

# Essays on rough and classical stochastic volatility

Sigurd Emil Rømer

PhD Thesis  
Submitted on March 10, 2022

This thesis has been submitted to the PhD School of The Faculty of  
Science, University of Copenhagen

Department of Mathematical Sciences  
Faculty of Science  
University of Copenhagen

Sigurd Emil Rømer  
Department of Mathematical Sciences  
University of Copenhagen  
Universitetsparken 5  
2100 København Ø  
Denmark

Email: sigurdroemer@gmail.com

Supervisor: Rolf Poulsen, University of Copenhagen

Assessment committee: Christa Cuchiero, University of Vienna  
Antoine Jacquier, Imperial College London  
David Skovmand, University of Copenhagen

Submission date: March 10, 2022

ISBN: 978-87-7125-054-1

# Preface

This thesis has been prepared in partial fulfilment of the requirements for the PhD degree at the Department of Mathematical Sciences, Faculty of Science, University of Copenhagen. The project has been supervised by professor Rolf Poulsen from the same department and was conducted in the period between March 2019 and March 2022.

I first wish to thank Rolf Poulsen for encouraging me to apply for a PhD position and for the many interesting discussions we had about the joint paper that we wrote. I would also like to thank Elisa Nicolato from Aarhus University for letting me stay for a week at the end of my PhD now that travelling abroad became difficult amid the Covid-19 pandemic. I much appreciate the discussions we had about rough volatility and I likewise do for the many other people that I discussed the topic with on e-mail. Lastly, I am thankful for the many helpful colleagues that I met at the department during my time.

The thesis consists of an introduction and three self-contained papers. Notation may change between the papers, though, it will always be properly introduced. It should therefore not cause any confusion.

*Sigurd Emil Rømer  
Østerbro, March 2022*

# Contents

Preface	i
Contents	ii
Abstract	iii
Resumé	iv
List of papers	v
Summary	vi
Sammenfatning	viii
1 Introduction	1
2 How does the volatility of volatility depend on volatility?	6
3 Hybrid multifactor scheme for stochastic Volterra equations with completely monotone kernels	28
4 Empirical analysis of rough and classical stochastic volatility models to the SPX and VIX markets	124

# Abstract

This thesis deals with stochastic volatility for financial markets with a special focus on the recent paradigm of *rough* volatility. We start with an investigation of how the volatility-of-volatility of the S&P 500 index depends on volatility. Our main conclusion is that volatility behaves more like a log-normal model than a square-root model. Moreover, we find that an accurate specification of the level-dependence matters for the predictive quality, including for the effective hedging of options. Next, we propose the *hybrid multifactor scheme* for the simulation of stochastic Volterra equations with completely monotone kernels, rough volatility especially. We prove convergence and develop efficient methods for computation of the VIX index for a number of volatility models: multifactor Volterra Bergomi, quadratic Volterra Heston, generalised CEV Volterra. We observe good numerical convergence except for a specific parameter choice under rough Heston where a large positivity bias appears. We then look at the problem of calibrating to SPX options and jointly to SPX and VIX options. The key observations are as follows: We find that the one-factor rough Bergomi model falls short on solving the SPX calibration problem in two ways: (1) it fails to sufficiently separate the volatility-of-volatility that is implied by option prices at short and long expiries, (2) it fails to create a term structure of smile (a)symmetry; the latter is needed as we find short-term option smiles to be more symmetric, generally. We propose an alternative volatility model driven by two Ornstein-Uhlenbeck processes that uses a non-standard transformation function. We demonstrate that our model can calibrate almost perfectly to SPX options and very well to SPX and VIX options jointly. This suggests that the SPX and VIX options markets can largely be reconciled with two-factor classical volatility, all without roughness and jumps.

# Resumé

Denne afhandling omhandler stokastisk volatilitet for finansielle markeder med et specielt fokus på det nylige paradigme *rough volatilitet*. Vi begynder med en undersøgelse af hvordan volatiliteten-af-volatiliteten på S&P 500 indekset afhænger af volatiliteten selv. Vi bemærker at volatiliteten opfører sig mere som en log-normal model end en kvadratrodsmodel. Endvidere bemærker vi at en korrekt specifikation af niveauafhængigheden er væsentligt for den prædiktive kvalitet, herunder til at hedge optioner effektivt. Derefter foreslår vi *hybrid multifaktor skemaet* til simulation af stokastiske Volterra ligninger med fuldstændigt monotone kernels, rough volatilitet specielt. Vi beviser konvergens og udleder effektive metoder til beregning af VIX indekset for en række volatilitetsmodeller: multifaktor Volterra Bergomi, kvadratisk Volterra Heston, generaliseret CEV Volterra. Vi observerer god numerisk konvergens bortset fra et specifikt parametervalg under rough Heston hvor et betydeligt positivitetsbias opstår. Vi tager derefter et kig på kalibreringsproblemet til SPX optioner og det til SPX og VIX optioner samlet. De vigtigste observationer er som følger: Vi bemærker at et-faktor rough Bergomi modellen kommer til kort på to måder i forhold til at løse SPX kalibreringsproblemet: (1) modellen kan ikke i tilstrækkelig grad afkoble den volatilitet-af-volatilitet der er underforstået i markedspriserne på henholdsvis korte og lange tidshorisonter, (2) modellen er ikke i stand til at skabe en tidsstruktur af (a)symmetri; det sidste er nødvendigt idet vi finder at optionssmil for korte tidshorisonter er mere symmetriske, generelt. Vi foreslår en alternativ model drevet af to Ornstein-Uhlenbeck processer der benytter en ikke-standard transformationsfunktion. Vi demonstrerer at vores model kan kalibrere næsten perfekt til SPX optioner og meget godt til SPX og VIX optioner samlet. Observationen antyder at SPX og VIX markederne kan forenes med to-faktor klassisk volatilitet, helt uden roughness og spring.

# List of papers

The thesis is based on three papers:

- Rømer, S. E., Poulsen, R., How does the volatility of volatility depend on volatility? *Risks*, 2020, **8**(2), 59.
- Rømer, S. E., Hybrid multifactor scheme for stochastic Volterra equations with completely monotone kernels. 2022. Preprint is available at SSRN:3706253.
- Rømer, S. E., Empirical analysis of rough and classical stochastic volatility models to the SPX and VIX markets. 2022. A version of this paper is under review in *Quantitative Finance*.

# Summary

This dissertation is based on three self-contained papers. Their abstracts are given below:

- **How does the volatility of volatility depend on volatility?** We investigate the state dependence of the variance of the instantaneous variance of the S&P 500 index empirically. Time-series analysis of realized variance over a 20-year period shows strong evidence of an elasticity of variance of the variance parameter close to that of a log-normal model, albeit with an empirical autocorrelation function that one-factor diffusion models fail to capture at horizons above a few weeks. When studying option market behavior (in-sample pricing as well as out-of-sample pricing and hedging over the period 2004–2019), messages are mixed, but systematic, model-wise. The log-normal but drift-free SABR (stochastic-alpha-beta-rho) model performs best for short-term options (times-to-expiry of three months and below), the Heston model—in which variance is stationary but not log-normal—is superior for long-term options, and a mixture of the two models does not lead to improvements.
- **Hybrid multifactor scheme for stochastic Volterra equations with completely monotone kernels.** We propose a hybrid scheme for the simulation of stochastic Volterra equations with completely monotone kernels. Our scheme is a mix of the hybrid scheme for Brownian semistationary processes of Bennedsen et al. [Financ. Stoch., **21**(4), 931-965, 2017] and the multifactor approximations of Abi Jaber et al. [SIAM J. Finan. Math., **10**(2), 309-349, 2019]. Merging the two methods allow us to both accurately capture singularities and efficiently track the inherent path dependence. We consider a forward process that is easily computable under our scheme and show how it facilitates simulation of the VIX index for a number of volatility models of the Volterra type. Numerical experiments indicate good convergence for rough Bergomi type models and the quadratic rough Heston model. Experiments on rough Heston, where we had to truncate values in zero, sometimes resulted in a large positive bias.
- **Empirical analysis of rough and classical stochastic volatility models to the SPX and VIX markets.** We conduct an empirical analysis of rough and classical stochastic volatility models to the SPX and VIX options markets. Our analysis focusses primarily on calibration quality and is split in two parts. In the first part, we perform a historical calibration to SPX options over the years 2004-2019 of a selection of models that include the one-factor rough Bergomi model. In the second part, we consider three calibration dates with low, typical, and

high volatility, but examine a wide selection of models and calibrate to both SPX options and jointly to SPX and VIX options. The key results are as follows: The rough Bergomi model fails to create a term structure of smile effect that is sufficiently pronounced for SPX options. Moreover, we discover that short-expiry SPX smiles generally are more symmetric than long-expiry smiles, a feature we neither find that the rough Bergomi model can reproduce. We propose an alternative volatility model driven by two Ornstein-Uhlenbeck processes that uses a non-standard transformation function. Calibrating it to SPX options we obtain almost perfect fits and calibrating it jointly to SPX and VIX options we obtain very decent fits. This suggests—contrary to what one might be led to believe based on much of the existing literature—that the joint SPX-VIX calibration problem is largely solvable with classical two-factor volatility, all without roughness and jumps.

# Sammenfatning

Afhandlingen baserer sig på tre selvstændige artikler. Resuméer gives nedenfor:

- **How does the volatility of volatility depend on volatility?** Vi undersøger tilstandsafhængigheden af variansen af den instantane varians af S&P 500 indekset empirisk. Tidsserieanalyse på realiseret varians over en 20-årig periode viser stærk evidens for en elasticitet af variansen af variansparameteren der er tæt på den for en log-normal model, dog med en empirisk autokorrelationsfunktion som et-faktor diffusionsmodeller ikke kan reproducere på tidshorisoner udover et par uger. Når vi undersøger adfærden på optionsmarkedet (in-sample prisning såvel som out-of-sample prisning og hedging over perioden 2004–2019), da er konklusionerne blandede, men systematiske, modelmæssigt. Den log-normale men driftfrie SABR (stokastisk-alpha-beta-rho) model performer best for kortsigtede optioner (tid-til-udløb på tre måneder eller derunder), Heston modellen—hvor variansen er stationær men ikke log-normal—er overlegen for langsigtede optioner, og en blanding af de to modeller giver ikke nogen forbedringer.
- **Hybrid multifactor scheme for stochastic Volterra equations with completely monotone kernels.** Vi foreslår et hybrid skema til simulation af stokastiske Volterra ligninger med fuldstændigt monotone kernels. Vores skema er en blanding af hybrid skemaet for Brownske semistationære processer af Bennedsen et al. [Financ. Stoch., **21**(4), 931-965, 2017] og multifaktor approksimationerne af Abi Jaber et al. [SIAM J. Finan. Math., **10**(2), 309-349, 2019]. Ved at slå de to metoder sammen er vi i stand til både at håndtere singulariteter og at holde styr på den iboende stiafhængighed. Vi betragter en forward process der er let at udregne under vores skema og viser hvorledes den faciliterer simulation af VIX indekset for en række Volterra volatilitetsmodeller. Numeriske eksperimenter indikerer god konvergens for rough Bergomi type modeller og den kvadratiske rough Heston model. Eksperimenter med rough Heston, hvor vi måtte trunkere værdierne i nul, resulterede til tider i et stort positivitetsbias.
- **Empirical analysis of rough and classical stochastic volatility models to the SPX and VIX markets.** Vi udfører en empirisk analyse af rough og klassiske stokastiske volatilitetsmodeller til SPX og VIX optionsmarkederne. Vores analyse fokuserer primært på kalibreringskvalitet og består af to dele. I den første del udfører vi en historisk kalibrering til SPX optioner over årene 2004-2019 af et udvalg af modeller der inkluderer et-faktor rough Bergomi modellen. I den anden del betragter vi tre kalibreringsdage med henholdsvis lav, typisk, og høj volatilitet,

men undersøger en bred selektion af modeller og kalibrerer til både SPX optioner og samlet til SPX og VIX optioner. De vigtigste resultater er som følger: Rough Bergomi modellen er ikke i stand til at skabe en kurve for smileffekten der er tilstrækkeligt markant for SPX optioner. Endvidere observerer vi at smilene for kortsigtede SPX optioner generelt er mere symmetriske end dem for langsigtede optioner. Rough Bergomi modellen er ligeledes ikke i stand til at genskabe dette kendetegn. Vi foreslår i stedet en klassisk volatilitetsmodel drevet af to Ornstein-Uhlenbeck processer der benytter en ikke-standard transformationsfunktion. Ved kalibrering til SPX optioner opnår vi næsten perfekte fits og ved kalibrering til SPX og VIX optioner meget anstændige fits. Dette antyder—i modsætning til hvad man kunne forledes til at tro baseret på meget af den eksisterende litteratur—at det samlede SPX-VIX kalibreringsproblem mere eller mindre kan løses med klassisk to-faktor volatilitet, helt uden roughness og spring.

# Chapter 1

## Introduction

The goal of the PhD project has been to contribute to the literature on stochastic volatility for financial markets with a special focus on the recent paradigm of *rough* volatility. Our work is partly on empirical aspects (we look at realized volatility and derivatives prices), partly on simulation of stochastic rough, and more generally Volterra, models. To understand how our contributions relate to the existing literature, it is worth giving some highlights of the history of volatility modelling, starting with the Black-Scholes model of [6] (published 1973) going all the way up to today's rough volatility models.

In the Black-Scholes model the asset price is assumed to have constant volatility and no jumps. Due to its simplicity there is an explicit (log-normal) solution for the asset price. Analytical pricing formulas may then be derived for many option types (European calls and puts most notably) making implementation easy and robust. Its simplicity, however, is also its downfall. A constant volatility model with no jumps is inconsistent with (1) 'skews' or 'smiles' when looking at (Black-Scholes) implied volatility across strikes, (2) the fatter than Gaussian log-returns observed for many assets. To reproduce these features, more realistic *stochastic* volatility models have since appeared. The Heston model of [13] (published 1993) is an early example that has been widely adopted in banks due to the fast Fourier pricing methods that apply and which are a consequence of its affine structure. Another example is the SABR (stochastic-alpha-beta-rho) model of [12] (published 2002) which likewise has been widely adopted also largely due to tractability; there is an analytical approximation formula for implied volatility, commonly known as 'Hagan's formula'. Multifactor stochastic volatility models that offer more flexibility have also appeared. The  $n$ -factor Bergomi model is a well-known example; the two-factor and discrete  $n$ -factor versions are formulated in [5] (published 2005).

Another modelling approach is to introduce jumps. While some might say that jumps are a natural feature of financial prices, the authors of [7] argue with the use of high-frequency data that by far most of the price movements on many assets can be explained by volatility, not jumps. In this project, we shall limit ourselves to pure stochastic volatility which then is justified.

Lastly, we come to our main focus, *rough volatility*. Motivated by empirical observations

of realized volatility, the authors of [10] (published 2018) show that volatility trajectories are less smooth than those of Brownian motion, more generally SDE's. Combined with the evidence of [7] that the jump-component is small, it suggests to look for a continuous-path stochastic volatility model that lies beyond classical ones (Heston, SABR, and the  $n$ -factor Bergomi model, are all driven by SDE's and are thus 'classical' in our terminology). The authors of [10] more precisely argue that log-volatility behaves much like *fractional Brownian motion* (fBm) with a *Hurst exponent*  $H$  around 0.1. Fractional Brownian motion is a Gaussian process that generalizes Brownian motion to allow the increments to correlate. The correlation is negative if  $H < \frac{1}{2}$ , positive if  $H > \frac{1}{2}$ , and zero if  $H = \frac{1}{2}$  giving ordinary Brownian motion;  $H$  is restricted to  $(0, 1)$ . The sample paths are Hölder continuous of all orders less than  $H$ . When the process is less smooth than Brownian motion (i.e.  $H < \frac{1}{2}$ ) we call it *rough*. The observation  $H \approx 0.1$  therefore suggests a process with very erratic (rough) sample paths. The authors propose the RFSV ('Rough Fractional Stochastic Volatility') model that effectively is based on fBm and which they find is consistent with many empirical observations of realized volatility.

The RFSV model is a special case of the FSV ('Fractional Stochastic Volatility') model of [8] (published 1998), where the authors in contrast choose  $H > \frac{1}{2}$  to allow for long-memory (a feature of volatility that effectively says that the autocorrelations should decay 'slowly' at long time horizons and which there are indications of in e.g. [3]). The RFSV model does not allow for long-memory as such but as shown in [10] statistical tests applied to sample paths generated by the model may result in false positive detections of long-memory. In [2], the rough Bergomi pricing model was derived by application of a deterministic change of measure to the RFSV model. The authors show that the model can be remarkably consistent with quoted SPX options. A central feature of the volatility surface for European calls and puts is the term-structure of (at-the-money) skew; we write  $\text{Skew}(T)$  where  $T$  is the expiry. For SPX options, we often observe  $\text{Skew}(T) \sim cT^\alpha$  where  $c$  is an anonymous constant and  $\alpha \approx -0.4$ . The rough Bergomi model reproduces this relationship with  $\alpha = H - \frac{1}{2}$  and is therefore consistent with  $H \approx 0.1$ . Classical one-factor models such as Heston are unable to reproduce the power-law term structure of skew; see [9, pp. 35] for an approximate skew formula for the Heston model.

With all the evidence pointing towards rough volatility, it is no wonder that the topic has gained so much attention in recent years.

In the remaining chapters, we will look at volatility through the lenses of both classical and rough volatility as well as cover simulation aspects. We outline the chapters below.

In Chapter 2, we limit ourselves to classical volatility. We consider the question of how the volatility-of-volatility of the S&P 500 index depends on volatility. We look at both realized volatility and derivatives prices. In terms of the latter, we conduct calibration and hedge experiments, comparing mainly the (square-root diffusion) Heston model with the (log-normal) SABR model. Our evidence indicates that the volatility-of-volatility depends on volatility more like that of a log-normal model than a square-root one. Moreover, we find that an accurate specification of the volatility-of-volatility level-dependence matters for the predictive quality, in particular, for the effective hedging of

options. On a side-note, calibrating a one-factor classical volatility model (with linear drift) to realized volatility, we find that the autocorrelations fail to match empirical estimates at horizons beyond a few weeks. This suggests the need for a more complex 'mean-reversion' specification such as that of rough, or multifactor classical, volatility.

In Chapter 3, we consider the simulation of *stochastic Volterra equations*. These are equations of the form

$$X_t = g_0(t) + \int_0^t K(t-s)b(X_s)ds + \int_0^t K(t-s)\sigma(X_s)dW_s, \quad t \geq 0, \quad (1.1)$$

where  $g_0, K, b, \sigma$  are deterministic functions,  $(W_t)_{t \geq 0}$  is Brownian motion, and we look for a solution  $(X_t)_{t \geq 0}$ . With the fractional kernel  $K(t) = t^{H-\frac{1}{2}}$  where  $H \in (0, \frac{1}{2})$ , we get, under mild conditions, a process that has the same regularity, i.e. smoothness, as (rough) fractional Brownian motion with Hurst exponent  $H$ . The main challenges from a simulation perspective are: (a) The convolutional nature of (1.1), which for a naive Euler discretisation results in costs  $O(n^2)$  where  $n$  is the number of steps. (b) That  $K$  is approximated poorly by such Euler (i.e. piecewise constant) approximations if singular.

Many alternative discretisations of Volterra processes exist in the literature. Two important examples are the hybrid scheme of [4] and the multifactor approximations of [1]. While the models covered in these papers do not fully overlap with (1.1), the ideas to a large extent generalise to that equation. Under the hybrid scheme of [4]  $K$  is used without error near the origin and approximated by a step function elsewhere. Extrapolated to (1.1), the coefficient processes  $b(X.)$ ,  $\sigma(X.)$  are approximated piecewise constant everywhere. For the multifactor approximations of [1]  $K$  is assumed *completely monotone*, which covers most conceivable positive, decreasing, and smooth functions, including the rough fractional kernel  $K(t) = t^{H-\frac{1}{2}}$ . The kernel  $K$  is approximated by a weighted sum of exponential functions which results in a multifactor Markovian model. This model can then be discretised with an Euler scheme. Our contribution is to merge the two ideas by using  $K$  without error near the origin and to approximate it by a sum-of-exponentials elsewhere. Our reasoning is that it is costly to capture a singularity with exponentials as they are non-singular, whereas they can be very efficient for the non-singular part. This is backed up by our numerical experiments. We provide a comprehensive guide to the resulting discretisation which we call the *hybrid multifactor scheme*.

In Chapter 4, we perform an empirical analysis of classical and rough volatility models to the SPX and VIX markets. Our analysis focuses mainly on calibration quality and is motivated by a lack of thorough testing of rough volatility pricing models in the existing literature, especially fair comparisons against classical models. We test a wide selection of models, including the Heston and rough Bergomi models, a quadratic rough Heston model akin to [11], and several two-factor models. Our results shed new light on what is required of a continuous-path stochastic volatility model to calibrate to the SPX and VIX options markets. Some key observations are this: While we find that rough Bergomi calibrates better to SPX options than classical Heston, it falls short in two ways. The first is that, as we discover, short-term SPX smiles are more symmetric than long-term smiles, generally. We are unable to reproduce this (a)symmetry term-structure with one-factor models such as rough Bergomi. We even find that the widely reported relationship

$\text{Skew}(T) \sim cT^\alpha$  sometimes do not hold for the market data due to the symmetry of the short term smiles. The second way rough Bergomi falls short in calibrating to SPX options, is that it fails to sufficiently decouple the smile effect at short and long expiries. Extending to a two-factor volatility framework, we obtain mostly better results, though, we find no meaningful differences between comparable rough and classical models. We construct a classical two-factor volatility model that can calibrate very well to even the joint SPX-VIX calibration problem. This demonstrates that the joint problem can be solved without roughness and jumps. Our analysis is extensive and offers many other insights into the volatility dynamics that are implied by quoted SPX and VIX options.

## References

- [1] Abi Jaber, E., El Euch, O., Multifactor approximation of rough volatility models. *SIAM Journal on Financial Mathematics*, 2019, **10**(2), 309–349.
- [2] Bayer, C., Friz, P., Gatheral, J., Pricing under rough volatility. *Quantitative Finance*, 2016, **16**(6), 887–904.
- [3] Bennedsen, M., Lunde, A., Pakkanen, M. S., Decoupling the short- and long-term behavior of stochastic volatility. *Journal of Financial Econometrics*, 2021, 1–46.
- [4] Bennedsen, M., Lunde, A., Pakkanen, M. S., Hybrid scheme for Brownian semistationary processes. *Finance and Stochastics*, 2017, **21**, 931–965.
- [5] Bergomi, L., Smile dynamics II. *Risk*, October 2005, pp. 67–73.
- [6] Black, F., Scholes, M., The pricing of options and corporate liabilities. *The Journal of Political Economy*, 1973, **81**(3), 637–654.
- [7] Christensen, K., Oomen, R. C. A., Podolskij, M., Fact or friction: Jumps at ultra high frequency. *Journal of Financial Economics*, 2014, **114**(3), 576–599.
- [8] Comte, F., Renault, E., Long memory in continuous-time stochastic volatility models. *Mathematical Finance*, 1998, **8**(4), 291–323.
- [9] Gatheral, J., *The Volatility Surface: A Practitioner’s Guide*, John Wiley & Sons, New Jersey, 2006.
- [10] Gatheral, J., Jaisson, T., Rosenbaum, M., Volatility is rough. *Quantitative Finance*, 2018, **18**(6), 933–949.
- [11] Gatheral, J., Jusselin, P., Rosenbaum, M., The quadratic rough Heston model and the joint S&P 500/Vix smile calibration problem. *Risk*, May 2020.
- [12] Hagan, P. S., Kumar, D., Lesniewski, A. S., Woodward, D. E., Managing smile risk. *Wilmott Magazine*, January 2002, pp. 84–108.
- [13] Heston, S. L., A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 1993, **6**(2), 327–343.

# Chapter 2

## How does the volatility of volatility depend on volatility?

Sigurd Emil Rømer

Rolf Poulsen<sup>1</sup>

### Abstract

We investigate the state dependence of the variance of the instantaneous variance of the S&P 500 index empirically. Time-series analysis of realized variance over a 20-year period shows strong evidence of an elasticity of variance of the variance parameter close to that of a log-normal model, albeit with an empirical autocorrelation function that one-factor diffusion models fail to capture at horizons above a few weeks. When studying option market behavior (in-sample pricing as well as out-of-sample pricing and hedging over the period 2004–2019), messages are mixed, but systematic, model-wise. The log-normal but drift-free SABR (stochastic-alpha-beta-rho) model performs best for short-term options (times-to-expiry of three months and below), the Heston model—in which variance is stationary but not log-normal—is superior for long-term options, and a mixture of the two models does not lead to improvements.

*Keywords:* Stochastic volatility; Elasticity of variance of variance; Heston; SABR.

### 2.1 Introduction

The aim of this paper is to investigate the elasticity of variance of variance of the S&P 500 index. As that concept is not part of the standard financial nomenclature<sup>2</sup>, let us introduce the general set-up. We consider a probability space  $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$  whose filtration is generated by two independent Brownian motions  $W_{1,t}$  and  $\bar{W}_{\perp,t}$ . Letting

---

<sup>1</sup> Department of Mathematical Sciences, University of Copenhagen.

<sup>2</sup> A Google search on "elasticity of variance of variance" in late March 2020 gave only seven hits.

## 2.2. EVIDENCE FROM REALIZED VARIANCE

---

$S_t$  denote the value of the S&P 500 index at time  $t$ , we model dynamics under the real world probability measure  $\mathbb{P}$  as:

$$dS_t = \mu_t^s S_t dt + \sqrt{V_t} S_t dW_{1,t} \quad (2.1)$$

with

$$dV_t = \mu_t^v dt + \eta V_t^{\lambda_v} dW_{2,t} \quad (2.2)$$

and where we have defined the Brownian motion  $W_{2,t} = \rho W_{1,t} + \sqrt{1 - \rho^2} W_{\perp,t}$  for a  $\rho \in [-1, 1]$ , meaning that  $\langle dW_{1,t}, dW_{2,t} \rangle = \rho dt$ . We assume that  $\mu_t^s$  and  $\mu_t^v$  are general (but suitable) adapted processes. We refer to  $V_t$  as the (instantaneous) variance. Our main focus will be the elasticity parameter<sup>3</sup>  $\lambda_v$ , which controls the level dependence of variance-of-variance with respect to the instantaneous variance  $V_t$  itself. We will see that determining  $\lambda_v$  is not merely an academic exercise. Empirically the parameter choice is important for model performance, including the effective hedging of options.

There is an abundance of both theoretical and empirical work on stochastic volatility in financial markets. However, as succinctly formulated by [10], "Often it is observed that a specific stochastic volatility model is chosen not for particular dynamical features, but instead for convenience and ease of implementation". In this short paper we aim for an empirical cross-examination of models, methods, and markets. It means that either of the separate analyses may be described—or dismissed—as "quite partial", but we believe that the sum of their parts brings to the fore some insights that were not hitherto available. In Section 2.2 we estimate the elasticity from a time series of realized variance, thus extending the analysis in [24], where only pre-specified values of elasticity were considered. We end the section with an investigation of features that are not captured by one-factor diffusion-type stochastic volatility models, or long-memory or long-range dependence in particular. In Section 2.3 we turn our attention to option prices to investigate the two—by far—most widely used models, the square-root model from [14] and the log-normal SABR (stochastic-alpha-beta-rho) model from [19], as well a simple hybrid of the two models. We study in-sample calibration issues along the lines of [13] as well as prediction and hedge performance, adding to the analyses in [22] and [21].

## 2.2 Evidence from realized variance

In this section we present four different methods for estimating  $\lambda_v$  from time series data, investigate their biases in a simulation experiment, and finally show the estimates on publicly available empirical data—the realized variance obtained from the Oxford-Man Institute's "realized library".<sup>4</sup> In [24] it is demonstrated that it is possible to discriminate between common diffusion-type stochastic volatility models when measuring realized variance from a 5-minute observation frequency, so we chose that data (denoted "rv5" in the files) among the many time series in the Oxford-Man data.<sup>5</sup> Our dataset thus

<sup>3</sup> Mathematically, the elasticity of a function  $f$  is  $xf'(x)/f(x)$ . The *variance* of variance  $V_t$  is  $\eta^2 V_t^{2\lambda_v}$ , and thus pedants can rightly claim that  $2\lambda_v$  is the elasticity of variance of variance.

<sup>4</sup> <https://realized.oxford-man.ox.ac.uk/>

<sup>5</sup> The use of 5-minute observations to measure instantaneous variance is common and we think it is a reasonable compromise, but we should stress that it does not tell the whole story about what goes on at higher frequencies; see for instance [9].

CHAPTER 2. HOW DOES THE VOLATILITY OF VOLATILITY DEPEND ON  
VOLATILITY?

---

consists of daily observations between the 3rd of January 2000 and the 3rd of September 2019. After filtering out non-positive variance estimates and removing the day of the Flash Crash (6th of May 2010) we were left with a total of 4935 observations. We should also remark that since we focus on the  $\lambda_v$ -parameter, the estimation methods will all solely focus on estimating the variance (or volatility) process. That is, we leave out a joint estimation also including the index price.

Let us now start by assuming a mean-reverting variance model of the form

$$dV_t = \kappa(\theta - V_t)dt + \eta V_t^{\lambda_v} dW_{2,t} \quad (2.3)$$

with  $\kappa, \theta > 0$  being additional parameters. To estimate the parameters we can apply an Euler discretisation and then use maximum likelihood. Let  $t_0 < t_1 < \dots < t_n$  with  $n = 4934$  denoting the observation time points. We will, for simplicity, assume the time points are equidistant with step sizes of  $\Delta t = 1$  trading day with 252 trading days assumed per year. We then approximate

$$\Delta V_{t_i} = \kappa(\theta - V_{t_i})\Delta t + \eta V_{t_i}^{\lambda_v} \Delta W_{2,t_i} \quad (2.4)$$

for  $i = 0, 1, \dots, n-1$ , where we write  $\Delta X_{t_i} = X_{t_{i+1}} - X_{t_i}$  for a general process  $X$ . Under the approximative model (2.4), the joint density of  $(V_{t_1}, \dots, V_{t_n})$  is a product of conditional densities, which are all Gaussian. To obtain maximum approximate likelihood estimates we plug in the observed time series and numerically maximize the log of this joint density, the log-likelihood function, over the model parameters<sup>6</sup> and use the Hessian matrix of the log-likelihood at the maximum to give us standard errors. Another model specification is to use a mean-reverting model for the *volatility process*  $\sigma_t := \sqrt{V_t}$  instead,

$$d\sigma_t = \kappa(\theta - \sigma_t)dt + \eta \sigma_t^{\lambda_\sigma} dW_{2,t}. \quad (2.5)$$

It's formula applied to the squared solution of (2.5) reveals that while the mean-reverting volatility and variance models are not equivalent as the functional forms of the drifts are different, we do have

$$dV_t = \dots dt + 2\eta V_t^{(\lambda_\sigma + 1)/2} dW_{2,t} \quad (2.6)$$

so that  $\lambda_v = (\lambda_\sigma + 1)/2$  and the maximum approximate likelihood technique applied to the volatility process gives us an alternative way to estimate the elasticity.<sup>7</sup> The final two estimators we will consider are based on the concept of quadratic variation. Following proposition 4.21 of [15] the quadratic variation of a continuous semimartingale  $X$  over the interval  $[0, t]$ , here denoted  $\langle X, X \rangle_t$ , can be characterised as follows: Let  $0 = t_0^n < t_1^n < \dots < t_{p(n)}^n = t$  with  $p(n) \in \mathbb{N}$  for  $n = 1, 2, \dots$  be an increasing sequence of partitions of  $[0, t]$  satisfying  $\sup_{1 \leq i \leq p(n)} |t_i^n - t_{i-1}^n| \rightarrow 0$  as  $n \rightarrow \infty$ . Then

$$\langle X, X \rangle_t = \lim_{n \rightarrow \infty} \sum_{i=1}^{p(n)} \left( X_{t_i^n} - X_{t_{i-1}^n} \right)^2 \quad (2.7)$$

<sup>6</sup> We exclude here the  $\theta$ -parameter by instead fixing it at the time series mean of  $V_t$ . We find this produces a more reasonable estimation.

<sup>7</sup> In case  $V_t$  from Equation (2.3) can hit zero, applying the  $\sqrt{x}$ -function to get back an elasticity for  $\sigma_t$  is problematic. However, all our estimators suggest  $\lambda_v$  is above  $\frac{1}{2}$ , in which case zero (see for instance [3]) is unattainable, so we will use the relationship  $\lambda_v = (\lambda_\sigma + 1)/2$  without worry.

## 2.2. EVIDENCE FROM REALIZED VARIANCE

---

with convergence in probability. For the  $V_t$  in the general model (2.2) the quadratic variation is:

$$\langle V, V \rangle_t = \int_0^t \eta^2 V_s^{2\lambda_v} ds. \quad (2.8)$$

Combining (2.7) and (2.8) leads to the approximation  $(V_{t_i} - V_{t_{i-1}})^2 \approx \eta^2 V_{t_{i-1}}^{2\lambda_v} \Delta t$ , or in logarithm terms

$$\log((V_{t_i} - V_{t_{i-1}})^2) \approx 2 \log(\eta) + 2\lambda_v \log(V_{t_{i-1}}) + \log(\Delta t). \quad (2.9)$$

This means we can estimate  $\lambda_v$  and  $\eta$  from a linear regression between the samples  $X_i := \log(V_{t_{i-1}})$  and  $Y_i := \log((V_{t_i} - V_{t_{i-1}})^2)$ . Applying the same reasoning to the squared increments of the volatility process,

$$\log((\sigma_{t_i} - \sigma_{t_{i-1}})^2) \approx 2 \log(\eta) + 2\lambda_\sigma \log(\sigma_{t_{i-1}}) + \log(\Delta t), \quad (2.10)$$

gives way to estimate  $\lambda_\sigma$  and  $\eta$  by a linear regression.

The four estimators presented all suffer from several sources of bias: discretization bias, small sample bias, and a possible error from model misspecification; if the variance has linear drift, then the volatility does not, and vice versa. Hence to validate the estimators we conducted a simulation experiment, whose results, shown in Table 2.1, we shall now briefly describe. For the experiment we assumed either the mean-reverting volatility or the mean-reverting variance model to be the true model, and simulated 1000 sample paths with 2520 steps per year across a total of 4935 days, thus mimicking the empirical set-up. Using only daily observations, we then applied each of the four methods to estimate  $\lambda_v$  and  $\lambda_\sigma$  and averaged across the paths. We performed this entire experiment for different choices of  $\lambda_v \in [\frac{1}{2}, 1]$  and  $\lambda_\sigma \in [\frac{1}{2}, 1]$ . For a realistic set-up, the remaining parameters (except for  $\theta$  under (2.3)) were chosen by running the relevant likelihood estimation on the Oxford-Man dataset with  $\lambda_v$  or  $\lambda_\sigma$  held fixed at the true value. The experiment shows that both regression estimators are fairly robust under the two different models as well as across different elasticities. The same is true when using maximum approximate likelihood on Equation (2.5). For these three estimators, the bias when estimating  $\lambda_v$  is mostly less than 0.05. Maximum approximate likelihood estimation on Equation (2.3) performed much worse, with a consistent downwards bias of 0.1–0.2 depending on the true modeling assumptions and the elasticity used.

In Table 2.2 we show estimates on the actual dataset using each of the four methods. We include the  $R^2$  statistic for the regression methods and show standard errors in parenthesis. There is a high correspondence between the elasticities estimated under all methods—except using likelihood on Equation (2.3), which we just saw to be problematic. We found  $\lambda_v$ -values in the range 0.91–0.97. In this range, the bias is still mostly less than 0.05 and the estimated standard errors are less than 0.02, so we are quite confident that the elasticity is in that range. The estimates in Table 2.2 also show that instantaneous variance is characterized by a combination of strong mean-reversion and high volatility of volatility (both  $\kappa$  and  $\eta$  are high compared to many other sources), which—if one were to think briefly beyond the realm of diffusion models—could point towards so-called rough volatility models as suggested by [12].<sup>8</sup>

<sup>8</sup> A very useful source for the rapidly expanding field of rough volatility is <https://sites.google.com/site/roughvol/home>.

**Table 2.1:** The accuracy when estimating the elasticity parameters  $\lambda_v$  and  $\lambda_\sigma$  under various modelling assumptions (mean-reverting variance or volatility) and various estimation techniques (maximum approximate likelihood estimation and regressions). The experiment is conducted as follows: Assuming either the model (2.3) or (2.5) we first simulated  $N = 1000$  independent paths across 4935 days with 10 simulation steps per day. Here we use the locally log-normal scheme from [2] as presented in [18]. Using only the 4935 daily values we then path-by-path estimated the  $\lambda_v$  and  $\lambda_\sigma$  parameters using each of the four estimation techniques. This gives us 1000 estimates for each, which we write as  $\{\hat{\lambda}_{v,i}\}_{i=1}^N$  and  $\{\hat{\lambda}_{\sigma,i}\}_{i=1}^N$  respectively. The goal is then to estimate the means  $E(\hat{\lambda}_{v,1})$  and  $E(\hat{\lambda}_{\sigma,1})$  for each estimation technique. These values should then be compared to the true values used to simulate the paths (columns 4 and 9 below) to see how biased each method is. The means can naturally be estimated by the averages  $\bar{\lambda}_\sigma = \frac{1}{N} \sum_{i=1}^N \hat{\lambda}_{\sigma,i}$  (columns 5-8) and  $\bar{\lambda}_v = \frac{1}{N} \sum_{i=1}^N \hat{\lambda}_{v,i}$  (columns 10-13). To get an idea of how reliably  $\bar{\lambda}_v$  and  $\bar{\lambda}_\sigma$  estimate the means of each method we can also compute standard errors as  $\hat{v}_v/\sqrt{N}$  and  $\hat{v}_\sigma/\sqrt{N}$  where we have defined  $\hat{v}_v = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (\hat{\lambda}_{v,i} - \bar{\lambda}_v)^2}$  and  $\hat{v}_\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (\hat{\lambda}_{\sigma,i} - \bar{\lambda}_\sigma)^2}$ . While we do not show the standard errors we can report that they were all less than 0.01.

True model:  $dV_t = \kappa(\theta - V_t)dt + \eta V_t^{\lambda_v} dW_{2,t}$

$\kappa$	$\theta$	$\eta$	$\lambda_v$	$\bar{\lambda}_v$	$\hat{v}_v$	$\lambda_\sigma$	$\bar{\lambda}_\sigma$	$\hat{v}_\sigma$	$\lambda_v$	$\bar{\lambda}_v$	$\hat{v}_v$	$\lambda_\sigma$	$\bar{\lambda}_\sigma$	$\hat{v}_\sigma$
25.56	0.16 <sup>2</sup>	2.31	<b>0.50</b>	0.35	0.63	0.59	0.68	<b>0.00</b>	-0.31	0.25	0.19	0.36	0.44	0.36
22.34	0.16 <sup>2</sup>	3.27	<b>0.60</b>	0.44	0.68	0.65	0.72	<b>0.20</b>	-0.12	0.36	0.29	0.44	0.57	0.44
20.25	0.16 <sup>2</sup>	4.82	<b>0.70</b>	0.54	0.75	0.72	0.78	<b>0.40</b>	0.08	0.49	0.43	0.57	0.68	0.57
18.64	0.16 <sup>2</sup>	7.33	<b>0.80</b>	0.65	0.82	0.79	0.86	<b>0.60</b>	0.29	0.64	0.58	0.71	0.87	0.71
17.25	0.16 <sup>2</sup>	11.49	<b>0.90</b>	0.75	0.89	0.87	0.93	<b>0.80</b>	0.50	0.79	0.73	0.87	1.02	0.87
16.01	0.16 <sup>2</sup>	18.52	<b>1.00</b>	0.86	0.96	0.93	1.01	<b>1.00</b>	0.71	0.93	0.86	1.02	1.02	1.02

True model:  $d\sigma_t = \kappa(\theta - \sigma_t)dt + \eta\sigma_t^{\lambda_\sigma} dW_{2,t}$

$\kappa$	$\theta$	$\eta$	$\lambda_v$	$\bar{\lambda}_v$	$\hat{v}_v$	$\lambda_\sigma$	$\bar{\lambda}_\sigma$	$\hat{v}_\sigma$	$\lambda_v$	$\bar{\lambda}_v$	$\hat{v}_v$	$\lambda_\sigma$	$\bar{\lambda}_\sigma$	$\hat{v}_\sigma$
40.29	0.13	2.05	<b>0.75</b>	0.55	0.72	0.69	0.74	<b>0.50</b>	0.09	0.44	0.37	0.48	0.58	0.48
38.98	0.13	2.50	<b>0.80</b>	0.62	0.77	0.74	0.79	<b>0.60</b>	0.23	0.53	0.48	0.58	0.68	0.58
37.69	0.14	3.07	<b>0.85</b>	0.69	0.81	0.79	0.84	<b>0.70</b>	0.38	0.63	0.59	0.68	0.77	0.68
36.39	0.14	3.79	<b>0.90</b>	0.76	0.86	0.85	0.89	<b>0.80</b>	0.51	0.73	0.69	0.77	0.87	0.77
35.09	0.14	4.72	<b>0.95</b>	0.82	0.91	0.90	0.94	<b>0.90</b>	0.64	0.82	0.79	0.87	0.97	0.87
33.76	0.14	5.92	<b>1.00</b>	0.90	0.96	0.95	0.98	<b>1.00</b>	0.79	0.93	0.89	0.97	0.97	0.97

**Column descriptions:**\* ML MR var. = (average) maximum approximate likelihood estimate assuming variance is mean reverting;\* ML MR vol. = (average) maximum approximate likelihood estimate assuming volatility is mean reverting;\* Regression on var. = (average) regression estimate when using the quadratic variation of variance;\* Regression on vol. = (average) regression estimate when using the quadratic variation of volatility.

We end this section by looking into some of the problems with modeling volatility by a one-dimensional diffusion process. First, the long-memory properties of volatility that have been widely documented; an example is the paper [4]. To have a precise discussion let us introduce some notation: Consider a covariance-stationary process  $X$  and pick an arbitrary time point  $t$ . We then define the autocorrelation function of the  $X$ -process at lag  $h$  as:

$$\text{ACF}(h) = \frac{\text{Cov}(X_t, X_{t+h})}{\text{Var}(X_t)}. \quad (2.11)$$

Several slightly different definitions of long-memory can be found in the literature. As an example, in [4] the autocorrelation function of the log-volatility is assumed to decay as  $\text{ACF}(h) \sim c|h|^{-\beta}$  for  $|h| \rightarrow \infty$ , where  $c, \beta > 0$  are constants. The volatility process is then said to have long-memory exactly if  $\int_0^\infty |\text{ACF}(h)|dh = \infty$ , which happens when  $\beta \in (0, 1)$ . In our context, with daily observations, a reasonable test for long-memory is therefore to see if the sums

$$\sum_{i=0}^k \widehat{\text{ACF}}(i/252), \text{ for } k = 0, 1, 2, \dots \quad (2.12)$$

converge, where  $\widehat{\text{ACF}}(h)$  denotes an empirical estimate of (2.11).

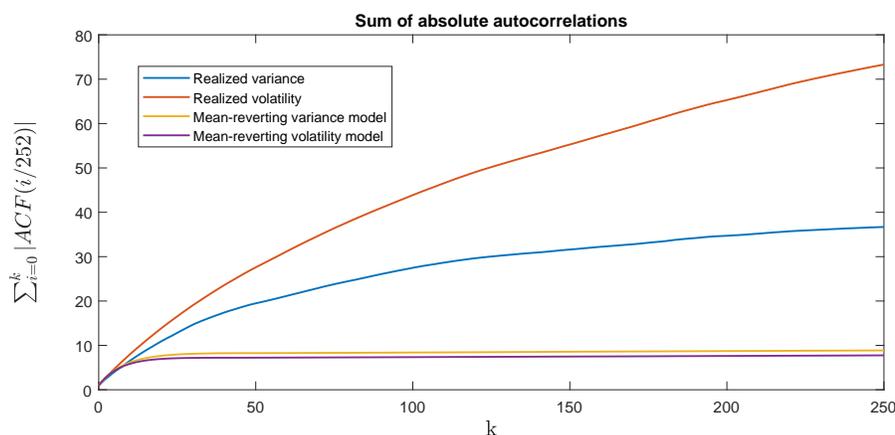
In Figure 2.1, we show the sums in (2.12) for lags up to around 1 year. Specifically, the blue and orange lines show the sums (2.12) for the realized variance and volatility, respectively, both obtained from the Oxford-Man dataset. The lines indicate that there is detectable autocorrelation even at horizons of up to one year (although it is most pronounced looking at realized *volatility*). This can then be compared to what our diffusion model specifications from (2.3) and (2.5) are able to produce. To this end, we conducted a simulation experiment, where we simulated each model in the same way as done for the bias-experiment shown in Table 2.1, though this time using the estimated parameters from Table 2.2. For each such simulated path we then computed the empirical autocorrelation function  $\widehat{\text{ACF}}(h)$  at various lags  $h$ . Repeating the simulations 10,000 times, averaging the autocorrelations and computing the running sums in Equation (2.12), we get the yellow line (simulation of Equation (2.3)) and purple line (simulation of Equation (2.5)). Here it is clear that while the autocorrelations closely match the empirical ones for the first few lags, the fit is very bad at longer horizons—where in fact “longer” does not have to mean more than a few weeks.

To interpret the figure we once again cite [4], where it was assumed that the autocorrelation function of log-volatility behaves as  $1 - \text{ACF}(h) \sim c|h|^{2\alpha+1}$  when  $|h| \rightarrow 0$  and where  $c > 0$  and  $\alpha \in (-1/2, \infty)$  are constants. Volatility is then said to be rough when  $\alpha \in (-1/2, 0)$ . That is, roughness is exactly related to how the autocorrelation function behaves for very short lags. With this knowledge and together with our high  $\kappa$  and  $\eta$  estimates we interpret the good fits for short lags in Figure 2.1 as evidence that our one-factor diffusion models have attempted to mimic roughness. The model-structure then does not allow us to simultaneously capture the memory in the process at medium

CHAPTER 2. HOW DOES THE VOLATILITY OF VOLATILITY DEPEND ON  
VOLATILITY?

**Table 2.2:** Estimated parameters under each of the four estimation methods discussed. Standard errors are shown in parenthesis where applicable.

Method	$\kappa$	$\theta$	$\eta$	$\lambda_v$	$\lambda_\sigma$	$R^2$
Max. approx. likelihood on Equation (2.3)	18.62 (1.07)	0.16 <sup>2</sup>	7.37 (0.27)	0.80 (0.01)	0.60 (0.02)	
Max. approx. likelihood on Equation (2.5)	36.01 (2.65)	0.14 (0.01)	4.04 (0.16)	0.91 (0.01)	0.83 (0.02)	
Regression on Equation (2.9)			3.92 (0.27)	0.93 (0.02)	0.85 (0.03)	0.42
Regression on Equation (2.10)			2.35 (0.16)	0.97 (0.01)	0.94 (0.03)	0.17



**Figure 2.1:** Cumulative sums of the empirical (absolute-value) autocorrelations. Blue and orange lines show the sums for the realized variance and volatility and that using the Oxford-Man dataset, respectively. The yellow and purple lines show the sums for the model (2.3) and (2.5), respectively. These values are computed in a simulation experiment using the parameters from Table 2.2. For these curves, the standard errors of each autocorrelation value were all less than 0.01.

and long lags.

An extension that would allow for both roughness and long-memory would be to model  $V_t$  (alternatively  $\sigma_t$ ) as a mean-reverting stochastic Volterra equation (SVE).<sup>9</sup> Specifically, one could model

$$V_t = V_0 + \int_0^t K(t-s)\kappa(\theta - V_s) ds + \int_0^t K(t-s)\eta V_s^{\lambda_v} dW_{2,s}, \quad (2.13)$$

where  $K$  is a kernel function that if chosen appropriately and exactly would allow  $V_t$  to display both of these properties. Although an estimation of the more general Equa-

<sup>9</sup> The reader can consult [1] for some mathematical theory on stochastic Volterra equations.

## 2.3. EVIDENCE FROM OPTION PRICES

---

tion (2.13) would certainly be a worthwhile pursuit, we leave it as an open hypothesis whether or not this will change the  $\lambda_v$ -estimates. Two other examples—among several—of models that can capture long-memory are [25] and [5].

Another objection to our models is the lack of jumps in the asset price. However, as shown in [8] using high-frequency data on numerous different financial assets, the jump proportion of the total variation of the asset price is generally small compared to the volatility component—at least when market micro-structure effects are properly accounted for. It is therefore very possible that adding jumps to the asset price will only slightly change the  $\lambda_v$ -estimates. On the other hand, one could also consider jumps in the volatility process. Then again, with the current evidence in favor of rough volatility one could hypothesize that the most variation in the true volatility process can also be explained by continuous but rough (i.e., explosive) movements and not jumps. As with the long-memory objections, we leave a proper analysis for future research.

## 2.3 Evidence from option prices

We now turn our attention to options, more specifically to the market for European call and put options on the S&P 500 index. In an ideal world, one would specify flexible parametric structures for the drift function in Equation (2.2) and for the market price of volatility risk, solve the associated pricing PDE, and then choose the parameter-values that minimize the distance—suitably measured—between market data and model prices. However, genuinely efficient methods for option price calculations exist only for a few quite specific models. We shall restrict our interest to the two most common models, the square-root model from [14] that sparked a revolution in affine models and transform methods, and the SABR (stochastic-alpha-beta-rho) model from [19] that made stochastic volatility a household object in banks.

We used the *End-of-Day Options Quote Data* obtained from <https://datashop.cboe.com>. This dataset contains bid and ask quotes on SPX European options at 15:45 Eastern Time (ET) and again at the close of the market. We used the 15:45 quotes for liquidity reasons and applied a number of filters to (a) clean the data and (b) compute the zero coupon bond and dividend yields implied by the put-call parity (and used in the subsequent analysis). Finally, we used smoothing and interpolation techniques on the mid quotes to obtain prices on a continuous set of strikes on each of the fixed set of expiries 1, 3, 6, 12, 18, and 30 months.<sup>10</sup> The cleaned dataset contains observations on 3783 trading days between the 3rd of May 2004 and the 15th of May 2019.

We will now briefly define the pricing models we use in our experiments. The models are therefore stated under an equivalent risk-neutral probability measure  $\mathbb{Q}$ . Note that Girsanov’s theorem tells us that the elasticity parameter is unaffected by equivalent measure changes. Thus, there is no a priori conflict between  $\lambda_v$ -estimates obtained from time series observations of  $V_t$  (“under  $\mathbb{P}$ ”) from the previous section and risk-neutral values affecting option prices that we shall be looking at in this section. For simplicity we will abuse notation and write  $W_{1,t}$  and  $W_{2,t}$  to also denote two  $\mathbb{Q}$ -Brownian motions s.t.

<sup>10</sup> We used the arbitrage free smoothing spline from [11] on expiry-slice.

$\langle dW_{1,t}, dW_{2,t} \rangle = \rho dt$  for a  $\rho \in [-1, 1]$ . Denoting the interest rate by  $r$  and the dividend yield by  $q$  we then have

$$dS_t = (r - q)S_t dt + \sqrt{V_t} S_t dW_{1,t}. \quad (2.14)$$

Under the Heston model of [14], the variance process  $(V_t)_{t \geq 0}$  is then modeled as a square-root diffusion of the form

$$dV_t = \kappa(\bar{v} - V_t)dt + \eta\sqrt{V_t}dW_{2,t} \quad (2.15)$$

with  $\kappa, \bar{v}, \eta, V_0 > 0$ . The model thus has elasticity  $\lambda_v = \frac{1}{2}$ , which (Feller condition issues notwithstanding) is equivalent to  $\lambda_\sigma = 0$ . We compute option prices using numerical integration with the techniques from [17]. The SABR model from [19] (here with their  $\beta = 1$ ) instead assumes

$$d\sigma_t = \alpha\sigma_t dW_{2,t} \quad (2.16)$$

with  $\alpha, \sigma_0 > 0$ . To compute option prices we use here the approximation formula also found in [19]. In elasticity-terms, the log-normal SABR model has  $\lambda_v = \lambda_\sigma = 1$ . Finally, we consider a third hybrid or mixture model, where we assume that the price of any vanilla European option is the average of the prices under both the Heston and SABR models. Specifying such a mixture model is exactly equivalent to specifying the marginal distributions of  $S_t$  for all  $t > 0$  but nothing more. The mixture model allows us to incorporate a marginal distribution for  $S_t$  somewhere between Heston and SABR. The model is *underspecified*, as it does not contain information on the dynamical structure of the asset price  $S_t$ ; see the paper [20] on mixture models and their pitfalls. However, because vanilla option prices under the mixture model are convex combinations of prices from the arbitrage-free models Heston and SABR, the mixture model is free from static arbitrages; it could thus at the very least be supported by a local volatility model for  $S$ .<sup>11</sup>

As explained, our choice of pricing models has been restricted to those that are tractable. As a downside, each model has its own problems in terms of matching the empirical stylized facts of volatility. Some of these problems are related to the topic of this paper, i.e., what should the elasticity of variance-of-variance be. Specifically, in Section 2.1 we estimated  $\lambda_v \approx 1$  while Heston has  $\lambda_v = 1/2$ . Likewise, in [24] it is shown that while a log-normal model fits the marginal distribution of volatility well, Heston does not. Given our topic, having differences in  $\lambda_v$  is of course exactly as desired. In an ideal controlled experiment this would be the only difference between the models. This is not quite true. The Heston model has mean-reversion and SABR does not. Furthermore, any inability to match empirical facts, even if shared by both models, could make other parts of the models and the results move in unexpected and hard-to-explain ways. As an example, both models suffer from the inability to match the entire term structure of autocorrelations (as shown in Section 2.1). With all of this taken together, it would be naive to think that we can reasonably fit and/or accurately model multiple expiries at

<sup>11</sup> [16] presents a fully specified mean-reverting, lognormal-ish volatility model that allows for explicit call option prices. However, in the author's own words, "The bad news is that the integrand requires the infinite sum at (37). I truncate that sum, and make some use of Mathematica's `Parallelize`. This case is really computationally tedious: results can take a half-hour or more." Hence the model is not suitable for our analysis.

once. In an attempt to mend some of these problems we therefore chose to model each expiration-slice with a separate model.<sup>12</sup> The hope is that this will help the models capture the temporal properties that are the most important to modeling the particular expiration in question. Ideally this would also mean that the models approximately only differ by their elasticity. The reality is of course not so simple. We will make our best attempt at analyzing the results despite these a-priori objections.

### 2.3.1 Calibration and in-sample model performance

In this subsection we perform a calibration experiment, where we calibrate the models separately on each expiry. We perform the calibration on a given expiry by minimizing the mean absolute error in Black–Scholes implied volatilities across a number of observed contracts,  $n$ . Let  $\hat{\sigma}_i^{obs}$  denote the implied volatility of the  $i$ 'th contract as observed on the market, and let  $\hat{\sigma}_i^{model}$  denote the corresponding implied volatility for the model in question and that for a particular choice of parameters. We thus calibrate each model by minimizing

$$\frac{1}{n} \sum_{i=1}^n |\hat{\sigma}_i^{obs} - \hat{\sigma}_i^{model}| \quad (2.17)$$

with respect to the model parameters. On each day and for each expiry we aim to calibrate to 7 options: those with Black–Scholes call Deltas 0.05, 0.2, 0.3, 0.5, 0.7, 0.8, and 0.95.<sup>13</sup> When calibrating the Heston model to a single expiry-slice it is difficult to separate the speed of mean-reversion parameter  $\kappa$  from the volatility-of-volatility parameter  $\eta$ , as well as separating the current instantaneous variance  $V(t)$  from its long-term mean  $\bar{v}$ ; see also [13]. Hence we use the following two-step procedure: first, optimize the fit to all implied volatilities over  $(\kappa, \eta, \rho, V(t), \bar{v})$ ; second, keep the  $\kappa$ -estimate from the first step fixed and then optimize each expiry-slice separately and under the constraint  $V(t) = \bar{v}$ .

In the left-hand side of Table 2.3 we show average calibration errors in basis points (bps) as well as their standard deviations (shown in parentheses). The main observation is that on average, SABR calibrates better for short expiries (less than 6 months) and Heston calibrates better for long expiries (more than one year). Specifically, the average error for the 1 month expiry is 28.7 bps for Heston and 11.2 bps for SABR, whereas for the 2.5 years' expiry, the average error is 14.2 bps for Heston and 23.6 bps for SABR.<sup>14</sup> To understand this result, remember (see [23]) that call option prices are determined by the (risk-neutral) distribution of instantaneous variance integrated over the life-time of

<sup>12</sup> The  $r$  and  $q$  values of Equation (2.14) will therefore be the interest rate yield and dividend yield for the particular expiration in question.

<sup>13</sup> To be precise: on a given day and for a given expiry we first attempt to interpolate the values 0.3, 0.5, and 0.7 (if that is not possible we discard that expiry on that day). Next we attempt to choose Black–Scholes deltas closest to the remaining shown values on each of the intervals  $[0.05, 0.15]$ ,  $[0.15, 0.30]$ ,  $[0.7, 0.85]$ , and  $[0.85, 0.95]$ . Depending on the available range of strikes we thus in practice (on a smaller number of days) calibrate to fewer than seven quotes.

<sup>14</sup> For comparison, in our dataset (and before interpolation and smoothing) the medians (across days) of the smallest bid–ask spreads within  $\pm 10\%$  of at-the-money and for each of the expiration groups 0–3 months, 3–12 months, and 12–36 months are 17, 32, and 47 bps of implied volatility, respectively.

CHAPTER 2. HOW DOES THE VOLATILITY OF VOLATILITY DEPEND ON  
VOLATILITY?

**Table 2.3:** The left part of the table shows average calibration errors in basis points. Standard deviations are shown in parentheses. The right part of the table shows the average of the per-day differences in calibration errors between each pair of models. I.e., the column "Heston vs. SABR" shows statistics on the time series constructed, where on each day we subtract the Heston error from the SABR (stochastic-alpha-beta-rho) error.

Expiry	Heston	SABR	Mixture	Heston	Heston	SABR
				vs. SABR	vs. Mixture	vs. Mixture
1 month	28.7	11.2	18.4	17.4	10.3	-7.1
	(14.8)	(9.9)	(10.3)	(15.2)	(7.3)	(8.4)
3 months	22.7	9.1	13.1	13.6	9.5	-4.1
	(10.1)	(7.7)	(5.7)	(14.6)	(7.3)	(8.5)
6 months	14.8	11.8	9.9	2.9	4.8	1.9
	(8.9)	(8.9)	(6.0)	(14.4)	(8.0)	(7.6)
1 year	9.5	17.7	11.2	-8.2	-1.7	6.4
	(7.3)	(9.9)	(8.7)	(8.5)	(6.0)	(4.4)
1.5 years	10.3	20.9	14.7	-10.6	-4.5	6.2
	(9.2)	(10.7)	(10.0)	(7.1)	(4.3)	(3.5)
2.5 years	14.2	23.6	18.4	-9.4	-4.3	5.2
	(10.0)	(12.1)	(10.9)	(7.5)	(4.4)	(3.5)

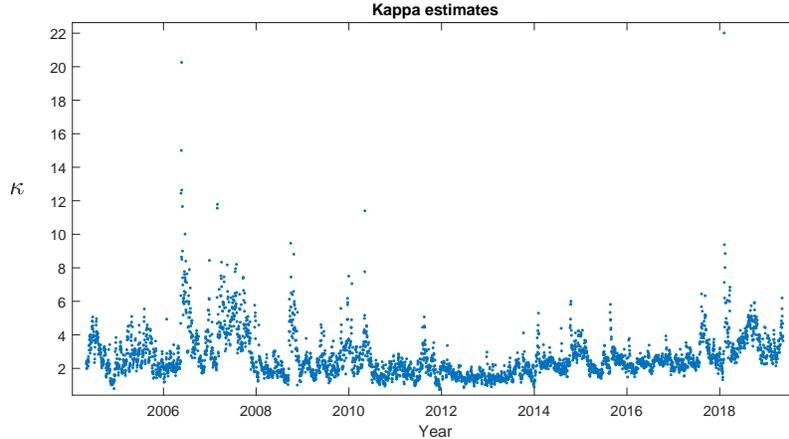
the option ( $\tau$ ),

$$\frac{1}{\tau} \int_t^{t+\tau} \sigma_u^2 du. \quad (2.18)$$

For short times to expiry (small  $\tau$ ), the dominant feature in determining this distribution is the  $dW$ -term in the dynamics of instantaneous volatility, while for longer times to expiry temporal dependence (such as mean-reversion) becomes more important. The empirical analysis in Section 2.2 shows that the SABR model (with  $\lambda_v = 1$ ) captures the functional form of the variance of instantaneous variance better than the Heston model (that has  $\lambda_v = \frac{1}{2}$ ). On the other hand, that analysis also shows that instantaneous variance has a quite significant mean-reversion, which is something that the Heston model captures but the SABR model does not. As Figure 2.2 shows, there is, however, a subtlety to this; the risk-neutral ( $\mathbb{Q}$ ) speed of mean-reversion is considerably lower (2–4 typically) than the real-world ( $\mathbb{P}$ ) estimates (18.62 from Table 2.2 is directly comparable).<sup>15</sup> It is widely documented in the literature that the short-expiry at-the-money implied volatility is typically higher than realized volatility, i.e., long-term levels are different between  $\mathbb{P}$  and  $\mathbb{Q}$  ( $\bar{v} > \theta$ ), which can be explained by investor risk-aversion as a stochastic volatility model is, in option pricing terms, incomplete. The effect of investor preferences on the speed of mean-reversion is however less well documented. With that mean-reversion vs. elasticity reasoning in mind, a natural conjecture would be that a mixture model, a convex combination of SABR and Heston, would outperform both these models; that there would be a benefit from model diversification. However, such an effect is far from

<sup>15</sup> [7] show that for square-root processes measure changes that change both the long-term level and the speed of mean-reversion are allowed.

### 2.3.1. CALIBRATION AND IN-SAMPLE MODEL PERFORMANCE



**Figure 2.2:** Estimates of the Heston model’s  $\kappa$  under the risk-neutral measure  $\mathbb{Q}$  obtained from the two-step calibration procedure.

evident in the data. With the 6-month horizon as the only exception, the errors from the mixture model fall between those of SABR and Heston—often right in the middle.

Average errors do not tell the whole story; the variation (between days) of calibration errors (as measured by their standard deviations) are of the same order of magnitude as the averages. In the right-hand part of Table 1, we give paired comparisons of the models. On each day (and for all three combinations) we subtract one model’s calibration error from another model’s and report averages (that could be calculated from the left-hand part of the table) and standard deviations (that cannot). Let us first note that this shows that the differences in average that we comment on are all highly statistically significant (thus we have not cluttered the table with indications of this). However, we also note that the differences display a variation that is quite similar to the individual calibration errors themselves. Hence when the average Heston error is 14.2 bps for 2.5-year options and it is 23.6 bps for SABR, it by no means implies that Heston calibrates around 10 bps better each and every day, as the standard deviation of the difference is 7.5 bps.

As a final in-sample investigation, let us look at the time evolution of the calibration errors and the model parameters. In Figures 2.3–2.5 we therefore show such results for, respectively, 1 month, 6 months, and 2.5 years expiries. The calendar and expiry time variation that one would expect from the results in Table 2.3 is evident. Around the 2008–2009 financial crisis we see deterioration in model performance in various guises; Heston calibration errors for 1 month options more than double, and the model never really recovers; the crisis leads to extreme (negative) correlation for SABR. But overall, nothing off-the-scale happened. One thing that stands out visually is how high the Heston’s variance-of-variance parameter  $\eta$  (left-hand side columns, 2nd panel) correlates with the level of volatility (left-hand columns, 1st panel); the average (across expiries) correlation is 0.53. This is consistent with our previous elasticity estimation: The Heston model’s elasticity of  $\frac{1}{2}$  does not allow the variance of variance to react as strongly

to changes in the variance level as empirically observed in time-series ( $\lambda_v$  around 0.91–0.97), and that manifests itself as changes in the  $\eta$ -estimate. For SABR, the (volatility, vol-of-vol) correlation is mildly negative (–0.28 on average), which is also consistent with that model slightly overstating the elasticity. A final note to make (based on the top left panel in Figure 2.3) is that even though implied, at-the-money volatility goes to instantaneous volatility  $\sigma_t$  as time to expiry goes to 0, a one month expiry is not sufficiently small for this asymptotic result to have kicked in.

### 2.3.2 Predictions and hedging; Out-of-sample performance

Figures 2.3–2.5 show that calibrated model parameters change both over calendar time and across expiries. A cynic would say, "Therefore they are not *parameters*—which is a crucial assumption in analytical work with the models, e.g., derivation of option price formulas. So: back to the drawing-board". Our defense against that argument is pragmatism as formulated in the famous quote from statistician George Box that "all models are wrong, some are useful". Thus we now investigate the practical usefulness of the SABR and Heston models. More specifically, we look at the quality of predictions and at how helpful they are for constructing hedge portfolios; two aspects that are central to financial risk management. Both investigations are done in out-of-sample fashion; the predictions or portfolios made at time  $t$  use only information that is available at time  $t$ .

To study prediction quality, we conducted the following experiment: on each trading day we calibrated the parameters and volatility. We then moved forward in time, updated observable market variables, such as the index price as well as yields, and recalibrated the volatility (which is allowed to change in the model)—but *not* the parameters. We performed the experiment by moving 1 to 20 trading days ahead (the horizon) and considering all possible starting dates. Figure 2.6 shows the results, and more specifically average absolute errors at different horizons and for different expiries for the SABR and Heston models.<sup>16</sup> The results are consistent with what we have observed so far. SABR works better for short expiries (where errors are generally larger), while Heston does better for longer expiries. Average absolute errors are well fitted by square-root functions (which is how we would a priori expect standard deviation-like quantities to grow with time), and the difference between the models is stable across horizons.

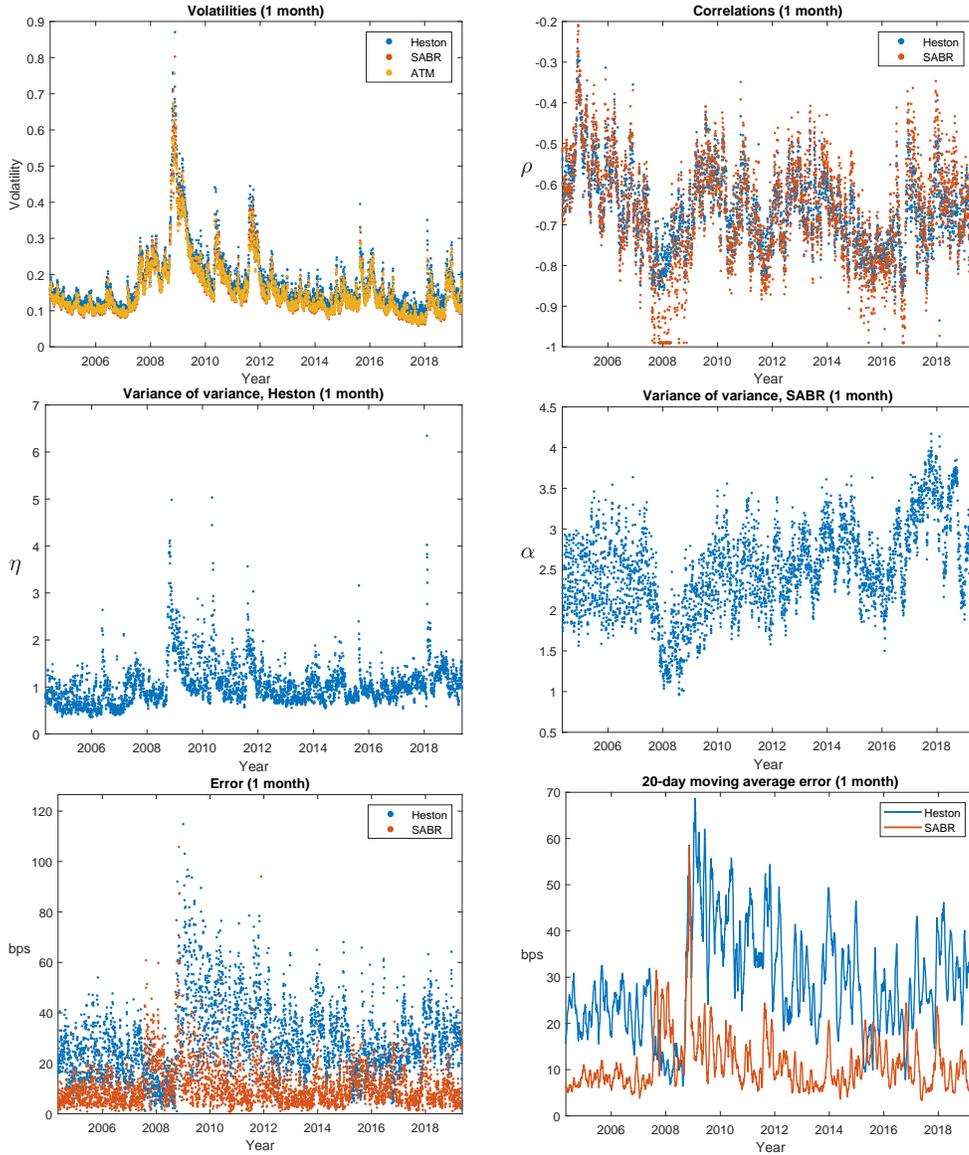
Finally, we turn to the hedge performance of the models. We will attempt to hedge out-of-the-money options by trading appropriately in the underlying asset, the risk-free asset as well as an at-the-money option of the same type as the out-of-the-money option (i.e., call or put). We performed the analysis on out-of-the-money options with strikes corresponding to Black–Scholes Delta values of 0.1, 0.3, 0.7, and 0.9. Thus, if the Delta is below 0.5 we hedged a call option and otherwise hedged a put option.

The details of the experiment are as follows: on each day we sell the out-of-the-money option and form a portfolio that according to the calibrated model perfectly hedges this option. Let us write  $h_t := (h_t^b, h_t^s, h_t^a, h_t^o)$  to denote the entire portfolio time  $t$ , where  $h_t^b$

---

<sup>16</sup> Since the mixture model is underspecified and therefore does not imply a specific dynamic structure, we cannot move the model forward in time without further assumptions. We therefore excluded it from this experiment.

### 2.3.2. PREDICTIONS AND HEDGING; OUT-OF-SAMPLE PERFORMANCE



**Figure 2.3:** Calibration to the 1 month expiry. To improve visibility, we have in the bottom left plot excluded the extreme errors on the 21st of November 2008 of 260 bps (Heston) and 352 bps (SABR).

CHAPTER 2. HOW DOES THE VOLATILITY OF VOLATILITY DEPEND ON VOLATILITY?

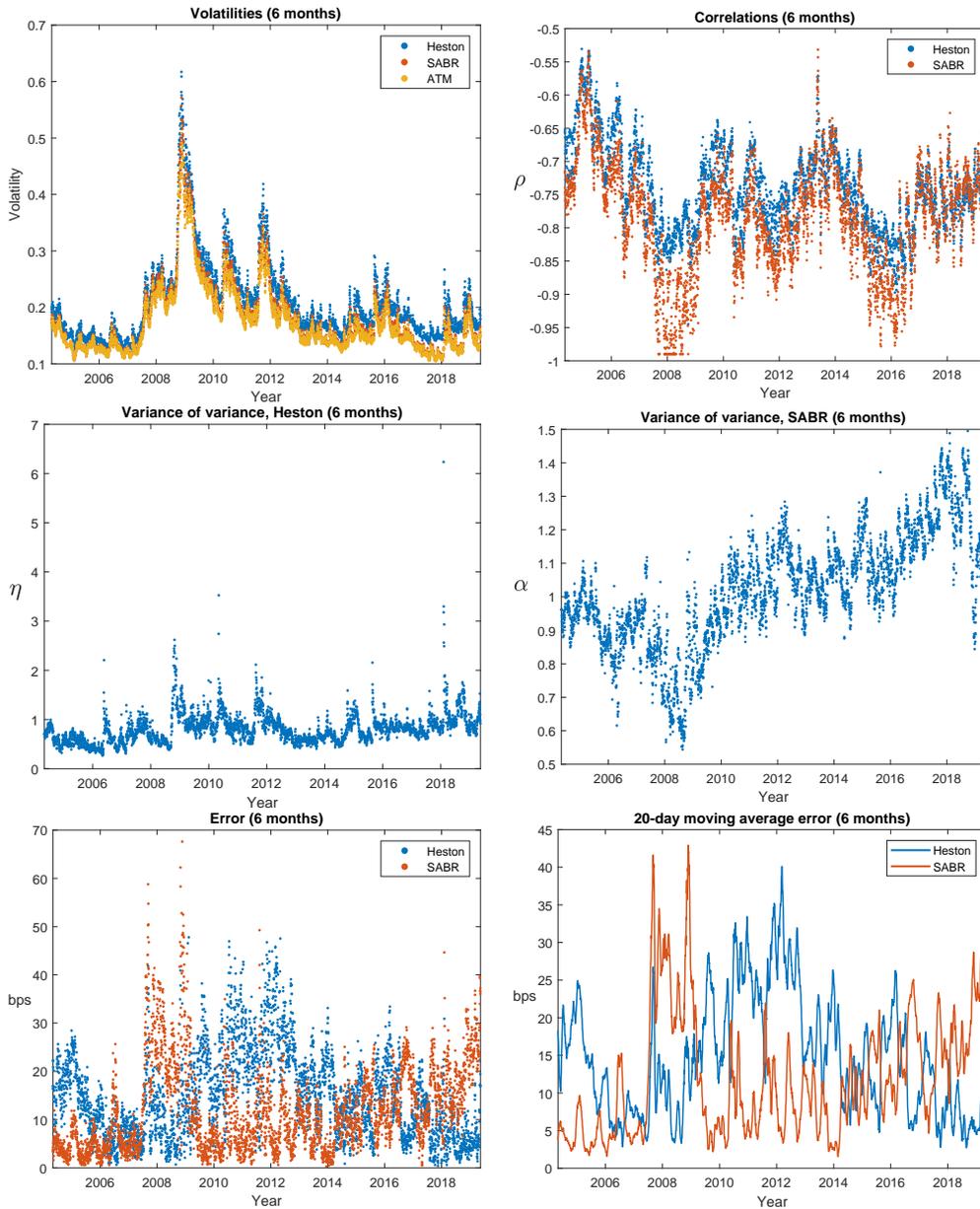


Figure 2.4: Calibration to the 6 month expiry.

