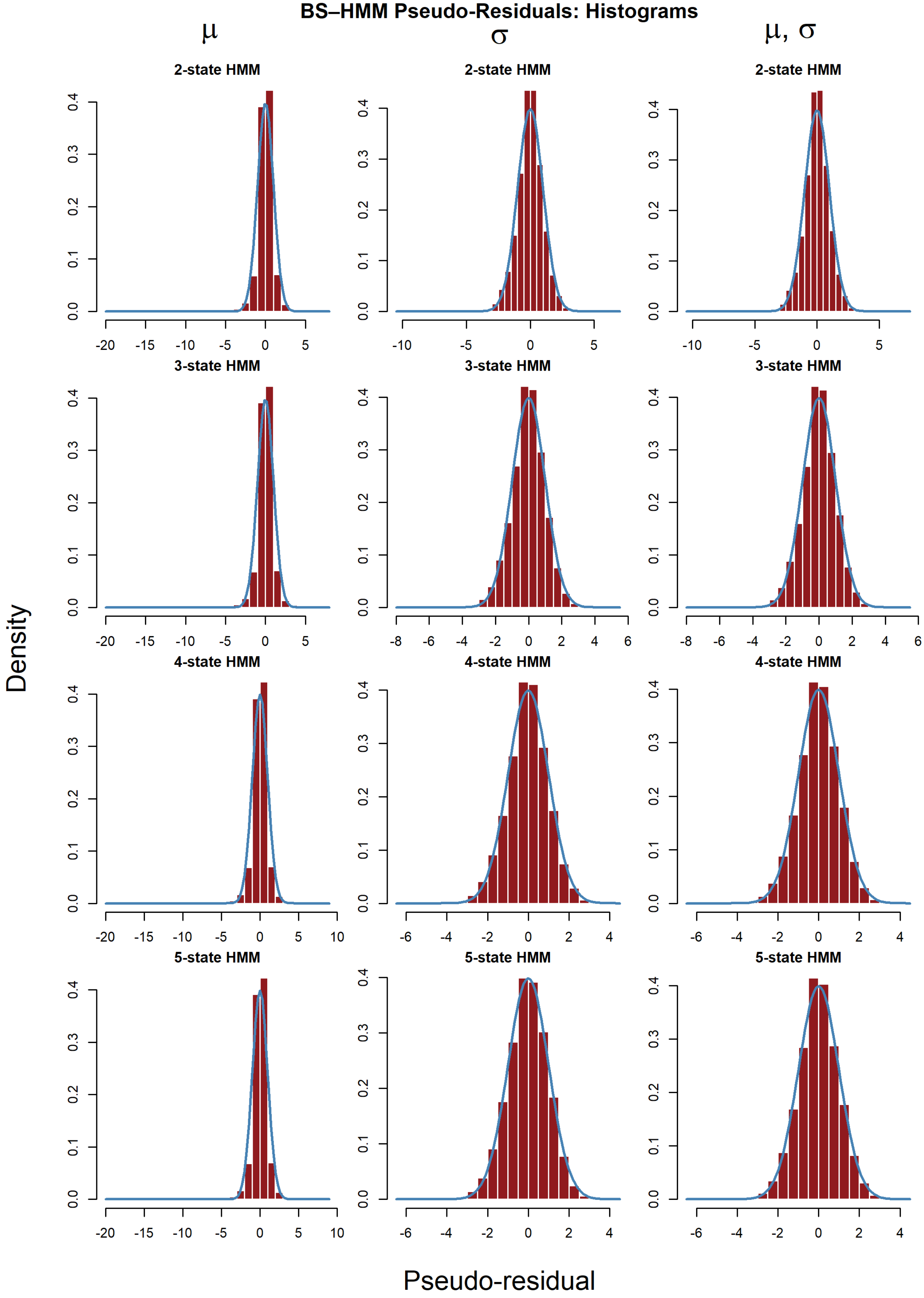


Figure A.3.4: Histograms for the BS-HMMs. We trimmed the residuals for inspection

