Investment Strategy for VIX Futures based on a Bayesian Approach to Term Structure Modelling

Oskar Oostdam

Technische Universiteit Delft

TU Delft

Investment Strategy for VIX Futures based on a Bayesian Approach to Term Structure Modelling

by

Oskar Oostdam

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Wednesday August 25, 2021 at 1:30 PM.

Student number:	4574281	
Project duration:	November 16, 2020 – August 25, 2021	
Thesis committee:	Dr. Ir. J. Bierkens,	TU Delft, supervisor
	Dr. A. Papapantoleon,	TU Delft
	Dr. C.S.L. de Graaf,	Aegon Asset Management, supervisor

Abstract

The VIX index, which is the expected volatility of the S&P 500 index in 30 days, is of interest to a lot of investors on the US financial market. Allowing the volatility of the financial market to be used as a trading tool gives rise to interesting investment opportunities, such as hedging and speculation. In this thesis we will be creating an investment strategy on VIX futures by modelling the term structure with a Bayesian approach. Using Markov Chain Monte Carlo (MCMC) methods, we will simulate a posterior distribution of our term structure, which results in credible intervals of our futures prices. We will compare the efficiency of Metropolis-Hastings algorithms against the No-U-Turn Sampler, which is a Hamiltonian Monte Carlo algorithm. Eventually we find that the No-U-Turn Sampler significantly outperforms the Metropolis-Hastings algorithms. The resulted credible intervals of our futures prices will be used to determine whether a contract is overvalued or undervalued. The strategy consists of taking a combination of long and short positions on VIX futures contracts which we consider to be mispriced. We will therefore take a long position on undervalued VIX futures, while taking a short position on overvalued VIX futures. We eventually find that this investment strategy is very risky due to the high volatile behaviour of the VIX index.

Contents

1 Introduction 1

2 The VIX Index 3
2.1 Calculation of the VIX index . . . 4
2.2 VIX products . . . 6
2.2.1 VIX futures . . . 7
2.3 Investment strategies with the VIX . . . 9

3 Bayesian Parameter Estimation and Markov Chain Monte Carlo Methods 10
3.1 An introduction to Bayesian statistics . . . 10
3.1.1 Choosing a prior distribution . . . 12
3.2 Metropolis-Hastings algorithm . . . 13
3.2.1 Independent Metropolis-Hastings algorithm . . . 14
3.2.2 Random walk Metropolis-Hastings algorithm . . . 14
3.2.3 Gibbs sampler . . . 16
3.3 Hamiltonian Monte Carlo algorithm . . . 19
3.3.1 Hamiltonian dynamics . . . 20
3.3.2 Construction of the algorithm . . . 21
3.3.3 No-U-Turn Sampler . . . 23

4 Modelling the VIX Term Structure of Futures 28
4.1 Stylized facts about volatility and the VIX index . . . 28
4.2 Term structure model . . . 29
4.2.1 Derivation of the futures contract price . . . 30
4.2.2 Bayesian parameter estimation . . . 36
4.2.3 Metropolis-Hastings sampling scheme . . . 39
4.2.4 Drawbacks of the term structure model . . . 40
4.3 Error correction model . . . 42
4.3.1 Model formulation . . . 43
4.3.2 Bayesian parameter estimation . . . 44
4.4 Strategy . . . 46
4.4.1 Volatility targeting . . . 47
4.4.2 Sharpe ratio . . . 47
4.4.3 Evaluation methods . . . 48

5 Results 51
5.1 Term structure model simulations . . . 52
5.1.1 Metropolis-Hastings algorithm . . . 52
5.1.2 No-U-Turn Sampler . . . 56
5.1.3 Simulation results of the data set . . . 59
5.2 Error correction model simulations . . . 60
5.2.1 Gibbs Sampler . . . 61
5.2.2 No-U-Turn Sampler . . . 64
5.2.3 Simulation results of the data set . . . 66
5.3 Strategy results . . . 68
5.3.1 Credible region selection . . . 70
5.3.2 Volatility targeting . . . 72
5.3.3 Further evaluation of the strategy . . . 74

6 Discussion 78

1 Introduction

Financial volatility is a fundamental element when it comes to trading on financial markets. Although many different definitions have been given to volatility, Poon (2005) nicely describes volatility as the spread of all likely outcomes of an uncertain variable (p. 1). Generally, regarding mathematical definitions of volatility, the standard deviation of the uncertain variable is used as volatility. Understanding the volatile behaviour of certain financial instruments allow investors and strategists to identify the amount of risk they are exposed to. For instance, when trading on the stock market, the investor would like to understand the uncertainty behaviour of the stock before he wishes to invest. This way the investor can consider investing in the stock based on the amount of risk he is exposed to, given a certain expected return of the stock.

To give investors more insight in the expected volatile behaviour of the market, the Chicago Board Options Exchange (CBOE) introduced in 1993 the volatility index called the VIX, which was designed by Whaley (1993). This volatility index is a representation of the expected volatility of the S&P 500 index in 30-days, which is why the VIX index is also commonly known as the "investor fear gauge". The calculation of the VIX index is based on the value of the available options that are currently traded on the market. One way to extract volatility based on option prices is to assume a certain option pricing model beforehand. By knowing the current market price of an option, it is possible to look at what the input of the volatility in the option pricing model should be to result in the current market price. Extracting volatility this way is also known as implied volatility. The first version of the VIX index in 1993 was based upon the implied volatility of the available options of the S&P 500 index, but later in 2003 the methodology changed and was based on the expected variance swap rate, which is the strike price of a certain derivative on the realized variance of the S&P 500 index. This methodology also uses the market prices of the available options of the S&P 500 index to calculate the VIX index, but in a different way.

Although the VIX gives an idea of the expected volatile behaviour of the market, it also allows for various strategic opportunities for investors. The VIX index itself is not an asset one can simply buy on the market. However, derivatives and ETFs (Exchange-Traded Funds) of the VIX index are widely available. Lots of different types of futures and options are traded on the CBOE, which allow investors to correlate their portfolio with the VIX index or even speculate on the volatility of the market. It is well-known that the volatility and stock price have a negative correlation. When the volatility is high and the market is uncertain, the stock price often seems to decline. The existence of negative correlation between the VIX index and the S&P 500 index has been shown as well, which allow investors to diversify their portfolio with VIX derivatives (Daigler & Rossi, 2006). Including VIX derivatives in a portfolio with stocks that are in the S&P 500 index may decrease the downfall of the portfolio value when the S&P 500 index declines. When the value of the stocks in the portfolio decrease, the VIX index could increase due to the negative correlation with the S&P 500 index. Hence, adding a long position in call options of the VIX index could greatly reduce the risk of this portfolio.

In this thesis we will attempt to create an investment strategy on VIX futures. We will first create a model for the term structure of VIX futures, which then allows us to design an investment strategy based on our model for the term structure. We begin by creating a model for the term structure based on the theoretical price of a futures contract, which we refer to as the "Term Structure model". Then we will introduce a second model that attempts to fix the shortcomings of the initial term structure model, which we call the "Error Correction model". For both models we will use a Bayesian approach for the estimation of the parameters. This allows us to obtain a credible interval of our VIX futures term structure. Using these intervals that our models provide, we will create an investment strategy that tells us when to take a long or short position in a certain VIX futures contract.

The research is conducted in cooperation with Aegon Asset Management, which is a brand of the financial services company Aegon N.V. The motivation behind this research was a request for an Alternative Risk Premia (ARP) investment strategy using VIX futures. In general, risk premium is the extra return for an investor after taking a certain risk. Now, alternative risk premia strategies are concerned with risk premia that are an alternative to traditional risk premia investments such as equity or fixed-income (Hamdan et al., 2016). These strategies therefore attempt to create a profit in a non-traditional way, which are usually in a form of a long/short investment. Our investment strategy will attempt to create a profit by taking long positions on undervalued VIX futures and short positions on overvalued VIX futures.

We will begin this thesis by giving more information of the VIX index in Chapter 2. There we explain how the VIX index is exactly calculated, what kind of financial products of the VIX index are traded and how the VIX index is used for investment strategies. In Chapter 3 we will give a general introduction to Bayesian parameter estimation and Markov Chain Monte Carlo (MCMC) methods. The MCMC methods we will focus on are the Metropolis-Hasting algorithm and the Hamiltonian Monte Carlo algorithm. In Chapter 4 we will introduce the term structure model, the error correction model and the investment strategy. In Chapter 5 we will go over the simulations of the models and evaluate the results of our investment strategy. Finally, Chapter 6 will contain a discussion and recommendations for further research.

2 The VIX Index

When the VIX was first introduced in 1993, the calculation of the VIX index was based on the implied volatility of the 30-day S&P 100 index at-the-money options, which are options that expire in 30 days and have a strike price identical to the current market price. The assumed underlying option pricing model that was used to derive the implied volatility was the well-known Black-Scholes model by Black and Scholes (1973). As Whaley (2009) explains, the option market back then behaved in a different way. The S&P 100 index options were the most actively traded options in the United States with taking 75% of the total index option volume in 1992, while the S&P 500 index options only took up 16.1% of total volume (see Figure 2.1[1]). Not only the S&P 100 index options were the most popular options, but also specifically at-the-money options were actively traded. Typically options with a strike price far away from the current stock price were not traded as often, which we call out-of-the money options when it has no intrinsic value (the value of the option if it would be exercised at this moment). This could lead to these options having large bid-ask spreads and stale price quotes, hence including out-of-the money options in the VIX calculation could lead to an inaccurate representation of the expected market volatility.

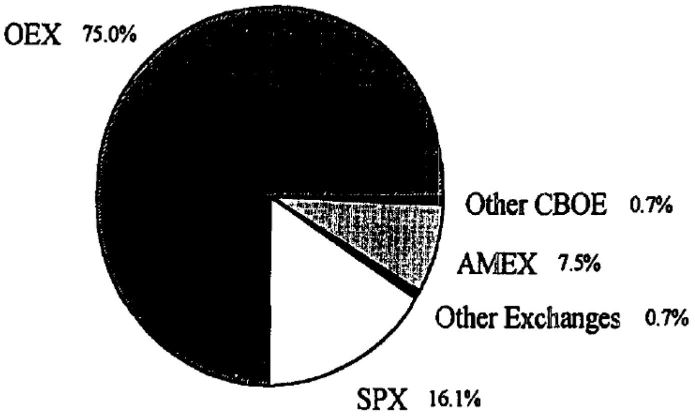


Figure 2.1: A comparison of proportions of the option volume of popular indices in 1992 (Whaley, 1993, p. 72).

Over the next 10 years, the behaviour of investors on index option market in the United States was changing. Whaley (2009) explains that in this period there were two fundamental changes in the market. First, the popularity of the S&P 500 index options started increasing over the years. To give an idea on how the interest of investors in these options increased, during the first 10 months of 2008 the S&P 500 index options were traded about 12.7 times as frequently as S&P 100 index options (Whaley, 2009, p. 99). It is not entirely known why the increase in popularity of these options occurred, but one detail that might have had an influence is the difference in option style between these two indices. Since the most actively traded option contracts of the S&P 100 index were American-style, which means that the holder of this option can exercise the contract at any time before the expiration date, it may be harder for investors to value these type of contracts. Since the most popular options on the S&P 500 index are European-style, which means that the holder of this option can only exercise the contract on the expiration date, these types of contracts are easier to value and could be more favoured among investors.

The second fundamental change in the option market was the adjustment in trading motives of investors. In 1992 the trading volumes of S&P 100 index call and put options were approximately balanced. On a daily average, the S&P 100 call option contracts were traded 120,475 times and the put option contracts 125,302 times (Whaley, 2009, p. 99). In the following years, this balance started to shift more towards the put option contracts. Investors were now adding more at-the-money and out-of-the-money put option contracts to their portfolio to insure themselves against possible declines in the stock market. To give an idea of how the balance between put and call options shifted over time, during the first 10 months of 2008 the average amount of traded S&P 500 call option contracts were 525,460 and

[1] The OEX is the ticker symbol of the S&P 100 index and the SPX is the ticker symbol of the S&P 500 index.

for put options 909,748 contracts (Whaley, 2009, p. 99). This makes the demand of put options in 2008 to be over 73% higher than the demand of call options. Since the difference in volume between call and put options was not taken into account in the calculation of the VIX, the index was becoming more and more inaccurate and needed a transformation.

These two fundamental changes in the option market motivated the CBOE to start developing a new methodology for the calculation of the VIX index. In 2003 the CBOE introduced a new calculation method for the VIX index in an attempt to improve the representation of the expected 30-day volatility of the U.S. markets. One important change was that the VIX index is now derived of the options prices of the S&P 500 index instead of the S&P 100 index. Furthermore, now the out-of-the money options are taken into account in the calculation of the VIX index as well, since specifically the out-of-the money put option prices give a lot of information regarding expected market volatility. When the demand and value of out-of-the money put option contracts increase, then investors are trying to increase the insurance against a larger decline in the S&P 500 index, which implies more uncertainty among investors. The lower the strike price of the put option, the greater of a decline in the market is expected.

At first, only the monthly options were used in the calculation of the VIX index, but in 2014 also the weekly options got included. The VIX index uses interpolation of the two options with the expiration dates closest to 30 days. Adding weekly S&P 500 options in the calculation of the VIX gives a more accurate representation of the volatility, since the time interval between the expiration date of the two options used in the interpolation now becomes one week instead of one month. A more obvious reason for including weekly options was the removal of the one week in each month where extrapolation was used in the VIX calculation. One requirement for the two options that were chosen for the VIX calculation was that the expiration date had to be at least one week or later due to the different option price behaviour near maturity. When there was a monthly option contract on the market expiring in less than 1 week, the two options that were used for the VIX calculation both expired in more than 30 days, hence extrapolation was used. Till this day the methodology developed in 2003 has remained unchanged and is still used as the current VIX index calculation.

2.1 Calculation of the VIX index

The current calculation of the VIX index is an approximation of the square root of the risk-neutral expectation of the realized variance of the S&P 500 index over the next 30 days. Realized variance is the variance of a stock price based on historical data given a certain period. In mathematical terms, if we define the variance of a stock price to be σ_t^2 at time t, then the realized variance over a period from 0 to T would simply be:

$$RV_T = \frac{1}{T}\int_0^T \sigma_t^2 dt.$$

The VIX index attempts to approximate the square root of the risk-neutral expectation of the realized variance in 30 days, hence

$$VIX^2 \approx \mathbb{E}_{\mathbb{Q}}[RV_{30}], \tag{2.1}$$

with $\mathbb{Q}$ as the risk-neutral measure. The risk-neutral measure $\mathbb{Q}$ makes sure that the price of all assets under this measure is its expected discounted payoff. This allows for the pricing of financial products under the assumption that there is no arbitrage on the market and that the market is complete. Arbitrage is the situation when a portfolio with no value has zero probability to lose value, while having a strictly positive probability of increasing in value. This phenomenon is sometimes also referred to as "free lunch". Furthermore, a market is complete when every risk position can be replicated with available securities. Since each individual asset has its own risk profile, the pricing of assets with real-world probabilities would need to be done separately for each asset. However, under the assumption that we have an arbitrage-free and complete market, the risk-neutral measure allows us to price all assets under the same measure.

The same way of looking at (2.1) is by comparing it to the risk-neutral valuation of a financial security. In this case here, the risk-neutral expectation of the realized variance is equivalent as the fair variance strike price of a 30-day variance swap. A variance swap is a type of derivative that allows the holder to receive a payoff of the realized variance minus some prearranged variance strike. Using the risk-neutral measure, this type of contract can simply be valued as

$$V = \mathbb{E}_{\mathbb{Q}}[e^{-30r}(RV_{30} - K)],$$

with r the risk-free interest rate. The variance strike price, also known as the variance swap rate, would be considered a fair price when $V = 0$, hence K as in (2.1) would be the fair variance swap rate. Therefore, it is also said that the squared VIX approximates the 30-day variance swap rate. Demeterfi et al. (1999) show that the fair variance swap rate can be expressed in terms of out-of-the money options as (p. 19):

$$\mathbb{E}_{\mathbb{Q}}[RV_T] = \frac{2}{T}\left(e^{rT}\int_0^{S_*}\frac{P(K)}{K^2}dK + e^{rT}\int_{S_*}^{\infty}\frac{C(K)}{K^2}dK + \epsilon\right), \tag{2.2}$$

with

$$\epsilon = \log\left(\frac{S_0 e^{rT}}{S_*}\right) - \left(\frac{S_0 e^{rT}}{S_*} - 1\right).$$

Here $P(K)$ and $C(K)$ represent the price of a put and call option of the underlying stock S with strike price K respectively. Furthermore, S_0 represents the spot price and S_* the strike price that is the boundary of out-of-the money call options and out-of-the money put options. Note that when this boundary is chosen as the risk-neutral forward price $S_* = e^{rT}S_0$, then $\epsilon = 0$.

The current VIX index calculation is based on the result of (2.2). Since there are only a discrete amount of strike prices for option contracts available on the market, the calculation uses a numerical approximation with the tradable option contracts to approximate the integrals. The general formula used for the calculation of the current VIX index using options with expiry date T is (CBOE, 2019):

$$\sigma^2 = \frac{2}{T}\sum_i \frac{\Delta K_i}{K_i^2} e^{rT} Q_t(K_i) - \frac{1}{T}\left(\frac{F}{K_0} - 1\right)^2, \tag{2.3}$$

with

- T: Time to expiration. The time to expiration is taken in calendar days and is adjusted each minute. The value of T is the proportion of the amount of minutes to expiration in a year, hence
$$T = \frac{\text{Minutes till expiration}}{\text{Minutes in a year}}.$$
- r: Risk-free interest rate to expiration. The risk-free interest rate used for the VIX index calculation is based on the U.S. Treasury yield curve rates, which is the annual return of U.S. Treasury bonds based on the current bond prices. To obtain a representation of the interest rate to a specific expiration date, a cubic spline interpolation of the yield curve is used.
- F: Forward index level. This is the expected spot price of the VIX at the expiration date based on option contracts. For a certain expiry date T the forward index is derived from a pair of call and put options where their midpoint prices (the middle of the bid and ask price) have the smallest absolute difference between them. Then using the put-call parity, one arrives at the equation:
$$F = K + e^{rT}(C - P).$$
- K_0: The first strike price below the forward index level F.
- K_i: The strike price of the ith out of the money of option. The strike belongs to the call option when $K_i > K_0$, to the put option when $K_i > K_0$ and both to the call and put option when $K_i = K_0$.
- ΔK_i: Interval of strike price i, defined as
$$\Delta K_i = \frac{K_{i+1} - K_{i-1}}{2}.$$
For the lowest (highest) strike price, simply the difference with the next (previous) strike price is used.

- $Q(K_i)$: The midpoint price of the option belonging to strike price K_i. For $Q(K_0)$, the average of the midpoint prices of the corresponding call and put options is used.

The sum in (2.3) represents the numerical approximation of the integrals and the remaining term represents an approximation of ϵ. To understand the latter, we need to use the Taylor expansion of $\log(x)$ at 1 to obtain:

$$\log\left(\frac{S_0 e^{rT}}{S_*}\right) = 0 + \left(\frac{S_0 e^{rT}}{S_*} - 1\right) - \frac{1}{2}\left(\frac{S_0 e^{rT}}{S_*} - 1\right)^2 + \mathcal{O}\left(\left(\frac{S_0 e^{rT}}{S_*}\right)^3\right),$$

hence

$$\begin{aligned} \epsilon &= \log\left(\frac{S_0 e^{rT}}{S_*}\right) - \left(\frac{S_0 e^{rT}}{S_*} - 1\right) &(2.4)\\ &\approx -\frac{1}{2}\left(\frac{S_0 e^{rT}}{S_*} - 1\right)^2. &(2.5) \end{aligned}$$

For the VIX calculation, the boundary is taken as $S_* = K_0$ and the forward price $S_0 e^{rT}$ is approximated with the forward index level F, which explains the the remaining term in (2.3).

With the calculated σ^2, the VIX index is defined as $\sigma \times 100$. The VIX index attempts to represent the 30-day expected volatility of the S&P 500 index, using the value of the options of the S&P 500 index with a 30-day expiration. However, there will (almost) never be an option contract available on the market that expires in exactly 30 days. To take this into account, the VIX index calculation uses the option contracts of the 2 closest expiration dates, so-called the near-term and next-term options. For both the near-term and next-term options, we calculate σ^2 and then take the 30-day weighted average. Therefore, we finally obtain the calculation of the VIX index as (CBOE, 2019):

$$VIX = 100 \times \sqrt{\left(T_{near}\sigma^2_{near}\left(\frac{N_{next} - N_{30}}{N_{next} - N_{near}}\right) + T_{next}\sigma^2_{next}\left(\frac{N_{30} - N_{near}}{N_{next} - N_{near}}\right)\right)\frac{N_{365}}{N_{30}}},$$

with

- T_{near}, T_{next}: Time to expiration of the near-term and next-term options respectively.
- $\sigma^2_{near}, \sigma^2_{next}$: The general VIX calculation from (2.3) for the near-term and next-term options respectively.
- $N_{near}, N_{next}, N_{30}, N_{365}$: The number of minutes until the expiration of the near-term options, the number of minutes until the expiration of the next-term options, the amount of minutes in 30 days (43,200 minutes) and the amount of minutes in 365 days (525,600 minutes) respectively.

The type of options that are used in the calculation of the VIX index are the weekly options and the monthly options. The weekly option contracts have expiration dates each week, except for the third Friday of each month. The monthly option contacts have expiration dates on the third Friday of each month. One of the differences between the weekly and the monthly contracts is the time of the day at which the contract is expired. Weekly contracts expire at the end of the trading day (4:00 p.m. ET), while monthly contracts expire at the start of the trading day (9:30 a.m. ET). Since there is an expiration date each Friday for some option contracts, the gap between the near-term and next-term expiration dates will always be seven days. The expiration dates that are chosen for the VIX calculation are the option contracts with 24 to 30 calendar days to expiration for the near-term options and 31 to 37 calendar days to expiration for the next-term options.

2.2 VIX products

Since the VIX index is just a representation of the expected volatility of the S&P 500 index, it is not a product that one can simply buy. An investor is unable to just go to the market and buy a few assets of the VIX index. However, it is possible to buy derivatives, ETFs (Exchange-Traded Funds) and ETNs

(Exchange-Traded Notes) of the VIX index. These products make it possible for investors to correlate their portfolio with the VIX index. Reasons for buying VIX products may be for diversification of a portfolio that is correlated to the S&P 500 index or simply just for speculation of the uncertainty in the market.

When one buys a futures or option contract of the VIX index, the payoff will be a cash settlement based upon the current VIX index. The payoff will not use the exact value of the VIX index, but it will be close to it. As explained in Section 2.1, the calculation of VIX index uses the two options of the S&P 500 index with the time to expiry closest to 30 days. At the opening of the final settlement day of a derivative, it is the only moment in the month that an option contract of the S&P 500 index is available that expires in exactly 30 days. Now the settlement value of VIX derivatives is based upon the SOQ (Special Opening Quotation) of the VIX index. The SOQ is a special calculation of the VIX index which uses only uses the prices of these S&P 500 option contracts that expire in exactly 30 days. In addition, the SOQ calculation uses the values of the option contracts that have actually been traded during the opening of the day, instead of the midpoint values. However, when an option is not traded during the opening, then the midpoint price is used in the calculation.

2.2.1 VIX futures

A futures contract is a contract an investor can buy on the market which allows him to obtain the underlying asset or value for a specific price at some specified date in the future. An oil manufacturer for example may choose to sell their oil barrels by selling futures contracts each month. At the date of expiry, the manufacturer has to deliver the oil barrels to the holders of the futures contracts. The manufacturer may choose to take this approach when the oil price in the future is uncertain. By fixing a price and a demand to sell their oil each month, this can reduce the risk for the manufacturer. A reason for investors to buy futures contracts may also be to fix the price of the underlying asset to avoid facing the risk of sudden price movements. While this may be one of the reasons to buy futures, another reason to buy futures is to use the price movement of the underlying asset to an advantage. By buying a futures contract of a commodity and selling the contract before expiry, an investor is able to use the price movement of the commodity without physically receiving the asset at expiry. This way an investor is able to invest in oil without physically buying barrels of oil.

The first type of tradable derivatives of the VIX index were the futures contracts, which got introduced by the CBOE in 2004. The holder of a VIX futures contract will receive a cash settlement of the SOQ at the expiration date of the contract. At first only monthly contracts were listed on the market, but in 2015 also weekly VIX futures contracts were introduced. The CBOE may list up futures contracts with expiration dates up to:

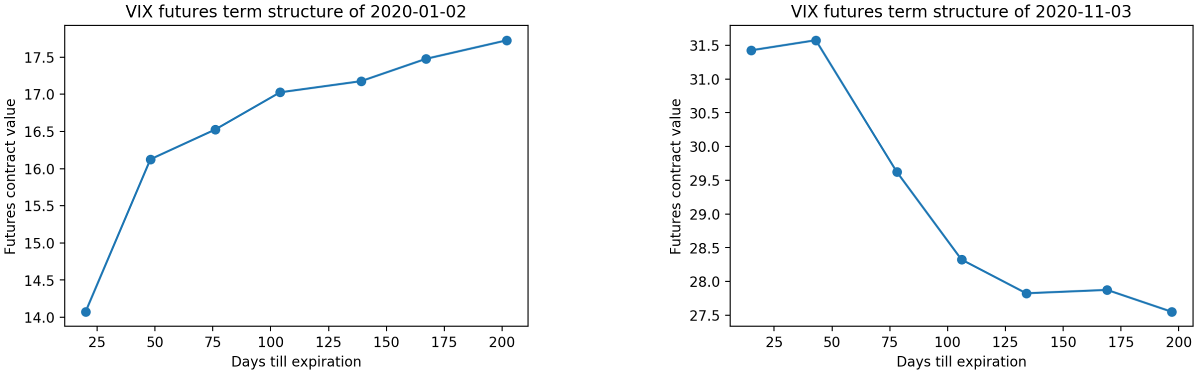
- 6 near-term weeks.
- 9 near-term months.
- 5 near-term months on the February quarterly cycle (February, May, August, November).

The expiration date of the VIX futures contracts is always on a Wednesday. For the monthly and February quarterly cycle contracts, the expiration date is the Wednesday that is 30 days prior to the third Friday of that next month. The reason behind this specific expiration date has to do with the fact that the expiration dates of the monthly S&P 500 index futures and options are on the third Friday of the month (exactly 30 days later). The weekly contracts are the Wednesdays of the weeks when a monthly or February quarterly cycle contract is not expiring.

Besides the normal VIX futures that are on the market, there is also a possibility to trade mini VIX futures contracts. These contracts are 10% the size of a normal VIX futures contract, which makes these contracts more appealing to smaller investors. Normal VIX futures contracts have a multiplier of $1000, which means that the payoff of a futures contract will be multiplied by 1000. This also implies that the price of a futures contract is multiplied by $1000, which can make futures contracts quite expensive. However, investors do not always need to pay the full amount of the futures contract. Market exchanges may set a initial margin (minimum of 50%) on the futures contract, which is the percentage of the value of the futures contract that an investor must have in his account. Mini VIX futures only have a multiplier of $100, which make these futures more attractive to smaller investors. For mini VIX futures the CBOE lists up contracts with the same expiration dates as the normal VIX futures.

Although these mini VIX futures allow for flexibility among investors, the normal VIX futures contracts are by far the most traded contracts. On June 18th 2021 for example, the trading volume of the normal VIX futures contracts were 274,522 contracts, while the mini VIX futures contracts had a trading volume of only 19,054 contracts (CBOE, 2021). The most often traded futures contracts are the monthly contracts which are the closest to the expiration date. The closer the contract is to expiration, the more information the investor has regarding the expected value of the VIX index at maturity. On June 18th 2021 for example, the trading volume of the futures contract of July were 159,384 contracts, the futures contract of August were 72,224 contracts and the futures contract of September were 19,867 contracts (CBOE, 2021). The liquidity of the market may be important when it comes to trading futures. When the liquidity of a futures contract is too low, the trader might not get the desired price for their futures contract due to a wide bid-ask spread.

When an investor is interested in the values of the current available futures contracts, he will look at the term structure. This is a representation of the prices of the available futures contracts at different expiration months, as in Figure 2.2.



(a) Term structure of January 2nd, 2020

(b) Term structure of November 3rd, 2020

Figure 2.2: Term structures of VIX futures

The term structure of a futures contract is most of the time in one of two states: in contango or in normal backwardation. A market is in contango when the prices of the futures contracts are higher than the expected spot price of the underlying asset. When the market is in the state of contango, an investor is willing to pay a higher premium to obtain the underlying asset in the future instead of immediately buying it. On the contrary, the sellers of the futures contracts are willing to sell the contracts in return for a higher premium. This premium is usually referred to as the cost of carry. When investing in commodities, such as oil, gold, silver etc., the cost of carry can be compared to the physical storage costs of the product. Consequently we will see an increasing futures curve over time in the term structure, for example as in Figure 2.2a. Conversely, the market is in normal backwardation when the prices of the futures contracts are lower than the expected spot price of the underlying asset. In this state of the market the investors of the futures contracts want to pay less than the expected spot price. Investors now want to be compensated for buying a futures contract, which implies that investors have a higher risk for fixing the price of the underlying asset at expiry. Consequently we will see a decreasing futures curve over time in the term structure, for example as in Figure 2.2b.

Hicks (1946) explains that there are two types of dominant participants in the futures market: speculators and hedgers (p. 138). The speculator will try to make a profit based upon the price movement of the futures price, while the hedger attempts to eliminate as much risk as possible. When the market is in contango, then it is preferable for the speculators to have a short position in futures. This is because the futures contract value will be expected to decrease when the time to expiration becomes shorter, since the spot VIX index is below the current futures price. Conversely, when the market is in normal backwardation, the futures contract is expected to increase in value over time, since the spot VIX index is above the current futures price. A speculator will then look to take a long position in a futures contract.

The main motivation for hedgers to buy futures is to reduce their risk position. When an investor takes an opposite position of the speculator, then this will be for reasons to reduce his risk.

2.3 Investment strategies with the VIX

Lots of different investment strategies that involve trading derivatives of the VIX index have been researched in literature. Auinger (2015) gives a clear summary on popular ideas regarding VIX strategies among investors. The most standard investment strategy of the VIX index revolves around exploiting the negative correlation between the VIX index and the S&P 500 index. By creating a portfolio of stocks from the S&P 500 index, one can diversify their portfolio by adding a VIX index derivative with negative correlation to the portfolio. When the S&P 500 index declines, the total loss of the portfolio will most likely be less, since the added derivative of the VIX index may increase in value. Daigler and Rossi (2006) found that adding VIX index derivatives to a S&P 500 portfolio significantly reduces risk, without having that much of an impact on the return of the portfolio. Following the example of Auinger (2015), suppose we have a portfolio of $100.000, containing:

- $90.000 (90%): S&P 500 index
- $10.000 (10%): VIX index

Note that we assume here that we can create a portfolio that perfectly represents the S&P 500 and VIX index. During the period of July 7th and September 28 in 1998, the indices changed in value as:

- S&P 500 index: 1184 → 956 (19% decrease)
- VIX index: 16.23 → 44.28 (150% increase)

If one invested $100.000 in this portfolio at July 7th, 1998, then one loses approximately $17.000 from the position on the S&P 500 index, while gaining approximately $15.000 from the position on the VIX index. Thus, even a small allocation of 10% of the VIX index in the portfolio could help in reducing the risk in a S&P 500 index portfolio. Dash and Moran (2005) found that an allocation of the VIX index between 0% and 10% resulted in the most efficient portfolios, since the VIX index itself is very volatile. As we saw in the example above, an increase of 150% in the VIX index can have huge impacts on portfolios where the VIX index is a large proportion of it. Investment strategies in general are not only considered efficient when they make a lot of profit, but also when there is not a lot of risk involved. The consequence of the behaviour of the VIX index is that one should limit their budget allocation of the VIX index in a portfolio.

Dash and Moran (2005) created an investment strategy that used different VIX index allocations over time based on the volatility of the VIX index. When the VIX index in the last quarter increased more than 20%, then 0% was allocated to the VIX index in the portfolio, while 10% was allocated to the VIX index when it decreased by more than 20%. When the VIX index in the last quarter had an increase or decrease below 20%, then the allocation of the VIX index in the portfolio was 5%.

An example of basis trading is presented by Simon and Campasano (2014). The basis of a futures contract is the spot price minus the market price of the futures contract. Hence, we have a negative basis when the market is in contango, while we have a positive basis when the market is in normal backwardation. The investment strategy takes short positions in futures when the basis is negative, while taking long positions when the basis is positive. Simultaneously, this strategy attempts to hedge its risk by taking positions in mini-S&P 500 futures.

Another popular trading strategy is calendar spread trading. which revolves around buying the futures contract with 1 month time to expiration, while simultaneously selling a futures contract with a later time to expiration (Nordén & Hou, 2018). This strategy takes advantage of the pricing differences between the futures contracts with different time to expiration. This trading strategy is specifically popular for VIX futures, since the spread between different futures contracts is invariant of the high volatile behaviour of the VIX index.

3 Bayesian Parameter Estimation and Markov Chain Monte Carlo Methods

Markov Chain Monte Carlo (MCMC) methods allow the ability to simulate from probability distributions that may be difficult to sample from. In Bayesian statistics, one usually looks to sample from a posterior distribution. These posterior distributions may become very difficult to sample from when more prior distributions are assumed. With the help of MCMC methods, one is able to create samples from this posterior distribution. In Section 3.1 we will first present a small introduction to Bayesian statistics and the use of priors. Afterwards, in Section 3.2 we will discuss the most general MCMC method, which is the Metropolis-Hastings algorithm. Finally, in Section 3.3 we will introduce an improved MCMC method, which is the Hamiltonian Monte Carlo algorithm.

3.1 An introduction to Bayesian statistics

When it comes to estimating parameters in a statistical model, there are usually two main ways to approach this problem. This is the well-known division of the frequentist and the Bayesian approach in statistical inference. The frequentist has the interpretation that the probability of an event is the frequency that this event occurs in the data. For example, a common frequentist method for estimating parameters is the maximum likelihood estimation. This well-known estimation method chooses the parameters of a distribution that maximize the likelihood given the data, which therefore nicely describes the frequentist interpretation. With frequentist methods one usually ends up with one fixed estimated value for the parameter it is estimating. However, what is different in Bayesian statistics is that one will assume that the parameters are actually random variables. The Bayesian statistician has a subjective interpretation on probability. He will assume some prior belief about the parameters beforehand, which we call the prior distribution. Using this prior distribution of the parameters, the Bayesian statistician will update this prior belief into the posterior belief when more data is observed. We call this posterior belief the posterior distribution.

In Bayesian statistics, one is looking for the distribution of the parameters θ given the data y, or $p(\theta|y)$. This distribution is called the posterior distribution, which can be determined using Bayes' Theorem (Rachev et al., 2008, Chapter 2):

$$\begin{aligned}p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)}\\ &= \frac{L(\theta|y)p(\theta)}{p(y)}\\ &\propto L(\theta|y)p(\theta),\end{aligned}$$

with $L(\theta|y)$ being the likelihood of θ given the data. Often it is useful to write the posterior distribution proportional to some function only depending on θ, which is the likelihood function times the prior distribution. Note that the likelihood $L(\theta|y)$ represents the information we have of θ given the data, while $p(\theta)$ is the prior distribution that we assumed beforehand. The frequentist would try to obtain θ by finding θ that maximizes this likelihood function, while the Bayesian statistician obtains the posterior distribution by also including their prior beliefs about the parameters.