### 2.3.2. PREDICTIONS AND HEDGING; OUT-OF-SAMPLE PERFORMANCE

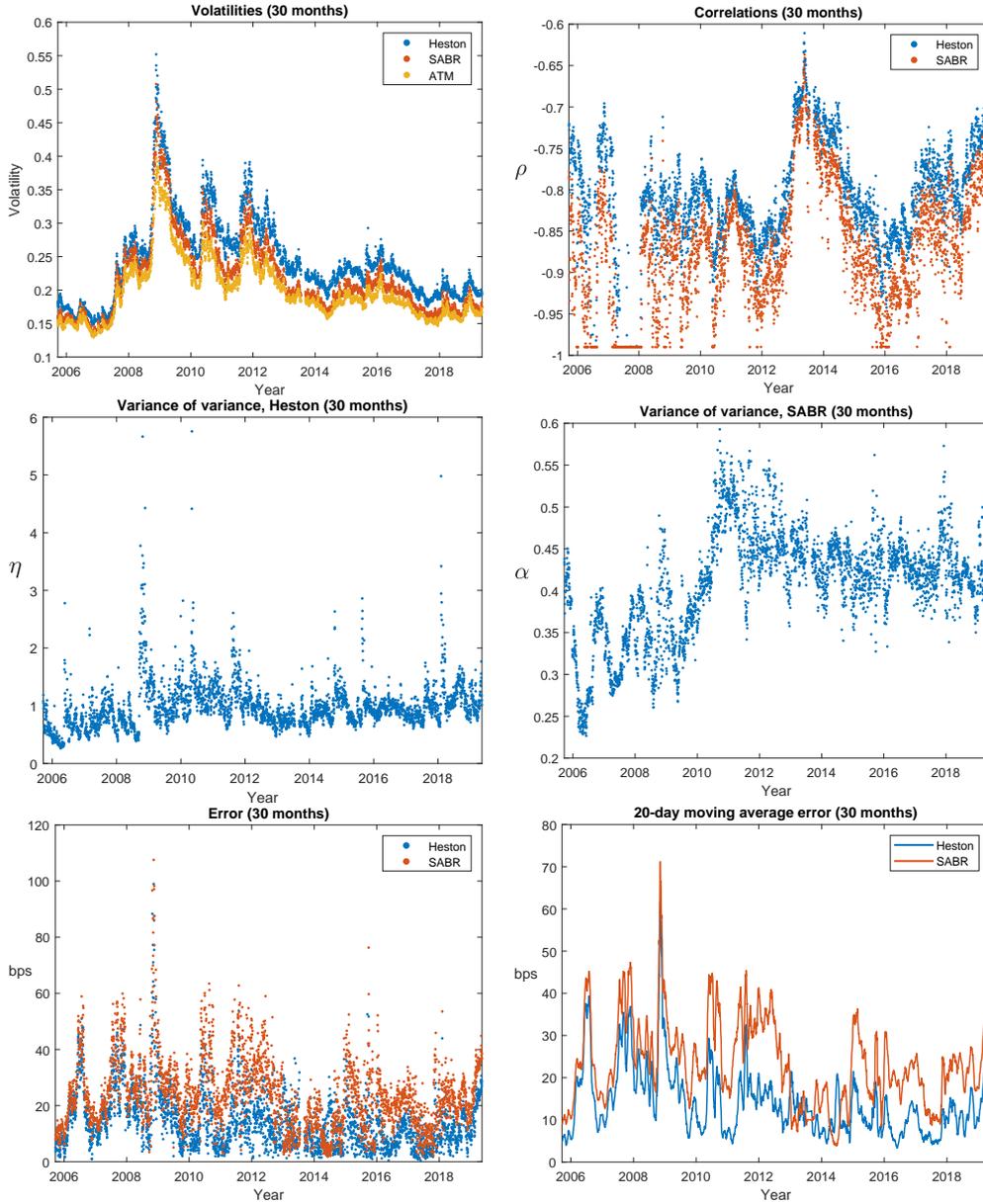


Figure 2.5: Calibration to the 30 months (2.5 years) expiry.

denotes the number of units in the risk-free asset,  $h_t^s$  the number of units in the underlying asset,  $h_t^a$  the number of units in the at-the-money option, and  $h_t^o$  the number of units in the out-of-the-money option. We set  $h_t^o = -1$  as mentioned. Let us also write  $\pi_t^o$  and  $\pi_t^a$  to denote the observed values of the out- and at-the-money option, respectively.

For compactness of presentation, let us consider a general hedge model

$$dS_t = S_t(r - q)dt + S_t\sqrt{V_t}dW_{1,t} \quad (2.19)$$

$$dV_t = a(t, S_t, V_t)dt + b(t, S_t, V_t)\eta\sqrt{V_t}dW_{2,t} \quad (2.20)$$

where  $\langle dW_{1,t}, dW_{2,t} \rangle = \rho dt$  and where  $a$  and  $b$  are functions. We get Heston with  $a(t, s, v) = \kappa(\bar{v} - v)$ ,  $b(t, s, v) = 1$ , and SABR with  $a(t, s, v) = \alpha^2 v$ ,  $b(t, s, v) = \sqrt{v}$  and  $\eta = 2\alpha$ . We can write the price of the out-of-the-money option at time  $t$  as  $F(t, S_t, V_t; \theta)$  for an appropriate function  $F$  and a parameter vector  $\theta$  depending on model specifics;  $\theta := (\eta, \rho, \bar{v}, r, q)$  for Heston and  $\theta := (\alpha, \rho, r, q)$  for SABR. Similarly, the price of the at-the-money option can be written as  $G(t, S_t, V_t; \theta)$  for an appropriate function  $G$ . To determine the perfectly replicating portfolio under the hedge model we start by computing our models *Delta* and *Vega*, which for the out-of-the-money option with function  $F(t, s, v; \theta)$  will be defined as  $F^s := \frac{\partial F}{\partial s}$  and  $F^v := \frac{\partial F}{\partial v}$ , respectively.<sup>17</sup> Our hedge will now consist of

$$h_t^a = \frac{F_t^v}{G_t^v} \quad (2.21)$$

in the at-the-money option, and

$$h_t^s = F_t^s - h_t^a \cdot G_t^s \quad (2.22)$$

in the underlying asset. The joint portfolio will then be kept self-financing with the risk-free asset. With these choices the associated value process will exactly be a function of the current state  $(t, S_t, V_t)$ . Computing the dynamics of it using Ito's Formula and applying the principle of no arbitrage proves that  $h_t$  is in fact a perfect hedge of the out-of-the-money option—assuming we are using the correct model. For the mixture model the whole problem of finding a perfectly replicating portfolio is ill-defined, since the dynamical structure is unspecified. We therefore instead test a mixed *portfolio* that is the average of the portfolios under Heston and SABR. While this portfolio may not perfectly replicate the option it will be a valid self-financing strategy, which we can compare to the pure Heston and SABR strategies.

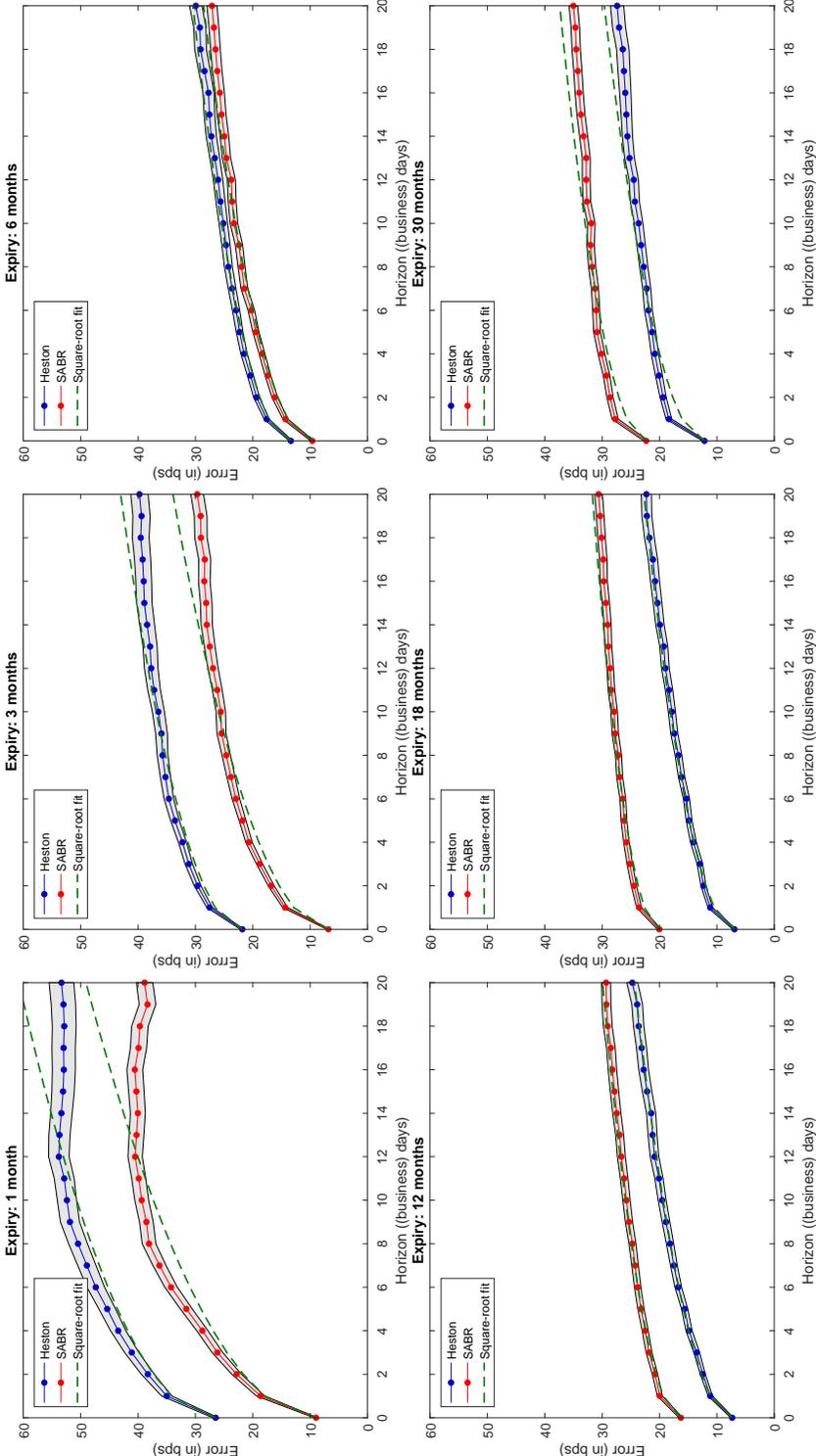
Consider now discrete hedging between two trading days  $t_i$  and  $t_{i+1}$  with  $\Delta t = t_{i+1} - t_i = 1/252$  assumed for simplicity. Letting  $V_t^h$  denote the actual value of the portfolio at time  $t$  we record (with a few discrete approximations) the change in the value process from  $t_i$  to  $t_{i+1}$  as

$$\Delta V_{t_i}^h = h_{t_i}^b B_{t_i} r_{t_i} \Delta t + h_{t_i}^s \Delta S_{t_i} + h_{t_i}^a \Delta \pi_{t_i}^a + h_{t_i}^o \Delta \pi_{t_i}^o + h_{t_i}^s S_{t_i} q_{t_i} \Delta t. \quad (2.23)$$

---

<sup>17</sup> As done here, we will often suppress the input arguments to simplify the notation. We will also write  $F_t^s := \frac{\partial}{\partial s} F(t, S_t, V_t; \theta)$ , etc. when we need to stress the time point used. We will use similar notation to denote other derivatives of  $F$  and do all of this also for the function  $G(t, s, v; \theta)$ .

2.3.2. PREDICTIONS AND HEDGING; OUT-OF-SAMPLE PERFORMANCE



**Figure 2.6:** Model prediction quality. The graphs show average absolute errors at different horizons (x-axis) and for different expiries (panels). Blue is Heston and Red is SABR. The shaded areas are 95% confidence bands, the dotted green curves are the best-fitting  $\sqrt{x}$ -functions.

CHAPTER 2. HOW DOES THE VOLATILITY OF VOLATILITY DEPEND ON  
VOLATILITY?

where  $B_t$  is the value of the risk-free asset. We then define the (relative and discounted) hedge error as

$$\text{Hedge error (relative and discounted)} = 100 \times \frac{e^{-r_{t_i} \Delta t} \Delta V_{t_i}^h}{\pi_{t_i}^o}. \quad (2.24)$$

and summarize the performance of each model by taking the standard deviation of this across all trading days.

**Table 2.4:** Hedge errors (in the scaled standard deviation sense of Equation (2.24)) from daily hedging for each expiry, model and moneyness (measured by the Black–Scholes call Delta  $\Delta_{BS}$ ). The symbols ‡ and † indicate significance at the 1% and 5% levels for testing if the standard deviation of Heston and the mixed portfolio, respectively, are different from SABR. The symbols therefore only appear for rows related to Heston and the mixed portfolio. Significance is tested using the Brown–Forsythe test of [6] with the central locations estimated by the medians.

Standard Deviation of Hedge Error (Daily Hedging)					
Expiry	Model	$\Delta_{BS} = 0.9$	$\Delta_{BS} = 0.7$	$\Delta_{BS} = 0.3$	$\Delta_{BS} = 0.1$
1 month	Heston	9.8	3.2	6.2 ‡	16.6 ‡
	SABR	9.3	3.2	4.3	14.7
	Mixed portfolio	9.3	3.2	5.1 ‡	15.2
3 months	Heston	5.3	1.4	2.5 ‡	9.5 ‡
	SABR	5.2	1.4	1.8	8.8
	Mixed portfolio	5.2	1.4	2.1 ‡	8.9
6 months	Heston	3.7	1.0	1.4 ‡	5.7
	SABR	3.7	1.0	1.2	5.5
	Mixed portfolio	3.7	1.0	1.3	5.4
1 year	Heston	3.1	0.8	1.0	4.5 ‡
	SABR	3.1	0.8	1.0	4.8
	Mixed portfolio	3.1	0.8	1.0	4.6 †
1.5 years	Heston	2.5	0.6	0.9 ‡	3.5 ‡
	SABR	2.5	0.7	1.0	3.9
	Mixed portfolio	2.5	0.7	0.9	3.6 ‡
2.5 years	Heston	3.7	0.8	1.1	4.4 ‡
	SABR	3.8	0.8	1.2	4.7
	Mixed portfolio	3.7	0.8	1.1	4.5

In Table 2.4 we show the results across expiries, moneyness, and models. We first note that the more out-of-the-money the target option is, the more difficult it is to hedge; not surprising, particularly because the at-the-money option is one of the hedge instruments. Short-expiry options are more difficult to hedge; above 6 months expiry standard deviations are quite stable. For low-strike options (i.e., out-of-the-money puts) the differences between models are small: at most 0.5 percentage (the 1 month, lowest-strike case), and none of the differences are statistically significant at the 5%-level. But there is an asymmetry. For high strikes (i.e., for out-of-the-money calls) differences in hedge errors are statistically significant across models; SABR outperforms Heston for expiries of three months and below, but is beaten by Heston for expiries above one year, albeit

## 2.4. CONCLUSIONS

---

with a lower absolute margin. We also see that except in a single, statistically insignificant case (highest strike, 6 months expiry) the mixture strategy is dominated in terms of hedge performance by either Heston or SABR, i.e., there are no benefits from model "averaging" or "diversification". Results (not reported) are qualitatively similar in a hedge experiment where Vega-hedging is done only weekly.

## 2.4 Conclusions

The one-word answer to the question in the title would be "lognormal-ish". We find that the dynamics of the instantaneous variance of the S&P 500 index is best described by a model with elasticity slightly below one, i.e., a model with a close-to lognormal volatility structure. However, one-factor diffusion models fail to capture the empirical auto-correlation structure of instantaneous volatility at horizons above a few weeks; long-memory even in the rather short run, as it were. For option pricing, the lognormal SABR model performs best for short-expiry options (expiries of 1–3 months) and the Heston model performs best for expiries of one year and longer. We ascribe the latter effect to the Heston model's mean-reversion—which one should treat with care as its force is quite different under  $\mathbb{P}$  and  $\mathbb{Q}$ . A simple mixture of the two models does not yield benefits. These results are robust across time (15 years of option data) and both in-sample and out-of-sample, including predictions and hedging.

# References

- [1] Abi Jaber, E., Larsson, M., Pulido, S., Affine Volterra processes. *Annals of Applied Probability*, 2019, **29**(5), 3155–3200.
- [2] Andersen, L. B. G., Brotherton-Ratcliffe, R., Extended LIBOR market models with stochastic volatility. *Journal of Computational Finance*, 2005, **9**(1), 1–40.
- [3] Andersen, L. B. G., Piterbarg, V. V., Moment explosions in stochastic volatility models. *Finance and Stochastics*, 2007, **11**, 29–50.
- [4] Bennedsen, M., Lunde, A., Pakkanen, M. S., Decoupling the short- and long-term behavior of stochastic volatility (2016). Working paper. Available online at SSRN:2846756 (accessed 2 June 2020).
- [5] Borland, L., Bouchaud, J.-P., On a multi-scale statistical feedback model for volatility fluctuations. *Journal of Investment Strategies*, 2011, **1**(1), 65–104.
- [6] Brown, M. B., Forsythe, A. B., Robust tests for the equality of variances. *Journal of the American Statistical Association*, 1974, **69**(346), 364–367.
- [7] Cheridito, P., Filipović, D., Kimmel, R. L., Market price of risk specifications for affine models: Theory and evidence. *Journal of Financial Economics*, 2007, **83**(1), 123–170.
- [8] Christensen, K., Oomen, R. C. A., Podolskij, M., Fact or friction: Jumps at ultra high frequency. *Journal of Financial Economics*, 2014, **114**(3), 576–599.
- [9] Christensen, K., Thyrgaard, M., Veliyev, B., The realized empirical distribution function of stochastic variance with application to goodness-of-fit testing. *Journal of Econometrics*, 2019, **212**(2), 556–583.
- [10] Felpel, M., Kienitz, J., McWalter, T. A., Effective stochastic volatility: Applications to ZABR-type models (2020). Working paper. Available online at SSRN:3518141 (accessed on 2 June 2020).
- [11] Fengler, M. R., Arbitrage-free smoothing of the implied volatility surface. *Quantitative Finance*, 2009, **9**(4), 417–428.
- [12] Gatheral, J., Jaisson, T., Rosenbaum, M., Volatility is rough. *Quantitative Finance*, 2018, **18**(6), 933–949.

## REFERENCES

---

- [13] Guillaume, F., Schoutens, W., Use a reduced Heston or reduce the use of Heston? *Wilmott Journal*, 2010, **2**(4), 171–192.
- [14] Heston, S. L., A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 1993, **6**(2), 327–343.
- [15] Le Gall, J.-F., *Brownian Motion, Martingales, and Stochastic Calculus*, Springer, Berlin, 2016.
- [16] Lewis, A. L., Exact solutions for a GBM-type stochastic volatility model having a stationary distribution. *Wilmott Magazine*, May 2019, pp. 20–41.
- [17] Lord, R., Kahl, C., Optimal Fourier inversion in semi-analytical option pricing. Tinbergen Institute Discussion Paper No. 2006-066/2. Available online at SSRN:921336 (accessed 2 June 2020).
- [18] Lord, R., Koekkoek, R., van Dijk, D., A comparison of biased simulation schemes for stochastic volatility models. *Quantitative Finance*, 2010, **10**(2), 177–194.
- [19] Hagan, P. S., Kumar, D., Lesniewski, A. S., Woodward, D. E., Managing smile risk. *Wilmott Magazine*, January 2002, pp. 84–108.
- [20] Piterbarg, V. V., Mixture of models: A simple recipe for a ... hangover? (2003). Working paper. Available online at SSRN:393060 (accessed 2 June 2020).
- [21] Poulsen, R., Elasticity of variance of variance. *Wilmott Magazine*, January 2020, pp. 30–32.
- [22] Rebonato, R. What does today’s smile imply about future volatilities? *Journal of Derivatives*, 2020, **27**(2), 26–44.
- [23] Romano, M., Touzi, N., Contingent claims and market completeness in a stochastic volatility model. *Mathematical Finance*, 1997, **7**(4), 399–412.
- [24] Tegnér, M., Poulsen, R., Volatility is log-normal—But not for the reason you think. *Risks*, 2018, **6**(2), 46.
- [25] Zumbach, G., Volatility processes and volatility forecast with long memory. *Quantitative Finance*, 2004, **4**(1), 70–86.

# Chapter 3

## Hybrid multifactor scheme for stochastic Volterra equations with completely monotone kernels

Sigurd Emil Rømer

### Abstract

We propose a hybrid scheme for the simulation of stochastic Volterra equations with completely monotone kernels. Our scheme is a mix of the hybrid scheme for Brownian semistationary processes of Bennedsen et al. [Financ. Stoch., **21**(4), 931-965, 2017] and the multifactor approximations of Abi Jaber et al. [SIAM J. Finan. Math., **10**(2), 309-349, 2019]. Merging the two methods allow us to both accurately capture singularities and efficiently track the inherent path dependence. We consider a forward process that is easily computable under our scheme and show how it facilitates simulation of the VIX index for a number of volatility models of the Volterra type. Numerical experiments indicate good convergence for rough Bergomi type models and the quadratic rough Heston model. Experiments on rough Heston, where we had to truncate values in zero, sometimes resulted in a large positive bias.

*Keywords:* Stochastic Volterra equations; Complete monotonicity; Simulation; Rough volatility; Option pricing.

### 3.1 Introduction

We present a simulation scheme for a class of stochastic Volterra equations. We say that  $X = (X_t)_{t \geq 0}$  solves a stochastic Volterra equation (SVE) if it satisfies an equation of the form

$$X_t = g_0(t) + \int_0^t K(t-s)b(X_s)ds + \int_0^t K(t-s)\sigma(X_s)dW_s, \quad t \geq 0, \quad (3.1)$$

### 3.1. INTRODUCTION

---

where  $W = (W_t)_{t \geq 0}$  is Brownian motion,  $K \in L_{\text{loc}}^2(\mathbb{R}_+)$ , and  $g_0 : \mathbb{R}_+ \rightarrow \mathbb{R}$ ,  $b, \sigma : \mathbb{R} \rightarrow \mathbb{R}$  are continuous. We consider only continuous and adapted solutions  $X$  that live on some filtered probability space  $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$  where the filtration  $(\mathcal{F}_t)_{t \geq 0}$  is assumed to satisfy the usual hypothesis. Throughout we define Brownian motion relative to the given filtration in the sense of [37, Definition 3.11]. Existence and uniqueness results can be found in [3]; we recite parts of them in Appendix 3.B. In Appendix 3.A, we also provide some theory on convolutions and resolvents which will be used, explicitly or not, as we go along. We will sometimes abuse notation and write  $b_t = b(X_t)$ ,  $\sigma_t = \sigma(X_t)$ ,  $t \geq 0$ .<sup>1</sup>

Our paper is motivated by an observation from quantitative finance showing that volatility on many financial assets exhibits *roughness* [8, 22]. A stochastic process is rough if the sample paths are less Hölder continuous than those of Brownian motion. A SVE can generate roughness when  $K(t) \sim ct^\alpha$  in the limit  $t \rightarrow 0_+$  where  $c \in \mathbb{R} \setminus \{0\}$ ,  $\alpha \in (-\frac{1}{2}, 0)$ ; by ' $\sim$ ' we signify informally that the left-hand side behaves like the right-hand side in the given limit. Affine SVE's have attracted special attention as they, when used to model the instantaneous variance, allow efficient option pricing by Fourier methods [2, 4, 5, 16, 17, 18, 25]. Generally, however, we rely on Monte Carlo for pricing. The rough Bergomi model of [7] is a well-known non-affine model where the asset price  $S = (S_t)_{t \geq 0}$  is modelled as

$$dS_t = S_t \sqrt{V_t} dW_{1,t}, \quad S_0 > 0, \quad (3.2)$$

$$V_t = \xi_0(t) \exp \left( \eta \sqrt{2\alpha + 1} \int_0^t (t-s)^\alpha dW_{2,s} - \frac{\eta^2}{2} t^{2\alpha+1} \right), \quad t \geq 0, \quad (3.3)$$

where  $\xi_0 : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ ,  $\alpha \in (-\frac{1}{2}, 0)$ ,  $\eta > 0$ , and  $(W_{1,t}, W_{2,t})_{t \geq 0}$  is Brownian motion with  $dW_{1,t} dW_{2,t} = \rho dt$ ,  $\rho \in [-1, 1]$ .<sup>2</sup>

Although the rough Bergomi model can calibrate very well to SPX options (see [7]), naive simulation of  $V = (V_t)_{t \geq 0}$  is slow; the authors of [7] simulate  $V$  exactly (note that  $\log(V_t)$  is Gaussian) which has a computational complexity of  $O(n^2)$ ,  $n$  being the number of steps, even rising to  $O(n^3)$  if the covariance-matrix-factorisation is accounted for.

The literature on the simulation of Volterra processes has grown significantly since then leading to many new insights and improved methods [1, 2, 4, 9, 13, 20, 21, 30, 32, 40, 46]. An important contribution is [9] where simulation of Brownian semistationary ( $\mathcal{BSS}$ ) and truncated  $\mathcal{BSS}$  ( $\mathcal{TBSS}$ ) processes is considered. The latter are processes of the form

$$X_t = \int_0^t K(t-s) \sigma_s dW_s, \quad t \geq 0, \quad (3.4)$$

where  $(\sigma_t)_{t \geq 0}$  is some stochastic process and  $K$  is a deterministic function that behaves as  $K(t) \sim ct^\alpha$  in the limit  $t \rightarrow 0_+$  where  $c \in \mathbb{R}_+$ ,  $\alpha \in (-\frac{1}{2}, \frac{1}{2}) \setminus \{0\}$ ; consult the paper

<sup>1</sup> We shall also remark that we only consider continuous versions of any processes that are of the form  $Z_t = \int_0^t \tilde{K}(t-s) a_{1,s} ds + \int_0^t \tilde{K}(t-s) a_{2,s} d\tilde{W}_s$ ,  $t \geq 0$ , where  $\tilde{K} \in L_{\text{loc}}^2(\mathbb{R}_+)$ ,  $\tilde{W} = (\tilde{W}_t)_{t \geq 0}$  is Brownian motion, and  $(a_{1,t})_{t \geq 0}$ ,  $(a_{2,t})_{t \geq 0}$  are given locally bounded and progressively measurable processes. We refer to the comment in [4, Section A.1] for the existence of a continuous version of  $(Z_t)_{t \geq 0}$ .

<sup>2</sup> The model is stated in risk-neutral terms with zero interest rates and dividends assumed; the latter is without loss of generality up to deterministic values.

CHAPTER 3. HYBRID MULTIFACTOR SCHEME FOR STOCHASTIC  
VOLTERRA EQUATIONS WITH COMPLETELY MONOTONE KERNELS

---

for all the assumptions. The Volterra integral of (3.3) is nested in  $\sigma_t = 1$ ,  $K(t) = t^\alpha$ . A naive discretisation of (3.4) would be to approximate both  $(\sigma_t)_{t \geq 0}$  and  $K$  by step functions. However, as the authors of [9] show, that works poorly when  $K$  is singular. They propose instead a hybrid scheme where the power-law part of  $K$  is used without error for some  $\kappa$  steps near the origin and only otherwise approximated piecewise constant. For the singular case ( $\alpha < 0$ ) the numerical solution turns out to be accurate for even low, but non-zero,  $\kappa$ 's. Consequently, the cost related to the power-law part of  $K$  is small. The remainder of the approximations amounts to a discrete convolution. With a loop the cost of that is  $O(n^2)$ , though, when  $(\sigma_t)_{t \geq 0}$  can be pre-simulated, the fast Fourier transform (FFT) can be used lowering it to  $O(n \log_2 n)$ . The FFT applies e.g. for the rough Bergomi model and their scheme therefore beats exact simulation for it.

Another important contribution is [4] where focus is on a class of SVE driven volatility models, but see also [2]. In these papers, it is suggested that we approximate  $K(t) \approx \sum_{i=1}^m c_i e^{-\gamma_i t}$  for coefficients  $(c_i, \gamma_i)_{i=1}^m$ ,  $m \in \mathbb{N}$ . The approximation is justified essentially if and only if  $K$  is *completely monotone* (CM).<sup>3</sup> The CM assumption covers most conceivable positive, decreasing, and smooth functions and is therefore mild in a volatility context. If we replace  $K$  in (3.1) by the sum-of-exponentials approximation, we obtain a  $m$ -factor Markovian model—also called a *lifted* model. Simulation with an Euler scheme results in costs  $O(mn)$  scaling even better, linearly, in  $n$ . That analysis, however, is in general too simplistic as it disregards realistic values of  $m$  and  $n$ . Indeed, as exponential functions are non-singular, a large  $m$  may be needed to capture a singular  $K$ , and even if the given singularity is captured by the exponentials, a large  $n$  may be needed for convergence. Therefore, although the multifactor approximation scales better in  $n$  as such, the scheme need not be efficient for practical values of  $m$  and  $n$ .

To exploit only the best aspects of each method, we in this work propose to merge the hybrid scheme of [9] with the multifactor approximations of [2, 4]. More precisely, we suggest to handle  $K$  without error for a number of steps  $\kappa$  near the origin and to approximate the remainder by a sum-of-exponentials. We call our method the *hybrid multifactor scheme*. As we will demonstrate, the hybrid multifactor scheme is more accurate than a pure multifactor approximation and faster than the hybrid scheme of [9].

Another contribution of ours is a detailed review of methods for finding the exponential terms, including those of [2, 4]. An efficient kernel approximation is vital as the computational costs are related directly to the number of exponentials. Our research has here led us to the method of [10] which, to our knowledge, have not previously been used in the context of SVE's. By example on the rough fractional kernel, we demonstrate that it outperforms the methods of [2, 4] by a notable margin. Inspired by [3], we consider also a *forward process*  $g = (g_t)_{t \geq 0}$  defined by

$$g_t(\tau) = g_0(t + \tau) + \int_0^\tau K(t + \tau - s)b(X_s)ds + \int_0^\tau K(t + \tau - s)\sigma(X_s)dW_s, \quad (3.5)$$

---

<sup>3</sup> A real-valued function  $f$  is completely monotone (CM) on  $(0, \infty)$  if it is infinitely differentiable on the same set and  $(-1)^n f^{(n)}(t) \geq 0$  for all  $t > 0$  and  $n \in \mathbb{N}_0$ . If  $f$  is also continuous at zero, we say it is CM on  $[0, \infty)$ . Whenever we write CM without specifying the interval, we refer to  $(0, \infty)$  only.

### 3.2. THE HYBRID MULTIFACTOR SCHEME

---

for  $t, \tau \geq 0$  and which nests  $X_t = g_t(0)$ . We will show how our scheme facilitates easy computation of  $g$  and will demonstrate how one can use that to compute the VIX index for a number of volatility models of the Volterra type. Moreover, we shall prove strong convergence for the scheme assuming that  $b, \sigma$  are Lipschitz continuous. A Matlab implementation is made available at: [https://github.com/sigurdroemer/hybrid\\_multifactor\\_scheme](https://github.com/sigurdroemer/hybrid_multifactor_scheme).

The paper is structured as follows: In Section 3.2, we outline the scheme and discuss it against the ideas of [2, 4, 9]. In Section 3.3, we motivate the forward process and show how to compute it numerically under the scheme. In Section 3.4, we present our convergence result. Sections 3.5-3.7 are more practical. In Section 3.5, we review methods for choosing the sum-of-exponentials approximation, in Section 3.6, discuss simulation of the VIX index for three Volterra volatility models, and in Section 3.7, present a number of simulation and option pricing experiments. In Section 3.8, we conclude on our work and outline ideas for future research. We have delegated most of the proofs to the appendix.

**Notation:** For a matrix  $A \in \mathbb{C}^{k \times k}$ ,  $k \in \mathbb{N}$ , we write  $A^\top$  for the transpose,  $\bar{A}$  for the complex conjugate,  $A^* = (\bar{A})^\top$  for the conjugate transpose. We write also  $A = \{A_{i,j}\}_{i,j=1}^k$  where  $A_{i,j}$  refers to the element of row  $i$ , column  $j$ . Given  $x, y \in \mathbb{C}^k$  we write  $x \cdot y$  for their element-by-element product here assuming that they both are either row or column vectors. Unless written otherwise, we consider elements of  $\mathbb{C}^k$  as column vectors. Given  $x = (x_1, \dots, x_k) \in \mathbb{C}^k$ ,  $k \in \mathbb{N}$ , we write  $\|x\|_2 = \sqrt{|x_1|^2 + \dots + |x_k|^2}$  and  $\|x\|_\infty = \max\{|x_1|, \dots, |x_k|\}$  where  $|\cdot|$  is the complex modulus. When the specific norm on  $\mathbb{C}^k$  (especially  $\mathbb{R}^k$ ) is subordinate to our statements we write  $\|\cdot\|$  for an arbitrary one; recall that all norms on  $\mathbb{C}^k$  (in particular  $\mathbb{R}^k$ ) are equivalent—we use this in our convergence proofs. We write  $x^+ = \max\{0, x\}$ ,  $x \wedge y = \min\{x, y\}$ ,  $x \vee y = \max\{x, y\}$  for  $x, y \in \mathbb{R}$ . Given  $p \geq 1$  and an interval  $I \subset \mathbb{R}$  we write  $L^p(I)$  for the space of real-valued functions whose  $p$ 'th moment is absolutely integrable on  $I$ . The corresponding norm is written

$$\|f\|_{L^p(I)} = \left( \int_I |f(t)|^p dt \right)^{1/p}, \quad f \in L^p(I).$$

By  $L^p_{\text{loc}}(I)$  we denote the set of real-valued functions that are  $L^p$  integrable on all compact subsets of  $I$ . Given a filtration  $(\mathcal{F}_t)_{t \geq 0}$  on a probability space, we write for any  $t \geq 0$ ,  $E_t(\cdot) = E(\cdot | \mathcal{F}_t)$  and  $\text{Var}_t(\cdot) = \text{Var}(\cdot | \mathcal{F}_t)$  for the conditional expectation and variance operators. By  $\Delta_h$ ,  $h \geq 0$ , we denote the shift operator, i.e. for a function  $f$ :  $\Delta_h f(t) = f(t+h)$ . If  $f$  is right-continuous on  $\mathbb{R}_+$  and of locally bounded variation we write  $df$  for the measure that is induced by its distributional derivative so  $f(t) = f(0) + \int_{[0,t]} df(s)$ ,  $t \geq 0$ . The convolution operator is written  $*$ . We define it rigorously in Appendix 3.A.

## 3.2 The hybrid multifactor scheme

We assume that (3.1) has a solution  $X$  which we now fix. The goal is to simulate  $X$  over the time interval  $[0, T]$  for some  $T > 0$ . For  $n \in \mathbb{N}$  define  $t_i := \frac{i}{n}$ ,  $i = 0, 1, \dots, [nT]$ , as well as  $\Delta_n := n^{-1}$  and  $t_n^- := [nt]/n$ ,  $t_n^+ := \lceil nt \rceil / n$ , for  $t \geq 0$ . As lies at the core of our

CHAPTER 3. HYBRID MULTIFACTOR SCHEME FOR STOCHASTIC  
VOLTERRA EQUATIONS WITH COMPLETELY MONOTONE KERNELS

---

scheme, we will approximate  $K$  by the function

$$K_{mn}(t) := 1_{(t \leq \kappa \Delta_n)} K(t) + 1_{(t > \kappa \Delta_n)} K_m(t), \quad t > 0, \quad (3.6)$$

where  $\kappa \in \{0, 1, \dots, \lfloor nT \rfloor\}$  and  $K_m(t) := \sum_{i=1}^m c_i^m e^{-\gamma_i^m t}$ ,  $t > 0$ ,  $m \in \mathbb{N}$ ,  $c^m = (c_1^m, \dots, c_m^m)^\top$ ,  $\gamma^m = (\gamma_1^m, \dots, \gamma_m^m)^\top \in \mathbb{R}^m$ . To validly simulate  $(X_t)_{t \in [0, T]}$  we need  $K_{mn}$  to approximate  $K$  on  $[0, T]$ , equivalently  $K_m$  to approximate  $K$  on  $[\kappa \Delta_n, T]$ .<sup>4</sup> We will need to extend the approximation interval when we later consider computation of the forward process  $g$ . The approximation by  $K_m$  is justified, at least heuristically, if  $K$  is CM: Note by [51, Corollary 7.12] that  $K$  being CM is equivalent to it having the representation

$$K(t) = \int_{[0, \infty)} e^{-\gamma t} \mu(d\gamma), \quad t > 0, \quad (3.7)$$

for a non-negative Borel measure  $\mu$ .<sup>5</sup> Consider now the possibility of approximating  $\mu \approx \hat{\mu}_m := \sum_{i=1}^m c_i^m \delta_{\gamma_i^m}$  for a given  $m$  and appropriate coefficients  $(c^m, \gamma^m)$ ; here  $\delta_{\gamma_i^m}$  denotes the Dirac measure with point mass in  $\gamma_i^m$ ,  $i = 1, \dots, m$ . If the coefficients  $(c^m, \gamma^m)$  are chosen non-negative—this is natural given that  $\mu$  is non-negative and the integral in (3.7) is over  $[0, \infty)$ —we obtain, replacing  $\mu$  by  $\hat{\mu}_m$  in (3.7), the function  $K_m$ . If  $K$  is CM, it is then conceivable that we should be able to find a sequence  $\{(c^m, \gamma^m)\}_{m=1}^\infty$  so  $K_m$  (and thereby  $K_{mn}$ ) converges to  $K$  on  $[0, T]$  in some given sense in the limit  $m \rightarrow \infty$ .<sup>6</sup>

As claimed earlier, the CM assumption is mild for volatility modelling. Examples are the exponential kernel  $K(t) = ce^{-\lambda t}$ ,  $c, \lambda \geq 0$ , the rough fractional kernel  $K(t) = ct^\alpha$ ,  $\alpha \in (-\frac{1}{2}, 0]$ ,  $c \geq 0$ , and the shifted power-law kernel  $K(t) = (1+t)^\beta$ ,  $\beta \leq 0$ ; check the definition in footnote 3. Since complete monotonicity is preserved by products and sums [43, Theorem 1], we can easily construct more flexible kernels from basic building blocks such as these. It should be noted that CM functions are always non-negative, non-increasing, and smooth ( $C^\infty$ ); check again the definition. Although our scheme effectively can only be justified if  $K$  is CM, we will not explicitly invoke the assumption unless needed for theoretical results or otherwise. For now it is simply assumed that  $K_{mn}$  for given  $m, n$  and  $(c^m, \gamma^m)$  approximates  $K$  on  $[0, T]$  in some unspecified sense.

In what follows, we derive the scheme. This goes in three steps: First we replace  $K$  by  $K_{mn}$  which results in a number of auxiliary stochastic processes. Next, we discretise all coefficient processes that appear with Euler approximations. This will give us a numerical solution on the time points  $t_0, t_1, \dots, t_{\lfloor nT \rfloor}$ . Next we extend the solution to the full interval  $[0, T]$ . We end with a discussion of the computational costs and accuracy.

<sup>4</sup> The results that we present in Section 3.4 suggests that  $K \approx K_{mn}$  should hold in the  $L^p([0, T])$  sense for some  $p > 2$ ; we suspect that this could be strengthened to  $p = 2$ . It is as expected that the kernel approximation should be valid in some  $L^p([0, T])$  sense as  $K$  only appears inside integrals (ordinary or stochastic). When we say that  $K_{mn}$  or  $K_m$  approximates  $K$  on some interval it should thus *not* be understood in a pointwise sense (unless otherwise stated); it is then meaningful to e.g. state that  $K_{mn} = K_m$  (say  $\kappa = 0$ ) should approximate  $K$  on  $[0, T]$  even if  $K$  is singular at the origin.

<sup>5</sup> Equation (3.7) extends to  $t = 0$  if  $K$  is CM on  $[0, \infty)$  in which case  $\mu$  is finite [51, Theorem 7.11]. Note that if  $K$  is CM on  $[0, \infty)$  then it is non-singular. This should be kept in mind for later discussions.

<sup>6</sup> We have been unable to find rigorous results stating that a sequence  $\{(c^m, \gamma^m)\}_{m=1}^\infty$  exists for an arbitrary CM kernel  $K$  so  $K_m$  converges to it in some given sense. The papers [2, 4] contain limited results. For a general CM kernel we rely on our heuristic justification.

### 3.2.1 Applying the kernel approximation

Replacing  $K$  by  $K_{mn}$  in (3.1), we obtain the equation

$$X_t^{mn} = g_0(t) + \int_0^t K_{mn}(t-s)b(X_s^{mn})ds + \int_0^t K_{mn}(t-s)\sigma(X_s^{mn})dW_s, \quad t \geq 0, \quad (3.8)$$

for a stochastic process  $X^{mn} = (X_t^{mn})_{t \geq 0}$ . A solution to (3.8) need not exist, especially not on the same probability space as  $X$ . To ease the exposition, we will nevertheless precisely assume that a continuous adapted solution  $X^{mn}$  exists on the same filtered probability space as  $X$  and which is driven by the same Brownian motion—thus we write only  $W$ . This will be true under the assumptions of our convergence result.

We now rewrite (3.8) to facilitate further numerical approximation. We will abuse notation and write  $b_t^{mn} = b(X_t^{mn})$ ,  $\sigma_t^{mn} = \sigma(X_t^{mn})$  for  $t \geq 0$ . Plugging (3.6) into (3.8):

$$X_t^{mn} = g_0(t) + \int_0^{(t-\kappa\Delta_n)^+} K_m(t-s)b_s^{mn} ds + \int_0^{(t-\kappa\Delta_n)^+} K_m(t-s)\sigma_s^{mn} dW_s \quad (3.9)$$

$$+ \int_{(t-\kappa\Delta_n)^+}^t K(t-s)b_s^{mn} ds + \int_{(t-\kappa\Delta_n)^+}^t K(t-s)\sigma_s^{mn} dW_s, \quad t \geq 0. \quad (3.10)$$

Define

$$U_{i,t}^{mn} := \int_0^t e^{-\gamma_i^m(t-s)} b_s^{mn} ds + \int_0^t e^{-\gamma_i^m(t-s)} \sigma_s^{mn} dW_s, \quad i = 1, \dots, m, \quad t \geq 0. \quad (3.11)$$

Since  $b, \sigma$  are assumed continuous, the above are well-defined as continuous semimartingales. We derive their dynamics for later use: Let  $i \in \{1, \dots, m\}$  and define for  $t \geq 0$ ,  $Y_{i,t}^{mn} := \int_0^t e^{\gamma_i^m s} b_s^{mn} ds + \int_0^t e^{\gamma_i^m s} \sigma_s^{mn} dW_s$  so  $U_{i,t}^{mn} = e^{-\gamma_i^m t} Y_{i,t}^{mn}$ . By Ito's lemma then

$$dU_{i,t}^{mn} = -\gamma_i^m U_{i,t}^{mn} dt + e^{-\gamma_i^m t} dY_{i,t}^{mn} \quad (3.12)$$

$$= (b_t^{mn} - \gamma_i^m U_{i,t}^{mn}) dt + \sigma_t^{mn} dW_t. \quad (3.13)$$

Note that

$$\int_0^{(t-\kappa\Delta_n)^+} K_m(t-s)b_s^{mn} ds + \int_0^{(t-\kappa\Delta_n)^+} K_m(t-s)\sigma_s^{mn} dW_s \quad (3.14)$$

$$= \sum_{i=1}^m c_i^m \left( \int_0^{(t-\kappa\Delta_n)^+} e^{-\gamma_i^m(t-s)} b_s^{mn} ds + \int_0^{(t-\kappa\Delta_n)^+} e^{-\gamma_i^m(t-s)} \sigma_s^{mn} dW_s \right) \quad (3.15)$$

$$= \sum_{i=1}^m c_i^m e^{-\gamma_i^m \kappa\Delta_n} \left( \int_0^{(t-\kappa\Delta_n)^+} e^{-\gamma_i^m(t-\kappa\Delta_n-s)} (b_s^{mn} ds + \sigma_s^{mn} dW_s) \right) \quad (3.16)$$

$$= \sum_{i=1}^m c_i^m e^{-\gamma_i^m \kappa\Delta_n} U_{i,(t-\kappa\Delta_n)^+}^{mn}, \quad t \geq 0. \quad (3.17)$$

The last line should be clear for  $t \geq \kappa\Delta_n$  where  $(t - \kappa\Delta_n)^+ = t - \kappa\Delta_n$ . To make sense of it for  $t \in [0, \kappa\Delta_n]$  where  $(t - \kappa\Delta_n)^+ = 0$  note that  $U_{i,0}^{mn} = 0$ ,  $i = 1, \dots, m$ .

Combining (3.9)-(3.10) and (3.14)-(3.17), we conclude

$$X_t^{mn} = g_0(t) + \sum_{i=1}^m c_i^m e^{-\gamma_i^m \kappa \Delta_n} U_{i,(t-\kappa\Delta_n)^+}^{mn} \quad (3.18)$$

$$+ \int_{(t-\kappa\Delta_n)^+}^t K(t-s) b_s^{mn} ds + \int_{(t-\kappa\Delta_n)^+}^t K(t-s) \sigma_s^{mn} dW_s, \quad t \geq 0. \quad (3.19)$$

The above will form the basis of our numerical approximation of  $X$ .

### 3.2.2 Discretising the equations

In this and the next subsection, we will construct numerical solutions of  $X_t^{mn}$  and  $U_t^{mn} := (U_{1,t}^{mn}, \dots, U_{m,t}^{mn})^\top$  for  $t \in [0, T]$ . They will be denoted  $\hat{X}_t^{mn}$  and  $\hat{U}_t^{mn} := (\hat{U}_{1,t}^{mn}, \dots, \hat{U}_{m,t}^{mn})^\top$ ,  $t \in [0, T]$ . In the current subsection, we apply Euler approximations to (3.12)-(3.13) and (3.18)-(3.19) to produce solution values  $(\hat{X}_t^{mn}, \hat{U}_t^{mn})$  on the grid points  $t \in \{t_0, t_1, \dots, t_{[nT]}\}$ . In the next subsection, we extend to a general time point in  $[0, T]$ .

To simplify notation, we will, given a numerical solution  $\hat{X}_t^{mn}$  for some  $t \in [0, T]$ , write  $\hat{b}_t^{mn} := b(\hat{X}_t^{mn})$  and  $\hat{\sigma}_t^{mn} := \sigma(\hat{X}_t^{mn})$ . Define now  $\Delta W_i^n := W_{t_{i+1}} - W_{t_i}$ ,  $i = 0, 1, \dots, [nT] - 1$ . Let  $j \in \{1, \dots, m\}$ . An explicit Euler scheme for (3.12)-(3.13) results in

$$\hat{U}_{j,t_{i+1}}^{mn} = \hat{U}_{j,t_i}^{mn} + \left( \hat{b}_{t_i}^{mn} - \gamma_j^m \hat{U}_{j,t_i}^{mn} \right) \Delta_n + \hat{\sigma}_{t_i}^{mn} \Delta W_i^n, \quad i = 0, 1, \dots, [nT] - 1, \quad (3.20)$$

where  $\hat{U}_{j,0}^{mn} = 0$  is the initial condition. Alternatively, we can let  $U_{j,t}^{mn}$  be implicit in (3.13). This results in the explicit-implicit equation:<sup>7</sup>

$$\hat{U}_{j,t_{i+1}}^{mn} = \frac{1}{1 + \gamma_j^m \Delta_n} \left( \hat{U}_{j,t_i}^{mn} + \hat{b}_{t_i}^{mn} \Delta_n + \hat{\sigma}_{t_i}^{mn} \Delta W_i^n \right), \quad i = 0, 1, \dots, [nT] - 1. \quad (3.21)$$

The choice (3.21) coincides with that used in [2] for the rough Heston model. Since implicit terms tend to improve numerical stability, see e.g. Chapter 14 of [44], the numerical experiments of Section 3.7 are based on (3.21). The convergence result we present in Section 3.4 though uses (3.20) as we found the proof more manageable with that equation.

For  $\kappa > 0$ , define now

$$\tilde{W}_{i,k}^n = \int_{t_i}^{t_{i+1}} K(t_{i+k} - s) dW_s, \quad i = 0, 1, \dots, [nT] - 1, \quad k = 1, 2, \dots, \kappa, \quad (3.22)$$

and

$$w_k = \int_{\frac{k-1}{n}}^{\frac{k}{n}} K(s) ds, \quad k = 1, \dots, \kappa. \quad (3.23)$$

---

<sup>7</sup> Technically we need  $1 + \gamma_j^m \Delta_n \neq 0$ . This is trivially satisfied if  $\gamma_j^m \geq 0$  as is natural given (3.7).

### 3.2.2. DISCRETISING THE EQUATIONS

Inspired by (3.18)-(3.19), we for  $i = 0, 1, \dots, \lfloor nT \rfloor$ , define and rewrite

$$\hat{X}_{t_i}^{mn} := g_0(t_i) + \sum_{j=1}^m c_j^m e^{-\gamma_j^m \kappa \Delta_n} \hat{U}_{j, (t_i - \kappa \Delta_n)^+}^{mn} \quad (3.24)$$

$$+ \int_{(t_i - \kappa \Delta_n)^+}^{t_i} K(t_i - s) \hat{b}_{s_n}^{mn} ds + \int_{(t_i - \kappa \Delta_n)^+}^{t_i} K(t_i - s) \hat{\sigma}_{s_n}^{mn} dW_s \quad (3.25)$$

$$= g_0(t_i) + \sum_{j=1}^m c_j^m e^{-\gamma_j^m \kappa \Delta_n} \hat{U}_{j, (t_i - \kappa \Delta_n)^+}^{mn} \quad (3.26)$$

$$+ \sum_{k=1}^{\min\{i, \kappa\}} \left( \hat{b}_{t_{i-k}}^{mn} \int_{t_{i-k}}^{t_{i-k+1}} K(t_i - s) ds + \hat{\sigma}_{t_{i-k}}^{mn} \int_{t_{i-k}}^{t_{i-k+1}} K(t_i - s) dW_s \right) \quad (3.27)$$

$$= g_0(t_i) + \sum_{j=1}^m c_j^m e^{-\gamma_j^m \kappa \Delta_n} \hat{U}_{j, t_{i-\kappa}}^{mn} + \sum_{k=1}^{\min\{i, \kappa\}} \left( \hat{b}_{t_{i-k}}^{mn} w_k + \hat{\sigma}_{t_{i-k}}^{mn} \tilde{W}_{i-k, k}^n \right). \quad (3.28)$$

By convention, we interpret the sum of (3.27) and the second sum of (3.28) as zero when  $i = 0$  or  $\kappa = 0$ . We likewise interpret the first sum of (3.28) as zero when  $i < \kappa$ .

Given access to the Brownian increments and the stochastic integrals of (3.22), we can iterate and evaluate (3.20) or (3.21) and (3.28) to obtain  $(\hat{X}_{t_i}^{mn}, \hat{U}_{t_i}^{mn})$  for  $i = 0, 1, \dots, \lfloor nT \rfloor$ .

To sample the stochastic terms  $\Delta W_i^n$  and  $\tilde{W}_{i, k}^n$  for  $i = 0, 1, \dots, \lfloor nT \rfloor - 1$ , and  $k = 1, \dots, \kappa$ , we follow [9] and note that they are contained in the  $\kappa + 1$  dimensional vectors

$$\mathbf{W}_i^n := (\Delta W_i^n, \tilde{W}_{i, 1}^n, \tilde{W}_{i, 2}^n, \dots, \tilde{W}_{i, \kappa}^n)^\top, \quad i = 0, \dots, \lfloor nT \rfloor - 1, \quad (3.29)$$

which are i.i.d. mean-zero Gaussians. For sampling we need the covariance matrix which we denote by  $\Sigma = \{\Sigma_{j, k}\}_{j, k=1}^{\kappa+1}$ . The entries can be found with the Ito isometry: Take an arbitrary  $i \in \{0, 1, \dots, \lfloor nT \rfloor - 1\}$ . Then  $\Sigma_{1, 1} = E[(\Delta W_i^n)^2] = \frac{1}{n}$  and for  $k = 2, \dots, \kappa + 1$ :

$$\Sigma_{1, k} = E \left[ \int_{t_i}^{t_{i+1}} dW_s \int_{t_i}^{t_{i+1}} K(t_{i+k-1} - s) dW_s \right] = \int_{\frac{k-2}{n}}^{\frac{k-1}{n}} K(s) ds. \quad (3.30)$$

For  $j, k = 2, \dots, \kappa + 1$ , so  $j \leq k$ :

$$\Sigma_{j, k} = E \left[ \int_{t_i}^{t_{i+1}} K(t_{i+j-1} - s) dW_s \int_{t_i}^{t_{i+1}} K(t_{i+k-1} - s) dW_s \right] \quad (3.31)$$

$$= \int_{\frac{j-2}{n}}^{\frac{j-1}{n}} K(s) K(s + (k-j)/n) ds. \quad (3.32)$$

The remaining values follow by symmetry. Note that the integrals of (3.23) coincide with those of (3.30).

While semi-analytical expressions for (3.30)-(3.32) can be found in [9] for the fractional

kernel, we generally have to use numerical integration. As standard integration algorithms may struggle for singular integrands let us suggest the following when  $K$  is singular: We will make the mild assumption that  $K(t) = \sum_{i=1}^d c_i t^{\alpha_i} K_i(t)$ ,  $t > 0$ , where  $d \in \mathbb{N}$  and for  $i = 1, 2, \dots, d$ ,  $K_i \in L_{\text{loc}}^2(\mathbb{R}_+)$  are CM and non-singular,  $c_i \in \mathbb{R}_+$ ,  $\alpha_i \in (-\frac{1}{2}, 0]$ . The terms of (3.30)-(3.32) that can have singular integrands then all reduce to a series of integrals of the form  $\int_0^b f(s)ds$  where  $b > 0$  and  $f$  is a function that behaves as  $f(s) \sim s^\alpha h(s)$  in the limit  $s \rightarrow 0_+$  where  $\alpha \in (-1, 0]$  and  $h$  is a non-singular and continuous function. For robust numerical integration let us then suggest to choose a small number  $\delta \in (0, b]$  and approximate

$$\int_0^b f(s)ds = \int_0^\delta f(s)ds + \int_\delta^b f(s)ds \approx \frac{1}{\alpha+1} \delta^{\alpha+1} \frac{h(0) + h(\delta)}{2} + \int_\delta^b f(s)ds$$

where the second term should be digestible for a standard integration algorithm.

### 3.2.3 Extending the solution

We extend to an arbitrary time point as

$$\hat{X}_t^{mn} = \hat{X}_{t_n^-}^{mn}, \quad \hat{U}_{i,t}^{mn} = \hat{U}_{i,t_n^-}^{mn}, \quad i = 1, \dots, m, \quad 0 \leq t \leq T.$$

The above avoids foresight in the solution values as is critical for some applications.

### 3.2.4 Algorithm

We summarise the scheme below.

**Definition 3.2.1.** (The hybrid multifactor scheme) *Set  $\hat{X}_0^{mn} = g_0(0)$  and  $\hat{U}_{i,0}^{mn} = 0$  for  $i = 1, \dots, m$ . Perform the following steps for  $i = 1, \dots, \lfloor nT \rfloor$ :*

1. Simulate  $\mathbf{W}_{i-1}^n \sim \mathcal{N}(0, \Sigma)$ .
2. Update the  $b$ - and  $\sigma$ -processes:

$$\hat{b}_{i-1}^{mn} = b(\hat{X}_{i-1}^{mn}), \quad \hat{\sigma}_{i-1}^{mn} = \sigma(\hat{X}_{i-1}^{mn}).$$

3. Update the  $U$ -factors, for  $j = 1, \dots, m$ , as

$$\hat{U}_{j,i}^{mn} = \hat{U}_{j,i-1}^{mn} + \left( \hat{b}_{i-1}^{mn} - \gamma_j^m \hat{U}_{j,i-1}^{mn} \right) \Delta_n + \hat{\sigma}_{i-1}^{mn} \Delta W_{i-1}^n \quad (3.33)$$

or, assuming  $1 + \gamma_j^m \Delta_n \neq 0$ , as

$$\hat{U}_{j,i}^{mn} = \frac{1}{1 + \gamma_j^m \Delta_n} \left( \hat{U}_{j,i-1}^{mn} + \hat{b}_{i-1}^{mn} \Delta_n + \hat{\sigma}_{i-1}^{mn} \Delta W_{i-1}^n \right). \quad (3.34)$$

4. Update the  $X$ -process as

$$\hat{X}_i^{mn} = g_0(t_i) + \sum_{j=1}^m c_j^m e^{-\gamma_j^m \kappa \Delta_n} \hat{U}_{j,i-\kappa}^{mn} + \sum_{k=1}^{\min\{i,\kappa\}} \left( \hat{b}_{i-k}^{mn} w_k + \hat{\sigma}_{i-k}^{mn} \tilde{W}_{i-k,k}^n \right) \quad (3.35)$$

where the first sum should be interpreted as zero if  $i < \kappa$ , the second sum if  $\kappa = 0$ .

### 3.2.5. COMPUTATIONAL COSTS AND ACCURACY

---

Extend the solution to an arbitrary time point as:

$$\hat{X}_t^{mn} = \hat{X}_{t_n^-}^{mn}, \quad \hat{U}_{i,t}^{mn} = \hat{U}_{i,t_n^-}^{mn}, \quad i = 1, \dots, m, \quad 0 \leq t \leq T. \quad (3.36)$$

□

Let us give two initial remarks on a practical implementation.

Firstly, to implement bullet point 1, note the following: Because  $\Sigma$  is a covariance matrix it is symmetric and positive semi-definite. Simulation of  $\mathbf{W}_{i-1}$  for a given  $i$  can then be achieved by first factorising  $\Sigma = AA^\top$  where  $A$  is a  $(\kappa+1) \times (\kappa+1)$  dimensional lower triangular matrix.<sup>8</sup> It then follows by standard results that simulating  $Z_{i-1} \sim \mathcal{N}(0, I_{\kappa+1})$ , where  $I_{\kappa+1}$  is the  $\kappa+1$  dimensional identity matrix, and setting  $\mathbf{W}_{i-1} = AZ_{i-1}$  returns a sample from the correct distribution. We can expect the factorisation to cost little time as it only needs to be done once even if multiple paths are to be simulated and because  $\kappa$ , as the reader will see, is recommended low, though not necessarily zero.

Secondly, it should be noted that even if  $X_t$  lives in a restricted region of  $\mathbb{R}$ , the numerical solution  $\hat{X}_t^{mn}$  is unrestricted as we use a locally Gaussian approximation. Sometimes it is important that  $\hat{X}_t^{mn}$  satisfies the same restrictions as  $X_t$ . One example is that of modelling the instantaneous variance or volatility of an asset where a typical equation, if we model either directly, would have a non-negative solution. For the classical case [39] gives a thorough analysis of non-negativity fixes for a class of stochastic volatility models driven by SDE's. We will not perform a similar detailed analysis of how to restrict values under our scheme (to the non-negative numbers or otherwise). A simple, though not necessarily efficient, solution is to truncate values in bullet point 4 of the scheme.

### 3.2.5 Computational costs and accuracy

We discuss the computational costs and accuracy. We start with a discussion of the choice of  $\kappa$  and consider the extremal values first. Setting  $\kappa = 0$  we get

$$\hat{X}_{t_i}^{mn} = g_0(t_i) + \sum_{j=1}^m c_j^m \hat{U}_{j,t_i}^{mn}, \quad i = 0, 1, \dots, \lfloor nT \rfloor, \quad (3.37)$$

which is a discrete version of a Markovian lift of the model. The computational complexity is  $O(mn)$ . Setting  $\kappa = \lfloor nT \rfloor$  gives

$$\hat{X}_{t_i}^{mn} = g_0(t_i) + \sum_{k=1}^i \left( \hat{b}_{t_i-k}^{mn} w_k + \hat{\sigma}_{t_i-k}^{mn} \tilde{W}_{i-k,k}^n \right), \quad i = 0, 1, \dots, \lfloor nT \rfloor, \quad (3.38)$$

where the sum should be interpreted as zero when  $i = 0$ . The above represents exact simulation up to approximating the coefficient processes piecewise constant. To compute (3.38) at all time points, in general, costs  $O(n^2)$ . If we include simulation of the  $\mathbf{W}_i$ 's, it even rises to  $O(n^3)$  as each of the computations  $AZ_i$ ,  $i = 0, 1, \dots, \lfloor nT \rfloor - 1$ , costs  $O(n^2)$ .

<sup>8</sup> This is known as the Cholesky factorisation for the positive definite case and it is unique. The factorisation is also possible in the semi-definite case though it need not be unique [26, Chapter 4.2].

The choice of  $\kappa$  involves a trade-off between accuracy and computational costs. While higher values improve the approximation of  $K$  near the origin it also increases the dimensions of  $\mathbf{W}_i^n$ ,  $i = 0, \dots, \lfloor nT \rfloor - 1$ , and adds more terms to the second sum of (3.35); as noted, for  $\kappa = \lfloor nT \rfloor$  the cost is even cubic in  $n$ . For singular kernels, the experiments of Section 3.7 suggest that we should choose  $\kappa$  non-zero but low. Indeed, in the singular case, we find a dramatic reduction in the error using  $\kappa = 1$  instead of  $\kappa = 0$ , but little improvement with higher values. For non-singular kernels we find that  $\kappa = 0$  suffices. We should therefore prefer the hybrid multifactor scheme over the pure multifactor approximation as it, at least in the singular case, achieves a much lower error at almost the same cost; of course when  $\kappa = 0$  our scheme *is* the pure multifactor approximation. Note also that for  $\kappa$  fixed, not necessarily zero, the computational complexity is still  $O(mn)$ .

It remains, among other things, to discuss the practical aspects of the sum-of-exponentials approximation  $K_m$ . Since  $m$  is related directly to the computational costs, we want an accurate approximation to  $K$  on the relevant domain with as few terms as possible. It is for now unclear how this should be done. Moreover, given some hopefully efficient method for choosing  $K_m$  it is unclear what  $m$ 's we can expect for a reasonable approximation and therefore if the scheme is practically feasible. Lastly, it is unclear how  $m$  should, or *will* as it turns out, relate to  $n$  and therefore how the complexity  $O(mn)$  can be compared with other schemes (we are thinking of the hybrid scheme of [9] in particular) that only refine the solution in a number of steps  $n$ . We discuss the above in what follows and will to that end draw on numerical evidence from the future Sections 3.5 and 3.7.

For our discussion we will, as we also suggest, implicitly assume  $\kappa > 0$  when  $K$  is singular,  $\kappa = 0$  otherwise. Based on the review we give in Section 3.5, we recommend the method of [10] for choosing  $m$  and  $(c^m, \gamma^m)$ . We give an overview of the method as we propose to use it: Let  $n$  and an error tolerance  $\epsilon > 0$  be given. First we sample values of  $K$  on equidistant points on  $[\kappa\Delta_n, T]$ . We assume that the number sampling points, up to rounding, is a constant times the number of simulation steps on  $[\kappa\Delta_n, T]$ . When we later consider computation of the forward values the interval will be changed to  $[\kappa\Delta_n, T^*]$  for a time horizon  $T^* \geq T$ ; the current discussion does not depend on this generalisation. The method of [10] can return  $m$  (we write also  $m(n, \epsilon)$ ) and coefficients  $(c^m, \gamma^m)$  so the normalised  $l^2$ -error between  $K$  and  $K_m$  on the sampled points is less than or close to  $\epsilon$ .<sup>9</sup> Numerical evidence suggests that the resulting approximation is close to optimal in the sense that  $m(n, \epsilon)$  is about the minimal number of terms that ensures a normalised  $l^2$ -error below  $\epsilon$ . For reasonable  $n$  and  $\epsilon$ , we obtain  $m(n, \epsilon)$  in the range 1–10; this includes singular kernels. Consequently, we can expect the scheme's practical cost to be manageable in terms of  $m$  when  $K_m$  is chosen with the method of [10].

We turn our attention to the scheme's complexity when  $K_m$  is chosen as described. Note that we may now write the complexity as  $O(m(n, \epsilon)n)$ . For  $\hat{X}_t^{mn}$  to converge to  $X_t$ , intuition tells us that we need both  $n \rightarrow \infty$  (to refine the scheme's step size and sample more points on  $K$  per unit of time<sup>10</sup>) and  $\epsilon \rightarrow 0$  (to refine the kernel approxi-

<sup>9</sup> Let  $y, \hat{y} \in \mathbb{R}^k$ ,  $k \in \mathbb{N}$ ,  $y \neq 0$ . We define the normalised  $l^2$ -error of  $\hat{y}$  against  $y$  by  $\|\hat{y} - y\|_2 / \|y\|_2$ .

<sup>10</sup> Because  $[\kappa\Delta_n, T]$  is bounded in  $n$ , the number of sampling points per time unit goes to  $\infty$  as  $n \rightarrow \infty$ .

### 3.2.5. COMPUTATIONAL COSTS AND ACCURACY

mation on the given points). Consider now how we can expect  $m(n, \epsilon)$  to depend on  $n$  and  $\epsilon$ : For fixed  $\epsilon$ , we can in the limit  $n \rightarrow \infty$ , for a singular  $K$ , expect a larger  $m(n, \epsilon)$  primarily to keep the normalised  $l^2$ -error below  $\epsilon$  as  $[\kappa\Delta_n, T]$  grows towards the singularity where exponential functions work relatively worse. This effect will not exist for non-singular kernels where we, per the recommendation for  $\kappa$ , will fit to the full interval  $[0, T]$ . For both singular and non-singular kernels, we can for a larger  $n$  expect a further increase in  $m(n, \epsilon)$  as the number of sampling points per time unit goes up; evidence though suggests that this effect is very small. For fixed  $n$ , we can in the limit  $\epsilon \rightarrow 0$ , in both cases, expect  $m(n, \epsilon)$  to grow to lower the error on the (here fixed) interval  $[\kappa\Delta_n, T]$ .

Although a natural conjecture is that both  $n$  and  $\epsilon$  should tend (to  $\infty$  resp. 0) for convergence, an experiment of Section 3.7 suggests that  $\hat{X}_t^{mn}$  converges to  $X_t$  significantly faster in  $\epsilon$  than in  $n$ . It is therefore our view that it practically suffices to keep  $\epsilon$  fixed at some low but reasonable value and only consider the convergence in  $n$ . To arrive at a complexity that we can compare with e.g. the hybrid scheme of [9] it then only remains to determine how  $m(n, \epsilon)$  scales in  $n$ . An experiment of Section 3.5 to this end suggest for a fixed  $\epsilon$  that  $m(n, \epsilon) = O(\log_2 n)$  when  $K$  is singular in the power-law sense,  $m(n, \epsilon) = O(1)$  when it is non-singular.<sup>11</sup> Consequently we can expect a practical complexity of  $O(n \log_2 n)$  when  $K$  is power-law singular and of  $O(n)$  when it is non-singular.

In what follows, we shall compare our scheme against the hybrid scheme of [9]. We therefore switch focus to the  $\mathcal{TBSS}$  process of (3.4) and temporarily let  $X_t$  refer to the stochastic integral of that equation. In [9] the notion that the kernel behaves like a power-law in the short-time limit is formalised by assuming  $K(t) = t^\alpha L_K(t)$  where  $L_K$  is slowly varying at zero. The latter implies that  $L_K$  is dominated by the power-law near the origin. Let  $\hat{X}_t^n$  denote the numerical approximation of  $X_t$  under their scheme. Then

$$\hat{X}_{t_i}^n = \sum_{k=\kappa+1}^i K\left(\frac{b_k^*}{n}\right) \hat{\sigma}_{t_i-k}^n \Delta W_{i-k}^n + \sum_{k=1}^{\min\{i, \kappa\}} L_K\left(\frac{k}{n}\right) \hat{\sigma}_{t_i-k}^n \tilde{W}_{i-k, k}^n, \quad i = 0, 1, \dots, \lfloor nT \rfloor, \quad (3.39)$$

where  $\kappa \in \{0, 1, \dots, \lfloor nT \rfloor\}$ ,  $\hat{\sigma}_t^n$  is an approximate or exact sample of  $\sigma_t$  for a given  $t$ , and  $b_k^* \in [k-1, k] \setminus \{0\}$ ,  $k = \kappa+1, \kappa+2, \dots$ , are optimal values from their Proposition 2.8. By convention the first sum is zero when  $i < \kappa+1$ , the second sum when  $i = 0$  or  $\kappa = 0$ . Also, in this context:

$$\tilde{W}_{i, k}^n = \int_{t_i}^{t_{i+1}} (t_{i+k} - s)^\alpha dW_s, \quad i = 0, 1, \dots, \lfloor nT \rfloor - 1, \quad k = 1, 2, \dots, \kappa.$$

Their scheme generalises in the obvious way to SVE's of the form (3.1).

From the second sum of (3.39), we see that they handle  $K$  near the origin much like we do—a slight difference is that they treat  $L_K$  piecewise constant whereas we treat  $K$  without error in its entirety. The authors of [9] too find small non-zero  $\kappa$ 's to be sufficient when  $\alpha < 0$ . Thus the second sum of (3.39) is of negligible cost. Their scheme mainly differs from ours in that the remainder of the stochastic integral in (3.4) is approximated

<sup>11</sup> Our experiments do not allow us to conjecture what the complexity is for kernels that are singular in a non power-law way.

by a Riemann sum; we refer to the first sum in (3.39). To compute the sums across all time points in general costs  $O(n^2)$ . However, if  $\{\hat{\sigma}_{t_i}^n\}_{i=0}^{\lfloor nT \rfloor - 1}$  can be pre-simulated that can be lowered to  $O(n \log_2 n)$  with the FFT algorithm; the reason is that we may view the first sum in (3.39) across  $i = 0, 1, \dots, \lfloor nT \rfloor$ , as a discrete convolution. The FFT algorithm applies e.g. in the Gaussian case where  $\sigma_t$  is deterministic, in particular constant.<sup>12</sup>

The experiments of Section 3.7 suggest that the scheme of [9] is about equally accurate to ours for the same  $n$ . To rank them in terms of performance, it then suffices to compare their computational costs. In the general case where the FFT cannot be used, the hybrid multifactor scheme is notably faster with running times of  $O(n)$  for non-singular  $K$  and of  $O(n \log_2 n)$  for power-law singular  $K$  whereas the scheme of [9] displays costs  $O(n^2)$ . Our scheme is likewise faster when  $K$  is non-singular and the FFT can be used ( $O(n)$  versus  $O(n \log_2 n)$  running time). At a glance, we should observe similar performance both with complexity  $O(n \log_2 n)$  when the FFT can be used and  $K$  is power-law singular. However, as the numerical tests of Section 3.7 show, the constant in front of  $O(n \log_2 n)$  is smaller under our scheme. The reason, we believe, is that  $m(n, \epsilon)$  even in the singular case scales only very slowly in  $n$ ; we refer to Section 3.5 for details. We conclude that our scheme should be preferred in all the discussed cases.

We end by highlighting a potential problem that could materialise when  $K$  is singular and which is related to the piecewise constant approximations we apply to the coefficient processes  $b_t$  and  $\sigma_t$ : Consider the stochastic integral  $\int_{t_{i-1}}^{t_i} K(t_i - s) \sigma_s dW_s$  for some  $i \in \{1, 2, \dots, \lfloor nT \rfloor\}$  and think of  $s \mapsto K(t_i - s)$  as giving weights to the terms " $\sigma_s dW_s$ ". When  $K$  is singular by far most of the weight is placed on the upper end of the integration interval  $[t_{i-1}, t_i]$ . It follows that the estimate

$$\int_{t_{i-1}}^{t_i} K(t_i - s) \sigma_s dW_s \approx \sigma_{t_{i-1}} \tilde{W}_{i-1,1}^n \quad (3.40)$$

could be very inaccurate. The same comment applies to our drift approximations, though, it is likely to be less of a problem as the drift-integral (by definition) is ordinary. Despite the warning, our experiments suggests that the above is only sometimes a notable problem. We refer to Section 3.7 for details.

### 3.3 The forward process

We discuss the forward process  $g$ . We start with some theoretical considerations to improve our understanding of the object. Our main results are that it under certain

<sup>12</sup> The discrete convolution between two vectors  $x$  and  $y$  can be written  $x * y = \text{DFT}^{-1}(\text{DFT}(x) \cdot \text{DFT}(y))$  where DFT is the discrete Fourier transform,  $\text{DFT}^{-1}$  its inverse; this is subject to zero padding. We refer to [41, Chapter 3] for details. If the lengths are of order  $O(n)$  the cost with the FFT is  $O(n \log_2 n)$ . However, as  $x$  and  $y$  has to be inputted in their entirety, we cannot in our context when  $\sigma_t$  is state-dependent use the FFT to achieve costs  $O(n \log_2 n)$  when evaluating the first sum of (3.39) across all time points. For that we need  $\hat{\sigma}_{t_i}^n, i = 0, \dots, \lfloor nT \rfloor - 1$ , pre-simulated. Although a little pedantic, let us also state the following to avoid misinterpretations: In the main text, we say that the FFT can (cannot) be used to mean that it can (cannot) be used with the vectors for the given convolution inputted in their entirety *at once*; technically we can always use the FFT iteratively to compute the convolution, though, the cost of that is worse than  $O(n^2)$ .

conditions represents a Markov state and under other conditions are the conditional expectations of  $X$ . We end by showing how to compute it under the scheme.

### 3.3.1 Markov property, conditional expectations, and more

For our theoretical results, we need assumptions that guarantee a solution to (3.1). We start with a hypothesis for  $K$ .

**Hypothesis ( $H_0$ ):**  $K \in L^2_{\text{loc}}(\mathbb{R}_+)$  and there exists a  $\tilde{\gamma} \in (0, 2]$  so  $\int_0^h K(t)^2 dt = O(h^{\tilde{\gamma}})$  and  $\int_0^{\tilde{T}} (K(t+h) - K(t))^2 dt = O(h^{\tilde{\gamma}})$  for every  $\tilde{T} \in [0, \infty)$ .

