universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Bayesian Consistency for a Class of Stochastic Volatility Models“

verfasst von / submitted by

Noah Leander Wolfahrt, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien / Vienna, 2022

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 821
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Mathematik UG2002
Betreut von / Supervisor:	Ass.-Prof. Dr. Julio Daniel Backhoff-Veraguas

I would like to thank my supervisor thoroughly for his extensive advice and patience, as well as everyone else who believed in me all these years.

Abstract [EN]

The following master's thesis aims to prove consistency for a specific class of stochastic volatility models. Here, the log-price of a stock is assumed to behave as in a multinomial tree model, however we allow for a 1-step dependency that extends the model's scope. Furthermore, it is possible to include into the model the observation of a finite number of financial derivatives. Applying a Monte Carlo-like approach, no parameters need to be estimated. Instead, a finite number of these models with given parameters is combined and weighted by their likelihood given the observed data over time. We prove that this Bayesian model approach eventually converges to the submodel that obtains the lowest reverse relative entropy among all other models. Finally, we provide the reader with a short implementation of a simplified version of this Bayesian model and discuss the results.

Abstract [DE]

In der folgenden Masterarbeit soll die Konsistenz für eine bestimmte Klasse von stochastischen Volatilitätsmodellen bewiesen werden. Dabei wird davon ausgegangen, dass sich der logarithmierte Preis einer Aktie wie in einem Multinomial-Modell verhält, wobei wir jedoch eine einstufige Abhängigkeit zulassen, die das Modell erweitert. Außerdem ist es möglich, eine endliche Anzahl von Finanzderivaten in das Modell miteinzubeziehen. Bei Anwendung eines Monte-Carlo-ähnlichen Ansatzes müssen keine Parameter geschätzt werden. Stattdessen wird eine endliche Anzahl dieser Modelle mit gegebenen Parametern kombiniert und mit ihrer Wahrscheinlichkeit angesichts der beobachteten Daten im Zeitverlauf bewertet und gewichtet. Wir beweisen, dass dieser Bayes'sche Ansatz schließlich zu dem Teilmodell konvergiert, das unter allen anderen Modellen die geringste „umgekehrte relative Entropie“ aufweist. Am Ende stellen wir dem Leser eine kurze Implementierung einer vereinfachten Version dieses Bayes'schen Modells zur Verfügung und diskutieren die Ergebnisse.

Contents

Chapter 1

Introduction

What is the modern approach regarding the mathematical modeling of modern financial problems? Mostly, some parametric model is implemented, and the parameters are estimated by historical data. There are several issues to be considered using this method. Firstly, it is anything but obvious how the estimation error should be accounted for in the model. The future impact of the error is mostly unclear, however often times it is simply implicitly postulated that it is of negligible significance. Secondly, the further we travel into the future, the more precise our parameters should become. This will always lead to contradictions. At one time, the model might tell us to buy a certain asset right now, at another time, it tells us we should have sold the asset at that time since one or more parameters have changed. Yet we continue to act according to our model, although we know it might give us different advice for the current moment if we wait long enough. Thirdly, the model is rigid. Its fundamental form is considered as given beforehand, the data changes only the parameters, not the model itself. In this thesis, another approach is presented that aims to alleviate these issues.

For now, our goal is to model the course of an asset, say a stock. The most crucial model we will consider is the Stochastic Volatility Model (SVM), see Figure 1.1.

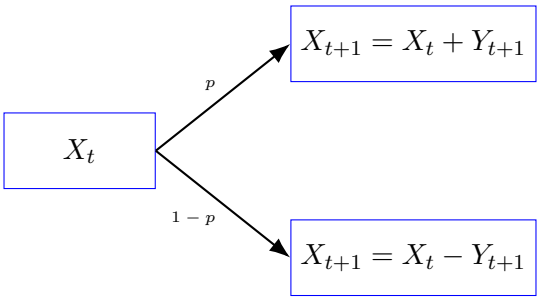

Figure 1.1: One time step progression of an asset in a Stochastic Volatility Model.

Resembling a binomial model, it is actually much more sophisticated by additionally accounting for a time-dependent stochastic volatility. That is, its current volatility influences the volatility in the next step, both in probability and magnitude. Here the log-price of the share is denoted by X_t, and the volatility is modelled by a Markov chain Y_t. The direction of the asset, i.e. the sign of the volatility, is modelled as a family of independent, identically distributed Bernoulli variables ξ_t with parameter p. At a later stage, we are going to take into account observations of a finite amount of financial derivatives such as options.

Now the fundamental idea is to consider a finite universe of $J > 1$ SVMs that is already given to us. At each time step t, we assign a probability $\pi_j(t)$ to each model j, which can be interpreted as a kind of reliability measure. Here, everything can be computed by summing over the weighted models. For instance, suppose we have computed the fair price $\varphi^j(t)$ of an option in each model, then our best guess to the true value is $\sum_j \pi_j(t)\varphi^j(t)$. Likewise, the distribution of X_t can be computed in a similar fashion. Note that these probabilities $\pi_j(t)$ change over time, as we observe more and more data, which intuitively might make more sense than a dynamical parameter that is assumed to model the same unchanging quantity for the time span of our purpose.

What are the advantages of this approach? As we calculate exact values, we do not have an estimation error in the classical sense. Furthermore, as probabilities often do change over time, we avoid contradictions as the ones mentioned before: our parameters, which describe the same quantity over time, do not change, however probabilities can and do change when more data is considered. Lastly, the model is not entirely rigid. For instance, over time we allow for a different emphasis of the observed derivative prices with respect to the asset prices.

The underlying idea of "estimating nothing" was first discussed by Duembgen and Rogers in [2]. We will utilize their approach of aggregating multiple SVMs into one Bayesian Stochastic Volatility supermodel, and show that the supermodel is consistent, i.e. it converges to the one submodel that fits best to the historical data. The thesis is structured as follows. In Chapter 1, we give an overview and introduce the most important concepts in order to formalize our model. Here, we also introduce the concept of reverse relative entropy to formalize the notion of best-fitting. Chapter 2 discusses various models and introduces the SVM and its Bayesian counterpart. Chapter 3 covers the essence of the paper, the consistency proofs under various assumptions. Finally, in Chapter 4 we present the results of a short Python implementation of the Bayesian SVM.

Chapter 2

Preliminaries

2.1 Motivation

Inspired by [2], we would like to study a single asset whose log-price at time t will be denoted by a stochastic process X_t and which is observed at discrete time steps $t = 0, 1, 2, \ldots$ only. We propose the asset moves according to

$$X_{t+1} - X_t = \xi_t Y_t.$$

In this equation, $Y_t > 0$ accounts for the volatility of the asset price, which we assume to be a Markov chain, and each $\xi_t \in \{-1, 1\}$ is an independent random sign that determines the direction our asset is moving towards. Furthermore, we observe A derivatives depending on the current value X_t, and we denote the market price of the derivative a at time t by $\Phi_a(t)$. Hence we equip the model with pricing functions $(\varphi_a)_{a=1}^A$, and it is the difference between the theoretical and observed prices that we also want to account for in our model. Formally, the error terms are measured by

$$Q\left(\varphi_a(t)\right) := \sum_{a=1}^{A} |\varphi_a(t) - \Phi_a(t)|^2.$$

Of course, the particular choice of Q could be altered to meet different demands. Now the fundamental idea is not to estimate any parameters but instead consider a given finite universe of J models and assign a reliability measure, i.e. a probability value, to each. Consequentially, all models have to be reevaluated at each time step t, since we observe more and more data over time.

How do we compute probabilities for each model j? Consider the log-likelihood function generated by our observed data, i.e.

$$\ell_j(t) = \sum_{s=1}^{t} \log\left[\mathbb{P}_j(X_s = x_s \mid X_{s-1} = x_{s-1})\right] - Q(\varphi_a^j(s)).$$

We can now quantify the reliability of model j by assigning a probability

$$\pi_j(t) = \frac{\exp(\ell_j(t))}{\sum_k \exp(\ell_k(t))}$$

to each model j. At this point, everything can be calculated easily. The distribution of X_{t+1} at time t is given by

$$\mathbb{P}(X_{t+1} = x_{t+1}) = \sum_{j=1}^{J} \pi_j(t)\mathbb{P}_j(X_{t+1} = x_{t+1} \mid X_t = x_t),$$

whereas we will consider the price of derivative a to be

$$\sum_{j=1}^{J} \pi_j(t)\varphi_a^j(t).$$

It should be highlighted that no parameters have been estimated. Instead, different models have been compared and combined, and due to the assigned model probabilities, prices can be calculated exactly. Moreover, the likelihood incorporates not only the observed market price of the underlying asset but also the market prices of financial derivatives. This comparatively novel Bayesian approach can be seen as some kind of particle filtering or sequential Monte Carlo method. The Bachelor thesis [6] discusses the essential ideas from [2] for simplified trinomial models. What is new in this thesis is the discussion of a multinomial tree resembling model, with the added difficulty that the size of increments contains a degree of past-dependence (stochastic volatility) that is absent in multinomial trees. At the end, we present a simple implementation based on real financial data programmed in Python.

We resume by recalling the fundamental concepts needed to rigorously construct our model.

2.2 Markov Chains

In this section, we will implicitly consider an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which all random variables are defined, if not stated otherwise. Similarly, Markov chains are implicitly assumed to be time-homogenous and take values in a finite state space E.

2.2.1 Basic Definitions

This section recalls the most important definitions from [7].

Definition 2.2.1 (Markov Chain)**.** A discrete-time stochastic process $(Y_t)_{t\geq 0}$ on a state space E is a Markov chain with initial distribution λ and transition matrix P if

(i) $\mathbb{P}(Y_0 = y) = \lambda(y)$,

(ii) $\mathbb{P}(Y_{t+1} = y_{t+1} \mid Y_0 = y_0, \ldots, Y_t = y_t) = P(y_t, y_{t+1})$.

The second condition is called the Markov property. As a shorthand notation, we say $(Y_t)_{t\geq 0}$ is Markov(λ, P).