Obtaining the posterior distribution of a parameter is usually very straightforward in Bayesian statistics. Work out Bayes' rule for your posterior distribution and one ends up with the posterior, or most of the time a function proportional to the posterior. When we have a function proportional to the posterior, it can sometimes be easy to figure out which constant to choose to multiply the proportional function to obtain the posterior. Take for example the posterior distribution $p(\theta|y) \propto \theta^{\alpha-1}e^{-\beta\theta}\mathbb{1}_{\theta\geqslant 0}$. In order to figure out what the constant is that one has to multiply to obtain a proper distribution, we have to find the constant c such that

$$c\int_0^\infty \theta^{\alpha-1}e^{-\beta\theta}d\theta = 1.$$

However, without too much thought we can figure out what c must be when we take a closer look at the posterior distribution. We can recognize that the posterior is also proportional to the probability density function of a gamma distribution with parameters α and β. Since the gamma distribution is a properly defined distribution, we instantly know that $c = \frac{\beta^\alpha}{\Gamma(\alpha)}$, where Γ is the gamma function. Hence the posterior $\theta|y$ is a gamma distribution with parameters α and β.

Although the posterior distribution may not always be part of a recognizable probability distribution family, MCMC methods allow us to sample from the posterior if we know a function proportional to the posterior distribution. Frequentists statistical methods however, may have a more difficult time to find the parameter estimates. The maximum likelihood estimator for instance, can be difficult to find when the likelihood function is complicated and has multiple parameters. When an analytical maximum of the likelihood function can not be found, one uses numerical approximation methods to find the maximum parameter values. When the likelihood function gets more complicated and more parameters get involved, numerical methods have a harder time finding the maximum. Even when a maximum value of the likelihood function is found, one needs to be very sure that the numerical method did not get trapped in a local maximum.

Example 3.1. *In this example we will demonstrate how to derive the posterior distribution from a simple model. Suppose we want to model the returns of a stock price $y = y_1, \dots, y_n$ as i.i.d. samples from a normal distribution $y_i \sim \mathcal{N}(\mu, \sigma^2)$ for $i = 1, \dots, n$. The idea is now to choose prior distributions for μ and σ based on our prior beliefs of these parameters. Let us for now assume that σ^2 is a known value. We could assume μ to be simply a normal distribution centered around zero with a large variance, hence*

$$\mu \sim \mathcal{N}(\eta, \tau^2).$$

With the chosen prior distribution of μ, we obtain the posterior distribution:

$$\begin{aligned}
p(\mu|y) &\propto p(y|\mu)p(\mu) \\
&\propto e^{-\frac{\sum_{i=1}^n (y_i-\mu)^2}{2\sigma^2}} e^{-\frac{(\mu-\eta)^2}{2\tau^2}} \\
&= e^{-\frac{\mu^2(n\tau^2+\sigma^2)-2\mu(\tau^2\sum_{i=1}^n y_i+\eta\sigma^2)}{2\sigma^2\tau^2}} e^{-\frac{\sum_{i=1}^n y_i^2}{2\sigma^2}-\frac{\eta^2}{2\tau^2}} \\
&\propto e^{-\frac{\mu^2-2\mu(\tau^2\sum_{i=1}^n y_i+\eta\sigma^2)(n\tau^2+\sigma^2)^{-1}}{2\sigma^2\tau^2(n\tau^2+\sigma^2)^{-1}}} \\
&\propto e^{-\frac{\mu^2-2\mu(\tau^2\sum_{i=1}^n y_i+\eta\sigma^2)(n\tau^2+\sigma^2)^{-1}}{2\sigma^2\tau^2(n\tau^2+\sigma^2)^{-1}}} e^{-\frac{(\tau^2\sum_{i=1}^n y_i+\eta\sigma^2)^2(n\tau^2+\sigma^2)^{-2}}{2\sigma^2\tau^2(n\tau^2+\sigma^2)^{-1}}} \\
&= e^{-\frac{(\mu-(\tau^2\sum_{i=1}^n y_i+\eta\sigma^2)(n\tau^2+\sigma^2)^{-1})^2}{2\sigma^2\tau^2(n\tau^2+\sigma^2)^{-1}}},
\end{aligned}$$

which implies the posterior of μ to be:

$$\mu|y \sim \mathcal{N}((\tau^2 \sum_{i=1}^n y_i + \eta\sigma^2)(n\tau^2 + \sigma^2)^{-1}, \sigma^2\tau^2(n\tau^2 + \sigma^2)^{-1}).$$

Note that both the prior and the posterior of μ are a normal distribution. When we choose a prior such that the posterior is from the same probability distribution family, then we call this prior a conjugate prior.

Example 3.2. *Continuing from Example 3.1, we can do the same for σ^2. Let us assume now that μ is a known value. It turns out that an inverse-gamma prior for σ^2 is a conjugate prior as well. Hence, let us assume*

$$\sigma^2 \sim \mathcal{IG}(\alpha, \beta).$$

With this prior distribution, we obtain the posterior distribution:

$$\begin{aligned}
p(\sigma^2|y) &\propto p(y|\sigma^2)p(\sigma^2) \\
&\propto \sigma^{-n} e^{-\frac{\sum_{i=1}^n (y_i-\mu)^2}{2\sigma^2}} \sigma^{-2\alpha-2} e^{-\frac{\beta}{\sigma^2}} \\
&\propto \sigma^{-2(\frac{n}{2}+\alpha)-2} e^{-\frac{\frac{1}{2}\sum_{i=1}^n (y_i-\mu)^2+\beta}{\sigma^2}},
\end{aligned}$$

which implies the posterior of σ^2 to be:

$$\sigma^2|y \sim \mathcal{IG}(\frac{n}{2} + \alpha, \frac{1}{2}\sum_{i=1}^n (y_i - \mu)^2 + \beta).$$

3.1.1 Choosing a prior distribution

In general, when one observes more and more data, the assumed prior distribution will have a smaller effect on the obtained posterior distribution. To make this prior distribution have a larger impact on the posterior distribution, one can decide to choose an informative prior on the parameters. A prior is said to be more informative when the distribution will tell you more information on what the parameter value will be. Every piece of information one has about the behaviour of the parameters could be included in the prior distribution to make the priors more informative, which in turn will guide you to find the correct posterior distribution. But suppose we have no idea how these parameters even behave, then a solution would be to assume non-informative priors for these parameters. Non-informative priors allow prior distributions to be assumed for parameters that we have no information of. Lots of different non-informative priors exist and it is up to the statistician to choose the prior that matches his prior beliefs. If we would have a parameter θ and the only information we have on the behaviours of this parameter is that $\theta \geqslant 0$, then a non-informative prior one could choose is simply the uniform distribution on $\mathbb{R}^+$, hence $p(\theta) \propto \mathbb{1}_{\theta \geqslant 0}$. This allows for a minimal impact of our prior distribution on the posterior, since our prior belief is that θ can be any positive real number with the same probability.

Example 3.3. *Let us continue from Example 3.1 and let us set up an experiment to test different priors. We will assume y to be samples from $\mathcal{N}(2,1)$. Suppose our prior belief is that we know nothing about the behaviour of μ. An idea could be to choose a prior distribution for μ such that little prior information is assumed. We can do this by allowing μ to have a large variance, hence $\eta = 0$ and $\tau^2 = 10000$ would suffice.*

Now another option is to choose an informative prior. Now suppose we have the strong prior belief that the return of the stock is typically 1%. This prior belief implies that we can take $\eta = 1$ in our prior distribution. Let us assume our belief is so strong that the returns are 1%, that we choose $\tau = 0.01$. Note that the actual distribution of the stock has a mean of 2, hence we choose here an informative prior that is inaccurate for the sake of the example. Let us now draw up to 100 samples of y and see if we can explain the behaviour of the posterior distributions of both priors.

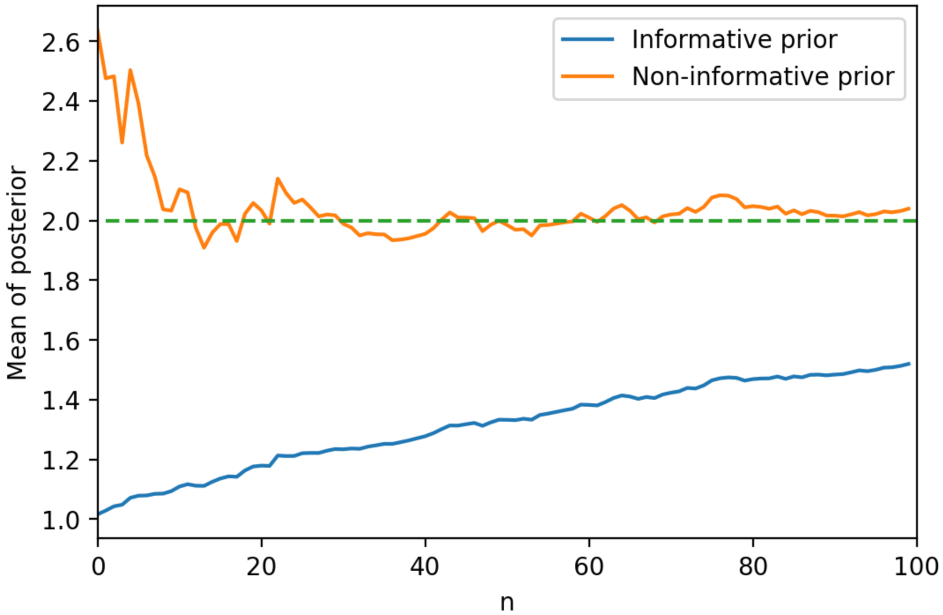


Figure 3.1: The mean of the posterior distributions for both priors.

In Figure 3.1 we see that the non-informative prior has no problems converging to the correct mean of y. Since we chose this prior to be extremely non-informative, the posterior is almost not at all influenced by the chosen prior, but mostly by the data itself. In contrary, the informative prior is converging very slow to the correct mean of y. The reason for this is that the chosen prior is highly informative, but *inaccurate. The very strong prior belief that the mean is around 1 is heavily impacting the result of the posterior distribution.*

3.2 Metropolis-Hastings algorithm

MCMC methods are designed to allow one to sample from complicated probability distributions. Robert and Casella (2004, p. 268) define MCMC methods as:

Definition 3.1. *A Markov Chain Monte Carlo (MCMC) method for the simulation of a distribution π is any method producing an ergodic Markov chain $(X_{(t)})$ whose stationary distribution is π, that is if $X_n \sim \pi$, then $X_{n+1} \sim \pi$.*

Ergodicity of a Markov chain implies that there exists an integer N such that all states in the Markov chain can be reached in $\leqslant N$ steps. More specific details on Markov chains can be found in Chapter 6 of Robert and Casella (2004). Now to sample from the posterior distribution, MCMC methods will use the prior distributions and the data that is observed. It will then create a Markov chain with the posterior distribution as its stationary distribution, which will be the sample of the posterior distribution.

The Metropolis-Hastings algorithm is such an MCMC method, introduced by Metropolis et al. (1953) and later extended to a more general method by Hastings (1970). This algorithm is considered to be the most general MCMC method. The algorithm makes use of a conditional density $q(z|x)$, which we call the proposal density. Using the previous value in the Markov chain, it will draw a sample from the proposal density and accept the draw with a certain probability. Suppose the probability density function we want to draw samples from is $\pi(x)$, also known as the target distribution, then the Metropolis-Hastings algorithm suggests the following sampling method:

Algorithm 1: Metropolis-Hastings algorithm

```
Given $X_0, N$:
for $n = 1$ to $N$ do
    Sample $z \sim q(z|X_{n-1})$.
    Set $X_n \leftarrow X_{n-1}$.
    With probability $\alpha = \min\left\{1, \frac{\pi(z)q(X_{n-1}|z)}{\pi(X_{n-1})q(z|X_{n-1})}\right\}$, set $X_n \leftarrow y$.
end
```

Note that the acceptance probability α will not have problems with division by zero. By the ergodic property of the produced Markov chain, we have that $\pi(X_{n-1}) > 0$. The only way this may not be the case is when we choose an initial value X_0 when $\pi(X_0) = 0$, hence if we choose X_0 such that $\pi(X_0) \neq 0$, we will always have $\pi(X_{n-1}) > 0$. We also clearly have that $q(z|X_{n-1}) > 0$, since y is a drawn value from the proposal distribution q. What is favorable about this algorithm is that one only needs to know a function that is proportional to the target density π. This is because the value of the acceptance probability α will not change when we use a function proportional to π, since the multiplied constant in the target distribution will of course vanish when the term $\frac{\pi(z)}{\pi(X_{n-1})}$ is calculated.

When one uses the Metropolis-Hastings algorithm, one has to determine the initial value X_0 of the Markov chain. This can be a difficult task, since you do not always know beforehand what the accurate starting values could be. When the starting value is not resembling the actual parameter space, then the first few draws of the MCMC sampler will not accurately represent a simulation of the target distribution π. However, the MCMC method does guarantee the ergodic Markov chain with π as stationary distribution throughout the simulation. Since we do want to end up with independent samples of the target distribution, one might consider removing the first few samples of the chain to remove any dependence of the chosen initial value. This is called the burn-in period of the simulation. The ergodicity of the Markov chain allows for the removal of these samples without affecting the underlying distribution of the samples. The amount of samples that have to be removed at the beginning of the chain depends on the convergence of the MCMC method to the target distribution. Therefore, one has to manually check at which point the starting value is not significantly impacting the sampled values. Usually, one overestimates the amount of samples to be removed from the chain just to be sure there are no unwanted dependencies.

After the removal of the samples from the burn-in period, one ends up with the samples of the target distribution. In order to analyse the results of the drawn samples, one usually looks at mean of the drawn samples, which is called the posterior mean. A way to interpret the spread of the posterior distribution is to look at the credible region, which is the region of values in which the target distribution will be in with a certain probability. Theoretically, these credible regions will asymptotically be equivalent with the confidence intervals that the frequentists use. This is a consequence of the Bernstein-von Mises theorem, of which more information on this theorem can be found in Chapter 10.2 of Van der Vaart (2000).

There are various types of Metropolis-Hastings algorithms that choose their own type of proposal distribution for $q(z|x)$. The most common algorithms are the independent Metropolis-Hastings algorithm and the random-walk Metropolis-Hastings algorithm. The independent Metropolis-Hastings algorithm uses a proposal distribution that can be perceived as a global proposal, since this proposal distribution tries to explore the global parameter space at each iteration. On the other hand, the random-walk Metropolis-Hastings algorithm uses a proposal distribution that can be perceived as a local proposal, since this proposal distribution explores the parameter space locally of the previous sampled values in the Markov chain. Finally, a special case of a Metropolis-Hastings algorithm is the Gibbs sampler. One can use this sampler when the posterior distributions of all parameters are known, which allows the acceptance rate to always be equal to one. We will briefly describe these methods in the following sections.

3.2.1 Independent Metropolis-Hastings algorithm

The independent Metropolis-Hastings algorithm uses a proposal distribution that is independent of the previous value x. Instead of using a proposal distribution $q(z|x)$ as in Algorithm 3, we now use a proposal distribution that does not depend on the parameter x, hence $q(z)$. The independent Metropolis-Hastings algorithm will then have the acceptance rate:

$$\alpha = \min\left\{1, \frac{\pi(z)q(X_{n-1})}{\pi(X_{n-1})q(z)}\right\}.$$

Proposal distributions that are closely related to the target distribution π will result in faster convergence of the algorithm. This can clearly be seen in the acceptance rate α, since $q = \pi$ will result in the acceptance rate always being 1. In this case all samples will be drawn from the target distribution and will always be accepted. A method one could use to find an independent proposal distribution that resembles the target distribution is to choose a distribution from some probability distribution family and then try to optimize the parameters (such as the location or scale parameters). The parameters will be optimized in such a way that the highest acceptance rate is achieved. This can be done by running the independent Metropolis-Hastings algorithm for different values of the parameters for a certain amount of iterations and then select the parameter values with the highest average acceptance rates.

However, it may be difficult to find a proposal distribution that is closely related to the target distribution, since most of the time the target distribution is only known up to a multiplicative constant. Posterior distributions may turn out to be very complicated, which happens with high-dimensional models for example. Finding an accurate proposal distribution that one can easily sample from may then be very difficult or even impossible. If one can not find a decent independent proposal that resembles the target distribution, then there may be a problem that the global parameter space does not get correctly explored. Then it may be favored to look for more a more local proposal distribution, such as the proposal distribution that the random walk Metropolis-Hastings algorithm uses.

3.2.2 Random walk Metropolis-Hastings algorithm

The random walk Metropolis-Hastings algorithm uses the previous sampled value of the parameter to locally explore the neighbourhood of the parameter space. The random walk proposal distribution gets chosen such that

$$z = X_{n-1} + \epsilon_n,$$

with $\epsilon_n \sim g$ being some random perturbation, where g is some distribution. The proposal distribution q is now of the form:

$$q(z|X_{n-1}) = g(z - X_{n-1}),$$

which means that the distribution g is now translated towards the previous sampled value of the parameter in the simulation. When we choose g to be a symmetric distribution, thus when $g(z) = g(-z)$, then the random walk Metropolis-Hastings algorithm has an acceptance rate:

$$\begin{aligned}\alpha &= \min\left\{1, \frac{\pi(z)q(X_{n-1}|z)}{\pi(X_{n-1})q(z|X_{n-1})}\right\} \\ &= \min\left\{1, \frac{\pi(z)g(X_{n-1} - z)}{\pi(X_{n-1})g(z - X_{n-1})}\right\} \\ &= \min\left\{1, \frac{\pi(z)g(-(z - X_{n-1}))}{\pi(X_{n-1})g(z - X_{n-1})}\right\} \\ &= \min\left\{1, \frac{\pi(z)}{\pi(X_{n-1})}\right\}.\end{aligned}$$

Usually, g is chosen as a normal or uniform distribution. Therefore, a common random walk proposal distribution could be:

$$\begin{aligned}z &\sim q(z|X_{n-1}) \\ &\sim \mathcal{N}(X_{n-1}, \sigma^2).\end{aligned} \tag{3.1}$$

At each iteration, the neighbourhood of the previous sampled parameter value is explored. Although the parameters only get locally explored, the ergodicity of the Markov chain allows for global exploration of the parameter space throughout the simulation. However, one must choose the random walk proposal with care. When we look at the proposal distribution in (3.1) for example, we see that we still have to specify σ^2. When we choose σ^2 too small, then it may be hard for the simulation to explore the entire parameter space. This does result in a very high acceptance rate of the proposed values, since $\pi(z)$ and $\pi(X_{n-1})$ will approximately be similar values. However, it is still not a good proposal distribution, since it has a hard time exploring the global parameter space. Therefore, contrary to the proposal distribution of the independent Metropolis-Hastings algorithm, one does not look for the highest acceptance rate when selecting a random walk proposal distribution. The variance σ^2 should also not be too large, since then a lot of sampled values for the proposal distribution will be rejected, hence the acceptance rate should also not be too low. One has to manually select a random walk proposal that avoids both of these problems. There does not explicitly exist a method that helps you find the optimal acceptance rate for the sampler. However interestingly enough, Roberts et al. (1997) explain that an acceptance rate of 0.234 for a random walk Metropolis-Hastings algorithm results in the maximum efficiency of the algorithm under quite general conditions.

Example 3.4. *In this example we will show that the random walk Metropolis-Hastings algorithm may have a harder time exploring the global parameter space when the variance of the proposal distribution is too low. Following Example 12.7 from Robert and Casella (2004, p. 475), take the following bimodal distribution:*

$$p(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \frac{4(x-0.3)^2 + 0.01}{4(1+0.09) + 0.01}, \tag{3.2}$$

which is the probability density function:

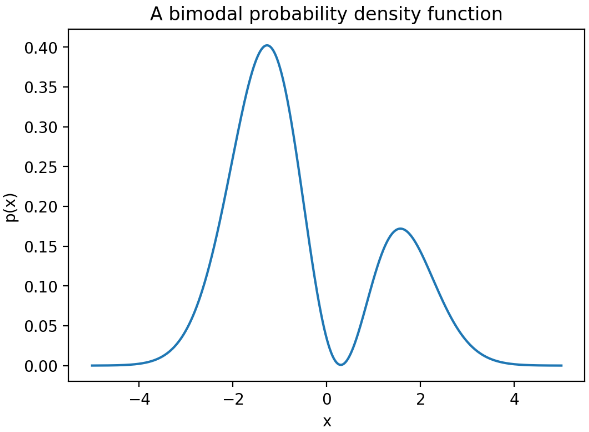
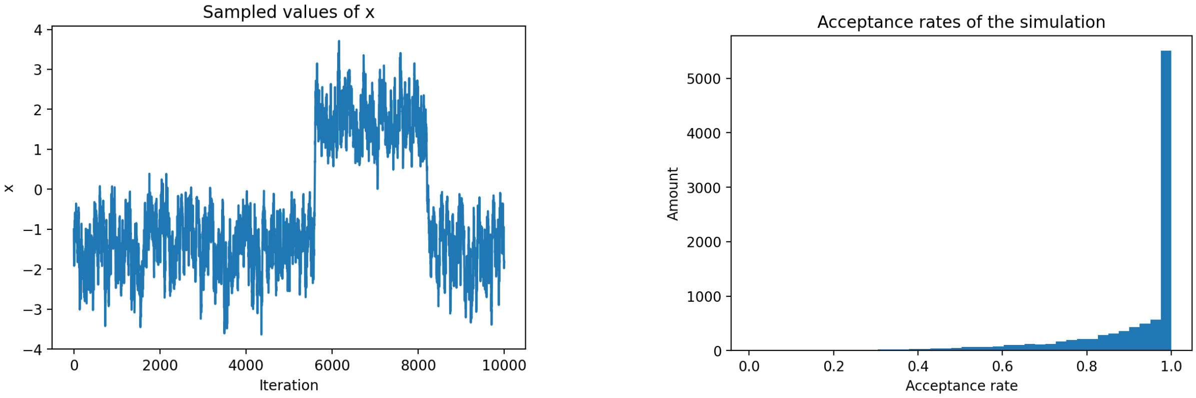


Figure 3.2: The probability density function from (3.2).

Suppose the density function in Figure 3.2 is the parameter space which we want to explore using a random walk proposal. When the variance of the proposal distribution is too small, then the sampler may take a long time to swap from one mode to the other mode. Let us take the proposal distribution:

$$q(z|x) \sim \mathcal{N}(x, 0.05).$$

Let us start the simulation at $x_0 = -1$, thus we start in the left mode of the distribution. After running the random walk Metropolis-Hastings algorithm for 10000 iterations, we obtain the following results:



(a) Progression of the Markov chain during the simulation.

(b) Histogram of the acceptance rates of the proposed values.

Figure 3.3: Results of the random walk Metropolis-Hastings algorithm after 10000 iterations.

We see in Figure 3.3a that it takes more than 5000 iterations to go from the left mode to the right mode, which shows that the algorithm is really struggling to explore the global parameter space with this proposal distribution. The corresponding acceptance rates of the simulation in Figure 3.3b clearly show that most of the proposed samples have a chance of getting accepted, which is a sign of slow movement throughout the parameter space.

3.2.3 Gibbs sampler

A special case of a Metropolis-Hastings algorithm one can use is Gibbs sampler, named after Josiah Willard Gibbs and introduced by Geman and Geman (1984). This sampler allows you to easily sample from high-dimensional target distributions, without having to worry about selecting an appropriate proposal distribution. Gibbs sampler will always propose values with an acceptance rate of one, hence all drawn values from the proposal distribution will be selected for the Markov chain. For each iteration, Gibbs sampler will go through all dimensions of the target distribution and draw from a specific proposal distribution, which is the conditional probability of the current dimension given the values of the other dimensions. However, it is not always possible to derive the conditional probability distribution function for all dimensions, hence Gibbs sampler can only be used when we know these distributions.

Suppose we have a K random variables in the target distribution, hence the nth sample of the target distribution is $X_n = (X_n^{(1)}, X_n^{(2)}, \ldots, X_n^{(K)})$. Gibbs sampler will draw a sample for each of the K random variables in order. For the draw of the kth variable, the proposal distribution uses the drawn values for $X_n^{(1)}, X_n^{(2)}, \ldots, X_n^{(k-1)}$ of the current iteration and the drawn values of $X_{n-1}^{(k+1)}, X_{n-1}^{(k+2)}, \ldots, X_{n-1}^{(K)}$ of the previous iteration. The proposal distribution for the kth random variable in Gibbs sampler becomes:

$$p(z|X_n^{(1)}, X_n^{(2)}, \ldots, X_n^{(k-1)}, X_{n-1}^{(k+1)}, X_{n-1}^{(k+2)}, \ldots, X_{n-1}^{(K)}),$$

which results in the algorithm for Gibbs sampler:

Algorithm 2: Gibbs sampler

```
Given $X_0, N$:
for $n$ = 1 to N do
    Sample $X_n^{(1)} \sim p(z|y, X_{n-1}^{(2)}, X_{n-1}^{(3)}, \ldots, X_{n-1}^{(K)})$.
    Sample $X_n^{(2)} \sim p(z|y, X_n^{(1)}, X_{n-1}^{(3)}, X_{n-1}^{(4)}, \ldots, X_{n-1}^{(K)})$.
    $\vdots$
    Sample $X_n^{(k)} \sim p(z|y, X_n^{(1)}, X_n^{(2)}, \ldots, X_n^{(k-1)}, X_{n-1}^{(k+1)}, X_{n-1}^{(k+2)}, \ldots, X_{n-1}^{(K)})$.
    $\vdots$
    Sample $X_n^{(K)} \sim p(z|y, X_n^{(1)}, X_n^{(2)}, \ldots, X_n^{(K-1)})$.
end
```

Note that we do not need to accept the drawn value with a certain acceptance rate α, since Gibbs sampler makes sure that we always have $\alpha = 1$ for each drawn value. This can be seen by writing out the acceptance probability for the proposal distribution of $X_n^{(k)}$. For notation, let us define $\mathbf{X}$ as the multivariate sample in the Markov chain before sampling $X_n^{(k)}$ and define $\mathbf{X}^*$ as multivariate sample in the Markov chain after sampling $X_n^{(k)}$, hence

$$\begin{aligned}\mathbf{X} &:= (X_{n-1}^{(1)}, \ldots, X_{n-1}^{(k)}, X_n^{(k+1)}, \ldots, X_n^{(K)})\\ \mathbf{X}^* &:= (X_{n-1}^{(1)}, \ldots, X_{n-1}^{(k-1)}, X_n^{(k)}, \ldots, X_n^{(K)}).\end{aligned}$$

Furthermore, let $\mathbf{X}_{-j}$ be the vector $\mathbf{X}$ with the jth element removed. We can now write out the acceptance probability for the proposal distribution of $X_n^{(k)}$ as:

$$\begin{aligned}\alpha &= \min\left\{1, \frac{\pi(\mathbf{X}^*)q(\mathbf{X}|\mathbf{X}^*)}{\pi(\mathbf{X})q(\mathbf{X}^*|\mathbf{X})}\right\}\\ &= \min\left\{1, \frac{p(\mathbf{X}^*)p(X_{n-1}^{(k)}|\mathbf{X}^*_{-k})}{p(\mathbf{X})p(X_n^{(k)}|\mathbf{X}_{-k})}\right\}\\ &= \min\left\{1, \frac{p(X_n^{(k)}|\mathbf{X}^*_{-k})p(\mathbf{X}^*_{-k})p(X_{n-1}^{(k)}|\mathbf{X}^*_{-k})}{p(X_{n-1}^{(k)}|\mathbf{X}_{-k})p(\mathbf{X}_{-k})p(X_n^{(k)}|\mathbf{X}_{-k})}\right\}\\ &= \min\left\{1, 1\right\} \qquad (3.3)\\ &= 1.\end{aligned}$$

Note that in step (3.3) we use the fact that $\mathbf{X}^*_{-k} = \mathbf{X}_{-k}$ by definition.

Example 3.5. *In this example we will implement a simple Gibbs sampler algorithm. Suppose we assume $y = (y_1, y_2, \dots, y_N)$ to be samples from $\mathcal{N}(2, 1)$. We will look at the following hierarchical model:*

$$\begin{aligned} y_i &\sim \mathcal{N}(\mu, \sigma^2) \quad \textit{for } i = 1, 2, \dots, N \\ \mu &\sim \mathcal{N}(0, 10000) \\ \sigma^2 &\sim \mathcal{IG}(0.001, 0.001). \end{aligned}$$

We have chosen the prior distributions of μ and σ^2 to be independent and non-informative. In Example 3.1 and 3.2 we have used similar prior distributions and from these examples we already know that the posterior of μ is:

$$\mu|y, \sigma^2 \sim \mathcal{N}\left(\frac{10000\sum_{i=1}^N y_i}{10000N + \sigma^2}, \frac{10000\sigma^2}{10000N + \sigma^2}\right),$$

and the posterior of σ^2 is:

$$\sigma^2|y, \mu \sim \mathcal{IG}(\frac{N}{2} + 0.001, \frac{1}{2}\sum_{i=1}^N (y_i - \mu)^2 + 0.001).$$

Using these posteriors, our Gibbs sampler is the following algorithm:

Algorithm 3: Gibbs sampler for Example 3.5

```
Given $\mu_0, \sigma_0^2, M$:
for $m = 1$ to $M$ do
    Sample $\mu_m \sim \mathcal{N}\left(\frac{10000\sum_{i=1}^N y_i}{10000N+\sigma^2_{m-1}}, \frac{10000\sigma^2_{m-1}}{10000N+\sigma^2_{m-1}}\right)$.
    Sample $\sigma^2_m \sim \mathcal{IG}(\frac{N}{2} + 0.001, \frac{1}{2}\sum_{i=1}^N (y_i - \mu_m)^2 + 0.001)$.
end
```

We will sample $N = 500$ values for y from $\mathcal{N}(2, 1)$ and use Gibbs sampler for $M = 50000$ iterations. We start the Markov chain with the initial values $\mu_0 = 0$ and $\sigma_0^2 = 3$. In order to remove any dependency of the selection of the initial values, we will apply a burn-in period of 5000 samples. The posterior mean and 95% credible region of the parameters are:

Table 1: Posterior mean and 95% credible region of μ and σ^2

	Posterior mean	**95% Credible region**
μ	2.057	(1.970, 2.145)
σ^2	0.984	(0.870, 1.113)

Visually, we have to following posterior distributions:

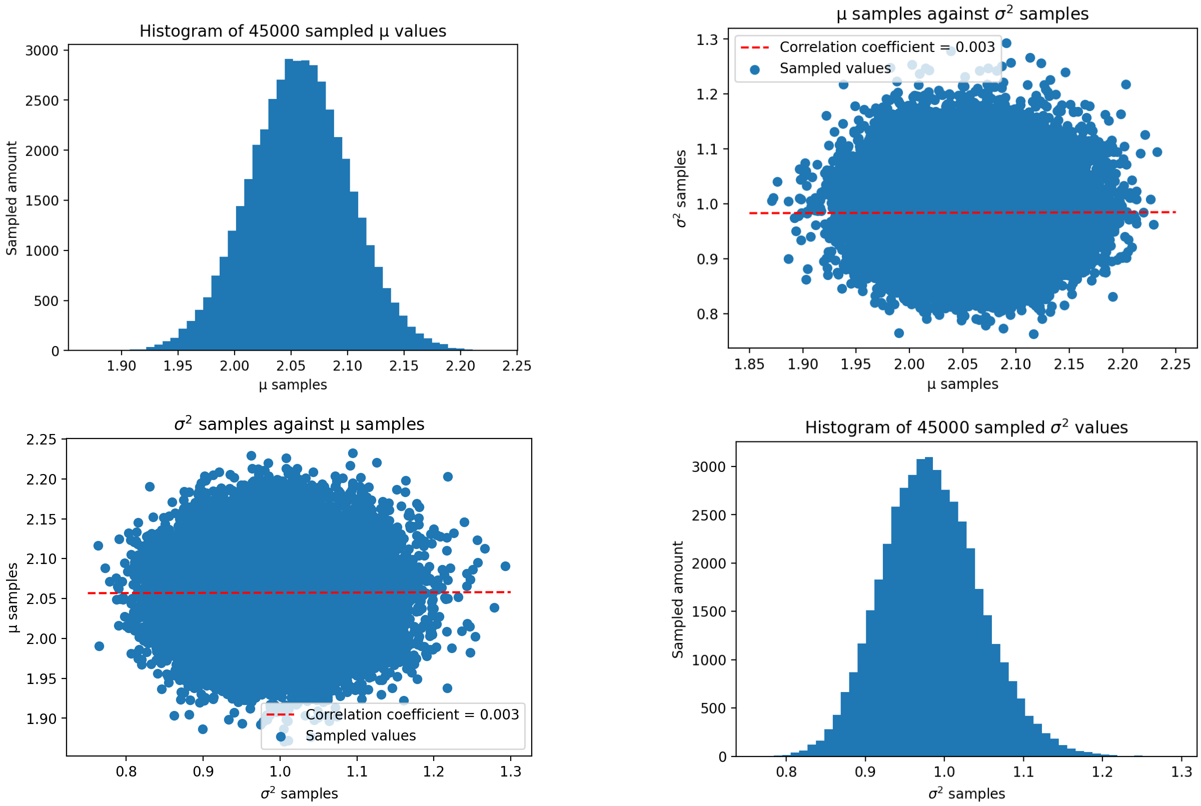


Figure 3.4: Correlation matrix of the posteriors.

As expected, the posterior distributions for μ and σ^2 are centered around 2 and 1 respectively. The posterior distributions seem to have no correlation with each other.

3.3 Hamiltonian Monte Carlo algorithm

The Hamiltonian Monte Carlo (also known as Hybrid Monte Carlo) algorithm was first introduced by Duane et al. (1987). This method combines the use of the Metropolis-Hastings algorithm and Hamiltonian dynamics to create a powerful MCMC method. This method is an improvement of the random walk Metropolis-Hastings algorithm, which is known to have problems for slow convergence for high-dimensional models when a good proposal distribution can not be found. When the proposal distribution is forced to be a random walk with low variance, then it takes a long time to explore the global parameter space. The Hamiltonian Monte Carlo algorithm attempts to avoid this problem by efficiently selecting the next proposed value with the help of Hamiltonian dynamics. The cost for a single sample from the target distribution π of dimension D using a random-walk Metropolis-Hastings algorithm is $\mathcal{O}(D^2)$, while using the Hamiltonian Monte Carlo algorithm results in a cost of $\mathcal{O}(D^{\frac{5}{4}})$ (Hoffman & Gelman, 2014).

The Hamiltonian Monte Carlo algorithm makes use of the Hamiltonian dynamics for a position vector q and a momentum vector p. The variable one is interested in for constructing the MCMC method is the position variable q, which gets updated each iteration using the momentum vector p. With the help of Hamilton's equations, the dynamics of q and p can be expressed in the Hamiltonian function $H(q, p)$, which is the sum of total energy of the system (potential energy + kinetic energy). In Section 3.3.1 we will discuss these dynamics in detail. By describing the Hamiltonian H in terms of a probability distribution, one can construct an MCMC algorithm which is able to sample from this distribution. We will describe the construction of the MCMC algorithm in Section 3.3.2. For both of these sections, we will base our notation and explanations on Chapter 5 of Brooks et al. (2011). Finally, in Section 3.3.3 we will discuss an improvement of the Hamiltonian Monte Carlo algorithm, which is the No-U-Turn Sampler.