We view ( $H_0$ ) as a mild condition. It is e.g. satisfied by the rough fractional kernel  $K(t) = ct^\alpha$ ,  $c \geq 0$ ,  $\alpha \in (-\frac{1}{2}, 0]$ , with  $\tilde{\gamma} = 2\alpha + 1$ , and by any kernel that is locally Lipschitz on  $[0, \infty)$  with  $\tilde{\gamma} = 1$ ; we refer to [5, Example 2.3] for these facts. Note that kernels that are continuously differentiable on  $[0, \infty)$ , e.g. kernels that are CM on  $[0, \infty)$ , are locally Lipschitz on  $[0, \infty)$ . From [5, Example 2.3] we also have the following: If  $K_1$  satisfies ( $H_0$ ) with some  $\tilde{\gamma}$  and  $K_2$  is locally Lipschitz then  $K_1 K_2$  satisfies ( $H_0$ ) with the same  $\tilde{\gamma}$ . Also, if  $K_1$  and  $K_2$  satisfies ( $H_0$ ) with some common  $\tilde{\gamma}$  then so does  $K_1 + K_2$ .

Define

$$H^\phi := \{g_0 : \mathbb{R}_+ \rightarrow \mathbb{R}, \text{ locally H\"older continuous of all orders strictly less than } \phi\}.$$

By Theorem A.1 of [3] (for reference see also Theorem 3.B.1 in our appendix) the below assumptions guarantee a unique continuous strong (under part (i)) or continuous weak (under part (ii)) solution  $X$  to (3.1).

**Assumption 3.3.1.**  $K$  satisfies ( $H_0$ ) for some  $\tilde{\gamma} \in (0, 2]$  and  $g_0 \in \mathcal{H}^{\tilde{\gamma}/2}$ . Additionally at least one of the following statements hold: (i)  $b$  and  $\sigma$  are Lipschitz continuous,<sup>13</sup> (ii)  $b$  and  $\sigma$  are continuous with linear growth and  $K$  has a resolvent of first kind  $L$ .

Invoke Assumption 3.3.1 and fix a continuous solution  $X$  which exists. Define

$$\zeta_t(u) := g_0(u) + \int_0^t K(u-s)b(X_s)ds + \int_0^t K(u-s)\sigma(X_s)dW_s, \quad 0 \leq t \leq u. \quad (3.41)$$

Since  $g_t(\tau) = \zeta_t(t+\tau)$  for  $t, \tau \geq 0$ , the above is just a reformulation of  $g$ . Both are well-defined for the following reasons: The stochastic integral of (3.41) is well-defined as

$$\int_0^t K(u-s)^2 \sigma(X_s)^2 ds < \infty, \quad 0 \leq t \leq u,$$

which holds since  $X$  and  $\sigma$  are continuous and  $K \in L^2_{\text{loc}}(\mathbb{R}_+)$ . The Lebesgue integral of (3.41) is well-defined as  $X$  and  $b$  are continuous and  $K \in L^1_{\text{loc}}(\mathbb{R}_+)$  especially.

While  $(g_t(\tau))_{t \geq 0}$  for a fixed  $\tau$  is not generally a semimartingale, the process  $(\zeta_t(u))_{t \in [0, u]}$

<sup>13</sup> Lipschitz continuity in particular implies linear growth, i.e. the existence of a positive constant  $C$  so for all  $x \in \mathbb{R}$ :  $|b(x)| + |\sigma(x)| \leq C(1 + |x|)$ . We use this fact in many of our proofs.

for a fixed  $u$  is. Despite this, we will formulate our results entirely in terms of  $g$ .

In the following, we argue that  $g$  may be Markov. It will be useful to treat the case of a non-negative solution  $X$  separately and we therefore briefly consider additional assumptions that guarantee the existence of such. We start with an extra hypothesis for  $K$ .

**Hypothesis  $(H_1)$ :**  $K$  is non-negative, not identically zero, non-increasing and continuous on  $(0, \infty)$ , and its resolvent of first kind  $L$  is non-negative and non-increasing in the sense that  $s \mapsto L([s, s + t])$  is non-increasing for all  $t \geq 0$ .

We shall always use  $(H_1)$  together with  $(H_0)$ . The assumptions that  $K$  is non-negative, not identically zero, and non-increasing on  $(0, \infty)$ , combined with the condition  $K \in L_{\text{loc}}^2(\mathbb{R}_+)$  of  $(H_0)$  by [27, Theorem 5.5.5] guarantees that the resolvent  $L$  of  $(H_1)$  exists.

The hypothesis  $(H_1)$  is satisfied by all  $K \in L_{\text{loc}}^2(\mathbb{R}_+)$  that are CM on  $(0, \infty)$  and not identically zero; use [27, Theorem 5.5.4]. This was also pointed out in [5, Example 3.7].

Consider now the following condition for  $g_0$  which is meaningful under  $(H_0)$ – $(H_1)$ :

$$\Delta_h g_0 - (\Delta_h K * L)(0)g_0 - d(\Delta_h K * L) * g_0 \geq 0, \quad h \geq 0. \quad (3.42)$$

Define also:

$$\mathcal{H}_+^{\tilde{\gamma}/2} := \{g_0 \in \mathcal{H}^{\tilde{\gamma}/2} \text{ satisfying (3.42) and } g_0(0) \geq 0\}.$$

If Assumption 3.3.1 holds,  $K$  satisfies  $(H_1)$ , and also  $g_0 \in \mathcal{H}_+^{\tilde{\gamma}/2}$ ,  $b(0) \geq 0$ ,  $\sigma(0) = 0$ , then (3.1) has a continuous *non-negative* weak solution  $X$  as follows by Theorem 3.B.2.

The reader may think that (3.42) looks somewhat abstract (we would concur). In [3, Example 2.2] the following was however proved: Say that  $K$  satisfies  $(H_0)$ – $(H_1)$  and that  $g_0 = X_0 + K * \theta$  with  $X_0 \geq 0$  and  $\theta \in L_{\text{loc}}^2(\mathbb{R}_+)$  such that  $\theta(t)dt + X_0L(dt)$  is a non-negative measure. Then  $g_0 \in \mathcal{H}_+^{\tilde{\gamma}/2}$ . Explicit expressions for  $L$  can be found in [5, Table 1] for some specific kernels that include the fractional one. Non-negativity of  $\theta(t)dt + X_0L(dt)$  can then in many cases be ensured with an explicit inequality for  $\theta$ .

Define for  $t_0, t \geq 0$ :  $X_t^{t_0} := X_{t_0+t}$ ,  $g_t^{t_0} := g_{t_0+t}$ ,  $W_t^{t_0} := W_{t_0+t} - W_{t_0}$ ,  $\mathcal{F}_t^{t_0} := \mathcal{F}_{t_0+t}$ .

The result below is key to make our point about  $g$  being Markov. It effectively generalises [3, Theorem 3.1] for the Volterra Heston model. It is not a complete generalisation by technicality as the authors of [3] consider the joint system of an asset price and its instantaneous variance—as modelled by the Volterra Heston equation—and we do not.<sup>14</sup>

**Theorem 3.3.2.** *Invoke Assumption 3.3.1 and let  $X$  be a continuous solution of (3.1). Fix  $t_0 \geq 0$ . Then*

$$X_t^{t_0} = g_{t_0}(t) + \int_0^t K(t-s)b(X_s^{t_0})ds + \int_0^t K(t-s)\sigma(X_s^{t_0})dW_s^{t_0}, \quad t \geq 0, \quad (3.43)$$

<sup>14</sup> The  $g$ -process of [3] is expressed as a conditional expectation but does match that of our paper.

### 3.3.1. MARKOV PROPERTY, CONDITIONAL EXPECTATIONS, AND MORE

where  $(W_t^{t_0})_{t \geq 0}$  is Brownian motion with respect to  $(\mathcal{F}_t^{t_0})_{t \geq 0}$  which also satisfies the usual hypothesis and  $g_{t_0} \in \mathcal{H}^{\tilde{\gamma}/2}$  is  $\mathcal{F}_{t_0}$ -measurable. More generally, we have for  $t, \tau \geq 0$ :

$$g_t^{t_0}(\tau) = g_{t_0}(t + \tau) + \int_0^t K(t + \tau - s)b(X_s^{t_0})ds + \int_0^t K(t + \tau - s)\sigma(X_s^{t_0})dW_s^{t_0}. \quad (3.44)$$

Assume additionally that  $K$  satisfies  $(H_1)$  and that  $g_0 \in \mathcal{H}_+^{\tilde{\gamma}/2}$ ,  $b(0) \geq 0$ ,  $\sigma(0) = 0$ , and  $X$  is non-negative. Then  $g_{t_0} \in \mathcal{H}_+^{\tilde{\gamma}/2}$ .

*Proof.* The proof is a simple extension of that for [3, Theorem 3.1]. We refer to the appendix for the details.  $\square$

As in the theorem, invoke Assumption 3.3.1 and let  $X$  be a continuous solution of (3.1) on some filtered probability space  $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$ . Let also  $t_0 \geq 0$ . Note that conditionally on  $\mathcal{F}_{t_0}$ , Equation (3.43) for  $X^{t_0} = (X_t^{t_0})_{t \geq 0}$  viewed with the shifted filtration  $(\mathcal{F}_t^{t_0})_{t \geq 0}$  resembles Equation (3.1) for  $X$  but with an updated initial curve  $g_{t_0}$ . It should here be noted that  $g_{t_0}$  is  $\mathcal{F}_{t_0}$ -measurable and lies in  $\mathcal{H}^{\tilde{\gamma}/2}$  like  $g_0$  is required to and that  $W^{t_0} = (W_t^{t_0})_{t \geq 0}$  is Brownian motion under the shifted filtration. It suggests that the dynamics of  $X^{t_0}$  conditionally on  $\mathcal{F}_{t_0}$  can also be obtained by restarting the process in the original equation (3.1) with ' $g_{t_0}$  in place of  $g_0$ '. In other words, Theorem 3.3.2 alludes to  $g$  as a Markov state. We discuss this prospect in more detail in the coming paragraphs.

As a start assume additionally that  $b, \sigma$  are Lipschitz continuous, i.e. that specifically part (i) of Assumption 3.3.1 holds. We then conjecture that the suggested Markov property holds: By Theorem 3.B.1 part (i), which now for sure applies, the equation

$$\tilde{X}_t^{t_0} = g_{t_0}(t) + \int_0^t K(t - s)b(\tilde{X}_s^{t_0})ds + \int_0^t K(t - s)\sigma(\tilde{X}_s^{t_0})dW_s^{t_0}, \quad t \geq 0, \quad (3.45)$$

has a unique continuous strong solution  $\tilde{X}^{t_0} = (\tilde{X}_t^{t_0})_{t \geq 0}$  on the filtered probability space  $(\Omega, \mathcal{F}, (\mathcal{F}_t^{t_0})_{t \geq 0}, \mathbb{P})$  and where we stress that  $W^{t_0}$  in (3.45) is the same as in (3.43). By strong uniqueness  $\tilde{X}^{t_0}$  coincides with  $X^{t_0}$ . The inputs we have used in Theorem 3.B.1 to conclude the solution  $\tilde{X}^{t_0}$  which are not fixed are  $g_{t_0}$  which is  $\mathcal{F}_{t_0}$ -measurable and  $W^{t_0}$  which is independent of  $\mathcal{F}_{t_0}$ . It follows that we can expect the distribution of  $X^{t_0}$  conditionally on  $\mathcal{F}_{t_0}$  to be fully determined by  $g_{t_0}$ . Stated differently:  $g$  should be a Markov state for  $X$ . By (3.44) we can also expect  $g$  to be a Markov state for itself.<sup>15</sup>

A different strategy to establish the Markov property relies on weak uniqueness. Such an argument was carried out in [3] for the Volterra Heston model. We briefly outline it in our own words but only for the Volterra Heston equation itself, i.e. not the joint system with an asset price: The Volterra Heston equation is defined by assuming that  $K$  satisfies  $(H_0)$ – $(H_1)$  and that  $g_0 \in \mathcal{H}_+^{\tilde{\gamma}/2}$ ,  $b(x) = -\lambda x$ ,  $\sigma(x) = \nu\sqrt{x}$ ,  $x \geq 0$ ,  $(\lambda, \nu) \in \mathbb{R}_+^2$ . A continuous non-negative *weak* solution  $X$  exists by Theorem 3.B.2. From the last part of Theorem 3.3.2 (alternatively their Theorem 3.1) we have  $g_{t_0} \in \mathcal{H}_+^{\tilde{\gamma}/2}$  which means that (3.45) also has a continuous non-negative weak solution  $\tilde{X}^{t_0}$ . Although it need

<sup>15</sup> One should ideally formalise beyond what is provided here, though, we do not pursue it. For reference, a formal proof of the Markov property for SDE's with  $b, \sigma$  Lipschitz is that of [53, Theorem 7.1.2].

not be driven by the same Brownian motion as  $X$ , or even live on the same probability space, as it is weak, the authors prove *weak uniqueness* for  $X$  and  $g$  by characterising the Fourier-Laplace transform, respectively, the characteristic functional. They are then able to conclude that  $g$  is a Markov state.

We find it likely that the Markov property should hold in much larger generality than the Lipschitz and Volterra Heston cases, though, the existence and uniqueness results that are currently available in the literature puts a limit on what can be argued for now.

The other useful property of the forward process is that whenever the drift is deterministic (in the sense that  $b = 0$ ) it also represents the conditional expectations of  $X$ :

**Theorem 3.3.3.** *Invoke Assumption 3.3.1 and assume that  $b = 0$ . Let  $X$  be a continuous solution of (3.1). Then  $E_t(X_{t+\tau}) = g_t(\tau)$ ,  $t, \tau \geq 0$ .*

*Proof.* By linear growth of  $\sigma$ , the moment bound (3.141), and that  $K \in L^2_{\text{loc}}(\mathbb{R}_+)$ , we have with the use of Jensen's inequality and Tonelli's theorem

$$E \left( \int_0^t K(u-s)^2 \sigma(X_s)^2 ds \right) \leq C \left( 1 + \sup_{s \leq t} E(|X_s|^2) \right) \|K\|_{L^2([0,u])}^2 < \infty, \quad 0 \leq t \leq u,$$

where  $C$  is a positive (finite) constant. Consequently, for a fixed  $u \geq 0$ , the process

$$\left( \int_0^t K(u-s) \sigma(X_s) dW_s \right)_{t \in [0,u]} \quad (3.46)$$

is a true martingale. It follows that

$$E_t(X_u) - g_0(u) = E_t \left( \int_0^u K(u-s) \sigma(X_s) dW_s \right) = \int_0^t K(u-s) \sigma(X_s) dW_s, \quad 0 \leq t \leq u.$$

To complete the proof, rearrange, recognise  $\zeta_t(u) = g_t(u-t)$ , and set  $\tau = u-t$ .  $\square$

If we are in a financial context and  $X$  is the instantaneous variance of an asset, its conditional expectations are known to as *forward variances*. Integrals of these are related directly to important financial objects like variance swaps and the VIX index. We consider a class of instantaneous variance models of the form (3.1) with  $b = 0$  in Section 3.6.3.

The next theorem shows that  $b = 0$  is not as restrictive as one might initially think.

**Theorem 3.3.4.** *Consider the equation*

$$X_t = X_0 + \int_0^t K(t-s) \lambda (\theta(s) - X_s) ds + \int_0^t K(t-s) \sigma(X_s) dW_s, \quad t \geq 0, \quad (3.47)$$

where  $X_0 = x_0 \in \mathbb{R}$ ,  $\lambda > 0$ ,  $K$  satisfies  $(H_0)$  for some  $\tilde{\gamma} \in (0, 2]$ ,  $\theta \in L^2_{\text{loc}}(\mathbb{R}_+)$ , and  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  is continuous of linear growth. Then  $g_0 := X_0 + \lambda K * \theta \in \mathcal{H}^{\tilde{\gamma}/2}$ . If also:

(i)  $\sigma$  is Lipschitz continuous then (3.47) has a unique continuous strong solution  $X$ .

### 3.3.1. MARKOV PROPERTY, CONDITIONAL EXPECTATIONS, AND MORE

---

(ii)  $K$  has a resolvent of first kind  $L$  then (3.47) has a continuous weak solution  $X$ .

Let (i) or (ii) additionally hold and  $X$  be a continuous solution of (3.47). Then

$$X_t = \tilde{g}_0(t) + \frac{1}{\lambda} \int_0^t R_\lambda(t-s) \sigma(X_s) dW_s, \quad t \geq 0, \quad (3.48)$$

where  $R_\lambda \in L_{loc}^2(\mathbb{R}_+)$  is the resolvent of second kind of  $\lambda K$  and  $\tilde{g}_0 := X_0 + R_\lambda * (\theta - X_0)$ . Moreover, it holds that  $R_\lambda$  satisfies  $(H_0)$  with the same  $\tilde{\gamma}$  as  $K$  and  $\tilde{g}_0 \in \mathcal{H}^{\tilde{\gamma}/2}$ .

Say in addition that  $K$  is CM and not identically zero,  $\sigma(0) = 0$ ,  $\lambda\theta(t)dt + X_0L(dt)$  is a non-negative measure, and  $X_0 \geq 0$ . Then  $K$  has a resolvent of first kind  $L$  and satisfies  $(H_1)$ ,  $g_0 \in \mathcal{H}_+^{\tilde{\gamma}/2}$ , and (3.47) has a continuous non-negative weak solution  $X$ .

Furthermore, it holds that  $R_\lambda$  is CM and not identically zero with a resolvent of first kind given by  $\tilde{L}(dt) = \lambda^{-1}L(dt) + dt$  for which it satisfies  $(H_1)$ , that  $(\theta(t) - X_0)dt + X_0\tilde{L}(dt)$  is a non-negative measure, and  $\tilde{g}_0 \in \mathcal{H}_+^{\tilde{\gamma}/2}$ .

*Proof.* We refer to the appendix. □

*Remark 3.3.1.* Similar results already exists for the rough Heston equation elsewhere in the literature; see e.g. [18, Proposition 6.3].

Theorems 3.3.3-4 suggest that if we have an equation with linear drift of the form (3.47) and desire numerical computation of the conditional expectations, it may be worthwhile to reformulate it to (3.48). More precisely, it then suffices to evaluate the forward values to which end we provide a numerical approximation in the next subsection. Note that many properties of original equation (3.47) hold for (3.48) also: The initial curve  $\tilde{g}_0$  is e.g. yet locally Hölder continuous of orders less than  $\tilde{\gamma}/2$  and  $\frac{1}{\lambda}R_\lambda$  is (like  $K$ ) locally square integrable and satisfies  $(H_0)$  with the same  $\tilde{\gamma}$ ; the  $\frac{1}{\lambda}$  scaling does not change this. The results of Sections 3.4 and 3.7 show that  $\tilde{\gamma}$  controls the convergence rate for the strong error under the hybrid multifactor scheme, though, there are many nuances to this statement, see later. Consequently, whether we solve (3.47) or (3.48) with the scheme we can expect the same asymptotic convergence speed. For this whole discussion, it should be noted that if  $K$  is CM then so is  $\frac{1}{\lambda}R_\lambda$  [27, Theorem 5.3.1] and therefore if we can apply the hybrid multifactor scheme to (3.47) we can also apply it to (3.48).<sup>16</sup>

For practical use of (3.48) we need to know  $R_\lambda$ . While analytical expressions are not always available, they are in certain cases. Consider the rough fractional kernel

$$K(t) = \frac{t^{\alpha-1}}{\Gamma(\alpha)}, \quad t > 0, \quad \alpha \in (1/2, 1], \quad (3.49)$$

which satisfies  $(H_0)$ . As is common in some parts of the literature, we have let the roughness parameter  $\alpha$  lie in  $(\frac{1}{2}, 1]$  instead of  $(-\frac{1}{2}, 0]$ . By [5, Table 1] (see also Proposition 3.A.1 in our appendix) the resolvent of second kind of  $\lambda K$  under (3.49) is given by

$$R_\lambda(t) = \lambda t^{\alpha-1} E_{\alpha,\alpha}(-\lambda t^\alpha), \quad t > 0, \quad (3.50)$$

<sup>16</sup> Note that (3.47) is nested in (3.1) with  $g_0 = X_0 + \lambda K * \theta$ ,  $b(x) = -\lambda x$ , and where  $\sigma, K$  are the same.

where

$$E_{\alpha,\beta}(z) = \sum_{n=0}^{\infty} \frac{z^n}{\Gamma(\alpha n + \beta)}, \quad z, \alpha, \beta \in \mathbb{C}, \quad \operatorname{Re}(\alpha), \operatorname{Re}(\beta) > 0, \quad (3.51)$$

is the so-called *Mittag-Leffler* function.<sup>17</sup>

For the general case, note that  $R_\lambda$  by definition solves

$$R_\lambda = \lambda K - \lambda K * R_\lambda. \quad (3.52)$$

Consequently, if we assume that  $K$  is CM on  $[0, \infty)$ , we can obtain  $R_\lambda$  by solving (3.52) with the scheme where  $\lambda K$  (also) takes the role of ' $g_0$ ' compared with (3.1). If  $K$  is CM on  $(0, \infty)$  and singular, we are bound to need a modified version of the scheme to properly handle the initial curve which now also is singular. We leave this for a future project.

In what follows, we shed more light on how  $\frac{1}{\lambda}R_\lambda$  relates to  $K$ . We will assume that  $K$  is CM and will also use that  $K \in L^2_{\text{loc}}(\mathbb{R}_+)$  as is nested in  $(H_0)$ . Note first that (3.52) implies  $\frac{1}{\lambda}R_\lambda = K - K * R_\lambda$ . Since  $K$  and  $\frac{1}{\lambda}R_\lambda$  are CM they are non-negative. It follows that  $0 \leq \frac{1}{\lambda}R_\lambda \leq K$ , i.e. the scaled resolved kernel is bounded by the original one. But we can say more: By [27, Theorem 5.3.1] we have  $R_\lambda \in L^1(\mathbb{R}_+)$  which, because  $R_\lambda$  is non-negative and non-increasing as follows by complete monotonicity, implies  $R_\lambda(t) \rightarrow 0$  in the limit  $t \rightarrow \infty$ .<sup>18</sup> Moreover, as  $K \in L^2_{\text{loc}}(\mathbb{R}_+)$  it holds that  $R_\lambda \in L^2_{\text{loc}}(\mathbb{R}_+)$  by Remark 3.A.1 of the appendix. Choose now a number  $t^* > 0$  so  $R_\lambda(t) \leq 1$  for all  $t \geq t^*$  as is possible since  $R_\lambda$  decays to zero. By  $R_\lambda \in L^2_{\text{loc}}(\mathbb{R}_+)$  and  $R_\lambda \in L^1(\mathbb{R}_+)$  it then follows that  $\|R_\lambda\|_{L^2(\mathbb{R}_+)}^2 \leq \|R_\lambda\|_{L^2([0,t^*])}^2 + \|R_\lambda\|_{L^1([t^*,\infty))}^2 < \infty$  and we conclude that  $\frac{1}{\lambda}R_\lambda \in L^2(\mathbb{R}_+)$ .

The above shows that if  $K \in L^2_{\text{loc}}(\mathbb{R}_+)$  is CM then  $\frac{1}{\lambda}R_\lambda$  is dampened compared to  $K$ , both pointwise and in terms of square integrability. For the latter, note that  $K$  defined by (3.49) has  $\|K\|_{L^2(\mathbb{R}_+)} = \infty$  (recall that  $\int_1^\infty x^\beta dx = \infty$  when  $\beta \geq -1$ ) whereas we just argued that  $\frac{1}{\lambda}R_\lambda \in L^2(\mathbb{R}_+)$  always when  $K$  is CM and locally square integrable. To highlight a practical implication, consider (3.47)-(3.48) in the Gaussian case where  $\sigma$  is constant. Under the assumptions on  $K$  then  $\sup_{t \geq 0} \operatorname{Var}(X_t) \propto \|R_\lambda\|_{L^2(\mathbb{R}_+)}^2 < \infty$  always when  $\lambda > 0$  whereas  $\sup_{t \geq 0} \operatorname{Var}(X_t) \propto \|K\|_{L^2(\mathbb{R}_+)}^2$  when  $\lambda = 0$  and which could equal  $\infty$  as we noted for (3.49). We suspect that uniformly (or equivalently: *asymptotically*) bounded variance holds more generally when  $\lambda > 0$  although we will not attempt a proof.

Theorem 3.3.4 and the above discussion suggests that when  $K$  is CM, the effect of linear drift is mainly that of dampening the original kernel, perhaps most importantly to get square integrability, *and* to create a mean term structure via the relation  $\tilde{g}_0 = X_0 + R_\lambda * (\theta - X_0)$ . However, there is no loss of generality in modelling directly on the form (3.1) with a general initial curve  $g_0 \in \mathcal{H}^{\tilde{\gamma}/2}$  and the restrictions  $b = 0$  and  $K \in L^2(\mathbb{R}_+)$ . In fact, we expect this to be more practical as the conditional expectations, if we desire the numerical computation of these, then immediately coincide with

<sup>17</sup> For our numerical experiments, we use the implementation <https://www.mathworks.com/matlabcentral/fileexchange/48154-the-mittag-leffler-function> coded by Roberto Garrappa (retrieved December 1, 2021).

<sup>18</sup> If not  $\lim_{t \rightarrow \infty} R_\lambda(t) = 0$  then  $R_\lambda(t) \geq c$  for some  $c > 0$  for all  $t > 0$  which contradicts  $R_\lambda \in L^1(\mathbb{R}_+)$ .

### 3.3.1. MARKOV PROPERTY, CONDITIONAL EXPECTATIONS, AND MORE

---

the forward values without having to first determine  $R_\lambda$  and rewrite (3.47) to (3.48).

Given that the forward process in many cases represents a Markov state and/or the conditional expectations of the underlying process, we see there can be good value in being able to compute it. We will need the forward process in Section 3.6 when we show how to compute the VIX index for a number of volatility models of the Volterra type.

There is a final theorem that we wish to state in this subsection. Before we can show it, we need another assumption on  $K$ .

**Hypothesis ( $H_2^*$ ):**  $K$  is CM on  $(0, \infty)$  with representation (3.7) for a non-negative Borel measure  $\mu$  that satisfies  $\int_{[0, \infty)} (1 \wedge \gamma^{-\frac{1}{2}}) \mu(d\gamma) < \infty$ .

We have labelled the assumption ( $H_2^*$ ) as the integrability criterion is similar to that of hypothesis ( $H_2$ ) of [3]; the latter in fact implies the former. It should be noted that ( $H_2^*$ ) is satisfied by any kernel that is CM on  $[0, \infty)$  since  $\mu$  then is finite as remarked in footnote 5 and  $(1 \wedge \gamma^{-\frac{1}{2}}) \leq 1$  for all  $\gamma \geq 0$ . By Table 1 of [3] ( $H_2^*$ ) is also satisfied by the gamma kernel which we define by

$$K(t) = e^{-\lambda t} \frac{t^{\alpha-1}}{\Gamma(\alpha)}, \quad t > 0, \quad \lambda \geq 0, \quad \alpha \in (1/2, 1]. \quad (3.53)$$

It is therefore our expectation that ( $H_2^*$ ) is a mild assumption on  $K$ .

We can now state the theorem. Under the given conditions, it shows that  $g_t$ , in particular  $X_t$ , can be expressed in terms of an infinite set of mean-reverting semimartingales. We will need the result in Section 3.6.3 when we derive an equation for the VIX index for a class of Volterra volatility models with deterministic drift. The theorem also hints at the numerical approximation of  $g$  that we will present in the next subsection. Similar but less general results exist in [2, 3, 4]; for the purely Gaussian case see also [14, 29].

**Theorem 3.3.5.** *Invoke Assumption 3.3.1 and assume that  $K$  satisfies ( $H_2^*$ ). Let  $X$  be a continuous solution of (3.1). Define*

$$U_t(\gamma) := \int_0^t e^{-\gamma(t-s)} b(X_s) ds + \int_0^t e^{-\gamma(t-s)} \sigma(X_s) dW_s, \quad t, \gamma \geq 0.$$

For any  $\gamma \geq 0$  then

$$dU_t(\gamma) = (b(X_t) - \gamma U_t(\gamma)) dt + \sigma(X_t) dW_t, \quad U_0(\gamma) = 0. \quad (3.54)$$

Furthermore

$$g_t(\tau) = g_0(t + \tau) + \int_{[0, \infty)} e^{-\gamma\tau} U_t(\gamma) \mu(d\gamma), \quad t, \tau \geq 0. \quad (3.55)$$

In particular

$$X_t = g_0(t) + \int_{[0, \infty)} U_t(\gamma) \mu(d\gamma), \quad t \geq 0. \quad (3.56)$$

*Proof.* See the appendix. □

Note that (3.56) resembles (3.37) (our approximation of  $X$  when  $\kappa = 0$ ). The approximation we momentarily will suggest for  $g$  will analogously resemble (3.55) when  $\kappa = 0$ .<sup>19</sup>

### 3.3.2 Numerical approximation

We will denote our approximation of  $g_t(\tau)$  by  $\hat{g}_t^{mn}(\tau)$ . It will not be justified for any  $\tau \geq 0$  even if  $t \in [0, T]$ . Instead the set of valid  $\tau$ 's will depend on the domain where  $K_{mn}$  approximates  $K$ . To allow explicit control of the set of valid  $\tau$ 's, we introduce a *forward horizon*  $T^* \geq T$  and will now assume that  $K_{mn}$  approximates  $K$  on  $[0, T^*]$ , i.e. that  $K_m$  approximates  $K$  on  $[\kappa\Delta_n, T^*]$ . The definition of  $g_t(\tau)$  (c.f. (3.5)) for given  $t, \tau \geq 0$  uses values of  $K$  on the interval  $[\tau, \tau + t]$ . Therefore if  $K_{mn}$  approximates  $K$  on  $[0, T^*]$  and we have simulated  $(\hat{X}_t^{mn}, \hat{U}_t^{mn})_{t \in [0, T]}$ , we should be able to approximate  $g_t(\tau)$  for values  $t, \tau \geq 0$  that satisfy  $t \leq T$  and  $\tau + t \leq T^*$ . Below we define formally the set of points  $(t, \tau)$  where the approximation  $g_t(\tau) \approx \hat{g}_t^{mn}(\tau)$  will be justified:

$$A := \{(t, \tau) \in \mathbb{R}_+^2 : t \leq T \text{ and } \tau + t \leq T^*\}. \quad (3.57)$$

To define the numerical solution  $\hat{g}^{mn} = (\hat{g}_t^{mn}(\tau))_{(t, \tau) \in A}$ , we consider first the forward values that arise under (3.8):

$$g_t^{mn}(\tau) := g_0(t + \tau) + \int_0^t K_{mn}(t + \tau - s) (b_s^{mn} ds + \sigma_s^{mn} dW_s), \quad t, \tau \geq 0. \quad (3.58)$$

The above is well-defined because  $X^{mn}$  is continuous and adapted, and since  $b, \sigma$  are also continuous, and  $K_{mn} \in L_{\text{loc}}^2(\mathbb{R}_+)$  which holds due to the fact that  $K, K_m \in L_{\text{loc}}^2(\mathbb{R}_+)$ .

Say that  $(t, \tau) \in A$  are so  $\tau \geq \kappa\Delta_n$ . Then

$$g_t^{mn}(\tau) = g_0(t + \tau) + \int_0^t K_m(t + \tau - s) b_s^{mn} ds + \int_0^t K_m(t + \tau - s) \sigma_s^{mn} dW_s \quad (3.59)$$

$$= g_0(t + \tau) + \sum_{i=1}^m c_i^m e^{-\gamma_i^m \tau} U_{i,t}^{mn}. \quad (3.60)$$

Next let  $(t, \tau) \in A$  such that  $\tau \in [0, \kappa\Delta_n]$ . Then

$$g_t^{mn}(\tau) = g_0(t + \tau) + \int_0^{(t+\tau-\kappa\Delta_n)^+} K_m(t + \tau - s) (b_s^{mn} ds + \sigma_s^{mn} dW_s) \quad (3.61)$$

$$+ \int_{(t+\tau-\kappa\Delta_n)^+}^t K(t + \tau - s) (b_s^{mn} ds + \sigma_s^{mn} dW_s) \quad (3.62)$$

$$= g_0(t + \tau) + \sum_{i=1}^m c_i^m e^{-\gamma_i^m \kappa\Delta_n} U_{i,(t+\tau-\kappa\Delta_n)^+}^{mn} \quad (3.63)$$

---

<sup>19</sup> Equations (3.55)-(3.56) also allude to  $(U_t(\gamma))_{\gamma \geq 0}$  as a Markov state for  $g_t$  and  $X_t$ . We refrain from making that precise but the reader should note that the Volterra Heston equation is treated in [3].

### 3.4. STRONG CONVERGENCE

$$+ \int_{(t+\tau-\kappa\Delta_n)^+}^t K(t+\tau-s)(b_s^{mn} ds + \sigma_s^{mn} dW_s). \quad (3.64)$$

Let  $i \in \{0, 1, \dots, \lfloor nT \rfloor\}$  and say that  $\tau = l\Delta_n$  for  $l \in \{0, 1, \dots, \kappa\}$ . By (3.61)-(3.64) then

$$\begin{aligned} g_{t_i}^{mn}(\tau) &\approx g_0(t_{i+l}) + \sum_{j=1}^m c_j^m e^{-\gamma_j^m \kappa \Delta_n} U_{j, (t_{i+l}-\kappa\Delta_n)^+}^{mn} \\ &+ \sum_{k=1}^{\min\{i, \kappa-l\}} \left( b_{t_{i-k}}^{mn} \int_{t_{i-k}}^{t_{i-k+1}} K(t_{i+l}-s) ds + \sigma_{t_{i-k}}^{mn} \int_{t_{i-k}}^{t_{i-k+1}} K(t_{i+l}-s) dW_s \right) \\ &= g_0(t_{i+l}) + \sum_{j=1}^m c_j^m e^{-\gamma_j^m \kappa \Delta_n} U_{j, (t_{i+l}-\kappa\Delta_n)^+}^{mn} + \sum_{k=1}^{\min\{i, \kappa-l\}} \left( b_{t_{i-k}}^{mn} w_{k+l} + \sigma_{t_{i-k}}^{mn} \tilde{W}_{i-k, k+l}^n \right) \end{aligned}$$

where the second sums should be interpreted as zero when  $i = 0$  or  $l = \kappa$ .

The just stated derivations and approximations forms the basis of our numerical approximation  $\hat{g}_t^{mn}(\tau)$ , which as noted is valid for  $(t, \tau) \in A$ . We define it formally below:

**Definition 3.3.1.** Let  $i \in \{0, 1, \dots, \lfloor nT \rfloor\}$  and  $\tau$  be such that  $(t_i, \tau) \in A$ . Define then

$$\hat{g}_{t_i}^{mn}(\tau) := \begin{cases} g_0(t_i + \tau) + \sum_{j=1}^m c_j^m e^{-\gamma_j^m \tau} \hat{U}_{j, t_i}^{mn}, & \tau \geq \kappa\Delta_n, \\ g_0(t_{i+l}) + \sum_{j=1}^m c_j^m e^{-\gamma_j^m \kappa \Delta_n} \hat{U}_{j, (t_{i+l}-\kappa\Delta_n)^+}^{mn} \\ + \sum_{k=1}^{\min\{i, \kappa-l\}} \left( \hat{b}_{t_{i-k}}^{mn} w_{k+l} + \hat{\sigma}_{t_{i-k}}^{mn} \tilde{W}_{i-k, k+l}^n \right), & \tau = l\Delta_n, \quad l \in \{0, 1, \dots, \kappa\}, \\ \frac{\Delta_n - (\tau - \tau_n^-)}{\Delta_n} \hat{g}_{t_i}^{mn}(\tau_n^-) + \frac{\tau - \tau_n^-}{\Delta_n} \hat{g}_{t_i}^{mn}(\tau_n^+), & \tau \in [0, \kappa\Delta_n] \end{cases}$$

and extend the definition as

$$\hat{g}_t^{mn}(\tau) := \hat{g}_{t_n^-}^{mn}(\tau), \quad (t, \tau) \in A. \quad (3.65)$$

□

The piecewise linear interpolation that is used to define  $\hat{g}_{t_i}^{mn}(\tau)$  for  $i \in \{0, 1, \dots, \lfloor nT \rfloor\}$  and  $\tau \in [0, \kappa\Delta_n \wedge (T^* - t_i)]$  is justified under Assumption 3.3.1 if  $X$  is continuous since  $g_{t_i}(\cdot)$  then likewise is by Theorem 3.3.2. The reader may check that Definition 3.3.1 is consistent with itself where there is overlap in the specification of  $\hat{g}^{mn}$ . It is also consistent with  $\hat{X}^{mn}$  since  $\hat{g}_t^{mn}(0) = \hat{X}_t^{mn}$  for  $t \in [0, T]$ . Note also that if we need values of  $g_t(\tau)$  for all  $(t, \tau) \in [0, T] \times [0, \tau^*]$  for some  $\tau^* \geq 0$ , we can set  $T^* = T + \tau^*$ ; c.f. (3.57).

*Remark 3.3.2.* For the alternative process (3.41) the approximation  $\hat{\zeta}_t^{mn}(u) := \hat{g}_t^{mn}(u - t) \approx \zeta_t(u)$  is justified for points  $(t, u)$  that satisfy  $0 \leq t \leq T$  and  $t \leq u \leq T^*$ .

## 3.4 Strong convergence

We present our convergence result. The main assumptions are collected below.

CHAPTER 3. HYBRID MULTIFACTOR SCHEME FOR STOCHASTIC  
VOLTERRA EQUATIONS WITH COMPLETELY MONOTONE KERNELS

---

**Assumption 3.4.1.**  $K$  is CM and satisfies  $(H_0)$  for some  $\tilde{\gamma} \in (0, 2]$ . There exists a  $\beta > 1$  so  $\|K\|_{L^{2\beta}([0, T^*])} < \infty$  and  $\lim_{m \rightarrow \infty} \|K - K_m\|_{L^{2\beta}([0, T^*])} = 0$ . Moreover  $g_0$  is locally Hölder continuous of order  $\tilde{\gamma}/2$  and  $b$  and  $\sigma$  are Lipschitz continuous, i.e. there exists a positive constant  $C$  so  $|b(x) - b(y)| + |\sigma(x) - \sigma(y)| \leq C|x - y|$  for all  $x, y \in \mathbb{R}$ .

We need  $g_0$  to be locally Hölder continuous of order  $\tilde{\gamma}/2$ —and not just all orders strictly smaller as in Assumption 3.3.1—to avoid weakening the convergence rate of Theorem 3.4.3 (to be presented) by an arbitrarily small number. Together with the assumption that  $K$  satisfies  $(H_0)$  and Lipschitz continuity of  $b, \sigma$  it implies that there exists a unique continuous strong solution  $X$  to (3.1) as follows by Theorem 3.B.1. The assumption that there exists a  $\beta > 1$  so  $\|K\|_{L^{2\beta}([0, T^*])} < \infty$  and  $\lim_{m \rightarrow \infty} \|K - K_m\|_{L^{2\beta}([0, T^*])} = 0$  could be criticized. While the integrability condition is satisfied by most conceivable kernels<sup>20</sup>, we are unaware of results that under general conditions guarantee the existence of a sequence  $\{(c^m, \gamma^m)\}_{m=1}^\infty$  so  $\lim_{m \rightarrow \infty} \|K - K_m\|_{L^{2\beta}([0, T^*])} = 0$  for a  $\beta > 1$ . Proposition 3.3 and Remark 3.4 of [4] give conditions on  $K$  and  $\{(c^m, \gamma^m)\}_{m=1}^\infty$  that ensure convergence with  $\beta = 1$ . We are nevertheless hopeful that our assumptions could be relaxed to  $\beta = 1$  or that their results could be extended to  $\beta > 1$  under reasonable conditions.

The below shows that  $K_{mn}$  under Assumption 3.4.1 satisfies  $(H_0)$  with rate  $\tilde{\gamma} \wedge 1$ .

**Corollary 3.4.2.** Say that  $K$  is CM and satisfies  $(H_0)$  for some  $\tilde{\gamma} \in (0, 2]$ . Let  $m, n \in \mathbb{N}$ . Then  $K_{mn} \in L^2_{loc}(\mathbb{R}_+)$ ,  $\int_0^h K_{mn}(t)^2 dt = O(h^{\tilde{\gamma} \wedge 1})$ , and  $\int_0^{\tilde{T}} (K_{mn}(t+h) - K_{mn}(t))^2 dt = O(h^{\tilde{\gamma} \wedge 1})$  for every  $\tilde{T} \in [0, \infty)$ .

*Proof.* The result is a direct implication of Proposition 3.F.2 of the appendix.  $\square$

*Remark 3.4.1.* The rate  $\tilde{\gamma} \wedge 1$  of Corollary 3.4.2 can be strengthened to 1 when  $\kappa = 0$  as then  $K_{mn} = K_m$  where  $K_m$  is locally Lipschitz on  $[0, \infty)$ ; c.f. [5, Example 2.3 part (i)].

Combining Corollary 3.4.2 and Theorem 3.B.1 we note that Assumption 3.4.1 for any  $m, n$  guarantees also the existence of a unique continuous adapted solution  $X^{mn}$  to (3.8) on the same probability space as  $X$  driven by the same Brownian motion  $W$ . Using moreover continuity of  $b, \sigma$  we conclude that  $g^{mn}$  of (3.58) likewise is well-defined on the same probability space. We need this so we can meaningfully compute the strong error.

*Remark 3.4.2.* To avoid misunderstandings, we stress that we for Theorem 3.4.3 stated below keep the following quantities fixed:  $g_0, K, b, \sigma, T, T^*, \kappa, \{(c^m, \gamma^m)\}_{m=1}^\infty, \beta, \tilde{\gamma}$ . The constants  $C_p$  and  $C_{p,m}$  that appear in the theorem therefore implicitly depend on them.

Our main result is that below.

**Theorem 3.4.3.** Invoke Assumption 3.4.1 and assume that Equation (3.33) of Definition 3.2.1 is used. For  $p \geq \frac{2\beta}{\beta-1}$  and  $m \in \mathbb{N}$  there then exist positive constants  $C_p$  and  $C_{p,m}$ , the first of which depends on  $p$  only, the latter on both  $p$  and  $m$ , so

$$\sup_{(t, \tau) \in A} E [|g_t(\tau) - \hat{g}_t^{mn}(\tau)|^p] \leq C_p \|K - K_{mn}\|_{L^{2\beta}([0, T^*])}^p + C_{p,m} n^{-\frac{p}{2}(\tilde{\gamma} \wedge 1)}, \quad n \in \mathbb{N}. \quad (3.66)$$

<sup>20</sup> Consider  $K(t) = ct^\alpha$  with  $c \geq 0$  and  $\alpha \in (-\frac{1}{2}, 0]$ . Then  $\|K\|_{L^{2\beta}([0, T^*])} < \infty$  when  $2\alpha\beta > -1$ .

For  $\alpha \in (-1/2, 0)$  the values  $\beta \in (1, -\frac{1}{2\alpha})$  work. For  $\alpha = 0$  arbitrary  $\beta > 1$  can be used. As the integrability condition is preserved under multiplication by functions that are continuous on  $[0, \infty)$  and sums, we consider it mild.

### 3.4. STRONG CONVERGENCE

---

*Proof.* We provide a full proof in Appendix 3.G. Below we give a (very) brief summary as we need it to discuss the theorem: Let  $p \geq \frac{2\beta}{\beta-1}$  and  $m, n \in \mathbb{N}$ . By Jensen's inequality:

$$\begin{aligned} \sup_{(t,\tau) \in A} E [|g_t(\tau) - \hat{g}_t^{mn}(\tau)|^p] &\leq 2^{p-1} \sup_{(t,\tau) \in A} E [|g_t(\tau) - g_t^{mn}(\tau)|^p] \\ &\quad + 2^{p-1} \sup_{(t,\tau) \in A} E [|g_t^{mn}(\tau) - \hat{g}_t^{mn}(\tau)|^p]. \end{aligned}$$

As we show in the appendix, we may bound the above terms as

$$\begin{aligned} 2^{p-1} \sup_{(t,\tau) \in A} E [|g_t(\tau) - g_t^{mn}(\tau)|^p] &\leq C_p \|K - K_{mn}\|_{L^{2\beta}([0, T^*])}^p, \\ 2^{p-1} \sup_{(t,\tau) \in A} E [|g_t^{mn}(\tau) - \hat{g}_t^{mn}(\tau)|^p] &\leq C_{p,m} n^{-\frac{p}{2}(\tilde{\gamma} \wedge 1)}. \end{aligned}$$

The result follows.  $\square$

The below is more or less immediate.

**Theorem 3.4.4.** *Invoke Assumption 3.4.1 and assume that Equation (3.33) of Definition 3.2.1 is used. For  $p \geq 1$  then*

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \sup_{(t,\tau) \in A} E [|g_t(\tau) - \hat{g}_t^{mn}(\tau)|^p] = 0. \quad (3.67)$$

*Proof.* Note for arbitrary  $m, n$  that  $\|K - K_{mn}\|_{L^{2\beta}([0, T^*])} \leq \|K - K_m\|_{L^{2\beta}([0, T^*])}$ . Equation (3.67) then follows for any  $p \geq \frac{2\beta}{\beta-1} > 2$  by combining this with (3.66) and taking limits. That (3.67) holds for a general  $p \geq 1$  follows by the fact that convergence in  $L^p$  implies convergence in  $L^q$  for any  $1 \leq q < p < \infty$  on probability spaces.  $\square$

We discuss the results in the following.

Note first that the left-hand side of (3.66) is estimated by two terms. The proof shows that the first term represents the error between  $g$  and  $g^{mn}$  (the error incurred by approximating  $K$  by  $K_{mn}$ ), the second term that between  $g^{mn}$  and  $\hat{g}^{mn}$  (the error incurred by discretising). As the first term is bounded by  $C_p \|K - K_m\|_{L^{2\beta}([0, T^*])}^p$  it suggests for convergence a double limit first in  $n \rightarrow \infty$  to make the discretisation error vanish, then in  $m \rightarrow \infty$  to make the kernel error vanish; this is also reflected in Theorem 3.4.4. There is no rate estimate for the first term and we neither expect such without additional assumptions on  $\{(c^m, \gamma^m)\}_{m=1}^\infty$ . The convergence rate for the discretisation error is  $\frac{1}{2}(\tilde{\gamma} \wedge 1)$ ; we always discuss it after normalisation by the moment  $p$ . For a kernel that is CM on  $[0, \infty)$ , non-singular especially, may set  $\tilde{\gamma} = 1$  to attain a convergence rate of  $\frac{1}{2}$ . For the rough fractional kernel  $K(t) = ct^\alpha$ ,  $c \geq 0$ ,  $\alpha \in (-\frac{1}{2}, 0]$ , we may set  $\tilde{\gamma} = 2\alpha + 1$  and thus attain a rate of  $\alpha + \frac{1}{2}$ . We refer to our discussion on page 41 for the validity of the  $\tilde{\gamma}$ -values; for the tightest estimates, we consider only the largest known valid  $\tilde{\gamma}$ 's. More generally the convergence rate is related to the regularity of the spot process: Indeed,  $X$  is locally Hölder continuous of all orders strictly less than  $\tilde{\gamma}/2$  by Theorem 3.B.1.

The rate estimate of  $\frac{1}{2}$  when  $K$  is CM on  $[0, \infty)$  coincides with that of the Euler scheme for SDE's; c.f. [45, Theorem 10.2.2] although it only covers the case of  $p = 1$ .<sup>21</sup> This is meaningful: When  $\kappa = 0$  the factors  $U^{mn} = (U_t^{mn})_{t \geq 0}$  form a multidimensional SDE. The reason is that  $X_t^{mn} = g_0(t) + \sum_{i=1}^m c_i^m U_{i,t}^{mn}$  which is plugged into the coefficient functions  $b$  and  $\sigma$  then is a simple time-dependent (according to  $g_0$ ) function of  $U_t^{mn}$ . Since also  $g_t^{mn}(\tau) = g_0(t + \tau) + \sum_{i=1}^m c_i^m e^{-\gamma_i^m \tau} U_{i,t}^{mn}$  is a simple time-dependent function of  $U_t^{mn}$ ,  $\hat{g}_t^{mn}(\tau)$  similarly of  $\hat{U}_t^{mn}$ , the error between  $g^{mn}$  and  $\hat{g}^{mn}$  is that of an Euler scheme for a SDE.<sup>22</sup> The same logic does not apply when  $\kappa > 0$  as  $X_t^{mn}$  and  $g_t^{mn}(\tau)$ , respectively  $\hat{X}_t^{mn}$  and  $\hat{g}_t^{mn}(\tau)$ , then are functions of more stochastic terms than only  $U^{mn}$ , respectively  $\hat{U}^{mn}$ . However, since we have assumed  $K$  non-singular it makes sense that setting  $\kappa > 0$  should not change the asymptotic behaviour. That is, we also find it meaningful that we should attain the Euler rate  $\frac{1}{2}$  when  $\kappa > 0$  and  $K$  is CM on  $[0, \infty)$ .

The rate estimate  $\alpha + \frac{1}{2}$  for the discretisation error under the rough fractional kernel coincides with that derived in [9] for their hybrid scheme and in [46] for their Euler scheme. It should however be mentioned that the convergence rate  $\frac{1}{2}(\tilde{\gamma} \wedge 1)$  for the discretisation error under our scheme can be improved when  $\tilde{\gamma} < 1$  (such as when  $K(t) = t^\alpha$ ,  $\alpha \in (-\frac{1}{2}, 0)$ ) and  $\kappa = 0$  and  $g_0$  is locally Hölder continuous of order  $\frac{1}{2}$ : Invoke the three mentioned extra assumptions, the latter only so the approximation error of  $g_0$  does not dominate. A careful reading of our proof then shows that the convergence rate  $\frac{1}{2}(\tilde{\gamma} \wedge 1)$  for the discretisation error can be strengthened to  $\frac{1}{2}$ . The result is unsurprising as we for  $\kappa = 0$  have  $K_{mn} = K_m$  so, regardless of the properties of  $K$ , the kernel for the discretisation error,  $K_{mn}$ , is CM on  $[0, \infty)$ . A naive interpretation is that we should set  $\kappa = 0$  always to achieve the Euler rate  $\frac{1}{2}$  even when  $\tilde{\gamma} < 1$ , assuming also local  $\frac{1}{2}$ -Hölder continuity of  $g_0$ . However, as we see it, there is no free lunch. The error that results from improper treatment of  $K$  near the origin when singular is instead, in a sense, transferred to the first error term of (3.66). In any case, as we have also mentioned several times, the choice  $\kappa > 0$  is recommended for a practical implementation when  $K$  is singular.

In Section 3.7, we examine numerically the strong convergence rate for the Gaussian case, though, under the explicit-implicit equation (3.34) and with  $m = m(n, \epsilon)$  as a function of  $n$  where  $\epsilon$  is fixed.

### 3.5 Sum-of-exponentials approximation

In this section, we examine five methods for choosing the sum-of-exponentials approximation  $K_m(t) = \sum_{i=1}^m c_i^m e^{-\gamma_i^m t}$ . There are three subsections. In the first we describe the methods, in the second compare them numerically in terms of speed and accuracy, and, in the third, investigate numerically the recommended method, that of [10], in more detail. Our focus will mainly, though not exclusively, be on the rough fractional kernel.

---

<sup>21</sup> In Theorem 3.4.3 we instead require  $p \geq \frac{2\beta}{\beta-1} > 2$  but this could simply be a limitation of our proof.

<sup>22</sup> Technically, for a full argument, it should also be noted that  $g_0$  is locally Hölder continuous of order  $\frac{1}{2}$  when  $\tilde{\gamma} = 1$  so the error that arises from approximation of  $g_0$  does not worsen the  $\frac{1}{2}$  convergence rate; recall that we e.g. truncate the solution values down to the nearest grid point in (3.36) and (3.65).

### 3.5.1 Description of methods

We start with the paper [4] where multifactor approximations are discussed for a class of rough volatility models with kernels of the form

$$K(t) = c_\alpha t^\alpha, \quad t > 0, \quad c_\alpha = \Gamma(1 + \alpha)^{-1}, \quad \alpha \in (-1/2, 0). \quad (3.68)$$

For a given  $m$ , the authors choose a set of auxiliary exponents  $\eta^m = (\eta_0^m, \eta_1^m, \dots, \eta_m^m) \in \mathbb{R}^{m+1}$  that satisfy  $0 = \eta_0^m < \eta_1^m < \dots < \eta_m^m$  and set the coefficients as

$$c_i^m = \int_{\eta_{i-1}^m}^{\eta_i^m} \mu(d\gamma), \quad \gamma_i^m = \frac{1}{c_i^m} \int_{\eta_{i-1}^m}^{\eta_i^m} \gamma \mu(d\gamma), \quad i = 1, \dots, m, \quad (3.69)$$

where

$$\mu(d\gamma) = \frac{\gamma^{-\alpha-1}}{\Gamma(1 + \alpha)\Gamma(-\alpha)} d\gamma. \quad (3.70)$$

The measure (3.70) is that of the representation (3.7) for the kernel (3.68). Explicit expressions for  $(c^m, \gamma^m)$  are obtained by plugging (3.70) into (3.69) and integrating. The problem is reduced to that of choosing  $\eta^m$ . For a given  $T > 0$ , the authors show that

$$\|K - K_m\|_{L^2([0, T])} \leq f_m^{(2)}(\eta^m) \quad (3.71)$$

where

$$f_m^{(2)}(\eta^m) = \frac{T^{5/2}}{2\sqrt{5}} \sum_{i=1}^m \int_{\eta_{i-1}^m}^{\eta_i^m} (\gamma - \gamma_i^m)^2 \mu(d\gamma) + \frac{1}{(\alpha + \frac{1}{2})\Gamma(1 + \alpha)\Gamma(-\alpha)\sqrt{2}} (\eta_m^m)^{-(\alpha + \frac{1}{2})}.$$

Thus convergence in  $L^2([0, T])$  norm is ensured if  $\eta^m$  is chosen so

$$\sum_{i=1}^m \int_{\eta_{i-1}^m}^{\eta_i^m} (\gamma - \gamma_i^m)^2 \mu(d\gamma) \rightarrow 0 \quad \text{and} \quad \eta_m^m \rightarrow \infty \quad \text{as} \quad m \rightarrow \infty. \quad (3.72)$$

The equidistant choice  $\eta_i^m = i\pi_m$ ,  $i = 0, 1, \dots, m$ , for a number  $\pi_m > 0$  is explored in detail. Under that choice the authors minimise  $f_m^{(2)}(\eta^m)$  with respect to  $\pi_m$  to obtain

$$\pi_m = \frac{m^{-\frac{1}{5}}}{T} \left( \frac{\sqrt{10}(-2\alpha)}{4 - 2\alpha} \right)^{\frac{2}{5}} \quad (3.73)$$

and

$$\|K - K_m\|_{L^2([0, T])} \leq C_\alpha m^{-\frac{4}{5}(\alpha + \frac{1}{2})} \quad (3.74)$$

where  $C_\alpha$  is a positive constant depending on  $\alpha$ ; the inequality (3.74) shows that convergence is ensured under the choice (3.73). We include this method for choosing  $(c^m, \gamma^m)$  in our experiments and refer to it as **AJEE2019**. To obtain a higher degree of accuracy, it is alternatively suggested that we set  $\eta^m$  by numerically minimising  $f_m^{(2)}(\eta^m)$  subject to  $0 = \eta_0^m < \eta_1^m < \dots < \eta_m^m$ . That is, to minimise without additional constraints on  $\eta^m$ . We likewise include this method in our test and refer to it as **AJEE2019-optim**.

CHAPTER 3. HYBRID MULTIFACTOR SCHEME FOR STOCHASTIC  
VOLTERRA EQUATIONS WITH COMPLETELY MONOTONE KERNELS

---

In [2], Eduardo Abi Jaber considers a Markovian approximation of the rough Heston model and therefore also focusses on the kernel (3.68). The coefficients are set as

$$c_i^m = \frac{(r_m^{-\alpha} - 1)r_m^{\alpha(1+m/2)}}{\Gamma(1+\alpha)\Gamma(1-\alpha)} r_m^{-\alpha i}, \quad \gamma_i^m = \frac{\alpha}{\alpha-1} \frac{r_m^{1-\alpha} - 1}{r_m^{-\alpha} - 1} r_m^{i-1-m/2}, \quad i = 1, \dots, m, \quad (3.75)$$

in terms of a free parameter  $r_m > 1$ . The above is equivalent to the geometric partitioning  $\eta_i^m = r_m^{i-m/2}$ ,  $i = 0, 1, \dots, m$ , for the auxiliary exponents of [4]. It is shown that  $\lim_{m \rightarrow \infty} \|K - K_m\|_{L^2([0, T])} = 0$  for any  $T > 0$  if  $\{r_m\}_{m=1}^\infty$  is chosen so  $r_m \rightarrow 1$  and  $m \log(r_m) \rightarrow \infty$  in the limit  $m \rightarrow \infty$ . Though explicit sequences  $\{r_m\}_{m=1}^\infty$  yielding convergence may be constructed,<sup>23</sup> the author finds the fixed choice  $m = 20$  and  $r_{20} = 2.5$  to work well for practical purposes. Alternatively, it is for a given  $m$  suggested that we solve

$$\min_{r_m > 1} \sum_{i=1}^N w_i \|K - K_m\|_{L^2([\delta, T_i])} \quad (3.76)$$

where  $N$  is a positive integer,  $(w_1, \dots, w_N) \in \mathbb{R}^N$  weights,  $(T_1, \dots, T_N) \in \mathbb{R}_+^N$  time points, and  $\delta = 0$ . We include the method defined by (3.76) in our experiments, except we use it with  $N = 1$  and let  $\delta \geq 0$  be more general; we refer to it as **AJ2019-optim**.

The methods that we have introduced so far have several limitations and are generally suboptimal: Firstly, by their construction they all restrict  $(c^m, \gamma^m)$  to limited regions of  $\mathbb{R}^{2m}$ . Secondly, those of [4] minimise an upper bound on the error. Thirdly, that of [2] is tailored to the rough fractional kernel and is therefore not applicable in general. Lastly, although the authors of [4] argue that their approach generalises to other kernels, we note that it relies on an explicit expression for  $\mu$  which may not always be available.<sup>24</sup>

In search of higher accuracy and more generality, we include an additional method in our tests which we refer to as  **$l^2$ -optim**: We consider a general CM kernel  $K$  to be approximated on an interval  $[\delta, T]$  where  $0 \leq \delta < T$ . In the singular case we need to assume  $\delta > 0$ , though, of course, it can be chosen arbitrarily small. Let  $m, n \in \mathbb{N}$  and define the equidistant points  $t_i = \delta + (T - \delta) \frac{i-1}{n-1}$ ,  $i = 1, \dots, n$ . Define also  $y_i = K(t_i)$ ,  $i = 1, \dots, n$ , and set  $y = (y_1, \dots, y_n)^\top$ . What we propose is then to minimise the  $l^2$ -error between  $K$  and  $K_m$  across the sampled points. That is, to set the coefficients as

$$(c^m, \gamma^m) := \operatorname{argmin}_{c, \gamma \in \mathbb{R}^m} \|y - A(\gamma)c\|_2^2 \quad \text{subject to} \quad 0 \leq \gamma_1 < \gamma_2 < \dots < \gamma_m, \quad (3.77)$$

where for a vector  $\gamma = (\gamma_1, \dots, \gamma_m) \in \mathbb{R}^m$  we define  $A(\gamma)$  as the  $n \times m$  dimensional matrix with  $(i, j)$ 'th entry  $e^{-\gamma_j t_i}$ . When  $\gamma$  is held fixed (3.77) reduces to a least squares minimisation problem with respect to  $c$ . The minimum norm solution is  $c = A^+(\gamma)y$  where  $A^+(\gamma)$  is the pseudoinverse of  $A(\gamma)$  [36, pp. 291]. We may therefore instead solve

$$\gamma^m := \operatorname{argmin}_{\gamma \in \mathbb{R}^m} \|y - A(\gamma)c(\gamma)\|_2^2 \quad \text{subject to} \quad 0 \leq \gamma_1 < \gamma_2 < \dots < \gamma_m, \quad (3.78)$$

<sup>23</sup> In the paper, the sequence  $r_m = 1 + 10m^{-0.9}$  is tested. Their Figure 2 however shows that upwards of  $m = 100$  terms are needed for convergence of the volatility smiles for an example where  $\alpha = -0.4$ .

<sup>24</sup> We refer to their Remark 3.4 where the authors for a general CM kernel  $K$  (satisfying a certain integrability condition) state that setting the coefficients as in (3.69) for a  $\eta^m$  satisfying (3.72) yet results in  $\lim_{m \rightarrow \infty} \|K - K_m\|_{L^2([0, T])} = 0$ . However, as we note, to evaluate (3.69) we must know  $\mu$ .

### 3.5.1. DESCRIPTION OF METHODS

---

where  $c(\gamma) := A^+(\gamma)y$  and set  $c^m := c(\gamma^m)$ . This is how we implement the method.<sup>25</sup>

Lastly, we shall examine the method of [10] where the authors, in a very general setting, consider the efficient approximation of functions by exponential sums. We refer to their method as **BM2005**. To our knowledge, it has not been used in our context before and we will therefore be rather elaborate on how it works and specialises to the CM case.

Basic setup: We consider a continuous function  $\tilde{f} : [a, b] \rightarrow \mathbb{C}$ ,  $a, b \in \mathbb{R}$ ,  $a < b$ , to be approximated by a sum of exponentials.<sup>26,27</sup> We will normalise the situation and work with the equivalent mapping  $f : [0, 1] \rightarrow \mathbb{C}$  defined by  $f(x) := \tilde{f}(a + (b - a)x)$ ,  $x \in [0, 1]$ . The goal is to find  $m \in \mathbb{N}$  and  $c = (c_1, \dots, c_m), \gamma = (\gamma_1, \dots, \gamma_m) \in \mathbb{C}^m$  so  $f(x) \approx \sum_{i=1}^m c_i e^{-\gamma_i x}$ ,  $x \in [0, 1]$ . For the original function  $\tilde{f}$  then:

$$\tilde{f}(x) = f\left(\frac{x-a}{b-a}\right) \approx \sum_{i=1}^m \tilde{c}_i e^{-\tilde{\gamma}_i x}, \quad x \in [a, b], \quad \tilde{c}_i := c_i e^{\gamma_i \frac{a}{b-a}}, \quad \tilde{\gamma}_i := \frac{\gamma_i}{b-a}, \quad i = 1, \dots, m.$$

In what follows, we outline the method of [10] for selecting  $m$  and  $(c, \gamma)$ . We will comment on the CM case as we go along, and will to that end introduce the following assumption:

**Assumption 3.5.1.**  *$\tilde{f}$  is the restriction to the domain  $[a, b]$  of a function  $K$  that is CM on an interval that covers  $[a, b]$  and which has representation (3.7) for a measure  $\mu \geq 0$ .*

Naturally, we will under the assumption need to assume  $a \geq 0$  and if  $K$  is CM on  $(0, \infty)$  only, more restrictively that  $a > 0$ . Assumption 3.5.1 implies for  $x \in [0, 1]$  that

$$f(x) = \tilde{f}(a + (b - a)x) = \int_{[0, \infty)} e^{-\gamma(a + (b - a)x)} \mu(d\gamma) = \int_{[0, \infty)} e^{-\gamma(b - a)x} \hat{\mu}(d\gamma), \quad (3.79)$$

where  $\hat{\mu} := e^{-(\cdot)a} \cdot \mu$  (like  $\mu$ ) is non-negative. The representation (3.79) justifies a sum-of-exponentials approximation to  $f$  with real-valued non-negative coefficients. We explicitly state whenever we work under Assumption 3.5.1. Otherwise we work in the general setup.

Because the method of [10] is formulated for a discrete set of points and not directly functions, we start by sampling  $f$  at  $2N + 1$ ,  $N \in \mathbb{N}$ , equidistant points on its domain as

$$h_k := f\left(\frac{k}{2N}\right), \quad k = 0, 1, \dots, 2N.$$

For future use, let  $h := (h_0, h_1, \dots, h_{2N})$ . For practical purposes, we consider the problem of finding  $m$  and coefficients  $(c, \gamma)$  so

$$h_k \approx \sum_{i=1}^m c_i e^{-\gamma_i \frac{k}{2N}}, \quad k = 0, 1, \dots, 2N. \quad (3.80)$$

---

<sup>25</sup> We use Matlab's *fmincon* function to solve the optimisation problem; we also use it for AJEE2019-optim and AJ2019-optim.

<sup>26</sup> It might seem surprising that we should be able to approximate more general functions with exponentials. However, in this setting, the coefficients will be allowed complex. Theorem 2 of [10], which we recite later, adds more rigour. It also suggests that perhaps we could do multifactor approximations for SVE's with non CM but continuous kernels if we work in  $\mathbb{C}$ . We leave that for a future project.

<sup>27</sup> To approximate a function on  $(a, b]$  that is singular in  $a$ , we should consider it on  $[a + \delta, b]$  for some small  $\delta > 0$ .

CHAPTER 3. HYBRID MULTIFACTOR SCHEME FOR STOCHASTIC  
VOLterra EQUATIONS WITH COMPLETELY MONOTONE KERNELS

---

If (3.80) holds and  $N$  is large enough, i.e. we have oversampled  $f$ , we can by continuity expect  $f(x) \approx \sum_{i=1}^m c_i e^{-\gamma_i x}$  for a general  $x \in [0, 1]$ , which means that we have solved the original problem. Define  $\rho_i := \exp(-\frac{\gamma_i}{2N})$ ,  $i = 1, \dots, m$ , and  $\rho := (\rho_1, \dots, \rho_m)$ . We may, and indeed will, then reformulate the problem (3.80) to that of finding  $m$  and  $(c, \rho)$  so

$$h_k \approx \sum_{i=1}^m c_i \rho_i^k, \quad k = 0, 1, \dots, 2N. \quad (3.81)$$

We recover the original formulation as  $\gamma_i = -2N \log(\rho_i)$ ,  $i = 1, \dots, m$ .

The *Hankel* matrix defined by  $h$  is central to the authors' method. More precisely, we consider the matrix  $H = \{H_{i,j}\}_{i,j=1}^{N+1} := \{h_{i+j-2}\}_{i,j=1}^{N+1}$  or written in matrix form:<sup>28</sup>

$$H = \begin{pmatrix} h_0 & h_1 & \dots & h_{N-1} & h_N \\ h_1 & \ddots & \ddots & \ddots & h_{N+1} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ h_{N-1} & \ddots & \ddots & \ddots & h_{2N-1} \\ h_N & h_{N+1} & \dots & h_{2N-1} & h_{2N} \end{pmatrix}.$$

We need to apply a number of non-standard concepts to  $H$  that we briefly introduce: For a matrix  $A \in \mathbb{C}^{n \times n}$ ,  $n \in \mathbb{N}$ , the *con-eigenvalue problem* is that of solving the equation

$$Au = \sigma \bar{u}$$

for  $u \in \mathbb{C}^n$ ,  $u \neq 0$ , and  $\sigma \in \mathbb{C}$ .<sup>29</sup> For a solution  $\{\sigma, u\}$  we refer to  $\sigma$  as a *con-eigenvalue* and to  $u$  as a *con-eigenvector* [31, Definition 4.6.5]. For a con-eigenvalue  $\sigma$  all other complex numbers of the same modulus are also con-eigenvalues [31, pp. 245]. We refer to a real non-negative con-eigenvalue  $\sigma$  representing all con-eigenvalues of the same modulus as a *c-eigenvalue*; to avoid confusion we stress that the 'c' in this context does *not* refer to the weights of (3.81). A con-eigenvector corresponding to a c-eigenvalue is referred to as a *c-eigenvector* and a solution  $\{\sigma, u\}$  where  $\sigma$  is a c-eigenvalue as a *c-eigenpair*. The c-eigenpairs of  $H$  play a central role for the approximation of  $h$  under the method of [10].

The result below gives a useful characterisation of the c-eigenvalues.

**Proposition 3.5.2.** (*Proposition 4.6.6 of [31]*) *Let  $A \in \mathbb{C}^{n \times n}$  and  $\sigma$  be a real and non-negative number. Then  $\sigma$  is a c-eigenvalue of  $A$  if and only if  $\sigma^2$  is an eigenvalue of  $\overline{AA}$ .*

In what follows, we discuss the problem of finding c-eigenpairs of  $H$ . We focus on locating pairs for all c-eigenvalues only. The c-eigenvalues are of primary concern because, as the reader will see, they control the accuracy of the approximation to  $h$ . Following [10] we

<sup>28</sup> Definition:  $A \in \mathbb{C}^{(n+1) \times (n+1)}$ ,  $n \in \mathbb{N}_0$ , is a Hankel matrix if  $A = \{A_{i,j}\}_{i,j=1}^{n+1} = \{a_{i+j-2}\}_{i,j=1}^{n+1}$  for a sequence of numbers  $a_0, a_1, \dots, a_{2n}$ .

<sup>29</sup> The reader should note that some of the definitions and results we recite from [31] have been slightly reformulated to remain consistent with [10].

### 3.5.1. DESCRIPTION OF METHODS

---

consider the *Takagi* factorisation [31, Corollary 4.4.4] which exists for any symmetric complex matrix. The factorisation applied to  $H$ —that indeed is symmetric—results in

$$H = \bar{U}\Sigma U^* \quad (3.82)$$

where  $U$  is a unitary complex matrix and  $\Sigma = \text{diag}(\sigma_0, \dots, \sigma_N)$  is a real-valued diagonal matrix with non-negative entries. To see why the factorisation is useful, note that

$$\bar{H}H = \bar{U}\Sigma U^* \bar{U}\Sigma U^* = U\Sigma U^\top \bar{U}\Sigma U^* = U\Sigma^2 U^*. \quad (3.83)$$

Consequently, the matrices  $\bar{H}H$  and  $\Sigma^2$  are similar. It follows by [31, Corollary 1.3.4] that the eigenvalues of  $\bar{H}H$  coincide with those of  $\Sigma^2$  which are  $\{\sigma_0^2, \dots, \sigma_N^2\}$ . From Proposition 3.5.2, we deduce that the c-eigenvalues of  $H$  are the numbers  $\{\sigma_0, \dots, \sigma_N\}$  of the Takagi factorisation. Rewriting (3.82) to

$$HU = \bar{U}\Sigma$$

we also note that the columns of  $U$  are corresponding c-eigenvectors. Therefore, to find c-eigenpairs for all c-eigenvalues it suffices to perform a Takagi factorisation.

Nevertheless, as we will argue in the following, a regular eigendecomposition is enough in the CM case. We start with the below result.

**Proposition 3.5.3.** *Under Assumption 3.5.1 the eigenvalues of  $H$  are real and non-negative.*

*Proof.* Because  $K$  is CM it is real-valued. Then so is  $f$  and it follows that  $H$  is both symmetric and real and therefore that the eigenvalues are real too [31, pp. 38]. The eigenvalues are non-negative because  $H$  is positive semi-definite<sup>30</sup> which is a consequence of the representation (3.79); we prove the details in the appendix.  $\square$

Let us temporarily invoke Assumption 3.5.1: Then because  $H$  is real and symmetric, we can perform an eigendecomposition [31, Theorem 4.1.5] to write

$$H = U\Sigma U^\top \quad (3.84)$$

where  $\Sigma = \text{diag}(\sigma_0, \dots, \sigma_N)$  is a real-valued diagonal matrix consisting of the eigenvalues of  $H$  and  $U$  is an orthonormal real matrix of corresponding eigenvectors. Since the eigenvalues  $\{\sigma_0, \dots, \sigma_N\}$  are real and non-negative as follows from Proposition 3.5.3, and  $U$  is real, the eigendecomposition (3.84) coincides with the Takagi factorisation (3.82). In particular, the eigenvalues and c-eigenvalues coincide. We conclude that it in the CM case suffices to perform an eigendecomposition to find c-eigenpairs for all c-eigenvalues.

Consider now the following theorem which is taken from [10]. Given certain c-eigenpairs of  $H$  it alludes to an approximation of  $h$  of the form (3.81). For the result we need the definition  $P_v(z) := \sum_{k=0}^N v_k z^k$ ,  $z \in \mathbb{C}$ ,  $v = (v_0, v_1, \dots, v_N) \in \mathbb{C}^{N+1}$ .

<sup>30</sup> Say that  $H$  is positive semi-definite. Let  $\sigma$  be an eigenvalue. Because both  $H$  and  $\sigma$  are real, we may find a real eigenvector  $u$ . Then  $Hu = \sigma u$  where all components are real. This implies  $u^\top Hu = \sigma u^\top u$  where  $u^\top Hu \geq 0$  since  $H$  is positive semi-definite and  $u^\top u \geq 0$ . It follows that  $\sigma \geq 0$ .

**Theorem 3.5.4.** (Theorem 2 of [10]) Let  $\{\sigma, u\}$  be a  $c$ -eigenpair of  $H$  and assume that  $P_u$  has  $N$  distinct roots which we denote by  $\chi = (\chi_1, \dots, \chi_N) \in \mathbb{C}^N$ . Then, for each integer  $L > 2N$ , there exists a unique vector  $\omega = (\omega_1, \dots, \omega_N) \in \mathbb{C}^N$  so

$$h_k = \sum_{i=1}^N \omega_i \chi_i^k + \sigma d_k^{(L)}, \quad k = 0, 1, \dots, 2N, \quad (3.85)$$

where  $d^{(L)} = (d_0^{(L)}, \dots, d_{L-1}^{(L)})^\top \in \mathbb{C}^L$  is that of [10, Equation (18)] and which has unit  $l^2$ -norm.

*Remark 3.5.1.* Note that while  $\omega$  is not explicitly stated in Theorem 3.5.4, it is implicitly given as the unique solution of the Vandermonde system defined by the first  $N$  equations of (3.85). The system is invertible since  $\{\chi_1, \dots, \chi_N\}$  are assumed distinct; we refer to [31, pp. 29].

*Remark 3.5.2.* We stress that the theorem only applies when  $P_u$  has  $N$  distinct roots. When we in the following discuss approximations of  $h$  inspired by the result, we will therefore exactly assume that  $N$  distinct roots exists. According to Remark 5 of [10] the assumption can be relaxed some. Also, the algorithm we end up with can be adjusted when there are fewer roots although we then, of course, cannot appeal to the theorem.