Remark 2.2.2. We can compute the transition probabilities for two states and t steps by computing P^t, i.e. the t-th power of the stochastic matrix P. Then $P^t(y, \tilde{y})$ denotes the probability of reaching $\tilde{y}$ from y in exactly t steps.

Definition 2.2.3 (Irreducibility)**.** For $y, \tilde{y} \in E$, we say that y leads to $\tilde{y}$ if $\mathbb{P}(Y_t = \tilde{y} \text{ for some } t \geq 0 \mid Y_0 = y) > 0$. This will be denoted by $y \to \tilde{y}$. Similarly, we write $y \leftrightarrow \tilde{y}$ if both $y \to \tilde{y}$ and $\tilde{y} \to y$. One can show that $\leftrightarrow$ is an equivalence relation on E, thus partioning our state space into equivalence classes. If E consists only of one equivalence class under $\leftrightarrow$, we call the Markov chain irreducible.

Remark 2.2.4. Alternatively, we can characterize $y \to \tilde{y}$ by requiring the existence of a t such that $\mathbb{P}(Y_t = \tilde{y} \mid Y_0 = y) > 0$. Both definitions are equivalent to one another, as can be seen as follows:

$$\mathbb{P}(Y_t = \tilde{y} \mid Y_0 = y) \leq \mathbb{P}\big(\bigcup_{s=1}^{t} \{Y_s = \tilde{y}\} \mid Y_0 = y\big) \leq \sum_{s=1}^{t} \mathbb{P}(Y_s = \tilde{y} \mid Y_0 = y).$$

Definition 2.2.5 (Aperiodicity)**.** Let $(Y_t)_{t\geq 0}$ be Markov(λ, P). If for a state $y \in E$ it holds that $P^t(y, y) > 0$ for all sufficiently large t, we call y aperiodic. If all states are aperiodic, we call the Markov chain aperiodic. The Markov chain in Figure 2.1 is aperiodic, and the one in Figure 2.2 has a period of 3, since each state can only be reached from itself in a number of steps that is a multiple of three.

Definition 2.2.6 (Recurrence)**.** Let $(Y_t)_{t\geq 0}$ be Markov(λ, P). We call the state $y \in E$ recurrent if almost surely, the chain returns to y infinitely often, i.e. $\mathbb{P}(Y_t = y \text{ for infinitely many } t \mid Y_0 = y) = 1$. Additionally, we define the hitting time of a state y by

$$T_y := \inf\{t > 0 : Y_t = y\}.$$

If we have $\mathbb{E}(T_y \mid Y_0 = y) < \infty$ (where the notation conveys the underlying assumption that the chain starts at y), then we say y is positive recurrent.

Definition 2.2.7 (Invariant distribution)**.** Let $(Y_t)_{t\geq 0}$ be a Markov chain with transition matrix P. A probability distribution π on E can be considered as a row vector $(\pi(y) : y \in E)$. It is called an invariant distribution if

$$\sum_{\tilde{y} \in E} \pi(\tilde{y}) P(\tilde{y}, y) = \pi(y) \qquad \text{for all } y \in E,$$

or in short, $\pi P = \pi$.

Remark 2.2.8. Multiple Markov chains can have the same underlying invariant distribution. For instance, consider the Markov chains in Figure 2.1, Figure 2.2 and Figure 2.3. All three chains are defined on the same state space $E = \{1, 2, 3\}$. Although their transition probabilities are different, the probability distribution $\pi = (1/3, 1/3, 1/3)$ is invariant for all chains (although for the latter, its invariant distribution is not unique).

2.2.2 Useful Properties

Now that we have introduced Markov chains, we recall some well-known properties from [7] that we will use later. For a more extensive outlook, we refer to the original source.

Lemma 2.2.9. *Let $(Y_t)_{t\geq 0}$ be an irreducible Markov(λ, P) with at least one aperiodic state. Then for all states $y, \tilde{y} \in E$, we have $P^t(y, \tilde{y}) > 0$ for all sufficiently large t. In particular, all states are aperiodic.*

Proof. Since P is irreducible, there exist $r, s \geq 0$ such that $P^r(y, x), P^s(x, \tilde{y}) > 0$ for any $x \in E$. Therefore, if x denotes an aperiodic state,

$$P^{r+q+s}(y, \tilde{y}) \geq P^r(y, x)P^q(x, x)P^s(x, \tilde{y}) > 0.$$

□

Theorem 2.2.10. *Let $(Y_t)_{t\geq 0}$ be an irreducible and positive recurrent Markov(λ, P). Then there exists a unique invariant distribution. Furthermore, if the chain is aperiodic, we have*

$$\mathbb{P}(Y_t = y) \xrightarrow[t\to\infty]{} \pi(y) \qquad \text{for all } y \in E.$$

Proof. We only provide a sketch of the proof given in [7, Thm 1.7.7]. The idea is to show that the expected time $\gamma^x(y)$ spent in a state y between visits to x, formally

$$\gamma^x(y) = \mathbb{E}\left[\sum_{s=1}^{T_x} \mathbb{1}_{\{X_s=y\}} \mid X_0 = x\right],$$

is invariant, that is, it satisfies $\gamma^x P = \gamma^x$ for all $x \in E$. Furthermore, one has to show that

$$\sum_{y\in E} \gamma^x(y) =: m_x < \infty,$$

since then $\pi_y := \gamma^x(y)/m_x$ defines an invariant distribution. □

The following properties are not stated in [7], however they are useful for later purposes. They can be deduced using simple arguments.

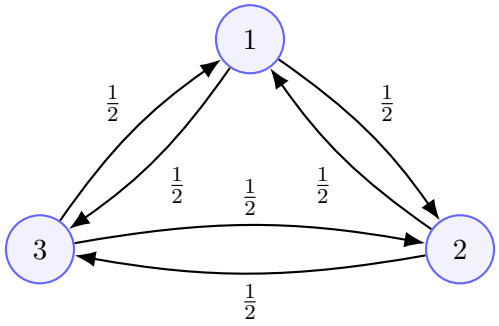

Figure 2.1: Example of an irreducible, aperiodic Markov chain.

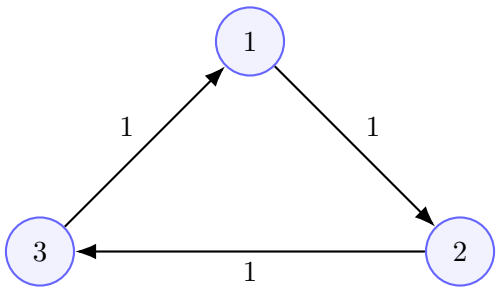

Figure 2.2: Example of an irreducible Markov chain with period 3.

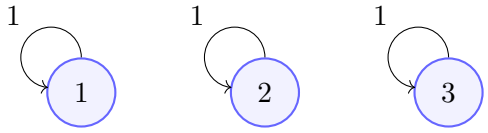

Figure 2.3: Example of a Markov chain with three communicating classes.

Definition 2.2.11. Let $(Y_t)_{t\geq 0}$ be Markov(λ, P). The corresponding concatenated chain Z_t is defined by

$$Z_t := \begin{pmatrix} Y_t \\ Y_{t-1} \end{pmatrix}, \qquad t \geq 1,$$

which acts on the state space

$$\tilde{E} := \{(x, y) \in E \times E : P(y, x) > 0\}.$$

Proposition 2.2.12. *Let $(Y_t)_{t\geq 0}$ be an irreducible and aperiodic Markov(λ, P). Then $(Z_t)_{t\geq 1}$ defined as in Definition 2.2.11 is an irreducible and aperiodic Markov chain.*

Proof. We split the proof into three parts.

(i) $\tilde{\lambda}$ is given by $\tilde{\lambda}((y, x)) = \lambda(x)P(x, y)$. The transition matrix $\tilde{P}$ is given by

$$\tilde{P}((y_1, y_2), (y_3, y_4)) = \begin{cases} P(y_1, y_2)P(y_3, y_4), & y_1 = y_4, \\ 0, & y_1 \neq y_4. \end{cases}$$

The Markov property is trivially preserved and $(Z_t)_{t\geq 1}$ is Markov$(\tilde{\lambda}, \tilde{P})$.

(ii) Irreducibility: without loss of generality, take arbitrary $(\tilde{y}, y), (y_1, y_0) \in \tilde{E}$. The desired property then follows from the Markov property and Lemma 2.2.9:

$$\begin{aligned} &\mathbb{P}(Z_t = (\tilde{y}, y)^\top \text{ for some } t \mid Z_1 = (y_1, y_0)^\top) \\ &\quad = \mathbb{P}(Y_t = \tilde{y}, Y_{t-1} = y \text{ for some } t \mid Y_1 = y_1, Y_0 = y_0) \\ &\quad = \mathbb{P}(Y_t = \tilde{y}, Y_{t-1} = y \text{ for some } t \mid Y_1 = y_1) \\ &\quad \geq P^{t-2}(y_1, y)P(y, \tilde{y}). \end{aligned}$$

We now choose t such that the first factor is positive.

(iii) Aperiodicity: let $z = (y, \tilde{y}) \in \tilde{E}$. We have

$$\tilde{P}^t(z, z) = \tilde{P}^t((y, \tilde{y}), (y, \tilde{y})) = P(\tilde{y}, y)P^{t-1}(y, \tilde{y})P(\tilde{y}, y) > 0$$

for all n large enough, by Lemma 2.2.9.

□

Remark 2.2.13. Let $(Y_t)_{t\geq 0}$ be an irreducible Markov chain acting on a finite state space E. Then there exists at least one positive recurrent state $y \in E$. As per the irreducibility, all states must be positive recurrent.