3.3.1 Hamiltonian dynamics

The Hamiltonian dynamics describe the behaviour of the position variable q and momentum variable p. A physical interpretation of Hamiltonian dynamics can be described with an object moving over some surface in a 3-dimensional space. The vectors q and p will now be 2-dimensional, with q representing the position of the object and p representing the momentum (mass times the velocity) of the object. In this setting, the total mechanical energy in the system consists of the potential energy $U(q)$ and the kinetic energy $K(p)$. The potential energy is the energy that an object has due to its position in relation to other objects, such as gravitational energy. The potential energy is proportional to the height of the surface at its current position. The kinetic energy is the energy of an object based on its motion and is equal to $\frac{|p|^2}{2m}$, with m the mass of the object. When an object is moving upwards on the surface, then the current momentum of the object allows it to keep moving upwards. This will decrease the kinetic energy and increase the potential energy, until the upwards momentum of the object is zero, which will then make the object move downwards from the surface.

Suppose we have the d-dimensional position vector q and d-dimensional momentum vector p. The Hamiltonian function $H(q,p)$ is the sum of the potential energy and kinetic energy, thus:

$$H(q,p) = U(q) + K(p).$$

The dynamics of q and p can be described with Hamilton's equations:

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i} \tag{3.4}$$

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}, \tag{3.5}$$

for $i = 1, 2, \ldots, d$. For the MCMC setting, we want to set the potential energy equal to minus the log probability density of the target distribution. This way we will create a physical system of the target distribution, which we will allow our samples to move in according the dynamics given by Hamilton's equations. The kinetic energy is usually defined as $\frac{p^T C^{-1} p}{2}$, where C is a symmetric positive-definite matrix. This is equivalent to minus the log probability density of a multivariate normal distribution with mean 0 and covariance matrix C plus a constant. Thus we end up with:

$$U(q) = -\log(\pi(q)) \tag{3.6}$$

$$K(p) = \frac{p^T C^{-1} p}{2}, \tag{3.7}$$

with π the target distribution. Replacing (3.7) in (3.4) and (3.5) results in the dynamics:

$$\frac{dq_i}{dt} = [C^{-1}p]_i \tag{3.8}$$

$$\frac{dp_i}{dt} = -\frac{\partial U}{\partial q_i}, \tag{3.9}$$

for $i = 1, 2, \ldots, d$. Note that we can alternatively combine q and p into the $2d$-dimensional vector $z = (q,p)$ with dynamics:

$$\frac{dz}{dt} = J \nabla H(z),$$

with the $(2d \times 2d)$ matrix J as:

$$J = \begin{bmatrix} 0_{d\times d} & I_{d\times d} \\ -I_{d\times d} & 0_{d\times d} \end{bmatrix}. \tag{3.10}$$

The Hamiltonian dynamics have a few useful properties. First of all, Hamiltonian dynamics are time-reversible, which means that the mapping T_s that maps $(q(t), p(t))$ to $(q(t+s), p(t+s))$ is one-to-one. The inverse mapping is then simply obtained by negating the dynamics in (3.8) and (3.9). This property is useful for MCMC sampling using these Hamiltonian dynamics, since the stationary distribution of the Markov chain now remains invariant.

Secondly, these Hamiltonian dynamics keep the Hamiltonian invariant, also known as the law of conservation of energy in physics. This can be seen by writing out $\frac{dH}{dt}$ using the chain rule as:

$$\begin{aligned}\frac{dH}{dt} &= \sum_{i=1}^{d} \frac{dq_i}{dt}\frac{\partial H}{\partial q_i} + \frac{dp_i}{dt}\frac{\partial H}{\partial p_i} \\ &= \sum_{i=1}^{d} \frac{\partial H}{\partial p_i}\frac{\partial H}{\partial q_i} - \frac{\partial H}{\partial q_i}\frac{\partial H}{\partial p_i} \\ &= 0.\end{aligned}$$

In Section 3.3.2 we will see that the algorithm requires a Metropolis-Hastings update, which has an acceptance rate of 1 when the Hamiltonian is invariant. However, in practice we will use a numerical approximation method that makes the Hamiltonian approximately invariant.

Thirdly, the Hamiltonian dynamics preserve in volume in the (q, p) space, which implies that the image of the mapping T_s from some region in the (q, p) space with volume V will also have a volume of V. Volume conservation is equivalent with the absolute value of the determinant of the Jacobian of this mapping being equal to one. This follows from applying a change of variables inside an integral. For the Metropolis-Hastings step in the Hamiltonian Monte Carlo algorithm, we will see that this implies that we do not need to take the Jacobian into account when calculating the acceptance probability.

Finally, a stronger condition of volume conservation is symplecticness of the Hamiltonian dynamics. Suppose B_s is the Jacobian matrix of the mapping T_s, then the dynamics are symplectic when

$$B_s^T J^{-1} B_s = J^{-1},$$

with J as in (3.10). Taking the determinant on both sides implies that $|\det(B_s)| = 1$, which is equivalent with volume conservation.

3.3.2 Construction of the algorithm

Using the Hamiltonian dynamics as described in the previous section, we can now set up the Hamiltonian Monte Carlo algorithm. Just like the Metropolis-Hastings algorithm, we first want to propose a value for the next value of the Markov chain and then accept this proposed value with a certain acceptance probability. However, this proposed value is not a draw from a proposal distribution, but it is a simulation with a time discretization of L time steps of step size ϵ of the Hamiltonian dynamics. The value that we then end up with is the proposed value, which we will then accept with a certain probability.

Let us first look at the target distribution which we want to sample from. The Hamiltonian Monte Carlo algorithm attempts to sample from the target density:

$$\begin{aligned}\pi^*(q, p) &\propto e^{-H(q,p)} \qquad (3.11)\\ &\propto e^{-U(q)}e^{-K(p)} \\ &\propto \pi(q)e^{-\frac{p^T C^{-1} p}{2}}, \qquad (3.12)\end{aligned}$$

with $\pi(q)$ the distribution which want to obtain samples from. Note that q and p are independent in this target density, thus although we are sampling values for both q and p simultaneously, the sampled values for q are from $\pi(q)$. The motivation to choose this target distribution comes from statistical mechanics, where (3.11) represents a canonical distribution with a temperature of 1. The Hamiltonian is used as the energy function for the joint state of the position and momentum in the canonical distribution.

The first step of a single iteration in the Hamiltonian Monte Carlo algorithm consists of sampling a momentum vector p from a zero-mean normal distribution with covariance matrix C. This is a draw from the conditional distribution of p given q, since from (3.12) we have:

$$\begin{aligned}\pi^*(p|q) &= \pi^*(p) \\ &\propto e^{-\frac{p^T C^{-1} p}{2}},\end{aligned}$$

hence $p|q \sim \mathcal{N}(0, C)$. As seen with Gibbs sampler in Section 3.2.3, we can interpret this draw as a sample of p with acceptance probability 1. Usually C is chosen as a diagonal matrix, which implies independence between the momenta. Let us for simplicity assume that from now on C is a diagonal matrix.

The second step consists of simulating L steps of the Hamiltonian dynamics with step size ϵ. For the simulation of the Hamiltonian dynamics, we will need the Leapfrog integration method, which is a second-order numerical integration method to approximate the dynamics in (3.8) and (3.9). Using a time discretization of step size ϵ, the Leapfrog integrator consists of the following three updates:

$$\begin{aligned}
p_i(t + \frac{\epsilon}{2}) &= p_i(t) - (\frac{\epsilon}{2})\frac{\partial U}{\partial q_i}(q(t)), \\
q_i(t + \epsilon) &= q_i(t) + \epsilon \frac{p_i(t + \frac{\epsilon}{2})}{C_i}, \\
p_i(t + \epsilon) &= p_i(t + \frac{\epsilon}{2}) - (\frac{\epsilon}{2})\frac{\partial U}{\partial q_i}(q(t + \epsilon)),
\end{aligned}$$

with C_i the ith element on the diagonal of C. Important features of the Leapfrog method is that the time reversibility and symplecticness of the Hamiltonian dynamics are remained after approximation. However, the Hamiltonian is not invariant after approximation, since the vectors q and p are approximated to behave according to the Hamiltonian dynamics. Since the Hamiltonian will be different after L steps of the Leapfrog method, we will need to apply a Metropolis-Hastings step to accept the proposed values with a certain acceptance probability. The Hamiltonian after L steps of Leapfrog integration will of course be close to the actual Hamiltonian, which usually results in a high acceptance probability. The proposed values will be the vector $(\tilde{q}, -\tilde{p})$, with the vectors $\tilde{q}$ and $\tilde{p}$ we obtain after L steps of the Leapfrog algorithm. Note that we propose the negated vector $-\tilde{p}$, which makes the proposal distribution that we use for the acceptance probability symmetric. The proposed values of q and p do not come from a probability distribution, but are deterministic and determined by L steps of the Leapfrog integrator. When we negate the momentum vector $\tilde{p}$ after using L steps of the Leapfrog integrator and obtain the proposed values $(\tilde{q}, -\tilde{p})$, we can again use L steps of the Leapfrog integrator to end up with (p, q) again, hence negating p results in a symmetric proposal. Now since we have a symmetric proposal and since the Hamiltonian and Leapfrog integrator are volume preserving, we obtain the acceptance rate:

$$\alpha = \min\left\{1, \frac{e^{-H(\tilde{q},\tilde{p})}}{e^{-H(q,p)}}\right\}.$$

Al together, we end up with the Hamiltonian Monte Carlo algorithm:

Algorithm 4: Hamiltonian Monte Carlo

```
Given $q^0, C, \epsilon, L, M$:
for $m = 1$ to $M$ do
    Sample $p^0 \sim \mathcal{N}(0, C)$.
    Set $q^m \leftarrow q^{m-1}, \tilde{q} \leftarrow q^{m-1}, \tilde{p} \leftarrow p^0$.
    for $i = 1$ to $L$ do
        Set $\tilde{q}, \tilde{p} \leftarrow$ Leapfrog$(\tilde{q}, \tilde{p}, \epsilon)$.
    end
    With probability $\alpha = \min\left\{1, e^{-H(\tilde{q},\tilde{p})+H(q,p)}\right\}$, set $q^m \leftarrow \tilde{q}, p^m \leftarrow -\tilde{p}$.
end

function Leapfrog$(q, p, \epsilon)$:
    Set $\tilde{p} \leftarrow p - \frac{\epsilon}{2}\nabla_q U(q)$.
    Set $\tilde{q} \leftarrow q + \epsilon C^{-1}\tilde{p}$.
    Set $\tilde{p} \leftarrow \tilde{p} - \frac{\epsilon}{2}\nabla_q U(\tilde{q})$.
return
```

Here ∇_q represents the gradient with respect to q, hence in order to use the Hamiltonian Monte Carlo algorithm, we need to know the gradient of the logarithm of the target density beforehand. The performance of the Hamiltonian Monte Carlo algorithm strongly depends on tuning the Leapfrog parameters ϵ and L. When ϵ is taken too small, then the parameter space will be explored too slowly and the algorithm will need a lot of computation steps. On the other hand, if ϵ is taken too large, then the Leapfrog approximation will be inaccurate, which results in the Hamiltonian of the proposed values being far away from the actual Hamiltonian, hence this results in a low acceptance rate. Similarly for L, when we do not take enough amount of steps, then the parameter space will be explored too slowly. When we take too many steps, then the trajectory may reverse in direction. This means we wasted computation time by calculating too many Leapfrog integration steps, since we could have reached the proposed position at some earlier stage in the Leapfrog trajectory.

3.3.3 No-U-Turn Sampler

Finding the optimal values for ϵ and L may be difficult in practice. Although a correct value for ϵ may be easier to find, the amount of steps L for which the trajectory path is not too long or too short requires more effort. One usually has to perform multiple preliminary runs to test which values of these hyperparameters are good enough. A solution to this problem could be to use a No-U-Turn Sampler, introduced by Hoffman and Gelman (2014). This algorithm removes the need to predetermine the number of Leapfrog steps L in the Hamiltonian Monte Carlo algorithm.

The idea behind the No-U-Turn Sampler is to find some condition that tells us that the trajectory path is long enough, which then tells up to stop taking Leapfrog integration steps. For the No-U-Turn Sampler, this stopping criterion is the moment that the trajectory path starts to double back on itself, hence when the trajectory path makes a U-turn. In terms of the position and momentum vectors q and p, we could say the stopping criterion would be when the distance between $\tilde{q}$ and q starts decreasing, or equivalently when:

$$\begin{aligned}\frac{d}{dt}\frac{(\tilde{q}-q)^T(\tilde{q}-q)}{2} &= (\tilde{q}-q)\frac{d}{dt}(\tilde{q}-q)\\ &= (\tilde{q}-q)\tilde{p}\\ &< 0.\end{aligned}$$

The problem with this criterion is that the criterion will not imply the same stopping criterion when we want to move from $(\tilde{q},\tilde{p})$ to (q,p), hence this criterion is not time reversible. This would not guarantee that the Markov chain is sampled from the target distribution, hence we will need another stopping criterion.

The No-U-Turn sampler uses a repeated doubling procedure designed by Neal (2003) to solve this problem. This procedure starts the simulation of the Hamiltonian dynamics in the previous value of the Markov chain, but it has a probability of 0.5 to continue the trajectory forwards $(+\epsilon)$ and a probability of 0.5 to continue the trajectory backwards $(-\epsilon)$. The doubling procedure starts with moving 1 forwards or backwards, then move 2 forwards or backwards, then move 4 forwards or backwards and so on. Thus the amount of steps get doubled each time, hence the name of the doubling procedure. If we start with the amount of previous doubling steps $j = 0$, then for each doubling step we take 2^j Leapfrog steps forwards or backwards. This process builds up a balanced binary tree, which is a binary tree where the left and right subtrees have a difference in height of less than or equal to 1. The leaf nodes of the binary tree represent the trajectory path in the (q,p) space. The resulting balanced tree after taking 2^j Leapfrog steps in some direction has a height of $j+1$. In Figure 3.5 we see an example of doubling procedure simulation up to doubling 3 times.

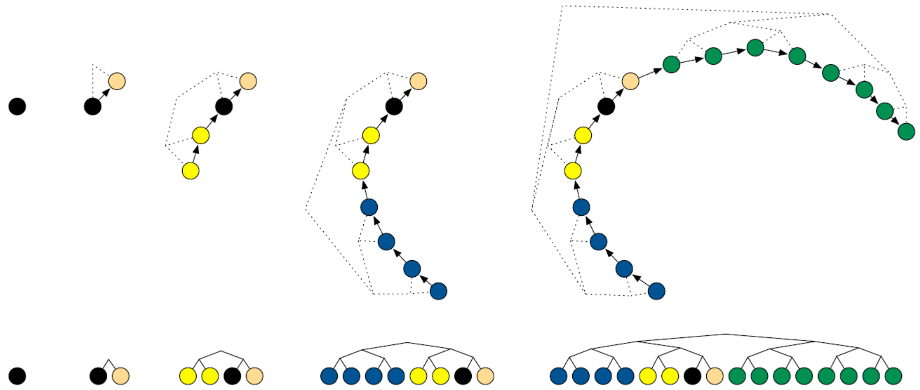


Figure 3.5: Example of the doubling procedure in 2 dimensions starting in the black node. The trajectory after doubling j times moves 2^j Leapfrog steps forwards or backwards, which is uniformly determined. The first subtree moves forward (orange node), the second moves backwards (yellow nodes), the third moves backwards (blue nodes) and the fourth moves forwards (green nodes). The resulted balanced binary tree can be seen below the trajectory path (Hoffman & Gelman, 2014, p. 5).

The doubling procedure stops when the trajectory from the leftmost point to the rightmost point of any balanced subtree with height > 0 within the total balanced binary tree has made a u-turn. After doubling j times, the created balanced binary tree has a height of $j+1$, which results in a total of $2^{j+1}-1$ balanced subtrees with height > 0 existing within the binary tree. Thus, if for any of these $2^{j+1}-1$ subtrees we have that the trajectory from leftmost point to the rightmost point has made a u-turn, then we stop the doubling procedure. We will denote the leftmost points as q^- and p^- and the rightmost points as q^+ and p^+. Thus, the doubling procedure will stop when we have in any balanced subtree that:

$$(q^+ - q^-)^T p^- < 0 \quad \text{or} \quad (q^+ - q^-)^T p^+ < 0. \tag{3.13}$$

Eventually, we end up with a trajectory of Leapfrog steps in the (q,p) space. Let us denote the collection of the (q,p) values in this trajectory as the set $\mathcal{B}$. The idea now is that the next value in the Markov chain will be a uniform sample from the collection of simulated (q,p) values. However, for the No-U-Turn Sampler we will not draw a uniform sample from all simulated values in $\mathcal{B}$, but only from a subset of a few selected candidates $\mathcal{C} \subseteq \mathcal{B}$. To create this candidate set $\mathcal{C}$, one has to remove values of the simulated trajectory based on two conditions:

1. Let us introduce a so-called slice variable u, for which we have the uniform distribution:
$$u \sim \mathcal{U}(0, e^{-H(q,p)}). \tag{3.14}$$
The first condition is that we do not include the simulated values $\tilde{q}$ and $\tilde{p}$ in $\mathcal{C}$ for which $e^{-H(\tilde{q},\tilde{p})} < u$.
2. The second condition is that we remove the last 2^j steps of the latest simulated subtree doubling procedure step. However, when the stopping criterion only holds for the leftmost and rightmost values of the entire balanced binary tree, then we do not remove the last 2^j steps from $\mathcal{C}$.

More theoretical motivation behind selection of these candidates can be found in Section 3.1.1 of Hoffman and Gelman (2014).

One final problem with the No-U-Turn Sampler is that our error in the simulation could in some situations become extremely large, which implies that the simulated trajectory of the doubling procedure in these situations could have an extremely low probability. A situation where this could happen is when we take the step size ϵ too large, which could make the Leapfrog integrator steps inaccurate and result in a low probability trajectory. In order to make sure that our trajectory does not contain inaccurate simulated values, we add an extra stopping condition of the doubling procedure. We stop the doubling procedure when the trajectory includes a sample $(\tilde{q},\tilde{p})$ such that:

$$H(\tilde{q},\tilde{p}) + \log(u) > \Delta_{\max},$$

with $\Delta_{\max}$ a large constant and u the slice variable as in (3.14). We use this condition, since this will prevent $H(\tilde{q},\tilde{p})$ from becoming too large, which implies the values $(\tilde{q},\tilde{p})$ to have a low probability with the target distribution π^*. Adding $\log(u)$ in this condition allows us to not accidentally stop the doubling procedure when the first node of the trajectory already has a low probability. Since u is uniform between 0 and $e^{-H(q,p)}$ with (q,p) as the first node of the trajectory, we already have that $H(q,p)$ is a high value when (q,p) has a low probability. It is recommended to take a large value for $\Delta_{\max}$ like 1000, such that the simulation is not stopped when the trajectory is accurate, but that the simulation is only stopped when extreme inaccurate samples are simulated.

Finally, when we put all these conditions and procedures together, we end up with the No-U-Turn Sampler algorithm:

Algorithm 5: No-U-Turn Sampler

```
Given $q^0, C, \epsilon, M$:
for $m$ = 1 to $M$ do
    Sample $p^0 \sim \mathcal{N}(0, C)$.
    Sample $u \sim \mathcal{U}(0, e^{-H(q^{m-1}, p^{m-1})})$.
    Set $q^- \leftarrow q^{m-1}, q^+ \leftarrow q^{m-1}, p^- \leftarrow p^0, p^+ \leftarrow p^0, j = 0, \mathcal{C} = \{(q^{m-1}, p^{m-1})\}, s = 1$.
    while $s$ = 1 do
        Sample $v_j \sim \mathcal{U}(\{-1, 1\})$.
        if $v_j = 1$ then
            $q^-, p^-, -, -, C', s' \leftarrow$ BuildTree$(q^-, p^-, u, v_j, j, \epsilon)$.
        else
            $-, -, q^+, p^+, C', s' \leftarrow$ BuildTree$(q^+, p^+, u, v_j, j, \epsilon)$.
        end
        if $s'$ = 1 then
            $\mathcal{C} \leftarrow \mathcal{C} \cup \mathcal{C}'$.
        end
        $s \leftarrow s' \mathbb{1}[(q^+ - q^-)^T p^- \geqslant 0] \mathbb{1}[(q^+ - q^-)^T p^+ \geqslant 0]$.
        $j \leftarrow j + 1$.
    end
    Sample $(q^m, p^m)$ uniformly from $\mathcal{C}$.
end
```

In Algorithm 5 the set $\mathcal{C}'$ represents the selected candidates from one step in the doubling procedure, such that the values satisfying the first condition $e^{-H(\tilde{q},\tilde{p})} < u$ are removed. The variable s represents whether the stopping criterion in (3.13) has been reached at some point during the simulation. When $s = 1$, then the stopping criterion has not been reached yet and we continue the doubling procedure and we include all samples of $\mathcal{C}'$ in the set of candidates $\mathcal{C}$. However, when $s = 0$ after the doubling procedure, then we do not include the latest candidate set $\mathcal{C}'$ in $\mathcal{C}$ based on the second condition of the candidate set. Finally, we need to add the case when the stopping criterion holds for the leftmost and the rightmost values of the entire balanced binary tree. The second condition for of the candidate set states that in this case we do add $\mathcal{C}'$ to the candidates set $\mathcal{C}$. Hence, after the doubling procedure we update s one last time using the leftmost and rightmost values of the entire tree. If it turns out that $s = 0$, then we stop the doubling procedure after adding $\mathcal{C}'$ to $\mathcal{C}$.

The algorithm of the doubling procedure function `BuildTree` can be described as follows:

Algorithm 6: BuildTree function

```
function BuildTree$(q, p, u, v, j, \epsilon)$:
    if $j = 0$ then
        $\tilde{q}, \tilde{p} \leftarrow$ Leapfrog$(q, p, v\epsilon)$.
        $\mathcal{C}' \leftarrow \begin{cases} \{(\tilde{q}, \tilde{p})\} & \text{if u} \leqslant e^{-H(\tilde{q},\tilde{p})} \\ \varnothing & \text{else} \end{cases}$
        $s' \leftarrow \mathbb{I}[u < e^{\Delta_{\max} - H(\tilde{q},\tilde{p})}]$.
        return $\tilde{q}, \tilde{p}, \tilde{q}, \tilde{p}, \mathcal{C}', s'$.
    else
        $q^-, p^-, q^+, p^+, \mathcal{C}', s' \leftarrow$ BuildTree$(q, p, u, v, j - 1, \epsilon)$.
        if $v = -1$ then
            $q^-, p^-, -, -, \mathcal{C}'', s'' \leftarrow$ BuildTree$(q^-, p^-, u, v, j - 1, \epsilon)$.
        else
            $-, -, q^+, p^+, \mathcal{C}'', s'' \leftarrow$ BuildTree$(q^+, p^+, u, v, j - 1, \epsilon)$.
        end
        $s' \leftarrow s's''\mathbb{I}[(q^+ - q^-)^T p^- \geqslant 0]\mathbb{I}[(q^+ - q^-)^T p^+ \geqslant 0]$.
        $\mathcal{C}' \leftarrow \mathcal{C}' \cup \mathcal{C}''$.
        return $q^-, p^-, q^+, p^+, \mathcal{C}', s'$.
    end
return
```

The algorithm of the BuildTree function in Algorithm 6 is described as a recursive function, with base case $j = 0$. When the base case is reached, then the Leapfrog step is executed and then will be determined whether the simulated value will be added to the candidate set. For each j the BuildTree function will sample 2 values, which results in 2^j values being simulated. The BuildTree function has as input the initial starting point of the values q and p, the slice variable u, the direction $v \in \{-1, 1\}$, the amount of doubling steps j and the step size of the Leapfrog integrator ϵ. It will return the leftmost values q^- and p^-, the rightmost values q^+ and p^+, the candidate list $\mathcal{C}'$ and the variable s which represents whether the doubling procedure should be stopped.

The algorithm of the No-U-Turn sampler as described in Algorithm 5 is actually a simplified version of the algorithm. Problems with the simplified No-U-Turn Sampler algorithm are that 2^j position and momentum vectors have to be stored and that improved methods exist for sampling from $\mathcal{C}$ that result in larger jumps on average than uniform sampling. More complicated and efficient versions of the No-U-Turn sampler that try to tackle these problems already exist and are described in Hoffman and Gelman (2014). Here is also describe how one can adaptively tune the step size parameter ϵ to a correct value.

For the simulation of our models with the No-U-Turn Sampler, we will use the software Stan (Stan Development Team, 2021). This is an open source C++ package that includes an efficient implementation of the No-U-Turn Sampler, based on Hoffman and Gelman (2014). The implementation will perform Leapfrog steps until the criteria of a u-turn is reached or until the maximum tree depth is reached, which is 10 at default, hence $2^{10} - 1$ Leapfrog steps. The step size parameter ϵ is tuned during the burn-in period of the simulation. The covariance matrix C that is used to sample the initial momentum vector in each iteration is a diagonal matrix with positive diagonal values. The values on the diagonal are the variances of the parameters, which also get estimated during the burn-in period. The final selection of the parameter values from the candidates set is a sample from a multinomial distribution, where the values sampled in the second half of the trajectory have a higher probability of being selected.

Example 3.6. *Let us look at an example of the simulation of the candidate set $\mathcal{C}$ in the No-U-Turn Sampler using the doubling procedure in Figure 3.6.*

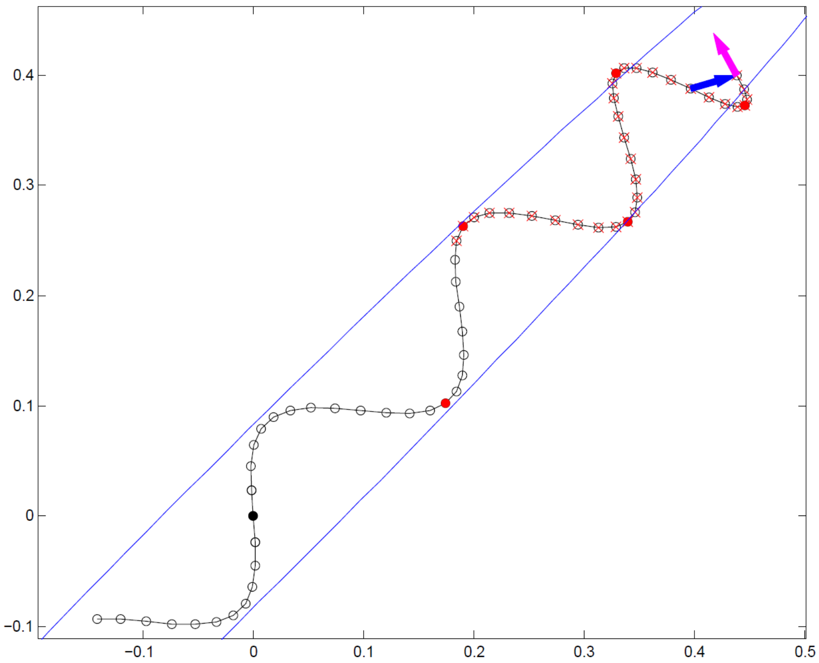


Figure 3.6: Selection of the candidate set $\mathcal{C}$ for a trajectory in the No-U-Turn Sampler using the doubling procedure (Hoffman & Gelman, 2014, p. 6).

In Figure 3.6 we start the trajectory at the solid black circle. From there we start the doubling procedure up to $j = 5$, which results in a trajectory of $2 \cdot 2^5 = 64$ simulated values throughout the parameter space, represented by the black circles. The blue lines represent a contour (constant value) of the target distribution. The open black circles represent the values that are selected in the candidate set $\mathcal{C}$, while the black circles with a red cross and the solid red circles represent the values of the trajectory that are not selected in the candidate set. The blue arrow represents the position vector of the rightmost node minus the leftmost node and the magenta vector the momentum vector of the right most node of a subtree of height 3 within the total balanced binary tree of the doubling procedure. The angle between the position and the momentum vector is larger than 90 degrees, hence the stopping criterion is reached in this subtree. Since the stopping criterion is reached in a subtree during the doubling procedure of $j = 5$, we have to remove all 2^5 values of the last step of the doubling procedure, which are represented by the black circles with a red cross. Now, we also have to remove the simulated values for which the joint probability is below the slice variable u, hence for which $u \leqslant e^{-H(\tilde{q},\tilde{p})}$. These values are marked with a solid red circle. Finally, the No-U-Turn sampler will uniformly sample the next value in the Markov chain from the remaining black circles.

4 Modelling the VIX Term Structure of Futures

Modelling the term structure of VIX futures allows us to get a better understanding of the current behaviour of the market. For example, with a term structure model we can figure out whether the market is in contango or in normal backwardation, which may alter the decisions we make in an investment strategy. The model of the term structure itself will give us an idea of how the term structure in theory should behave. When for some futures contracts the market significantly deviates from the theoretical term structure, we could argue that this futures contract is incorrectly priced by the market. Inaccurate pricing of financial products by the market is certainly of interest to investors and strategists, since this longs for opportunities to make money. Financial products that have a low price in the market, while the actual (or theoretical) value of the product is higher, is a good investment opportunity for a long position. These cheap products are expected to move upwards in value over time, since we expect the market to notice at some point that this product is a cheap purchase, which in turn will make the market correct itself to the right price.

In this Chapter we will attempt to create such a term structure model and investment strategy that could take an advantage of incorrectly priced VIX futures. First in Section 4.1 we will look at general modelling ideas of volatility and the VIX index, which will give us an idea on how to model the VIX index. We will then introduce our first model in Section 4.2 and explain how we estimate our parameters. In Section 4.3 we will adapt the term structure model into the error correction model in an attempt to fix the drawbacks of the initial model. For both models we will use Bayesian parameter estimation methods to end up with credible regions of our futures prices in the term structure. Ultimately the idea of these credible regions is to use them in our strategy to detect undervalued or overvalued futures contracts on the market. When the market price of a futures contract lies outside the credible region of the theoretical term structure, we could argue that this contract is incorrectly priced. How we eventually set up an investment strategy with this model is discussed in Section 4.4.

4.1 Stylized facts about volatility and the VIX index

The modelling of financial volatility has been given a lot of attention in the last few decades. While the well-known Black-Scholes model (Black & Scholes, 1973) assumes the volatility to be a constant, more recent models provide evidence on the effectiveness of the heteroscedasticity and stochastic volatility assumptions. In contrast to homoscedasticity, which assumes volatility to remain the same over time, heteroscedasticity refers to the behaviour of volatility varying over time. For example, the constant elasticity of variance (CEV) model by Cox (1975) is a more general stochastic differential equation that deals with heteroscedasticity. Similarly, the autoregressive conditional heteroscedasticity (ARCH) model by Engle (1982) attempts to take heteroscedasticity into account by describing the variance of the process as a function of the previous squared error terms. This model was later generalized as the generalised autoregressive conditional heteroscedasticity (GARCH) model by Bollerslev (1986), which describes the variance by also including the previous variance terms. A final example of a well-known stochastic volatility model is the Heston model by Heston (1993), which assumes the volatility to be a stochastic process itself. All of these models concentrate on the assumption that the behaviour of volatility is indeterministic.

Financial volatility typically appears to behave according to certain characteristics. The mean-reverting behaviour of volatility seems to be straightforward: a period of high volatility will eventually return to a normal volatility level and a period of low volatility will eventually be followed by a rise in volatility (Majmudar & Banerjee, 2004). Another stylized fact of volatility behaviour is volatility clustering, which explains the phenomenon of longer periods where there is high volatility and longer periods where there is low volatility. These characteristics are clearly visible in Figure 4.1.

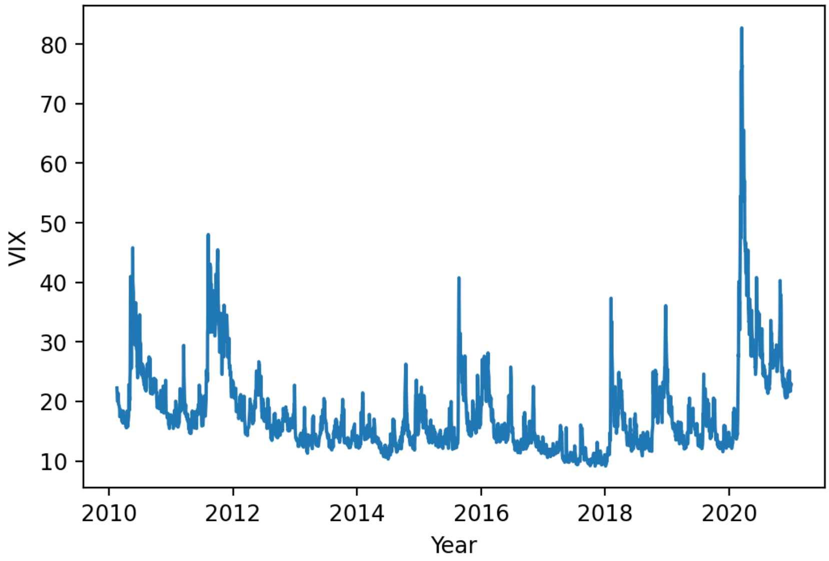


Figure 4.1: VIX index between 2010-2021

Another characteristic one might notice in Figure 4.1 is the presence of jumps. Jumps in stock prices, and therefore in volatility, may occur when suddenly new critical information has come to light or a certain event is happening. Examples of these occurrences could be an unexpected negative earnings report or a terrorist attack. In situations as these, the investors may become uncertain about the value of the financial instrument and will try to respond quickly, resulting into a jump of the price. In Figure 4.1 one can clearly notice a jump in the VIX index at the beginning of 2020. This sudden jump in the VIX index is related to the COVID-19 outbreak, which started spreading across the world around this period (Albulescu, 2020). The improvement of including jumps into the modelling of volatility index dynamics has been shown by Dotsis et al. (2007). However, when it comes to modelling future prices, the improvement of including jumps is not evident at all. With the empirical research performed by Dupoyet et al. (2011) it was demonstrated that the processes without jumps outperformed the processes with jumps on future pricing, even though the processes with jumps were a better fit for the VIX index than the processes without jumps. When a jump occurred in the VIX index, the models with jumps turned out to overreact on the VIX futures prices, while the market reacted rather calm.

4.2 Term structure model

We will initiate our model by assuming the dynamics of the VIX index to be a mean-reverting constant elasticity of variance model without jumps, which we from now on name the CEV model. This assumption is based on the research on VIX futures pricing models performed by Dupoyet et al. (2011). There it was shown that the CEV model consistently outperformed other common mean-reverting and jump-diffusion models, such as the CEV model with jumps and the Cox–Ingersoll–Ross (CIR) model (Cox et al., 1985) with and without jumps in future pricing. Although they do use a different method of parameter estimation, we will explain why our method does not need to include jumps into the model.

The CEV model describes the VIX index V_t to be the following stochastic differential equation:

$$dV_t = (\alpha - \kappa V_t)dt + \sigma V_t^{\gamma} dW_t, \tag{4.1}$$

where $\alpha, \kappa, \sigma, \gamma \geqslant 0$ are constants and W_t is a Brownian motion under the probability measure P on the probability space $(\Omega, \mathcal{F}, P)$. There are a few interesting characteristics of (4.1). First of all, this stochastic differential equation has the mean-reverting factor $(\alpha - \kappa V_t)dt$. This factor implies a reversion of the process to the mean $\frac{\alpha}{\kappa}$ with theoretical speed of reversion κ. Secondly, note that the variance term $\sigma(V_t)^{\gamma} dW_t$ allows the variance to depend on the value of the VIX. Finally, the γ parameter attempts to capture the so-called *leverage effect.* This effect reflects the correlation between the market and the volatility. When the value of stocks in equity markets decrease in value, then the leverage of the company relatively increases, which results in a higher volatility. Therefore in equity markets we would expect a $\gamma < 1$. In commodity markets we see an inverse leverage effect occurring, where an increase in the commodity market actually implies an increase in volatility. Therefore in commodity markets we would expect a $\gamma > 1$. Since the VIX index in itself is not a tradable market, the motivation behind the leverage effect does not apply here. However, one can notice that periods of high volatility of the VIX usually occur when the VIX reaches high levels, while the VIX has low volatility when it is on a lower level. This behaviour seems similar to commodity markets, hence we would expect a $\gamma > 1$. In the estimated CEV model (with and without jumps) by Dupoyet et al. (2011) of the VIX, they found a value of approximately 1.6 for γ.