For a  $c$ -eigenpair  $\{\sigma, u\}$  so  $P_u$  has  $N$  distinct roots and an integer  $L > 2N$ , the theorem alludes to the approximation

$$h_k \approx \sum_{i=1}^N \omega_i \chi_i^k, \quad k = 0, 1, \dots, 2N. \quad (3.86)$$

Let  $\hat{h}_k^{(0)} := \sum_{i=1}^N \omega_i \chi_i^k$ ,  $k = 0, 1, \dots, 2N$ ,  $\hat{h}^{(0)} := (\hat{h}_0^{(0)}, \dots, \hat{h}_{2N}^{(0)})$ . Then (3.86) is valid in the following sense: By (3.85) and that  $d^{(L)}$  has unit  $l^2$ -norm, we have  $\|h - \hat{h}^{(0)}\|_2 \leq \sigma$  and may thus control the error by selecting a valid  $c$ -eigenpair with a low enough  $\sigma$ . If  $H$  is singular then  $\sigma = 0$  is a ( $c$ -)eigenvalue which makes the lowest possible error attainable if an applicable  $c$ -eigenvector can be found. If  $H$  is non-singular, we can make it singular by extending  $h$ ,<sup>31</sup> though, it is unclear if a corresponding valid  $c$ -eigenvector can be found.

Consider also the approximation that arises if we set the weights, here denoted  $c^{LS} = (c_1^{LS}, \dots, c_N^{LS}) \in \mathbb{C}^N$ , as the solution of the below least squares problem:

$$c^{LS} := \underset{(\hat{c}_1, \dots, \hat{c}_N) \in \mathbb{C}^N}{\operatorname{argmin}} \sum_{k=0}^{2N} \left( h_k - \sum_{i=1}^N \hat{c}_i \chi_i^k \right)^2. \quad (3.87)$$

Define  $\hat{h}_k^{(1)} := \sum_{i=1}^N c_i^{LS} \chi_i^k$ ,  $k = 0, 1, \dots, 2N$ ,  $\hat{h}^{(1)} := (\hat{h}_0^{(1)}, \dots, \hat{h}_{2N}^{(1)})$ . Since  $c^{LS}$  minimises the  $l^2$ -error, we conclude that  $\|h - \hat{h}^{(1)}\|_2 \leq \|h - \hat{h}^{(0)}\|_2 \leq \sigma$ . Also, because  $c^{LS}$  is independent of  $L$ , the construction of  $\hat{h}^{(1)}$  only requires us to choose a valid  $c$ -eigenpair  $\{\sigma, u\}$ .

<sup>31</sup> More precisely: Add two extra elements  $h_{2N+1}$  and  $h_{2N+2}$  to  $h$ . Then, with a cofactor expansion, select them so the corresponding Hankel matrix has determinant zero. We add two elements because the extended  $h$  vector must have an odd number of elements as required for the Hankel matrix.

### 3.5.1. DESCRIPTION OF METHODS

---

However, even if  $\hat{h}^{(1)}$  results in a lower error than  $\hat{h}^{(0)}$ , we can expect both approximations to be inefficient because  $N$ , which is also the number of terms in the approximating sums, must be large to oversample  $f$ . The authors suggest a procedure for reducing the number of terms: Assume the c-eigenvalues are ordered as

$$\sigma_0 \geq \sigma_1 \geq \dots \geq \sigma_N \geq 0.$$

Fix a c-eigenvalue  $\sigma = \sigma_m$  for some index  $m$  (say  $m \geq 1$  for now) and let  $u \in \mathbb{C}^{N+1}$  be a corresponding c-eigenvector. We assume again that  $P_u$  has  $N$  distinct roots denoted  $\chi = (\chi_1, \dots, \chi_N) \in \mathbb{C}^N$ . The goal is now an approximation of  $h$  that uses fewer terms than  $\hat{h}^{(1)}$  but results in a  $l^2$ -error that is only slightly worse. For this purpose, we carefully select  $m$  of the roots which we denote by  $\{\rho_1, \dots, \rho_m\} \subset \{\chi_1, \dots, \chi_N\}$  (we explain how in a moment) and let  $c = (c_1, \dots, c_m) \in \mathbb{C}^m$  solve the below least squares problem:

$$c := \operatorname{argmin}_{(\hat{c}_1, \dots, \hat{c}_m) \in \mathbb{C}^m} \sum_{k=0}^{2N} \left( h_k - \sum_{i=1}^m \hat{c}_i \rho_i^k \right)^2. \quad (3.88)$$

Define  $\hat{h}_k^{(2)} := \sum_{i=1}^m c_i \rho_i^k$ ,  $k = 0, 1, \dots, 2N$ ,  $\hat{h}^{(2)} := (\hat{h}_0^{(2)}, \dots, \hat{h}_{2N}^{(2)})$ . Although the  $l^2$ -error under  $\hat{h}^{(2)}$  can only deteriorate compared to  $\hat{h}^{(1)}$ , the authors of [10] make the empirical observation that if  $\{\rho_1, \dots, \rho_m\}$  are chosen in the 'significant region' of  $\mathbb{C}$  where the associated weights are large, the error need not worsen much. One may then hope that the  $l^2$ -error is still roughly bounded by  $\sigma = \sigma_m$ . In some cases the 'significant region' can be reasoned from properties of  $f$ . Under Assumption 3.5.1, Equation (3.79) e.g. suggest to consider roots in  $(0, 1]$  only as that corresponds to exponents in  $[0, \infty)$ ; recall that  $\gamma_i = -2N \log(\rho_i)$ ,  $i = 1, \dots, m$ . In our experiments with CM functions, we consistently observe exactly  $m$  distinct roots in  $(0, 1]$ , which makes the root selection obvious.

*Remark 3.5.3.* The approximation  $\hat{h}^{(2)}$  can also be constructed if there are less than  $N$  distinct roots, except, as noted, we then cannot appeal to Theorem 3.5.4. If some roots have a higher multiplicity than one, we can consider distinct ones only, and, if there are fewer than  $m$  distinct ones overall, we may select however many are available resulting in a shorter sum. If there are no roots at all, we can always approximate  $h$  by a constant.<sup>32</sup>

We see that in choosing  $m$  and thus  $\sigma_m$ , we can, if the roots are chosen in the 'significant region', under the suggested hypothesis, expect a  $l^2$ -error with  $\hat{h}^{(2)}$  that is below or at least close to  $\sigma_m$ . As  $m$  is also the number of terms in the sum-of-exponentials approximation, and because we desire an efficient approximation with as few terms as possible, we should make sure to choose the smallest index  $m$  so  $\sigma_m$  is below the desired error level. However, instead of selecting an index  $m$  directly, let us for practical purposes suggest to first choose an  $\epsilon > 0$  which will represent a normalised  $l^2$ -error tolerance and set  $m$  as the smallest index so  $\sigma_m \leq \epsilon \|h\|_2$ . If it truly holds that  $\|h - \hat{h}^{(2)}\|_2$  is less than or close to  $\sigma_m$ , we can expect the normalised  $l^2$ -error, which we define as  $\|h - \hat{h}^{(2)}\|_2 / \|h\|_2$ , to be less than or close to  $\epsilon$ ; we here rather reasonably assume that  $h$  is not the zero vector.

---

<sup>32</sup> We can do likewise if we, for whatever reason, decide to let  $m = 0$ .

CHAPTER 3. HYBRID MULTIFACTOR SCHEME FOR STOCHASTIC  
VOLTERRA EQUATIONS WITH COMPLETELY MONOTONE KERNELS

---

An noteworthy observation of [10] is that the  $c$ -eigenvalues  $\sigma_0, \sigma_1, \dots, \sigma_N$ , decay rapidly in the index for many problems. Consequently, we often do not need  $m$  to be very large for a good approximation. Additionally, as we will demonstrate numerically for CM functions, we even find that the solution  $\hat{h}^{(2)}$  is close to optimal in the sense that the resulting  $m$  roughly is the minimal choice that ensures a normalised  $l^2$ -error below  $\epsilon$ .<sup>33</sup>

From our discussion, we propose the following algorithm which specialises that of [10] to the CM case. A few corner cases are not covered. We comment on how they can be handled in Remark 3.5.4 further down:

Say  $f : [0, 1] \rightarrow \mathbb{R}$  is constructed from a CM function in the sense of Assumption 3.5.1. Let  $\epsilon > 0$ ,  $N \in \mathbb{N}$ . Then do as follows:

1. Set  $h_k = f(\frac{k}{2N})$ ,  $k = 0, 1, \dots, 2N$ . Construct the Hankel matrix  $H = \{h_{i+j-2}\}_{i,j=1}^{N+1}$ .
2. Order the eigenvalues of  $H$  as  $\sigma_0 \geq \sigma_1 \geq \dots \geq \sigma_N \geq 0$ . Choose the smallest index  $m$  so  $\sigma_m \leq \|h\|_2 \epsilon$ . Find a real eigenvector  $u \in \mathbb{R}^{N+1}$  that corresponds to  $\sigma_m$ .
3. Select  $m$  distinct roots  $\{\rho_1, \dots, \rho_m\}$  of  $P_u$  in the interval  $(0, 1]$ .
4. Solve the least squares problem (3.88) for the weights  $\{c_1, \dots, c_m\}$ .<sup>34</sup>
5. Set  $\gamma_i = -2N \log(\rho_i)$ ,  $i = 1, \dots, m$ .

The weights  $\{c_1, \dots, c_m\}$  and exponents  $\{\gamma_1, \dots, \gamma_m\}$  may be translated to those of  $\tilde{f} = K|_{[a,b]}$  as explained on page 55.

*Remark 3.5.4.* On step 2: It could happen that the smallest eigenvalue  $\sigma_N$  satisfies  $\sigma_N > \|h\|_2 \epsilon$  which makes the step impossible. While the often observed rapid decay of the eigenvalues suggests it is unlikely to happen, one could handle it by extending  $h$  to make the associated Hankel matrix singular (c.f. footnote 31). On step 3: If there are fewer than  $m$  distinct roots in  $(0, 1]$ , we may choose however many are available to get an approximation with fewer terms. If there are more than  $m$ , one must prioritise. If there are no roots or  $m = 0$  we can default to approximate  $h$  by a constant.

*Remark 3.5.5.* A few comments are in order on a practical implementation: Firstly, because  $N$  must be large to oversample  $f$ , the matrix  $H$  will also be relatively large. Consequently, recovering all eigenvalues in step 2 may significantly slow down the algorithm. However, because the eigenvalues tend to decay rapidly in the index, we often only need to iterate a few times starting from the largest until an index  $m$  is found so  $\sigma_m \leq \|h\|_2 \epsilon$ ; we refer to [47] for an overview of methods that can recover eigenpairs in order of the eigenvalues.<sup>35</sup> Secondly, to find roots in step 3, there are several approaches. We do as follows for our experiments: Let  $N_z \in \mathbb{N}$ . Then note down all sign changes in  $P_u(z)$  over the points  $z \in \{\frac{i}{N_z}\}_{i=0}^{N_z}$  to isolate intervals where there must be roots. Lastly run a local optimizer to find a root in each interval. If  $N_z$  is large enough, the procedure should recover all roots. We set  $N_z = 10^4$  for our experiments.

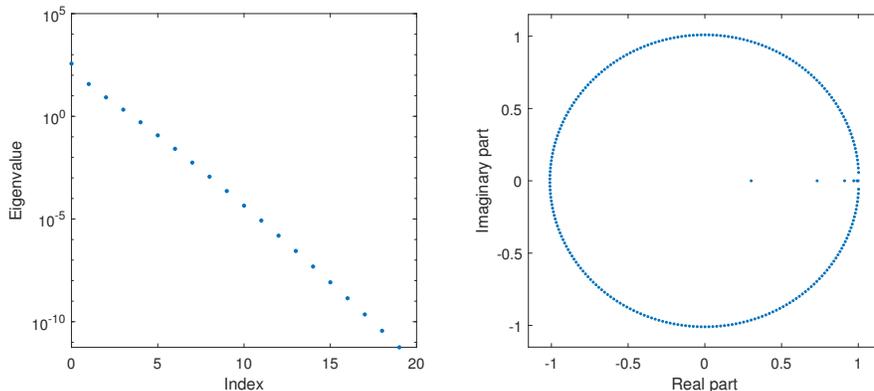
<sup>33</sup> The authors of [10] also provide arguments why the  $m$  of their algorithm is approximately minimal.

<sup>34</sup> The weights will be real-valued because both  $h$  and  $\rho$  are real.

<sup>35</sup> Regardless, for our numerical tests, we do as follows: We use Matlab's *eigs* function to compute, at once, eigenpairs for the 20 largest eigenvalues and then select the appropriate index  $m$ . We found the cost of this approach to be negligible for our experiments and never had the need for more values.

### 3.5.2. NUMERICAL COMPARISON

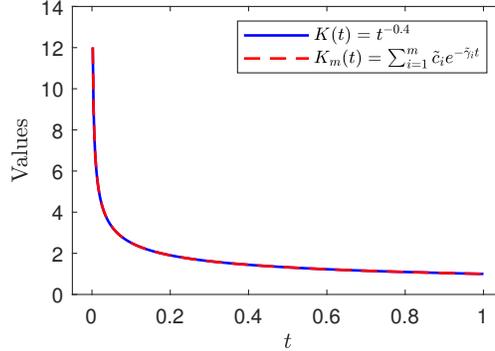
Let us briefly illustrate the method on the kernel  $K(t) = t^{-0.4}$  which we will limit to the domain  $[\frac{1}{500}, 1]$ . That we exclude the singularity is as required for the method for a singular function but also as recommended for the hybrid multifactor scheme for this case. We set  $N = 250$  and thus sample 501 equidistant points on the domain and then construct the Hankel matrix  $H$ . In Figure 3.1 (left) we show the 20 largest eigenvalues of  $H$ . The values decay exponentially which suggests that a good approximation with a few terms is possible. Indeed, setting  $\epsilon = 10^{-3}$ , the smallest index  $m$  so  $\sigma_m \leq \|h\|_2 \epsilon$  is  $m = 6$ . With that choice we then locate a real eigenvector  $u$  for  $\sigma_6$ . We note that the corresponding polynomial  $P_u$  has 250 distinct roots, which we show in the right plot. Precisely six of them lie in  $(0, 1]$ . Selecting those six roots, we solve the least squares problem (3.88) for the weights. The final solution (after rescaling) comes out as  $K_m(t) = \sum_{i=1}^m \tilde{c}_i e^{-\tilde{\gamma}_i t}$  where  $\tilde{c} = (8.54, 4.28, 2.44, 1.55, 1.23, 1.37)$  and  $\tilde{\gamma} = (599.72, 156.52, 46.90, 14.89, 4.03, 0.33)$ . The solution reflects the fact that the rough fractional kernel requires both fast and slowly decaying terms as (3.70) also suggests. The normalised  $l^2$ -error across the 501 sampled points is  $6.10 \cdot 10^{-4}$  and which is less than the desired  $10^{-3}$ . This shows that it suffices to select only  $m$  roots in  $(0, 1]$ , at least for this example. We show the fit in Figure 3.2.



**Figure 3.1:** Left: The 20 largest eigenvalues of  $H$  as constructed for  $K(t) = t^{-0.4}$  with domain  $[\frac{1}{500}, 1]$  where  $N = 250$ . Right: The 250 roots of  $P_u$  where  $u$  is a real eigenvector of  $H$  that is associated with the seventh largest eigenvalue  $\sigma_6$ . There are six roots in  $(0, 1]$ ; three are very close to 1 and are therefore difficult to distinguish visually in the plot.

### 3.5.2 Numerical comparison

In what follows, we compare the methods numerically. We use the rough fractional kernel (3.68) with domain  $[\frac{1}{500}, 1]$  and  $\alpha = -0.4$  as our test case. For  $l^2$ -optim and BM2005 we sample 501 equidistant points. For AJ2019-optim and  $l^2$ -optim we set  $\delta = \frac{1}{500}$ . We let AJEE2019 generate initial guesses for AJEE2019-optim, use  $r_m = 2.5$  as an initial guess for AJ2019-optim, and let AJ2019-optim generate those for  $l^2$ -optim. For BM2005 we consider choices of  $m$  (and not  $\epsilon$ ) to remain consistent with the other methods.



**Figure 3.2:** Approximation of  $K(t) = t^{-0.4}$  under BM2005.

In Figure 3.3, we show the fits under each method for different  $m$ . We note first that AJEE2019 and AJEE2019-optim are unable to capture the steep part of the rough fractional kernel with even several hundred terms. Both are outperformed by AJ2019-optim, which results in a very decent approximation using about 20 terms.<sup>36</sup> We note that  $l^2$ -optim produces the most accurate approximation for a given  $m$ , which is as expected given that it by construction is optimal in the  $l^2$ -sense; numerical issues may though arise, see later. With that method only about 5 terms are needed for a visually indistinguishable approximation. Remarkably, BM2005, which does not optimise over  $\gamma$ , performs effectively the same as  $l^2$ -optim. Further results to be presented show that BM2005 produces close to optimal solutions across a wide range of kernels and settings.

We perform an additional experiment to examine the convergence behaviour in more detail and also the running times. In Figure 3.4, we show the normalised  $l^2$ -error across the 501 sampled points, respectively, the running time, against  $m$ .<sup>37</sup> The left plot shows that  $l^2$ -optim and BM2005 converge the fastest by a notable margin again. However, the convergence of  $l^2$ -optim is unstable at larger  $m$ 's. A likely reason is that (3.78) becomes more challenging to solve numerically in higher dimensions as we can expect more local minima. For low values of  $m$  (where  $l^2$ -optim is numerically stable) note also how remarkably close BM2005 is to the optimal  $l^2$ -error. Combined with the numerical instability of  $l^2$ -optim, we should thus favour BM2005 based on the convergence behaviour alone. In terms of speed, we also see that  $l^2$ -optim is much slower taking around 0.5 seconds for  $m$ 's in the range 5-10, whereas BM2005 only sets us back around 0.01 seconds.

Our experiments suggest that BM2005 should be preferred as it results in a close to optimal approximation with a low computational time. Additionally, it is more practical as one can input an error tolerance  $\epsilon$  and have the number of terms  $m$  be computed endogenously. This also makes the method easier to use and more robust as  $m$  need not be manually readjusted or reoptimised for different kernels and intervals. Lastly, as

<sup>36</sup> With  $m = 20$ , we obtain  $r_{20} = 2.69$ , which almost is the fixed choice  $r_{20} = 2.50$  suggested in [2].

<sup>37</sup> All our computational times are recorded on a laptop that runs a 1.6-3.4 GHz Intel Core i5 8250U CPU with 4 cores (8 logical processors) and 8 GB RAM. The code is run in Matlab 2019a.

opposed to the approach taken in [2, 4], it works in the general case where  $\mu$  is unknown.

### 3.5.3 Robustness, optimality, and scaling of $m$ under BM2005

In this subsection, we examine BM2005 in more detail. The main purposes are to check if it is robust to different inputs and settings, to verify more thoroughly the supposed optimality, and to provide part of the evidence needed to back up the claims we made in Section 3.2 about the complexity of our scheme when combined with the method.

We start with an experiment where we run BM2005 for a wide range of inputs. We will consider two kernel types. The first is the rough fractional kernel  $K(t) = t^\alpha$  for which we fit to intervals of the form  $[\frac{1}{n}, 1]$  for a given  $n$  and always sample  $n + 1$  points. This corresponds to setting  $\kappa = 1$  and  $T^* = 1$  for the hybrid multifactor scheme; the choice of  $\kappa$  is as recommended and  $T^* = 1$ , we believe, a reasonable time horizon. We then run BM2005 with the following 980 combinations of settings and parameters:

$$(n, \epsilon, \alpha) \in \{100, 250, 500, 1000\} \times \{10^{-1}, 10^{-2}, \dots, 10^{-5}\} \times \{-0.49, -0.48, \dots, -0.01\}.$$

We consider also the non-singular kernel  $K(t) = (1 + t)^\beta$  for which we fit to the interval  $[0, 1]$  and sample  $n + 1$  points for a given  $n$ . For this kernel, we run BM2005 with the following 1000 combinations of settings and parameters:

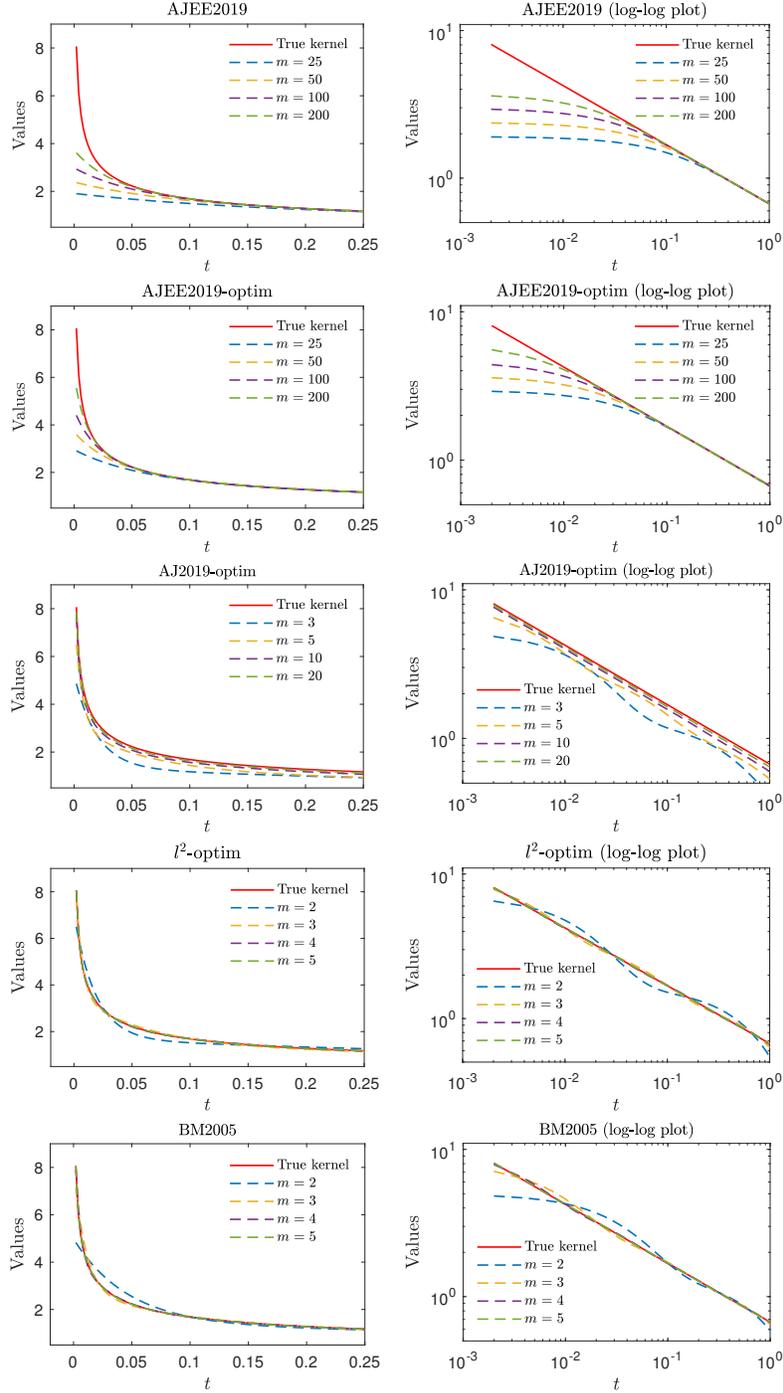
$$(n, \epsilon, \beta) \in \{100, 250, 500, 1000\} \times \{10^{-1}, 10^{-2}, \dots, 10^{-5}\} \times \{-50, -49, \dots, -1\}.$$

We present the results of the experiment in Table 3.1 split by the kernel type. In rows 3 and 4 we show the number of times we record the polynomial from step 3 of the algorithm to have  $N$  distinct roots in  $\mathbb{C}$ , respectively,  $m$  distinct roots in  $(0, 1]$ .<sup>38</sup> We observe, in all cases, precisely  $N$  distinct roots in  $\mathbb{C}$ . Consequently, Theorem 3.5.4 is in effect across the entire dataset. Additionally, we always observe  $m$  roots in  $(0, 1]$ . There is then no ambiguity about which roots to select. In rows 5-7 we show statistics on the computational time. The method runs in about one hundred of a second, which is in line with the numbers of the previous subsection and shows that the algorithm also more generally is very fast. In rows 8-10 we show statistics on the ratios  $\|h - \hat{h}^{(2)}\|_2 / (\epsilon \|h\|_2)$ . We observe values in the range 0.00-1.03 which demonstrates that the algorithm works as intended in that it constructs solutions with normalised  $l^2$ -errors below or close to  $\epsilon$ .

The last rows of the table are included to argue, with more rigour, that BM2005 is close to optimal. In this context, we will say that a solution is optimal if it has a minimal  $m$  and coefficients so the normalised  $l^2$ -error is below  $\epsilon$ . For each observation that underlie the table, we compute an optimal solution as follows: We iterate over increasing values of  $m$ , starting from  $m = 1$ , and each time run  $l^2$ -optim on the interval  $[\frac{1}{n}, 1]$  when  $K(t) = t^\alpha$  and on  $[0, 1]$  when  $K(t) = (1 + t)^\beta$  always using  $n + 1$  sampling points; we use initial guesses produced by BM2005 with corresponding settings. The first solution with a normalised  $l^2$ -error below  $\epsilon$  is optimal. We denote its approximation to  $h$  by  $\hat{h}^{(3)}$ .

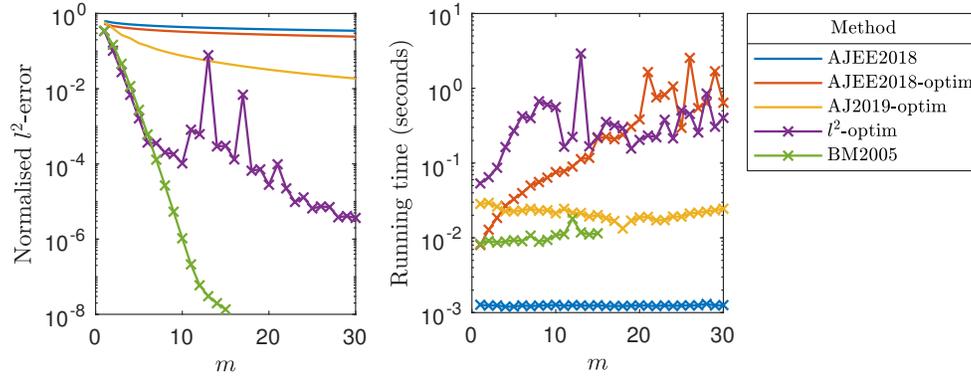
<sup>38</sup> We measure two roots as equal if the modulus of their difference is less than  $10^{-6}$ . The code is run in double precision which has 16 significant digits.

CHAPTER 3. HYBRID MULTIFACTOR SCHEME FOR STOCHASTIC  
VOLTERRA EQUATIONS WITH COMPLETELY MONOTONE KERNELS



**Figure 3.3:** Approximations of  $K$  defined by (3.68) with  $\alpha = -0.4$  and domain  $[\frac{1}{500}, 1]$ . Approximating functions are of the form  $K_m(t) = \sum_{i=1}^m c_i e^{-\gamma_i t}$  where various methods are used to find the coefficients  $(c_i, \gamma_i)_{i=1}^m$ . Red lines refer to  $K$ , the rest to  $K_m$  for different  $m$ .

### 3.5.3. ROBUSTNESS, OPTIMALITY, AND SCALING OF $M$ UNDER BM2005



**Figure 3.4:** Errors and running times against the number of exponential terms  $m$ . To improve visibility BM2005 is shown for smaller  $m$ 's only. Running times are averaged over 1000 runs; standard errors are less than 5% of the averages.

In rows 11-13, we show statistics on the ratios  $\|h - \hat{h}^{(3)}\|_2 / (\epsilon \|h\|_2)$ . The values are overall comparable to those obtained under BM2005, though, by construction never above 1. In the last rows we compare the number of exponential terms  $m$  produced by BM2005 with the number produced by the optimal solution. More precisely, we show counts of the differences in  $m$  as they fall into different categories. We see that the number of terms for BM2005 in all cases fall within 1 of the optimal solution and mostly is either equal or 1 higher. The few cases where BM2005 results in fewer terms correspond to those few cases where  $\|h - \hat{h}^{(2)}\|_2 / \|h\|_2$  is (slightly) above  $\epsilon$ . Since BM2005 results in normalised  $l^2$ -errors below or close to  $\epsilon$  and has  $m$ 's close to the optimal number, we conclude that the method, in fact, is almost optimal. In Table 3.2 we show a few specific examples that compare BM2005 with the optimal solution in terms of errors and  $m$ .

*Remark 3.5.6.* Overall BM2005 results in some striking and consistent empirical observations. This includes the exponential decay of the eigenvalues shown in Figure 3.1 (see also Figure 3.5), that we observe exactly  $N$  distinct roots in  $\mathbb{C}$ , exactly  $m$  in  $(0, 1]$ , that the roots in  $(0, 1]$  suffices for a  $l^2$ -error roughly below  $\sigma_m$ , and lastly that the approximations are almost optimal. We find it likely that the observations could be backed up by additional theoretical results, though, we leave a deeper investigation for future work.

In the following, we provide part of the necessary numerical evidence to back up our claims about the complexity of the hybrid multifactor scheme when combined with BM2005. We consider again the kernels  $K(t) = t^\alpha$  and  $K(t) = (1 + t)^\beta$  and for a given  $n$  fit to, respectively, the intervals  $[\frac{1}{n}, 1]$  and  $[0, 1]$  using  $n + 1$  sampling points. We will study how  $m$ , which we also write as  $m(n, \epsilon)$ , depends on  $n$ , and for completeness  $\epsilon$ .

In Figure 3.6 we show  $m(n, \epsilon)$  as a function of  $n$  keeping  $\epsilon = 10^{-3}$  fixed. We note for the rough fractional kernels that  $m(n, \epsilon) \approx c \log_2 n$  for constants  $c$  and for the (non-singular) shifted power-law kernels that  $m(n, \epsilon) = O(1)$ . However, for the former kernel type we also note that the  $c$ 's are fairly small meaning that  $m(n, \epsilon)$  scales only very slowly in  $n$ . In fact, going from  $n = 2^4 = 16$  to  $n = 2^{14} = 16384$  (a factor 1024 in-

CHAPTER 3. HYBRID MULTIFACTOR SCHEME FOR STOCHASTIC  
VOLTERRA EQUATIONS WITH COMPLETELY MONOTONE KERNELS

**Table 3.1:** Test of BM2005 across a wide range of configurations and kernels. We refer to the main text for a description of the experiment.

Test of BM2005		$K(t) = t^\alpha$	$K(t) = (1+t)^\beta$
# Observations		980	1000
# Observations with $N$ distinct roots in $\mathbb{C}$		980	1000
# Observations with $m$ distinct roots in $(0, 1]$		980	1000
Running times (in seconds)	Min.	0.003	0.003
	Median	0.008	0.006
	Max.	0.037	0.020
Error ratios (BM2005): $\frac{\ h-\hat{h}^{(2)}\ _2}{\epsilon\ h\ _2}$	Min.	0.04	0.00
	Median	0.41	0.09
	Max.	1.03	0.92
Error ratios (optimal solution): $\frac{\ h-\hat{h}^{(3)}\ _2}{\epsilon\ h\ _2}$	Min.	0.04	0.01
	Median	0.46	0.15
	Max.	1.00	1.00
Number of terms $m$ under BM2005 minus number of terms under the optimal solution (# observations per category)	$> 1$	0	0
	1	238	233
	0	736	767
	$-1$	6	0
	$< -1$	0	0

crease) the value of  $m(n, \epsilon)$  only goes up from about 2–3 to 7–10 (a factor 3 increase roughly). Although unreported, we observe the same scaling behaviour for other fixed  $\epsilon$ 's.

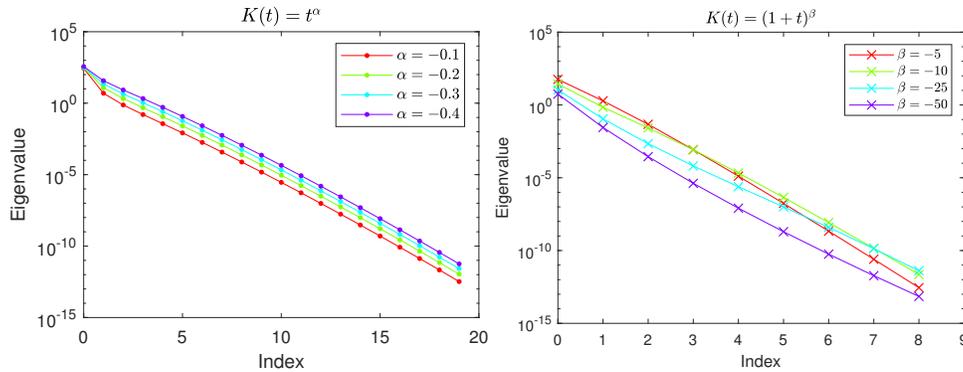
We hypothesize that  $m(n, \epsilon) = O(1)$  holds for *all* non-singular (CM) kernels,  $m(n, \epsilon) = O(\log_2 n)$  for *all* power-law singular (CM) kernels, and that more generally when fitting to intervals of the form  $[\kappa\Delta_n, T^*]$  in the setup of our scheme where  $\kappa = 0$  when  $K$  is non-singular,  $\kappa > 0$  otherwise, and the number of sampling points, up to rounding, is proportional to the number of steps of length  $\Delta_n$  over the interval. Our reasoning is as follows: That we observe  $m(n, \epsilon) = O(1)$  for the shifted power-law kernel where the interval is fixed suggests more generally that adding more sampling points to an interval, the number of exponentials  $m(n, \epsilon)$  eventually stops growing. As the rough fractional kernel can be viewed as a shifted power-law when the domain is bounded away from zero, the  $m(n, \epsilon) = O(\log_2 n)$  scaling is likely the result of the fitting interval (here  $[\frac{1}{n}, 1]$ ) moving towards the *power-law* singularity. We believe it is a mild extrapolation to suggest that the two scaling relationships should hold for more general intervals of the form  $[\kappa\Delta_n, T^*]$  in the given setup. However, it would be speculative to suggest, based on our results, that  $m(n, \epsilon) = O(\log_2 n)$  should hold for kernels that are singular in a non power-law way.

Let us briefly examine how  $m(n, \epsilon)$  depends on  $\epsilon$ . We refer to Figure 3.7 where we fix  $n = 500$  and vary  $\epsilon$ . We observe for all kernels logarithmic scaling of the form  $m(n, \epsilon) \approx c \log_{10}(\epsilon)$  where the  $c$ 's are negative. The scaling in  $\epsilon$  is mild, just like that in  $n$ ; for  $\epsilon = 10^{-6}$  we observe around 10 terms for the rough fractional kernels and around

### 3.6. SIMULATING THE VIX INDEX

**Table 3.2:** Comparison of BM2005 and the optimal solution for different kernels and error tolerances. Errors are reported as normalised  $l^2$ -error. We refer to the main text for details.

$\epsilon$	$K(t) = t^{-0.4}$				$K(t) = t^{-0.1}$			
	BM2005		Optimal		BM2005		Optimal	
	error	$m$	error	$m$	error	$m$	error	$m$
$10^{-1}$	$4.58 \cdot 10^{-2}$	3	$2.75 \cdot 10^{-2}$	3	$1.80 \cdot 10^{-2}$	2	$5.54 \cdot 10^{-2}$	1
$10^{-2}$	$2.75 \cdot 10^{-3}$	5	$6.88 \cdot 10^{-3}$	4	$5.51 \cdot 10^{-3}$	3	$3.63 \cdot 10^{-3}$	3
$10^{-3}$	$6.10 \cdot 10^{-4}$	6	$3.70 \cdot 10^{-4}$	6	$3.31 \cdot 10^{-4}$	5	$8.74 \cdot 10^{-4}$	4
$10^{-4}$	$2.69 \cdot 10^{-5}$	8	$2.69 \cdot 10^{-5}$	8	$7.24 \cdot 10^{-5}$	6	$7.23 \cdot 10^{-4}$	6
$10^{-5}$	$5.41 \cdot 10^{-6}$	9	$5.44 \cdot 10^{-6}$	9	$3.09 \cdot 10^{-6}$	8	$3.09 \cdot 10^{-6}$	8



**Figure 3.5:** The first few eigenvalues of  $H$  shown in descending order where  $N = 250$  computed for different kernels considered on  $[\frac{1}{500}, 1]$  (left plot) and  $[0, 1]$  (right plot).

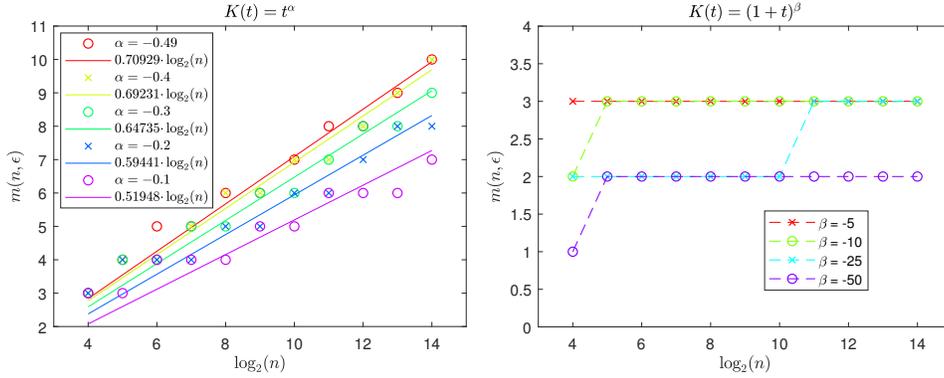
5 for the shifted power-laws. We can report comparable observations for other fixed  $n$ 's.

## 3.6 Simulating the VIX index

We discuss simulation of the VIX index for a series of volatility models of the Volterra type. Our motivation is the pricing of VIX options with Monte Carlo. Throughout we assume a risk-neutral model of the form (3.2) for some instantaneous variance  $V$ . We will assume that  $V$  is locally bounded and progressively measurable (e.g. locally bounded, left- or right-continuous, and adapted), non-negative, and satisfies  $\sup_{t \leq T} E[V_t] < \infty$  for all  $T > 0$ . The assumptions will hold for any of the models of the coming subsections. It should be noted that setting

$$S_t = S_0 \exp \left( \int_0^t \sqrt{V_s} dW_{1,s} - \frac{1}{2} \int_0^t V_s ds \right), \quad t \geq 0, \quad (3.89)$$

ensures that  $S$  satisfies (3.2) as follows by Ito's lemma. The pricing model (3.2) is thus guaranteed to have a solution once a valid instantaneous variance is specified.



**Figure 3.6:** Number of terms  $m(n, \epsilon)$  when using BM2005 with  $\epsilon = 10^{-3}$  to fit kernels  $K$  on intervals of the form  $[\frac{1}{n}, 1]$  (left plot) and  $[0, 1]$  (right plot) using  $n + 1$  sampling points. Actual errors are all below the target. Left:  $K(t) = t^\alpha$ . Observed values are shown as points, regressions of the form  $m(n, \epsilon) = c \log_2(n)$  for constants  $c$  as lines. Right:  $K(t) = (1 + t)^\beta$ .

We define the VIX index squared by

$$\text{VIX}_t^2 := \frac{100^2}{\Delta_{\text{vix}}} \int_0^{\Delta_{\text{vix}}} \tilde{\xi}_t(\tau) d\tau, \quad t \geq 0, \quad \Delta_{\text{vix}} := \frac{1}{12}, \quad (3.90)$$

where  $\tilde{\xi}_t(\tau) := E_t(V_{t+\tau})$ ,  $t, \tau \geq 0$ , are the forward variances and  $(\mathcal{F}_t)_{t \geq 0}$  is the market filtration which we will assume satisfies the usual hypothesis; we use ' $\sim$ ' as a superscript to avoid conflict with the common notation  $\xi_t(u) = E_t(V_u)$ ,  $0 \leq t \leq u$ . The definition (3.90) roughly coincides with that of the VIX whitepaper (downloadable via <https://www.cboe.com/education/research>, accessed July 25, 2021); we refer to e.g. [23]. A VIX option is either a put or call option on  $\text{VIX}_T$  for some expiry  $T > 0$ .

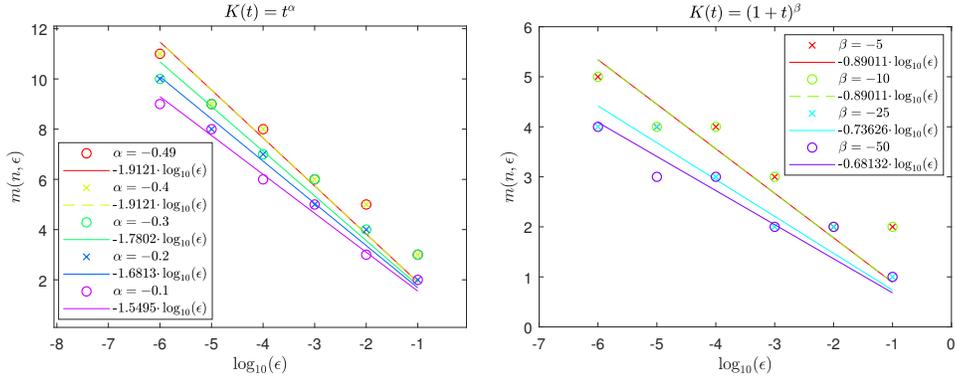
We will consider three classes of models for  $V$ . They will all be driven by SVE's with deterministic drift, though, it should be kept in mind that this is without loss of generality up linear drift in the sense of Theorem 3.3.4. As will become clear, to compute the VIX index, we recommend different methods for each model class. However, one starting point for an approximation that we will discuss, and which applies regardless of the model, is to start with a trapezoidal rule applied to (3.90) as

$$\text{VIX}_t^2 \approx \frac{100^2}{n_v} \sum_{i=0}^{n_v} a_i \tilde{\xi}_t(\tau_i), \quad t \geq 0, \quad (3.91)$$

where  $n_v \geq 2$  is an integer and  $a_0 = a_{n_v} = \frac{1}{2}$ ,  $a_i = 1$ ,  $i = 1, 2, \dots, n_v - 1$ ,  $\tau_i = i \frac{\Delta_{\text{vix}}}{n_v}$ ,  $i = 0, 1, \dots, n_v$ .<sup>39</sup> Sampling VIX then amounts to sampling a finite number of forward variances.

<sup>39</sup> Convergence is ensured in the limit  $n_v \rightarrow \infty$  assuming only that  $\tilde{\xi}_t(\cdot)$  on  $[0, \Delta_{\text{vix}}]$  is piecewise continuous and bounded with finitely many discontinuities since (3.90) then is valid as a Riemann integral; recall that the trapezoidal rule averages the left and right Riemann sums. Continuity of

### 3.6.1. MULTIFACTOR VOLTERRA BERGOMI MODEL



**Figure 3.7:** Number of terms  $m(n, \epsilon)$  in the sum-of-exponentials approximation when using BM2005 to fit kernels  $K$  on the interval  $[\frac{1}{n}, 1]$  (left plot) and  $[0, 1]$  (right plot) using  $n + 1$  sampling points where  $n = 500$ . Actual errors are all below the target. Observed values are shown as points, regressions of the form  $m(n, \epsilon) = c \log_{10}(\epsilon)$  for constants  $c$  as lines.

### 3.6.1 Multifactor Volterra Bergomi model

We consider a multifactor Volterra Bergomi model defined by

$$V_t = \xi_0(t) \exp \left( \eta \sum_{i=1}^d \theta_i X_{i,t} - \frac{\eta^2}{2} \theta^\top \Lambda(t) \theta \right), \quad (3.92)$$

$$X_{i,t} = \int_0^t K_i(t-s) dW_{i+1,s}, \quad t \geq 0, \quad i = 1, \dots, d, \quad (3.93)$$

where  $d \in \mathbb{N}$ ,  $\xi_0 : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  is left- or right-continuous,  $\sup_{t \leq T} |\xi_0(t)| < \infty$  for all  $T > 0$ ,  $\theta = (\theta_1, \dots, \theta_d)^\top \in [0, 1]^d$ ,  $\sum_{i=1}^d \theta_i = 1$ ,  $\eta \geq 0$ ,  $K_i \in L_{loc}^2(\mathbb{R}_+)$ ,  $i = 1, \dots, d$ ,  $W = (W_{1,t}, W_{2,t}, \dots, W_{d+1,t})_{t \geq 0}$  is  $d + 1$  dimensional correlated Brownian motion, and  $\Lambda(t) \in \mathbb{R}^{d \times d}$  is the covariance matrix of  $X = (X_{1,t}, \dots, X_{d,t})$  for  $t \geq 0$ . The  $\Lambda$ -function is computable with Ito's isometry and numerical integration. We assume that the filtration  $(\mathcal{F}_t)_{t \geq 0}$  is generated by  $W$  and augmented so the usual hypothesis holds. The model nests the rough and two-factor Bergomi models of [7, 11]. Note that our assumptions on  $\xi_0$  allows it to be a step function as is a common parameterisation for these models.

Define

$$g_{i,t}(\tau) = \int_0^t K_i(t + \tau - s) dW_{i+1,s}, \quad t, \tau \geq 0, \quad i = 1, \dots, d, \quad (3.94)$$

$\tilde{\xi}_t(\cdot)$  will hold for the quadratic Volterra Heston model of Section 3.6.2 and the generalised CEV Volterra volatility model of Section 3.6.3. The multifactor Volterra Bergomi model of Section 3.6.1 is easily parameterised so  $\tilde{\xi}_t(\cdot)$  on  $[0, \Delta_{\text{vix}}]$  is piecewise continuous and bounded with finitely many discontinuities.

which are the forward values for the components of  $X$ . The forward *variances* are expressible in terms of the forward *values* as:

$$\tilde{\xi}_t(\tau) = \xi_0(t + \tau) \exp \left( \eta \sum_{i=1}^d \theta_i \int_0^t K_i(t + \tau - s) dW_{i+1,s} - \frac{\eta^2}{2} \theta^\top \Lambda(t + \tau) \theta \right) \quad (3.95)$$

$$E_t \left( \exp \left( \eta \sum_{i=1}^d \theta_i \int_t^{t+\tau} K_i(t + \tau - s) dW_{i+1,s} \right) \right) \quad (3.96)$$

$$= \xi_0(t + \tau) \exp \left( \eta \sum_{i=1}^d \theta_i g_{i,t}(\tau) - \frac{\eta^2}{2} \theta^\top (\Lambda(t + \tau) - \Lambda(\tau)) \theta \right), \quad t, \tau \geq 0. \quad (3.97)$$

*Remark 3.6.1.* By Theorem 3.3.2, the mappings  $g_{i,t}(\cdot)$ ,  $i = 1, \dots, d$ , are continuous on  $\mathbb{R}_+$ . Thus, if we parameterise  $\xi_0$  (say) bounded and piecewise continuous with finitely many discontinuities, then (3.91) is valid as an approximation of (3.90); c.f. footnote 39.

For the present model, we consider only sampling methods that are based on (3.91). To sample  $\text{VIX}_T$  for some  $T > 0$  we must then jointly sample  $\tilde{\xi}_T(\tau_j)$ ,  $j = 0, \dots, n_v$ . By (3.95)-(3.97) that boils down to sampling jointly  $g_{i,T}(\tau_j)$  for  $i = 1, \dots, d$ ,  $j = 0, \dots, n_v$ , or  $\sum_{i=1}^d \theta_i g_{i,T}(\tau_j)$  for only  $j = 0, \dots, n_v$ . We outline three methods to this end.

**Exact sampling:** We sample  $\sum_{i=1}^d \theta_i g_{i,T}(\tau_j)$ ,  $j = 0, 1, \dots, n_v$ , directly from its  $n_v + 1$  dimensional normal distribution. The covariances are computable with Ito's isometry and numerical integration. The means are zero.

The next two methods rely on the hybrid multifactor approach. For these we shall therefore assume that  $K_i$ ,  $i = 1, \dots, d$ , are CM.

**Hybrid multifactor scheme:** We use the hybrid multifactor scheme to simulate  $(X_{i,t})_{0 \leq t \leq T}$ ,  $i = 1, \dots, d$ , and then evaluate  $(g_{i,T}(\tau_j))_{j=0}^{n_v}$ ,  $i = 1, \dots, d$ , with Definition 3.3.1. The kernel approximations must be valid on, at minimum, the interval  $[0, T + \Delta_{\text{vix}}]$ .

**Hybrid multifactor approach with direct sampling:** We use the hybrid multifactor approach, but not the scheme itself, to sample directly the variables that are necessary to compute the forward values  $(g_{i,T}(\tau_j))_{j=0}^{n_v}$ ,  $i = 1, \dots, d$ , as in Definition 3.3.1: Let  $\kappa \in \mathbb{N}_0$ ,  $n \in \mathbb{N}$ , and assume for simplicity that  $\kappa \Delta_n \leq \Delta_{\text{vix}}$  where we keep the definition  $\Delta_n := n^{-1}$ . Say furthermore for  $i = 1, \dots, d$ , that  $K_i(t) \approx 1_{(t \leq \kappa \Delta_n)} K_i(t) + 1_{(t > \kappa \Delta_n)} \sum_{j=1}^{m_i} c_{ij} e^{-\gamma_{ij} t}$  for  $t \in [0, T + \Delta_{\text{vix}}]$  where  $(c_{ij}, \gamma_{ij})_{j=1}^{m_i} \in \mathbb{R}^{2m_i}$ ,  $m_i \in \mathbb{N}$ . Define

$$U_{ij,T} := \int_0^T e^{-\gamma_{ij}(T-s)} dW_{i+1,s}, \quad i = 1, \dots, d, \quad j = 1, \dots, m_i. \quad (3.98)$$

Let  $i \in \{1, \dots, d\}$ . The derivations of Section 3.3 for  $\kappa = 0$  suggest

$$g_{i,T}(\tau) \approx \sum_{j=1}^{m_i} c_{ij} e^{-\gamma_{ij} \tau} U_{ij,T}, \quad \tau \in [0, \Delta_{\text{vix}}], \quad (3.99)$$

### 3.6.1. MULTIFACTOR VOLTERRA BERGOMI MODEL

---

and for  $\kappa = 1$  that

$$g_{i,T}(\tau) \approx \sum_{j=1}^{m_i} c_{ij} e^{-\gamma_{ij}\tau} U_{ij,T}, \quad \tau \in [\Delta_n, \Delta_{\text{vix}}], \quad (3.100)$$

$$g_{i,T}(\tau) \approx \frac{\Delta_n - \tau}{\Delta_n} X_{i,T} + \frac{\tau}{\Delta_n} g_{i,T}(\Delta_n), \quad \tau \in [0, \Delta_n]. \quad (3.101)$$

To obtain samples of  $(g_{i,T}(\tau_j))_{j=0}^{n_v}$ ,  $i = 1, \dots, d$ , for  $\kappa = 0$ , we can thus sample the variables (3.98) from their joint normal distribution and use (3.99). For  $\kappa = 1$ , we can use (3.100)-(3.101) if we additionally as part of the joint sample include the, also Gaussian, terms  $X_{i,T}$ ,  $i = 1, \dots, d$ . The idea generalises to  $\kappa > 1$  though yet more Gaussians must be sampled. Since, as noted,  $\kappa = 0$  or  $\kappa = 1$  generally suffices, the latter for the singular case even, we do not spell out the details for  $\kappa > 1$ . It is useful to note that the covariances among (3.98) are analytical: Let  $\nu_1, \nu_2 \in \mathbb{R}$ ,  $\nu_1 + \nu_2 \neq 0$ , and say that  $(Z_{1,t}, Z_{2,t})_{t \geq 0}$  is two-dimensional Brownian motion such that  $dZ_{1,t} dZ_{2,t} = \rho dt$ ,  $\rho \in [-1, 1]$ . Then

$$E \left( \int_0^T e^{-\nu_1(T-t)} dZ_{1,t} \int_0^T e^{-\nu_2(T-t)} dZ_{2,t} \right) = \frac{\rho}{\nu_1 + \nu_2} \left( 1 - e^{-(\nu_1 + \nu_2)T} \right). \quad (3.102)$$

When  $\nu_1 + \nu_2 = 0$  the left-hand side equals  $\rho T$ . If other covariances are required they will in general have to be computed with Ito's isometry and numerical integration.

We briefly compare the methods from a high level perspective. A numerical comparison is provided in Section 3.7 for the pricing of VIX options.<sup>40</sup> The evidence provided there suggest  $n_v \approx 16$  or 32 about suffices for all three methods. For our discussion, we will therefore assume that such  $n_v$ 's are chosen. One observation we have made is that if VIX options are to be priced at a single expiry then running the scheme from scratch is slower by a good margin compared to sampling exactly or directly under the hybrid multifactor approach. This is likely because  $n_v$  and  $m_i$ ,  $i = 1, \dots, d$ , need not be that large so the sampling dimensions of the former methods is small compared to the cost of first simulating the scheme, then computing the forward values with Definition 3.3.1.

The ranking of the methods could change when option prices are desired at more expiries. The reason is that the exact and direct sampling methods have to be completely rerun to compute values at different time points. In contrast, having run the scheme over some fixed time interval, we only need to reuse Definition 3.3.1 to evaluate forward values at additional time points. Our numerical evidence suggest that we must price options on least a handful of expiries before running the scheme from scratch is faster than the direct sampling methods. The feasibility of this however also depends on the degree to which we can use the same total number of steps for all expiries in some given fixed interval. We also find that if the hybrid multifactor scheme has been run for other purposes already, the extra cost of evaluating the VIX index is much less than the other methods. Such a situation arises when options on  $S$  and the VIX index both need to be priced and is relevant for the joint SPX-VIX calibration problem; a recent paper on the latter is [24]. Lastly, our experiments suggest that exact and direct sampling under the

<sup>40</sup> Technically we test on a two-factor *mixed* rough Bergomi model, though, we believe the overall conclusions are applicable in more generality. We explain how mixtures can be handled further down.

hybrid multifactor approach are about equally efficient, the latter slightly faster.

By linearity of conditional expectations, the derivations and sampling methods of the present subsection transfer without complications to models for  $V$  that are mixtures with log-normal terms of the form (3.92).<sup>41</sup> We exemplify how that works in Section 3.7.3.

In terms of the existing literature, we note that exact sampling was also considered in [33] for a comparable class of log-normal Volterra models and in [34] for the rough Bergomi model; the latter work contains additional methods that are disjoint from those examined here.

### 3.6.2 Quadratic Volterra Heston model

We define the quadratic Volterra Heston model by

$$V_t = a(Z_t - b)^2 + c, \quad (3.103)$$

$$Z_t = g_0(t) + \int_0^t K(t-s)\eta\sqrt{V_s}dW_{2,s}, \quad t \geq 0, \quad (3.104)$$

where  $a, c, \eta \geq 0$ ,  $b \in \mathbb{R}$ ,  $K \in L^2_{\text{loc}}(\mathbb{R}_+)$  satisfies  $(H_0)$  for some  $\tilde{\gamma} \in (0, 2]$ ,  $g_0 \in \mathcal{H}^{\tilde{\gamma}/2}$ ,  $W_2 = (W_{2,t})_{t \geq 0}$  is Brownian motion,  $dW_{1,t}dW_{2,t} = \rho dt$ ,  $\rho \in [-1, 1]$ . As the diffusion coefficient of (3.104) is Lipschitz continuous (in ' $Z_s$ ') there by Theorem 3.B.1 exists a unique continuous strong solution  $(V, Z)$  which we fix for the rest of the subsection.<sup>42</sup> We let the filtration  $(\mathcal{F}_t)_{t \geq 0}$  be that generated by  $(W_1, W_2)$  and augmented so the usual hypothesis holds. The model generalises the quadratic *rough* Heston model of [24].

The theorem below will be useful for computation of the VIX index.

**Theorem 3.6.1.** *Under the stated setup and assumptions, it holds that*

$$\tilde{\xi}_t(\tau) = f_t(\tau) + a\eta^2 \int_0^\tau K(\tau-s)^2 \tilde{\xi}_t(s) ds, \quad t, \tau \geq 0, \quad (3.105)$$

where

$$f_t(\tau) = a(g_t(\tau) - b)^2 + c, \quad (3.106)$$

$$g_t(\tau) = g_0(t + \tau) + \int_0^t K(t + \tau - s)\eta\sqrt{V_s}dW_{2,s}, \quad t, \tau \geq 0. \quad (3.107)$$

Equation (3.105) has the unique solution

$$\tilde{\xi}_t(\tau) = f_t(\tau) - \int_0^\tau \tilde{R}(\tau-s)f_t(s)ds, \quad t, \tau \geq 0, \quad (3.108)$$

---

<sup>41</sup> Log-normal models tend to produce almost flat VIX smiles whereas market smiles are typically upward sloping, see [33]. Log-normal mixtures precisely allow the creation of a skew; c.f. [12, 15, 28].

<sup>42</sup> Proof of Lipschitz claim: Define  $f(x) = \sqrt{a(x-b)^2 + c}$ ,  $x \in \mathbb{R}$ . Say  $c = 0$ . Lipschitz continuity then follows from the reverse triangle inequality:  $|f(x) - f(y)| = |\sqrt{a}|x-b| - |y-b|| \leq \sqrt{a}|x-y|$ ,  $x, y \in \mathbb{R}$ . Let  $c > 0$ . One may then show that  $\sup_{x \in \mathbb{R}} |f'(x)| = \sqrt{a}$  from which we conclude Lipschitz continuity by use of the mean-value theorem.

### 3.6.2. QUADRATIC VOLTERRA HESTON MODEL

---

where  $\tilde{R} \in L^1_{\text{loc}}(\mathbb{R}_+)$  is the resolvent of second kind of  $\tilde{K} := -a\eta^2 K^2$ . The mapping  $\tilde{\xi}_t(\cdot)$  is continuous on  $\mathbb{R}_+$  for any  $t \geq 0$ . Furthermore

$$\text{VIX}_t^2 = \frac{100^2}{\Delta_{\text{vix}}} \int_0^{\Delta_{\text{vix}}} [1 - \bar{R}(\Delta_{\text{vix}} - \tau)] f_t(\tau) d\tau, \quad t \geq 0, \quad (3.109)$$

where  $\bar{R}(t) = \int_0^t \tilde{R}(s) ds$ ,  $t \geq 0$ .

*Proof.* Let  $t, \tau \geq 0$ . By  $g_t(\tau) = E_t(Z_{t+\tau})$  as follows from Theorem 3.3.3, we have

$$E_t(Z_{t+\tau}^2) = \text{Var}_t(Z_{t+\tau}) + g_t(\tau)^2, \quad (3.110)$$

where the moment exists by (3.141). From the Ito isometry, we obtain

$$\text{Var}_t(Z_{t+\tau}) = E_t \left( \left[ \int_t^{t+\tau} K(t+\tau-s)\eta\sqrt{V_s} dW_{2,s} \right]^2 \right) \quad (3.111)$$

$$= \eta^2 \int_0^\tau K(\tau-s)^2 \tilde{\xi}_t(s) ds, \quad (3.112)$$

where we have also used Fubini's theorem and applied a change of variables.

Expanding the square in (3.103):

$$V_{t+\tau} = aZ_{t+\tau}^2 - 2abZ_{t+\tau} + ab^2 + c. \quad (3.113)$$

Taking conditional expectation, using (3.110)-(3.112), and  $g_t(\tau) = E_t(Z_{t+\tau})$ :

$$\tilde{\xi}_t(\tau) = ag_t(\tau)^2 - 2abg_t(\tau) + ab^2 + c + a\eta^2 \int_0^\tau K(\tau-s)^2 \tilde{\xi}_t(s) ds \quad (3.114)$$

$$= f_t(\tau) + a\eta^2 \int_0^\tau K(\tau-s)^2 \tilde{\xi}_t(s) ds. \quad (3.115)$$

For a fixed  $t$  the above is a linear Volterra equation for  $\tilde{\xi}_t(\cdot)$ . By [27, Theorem 2.3.5] a unique locally integrable solution exists if  $K^2$  and  $f_t(\cdot)$  are locally integrable. The first follows by the assumption  $K \in L^2_{\text{loc}}(\mathbb{R}_+)$ , the second because  $g_t(\cdot)$  and thus  $f_t(\cdot)$  are continuous on  $\mathbb{R}_+$ ; c.f. Theorem 3.3.2. Theorem 2.3.5 of [27] then gives the solution as

$$\tilde{\xi}_t(\tau) = f_t(\tau) - \int_0^\tau \tilde{R}(\tau-s) f_t(s) ds, \quad t, \tau \geq 0, \quad (3.116)$$

where  $\tilde{R} \in L^1_{\text{loc}}(\mathbb{R}_+)$  is the resolvent of second kind of  $\tilde{K} = -a\eta^2 K^2$ . From [27, Theorem 2.3.5] we also get that  $\tilde{\xi}_t(\cdot)$  is continuous since  $f_t(\cdot)$  is.

Lastly, by (3.116), Fubini's theorem, and the definition of  $\bar{R}$ :

$$\begin{aligned} \text{VIX}_t^2 &= \frac{100^2}{\Delta_{\text{vix}}} \int_0^{\Delta_{\text{vix}}} \tilde{\xi}_t(\tau) d\tau \\ &= \frac{100^2}{\Delta_{\text{vix}}} \int_0^{\Delta_{\text{vix}}} \left( f_t(\tau) - \int_0^\tau \tilde{R}(\tau-s) f_t(s) ds \right) d\tau \end{aligned}$$

$$\begin{aligned}
&= \frac{100^2}{\Delta_{\text{vix}}} \left[ \int_0^{\Delta_{\text{vix}}} f_t(\tau) d\tau - \int_0^{\Delta_{\text{vix}}} \int_s^{\Delta_{\text{vix}}} \tilde{R}(\tau - s) f_t(s) d\tau ds \right] \\
&= \frac{100^2}{\Delta_{\text{vix}}} \left[ \int_0^{\Delta_{\text{vix}}} f_t(\tau) d\tau - \int_0^{\Delta_{\text{vix}}} \bar{R}(\Delta_{\text{vix}} - s) f_t(s) ds \right] \\
&= \frac{100^2}{\Delta_{\text{vix}}} \int_0^{\Delta_{\text{vix}}} [1 - \bar{R}(\Delta_{\text{vix}} - \tau)] f_t(\tau) d\tau, \quad t \geq 0.
\end{aligned}$$

□

*Remark 3.6.2.* Since  $\tilde{\xi}_t(\cdot)$  is continuous on  $\mathbb{R}_+$  as follows from the theorem, it holds that (3.91) is valid as an approximation of (3.90).

We see three main ways to use Theorem 3.6.1 for computation of the VIX index. We present them in the following. In all cases, we will assume that  $K$  is CM and that  $Z$  has been simulated with the hybrid multifactor scheme across some time horizon which shall remain unspecified; by (3.103) we have values of  $V$  as well. Depending on the simulation horizon and the forward horizon to which we have fitted the kernel approximation, we have also values of  $g$  and  $f$  via Definition 3.3.1 and (3.106). The first method we consider is to apply a trapezoidal rule to (3.109) as

$$\text{VIX}_t^2 \approx \frac{100^2}{n_v} \sum_{i=0}^{n_v} a_i [1 - \bar{R}(\Delta_{\text{vix}} - \tau_i)] f_t(\tau_i), \quad t \geq 0, \quad (3.117)$$

where again  $n_v \geq 2$  is an integer and  $a_0 = a_{n_v} = \frac{1}{2}$ ,  $a_i = 1$ ,  $i = 1, \dots, n_v - 1$ ,  $\tau_i = i \frac{\Delta_{\text{vix}}}{n_v}$ ,  $i = 0, 1, \dots, n_v$ . The approximation is valid because  $f_t(\cdot)$  and  $\bar{R}$  are continuous, the former since  $g_t(\cdot)$  is continuous (we refer to Theorem 3.3.2), the latter by absolute continuity of the Lebesgue integral for (locally) integrable functions such as  $\tilde{R}$ . The values

$$\bar{R}(\Delta_{\text{vix}} - \tau_i) = \int_0^{\Delta_{\text{vix}} - \tau_i} \tilde{R}(s) ds, \quad i = 0, 1, \dots, n_v, \quad (3.118)$$

are computable with numerical integration if we can evaluate  $\tilde{R}$ . The cost is negligible in a Monte Carlo setting as the terms are deterministic.

As noted, resolvents such as  $\tilde{R}$  do not always have analytical expressions. Consider though the gamma kernel whose definition we restate below:

$$K(t) = e^{-\lambda t} \frac{t^{\alpha-1}}{\Gamma(\alpha)}, \quad t > 0, \quad \lambda \geq 0, \quad \alpha \in (1/2, 1]. \quad (3.119)$$

Under (3.119) we have

$$\tilde{K}(t) = -a\eta^2 K(t)^2 = c^* e^{-\lambda^* t} \frac{t^{\alpha^*-1}}{\Gamma(\alpha^*)}, \quad t > 0,$$

where  $\alpha^* = 2\alpha - 1$ ,  $c^* = -a\eta^2 \Gamma(\alpha^*) \Gamma(\alpha)^{-2}$ ,  $\lambda^* = 2\lambda$ . It follows that  $\tilde{K}$  also is a (scaled) gamma kernel. By [5, Table 1] or Proposition 3.A.1, we then, in fact, have an expression:

$$\tilde{R}(t) = c^* e^{-\lambda^* t} t^{\alpha^*-1} E_{\alpha^*, \alpha^*}(-ct^{\alpha^*}), \quad t > 0. \quad (3.120)$$

### 3.6.3. GENERALISED CEV VOLTERRA MODEL

---

For the general case, we could, like what we suggested for  $R_\lambda$  in Section 3.3, try to solve numerically with the scheme the equation  $\tilde{R} = \tilde{K} - \tilde{K} * \tilde{R}$  that defines  $\tilde{R}$ ; note that  $-\tilde{K} = a\eta^2 K^2$  is CM given that  $K$  is as complete monotonicity is preserved under products. As remarked in Section 3.3, there is however an unresolved problem in terms of how to handle the possibly singular initial curve that appears (here via  $\tilde{K}$ ). We are hopeful that a modified version of the scheme could be constructed to handle it.

In Section 3.7, we test (3.117) for the pricing of VIX options. The results that we present there show that convergence is reasonably fast in  $n_v$  and that values around 64 roughly suffices. Despite this, let us briefly outline two alternative methods that are based on (3.91): The first method is to compute the relevant forward variances with (3.108) using knowledge of  $\tilde{R}$  and  $f_t$  combined with a discretisation of the integral that appears. The second method is to compute the relevant forward variances by solving (3.105) with the hybrid multifactor scheme. We expect a discretisation of (3.108) combined with (3.91) to be less efficient than (3.117) as the former involves nested integration; so does (3.117) but one dimension is pre-computable. Also, although solving (3.105) does not rely on knowledge of  $\tilde{R}$ , we expect it to be even less efficient as we have to run the scheme.

### 3.6.3 Generalised CEV Volterra model

We consider the model

$$V_t = \xi_0(t) + \int_0^t K(t-s)\eta(V_s)dW_{2,s}, \quad t \geq 0, \quad (3.121)$$

where  $K \in L_{\text{loc}}^2(\mathbb{R}_+)$  is not identically zero and satisfies  $(H_0)$  for some  $\tilde{\gamma} \in (0, 2]$  and  $(H_2^*)$ . Using [27, Theorem 5.5.4], the assumptions imply that  $K$  also satisfies  $(H_1)$  as we will need in the following. We shall furthermore assume that  $\eta : \mathbb{R} \rightarrow \mathbb{R}$  is continuous of linear growth so  $\eta(0) = 0$ , that  $W_2 = (W_{2,t})_{t \geq 0}$  is Brownian motion with  $dW_{1,t}dW_{2,t} = \rho dt$ ,  $\rho \in [-1, 1]$ , and that  $\xi_0 \in \mathcal{H}_+^{\tilde{\gamma}/2}$ . We fix a non-negative continuous weak solution  $V$  which exists by Theorem 3.B.2 and let the market filtration  $(\mathcal{F}_t)_{t \geq 0}$  be that generated by  $(W_1, W_2, V)$  and augmented; for a strong solution we can let it be generated by  $(W_1, W_2)$  and augmented. We call (3.121) a *generalised* CEV Volterra volatility model as we allow  $\eta$  more general than the CEV case where  $\eta(x) = x^\beta$  for some  $\beta \in [0, 1]$ .

By Theorem 3.3.3 we have

$$\tilde{\xi}_t(\tau) = \xi_0(t+\tau) + \int_0^t K(t+\tau-s)\eta(V_s)dW_{2,s}, \quad t, \tau \geq 0. \quad (3.122)$$

Define

$$U_t(\gamma) := \int_0^t e^{-\gamma(t-s)}\eta(V_s)dW_{2,s}, \quad t, \gamma \geq 0.$$

We may then express  $\text{VIX}_t^2$  as a functional of  $(U_t(\gamma))_{\gamma \in [0, \infty)}$  as shown below.

**Theorem 3.6.2.** *Under the stated setup and assumptions it for  $t \geq 0$  holds that*

$$\text{VIX}_t^2 = \frac{100^2}{\Delta_{\text{vix}}} \left[ \int_0^{\Delta_{\text{vix}}} \xi_0(t+\tau)d\tau + U_t(0)\mu(\{0\})\Delta_{\text{vix}} + \int_{(0, \infty)} \frac{1}{\gamma}(1 - e^{-\gamma\Delta_{\text{vix}}})U_t(\gamma)\mu(d\gamma) \right].$$

CHAPTER 3. HYBRID MULTIFACTOR SCHEME FOR STOCHASTIC  
VOLTERRA EQUATIONS WITH COMPLETELY MONOTONE KERNELS

---

*Proof.* Note that (3.122) is the forward process for  $V$ . Then using Theorem 3.3.5 and Fubini's theorem:<sup>43</sup>

$$\begin{aligned} \int_0^{\Delta_{\text{vix}}} \tilde{\xi}_t(\tau) d\tau &= \int_0^{\Delta_{\text{vix}}} \xi_0(t + \tau) d\tau + \int_0^{\Delta_{\text{vix}}} \left( \int_{[0, \infty)} e^{-\gamma\tau} U_t(\gamma) \mu(d\gamma) \right) d\tau \\ &= \int_0^{\Delta_{\text{vix}}} \xi_0(t + \tau) d\tau + \int_{[0, \infty)} \left( \int_0^{\Delta_{\text{vix}}} e^{-\gamma\tau} d\tau \right) U_t(\gamma) \mu(d\gamma) \\ &= \int_0^{\Delta_{\text{vix}}} \xi_0(t + \tau) d\tau + U_t(0) \mu(\{0\}) \Delta_{\text{vix}} + \int_{(0, \infty)} \frac{1}{\gamma} (1 - e^{-\gamma\Delta_{\text{vix}}}) U_t(\gamma) \mu(d\gamma). \end{aligned}$$

The result follows.  $\square$

The formula of Theorem 3.6.2 resembles the numerical approximation of  $\text{VIX}_t^2$  that we will momentarily suggest for the case of a pure multifactor approximation. To explain our approximation, also for the general case, we invoke the below additional setup.

*Additional setup:* We assume the hybrid multifactor scheme of Definition 3.2.1 as applied with  $X = V$ ,  $W = W_2$ ,  $g_0 = \xi_0$ ,  $b = 0$ ,  $\sigma = \eta$ , and  $K$  as given in the current subsection. We retain the setup and notation of Section 3.2. In particular, we consider also a simulation interval  $[0, T]$  for some  $T > 0$ , though, we will more restrictively assume that the forward horizon satisfies  $T^* \geq T + \Delta_{\text{vix}}$  to justify our approximations of  $\text{VIX}_t^2$  for any  $t \in [0, T]$ . In other words, we will need  $K_{mn}$  to approximate  $K$  on at minimum the interval  $[0, T + \Delta_{\text{vix}}]$ ; this will be made rigorous in Theorem 3.6.3 further down (our convergence result). For any  $m, n$  we especially also keep the definitions of  $X_t^{mn} =: V_t^{mn}$ ,  $U_t^{mn} = (U_{1,t}^{mn}, \dots, U_{m,t}^{mn})^\top$ ,  $g_t^{mn}(\tau) =: \tilde{\xi}_t^{mn}(\tau)$  for  $t, \tau \geq 0$  and of  $\hat{X}_t^{mn} =: \hat{V}_t^{mn}$ ,  $\hat{U}_t^{mn} = (\hat{U}_{1,t}^{mn}, \dots, \hat{U}_{m,t}^{mn})^\top$ ,  $\hat{g}_t^{mn}(\tau) =: \hat{\xi}_t^{mn}(\tau)$  for  $(t, \tau) \in A$ . To ease the exposition, we will assume  $\kappa \Delta_n \leq \Delta_{\text{vix}}$ . We will also assume that  $V^{mn} = (V^{mn})_{t \geq 0}$  for any  $m, n$  has a unique continuous adapted solution on the same probability space as  $V$  and which is driven by the same Brownian motion  $W_2$ . This will be true under the assumptions of our convergence result, where we will assume more strictly that  $\eta$  is Lipschitz; c.f. Theorem 3.B.1 combined with Corollary 3.4.2 for why that suffices.

By Corollary 3.4.2 and Theorem 3.3.3, the below holds and is well-defined:

$$\tilde{\xi}_t^{mn}(\tau) = E_t(V_{t+\tau}^{mn}) = \xi_0(t + \tau) + \int_0^t K_{mn}(t + \tau - s) \eta(V_s^{mn}) dW_{2,s}, \quad t, \tau \geq 0.$$

We then define

$$\text{VIX}_{mn,t}^2 := \frac{100^2}{\Delta_{\text{vix}}} \int_0^{\Delta_{\text{vix}}} \tilde{\xi}_t^{mn}(\tau) d\tau, \quad t \geq 0. \quad (3.123)$$

Note that we cannot claim  $V^{mn}$  to be non-negative since  $K_{mn}$  need not satisfy  $(H_1)$  and which is needed for Theorem 3.B.2. As an example  $K_{mn}$  need not be continuous on  $(0, \infty)$ . Consequently, we cannot interpret  $V^{mn}$  as an instantaneous variance and

---

<sup>43</sup> The integrability required by Fubini's theorem follows by that implied in Theorem 3.3.5.

### 3.6.3. GENERALISED CEV VOLTERRA MODEL

therefore  $\text{VIX}_{mn,t}^2$ , although well-defined, cannot in general be interpreted as a VIX index squared in the sense of the official VIX white paper, neither can we claim it to be non-negative. However, we *can* still expect  $\tilde{\xi}_t^{mn}(\tau)$  to approximate  $\tilde{\xi}_t(\tau)$  for  $(t, \tau) \in A$ ; Lemma 3.G.2 of the appendix makes this rigorous under Assumption 3.4.1. It follows that we can also expect  $\text{VIX}_{mn,t}^2$  to approximate  $\text{VIX}_t^2$  for  $t \in [0, T]$ ; recall that  $T^* \geq T + \Delta_{\text{vix}}$ .