Proposition 2.2.14. *Let $(Y_t)_{t\geq 0}$ be an irreducible and aperiodic Markov chain with transition matrix P and unique invariant distribution π. Let $(Z_t)_{t\geq 1}$ denote the corresponding concatenated chain as defined in Definition 2.2.11. Then $(Z_t)_{t\geq 1}$ has a unique invariant distribution μ, and*

$$\mu(z) = \mu((y, \tilde{y})) = P(\tilde{y}, y)\pi(\tilde{y}), \qquad z = (y, \tilde{y}) \in \tilde{E}.$$

Proof. By Proposition 2.2.12 and Remark 2.2.13, $(Z_t)_t$ is irreducible, aperiodic and positive recurrent. Therefore by the Markov property and Theorem 2.2.10, we obtain with $z = (y, \tilde{y}), z_1 = (y_1, y_0)$:

$$\begin{aligned}
&\lim_{t\to\infty} \mathbb{P}(Z_t = z \mid Z_1 = z_1) \\
&\quad = \lim_{t\to\infty} \mathbb{P}(Y_t = y, Y_{t-1} = \tilde{y} \mid Y_1 = y_1, Y_0 = y_0) \\
&\quad = \lim_{t\to\infty} \mathbb{P}(Y_t = y, Y_{t-1} = \tilde{y} \mid Y_1 = y_1) \\
&\quad = \lim_{t\to\infty} \mathbb{P}(Y_t = y \mid Y_{t-1} = \tilde{y}, Y_1 = y_1) \lim_{t\to\infty} \mathbb{P}(Y_{t-1} = \tilde{y} \mid Y_1 = y_1) \\
&\quad = \lim_{t\to\infty} \mathbb{P}(Y_t = y \mid Y_{t-1} = \tilde{y}) \lim_{t\to\infty} \mathbb{P}(Y_{t-1} = \tilde{y}) \\
&\quad = P(\tilde{y}, y)\pi(\tilde{y}).
\end{aligned}$$

□

2.2.3 Ergodicity

The definitions, statements and proofs in this section are also taken from [7].

Definition 2.2.15. Let $(Y_t)_{t\geq 0}$ be a Markov chain. We call it ergodic if it is aperiodic and positive recurrent.

Definition 2.2.16. Let $(Y_t)_{t\geq 0}$ be a Markov chain. In the following, we will denote the number of visits to a state $y \in E$ before time $s \geq 0$ by

$$V_y(s) = \sum_{t=0}^{s-1} \mathbb{1}_{\{Y_t = y\}}.$$

Proposition 2.2.17. *Let $(Y_t)_{t\geq 0}$ be an irreducible and positive recurrent Markov chain. Then the proportion of time spent in a state $y \in E$ converges towards its probability under the unique invariant distribution π:*

$$\frac{V_y(t)}{t} \xrightarrow[t\to\infty]{} \pi(y) \quad \textit{for all } y \in E \qquad \textit{a.s.}$$

Proof. We refer to [7, Thm 1.10.2]. □

Theorem 2.2.18 (Ergodic Theorem for Markov Chains)**.** *Let $(Y_t)_{t\geq 0}$ be an irreducible and positive recurrent Markov chain with transition matrix P. Then there exists a unique invariant distribution π and for any bounded $f : E \to \mathbb{R}$, it holds that*

$$\frac{1}{t}\sum_{k=0}^{t-1} f(Y_k) \xrightarrow[t\to\infty]{} \mathbb{E}_\pi[f] = \sum_{y\in E} \pi(y) f(y) \qquad \textit{a.s.}$$

Proof. By Theorem 2.2.10, there exists a unique invariant distribution π. Now without loss of generality, assume that $\|f\|_\infty \leq 1$. Now for any $F \subseteq E$, we have

$$\begin{aligned}
&\left|\frac{1}{t}\sum_{k=0}^{t-1} f(Y_k) - \mathbb{E}_\pi[f]\right| \\
&= \left|\sum_{y\in E}\sum_{k=0}^{t-1} \frac{\mathbb{1}_{\{Y_k=y\}}}{t} f(y) - \sum_{y\in E}\pi(y)f(y)\right| \\
&= \left|\sum_{y\in E}\left(\frac{V_y(t)}{t} - \pi(y)\right) f(y)\right| \\
&\leq \sum_{y\in F}\left|\frac{V_y(t)}{t} - \pi(y)\right| + \sum_{y\notin F}\left|\frac{V_y(t)}{t} - \pi(y)\right| \\
&\leq \sum_{y\in F}\left|\frac{V_y(t)}{t} - \pi(y)\right| + \sum_{y\notin F}\left(\frac{V_y(t)}{t} + \pi(y)\right) = (\star).
\end{aligned}$$

At this point, we note that $\sum_{y\in E} V_y(t) = t$, and hence

$$\begin{aligned}
(\star) &= \sum_{y\in F}\left|\frac{V_y(t)}{t} - \pi(y)\right| + \left(1 - \sum_{y\in F}\frac{V_y(t)}{t}\right) + \sum_{y\notin F}\pi(y) \\
&= \sum_{y\in F}\left|\frac{V_y(t)}{t} - \pi(y)\right| + \sum_{y\in F}\left(\pi(y) - \frac{V_y(t)}{t}\right) + 2\sum_{y\notin F}\pi(y) \\
&\leq 2\sum_{y\in F}\left|\frac{V_y(t)}{t} - \pi(y)\right| + 2\sum_{y\notin F}\pi(y).
\end{aligned}$$

Let $\varepsilon > 0$. Now we can choose F finite such that

$$\sum_{y\notin F}\pi(y) < \varepsilon/4$$

and by Proposition 2.2.17, there exists $N = N(\omega)$ such that for all $t \geq N$, it holds

$$\sum_{y\in F}\left|\frac{V_y(t)}{t} - \pi(y)\right| < \varepsilon/4.$$

Putting everything together, we obtain the desired result. □

Theorem 2.2.19 (Strong Law of Large Numbers)**.** *Let $(Y_n)_{n\geq 0}$ be a sequence of independent, identically distributed random variables with finite first moment. Then*

$$\frac{1}{n}\sum_{k=0}^{n-1} Y_k \xrightarrow[n\to\infty]{} \mathbb{E}(Y_1) \qquad a.s.$$

Proof. Etemadi presented a proof in [3]. □

Remark 2.2.20. At this point, we would like to emphasize the similarity of Theorems 2.2.18 and 2.2.19: both yield the convergence of the average to a certain kind of mean. We will make use of these statements simultaneously later on.

2.3 Entropy

For this section, consider a measurable space $(\Omega, \mathcal{F})$ on which all measures are defined.

Definition 2.3.1 (Absolute continuity)**.** Let μ, ν be probability measures. We say ν is absolutely continuous with respect to μ and write $\nu \ll \mu$, if

$$\forall A \in \mathcal{F}: \qquad \mu(A) = 0 \quad \Rightarrow \quad \nu(A) = 0.$$

Definition/Theorem 2.3.1 (Radon-Nikodym)**.** *Let $\nu \ll \mu$. Then there exists an $\mathcal{F}$-measurable function $f : \Omega \to [0, \infty)$ such that for any $A \in \mathcal{F}$, we have*

$$\nu(A) = \mathbb{E}_\mu[\mathbb{1}_A f].$$

We call f the Radon-Nikodym derivative of ν with respect to μ and write $f = \frac{d\nu}{d\mu}$.

Proof. See [1, Thm 8.6] for a comprehensive proof. □

Example 2.3.2. Let $p, \tilde{p} \in (0, 1)$ and consider two independent Bernoulli variables $B \sim \mathrm{Ber}(p), \tilde{B} \sim \mathrm{Ber}(\tilde{p})$. Then $\mathrm{Law}(B) \ll \mathrm{Law}(\tilde{B})$ and the Radon-Nikodym derivative f is given by

$$f = \frac{\mathrm{d}B}{\mathrm{d}\tilde{B}} = \frac{p}{\tilde{p}} \mathbb{1}_{\{1\}} + \frac{1-p}{1-\tilde{p}} \mathbb{1}_{\{0\}}.$$

Proof. Simple calculations lead to the statement: for instance, let $A = \{1\}$, then

$$\tilde{\mathbb{E}}[\mathbb{1}_{\{1\}} f] = \tilde{p} f(\{1\}) = p = B(\{1\}).$$

□

Proposition 2.3.3. *Let Ω be at most countably infinite and let $\nu \ll \mu$ be two probability measures. Then the Radon-Nikodym derivative of ν with respect to μ is given by*

$$\frac{d\nu}{d\mu} = \sum_{\substack{\omega \in \Omega: \\ \mu(\omega) > 0}} \frac{\nu(\omega)}{\mu(\omega)} \mathbb{1}_{\{\omega\}}.$$

Proof. For any $A \subseteq \Omega$,

$$\begin{aligned}
\mathbb{E}_\mu\left[\mathbb{1}_A \sum_{\substack{\omega \in \Omega \\ \mu(\omega) > 0}} \frac{\nu(\omega)}{\mu(\omega)} \mathbb{1}_{\{\omega\}}\right] &= \mathbb{E}_\mu\left[\sum_{\substack{\omega \in A \\ \mu(\omega) > 0}} \frac{\nu(\omega)}{\mu(\omega)} \mathbb{1}_{\{\omega\}}\right] \\
&= \sum_{\substack{\omega \in A \\ \mu(\omega) > 0}} \frac{\nu(\omega)}{\mu(\omega)} \mathbb{E}_\mu[\mathbb{1}_{\{\omega\}}] \\
&= \sum_{\substack{\omega \in A \\ \mu(\omega) > 0}} \nu(\omega) \\
&= \nu(A \cap \{\omega : \mu(\omega) > 0\}) \\
&= \nu(A),
\end{aligned}$$

since $\nu \ll \mu$. □

Lemma 2.3.4 (Jensen's inequality)**.** *Let $a_1, a_2, \ldots > 0$ such that $\sum_i a_i = 1$ and let $x_1, x_2, \ldots \in \mathbb{R}$. Given a concave function φ, it holds that*

$$\varphi\Big(\sum_{i\geq 1} a_i x_i\Big) \geq \sum_{i\geq 1} a_i \varphi(x_i).$$

Consequentially, for a convex function $\tilde{\varphi}$ and real numbers $x_1, \ldots, x_n \in \mathbb{R}$, we have

$$\tilde{\varphi}\Big(\frac{1}{n}\sum_{i=1}^{n} x_i\Big) \leq \frac{1}{n}\sum_{i=1}^{n} \tilde{\varphi}(x_i).$$

Proof. [1, Lem 8.10] provides a concise proof. □

Definition 2.3.5. Let $\nu \ll \mu$. We can define the reverse relative entropy, or entropy for short, of ν given μ in terms of its Radon-Nikodym derivative by

$$\mathbb{H}(\nu \mid \mu) := \mathbb{E}_\nu\left[\log \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right].$$

Remark 2.3.6. Define a convex function $f(x) = x\log(x)\mathbb{1}_{\{x\geq 0\}}$. Then by Jensen,

$$\begin{aligned}\mathbb{E}_\nu\left[\log \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right] &= \mathbb{E}_\mu\left[f\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right)\right] \\ &\geq f\left(\mathbb{E}_\mu\left[\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right]\right) \\ &= f(1) = 0,\end{aligned}$$

so the entropy $\mathbb{H}(\nu \mid \mu) \in [0, \infty]$ is well-defined, non-negative and possibly infinite.