In order to derive the price of a futures contract based on (4.1), we will have to find the dynamics of the CEV model under the risk-neutral measure Q. The current process is defined under the real-world measure P, thus we will have to find a transformation from P to Q. We will mainly follow the derivation and motivation of Grünbichler and Longstaff (1996), but some more details on certain derivations come from chapter 3 of Lutz (2009). Although in the paper of Grünbichler and Longstaff (1996) the futures price is derived from a CIR model, we can use the same derivation using the CEV model. This is because at some point in the derivation the γ parameter will dissolve when we take the expectation. Since the CIR model is just a special case of the CEV model ($\gamma = \frac{1}{2}$), we will end up with the same futures price.

4.2.1 Derivation of the futures contract price

Now let $A(t, T, V_t)$ be the value of a futures contract at time t, which has a payoff of V_T at some expiry date $T > t$. Then by using (4.1) and Itô's Lemma (Itô, 1944), we obtain the dynamics:

$$dA = \left(\frac{\partial A}{\partial t} + (\alpha - \kappa V_t)\frac{\partial A}{\partial V_t} + \frac{1}{2}\sigma^2 V_t^{2\gamma}\frac{\partial^2 A}{\partial V_t^2}\right) dt + \sigma V_t^{\gamma}\frac{\partial A}{\partial V_t} dW_t. \tag{4.2}$$

The next step for obtaining the measure transformation is to determine the risk premium of the volatility process. The risk premium is the expected return of the asset minus the risk-free rate, thus it is the compensation that an investor receives for taking a risk by investing in the asset instead in the risk-free products. Grünbichler and Longstaff (1996) explain that it is a common assumption to take the expected premium for volatility risk to be proportional with the volatility itself, hence ξV for a constant ξ. As explained in Lutz (2009), one can then use the risk premium of the volatility to define the Brownian motion $\tilde{W}_t$ of the risk-neutral process under Q as:

$$\tilde{W}_t = W_t + \frac{\xi V_t}{\sigma V_t^{\gamma}}. \tag{4.3}$$

Using this Brownian motion transformation, we now obtain the risk-neutral dynamics of the futures contract:

$$dA = \left(\frac{\partial A}{\partial t} + (\alpha - (\kappa + \xi)V_t)\frac{\partial A}{\partial V_t} + \frac{1}{2}\sigma^2 V_t^{2\gamma}\frac{\partial^2 A}{\partial V_t^2}\right) dt + \sigma V_t^{\gamma}\frac{\partial A}{\partial V_t} d\tilde{W}_t. \tag{4.4}$$

However, to proof formally that $\tilde{W}_t$ is in fact the Brownian motion under the risk-neutral measure Q, we will require a few more steps. First of all, we will need to proof that $\tilde{W}_t$ is a standard Brownian motion under some constructed probability measure Q. Secondly, we will have to show that this measure Q is in fact the risk-neutral measure we are looking for. Finally, we will be able to use the dynamics of our process under this risk-neutral measure to obtain the futures price.

To start off, we will first need to state Novikov's condition (Novikov, 1972) and Girsanov's Theorem (Girsanov, 1960).

Lemma 4.1. *(Novikov's condition) Let $\{\chi_t\}$ be an adapted process to the natural filtration $\{F_t\}$ and furthermore define Z_t to be:*

$$Z_t = e^{\int_0^t \chi_s dW_s - \frac{1}{2}\int_0^t \chi_s^2 ds}. \tag{4.5}$$

If for all $t \geqslant 0$

$$\mathbb{E}\left[e^{\int_0^t \frac{1}{2}\chi_s^2 ds}\right] < \infty, \tag{4.6}$$

then $\{Z_t\}$ is a martingale w.r.t. $\{F_t\}$.

Proof. A proof of this lemma can be found in Section 3.5.D of Karatzas and Shreve (1998). □

Theorem 4.1. *(Girsanov's Theorem) Let $\{\chi_t\}$ be an adapted process to the natural filtration $\{F_t\}$. Define Z_t as in (4.5) and take measure Q on $(\Omega, \mathcal{F})$ such that the Radon-Nikodyn derivative of Q w.r.t. the real-world measure P is equal to Z_t, or equivalently*

$$\left(\frac{dQ}{dP}\right)_{\mathcal{F}_t} = Z_t \leftrightarrow Q(F) = \mathbb{E}_P[Z_t \mathbb{1}_F] \qquad \forall F \in \mathcal{F}_t,\ \forall t \geqslant 0. \tag{4.7}$$

Now let $\hat{W}_t$ be the process

$$\hat{W}_t = W_t + \int_0^t \chi_s ds,$$

then if Novikov's condition from Lemma 4.1 holds for $\{\chi_t\}$, then $\hat{W}_t$ is a Brownian motion under the probability measure Q.

Proof. A proof of Girsanov's Theorem can be found in Section 3.5.B of Karatzas and Shreve (1998). □

Note that Novikov's condition is referred to as inequality in (4.6), rather than Lemma 4.1 itself. We can now set up the following corollary to obtain the measure transformation from the real-world measure to the risk-neutral measure:

Corollary 4.1. *Suppose V_t is the stochastic differential equation as in (4.1) and W_t is a a Brownian motion under the probability measure P on the probability space $(\Omega, \mathcal{F}, P)$, then $\tilde{W}_t = W_t + \frac{\xi V_t}{\sigma V_t^\gamma}$ is a Brownian motion under the risk-neutral probability measure Q, with Q as*

$$\left(\frac{dQ}{dP}\right)_{\mathcal{F}_t} = Z_t \leftrightarrow Q(F) = \mathbb{E}_P[Z_t \mathbb{1}_F] \qquad \forall F \in \mathcal{F}_t,\ \forall t \geqslant 0, \tag{4.8}$$

with Z_t as

$$Z_t = e^{\int_0^t \chi_s dW_s - \frac{1}{2}\int_0^t \chi_s^2 ds}$$

and $\chi_t = \frac{\xi V_t}{\sigma V_t^\gamma}$.

Proof. Usually, the first step would be to confirm Novikov's condition for $\chi_t = \frac{\xi V_t}{\sigma V_t^\gamma}$ and then use Girsanov's Theorem to show that $\tilde{W}_t$ is a Brownian motion under Q. However, Confirming Novikov's condition for this particular stochastic process requires a challenging derivation, which can be found in Appendix A of Cheridito et al. (2007). For this reason we will derive the risk-neutral process in an alternative way. We will show that the Brownian motion transformation $\tilde{W}_t$ is the risk-neutral transformation of V_t by setting up a risk-free portfolio and show that we end up with the same risk-neutral process.

The CEV model as in (4.1) is part of the one-factor short rate models, which are mostly used to model interest rates. The term one-factor refers to the single stochastic factor in the process, which is just V_t in this case. Following Vasicek (1977), we will create a portfolio containing two zero-coupon bonds and eventually find the pricing PDE of the bonds that follows the hedging strategy of this portfolio. The payoff of these bonds will be V_T at time T. Note that these bonds are essentially just futures contracts, since the payoff is the value underlying asset. We denote the value of these futures at time t with expiration date T again as $A(t, T, V_t)$.

Suppose at time t our portfolio consists of two futures contracts maturing at T_1 and T_2, thus $A(t, T_1, V_t)$ and $A(t, T_2, V_t)$. As we similarly have seen in (4.2), we can use Itô's Lemma to obtain the returns of the contract as:

$$\frac{dA}{A} = \mu_A(t, T, V_t)dt + \sigma_A(t, T, V_t)dW_t,$$

with

$$\mu_A(t,T,V_t) = \frac{1}{A}\left(\frac{\partial A}{\partial t} + (\alpha - \kappa V_t)\frac{\partial A}{\partial V_t} + \frac{1}{2}\sigma^2 V_t^{2\gamma}\frac{\partial^2 A}{\partial V_t^2}\right), \tag{4.9}$$

$$\sigma_A(t,T,V_t) = \frac{1}{A}\sigma V_t^{\gamma}\frac{\partial A}{\partial V_t}. \tag{4.10}$$

W.l.o.g. suppose at time t our portfolio $\Pi(t)$ has a value of \$1 and contains $\Delta_t^{(1)}$ and $\Delta_t^{(2)}$ dollar of $A(t,T_1,V_t)$ and $A(t,T_2,V_t)$ respectively. Thus our portfolio value $\Pi(t)$ at time t is:

$$\Pi(t) = \Delta_t^{(1)}A(t,T_1,V_t) + \Delta_t^{(2)}A(t,T_2,V_t).$$

The returns of this portfolio will then look like

$$\begin{aligned}\frac{d\Pi}{\Pi} &= \Delta_t^{(1)}\frac{dA(t,T_1,V_t)}{A(t,T_1,V_t)} + \Delta_t^{(2)}\frac{dA(t,T_2,V_t)}{A(t,T_2,V_t)}\\ &= \left[\Delta_t^{(1)}\mu_A(t,T_1,V_t) + \Delta_t^{(2)}\mu_A(t,T_2,V_t)\right]dt + \left[\Delta_t^{(1)}\sigma_A(t,T_1,V_t) + \Delta_t^{(2)}\sigma_A(t,T_2,V_t)\right]dW_t. \end{aligned} \tag{4.11}$$

In order to create a risk-free strategy with this portfolio, we will need to remove the uncertainty from the portfolio dynamics, hence the portfolio will be risk-free when

$$\Delta_t^{(1)}\sigma_A(t,T_1,V_t) + \Delta_t^{(2)}\sigma_A(t,T_2,V_t) = 0,$$

hence take $\Delta_t^{(1)}$ and $\Delta_t^{(2)}$ such that

$$\frac{\Delta_t^{(1)}}{\Delta_t^{(2)}} = -\frac{\sigma_A(t,T_2,V_t)}{\sigma_A(t,T_1,V_t)}. \tag{4.12}$$

When we choose these amounts, the uncertainty in (4.11) vanishes and we are left with

$$\frac{d\Pi}{\Pi} = \left[\Delta_t^{(1)}\mu_A(t,T_1,V_t) + \Delta_t^{(2)}\mu_A(t,T_2,V_t)\right]dt.$$

Now since this portfolio is risk-free, we know by the no-arbitrage assumption that this portfolio must grow with the risk-free rate r. Since our portfolio starts with a value of \$1, we have:

$$\frac{d\Pi}{\Pi} = rdt$$
$$\leftrightarrow$$
$$\Delta_t^{(1)}\mu_A(t,T_1,V_t) + \Delta_t^{(2)}\mu_A(t,T_2,V_t) = r.$$

Since $(\Delta_t^{(1)} + \Delta_t^{(2)}) = 1$ due to the portfolio having a value of \$1 at time t, we obtain

$$\Delta_t^{(1)}\mu_A(t,T_1,V_t) + \Delta_t^{(2)}\mu_A(t,T_2,V_t) - (\Delta_t^{(1)} + \Delta_t^{(2)})r = 0$$
$$\leftrightarrow$$
$$\frac{\Delta^{(1)}}{\Delta^{(2)}} = -\frac{\mu_A(t,T_2,V_t) - r}{\mu_A(t,T_1,V_t) - r}. \tag{4.13}$$

Using (4.12) and (4.13), we obtain:

$$\frac{\mu_A(t,T_2,V_t) - r}{\sigma_A(t,T_2,V_t)} = \frac{\mu_A(t,T_1,V_t) - r}{\sigma_A(t,T_1,V_t)}.$$

Since we arbitrarily chose T_1 and T_2, we know that these fractions are independent of the expiration date T. Let us therefore now define q as

$$q(t,V_t) = \frac{\mu_A(t,T,V_t) - r}{\sigma_A(t,T,V_t)}, \quad T \geqslant s. \tag{4.14}$$

Rewriting (4.14) results in:

$$\mu_A(t,T,V_t) - r = q(t,V_t)\sigma_A(t,T,V_t). \tag{4.15}$$

Let us now substitute $\mu_A(t,T,V_t)$ with (4.9). Note that the term $\mu_A(t,T,V_t) - r$ in $q(t,V_t)$ represents the risk premium for volatility, hence we can replace this with ξV_t for a constant ξ. With all these terms replaced in (4.15), we obtain the risk-neutral pricing PDE:

$$\frac{\partial A}{\partial t} + (\alpha - \kappa V_t)\frac{\partial A}{\partial V_t} + \frac{1}{2}\sigma^2 V_t^{2\gamma}\frac{\partial^2 A}{\partial V_t^2} - rA = \xi V_t \frac{\partial A}{\partial V_t}$$

$$\leftrightarrow$$

$$\frac{\partial A}{\partial t} + (\alpha - (\kappa+\xi)V_t)\frac{\partial A}{\partial V_t} + \frac{1}{2}\sigma^2 V_t^{2\gamma}\frac{\partial^2 A}{\partial V_t^2} - rA = 0, \tag{4.16}$$

From the Feynmac-Kac Theorem, we clearly see that the pricing equation in (4.16) implies the stochastic process for V_t to be (Oosterlee & Grzelak, 2019, Chapter 3.2):

$$dV_t = (\alpha - (\kappa+\xi)V_t)dt + \sigma V_t^{\gamma} d\hat{W}_t, , \tag{4.17}$$

with $\hat{W}_t$ as a Brownian motion under the risk-neutral measure Q. We know this is the risk-neutral measure, since the pricing PDE follows from the risk-free portfolio that we created. We see that (4.17) is the result of $\hat{W}_t = \tilde{W}_t$ as Brownian motion transformation on the original process, hence we know that the measure Q that follows from Girsanov's Theorem using $\chi_t = \frac{\xi V_t}{\sigma V_t^{\gamma}}$ results in the risk-neutral measure. □

Corollary 4.2. *Suppose V_t is the stochastic differential equation as in (4.1), then the price of a futures contract at time t with expiration date T is*

$$\begin{aligned} F(t,T,V_t) &= \mathbb{E}_Q[V_T] \\ &= e^{-\beta(T-t)}V_t + \frac{\alpha}{\beta}(1 - e^{-\beta(T-t)}), \end{aligned}$$

with $\beta = \kappa + \xi$ and Q the risk-neutral measure.

Proof. From Corollary 4.1, we know that the Brownian motion transformation $\tilde{W}_t = W_t + \frac{\xi V_t}{\sigma V_t^{\gamma}}$ results in the risk-neutral process of V_t, which is:

$$dV_t = (\alpha - \beta V_t)dt + \sigma V_t^{\gamma} d\tilde{W}_t, \tag{4.18}$$

where $\beta = \kappa + \xi$. Now the theoretical value of a futures contract is $\mathbb{E}_Q[V_T]$, which follows from Cox et al. (1981). They show there that one can create a risk-free strategy that continuously buys and sells futures contracts and that simultaneously borrows and invests in the risk-free rate. This strategy results in a risk-free payoff of $e^{r(T-t)}V_T$, hence in the risk-neutral world the futures contract must have a value of $e^{-r(T-t)}\mathbb{E}_Q[e^{r(T-t)}V_T] = \mathbb{E}_Q[V_T]$.

To find the futures price, let us first apply Itô's lemma on the transformation $g(t,V_t) = e^{\beta t}V_t$, which results in:

$$\begin{aligned} dg(T,V_T) &= \left(\beta e^{\beta T}V_T + (\alpha - \beta V_T)e^{\beta T}\right)dT + \sigma V_T^{\gamma} e^{\beta T} d\tilde{W}_T \\ &= \alpha e^{\beta T} dT + \sigma V_T^{\gamma} e^{\beta T} d\tilde{W}_T \\ &= \frac{\alpha}{\beta}(e^{\beta T} - e^{\beta t}) + \sigma V_T^{\gamma} e^{\beta T} d\tilde{W}_T, \end{aligned}$$

hence

$$e^{\beta T}V_T - e^{\beta t}V_t = \frac{\alpha}{\beta}(e^{\beta T} - e^{\beta t}) + \sigma V_T^{\gamma} e^{\beta T} d\tilde{W}_T$$

$$\leftrightarrow$$

$$V_T = e^{-\beta(T-t)}V_t + \frac{\alpha}{\beta}(1 - e^{-\beta(T-t)}) + \sigma V_T^{\gamma} e^{-\beta(T-t)} d\tilde{W}_T.$$

Now the theoretical price of a futures contract at time t with expiration date T is

$$\begin{aligned}F(t,T,V_t) &= \mathbb{E}_{\mathbb{Q}}[V_T]\\ &= V_t e^{-\beta(T-t)} + \frac{\alpha}{\beta}(1 - e^{-\beta(T-t)}).\end{aligned}$$

□

The result from Corollary 4.2 will essentially be the basis model we will be using for our futures term structure model. The model will therefore depend on the two variables α and β, which we will estimate with Bayesian estimation methods in the next section. Note that the model does not depend on the σ and γ parameters of the process, since we saw in the proof of Corollary 4.2 that the risk-neutral expectation make these parameters vanish. The reason we do not need to include jumps in the CEV model is because the theoretical futures price from Corollary 4.2 when we include jumps would only replace α with $\alpha + \mu\lambda$, with λ the jump probability and μ the jump size parameters (Dupoyet et al., 2011). Therefore, we can estimate $\alpha + \mu\lambda$ as 1 parameter and end up with the same results. Also note that V_t in the model is not a variable to be estimated, since this is the initial value of the process and can simply be observed. The risk-neutral process of V_t as in (4.18) is similar to the original process in the real-world measure as in (4.1), but only the constant κ is replaced with the constant $\beta = \kappa + \xi$. Therefore, the resulting risk-neutral process is still a CEV model and should have similar properties. The process has a reversion to the mean $\frac{\alpha}{\beta}$ with a theoretical speed of mean reversion represented by β.

Let us look at a few visual representations of our model and how our parameters influence the term structure. Suppose we start at $t = 0$ and $V_0 = 10$, with parameter values $\alpha = 30$ and $\beta = 2$. This results in the term structure:

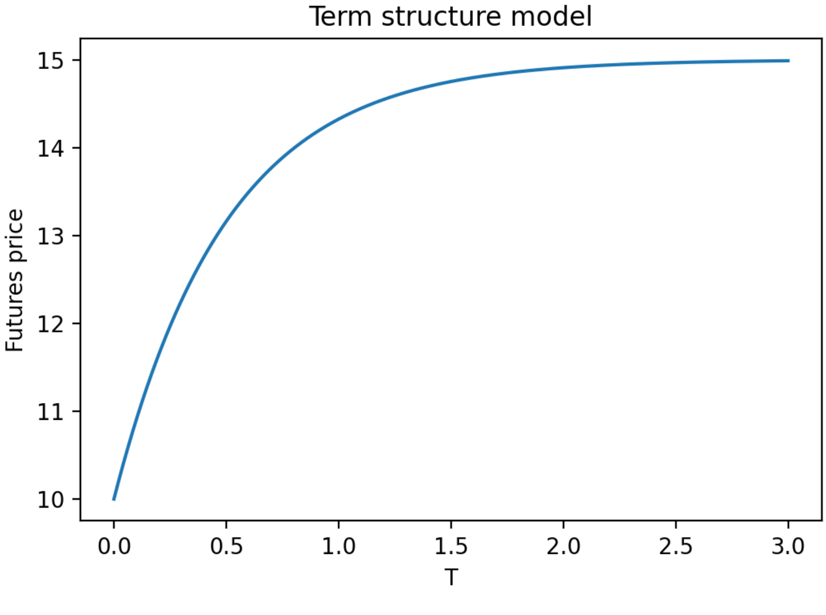


Figure 4.2: Term structure model with $\alpha = 30$ and $\beta = 2$

One may notice in Figure 4.2 that the futures price is $V_0 = 10$ at $T = 0$ and that the term structure seems to converge to $\frac{\alpha}{\beta} = 15$ when the time to expiration increases. The reason why a futures contract with a low time to expiration is close to the spot price is because the futures price payoff is the value of the underlying asset V_t. When a contract is about to expire, then the probability of the payoff being a lot different than the spot price becomes really low, since we do not expect V_t to significantly change in value in a short period of time. For the futures contracts with a long time to expiration, it seems that the long term mean has a higher impact on the futures price than the spot price. The spot price of V_t will have less impact on futures prices with longer time to expiration, since there simply is more uncertainty in what will happen with V_t at time T. The expectation of V_T will therefore be closer to the long term mean when T increases. From the formula of the futures price at time t we can also see these effects occurring:

$$\lim_{T\to t} F(t,T,V_t) = e^{-\beta(t-t)}V_t + \frac{\alpha}{\beta}(1-e^{-\beta(t-t)}) = V_t,$$
$$\lim_{T\to\infty} F(t,T,V_t) = 0 * V_t + \frac{\alpha}{\beta}(1-0) = \frac{\alpha}{\beta}.$$

Suppose now we start at $t = 0$ with $V_0 = 10$ and we would like to see the influence of the speed of mean reversion parameter β. Let us fix the long term mean $\frac{\alpha}{\beta}$ at 15 and alter the value of the speed of mean reversion β. Note that we also have to adjust the value of α to remain at constant long term mean level.

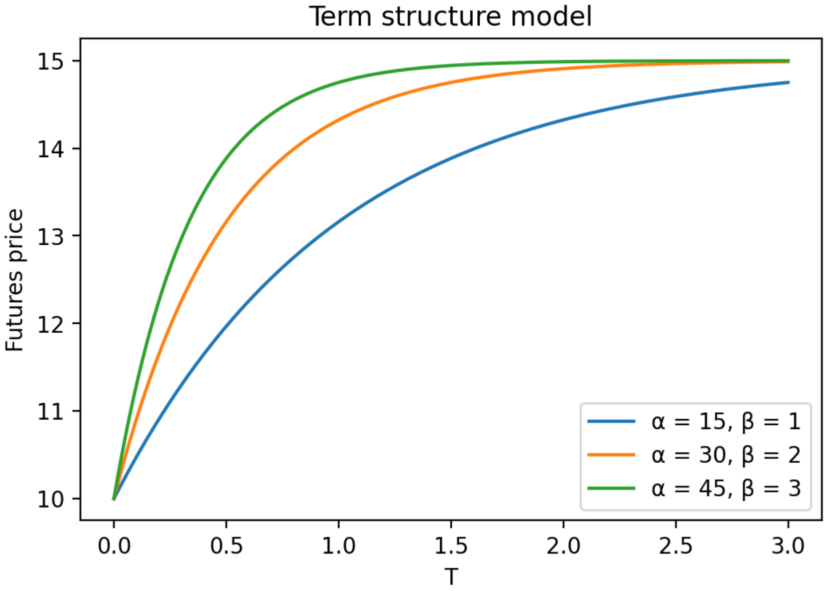


Figure 4.3: Term structure model with a constant long term mean, but different speeds of mean reversion.

In Figure 4.3 we clearly see the effect of the speed of mean reversion parameter. When the speed of mean reversion increases, the futures price seem to also revert faster to the long term mean. This is the logical effect that we expect to happen when the process V_t reverts faster to its long term mean. We then have more uncertainty on what happens with V_T, hence the expected value of the process gets closer to the long term mean, which implies that the futures price gets closer to the long term mean.

For the final visualisation, suppose we now fix the speed of mean reversion β at 2 and alter the value of the long term mean $\frac{\alpha}{\beta}$, thus we only alter the value of α. Let us again start at $t = 0$ with $V_0 = 10$.

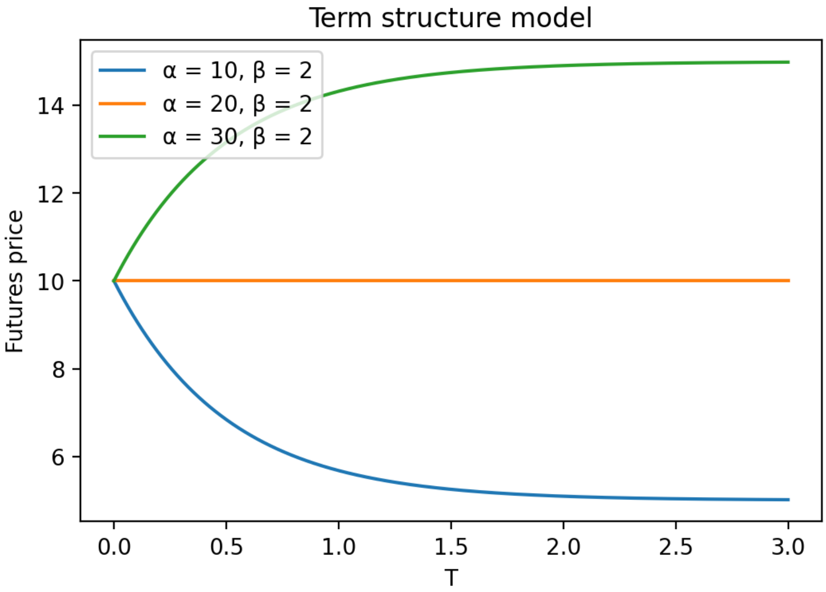


Figure 4.4: Term structure model with a constant speed of mean reversion, but different long term means.

Figure 4.4 shows the two different type of curves our model can make, which is a monotone increasing and a monotone decreasing curve starting from V_0. In financial terms, the market state is in contango when the term structures is an increasing curve and the market is in normal backwardation when it is a decreasing curve. What these states of the market imply for different type of investors is explained in Section 2.2.1. We see that the boundary between a contango market and a normal backwardation market is when the spot price is equal to the long term mean of the process. Hence, $\frac{\alpha}{\beta} > V_0$ implies that the market is in contango, while $\frac{\alpha}{\beta} < V_0$ implies the market to be in normal backwardation. These properties of the model may be useful when it comes to creating an investment strategy. After estimating our parameters of the model, we can determine the current market state and select the corresponding investment strategy.

4.2.2 Bayesian parameter estimation

The next step will be to find an estimation method for our parameters in the term structure model. The idea of our model is to use Bayesian statistics to find an estimate of our futures price. In the previous section in Corollary 4.2, we showed that the theoretical futures price at time t is the function:

$$F(t, T, V_t) = e^{-\beta(T-t)}V_t + \frac{\alpha}{\beta}(1 - e^{-\beta(T-t)}).$$

Ultimately we want to use this theoretical futures price as the basis of our term structure model. We understand that the theoretical futures price will be different from the current market prices, however, we will make the assumption that the current market prices are samples from a distribution centered around the theoretical futures price. Hence, if M is the market price of some futures contract with a theoretical futures price F, then we assume that M deviates from the theoretical futures price as:

$$M = F + \epsilon,$$

with ϵ being some random variable that represents the error. In our model we will assume our errors to behave as a zero-mean normal distribution with variance σ^2, hence we would have $y \sim \mathcal{N}(F, \sigma^2)$.

Now the parameters that we have to estimate of the theoretical futures price are α and β. With frequentist statistical methods, the idea is to find the optimal values of α, β and σ^2 such that the assumed normal distribution best represents the data y. One rather simple way to do this is to estimate α, β and σ^2 with a maximum likelihood estimator. This way we will end up with one value for α, β and σ^2. After this, we can construct confidence intervals based on our estimated parameters and the assumption of the normal distribution. However, the latter assumption may be rather dubious. We have no idea if the current market prices actually behave according to this distribution. We could attempt to tune the distribution of ϵ to make it more representing to the data M. In some situations, introducing a student's t-distribution or a generalized normal distribution and tuning its parameters may help overcoming the shortcomings in the tails in comparison with the assumed normal distribution. One could even try to apply the skewed variants of these distributions to optimally fit the data. The frequentist will then apply hypothesis tests on the assumed distributions and attempts to find a significant distribution.

Here is where Bayesian statistics can help us out. Although we still assume that the data y behaves as a normal distribution centered around F, we also assume that our parameters that we estimate behave as random variables. By assuming prior distributions for the parameters α, β and σ^2, we can set up a sampling scheme using MCMC algorithms. This allows us to sample from the posterior distribution of the parameters given the data, which gives us a simulation of futures prices. We now do not end up with confidence intervals, but with credible intervals determined by our simulation. We do now also have to make assumptions for our prior distributions for α, β and σ^2. However, the impact of our chosen prior distributions on the simulation depends on how informative these prior distributions are, which allows us to determine the impact of our prior beliefs. If we have strong prior beliefs on certain parameters, then we can allow the prior distributions to have a large impact, but if we are not certain on the behaviour of our parameters, we can simply choose a non-informative prior.

The shape of the term structure can be different from day to day, which would make it difficult to find accurate values for α, β and σ^2 that accurately represent the term structure over some time period. Therefore, we will be estimating these parameters daily with the current market prices of futures contracts of the upcoming expiration months. In our model for the term structure, we will be using the monthly futures contracts with the nearest E expiration dates. We will call the futures contracts with the kth nearest expiration date the k-month futures contract. Altogether, we will estimate the parameters of α and β daily, based on the current market prices of E futures contracts.

Let us start by formally introducing our term structure model. Since we are modelling the term structure for individual days, we will not assign a specific parameter for which day we are modelling the term structure. The values of α and β used to be constant values which were determined by the CEV model, however, these values are now turned into parameters that we have to estimate. Therefore, we will add the parameters to the function of the futures price as $F(t, T, V_t, \alpha, \beta)$. Without loss of generality, we also assume from now on that we estimate our futures price at time $t = 0$, which means that the notation $F(T, V_0, \alpha, \beta)$ suffices for the notation of the futures price, where V_0 is also called the spot price of the VIX index. Now, let $T_1, T_2, \ldots, T_E$ represent time to expiration of the 1-E month futures contracts that are traded on a specific day. Furthermore, let us define $\mathbf{F}(\alpha, \beta)$ be the $(E \times 1)$ vector:

$$\mathbf{F}(\alpha, \beta) = \begin{bmatrix} F(T_1, V_0, \alpha, \beta) \\ F(T_2, V_0, \alpha, \beta) \\ \vdots \\ F(T_E, V_0, \alpha, \beta)) \end{bmatrix}.$$

The vector $\mathbf{F}(\alpha, \beta)$ now represents the theoretical values of the E futures contracts that we want to fit as good as possible to the current market prices with α and β. Now suppose the $(E \times 1)$ vector M represents the market prices of the 1-E month futures contracts. If we assume the market prices are independently sampled from a normal distribution centered around $\mathbf{F}(\alpha, \beta)$, we have:

$$M|\alpha, \beta, \sigma^2 \sim \mathcal{N}(\mathbf{F}(\alpha, \beta), \sigma^2 I), \tag{4.19}$$

with I a $(E \times E)$ identity matrix. This results in the likelihood:

$$p(M|\alpha, \beta, \sigma^2) \propto \sigma^{-E} e^{-\frac{(M-\mathbf{F}(\alpha,\beta))^T (M-\mathbf{F}(\alpha,\beta))}{2\sigma^2}}. \tag{4.20}$$

The next step would be to set up prior distributions for our parameters α, β and σ^2. We do have a small amount of prior information available of these parameters. We know that α and β are the parameters of a CEV model and hence should be positive. Obviously, the variance parameter σ^2 should be a positive value as well. Another piece of prior information of our parameters we have is that $\frac{\alpha}{\beta}$ should represent the long term mean of the process. However, to keep the model simple we assume the prior distributions for α and

β to be independent. This does not necessarily mean that the corresponding posterior distributions are independent as well. All we do is make the prior distribution less informative by not including dependence between the parameters. Any other sort of prior information on the behaviour of our parameters seems to be out of our reach, therefore it would be a good idea to select non-informative priors.

By clever selection of our prior distributions, we may be able to obtain posterior distributions from recognizable probability distribution families. Although this is not much of importance if we would use a No-U-Turn Sampler, it will be of importance when we use a Metropolis-Hastings algorithm. By proposing values for the parameters in the algorithm from the posterior distributions, we can use Gibbs sampler to always have an acceptance rate of 1 of the proposed value in the algorithm. However, since $\mathbf{F}(\alpha, \beta)$ is a complicated function, it seems difficult to see which prior distributions for α and β we can choose to apply Gibbs sampler. For the parameters α and β we will therefore simply assume a log-normal prior distribution, which is a common choice for positive parameters. Hence, we have the independent distributions:

$$\ln(\alpha) \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2), \tag{4.21}$$

$$\ln(\beta) \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2). \tag{4.22}$$

By assuming a large variance on these prior distributions, we will make sure that we have non-informative priors, which represents our prior belief of α and β. Now the values for μ_α and μ_β are rather irrelevant, since we have such a flat distribution for the priors when we choose a large variance. However, we do know that $\frac{\alpha}{\beta}$ should represent the long term mean in the CEV model. We could integrate this little bit of prior information into the prior distributions by alternating the values of μ_α and μ_β to the long term mean of the VIX index. But again, the impact of these values will be minimal to the posterior, hence we are rather free in choosing these hyperparameter values. Therefore, suitable values for the hyperparameters in (4.21) and (4.22) based on our prior beliefs are:

$$\begin{aligned}
\mu_\alpha &= \ln(50) \\
\sigma_\alpha^2 &= 10000 \\
\mu_\beta &= \ln(3) \\
\sigma_\beta^2 &= 10000.
\end{aligned}$$

Based on the likelihood in (4.20), it is easy to see after a few derivation steps that the inverse-gamma distribution for the variance parameter σ^2 is a conjugate prior, hence the posterior is also an inverse-gamma distribution. This implies that the posterior of σ^2 can simply be drawn from a known probability distribution, which allows us to use a Gibbs sampler step in the sampling algorithm. Thus, if we choose the independent prior distribution for σ^2 to be:

$$\sigma^2 \sim \mathcal{IG}(a, b),$$

then we end up with the posterior distribution:

$$\begin{aligned}
p(\sigma^2 | M, \alpha, \beta) &\propto p(M|\alpha, \beta, \sigma^2) p(\alpha) p(\beta) p(\sigma^2) \\
&\propto \sigma^{-E} e^{-\frac{(M-\mathbf{F}(\alpha,\beta))^T (M-\mathbf{F}(\alpha,\beta))}{2\sigma^2}} \sigma^{-2a-2} e^{-\frac{b}{\sigma^2}} \\
&\propto \sigma^{-2(a+\frac{E}{2})-2} e^{-\frac{\frac{1}{2}(M-\mathbf{F}(\alpha,\beta))^T (M-\mathbf{F}(\alpha,\beta))+b}{\sigma^2}},
\end{aligned}$$

hence we end up with the posterior distribution:

$$\sigma^2 | M, \alpha, \beta \sim \mathcal{IG}(a + \frac{E}{2}, \frac{1}{2}(M - \mathbf{F}(\alpha, \beta))^T (M - \mathbf{F}(\alpha, \beta)) + b).$$

Based on our uncertain prior belief on the behaviour of σ^2, we can choose values of a and b such that we obtain a non-informative prior distribution. A suitable choice for these hyperparameters would therefore be

$$\begin{aligned}
a &= 0.001, \\
b &= 0.001.
\end{aligned}$$

Altogether, we obtain the following hierarchical structure for our term structure model:

$$\begin{aligned} M \mid \alpha, \beta, \sigma &\sim \mathcal{N}(\mathbf{F}(\alpha,\beta), \sigma^2 I) \\ \ln(\alpha) &\sim \mathcal{N}(\ln(50), 10000) \\ \ln(\beta) &\sim \mathcal{N}(\ln(3), 10000) \\ \ln(\sigma) &\sim \mathcal{IG}(0.001, 0.001), \end{aligned}$$

with the target distribution:

$$\begin{aligned} \pi(\alpha,\beta,\sigma^2) &= p(\alpha,\beta,\sigma^2|M) \\ &\propto p(M|\alpha,\beta,\sigma^2)p(\alpha,\beta,\sigma^2) \\ &= p(M|\alpha,\beta,\sigma^2)p(\alpha)p(\beta)p(\sigma^2) \\ &\propto \sigma^{-E} e^{-\frac{(M-\mathbf{F}(\alpha,\beta))^T(M-\mathbf{F}(\alpha,\beta))}{2\sigma^2}} \frac{1}{10000\alpha} e^{-\frac{(\ln(\alpha)-\ln(50))^2}{20000}} \frac{1}{10000\beta} e^{-\frac{(\ln(\beta)-\ln(3))^2}{20000}} \\ &\quad \sigma^{-2(0.001+\frac{E}{2})} e^{-\frac{\frac{1}{2}(M-\mathbf{F}(\alpha,\beta))^T(M-\mathbf{F}(\alpha,\beta))+0.001}{\sigma^2}}. \end{aligned}$$

We can use MCMC methods such as the Metropolis-Hastings algorithm and the No-U-Turn Sampler to sample from this target distribution. The No-U-Turn Sampler simply requires the logarithm of the target distribution $\pi(\alpha,\beta,\sigma^2)$ to be known up to a constant. The Metropolis-Hastings algorithm requires us to set up a sampling scheme and select appropriate proposal distributions, which we will discuss in the following section.