To derive our numerical approximation, note that

$$\begin{aligned} \text{VIX}_{mn,t}^2 = \frac{100^2}{\Delta_{\text{vix}}} & \left[ \int_0^{\kappa\Delta_n} \tilde{\xi}_t^{mn}(\tau) d\tau + \int_{\kappa\Delta_n}^{\Delta_{\text{vix}}} \xi_0(t + \tau) d\tau \right. \\ & \left. + \int_{\kappa\Delta_n}^{\Delta_{\text{vix}}} \int_0^t K_m(t + \tau - s) \eta(V_s^{mn}) dW_{2,s} d\tau \right], \quad t \geq 0, \end{aligned}$$

where

$$\begin{aligned} \int_{\kappa\Delta_n}^{\Delta_{\text{vix}}} \int_0^t K_m(t + \tau - s) \eta(V_s^{mn}) dW_{2,s} d\tau &= \int_{\kappa\Delta_n}^{\Delta_{\text{vix}}} \left( \sum_{i=1}^m c_i^m e^{-\gamma_i^m \tau} U_{i,t}^{mn} \right) d\tau \\ &= \sum_{\substack{i=1, \dots, m \\ \gamma_i^m \neq 0}} \frac{c_i^m}{\gamma_i^m} (e^{-\gamma_i^m \kappa\Delta_n} - e^{-\gamma_i^m \Delta_{\text{vix}}}) U_{i,t}^{mn} \\ &\quad + \sum_{\substack{i=1, \dots, m \\ \gamma_i^m = 0}} c_i^m (\Delta_{\text{vix}} - \kappa\Delta_n) U_{i,t}^{mn}, \quad t \geq 0. \end{aligned}$$

It follows that we for  $t \geq 0$  may write

$$\begin{aligned} \text{VIX}_{mn,t}^2 = \frac{100^2}{\Delta_{\text{vix}}} & \left[ \int_0^{\kappa\Delta_n} \tilde{\xi}_t^{mn}(\tau) d\tau + \int_{\kappa\Delta_n}^{\Delta_{\text{vix}}} \xi_0(t + \tau) d\tau \right. \\ & \left. + \sum_{\substack{i=1, \dots, m \\ \gamma_i^m \neq 0}} \frac{c_i^m}{\gamma_i^m} (e^{-\gamma_i^m \kappa\Delta_n} - e^{-\gamma_i^m \Delta_{\text{vix}}}) U_{i,t}^{mn} + \sum_{\substack{i=1, \dots, m \\ \gamma_i^m = 0}} c_i^m (\Delta_{\text{vix}} - \kappa\Delta_n) U_{i,t}^{mn} \right]. \end{aligned}$$

We then define our numerical approximation by

$$\widehat{\text{VIX}}_{mn,t_i}^2 := \frac{100^2}{\Delta_{\text{vix}}} \left[ \frac{\Delta_n}{2} \sum_{k=1}^{\kappa} \left[ \hat{\xi}_{t_i}^{mn}((k-1)\Delta_n) + \hat{\xi}_{t_i}^{mn}(k\Delta_n) \right] + \int_{\kappa\Delta_n}^{\Delta_{\text{vix}}} \xi_0(t_i + \tau) d\tau \right] \quad (3.124)$$

$$\left[ \sum_{\substack{j=1, \dots, m \\ \gamma_j^m \neq 0}} \frac{c_j^m}{\gamma_j^m} (e^{-\gamma_j^m \kappa\Delta_n} - e^{-\gamma_j^m \Delta_{\text{vix}}}) \hat{U}_{j,t_i}^{mn} + \sum_{\substack{j=1, \dots, m \\ \gamma_j^m = 0}} c_j^m (\Delta_{\text{vix}} - \kappa\Delta_n) \hat{U}_{j,t_i}^{mn} \right] \quad (3.125)$$

for  $i \in \{0, 1, \dots, \lfloor nT \rfloor\}$  and extend it as

$$\widehat{\text{VIX}}_{mn,t}^2 := \widehat{\text{VIX}}_{mn,t_n}^2, \quad 0 \leq t \leq T. \quad (3.126)$$

The first sum in (3.124) is the trapezoidal rule for  $\int_0^{\kappa\Delta_n} \hat{\xi}_{t_i}^{mn}(\tau) d\tau$  and we interpret it as zero when  $\kappa = 0$ . It is justified by continuity of  $\hat{\xi}_{t_i}^{mn}(\cdot)$  as follows by Theorem 3.3.2 and Corollary 3.4.2. The values of the trapezoidal discretisation are computable with Definition 3.3.1. The second term in (3.124) is left without further approximation as it is straightforward to evaluate with numerical integration independent of the given sample.

If  $\kappa, n$  are chosen so  $\kappa\Delta_n = \Delta_{\text{vix}}$  then (3.124)-(3.125) collapses to the trapezoidal rule (3.91) (with  $\hat{\xi}_t^{mn}$  in place of  $\tilde{\xi}_t$ ). Since  $\kappa$  is recommended low for the kernel approximation ( $\kappa = 0$  if non-singular,  $\kappa = 1$  otherwise) and since (3.124)-(3.125) exploits the fact that the exponential terms are analytically integrable, we expect it to be more efficient.

*Remark 3.6.3.* There is no guarantee that  $\widehat{\text{VIX}}_{mn,t}^2$  is non-negative (as with  $\text{VIX}_{mn,t}^2$ ) even if the notation suggests so. In practise, we thus recommend to truncate it in zero.

We provide a convergence result for the Lipschitz case below.

**Theorem 3.6.3.** *Invoke the stated setup and assumptions. Assume moreover that Equation (3.33) of Definition 3.2.1 is used, that  $\eta$  is Lipschitz continuous, and that  $\xi_0$  is locally Hölder continuous of order  $\tilde{\gamma}/2$ . Furthermore let there exist a  $\beta > 1$  so  $\|K\|_{L^{2\beta}([0, T+\Delta_{\text{vix}}])} < \infty$  and  $\lim_{m \rightarrow \infty} \|K - K_m\|_{L^{2\beta}([0, T+\Delta_{\text{vix}}])} = 0$ . Let  $p \geq \frac{2\beta}{\beta-1}$ ,  $m \in \mathbb{N}$ . There then exist positive constants  $C_p$  and  $C_{p,m}$ , the first of which depends on  $p$  only, the latter on both  $p$  and  $m$ ,<sup>44</sup> so for all  $n \in \mathbb{N}$  with  $\kappa\Delta_n \leq \Delta_{\text{vix}}$ :*

$$\sup_{t \in [0, T]} E \left[ |\text{VIX}_t^2 - \widehat{\text{VIX}}_{mn,t}^2|^p \right] \leq C_p \|K - K_{mn}\|_{L^{2\beta}([0, T+\Delta_{\text{vix}}])}^p + C_{p,m} n^{-\frac{p}{2}(\tilde{\gamma} \wedge 1)}.$$

Furthermore, for any  $p \geq 1$ :

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \sup_{t \in [0, T]} E \left[ |\text{VIX}_t^2 - \widehat{\text{VIX}}_{mn,t}^2|^p \right] = 0.$$

*Proof.* We refer to the appendix. □

## 3.7 Numerical experiments

In this section, we examine the scheme numerically. We look first at the strong error under the Gaussian process

$$X_t = \int_0^t K(t-s) dW_s, \quad t \geq 0, \quad (3.127)$$

for the kernels  $K(t) = t^\alpha$ ,  $\alpha \in (-\frac{1}{2}, 0]$ , and  $K(t) = (1+t)^\beta$ ,  $\beta \leq 0$ . Thereafter, we consider option pricing under models of the form (3.2) with standard Monte Carlo.

---

<sup>44</sup> As for Theorem 3.4.3 the constants generally also depend on the fixed setup listed in Remark 3.4.2.

### 3.7. NUMERICAL EXPERIMENTS

We will cover four models for  $V$ , including rough Bergomi and a two-factor mixed rough Bergomi model defined by

$$V_t = \xi_0(t) \left( \theta \exp \left( \eta \sqrt{2\alpha + 1} \int_0^t (t-s)^\alpha dW_{2,s} - \frac{\eta^2}{2} t^{2\alpha+1} \right) \right. \quad (3.128)$$

$$\left. + (1 - \theta) \exp \left( \nu \sqrt{2\beta + 1} \int_0^t (t-s)^\beta dW_{3,s} - \frac{\nu^2}{2} t^{2\beta+1} \right) \right), \quad t \geq 0, \quad (3.129)$$

where  $\xi_0 : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  is left- or right-continuous,  $\sup_{t \leq T} |\xi_0(t)| < \infty$  for all  $T > 0$ ,  $\theta \in [0, 1]$ ,  $\eta, \nu \geq 0$ ,  $\alpha, \beta \in (-\frac{1}{2}, 0]$ , and  $(W_{1,t}, W_{2,t}, W_{3,t})_{t \geq 0}$  is Brownian motion with  $dW_{1,t}dW_{2,t} = \rho_{12}dt$ ,  $dW_{1,t}dW_{3,t} = \rho_{13}dt$ ,  $dW_{2,t}dW_{3,t} = \rho_{23}dt$ , and  $\rho_{12}, \rho_{13}, \rho_{23} \in [-1, 1]$  so the associated correlation matrix is positive semi-definite. The market filtration is assumed to be that generated by  $(W_{1,t}, W_{2,t}, W_{3,t})_{t \geq 0}$  and augmented. We consider also a rough Heston model defined by (3.121) with  $K(t) = \Gamma(1 + \alpha)^{-1}t^\alpha$ ,  $\alpha \in (-\frac{1}{2}, 0]$ , and  $\eta(x) = \nu\sqrt{x}$ ,  $x \geq 0$ , for  $\nu \in \mathbb{R}_+$ .<sup>45</sup> Lastly, we examine a quadratic rough Heston model defined by (3.103)-(3.104) with  $K(t) = \Gamma(1 + \alpha)^{-1}t^\alpha$ ,  $\alpha \in (-\frac{1}{2}, 0]$ . For all models, we simplify by assuming  $\xi_0$  and  $g_0$  flat. This is more than enough to ensure that the equations have solutions including a non-negative one for the rough Heston equation.

In terms of numerics, we assume the following *unless otherwise stated*: To fit the sum-of-exponentials approximation we use BM2005 with  $\epsilon = 10^{-3}$  and generally fit to the interval  $[\kappa\Delta_n, T^*]$ . However, when  $\kappa = 0$  and  $K$  is singular, we fit to the interval  $[\Delta_n, T^*]$  as we cannot sample the singularity. We choose  $N$  so the sampling points, up to rounding, coincide with the extended simulation grid  $0, \frac{1}{n}, \frac{2}{n}, \dots$ . We always choose  $T^*$  as small as possible but yet so the relevant computations are valid. We use the explicit-implicit equation (3.34) and simulate  $S$  with a log-Euler scheme. To simulate  $V$  under rough Heston, we additionally truncate values in zero to ensure non-negativity; this also avoids problems when we compute the diffusion coefficient which has a square root. We always report option prices in Black-Scholes implied volatility. For a VIX option with expiry  $T$ , we use the corresponding VIX futures as the underlying when inverting the Black-Scholes equation. We will likewise estimate the futures price  $E(\text{VIX}_T)$  with Monte Carlo.

Our test models cover three important cases: Rough Bergomi covers the Gaussian case where there is no discretisation error from the diffusion coefficient. Rough Heston covers the case of state dependent diffusion coefficient and truncation in zero. Quadratic rough Heston covers the case of a state dependent diffusion coefficient and no truncation. On the latter, note that  $Z$ 's diffusion coefficient is bounded below by a strictly positive number when  $c, \eta > 0$  (as we will assume). Then, under that assumption, Theorem 3.B.2 does not apply and we can generally expect that it lives on the entire real line. The reason we have included the mixed two-factor rough Bergomi model (3.128)-(3.129) in our experiments is to test the pricing of VIX options in a two-factor model with skew.

<sup>45</sup> Throughout we have justified our approximations for strong solutions. For the rough Heston equation we cannot appeal to Theorem 3.B.1 for such as the square-root diffusion is non-Lipschitz. This however does not mean that the equation does not have a strong solution, we just do not know. Either way it will not stop us from simulating it with the hybrid multifactor scheme. Fortunately the evidence to be presented suggests that we have convergence for option prices: our Monte Carlo estimates for options on  $S$  converge to numbers produced by an applicable Fourier pricing method.

We consider the following parameters:

$$\text{Rough Bergomi : } (\alpha, \eta, \rho, \xi_0) = (-0.45, 3.06, -1, 0.16^2) \quad (3.130)$$

$$\text{Mixed rough Bergomi : } (\alpha, \beta, \theta, \eta, \nu, \rho_{23}, \xi_0) = (-0.45, -0.35, 0.3, 3, 1, 0.75, 0.15^2) \quad (3.131)$$

$$\text{Rough Heston (case A) : } (\alpha, \nu, \rho, \xi_0) = (-0.38, 0.29, -0.67, 0.15^2) \quad (3.132)$$

$$\text{Rough Heston (case B) : } (\alpha, \nu, \rho, \xi_0) = (-0.45, 0.41, -0.67, 0.15^2) \quad (3.133)$$

$$\text{Quadratic rough Heston : } (\alpha, \eta, \rho, a, b, c, g_0) = (-0.45, 0.93, 1, 0.36, 0.12, 0.0023, 0.037). \quad (3.134)$$

The parameter sets (3.132)-(3.133) are, up to rounding and disregarding  $\xi_0$ , taken from [16] where they are found by calibration to SPX options. They therefore represent realistic test cases, though, of varying difficulty; the case A equation with  $\alpha = -0.38$  and  $\nu = 0.29$  is less rough and has a smaller volatility-of-volatility than the case B equation where  $\alpha = -0.45$  and  $\nu = 0.41$ . To simplify our analysis, we consider only a single parameter set for each of the other models. The sets (3.130) and (3.134) are chosen by fixing  $\alpha = -0.45$  and calibrating to the expiry  $T = 0.1$  volatility smile for calls and puts on  $S_T$  as generated by the case B rough Heston model (the difficult case). When we compare the convergence behaviour between the models defined by (3.130), (3.133), and (3.134), it is then more likely that any differences reflect difficulties of the various types of SVE's, less so the specific parameters. The set (3.131) is not calibrated, but hand-picked to ensure a skew for VIX options.

### 3.7.1 Strong error in the Gaussian case

We consider simulation of (3.127) and the forward values  $g_t(\tau) = \int_0^t K(t + \tau - s) dW_s$ ,  $t, \tau \geq 0$ . We start with the influence of  $\kappa$  on the approximation of  $X$ : In Figure 3.8, we show, for a fixed Brownian path  $(W_t)_{t \in [0,1]}$ , paths of  $(X_t)_{t \in [0,1]}$  using exact simulation and the hybrid multifactor scheme with  $\kappa = 0$  and  $\kappa = 1$ . We have used 250 steps for the scheme. Trajectories are shown for the rough fractional kernel  $K(t) = t^{-0.4}$  and the shifted power-law kernel  $K(t) = (1+t)^{-20}$ . For the former,  $\kappa = 0$  produces a poor approximation with too little variability whereas  $\kappa = 1$  gives an almost perfect match. As noted, the same observation was made in [9] for their hybrid scheme. The choice  $\kappa = 0$  on the other hand seems sufficient for the (non-singular) shifted power-law kernel.

In what follows, we examine more systematically how the strong error depends on  $\kappa$ . For simplicity, we will only consider the error at the terminal time point  $T$ . We will look at the root-mean-squared-error (RMSE), which to avoid ambiguity, we define for  $X_T$  by:

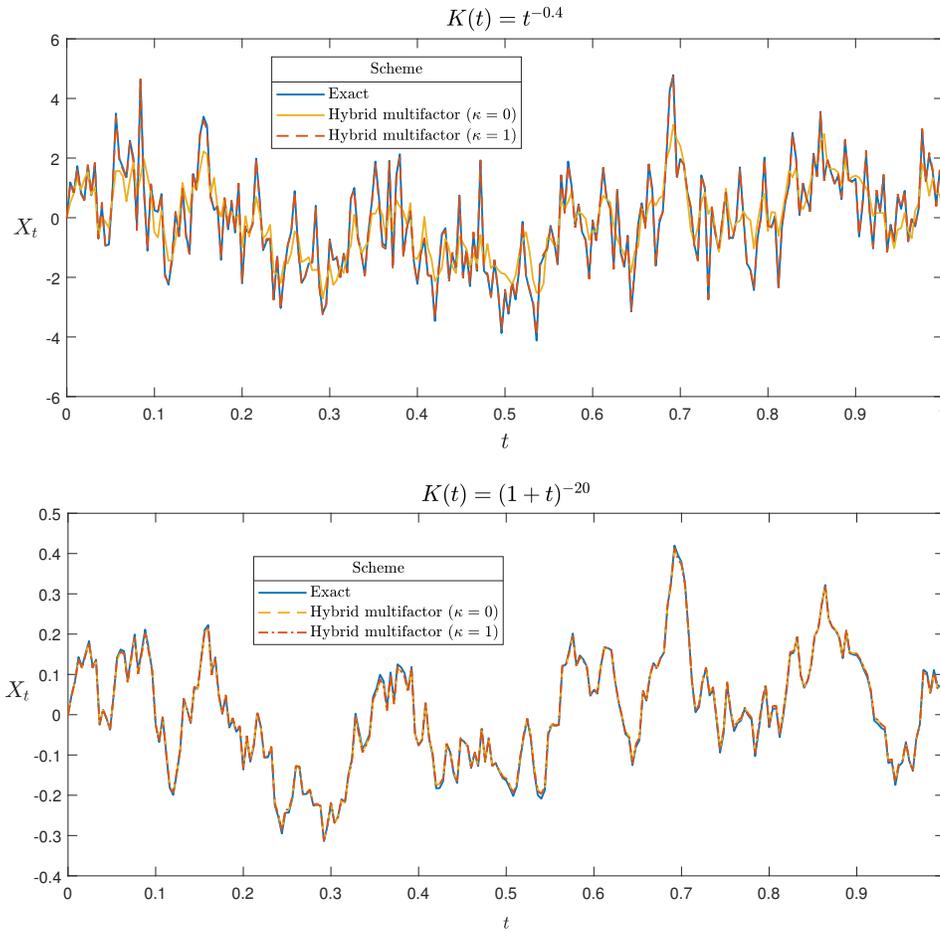
$$\left( E \left[ |X_T - \hat{X}_T^{mn}|^2 \right] \right)^{\frac{1}{2}}. \quad (3.135)$$

In Figure 3.9, we show Monte Carlo estimates of the RMSE for  $X_T$  with  $T = 1$  normalised by the standard deviation of  $X_T$ . For the rough fractional kernel we show values under the hybrid multifactor scheme (lines with crosses) and the hybrid scheme of [9] (lines with circles). For the shifted power-law kernel, we only show values under the hybrid multifactor scheme as the  $\alpha = 0$  case is excluded in [9] for the  $\mathcal{TBSS}$  process (3.4).

The results are much the same. For the rough fractional kernel, we again observe a

### 3.7.1. STRONG ERROR IN THE GAUSSIAN CASE

---



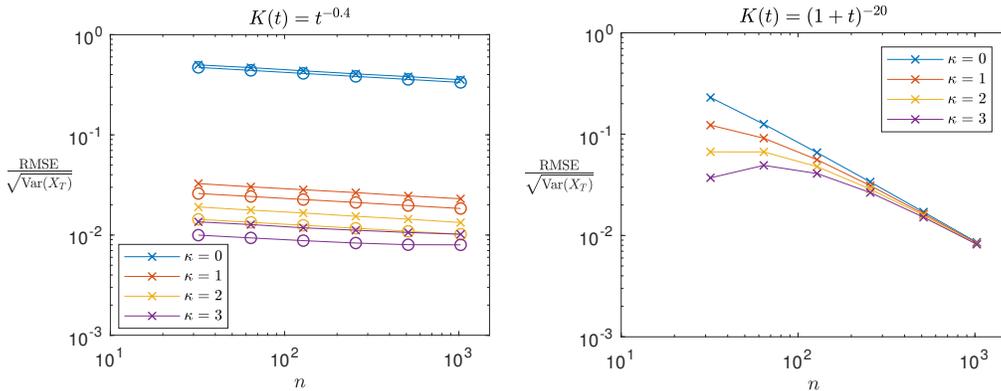
**Figure 3.8:** Trajectories of  $(X_t)_{t \in [0,1]}$  defined by (3.127) for different kernels and simulation methods. The Brownian path  $(W_t)_{t \in [0,1]}$  is held fixed. Blue: exact simulation. Yellow and red: hybrid multifactor scheme with  $\kappa = 0$  and  $\kappa = 1$ . We use 250 simulation steps in all cases.

large improvement going from  $\kappa = 0$  to  $\kappa = 1$ , but only small improvements with higher  $\kappa$ 's. For the shifted power-law kernel, we likewise again observe no worthwhile benefit in choosing  $\kappa > 0$  instead of  $\kappa = 0$  (except for small values of  $n$ ).

There are more observations to make from Figure 3.9. The first relates to the convergence rate. Note that we technically cannot appeal to the rate  $\frac{1}{2}(\tilde{\gamma} \wedge 1)$  that Theorem 3.4.3 suggests. There are three reasons for this: (1) the convergence rate is for the discretisation error and does not cover that of the kernel approximation whereas (3.135) includes both, (2) the way we use the scheme  $m = m(n, \epsilon)$  depends on  $n$  whereas the rate for the discretisation error of Theorem 3.4.3 is meaningful for a fixed  $m$  only, (3) we are using the explicit-implicit equation (3.34). Despite this, numerical evidence (to be presented) suggests that we yet obtain a convergence rate of  $\frac{1}{2}(\tilde{\gamma} \wedge 1)$  (i.e.  $\alpha + \frac{1}{2}$ ) for  $X$  when  $K$  is the rough fractional kernel; in other cases evidence suggests that the convergence rate can be strengthened. Figure 3.9 reflects *qualitatively* the idea that the convergence rate for  $X$  is related positively with the regularity of the process as it is much faster under the shifted power-law kernel than under the rough fractional kernel.

Another worthwhile observation is that the convergence rate looks to be independent of  $\kappa$ . This is not too surprising given Theorem 3.4.3 when we consider the rough fractional kernel and compare  $\kappa$ 's strictly greater than zero, and when we consider the shifted power-law kernel; recall though the mentioned caveats for interpreting Theorem 3.4.3 in our context. What is surprising, is that the convergence rate also looks to be independent of  $\kappa$  when we compare  $\kappa = 0$  with  $\kappa > 0$  for the rough fractional kernel; c.f. our discussion in Section 3.4 in that we for  $\kappa = 0$  can strengthen the rate to  $\frac{1}{2}$  for the discretisation error if  $g_0$ , like here, is locally Hölder continuous of at least order  $\frac{1}{2}$ . This shows that there is no benefit in choosing  $\kappa = 0$  over  $\kappa > 0$  *also* in terms of the convergence rate.

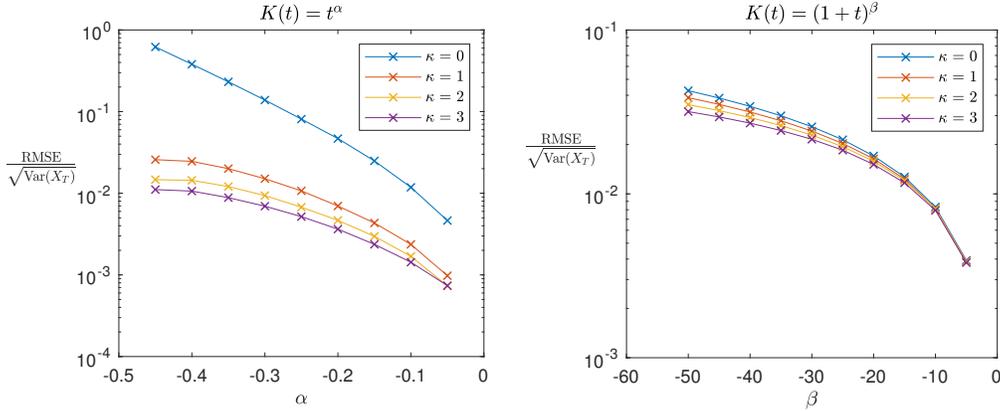
Lastly, note that the hybrid multifactor scheme and the hybrid scheme of [9] behave comparably for the same  $(n, \kappa)$ . We believe this is meaningful as the schemes handle the singularity similarly and both employ Euler approximations for the remainder, though, in different ways. As discussed, they instead, generally, differ in terms of running time.



**Figure 3.9:** Strong error for  $X_T$  defined by (3.127) with  $T = 1$ . Estimates are based on 1 million samples. Lines with crosses: hybrid multifactor scheme. Lines with circles: hybrid scheme of [9] (only for the left plot). Standard errors are less than 2% of the shown values.

### 3.7.1. STRONG ERROR IN THE GAUSSIAN CASE

We briefly examine the influence of  $\kappa$  on the error across values of  $\alpha$ , respectively,  $\beta$ , for the kernels  $K(t) = t^\alpha$  and  $K(t) = (1+t)^\beta$ . We refer to Figure 3.10, where we keep  $\lfloor nT \rfloor = 250$  fixed. For the former kernel, the error, as expected, worsens as  $\alpha \rightarrow -\frac{1}{2}$  and the benefit of choosing  $\kappa > 0$  becomes larger. However, even with  $\alpha = -0.05$  there is yet a very noticeable improvement using  $\kappa = 1$  instead of  $\kappa = 0$ . For the shifted power-law kernel, we find no worthwhile improvements in setting  $\kappa > 0$  even for  $\beta$ 's near  $-50$ . We conclude that our recommendations for  $\kappa$  are robust across a wide range of parameters.



**Figure 3.10:** Strong error for  $X_T$  defined by (3.127) with  $T = 1$  under the hybrid multifactor scheme with  $\lfloor nT \rfloor = 250$ . Estimates are based on 1 million samples. Standard errors are less than 4% (left plot) and 0.1% (right plot) of the shown values.

In what follows, we examine the convergence rate in more detail. We will look at that of  $\hat{X}_T^{mn}$  and  $\hat{g}_T^{mn}(\tau)$  towards respectively  $X_T$  and  $g_T(\tau)$  for  $T = 1$ ,  $\tau = 0.1$ , with  $K(t) = t^\alpha$ . We keep  $\kappa = 1$  as recommended. Disregarding the many caveats, Theorem 3.4.3 hints at the below convergence behaviour in  $n$  for the RMSE of  $\hat{X}_T^{mn}$  against  $X_T$ :

$$\left( E \left[ |X_T - \hat{X}_T^{mn}|^2 \right] \right)^{\frac{1}{2}} \approx O(n^{-(\alpha + \frac{1}{2})}). \quad (3.136)$$

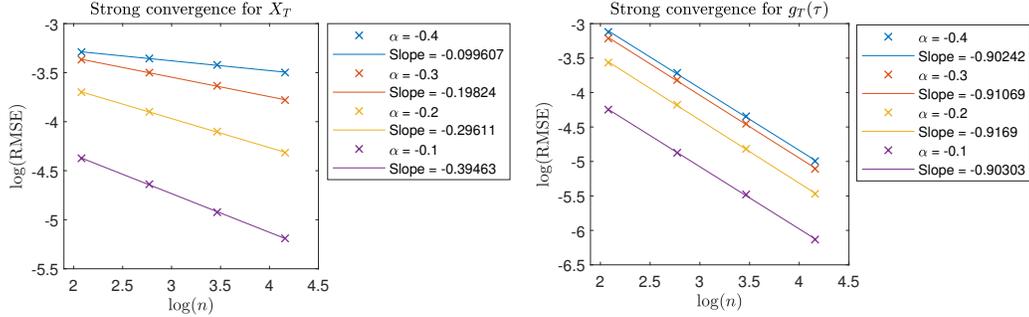
The theorem alludes to the same convergence behaviour for  $g_T(\tau)$ . However, as will become clear momentarily, numerical evidence suggest that the RMSE of  $\hat{g}_T^{mn}(\tau)$  against  $g_T(\tau)$ , for the given example, is better described by:

$$\left( E \left[ |g_T(\tau) - \hat{g}_T^{mn}(\tau)|^2 \right] \right)^{\frac{1}{2}} \approx O(n^{-1}). \quad (3.137)$$

In Figure 3.11, we show the  $\log(\text{RMSE})$  computed with Monte Carlo against  $\log(n)$  for  $\hat{X}_T^{mn}$  and  $\hat{g}_T^{mn}(\tau)$ . We sample the pathwise errors  $X_T - \hat{X}_T^{mn}$  and  $g_T(\tau) - \hat{g}_T^{mn}(\tau)$  by sampling jointly  $X_T$  and  $g_T(\tau)$  together with the Gaussian terms that are needed to construct  $\hat{X}_T^{mn}$  and  $\hat{g}_T^{mn}(\tau)$  under the scheme. Slope estimates are added by linear regression.

The slope estimates for convergence of  $\hat{X}_T^{mn}$  towards  $X_T$  are largely in line with (3.136) despite the many caveats when interpreting Theorem 3.4.3 in our context. For the forward values, we observe, regardless of  $\alpha$ , slopes around  $-0.9$ , close to  $-1$ . This suggests

that (3.137) roughly holds. While a rate close to 1 goes against that suggested by Theorem 3.4.3, there are theoretical reasons why it makes sense: Recall that we for  $n$  large enough, for this example, have  $\hat{g}_T^{mn}(\tau) = \sum_{i=1}^m c_i^m e^{-\gamma_i^m \tau} \hat{U}_{i,T}^{mn}$  and similarly for  $g_T^{mn}(\tau)$ . As  $U^{mn}$  solves a multivariate SDE with constant diffusion coefficients, the Euler and Milstein schemes for it coincide; compare pages 340 and 345 of [45]. Thus it is at least *meaningful* that we should attain the Milstein rate for SDE's which is 1;<sup>46</sup> c.f. [45, Theorem 10.3.5] for the absolute ( $p = 1$ ) error. The same logic does not apply to  $\hat{X}_T^{mn}$  when  $\kappa > 0$  as it then, regardless of  $n$ , is expressed in terms of other variables than  $\hat{U}^{mn}$  only; consequently there is no conflict in not observing the Milstein rate for the spot value  $X_T$  in Figure 3.11. Based on our logic, one might though claim that we, in fact, should attain the Milstein rate for  $X_T$  when  $\kappa = 0$  where  $\hat{X}_T^{mn} = \sum_{i=1}^m c_i^m \hat{U}_{i,T}^{mn}$  and similarly for  $X_T^{mn}$ . This, however, is disproved by Figure 3.9 (left) where the convergence rate for  $X_T$  is the same for  $\kappa = 0$  and  $\kappa > 0$ . A likely reason is this: For  $X_T$  we use  $K$  near the singularity and thus, although the convergence rate of the discretisation error may be that of the Milstein scheme when  $\kappa = 0$ , the kernel error is much worse. We do not expect to lose the Milstein rate of convergence for  $g_T(\tau)$  even when  $\kappa = 0$  as we only use  $K$  away from the singularity for that variable (recall for our example that  $\tau > 0$ ). Figure 3.9 (right plot) shows that we for the spot process under the shifted power-law kernel also obtain a rate close to 1 which likewise is meaningful as the kernel is non-singular.



**Figure 3.11:** Convergence of  $\hat{X}_T^{mn}$  and  $\hat{g}_T^{mn}(\tau)$  towards  $X_T$  (defined by (3.127)) and  $g_T(\tau)$  where  $T = 1$ ,  $\tau = 0.1$ ,  $K(t) = t^\alpha$ ,  $\kappa = 1$ . Estimates are based on 100 000 samples. Points show observed errors. Standard errors of the RMSE are within 1%. Lines show regressions of the form  $\log(\text{RMSE}) = a + b \log(n)$  for constants  $a$  and  $b$ . Slope estimates in the legend refer to  $b$ .

We return to the claims that we made about the complexity of the scheme when combined with BM2005. More precisely, we will back up the only remaining point to finish our argument that the complexity of the hybrid multifactor scheme is  $O(n)$  when  $K$  is (CM and) non-singular,  $O(n \log_2 n)$  when it is (CM and) power-law singular. That is, we will argue that it practically suffices to keep  $\epsilon$  fixed and only consider the convergence in  $n$ . We therefore now explore the joint influence of  $n$  and  $\epsilon$  on the approximations. In Table 3.3 we show the strong error for  $X_T$  with  $T = 1$  for different combinations of  $(n, \epsilon)$  for the kernel  $K(t) = t^{-0.4}$  where  $\kappa = 1$ . We see that the approximations convergence rapidly in  $\epsilon$  which confirms our claim. Even the choice  $\epsilon = 10^{-3}$  seems sufficient.

<sup>46</sup> We can expect the discretisation error (here that between  $\hat{U}_T^{mn}$  and  $U_T^{mn}$ ) to dominate for the total error when  $K_{mn}$  approximates  $K$  well on the domain where it is used such as for this example.

### 3.7.1. STRONG ERROR IN THE GAUSSIAN CASE

---

Although unreported, we observe the same, i.e. that convergence quickly saturates in  $\epsilon$ , for other  $\alpha$ 's and for the kernel  $K(t) = (1+t)^\beta$  across different  $\beta$ 's where we keep  $\kappa = 0$ .

**Table 3.3:** Strong error for  $X_T$  defined by (3.127) where  $K(t) = t^{-0.4}$  as computed under the hybrid multifactor scheme with  $\kappa = 1$  and  $T = 1$ . The RMSE is that defined by (3.135). Estimates are based on 100 000 samples. Standard errors are less than 0.0002.

$\frac{\text{RMSE}}{\sqrt{\text{Var}(X_T)}}$		$n$					
		16	32	64	128	256	512
$\epsilon$	$10^{-1}$	0.0334	0.0313	0.0360	0.0506	0.0270	0.0340
	$10^{-2}$	0.0348	0.0320	0.0291	0.0278	0.0254	0.0241
	$10^{-3}$	0.0348	0.0326	0.0303	0.0283	0.0266	0.0246
	$10^{-4}$	0.0350	0.0327	0.0305	0.0284	0.0267	0.0248
	$10^{-5}$	0.0351	0.0327	0.0305	0.0285	0.0267	0.0249

We examine now the computational times. We will compare against the hybrid scheme of [9]. In Table 3.4, we show running times for the simulation of 10 000 paths of  $(X_t)_{t \in [0, T]}$  for different  $\lfloor nT \rfloor$  where  $K(t) = t^{-0.4}$ ,  $T = 1$ ,  $\kappa = 1$ . We exclude the time used to sample the underlying standard normals  $\{Z_i\}_{i=0}^{\lfloor nT \rfloor - 1}$  defined on page 37. There is a practical justification for this: when calibrating an option pricing model to market data with Monte Carlo, the simulations are often rerun with a fixed set of random numbers.

Columns 2–3 show running times for the hybrid multifactor scheme and the hybrid scheme of [9] implemented with the FFT. Running times for the hybrid multifactor scheme scale about linearly in  $n$ . Although the theoretical running time is  $O(n \log_2 n)$  recall that the constant in front of  $m(n, \epsilon) = O(\log_2 n)$  is small which could explain it. Running times under the FFT based hybrid scheme of [9] increase faster than linearly but less than quadratically (at least for large  $\lfloor nT \rfloor$ 's). This is in line with its theoretical running time which also is  $O(n \log_2 n)$ . The ratio between the two (column 5) increases from a factor 1.10 in favour of the hybrid multifactor scheme with 32 steps to a factor 3.46 with 2048 steps. Since the schemes are comparably accurate for the same  $\lfloor nT \rfloor$ , c.f. Figure 3.9 (left), we conclude that the hybrid multifactor scheme is more efficient.

As noted, we cannot, for the hybrid scheme of [9], use the FFT to achieve costs  $O(n \log_2 n)$  for the  $\mathcal{TBSS}$  process (3.4) when we are unable to pre-simulate  $(\sigma_t)_{t \in [0, T]}$ . Consequently, if we generalise the scheme of [9] to a state-dependent SVE, we can only expect costs  $O(n^2)$ . For a more fair comparison in the general case, we therefore also implemented the scheme of [9] with the convolution of (3.39) computed using a loop in the time dimension. We show the running times for this loop based implementation in column 4 and the ratio to the hybrid multifactor scheme in column 6. We see that the scaling indeed is close to quadratic. The ratio of running times increases from 1.54 in favour of the hybrid multifactor scheme (with 32 steps) to a substantial ratio of 192.80 (with 2048 steps). We conclude that the idea behind [9] is very costly for a state-dependent SVE.

CHAPTER 3. HYBRID MULTIFACTOR SCHEME FOR STOCHASTIC  
VOLTERRA EQUATIONS WITH COMPLETELY MONOTONE KERNELS

**Table 3.4:** Running times (in milliseconds) for the simulation of 10 000 paths of  $(X_t)_{t \in [0, T]}$  defined by (3.127) with  $K(t) = t^{-0.4}$  for different  $\lfloor nT \rfloor$  where  $T = 1$ ,  $\kappa = 1$ . Time spent sampling the underlying standard normals is excluded. Column 2: hybrid multifactor scheme. Column 3: hybrid scheme of [9] where the convolution in (3.39) is computed with the FFT. Column 4: hybrid scheme of [9] where the convolution is computed with a loop in the time-dimension. Columns 5 and 6: ratios between columns 3 and 2, respectively, columns 4 and 2. Running times are averaged over 1000 runs. Standard errors are within 3% of the averages.

$\lfloor nT \rfloor$	Hybrid multi -factor	Hybrid TBSS	Hybrid TBSS (loop)	Ratio vs. Hybrid TBSS	Ratio vs. Hybrid TBSS (loop)
32	32	19	27	1.10	1.54
64	24	38	92	1.60	3.86
128	40	75	353	1.88	8.83
256	65	139	1332	2.14	20.53
512	120	282	5346	2.36	44.63
1024	242	767	21194	3.17	87.65
2048	468	1616	90172	3.46	192.80

### 3.7.2 Rough Bergomi

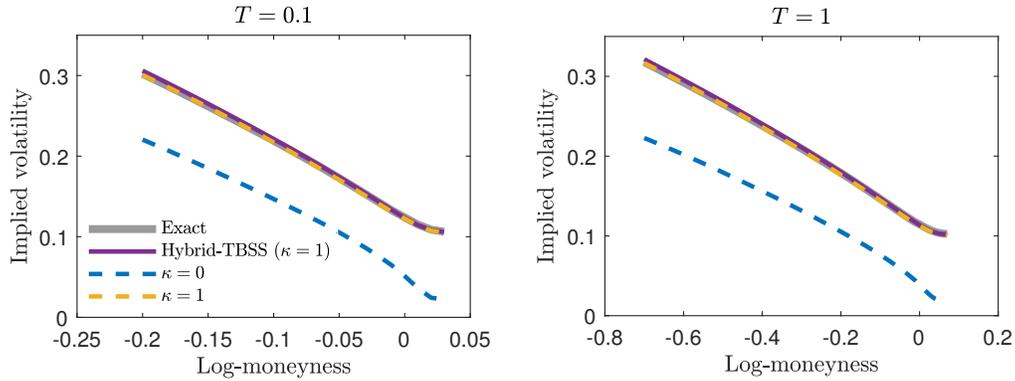
We consider now the rough Bergomi model with parameters (3.130). In Figure 3.12, we show Monte Carlo estimates of European calls and puts on  $S_T$  for the expiries  $T = 0.1$  and  $T = 1$  under various simulation methods (exact simulation of  $V$  included) and  $\kappa$ 's. We have used  $\lfloor nT \rfloor = 500$  steps. The first axes show the *log-moneyness* defined by  $\log(\bar{K}/S_0)$  where  $\bar{K}$  is the strike. From the plots, we confirm that  $\kappa = 1$  is sufficient for the rough fractional kernel also in terms of pricing. As we will only deal with the rough fractional kernel from here on, we simplify by fixing  $\kappa = 1$  for the rest of Section 3.7.

In Figure 3.13, we examine the convergence in the number of steps for the expiry  $T = 0.1$ . The left plot shows the volatility smiles under the hybrid multifactor scheme; the red line highlights a benchmark solution computed with  $\lfloor nT \rfloor = 10\,000$  steps. We note that the smiles converge reasonably fast; only a few hundred steps seems to suffice. In the right plot, we show the convergence of the absolute error (i.e. bias) in implied volatility versus the benchmark solution for different fixed moneyness values; dashed lines show the hybrid multifactor scheme, solid lines the hybrid scheme of [9]. We note that the schemes converge about equally fast in line with the conclusion from the previous subsection.

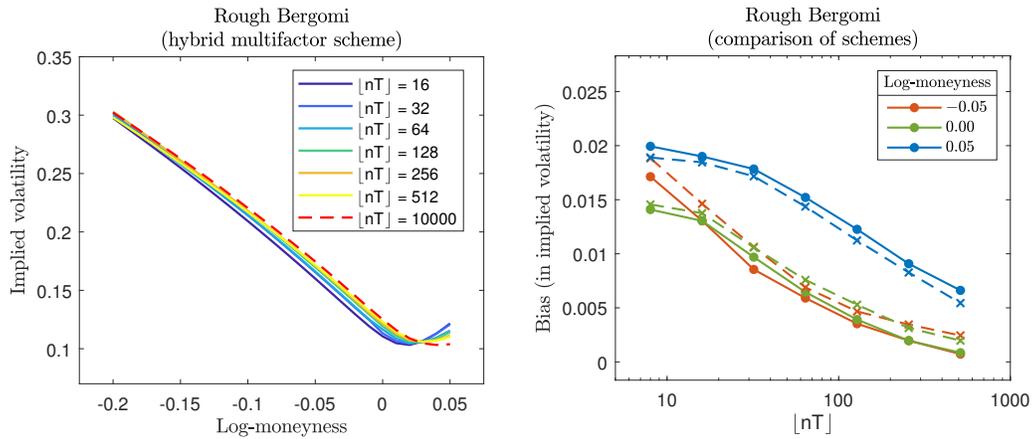
### 3.7.3 Mixed two-factor rough Bergomi

We consider the pricing of VIX options under the two-factor mixed rough Bergomi model (3.128)-(3.129). In Table 3.5, we present the results of an experiment where we have used Monte Carlo to price the VIX option with strike 25 and expiry  $T = 0.1$  always using the trapezoidal rule (3.91) to compute the VIX index. Values are shown for different  $n_v$ 's and methods for sampling the forward variances. In the left part of the table, we show implied volatilities, in the right part computational times, though, the latter excludes

### 3.7.3. MIXED TWO-FACTOR ROUGH BERGOMI



**Figure 3.12:** Volatility smiles for options on  $S_T$  under rough Bergomi with parameters (3.130). Prices are computed with Monte Carlo using  $\lfloor nT \rfloor = 500$  steps. Lines differ by method used to simulate the Gaussian Volterra integral of (3.3). Grey: Exact simulation. Magenta: Hybrid scheme of [9] with  $\kappa = 1$ . Blue and yellow: Hybrid multifactor scheme with  $\kappa = 0$  and  $\kappa = 1$ . Estimates are based on 100 000 samples. Standard errors are within 2% of the shown values.



**Figure 3.13:** Monte Carlo estimation of European calls and puts on  $S_T$  with expiry  $T = 0.1$  under rough Bergomi. Left: Convergence under the hybrid multifactor scheme. Right: Convergence at selected moneyness values. Dashed lines: Hybrid multifactor scheme. Solid lines: Hybrid scheme of [9]. Model parameters are those of (3.130) and we keep  $\kappa = 1$  for both schemes. Estimates are based on 100 000 samples. Standard errors are within 1% of the values.

the time spent sampling the underlying standard normal variables. Covariances, except those of the form (3.102), are computed with numerical integration.

By linearity of conditional expectations and (3.95)-(3.97) we for  $t, \tau \geq 0$  deduce

$$\tilde{\xi}_t(\tau) = \xi_0(t + \tau) \left\{ \theta \exp \left( \eta g_{1,t}(\tau) - \frac{\eta^2}{2} [(t + \tau)^{2\alpha+1} - \tau^{2\alpha+1}] \right) \right. \quad (3.138)$$

$$\left. + (1 - \theta) \exp \left( \nu g_{2,t}(\tau) - \frac{\nu^2}{2} [(t + \tau)^{2\beta+1} - \tau^{2\beta+1}] \right) \right\} \quad (3.139)$$

where

$$g_{1,t}(\tau) := \sqrt{2\alpha + 1} \int_0^t (t + \tau - s)^\alpha dW_{2,s},$$

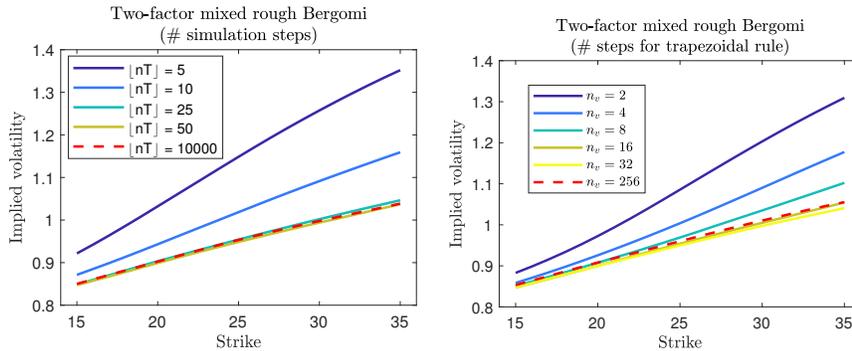
$$g_{2,t}(\tau) := \sqrt{2\beta + 1} \int_0^t (t + \tau - s)^\beta dW_{3,s}, \quad t, \tau \geq 0.$$

Let us explain Table 3.5 in more detail: For all methods, we first sample, possibly with error, the values  $g_{1,T}(\tau_i), g_{2,T}(\tau_i), i = 0, 1, \dots, n_v$ . Thereafter, we plug them into (3.138)-(3.139) to obtain the forward variances  $\tilde{\xi}_T(\tau_i), i = 0, 1, \dots, n_v$ , or, at least, approximations thereof. Next, we plug the forward variances into (3.91) to obtain the corresponding samples of the VIX index. Computing the payoffs of the VIX option and averaging we obtain price estimates. We then invert the Black-Scholes equation to get the implied volatility. Columns 2 and 5 show values where  $g_{1,T}(\tau_i), g_{2,T}(\tau_i), i = 0, 1, \dots, n_v$ , are sampled exactly from their joint  $2(n_v + 1)$  dimensional normal distribution. Columns 3 and 6 show values under an implementation of the hybrid multifactor approach with direct sampling. More precisely: For a given  $n_v$  we set  $\Delta_n = \frac{\Delta_{\text{vix}}}{n_v}$ . We then use BM2005 to find for  $i = 1, 2$ , integers  $m_i \in \mathbb{N}$  and coefficients  $(c_{ij}, \gamma_{ij})_{j=1}^{m_i} \in \mathbb{R}^{2m_i}$  so  $t^\alpha \approx \sum_{j=1}^{m_1} c_{1j} e^{-\gamma_{1j}t}$  and  $t^\beta \approx \sum_{j=1}^{m_2} c_{2j} e^{-\gamma_{2j}t}$  for  $t \in [\Delta_n, T + \Delta_{\text{vix}}]$ . For this application of BM2005, we always sample 501 points on  $[\Delta_n, T + \Delta_{\text{vix}}]$  and keep  $\epsilon = 10^{-3}$ . Defining

$$U_{ij,T} := \int_0^T e^{-\gamma_{ij}(T-s)} dW_{i+1,s}, \quad i = 1, 2, \quad j = 1, \dots, m_i,$$

we sample  $g_{1,T}(0), g_{2,T}(0), (U_{1j})_{j=1}^{m_1}, (U_{2j})_{j=1}^{m_2}$ , from their dimension  $2 + m_1 + m_2$  joint normal distribution. Thereafter, we compute the forward values using (3.100)-(3.101). Column 4 shows prices where  $(g_{1,t}(0))_{t \in [0,T]}$  and  $(g_{2,t}(0))_{t \in [0,T]}$  are simulated with the hybrid multifactor scheme using  $\lfloor nT \rfloor = 50$  steps and the forward values afterwards are computed with Definition 3.3.1. Column 7 shows the running times under this approach except the time spent simulating  $(g_{1,t}(0))_{t \in [0,T]}$  and  $(g_{2,t}(0))_{t \in [0,T]}$  is excluded. The shown numbers are averages over 1000 runs each of which is based on 10 000 samples.

### 3.7.3. MIXED TWO-FACTOR ROUGH BERGOMI



**Figure 3.14:** VIX option smiles for the expiry  $T = 0.1$  as estimated with Monte Carlo under the two-factor mixed rough Bergomi model with parameters (3.131). The model is simulated with the hybrid multifactor scheme where  $\kappa = 1$ . The VIX index is computed with the trapezoidal rule (3.91). Estimates are based on 1 million paths. Standard errors are below 1% of the shown values. Left: Convergence in  $\lfloor nT \rfloor$  where  $n_v = 32$ . Right: Convergence in  $n_v$  where  $\lfloor nT \rfloor = 50$ .

**Table 3.5:** Implied volatilities and computational times (in milliseconds) for pricing the strike 25 and expiry 0.1 VIX option under the two-factor mixed rough Bergomi model with parameters (3.131) for different numbers of integration steps  $n_v$ . Numbers are averaged over 1000 runs each of which contains 10 000 samples. Standard errors are less than 0.1% of the reported volatilities and less than 1% of the reported running times. See the main text for more details.

	Implied volatility			Computational time		
	Exact sim.	Hybrid multi-factor (direct sim.)	Hybrid multi-factor	Exact sim.	Hybrid multi-factor (direct sim.)	Hybrid multi-factor (excl. sim.)
$n_v = 2$	1.09	1.09	1.09	70	55	9
$n_v = 4$	1.01	1.01	1.01	67	49	5
$n_v = 8$	0.97	0.97	0.97	73	53	7
$n_v = 16$	0.96	0.96	0.96	96	58	12
$n_v = 32$	0.95	0.95	0.95	120	70	20
$n_v = 256$	0.95	0.95	0.95	2278	247	131

From the table, we note that the methods converge equally fast in  $n_v$  and that  $n_v = 16$  or  $n_v = 32$  about suffices; see also Figure 3.14 (right plot) where we show the convergence in  $n_v$  of the full volatility smiles under the hybrid multifactor scheme. We refer to the left plot of Figure 3.14 for why  $\lfloor nT \rfloor = 50$  is a reasonable choice for running the scheme for this example.<sup>47</sup> To rank the methods in terms of performance it therefore suffices to compare their computational times. To this end, we note that exact sampling and direct sampling under the hybrid multifactor approach are about equally fast for recommended

<sup>47</sup> The plot suggests that  $\lfloor nT \rfloor = 25$  also works. However, for a practical implementation it is recommended to err on side of caution. Thus we have chosen  $\lfloor nT \rfloor = 50$ .

values of  $n_v$  (i.e. 16 or 32) with running times in the (very) rough neighbourhood of 100 milliseconds, though, the latter is slightly faster (by just shy of a factor two). From the table's last column, we also see that if the hybrid multifactor scheme has already been run for other purposes, the extra costs of evaluating the VIX index is much lower at around 20 milliseconds. It took 410 milliseconds (average over 1000 runs, standard error is 2 milliseconds) to simulate  $(g_{1,t}(0))_{t \in [0, T]}$  and  $(g_{2,t}(0))_{t \in [0, T]}$ . Therefore, to the extent that we can keep  $\lfloor nT \rfloor = 50$  for expiries in  $[0, T]$ , we can expect to need to price options on (very) roughly  $\frac{410}{100-20} \approx 5$  expiries in the interval before it is worthwhile to run the scheme from scratch. Although the calculation is very approximate and depends on the specific example, it shows that it is not entirely unrealistic that it could be worthwhile to run the scheme from start, also if not additionally for other purposes.

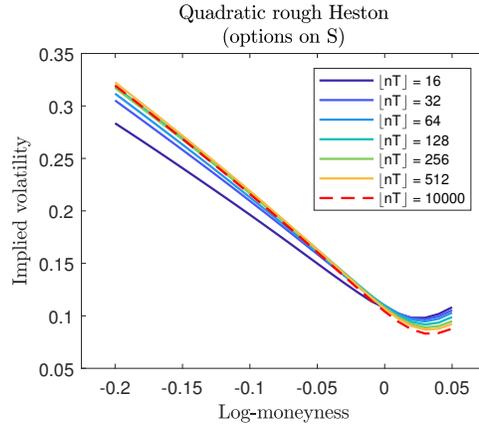
### 3.7.4 Quadratic and non-quadratic rough Heston

We consider the Monte Carlo pricing of options on  $S_T$  and  $VIX_T$  under the rough Heston models for the expiry  $T = 0.1$ . We simulate  $V$  with the hybrid multifactor scheme, directly or via  $Z$ ; for rough Heston we truncate values of  $V$  in zero. For computation of  $VIX_T$  we use (3.124)–(3.126) for rough Heston and (3.117) for quadratic rough Heston.

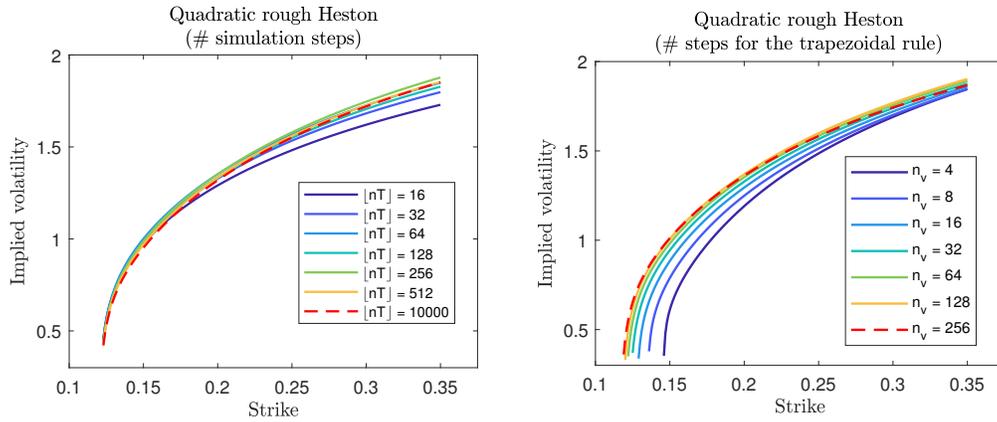
We start with the quadratic rough Heston model. In Figure 3.15, we show the convergence for options on  $S_T$  in the number of simulation steps  $\lfloor nT \rfloor$ . We note that convergence is reasonably fast in that a few hundred steps about suffices. In Figure 3.16, we show the convergence for options on  $VIX_T$  in values of  $\lfloor nT \rfloor$  (left) keeping  $n_v = 64$ , and  $n_v$  (right) keeping  $\lfloor nT \rfloor = 64$ . Convergence likewise appears reasonable and is roughly comparable to that observed for the two-factor mixed rough Bergomi model though slightly worse.

Lastly, we consider pricing under the rough Heston model. We refer to Figure 3.17. The plots in the top row show prices for options on  $S_T$  and include benchmark solutions (red lines) that are computed with Fourier pricing along the lines of [16]. We note that while convergence is reasonably fast in case A, it is impractically slow in case B, both for options on  $S_T$  and  $VIX_T$ . For example, in case B, using even 10 000 steps we obtain a bias of several percentage points of implied volatility when looking at options on  $S_T$ . It is as expected that convergence is slower when the solution is more erratic, i.e. when  $\alpha$  is low, and when the volatility-of-volatility parameter,  $\nu$ , is high. Because we use a locally Gaussian approximation and truncate  $V$  in zero, it too is meaningful that prices are biased in the positive direction. We can expect the positivity bias to vanish slower under (very) rough volatility since  $\text{Var}(\tilde{W}_{i-1,1}^n) \propto \Delta_n^{2\alpha+1}$  which means that the probability of negative values decays slower as  $n \rightarrow \infty$  when  $\alpha$  is low.

### 3.7.4. QUADRATIC AND NON-QUADRATIC ROUGH HESTON

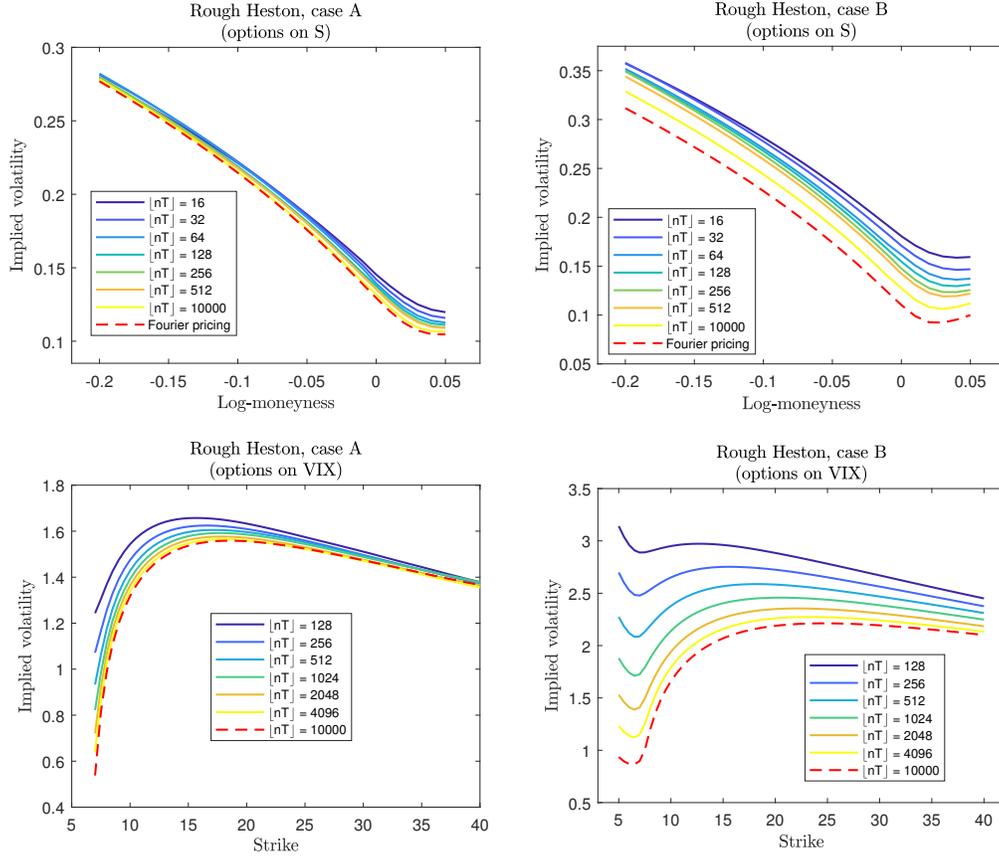


**Figure 3.15:** Volatility smiles for options on  $S_T$  estimated with Monte Carlo for the expiry  $T = 0.1$  under quadratic rough Heston with parameters (3.134). We use 200 000 paths. The model is simulated with the hybrid multifactor scheme using  $\kappa = 1$ . Standard errors are less than 2% of the shown values.



**Figure 3.16:** Volatility smiles for options on  $VIX_T$  using Monte Carlo for the expiry  $T = 0.1$  under quadratic rough Heston. We use 200 000 paths. The model is simulated with the hybrid multifactor scheme using  $\kappa = 1$ . Standard errors are less than 2%. Left: Convergence in  $[nT]$  where  $n_v = 64$ . Right: Convergence in  $n_v$  where  $[nT] = 64$ .

CHAPTER 3. HYBRID MULTIFACTOR SCHEME FOR STOCHASTIC  
VOLTERRA EQUATIONS WITH COMPLETELY MONOTONE KERNELS



**Figure 3.17:** Volatility smiles for options on  $S_T$  (top row) and  $VIX_T$  (bottom row) for the expiry  $T = 0.1$  under rough Heston with parameters (3.132)-(3.133). Values are estimated with Monte Carlo except for the red lines in the top row which are computed with Fourier pricing along the lines of [16]. For the Monte Carlo estimates simulation is performed using the hybrid multifactor scheme with  $\kappa = 1$ . Standard errors of the Monte Carlo estimates are within 1%.

### 3.8 Conclusion and future research

We have in this paper merged the ideas of [2, 4] with those of [9] to produce the hybrid multifactor scheme for the simulation of SVE's with completely monotone kernels. The scheme combines exact treatment of the kernel function near the origin with a sum-of-exponentials approximation for the remainder. In relation to the latter, we have specialised the method of [10] to completely monotone functions and demonstrated, by example on the rough fractional kernel, that it results in many fewer exponential terms than the techniques of [2, 4]. We have introduced a forward process akin to [3] and shown how it can be computed efficiently under the scheme. We have also shown how the forward values can be used for efficient computation of the VIX index for several instantaneous variance models: multifactor Volterra Bergomi, quadratic Volterra Heston, generalised CEV Volterra. Numerical experiments showed good convergence for equa-

### 3.8. CONCLUSION AND FUTURE RESEARCH

---

tions such as those of the rough Bergomi and quadratic rough Heston models. For rough Bergomi in particular, we have shown that our scheme is faster than the hybrid scheme of [9] and more accurate than a pure multifactor approximation. Experiments on rough Heston, however, resulted in a large positive bias in the case of a high volatility-of-volatility and a high level of roughness. We hypothesized that the problem arises as the combined effect of the volatility process being very volatile, approximating the coefficient processes piecewise constant, and having to truncate values in zero. Lastly, we have proved strong convergence under the assumption that the coefficient functions are Lipschitz continuous.

A number of ideas for future research come to mind. Firstly, to improve the slow convergence for rough Heston we could look at higher order methods. The paper [46] e.g. contains a Milstein scheme for SVE's, though, they do not approximate the kernel function as we have done in this paper and which we believe is more efficient. One could also try Richardson extrapolation; results on the weak convergence rate under rough volatility are limited but see [6]. The recent scheme of [21] for rough affine forward variance models also shows promising results for the rough Heston equation. Secondly, we believe our scheme could be adapted to other Volterra equations such as the fractional Riccati equations associated with rough Heston and other affine Volterra models, c.f. [5]. Thirdly, recall that our scheme, in many cases, outputs an approximate Markov state via the forward curve summarised in a small number of state variables. The variables are provided at no extra cost besides running the scheme. Consequently, we believe our scheme is well-suited for (e.g.) the least squares Monte Carlo method of [38]—there one is required to perform cross-path regressions on precisely the Markov state.

## References

- [1] Alfonsi, A., Kebaier, A., Approximation of stochastic Volterra equations with kernels of the completely monotone type (2021). Working paper. Available online at arXiv:2102.13505v1 (accessed January 27, 2022).
- [2] Abi Jaber, E., Lifting the Heston model. *Quantitative Finance*, 2019, **19**(12), 1995–2013.
- [3] Abi Jaber, E., El Euch, O., Markovian structure of the Volterra Heston model. *Statistics and Probability Letters*, 2019, **149**, 63–72.
- [4] Abi Jaber, E., El Euch, O., Multifactor approximation of rough volatility models. *SIAM Journal on Financial Mathematics*, 2019, **10**(2), 309–349.
- [5] Abi Jaber, E., Larsson, M., Pulido, S., Affine Volterra processes. *The Annals of Applied Probability*, 2019, **29**(5), 3155–3200.
- [6] Bayer, C., Hall, E. J., Tempone, R., Weak error rates for option pricing under linear rough volatility (2021). Working paper. Available online at arXiv:2009.01219v3 (accessed January 27, 2022).
- [7] Bayer, C., Friz, P., Gatheral, J., Pricing under rough volatility. *Quantitative Finance*, 2016, **16**(6), 887–904.
- [8] Bennedsen, M., Lunde, A., Pakkanen, M. S., Decoupling the short- and long-term behavior of stochastic volatility. *Journal of Financial Econometrics*, 2021, 1–46.
- [9] Bennedsen, M., Lunde, A., Pakkanen, M. S., Hybrid scheme for Brownian semistationary processes. *Finance and Stochastics*, 2017, **21**(4), 931–965.
- [10] Beylkin, G., Monzón, L., On approximation of functions by exponential sums. *Applied and Computational Harmonic Analysis*, 2005, **19**(1), 17–48.
- [11] Bergomi, L., Smile dynamics II. *Risk*, October 2005, pp. 67–73.
- [12] Bergomi, L., Smile dynamics III. *Risk*, October 2008, pp. 90–96.
- [13] Bonesini, O., Callegaro, G., Jacquier, A., Functional quantization of rough volatility and applications to the VIX (2021). Working paper. Available online at arXiv:2104.04233v1 (accessed January 27, 2022).

## REFERENCES

---

- [14] Carmona, P., Coutin, L., Fractional Brownian motion and the Markov property. *Electronic Communications in Probability*, 1998, **3**, 95–107.
- [15] De Marco, S., Volatility derivatives in (rough) forward variance models. *Stochastic Analysis and its Applications*, Oaxaca, May 2018.
- [16] El Euch, O., Gatheral, J., Rosenbaum, M., Roughening Heston. *Risk*, May 2019.
- [17] El Euch, O., Rosenbaum, M., Perfect hedging in rough Heston models. *The Annals of Applied Probability*, 2018, **28**(6), 3813–3856.
- [18] El Euch, O., Rosenbaum, M., The characteristic function of rough Heston models. *Mathematical Finance*, 2019, **29**(1), 3–38.
- [19] Ethier, S. N., Kurtz, T. G., *Markov Processes: Characterization and Convergence*, John Wiley & Sons, New York, 1986.
- [20] Fukasawa, M., Hirano, A., Refinement by reducing and reusing random numbers of the hybrid scheme for Brownian semistationary processes. *Quantitative Finance*, 2021, **21**(7), 1127–1146.
- [21] Gatheral, J., Efficient simulation of affine forward variance models (2021). Working paper. Available online at SSRN:3876680 (accessed January 27, 2022).
- [22] Gatheral, J., Jaisson, T., Rosenbaum, M., Volatility is rough. *Quantitative Finance*, 2018, **18**(6), 933–949.
- [23] Gatheral, J., *The Volatility Surface: A Practitioner’s Guide*, John Wiley & Sons, New Jersey, 2006.
- [24] Gatheral, J., Jusselin, P., Rosenbaum, M., The quadratic rough Heston model and the joint S&P 500/VIX smile calibration problem. *Risk*, May 2020.
- [25] Gatheral, J., Radoičić, R., Rational approximation of the rough Heston solution. *International Journal of Theoretical and Applied Finance*, 2019, **22**(3), 1950010.
- [26] Golub, G. H., van Loan, C. F., *Matrix Computations* (3rd ed.), The Johns Hopkins University Press, Baltimore and London, 1996.
- [27] Gripenberg, G., *Volterra Integral and Functional Equations*, Cambridge University Press, New York, 1990.
- [28] Guyon, J., On the joint calibration of SPX and VIX options. QuantMinds, Lisbon, 16 May, 2018.
- [29] Harms, P., Stefanovits, D., Affine representations of fractional processes with applications in mathematical finance. *Stochastic Processes and their Applications*, 2019, **129**(4), 1185–1228.
- [30] Harms, P., Strong convergence rates for Markovian representations of fractional processes. *Discrete and Continuous Dynamical Systems Series B*, 2021, **26**(10), 5567–5579.

- 
- [31] Horn, R. A., Johnson, C. R., *Matrix Analysis*, Cambridge University Press, New York, 1985.
- [32] Horvath, B., Jacquier, A., Muguruza, A., Functional central limit theorems for rough volatility (2019). Working paper. Available online at arXiv:1711.03078v3 (accessed January 27, 2022).
- [33] Horvath, B., Jacquier, A., Tankov, P., Volatility options in rough volatility models. *SIAM Journal on Financial Mathematics*, 2020, **11**(2), 437–469.
- [34] Jacquier, A., Martini, C., Muguruza, A., On VIX futures in the rough Bergomi model. *Quantitative Finance*, 2018, **18**(1), 45–61.
- [35] Kobayashi, K., Stochastic calculus for a time-changed semimartingale and the associated stochastic differential equations. *Journal of Theoretical Probability*, 2011, **24**, 789–820.
- [36] Kincaid, D., Cheney, W., *Numerical Analysis: Mathematics of Scientific Computing* (3rd ed.), Brooks/Cole, California, 2002.
- [37] Le Gall, J.-F., *Brownian Motion, Martingales, and Stochastic Calculus*, Berlin, Springer, 2016.
- [38] Longstaff, F. A., Schwartz, E. S., Valuing American options by simulation: A simple least-squares approach. *The Review of Financial Studies*, 2001, **14**(1), 113–147.
- [39] Lord, R., Koekkoek, R., van Dijk, D., A comparison of biased simulation schemes for stochastic volatility models. *Quantitative Finance*, 2010, **10**(2), 177–194.
- [40] Ma, J., Wu, H., A fast algorithm for simulation of rough volatility models. Published online. *Quantitative Finance*, 2021, <https://doi.org/10.1080/14697688.2021.1970213>.
- [41] Mallat, S., *A Wavelet Tour of Signal Processing* (3rd ed.), Elsevier, Amsterdam, 2009.
- [42] Mathai, A. M., Haubold, H. J., *Special Functions for Applied Scientists*, Springer, New York, 2008.
- [43] Miller, K. S., Samko, S. G., Completely monotonic functions. *Integral Transforms and Special Functions*, 2001, **12**(4), 389–402.
- [44] Platen, E., Bruti-Liberati, N., *Numerical Solution of Stochastic Differential Equations with Jumps in Finance*, Springer, Berlin-Heidelberg, 2010.
- [45] Kloeden, P. E., Platen, E., *Numerical Solution of Stochastic Differential Equations* (corrected 3rd printing), Springer, Berlin-Heidelberg, 1999.
- [46] Richard, A., Tan, X., Yang, F., Discrete-time simulation of stochastic Volterra equations. *Stochastic Processes and their Applications*, 2021, **141**, 109–138.
- [47] Saad, Y., *Numerical Methods for Large Eigenvalue Problems* (2nd ed.), Society for Industrial and Applied Mathematics, 2011.

- [48] Schilling, R., Measures, *Integrals and Martingales*, Cambridge University Press, New York, 2005.
- [49] Schmüdgen, K., *The Moment Problem*, Springer, Cham, 2017.
- [50] Veraar, M., The stochastic Fubini theorem revisited. *Stochastics*, 2012, **84**(4), 543–551.
- [51] Wendland, H., *Scattered Data Approximation*, Cambridge University Press, New York, 2005.
- [52] Zhang, X., Euler schemes and large deviations for stochastic Volterra equations with singular kernels. *Journal of Differential Equations*, 2008, **244**(9), 2226–2250.
- [53] Øksendal, B., *Stochastic Differential Equations*, Springer, Berlin Heidelberg, 2010.

# Appendix

## 3.A Convolutions and resolvents

We recite some theory of convolutions and resolvents.

Let  $F$  be a measurable function on  $\mathbb{R}_+$  and  $\nu$  a measure on  $\mathbb{R}_+$  of locally bounded variation. When valid we define  $F * \nu$  and  $\nu * F$  by  $(F * \nu)(t) = (\nu * F)(t) = \int_{[0,t]} F(t-s)\nu(ds)$  for  $t > 0$  and extend to  $t = 0$  by right continuity. For an additional measurable function  $G$  on  $\mathbb{R}_+$ , we write  $F * G = F * (Gdt)$  when well-defined. By [27, Theorem 3.6.1, Corollary 3.6.2] we have the following: Given  $F \in L_{\text{loc}}^p(\mathbb{R}_+)$  for  $p \geq 1$ , it holds that  $\nu * F$  is well-defined and lies in  $L_{\text{loc}}^p(\mathbb{R}_+)$ . Furthermore, given  $F, G \in L_{\text{loc}}^1(\mathbb{R}_+)$ , we have  $(F * G) * \nu = F * (G * \nu)$  almost everywhere.

Let  $K \in L_{\text{loc}}^1(\mathbb{R}_+)$ . The resolvent of *first kind* of  $K$  is defined as the measure  $L$  on  $\mathbb{R}_+$  of locally bounded variation that satisfies  $K * L = L * K = 1$ . The resolvent of first kind need not exist but is unique if it does [27, Theorem 5.5.2]. Its existence is guaranteed if  $K$  is non-negative and non-increasing on  $(0, \infty)$  and is not identically zero [27, Theorem 5.5.5]. The resolvent of *second kind* is defined as the element  $R \in L_{\text{loc}}^1(\mathbb{R}_+)$  that solves  $K * R = R * K = K - R$ . It always exists and is unique [27, Theorem 2.3.1].

*Remark 3.A.1.* Let  $K \in L_{\text{loc}}^p(\mathbb{R}_+)$  for some  $p \geq 1$  and  $R \in L_{\text{loc}}^1(\mathbb{R}_+)$  be its resolvent of second kind. By [27, Corollary 2.2.3 part (i)] we have  $K * R \in L_{\text{loc}}^p(\mathbb{R}_+)$ . Consequently  $R = K - K * R \in L_{\text{loc}}^p(\mathbb{R}_+)$ .

Examples of resolvents can be found in Table 1 of [5], though, they are given without explicit proof. As we use it in the main text, we provide below an explicit proof for the resolvent of second kind of the gamma kernel:

**Proposition 3.A.1.** *Let  $K(t) = ce^{-\lambda t} \frac{t^{\alpha-1}}{\Gamma(\alpha)}$ ,  $t > 0$ ,  $c \in \mathbb{R}$ ,  $\lambda \geq 0$ ,  $\alpha \in (0, 1]$ . For  $1 \leq \beta < \frac{1}{1-\alpha}$  with the interpretation  $1/0 = \infty$  we have  $K, R \in L_{loc}^\beta(\mathbb{R}_+)$  where  $R$  is the resolvent of second kind of  $K$ . Also*

$$R(t) = ce^{-\lambda t} t^{\alpha-1} E_{\alpha, \alpha}(-ct^\alpha), \quad t > 0, \quad (3.140)$$

where  $E_{\alpha, \alpha}(\cdot)$  is defined by (3.51).