Corollary 2.3.7. For any measure μ, we have $\mathbb{H}(\mu \mid \mu) = 0$.

Example 2.3.8. In the situation of Example 2.3.2, we can calculate the entropy of B given $\tilde{B}$:

$$\begin{aligned}\mathbb{H}(B \mid \tilde{B}) &= \mathbb{E}\left[\log \frac{\mathrm{d}B}{\mathrm{d}\tilde{B}}\right] \\ &= \mathbb{E}\left[\log\left(\frac{p}{\tilde{p}}\mathbb{1}_{\{1\}} + \frac{1-p}{1-\tilde{p}}\mathbb{1}_{\{0\}}\right)\right] \\ &= p\log\frac{p}{\tilde{p}} + (1-p)\log\frac{1-p}{1-\tilde{p}}.\end{aligned}$$

Proposition 2.3.9. *Let Ω be at most countably infinite and let $\nu \ll \mu$ be two probability measures. Then the entropy of μ given ν is given by*

$$\mathbb{H}(\nu \mid \mu) = \sum_{\omega\in\Omega} \nu(\omega)\log\frac{\nu(\omega)}{\mu(\omega)}.$$

Proof.

$$\begin{aligned}\mathbb{H}(\nu \mid \mu) &= \mathbb{E}_\nu\left[\log \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right] \\ &= \mathbb{E}_\nu\left[\log \sum_{\omega\in\Omega} \frac{\nu(\omega)}{\mu(\omega)} \mathbb{1}_{\{\omega\}}\right] \\ &= \sum_{\omega\in\Omega} \nu(\omega) \log \frac{\nu(\omega)}{\mu(\omega)}.\end{aligned}$$

□

Chapter 3

Estimating Nothing

3.1 Model Theory

In this section, we discuss the theoretical model concepts of [2].

3.1.1 Likelihood

Definition 3.1.1 (Statistical Model)**.** A family of probability spaces $(\Omega, \mathcal{F}, \mathbb{P}_\theta)_{\theta\in\Theta}$, consisting of a measurable space and family of probability measures indexed by a non-empty set Θ, is called a statistical model. We will proceed to call a statistical model discrete (resp. continuous) if Ω is discrete (resp. a Borel-measurable subset of $\mathbb{R}^n$). In either of these cases, $(\Omega, \mathcal{F}, (\mathbb{P}_\theta : \theta \in \Theta))$ is a standard model.

Definition 3.1.2 (Likelihood function)**.** Given a statistical standard model, we define the corresponding likelihood function by

$$\begin{aligned}\kappa : \Omega \times \Theta &\to [0,1] : \\ (\omega, \theta) &\mapsto \mathbb{P}_\theta(\{\omega\}).\end{aligned}$$

Given $\omega \in \Omega$, the mapping

$$\begin{aligned}\kappa_\omega = \kappa(\omega, \cdot) : \Theta &\to [0,1] : \\ \theta &\mapsto \kappa(\omega, \theta)\end{aligned}$$

is the likelihood function of the observed value $\omega \in \Omega$. Moreover, the function

$$\ell_\omega : \quad \theta \mapsto \log \kappa_\omega(\theta)$$

is the log-likelihood function of κ_ω to the observed value $\omega \in \Omega$.

3.1.2 Types of Models

Definition 3.1.3 (Binomial Model)**.** Let $p \in (0,1), a, b \in \mathbb{R}, a \neq b$ and let $(B_t)_{t\geq 1}$ be a sequence of independent, identically distributed Ber(p)-variables on $\{a, b\}$, i.e.

$$\mathbb{P}(B_t = a) = 1 - \mathbb{P}(B_t = b) = p.$$

Let $X_0 = x_0 \in \mathbb{R}$. The binomial model is a discrete-time stochastic process $(X_t)_{t\geq 0}$ on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$, with $\mathcal{F}_t = \sigma(X_1, \ldots, X_t)$ such that

$$X_t = x_0 + \sum_{k=1}^{t} B_k, \qquad t \geq 1.$$

A visualization of the model can be seen in Figure 3.1.

Definition 3.1.4 (Multinomial Model)**.** Let $p_1, p_2, ... \in (0,1)$ with $\sum_i p_i = 1$ and let $(M_t)_{t\geq 1}$ be a sequence of independent, identically distributed discrete random variables on a finite space $E \subset \mathbb{R}$ such that

$$\mathbb{P}(M_t = x_i) = p_i, \qquad x_i \in E.$$

Let $X_0 = x_0 \in \mathbb{R}$. The multinomial model is a discrete stochastic process $(X_t)_{t\geq 1}$ on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$, with $\mathcal{F}_t = \sigma(X_1, \ldots, X_t)$ such that

$$X_t = x_0 + \sum_{k=1}^{t} M_k, \qquad t \geq 1.$$

We call a multinomial model symmetric if $\mathbb{P}(M_t = x) = \mathbb{P}(M_t = -x)$ for all $x \in E$.

Definition 3.1.5 (Stochastic Volatility Model)**.** Let $p \in (0,1)$ and let $(\xi_t)_{t\geq 1}$ be a sequence of independent, identically distributed Bernoulli variables on $\{1, -1\}$ with parameter p, that is

$$\mathbb{P}(\xi_t = 1) = 1 - \mathbb{P}(\xi_t = -1) = p.$$

Furthermore, let $(Y_t)_{t\geq 1}$ be an irreducible, aperiodic and positive recurrent Markov chain on a finite state space $E \subset (0, \infty)$, independent of the family $(\xi_t)_t$. Suppose $X_0 = x_0 \in \mathbb{R}$. The stochastic volatility model is a discrete-time stochastic process $(X_t)_{t\geq 0}$ on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$, with $\mathcal{F}_t = \sigma(X_1, \ldots, X_t)$ such that

$$X_t - X_{t-1} = \xi_t Y_t, \qquad t \geq 1.$$

A visualization of the model can be seen in Figure 3.2.

In the following, we will mostly consider stochastic volatility models. One might suppose that this includes only a small subclass of multinomial models. However, the only proper restriction is an implicit symmetry assumption, as the following remark illustrates.

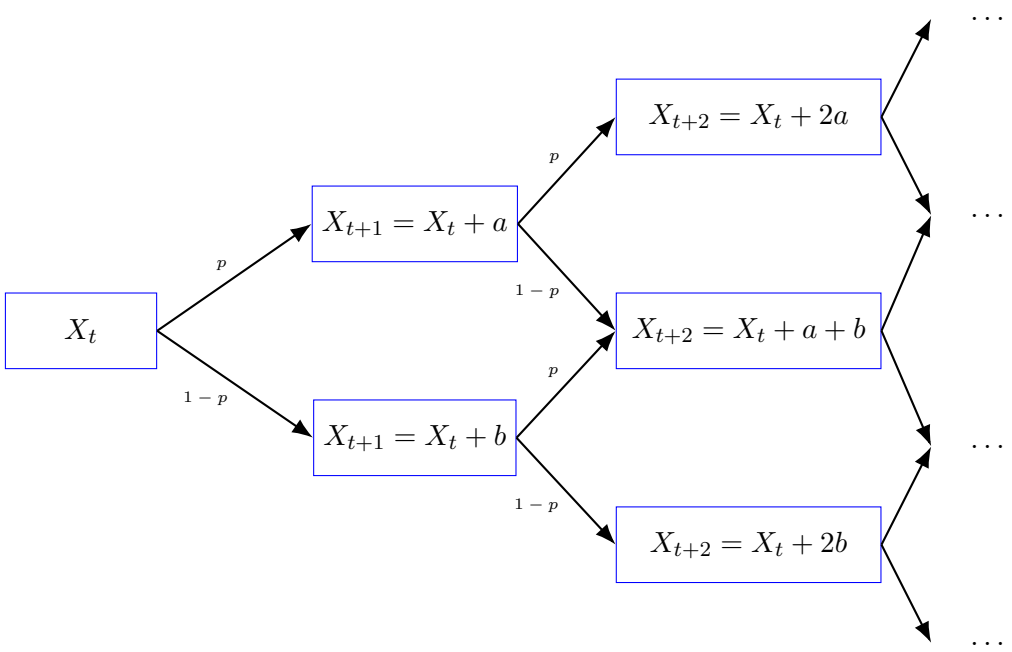

Figure 3.1: Progression of an asset under a binomial model.