4.2.3 Metropolis-Hastings sampling scheme

Using the hierarchical structure of the term structure model as described in the previous section, we apply the Metropolis-Hastings algorithm to set up a sampling scheme of the target distribution $\pi(\alpha,\beta,\sigma^2)$. We have seen that the posteriors of α and β are both not convenient distributions that can be easily sampled from. Thus, we will use a random-walk Metropolis-Hastings algorithm to be able to sample from α and β. However, the posterior of σ^2 is simply an inverse-gamma distribution, which allows us to use Gibbs sampler for σ^2. Both of these MCMC sampling methods are explained in Chapter 3.

For the random-walk proposal distribution for (α,β), we simply use a normal distribution centered around the previous sampled value, hence

$$q(z \mid (\alpha_{m-1},\beta_{m-1})) \sim \mathcal{N}((\alpha_{m-1},\beta_{m-1}), \sigma_q^2 I),$$

with I as the (2×2) identity matrix and σ_q^2 being the variance which we have to select in such a way that our simulation runs smoothly. As discussed in Section 3.2.2, one should carefully tune the parameter σ_q^2 in such a way that the random-walk step size is not too small nor too large. Problems may occur with the convergence speed of the Metropolis-Hastings algorithm when one chooses the wrong variance σ_q^2. Since the proposal distribution is a symmetric distribution, we have that the acceptance probability only depends upon $\pi(\cdot,\sigma_{m-1}^2)$ as:

$$\alpha = \min\left\{1, \frac{\pi(z,\sigma_{m-1}^2)}{\pi(\alpha_{m-1},\beta_{m-1},\sigma_{m-1}^2)}\right\}$$

By combining the random walk proposal and Gibbs sampler, we can create the following sampling algorithm:

Algorithm 7: Term structure model sampling algorithm

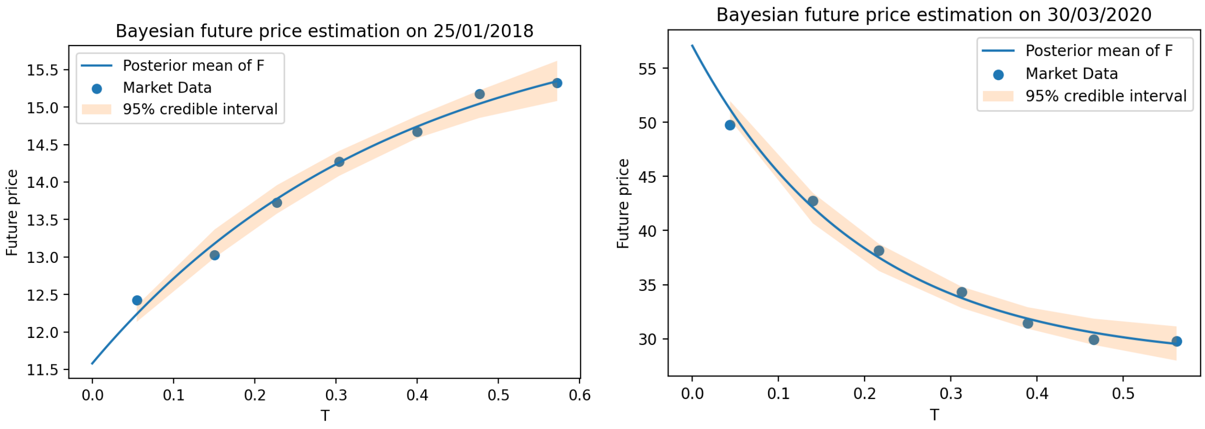
```
Given $\alpha_0, \beta_0, \sigma_0^2, \sigma_q^2, E, M$:
for $m$ = 1 to $M$ do
    Sample $z \sim \mathcal{N}(\alpha_{m-1}, \beta_{m-1}, \sigma_q^2)$.
    Set $(\alpha_m, \beta_m) \leftarrow (\alpha_{m-1}, \beta_{m-1})$.
    With probability $\alpha = \min\left\{1, \frac{\pi(z, \sigma_{m-1}^2)}{\pi(\alpha_{m-1}, \beta_{m-1}, \sigma_{m-1}^2)}\right\}$, set $(\alpha_m, \beta_m) \leftarrow z$.
    Sample $\sigma_m^2 \sim \mathcal{IG}(0.001 + \frac{E}{2}, \frac{1}{2}(M - \mathbf{F}(\alpha_m, \beta_m))^T (M - \mathbf{F}(\alpha_m, \beta_m)) + 0.001)$.
end
```

Before starting the simulation, we have to initiate the Markov chain for the parameters. Since we will apply a burn-in period in the beginning of the simulation, the initial values will minimally impact the stationary distribution of the Markov chain. Therefore, we can simply choose starting values such as:

$$\begin{aligned}\alpha_0 &= 50\\ \beta_0 &= 3\\ \sigma^2 &= 1.\end{aligned}$$

4.2.4 Drawbacks of the term structure model

The current model for term structure is on most days nicely able to fit the term structure. However, in certain situations this model is not able to give an accurate representation of the term structure. Although we will later on go over the simulation results of the term structure model in Section 5.1, we can already look at some results of the No-U-Turn Sampler to get an insight in the shortcomings of the current model. Let us look a few examples where the term structure model makes an accurate fit:



(a) Contango market situation.

(b) Normal backwardation market situation.

Figure 4.5: Example of days where the term structure model is performing well. The simulation is performed using a No-U-Turn Sampler.

In Figure 4.5a and 4.5b our term structure model is represented by the posterior mean of the theoretical futures price F using the simulated values of α and β, including the 95% credible intervals. Here we see perfectly normal behaviours of the futures market. Given the spot price of the VIX index, the market expects the VIX index to move upwards or downwards over time, resulting in similar behaviour for the prices of the futures contracts. However, let us now look at two situations in which our model seems to fail to model the term structure.

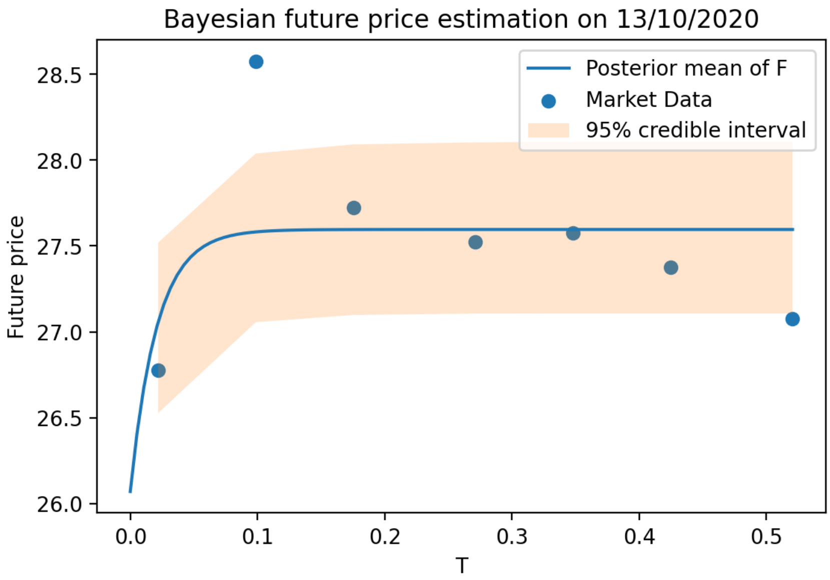


Figure 4.6: Situation when an event is taking place in the market, that increases the price of a specific futures contract. The increase in price of the 2-month futures contract is not an outlier, but the result of the 2020 U.S. presidential election. The simulation is performed using a No-U-Turn Sampler.

The first situation where our term structure model seems to fail is when some specific event is taking place in the market near the expiration date of a futures contract, which will guarantee a certain behaviour of the VIX index. For example, in Figure 4.6 we see a sudden increase in the price of the 2-month futures contract. The behaviour of the price of this specific futures contract can be explained by the event that is taking place near the expiration date of this contract, which is the 2020 U.S. elections that took place on November 3rd, 2020. An investor knows beforehand that the U.S. markets will react to the result of the elections, hence an increase in volatility is definitely expected near the election date. The 2-month futures contract in Figure 4.6 expires on November 18th, 2020. Therefore, an increase in the price of this futures contract is expected behaviour, since the volatility and hence the VIX index is expected to be higher near this period.

However, our term structure model is not designed to take these events into account. The term structure model can simply not capture the movements of these specific contracts. Although here the term structure model is still able to model the rest of the futures contracts, the problem is that it considers the 2-month futures contract as overpriced. The term structure model tells us that the 2-month futures contract is a strong outlier, based on our credible intervals. If an investor would use this model, then he would be tempted to take a short position on this futures contract, since it is considered overpriced. However, we preemptively know that the VIX index has a higher value around the expiration date, which would mean that this futures contract does not have to be overpriced at all.

The second situation in which the term structure model seems to fail can be illustrated in the following figure:

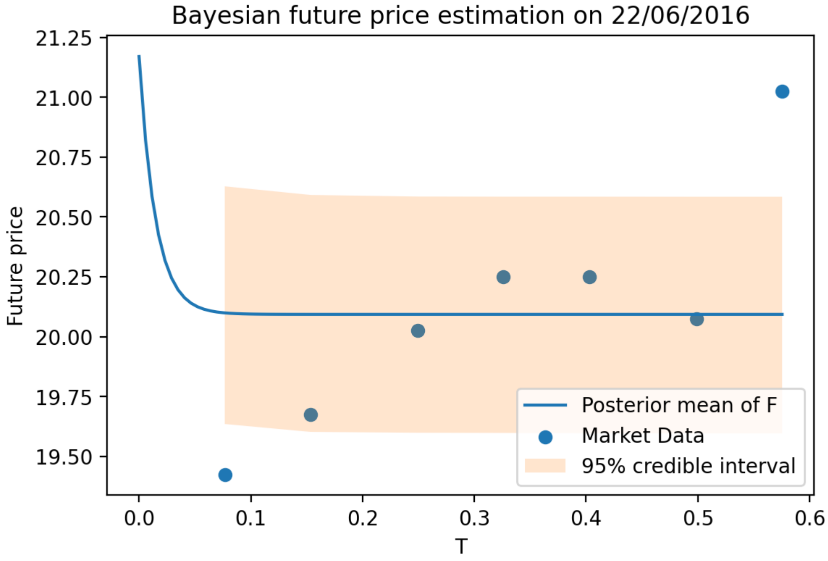


Figure 4.7: Situation where the prices of the futures contracts behave in a non-monotone way with respect to the spot price of the VIX index. The simulation is performed using a No-U-Turn Sampler.

In Figure 4.7 we clearly see a bizarre result from our term structure model. This behaviour of our model undoubtedly misrepresents the current market data of the futures prices. We can explain why our model is behaving this way by looking at the spot price of the VIX index V_0. In Figure 4.7, the spot price of the VIX index on the 22nd of June in 2020 is 21.17, which we also see for $T = 0$ in the figure. If we look at the theoretical futures price from Corollary 4.2, we may notice that the futures price is a monotone increasing or monotone decreasing function. When $V_0 > \frac{\alpha}{\beta}$, then the theoretical futures price is a monotone decreasing function, while $V_0 < \frac{\alpha}{\beta}$ implies it to be a monotone increasing function. The term structure model will always attempts to fit some monotone curve from V_0 to $\frac{\alpha}{\beta}$ based on the parameters α and β. When the spot price of the VIX index of a certain day is in the unfortunate position of not having a monotone curve between V_0 and the long term mean that represents the current market data, then results similar as in this figure will be obtained.

4.3 Error correction model

We saw in the previous section that the term structure model still has its limitations when it comes to the modelling of the term structure. The capability to only model monotone functions between V_0 and $\frac{\alpha}{\beta}$ resulted that on certain days the term structure with unusual behaviour could not accurately be represented with our model. We have also seen that the term structure model is unable to detect whether there is a certain event happening near the expiry date of the future. In this section we would like to introduce an additional model that will help us repair the shortcomings of our term structure model. By looking at the error that our term structure model made with comparison to the actual market data, we will try to perform some correction on the model, which is why we call this model the error correction model.

An important aspect of the error correction model is the difference of intentions with the term structure model when it comes to fitting the current market data. The term structure model will attempt to find the optimal values for α and β such that our model represents the current market data as good as possible. However, the intention of the error correction model is to optimally fit to the error of our term structure model. At first, it may look like these intentions are equivalent, however, the way we will fit the error of the term structure model creates different intentions for the error correction model. The way we will fit the error of our previous model is by introducing a few regression factors that we assume investors would use to determine whether a futures contract is considered overvalued or undervalued. We will apply these regression factors on data of the error of the term structure model of a specific futures contract instead of the entire term structure of E contracts. By doing so, we will for each individual futures contract investigate which factors had influenced the price of that specific contract in the past, which then allows us to determine whether this contract shows signs of being overpriced or underpriced by investors. Hence, the most important intention of the error correction model will be to investigate the influence of certain factors on the pricing behaviour of investors.

4.3.1 Model formulation

The error correction model will be a regression analysis on the error that the term structure model made on a specific futures contract. Let us first introduce the model by looking at how we would model the error y of the term structure model of a single futures contract on a single day. Suppose F represents the posterior mean of the theoretical futures price with the simulated parameters of the term structure model of that specific futures contract and M represents the market price of that specific futures contract, then we define the error of our term structure model on this specific futures contract as

$$y = F - M.$$

The idea is to use regression factors in attempt to model the error y. After we modeled the error y, we can subtract it from the posterior mean F to obtain a corrected version of the futures price. The regression factors that we will be using is based upon the final section of Dupoyet et al. (2011, pp. 333-335). In that paper a small analysis is performed on the error of a futures pricing model with a few regression factors. Now the idea of our model is, instead of only analysing our model, to use these regression factors to actually adjust the initial term structure model. These regression factors will represent the factors that an investor would look at when pricing the specific futures contract. We will use the following regression factors to investigate the behaviour of the y:

- An intercept. When the term structure model is consistently overpricing (or underpricing) a certain futures contract, then the intercept will catch this behaviour of our model. In this case, investors would be willing to pay more (or pay less) for a specific futures contract than the theoretical futures price. This could imply the existence of some event that happens near the expiration date of the futures contract, or some other reason why this specific futures contract would have more (or less) value on the expiration date.
- Time to the expiration of our futures contract T. By investigating the influence of the time to expiration of the futures contract, we will investigate whether investors have different pricing behaviours for contracts with different expiration dates.
- Level of our futures price $\log(M)$. With this factor, we will investigate whether investors have different pricing behaviours based on the level of the futures price.
- Difference of the spot VIX index and the futures price $\frac{M-V_0}{V_0}$. If this difference is rather large, than investors expect the VIX index to have a larger change between now and the expiration date, which implies the investor to expect a higher speed of mean reversion of the VIX index. Including this factor in the model will therefore allow us to investigate the pricing behaviour of investors based on the speed of mean reversion of the term structure.
- Change of the spot VIX index $\frac{V_0-V_{-1}}{V_{-1}}$, with V_{-1} representing the spot price of yesterday. This factor allows us to investigate the behaviour of investors when jumps occur in the VIX index.

When any of these factors show significant behaviour in the error of the term structure model, then this will tell us that the term structure model is reacting different to these factors than the investors when the futures contracts are priced. When these factors do not show significant behaviour, then the term structure model shows similar pricing behaviour as the investors with these factors. All these factors together create the multiple regression for the error y of the specific futures contract on a specific day:

$$y = \beta_0 + \beta_1 T + \beta_2 \log(M) + \beta_3 \frac{M - V_0}{V_0} + \beta_4 \frac{V_0 - V_{-1}}{V_{-1}} + \epsilon,$$

where ϵ is a random variable and $\beta_0, \ldots, \beta_4$ are constants. In the error correction model, we will assume ϵ to be a random variable from a zero-mean normal distribution with variance σ^2, hence $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

Let us now define Y as the $(N \times 1)$ vector containing the errors of the first model of a specific futures contract of the previous N days. It is important to note that the errors in Y are from the same specific futures contract that we chose to investigate. This way we can investigate the behaviour of the regression factors for each individual futures contract. For the value of N we will be using all possible data that we have available. At a specific day, there will be more data available of the error of the 1-month futures contract than of the E-month futures contract. Since the term structure model uses data up to the nearest E futures contract, we know that this specific 1-month futures already has been priced by the term structure model in the past $E - 1$ months, while the E-month futures contract has not been priced for one month yet. Therefore, each day the total amount of data we have for a specific futures contract can be different, which is denoted with N. Let us define X as the $(N \times 5)$ regression matrix, with the regression factors of the ith error in Y on the ith row, hence X is defined such that

$$Y = X\beta + \epsilon,$$

with $\beta = \left[\beta_0, \dots, \beta_4\right]^T$ and ϵ some N-dimensional random variable, which we define in our model to be a zero-mean multivariate normal distribution as $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$.

The drawback of the way we formulate this model, is that we are not able to include the credible intervals of the term structure model into the error correction model, since we use the posterior mean of the term structure model. This will mean that the uncertainty of our posterior mean will not be included in the correction of the model. The posterior mean only provides a basic structure of the term structure, which will then be corrected with the help of the regression factors.

4.3.2 Bayesian parameter estimation

As we have seen in the term structure model, one is able to apply frequentist methods to the model to estimate its parameters. However, we would then need to look at different error distributions and apply statistical hypothesis tests in order to find a significant distribution. Bayesian methods allow us to assume the parameters β and σ^2 to be random variables and simulate the values of these parameters. Hence, we obtain the likelihood for Y as:

$$\begin{aligned} Y \mid \beta, \sigma^2 &\sim \mathcal{N}(X\beta, \sigma^2 I) \\ &\propto \sigma^{-N} e^{-\frac{(Y-X\beta)^T(Y-X\beta)}{2\sigma^2}}. \end{aligned}$$

The next step for Bayesian parameter estimation will be to assume prior distributions for our parameters. However, we do not have any prior information on the parameters, besides that σ^2 should be a positive value. Therefore, a good idea would be to select non-informative prior distributions for our parameters. Following Rachev et al. (2008, pp. 46-47), we will assume the prior distribution:

$$p(\beta, \sigma^2) \propto \frac{1}{\sigma^2}.$$

This well-known prior distribution is also known as Jeffreys prior for location and scale parameters (Jeffrey, 1946). Note that this is a non-informative prior distribution for both parameters, since we assume β to be uniform over $\mathbb{R}$ and σ^2 to be very wide spread as well. An interesting property of this prior distribution is that this is an improper prior distribution for (β, σ^2), which means $p(\beta, \sigma^2)$ is not a valid distribution function, or that we can not find a constant c such that

$$c \int_{-\infty}^{\infty} \int_{0}^{\infty} \frac{1}{\sigma^2} d\sigma^2 d\beta = 1.$$

However, improper priors can still be used as prior distributions, as long as the posterior distribution remains a proper distribution. We will see that the posterior distributions that we end up with will be proper distributions, hence we are allowed to use this improper prior distribution. For the posterior distribution of (β, σ^2), we have the target distribution:

$$\begin{aligned} \pi(\beta, \sigma^2) &= p(\beta, \sigma^2) \\ &\propto p(Y|\beta, \sigma^2) p(\beta, \sigma^2) \\ &\propto \sigma^{-N-2} e^{-\frac{1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)} \end{aligned}$$

The posterior of β is now the following:

$$\begin{aligned}
p(\beta|y,\sigma^2) &\propto e^{-\frac{1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)} \\
&\propto e^{-\frac{1}{2\sigma^2}(-2Y^TX\beta+\beta^TX^TX\beta)} \\
&= e^{-\frac{1}{2\sigma^2}(-2Y^TX(X^TX)^{-1}(X^TX)\beta+\beta^TX^TX\beta)} \\
&\propto e^{-\frac{1}{2\sigma^2}(-2Y^TX(X^TX)^{-1}(X^TX)\beta+\beta^TX^TX\beta+((X^TX)X^TY)^T(X^TX)((X^TX)X^TY)} \\
&= e^{-\frac{1}{2\sigma^2}(\beta-\hat{\beta})^T(X^TX)(\beta-\hat{\beta})},
\end{aligned}$$

with

$$\hat{\beta} = (X^TX)^{-1}X^TY,$$

hence

$$\beta|Y,\sigma^2 \sim \mathcal{N}(\hat{\beta}, \sigma^2(X^TX)^{-1}).$$

Note that $\hat{\beta}$ is equivalent to the least squares estimator of a multiple regression. For σ^2, we end up with the posterior distribution:

$$\begin{aligned}
p(\sigma^2|Y,\beta) &\propto \sigma^{-N}e^{-\frac{1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)}\sigma^{-2} \\
&\propto (\sigma^2)^{-\frac{N}{2}-1}e^{-\frac{1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)}
\end{aligned}$$

hence

$$\sigma^2|Y,\beta \sim \mathcal{IG}(\frac{N}{2}, \frac{(Y-X\beta)^T(Y-X\beta)}{2}).$$

Both posterior distributions are proper distributions, hence the assumption of an improper prior distribution is still valid. Since the posterior distributions are distributions from recognizable distribution families, we can easily sample from the posterior distributions. This allows us to set up a simple Gibbs sampler algorithm to sample from the posterior distribution, which is as follows:

Algorithm 8: Error correction model sampling algorithm

```
Given $\beta_{(0)}, \sigma_0^2, M$:
for $m = 1$ to $M$ do
    Sample $\beta_{(m)} \sim \mathcal{N}((X^TX)^{-1}X^TY, \sigma_{m-1}^2(X^TX)^{-1})$.
    Sample $\sigma_m^2 \sim \mathcal{IG}(\frac{N}{2}, \frac{(Y-X\beta_{(m)})^T(Y-X\beta_{(m)})}{2})$.
end
```

Note that in algorithm 8 we use the notation $\beta_{(m)}$ to make sure we are not confusing the Markov chain for β with the individual regression parameters $\beta_0, \ldots, \beta_4$. Before starting the simulation, we have to initiate the Markov chain for the parameters. Since we will apply a burn-in period in the beginning of the simulation, the initial values will minimally impact the stationary distribution of the Markov chain. Therefore, we can simply choose starting values such as:

$$\beta_{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\sigma_0^2 = 1.$$

4.4 Strategy

Now that we have created two models for the modelling of the term structure, we will attempt to use these models to create an investment strategy based on VIX futures. The idea of the strategy is very simple, we buy VIX futures that are undervalued and we short sell when they are overvalued. This implies that we are holding two different positions in futures:

- A long position. The investor that takes a long position buys the futures contract for a certain price at the start of the position and will later on sell the contract or let the contract expire. When the contract expires, he receives a certain cash settlement that is based upon the value of the VIX index.

- A short position. The investor that takes a short position receives the value of the futures contract at the start of the position. On a later date, the investor is obliged to repurchase the futures contract or pay out the cash settlement.

The simulation of the models that we created will result in a simulated posterior distribution of our futures price. We will make the assumption that this posterior distribution is the actual distribution of how the market would price these futures contracts. Now by making use of the credible intervals of the posterior distribution, we can determine whether we should take a long or a short position on the futures contract. The idea of our model is to assume a futures contract is overpriced when the market price lies above the range over the credible interval, while we assume that the futures contract is underpriced when the price lies below the credible interval. We will take a short position when the futures contract is overpriced, while we will take a long position when the future is underpriced. The idea is that in theory the price of the futures contract should be within the credible interval with a certain probability. When the futures contract lies outside the interval, then we will assume that investors will notice that this futures contract is incorrectly priced and hence the market will correct itself. The time interval in which we could obtain a profit is the period between a futures contract becoming overpriced or underpriced and the market adjusting itself to the correct valuation of the futures price. This means that our strategy will be to hold a position in the futures contract when it lies outside the credible interval of the posterior distribution and sell our position when the futures contract is back in the credible interval.

In our strategy we would only like to take positions in futures contracts that have high liquidity. In general, liquidity refers to the difficulty of buying and selling a financial product or asset. When the liquidity of futures contracts is high, then this means that lots of contracts of this future are actively traded on the market, resulting in a small bid-ask spread. However, when the liquidity of a futures contract is low, then this means that the contract is not actively traded at all, which results in a wide bid-ask spread. When the bid-ask spread is wide, it is less beneficial for an investor to buy or sell the futures contract, since the buy and sell offers have a less favorable value for the investor. As discussed in Section 2.2.1, the VIX futures contracts with lower time to expiration are being traded most often, while futures contract with further expiration dates are being traded less. To avoid a large bid-ask spread of VIX futures, we will in our strategy only trade the 1-4 month futures contracts. Each day, we will check for each of these four contracts whether the current market price lies outside the credible interval of the posterior distribution of the futures price. Again, when an outlier is detected, the strategy will respond by taking a long or short position in the contract, until the the market price is back in between the credible interval. Although we will simulate the models for the term structures using more than four futures contracts, we will only take positions in the first four contracts.

The amount of futures that we buy or sell when the strategy tells us to take a long or short position in the futures contract is a factor that we can optimize. A simple method to start with is to allocate a certain amount of money to each of the four futures contracts, for which we will not take a position worth more than this amount. The total amount of money that we will limit ourselves to in our strategy is called the budget. For example, a reasonable strategy is to assume that an investor has a budget of a certain amount, which we split evenly across the four futures contracts to invest with. This can of course be made more complicated in order to optimize the strategy performance. In Section 4.4.1 we will discuss a method of the allocation of money to the futures contracts to stabilize the volatility of the strategy returns, which is called volatility targeting. Furthermore, in Section 4.4.2 we will present the Sharpe ratio, which is a general performance measure of an investment strategy. Finally, in Section 4.4.3

we will discuss some evaluation methods that are generally used to evaluate the results of an investment strategy.

4.4.1 Volatility targeting

Volatility targeting refers to the method of scaling the returns of the strategy based upon the volatility of these returns. By doing so, one attempts to stabilize the volatility of the strategy of the returns by scaling down the returns when the volatility becomes higher, while scaling up the returns when the volatility becomes lower. Volatility targeting has in general been proven to increase the portfolio returns and reduce the risk in comparison with the original portfolio in certain situations. Moreira and Muir (2017) show for example that volatility targeting of US equity portfolios increase their Sharpe ratios. Scaling of the returns is done by scaling the bet size (investment amount) of the strategy. Although the percentage return of the invested amount remains the same, one leaves out a part of the investment by scaling the bet size down, which results in lower returns of the strategy. This may be useful in periods of high volatility, when one needs to manage the risk that is involved when losing money with the strategy. Downscaling the betsize in this situation may help in reducing the size of the tail of the strategy returns. Conversely, one adds extra money to the investment when upscaling the betsize to increase the returns of the strategy. This may be useful in periods of low volatility, when one can afford to take extra risks.

There are various methods that exist for volatility targeting in portfolio strategies. For example, Harvey et al. (2018) propose a method of scaling the returns based upon a predetermined target volatility that the returns should end up with. However, this method makes use of ex post scaling, which uses information of the entire sample period to scale the returns. In practice, a method like this seems impossible to implement and analyse throughout the sample period. We will therefore make use of so-called conventional volatility targeting, as explained in Bongaerts et al. (2020). This method rescales the returns r_t at month t as follows:

$$r_t^{\text{scaled}} = r_t \times \frac{\sigma^{\text{target}}}{\hat{\sigma}_{t-1}},$$

where σ^{target} is the realized volatility up to and including month $t-1$, while $\hat{\sigma}_{t-1}$ is the realized volatility of month $t-1$. The idea of this volatility targeting method is that we compare the overall volatility of the strategy returns with the current volatility. When the current volatility is lower than the overall volatility over the strategy, or equivalently $\hat{\sigma}_{t-1} < \sigma^{\text{target}}$, then the fraction $\frac{\sigma^{\text{target}}}{\hat{\sigma}_{t-1}} > 1$, which implies we are upscaling the returns of the strategy. This makes sense, since in periods of low volatility we can afford to increase the bet size. On the other hand, when $\hat{\sigma}_{t-1} > \sigma^{\text{target}}$, then the fraction $\frac{\sigma^{\text{target}}}{\hat{\sigma}_{t-1}} < 1$, which implies we are downscaling the returns of the strategy. Reducing the bet size also makes sense in periods of high volatility, since it reduces the risk and the size of the tail of the returns.

Rescaling the returns only once a month might be too slow when we trade VIX futures. Volatility indices such as the VIX index are known to have jumps, which may suddenly change the current volatility of the strategy returns. Therefore, we will adjust this method by rescaling the returns on a daily basis. We can now determine the current volatility based upon the previous K days, where K is a hyperparameter that can be tuned using historical data.

4.4.2 Sharpe ratio

When investigating which investment strategy to choose, it is usually not good enough to select the strategy with the highest profits over some backtest period. Although high profits are of course preferable in an investment strategy, one also has to take the risk of the strategy into account. If the returns of a strategy are very high, but one might lose all of his invested money one day, then he might want to reconsider this strategy. A common performance measure that is used to measure the investment strategy performance is the Sharpe ratio, which was first introduced by Sharpe (1966). It is a measure that takes the average excess return (strategy return minus risk-free rate) of the strategy into account, while also taking the volatility of the strategy into account. Suppose rf_t is the risk-free rate at time t and the excess returns $r_t - rf_t$ are normally distributed with mean μ and variance σ^2, then the Sharpe ratio is defined as

$$SR = \frac{\mu}{\sigma}.$$

The higher the value of the Sharpe ratio, the better the performance of the strategy according to this measure. Since we use excess returns, a negative value of the Sharpe ratio ($SR < 0$) implies that the strategy is returning less on average than an investment in the risk-free rate, while a positive value of the Sharpe ratio ($SR > 0$) implies the strategy is returning more on average than an investment in the risk-free rate. An average return higher than the risk-free rate does not automatically imply that an investment in the strategy is better than an investment in the risk-free rate. If the volatility of the strategy returns are extremely high, then it may not be worth for an investor to take the risk to invest in the strategy. Therefore, the Sharpe ratio should not only be positive, but should also be not close to 0.

When we empirically estimate μ and σ, we obtain the empirical Sharpe ratio

$$\hat{SR} = \frac{\hat{\mu}}{\hat{\sigma}},$$

with

$$\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} r_i - rf_i \\ \hat{\sigma}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}((r_i - rf_i) - \hat{\mu})^2. \end{aligned}$$

Using the Delta method, one finds the follow asymptotic distribution of the estimator (Lo, 2002):

$$(\hat{SR} - SR) \xrightarrow{d} \mathcal{N}(0, \frac{1 + \frac{1}{2}SR^2}{n}). \tag{4.23}$$

Using (4.23) we can perform statistical hypothesis tests to determine the significance of our estimated Sharpe ratio. Until now, we still assumed our excess returns to behave as a normal distribution. However, in practice this assumption is almost always invalid. In reality the returns of a strategy often seem to be skewed and have fatter tails. The asymptotic distribution in (4.23) may therefore be unrealistic when it comes to hypothesis testing. Mertens (2002) showed that the assumed normal distribution may be discarded and replaced with any distribution, with the corresponding asymptotic distribution:

$$(\hat{SR} - SR) \xrightarrow{d} \mathcal{N}(0, \frac{1 + \frac{1}{2}SR^2 - \gamma_3 SR + \frac{\gamma_4 - 3}{4}SR^2}{n}), \tag{4.24}$$

where γ_3 and γ_4 represent the third standardized moment (skewness) and the fourth standardized moment (kurtosis) of the distribution of the returns respectively. When we empirically estimate the variance of the asymptotic distribution in (4.24) together with an empirical estimate of the skewness and kurthosis, we end up with the Probabilistic Sharpe Ratio (PSR). Given a certain benchmark Sharpe ratio SR^*, the Probabilistic Sharpe Ratio becomes (Bailey & Lopez de Prado, 2012):

$$\begin{aligned} \widehat{PSR}(SR^*) &= \mathbb{P}(SR^* \leqslant \hat{SR}) \\ &= \mathbb{P}(Z \leqslant \frac{(\hat{SR} - SR^*)\sqrt{n-1})}{\sqrt{1 - \hat{\gamma}_3\hat{SR} + \frac{\hat{\gamma}_4 - 1}{4}\hat{SR}^2}}), \end{aligned}$$

with $Z \sim \mathcal{N}(0,1)$ and the sample skewness and kurtosis:

$$\begin{aligned} \hat{\gamma}_3 &= \frac{\frac{1}{n}\sum_{i=1}^{n}((r_i - rf_i) - \hat{\mu})^3}{\hat{\sigma}^3} \\ \hat{\gamma}_4 &= \frac{\frac{1}{n}\sum_{i=1}^{n}((r_i - rf_i) - \hat{\mu})^4}{\hat{\sigma}^4}. \end{aligned}$$

4.4.3 Evaluation methods

When one create an investment strategy, he usually tries to backtest the strategy on a large historical data set to see if the strategy actually works. There are many ways one can evaluate the results of such a backtest, which also depends on what type of strategy one implements. In this section we will describe a few evaluation methods that we will use on our own backtest in Section 5.3.

An evaluation method that is often used to investigate the tails of the returns is the value at risk, or the VaR. This is a measure for the risk of the loss in a strategy and is often used for risk management purposes. With the value at risk we want to show that, for a certain probability α, we will not lose more than VaR_α during the strategy. For example, if our distribution of daily losses L is a standard normal distribution, then the value at risk for probability 0.975 is $\text{VaR}_{0.975}(L) \approx 1.96$, since the probability of the loss being higher than approximately 1.96 is 0.025, thus the probability of not having a loss higher than approximately 1.96 is 0.975.

A problem with the value at risk is that the distribution of the tail is not being taken into account, only the lowest value in the upper tail of the loss distribution represents the VaR. Hence, the VaR only represents the upper bound of the loss of a strategy given a certain probability. However, in the case with probability $(1-\alpha)$ where the loss of the strategy is above $\text{VaR}_\alpha(L)$, then the VaR tells us nothing about how high our loss will be. For example, if we are 95% certain that the loss of a strategy will not be higher than \$100, then one may think that this strategy might not have that much risk. However, if there is a 3% probability that the strategy suddenly loses \$10000, then the investor will think differently about this strategy. The investor could not retrieve this information from the value at risk alone. A clear visual example of this phenomenon occurring can be seen in the distribution of returns below:

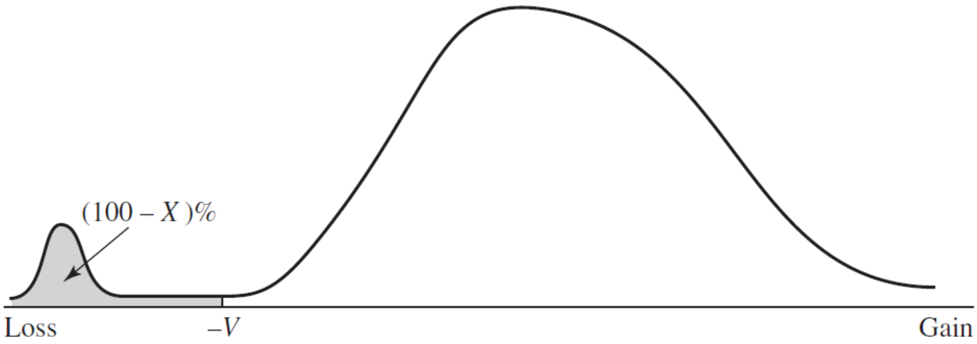


Figure 4.8: Example of the value at risk giving little information on the distribution of the tail (Hull, 2015, p. 273).