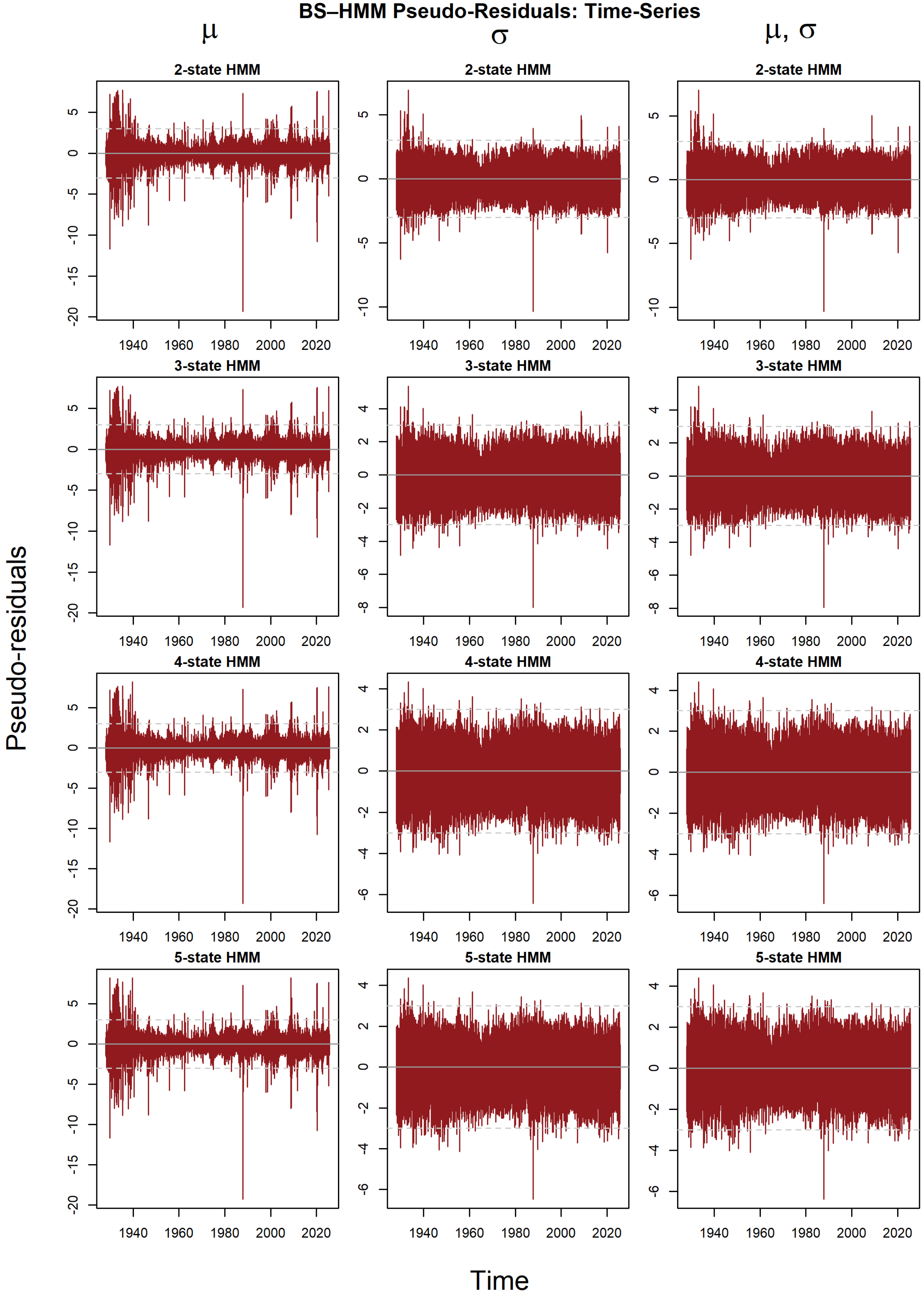


Figure A.3.5: *Histograms for the BS-HMMs. We trimmed the residuals for inspection*