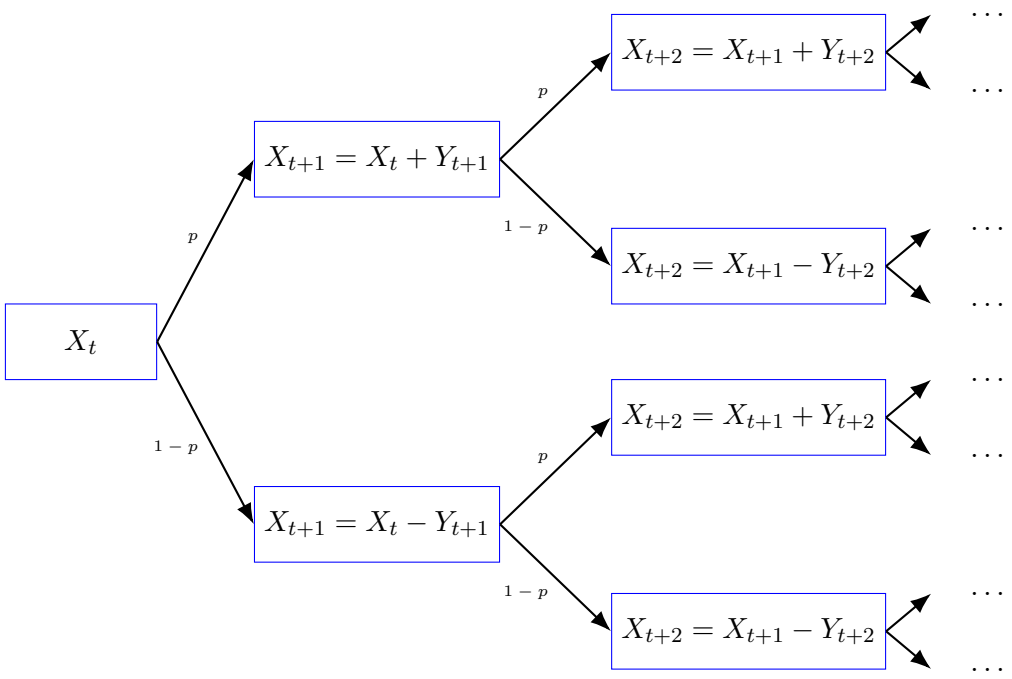

Figure 3.2: Progression of an asset under a stochastic volatility model.

Remark 3.1.6. Given a symmetric multinomial model $(M_t)_t$, we can construct a stochastic volatility model $(X_t)_t$ that approximates the former arbitrarily well. Due to reasons of symmetry, set $p := 1/2$, and let $(Y_t)_{t\geq 1} = (|M_t|)_{t\geq 1}$ be the Markov chain on the state space E^+ of positive volatility values. If the multinomial volatility is strictly positive almost surely, we are done. Otherwise, if the multinomial model can reach a volatility of zero with probability $p_0 > 0$, add a value of $\varepsilon > 0$ to the state space E^+ and set $P(y, \varepsilon) := p_0/2$ for each $y \in E^+ \cup \{\varepsilon\}$. Now since the underlying model is symmetric,

$$\mathbb{P}(|X_t - X_{t-1}| = \varepsilon) = 2\,\mathbb{P}(X_t - X_{t-1} = \varepsilon) = p_0.$$

These two models now differ at most by ε in each time step. So by letting $\varepsilon \to 0$, we have convergence in the following sense:

$$\forall t \geq 0 : \qquad |M_t - X_t| \to 0.$$

However, notice that the Markov chain induced by the multinomial is rather trivial (it is an independent, identically distributed sequence). By contrast, the family of stochastic volatility models is richer, allowing for 1-step dependencies in the Markov chain, which is beyond the scope of multinomials.
See also a visualization of this concept in Figure 3.3.

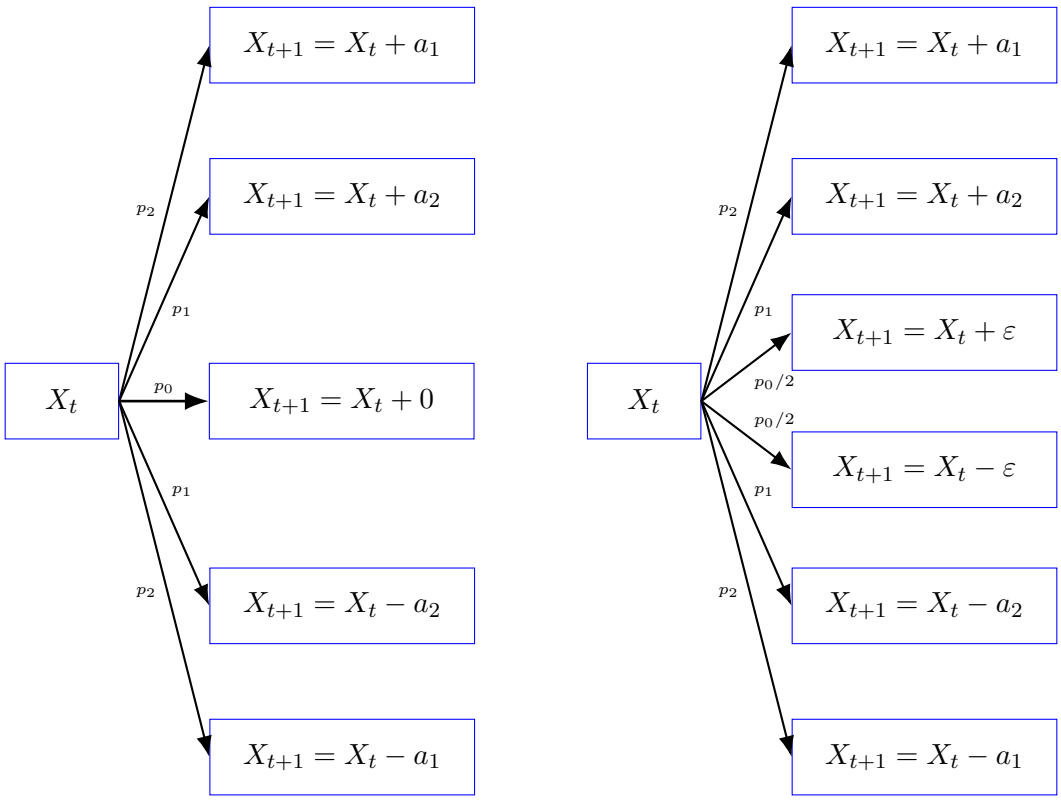

Figure 3.3: Comparison of a symmetric multinomial model (left) to its corresponding stochastic volatility model (right).

Lemma 3.1.7. *Suppose $X_0 = x_0 \in \mathbb{R}$ is known. Then $\mathcal{F}_t = \sigma(\xi_1, \dots, \xi_t, Y_1, \dots, Y_t)$ and $X_t = X_0 + \sum_{s \le t} \xi_s Y_s$.*

Proof. We will give an induction argument. The first step is trivial. Suppose we have proven the statement up to time $t-1$. If ξ_t and Y_t are known, then $X_t = X_{t-1} + \xi_t Y_t$ can be derived. Conversely, if X_t is known, we can compute $Y_t = |X_t - X_{t-1}|$ and $\xi_t = \operatorname{sign}(X_t - X_{t-1})$. □

3.2 Bayesian Approach

Let us suppose we have $J > 1$ (pairwise distinct) stochastic volatility models on a family of probability spaces $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \ge 0}, \mathbb{P}_j)_{j=1}^J$. The log-likelihood function of the particular model j should be

$$\begin{aligned}\ell_j(t) &= \sum_{k=1}^{t} \log \mathbb{P}_j(X_k = x_k \mid X_{k-1} = x_{k-1}, \dots, X_0 = x_0) \\ &= \sum_{k=1}^{t} \log \mathbb{P}_j(X_k = x_k \mid \mathcal{F}_{t-1}).\end{aligned}$$

Due to the independence of ξ_t from $\mathcal{F}_{t-1}$ and since Y_t is only dependent on Y_{t-1} as per the Markov property, we infer

$$\begin{aligned}\mathbb{P}_j(X_t = x_t \mid \mathcal{F}_{t-1}) &= \mathbb{P}_j(X_{t-1} + \xi_t Y_t = x_t \mid \mathcal{F}_{t-1}) \\ &= \mathbb{P}_j(\xi_t = \operatorname{sign}(x_t - x_{t-1}), Y_t = |x_t - x_{t-1}| \ \mid \mathcal{F}_{t-1}) \\ &= \mathbb{P}_j(\xi_t = \operatorname{sign}(x_t - x_{t-1}))\mathbb{P}_j(Y_t = |x_t - x_{t-1}| \ \mid Y_{t-1})).\end{aligned}$$

Let $Z_t := (Y_t, Y_{t-1})^\top$. Now we can rewrite the equation for $\ell_j(t)$ by substituting

$$\mathbb{P}_j(X_t = \cdot \mid \mathcal{F}_{t-1}) = F_j(\xi_t) G_j(Z_t),$$

where

$$F_j(x) = \begin{cases} p_j, & x = 1, \\ 1 - p_j, & x = -1, \end{cases} \qquad \text{and} \qquad G_j((\cdot, y)) = \mathbb{P}_j(Y_t = \cdot \mid Y_{t-1} = y).$$

Finally, we can evaluate the posterior distribution of each model j at time t by

$$\pi_j(t) = \frac{\exp(\ell_j(t))}{\sum_k \exp(\ell_k(t))}.$$

Now we have everything we need in order to construct a new model.

Definition 3.2.1 (Bayesian Stochastic Volatility Model)**.** Given $J > 1$ stochastic volatility models, the Bayesian stochastic volatility model is constructed as a stochastic volatility model by imposing

$$\mathbb{P}(\xi_{t+1} = 1) = 1 - \mathbb{P}(\xi_{t+1} = -1) = \bar{p}(t) := \sum_{j=1}^{J} \pi_j(t) p_j, \qquad t \ge 0,$$

$$\mathbb{P}(Y_{t+1} = \cdot \mid Y_t = y) = \sum_{j=1}^{J} \pi_j(t) G_j((\cdot, y)), \qquad y \in E, \ t \ge 0.$$

Chapter 4

Consistency

In this chapter, we analyze the Bayesian stochastic volatility model regarding its long-time behavior. At first, we show that if we actually have a model that describes the distribution of the observed data perfectly, our model disregards all other models in the long run. In the second and arguably more practical scenario, we suppose that no model can perfectly represent this distribution. We will see that the stochastic volatility model will converge to the model with the lowest entropy. One possible interpretation is as follows: if we were to consider each model invidually, we could quantify how much we would be "surprised" by the observed data. Our Bayesian model then will converge to the model that surprises us the least.