In Figure 4.8 we have $\text{VaR}_X(L) = V$, however, in the case with probability X when the loss is higher than V, we will have much higher losses than V. This means that the VaR of V may be misrepresenting the losses of the worst cases in the strategy. Therefore, we will be looking at another evaluation method, which is the expected shortfall. The expected shortfall is the expectation of the loss distribution in the case with probability α. Hence, the expected shortfall measures the expected loss one has for losses above $\text{VaR}_\alpha(L)$. In other words, the expected shortfall is the average value of risk for probabilities larger than α, or:

$$ES = \frac{1}{1-\alpha}\int_\alpha^1 \text{VaR}_x(L)dx.$$

Besides the expected shortfall, we will investigate a few other factors as well in order to evaluate the strategy (Lopez de Prado, 2018, Sections 14.3 and 14.4):

- **Budget:** The initial amount of money that we use to invest in the 1-4 month futures contracts.
- **Frequency of bets:** This is the amount of futures contracts we bought of a certain type. For example, we can look at the frequency for futures contracts with long and short positions.
- **Ratio of longs:** This is the proportion of all traded futures contracts that consist of long positions. In strategies where the both long and short positions are taken, we would expect a ratio of 0.5. When this is not the case, then either the backtest period is too short, or there the strategy has

some sort of bias towards taking a certain position. When this bias occurs, we should be able to explain why this is happening.

- **Ratio of expired contracts:** This is the proportion of all traded futures contracts that have expired during the strategy. When an expiry of a futures contract happens, it means that the market price of the futures contract was lying outside the credible interval of the posterior distribution up until expiry. This means that either on the last trading day the futures contract was underpriced or overpriced, or this means that the model incorrectly stated that the futures contract was underpriced or overpriced, while this was actually not the case.

- **Holding period:** The holding period refers to the amount of days that we hold a position in a certain futures contract.

- **Maximum drawdown:** The maximum drawdown is the maximum loss that occurs during the backtest period.

- **PnL (profit and loss)**: The total amount of money that the strategy returns over the entire backtest period. We can investigate whether certain futures contracts or positions generate more income than others.

5 Results

In this section we will go over the results of the simulations of both our models and the investment strategy. First in Section 5.1 we will discuss the simulation results of the term structure model. We will analyse the performance of the Metropolis-Hastings algorithm that we set up in Section 4.2.3 alongside with a No-U-Turn sampler. Afterwards in Section 5.2, we will discuss the results of the error correction model, for which we will need the simulation results of the term structure model. For this model, we simulated the Gibbs sampling scheme that we have set up in Section 4.3.2 alongside with a No-U-Turn sampler. Finally, we will look at the results of the investment strategy in Section 5.3 and see if our models are actually able to detect underpriced or overpriced futures contracts.

For all of the results in this section, we will use a data set of the closing prices of the VIX index between the first trading day of 2011 (January 3rd, 2011) and the last trading day of 2020 (December 31st, 2020), consisting of a total of 2518 trading days. Furthermore, we have data of the closing prices of the 1-7 month VIX futures contracts on this time interval. The VIX index behaved as followed during the time interval of our data set:

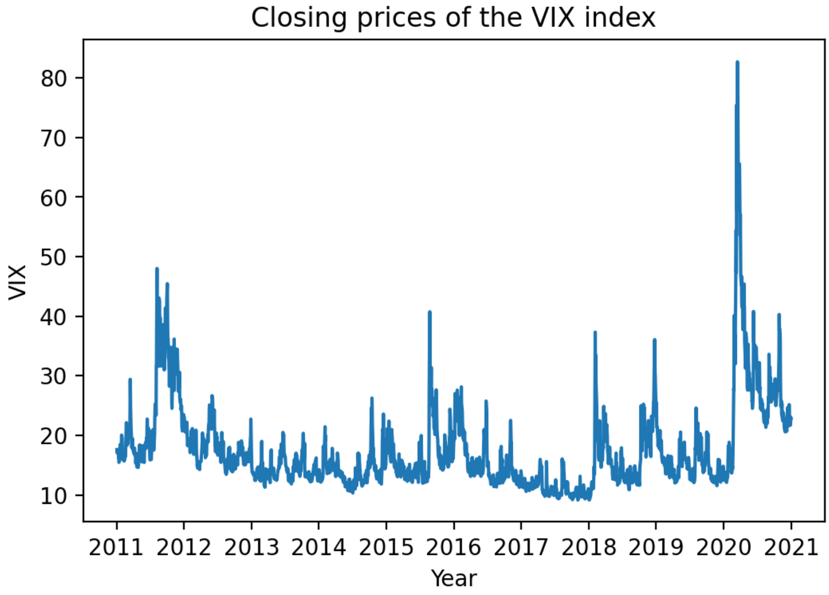


Figure 5.1: The VIX index between 2011-2021.

Note that in Figure 5.1 we see that the impact of the outbreak of COVID-19 is clearly present in the data set. This abnormal behaviour of the VIX index may become more challenging when it comes to modelling the term structure or implementing an investment strategy. Another interesting aspect to look at is the behaviour of the futures price of short term futures contracts in relation to long term futures contracts:

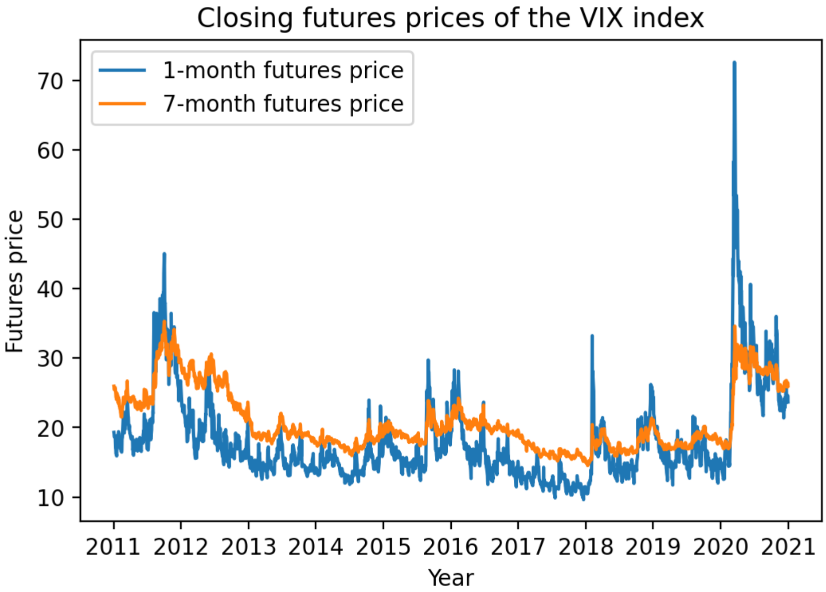


Figure 5.2: Closing futures prices of short term and long term VIX futures contracts between 2011-2021.

In Figure 5.2 we can see the effect of the mean-reverting behaviour of the VIX index on the futures price. The prices of the long term futures contracts seem to be more stable than the short term futures contracts. This is because the price of the 1-month futures contract is heavily impacted by the spot VIX index, while the price of the 7-month futures contract will be closer to the long term mean of the VIX index.

5.1 Term structure model simulations

In this section we will go over the results of the proposed Metropolis-Hasting algorithm and the No-U-Turn Sampler of the term structure model. Since we will be estimating our term structure model every day, thus on all 2518 trading days, our goal is to find the most efficient MCMC algorithm for our simulation.

5.1.1 Metropolis-Hastings algorithm

We will use the Metropolis-Hastings algorithm from Algorithm 7 of Section 4.2.3. We will use $M = 50000$ iterations of our Metropolis-Hastings algorithm and then apply a burn-in of 5000 samples. Remember that we are using a symmetric random-walk proposal as a normal distribution for (α, β) as:

$$q(z \mid (\alpha_{m-1}, \beta_{m-1})) \sim \mathcal{N}((\alpha_{m-1}, \beta_{m-1}), \sigma_q^2 I),$$

for which the value σ_q^2 needs to be fine-tuned in such a way that our algorithm becomes more efficient.

Suppose we want to model the term structure on 24th of April, 2011. On this day, the market is in contango and has the following term structure:

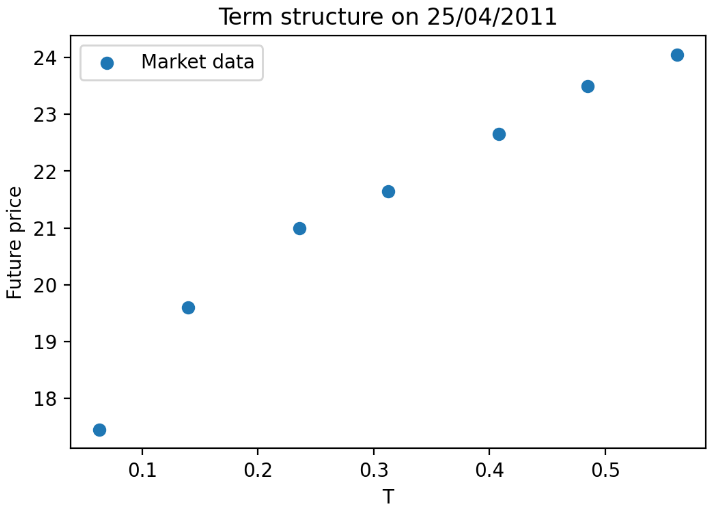
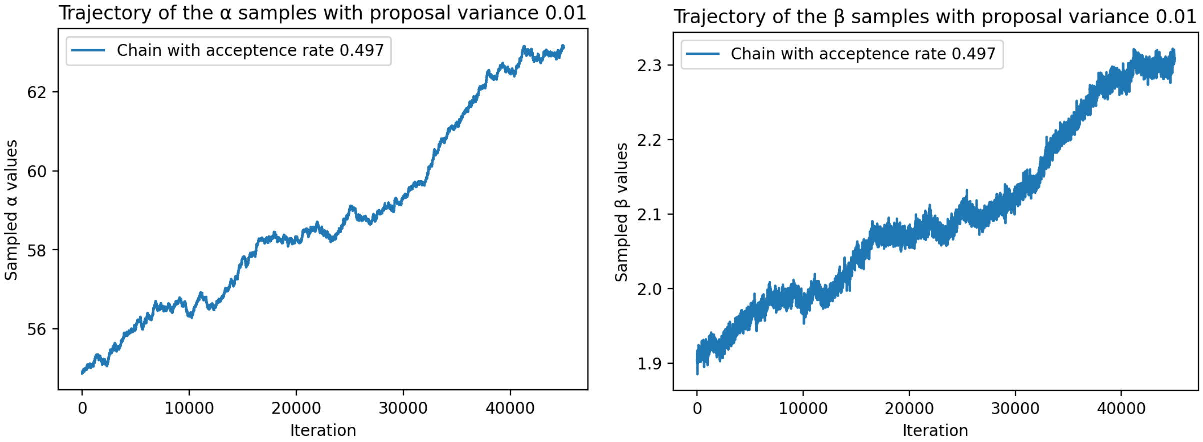


Figure 5.3: Term structure of VIX futures on 24th of April, 2011.

If we use $\sigma_q^2 = 0.01$, then we obtain the following simulation results for α and β after 50000 iterations:

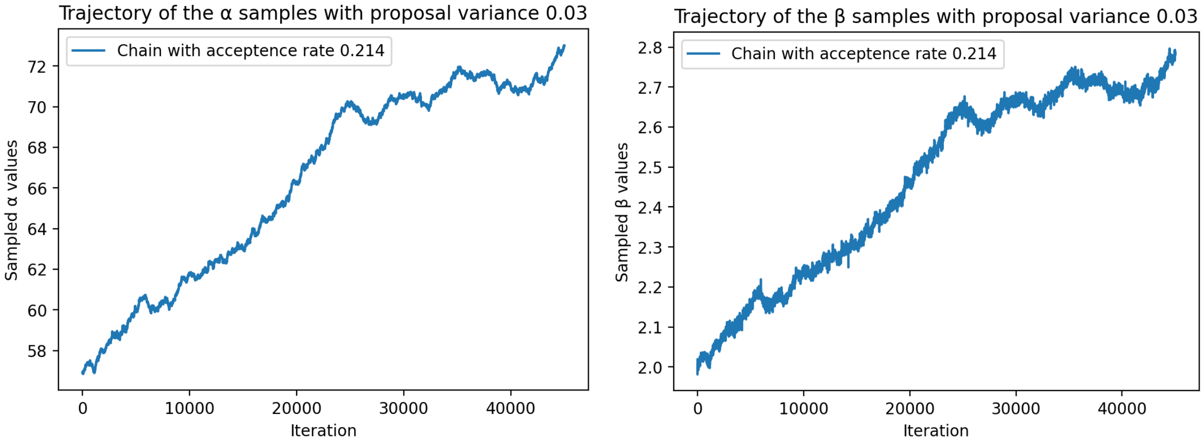


(a) Simulated α values.

(b) Simulated β values.

Figure 5.4: Simulation of our parameters using Metropolis-Hastings algorithm with proposal variance $\sigma_q^2 = 0.01$.

The problem of our results in Figure 5.4 is that we did not even reach any sort of convergence of our posterior distributions after 50000 iterations, although we already reject over 50% of our proposed samples. We can try to solve this problem by increasing the variance of the proposal distribution, since this result in large step sizes for the proposed values from the random-walk proposal. However, this will result in an even lower acceptance rate. A lower acceptance rate does not necessarily imply that the simulation is inefficient, since it may help the parameters move more efficient throughout the parameter space. However, the acceptance rate should not become too low, since we will then waste lots of computation time on rejecting proposed samples. When we now increase the variance of the proposal distribution, we obtain the trajectories of the simulations for α and β using $\sigma_q^2 = 0.03$:

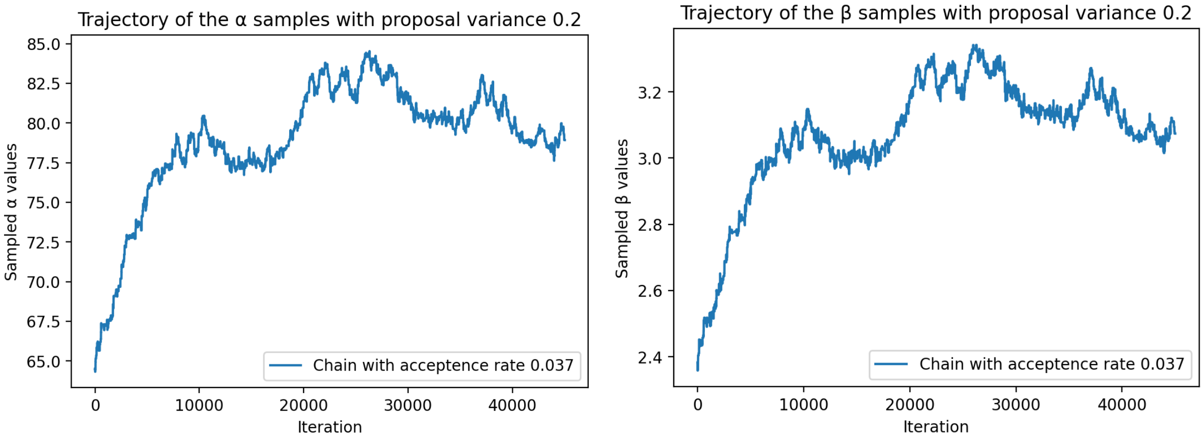


(a) Simulated α values.

(b) Simulated β values.

Figure 5.5: Simulation of our parameters using Metropolis-Hastings algorithm with proposal variance $\sigma_q^2 = 0.03$.

The results from Figure 5.12 show that we still have trouble obtaining convergence with a proposal variance of $\sigma_q^2 = 0.03$. We do see that the parameters are moving faster throughout the parameter space, however, apparently we still need to adjust the value of the proposal variance. Even after 50000 iterations we have no sign of convergence. We will need to increase the proposal variance even more, while we keep decreasing the acceptance probability. When we now again increase the variance of the proposal distribution, we obtain the trajectories of the simulations for α and β using $\sigma_q^2 = 0.2$:



(a) Simulated α values.

(b) Simulated β values.

Figure 5.6: Simulation of our parameters using Metropolis-Hastings algorithm with proposal variance $\sigma_q^2 = 0.2$.

Here we finally see some sort of convergence behaviour of our posterior distributions. However, note that the acceptance rate of the algorithm is 0.037 which would requires us to reject a lot of proposed values during the simulation. Figure 5.6 does recommend us to increase the burn-in of the sample by another 5000, hence we would have 10000 burn-in samples. The reason that our random-walk behaviour is acting this way has to do with the correlation between the parameters α and β. If we look in the figures below, we clearly see dependence in the posterior distributions:

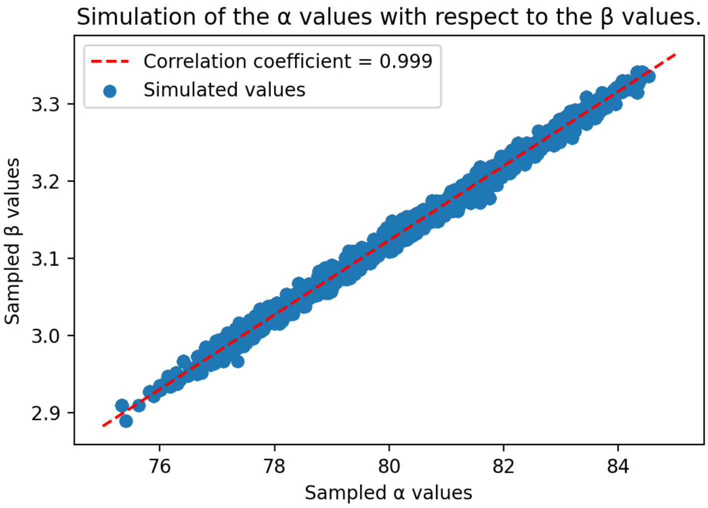
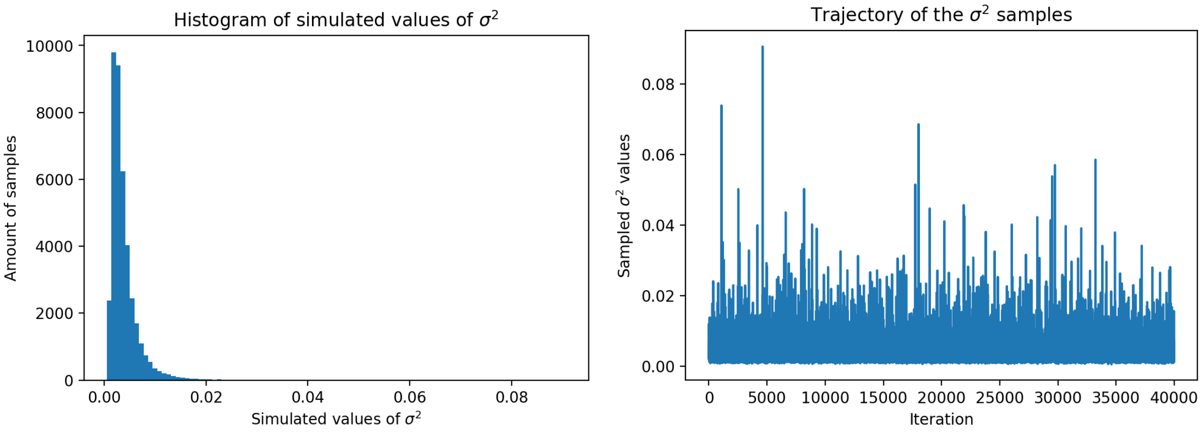


Figure 5.7: Simulation of the α values with respect to the simulation of the β values.

This high correlation between α and β is the result of the influence of the long term mean $\frac{\alpha}{\beta}$ on the theoretical futures price. The random-walk proposed value will most likely only be accepted when both proposed values of α and β move in the same direction, or equivalently when the long term mean $\frac{\alpha}{\beta}$ approximately stays the same. We also simulated the σ^2 parameter, which results in the following simulated values:



(a) Histogram of simulated σ^2 values. Note that the histogram includes values of σ^2 up to approximately 0.1, which can be seen more clear in the figure on the right.

(b) Trajectory of σ^2 values.

Figure 5.8: Simulation results of our simulated σ^2 parameter.

We can summarize the Pearson correlation coefficients between the simulated values of the parameters as:

Table 2: Pearson correlation coefficients between the simulated values. The coefficients with * are significant non-zero with significance level 0.05.

Corr$(\cdot,\cdot)$	α	β	σ^2
α	1*	0.999*	−0.0378*
β	0.999*	1*	−0.0380*
σ^2	−0.0378*	−0.0380*	1*

We can summarize the posterior distributions of the parameters as:

Table 3: Posterior mean and 95% credible region of α, β and σ^2.

	Posterior mean	**95% Credible region**
α	80.53	(76.13, 83.98)
β	3.148	(2.987, 3.325)
σ^2	0.003898	(0.001216, 0.01177)

Finally, we end up with the posterior for F:

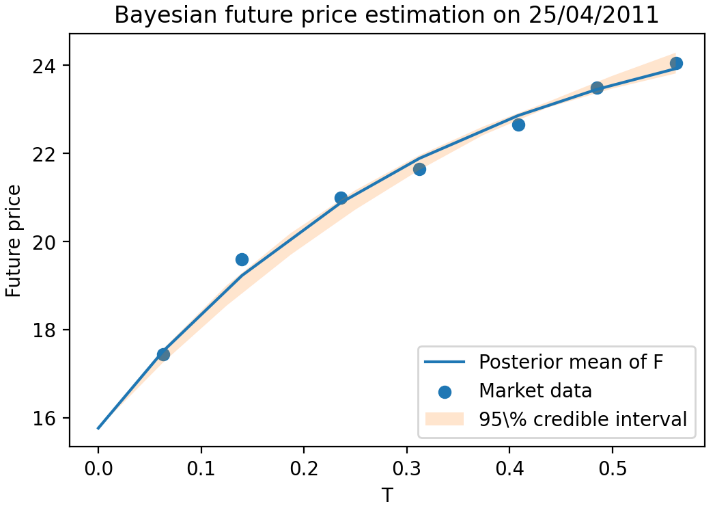
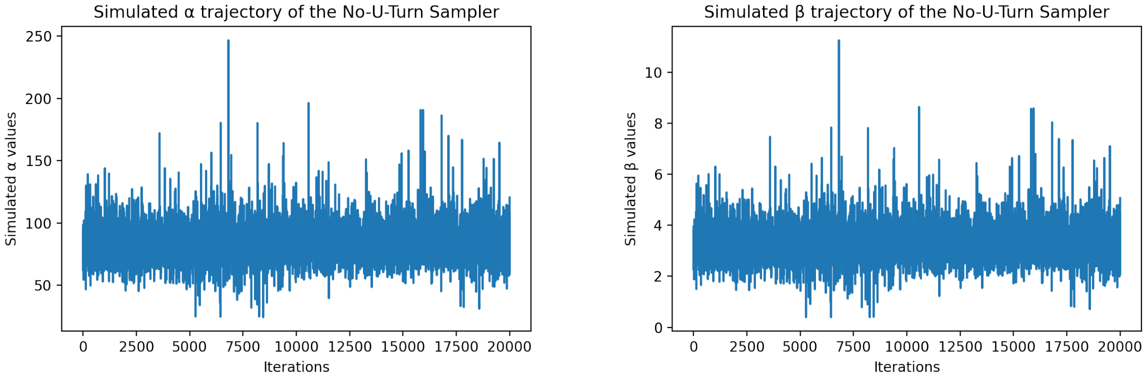


Figure 5.9: Posterior of F after the Metropolis-Hastings simulation.

Figure 5.9 shows very narrow credible intervals of our futures prices, which could be a sign that the parameter space has not been fully explored yet by the simulation.

5.1.2 No-U-Turn Sampler

In this section we will use the No-U-Turn sampler to sample from our posterior distribution of the term structure model. We will use the implementation of the No-U-Turn Sampler from the software Stan. This software allows us to simultaneously run four different Markov chains on our four CPU cores, with a length 10000 samples for each chain. We will regard the first 5000 samples of each chain as burn-in. After this simulation, we obtain the following simulated values for our parameters:



(a) Simulated α values.

(b) Simulated β values.

Figure 5.10: Simulation of our parameters using the No-U-Turn Sampler.

The No-U-Turn Sampler seems to have no problems with sampling from the correlated posteriors of α and β. We also clearly see in Figure 5.10 that more of the parameter space gets explored than in Figure 5.6. One may notice that the behaviour of both parameters seem to be similar, which can again be explained by the dependence of the posterior distributions of α and β. An illustration of the correlation of the sampled parameter values using the No-U-Turn Sampler can be seen in the figure below:

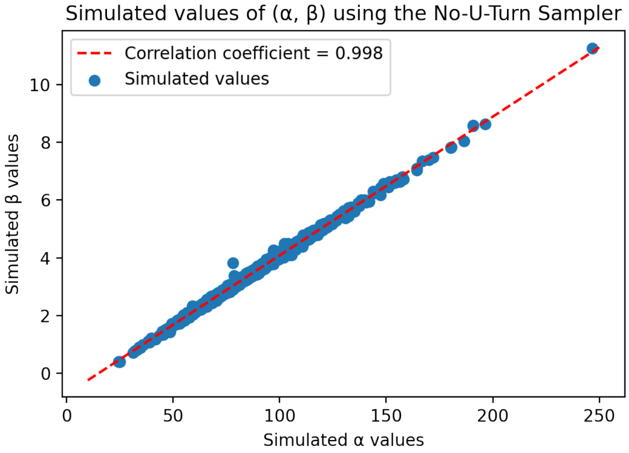
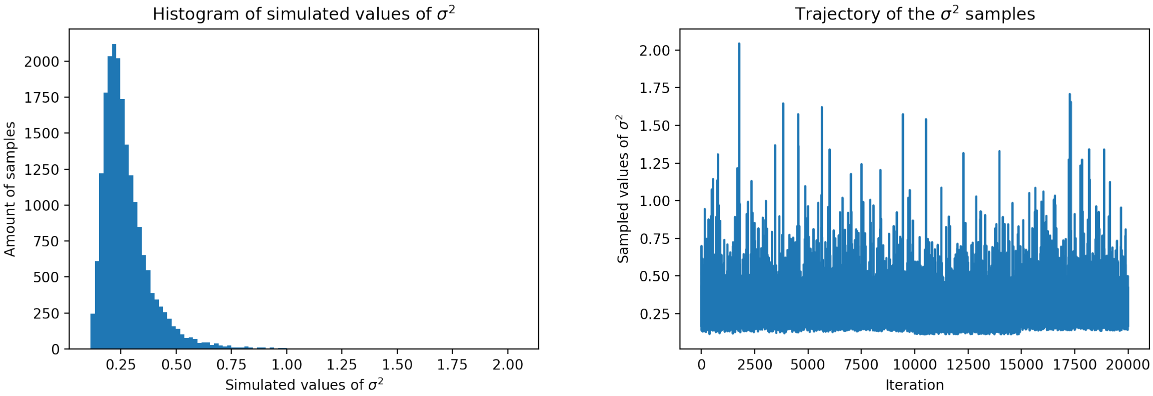


Figure 5.11: Simulation of the α values with respect to the simulation of the β values.

Comparing these results with the Metropolis-Hastings algorithm in Figure 5.7, we clearly see that the parameter space gets further explored, while we use only half of the amount of iterations. We have also simulated the σ^2 parameter, resulting in:



(a) Histogram of simulated σ^2 values. Note that the histogram includes values of σ^2 up to approximately 2, which can be seen more clear in the figure on the right.

(b) Trajectory of σ^2 values.

Figure 5.12: Simulation results of our simulated σ^2 parameter.

We can summarize the Pearson correlation coefficients between the simulated values of the parameters as:

Table 4: Pearson correlation coefficients between the simulated values. The coefficients with * are significant non-zero with significance level 0.05.

Corr$(\cdot,\cdot)$	α	β	σ^2
α	1*	0.998*	0.149*
β	0.998*	1*	0.146*
σ^2	0.149*	0.146*	1*

We can summarize the posterior distributions of the parameters as:

Table 5: Posterior mean and 95% credible region of α, β and σ^2.

	Posterior mean	**95% Credible region**
α	83.04	(64.48, 104.82)
β	3.27	(2.371, 4.318)
σ^2	0.276	(0.142, 0.572)

Finally, the resulting posterior distribution for F after the simulation is:

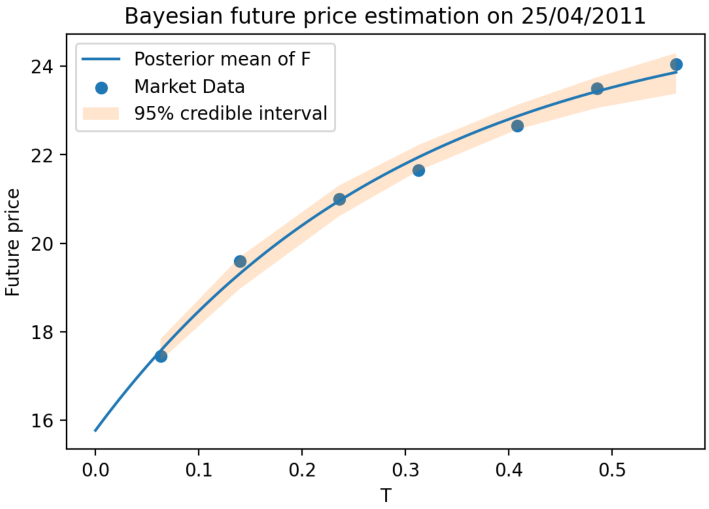
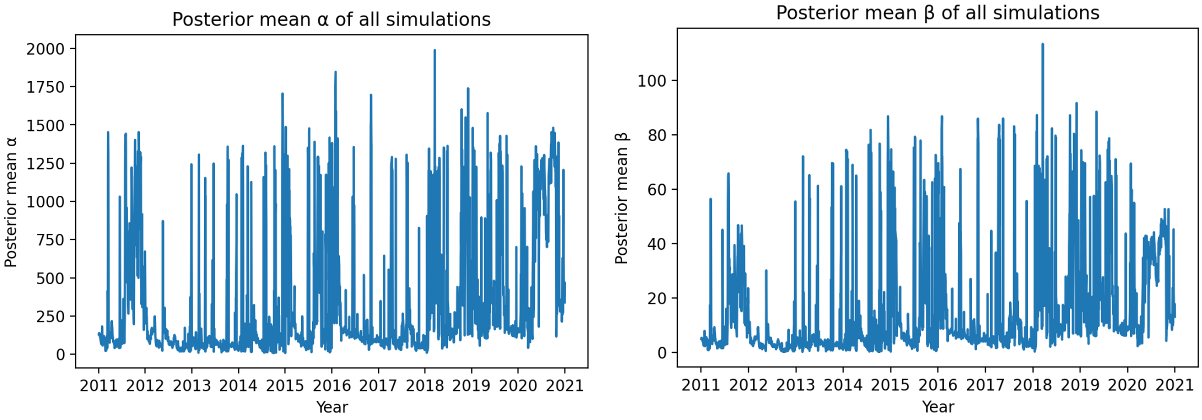


Figure 5.13: Posterior distribution of F after the simulation of the No-U-Turn Sampler.

We clearly see wider credible intervals in Figure 5.13 than the Metropolis-Hastings simulation in Figure 5.9, since we are exploring the parameter space deeper.

5.1.3 Simulation results of the data set

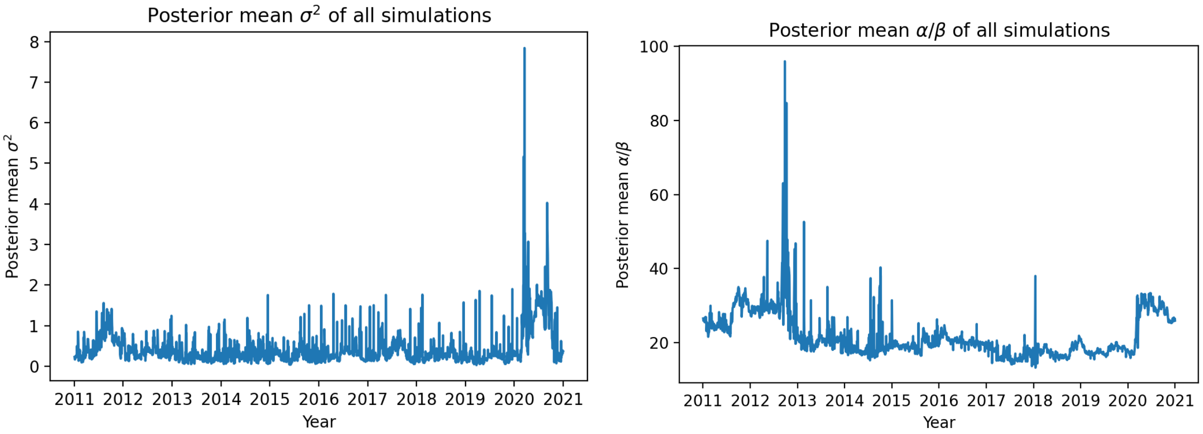
We will use the No-U-Turn Sampler for the simulation of term structure models for all 2518 trading days, since the No-U-Turn Sampler clearly outperforms the Metropolis-Hastings algorithm for our model. After the simulation, we obtain the following posterior means for the parameters α, β, σ^2 and the long term mean $\frac{\alpha}{\beta}$:



(a) Simulated posterior means of α.

(b) Simulated posterior means of β.

Figure 5.14: Simulated posterior means of our α and β over the data set using the No-U-Turn Sampler.



(a) Simulated posterior means of σ^2.

(b) Simulated posterior means of $\frac{\alpha}{\beta}$.

Figure 5.15: Simulated posterior means of σ^2 and $\frac{\alpha}{\beta}$ over the data set using the No-U-Turn Sampler.

Although the posterior means of α and β may differ a lot in Figure 5.14, we find that the long term mean fraction $\frac{\alpha}{\beta}$ in Figure 5.15 remains rather stable. High values of β (with corresponding high values of α) imply a high speed of reversion towards the long term mean. To see how well our term structure is fitting to the market prices, we will look at the mean absolute percentage error (MAPE) of our posterior means for each of the 7 contracts:

Table 6: Posterior mean and 95% credible region of α, β and σ^2.

Month	**MAPE (%)**
1	1.962
2	1.316
3	1.087
4	0.868
5	0.768
6	0.912
7	1.469

In the next section, we will compare these statistics with the error correction model. However, one has to keep in mind the the error correction model is not primarily modelling the term structure, but the error of the term structure model. Therefore, minimizing the MAPE is not the primary task of the error correction model.