*Proof.* Let  $1 \leq \beta < \frac{1}{1-\alpha}$ . Then  $K \in L_{loc}^\beta(\mathbb{R}_+)$  since  $\int_0^t s^{(\alpha-1)\beta} ds = \frac{1}{(\alpha-1)\beta+1} t^{(\alpha-1)\beta+1} < \infty$  when  $(\alpha-1)\beta > -1$  (i.e. when  $\beta < \frac{1}{1-\alpha}$ ) and  $t \mapsto |e^{-\lambda t}|$  is bounded on closed subintervals of  $\mathbb{R}_+$ . By Remark 3.A.1 also  $R \in L_{loc}^\beta(\mathbb{R}_+)$ . It remains to check that  $K * R = R * K = K - R$  with  $R$  given by (3.140). Let  $t > 0$ . The result follows by the below where we change variables and use [42, Equation (2.2.14), Theorem 2.2.1 part (i)]:

$$\begin{aligned} (R * K)(t) &= (K * R)(t) \\ &= c^2 e^{-\lambda t} \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} s^{\alpha-1} E_{\alpha, \alpha}(-cs^\alpha) ds \\ &= c^2 e^{-\lambda t} \frac{t^{2\alpha-1}}{\Gamma(\alpha)} \int_0^1 (1-z)^{\alpha-1} z^{\alpha-1} E_{\alpha, \alpha}(-ct^\alpha z^\alpha) dz \\ &= c^2 e^{-\lambda t} t^{2\alpha-1} E_{\alpha, 2\alpha}(-ct^\alpha) \\ &= c^2 e^{-\lambda t} t^{2\alpha-1} (-ct^\alpha)^{-1} (E_{\alpha, \alpha}(-ct^\alpha) - \Gamma(\alpha)^{-1}) \\ &= ce^{-\lambda t} \frac{t^{\alpha-1}}{\Gamma(\alpha)} - ce^{-\lambda t} t^{\alpha-1} E_{\alpha, \alpha}(-ct^\alpha) \\ &= K(t) - R(t). \end{aligned}$$

□

Let  $Z = \int_0^\cdot b_s ds + \int_0^\cdot a_s dW_s$  be a continuous semimartingale with  $a, b$  locally bounded and progressively measurable (e.g. continuous and adapted). For  $K \in L_{loc}^2(\mathbb{R}_+)$  we will write  $(K * dZ)_t = \int_0^t K(t-s) dZ_s$ ,  $t \geq 0$ , which is well-defined. Using stochastic Fubini [5, Lemma 2.1] we have also  $(\nu * (K * dZ))_t = ((\nu * K) * dZ)_t$  almost surely for every  $t \geq 0$ . Lastly, following the remarks of [4, Section A.1] there exists a continuous version of  $(K * dZ)_{t \geq 0}$ . Throughout the paper we work only with such continuous versions.

### 3.B Existence results for stochastic Volterra equations

We recite some existence results of [3] for the SVE (3.1). Their results generalise those of [5] to the case of a non-constant  $g_0$ . We start with two hypotheses for  $K$  (which we have also introduced in the main text):

**Hypothesis** ( $H_0$ ):  $K \in L^2_{\text{loc}}(\mathbb{R}_+)$  and there exists a  $\tilde{\gamma} \in (0, 2]$  so  $\int_0^h K(t)^2 dt = O(h^{\tilde{\gamma}})$  and  $\int_0^T (K(t+h) - K(t))^2 dt = O(h^{\tilde{\gamma}})$  for every  $T \in [0, \infty)$ .

**Hypothesis** ( $H_1$ ):  $K$  is non-negative, not identically zero, non-increasing and continuous on  $(0, \infty)$ , and its resolvent of first kind  $L$  is non-negative and non-increasing in the sense that  $s \mapsto L([s, s+t])$  is non-increasing for all  $t \geq 0$ .

We then have the following:

**Theorem 3.B.1.** (*one-dimensional version of [3, Theorem A.1]*) Assume that ( $H_0$ ) holds and that  $g_0$  is locally Hölder continuous of all orders less than  $\tilde{\gamma}/2$ . Then:

(i) If  $b$  and  $\sigma$  are Lipschitz continuous (3.1) admits a unique continuous strong solution  $X$ .

(ii) If  $b$  and  $\sigma$  are continuous with linear growth and  $K$  admits a resolvent of first kind  $L$  then (3.1) admits a continuous weak solution  $X$ .

In both cases,  $X$  is locally Hölder continuous of all orders strictly less than  $\tilde{\gamma}/2$  and

$$\sup_{t \leq T} E[|X_t|^p] < \infty, \quad p > 0, \quad T > 0. \quad (3.141)$$

**Theorem 3.B.2.** (*Theorem A.2 of [3]*) Assume that  $K$  satisfies ( $H_0$ ) – ( $H_1$ ) and that  $b$  and  $\sigma$  are continuous of linear growth so  $b(0) \geq 0$  and  $\sigma(0) = 0$ . Assume also that  $g_0$  is locally Hölder continuous of all orders less than  $\tilde{\gamma}/2$  with  $g_0(0) \geq 0$  and

$$\Delta_h g_0 - (\Delta_h K * L)(0)g_0 - d(\Delta_h K * L) * g_0 \geq 0, \quad h \geq 0.$$

Then (3.1) admits a non-negative continuous weak solution.

### 3.C Proof of Theorem 3.3.2

*Proof.* We reiterate that the proof mostly is only a slight extension of that from [3]. For the sake of completeness we nevertheless go through all the details.

Let  $t_0, t, \tau \geq 0$ . Then

$$\begin{aligned} g_t^{t_0}(\tau) &= g_0(t_0 + t + \tau) + \int_0^{t_0+t} K(t_0 + t + \tau - s)(b(X_s)ds + \sigma(X_s)dW_s) \\ &= g_0(t_0 + t + \tau) + \int_0^{t_0} K(t_0 + t + \tau - s)(b(X_s)ds + \sigma(X_s)dW_s) \\ &\quad + \int_{t_0}^{t_0+t} K(t_0 + t + \tau - s)(b(X_s)ds + \sigma(X_s)dW_s) \\ &= g_{t_0}(t + \tau) + \int_0^t K(t + \tau - s)(b(X_s^{t_0})ds + \sigma(X_s^{t_0})dW_s^{t_0}). \end{aligned}$$

In the last line, we have used the change-of-variables Lemma 2.3 of [35]; use the time change  $T_s = t_0 + s$ ,  $s \geq 0$ , in their notation. It is straightforward to show that  $(\mathcal{F}_t)_{t \geq 0}$  satisfies the usual hypothesis if  $(\mathcal{F}_t^{t_0})_{t \geq 0}$  does. It likewise follows with ease that  $(W_t^{t_0})_{t \geq 0}$  is Brownian motion and that  $(g_t^{t_0})_{t \geq 0}$  is adapted with respect to the shifted filtration.

Equation (3.43) follows by setting  $\tau = 0$  in the above.

We argue that  $g_{t_0} \in \mathcal{H}^{\tilde{\gamma}/2}$ : Fix  $p \geq 2$  and  $\epsilon \in (0, \frac{\tilde{\gamma}}{2})$ . Let also  $T > 0$  and  $t, h \geq 0$  so  $t, t+h \leq T$ . In what follows  $C_1, C_2, C_3$ , denote positive constants that may change from line to line, though, they can and will be chosen independent of  $t, h$ . Note that

$$\begin{aligned} g_{t_0}(t+h) - g_{t_0}(t) &= g_0(t_0+t+h) - g_0(t_0+t) \\ &\quad + \int_0^{t_0} (K(t_0+t+h-s) - K(t_0+t-s))(b(X_s)ds + \sigma(X_s)dW_s). \end{aligned}$$

Since  $g_0 \in \mathcal{H}^{\tilde{\gamma}/2}$ :

$$|g_0(t_0+t+h) - g_0(t_0+t)|^p \leq C_1 h^{(\frac{\tilde{\gamma}}{2}-\epsilon)p}.$$

By Jensen's inequality (repeated use), Minkowski's integral inequality [48, Theorem 13.14], linear growth, the moment bound (3.141),  $(H_0)$ :

$$\begin{aligned} &E \left[ \left| \int_0^{t_0} (K(t_0+t+h-s) - K(t_0+t-s))b(X_s)ds \right|^p \right] \\ &\leq C_2 E \left[ \left| \int_0^{t_0} (K(t_0+t+h-s) - K(t_0+t-s))^2 b(X_s)^2 ds \right|^{\frac{p}{2}} \right] \\ &\leq C_2 \left[ \int_0^{t_0} (E[|K(t_0+t+h-s) - K(t_0+t-s)|^p |b(X_s)|^p])^{\frac{2}{p}} ds \right]^{\frac{p}{2}} \\ &\leq C_2 \left[ \int_0^{t_0} |K(t_0+t+h-s) - K(t_0+t-s)|^2 (1 + E(|X_s|^p))^{\frac{2}{p}} ds \right]^{\frac{p}{2}} \\ &\leq C_2 \left( \int_0^{t_0} (K(t_0+t+h-s) - K(t_0+t-s))^2 ds \right)^{\frac{p}{2}} \\ &= C_2 \left( \int_t^{t+t_0} (K(s+h) - K(s))^2 ds \right)^{\frac{p}{2}} \\ &\leq C_2 \left( \int_0^{T+t_0} (K(s+h) - K(s))^2 ds \right)^{\frac{p}{2}} \\ &\leq C_2 h^{\frac{\tilde{\gamma}}{2}p}. \end{aligned}$$

### 3.C. PROOF OF THEOREM 3.3.2

---

Similarly, though using also the BDG inequality:

$$\begin{aligned} & E \left[ \left| \int_0^{t_0} (K(t_0 + t + h - s) - K(t_0 + t - s)) \sigma(X_s) dW_s \right|^p \right] \\ & \leq C_3 E \left[ \left| \int_0^{t_0} (K(t_0 + t + h - s) - K(t_0 + t - s))^2 \sigma(X_s)^2 ds \right|^{\frac{p}{2}} \right] \\ & \leq C_3 h^{\frac{\tilde{\gamma}}{2} p}. \end{aligned}$$

By the above and Jensen's inequality, it follows that

$$E (|g_{t_0}(t+h) - g_{t_0}(t)|^p) \leq Ch^{(\frac{\tilde{\gamma}}{2} - \epsilon)p},$$

where  $C$  is a positive constant that is independent of  $t, h$ .

By Kolmogorov's lemma [37, Theorem 2.9] applied to  $(g_{t_0}(s))_{s \leq T}$  there exists a modification that is Hölder continuous on  $[0, T]$  of all orders less than  $\tilde{\gamma}/2 - \epsilon - p^{-1}$ . Because  $\epsilon$  and  $p^{-1}$  can be chosen arbitrarily small and  $T > 0$  was arbitrary too, we conclude that  $g_{t_0}$  is locally Hölder continuous of all orders less than  $\tilde{\gamma}/2$  when viewed with domain  $\mathbb{R}_+$ .

We show that  $g_0 \in \mathcal{H}_+^{\tilde{\gamma}/2}$  under the extra assumptions: We will use [35, Lemma 2.3] without explicit mentioning. Let  $t_0, t, h \geq 0$ . Define  $Z = \int_0^\cdot (b(X_s) ds + \sigma(X_s) dW_s)$ . By Lemma B.2 and Remark B.3 of [3]:

$$\Delta_h K = (\Delta_h K * L)(0)K + d(\Delta_h K * L) * K. \quad (3.142)$$

Therefore:

$$(\Delta_h K * dZ) = (\Delta_h K * L)(0) \underbrace{K * dZ}_{=X-g_0} + d(\Delta_h K * L) * \underbrace{K * dZ}_{=X-g_0}.$$

Note then

$$X_{t+h}^{t_0} = g_0(t_0 + t + h) + (\Delta_h K * dZ)_{t_0+t} + \int_0^h K(h-s) dZ_{t_0+t+s} \quad (3.143)$$

$$= g_0(t_0 + t + h) + (\Delta_h K * L)(0)(X_t^{t_0} - g_0(t_0 + t)) \quad (3.144)$$

$$+ (d(\Delta_h K * L) * (X - g_0))_{t_0+t} + \int_0^h K(h-s) dZ_{t_0+t+s} \quad (3.145)$$

$$= g_0(t_0 + t + h) - (\Delta_h K * L)(0)g_0(t_0 + t) - (d(\Delta_h K * L) * g_0)_{t_0+t} \quad (3.146)$$

$$+ (\Delta_h K * L)(0)X_t^{t_0} + (d(\Delta_h K * L) * X)_{t_0+t} + \int_0^h K(h-s) dZ_{t_0+t+s} \quad (3.147)$$

$$\geq (\Delta_h K * L)(0)X_t^{t_0} + (d(\Delta_h K * L) * X)_{t_0+t} \quad (3.148)$$

$$+ \int_0^h K(h-s) b(X_{t+s}^{t_0}) ds + \int_0^h K(h-s) \sigma(X_{t+s}^{t_0}) dW_{t+s}^{t_0}, \quad (3.149)$$

where the inequality follows by (3.42).

Let  $\tau \geq 0$ . For the next part, see that

$$\begin{aligned} E_{t_0}(X_\tau^{t_0}) &= E_{t_0} \left( g_0(t_0 + \tau) + \int_0^{t_0 + \tau} K(t_0 + \tau - s)(b(X_s)ds + \sigma(X_s)dW_s) \right) \\ &= g_0(t_0 + \tau) + \int_0^{t_0} K(t_0 + \tau - s)(b(X_s)ds + \sigma(X_s)dW_s) \\ &\quad + E_{t_0} \left( \int_{t_0}^{t_0 + \tau} K(t_0 + \tau - s)b(X_s)ds \right), \end{aligned}$$

so

$$g_{t_0}(\tau) = E_{t_0} \left[ X_\tau^{t_0} - \int_0^\tau K(\tau - s)b(X_s^{t_0})ds \right], \quad \tau \geq 0. \quad (3.150)$$

We have here used that (3.46) is a true martingale, also under the present assumptions.

Define  $G_h^{t_0} = \Delta_h g_{t_0} - (\Delta_h K * L)(0)g_{t_0} - d(\Delta_h K * L) * g_{t_0}$ . We must show it non-negative. By (3.143)-(3.150), and that

$$(\Delta_h K * b(X^{t_0}))_t = (\Delta_h K * L)(0)(K * b(X^{t_0}))_t + (d(\Delta_h K * L) * K * b(X^{t_0}))_t,$$

as follows from (3.142), we obtain

$$\begin{aligned} G_h^{t_0}(t) &= E_{t_0} \left[ X_{t+h}^{t_0} - (K * b(X^{t_0}))_{t+h} - (\Delta_h K * L)(0)(X_t^{t_0} - (K * b(X^{t_0}))_t) \right. \\ &\quad \left. - (d(\Delta_h K * L) * (X^{t_0} - K * b(X^{t_0})))_t \right] \\ &\geq E_{t_0} \left[ (d(\Delta_h K * L) * X)_{t_0+t} + \int_0^h K(h-s)b(X_{t+s}^{t_0})ds \right. \\ &\quad \left. + \int_0^h K(h-s)\sigma(X_{t+s}^{t_0})dW_{t+s}^{t_0} + (\Delta_h K * L)(0)(K * b(X^{t_0}))_t \right. \\ &\quad \left. - (K * b(X^{t_0}))_{t+h} - (d(\Delta_h K * L) * (X^{t_0} - K * b(X^{t_0})))_t \right] \\ &= E_{t_0} \left[ \int_t^{t_0+t} X_{t_0+t-s}(d(\Delta_h K * L))(ds) + \int_0^h K(h-s)b(X_{t+s}^{t_0})ds \right. \\ &\quad \left. - \int_t^{t+h} K(t+h-s)b(X_s^{t_0})ds \right] \\ &= E_{t_0} \left[ \int_t^{t_0+t} X_{t_0+t-s}(d(\Delta_h K * L))(ds) \right] \geq 0. \end{aligned}$$

Positivity holds because  $X$  and  $d(\Delta_h K * L)$  are non-negative. See [3, Remark B.3] for the latter. Note also that  $g_{t_0}(0) = X_{t_0} \geq 0$ . This completes the proof.  $\square$

### 3.D Proof of Theorem 3.3.4

*Proof.* That  $g_0 = X_0 + \lambda K * \theta \in \mathcal{H}^{\tilde{\gamma}/2}$  can be proved as in part (ii) of [3, Example 2.2]. The existence results (i) and (ii) of the theorem then follow from Theorem 3.B.1.

### 3.D. PROOF OF THEOREM 3.3.4

---

Say that (i) or (ii) of Theorem 3.3.4 additionally holds and let  $X$  be continuous and solve (3.47). To ease what follows, we change notation to that of [5, Lemma 2.5]. Define  $F := X_0 + \lambda K * \theta$ ,  $B := -\lambda$ ,  $Z := \int_0^\cdot \sigma(X_s) dW_s$ . We then have

$$X = F + (KB) * X + K * dZ,$$

and using the lemma

$$X = F - R_B * F + E_B * dZ,$$

where  $R_B$  is the resolvent of second kind of  $-KB$  and  $E_B = K - R_B * K$ . By definition

$$-KB * R_B = -KB - R_B \implies \frac{1}{B} R_B = K * R_B - K = -E_B.$$

Abusing notation by writing also  $R_\lambda$  for the resolvent of second kind of  $\lambda K$  ( $= -KB$ ):

$$\begin{aligned} X &= X_0 + \lambda K * \theta - R_\lambda * (X_0 + \lambda K * \theta) + \frac{1}{\lambda} R_\lambda * (\sigma(X) dW) \\ &= X_0 + \lambda K * \theta - R_\lambda * X_0 - \underbrace{R_\lambda * (\lambda K)}_{= \lambda K - R_\lambda} * \theta + \frac{1}{\lambda} R_\lambda * (\sigma(X) dW) \\ &= X_0 + R_\lambda * (\theta - X_0) + \frac{1}{\lambda} R_\lambda * (\sigma(X) dW). \end{aligned}$$

Since  $\lambda K \in L_{\text{loc}}^2(\mathbb{R}_+)$  we have  $R_\lambda \in L_{\text{loc}}^2(\mathbb{R}_+)$  by Remark 3.A.1. This combined with

$$R_\lambda = \lambda(K - R_\lambda * K),$$

by [5, Example 2.3 (iii)-(v)] implies that  $R_\lambda$  satisfies  $(H_0)$  with the same  $\tilde{\gamma}$  as  $K$ . That  $\tilde{g}_0 = X_0 + R_\lambda * (\theta - X_0) \in \mathcal{H}^{\tilde{\gamma}/2}$  follows by the proof strategy of [3, Example 2.2 part (ii)].

We consider the second half of the theorem. Assume first additionally that  $K$  is CM and not identically zero. By [27, Theorem 5.5.4] this implies that  $K$  has a resolvent of first kind  $L$  and that it satisfies  $(H_1)$ . Say furthermore that  $\lambda\theta(t)dt + X_0L(dt)$  is a non-negative measure and  $X_0 \geq 0$ . By [3, Example 2.2 part(ii)] then  $g_0 \in \mathcal{H}_+^{\tilde{\gamma}/2}$ . Assuming lastly (in addition) that  $\sigma(0) = 0$ , we conclude by Theorem 3.B.2 that (3.47) has a continuous non-negative weak solution  $X$ .

That  $R_\lambda$  is CM follows by [27, Theorem 5.3.1] using that  $K$  is CM and local integrability. Note also that  $R_\lambda$  is not identically zero because if it were then by the relation  $(\lambda K) * R_\lambda = \lambda K - R_\lambda$  we must have  $K = 0$  which contradicts the assumption that  $K$  is not identically zero. Using [27, Theorem 5.5.4] we conclude that  $R_\lambda$  has a resolvent of first kind for which it satisfies  $(H_1)$ . Let now  $\tilde{L}$  be the measure on  $\mathbb{R}_+$  of locally bounded variation defined by  $\tilde{L}(dt) = \lambda^{-1}L(dt) + dt$ . Then by definition of  $R_\lambda$  and  $L$ :

$$R_\lambda * \tilde{L} = (\lambda K - R_\lambda * (\lambda K)) * (\lambda^{-1}L) + R_\lambda * 1 = 1 - R_\lambda * 1 + R_\lambda * 1 = 1.$$

It follows that  $\tilde{L}$  is the resolvent of first kind of  $R_\lambda$ . That  $(\theta(t) - X_0)dt + X_0\tilde{L}(dt)$  is a non-negative measure follows by  $\lambda\theta(t)dt + X_0L(dt)$  being non-negative. From this and [3, Example 2.2 part (ii)], we conclude lastly that  $\tilde{g}_0 = X_0 + R_\lambda * (\theta - X_0) \in \mathcal{H}_+^{\tilde{\gamma}/2}$ .  $\square$

### 3.E Proof of Theorem 3.3.5

*Proof.* The dynamics of  $(U_t(\gamma))_{t \geq 0}$  follow by Ito's lemma as applied for (3.12)-(3.13). The processes are well-defined by continuity of  $b, \sigma, X$ , and adaptedness of the latter.

By stochastic Fubini [50] it for  $t, \tau \geq 0$  holds that

$$\begin{aligned} g_t(\tau) &= g_0(t + \tau) + \int_0^t K(t + \tau - s)b(X_s)ds + \int_0^t K(t + \tau - s)\sigma(X_s)dW_s \\ &= g_0(t + \tau) + \int_0^t \int_{[0, \infty)} e^{-\gamma(t+\tau-s)} \mu(d\gamma) (b(X_s)ds + \sigma(X_s)dW_s) \\ &= g_0(t + \tau) + \int_{[0, \infty)} e^{-\gamma\tau} \left( \int_0^t e^{-\gamma(t-s)} (b(X_s)ds + \sigma(X_s)dW_s) \right) \mu(d\gamma) \\ &= g_0(t + \tau) + \int_{[0, \infty)} e^{-\gamma\tau} U_t(\gamma) \mu(d\gamma). \end{aligned}$$

The stochastic Fubini theorem applies because  $\mu$  is  $\sigma$ -finite and for any  $t, \tau \geq 0$ :

$$\int_{[0, \infty)} \left( \int_0^t e^{-2\gamma(t+\tau-s)} \sigma(X_s)^2 ds \right)^{\frac{1}{2}} \mu(d\gamma) < \infty, \quad (3.151)$$

$$\int_{[0, \infty)} \left( \int_0^t e^{-\gamma(t+\tau-s)} |b(X_s)| ds \right) \mu(d\gamma) < \infty. \quad (3.152)$$

We prove these claims in the following.

For  $\sigma$ -finiteness we argue as follows: Define  $A_n = [0, n]$  for  $n \in \mathbb{N}$  and note that  $\mathbb{R}_+ = \cup_{n \geq 1} A_n$ . Take any  $t > 0$  and  $n \geq 1$ . Then

$$e^{-nt} \mu(A_n) = \int_{A_n} e^{-nt} \mu(d\gamma) \leq \int_{A_n} e^{-\gamma t} \mu(d\gamma) \leq \int_{[0, \infty)} e^{-\gamma t} \mu(d\gamma) = K(t) < \infty.$$

Consequently  $\mu(A_n) < \infty$  and the claim follows.

We turn our attention to (3.151). Let  $t, \tau \geq 0$ . In what follows  $C$  denotes a positive constant that may change from line to line. By linear growth of  $\sigma$  and continuity of  $X$ :

$$\begin{aligned} & \int_{[0, \infty)} \left( \int_0^t e^{-2\gamma(t+\tau-s)} \sigma(X_s)^2 ds \right)^{\frac{1}{2}} \mu(d\gamma) \\ & \leq C(1 + \sup_{s \leq t} X_s^2)^{\frac{1}{2}} \int_{[0, \infty)} \left\{ \left[ \frac{1}{2\gamma} e^{-2\gamma(t+\tau-s)} \right]_0^t \right\}^{\frac{1}{2}} \mu(d\gamma) \\ & \leq C \int_{[0, \infty)} \gamma^{-\frac{1}{2}} e^{-\gamma\tau} (1 - e^{-2\gamma t})^{\frac{1}{2}} \mu(d\gamma) \\ & \leq C \int_{[0, \infty)} \underbrace{\gamma^{-\frac{1}{2}} (1 - e^{-2\gamma t})^{\frac{1}{2}}}_{=: F(\gamma)} \mu(d\gamma). \end{aligned}$$

An application of L'Hôpital's rule gives  $F(0) := \lim_{\gamma \rightarrow 0^+} F(\gamma) = \sqrt{2t} < \infty$ . It follows that  $F$  is well-defined as a continuous function on  $[0, 1]$  and thus is uniformly bounded on it. On  $[1, \infty)$  it is bounded by 1. Since also  $F(\gamma) \leq \gamma^{-\frac{1}{2}}$  for all  $\gamma > 0$ , we deduce

$$F(\gamma) \leq \left( \sup_{\lambda \leq 1} F(\lambda) \vee 1 \right) \wedge \gamma^{-\frac{1}{2}}, \quad \gamma \geq 0.$$

We conclude

$$\int_{[0, \infty)} F(\gamma) \mu(d\gamma) \leq C \int_{[0, \infty)} (1 \wedge \gamma^{-\frac{1}{2}}) \mu(d\gamma) < \infty,$$

which shows the claim. A similar argument gives (3.152); it should here be noted that  $\gamma^{-1} \leq \gamma^{-\frac{1}{2}}$  for  $\gamma \geq 1$  which allows us to apply the integrability bound of  $(H_2^*)$  again.  $\square$

### 3.F Properties of $K_{mn}$

Proposition 3.F.2 further down states some useful properties of  $K_{mn}$ . We need the lemma below for its proof.

**Lemma 3.F.1.** *Let there be  $f, g : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  and  $\alpha, \beta, C_f, C_g, \bar{h}_f, \bar{h}_g > 0$  so*

$$\begin{aligned} f(h) &\leq C_f h^\alpha & \forall h \in [0, \bar{h}_f], \\ g(h) &\leq C_g h^\beta & \forall h \in [0, \bar{h}_g]. \end{aligned}$$

Then

$$f(h) + g(h) \leq C_{f+g} h^{\alpha \wedge \beta} \quad \forall h \in [0, \bar{h}_{f+g}],$$

where  $C_{f+g} = 2(C_f \vee C_g)$  and  $\bar{h}_{f+g} = \bar{h}_f \wedge \bar{h}_g \wedge 1$ .

*Proof.* The result follows by

$$f(h) + g(h) \leq C_f h^\alpha + C_g h^\beta \leq (C_f \vee C_g)(h^\alpha + h^\beta) \leq 2(C_f \vee C_g) h^{\alpha \wedge \beta}, \quad h \in [0, \bar{h}_{f+g}],$$

since  $\bar{h}_{f+g} \leq \bar{h}_f, \bar{h}_g$  and  $h^\alpha + h^\beta \leq 2h^{\alpha \wedge \beta}$  for  $h \leq \bar{h}_{f+g} \leq 1$ .  $\square$

**Proposition 3.F.2.** *Say that  $K$  is CM and satisfies  $(H_0)$  for some  $\tilde{\gamma} \in (0, 2]$ . Then  $K_{mn} \in L_{loc}^2(\mathbb{R}_+)$  for any  $m, n \in \mathbb{N}$ . Moreover, for any  $m \in \mathbb{N}$  there exist constants  $C_m, h_m > 0$ , so*

$$\int_0^h K_{mn}(t)^2 dt \leq C_m h^{\tilde{\gamma} \wedge 1}, \quad \forall n \in \mathbb{N}, \quad h \in [0, h_m]. \quad (3.153)$$

For also a given  $T \in [0, \infty)$  there are constants  $C_{m,T}, h_{m,T} > 0$ , so

$$\int_0^T (K_{mn}(t+h) - K_{mn}(t))^2 dt \leq C_{m,T} h^{\tilde{\gamma} \wedge 1}, \quad \forall n \in \mathbb{N}, \quad h \in [0, h_{m,T}]. \quad (3.154)$$

*Proof.* Because  $K_m$  is continuous on  $[0, \infty)$ , it follows that not only  $K$  but also  $K_m$  is locally square integrable. Then clearly  $K_{mn} \in L^2_{\text{loc}}(\mathbb{R}_+)$  where measurability follows since  $K$  and  $K_m$  are both measurable. As  $K_m$  is locally Lipschitz it satisfies  $(H_0)$  with rate 1; c.f. [5, Example 2.3 part (i)]. There are then constants  $\tilde{C}_m, \tilde{h}_m > 0$  that are independent of  $n$  so  $\int_0^h K_m(t)^2 dt \leq \tilde{C}_m h$  for all  $h \in [0, \tilde{h}_m]$ . Since  $K$  satisfies  $(H_0)$  with rate  $\tilde{\gamma}$  there also exist constants  $\tilde{C}, \tilde{h} > 0$  that are independent of  $n$  so  $\int_0^h K(t)^2 dt \leq \tilde{C} h^{\tilde{\gamma}}$  for all  $h \in [0, \tilde{h}]$ . Define  $C_m = 2(\tilde{C}_m \vee \tilde{C})$  and  $h_m = \tilde{h}_m \wedge \tilde{h} \wedge 1$ . Then by  $K_{mn}$  being a concatenation of  $K$  and  $K_m$ , Lemma 3.F.1, that  $C_m$  and  $h_m$  are independent of  $n$  too:

$$\int_0^h K_{mn}(t)^2 dt \leq \|K\|_{L^2([0,h])}^2 + \|K_m\|_{L^2([0,h])}^2 \leq C_m h^{\tilde{\gamma} \wedge 1} \quad \forall n \in \mathbb{N}, \quad h \in [0, h_m].$$

This completes part one of the proposition.

We prove the second part. Let  $T \in [\kappa, \infty)$  as suffices. Note that  $T \geq \kappa \Delta_n$  for any  $n$ . We will use this implicitly later. Because  $K$  satisfies  $(H_0)$  with rate  $\tilde{\gamma}$  there exist constants  $\tilde{C}_T, \tilde{h}_T > 0$  independent of  $n$  so  $\int_0^T (K(t+h) - K(t))^2 dt \leq \tilde{C}_T h^{\tilde{\gamma}}$  for all  $h \in [0, \tilde{h}_T]$ . As  $K_m$  satisfies  $(H_0)$  with rate 1 there analogously exist constants  $\tilde{C}_{m,T}, \tilde{h}_{m,T} > 0$  independent of  $n$  so  $\int_0^T (K_m(t+h) - K_m(t))^2 dt \leq \tilde{C}_{m,T} h$  for all  $h \in [0, \tilde{h}_{m,T}]$ .

When  $\kappa = 0$ , we obtain  $\int_0^T (K_{mn}(t+h) - K_{mn}(t))^2 dt = \int_0^T (K_m(t+h) - K_m(t))^2 dt \leq \tilde{C}_{m,T} h$  for all  $h \in [0, \tilde{h}_{m,T}]$  so (3.154) is satisfied with  $C_{m,T} = \tilde{C}_{m,T}$  and  $h_{m,T} = \tilde{h}_{m,T} \wedge 1$ . Say then  $\kappa > 0$ . This case requires more care. We will here assume that  $\tilde{h} = \tilde{h}_T = \tilde{h}_{m,T} \geq \kappa$  which is possible—also so  $\tilde{h}, \tilde{h}_T, \tilde{h}_{m,T}, \tilde{C}, \tilde{C}_T, \tilde{C}_{m,T}$ , yet are independent of  $n$  as we will assume.<sup>48</sup> Note that it now especially holds that  $\tilde{h} = \tilde{h}_T = \tilde{h}_{m,T} \geq \kappa \Delta_n$  for any  $n$  which will be used in the following. Let  $h \in [\kappa \Delta_n, \tilde{h}]$  and see that

$$\begin{aligned} \int_0^T (K_{mn}(t+h) - K_{mn}(t))^2 dt &= \int_{\kappa \Delta_n}^T (K_m(t+h) - K_m(t))^2 dt \\ &\quad + \int_0^{\kappa \Delta_n} (K_m(t+h) - K(t))^2 dt \\ &=: I_1(m, n, h) + I_2(m, n, h). \end{aligned}$$

Note that  $I_1(m, n, h) \leq \int_0^T (K_m(t+h) - K_m(t))^2 dt \leq \tilde{C}_{m,T} h$ . From the triangle and Jensen's inequalities, that  $\kappa \Delta_n \leq h \leq \tilde{h}_T = \tilde{h}$ , the definition

$$C_{m,T}^* := \sup_{t \in [0, 2\tilde{h}_T]} K_m(t)^2 < \infty,$$

<sup>48</sup> Say  $f : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  is bounded on closed intervals and for  $\alpha, C_f, \tilde{h}_f > 0$  satisfies  $f(h) \leq C_f h^\alpha$  for all  $h \in [0, \tilde{h}_f]$ . Take an arbitrary  $h_f^* > 0$ . Set  $C_f^* = \sup_{h \in (0, h_f^*]} f(h)/(h^\alpha) < \infty$ ; finiteness holds by the boundedness assumption on  $f$  and since  $f(h) \leq C_f h^\alpha$  for  $h$  near zero. Then  $f(h) \leq C_f^* h^\alpha$  for all  $h \in [0, h_f^*]$ . For the claim of the main text use this on  $\int_0^h K(t)^2 dt, \int_0^T (K(t+h) - K(t))^2 dt, \int_0^T (K_m(t+h) - K_m(t))^2 dt$ , viewed as mappings in  $h$ ; boundedness on closed intervals holds because  $K, K_m \in L^2_{\text{loc}}(\mathbb{R}_+)$ .

### 3.G. PROOF OF THEOREM 3.4.3

---

which is independent of  $n$ , we obtain

$$I_2(m, n, h) \leq 2\|\Delta_h K_m\|_{L^2([0, h])}^2 + 2\|K\|_{L^2([0, h])}^2 \leq 2C_{m, T}^* h + 2\tilde{C}h^{\tilde{\gamma}}.$$

It follows that

$$\int_0^T (K_{mn}(t+h) - K_{mn}(t))^2 dt \leq (\tilde{C}_{m, T} + 2C_{m, T}^*)h + 2\tilde{C}h^{\tilde{\gamma}}, \quad h \in [\kappa\Delta_n, \tilde{h}]. \quad (3.155)$$

Consider now the case  $h \in [0, \kappa\Delta_n)$ . Write first

$$\begin{aligned} \int_0^T (K_{mn}(t+h) - K_{mn}(t))^2 dt &= \int_{\kappa\Delta_n}^T (K_m(t+h) - K_m(t))^2 dt \\ &\quad + \int_{\kappa\Delta_n - h}^{\kappa\Delta_n} (K_m(t+h) - K(t))^2 dt \\ &\quad + \int_0^{\kappa\Delta_n - h} (K(t+h) - K(t))^2 dt \\ &=: J_1(m, n, h) + J_2(m, n, h) + J_3(n, h). \end{aligned}$$

Note that  $J_1(m, n, h) \leq \|\Delta_h K_m - K_m\|_{L^2([0, T])}^2 \leq \tilde{C}_{m, T} h$ . In similar spirit  $J_3(n, h) \leq \|\Delta_h K - K\|_{L^2([0, T])}^2 \leq \tilde{C}_T h^{\tilde{\gamma}}$ .

For the center term, we obtain, by use of the triangle and Jensen's inequalities, that  $C_{m, T}^*$  uniformly bounds  $K_m^2$  on  $[0, 2\tilde{h}_T]$ , that  $K$  is non-increasing as follows by complete monotonicity:

$$\begin{aligned} J_2(m, n, h) &\leq 2 \left( \|\Delta_h K_m\|_{L^2([\kappa\Delta_n - h, \kappa\Delta_n])}^2 + \|K\|_{L^2([\kappa\Delta_n - h, \kappa\Delta_n])}^2 \right) \\ &\leq 2 \left( C_{m, T}^* h + \|K\|_{L^2([0, h])}^2 \right) \\ &\leq 2 (C_{m, T}^* h + \tilde{C}h^{\tilde{\gamma}}). \end{aligned}$$

For  $h \in [0, \kappa\Delta_n]$  then:

$$\int_0^T (K_{mn}(t+h) - K_{mn}(t))^2 dt \leq (\tilde{C}_{m, T} + 2C_{m, T}^*)h + (\tilde{C}_T + 2\tilde{C})h^{\tilde{\gamma}}. \quad (3.156)$$

Combining (3.155) and (3.156), we conclude

$$\int_0^T (K_{mn}(t+h) - K_{mn}(t))^2 dt \leq (\tilde{C}_{m, T} + 2C_{m, T}^*)h + (\tilde{C}_T + 2\tilde{C})h^{\tilde{\gamma}}, \quad h \in [0, \tilde{h}].$$

An application of Lemma 3.F.1 yields the desired conclusion.  $\square$

## 3.G Proof of Theorem 3.4.3

We prove Theorem 3.4.3. As noted in Section 3.4, it holds under Assumption 3.4.1 that the equations for  $X$  and  $X^{mn}$  have unique continuous solutions on the same probability

space. We will use this without further comment in the derivations. Note also that

$$U_t^{mn} = \int_0^t (b_s^{mn} - \gamma^m \cdot U_s^{mn}) ds + \int_0^t \sigma_s^{mn} dW_s, \quad 0 \leq t \leq T, \quad (3.157)$$

and, under Equation (3.33) of Definition 3.2.1, recalling here (3.36), that

$$\hat{U}_t^{mn} = \int_0^{t_n^-} (\hat{b}_s^{mn} - \gamma^m \cdot \hat{U}_s^{mn}) ds + \int_0^{t_n^-} \hat{\sigma}_s^{mn} dW_s, \quad 0 \leq t \leq T. \quad (3.158)$$

In the above, we interpret  $b_s^{mn}, \hat{b}_s^{mn}, \sigma_s^{mn}, \hat{\sigma}_s^{mn}$ , as repeated row-by-row  $m$  times. We use the same interpretation in the proofs when necessary. The Jensen's, Hölder's, and Minkowski's integral inequalities will be used repeatedly. A good reference for these are [48]. We will additionally use the BDG<sup>49</sup> and Gronwall's inequalities. We refer to [37, Theorem 5.16] and [19, Theorem 5.1 (appendix)]. Note moreover that for  $\beta^* := \frac{\beta}{\beta-1}$  we have  $\beta^{-1} + (\beta^*)^{-1} = 1$ ; this will be used whenever we apply Hölder's inequality. Throughout we keep  $g_0, K, b, \sigma, T, T^*, \kappa, \{(c^m, \gamma^m)\}_{m=1}^\infty, \beta, \tilde{\gamma}$ , fixed. Our results, any bounding constants especially, therefore implicitly depend on them. Also, we will in multiple places for given  $p \geq \frac{2\beta}{\beta-1}$  and  $m \in \mathbb{N}$  work with positive constants  $C_p$  and  $C_{p,m}$ . To simplify, we will allow these to change from line to line. A careful reading of our proofs reveal that the  $C_p$ 's can be chosen to depend on  $p$  only whereas the  $C_{p,m}$ 's will depend on both  $p$  and  $m$ , that is, besides the fixed setup listed previously. Reading the proof, the reader may come to realise that the rate estimates for many of the drift terms can be improved if viewed in isolation. This however has no influence on the final convergence rates as we do not worsen them more than the comparable diffusion terms that appear. We have done it so we can argue by analogy between the drift and diffusion terms. This has shortened the proofs slightly. Our derivations are partly inspired by those of [46, 52].

We start with a result that bounds the moments of  $\hat{X}^{mn}$  and  $\hat{U}^{mn}$ .

**Lemma 3.G.1.** *Invoke Assumption 3.4.1 and assume that Equation (3.33) of Definition 3.2.1 is used. Let  $p \geq \frac{2\beta}{\beta-1}$  and  $m \in \mathbb{N}$ . There then exists a positive constant  $C_{p,m}$  so for all  $n \in \mathbb{N}$  and  $t \in [0, T]$ :  $E \left[ |\hat{X}_t^{mn}|^p \right] \vee E \left[ \|\hat{U}_t^{mn}\|^p \right] \leq C_{p,m}$ .*

*Proof.* Let  $p \geq \frac{2\beta}{\beta-1}$ ,  $m, n \in \mathbb{N}$ ,  $t \in [0, T]$ . Note that  $p \geq 2$  and  $\frac{p}{2\beta^*} \geq 1$ . We will use this when we apply Jensen's inequality in a number of places. To validly use Gronwall's inequality at the end of the proof, we use a localisation argument. Thus for  $N \in \mathbb{N}$  define  $\tau_N := \inf\{s \in [0, T] : |\hat{X}_s^{mn}| \vee \|\hat{U}_s^{mn}\| \geq N\}$ ,  $\hat{X}_{N,t}^{mn} := \hat{X}_t^{mn} \mathbf{1}_{(t < \tau_N)}$ ,  $\hat{U}_{N,i,t}^{mn} := \hat{U}_{i,t}^{mn} \mathbf{1}_{(t < \tau_N)}$ ,  $i = 1, \dots, m$ ,  $\hat{U}_{N,t}^{mn} := (\hat{U}_{N,1,t}^{mn}, \dots, \hat{U}_{N,m,t}^{mn})^\top$ .

<sup>49</sup> Note that while  $\{\int_0^t \tilde{K}(t+\tau-s)Y_s dW_s; t \geq 0\}$  for  $\tau \in \mathbb{R}_+$ ,  $\tilde{K} \in L_{\text{loc}}^2(\mathbb{R}_+)$ ,  $(Y_t)_{t \geq 0}$  progressive and locally bounded, is not generally a local martingale, the process  $\{\int_0^t \tilde{K}(u-s)Y_s dW_s; t \in [0, u]\}$  for a fixed  $u \geq 0$  is. This is what allows us to apply the BDG inequality in our context.

### 3.G. PROOF OF THEOREM 3.4.3

As we will argue:

$$|\hat{X}_{N,t}^{mn}|^p \leq \left| g_0(t_n^-) + \sum_{i=1}^m c_i^m e^{-\gamma_i^m \kappa \Delta_n} \hat{U}_{N,i,(t_n^- - \kappa \Delta_n)^+}^{mn} \right|^p \quad (3.159)$$

$$+ \left| \int_{(t_n^- - \kappa \Delta_n)^+}^{t_n^-} K(t_n^- - s) b(\hat{X}_{N,s}^{mn}) ds \right|^p \quad (3.160)$$

$$+ \left| \int_{(t_n^- - \kappa \Delta_n)^+}^{t_n^-} K(t_n^- - s) \sigma(\hat{X}_{N,s}^{mn}) dW_s \right|^p. \quad (3.161)$$

Indeed, when  $t < \tau_N$  both sides equal  $|\hat{X}_t^{mn}|^p = |\hat{X}_{t_n^-}^{mn}|^p$  (consult (3.24)-(3.25) for the right-hand side), and when  $t \geq \tau_N$  the left-hand side is zero, the right-hand-side, of course, non-negative. Note that

$$\max_{i=1,\dots,m} |c_i^m e^{-\gamma_i^m \kappa \Delta_n}| \leq \underbrace{\max_{i=1,\dots,m} \left\{ |c_i^m| (1 \vee e^{-\gamma_i^m \kappa}) \right\}}_{\text{independent of } n} < \infty. \quad (3.162)$$

Let  $i \in \{1, \dots, m\}$ . The above then holds because if  $\gamma_i^m \geq 0$  we have  $e^{-\gamma_i^m \kappa \Delta_n} \leq 1$  and if  $\gamma_i^m \leq 0$  then  $e^{-\gamma_i^m \kappa \Delta_n} \leq e^{-\gamma_i^m \kappa}$  since  $\Delta_n \leq 1$  regardless of  $n \in \mathbb{N}$ . By the triangle inequality, (3.162), the definition of the  $l^1$ -norm, equivalence of norms on  $\mathbb{R}^m$ :

$$\left| \sum_{i=1}^m c_i^m e^{-\gamma_i^m \kappa \Delta_n} \hat{U}_{N,i,(t_n^- - \kappa \Delta_n)^+}^{mn} \right|^p \leq C_{p,m} \left( \sum_{i=1}^m |\hat{U}_{N,i,(t_n^- - \kappa \Delta_n)^+}^{mn}| \right)^p \quad (3.163)$$

$$\leq C_{p,m} \|\hat{U}_{N,(t_n^- - \kappa \Delta_n)^+}^{mn}\|^p. \quad (3.164)$$

Applying Jensen's inequality to (3.159)-(3.161) and using (3.163)-(3.164), we deduce:<sup>50</sup>

$$E \left[ |\hat{X}_{N,t}^{mn}|^p \right] \leq C_{p,m} \left( |g_0(t_n^-)|^p + E \left[ \|\hat{U}_{N,(t_n^- - \kappa \Delta_n)^+}^{mn}\|^p \right] \right) \quad (3.165)$$

$$+ E \left[ \left| \int_{(t_n^- - \kappa \Delta_n)^+}^{t_n^-} K(t_n^- - s) b(\hat{X}_{N,s}^{mn}) ds \right|^p \right] \quad (3.166)$$

$$+ E \left[ \left| \int_{(t_n^- - \kappa \Delta_n)^+}^{t_n^-} K(t_n^- - s) \sigma(\hat{X}_{N,s}^{mn}) dW_s \right|^p \right]. \quad (3.167)$$

By Jensen's inequality, Hölder's inequality, linear growth,  $\|K\|_{L^{2\beta}([0, T^*])} < \infty$ :

$$\begin{aligned} & E \left[ \left| \int_{(t_n^- - \kappa \Delta_n)^+}^{t_n^-} K(t_n^- - s) b(\hat{X}_{N,s}^{mn}) ds \right|^p \right] \\ & \leq C_p E \left[ \left| \int_{(t_n^- - \kappa \Delta_n)^+}^{t_n^-} K(t_n^- - s)^2 b(\hat{X}_{N,s}^{mn})^2 ds \right|^{\frac{p}{2}} \right] \end{aligned}$$

<sup>50</sup> Note also that  $\hat{U}_{N,(t_n^- - \kappa \Delta_n)^+}^{mn} = \hat{U}_{N,(t_n - \kappa \Delta_n)^+}^{mn}$ .

$$\begin{aligned}
&\leq C_p \left( \int_{(t_n^- - \kappa \Delta_n)^+}^{t_n^-} |K(t_n^- - s)|^{2\beta} ds \right)^{\frac{p}{2\beta}} E \left( \left[ \int_{(t_n^- - \kappa \Delta_n)^+}^{t_n^-} (1 + |\hat{X}_{N,s}^{mn}|)^{2\beta^*} ds \right]^{\frac{p}{2\beta^*}} \right) \\
&\leq C_p \left( \int_0^{T^*} |K(s)|^{2\beta} ds \right)^{\frac{p}{2\beta}} E \left( \int_{(t_n^- - \kappa \Delta_n)^+}^{t_n^-} (1 + |\hat{X}_{N,s}^{mn}|)^p ds \right) \\
&\leq C_p \left( 1 + \int_0^t E [|\hat{X}_{N,s}^{mn}|^p] ds \right).
\end{aligned}$$

Likewise but using also the BDG inequality:

$$\begin{aligned}
E \left[ \left| \int_{(t_n^- - \kappa \Delta_n)^+}^{t_n^-} K(t_n^- - s) \sigma(\hat{X}_{N,s}^{mn}) dW_s \right|^p \right] &\leq C_p E \left( \left[ \int_{(t_n^- - \kappa \Delta_n)^+}^{t_n^-} K(t_n^- - s)^2 \sigma(\hat{X}_{N,s}^{mn})^2 ds \right]^{\frac{p}{2}} \right) \\
&\leq C_p \left( 1 + \int_0^t E [|\hat{X}_{N,s}^{mn}|^p] ds \right).
\end{aligned}$$

By (local Hölder) continuity  $g_0$  is uniformly bounded on  $[0, T]$ . Using this and the bounds that we derived for the terms of (3.166)-(3.167):

$$E [|\hat{X}_{N,t}^{mn}|^p] \leq C_{p,m} \left( 1 + E [|\hat{U}_{N,(t-\kappa \Delta_n)^+}^{mn}|^p] + \int_0^t E [|\hat{X}_{N,s}^{mn}|^p] ds \right). \quad (3.168)$$

Analogous to (3.159)-(3.161) using also the triangle and Jensen's inequalities:

$$\begin{aligned}
\|\hat{U}_{N,t}^{mn}\|^p &\leq \left\| \int_0^{t_n^-} b(\hat{X}_{N,s}^{mn}) ds - \int_0^{t_n^-} \gamma^m \cdot \hat{U}_{N,s}^{mn} ds + \int_0^{t_n^-} \sigma(\hat{X}_{N,s}^{mn}) dW_s \right\|^p \\
&\leq C_p \left[ \left\| \int_0^{t_n^-} b(\hat{X}_{N,s}^{mn}) ds \right\|^p + \left\| \int_0^{t_n^-} \gamma^m \cdot \hat{U}_{N,s}^{mn} ds \right\|^p + \left\| \int_0^{t_n^-} \sigma(\hat{X}_{N,s}^{mn}) dW_s \right\|^p \right].
\end{aligned} \quad (3.170)$$

By the definition of the  $l^1$ -norm, Jensen's inequality, equivalence of norms on  $\mathbb{R}^m$ :

$$E \left[ \left\| \int_0^{t_n^-} \gamma^m \cdot \hat{U}_{N,s}^{mn} ds \right\|^p \right] \leq C_{p,m} E \left[ \left\| \sum_{i=1}^m \int_0^{t_n^-} |\hat{U}_{N,i,s}^{mn}| ds \right\|^p \right] \leq C_{p,m} \int_0^{t_n^-} E [|\hat{U}_{N,s}^{mn}|^p] ds.$$

Bounding the expectation of the remaining two terms of (3.170) as we did to arrive at (3.168) (except there is no need for Hölder's inequality in this case), we obtain in total

$$E [|\hat{U}_{N,t}^{mn}|^p] \leq C_{p,m} \left( 1 + \int_0^t E [|\hat{U}_{N,s}^{mn}|^p] ds + \int_0^t E [|\hat{X}_{N,s}^{mn}|^p] ds \right). \quad (3.171)$$

Combining (3.168) and (3.171):

$$E [|\hat{X}_{N,t}^{mn}|^p] \leq C_{p,m} \left( 1 + \int_0^t E [|\hat{U}_{N,s}^{mn}|^p] ds + \int_0^t E [|\hat{X}_{N,s}^{mn}|^p] ds \right). \quad (3.172)$$

### 3.G. PROOF OF THEOREM 3.4.3

---

Define

$$h_N(t) := \sup_{s \leq t} \left( E \left[ |\hat{X}_{N,s}^{mn}|^p \right] \vee E \left[ \|\hat{U}_{N,s}^{mn}\|^p \right] \right).$$

By (3.171) and (3.172):

$$h_N(t) \leq C_{p,m} \left( 1 + \int_0^t h_N(s) ds \right).$$

From Gronwall's inequality<sup>51</sup> we then deduce  $h_N(t) \leq C_{p,m}$ . A careful reading of the derivations till this point shows that  $C_{p,m}$  can be chosen independent of  $n$ ,  $t$ , and  $N$ . Using this, Fatou's lemma, and some other tricks, we have<sup>52,53,54</sup>

$$\begin{aligned} C_{p,m} &\geq \liminf_{N \rightarrow \infty} h_N(t) \\ &= \liminf_{N \rightarrow \infty} \sup_{s \leq t} \left( E \left[ |\hat{X}_{N,s}^{mn}|^p \right] \vee E \left[ \|\hat{U}_{N,s}^{mn}\|^p \right] \right) \\ &\geq \sup_{s \leq t} \left( \liminf_{N \rightarrow \infty} \left\{ E \left[ |\hat{X}_{N,s}^{mn}|^p \right] \vee E \left[ \|\hat{U}_{N,s}^{mn}\|^p \right] \right\} \right) \\ &\geq \sup_{s \leq t} \left( \left\{ \liminf_{N \rightarrow \infty} E \left[ |\hat{X}_{N,s}^{mn}|^p \right] \right\} \vee \left\{ \liminf_{N \rightarrow \infty} E \left[ \|\hat{U}_{N,s}^{mn}\|^p \right] \right\} \right) \\ &\geq \sup_{s \leq t} \left( \left\{ E \left[ \liminf_{N \rightarrow \infty} |\hat{X}_{N,s}^{mn}|^p \right] \right\} \vee \left\{ E \left[ \liminf_{N \rightarrow \infty} \|\hat{U}_{N,s}^{mn}\|^p \right] \right\} \right). \end{aligned}$$

By footnote 53, that  $\tau_N = \infty$  (a.s.) for  $N$  large enough, we get using the maximum-norm:

$$\begin{aligned} C_{p,m} &\geq \sup_{s \leq t} \left( \left\{ E \left[ \liminf_{N \rightarrow \infty} |\hat{X}_{N,s}^{mn}|^p \right] \right\} \vee \left\{ E \left[ \|\liminf_{N \rightarrow \infty} \hat{U}_{N,s}^{mn}\|_\infty^p \right] \right\} \right) \\ &\geq \sup_{s \leq t} \left( E \left[ |\hat{X}_s^{mn}|^p \right] \vee E \left[ \|\hat{U}_s^{mn}\|_\infty^p \right] \right). \end{aligned}$$

Then by equivalence of norms

$$\sup_{s \leq t} \left( E \left[ |\hat{X}_s^{mn}|^p \right] \vee E \left[ \|\hat{U}_s^{mn}\|^p \right] \right) \leq C_{p,m},$$

where  $\|\cdot\|$  is an arbitrary norm on  $\mathbb{R}^m$ . Since  $C_{p,m}$  can be chosen independent of  $n$  and  $t$  (and  $N$ ), the proof is completed.  $\square$

In what follows, we bound the error between  $g$  and  $g^{mn}$ .

---

<sup>51</sup> The inequality applies because  $h_N$  by localisation is bounded on bounded intervals.

<sup>52</sup> For the third line note this: Let  $f_n : \mathbb{R} \rightarrow \mathbb{R}$ ,  $n \in \mathbb{N}$ ,  $A \subset \mathbb{R}$ . For  $n \in \mathbb{N}$  clearly  $\sup_{y \in A} f_n(y) \geq f_n(x)$  for any  $x \in A$ . Consequently,  $\liminf_{n \rightarrow \infty} \sup_{y \in A} f_n(y) \geq \liminf_{n \rightarrow \infty} f_n(x)$  for any  $x \in A$ . We conclude that  $\liminf_{n \rightarrow \infty} \sup_{x \in A} f_n(x) \geq \sup_{x \in A} \liminf_{n \rightarrow \infty} f_n(x)$  which is what use for the inequality.

<sup>53</sup> For the fourth line note this: For  $d \in \mathbb{N}$ , let  $x^n = (x_1^n, \dots, x_d^n) \in \mathbb{R}^d$  be a sequence in  $n \in \mathbb{N}$ . As a special case of footnote 52, then  $\liminf_{n \rightarrow \infty} \max_{i=1, \dots, d} |x_i^n| \geq \max_{i=1, \dots, d} \left( \liminf_{n \rightarrow \infty} |x_i^n| \right)$ .

<sup>54</sup> In the last line, we use Fatou's lemma [48, Theorem 9.11].

**Lemma 3.G.2.** *Invoke Assumption 3.4.1 and assume that Equation (3.33) of Definition 3.2.1 is used. For every  $p \geq \frac{2\beta}{\beta-1}$  there then exists a positive constant  $C_p$  so for all  $m, n \in \mathbb{N}$ :*

$$\sup_{(t,\tau) \in A} E [|g_t(\tau) - g_t^{mn}(\tau)|^p] \leq C_p \|K - K_{mn}\|_{L^{2\beta}([0, T^*])}^p.$$

*Proof.* Let  $p \geq \frac{2\beta}{\beta-1}$ ,  $m, n \in \mathbb{N}$ ,  $(t, \tau) \in A$ . Decompose first

$$\begin{aligned} & g_t(\tau) - g_t^{mn}(\tau) \\ &= \int_0^t [K(t+\tau-s) - K_{mn}(t+\tau-s)] b_s ds + \int_0^t K_{mn}(t+\tau-s) [b_s - b_s^{mn}] ds \\ & \quad + \int_0^t [K(t+\tau-s) - K_{mn}(t+\tau-s)] \sigma_s dW_s + \int_0^t K_{mn}(t+\tau-s) [\sigma_s - \sigma_s^{mn}] dW_s \\ &=: I_1^b + I_2^b + I_1^\sigma + I_2^\sigma. \end{aligned}$$

By Jensen's and Hölder's inequality, linear growth, the moment bound (3.141):

$$\begin{aligned} E [|I_1^b|^p] &= E \left[ \left| \int_0^t [K(t+\tau-s) - K_{mn}(t+\tau-s)] b_s ds \right|^p \right] \\ &\leq C_p \left( \int_0^t |K(t+\tau-s) - K_{mn}(t+\tau-s)|^{2\beta} ds \right)^{\frac{p}{2\beta}} E \left[ \left( \int_0^t |b_s|^{2\beta^*} ds \right)^{\frac{p}{2\beta^*}} \right] \\ &\leq C_p \left( \int_0^{T^*} |K(s) - K_{mn}(s)|^{2\beta} ds \right)^{\frac{p}{2\beta}} \left( \int_0^t (1 + E [|X_s|^p]) ds \right) \\ &\leq C_p \|K - K_{mn}\|_{L^{2\beta}([0, T^*])}^p. \end{aligned}$$

Additionally by Jensen's and Hölder's inequality, Lipschitz continuity of  $b$ , the fact that  $\|K_{mn}\|_{L^{2\beta}([0, T^*])}$  is bounded by a constant that is independent of  $m$  and  $n$ :<sup>55</sup>

$$\begin{aligned} E [|I_2^b|^p] &= E \left[ \left| \int_0^t K_{mn}(t+\tau-s) [b_s - b_s^{mn}] ds \right|^p \right] \\ &\leq C_p \left( \int_0^t |K_{mn}(t+\tau-s)|^{2\beta} ds \right)^{\frac{p}{2\beta}} E \left[ \left( \int_0^t |b_s - b_s^{mn}|^{2\beta^*} ds \right)^{\frac{p}{2\beta^*}} \right] \\ &\leq C_p \int_0^t E [|X_s - X_s^{mn}|^p] ds. \end{aligned}$$

<sup>55</sup> By the triangle inequality and that  $K_{mn}$  is a concatenation of  $K$  and  $K_m$ :  $\|K_{mn}\|_{L^{2\beta}([0, T^*])} \leq \|K\|_{L^{2\beta}([0, T^*])} + \|K_m\|_{L^{2\beta}([0, T^*])}$ . From the triangle inequality again:  $\|K_m\|_{L^{2\beta}([0, T^*])} \leq \|K_m - K\|_{L^{2\beta}([0, T^*])} + \|K\|_{L^{2\beta}([0, T^*])}$ . Then  $\|K_{mn}\|_{L^{2\beta}([0, T^*])} \leq 2\|K\|_{L^{2\beta}([0, T^*])} + \|K - K_m\|_{L^{2\beta}([0, T^*])}$ . The claim now follows because  $\|K\|_{L^{2\beta}([0, T^*])} < \infty$  and  $\lim_{m \rightarrow \infty} \|K - K_m\|_{L^{2\beta}([0, T^*])} = 0$ .

### 3.G. PROOF OF THEOREM 3.4.3

By analogous argumentation, though, using also the BDG inequality, we obtain

$$\begin{aligned} E[|I_1^\sigma|^p] &\leq C_p E\left(\left[\int_0^t [K(t+\tau-s) - K_{mn}(t+\tau-s)]^2 \sigma_s^2 ds\right]^{\frac{p}{2}}\right) \\ &\leq C_p \|K - K_{mn}\|_{L^{2\beta}([0, T^*])}^p, \end{aligned}$$

and

$$E[|I_2^\sigma|^p] \leq C_p E\left(\left[\int_0^t K_{mn}(t+\tau-s)^2 [\sigma_s - \sigma_s^{mn}]^2 ds\right]^{\frac{p}{2}}\right) \leq C_p \int_0^t E[|X_s - X_s^{mn}|^p] ds.$$

Combining all estimates, we have for arbitrary  $(t, \tau) \in A$ :

$$E[|g_t(\tau) - g_t^{mn}(\tau)|^p] \leq C_p \left( \|K - K_{mn}\|_{L^{2\beta}([0, T^*])}^p + \int_0^t E[|X_s - X_s^{mn}|^p] ds \right). \quad (3.173)$$

Define now

$$h(t) := \sup_{\substack{(s, \tau) \in A \\ s \leq t}} E[|g_s(\tau) - g_s^{mn}(\tau)|^p], \quad 0 \leq t \leq T. \quad (3.174)$$

Because the right-hand-side of (3.173) is non-decreasing in  $t$ , we have

$$h(t) \leq C_p \left( \|K - K_{mn}\|_{L^{2\beta}([0, T^*])}^p + \int_0^t E[|X_s - X_s^{mn}|^p] ds \right), \quad (3.175)$$

and therefore

$$h(t) \leq C_p \left( \|K - K_{mn}\|_{L^{2\beta}([0, T^*])}^p + \int_0^t h(s) ds \right), \quad 0 \leq t \leq T.$$

We wish to apply Gronwall's inequality to the above. To do so validly, we first need to argue that  $h$  is bounded on bounded intervals. Because it is non-decreasing, it suffices to argue that  $h(t) < \infty$  for any  $t$ . This follows from (3.175) using (1) the fact that  $\|K - K_{mn}\|_{L^{2\beta}([0, T^*])} \leq 2\|K\|_{L^{2\beta}([0, T^*])} + \|K_{mn}\|_{L^{2\beta}([0, T^*])} < \infty$  which follows by the triangle inequality and the first inequality of footnote 55, and (2) that  $\sup_{s \leq t} E[|X_s - X_s^{mn}|^p] \leq 2^{p-1} \sup_{s \leq t} (E[|X_s|^p] + E[|X_s^{mn}|^p]) < \infty$  for which we have used Jensen's inequality and the moment bound (3.141) for both  $X$  and  $X^{mn}$ .

Then by Gronwall's inequality we finally obtain

$$\sup_{(t, \tau) \in A} E[|g_t(\tau) - g_t^{mn}(\tau)|^p] \leq C_p \|K - K_{mn}\|_{L^{2\beta}([0, T^*])}^p,$$

which completes the proof.  $\square$

Next, we bound the error between  $\hat{X}^{mn}$  and  $X^{mn}$ , respectively,  $\hat{U}^{mn}$  and  $U^{mn}$ .

**Lemma 3.G.3.** *Invoke Assumption 3.4.1 and assume that Equation (3.33) of Definition 3.2.1 is used. For any  $p \geq \frac{2\beta}{\beta-1}$  and  $m \in \mathbb{N}$  there then exists a constant  $C_{p,m}$  so*

$$\sup_{t \in [0, T]} \left( E[|X_t^{mn} - \hat{X}_t^{mn}|^p] \vee E[|U_t^{mn} - \hat{U}_t^{mn}|^p] \right) \leq C_{p,m} n^{-\frac{p}{2}(\tilde{\gamma} \wedge 1)}, \quad n \in \mathbb{N}.$$

*Proof.* Let  $p \geq \frac{2\beta}{\beta-1}$ ,  $m, n \in \mathbb{N}$ ,  $t \in [0, T]$ . Define

$$Z_{U,t}^{mn} := U_t^{mn} - \hat{U}_t^{mn}, \quad Z_{X,t}^{mn} := X_t^{mn} - \hat{X}_t^{mn}. \quad (3.176)$$

By (3.157)-(3.158), we may write

$$Z_{U,t}^{mn} = \int_0^{t_n^-} [b_s^{mn} - \hat{b}_s^{mn}] ds + \int_{t_n^-}^t b_s^{mn} ds + \int_0^{t_n^-} -\gamma^m \cdot [U_s^{mn} - \hat{U}_s^{mn}] ds \quad (3.177)$$

$$+ \int_{t_n^-}^t -\gamma^m \cdot U_s^{mn} ds + \int_0^{t_n^-} [\sigma_s^{mn} - \hat{\sigma}_s^{mn}] dW_s + \int_{t_n^-}^t \sigma_s^{mn} dW_s \quad (3.178)$$

$$=: I_{U,1}^b + I_{U,2}^b + I_{U,1}^\gamma + I_{U,2}^\gamma + I_{U,1}^\sigma + I_{U,2}^\sigma. \quad (3.179)$$

By Jensen's inequality and Lipschitz continuity:

$$E [ \|I_{U,1}^b\|^p ] \leq C_{p,m} \int_0^{t_n^-} E [ |b_s^{mn} - \hat{b}_s^{mn}|^p ] ds \leq C_{p,m} \int_0^t E [ |Z_{X,s}^{mn}|^p ] ds.$$

Furthermore, using Jensen's inequality, Minkowski's integral inequality, linear growth, that  $\sup_{s \leq T} E [ |X_s^{mn}|^p ] \leq C_{p,m}$ .<sup>56</sup>

$$E [ \|I_{U,2}^b\|^p ] \leq C_{p,m} E \left( \left[ \int_{t_n^-}^t |b_s^{mn}|^2 ds \right]^{\frac{p}{2}} \right) \leq C_{p,m} \left( \int_{t_n^-}^t (1 + E [ |X_s^{mn}|^p ]^{\frac{2}{p}} ds \right)^{\frac{p}{2}} \leq C_{p,m} n^{-\frac{p}{2}}.$$

By definition of the  $l^1$ -norm, Jensen's inequality, equivalence of norms on  $\mathbb{R}^m$ :

$$E [ \|I_{U,1}^\gamma\|^p ] = E \left[ \left\| \int_0^{t_n^-} -\gamma^m \cdot [U_s^{mn} - \hat{U}_s^{mn}] ds \right\|^p \right] \leq C_{p,m} \int_0^t E [ \|Z_{U,s}^{mn}\|^p ] ds.$$

Using additionally Minkowski's integral inequality and  $\sup_{s \leq T} E [ \|U_s^{mn}\|^p ] \leq C_{p,m}$ .<sup>57</sup>

$$E [ \|I_{U,2}^\gamma\|^p ] \leq C_{p,m} E \left( \left[ \int_{t_n^-}^t \|U_s^{mn}\|^2 ds \right]^{\frac{p}{2}} \right) \leq C_{p,m} \left( \int_{t_n^-}^t [E [ \|U_s^{mn}\|^p ]^{\frac{2}{p}} ds \right)^{\frac{p}{2}} \leq C_{p,m} n^{-\frac{p}{2}}.$$

By the BDG and Jensen's inequalities, Lipschitz continuity:

$$E [ \|I_{U,1}^\sigma\|^p ] \leq C_{p,m} E \left( \left[ \int_0^{t_n^-} |\sigma_s^{mn} - \hat{\sigma}_s^{mn}|^2 ds \right]^{\frac{p}{2}} \right) \leq C_{p,m} \int_0^t E [ |Z_{X,s}^{mn}|^p ] ds.$$

<sup>56</sup> By Jensen's inequality:  $|X_s^{mn}|^p \leq 2^{p-1} (|X_s|^p + |X_s^{mn} - X_s|^p)$ . Lemma 3.G.2 and the inequality  $\|K - K_{mn}\|_{L^{2\beta}([0, T^*])} \leq \|K - K_m\|_{L^{2\beta}([0, T^*])}$  then yields  $\sup_{s \leq T} E [ |X_s - X_s^{mn}|^p ] \leq C_{p,m}$ . From (3.141) we also have  $\sup_{s \leq T} E [ |X_s|^p ] \leq C_p$ . The claim follows.

<sup>57</sup> By equivalence of norms on  $\mathbb{R}^m$ , the definition of the  $l^1$ -norm, Jensen's inequality:  $E [ \|U_s^{mn}\|^p ] \leq C_{p,m} \sum_{i=1}^m E [ |U_{i,s}^{mn}|^p ]$ . Let  $i \in \{1, \dots, m\}$ . By Jensen's and the BDG inequality, linear growth:  $E [ |U_{i,s}^{mn}|^p ] \leq 2^{p-1} (E [ |\int_0^s e^{-\gamma_i^m(s-z)} b(X_z^{mn}) dz|^p ] + E [ |\int_0^s e^{-\gamma_i^m(s-z)} \sigma(X_z^{mn}) dW_z|^p ] ) \leq C_{p,m} (1 + \int_0^s E [ |X_z^{mn}|^p ] dz)$ . From this, equivalence of norms on  $\mathbb{R}^m$ , and footnote 56, we obtain:  $\sup_{s \leq T} E [ \|U_s^{mn}\|^p ] \leq C_{p,m} (1 + \int_0^T E [ |X_s^{mn}|^p ] ds) \leq C_{p,m}$ .

### 3.G. PROOF OF THEOREM 3.4.3

Moreover, by the BDG inequality, Minkowski's integral inequality, linear growth, Jensen's inequality,  $\sup_{s \leq T} E [|X_s^{mn}|^p] \leq C_{p,m}$ :

$$E [||I_{U,2}^\sigma||^p] \leq C_{p,m} E \left( \left[ \int_{t_n^-}^t |\sigma_s^{mn}|^2 ds \right]^{\frac{p}{2}} \right) \leq C_{p,m} \left( \int_{t_n^-}^t (E [|\sigma_s^{mn}|^p])^{\frac{2}{p}} ds \right)^{\frac{p}{2}} \leq C_{p,m} n^{-\frac{p}{2}}.$$

Combining all estimates:

$$E [||Z_{U,t}^{mn}||^p] \leq C_{p,m} \left( n^{-\frac{p}{2}} + \int_0^t E [||Z_{X,s}^{mn}||^p] ds + \int_0^t E [||Z_{U,s}^{mn}||^p] ds \right). \quad (3.180)$$

We turn our attention to  $Z_{X,t}^{mn}$ . Assume first  $\kappa > 0$ . Note the ordering  $(t_n^- - \kappa\Delta_n)^+ \leq (t - \kappa\Delta_n)^+ \leq t_n^- \leq t$  which we will use in the following. Decompose and rewrite:<sup>58</sup>

$$\begin{aligned} Z_{X,t}^{mn} &= X_t^{mn} - \hat{X}_{t_n^-}^{mn} \\ &= g_0(t) - g_0(t_n^-) + \sum_{i=1}^m c_i^m e^{-\gamma_i^m \kappa \Delta_n} \left( U_{i,(t-\kappa\Delta_n)^+}^{mn} - \hat{U}_{i,(t_n^- - \kappa\Delta_n)^+}^{mn} \right) \\ &\quad + \int_{t_n^-}^t K(t-s) b_s^{mn} ds + \int_{(t-\kappa\Delta_n)^+}^{t_n^-} \left[ K(t-s) b_s^{mn} - K(t_n^- - s) \hat{b}_s^{mn} \right] ds \\ &\quad - \int_{(t_n^- - \kappa\Delta_n)^+}^{(t-\kappa\Delta_n)^+} K(t_n^- - s) \hat{b}_s^{mn} ds + \int_{t_n^-}^t K(t-s) \sigma_s^{mn} dW_s \\ &\quad + \int_{(t-\kappa\Delta_n)^+}^{t_n^-} \left[ K(t-s) \sigma_s^{mn} - K(t_n^- - s) \hat{\sigma}_s^{mn} \right] dW_s \\ &\quad - \int_{(t_n^- - \kappa\Delta_n)^+}^{(t-\kappa\Delta_n)^+} K(t_n^- - s) \hat{\sigma}_s^{mn} dW_s \\ &= g_0(t) - g_0(t_n^-) + \sum_{i=1}^m c_i^m e^{-\gamma_i^m \kappa \Delta_n} \left( U_{i,(t-\kappa\Delta_n)^+}^{mn} - \hat{U}_{i,(t-\kappa\Delta_n)^+}^{mn} \right) \\ &\quad + \int_{t_n^-}^t K(t-s) b_s^{mn} ds + \int_{(t-\kappa\Delta_n)^+}^{t_n^-} \left[ K(t-s) - K(t_n^- - s) \right] b_s^{mn} ds \\ &\quad + \int_{(t-\kappa\Delta_n)^+}^{t_n^-} K(t_n^- - s) \left[ b_s^{mn} - \hat{b}_s^{mn} \right] ds - \int_{(t_n^- - \kappa\Delta_n)^+}^{(t-\kappa\Delta_n)^+} K(t_n^- - s) \hat{b}_s^{mn} ds \\ &\quad + \int_{t_n^-}^t K(t-s) \sigma_s^{mn} dW_s + \int_{(t-\kappa\Delta_n)^+}^{t_n^-} \left[ K(t-s) - K(t_n^- - s) \right] \sigma_s^{mn} dW_s \\ &\quad + \int_{(t-\kappa\Delta_n)^+}^{t_n^-} K(t_n^- - s) \left[ \sigma_s^{mn} - \hat{\sigma}_s^{mn} \right] dW_s - \int_{(t_n^- - \kappa\Delta_n)^+}^{(t-\kappa\Delta_n)^+} K(t_n^- - s) \hat{\sigma}_s^{mn} dW_s \\ &=: I_X^{g_0} + I_X^U + I_{X,1}^b + I_{X,2}^b + I_{X,3}^b + I_{X,4}^b + I_{X,1}^\sigma + I_{X,2}^\sigma + I_{X,3}^\sigma + I_{X,4}^\sigma. \end{aligned}$$

<sup>58</sup> For the third equality note that  $\hat{U}_{i,(t-\kappa\Delta_n)^+}^{mn} = \hat{U}_{i,\{(t-\kappa\Delta_n)^+\}_n^-}^{mn} = \hat{U}_{i,(t_n^- - \kappa\Delta_n)^+}^{mn}$  for  $i = 1, \dots, m$ .