For an introduction to consistency and Bayesian statistics, we refer to [4].

Suppose the distribution underlying the observed data is given by

$$\mathbb{P}(X_t = \cdot \mid \mathcal{F}_{t-1}) = F(\xi_t)G(Z_t),$$

where

$$F(x) = \begin{cases} p, & x = 1, \\ 1-p, & x = -1, \end{cases} \qquad \text{and} \qquad G((\cdot, y)) = \mathbb{P}(Y_{t+1} = \cdot \mid Y_t = y).$$

We denote the unique invariant distribution of the Markov chain $(Y_t)_{t\geq 0}$ by ρ and the unique invariant distribution of the Markov chain $(Z_t)_{t\geq 1}$ by μ.

In order to begin, we need the notion of a "true model."

Definition 4.0.1 (True model)**.** Consider a Bayesian stochastic volatility model consisting of $J > 1$ submodels. If there exists a submodel j with $\mathbb{P}(X_t = x \mid \mathcal{F}_{t-1}) = \mathbb{P}_j(X_t = x \mid \mathcal{F}_{t-1})$ for all $x \in \mathbb{R}$ and $t \geq 1$, we call submodel j the true model or the model with the right parameters.

In the first section, we will analyze the case where a true model exists. Afterwards, we will assume the contrary. Finally, we will discuss the pricing of derivatives in either case.

4.1 True Model Exists

In this section, we will consider a collection of $J > 1$ stochastic volatility models and its corresponding Bayesian model. Suppose there exists one and only one true model, without loss of generality, say, model 1, with the true unique invariant distribution $\rho = \rho_1$ of the Markov chain $(Y_t)_{t\geq 0}$. By Proposition 2.2.14, the unique invariant distribution of the corresponding concatenated Markov chain $(Z_t)_{t\geq 0} = ((Y_{t+1}, Y_t))_{t\geq 0}$ is given by

$$\mu((y,\tilde{y})) = G_1((y,\tilde{y}))\rho_1(\tilde{y}) = G((y,\tilde{y})\rho(\tilde{y}).$$

At first, we need a short lemma to compute a limit later on.

Lemma 4.1.1. *If the observed values follow the parameters of submodel 1, then we have for all $k > 1$ that*

$$x_k := \lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \log \frac{F_k(\xi_s)G_k(Z_s)}{F_1(\xi_s)G_1(Z_s)} < 0.$$

Proof. We have

$$\begin{aligned} x_k &= \lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \log \frac{F_k(\xi_s)G_k(Z_s)}{F_1(\xi_s)G_1(Z_s)} \\ &= \lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \log \frac{F_k(\xi_s)}{F_1(\xi_s)} + \lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \log \frac{G_k(Z_s)}{G_1(Z_s)}. \end{aligned}$$

For the first term, note that by Theorem 2.2.19,

$$\lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \log \frac{F_k(\xi_s)}{F_1(\xi_s)} = \mathbb{E}_1\left[\log \frac{F_k(\xi_s)}{F_1(\xi_s)}\right].$$

Since log is a concave function, Lemma 2.3.4 gives

$$\begin{aligned} \mathbb{E}_1\left[\log \frac{F_k(\xi_s)}{F_1(\xi_s)}\right] &= p_1 \log \frac{p_k}{p_1} + (1-p_1)\log \frac{1-p_k}{1-p_1} \\ &< \log(p_k + 1 - p_k) \\ &= 0, \end{aligned}$$

where the strict inequality is due to the strict concavity of the logarithm and the fact that $p_k \neq p_1$. Likewise for the second term, Theorem 2.2.18 and Lemma 2.3.4

yield

$$\begin{aligned}\lim_{t\to\infty}\frac{1}{t}\sum_{s=1}^{t}\log\frac{G_k(Z_s)}{G_1(Z_s)} &= \mathbb{E}_\mu\left[\log\frac{G_k(Z_s)}{G_1(Z_s)}\right]\\ &= \sum_{\tilde{y}}\rho_1(\tilde{y})\sum_{y}G_1((y,\tilde{y}))\log\frac{G_k((y,\tilde{y}))}{G_1((y,\tilde{y}))}\\ &\leq \sum_{\tilde{y}}\rho_1(\tilde{y})\log\sum_{y}G_k((y,\tilde{y}))\\ &= 0.\end{aligned}$$

□

In the following theorem, we will show that a submodel which follows the real parameters exactly will always dominate all other submodels in the long run. This constitutes our first consistency proof.

Theorem 4.1.2. *Suppose model 1 is the unique true model. Then the Bayesian model converges to submodel 1, i.e.*

$$\lim_{t\to\infty}\pi_1(t) = 1.$$

Proof. Simple algebraic manipulation yields

$$\begin{aligned}\pi_1(t) &= \frac{\exp\left(\sum_s \log F_1(\xi_s)G_1(Z_s)\right)}{\sum_k \exp\left(\sum_s \log F_k(\xi_s)G_k(Z_s)\right)}\\ &= \left(\sum_k \exp\left(\sum_s \log\frac{F_k(\xi_s)G_k(Z_s)}{F_1(\xi_s)G_1(Z_s)}\right)\right)^{-1}\\ &= \left(1+\sum_{k>1}\exp\left(t\cdot\frac{1}{t}\sum_s\log\frac{F_k(\xi_s)G_k(Z_s)}{F_1(\xi_s)G_1(Z_s)}\right)\right)^{-1}.\end{aligned}$$

By Lemma 4.1.1, we know that letting $t\to\infty$, the exponent in each summand diverges to $-\infty$, and hence

$$\lim_{t\to\infty}\pi_1(t) = 1.$$

□

We will now go a step further and relax the assumption that the true model has to be unique. In the following, we suppose that multiple submodels, say the first $K \leq J$ submodels, consist of the right parameters. We proceed similarly as before, by computing a limit beforehand.

Lemma 4.1.3. *Let $k \leq K$. If the observed values follow the parameters of the first $K \leq J$ submodels, then*

$$x_j := \lim_{t\to\infty}\frac{1}{t}\sum_{s=1}^{t}\log\frac{F_j(\xi_s)G_j(Z_s)}{F_k(\xi_s)G_k(Z_s)}\begin{cases}=0, & j\leq K\\ <0, & j>K.\end{cases}$$

Proof. The first case is trivial. Let $j > K$, then resuming like we did before,

$$\begin{aligned} x_j &= \lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \log \frac{F_j(\xi_s)G_j(Z_s)}{F_k(\xi_s)G_k(Z_s)} \\ &= \mathbb{E}_k\left[\log \frac{F_j(\xi_s)}{F_k(\xi_s)}\right] + \mathbb{E}_\mu\left[\frac{G_j(Z_s)}{G_k(Z_s)}\right] \\ &= p_k \log \frac{p_j}{p_k} + (1-p_k)\log\frac{1-p_j}{1-p_k} \\ &\qquad + \sum_{\tilde{y}} \rho(\tilde{y}) \sum_{y} G_k((y,\tilde{y})) \log \frac{G_j((y,\tilde{y}))}{G_k((y,\tilde{y}))} \\ &< 0. \end{aligned}$$

□

Now for the proof of the more general case, where we show consistency for a supermodel with multiple submodels that follow the real parameters.

Theorem 4.1.4. *Suppose the first $K \leq J$ models consist of the right parameters. Then all other models will be disregarded in the long run, i.e.*

$$\lim_{t\to\infty} \sum_{j=1}^{K} \pi_j(t) = 1.$$

Proof. Let $j \leq K$, then

$$\begin{aligned} \pi_j(t) &= \frac{\exp\left(\sum_s \log F_j(\xi_s)G_j(Z_s)\right)}{\sum_k \exp\left(\sum_s \log F_k(\xi_s)G_k(Z_s)\right)} \\ &= \left(\sum_k \exp\left(\sum_s \log \frac{F_k(\xi_s)G_k(Z_s)}{F_j(\xi_s)G_j(Z_s)}\right)\right)^{-1} \\ &\to \left(1 + \ldots + 1 + \sum_{j>K} \exp(t\cdot x_j)\right)^{-1} \\ &= 1/K. \end{aligned}$$

It follows that the aggregated probabilities sum to 1.

□

Remark 4.1.5. If a (not necessarily unique) true model j exists, it trivially minimizes the entropy among all models:

$$\mathbb{H}(F \mid F_j) + \mathbb{H}(G \mid G_j) = \mathbb{H}(F \mid F) + \mathbb{H}(G \mid G) = 0.$$

4.2 General Case

What happens in the general case, where no submodel follows the real parameters exactly? It turns out that the minimizer of the reverse relative entropy is what the supermodel boils down to in the long run. Summing up, we can show consistency as well, where we approach the submodel that fits best to our data.

Theorem 4.2.1. *Suppose no submodel in the Bayesian stochastic volatility model has the right parameters. If model 1 uniquely minimizes*

$$\min_j \mathbb{H}(F \mid F_j) + \sum_{\tilde{y}} \rho(\tilde{y}) \;\; \mathbb{H}\big(G((\cdot, \tilde{y})) \mid G_j((\cdot, \tilde{y}))\big), \tag{4.1}$$

then the Bayesian model converges to submodel 1, i.e.

$$\lim_{t\to\infty} \pi_1(t) = 1.$$

Proof. We will try to reduce the problem to the simpler case where the statement is already proven.