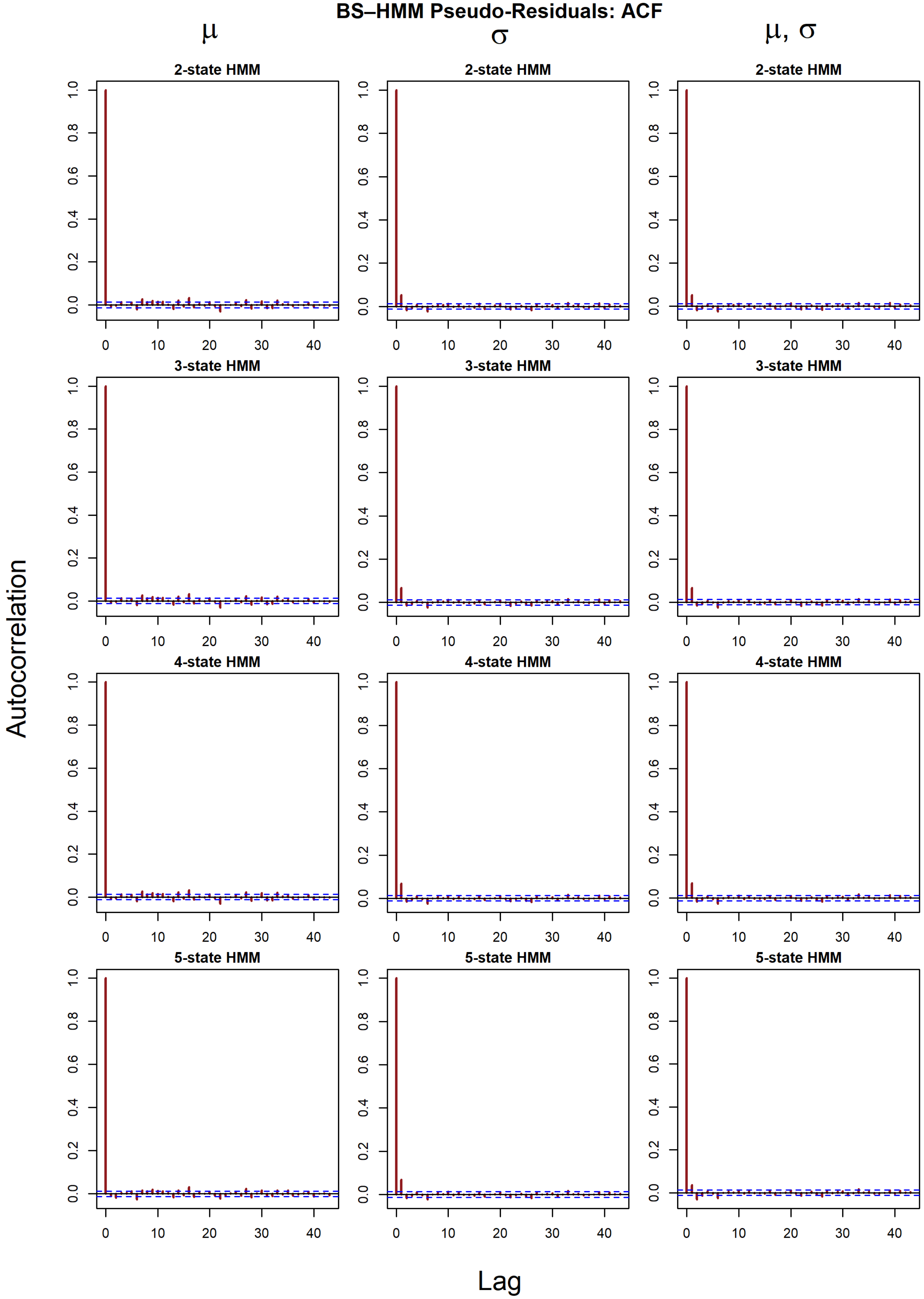


Figure A.3.6: ACF plot for the BS-HMMs.