For  $\kappa = 0$  we have  $Z_{X,t}^{mn} = I_X^{g_0} + I_X^U$ . Let  $\kappa \geq 0$  be arbitrary. Since  $g_0$  is locally Hölder continuous of order  $\tilde{\gamma}/2$ , there exists a positive constant  $C$  that depends on only  $T^*$  so

$$|g_0(t_1) - g_0(t_2)| \leq C|t_1 - t_2|^{\frac{\tilde{\gamma}}{2}}, \quad \forall t_1, t_2 \in [0, T^*].$$

Consequently,

$$|I_X^{g_0}|^p \leq C^p |t - t_n^-|^{\frac{\tilde{\gamma}}{2}p} \leq C_p n^{-\frac{p}{2}\tilde{\gamma}}.$$

Furthermore:

$$E [|I_X^U|^p] \leq C_{p,m} E \left[ \|U_{(t-\kappa\Delta_n)^+}^{mn} - \hat{U}_{(t-\kappa\Delta_n)^+}^{mn}\|^p \right] = C_{p,m} E \left[ \|Z_{U,(t-\kappa\Delta_n)^+}^{mn}\|^p \right].$$

We assume  $\kappa > 0$  again. By Jensen's and the Minkowski's integral inequality, linear growth,  $\sup_{s \leq T} E [|X_s^{mn}|^p] \leq C_{p,m}$ , ( $H_0$ ):

$$E [|I_{X,1}^b|^p] \leq C_p E \left[ \left[ \int_{t_n^-}^t K(t-s)^2 (b_s^{mn})^2 ds \right]^{\frac{p}{2}} \right] \quad (3.181)$$

$$\leq C_p \left( \int_{t_n^-}^t (E [|K(t-s)b_s^{mn}|^p])^{\frac{2}{p}} ds \right)^{\frac{p}{2}} \quad (3.182)$$

$$\leq C_p \left( \int_{t_n^-}^t K(t-s)^2 (E [|b_s^{mn}|^p])^{\frac{2}{p}} ds \right)^{\frac{p}{2}} \quad (3.183)$$

$$\leq C_p \left( \int_{t_n^-}^t K(t-s)^2 (1 + E [|X_s^{mn}|^p])^{\frac{2}{p}} ds \right)^{\frac{p}{2}} \quad (3.184)$$

$$\leq C_{p,m} \left( \int_0^{n^{-1}} K(s)^2 ds \right)^{\frac{p}{2}} \quad (3.185)$$

$$\leq C_{p,m} n^{-\frac{p}{2}\tilde{\gamma}}. \quad (3.186)$$

Analogous argumentation yields

$$\begin{aligned} E [|I_{X,2}^b|^p] &\leq C_{p,m} \left( \int_{(t-\kappa\Delta_n)^+}^{t_n^-} [K(t-s) - K(t_n^- - s)]^2 ds \right)^{\frac{p}{2}} \\ &\leq C_{p,m} \left( \int_0^{T^*} [K(s + (t - t_n^-)) - K(s)]^2 ds \right)^{\frac{p}{2}} \\ &\leq C_{p,m} n^{-\frac{p}{2}\tilde{\gamma}}. \end{aligned}$$

By the Jensen's and Hölder's inequalities, Lipschitz continuity,  $\|K\|_{L^{2\beta}([0, T^*])} < \infty$ :

$$\begin{aligned} E [|I_{X,3}^b|^p] &= E \left[ \left| \int_{(t-\kappa\Delta_n)^+}^{t_n^-} K(t_n^- - s) [b_s^{mn} - \hat{b}_s^{mn}] ds \right|^p \right] \\ &\leq C_{p,m} \left( \int_{(t-\kappa\Delta_n)^+}^{t_n^-} |K(t_n^- - s)|^{2\beta} ds \right)^{\frac{p}{2\beta}} E \left[ \left( \int_{(t-\kappa\Delta_n)^+}^{t_n^-} |b_s^{mn} - \hat{b}_s^{mn}|^{2\beta^*} ds \right)^{\frac{p}{2\beta^*}} \right] \end{aligned}$$

### 3.G. PROOF OF THEOREM 3.4.3

$$\leq C_{p,m} \int_0^t E [|Z_{X,s}^{mn}|^p] ds.$$

Analogous to (3.181)-(3.186) using Lemma 3.G.1 instead of the moment bound applied for those equations and that  $K$  is non-increasing:

$$E [|I_{X,4}^b|^p] \leq C_{p,m} \left( \int_{(t_n^- - \kappa \Delta_n)^+}^{(t - \kappa \Delta_n)^+} K(t_n^- - s)^2 ds \right)^{\frac{p}{2}} \leq C_{p,m} \left( \int_0^{\Delta_n} K(s)^2 ds \right)^{\frac{p}{2}} \leq C_{p,m} n^{-\frac{p}{2} \tilde{\gamma}}.$$

The terms  $E [|I_{X,i}^{\sigma}|^p]$  can be bounded similarly to  $E [|I_{X,i}^b|^p]$  for  $i = 1, 2, 3, 4$ , except we must also to use the BDG inequality. Combining the estimates, we, for arbitrary  $\kappa \geq 0$ , conclude

$$E [|Z_{X,t}^{mn}|^p] \leq C_{p,m} \left( n^{-\frac{p}{2} \tilde{\gamma}} + E [||Z_{U,(t-\kappa\Delta_n)^+}^{mn}||^p] + \int_0^t E [|Z_{X,s}^{mn}|^p] ds \right). \quad (3.187)$$

Equations (3.180) and (3.187) imply

$$h(t) \leq C_{p,m} \left( n^{-\frac{p}{2}(\tilde{\gamma} \wedge 1)} + \int_0^t h(s) ds \right), \quad h(t) := \sup_{s \leq t} (E [|Z_{X,s}^{mn}|^p] \vee E [||Z_{U,s}^{mn}||^p]).$$

Then proof is then completed by an application of Gronwall's inequality.<sup>59</sup>  $\square$

**Theorem 3.G.4.** (Theorem 3.4.3 of page 50 restated) Invoke Assumption 3.4.1 and assume that Equation (3.33) of Definition 3.2.1 is used. For  $p \geq \frac{2\beta}{\beta-1}$  and  $m \in \mathbb{N}$  there then exist positive constants  $C_p$  and  $C_{p,m}$ , the first of which depends on  $p$  only, the latter on both  $p$  and  $m$ , so

$$\sup_{(t,\tau) \in A} E [|g_t(\tau) - \hat{g}_t^{mn}(\tau)|^p] \leq C_p ||K - K_{mn}||_{L^{2\beta}([0,T^*])}^p + C_{p,m} n^{-\frac{p}{2}(\tilde{\gamma} \wedge 1)}, \quad n \in \mathbb{N}. \quad (3.188)$$

*Proof.* Let  $p \geq \frac{2\beta}{\beta-1}$ ,  $m, n \in \mathbb{N}$ . By Jensen's inequality:

$$\sup_{(t,\tau) \in A} E [|g_t(\tau) - \hat{g}_t^{mn}(\tau)|^p] \leq 2^{p-1} \sup_{(t,\tau) \in A} E [|g_t(\tau) - g_t^{mn}(\tau)|^p] \quad (3.189)$$

$$+ 2^{p-1} \sup_{(t,\tau) \in A} E [|g_t^{mn}(\tau) - \hat{g}_t^{mn}(\tau)|^p]. \quad (3.190)$$

From Lemma 3.G.2 we have already

$$\sup_{(t,\tau) \in A} E [|g_t(\tau) - g_t^{mn}(\tau)|^p] \leq C_p ||K - K_{mn}||_{L^{2\beta}([0,T^*])}^p. \quad (3.191)$$

It remains to bound the term in (3.190). Define  $Z_{g,t}^{mn}(\tau) := g_t^{mn}(\tau) - \hat{g}_t^{mn}(\tau)$ ,  $(t, \tau) \in A$ . In what follows, we bound  $E [|Z_{g,t}^{mn}(\tau)|^p]$  separately for each case of Definition 3.3.1:

<sup>59</sup> That  $h$  is bounded on bounded intervals follows in this context by use of the triangle inequality and that  $\sup_{s \leq T} (E [|X_s^{mn}|^p] \vee E [|\hat{X}_s^{mn}|^p] \vee E [||U_s^{mn}||^p] \vee E [||\hat{U}_s^{mn}||^p]) < \infty$ . The latter follows by Lemma 3.G.1 and that  $\sup_{s \leq T} E [|X_s^{mn}|^p] < \infty$  and  $\sup_{s \leq T} E [||U_s^{mn}||^p] < \infty$  as shown already.

Say  $(t, \tau) \in A$  with  $\tau \geq \kappa \Delta_n$ . Then

$$Z_{g,t}^{mn}(\tau) = g_t^{mn}(\tau) - \hat{g}_{t_n^-}^{mn}(\tau) = g_0(t + \tau) - g_0(t_n^- + \tau) + \sum_{i=1}^m c_i^m e^{-\gamma_i^m \tau} \left( U_{i,t}^{mn} - \hat{U}_{i,t_n^-}^{mn} \right).$$

By  $\hat{U}_{t_n^-}^{mn} = \hat{U}_t^{mn}$ , that  $g_0$  is locally Hölder continuous of order  $\tilde{\gamma}/2$ , Lemma 3.G.3:

$$E [|Z_{g,t}^{mn}(\tau)|^p] \leq C_{p,m} \left( n^{-\frac{p}{2}\tilde{\gamma}} + E \left[ \|U_t^{mn} - \hat{U}_t^{mn}\|^p \right] \right) \leq C_{p,m} n^{-\frac{p}{2}(\tilde{\gamma} \wedge 1)}.$$

Let  $(t, \tau) \in A$  and say that  $\kappa > 0$ ,  $\tau = l\Delta_n$ ,  $l \in \{0, 1, \dots, \kappa - 1\}$ . Note the ordering  $(t_n^- - (\kappa - l)\Delta_n)^+ \leq (t - (\kappa - l)\Delta_n)^+ \leq t_n^- \leq t$ . Then by (3.61)-(3.64) and Definition 3.3.1:

$$\begin{aligned} Z_{g,t}^{mn}(\tau) &= g_t^{mn}(\tau) - \hat{g}_{t_n^-}^{mn}(\tau) \\ &= g_0(t + \tau) - g_0(t_n^- + \tau) + \sum_{i=1}^m c_i^m e^{-\gamma_i^m \kappa \Delta_n} \left( U_{i,(t-(\kappa-l)\Delta_n)^+}^{mn} - \hat{U}_{i,(t_n^--(\kappa-l)\Delta_n)^+}^{mn} \right) \\ &\quad + \int_{(t-(\kappa-l)\Delta_n)^+}^t K(t + \tau - s) (b_s^{mn} ds + \sigma_s^{mn} dW_s) \\ &\quad - \int_{(t_n^--(\kappa-l)\Delta_n)^+}^{t_n^-} K(t_n^- + \tau - s) (\hat{b}_s^{mn} ds + \hat{\sigma}_s^{mn} dW_s) \\ &= g_0(t + \tau) - g_0(t_n^- + \tau) + \sum_{i=1}^m c_i^m e^{-\gamma_i^m \kappa \Delta_n} \left( U_{i,(t-(\kappa-l)\Delta_n)^+}^{mn} - \hat{U}_{i,(t_n^--(\kappa-l)\Delta_n)^+}^{mn} \right) \\ &\quad + \int_{t_n^-}^t K(t + \tau - s) b_s^{mn} ds + \int_{(t-(\kappa-l)\Delta_n)^+}^{t_n^-} K(t + \tau - s) [b_s^{mn} - \hat{b}_s^{mn}] ds \\ &\quad + \int_{(t-(\kappa-l)\Delta_n)^+}^{t_n^-} [K(t + \tau - s) - K(t_n^- + \tau - s)] \hat{b}_s^{mn} ds \\ &\quad - \int_{(t_n^--(\kappa-l)\Delta_n)^+}^{(t-(\kappa-l)\Delta_n)^+} K(t_n^- + \tau - s) \hat{b}_s^{mn} ds + \int_{t_n^-}^t K(t + \tau - s) \sigma_s^{mn} dW_s \\ &\quad + \int_{(t-(\kappa-l)\Delta_n)^+}^{t_n^-} K(t + \tau - s) [\sigma_s^{mn} - \hat{\sigma}_s^{mn}] dW_s \\ &\quad + \int_{(t-(\kappa-l)\Delta_n)^+}^{t_n^-} [K(t + \tau - s) - K(t_n^- + \tau - s)] \hat{\sigma}_s^{mn} dW_s \\ &\quad - \int_{(t_n^--(\kappa-l)\Delta_n)^+}^{(t-(\kappa-l)\Delta_n)^+} K(t_n^- + \tau - s) \hat{\sigma}_s^{mn} dW_s \\ &=: I_{g_0} + I_U + I_1^b + I_2^b + I_3^b + I_4^b + I_1^\sigma + I_2^\sigma + I_3^\sigma + I_4^\sigma. \end{aligned}$$

Because  $g_0$  is locally Hölder continuous of order  $\tilde{\gamma}/2$ :

$$|I_{g_0}|^p \leq C_p n^{-\frac{p}{2}\tilde{\gamma}}. \quad (3.192)$$

### 3.G. PROOF OF THEOREM 3.4.3

---

The definition of the  $l^1$ -norm, the triangle inequality, equivalence of norms on  $\mathbb{R}^m$ , Lemma 3.G.3, yields

$$E[|I_U|^p] \leq C_{p,m} E \left[ \|U_{(t-(\kappa-l)\Delta_n)^+}^{mn} - \hat{U}_{(t-(\kappa-l)\Delta_n)^+}^{mn}\|^p \right] \leq C_{p,m} n^{-\frac{p}{2}(\tilde{\gamma} \wedge 1)}. \quad (3.193)$$

By Jensen's inequality, Minkowski's integral inequality, linear growth,  $\sup_{s \leq T} E[|X_s^{mn}|^p] \leq C_{p,m}$ , that  $K$  is non-increasing,  $(H_0)$ :

$$E[|I_1^b|^p] \leq C_{p,m} \left( \int_{t_n^-}^t K(t+\tau-s)^2 ds \right)^{\frac{p}{2}} \leq C_{p,m} \left( \int_0^{n^{-1}} K(s)^2 ds \right)^{\frac{p}{2}} \leq C_{p,m} n^{-\frac{p}{2}\tilde{\gamma}}. \quad (3.194)$$

Using Jensen's and Hölder's inequalities, Lipschitz continuity,  $\|K\|_{L^{2\beta}([0, T^*])} < \infty$ , Lemma 3.G.3:

$$E[|I_2^b|^p] \leq C_{p,m} \int_{(t-(\kappa-l)\Delta_n)^+}^{t_n^-} E[|X_s^{mn} - \hat{X}_s^{mn}|^p] ds \leq C_{p,m} n^{-\frac{p}{2}(\tilde{\gamma} \wedge 1)}.$$

By Jensen's and Minkowski's integral inequalities, linear growth, Lemma 3.G.1,  $(H_0)$ :

$$\begin{aligned} E[|I_3^b|^p] &\leq C_{p,m} \left( \int_{(t-(\kappa-l)\Delta_n)^+}^{t_n^-} (K(t+\tau-s) - K(t_n^- + \tau - s))^2 ds \right)^{\frac{p}{2}} \\ &\leq C_{p,m} \left( \int_0^{T^*} (K(s + (t - t_n^-)) - K(s))^2 ds \right)^{\frac{p}{2}} \\ &\leq C_{p,m} n^{-\frac{p}{2}\tilde{\gamma}}. \end{aligned}$$

As we derived (3.194) but using Lemma 3.G.1 to bound the moment:

$$E[|I_4^b|^p] \leq C_{p,m} \left( \int_{(t_n^- - (\kappa-l)\Delta_n)^+}^{(t-(\kappa-l)\Delta_n)^+} K(t_n^- + \tau - s)^2 ds \right)^{\frac{p}{2}} \leq C_{p,m} \left( \int_0^{n^{-1}} K(s)^2 ds \right)^{\frac{p}{2}} \leq C_{p,m} n^{-\frac{p}{2}\tilde{\gamma}}.$$

We may bound  $E[|I_i^\sigma|^p]$  similarly to  $E[|I_i^b|^p]$  for  $i = 1, 2, 3, 4$ , except we should also use the BDG inequality. Combining the estimates, we conclude again

$$E[|Z_{g,t}^{mn}(\tau)|^p] \leq C_{p,m} n^{-\frac{p}{2}(\tilde{\gamma} \wedge 1)}. \quad (3.195)$$

Let now  $(t, \tau) \in A$  and  $\tau \in [0, \kappa\Delta_n]$  be more general. We still assume  $\kappa > 0$ . Then

$$Z_{g,t}^{mn}(\tau) = \frac{\Delta_n - (\tau - \tau_n^-)}{\Delta_n} (g_t^{mn}(\tau) - \hat{g}_t^{mn}(\tau_n^-)) + \frac{\tau - \tau_n^-}{\Delta_n} (g_t^{mn}(\tau) - \hat{g}_t^{mn}(\tau_n^+)).$$

Consequently, using the triangle and Jensen's inequalities, (3.195):

$$E[|Z_{g,t}^{mn}(\tau)|^p] \leq 4^{p-1} (E[|g_t^{mn}(\tau) - g_t^{mn}(\tau_n^-)|^p] + E[|g_t^{mn}(\tau_n^-) - \hat{g}_t^{mn}(\tau_n^-)|^p]) \quad (3.196)$$

$$+ E[|g_t^{mn}(\tau) - g_t^{mn}(\tau_n^+)|^p] + E[|g_t^{mn}(\tau_n^+) - \hat{g}_t^{mn}(\tau_n^+)|^p]) \quad (3.197)$$

$$\leq C_{p,m} \left( n^{-\frac{p}{2}(\tilde{\gamma} \wedge 1)} + E [|g_t^{mn}(\tau) - g_t^{mn}(\tau_n^-)|^p] \right) \quad (3.198)$$

$$+ E [|g_t^{mn}(\tau) - g_t^{mn}(\tau_n^+)|^p] \Big). \quad (3.199)$$

Note that

$$\begin{aligned} g_t^{mn}(\tau) - g_t^{mn}(\tau_n^-) &= g_0(t + \tau) - g_0(t + \tau_n^-) \\ &\quad + \int_0^t (K(t + \tau - s) - K(t + \tau_n^- - s)) b_s^{mn} ds \\ &\quad + \int_0^t (K(t + \tau - s) - K(t + \tau_n^- - s)) \sigma_s^{mn} dW_s. \end{aligned}$$

By Jensen's and the BDG inequality, local  $\frac{\tilde{\gamma}}{2}$ -Hölder continuity of  $g_0$ , Minkowski's integral inequality, linear growth,  $\sup_{s \leq T} E [|X_s^{mn}|^p] \leq C_{p,m}$ , ( $H_0$ ):

$$E [|g_t^{mn}(\tau) - g_t^{mn}(\tau_n^-)|^p] \quad (3.200)$$

$$\leq C_{p,m} \left( n^{-\frac{p}{2}\tilde{\gamma}} + \left[ \int_0^t (K(t + \tau - s) - K(t + \tau_n^- - s))^2 ds \right]^{\frac{p}{2}} \right) \quad (3.201)$$

$$\leq C_{p,m} \left( n^{-\frac{p}{2}\tilde{\gamma}} + \left[ \int_0^{T^*} (K(s + (\tau - \tau_n^-)) - K(s))^2 ds \right]^{\frac{p}{2}} \right) \quad (3.202)$$

$$\leq C_{p,m} n^{-\frac{p}{2}\tilde{\gamma}}. \quad (3.203)$$

By the same argumentation:

$$E [|g_t^{mn}(\tau) - g_t^{mn}(\tau_n^+)|^p] \leq C_{p,m} n^{-\frac{p}{2}\tilde{\gamma}}. \quad (3.204)$$

Combining (3.200)-(3.204) in (3.196)-(3.199), we conclude again

$$E [|Z_{g,t}^{mn}(\tau)|^p] \leq C_{p,m} n^{-\frac{p}{2}(\tilde{\gamma} \wedge 1)}. \quad (3.205)$$

Because we have established (3.205) for all  $(t, \tau) \in A$ , the proof is done.  $\square$

### 3.H Proof of Proposition 3.5.3 (continued)

*Proof.* We argue that  $H$  is positive semi-definite. Recall from (3.79) that

$$f(x) = \int_{[0, \infty)} e^{-\gamma(b-a)x} \hat{\mu}(d\gamma), \quad x \in [0, 1].$$

Define for  $\gamma \geq 0$  and  $k \in \mathbb{N}_0$  the terms  $\tilde{h}_k(\gamma) := e^{-\gamma(b-a)\frac{k}{2N}}$  and

$$\tilde{H}(\gamma) := \begin{pmatrix} \tilde{h}_0(\gamma) & \tilde{h}_1(\gamma) & \dots & \tilde{h}_{N-1}(\gamma) & \tilde{h}_N(\gamma) \\ \tilde{h}_1(\gamma) & \ddots & \ddots & \ddots & \tilde{h}_{N+1}(\gamma) \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \tilde{h}_{N-1}(\gamma) & \ddots & \ddots & \ddots & \tilde{h}_{2N-1}(\gamma) \\ \tilde{h}_N(\gamma) & \tilde{h}_{N+1}(\gamma) & \dots & \tilde{h}_{2N-1}(\gamma) & \tilde{h}_{2N}(\gamma) \end{pmatrix}.$$

### 3.I. PROOF OF THEOREM 3.6.3

---

By

$$h_k \stackrel{\text{def.}}{=} f\left(\frac{k}{2N}\right) = \int_{[0,\infty)} e^{-\gamma(b-a)\frac{k}{2N}} \hat{\mu}(d\gamma) = \int_{[0,\infty)} \tilde{h}_k(\gamma) \hat{\mu}(d\gamma), \quad k = 0, 1, \dots, 2N,$$

and  $H$  being the Hankel matrix constructed from  $\{h_0, \dots, h_{2N}\}$ , we deduce that

$$H = \int_{[0,\infty)} \tilde{H}(\gamma) \hat{\mu}(d\gamma). \quad (3.206)$$

Since  $\hat{\mu}$  is non-negative it by (3.206) suffices to show that  $\tilde{H}(\gamma)$  is positive semi-definite for any  $\gamma \geq 0$ . We to this end consider the so-called 'Hamburger moment problem': Let  $\{s_k\}_{k=0}^{\infty}$  be a sequence of real numbers. By Theorem 3.8 of [49, pp. 63] the Hankel matrices  $S_k := \{s_{i+j-2}\}_{i,j=1}^{k+1}$ ,  $k \in \mathbb{N}_0$ , are positive semi-definite if and only if

$$s_k = \int_{\mathbb{R}} y^k \lambda(dy), \quad k \in \mathbb{N}_0,$$

for a positive Radon measure  $\lambda$  on  $\mathbb{R}$  where the integrals must exist and be finite. To prove  $\tilde{H}(\gamma)$  positive semi-definite, we therefore only need to argue the representation

$$\tilde{h}_k(\gamma) = \int_{\mathbb{R}} y^k \lambda(dy), \quad k \in \mathbb{N}_0. \quad (3.207)$$

Equation (3.207) holds if we let  $\lambda$  be the Dirac measure with point mass in  $e^{-\gamma(b-a)\frac{1}{2N}}$ . We conclude that  $\tilde{H}(\gamma)$ ,  $\gamma \geq 0$ , are positive semi-definite. Then so is  $H$ . □

### 3.I Proof of Theorem 3.6.3

*Proof.* Let  $p \geq \frac{2\beta}{\beta-1}$ ,  $m, n \in \mathbb{N}$ , and assume that  $\kappa\Delta_n \leq \Delta_{\text{vix}}$ . Throughout  $C_p$  and  $C_{p,m}$  will denote positive constants that may change from line to line. A careful reading of the proof shows that  $C_{p,m}$  can be chosen independent of  $n$ ,  $C_p$  of both  $m$  and  $n$ .

By Jensen's inequality:

$$\sup_{t \in [0, T]} E \left[ |\text{VIX}_t^2 - \widehat{\text{VIX}}_{mn,t}^2|^p \right] \leq 2^{p-1} \sup_{t \in [0, T]} E \left[ |\text{VIX}_t^2 - \text{VIX}_{mn,t}^2|^p \right] \quad (3.208)$$

$$+ 2^{p-1} \sup_{t \in [0, T]} E \left[ |\text{VIX}_{mn,t}^2 - \widehat{\text{VIX}}_{mn,t}^2|^p \right]. \quad (3.209)$$

For the term of (3.208), we get with Jensen's inequality and Lemma 3.G.2:

$$\begin{aligned} \sup_{t \in [0, T]} E \left[ |\text{VIX}_t^2 - \text{VIX}_{mn,t}^2|^p \right] &\leq C_p \int_0^{\Delta_{\text{vix}}} \sup_{t \in [0, T]} E \left[ |\tilde{\xi}_t(\tau) - \tilde{\xi}_t^{mn}(\tau)|^p \right] d\tau \\ &\leq C_p \sup_{(t, \tau) \in A} E \left[ |\tilde{\xi}_t(\tau) - \tilde{\xi}_t^{mn}(\tau)|^p \right] \end{aligned}$$

$$\leq C_p \|K - K_{mn}\|_{L^{2\beta}([0, T + \Delta_{\text{vix}}])}^p.$$

For the term of (3.209), note first that

$$\begin{aligned} \frac{\Delta_{\text{vix}}}{100^2} \left[ \text{VIX}_{mn,t}^2 - \widehat{\text{VIX}}_{mn,t}^2 \right] &= \int_0^{\kappa \Delta_n} \left[ \tilde{\xi}_t^{mn}(\tau) - \frac{1}{2} (\hat{\xi}_t^{mn}(\tau_n^-) + \hat{\xi}_t^{mn}(\tau_n^+)) \right] d\tau \\ &\quad + \int_{\kappa \Delta_n}^{\Delta_{\text{vix}}} [\xi_0(t + \tau) - \xi_0(t_n^- + \tau)] d\tau \\ &\quad + \sum_{\substack{i=1, \dots, m \\ \gamma_i^m \neq 0}} \frac{c_i^m}{\gamma_i^m} (e^{-\gamma_i^m \kappa \Delta_n} - e^{-\gamma_i^m \Delta_{\text{vix}}}) [U_{i,t}^{mn} - \hat{U}_{i,t}^{mn}] \\ &\quad + \sum_{\substack{i=1, \dots, m \\ \gamma_i^m = 0}} c_i^m (\Delta_{\text{vix}} - \kappa \Delta_n) [U_{i,t}^{mn} - \hat{U}_{i,t}^{mn}] \\ &=: I_1 + I_2 + I_3 + I_4. \end{aligned}$$

For the above, the reader should recall that

$$\widehat{\text{VIX}}_{mn,t}^2 = \widehat{\text{VIX}}_{mn,t_n^-}^2, \quad \hat{U}_t^{mn} = \hat{U}_{t_n^-}^{mn}, \quad \hat{\xi}_t^{mn} = \hat{\xi}_{t_n^-}^{mn}, \quad t \in [0, T].$$

By Jensen's inequality:

$$\begin{aligned} E[|I_1|^p] &\leq C_p \int_0^{\kappa \Delta_n} \left( E[|\tilde{\xi}_t^{mn}(\tau) - \tilde{\xi}_t^{mn}(\tau_n^-)|^p] + E[|\tilde{\xi}_t^{mn}(\tau_n^-) - \hat{\xi}_t^{mn}(\tau_n^-)|^p] \right. \\ &\quad \left. + E[|\tilde{\xi}_t^{mn}(\tau) - \tilde{\xi}_t^{mn}(\tau_n^+)|^p] + E[|\tilde{\xi}_t^{mn}(\tau_n^+) - \hat{\xi}_t^{mn}(\tau_n^+)|^p] \right) d\tau. \end{aligned}$$

Because (3.200)-(3.204) from the proof of Theorem 3.G.4 generalises to arbitrary  $(t, \tau) \in A$ :

$$\sup_{(t, \tau) \in A} \left( E[|\tilde{\xi}_t^{mn}(\tau) - \tilde{\xi}_t^{mn}(\tau_n^-)|^p] + E[|\tilde{\xi}_t^{mn}(\tau) - \tilde{\xi}_t^{mn}(\tau_n^+)|^p] \right) \leq C_{p,m} n^{-\frac{p}{2} \tilde{\gamma}}.$$

In the proof of Theorem 3.G.4, the bound (3.205) was shown for arbitrary  $(t, \tau) \in A$ . It follows that

$$\sup_{(t, \tau) \in A} \left( E[|\tilde{\xi}_t^{mn}(\tau_n^-) - \hat{\xi}_t^{mn}(\tau_n^-)|^p] + E[|\tilde{\xi}_t^{mn}(\tau_n^+) - \hat{\xi}_t^{mn}(\tau_n^+)|^p] \right) \leq C_{p,m} n^{-\frac{p}{2} (\tilde{\gamma} \wedge 1)}.$$

Combining the above, we deduce:  $E[|I_1|^p] \leq C_{p,m} n^{-\frac{p}{2} (\tilde{\gamma} \wedge 1)}$ .

By local  $\frac{\tilde{\gamma}}{2}$ -Hölder continuity of  $\xi_0$ , Jensen's inequality, we have also

$$|I_2|^p \leq C_p \sup_{(t, \tau) \in A} |\xi_0(t + \tau) - \xi_0(t_n^- + \tau)|^p \leq C_p n^{-\frac{p}{2} \tilde{\gamma}}.$$

### 3.I. PROOF OF THEOREM 3.6.3

---

An application of Jensen's inequality and Lemma 3.G.3 then yields

$$\begin{aligned} \sup_{t \in [0, T]} E \left[ |\text{VIX}_{mn,t}^2 - \widehat{\text{VIX}}_{mn,t}^2|^p \right] &\leq C_p (E[|I_1|^p] + E[|I_2|^p] + E[|I_3|^p] + E[|I_4|^p]) \\ &\leq C_{p,m} \left( n^{-\frac{p}{2}(\tilde{\gamma} \wedge 1)} + E \left[ \| |U_t^{mn} - \hat{U}_t^{mn}| \|^p \right] \right) \\ &\leq C_{p,m} n^{-\frac{p}{2}(\tilde{\gamma} \wedge 1)}. \end{aligned}$$

Combining the estimates in (3.208)-(3.209), we obtain the desired result:

$$\sup_{t \in [0, T]} E \left[ |\text{VIX}_t^2 - \widehat{\text{VIX}}_{mn,t}^2|^p \right] \leq C_p \|K - K_{mn}\|_{L^{2\beta}([0, T + \Delta_{\text{vix}}])}^p + C_{p,m} n^{-\frac{p}{2}(\tilde{\gamma} \wedge 1)}.$$

The last part of the theorem can be proved as in the proof of Theorem 3.4.4.  $\square$

## Chapter 4

# Empirical analysis of rough and classical stochastic volatility models to the SPX and VIX markets

Sigurd Emil Rømer

### Abstract

We conduct an empirical analysis of rough and classical stochastic volatility models to the SPX and VIX options markets. Our analysis focusses primarily on calibration quality and is split in two parts. In the first part, we perform a historical calibration to SPX options over the years 2004-2019 of a selection of models that include the one-factor rough Bergomi model. In the second part, we consider three calibration dates with low, typical, and high volatility, but examine a wide selection of models and calibrate to both SPX options and jointly to SPX and VIX options. The key results are as follows: The rough Bergomi model fails to create a term structure of smile effect that is sufficiently pronounced for SPX options. Moreover, we discover that short-expiry SPX smiles generally are more symmetric than long-expiry smiles, a feature we neither find that the rough Bergomi model can reproduce. We propose an alternative volatility model driven by two Ornstein-Uhlenbeck processes that uses a non-standard transformation function. Calibrating it to SPX options we obtain almost perfect fits and calibrating it jointly to SPX and VIX options we obtain very decent fits. This suggests—contrary to what one might be led to believe based on much of the existing literature—that the joint SPX-VIX calibration problem is largely solvable with classical two-factor volatility, all without roughness and jumps.

*Keywords:* Rough volatility; Multifactor volatility; Calibration; SPX options; VIX options.

## 4.1 Introduction

In the influential paper [33], it is argued that volatility on a large number of financial assets is rough. A stochastic process is considered *rough* if the trajectories are less Hölder continuous than those of Brownian motion. Modelling volatility rough in practice implies explosive volatility paths and an at-the-money implied volatility skew that for many one-factor models is of the form  $\sim cT^{H-\frac{1}{2}}$  where  $T$  is the expiry,  $H \in (0, \frac{1}{2})$  the so-called Hurst exponent, and  $c$  is an anonymous constant. The power-law term structure of skew is remarkably consistent with typical SPX volatility surfaces where often  $H \approx 0.1$ . Several papers have by now also demonstrated that even simple one-factor rough volatility models can obtain excellent fits to the entire surface, all without jumps; see [7] for the rough Bergomi model or [22] for a rough Heston model. Theoretical works on rough volatility are [6, 8, 20, 23, 24, 25, 26, 27, 28, 29].

The modelling and numerical treatment of volatility derivatives have also received attention in the rough volatility literature [5, 32, 42, 44, 49, 58]. An important example is the VIX option (a European call or put on the VIX index). Since the VIX index is defined in terms of quoted SPX options<sup>1</sup>, the markets for SPX and VIX options are deeply connected. The analysis of [17] however also shows that each market contains some amount of distinct risk-neutral information and that predictive performance can be improved by taking advantage of the information contained in both. It is therefore highly worthwhile to look for models that can calibrate jointly to SPX and VIX options.

Numerous papers have been written on the joint calibration problem since VIX options first started trading in 2006. An often encountered challenge when calibrating jointly is to reconcile the typically very steep short-term SPX skews—which in a continuous-path setting imply a very high volatility-of-volatility—with the level of the VIX implied volatility which comparatively is often much lower. The problem has been described with more rigour in [36] where Julien Guyon shows that a certain inversion-of-convex-ordering property must be satisfied to fit both markets. For a continuous-path model it, in essence, boils down to a volatility process with a high volatility-of-volatility, very fast mean-reversion, and a sizeable negative correlation to the S&P 500 index.

Previous attempts with continuous-path models have had some success, although, the fits are not always perfect; the double CEV model introduced and calibrated in [31] is one example. Most (more or less) successful attempts have instead involved jumps; see e.g. [18, 48, 54]. However, the inclusion of jumps may be undesirable as it leads to indeterminacy in the construction of hedge portfolios. Continuous-path models are therefore preferable in general. The rough volatility finding is promising in this regard since rough volatility models naturally produce very fast mean-reversion and high volatility-of-volatility, which suggests that they would do better than classical models in solving the joint problem without jumps. An important related work is [32] where the authors calibrate their quadratic rough Heston model and obtain a decent fit to both markets, although they only consider short-term options as they use a simplified parameterisation.

---

<sup>1</sup> We refer to the VIX whitepaper downloadable via <https://www.cboe.com/education/research> (accessed July 27, 2021).

Despite the promising results, we believe the empirical literature is lacking when it comes to the testing of rough volatility pricing models. For example, while several calibrations to SPX options have been published, no systematic validation of the fits exist and we therefore believe it is still an open question if simple models such as rough Bergomi truly are consistent with the SPX volatility surface—also across market scenarios. Moreover, only a small selection of rough volatility models have been tested for the joint calibration problem. The present author is in fact only aware of the calibration in [32] of the quadratic rough Heston model (but only to short-term options) and the skewed rough Bergomi model calibrated in [37]; the latter though fails to reconcile the two markets.

The purpose of our paper is to help fill the gap of empirical work. Our analysis is divided in two parts and focuses primarily on calibration quality. However, a severely limiting factor when calibrating rough models is the slow computational speeds (relative to classical models) that result from the inherent path dependence. Fortunately, a series of recent papers [9, 43, 53] have shown how a pricing model can be significantly sped up with a neural network representation. In part one of our analysis, we therefore employ neural network techniques to calibrate a selection of models to SPX options over the years 2004-2019.<sup>2</sup> We include the rough Bergomi model, an extended rough Bergomi model, and add the original Heston model of [40] as a benchmark. The reader will see that these models are incapable of perfect fits across all market scenarios. In search of better results, and to explore the joint problem, we provide a second analysis where we test more advanced models, including several two-factor volatility models, and a quadratic rough Heston model.<sup>3</sup> We calibrate the models with Monte Carlo to SPX options and jointly to SPX and VIX options, albeit on a smaller set of dates, and analyse the results.

The paper is structured as follows: Sections 4.2-4.4 cover the first part of our analysis. We define the relevant models in Section 4.2, and, in Section 4.3, outline our neural network methodology and validate the approximations. In Section 4.4, we present and discuss the historical SPX calibrations. The second part of our analysis is provided in Section 4.5. We conclude in Section 4.6.

## 4.2 Setup and models

We consider a filtered probability space  $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{Q})$  where  $\mathbb{Q}$  is a risk-neutral pricing measure. The filtration  $(\mathcal{F}_t)_{t \geq 0}$  is assumed to be generated by two possibly correlated Brownian motions  $W_{1,t}$  and  $W_{2,t}$  and if not already adapted also by a stochastic process  $V_t$  which we call the *instantaneous variance*. We always let the filtration be augmented so the usual hypothesis holds. For notation, we write  $E_t(\cdot) := E(\cdot | \mathcal{F}_t)$  for conditional expectations. The underlying asset is assumed to have risk-neutral dynamics of the form

$$dS_t = S_t(r(t) - q(t))dt + S_t\sqrt{V_t}dW_{1,t} \quad (4.1)$$

where  $r(t)$  is the deterministic risk free interest rate and  $q(t)$  is the continuous deterministic dividend yield. Define  $\xi_t(u) = E_t(V_u)$ ,  $0 \leq t \leq u$ , which are the *forward variances*.

<sup>2</sup> The trained neural networks, including interfaces for Matlab, R, and Python, and other codes related to their construction, are made public at [https://github.com/sigurdroemer/rough\\_volatility](https://github.com/sigurdroemer/rough_volatility).

<sup>3</sup> Unless otherwise stated, we by the number of 'factors' refer to the number of Brownian motions that drive volatility.

### 4.2.1. HESTON

---

We denote the forward price of the underlying asset by  $F_{t,T} = S_t \exp(\int_t^T (r(s) - q(s)) ds)$ ,  $0 \leq t \leq T$ . Given a European call or put option with strike  $K$  and expiry  $T$  as observed at time  $t$ , we define its *log-moneyness* by  $k = \log(K/F_{t,T})$  and write  $\sigma_{BS}(k, T - t)$  for the Black-Scholes implied volatility. The arbitrage free price at time zero is

$$e^{-\int_0^T r(s) ds} E(\max\{\beta \cdot (S_T - K), 0\}) \quad (4.2)$$

where  $\beta = 1$  for a call option,  $\beta = -1$  for a put. Applying Ito's lemma to the log of the forward price and performing some simple rewritings, we may also write (4.2) as

$$F_{0,T} e^{-\int_0^T r(s) ds} E(\max\{\beta \cdot (\tilde{S}_T - \tilde{K}), 0\}) \quad (4.3)$$

where  $\tilde{K} = \frac{K}{F_{0,T}}$  and  $d\tilde{S}_t = \tilde{S}_t \sqrt{V_t} dW_{1,t}$ ,  $\tilde{S}_0 = 1$ . From (4.3) we see that it suffices to train our neural networks under the assumption of an initial asset price of 1 and zero interest rates and dividends. The general case is recovered by adjusting the strike and rescaling the price.

In what follows, we outline the models that we test in part one of our analysis.

#### 4.2.1 Heston

In the Heston model of [40], we have

$$dV_t = \kappa(v_\infty - V_t)dt + \eta\sqrt{V_t}dW_{2,t} \quad (4.4)$$

where  $V_0, \kappa, v_\infty, \eta > 0$ ,  $dW_{1,t}dW_{2,t} = \rho dt$ ,  $\rho \in [-1, 1]$ . The model has been popularised mostly due to the fact that the characteristic function of  $\log(S_t)$  is known analytically which in turn enables fast pricing with Fourier methods. For option pricing we use a Fourier based pricing formula from [50] as it appears in [51] though rewritten as an integral over  $\mathbb{R}_+$  and with an optimal dampening parameter.

#### 4.2.2 Rough Bergomi

The rough Bergomi model of [7] is defined by

$$V_t = \xi_0(t) \mathcal{E} \left( \eta \sqrt{2H} \int_0^t (t-s)^{H-\frac{1}{2}} dW_{2,s} \right), \quad t \geq 0,$$

where  $\eta > 0$ ,  $H \in (0, \frac{1}{2})$ ,  $dW_{1,t}dW_{2,t} = \rho dt$ ,  $\rho \in [-1, 1]$ , and we have defined  $\mathcal{E}(X) := e^{X - \frac{1}{2}E(X^2)}$  for a general mean-zero Gaussian variable  $X$ . We price options with Monte Carlo and to that end simulate  $V_t$  with the hybrid scheme of [12] and use the conditional estimator of [52] as a base estimate. We use a log-Euler scheme to simulate the part of  $S_t$  that is not integrated out by the conditioning and add the expectation of that as a control variable to our price estimator. We use 50 000 paths, half of which are antithetic.

### 4.2.3 Extended rough Bergomi

The rough Bergomi model can be extended so the volatility process is driven by two factors with different roughness levels: We define the extended rough (or fractional) Bergomi model by<sup>4</sup>

$$V_t = \xi_0(t)V_{1,t}V_{2,t}, \quad t \geq 0, \quad (4.5)$$

where

$$V_{1,t} = \mathcal{E} \left( \zeta \sqrt{2\alpha + 1} \int_0^t (t-s)^\alpha dW_{1,s} \right), \quad V_{2,t} = \mathcal{E} \left( \lambda \sqrt{2\beta + 1} \int_0^t (t-s)^\beta dW_{2,s} \right), \quad t \geq 0,$$

and  $\alpha, \beta \in (-\frac{1}{2}, \frac{1}{2})$ ,  $\zeta, \lambda \in \mathbb{R}$ . For this model, it is assumed that  $W_{1,t}$  and  $W_{2,t}$  are *independent*. When  $\alpha = \beta$  the model reduces to the ordinary rough (or fractional) Bergomi model with  $H = \alpha + \frac{1}{2}$  and

$$\rho = \frac{\zeta}{\sqrt{\zeta^2 + \lambda^2}}, \quad \eta = \sqrt{\zeta^2 + \lambda^2}. \quad (4.6)$$

We shall use the reparameterisation (4.6) also when  $\alpha \neq \beta$ . We may then translate back to the original formulation as  $\zeta = \eta\rho$  and  $\lambda = \eta\sqrt{1 - \rho^2}$ . We likewise use Monte Carlo to price options under (4.5). To this end, we simulate  $V_{1,t}$  and  $V_{2,t}$  as for rough Bergomi and use similar variance reduction techniques, except for the conditional Monte Carlo method which we could not carry over. We use anywhere between 25 000 and 100 000 paths depending on the shape of the initial forward variance curve—more paths when volatility is low as we found estimation difficult in those scenarios.

The extended model differs from the original in that the separation of the Hurst exponents  $\alpha + \frac{1}{2}$  and  $\beta + \frac{1}{2}$  between, respectively, the correlated and uncorrelated factor, allows for a corresponding separation of the explosion rates for the skew and curvature of the option smiles. To see why, note that

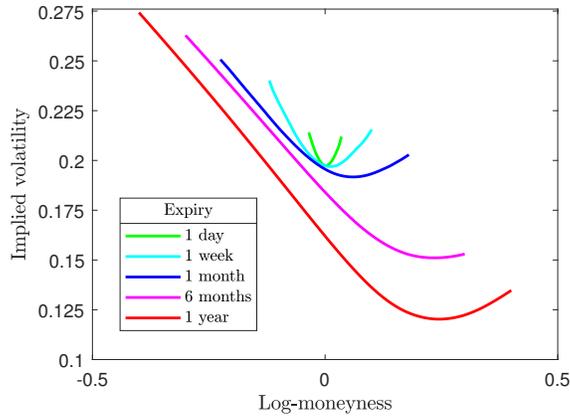
$$\text{Var}(\log V_{1,t}) = \zeta^2 t^{2\alpha+1}, \quad \text{Var}(\log V_{2,t}) = \lambda^2 t^{2\beta+1}, \quad t \geq 0,$$

so when  $\alpha > \beta$ , the correlated factor  $V_{1,t}$  dominates at long time horizons, the uncorrelated factor  $V_{2,t}$  at short horizons (vice-versa if  $\alpha < \beta$ ). As we illustrate in Figure 4.2.1 in the case of  $\alpha > \beta$ , the effect on option prices is as claimed. Compared to what the original model would produce, the option smiles are now more symmetric at short expiries and get increasingly skewed at longer horizons. The opposite would be observed if  $\alpha < \beta$ . The reader will see that the SPX volatility surface often have decoupled term structures for skew and curvature which is why we include the model.

## 4.3 Neural network approximations

We build neural network representations of our pricing models. We start by introducing neural networks, then review existing approaches on applying them to option pricing, and explain our own adaptation. Lastly, we train and validate our networks.

<sup>4</sup> The model is not our own invention as we were inspired by a no longer existing GitHub page by Ryan McCrickerd.



**Figure 4.2.1:** Volatility smiles under the extended rough Bergomi model with  $\alpha = 0.4$ ,  $\beta = -0.4$ ,  $\rho = -0.9$ ,  $\eta = 2.1$ ,  $\xi_0(t) = 0.20^2$ .

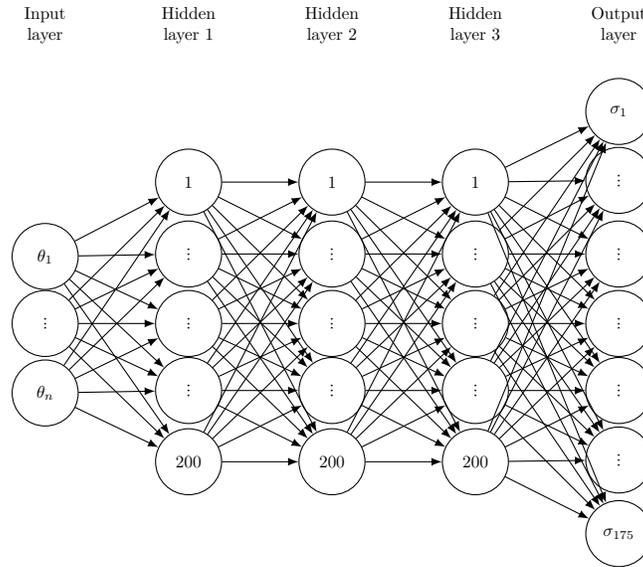
### 4.3.1 Neural networks

While neural networks are helpful for many purposes, we will introduce and motivate them in the context of approximating some multivariate function  $F : \mathbb{R}^N \rightarrow \mathbb{R}^M$ . We imagine that  $F$  is slow to evaluate and that we seek to replace it with a neural network that ideally will be faster but still sufficiently accurate. To this end, we consider a fully connected feed-forward neural network (henceforth just a 'neural network'). This is a mapping  $F_{\text{net}} : \mathbb{R}^N \rightarrow \mathbb{R}^M$  defined by its general architecture and a set of weights denoted  $w$ ; a network architecture is shown in Figure 4.3.1. To evaluate  $F_{\text{net}}$ , one traverses from left to right in the graph, and at each node, the input layer excluded, performs a computation of the form  $x \mapsto \sigma(a^\top x + b)$  where  $x$  is the output from the last layer (a column vector),  $a$  a coefficient column vector,  $b$  a scalar, and  $\sigma$  an (activation) function. The weights  $w$  are the collection of all coefficients  $a$  and  $b$ . The problem of finding a neural network approximation to  $F$  consists of finding an architecture and weights  $w$  so the error between  $F_{\text{net}}$  and  $F$  is small.

The weights are typically chosen as follows for a fixed architecture: First we generate a synthetic dataset of input-output pairs by evaluating  $F$  at various inputs  $x$  and storing the results  $y = F(x)$ . The inputs are often sampled from some distribution covering the relevant domain of  $F$ . For a random input sample  $X$ , we define the *generalisation error* by the number

$$E[\mathcal{L}(F_{\text{net}}(X; w), F(X))] \quad (4.7)$$

where  $\mathcal{L} : \mathbb{R}^M \times \mathbb{R}^M \rightarrow \mathbb{R}$  is a loss function and  $F_{\text{net}}(X; w)$  denotes the neural network evaluated in  $X$  with weights  $w$ . An empirical version of (4.7) can be constructed on the training data by averaging errors across the samples. To train the network, one can minimise the empirical version with respect to  $w$ . This is usually performed with stochastic gradient descent which is a gradient-based optimisation method. For the method one iterates in epochs, i.e. cycles, across the training data, first shuffling all samples, then



**Figure 4.3.1:** Our neural network design for implied volatilities on expiry slice in the interval  $[0, 0.008]$  for a model with  $n$  inputs parameters. Parameters are denoted  $\theta_i$  for  $i = 1, \dots, n$  and implied volatilities  $\sigma_i$  for  $i = 1, \dots, 175$ .

looping over smaller batches, covering all of them, each time updating the weights with a gradient estimate. We refer to [15, 35] for the finer details in the training of neural networks.

The literature contains many results that justify neural networks for function approximation: The Universal Approximation Theorem of [41] states (in essence) that a continuous function can be approximated to any desired precision on a compact domain using networks of a bounded number of layers (depth) but arbitrarily many neurons per layer (width). The paper [46] contains a dual result for networks of bounded width but arbitrary depth. A common observation in an option pricing context is nevertheless that only a few hidden layers are worthwhile. The authors of [9] e.g. report that their approximations did not consistently improve when going beyond four hidden layers. This is likely because the mapping from parameters to prices tends to be smooth so there is little benefit in introducing complex non-linearities beyond that achievable with a few layers.

### 4.3.2 Neural networks for options pricing

In the following, we formulate the approximation problem in relation to options pricing, calibration in particular, and review some existing approaches. We will frame the situation very abstractly. We therefore consider a general pricing model with some parameter space  $\Theta \subset \mathbb{R}^{d_1}$ ,  $d_1 \in \mathbb{N}$ , and let  $\Lambda \subset \mathbb{R}^{d_2}$ ,  $d_2 \in \mathbb{N}$ , denote a space of contract parameters

such as strike-expiry pairs for a number  $N_c \in \mathbb{N}$  of financial contracts. We consider a pricing function  $P : \Theta \times \Lambda \rightarrow \mathbb{R}^{N_c}$  that maps from model and contract parameters to a vector of prices which could be expressed in currency units or implied volatility. It is precisely such a function  $P$  that we wish to approximate with a neural network.

*Example:* Under Heston we may set  $\Theta = \mathbb{R}_+^4 \times [-1, 1]$  and for a vector  $\theta \in \Theta$  write  $\theta = (V_0, \kappa, v_\infty, \eta, \rho)$  in our previous notation. If we consider European calls and puts we could set  $\Lambda = \mathbb{R}_+^2$  and let  $\lambda = (K, T) \in \Lambda$  denote the strike and expiry of a given option. Then  $P(\theta, \lambda) = P(V_0, \kappa, v_\infty, \eta, \rho, K, T)$  would be its price. Alternatively, we could consider a fixed set of (say)  $N_c = 10$  options specified as strike-expiry pairs  $(K_i, T_i) \in \mathbb{R}_+^2$ ,  $i = 1, \dots, 10$ , set  $\Lambda = \emptyset$ , and let  $P : \Theta \rightarrow \mathbb{R}^{10}$  return a vector with their prices.

The setup and formulation of  $P$  can have important implications for the approximations and whether interpolation or extrapolation is needed on top. In the following we therefore discuss advantages and disadvantages of different neural network setups and definitions of  $P$ . All cases reviewed concern the pricing of European call and put options and  $P$  always returns prices in implied volatility; the same is true in our own setup.

We comment first on [9] where a *pointwise* learning approach is used. In the paper, the authors set  $\Lambda = \mathbb{R}_+^2$  with an element of the space referring to the strike and expiry of a single European call or put option. Their pricing function  $P : \Theta \times \Lambda \rightarrow \mathbb{R}$  returns the implied volatility of such a single option. They use a network with 4 hidden layers, 4096 neurons in each, which gives roughly 67 million weights. To focus the accuracy of the network on the relevant parts of the contract space, they sample strikes and expiries for the training data using a joint distribution constructed from observed SPX contracts.

In [43] an *image-based implicit* learning approach is suggested. For their numerical examples, the authors set  $\Lambda = \emptyset$  and let  $P : \Theta \rightarrow \mathbb{R}^{88}$  be the pricing function that returns implied volatilities for options on the 11 by 8 strike-expiry grid defined by the Cartesian product below (it is here assumed that  $S_0 = 1$ ):

$$\{0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3, 1.4, 1.5\} \times \{0.1, 0.3, 0.6, 0.9, 1.2, 1.5, 1.8, 2.0\}. \quad (4.8)$$

They use a network with 4 hidden layers, 30 neurons in each, resulting in about 7000 network weights. To price contracts that are not in the grid, appropriate interpolation and extrapolation can be used. The approach, of course, generalises to other contract grids. The method is called *image-based* as the network learns the mapping from model parameters to several points on the volatility surface ( $N_c > 1$ ). The word *implicit* refers to the fact that the pricing function implicitly depends on the chosen contracts ( $\Lambda = \emptyset$ ).

The authors of [43] argue that their method has a number of advantages. Let us point out a few as we see it: (1) The complexity is reduced as we only need to learn the volatility surface on a fixed set of points. This should facilitate the use of smaller networks. (2) While training, information on nearby contracts are considered jointly which allows the weights to adjust to the entire surface at once. This can speed up the training process. (3) The reduction in the input dimension and the increased information contained in an image, implies fewer training samples are needed. This is advantageous when combined with Monte Carlo as one can reuse simulated paths for multiple contracts in the grid.

A downside (compared to the pointwise method) is the need for interpolation which adds more errors and computational costs. However, in a calibration setting, the contracts to evaluate remain fixed as points in  $\mathbb{R}_+^2$  while the calibration runs. Consequently, we have found it possible to perform a number of pre-computations to speed up repeated evaluation of the interpolation scheme.<sup>5</sup> Because of this, and since we neither had problems controlling the interpolation error, we decided to use the image-based approach ourselves, although with a few adjustments due to a problem that we now explain.

More precisely, we find that it can be problematic to use a Cartesian product of contracts in the strike-expiry space such as (4.8). To explain the problem, say for example that we estimate  $P$  with Monte Carlo. We then argue that one should only include contracts in the grid that, with an acceptable amount of paths, can be accurately estimated. Indeed, if samples with significant errors enter into the training dataset this will likely have a negative influence on the accuracy and reliability of the trained network. One should therefore avoid options that are too far out-of-the-money. However, what is 'far out-of-the-money' is highly dependent on the strike, expiry, and model parameters. Consequently, with a grid in the form of a Cartesian product, one is forced to use relatively narrow strike bounds to control the Monte Carlo error across expiries and parameters.

One solution is to use a dynamic strike range. That is, we could look for functions  $\bar{K}_{min}, \bar{K}_{max} : \Theta \times \mathbb{R}_+ \rightarrow \mathbb{R}$  that satisfy  $\bar{K}_{min}(\theta, T) < \bar{K}_{max}(\theta, T)$  for  $(\theta, T) \in \Theta \times \mathbb{R}_+$  and which should specify reasonable strike bounds where  $\theta$  is a parameter vector,  $T$  an expiry. Such a technique is used in [53] where neural networks are trained for the SABR model of [39]. We though foresee several challenges: (1) The strike bounds of [53] use intricate properties of the SABR model and it is not obvious how to generalise them to other models. (2) The calibration task requires us to price all observed contracts, possibly after filtering, for any model parameter in the domain of optimisation. This is not in general possible with a dynamic moneyness range. That is, unless a robust extrapolation technique can be used, but we do not believe this is a trivial matter.

To balance the advantages and disadvantages, we choose a middle-of-the-road solution by using the image-based method, and that with dynamic moneyness bounds, but only allowing the bounds to depend on the expiry as  $\bar{K}_{min}(T)$  and  $\bar{K}_{max}(T)$ . In this way we can pre-filter the market data to the support of our networks, and yet, to some degree, accommodate how the probability distribution of  $S_T$  changes across the expiries  $T$ .

We show our choice of bounds in Figure 4.3.2. To specify the contracts for the image-based method, we fix the below 64 expiries

$$\mathcal{T} := \{0.002, 0.003, \dots, 0.01, 0.015, \dots, 0.05, 0.06, \dots, 0.20, 0.225, 0.25, 0.30, \dots, 0.5, 0.6, \dots, 3\}$$

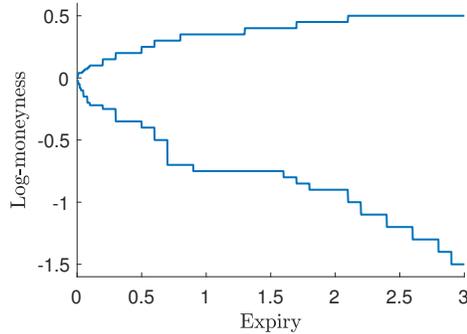
and use 25 uniformly spaced points in the log-moneyness dimension between the bounds. This results in 1600 contracts.

In [53] it was suggested to partition the volatility surface into expiry groups and to

---

<sup>5</sup> We were at least able to do so with our choice of interpolation method which we explain in a moment.

### 4.3.3. THE FORWARD VARIANCE CURVE



**Figure 4.3.2:** Moneyness region supported by our neural networks.

train separate networks for each. We ourselves find that this can help improve the approximations and keep the networks of a manageable size. A likely reason is that short- and long-term volatility smiles behave differently, especially under rough volatility, which suggests that they should be handled separately. We train six networks per model, each of which covers expiries in  $[0,0.008]$ ,  $(0.008,0.03]$ ,  $(0.03,0.12]$ ,  $(0.12,0.4]$ ,  $(0.4,1]$  and  $(1,3]$ .

Lastly, we outline our interpolation method: Say that we wish to interpolate the implied volatility of a contract with log-moneyness  $k$  and expiry  $T$  assumed within the bounds of Figure 4.3.2. We then interpolate as follows:

- Find the closest expiries  $T_1, T_2 \in \mathcal{T}$  so  $T_1 \leq T \leq T_2$
- Interpolate with a natural cubic spline the volatility smiles for the expiries  $T_1$  and  $T_2$  to obtain  $\sigma_{BS}(k, T_1)$  and  $\sigma_{BS}(k, T_2)$
- Apply linear interpolation between  $\sigma_{BS}^2(k, T_1)T_1$  and  $\sigma_{BS}^2(k, T_2)T_2$  in the time dimension to obtain  $\sigma_{BS}^2(k, T)T$  and thereby  $\sigma_{BS}(k, T)$

The natural cubic spline interpolation does not rule out static arbitrage. However, as follows by the work of [56], a necessary condition for no static arbitrage is that  $T \mapsto \sigma_{BS}^2(k, T)T$  is non-decreasing. Thus if implied volatilities are arbitrage free at the points  $(k, T_1)$  and  $(k, T_2)$  our choice of time-interpolation ensures that this is not violated.

An unreported numerical test showed that our interpolation method mostly ensures an absolute relative error in implied volatility below  $10^{-3}$ .

### 4.3.3 The forward variance curve

To handle the forward variance curve numerically, we assume it piecewise constant between the below 28 time points:

$$\{0, 0.0025, 0.005, \dots, 0.02, 0.04, \dots, 0.16, 0.28, \dots, 1, 1.25, 1.5, 1.75, 2, 3\}. \quad (4.9)$$

The grid roughly corresponds to daily sections for the first week, then weekly sections till the first two months, monthly sections for the remainder of the first year, quarterly

sections for the second year, and one section for the third year. We will sometimes abuse notation and denote the values by  $(\xi_1, \dots, \xi_{27})$ .

A few comments are in order on how to sample realistic curves for the training (and test) datasets: In [43], the curve  $\xi_0$  was assumed piecewise constant with 8 flat sections and each were sampled i.i.d. uniformly distributed. However, with the fine grid that we use this approach need not adequately cover the space of possible curves. Consider e.g. a time horizon  $T > 0$  and let  $\xi_0(\cdot)$  be piecewise constant between the time points  $t_i = \frac{T}{n}i$ ,  $i = 0, 1, \dots, n$  where  $n \in \mathbb{N}$ . If we sample each section i.i.d., the non-annualised variance swap quote with maturity  $T$ , here denoted  $VS(T)$ , converges to a fixed number by the law of large numbers, that is:

$$VS(T) := \int_0^T \xi_0(s) ds = \frac{T}{n} \sum_{i=1}^n \xi_0(t_i) \rightarrow TE(\xi_0(t_1)), \quad n \rightarrow \infty.$$

Indeed, the variability in the sampled variance swap has disappeared. To fix this problem we need to break the independence assumption when we sample adjacent sections.

Consider the Heston model where

$$VS(T) = v_\infty T + (V_0 - v_\infty) \frac{1 - e^{-\kappa T}}{\kappa}, \quad T \geq 0. \quad (4.10)$$

A possible solution is then to sample the Heston parameters  $\kappa, v_\infty, V_0$  and compute the piecewise constant forward variances as finite-difference derivatives on  $VS(T)$  in (4.10). To help the networks generalise to curves that are not based on the parametric form (4.10) one can add some smaller i.i.d. noise on top. We use this idea to sample most of our curves, although the technical details of our implementation are somewhat more involved. To add more variety, we also sample other types of curves, including i.i.d. ones as in [43]. The reader can consult the publicly available code for the details. We denote the distribution from which we sample  $(\xi_1, \dots, \xi_{27})$  by  $G_\xi$ .

### 4.3.4 Sampling the training (and test) data

In what follows, we explain how we sample and compute the underlying datasets. All prices are computed under the assumption of  $S_0 = 1$  and  $r(t) = q(t) = 0$ ,  $t \geq 0$ , which, as noted, suffices. For each model, we generated 136 000 samples for the training dataset and an additional 24 000 samples for a test dataset that we use to validate the final approximation error. This is double the amounts used in [43] for the rough Bergomi model. To ensure high-quality datasets, we discarded samples where option prices were not convex across strikes, where Merton's tunnel was violated, or where numerical integration for the Heston model failed to converge. For each discarded sample we generated a replacement. In Table 4.3.1 we show our choice of sampling distributions (distr.) for all models and parameters (par.). By  $\mathcal{U}$  we denote the uniform distribution. Columns  $a$  and  $b$  signify the sampling bounds.<sup>6</sup> For the rough Bergomi models we found

<sup>6</sup> The lower volatility bound is set to 7.5% for the extended rough Bergomi model and to 5% for the others. The reason is that despite using 100 000 paths in the case of low volatility, we for the extended model found it particularly difficult to estimate prices with sufficient accuracy for volatilities below 7.5%. If we did not adjust the sampling bound we would have discarded too many samples to finish the datasets in a reasonable amount of time.

### 4.3.5. HARDWARE AND SOFTWARE

that we could sample  $H$ , respectively,  $\beta$ , from an approximately uniform distribution and thereby roughly halve the total computation time. We write  $\mathcal{U}^*$  for this sampling method, though, the reader will again have to consult the publicly available code for the details. We sampled all parameters, except for the forward variances, with the Sobol sequence. As quasi-random numbers are more regularly spaced than pseudo-random numbers, the generalisation error (4.7) should then be better estimated by its empirical version. We expect this to lower the tendency of overfitting to the training data.

**Table 4.3.1:** Parameter distributions.

Heston				Rough Bergomi				Extended rough Bergomi			
Par.	Distr.	$a$	$b$	Par.	Distr.	$a$	$b$	Par.	Distr.	$a$	$b$
$\kappa$	$\mathcal{U}$	0	25	$H$	$\mathcal{U}^*$	0	0.5	$\alpha$	$\mathcal{U}$	-0.5	0.5
$\eta$	$\mathcal{U}$	0	10	$\eta$	$\mathcal{U}$	0.75	3.5	$\beta$	$\mathcal{U}^*$	-0.5	0.5
$\rho$	$\mathcal{U}$	-1	0	$\rho$	$\mathcal{U}$	-1	0	$\eta$	$\mathcal{U}$	0.75	3.5
$\sqrt{v_0}$	$\mathcal{U}$	0.05	1	$\sqrt{\xi}$	$G_\xi$	0.05	1	$\rho$	$\mathcal{U}$	-1	0
$\sqrt{v_\infty}$	$\mathcal{U}$	0.05	1					$\sqrt{\xi}$	$G_\xi$	0.075	1

### 4.3.5 Hardware and software

The training and test datasets were computed on a server running Intel Xeon Platinum 8175 3.1 GHz CPUs with 384 GB RAM and 48 physical cores (96 logical). It took around one week to finish them all. The neural networks were thereafter trained in Python 3.7.1 with the Keras 2.2.4 library; we used TensorFlow 1.13.1 as the backend. Computational times reported in the remainder of the paper are recorded in Matlab 2019a on a laptop running a 1.6-3.4 GHz Intel Core i5 8250U CPU with 4 cores (8 logical processors) and 8 GB RAM.

### 4.3.6 Hyperparameters

By hyperparameter optimisation, we refer to the act of minimising the generalisation error by tuning all other aspects of a neural network and its training than the weights themselves. Examples are the number of layers, the number of neurons per layer, the choice of optimiser and its settings, the batch size, and the choice of activation and loss function. Hyperparameter optimisation is very costly as it requires one to retrain the network across many different configurations. We therefore fixed most hyperparameters based on common choices found in the literature that generally appear to work well: We choose 3 hidden layers as used in a preprint version of [43], the Adam algorithm of [47] as our optimiser, Elu as our activation function except for a linear output layer, and root-mean-squared-error as the loss function. We scale the inputs and outputs as in [43] and use 200 neurons per layer which we found to be sufficient. For the training, we first let the optimiser run for 500 epochs with a batch size of 32 at which point both the training and test losses stabilised. However, by increasing the batch size to 5000 and continuing for another 200 epochs, we found that we could reduce the error by an

additional moderate amount.<sup>7</sup> The final errors on the test datasets turned out similar to those on the training datasets, which indicates that overfitting was not a problem.

### 4.3.7 Accuracy and speed

We examine the accuracy and speed of our networks. We shall here measure accuracy in terms of implied volatility. In Table 4.3.2 we show percentiles of the absolute relative errors between the trained networks and the test datasets. We note that the errors, generally, are below a few percent and mostly within a single percent. In Table 4.3.3, we, for the rough Bergomi models, show the relative standard errors; by this we refer to the standard errors of implied volatility divided by the estimated implied volatility. We see that also these errors are comfortably low. The errors versus the test datasets are generally larger than the standard errors which is unsurprising. On our filtered SPX dataset, the 25th and 75th percentiles of the relative bid-ask spreads are 1.36% and 4.25%; we here refer to the size of the implied volatility bid-ask spread divided by the implied volatility of the mid quote. The approximations are mostly within the bid-ask spreads and we are therefore confident in using our networks for the calibrations.

**Table 4.3.2:** Distribution of the absolute relative errors between the trained networks and the test datasets measured in implied volatility. Numbers in the header represent percentiles.

Model	50th	75th	95th	99th	99.9th	99.99th	Max.
Heston	0.09%	0.16%	0.51%	1.40%	4.97%	15.63%	219.82%
Rough Bergomi	0.18%	0.37%	1.03%	2.44%	7.52%	18.19%	165.17%
Extended rough Bergomi	0.40%	0.70%	1.38%	2.47%	5.69%	11.17%	51.18%

**Table 4.3.3:** Distribution of the relative standard errors on the training datasets measured in implied volatility. Numbers in the header represent percentiles.

Model	50th	75th	95th	99th	99.9th	99.99th	Max.
Rough Bergomi	0.13%	0.23%	0.44%	0.85%	2.06%	3.98%	278.51%
Extended rough Bergomi	0.51%	0.60%	0.72%	1.14%	2.47%	4.63%	256.49%

In terms of speed, we find that we can evaluate all networks once in about  $7.1 \cdot 10^{-4}$  seconds (average of 10 000 runs under the rough Bergomi model). When we calibrate, we however need to compute prices many times on a scattered set of points anywhere in the contract domain. This will cost us more time as we will need to interpolate. As noted, the points are though fixed and we therefore find it possible to perform a number of pre-computations to speed up repeated evaluation of the interpolation scheme. If we fix the

---

<sup>7</sup> We hypothesise that a batch size of 32 may allow for fast convergence initially but as one nears a minimum the gradient estimates may be too noisy for full convergence.

4166 contracts we have available on May 15, 2019, perform the pre-computations, and evaluate the neural networks and the interpolation scheme 10 000 times, we obtain an average running time of  $1.8 \cdot 10^{-3}$  seconds per evaluation (also based on rough Bergomi). The cost of interpolation is therefore small when spread across many evaluations.

## 4.4 Historical calibration to SPX options

In what follows, we present the historical calibrations to SPX options. We use a dataset obtained from <https://datashop.cboe.com> which consists of bid and ask quotes on SPX European call and put options on trading days between May 3, 2004, and May 15, 2019. The quotes are recorded at 15:45 Eastern time and we apply a number of filters to remove unreasonable and low liquidity quotes and to ensure a minimum number of strikes and expiries on each date. Interest rates and dividends are implied from the available put-call parities. In-the-money contracts and those outside the domain of Figure 4.3.2 are removed. After filtering, the dataset consists of 4 643 880 bid-ask pairs spread across 3775 trading days which corresponds to an average of 1230 observations per day.

Consider a given trading day and let  $\sigma_i^{bid}$ ,  $\sigma_i^{ask}$  and  $\sigma_i^{mid}$  denote the implied volatility of the bid, ask and mid quote for the  $i$ 'th observed option. Similarly, let  $\sigma_i^{model}(\theta)$  denote the implied volatility of the contract under a given model with parameter vector  $\theta$ . We define the weighted root-mean-square-error (wRMSE) by

$$\text{wRMSE}(\theta) := \sqrt{\sum_i w_i (\sigma_i^{mid} - \sigma_i^{model}(\theta))^2} \quad (4.11)$$

where the sum is over all contracts from the given date and  $w_i$  are weights that are normalised so  $\sum_i w_i = 1$ . We use the weights to make the error measure more robust. There is e.g. a considerable amount of time variation in the dataset in that the number of quoted expiries and their location changes a lot from 2004 to 2019. We therefore allocate 15 percent of the weight to expiries within 1 month, 35 percent to expiries between 1 and 6 months, and the rest to longer expiries. Additionally, we normalise so all expiries within the same expiry group carry the same total weight. To account for differences in liquidity, we further, up to normalisation, weigh each contract by

$$\frac{1}{0.01 + \sigma_i^{ask} - \sigma_i^{bid}}.$$

To calibrate the Heston model, we minimise precisely the wRMSE. However, for the rough volatility models, we find that  $\xi = (\xi_1, \dots, \xi_{27})$  often is overparameterised which can result in calibrated curves that look unrealistic. The main culprit is that there may be multiple grid points for the curve between the observed expiries. As a solution, we use the below algorithm to merge sections before we calibrate:

**Merging forward variance curve sections:** Let  $0 = t_0 < t_1 < \dots < t_n$  be the grid points between which the initial forward variance curve is assumed flat. We loop as  $i = n - 1, \dots, 1$  and if there are no observed expiries in  $(t_i, t_{i+1}]$  merge  $\xi_i = \xi_{i+1}$ .

Unfortunately, even merging sections as above, we still find that the calibrated curves can look unrealistic. A possible explanation is noise coming from the bid-ask spread. We therefore additionally use penalisation when we calibrate. More precisely, we minimise for these models  $wRMSE(\theta)^2 + \lambda C(\xi)$  with respect to  $\theta$  where  $\lambda \in \mathbb{R}_+$  and  $C(\xi) := \sum_{i=2}^{27} (\sqrt{\xi_i} - \sqrt{\xi_{i-1}})^2$  measures the non-smoothness of  $\xi$ . We find that  $\lambda = 0.002$  works well without worsening the fits much. Calibrating the models across the entire dataset gives average differences of only 4-5 bps (basis points) in wRMSE between  $\lambda = 0.002$  and  $\lambda = 0$  and we are therefore comfortable using  $\lambda = 0.002$  for our analysis.<sup>8</sup>

#### 4.4.1 Calibration results

We show the calibration errors and model parameters in Figures 4.4.1-3. For brevity, we have left out most of the Heston parameters. Daily closing values for the VIX index are provided for reference in the bottom of Figure 4.4.1.<sup>9</sup> The reader should note that for a model of the form (4.1), the VIX index squared is essentially<sup>10</sup> defined by

$$VIX_t^2 = 100^2 \left[ \frac{1}{\Delta} \int_t^{t+\Delta} \xi_t(u) du \right], \quad t \geq 0, \quad \Delta = \frac{1}{12}, \quad (4.12)$$

and thus, up to scale, represents the risk-neutral expected average spot variance over the next month.

We start with the calibration errors which are shown in Figure 4.4.1 as centralised 20-day moving averages. We note first that the rough volatility models perform much better than classical Heston. While the average error is around 80 bps for Heston, it is close to 50 bps for the rough Bergomi models. The error for Heston also deteriorates rapidly around the financial crisis of 2008-2009, remains elevated for several years thereafter, and occasionally reaches values as high as 160 bps. In contrast, errors for the rough models mostly stay below 80 bps and appear less sensitive to the volatility level. In terms of the latter, note that we for these models observe average errors in the ranges 60-61, 53-55 and 48-50 bps when the VIX index, respectively, is in the ranges 0-15, 15-25 and above 25. Under Heston the average errors are 69, 80 and 100 bps for the same VIX levels. It follows also that the *relative* errors for, at minimum, the rough models generally are larger when volatility is low. Evidence from sections 4.4.3 and 4.5 suggest that the volatility surface displays an increasingly complex structure at lower volatility levels. We believe this, at least partly, explains it. Another possible explanation is that we, due to our fixed moneyness bounds, can expect fewer contracts to be filtered when volatility is low.

While rough volatility clearly pays off looking at the calibration error, there are no notable differences overall between rough Bergomi and its extended version. The average difference of their calibration errors is only 1 bps. The standard deviation of the

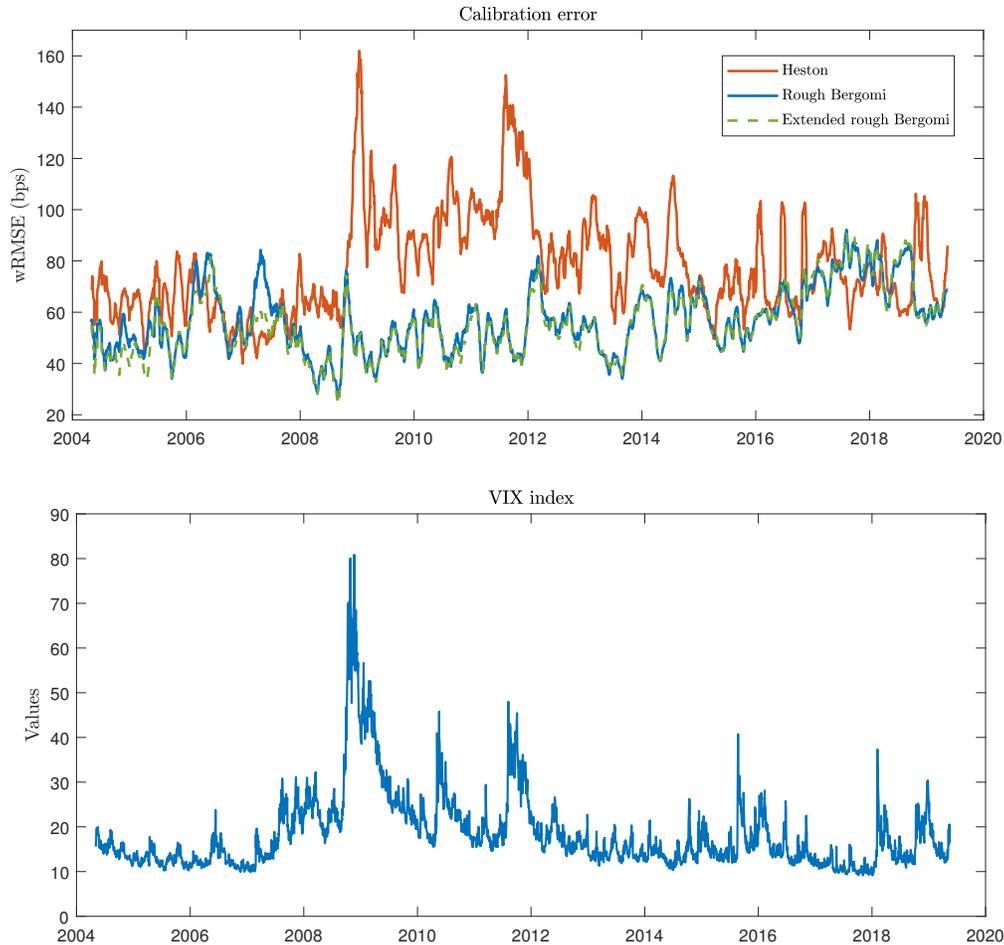
<sup>8</sup> All optimisations are performed with Matlab's `lsqnonlin` function and the *trust-region-reflective* algorithm.