(i) First, let us compute the entropy terms. By Example 2.3.2 and Example 2.3.8,

$$\mathbb{H}(F \mid F_1) = p \log \frac{p}{p_1} + (1-p) \log \frac{1-p}{1-p_1}.$$

Likewise, by Proposition 2.3.9,

$$\mathbb{H}(G((\cdot, \tilde{y})) \mid G_1((\cdot, \tilde{y}))) = \sum_y G((y, \tilde{y})) \log \frac{G((y, \tilde{y}))}{G_1((y, \tilde{y}))}.$$

(ii) Now we can rewrite equation (4.1): for all $k > 1$, we have

$$\begin{aligned} &p \log \frac{p}{p_1} + (1-p) \log \frac{1-p}{1-p_1} + \sum_{\tilde{y}} \rho(\tilde{y}) \sum_y G((y, \tilde{y})) \log \frac{G((y, \tilde{y}))}{G_1((y, \tilde{y}))} \\ &< p \log \frac{p}{p_k} + (1-p) \log \frac{1-p}{1-p_k} + \sum_{\tilde{y}} \rho(\tilde{y}) \sum_y G((y, \tilde{y})) \log \frac{G((y, \tilde{y}))}{G_k((y, \tilde{y}))}. \end{aligned}$$

Multiplying both sides by -1 and rearranging yields

$$\begin{aligned} &p \log \frac{p_k}{p} + (1-p) \log \frac{1-p_k}{1-p} + \sum_{\tilde{y}} \rho(\tilde{y}) \sum_y G((y, \tilde{y})) \log \frac{G_k((y, \tilde{y}))}{G((y, \tilde{y}))} \\ &- p \log \frac{p_1}{p} - (1-p) \log \frac{1-p_1}{1-p} - \sum_{\tilde{y}} \rho(\tilde{y}) \sum_y G((y, \tilde{y})) \log \frac{G_1((y, \tilde{y}))}{G((y, \tilde{y}))} < 0. \end{aligned} \tag{4.2}$$

Our goal is to proceed as in Lemma 4.1.1, that is we would like to show that $x_k < 0$. The statement then follows in the same way as in the simple case. As in the proof of Theorem 4.1.2,

$$\pi_1(t) = \left(1 + \sum_{k>1} \exp\left(t \cdot \frac{1}{t} \sum_s \log \frac{F_k(\xi_s) G_k(Z_s)}{F_1(\xi_s) G_1(Z_s)}\right)\right)^{-1}.$$

For the same reasons as before, namely the Law of Large Numbers and the Ergodic Theorem for Markov Chains, we have almost surely

$$\frac{1}{t} \sum_{s=1}^{t} \log \frac{F_k(\xi_s) G_k(Z_s)}{F_1(\xi_s) G_1(Z_s)} \xrightarrow[t\to\infty]{} x_k := \mathbb{E}_1\left[\log \frac{F_k(\xi)}{F_1(\xi)}\right] + \mathbb{E}_\mu\left[\log \frac{G_k(Z)}{G_1(Z)}\right].$$

(iii) Finally, in order to show that this term is indeed negative, note that

$$\begin{aligned} x_k &= \mathbb{E}_1\left[\log\frac{F_k(\xi)}{F(\xi)}\right] + \mathbb{E}_\mu\left[\log\frac{G_k(Z)}{G(Z)}\right] - \mathbb{E}_1\left[\log\frac{F_1(\xi)}{F(\xi)}\right] - \mathbb{E}_\mu\left[\log\frac{G_1(Z)}{G(Z)}\right] \\ &= p\log\frac{p_k}{p} + (1-p)\log\frac{1-p_k}{1-p} - p\log\frac{p_1}{p} - (1-p)\log\frac{1-p_1}{1-p} \\ &\quad + \sum_{\tilde{y}}\rho(\tilde{y})\sum_{y}G((y,\tilde{y}))\left(\log\frac{G_k((y,\tilde{y}))}{G((y,\tilde{y}))} - \log\frac{G_1((y,\tilde{y}))}{G((y,\tilde{y}))}\right) \\ &< 0 \end{aligned}$$

as we have seen in equation (4.2).

□

4.3 Option Pricing

In this section, we would like to expand the scope of our model, that is, we also want to factor in financial derivatives when considering which model is the most accurate. We will still consider Stochastic Volatility Models, however now our submodel likelihoods can also change based on the observation of the financial derivatives.

Suppose we observe $A \geq 1$ financial derivatives $f_1, \ldots, f_A$. By Lemma 3.1.7, the observation of the X_t is equivalent to observing ξ_t and Y_t, so at time $t \geq 0$, we have A random variables

$$f_1((\xi_{t+u+1}, Y_{t+u})_{u\geq 1}), \ldots, f_A((\xi_{t+u+1}, Y_{t+u})_{u\geq 1}).$$

Observe that we have assumed that these options do not depend on X_t explicitly. Moreover, suppose for all models j and the true model that

$$\mathbb{P}(\xi_t = 1) = \mathbb{P}_j(\xi_t = 1) = 1/2.$$

Then since ξ and Y are independent, all models are martingales:

$$\mathbb{E}_j[X_{t+1} \mid \mathcal{F}_t] = X_t + \mathbb{E}_j[\xi_{t+1}]\,\mathbb{E}_j[Y_{t+1} \mid \mathcal{F}_t] = X_t.$$

Furthermore, due to the specific structure of our model, the asset's future price increments are independent of the observed current price. Hence under model j, the price of option a should be

$$\varphi_a^j(t) := \mathbb{E}_j[f_a((\xi_{t+u+1}, Y_{t+u})_{u\geq 1}) \mid \mathcal{F}_t] = \mathbb{E}_j[f_a((\xi_{t+u+1}, Y_{t+u})_{u\geq 1})] = h_a^j(Y_t).$$

Let us denote by $\Phi_a(t)$ the observed price of option a at time t; in particular, this is the price of the option under any true model. Now we define our new likelihood as

$$\ell_j(t) = \sum_{k=1}^{t}\log(\mathbb{P}_j(X_k \mid \mathcal{F}_{k-1})) - Q(\varphi^j(k)),$$

$$Q(\varphi^j(t)) = \sum_{a=1}^{A}|\varphi_a^j(t) - \Phi_a(t)|^2.$$

Again, we would like to show that this kind of model is consistent. We recall the approach from before and postulate that by minimizing the reverse relative entropy regarding the observed data, we have found the best model available to us. The following theorem states that our supermodel does indeed converge to this very submodel.

Theorem 4.3.1. *If submodel 1 is the unique minimizer of*

$$\min_j \sum_{\tilde{y}} \rho(\tilde{y}) \Big(\mathbb{H}\left(G((\cdot,\tilde{y})) \mid G^j((\cdot,\tilde{y}))\right) + Q(h_a^j(\tilde{y})) \Big), \tag{4.3}$$

then the Bayesian model converges to submodel 1, i.e.

$$\lim_{t\to\infty} \pi_1(t) = 1.$$

Proof. We are now able to disregard the distribution of ξ though we still have to account for the option price differences:

(i) We recall a previous result regarding the entropy:

$$\mathbb{H}(G((\cdot,\tilde{y})) \mid G_1((\cdot,\tilde{y}))) = \sum_y G((y,\tilde{y})) \log \frac{G((y,\tilde{y}))}{G_1((y,\tilde{y}))}.$$

(ii) Let us now rewrite equation (4.3): for $j > 1$,

$$\begin{aligned} &\sum_{\tilde{y}} \rho(\tilde{y}) \Big(\sum_y G((y,\tilde{y})) \log \frac{G((y,\tilde{y}))}{G_1((y,\tilde{y}))} + Q(h_a^1(\tilde{y})) \Big) \\ &\quad < \sum_{\tilde{y}} \rho(\tilde{y}) \Big(\sum_y G((y,\tilde{y})) \log \frac{G((y,\tilde{y}))}{G_j((y,\tilde{y}))} + Q(h_a^j(\tilde{y})) \Big). \end{aligned}$$

(iii) Now our submodel probability becomes

$$\pi_1(t) = \left(1 + \sum_{k>1} \exp\left(\sum_s \log \frac{G_k(Z_s)}{G_1(Z_s)} - Q(\varphi^k)(s) + Q(\varphi^1)(s) \right) \right)^{-1}.$$

(iv) Using the Ergodic Theorem twice, we obtain

$$\begin{aligned} x_j &:= \lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^t \left(\log \frac{G_j(Z_s)}{G_1(Z_s)} - Q(\varphi^j)(s) + Q(\varphi^1)(s) \right) \\ &= \mathbb{E}_\mu \left[\log \frac{G_j(Z_s)}{G_1(Z_s)} \right] + \lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^t \left(Q(\varphi^1(t)) - Q(\varphi^j(t)) \right) \\ &= \mathbb{E}_\mu \left[\log \frac{G_j(Z_s)}{G_1(Z_s)} \right] + \mathbb{E}_\rho \left[Q(\varphi^1(t)) - Q(\varphi^j(t)) \right]. \end{aligned}$$

We are left to show that $x_j < 0$ for $j > 1$. Indeed,

$$\mathbb{E}_\mu\left[\log\frac{G_j(Z_s)}{G_1(Z_s)}\right] = \sum_{\tilde{y}}\rho(\tilde{y})\sum_{y}G((y,\tilde{y}))\log\frac{G_j((y,\tilde{y}))}{G_1((y,\tilde{y}))},$$
$$\mathbb{E}_\rho\left[Q(\varphi^j(t))\right] = \sum_{\tilde{y}}\rho(\tilde{y})Q(h_a^j(\tilde{y})),$$

and finally comparing the terms for different models yields the desired inequality. Proceeding in a similar fashion as before concludes the proof.

□

Chapter 5

Implementation

In this chapter, we provide results of a simple implementation of the presented Bayesian stochastic volatility model. For the underlying stock data, we used real daily Siemens AG opening stock prices [11] from August 17, 2021 up to August 16, 2022. For the Python [10] code, we refer to Appendix A. We utilized the Numpy [8], Pandas [9] and Matplotlib [5] packages for the subsequent plots.