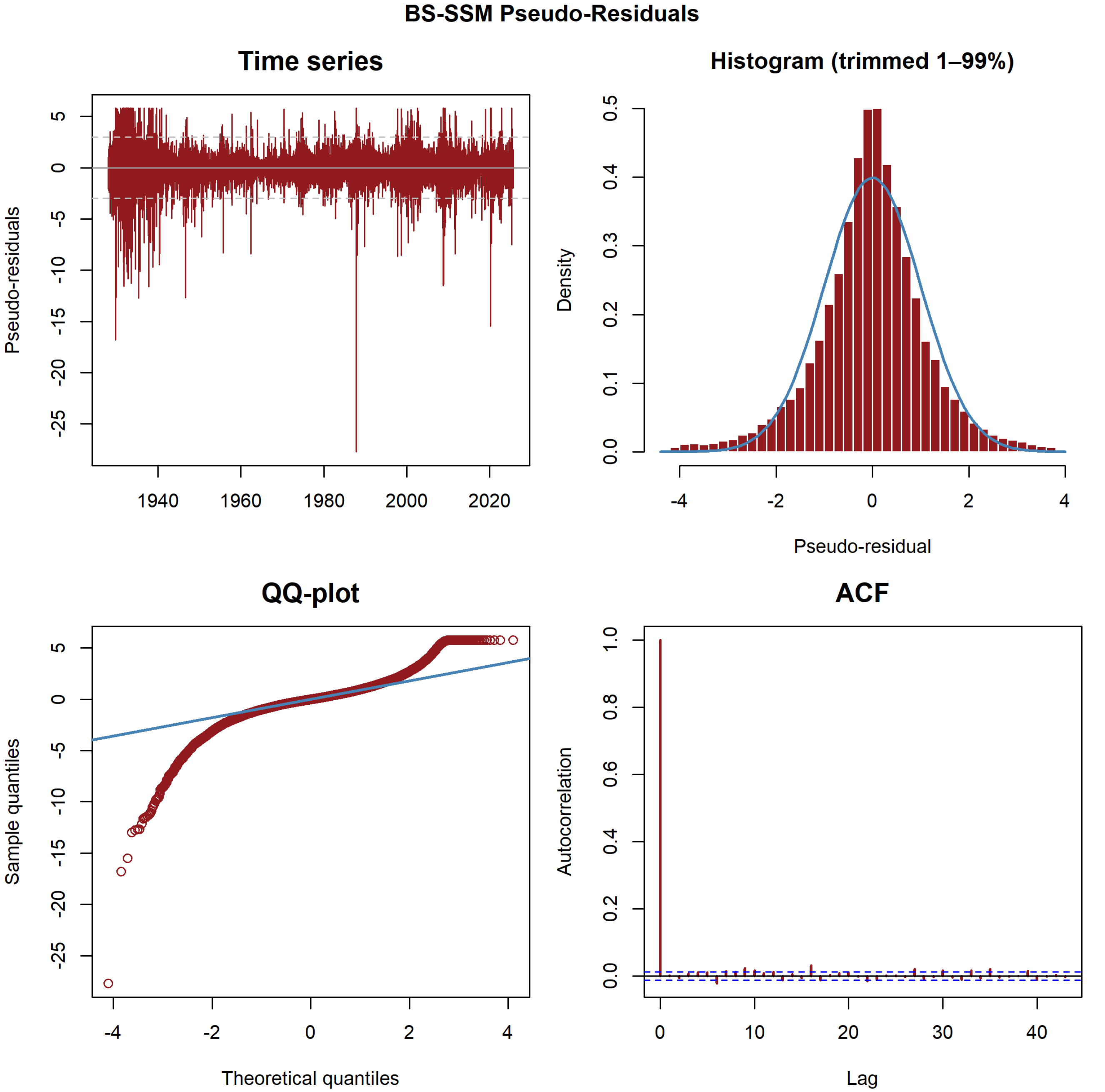


Figure A.3.7: *Pseudo-residuals for the BS-SSM. The histogram is trimmed for the purpose of inspection.*