5.2 Error correction model simulations

In this section we will investigate the results of the Gibbs sampler algorithm and the No-U-Turn Sampler of the error correction model. For this model we will eventually have to simulate our parameters for each of the seven contracts on all 2518 trading days. The efficiency of the MCMC algorithm that we choose is therefore very important. For the results, we will look at the term structure on 8th of October, 2020. On this specific day, we notice a specific future around November of 2020 having a significant higher market price than the other futures. As we discussed in Section 4.2.4, this outlier is caused by the US presidential election of 2020. The term structure looks as follows:

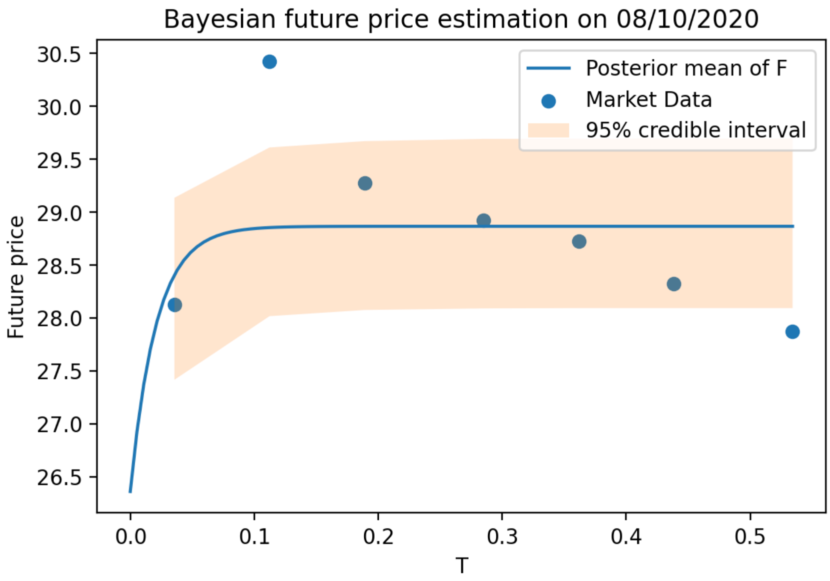
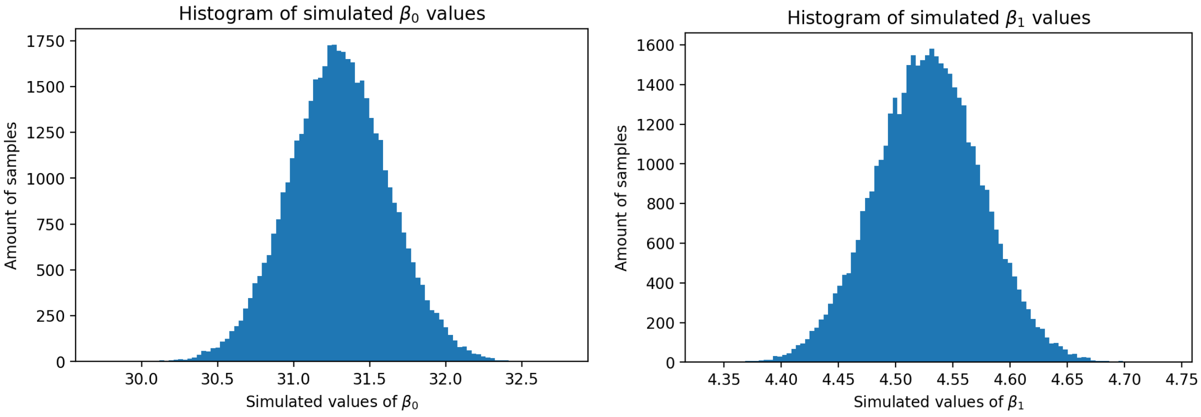


Figure 5.16: The term structure on 8th of October, 2020. We model the term structure with the term structure model and showcase the 95% credible intervals.

The term structure model will tell us that the 2-month futures contract is overvalued, simply because it has a much higher value than the other futures contracts. However, since an event is occurring near the expiration date, this futures contract may in reality be correctly priced. The error correction model will model the error of the term structure model of this specific futures contract with certain regression factors. After the simulation of both models, we will see whether the term structure model can actually identify the higher value of the 2-month futures contract.

5.2.1 Gibbs Sampler

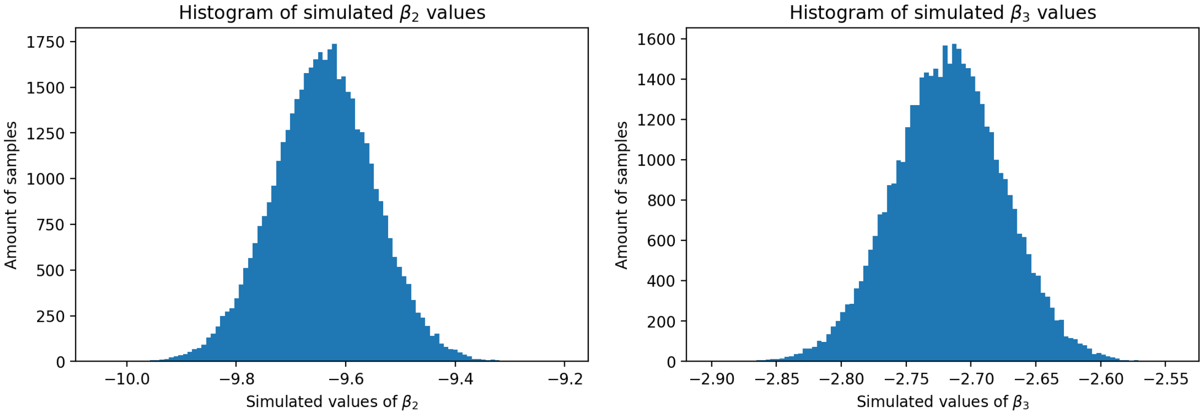
We will use the Gibbs sampler algorithm that we described in Algorithm 8 in Section 4.3.2. Again, we will use $M = 50000$ iterations and a burn-in of 5000 samples. Unlike the random-walk Metropolis-Hastings algorithm from the previous section, our Gibbs sampler algorithm has no problems with convergence to the posterior distribution in the first 5000 samples. This has to do with the fact that we can directly sample our parameters from the posterior distribution, while the Metropolis-Hastings algorithm requires an accepted sample from a proposal distribution. When we run the Gibbs sampler algorithm, we obtain the following sampled values for $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ and σ^2 for the 2-month futures contract:



(a) Histogram of simulated β_0 values

(b) Histogram of simulated β_1 values

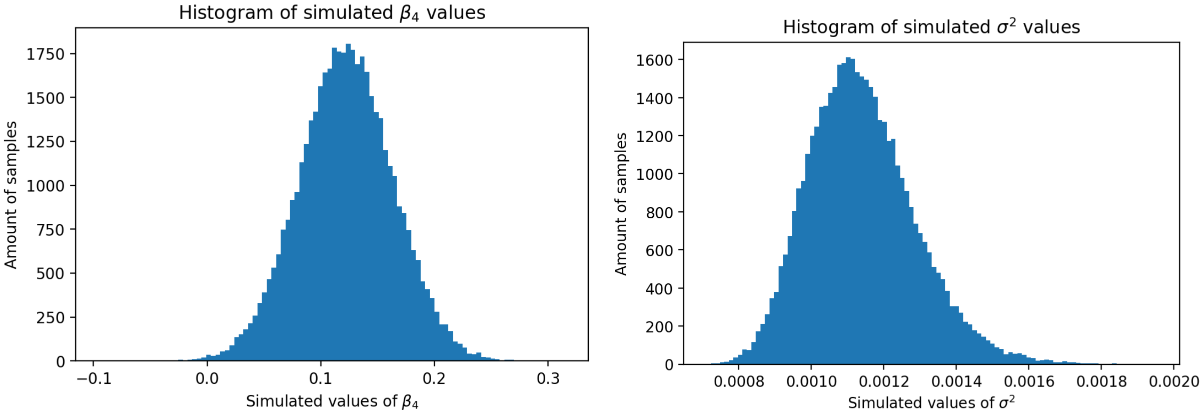
Figure 5.17: Histograms of the simulated values for β_0 and β_1.



(a) Histogram of simulated β_2 values

(b) Histogram of simulated β_3 values

Figure 5.18: Histograms of the simulated values for β_2 and β_3.



(a) Histogram of simulated β_4 values

(b) Histogram of simulated σ^2 values

Figure 5.19: Histograms of the simulated values for β_4 and σ^2.

We can summarize the Pearson correlation coefficients between the simulated values of the parameters as:

Table 7: Pearson correlation coefficients between the simulated values. The coefficients with * are significant non-zero with significance level 0.05.

Corr$(\cdot,\cdot)$	β_0	β_1	β_2	β_3	β_4	σ^2
β_0	1*	-0.691*	-0.999*	-0.676*	0.128*	0.00156
β_1	-0.691*	1*	0.6586*	0.876*	0.0751*	0.00767
β_2	-0.999*	0.6586*	1*	0.6477*	-0.138*	-0.00208
β_3	0.676*	0.876*	0.6477*	1*	0.106*	0.00417
β_4	0.128*	0.0751*	-0.138*	0.106*	1*	-0.00101
σ^2	0.00156	0.00767	-0.00208	0.00417	-0.00101	1*

The coefficients all seem to have some sort of significant correlation between each other, although not with the variance parameter σ^2. We can summarize the posterior distributions of the parameters as:

Table 8: Posterior mean and 95% credible region of β_0, β_1, β_2, β_3, β_4 and σ^2.

	Posterior mean	**95% Credible region**
β_0	31.30	(30.66, 31.95)
β_1	4.529	(4.437, 4.620)
β_2	-9.640	(-9.823, -9.460)
β_3	-2.717	(-2.801, -2.634)
β_4	0.122	(0.0413, 0.204)
σ^2	0.00114	(0.00089, 0.00146)

Finally, after using the Gibbs sampler algorithm for all seven contracts, the resulting posterior distribution for F after the simulation is:

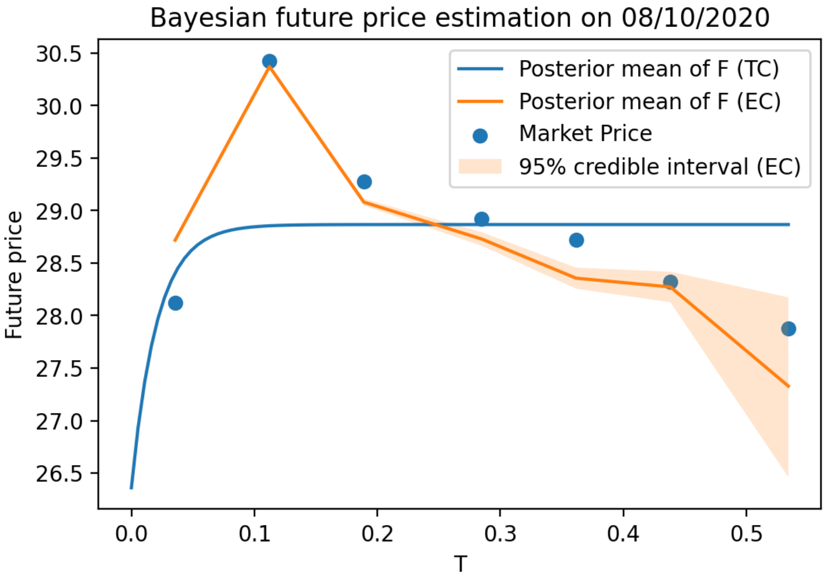
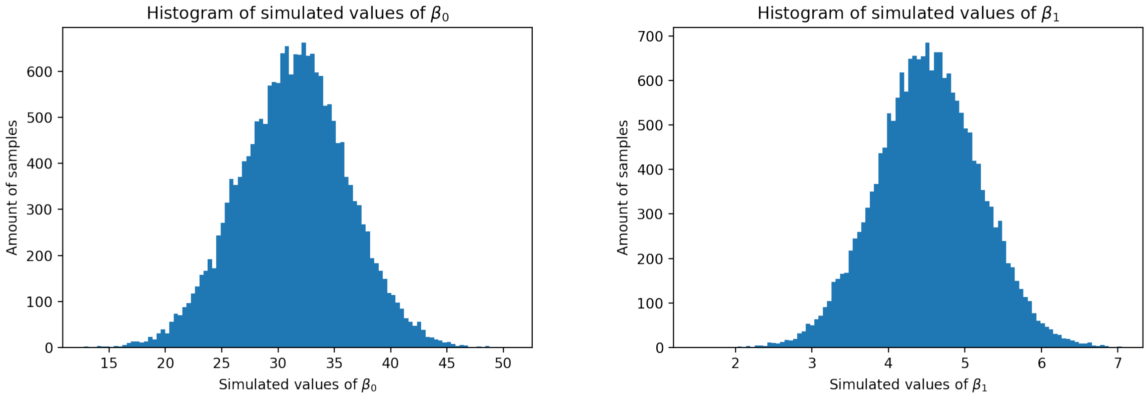


Figure 5.20: The term structure on 8th of October, 2020. Both the term structure model and the error correction model are presented, with the 95% credible intervals of the error correction model.

In Figure 5.20 we clearly see that the error correction model is able to already give a better posterior mean of F for the term structure than the term structure model. The error correction model seems to detect that the 2-month futures contract is worth more than the other contracts. However, the credible intervals of the simulation seem to be very narrow. This implies that the Gibbs sampler is not exploring the parameter space deep enough. The credible interval of the 7-month futures contract does seem to be wider, however this is a result of the little amount of data we have available of the 7-month futures contract, since it has only been on the market for a few days. Trying to fit a model with 5 regression parameters on a small data set may result in overfitting and hence inaccurate results of the 7-month futures contract.

5.2.2 No-U-Turn Sampler

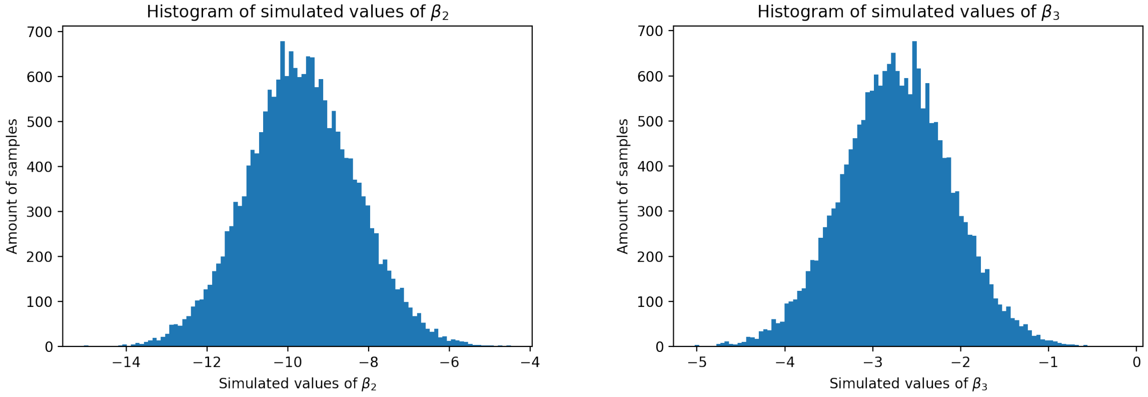
In the similar setting as the previous section, we will apply a No-U-Turn Sampler on the term structure of 8th of October, 2020. We will again analyse the posterior distributions of the parameters for the 2-month futures contract. We will again use the Stan software to run four different Markov chains simultaneously on our four CPU cores, with a length of 10000 for each chain and a burn-in period of 5000, such that we eventually end up with 20000 samples of our parameters. After the simulation, we obtain the following sampled values for $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ and σ^2 for the 2-month futures contract:



(a) Histogram of simulated β_0 values

(b) Histogram of simulated β_1 values

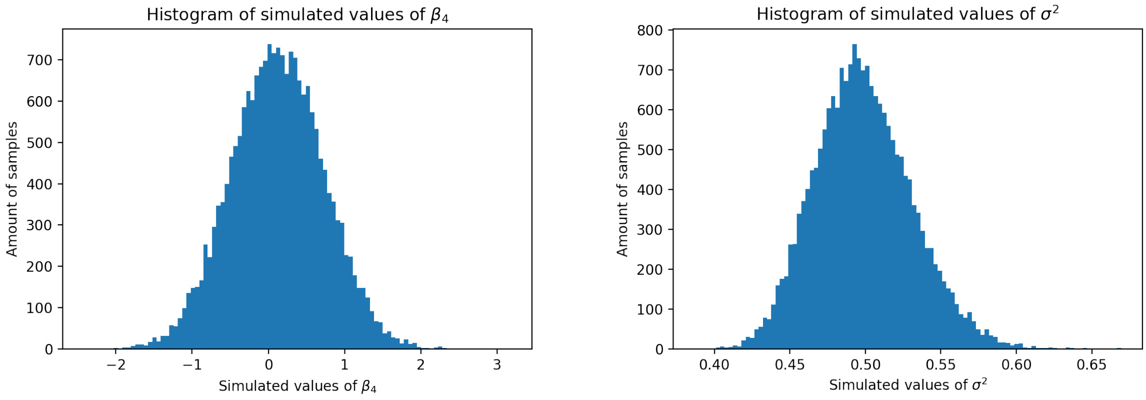
Figure 5.21: Histograms of the simulated values for β_0 and β_1.



(a) Histogram of simulated β_2 values

(b) Histogram of simulated β_3 values

Figure 5.22: Histograms of the simulated values for β_2 and β_3.



(a) Histogram of simulated β_4 values

(b) Histogram of simulated σ^2 values

Figure 5.23: Histograms of the simulated values for β_4 and σ^2.

We can summarize the Pearson correlation coefficients between the simulated values of the parameters as:

Table 9: Pearson correlation coefficients between the simulated values. The coefficients with * are significant non-zero with significance level 0.05.

Corr$(\cdot,\cdot)$	β_0	β_1	β_2	β_3	β_4	σ^2
β_0	1*	−0.677*	−0.999*	−0.670*	0.151*	−0.0066
β_1	−0.677*	1*	0.643*	0.872*	0.0606*	0.0108
β_2	−0.999*	0.643*	1*	0.6410*	−0.161*	0.00634
β_3	−0.670*	0.872*	0.6410*	1*	0.0798*	0.00508
β_4	0.151*	0.0606*	−0.161*	0.0798*	1*	−0.00478
σ^2	−0.0066	0.0108	0.00634	0.00508	−0.00478	1*

We can summarize the posterior distributions of the parameters as:

Table 10: Posterior mean and 95% credible region of β_0, β_1, β_2, β_3, β_4 and σ^2.

	Posterior mean	**95% Credible region**
β_0	31.36	(21.82, 40.72)
β_1	4.521	(3.191, 5.855)
β_2	-9.658	(-12.29, -6.983)
β_3	-2.723	(-3.948, -1.496)
β_4	0.1223	(-1.056, 1.308)
σ^2	0.4988	(0.4408, 0.567)

In general, we see that the posterior means of the parameters in Table 10 are very similar to the posterior means of the Gibbs sampler simulation in Table 8. However, the credible intervals of the posteriors distributions seem to be way larger for the No-U-Turn Sampler, which implies a deeper exploration of the parameter space with this method. Finally, after using the No-U-Turn Sampler for all seven contracts, the resulting posterior distribution for F after the simulation is:

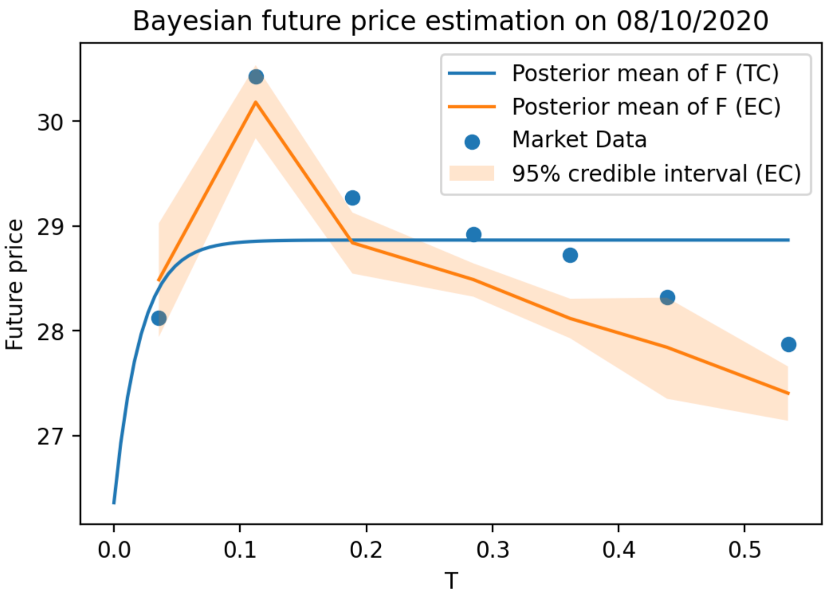
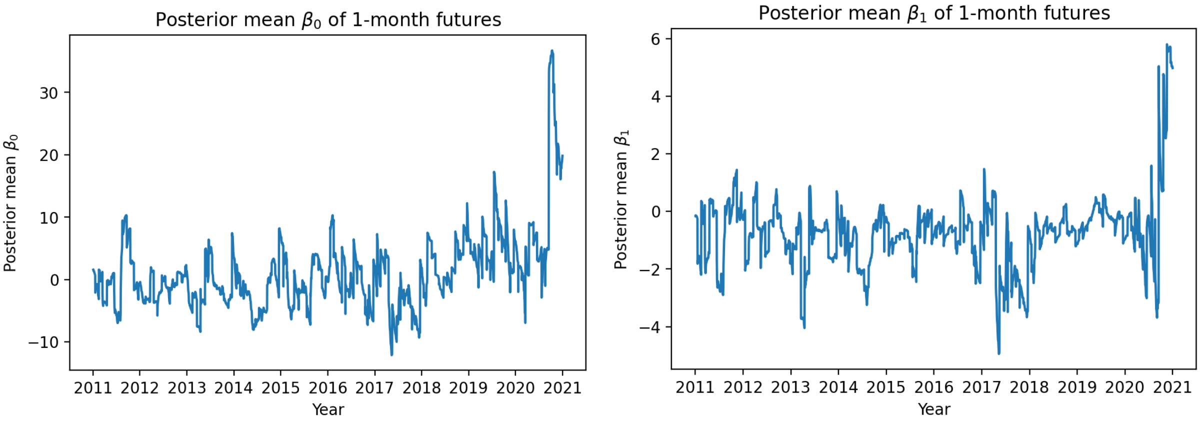


Figure 5.24: The term structure on 8th of October, 2020. Both the term structure model and the error correction model are presented, with the 95% credible intervals of the error correction model.

We see in Figure 5.24 that we clearly have larger credible intervals than when we used Gibbs sampler in Figure 5.20. However, we still see that 4 out of 7 futures contracts lie above the credible intervals. This can either imply that all these futures contracts are overpriced, or this means that the error correction model or the simulation is inaccurate. One way to find out if these contracts were actually overpriced, is by using the error correction model for the investment strategy, which we will investigate in Section 5.3.

5.2.3 Simulation results of the data set

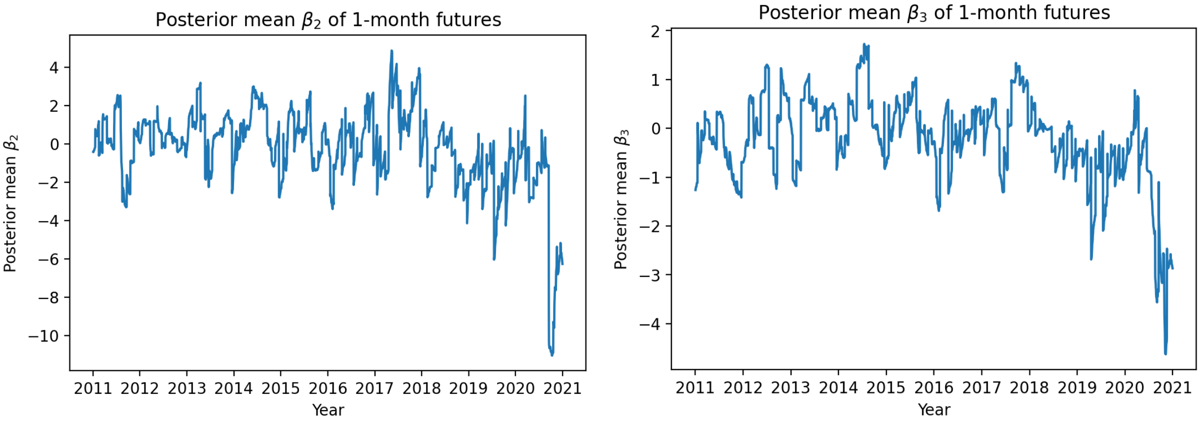
For the simulation of the error correction model on the entire data set, we will again use the No-U-Turn Sampler, since we saw that this sampler is able to search deeper in the parameter space than the Gibbs sampler. Since we will be modelling seven futures contracts on 2518 trading days, we will be simulating $7 \times 2518 = 17626$ futures contracts. We will therefore only be using only 2500 iterations in our four separate Markov chains in the Stan software with a burn-in period of 1250 samples, which results in 5000 samples of our parameters. Even with only 5000 samples, the simulation of all our futures contracts took approximately three days to complete. After the simulation, we obtain the following posterior means of our parameters for the 1-month futures contracts:



(a) Simulated posterior means of β_0.

(b) Simulated posterior means of β_1.

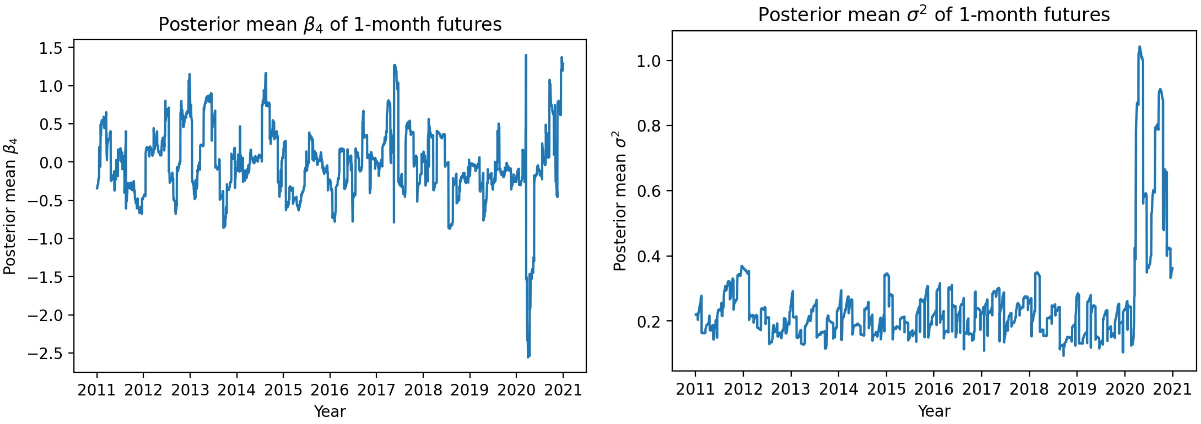
Figure 5.25: Simulated posterior means of our β_0 and β_1 over the data set of 1-month futures contracts using the No-U-Turn Sampler.



(a) Simulated posterior means of β_2.

(b) Simulated posterior means of β_3.

Figure 5.26: Simulated posterior means of our β_2 and β_3 over the data set of 1-month futures contracts using the No-U-Turn Sampler.



(a) Simulated posterior means of β_4.

(b) Simulated posterior means of σ^2.

Figure 5.27: Simulated posterior means of our β_4 and σ^2 over the data set of 1-month futures contracts using the No-U-Turn Sampler.

The simulated posterior means for the 2-7 month futures contract look very similar as in the figures above. Although it is not the primary task of the term structure model to fit the market prices as good as possible, we can still see if the error correction model is having a lot of impact on the term structure model. To see how well our corrected term structure is fitting to the market prices, we will again look at the mean absolute percentage errors (MAPE) of our posterior means for each of the 7 contracts:

Table 11: Posterior mean and 95% credible region of α, β and σ^2.

Month	TS MAPE (%)	TS MAPE (%)
1	1.962	2.122
2	1.316	1.235
3	1.087	0.812
4	0.868	0.635
5	0.768	0.645
6	0.912	0.796
7	1.469	14.92

In general, the error correction model seems to correct the term structure model towards the market prices. What we notice is that the MAPE of the 7-month futures contracts is suddenly 14.92. This has to do with overfitting of the term structure model of the 7-month contracts, since most of the time we do not have a lot of data of these contracts yet. Trying to fit five regression parameters on such a small data set may result in overfitting and inaccurate results.

5.3 Strategy results

In this section we will perform a backtest of the strategy as described in Section 4.4. Over the course of 2518 trading days, we will for each day check whether the 1-4 month futures contracts are undervalued or overvalued. Before we begin, we will have to make some assumptions in the strategy:

- The initial budget is 10000$ Although we will receive profits and losses from our strategy, we will assume to always have the budget of 10000$ available to invest.
- We will spread the available budget over the 1-4 month futures contracts to invest with, hence 2500$ for each individual futures contract. Although we initially do not need cash to take a short position, we will not take a short position of more than the allocated budget of 2500$.
- We will not take any form of transaction costs into account.

- We will not take the multiplier of VIX futures contracts into account, hence we take a multiplier of 1.
- For simplicity, we will assume that we can buy proportions of futures contracts.
- We can always buy/sell all of the desired futures contracts for their current market ask/bid prices.
- Since we use data of closing prices of VIX futures, we will determine whether a futures contract is undervalued or overvalued at the end of the day. When we see that the market price of the futures contract is in fact outside the credible interval of the posterior distribution, we will take a position on the contract on the opening of the market on the next day.
- When we purchase a futures contract, we use the ask price of the futures contract, while we use the bid price to sell the contract.
- We will assume that the bid-ask spread of the openings prices on the next day is of the same size as the bid-ask spread of the closing prices of the futures contract of today. The reason for this is since we have only data available of the bid-ask spread of the closing prices of the futures contracts. Although we do have data for the opening prices of the futures contracts, we do not have their corresponding bid-ask spreads, therefore we make the assumption that they are of the same size as yesterday's closing bid-ask spread.

The first step will be to use the credible intervals of the models that we generated in the previous sections. Due to long simulation times, we have saved the credible intervals of $50\%, 75\%, 90\%, 95\%$ and 99% for both models. To determine which intervals we will use, we can look at the Sharpe ratio of the backtest for the selection of various credible intervals of both models. When we use a larger credible intervals, the models will require the market price of underpriced or overpriced futures contracts to be further away from the theoretical derived price. This will imply that the strategy with a smaller credible intervals will hold the futures contracts for a longer time than the strategy with larger credible intervals. However, this does not necessarily mean that we make less trades when we use larger credible intervals. Although for larger credible intervals we have a stricter requirement for taking positions, we also have a stricter requirement for selling the position. The problem with small credible intervals might be that we determine a futures contract to be underpriced or overpriced too quickly, while it is actually correctly priced. This could reduce the returns of our model and consecutively reduce the Sharpe ratio. When the credible intervals are too large, it might happen that we invest too little in futures. This will imply that we have a lot of budget that is not invested, hence we could have invested this budget somewhere else and therefore the excess returns and Sharpe ratio will be lower.

In order to calculate the Sharpe ratio of the strategy, we need to subtract the risk-free rate from the strategy returns. A common assumption is to use the yield of a 10-year US treasury bond as the risk-free rate, which evolved between 2011-2021 as:

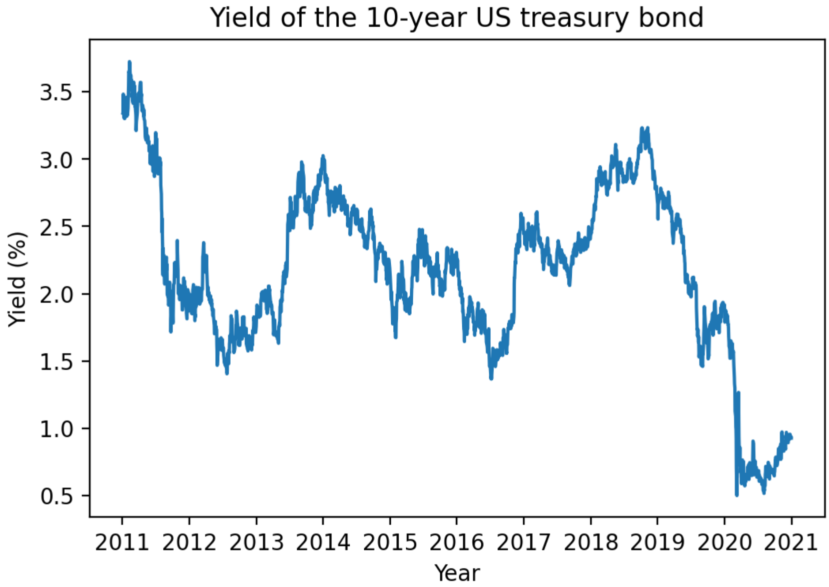


Figure 5.28: Yield of the 10-year US treasury bond between 2011-2020

5.3.1 Credible region selection

For our Sharpe ratio to be positive, we will require the strategy on average to have a higher return than the risk-free rate. Using the yield of the 10-year US treasury bond as the risk-free rate, we can derive the Sharpe ratios for both of our models with various credible regions. With TS as notation for the term structure model and EC as notation for the error correction model, we obtain the Sharpe ratios:

Table 12: Sharpe ratios and PSR(0) values of the term structure model for various credible intervals.

Credible region	**Sharpe ratio of TS model**	**PSR(0) of TS model**
50%	0.005199	0.6031
75%	0.005686	0.6126
90%	0.006802	0.634
95%	0.005878	0.6162
99%	0.003223	0.5643

Table 13: Sharpe ratios and PSR(0) values of the error correction model for various credible intervals

Credible region	**Sharpe ratio of EC model**	**PSR(0) of EC model**
50%	-0.002510	0.4500
75%	-0.002425	0.4517
90%	-0.002425	0.4517
95%	-0.002217	0.4558
99%	-0.002217	0.4558

As we see, the Sharpe ratios of the strategies are all really close to zero, which implies that the strategy is not that efficient. Another thing is that none of the Sharpe ratios are significantly greater than 0, thus we can not say for certain that we will on average have larger returns than the risk-free rate with this strategy. There does seem to be a difference in performance between the models. The term structure model seems to be working better for the investment strategy than the error correction model. For the error correction model, we have seen that we do not include the credible intervals of the term structure model in our model formulation, but only the posterior mean. This may result in losing important information on the posterior mean of the term structure model, that we do not use in the error correction model. Since the error correction model is formulated in such a way that we are not able to use these credible intervals, we will attempt to create a method that combines the intervals of both models. The idea of this method is that we extend the credible intervals of the error correction model by adding the uncertainty of the initial term structure model to them. Since the error correction model models the error $y = F - M$, with F the posterior mean of the term structure model and M the market prices, we end up with $F - y$ as the posterior of the corrected futures prices. Here F is the posterior mean, which is a constant vector. However, the term structure model has created credible intervals for this vector F that we are not including in this model. The idea now is to simply add the credible intervals of F to $F - y$.

Suppose the term structure model tells us that the posterior mean is TS_{mean} and that for some credible region of probability α the futures price lies between $[TS_\alpha^-, TS_\alpha^+]$ and for the error correction model $[EC_\alpha^-, EC_\alpha^+]$. We then assume that by combining these credible intervals, we have that the posterior of the error correction model lies between $[EC_\alpha^- - (TS_{\text{mean}} - TS_\alpha^-), EC_\alpha^+ + (TS_\alpha^+ - TS_{\text{mean}})]$. This way we will use the information of the uncertainty of both models as boundary for underpriced and overpriced futures contracts. The combination of these intervals does remove the mathematical interpretation of the probability of the credible region. However, we can still attempt to test the strategy on these new intervals and see if we can come up with better Sharpe ratios:

Table 14: Sharpe ratios of the error correction model with combined credible intervals.