At first, here is the evolution of the Siemens AG stock price over one year:

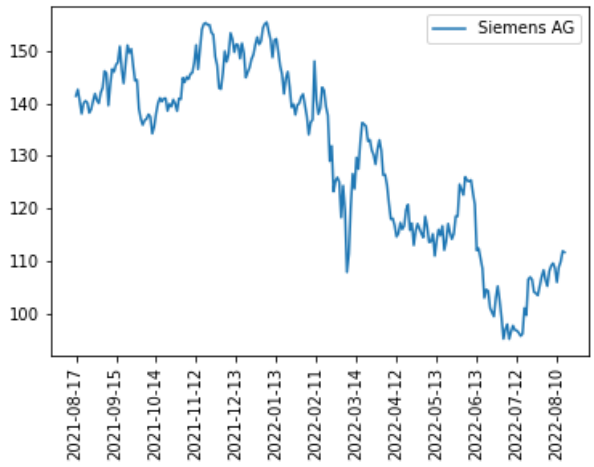

Figure 5.1: Siemens AG opening stock price in EUR.

We considered three stochastic volatility models, without any accountance of financial derivatives, i.e. $Q \equiv 0$. The parameters for the signs ξ_t are

$$p_1 = 0.4, \quad p_2 = 0.6, \quad p_3 = 0.85.$$

The Markov chains Y_t act on the state space $E = \{0.001,\ 0.003,\ 0.005,\ 0.01,\ 0.02\}$

and are subject to the following transition probabilities:

$$P_1 = \begin{bmatrix} 0.425 & 0.05 & 0.05 & 0.05 & 0.425 \\ 0.425 & 0.05 & 0.05 & 0.05 & 0.425 \\ 0.425 & 0.05 & 0.05 & 0.05 & 0.425 \\ 0.425 & 0.05 & 0.05 & 0.05 & 0.425 \\ 0.425 & 0.05 & 0.05 & 0.05 & 0.425 \end{bmatrix},$$

$$P_2 = \begin{bmatrix} 0.3 & 0.05 & 0.3 & 0.05 & 0.3 \\ 0.3 & 0.05 & 0.3 & 0.05 & 0.3 \\ 0.3 & 0.05 & 0.3 & 0.05 & 0.3 \\ 0.3 & 0.05 & 0.3 & 0.05 & 0.3 \\ 0.3 & 0.05 & 0.3 & 0.05 & 0.3 \end{bmatrix},$$

$$P_3 = \begin{bmatrix} 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{bmatrix}.$$

We end up with the following dynamic posterior model probabilities (note the log-scale):

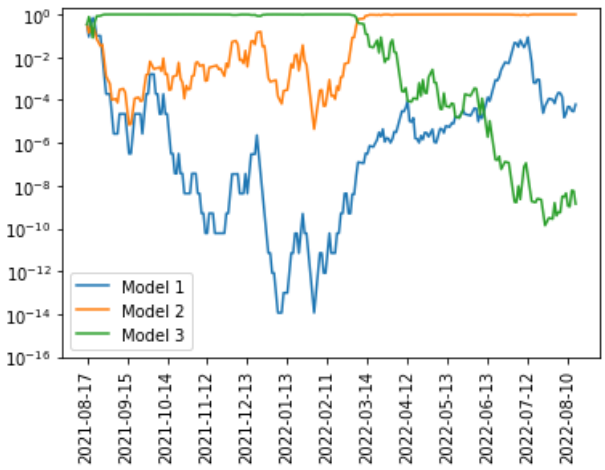

Figure 5.2: Evolution of posterior probabilities π_j over time.

In summary, we can see that the Bayesian approach allows for a diverse class of Stochastic Volatility Models to be considered that can fit to various situations. In the first half of the year, model 3 seems to be our best choice, but over time we realize model 2 has become more accurate. Whether this is due to seasonal reasons or complex global phenomena is up to the economist to decide, however we can clearly see the advantage that our Bayesian approach can handle different scenarios at once.

Chapter 6

Conclusion

In this thesis, we used the approach of [2] to construct a series of more and more sophisticated models and provided new consistency proofs.

We first enriched the classical binomial model by replacing its constant volatility with a Markov chain Y_t. By proceeding in this manner, we obtained a time-dependent volatility of random magnitude, something a simple binomial model typically lacks. We called the resulting model a Stochastic Volatility Model.

In a further step, we took a fixed number of SVMs and created a Bayesian SVM, a model that gives different weights to different SVMs according to the observed data, and combines them into one single model. This is a classical Bayesian approach, hence the name. For the Bayesian SVM, we showed consistency under various assumptions, showing that we must not necessarily have a perfect-fitting model in order to obtain convergence. Motivated by [4] and in order to formalize the notion of best fitting, we introduced the concept of reverse relative entropy. By doing so, we could measure which model describes the observed data most accurately, yielding a time-dependent measure $\pi_j(t)$.

In the end, we also considered observations of financial derivatives in order to assess our models more accurately. Under additional assumptions, we also showed consistency in a martingale case, which means that our asset moves up or down with the same probability. We have gone from a simple binomial model to a complex model with multiple decent features, resulting in a method that can fit to various situations. When we are facing different market situations, our model adapts accordingly and decides by itself how much different data should be valued.

Finally, we gave an illustrative example to demonstrate what this approach could look like in real life. Now, what is next? It could be interesting to study the consideration of financial derivatives in a more general setting, where we dismiss the martingale scenario. Moreover, the choice of Q, the error-term for our derivative observations, could be altered. Another idea is to follow [2], describing an approach where the historical data is weighted by their recency, which could lead to further consistency proofs or other crucial results.

Bibliography

[1] J. Cerny. Advanced probability theory. Lecture Notes, University of Vienna, 2016.

[2] M. Duembgen and L.C.G. Rogers. Estimate nothing. *Quantitative Finance*, 14(12):2065–2072, 2014.

[3] N. Etemadi. An elementary proof of the strong law of large numbers. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 55:119–122, 1981.

[4] S. Ghosal and A. van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, Cambridge, 2017.

[5] J. Hunter, D. Dale, E. Firing, M. Droettboom and the Matplotlib development team. *Matplotlib 3.5.3 documentation*. https://matplotlib.org/stable/index.html, 2012-2022.

[6] T. Kraler. Konsistenz im Rahmen des bayesschen Lernens in der Finanzmathematik. University of Vienna, 2022. Bachelor thesis.

[7] J.R. Norris. *Markov Chains*. Cambridge University Press, Cambridge, 15th edition, 2009.

[8] NumPy Developers. *NumPy documentation*. https://numpy.org/doc/stable/, 2005-2022.

[9] Pandas. *Pandas documentation*. https://pandas.pydata.org/docs/, 2022.

[10] Python Software Foundation. *Python 3.10.6 documentation*. https://docs.python.org/3/, 2022.

[11] Yahoo Finance. Siemens Aktiengesellschaft (SIE.DE) - Historical Data. https://finance.yahoo.com/quote/SIE.DE/history?p=SIE.DE. Last visited 2022/08/29, 2:41 p.m.

Appendix A

Python Code

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

#importing stock data
data = pd.read_csv('siemens_data.csv')
data.head()

length = len(data)

#setting parameters for Xi
par_xi = [0.4, 0.6, 0.85]

#State space for Markov chain Y
E = np.array([0.001, 0.003, 0.005, 0.010, 0.020])

def closest_state(y):
    diff = np.abs(np.abs(y) - E)
    minimizer = np.argmin(diff)
    return minimizer

def F(xi, i):
    res = par_xi[i]
    if xi == -1:
        res = 1 - res
    return res

def G(y, y0, i):
    index0 = int(np.where(E == y0)[0])
    index1 = int(np.where(E == y)[0])
    ret = 0
    if i == 0:
        ret = P1[index0, index1]
    elif i == 1:
        ret = P2[index0, index1]
    elif i == 2:
        ret = P3[index0, index1]
    return ret
```

```
def sign(x):
    if x < 0:
        ret = -1
    else:
        ret = 1
    return ret


#Transition probabilities for Markov chain Y

P1 = np.array([[0.425, 0.05, 0.05, 0.05, 0.425],
               [0.425, 0.05, 0.05, 0.05, 0.425],
               [0.425, 0.05, 0.05, 0.05, 0.425],
               [0.425, 0.05, 0.05, 0.05, 0.425],
               [0.425, 0.05, 0.05, 0.05, 0.425]])

P2 = np.array([[0.3, 0.05, 0.3, 0.05, 0.3],
               [0.3, 0.05, 0.3, 0.05, 0.3],
               [0.3, 0.05, 0.3, 0.05, 0.3],
               [0.3, 0.05, 0.3, 0.05, 0.3],
               [0.3, 0.05, 0.3, 0.05, 0.3]])

P3 = 1./5 * np.ones((5,5))


stock = np.log(data['Open'])

likely = np.zeros((length, 3))
likely[0][1] = 1./3
likely[0][2] = 1./3
likely[0][0] = 1./3

y0 = 0.001


#Actual calculation
for t in range(1, length):
    incr = stock[t] - stock[t-1]
    y1_index = closest_state(incr)
    y1 = E[y1_index]
    xi = sign(incr)

    for i in range(3):
        likely[t][i] = likely[t-1][i] + np.log(F(xi, i) * G(y1, y0, i))

    y0 = y1


#computing posterior distribution
posterior = np.zeros((length, 3))
posterior[0] = 1./3
for i in range(1, length):
```

```
    for j in range(3):
        posterior[i][j] = np.exp(likely[i][j])/(np.exp(likely[i][0]) +
            np.exp(likely[i][1]) + np.exp(likely[i][2]))


#creating labels for plot
xlabel = []
for i in range(len(data)):
    if i % 21 == 0:
        xlabel.append(data['Date'][i])


#plotting stock data
plt.plot()
plt.plot(data['Date'], data['Open'], label="Siemens AG")
plt.xticks(xlabel, rotation=90)
plt.legend()
plt.show()

#plotting posterior distribution
plt.plot()
plt.semilogy(data['Date'], posterior[:,0], label="Model 1")
plt.semilogy(data['Date'], posterior[:,1], label="Model 2")
plt.semilogy(data['Date'], posterior[:,2], label="Model 3")
plt.ylim(10**-16, 2)
plt.xticks(xlabel, rotation=90)
plt.legend()
plt.show()
```