(Model) (credible region)	**EC** 50%	**EC** 75%	**EC** 90%	**EC** 95%	**EC** 99%
TS 50%	0.006553	0.006675	0.006478	0.006662	0.006723
TS 75%	0.008133	0.007830	0.007708	0.007288	0.007631
TS 90%	0.009904	**0.01052**	0.010236	0.010083	0.009987
TS 95%	0.009939	0.009716	0.009700	0.009731	0.008936
TS 99%	0.008201	0.007858	0.008317	0.008106	0.007450

Again, the Sharpe ratios are very low and none of the Sharpe ratios seem to be significantly greater than 0, since all the PSR(0) values approximately lie between 0.60 and 0.70. We do see that the combined models have better performance than the separate models, such as the combined model of the 90% credible interval of the term structure model and the 75% credible interval of the error correction model with a Sharpe ratio of 0.01052 and PSR(0) value of 0.7020. However, still this Sharpe ratio is far from being statistical significantly better than the Sharpe ratios in Table 12 and Table 13. Although we can not choose any credible intervals with a significant Sharpe ratio, we will for reasons of showcasing the volatility targeting method and the other evaluation methods choose the model with the highest Sharpe ratio, which is the combined model of the 90% credible interval of the term structure model and the 75% credible interval of the error correction model.

The reason that the Sharpe ratios are very low and not significantly greater than 0, is since the strategy is simply not performing that well yet. This can clearly be seen when we look at the PnL of the strategy:

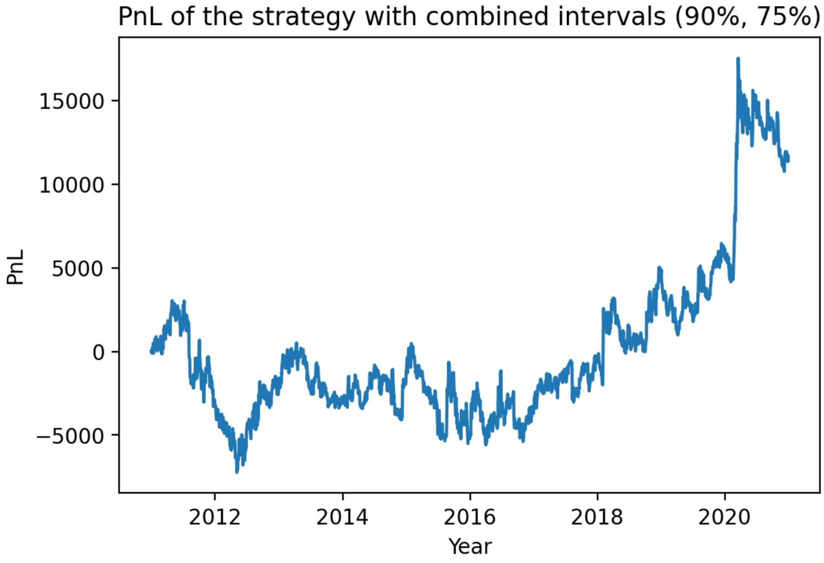


Figure 5.29: PnL of the strategy with combined credible regions of 90% of the term structure model and 75% of the error correction model.

The strategy suffers of a huge loss in the first 2011 and regaining it back in 2012. In the period from 2013-2016 the returns seem to move downwards again, while the strategy from 2017-2020 is performing really well. Around March of 2020, the long positions in the strategy seem to suddenly make a lot of profit. This has to do with the sudden increase in the VIX index as consequence of the COVID-19 outbreak. As we can see, it is definitely not clear what the expected behaviour is of the returns of the strategy. The strategy seems to be very risky and sudden jumps in the VIX index could result in huge losses or profits. We have to keep in mind that our budget is 10000$, hence a loss over 5000$ in one year is already 50% of our total budget. On the other hand, an increase of 10000$ during March of 2020 is a return of 100% on our budget. Therefore, since our strategy returns have such high volatility, it makes sense that the Sharpe ratio is not significantly greater than 0.

5.3.2 Volatility targeting

When we choose the model with the highest Sharpe ratio, which is the combined model with credible regions of 90% of the term structure model and 75% of the error correction model, we can apply the conventional volatility targeting method, where we scale the returns r_t as:

$$r_t^{\text{scaled}} = r_t \times \frac{\sigma^{\text{target}}}{\hat{\sigma}_{t-1}(K)}, \tag{5.1}$$

with σ^{target} the realized volatility of the returns up till day $t-1$ and $\hat{\sigma}_{t-1}(K)$ the realized volatility of the K days before day t. We can investigate the hyperparameter K by tuning this value such that we obtain the highest Sharpe ratio. If the value of K is too low, then the scale of our returns can be very volatile and unpredictable throughout the backtest, since we would base our scale on the volatility of just a few days. When K gets larger, we will reduce the effect of the volatility targeting method, since we will eventually end up with a scale of 1. If we tune K for values of a maximum one trading year (approximately 253 days), we obtain the Sharpe ratios:

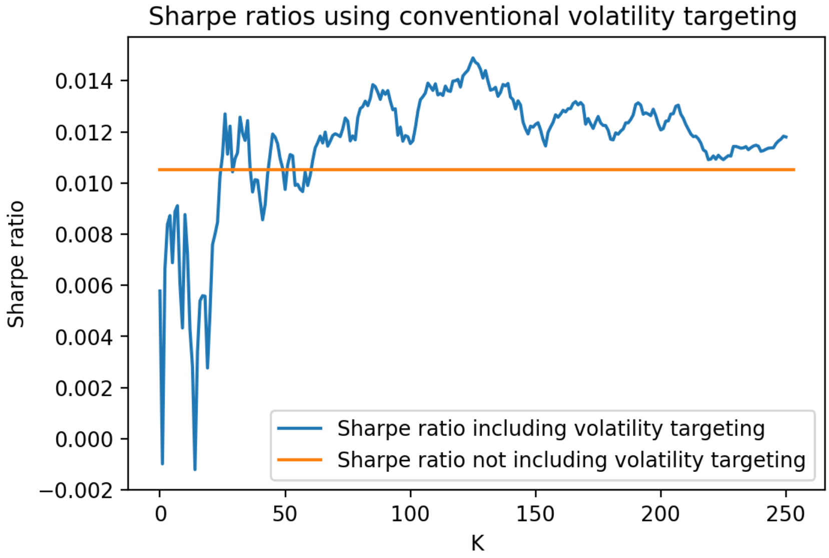


Figure 5.30: Sharpe ratios of the conventional volatility targeting method for various values of K.

We see that volatility targeting is able to increase the performance of our strategy and we see that values between for K between 100 and 150 are appropriate choices for the conventional volatility targeting of our strategy. For $K = 127$ we obtain an optimal Sharpe ratio of 0.01462 with a PSR(0) value of 0.7698, which is still not significantly greater than 0. For $K = 127$, the scale $\frac{\sigma^{\text{target}}}{\hat{\sigma}_{t-1}(K)}$ throughout the strategy is as follows:

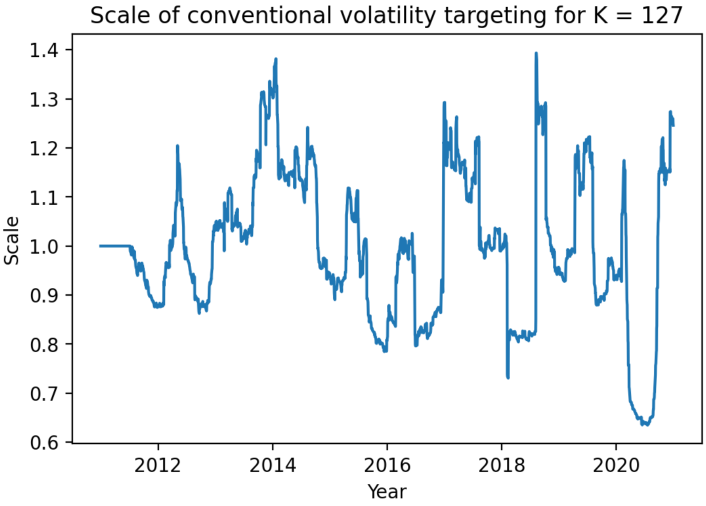


Figure 5.31: The scale of conventional volatility targeting of our strategy using $K = 127$.

Note that in Figure 5.31 we start conventional volatility targeting on the strategy after K days, hence the first 127 days we have a scale of 1. We can see that the strategy significantly scales down the returns around March 2020, which is expected since the strategy returns were very volatile around this period. If we now compare the original strategy with the strategy including volatility targeting volatility, we obtain the following statistics:

Table 15: Comparing the strategy without volatility targeting and the strategy including volatility targeting.

	Original strategy	**Including volatility targeting** ($K = 127$)
Sharpe ratio	0.01052	0.01489
PSR(0)	0.7020	0.7739
Realized volatility	0.03515	0.03367
Maximum drawdown	0.2794	0.2120
VaR$_{0.95}$	0.05168	0.04910
Expected Shortfall	0.07346	0.07132

The increase in Sharpe ratio value together with the increased $PSR(0)$ value imply a better performance of the strategy including volatility targeting. However, this difference in Sharpe ratios is not significant. The realized volatility is also decreased by a little, which implies that conventional volatility targeting results in a more stable strategy. The Maximum drawdown is decreased by a lot, which means that the scaling of the returns reduced the maximum daily loss in the strategy. However, a maximum loss of 21.20% is still a rather large amount to lose in one day. The VaR$_{0.95}$ and expected shortfall are also decreased with volatility targeting, which implies that the loss distribution is less fat tailed. Hence the conventional volatility targeting strategy can be seen as a less risky strategy, although the difference is not that large. Finally, the strategy including volatility targeting results in the following PnL:

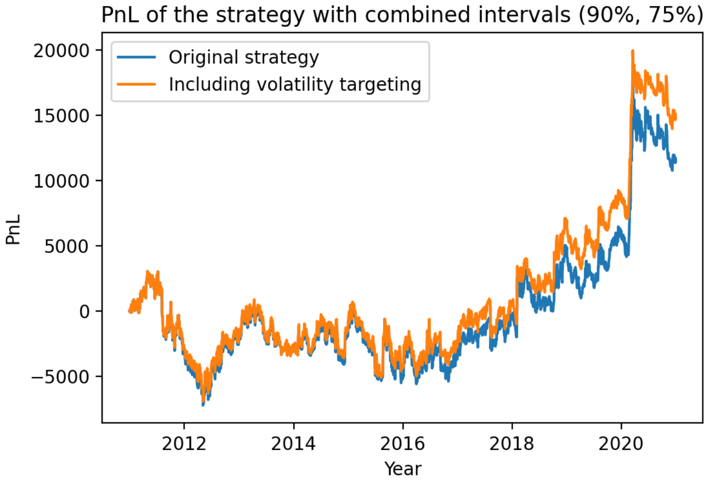


Figure 5.32: PnL comparison of the original strategy and the strategy including volatility targeting.

In general, there is improvement in the strategy when we use volatility targeting, however the improvement is not that significant.

5.3.3 Further evaluation of the strategy

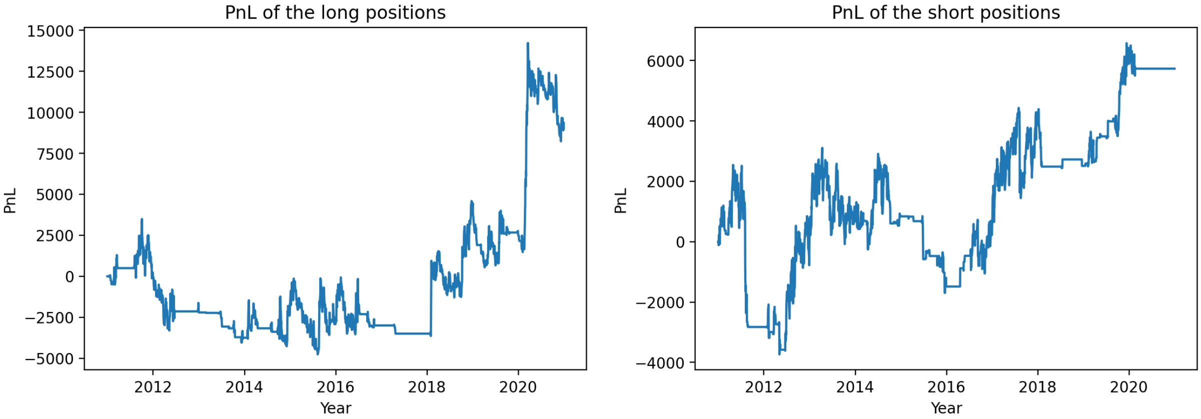
Let us continue evaluating some more factors of the strategy after choosing the $(90\%, 75\%)$ credible intervals and including volatility targeting in the strategy. When we investigate the trades that we made during the backtest period of the strategy, we can see what the behaviour is of certain type of trade that we made. Investigating this behaviour may be important for further research on VIX futures strategies. For example, if it turns out that certain trades are extremely valuable in this strategy, then one could try to create a strategy that revolves around this certain type of trade.

Let us first look at some statistics of the long and short positions in our strategy:

Table 16: Statistics of the long and short positions in our strategy.

	All long positions	**All short positions**
Amount of trades	319	324
Amount of expired contracts	54	66
Average holding period (days)	15.58	14.27
Average return	7.063e-4	5.017e-4
Realised volatility of returns	0.04399	0.03530

with corresponding returns:



(a) PnL of long positions.

(b) PnL of short positions.

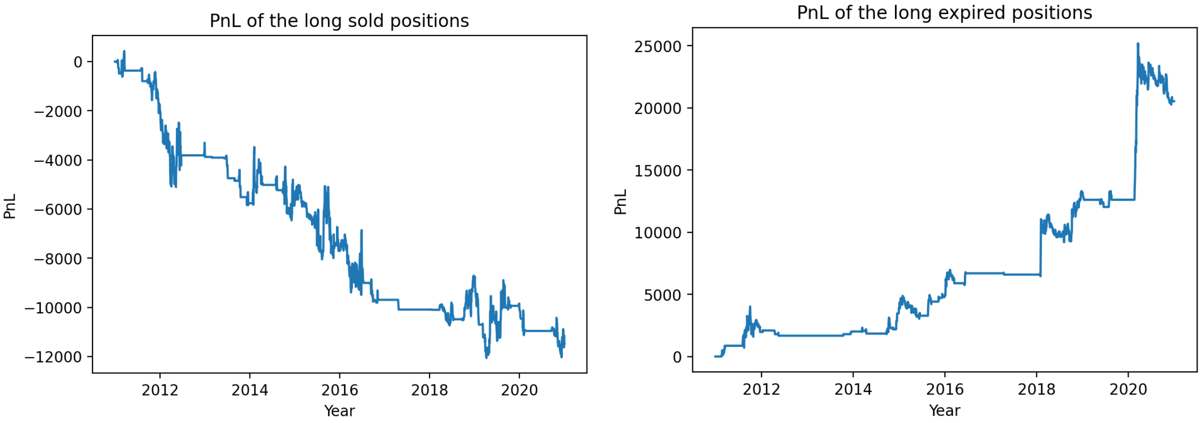
Figure 5.33: Seperated PnL of the long and short positions in the strategy.

In Table 16 we see that the amount of trades of both positions are about the same. The ratio of longs is $\frac{319}{319+324} \approx 0.496$, which implies no bias towards a certain position in our strategy. The average holding period and return seems to be similar as well. The volatility of the returns seems to be a little higher with long positions than short positions. This can be explained with the sudden positive increase of strategy returns around March 2020, which implies we were holding long positions in futures contracts around this time, since the VIX index increased in value as well. It may be argued that the amount of expired contracts is high in the strategy, since we would not expect to have contracts expiring a lot. When the contract expires, it means that either the futures contract is undervalued or overvalued on the last trading day according to the model. This is not necessarily inaccurate, however, when a contract is held for a long time and expires, then the market never bothered to correct the price of the futures contract, hence the model probably incorrectly claimed the futures contract to be undervalued or overvalued. Let us therefore make a distinction between expired contracts and positions that were sold before the expiration date:

Table 17: Statistics of the long sold, long expired, short sold and short expired positions in our strategy.

	Long sold	**Long expired**	**Short sold**	**Short expired**
Amount of trades	265	54	258	66
Average holding period (days)	11.16	37.26	10.64	28.50
Average return	-0.001493	0.004013	-0.002903	0.005294
Realised volatility of returns	0.03611	0.05353	0.02906	0.04211

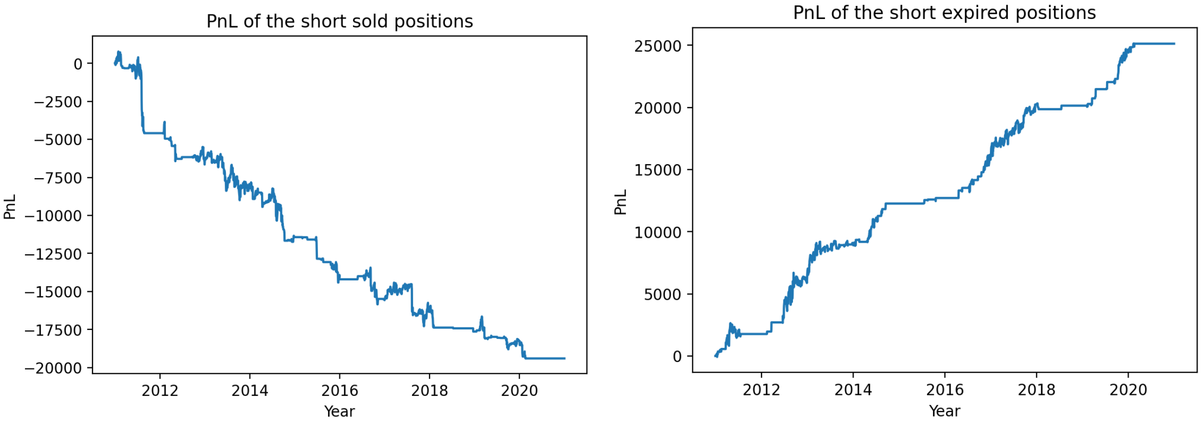
In Table 17 we see interesting results, since apparently the positions on the contracts that expire generate a much higher return than the contracts that get purchased back before expiry. In the PnL figures below, we can clearly see this happening:



(a) PnL of long sold positions.

(b) PnL of long expired positions.

Figure 5.34: PnL of the sold and expired long positions in the strategy.



(a) PnL of short sold positions.

(b) PnL of short expired positions.

Figure 5.35: PnL of the sold and expired short positions in the strategy.

We see in Figure 5.34 and 5.35 that the expired positions of the long positions generate a high return in certain periods, when we don't have a lot of short expired positions. The other way around is exactly the same, more short positions expire when we do not have many expired long positions. We can see that in 2012-2014, 2017 and 2019 we have not many expired long positions, while in these same years we see an increase in expired short positions. Is it possible to explain this behaviour of our strategy? When we look at the high and low volatile regions of the VIX index, we will notice some similarities with the results of our strategy:

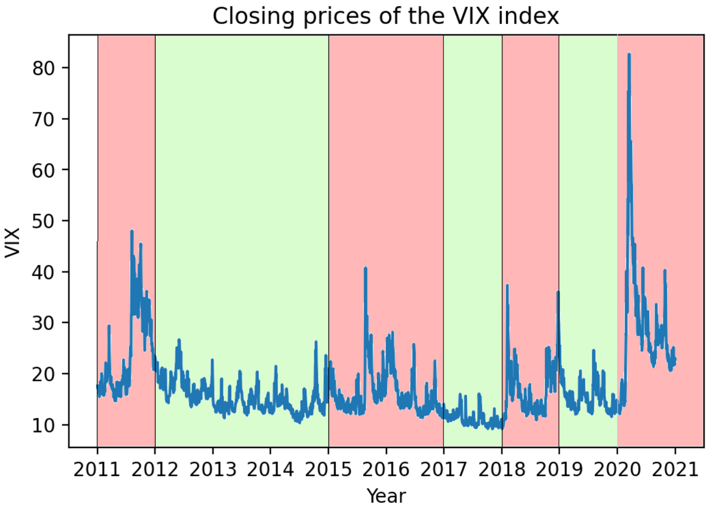


Figure 5.36: VIX index from 2011-2021 with the active period of expired long positions marked as red and the active period of expired short positions as green.

From Figure 5.36 we see that the VIX index during the activity of expired long positions has a high volatility, while during the activity of expired short positions the VIX index has a low volatility. The volatility clustering behaviour of the VIX index is what makes these expired positions are so profitable. If we have a long positions in high volatile regions, the VIX index will have increased values and hence our long positions can suddenly increase in value. In the low volatile regions that follow the high volatile regions, the VIX index is essentially "coolding down". The VIX index slowly decreases over time, which makes the short positions more profitable. Essentially, this tells us that the strategy is making more profits on the long term movement of the VIX index rather then the correction of mispriced futures.

6 Discussion

The goal of the thesis was to create an investment strategy using VIX futures. We have created an investment strategy that takes a long position on undervalued VIX futures and a short position on overvalued VIX futures. We determined whether a futures contract was undervalued or overvalued by creating a model for the term structure of VIX futures. By estimating the parameters of the model with a Bayesian approach, we obtain a credible interval of our futures prices, which we used to determine if a futures contract is currently mispriced. For the modelling of the term structure we created two models: the term structure model and the error correction model.

The term structure model is able to accurately represent the term structure of VIX futures in normal market situations, i.e. contango and normal backwardation market situations. This model is based upon a quite simple theoretical futures price formula with only two parameters. The advantage is that this allows for a financial interpretation of our parameters, since our parameters in the model represent the speed of mean reversion and the long term mean of the term structure. However, we saw that the term structure model can become very inaccurate when we deal with abnormal market situations. Certain events and a non-monotonic term structure were the cause for the term structure model to fail. Since this model is based upon a quite simple theoretical futures price formula with only two parameters, it is unable to accurately fit outliers and non-monotone data.

The error correction model attempts to overcome these shortcomings of the term structure model by modelling the previous errors of the term structure model. The idea is that this model is able to correct the term structure model based upon the behaviour of investors using certain regression factors. In the results we have seen that it was able to accurately identify an event in the term structure and correct the term structure model accordingly. However, the corresponding credible intervals do not always seem to accurately capture the data. The credible intervals are very important, since these intervals decide whether a VIX future is overpriced or underpriced. In the results of the investment strategy, we have seen that we obtain low Sharpe ratios when we only use the credible intervals of the error correction model, which tells us that our credible intervals are not very accurate. A reason may be the fact that we only use the posterior mean of the term structure model, which discards any measure of uncertainty of the term structure model. Due to the model formulation of the error correction model, we are not able to add the uncertainty of the term structure model in this model. However, we do see improvement in the investment strategy when we simply add up the credible intervals of both models, which allows us to have the uncertainty of both models integrated in the credible intervals. However, we do lose the mathematical interpretability of the credible intervals, since simply adding up the intervals does not result in a theoretical credible interval.

For both models we have simulated the parameters with a Metropolis-Hastings algorithm and a Hamiltonian Monte Carlo algorithm. We clearly saw that the No-U-Turn Sampler provided faster convergence and a deeper exploration of the parameter space. Hoffman and Gelman (2014) describes that it is expected behaviour that the No-U-Turn Sampler outperforms Metropolis-Hastings algorithms. Metropolis-Hastings algorithms may have problems with inefficient proposal distributions, as we saw in the results of the term structure model simulations. The correlated posterior distributions of α and β in the term structure model caused the random-walk proposed values to be rejected a lot. The No-U-Turn Sampler will not have this problem, since the sampling of the parameters is based on Hamiltonian dynamics. We do need to somehow find an efficient step size parameter ϵ, however, if we use the Stan software, this parameter will be tuned during the burn-in period.

The results of the investment strategy that we created show that the strategy is not performing very well. Although we do have a high return at the end of our backtest period, we also see large variances in the returns of the strategy and we see that the Sharpe ratios are very low. The high volatile behaviour of the VIX index cause the market value of the futures contracts to be very volatile as well. Jumps in the VIX index result in the same behaviour for the VIX futures as well. As a result, the strategy can suddenly have large losses or profits when we hold positions in VIX futures when the VIX index is becomes highly volatile. We did find that we could slightly reduce the risk of our strategy by applying volatility targeting methods. However, the returns of the strategy still remained to be very volatile due to the nature of our strategy and the behaviour of the VIX index.

After further evaluation of our investment strategy, we saw that the contracts that expire during the backtest period created the highest returns, while the positions on futures contracts that get sold before expiration have the lowest returns. This is contradicting our initial idea of creating a profit by selling our position when the market corrects the mispriced futures contract to the right price. Instead, the most profitable positions on futures contracts are the positions that eventually expire and return a cash settlement. We saw that long positions expire mostly during high volatile periods of the VIX, while short positions expire mostly during low volatile periods. Therefore, the strategy seems to make more profit using the long-term movement of the VIX index rather than exploiting mispriced futures contracts. The high volatile nature of the VIX index is probably the cause of this, since the large changes in the VIX index have more impact on the value of the futures contract than the correction of mispriced futures contracts to the right value. For example, when it takes ten days to correct a futures contract that was overvalued for $1, the price of the futures contract itself may have increased by $3 on its own due to an increase in the VIX index during this period. The return of our short position will now not necessarily depend that much on the correction of the market to the right price, but more on the underlying movement of the VIX index.

The investment strategy could be more realistic when we acquire data of the bid-ask spread for the opening prices of the VIX futures contracts. In the strategy we made the assumption that the bid-ask spread of the closing prices from the day before is of the same size as the bid-ask spread during the opening, since we do not have data of the bid-ask spread of the opening prices. Therefore, the buy and sell prices of our positions may be inaccurate. Another issue with buying the futures during the opening is that the price of the contract may be different than the closing price of the contract. The VIX index could change over night, which could result in different opening prices. When a contract is mispriced during the closing of the market, it may actually be correctly priced during the opening of the market, which would imply that we take a position on a futures contract that is not undervalued or overvalued. If we would work with data of VIX futures during the day, then we could avoid this problem.

For further research, one could try to find a way to combine the uncertainty of the term structure model and the error correction model in a correct way. Due to the formulation of our models, we were not able to simulate the models at the same time and combine the credible intervals. We do see promising results when we add up the credible intervals of both models, which motivates the need of the merging of the credible intervals of both models. However, if we want to further improve the investment strategy, one could look for hedging alternatives when taking a position in VIX futures. Dash and Moran (2005) explain that usually an allocation of 0% to 10% of the VIX index in a portfolio give the optimal results, due to the high volatile nature of the VIX index. Our strategy uses 100% of the VIX index, which resulted in high volatile returns.

References

Albulescu, C. (2020). Coronavirus and financial volatility: 40 days of fasting and fear. https://doi.org/10.2139/ssrn.3550630

Auinger, F. (2015). *The causal relationship between the s&p 500 and the vix index: Critical analysis of financial market volatility and its predictability*. Springer. https://doi.org/10.1007/978-3-658-08969-6

Bailey, D., & Lopez de Prado, M. (2012). The sharpe ratio efficient frontier. *Journal of Risk*, *15*(2), 3-44. https://doi.org/10.2139/ssrn.1821643

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *The Journal of Political Economy*, *81*(3), 637-654. https://doi.org/10.1086/260062

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, *31*(3), 307-327. https://doi.org/10.1016/0304-4076(86)90063-1

Bongaerts, D., Kang, X., & van Dijk, M. (2020). Conditional volatility targeting. *Financial Analysts Journal*, *76*(4), 54-71. https://doi.org/10.1080/0015198X.2020.1790853

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (2011). *Handbook of markov chain monte carlo*. CRC press. `https://www.mcmchandbook.net`

CBOE. (2019). *CBOE Volatility Index* (White paper). `https://cdn.cboe.com/resources/education/research_publications/vixwhite.pdf`

CBOE. (2021). CFE Futures Price & Volume Detail for Friday, June 18, 2021. `https://www.cboe.com/us/futures/market_statistics/daily/?dt=2021-06-18`

Cheridito, P., Filipović, D., & Kimmel, R. (2007). Market price of risk specifications for affine models: Theory and evidence. *Journal of Financial Economics*, *83*(1), 123-170. https://doi.org/10.1016/j.jfineco.2005.09.008

Cox, J. (1975). Notes on option pricing I: Constant elasticity of diffusions. (Unpublished manuscript, Stanford University)

Cox, J., Ingersoll, J., & Ross, S. (1981). The relation between forward prices and futures prices. *Journal of Financial Economics*, *9*(4), 321-346. https://doi.org/10.1016/0304-405X(81)90002-7

Cox, J., Ingersoll, J., & Ross, S. (1985). A theory of the term structure of interest rates. *Econometrica*, *53*(2), 385-407. https://doi.org/10.2307/1911242

Daigler, R., & Rossi, L. (2006). A portfolio of stocks and volatility. *The Journal of Investing*, *15*(2), 99-106. https://doi.org/10.3905/joi.2006.635636

Dash, S., & Moran, M. (2005). Vix as a companion for hedge fund portfolios. *The Journal of Alternative Investments*, *8*(3), 75-80. https://doi.org/10.3905/jai.2005.608034

Demeterfi, K., Derman, E., Kamal, M., & Zou, J. (1999). More than you ever wanted to know about volatility swaps. *Goldman Sachs*. `http://emanuelderman.com/wp-content/uploads/1999/02/gs-volatility_swaps.pdf`

Dotsis, G., Psychoyios, D., & Skiadopoulos, G. (2007). Continuous-time estimation, conditional characteristic function, implied volatility indices, volatility derivatives, vix futures. *Journal of Banking Finance*, *31*(12), 3584-3603. https://doi.org/10.1016/j.jbankfin.2007.01.011

Duane, S., Kennedy, A., Pendleton, B., & Roweth, D. (1987). Hybrid monte carlo. *Physics letters B*, *195*(2), 216-222. https://doi.org/10.1016/0370-2693(87)91197-X

Dupoyet, B., Daigler, R. T., & Chen, Z. (2011). A simplified pricing model for volatility futures. *Journal of Futures Markets*, *31*(4), 307-399. https://doi.org/10.1002/fut.20471

Engle, R. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, *50*(4), 987-1008. https://doi.org/10.2307/1912773

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, *6*(6), 721-741. https://doi.org/10.1109/TPAMI.1984.4767596

Girsanov, I. (1960). On transforming a certain class of stochastic processes by absolutely continuous substitution of measures. *Theory of Probability and its Applications*, *5*(3), 285-301. https://doi.org/10.1137/1105027

Grünbichler, A., & Longstaff, F. (1996). Valuing futures and options on volatility. *Journal of Banking Finance*, *20*(6), 985-1001. https://doi.org/10.1016/0378-4266(95)00034-8

Hamdan, R., Pavlowsky, F., Roncalli, T., & Zheng, B. (2016). A primer on alternative risk premia.

https://doi.org/10.2139/ssrn.2766850

Harvey, C., Hoyle, E., Korgaonkar, R., Rattray, S., Sargaison, M., & van Hemert, O. (2018). The impact of volatility targeting. *Journal of Portfolio Management*, *45*(1), 14-33. https://doi.org/10.3905/jpm.2018.45.1.014

Hastings, W. (1970). Equations of state calculations by fast computing machines. *Biometrika*, *57*(1), 97-109. https://doi.org/10.1093/biomet/57.1.97

Heston, S. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, *6*(2), 327–343. https://doi.org/10.1093/rfs/6.2.327

Hicks, J. (1946). *Value and capital: An inquiry into some fundamental principles of economic theory* (2nd ed.). Oxford: Clarendon Press.

Hoffman, M., & Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *The Journal of Machine Learning Research*, *15*(1), 1593-1623. https://www.jmlr.org/papers/volume15/hoffman14a/hoffman14a.pdf

Hull, J. (2015). *Risk management and financial institutions* (4th ed.). Wiley. https://www.wiley.com/en-us/Risk+Management+and+Financial+Institutions%2C+5th+Edition-p-9781119448099

Itô, K. (1944). Stochastic integral. *Proc. Imp. Acad.*, *20*(2), 519-524. https://doi.org/10.3792/pia/1195572786

Jeffrey, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *186*(1007), 453-461. https://doi.org/10.1098/rspa.1946.0056

Karatzas, I., & Shreve, S. (1998). *Brownian motion and stochastic calculus* (2nd ed.). Springer. https://doi.org/10.1007/978-1-4612-0949-2

Lo, A. (2002). The statistics of sharpe ratios. *Financial analysts journal*, *58*(4), 36-52. https://doi.org/10.2469/faj.v58.n4.2453

Lopez de Prado, M. (2018). *Advances in financial machine learning.* Wiley. https://www.wiley.com/en-us/Advances+in+Financial+Machine+Learning-p-9781119482086

Lutz, B. (2009). *Pricing of derivatives on mean-reverting assets.* Springer. https://doi.org/10.1007/978-3-642-02909-7

Majmudar, U., & Banerjee, A. (2004). VIX forecasting. *The 40th Annual Conference of the Indian Econometrics Society*. https://doi.org/10.2139/ssrn.533583

Mertens, E. (2002). *Comments on Variance of the IID estimator in Lo (2002)* (Working paper). http://www.elmarmertens.com/research/discussion

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E. (1953). Equations of state calculations by fast computing machines. *The journal of chemical physics*, *21*(6), 1087-1092. https://doi.org/10.1063/1.1699114

Moreira, A., & Muir, T. (2017). Volatility-managed portfolios. *The Journal of Finance*, *72*(4), 1611-1644. https://doi.org/10.1111/jofi.12513

Neal, R. (2003). Slice sampling. *The annals of statistics*, *31*(3), 705-767. https://doi.org/10.1214/aos/1056562461

Nordén, L., & Hou, A. (2018). Vix futures calendar spreads. *Journal of Futures Markets*, *38*(7), 822-838. https://doi.org/10.2139/ssrn.2968918

Novikov, A. (1972). On an identity for stochastic integrals. *Theory of Probability and its Applications*, *17*(4), 717-720. https://doi.org/10.1137/1117088

Oosterlee, C., & Grzelak, L. (2019). *Mathematical modeling and computation in finance.* World Scientific. https://doi.org/10.1142/q0236

Poon, S.-H. (2005). *A practical guide to forecasting financial market volatility.* John Wiley & Sons Ltd.

Rachev, S., Hsu, J., Bagasheva, B., & Fabozzi, F. (2008). *Bayesian methods in finance.* Wiley. https://www.wiley.com/en-us/Bayesian+Methods+in+Finance-p-9780470249246

Robert, C., & Casella, G. (2004). *Monte carlo statistical methods* (2nd ed.). Springer. https://doi.org/10.1007/978-1-4757-4145-2

Roberts, G., Gelman, A., & Gilks, W. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, *7*(1), 110-120.

https://doi.org/10.1214/aoap/1034625254

Sharpe, W. (1966). Mutual fund performance. *The Journal of business*, *39*(1), 119-138. https://doi.org/10.1086/294846

Simon, D., & Campasano, J. (2014). The vix futures basis: Evidence and trading strategies. *The Journal of Derivatives*, *21*(3), 54-69. https://doi.org/10.2139/ssrn.2094510

Stan Development Team. (2021). Stan Modeling Language Users Guide and Reference Manual, version 2.27. https://mc-stan.org

Van der Vaart, A. (2000). *Asymptotic statistics* (Vol. 3). Cambridge university press. https://doi.org/10.1017/CBO9780511802256

Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics*, *5*(2), 177-188. https://doi.org/10.1016/0304-405X(77)90016-2

Whaley, R. (1993). Derivatives on market volatility: Hedging tools long overdue. *The Journal of Derivatives*, *1*(1), 71-84. https://doi.org/10.3905/jod.1993.407868

Whaley, R. (2009). Understanding the vix. *The Journal of Portfolio Management*, *35*(3), 98-105. https://doi.org/10.3905/JPM.2009.35.3.098