<sup>9</sup> A single observation is missing from our VIX index dataset and is therefore excluded from any figures and calculations; the dataset is downloaded from [https://www.cboe.com/tradable\\_products/vix/vix\\_historical\\_data](https://www.cboe.com/tradable_products/vix/vix_historical_data) on May 30, 2021.

<sup>10</sup> See e.g. [34].

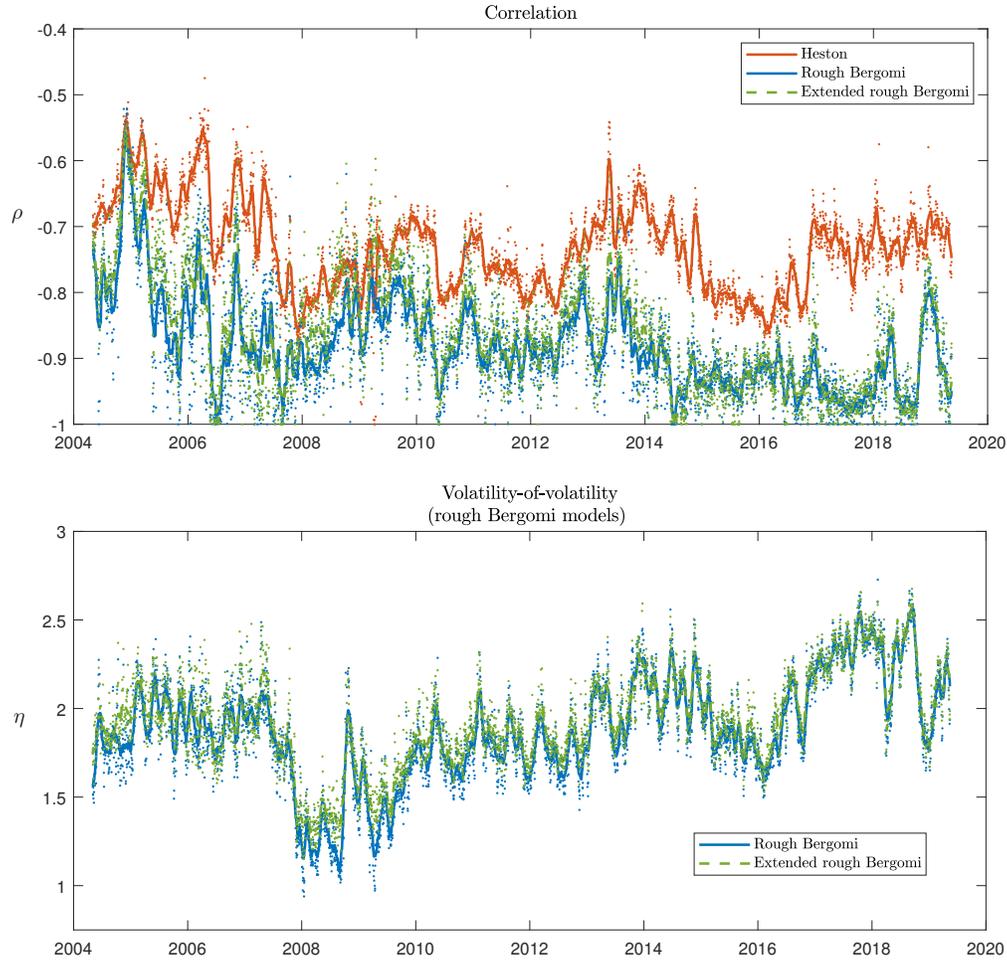
#### 4.4.1. CALIBRATION RESULTS

daily differences is 5 bps. This is also reflected in the  $\eta$  and  $\rho$  parameters that are almost equal between the models on most dates; as we will discuss, the models are effectively the same as calibrated. It therefore appears that there are no worthwhile gains to be made in separating the Hurst exponents between the two independent Brownian motions that drive volatility. This, however, does not mean that the fits are perfect or that one cannot construct better models. Additional results presented later will make this clear.



**Figure 4.4.1:** Calibration errors shown as 20-day centralised moving averages and VIX closing values.

We see some preliminary evidence of problems with the rough Bergomi models in Figure 4.4.2 (top) where we show the correlation parameters. Indeed, while  $\rho$  for Heston fluctuate mostly in the rather narrow range from  $-0.8$  to  $-0.6$  and stay a good distance away from the lower bound of  $-1$ , the parameter for the rough Bergomi models is more

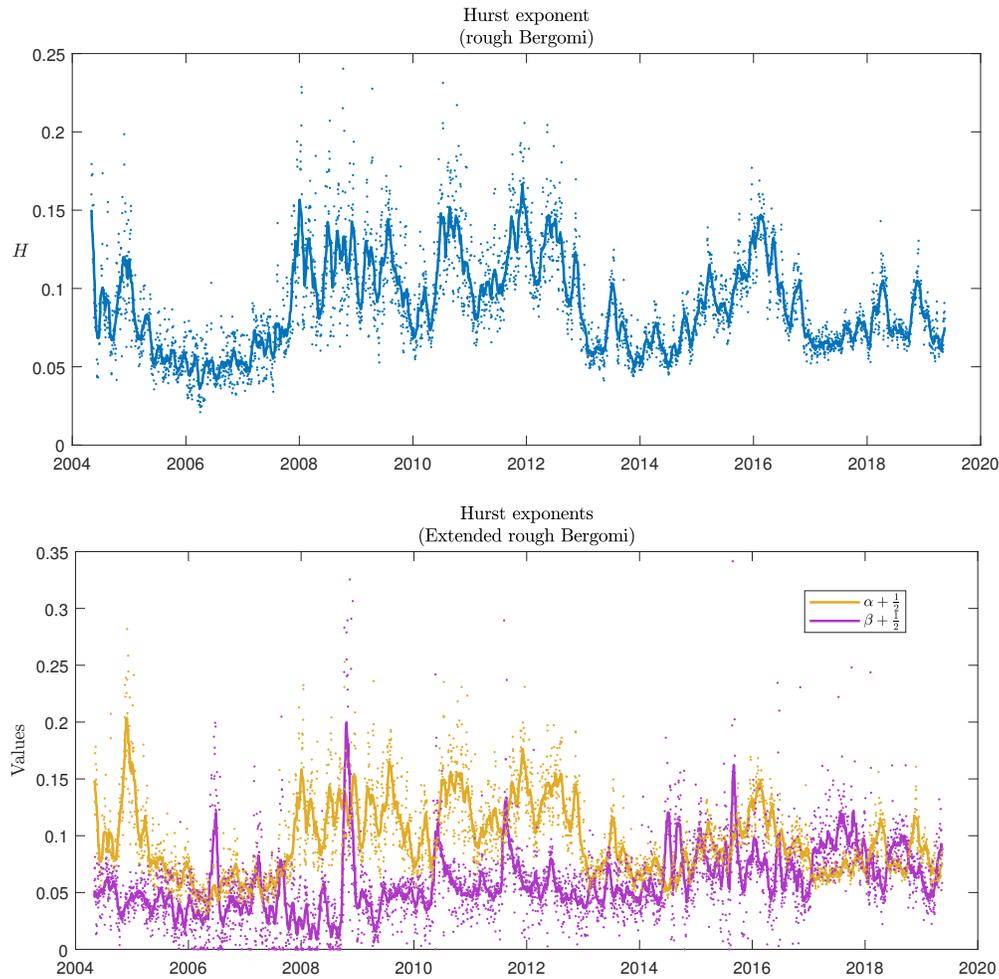


**Figure 4.4.2:** Calibrated correlation and volatility-of-volatility parameters. Solid lines show centralised 20-day moving averages. Dots show individual observations.

unstable and sometimes calibrates at or near  $-1$ . In terms of the instability, we can report that the standard deviation of daily changes in  $\rho$  is 0.0218 for Heston while it is 0.045 for both rough Bergomi models. In relation to the level, we believe extreme values such as  $\rho \approx -1$  suggest unrealistic dynamics for the S&P 500 index as it implies that the entire volatility surface is driven by essentially a single factor. That  $\rho$  sometimes is forced near its lower boundary could precisely indicate that the models lack flexibility.

We turn our attention to the Hurst exponent  $H$  under rough Bergomi; see Figure 4.4.3 (top). With values below 0.25 across the entire time series and an average of 0.09, it is clear that option prices, as viewed through the model, consistently suggest that volatility is very rough. Interestingly enough, the Hurst parameter correlates positively with the level of volatility; the correlation between  $H$  and VIX is 0.49. That volatility generally is less rough in periods of high volatility is a finding that also appears in [10] under the realized measure. Because  $H$  controls the speed at which volatility's dependence on the history of  $W_{2,t}$  dissipates, lower  $H$  implying a faster dissipation, it suggests

#### 4.4.1. CALIBRATION RESULTS



**Figure 4.4.3:** Calibrated Hurst exponents. Solid lines show centralised 20-day moving averages. Dots show individual observations; a few are left out to improve visibility.

that quoted SPX options imply a volatility process with weaker mean reversion when volatility is high. We believe this is meaningful from an economic point of view: typical high volatility events, think e.g. of the financial crisis of 2008-2009, arguably tend to influence markets for longer than events that correspond to smaller volatility increases.

The calibrated  $H$ -values could also be interpreted as a reflection of the volatility smoothness implied by the market. Note by [4, Example 2.3, Lemma 2.4] that the volatility sample paths under rough Bergomi are piecewise Hölder continuous of all orders less than  $H$ ; we say *piecewise* continuous because we have assumed  $\xi_0$  piecewise constant. This is a consequence of the rate of singular decay of the kernel function, or, in other words

the speed of 'mean-reversion' across very small time-scales.<sup>11</sup> It is therefore possible that a deeper explanation for the level dependence could be found by examining market microstructure models. We for example note that [21] shows a connection between certain features of high frequency trading and roughness. A critique of interpreting the calibrated Hurst exponents as representing the implied smoothness of volatility is that  $H$  for the fractional kernel is responsible for 'mean-reversion' at both short and long time scales. Therefore, unless prices are truly generated by a model with exactly the fractional kernel, the calibrated values are bound to reflect an imperfect trade-off between the mean-reversion implied at short and long time horizons. In fact, in [10] it is suggested to decouple the short and long time behaviour of the volatility autocorrelations.

We consider now the Hurst exponents under the extended rough Bergomi model: Note that the time series of  $\alpha + 1/2$  closely resembles that of  $H$  for rough Bergomi while that of  $\beta + \frac{1}{2}$  is somewhat different. The first observation is consistent with the fact that our calibrated rough Bergomi models are very similar. More precisely, for  $\rho$  near  $-1$ , as we observe,  $\lambda = \eta\sqrt{1 - \rho^2}$  is small in absolute value compared to  $\zeta = \eta\rho$  and thus, as calibrated, the extended model is very similar to the ordinary one but with Hurst exponent  $\alpha + \frac{1}{2}$ . It follows, that we should also be careful in drawing conclusions based on the time series of  $\beta + \frac{1}{2}$  as the effect of  $\beta$  on the extended model is small for such  $\rho$ 's. However, that we with relative consistency observe  $\alpha > \beta$  at least *indicates* that the short-term smiles, generally, are more symmetric than the long-term ones. This will be backed up by more robust evidence in Section 4.4.3. Also, as the reader will see in Section 4.5, there are better ways to decouple the term structures of skew and curvature.

Let us briefly comment on the volatility-of-volatility parameter  $\eta$  for the rough Bergomi models; see Figure 4.4.2 (bottom). Although the time evolutions looks somewhat stable—at least in comparison with the  $\rho$ -values—there is a systematic (negative) level-dependence with volatility: the correlation between  $\eta$  and VIX is respectively  $-0.47$  and  $-0.49$ . The same correlation is  $0.27$  for Heston. That the level-dependence should be somewhere between that of a square root and log-normal model is consistent with the conclusion of [57], though, the cited reference suggests that the truth is closer to the log-normal case.<sup>12</sup>

Lastly, we consider the calibration speed which we provide statistics on in Table 4.4.1. As can be observed, we may generally calibrate in less than a second on what can be considered a standard laptop as of the publication date. This shows that we, with neural networks, can implement rough volatility models in a realistic setting and obtain speeds that are feasible also in a real production environment.

#### 4.4.2 Predictive quality

We examine now the predictive quality of our models. We perform the following experiment: On each trading day and for each model we fix the calibrated parameters that are

<sup>11</sup> We will use the expression 'mean-reversion' liberally as e.g. rough Bergomi is not stationary. We shall use the word 'autocorrelation' with similar carelessness.

<sup>12</sup> That the correlation versus the VIX index is less (in absolute value) for Heston compared with the rough Bergomi models is likely because the  $\eta$  parameter is much more volatile under Heston. Indeed, the standard deviation of the daily percentage changes in  $\eta$  is  $0.23$  for Heston while it is  $0.05$  and  $0.06$  for the rough Bergomi models.

### 4.4.3. DECOUPLED TERM STRUCTURES OF SKEW AND CURVATURE

**Table 4.4.1:** Statistics on per day calibration times shown in seconds. Numbers in the header represent percentiles.

Model	Min.	5th	50th	95th	Max.
Heston	0.03	0.05	0.07	0.11	0.19
Rough Bergomi	0.08	0.16	0.37	1.01	3.24
Extended rough Bergomi	0.10	0.17	0.38	1.15	1.97

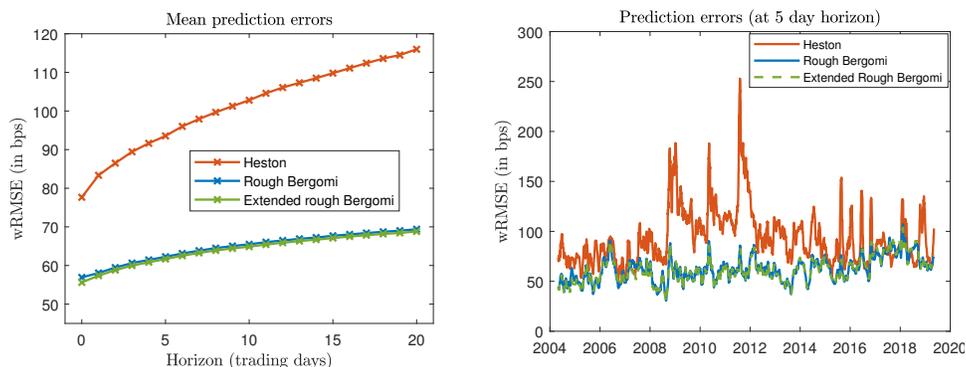
not state variables. We then recalibrate up to 20 trading days into the future allowing only the state variables to change. For Heston it means that we only recalibrate the instantaneous variance and for the rough volatility models it means that we only recalibrate the forward variance curve. Market variables such as the index price and yields are updated as observed on the given future day.

In Figure 4.4.4 (left), we show the mean errors at different horizons; standard errors are at or below 1 bps. The rough Bergomi models perform effectively the same which is hardly surprising given the in-sample calibrations. Compared to Heston their predictions are also more robust. For the rough Bergomi models, the error increases from about 55 bps in-sample to around 70 bps at the 20-day horizon. The Heston mean error grows from about 80 bps to almost 120 bps. The right part of Figure 4.4.4 shows the evolution of the 5-day horizon errors (as moving averages). The time series overall resembles the evolution of the in-sample errors, though, perhaps a bit more unstable, at least for the Heston model. The reason why the errors deteriorate at a faster rate for Heston in the left plot could possibly be due to the volatility-of-volatility level dependence being largely misspecified for that model whereas the log-normal case is more in line with reality; we again refer to [57]. One could also argue that it results from the volatility autocorrelation structure being better captured by rough volatility than classical (one-factor) volatility. However, we believe this is a less likely explanation since we allow the instantaneous variance, respectively the forward variance curve, to be freely recalibrated in our experiment. Of course, the differences in the overall error levels, the in-sample errors in particular, are likely for the most part precisely a consequence of this modelling difference.

### 4.4.3 Decoupled term structures of skew and curvature

In what follows, we analyse the in-sample calibrations in more detail. In Figure 4.4.5, we show the fits for a subset of the expiries on September 4, 2012, and January 13, 2017. The calibration errors on the first date are 64, 50, and 49 bps for, respectively, Heston, rough Bergomi and the extended rough Bergomi model, and, listed in the same order, are 70, 79, and 88 bps on the second date.<sup>13</sup> If we compare with the top of Figure 4.4.1, we see that, in terms of the errors, and for the rough volatility models, September 4, 2012, represents a typical example, whereas January 13, 2017, represents one of the more

<sup>13</sup> For the latter date, the ordinary rough Bergomi model performs slightly better than the extended version. Section 4.5 contains similar examples where nested models perform marginally better. We believe it can be explained by any combination of the following: smaller differences in the neural network approximations, Monte Carlo error, different initial guesses.



**Figure 4.4.4:** Left: Mean prediction errors at different horizons. Right: Prediction errors at the 5-day horizon (shown as 20-day centralised moving averages).

problematic dates. In Figure 4.4.6, we show the at-the-money skews on the market data. Power-law fits are added for comparison and we have used the definition:

$$\text{Skew}(T) := \left. \frac{\partial \sigma_{BS}(k, T)}{\partial k} \right|_{k=0}, \quad T > 0.$$

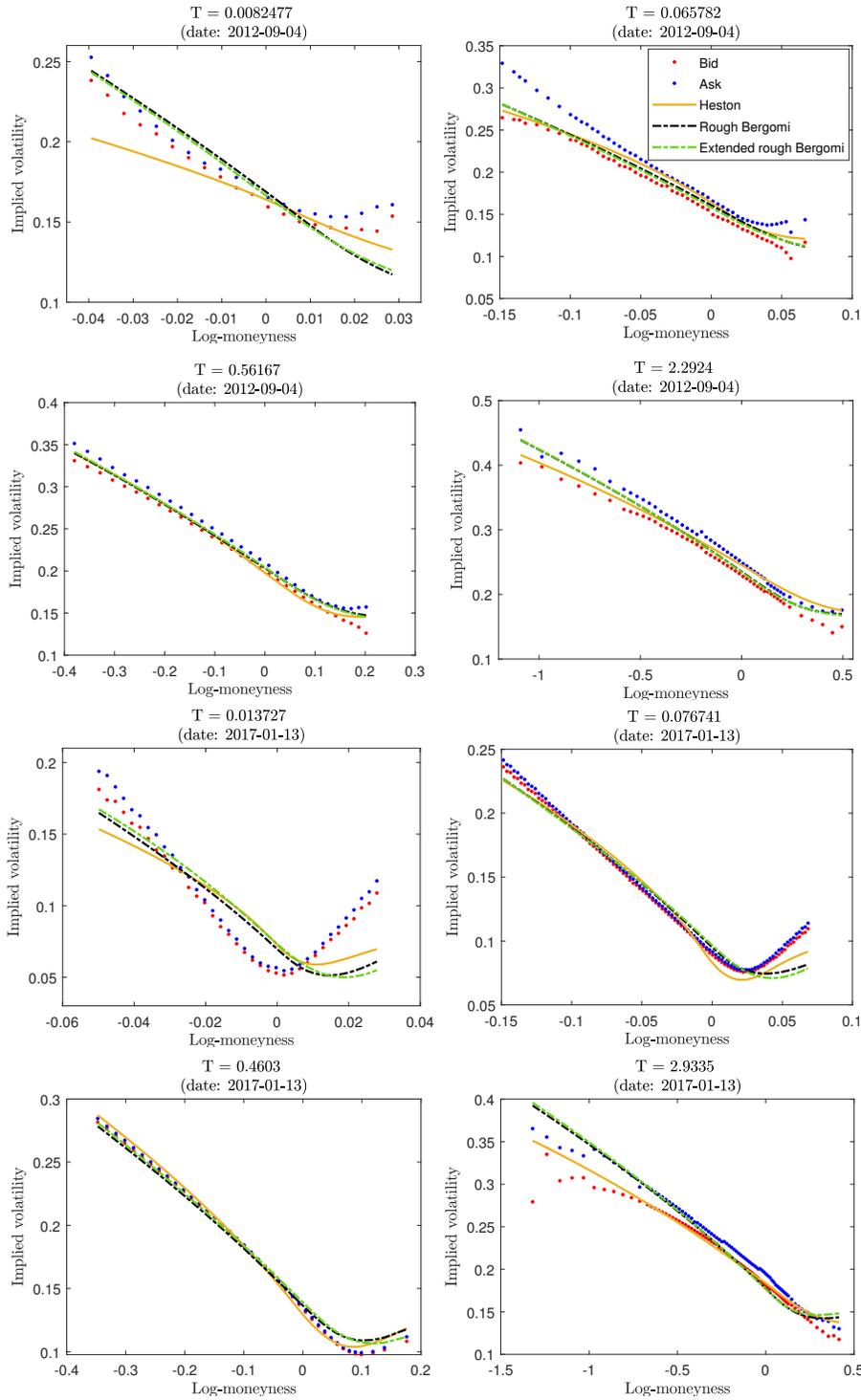
We consider first the fits on September 4, 2012, where the term structure of at-the-money skew is reasonably well represented by a power-law consistent with what typical one-factor rough volatility models display. The volatility smiles that we show in Figure 4.4.5 are also fitted relatively well by the rough Bergomi models on this date, though, improvements can still be made. The Heston model, however, cannot reproduce the power-law term structure of skew—see e.g. the functional form derived in [34, pp. 35]—and we believe this is why it results in a worse fit, in particular to the very short expiry.

In contrast, on January 13, 2017, even the rough models calibrate poorly. The main reason is that there on this date is a pronounced term structure of (a)symmetry as the short-term smiles are much more symmetric than the long-term ones. Unsurprisingly this feature is not reproduced by the one-factor rough Bergomi model, though, less obviously neither by its extended version. Figure 4.4.6 shows that the term structure of at-the-money skew also far from resembles a power-law. The short-term skews even flatten as the smiles become increasingly symmetric.<sup>14</sup> On a side-note, we shall remark that the very short-term smile on September 4, 2012, also is slightly more symmetric than what the rough Bergomi model(s) could produce. This hints at the possibility that the term structure of smile (a)symmetry is a structural feature of the volatility surface.

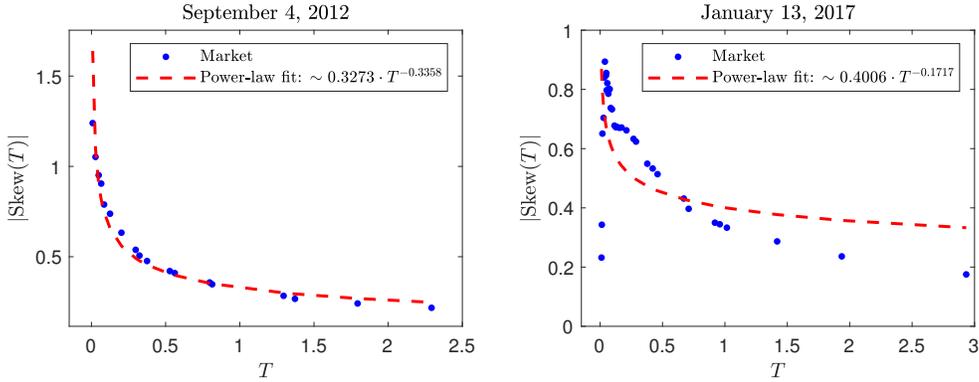
The calibration fits on January 13, 2017, especially, and the associated non-power-law term structure of observed skews demonstrate that typical one-factor rough volatility models are not always consistent with the SPX volatility surface. A likely explanation is that two volatility factors with different temporal properties and correlations to the index price are needed. This view is precisely the reason why we included the extended

<sup>14</sup> The non-exploding skew on January 13, 2017, though, need not imply that volatility is not rough (recall Figure 4.2.1).

### 4.4.3. DECOUPLED TERM STRUCTURES OF SKEW AND CURVATURE



**Figure 4.4.5:** Example calibrations from the historical SPX analysis. Only a subset of the expiries are shown for each date.



**Figure 4.4.6:** At-the-money skews. Market skews are extracted by fitting the SVI parameterisation of [30] to each volatility smile. Power-laws are fitted by log-linear regressions on the market skews.

rough Bergomi model in our analysis, though, as can be observed, it apparently is insufficient only to separate the Hurst exponents for each of the two independent underlying Brownian motions. The results of Section 4.5 will show that better fits can be achieved with proper two-factor volatility, i.e. using three Brownian motions to model  $(S_t, V_t)$ .

In the following, we provide a systematic investigation of the rough Bergomi fits across the entire dataset; we exclude the extended version from our analysis as it calibrates effectively the same. We do as follows on each trading day: First we fix the Hurst exponent  $H$  as already calibrated. Separately for each expiry we then recalibrate  $\rho$ ,  $\eta$ , and  $\xi_0$  which will be assumed flat. In this way we achieve almost perfect fits to each expiry slice. We can then use the recalibrated correlation and volatility-of-volatility parameters to study in closer detail how the model should be adjusted for a better fit.

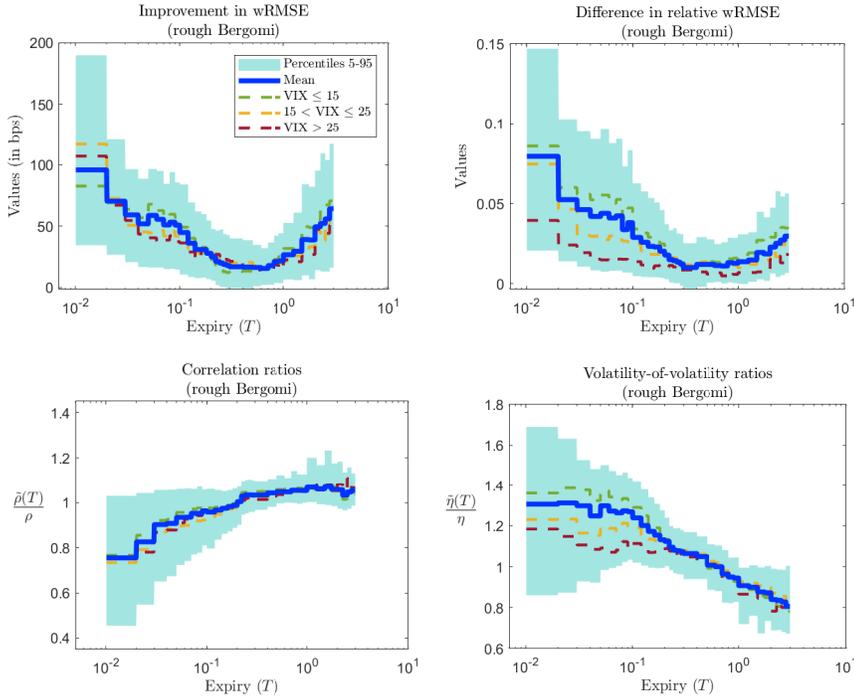
We present the results in Figure 4.4.7, where we for various quantities (to be introduced) show percentiles, means, and means split by the level of the VIX index. The expiries are placed into groups for the computations and the first axes are log-scaled. The plot in the top-left shows the wRMSE before recalibration minus that after.<sup>15</sup> We note that the recalibrations have generally lowered the errors by a good amount. The improvement is largest for short and long expiries which is unsurprising as a full calibration will seek to balance both ends. In the top-right, we show values for the relative wRMSE before recalibration minus that after. We have defined the relative wRMSE by:

$$\text{relative wRMSE}(\theta) = \sqrt{\sum_i w_i \left( \frac{\sigma_i^{mid} - \sigma_i^{model}(\theta)}{\sigma_i^{mid}} \right)^2}.$$

We note that the fits are improved by a good amount in relative terms also and we therefore believe that the recalibrations can meaningfully be used for our analysis.

<sup>15</sup> For a proper comparison, we always rescale the weights,  $w_i$ , so they sum to one for a given expiry.

#### 4.4.3. DECOUPLED TERM STRUCTURES OF SKEW AND CURVATURE



**Figure 4.4.7:** Expiry-by-expiry recalibrations. Shaded areas cover percentiles 5 to 95, solid lines show mean values, dashed lines mean values categorised by the level of the VIX index.

Let  $\rho$  be the correlation parameter from the full calibration on a given day and let  $\tilde{\rho}(T)$  denote the recalibrated version for an expiry  $T$ . Likewise let  $\eta$  denote the calibrated volatility-of-volatility parameter and let  $\tilde{\eta}(T)$  denote the recalibrated version. To structure our analysis, we will look at the ratios  $\tilde{\rho}(T)/\rho$  and  $\tilde{\eta}(T)/\eta$  to see in which direction we should shift the distribution of  $S_T$  for different time horizons  $T$ . The reader can find the results for the ratios in the bottom row of Figure 4.4.7. As can be observed, we, on average, need less correlation (more symmetry) and more volatility-of-volatility (more smile effect) for short expiries, vice versa at long expiries, and all relative to what the model could produce in the full calibration experiment. The observations are statistically significant as the standard errors of the means are all at or below 0.01.

The first observation shows that it *is* a structural feature that short expiry smiles are more symmetric. On the second observation, note that a faster decaying volatility autocorrelation function tends to increase the smile effect at short expiries *relative* to long expiries; see e.g. the expansion formula of [14].<sup>16</sup> A natural conclusion is then that

<sup>16</sup> That  $\text{Skew}(T) \approx cT^{H-\frac{1}{2}}$  for many one-factor rough models is an example of how 'mean-reversion' and thus, presumably, a stronger separation of short and long lag 'autocorrelations', here controlled by  $H$ , creates a more pronounced term structure of smile effect. As the January 13, 2017, smiles exemplify, we should though be careful to interpret the term structure of skew directly as a measure of the term structure of smile effect when there is also a significant term structure of (a)symmetry.

the autocorrelations implied by the calibrated rough Bergomi model fail to sufficiently separate the short and long time-scale properties. We can think of two possible reasons for this: Either the fractional kernel does not have enough flexibility to capture the implied autocorrelation structure, or it does but fails in our calibration experiment as a side-effect of the model's inability to reproduce the term structure of smile (a)symmetry. We believe the first explanation is most likely, though, both could hold simultaneously. The reason is that we in Section 4.5 calibrate a comparable one-factor rough Bergomi type model jointly to SPX and VIX options on a date where there is a negligent term structure of (a)symmetry and still conclude that the short and long time-scale properties are not adequately separated. At a glance, it is surprising that the fractional kernel, at least for the rough Bergomi model, supposedly lacks flexibility as it already has an apparently rich structure in that it induces mean-reversion at multiple time scales; we refer to the multifactor approximations of [1, 3] for the latter.<sup>17</sup> The results of Section 4.5 suggest that the volatility autocorrelation structure is better captured by a two-factor volatility model (even non-rough) or a quadratic rough Heston model.

The term structure of residual smile effect could also be a result of the log-normal distribution not being heavy-tailed enough to reproduce the short-term behaviour of volatility. The authors of [11] e.g. show that log-volatility is better described by a normal inverse Gaussian (NIG) distribution with heavier tails than a normal distribution. The results of Section 4.5 though show that, while there indeed does appear to be a distributional problem in some market scenarios in fitting to SPX options for the log-normal distribution, most of the residual calibration error can be eliminated by using a two-factor but yet log-normal model. We therefore believe that the poor SPX fits are mainly explainable by a combination of an inflexible volatility autocorrelation structure and the lack of a term structure for the smile (a)symmetry. The distributional problem is more apparent when we look at VIX options where log-normal models tend to produce almost flat smiles in contrast with the upward sloping ones typically observed in the market; see e.g. [42].<sup>18</sup>

Although the patterns of Figure 4.4.7 are robust across values of the VIX index, there is also some level-dependence. More precisely, when volatility is low, the term structure of  $\tilde{\eta}(T)/\eta$  becomes more pronounced and the improvement of recalibrating larger in relative terms, generally. It suggests that the calibration problem is more difficult when volatility is low. This is also reflected in Figure 4.4.5 where the VIX index is at 17.98 on September 4, 2012, and at 11.23 on January 13, 2017, where the fits are notably worse.

In conclusion, we have found evidence that the rough fractional kernel, as used in the rough Bergomi model, lacks flexibility in separating the short and long time scale prop-

---

<sup>17</sup> One may check that  $t^{H-1/2} = \int_0^\infty e^{-\gamma t} \mu(d\gamma)$  where  $\mu(d\gamma) = \gamma^{-H-1/2} (\Gamma(H+1/2)\Gamma(1/2-H))^{-1} d\gamma$ . As in the cited works, we may then express rough volatility as a superposition of an infinite set of semimartingales mean-reverting at both slow and fast exponential rates with weights set by  $\mu$ . Typical one-factor classical models correspond to a single exponential term. The Heston equation (4.4) can e.g. be rewritten as  $V_t = \xi_0(t) + \int_0^t e^{-\kappa(t-s)} \eta \sqrt{V_s} dW_{2,s}$ ; use Ito's lemma on  $X_t := e^{\kappa t} V_t$  and rewrite. However, even if the rough fractional kernel generates mean-reversion at multiple time-scales in the sense just described,  $\mu$  is controlled by the single parameter  $H$  which limits the flexibility.

<sup>18</sup> To solve this, the authors of [42] introduce a class of modulated Volterra processes for volatility and show that the VIX smiles can be captured using such a model with stochastic volatility-of-volatility.

erties of volatility implied by SPX options. Furthermore, we have found evidence that the short term smiles mostly are more symmetric than the long term ones. The latter suggests that we should model volatility two-factor where one factor should be less negatively correlated with the S&P 500 index and have more influence on short expiries. It would then likely also have to be noisier and mean-revert faster than the other factor.

## 4.5 Calibration of more advanced models

In search of a solution to the skew-curvature term structure problem and to examine the ability of rough volatility models for the joint SPX-VIX calibration problem, we in this section calibrate a series of (mostly) more advanced models, including two-factor volatility models, to SPX options and jointly to SPX and VIX options. However, with the increased model complexity, we find it is considerably more challenging to find reasonable fixed moneyness bounds to train neural networks. Because of this, we perform the calibrations directly with Monte Carlo and only calibrate to three dates, though, they will reflect different market scenarios. We leave it for future work to find a better neural network methodology.

### 4.5.1 Model selection

We maintain our assumption that the S&P 500 index has risk-neutral dynamics of the form (4.1). For the volatility models of the present section, we introduce a third Brownian motion  $(W_{3,t})_{t \geq 0}$  and now assume  $dW_{1,t}dW_{2,t} = \rho_{12}dt$ ,  $dW_{1,t}dW_{3,t} = \rho_{13}dt$ , and  $dW_{2,t}dW_{3,t} = \rho_{23}dt$  where  $\rho_{12}, \rho_{13}, \rho_{23} \in [-1, 1]$  are so the associated correlation matrix is positive semi-definite. We assume that the filtration  $(\mathcal{F}_t)_{t \geq 0}$  is generated by  $(W_{1,t}, W_{2,t}, W_{3,t})_{t \geq 0}$  and augmented. We will not always use all three Brownian motions.

All volatility models except one will be driven by Gaussian processes of the form

$$Y_t = \int_0^t K(t-s)dW_s, \quad t \geq 0, \quad (4.13)$$

where  $W_t$  is either  $W_{2,t}$  or  $W_{3,t}$  and  $K$  is a locally square integrable deterministic function. In the most flexible case, we consider the gamma kernel defined by

$$K(t) = e^{-\lambda t} t^\alpha, \quad t > 0, \quad \lambda \geq 0, \quad \alpha \in (-1/2, 0], \quad (4.14)$$

and which nests the rough fractional ( $\lambda = 0$ ) and exponential ( $\alpha = 0$ ) kernels. The latter results in an Ornstein-Uhlenbeck process if used in (4.13). We consider the gamma kernel for two reasons. Firstly, with  $\lambda > 0$ , we get bounded asymptotic variance in that  $\text{Var}(Y_t) = \int_0^t e^{-2\lambda s} s^{2\alpha} ds \rightarrow c^* < \infty$  in the limit  $t \rightarrow \infty$  for a constant  $c^*$ .<sup>19</sup> If  $\lambda = 0$ , we have instead  $\text{Var}(Y_t) = \int_0^t s^{2\alpha} ds = \frac{1}{2\alpha+1} t^{2\alpha+1} \rightarrow \infty$  in the same limit. Asymptotic bounded variance is desirable for volatility which reasonably can be assumed stationary. Secondly, the gamma kernel allows some degree of separate control over the singular part (via  $\alpha$ ) and the long term behaviour (via  $\lambda$ ).

<sup>19</sup> The singular part is integrable and the non-singular part is, up to scale, bounded by  $\int_0^t e^{-2\lambda s} ds = \frac{1}{2\lambda} (1 - e^{-2\lambda t}) \leq \frac{1}{2\lambda} < \infty$  for all  $t \geq 0$ . The claim follows by the monotone convergence theorem.

CHAPTER 4. EMPIRICAL ANALYSIS OF ROUGH AND CLASSICAL  
STOCHASTIC VOLATILITY MODELS TO THE SPX AND VIX MARKETS

---

**Table 4.5.1:** Model specifications.

Model identifier	Equations	Specification
S-RB-F	(4.15)-(4.17)	$\mu = 1, \theta_1 = 1, \eta := \eta_1, \rho := \rho_{12}, \lambda_1 = 0, \alpha := \alpha_1$
S-RB	(4.15)-(4.17)	$\mu = 1, \theta_1 = 1, \eta := \eta_1, \rho := \rho_{12}, \lambda := \lambda_1, \alpha := \alpha_1$
S-2F-B	(4.15)-(4.17)	$\mu = 1, \theta := \theta_1, \eta := \eta_1, \alpha_1 = \alpha_2 = 0$
S-2F-RB	(4.15)-(4.17)	$\mu = 1, \theta := \theta_1, \eta := \eta_1$
S-M-1F-RB	(4.15)-(4.17)	$\theta_1 = \theta_2 = 1, \rho := \rho_{12}, \lambda := \lambda_1, \alpha := \alpha_1$
S-M-2F-RB	(4.15)-(4.17)	No parameter restrictions or simplified notation
S-M-2F-RHyp	(4.22)-(4.24)	$\zeta_{2,0}(t) = \zeta_{1,0}(t) + \epsilon$ where $\epsilon \in \mathbb{R}$
QRH	(4.37)-(4.38)	$\eta = 1, \rho := \rho_{12}$
S-M-2F-QHyp	(4.39)-(4.42)	$\zeta_{2,0}(t) = \zeta_{1,0}(t) + \epsilon$ where $\epsilon \in \mathbb{R}, \alpha_1 = 0, \alpha_2 = 0$
S-M-2F-QRHyp	(4.39)-(4.42)	$\zeta_{2,0}(t) = \zeta_{1,0}(t) + \epsilon$ where $\epsilon \in \mathbb{R}$

\* With the S-prefix removed we refer to the same model with  $c = 0$ .

We will outline four general volatility models and will then consider several ones nested therein. For some models we will restrict the parameters when we calibrate to SPX options and some will not be calibrated to the joint problem. Overall, however, we believe our selection of models for both calibration problems is more than rich enough for a detailed and meaningful analysis. We provide an overview of the models that we calibrate in Table 4.5.1. For each we show an identifier (i.e. label) and specify its definition from a general model. Simplified parameter notation is introduced where meaningful.

We consider first a shifted mixture two-factor rough Bergomi like model defined by

$$V_t = \zeta_0(t) (\mu X_{1,t} + (1 - \mu) X_{2,t}) + c \quad (4.15)$$

$$X_{i,t} = \mathcal{E}(\eta_i \delta_i (\theta_i Y_{1,t} + (1 - \theta_i) Y_{2,t})) \quad (4.16)$$

$$Y_{i,t} = \int_0^t K_i(t-s) dW_{i+1,s}, \quad i = 1, 2, \quad t \geq 0, \quad (4.17)$$

where  $\zeta_0 : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ ,  $K_i(t) = e^{-\lambda_i t} t^{\alpha_i}$ ,  $t > 0$ ,  $\lambda_i \geq 0$ ,  $\alpha_i \in (-\frac{1}{2}, 0]$ ,  $i = 1, 2$ ,  $\eta_1, \eta_2, c \geq 0$ ,  $\mu, \theta_1, \theta_2 \in [0, 1]$ . Note that  $\xi_0(t) = \zeta_0(t) + c$  for  $t \geq 0$ . We choose the constants  $\delta_1$  and  $\delta_2$  so  $\eta_1$  and  $\eta_2$  represent normalised volatility-of-volatility parameters for  $X_{1,t}$  and  $X_{2,t}$ . To explain how exactly we set  $(\delta_1, \delta_2)$ , define first

$$g(t; \theta, \alpha_1, \alpha_2, \lambda_1, \lambda_2, \rho_{23}) := \theta^2 \int_0^t K_1(s)^2 ds + (1 - \theta)^2 \int_0^t K_2(s)^2 ds \quad (4.18)$$

$$+ 2\rho_{23}\theta(1 - \theta) \int_0^t K_1(s)K_2(s) ds, \quad t \geq 0, \quad \theta \in [0, 1]. \quad (4.19)$$

Let  $i \in \{1, 2\}$ . For the next part, recall that

$$\text{Var}(Z_1 + Z_2) = \text{Var}(Z_1) + \text{Var}(Z_2) + 2\text{Cov}(Z_1, Z_2) \quad (4.20)$$

#### 4.5.1. MODEL SELECTION

---

for sufficiently integrable random variables  $Z_1$  and  $Z_2$ . By the Ito isometry and the definition of  $g$ :

$$\text{Var}(\log(X_{i,t})) = (\eta_i \delta_i)^2 g(t; \theta_i, \alpha_1, \alpha_2, \lambda_1, \lambda_2, \rho_{23}).$$

We then set

$$\delta_i := g(1; \theta_i, \alpha_1, \alpha_2, \lambda_1, \lambda_2, \rho_{23})^{-\frac{1}{2}} \quad (4.21)$$

so  $\text{Var}(\log(X_{i,1})) = \eta_i^2$  always. It is worth noting that the function  $g$  can be written on semi-explicit form: Let  $i, j \in \{1, 2\}$  and assume that  $\lambda_i \neq 0$  or  $\lambda_j \neq 0$ . Then

$$\int_0^t K_i(s) K_j(s) ds = \int_0^t e^{-(\lambda_i + \lambda_j)s} s^{\alpha_i + \alpha_j} ds = (\lambda_i + \lambda_j)^{-\alpha_i - \alpha_j - 1} \gamma(\alpha_i + \alpha_j + 1, (\lambda_i + \lambda_j)t)$$

where  $\gamma(s, x) = \int_0^x u^{s-1} e^{-u} du$  denotes the lower incomplete gamma function. If  $\lambda_i = \lambda_j = 0$  we get  $\int_0^t K_i(s) K_j(s) ds = (\alpha_i + \alpha_j + 1)^{-1} t^{\alpha_i + \alpha_j + 1}$ . Plugging the formulas into (4.18)-(4.19) we obtain, at minimum, a semi-explicit expression for  $g$ .

We use Monte Carlo to price options under the model (4.15)-(4.17); we likewise do for the other models that will be presented in this subsection. To this end, we simulate  $Y_{1,t}$  and  $Y_{2,t}$  with the hybrid multifactor scheme of [58]. Plugging into (4.15)-(4.16), we obtain values of  $V_t$ . We compute the VIX index as part of the simulations as described in [58] and simulate  $S_t$  with a log-Euler scheme. For options on  $S_t$  we add its expected value as a control variate. For the other models of this section, we likewise use a log-Euler scheme for  $S_t$  and add its expected value as a control variate when we price options on it.

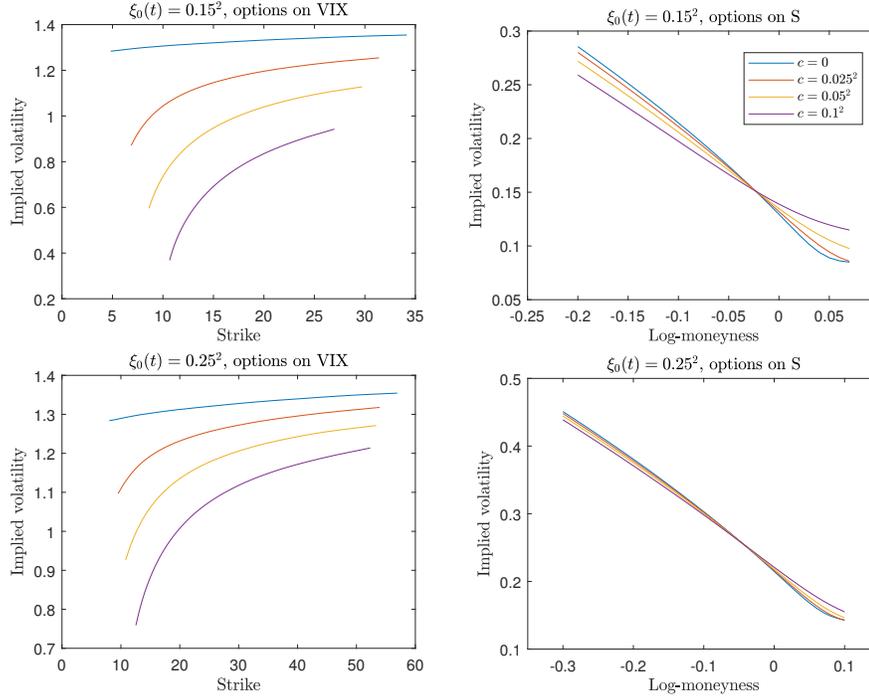
We justify the generality of the model (4.15)-(4.17) as follows: We have included two stochastic factors as the Section 4.4 results suggest. Furthermore, to create a skew for VIX options we allow  $\mu \in (0, 1)$  (i.e. a mixture distribution); the papers [13, 19, 37] also use this idea. As illustrated in Figure 4.5.1, a simple shift ( $c > 0$ ) can also generate a skew and we therefore include both features for the most flexibility.<sup>20</sup> The Heston++ model of [54, 55], with success, also uses a volatility shift, though time-dependent.

We will test six models nested in (4.15)-(4.17), that is, up to restricting  $c = 0$  or not; we refer to rows 1-6 of Table 4.5.1. The idea behind our labelling is as follows: We write 'S' for a shifted model ( $c \geq 0$  is unrestricted), 'M' for a mixture model ( $\mu \in [0, 1]$  is unrestricted), 'RB' for a rough Bergomi type model, 'B' for a non-rough Bergomi type model. Also, by '1F' and '2F', we refer to, respectively, one- and two-factor volatility. For the rough models, our labelling implicitly assumes that the (rough) gamma kernel (4.14) is used. That is, unless the letter 'F' is used, in which case we refer to the pure fractional kernel ( $\lambda = 0$ ). A careful inspection of the labels of Table 4.5.1 show that we do not always use all relevant letters. This is to limit the length of the names. We encourage the reader to become familiar with the model definitions, labels, and the simplified parameter notation as it will greatly enhance the reading of the rest of the paper.

---

<sup>20</sup> Large shifts relative to the volatility level can though flatten the smiles for options on  $S$ ; c.f. the top right plot.

CHAPTER 4. EMPIRICAL ANALYSIS OF ROUGH AND CLASSICAL STOCHASTIC VOLATILITY MODELS TO THE SPX AND VIX MARKETS



**Figure 4.5.1:** Option smiles at the expiry 0.1 under the shifted rough Bergomi model (S-RB-F) defined in Table 4.5.1 with parameters  $\alpha = -0.4$ ,  $\eta = 2.1$ ,  $\rho = -0.9$ , as shown for different values of  $\xi_0$ , which is assumed flat, and  $c$ ; recall that  $\xi_0(t) = \zeta_0(t) + c$  for this model.

To avoid overfitting the shift parameter, we fix  $c = 0$  when we calibrate to SPX options only. We then drop the S-prefix in the labels. Also, to simplify our analysis, we do not calibrate to the joint problem the one-factor models S-RB-F and S-RB or the shifted two-factor Bergomi model S-2F-B. These models, however, are nested in the remaining ones, which include a shifted two-factor rough Bergomi type model (S-2F-RB) and shifted one- and two-factor mixture rough Bergomi type models (S-M-1F-RB and S-M-2F-RB).

When it comes to the models with mixture terms, i.e. S-M-1F-RB, S-M-2F-RB, and their  $c = 0$  restrictions, some of our calibrations result in distributions of  $V_t$  that are too heavy tailed for accurate price estimation. To get around this problem, we borrow an idea from [45] and consider the hyperbolic transformation  $f_{\text{hyp}}(x) := x + \sqrt{x^2 + 1}$  as an alternative to the exponential transformation  $f_{\text{exp}}(x) := e^x$  used in (4.15)-(4.17). The two functions are very similar in that they both are monotone mappings from  $\mathbb{R}$  onto  $\mathbb{R}_+$  and behave alike near the origin; see their Equations (41), (42), and Figure 1. Importantly, however, the hyperbolic transformation grows much slower at large positive values and also decays slower at very negative values. Because of this, the authors' hyperbolic model for  $\sigma_t := \sqrt{V_t}$ , and which is defined as  $\sigma_t = \sigma_0 f_{\text{hyp}}(Y_t)$  where  $Y_t$  is an Ornstein-Uhlenbeck process, has thinner tails than the otherwise comparable log-normal

#### 4.5.1. MODEL SELECTION

---

model  $\sigma_t = \sigma_0 f_{\text{exp}}(Y_t)$  that was proposed in [59]; see e.g. their Figure 2.

In our context a natural model based on the hyperbolic transformation is

$$V_t = \mu f_{\text{hyp}}(X_{1,t}) + (1 - \mu) f_{\text{hyp}}(X_{2,t}) + c \quad (4.22)$$

$$X_{i,t} = \eta_i \delta_i (\theta_i Y_{1,t} + (1 - \theta_i) Y_{2,t}) \quad (4.23)$$

$$Y_{i,t} = \zeta_{i,0}(t) + \int_0^t K_i(t-s) dW_{i+1,s}, \quad i = 1, 2, \quad t \geq 0, \quad (4.24)$$

where  $\mu, c, \eta_1, \eta_2, \theta_1, \theta_2, K_1, K_2$  are as for the model (4.15)-(4.17), the constants  $(\delta_1, \delta_2)$  yet defined by (4.21), and  $\zeta_{i,0} : \mathbb{R}_+ \rightarrow \mathbb{R}$ ,  $i = 1, 2$ . There are no moment problems because  $f_{\text{hyp}}(x)$  grows linearly in the limit  $x \rightarrow \infty$  so those of  $V_t$  are bounded by appropriate moments of the Gaussians  $X_{1,t}$  and  $X_{2,t}$ ; the same logic applies for the *quadratic* hyperbolic model that will be defined on page 156. Define now

$$\zeta_{i,t}(u) := \zeta_{i,0}(u) + \int_0^t K_i(u-s) dW_{i+1,s}, \quad i = 1, 2, \quad 0 \leq t \leq u, \quad (4.25)$$

and note that

$$Y_{i,u} = \zeta_{i,t}(u) + \int_t^u K_i(u-s) dW_{i+1,s}, \quad i = 1, 2, \quad 0 \leq t \leq u, \quad (4.26)$$

where the first term is measurable with respect to  $\mathcal{F}_t$ , the second term independent.

We briefly justify the time-dependence in  $\zeta_{1,0}$  and  $\zeta_{2,0}$ : Let  $i \in \{1, 2\}$  and define for  $t, t_0 \geq 0$ :  $Y_{i,t}^{t_0} := Y_{i,t_0+t}$ ,  $W_{i+1,t}^{t_0} := W_{i+1,t_0+t} - W_{i+1,t_0}$ ,  $\mathcal{F}_t^{t_0} := \mathcal{F}_{t_0+t}$ . Fix  $t_0 \geq 0$ . By (4.26) then

$$\begin{aligned} Y_{i,t}^{t_0} &= \zeta_{i,t_0}(t_0+t) + \int_{t_0}^{t_0+t} K_i(t_0+t-s) dW_{i+1,s} \\ &= \zeta_{i,t_0}(t_0+t) + \int_0^t K_i(t-s) dW_{i+1,s}^{t_0}, \quad t \geq 0. \end{aligned}$$

Because  $(W_{i+1,t}^{t_0})_{t \geq 0}$  is Brownian motion with respect to  $(\mathcal{F}_t^{t_0})_{t \geq 0}$  and  $\zeta_{i,t_0}$  is  $\mathcal{F}_{t_0}$ -measurable, the distribution of  $(Y_{i,t}^{t_0})_{t \geq 0}$  conditional on  $\mathcal{F}_{t_0}$  is equivalent to that at time zero but with an updated initial term structure  $t \mapsto \zeta_{i,t_0}(t_0+t)$ . In particular, even if  $\zeta_{i,t_0}(t_0+\cdot)$  is flat for  $t_0 = 0$  it generally is not for  $t_0 > 0$ ; recall (4.25). Heuristically, we can therefore think of any time-dependence in the initial curve  $\zeta_{i,0}$  as an expression of the historical path-dependence that has occurred before time zero. It also means that movements in the curves  $\zeta_{1,t_0}(t_0+\cdot)$  and  $\zeta_{2,t_0}(t_0+\cdot)$  can be hedged under the model.

*Remark:* The same logic for (4.15)-(4.17) justifies both a time-dependent  $\xi_0$  curve, equivalently  $\zeta_0$ , and a time-dependent mixing parameter  $\mu$ . That we keep  $\mu$  constant is a simplification to avoid overfitting to the time-dependency when we calibrate. Similar arguments will justify time-dependencies in the remaining models that will be defined.

We consider a single but very general model nested in (4.22)-(4.24) for our calibrations, that is, up to setting  $c = 0$  or not; we refer to the label S-M-2F-RHyp in Table 4.5.1. To avoid overfitting to the time-dependency, we have restricted  $\zeta_{2,0}(t) = \zeta_{1,0}(t) + \epsilon$  for a number  $\epsilon \in \mathbb{R}$ . We again fix  $c = 0$  when we calibrate to SPX options only.

For (4.22)-(4.24), we also simulate  $(Y_{1,t}, Y_{2,t})$  with the hybrid multifactor scheme. Plugging into (4.22)-(4.23), we then obtain  $V_t$  and may thereafter simulate  $S_t$  with a log-Euler scheme. Computing the VIX index as part of the simulations is however less trivial and not covered by the discussions in [58]. We therefore explain our approach in the following.

We start with a trapezoidal rule applied to (4.12) as

$$\text{VIX}_t^2 \approx \frac{100^2}{n_v} \sum_{i=0}^{n_v} a_i \xi_t(t + \tau_i), \quad t \geq 0, \quad (4.27)$$

where  $n_v \geq 2$  is an integer and  $a_0 = a_{n_v} = \frac{1}{2}$ ,  $a_i = 1$ ,  $i = 1, \dots, n_v - 1$ ,  $\tau_i = \frac{i}{n_v} \Delta$ ,  $i = 0, \dots, n_v$ . To compute  $\text{VIX}_t$  as part of the simulations it then suffices to compute  $\xi_t(u)$  for specific  $u$ 's. Note that

$$\xi_t(u) = \mu E_t(f_{\text{hyp}}(X_{1,u})) + (1 - \mu) E_t(f_{\text{hyp}}(X_{2,u})) + c, \quad 0 \leq t \leq u. \quad (4.28)$$

Define

$$\chi_{i,t}(u) := \eta_i \delta_i (\theta_i \zeta_{1,t}(u) + (1 - \theta_i) \zeta_{2,t}(u)) \quad (4.29)$$

$$\nu_i(u - t) := \eta_i \delta_i \sqrt{g(u - t; \theta_i, \alpha_1, \alpha_2, \lambda_1, \lambda_2, \rho_{23})}, \quad i = 1, 2, \quad 0 \leq t \leq u. \quad (4.30)$$

By (4.23) and (4.26), we for  $i \in \{1, 2\}$  and  $(t, u)$  such that  $0 \leq t \leq u$  have

$$E_t(f_{\text{hyp}}(X_{i,u})) = E_t(f_{\text{hyp}}(\eta_i \delta_i \{\theta_i Y_{1,u} + (1 - \theta_i) Y_{2,u}\})) \quad (4.31)$$

$$= E_t \left( f_{\text{hyp}} \left( \eta_i \delta_i \{ \theta_i \zeta_{1,t}(u) + (1 - \theta_i) \zeta_{2,t}(u) \} \right. \right. \quad (4.32)$$

$$\left. \left. + \eta_i \delta_i \left\{ \theta_i \int_t^u K_1(u - s) dW_{2,s} \right. \right. \quad (4.33)$$

$$\left. \left. + (1 - \theta_i) \int_t^u K_2(u - s) dW_{3,s} \right\} \right). \quad (4.34)$$

Recall that  $\zeta_{1,t}(u)$  and  $\zeta_{2,t}(u)$  are  $\mathcal{F}_t$ -measurable and that the stochastic integrals of (4.33)-(4.34) are independent of  $\mathcal{F}_t$  and therefore are conditionally normally distributed as they are unconditionally. By this, (4.20), the Ito isometry, (4.29)-(4.30), we conclude that

$$E_t(f_{\text{hyp}}(X_{i,u})) = E(f_{\text{hyp}}(x + \nu_i(u - t)Z)) \Big|_{x=\chi_{i,t}(u)}, \quad i = 1, 2, \quad 0 \leq t \leq u, \quad (4.35)$$

where  $Z \sim \mathcal{N}(0, 1)$ . Given knowledge of the terms in (4.29)-(4.30), the problem then reduces to that of computing unconditional expectations of a transformed standard normal variable. Note that values of  $\nu_i(u - t)$  are readily computable via (4.30) and the

expression for  $g$  given on page 150. As the variables are deterministic, the cost of evaluating them is small when spread across many paths. Even better as  $\nu_i(u-t)$  only depends on the time-to-maturity  $u-t$ , we can reuse them when computing  $\text{VIX}_t$  at multiple time points  $t$ . In contrast, values of  $\chi_{i,t}(u)$  require knowledge of  $\zeta_{1,t}(u)$  and  $\zeta_{2,t}(u)$  which depend on  $\mathcal{F}_t$ . Fortunately, however, these terms are relatively easy to compute under the hybrid multifactor scheme as part of the simulations; we refer to [58] for details.

Let us then discuss how to evaluate (4.35) given knowledge of (4.29)-(4.30). Define

$$h(x, y) := \int_{\mathbb{R}} f_{\text{hyp}}(x + yz)\phi(z)dz, \quad (x, y) \in \mathbb{R}^2, \quad (4.36)$$

where  $\phi(z) := (2\pi)^{-\frac{1}{2}}e^{-z^2/2}$ ,  $z \in \mathbb{R}$ . Then  $E_t(f_{\text{hyp}}(X_{i,u})) = h(\chi_{i,t}(u), \nu_i(u-t))$  for  $i = 1, 2$ , and  $0 \leq t \leq u$ , which reduces the problem to that of evaluating the function  $h$ . To our knowledge, no analytical expression exists. We therefore use numerical integration with Matlab's *integral* function. However, evaluating  $h(\chi_{i,t}(u), \nu_i(u-t))$ ,  $i = 1, 2$ , with numerical integration across multiple time points  $t$ , forward variance maturities  $u$ , and across multiple sample paths, quickly becomes very time-consuming. We thus suggest a slightly different approach to evaluate the required terms that we shall now explain.

Let  $\mathcal{T} \subset \mathbb{R}_+$  be a set of expiries on which to price VIX options and say that we already have values of  $\nu_i(\tau_j)$ ,  $\chi_{i,t}(t + \tau_j)$  available for  $t \in \mathcal{T}$ ,  $i = 1, 2$ ,  $j = 0, 1, \dots, n_v$ , across the simulated paths. To compute the forward variances relevant to (4.27) at all desired dates  $t \in \mathcal{T}$  and across all paths, we must evaluate  $h(\chi_{i,t}(t + \tau_j), \nu_i(\tau_j))$  for  $t \in \mathcal{T}$ ,  $i = 1, 2$ ,  $j = 0, 1, \dots, n_v$ , likewise over all paths. Define the functions  $\bar{h}_{ij}(x) := h(x, \nu_i(\tau_j))$ ,  $x \in \mathbb{R}$ ,  $i = 1, 2$ ,  $j = 0, 1, \dots, n_v$ . To reduce the computational costs, we first pre-compute each function  $\bar{h}_{ij}$  on a smaller set of points on the relevant part of its domain. Next, we use linear interpolation between the computed values to obtain  $\bar{h}_{ij}(\chi_{i,t}(t + \tau_j)) = h(\chi_{i,t}(t + \tau_j), \nu_i(\tau_j))$  for  $i = 1, 2$ ,  $j = 0, 1, \dots, n_v$ ,  $t \in \mathcal{T}$ , for all paths. In our experience,  $n_v$  need not be very large; we have found  $n_v \approx 32$  to suffice. It is then only a smaller number of functions that we have to pre-compute points on.

We briefly elaborate on how precisely we pre-compute points on the functions  $\bar{h}_{ij}$ ,  $i = 1, 2$ ,  $j = 0, 1, \dots, n_v$ : Let  $i \in \{1, 2\}$  and  $j \in \{0, 1, \dots, n_v\}$ . To evaluate  $\bar{h}_{ij}(\chi_{i,t}(t + \tau_j))$  for all  $t \in \mathcal{T}$  across all paths, we first locate the minimum and maximum values of  $\chi_{i,t}(t + \tau_j)$  for all time points  $t \in \mathcal{T}$  and paths. Denote them by  $\chi_{ij}^{\min}$  and  $\chi_{ij}^{\max}$ . For a step size  $\Delta_x > 0$ , we set  $n_{ij} := \lceil (\chi_{ij}^{\max} - \chi_{ij}^{\min})/\Delta_x \rceil$ , define  $x_{ijk} := \chi_{ij}^{\min} + k\Delta_x$ ,  $k = 0, 1, \dots, n_{ij}$ , and evaluate  $y_{ijk} := \bar{h}_{ij}(x_{ijk})$ ,  $k = 0, 1, \dots, n_{ij}$ , with numerical integration. Next, we compute  $\bar{h}_{ij}(\chi_{i,t}(t + \tau_j))$  for any  $t \in \mathcal{T}$  and path by linear interpolation between the points  $(x_{ijk}, y_{ijk})_{k=0}^{n_{ij}}$ . For reasonable values of  $\Delta_x$  (we use  $\Delta_x = 0.01$ ) and realistic parameters, we find that (4.36) needs to be evaluated many fewer times than otherwise.

One concern with our approach is that  $\chi_{ij}^{\min}$ ,  $\chi_{ij}^{\max}$ ,  $i = 1, 2$ ,  $j = 0, 1, \dots, n_v$ , which determine the number of numerical integrations, are unbounded and only known at run-time. Fortunately, we have not found this to be a practical problem when reasonable model parameters are used. To bound the number of integrations one could nevertheless use fixed limits and look for an extrapolation formula.

Next, we consider a quadratic rough Heston model defined by

$$V_t = a(Z_t - b)^2 + c \quad (4.37)$$

$$Z_t = \zeta_0(t) + \int_0^t K(t-s)\eta\sqrt{V_s}dW_{2,s}, \quad t \geq 0, \quad (4.38)$$

where  $a, c, \eta \geq 0$ ,  $b \in \mathbb{R}$ ,  $K(t) = e^{-\lambda t}t^\alpha$ ,  $\lambda \geq 0$ ,  $\alpha \in (-\frac{1}{2}, 0]$ , and  $\zeta_0 : \mathbb{R}_+ \rightarrow \mathbb{R}$  is locally Hölder continuous of all orders strictly less than  $\alpha + \frac{1}{2}$ . A unique continuous strong solution exists by [2, Theorem A.1].<sup>21</sup> Define  $f^{\alpha,\lambda}(t) := \lambda t^{\alpha-1}E_{\alpha,\alpha}(-\lambda t^\alpha)$ ,  $t > 0$ ,  $\alpha \in (\frac{1}{2}, 1)$ ,  $\lambda > 0$ , where

$$E_{\alpha,\beta}(z) = \sum_{n \geq 0} \frac{z^n}{\Gamma(\alpha n + \beta)}, \quad z, \alpha, \beta \in \mathbb{C}, \quad \operatorname{Re}(\alpha), \operatorname{Re}(\beta) > 0,$$

is the so-called *Mittag-Leffler* function. The original quadratic rough Heston model of [32] then fits into the formulation (4.37)-(4.38) with  $K(t) = f^{\alpha,\lambda}(t)$ ,  $\zeta_0(t) = \int_0^t f^{\alpha,\lambda}(t-s)\theta_0(s)ds$ ,  $t > 0$ , where  $\theta_0$  is a deterministic function. We believe our version has comparable generality as  $\zeta_0$  in both cases control the term structure of  $(Z_t)_{t \geq 0}$  (directly or via  $\theta_0$ ) and since the  $\alpha$ 's (which are equivalent up to a shift) both control the singular part of  $K$  while the  $\lambda$ 's control the decay at larger values. We use the hybrid multifactor scheme of [58] to simulate the model. Computation of the VIX index is also described in that paper. For normalisation we set  $\eta = 1$  as suggested in [32] for the original model. The version we calibrate is stated in Table 4.5.1 with the label QRH. In keeping somewhat with the original formulation, we allow  $c \geq 0$  also when we calibrate to SPX options only.

Inspired by (4.22)-(4.24) and (4.37)-(4.38), we consider lastly a two-factor model that uses both the quadratic and hyperbolic transformations. Specifically,

$$V_t = \mu X_{1,t} + (1 - \mu)X_{2,t} + c \quad (4.39)$$

$$X_{i,t} = f_{\text{hyp}}(Z_{i,t}^2 - d_i) - f_{\text{hyp}}(-d_i), \quad (4.40)$$

$$Z_{i,t} = \eta_i \delta_i (\theta_i Y_{1,t} + (1 - \theta_i) Y_{2,t}), \quad (4.41)$$

$$Y_{i,t} = \zeta_{i,0}(t) + \int_0^t K_i(t-s)dW_{i+1,s}, \quad i = 1, 2, \quad t \geq 0, \quad (4.42)$$

where  $\mu, \theta_1, \theta_2 \in [0, 1]$ ,  $\eta_1, \eta_2, c \geq 0$ ,  $d_1, d_2 \in \mathbb{R}$ ,  $(\delta_1, \delta_2)$  is defined by (4.21),  $\zeta_{i,0} : \mathbb{R}_+ \rightarrow \mathbb{R}$ ,  $K_i(t) = e^{-\lambda_i t}t^{\alpha_i}$ ,  $t > 0$ ,  $\lambda_i \geq 0$ ,  $\alpha_i \in (-\frac{1}{2}, 0]$ ,  $i = 1, 2$ . The parameters  $d_1$  and  $d_2$  are introduced to allow control over the minima of the terms  $Z_{i,t}^2 - d_i$ ,  $i = 1, 2$ . The processes  $X_{1,t}$  and  $X_{2,t}$  are shifted so the minimum value they can attain is zero. We simulate the model akin to (4.22)-(4.24) and likewise use the approach outlined on pages 154-155 to compute the VIX index—it carries over with only trivial changes. We calibrate a version with gamma kernels and one with pure exponential kernels; they are defined with labels S-M-2F-QRHyp and S-M-2F-QHyp in Table 4.5.1. We again restrict  $\zeta_{2,0}(t) = \zeta_{1,0}(t) + \epsilon$  for a number  $\epsilon \in \mathbb{R}$  and also fix  $c = 0$  when we calibrate to SPX options only.

<sup>21</sup> To use the theorem we must check: (1) That  $K$  satisfies hypothesis  $(H_0)$  of the paper with their  $\gamma = 2\alpha + 1$ . This follows from parts (ii) and (iv) of [4, Example 2.3]. (2) That the diffusion coefficient of (4.38) is Lipschitz continuous in the value of  $Z_s$ . We here refer to [58].

### 4.5.2 Calibration results

We calibrate to the dates October 14, 2011, September 4, 2012, and January 13, 2017; the last two were also considered in Figure 4.4.5. The VIX index is respectively at 28.24, 17.98 and 11.23. We therefore cover the case of high, typical, and low volatility. Our VIX options dataset is, like our SPX dataset, obtained from <https://datashop.cboe.com>. We filter both datasets as we filtered the SPX dataset in Section 4.4 (in-the-money options are e.g. removed) except we do not filter the range of moneyness but instead the set of expiries. To select expiries, we consider the intervals defined by the time points  $\{0, 0.01, 0.02, 0.03, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1, 1.5, 2, 2.5, 3\}$  and in each select the expiry (if any) closest to the midpoint. This gives 10, 12 and 13 SPX expiries and 6, 5 and 7 VIX expiries for each of the dates (listed in the same order). To compute Black-Scholes implied volatility for VIX options, we use the same expiry VIX futures as the underlying. We extract the futures prices from the observed VIX option put-call parities.

For the joint calibration problem, we find that it works better to measure the error in terms of deviations from the bid-ask bounds. We therefore change to such an error measure: Consider a given date and let  $N_s$  and  $N_v$  denote the number of observed SPX options, respectively, VIX options. Let  $\sigma_{s,i}^{\text{bid}}$ ,  $\sigma_{s,i}^{\text{ask}}$  and  $\sigma_{s,i}^{\text{model}}$ , respectively,  $\sigma_{v,i}^{\text{bid}}$ ,  $\sigma_{v,i}^{\text{ask}}$  and  $\sigma_{v,i}^{\text{model}}$ , denote the bid, ask and model implied volatility for the  $i$ 'th observed SPX option, respectively, VIX option. We then measure the calibration error by the number

$$\sqrt{\theta \sum_{i=1}^{N_s} w_{s,i} F(\sigma_{s,i}^{\text{bid}}, \sigma_{s,i}^{\text{ask}}, \sigma_{s,i}^{\text{model}})^2 + (1-\theta) \sum_{i=1}^{N_v} w_{v,i} F(\sigma_{v,i}^{\text{bid}}, \sigma_{v,i}^{\text{ask}}, \sigma_{v,i}^{\text{model}})^2} \quad (4.43)$$

where  $\theta \in [0, 1]$  and we for a set of numbers  $x, y$  and  $z$  so  $x < y$  have defined

$$F(x, y, z) := \begin{cases} x - z & z < x \\ z - y & z > y \\ 0 & \text{otherwise} \end{cases}.$$

The terms  $\{w_{s,i}\}_{i=1}^{N_s}$ ,  $\{w_{v,i}\}_{i=1}^{N_v}$  represent weights which we choose so each expiry carries the same total amount but are otherwise normalised so  $\sum_{i=1}^{N_s} w_{s,i} = \sum_{i=1}^{N_v} w_{v,i} = 1$ . Stated in words, we measure the calibration error as the square root of a weighted average of squared deviations of the model implied volatilities from the bid-ask bounds. The parameter  $\theta$  allows us to control the relative weight of each market.

When we calibrate to SPX options we set  $\theta = 1$ . We then choose three initial guesses and minimise (4.43) under each; we here assume that the curves  $\zeta_0$ ,  $\zeta_{1,0}$ , and  $\zeta_{2,0}$ , whichever are relevant, are flat. We use that of the three solutions with the lowest error as an initial guess for a final calibration whose results we present and for which we assume the relevant curves piecewise constant between the observed SPX expiries.<sup>22</sup> For the joint problem we set  $\theta = 0.9$  which we find leads to a reasonable balance between the

<sup>22</sup> Although we need  $\zeta_0$  locally Hölder *continuous* of all orders less than  $\alpha + \frac{1}{2}$  for a solution to (4.38), we find no meaningful price differences when we replace the calibrated  $\zeta_0$  curves with approximating  $C^1$  curves.

two markets. We then run a single optimisation with the solution from the final SPX calibration as an initial guess. The final calibrations are run with 50 000 paths.

We present the calibration results in tables 4.5.2, 4.5.3, and 4.5.5. Fits are shown in figures 4.5.2–6. For brevity and visibility only a small selection of the expiries and models are included in the plots. However, since the chosen expiries include the shortest and longest ones on each date, and because the models that are included more or less represent the different error levels that can be found in the tables, it will suffice for our discussion. The reader may note that we do not report calibration errors for some of the (shifted) log-normal mixture models. The reason is, as alluded to earlier, that it sometimes is difficult to estimate prices with sufficient accuracy under these models. We discuss the issue in detail below. Only thereafter do we analyse the calibration results.

To illustrate the estimation problem, we examine, under the calibrated models, Monte Carlo estimation of the fair strike of a 1-year variance swap, i.e. the value of  $K_{\text{vs}} := E(\frac{1}{T} \int_0^T V_t dt)$  where  $T = 1$ . We focus on variance swaps as  $K_{\text{vs}}$  can be computed with other deterministic methods to compare with. If we cannot estimate a variance swap with sufficient accuracy, we can expect similar problems when we price other types of derivatives with Monte Carlo. The results of our experiment are presented in Table 4.5.4 where we show  $K_{\text{vs}}$  computed with deterministic methods (denoted  $\hat{K}_{\text{vs}}^{\text{det}}$ ) alongside statistics on a Monte Carlo estimator (denoted  $\hat{K}_{\text{vs}}^{\text{mc}}$ ).

We compute  $\hat{K}_{\text{vs}}^{\text{det}}$  as follows: For the models nested in (4.15)-(4.17), we use the formula  $K_{\text{vs}} = \frac{1}{T} \int_0^T \zeta_0(t) dt + c =: \hat{K}_{\text{vs}}^{\text{det}}$ , which we can (and will) evaluate without error as  $\zeta_0$  is assumed piecewise constant. For the models nested in (4.22)-(4.24) and (4.39)-(4.42), we discretise  $\frac{1}{T} \int_0^T E(V_t) dt$  with a trapezoidal rule and compute values of  $E(V_t)$  with numerical integration along the lines of pages 154-155. We use the same trapezoidal rule for the quadratic rough Heston model but compute values of  $E(V_t)$  by solving Equation (105) of [58] with the hybrid multifactor scheme. To construct a single Monte Carlo sample  $\hat{K}_{\text{vs}}^{\text{mc}}$ , we simulate  $(V_t)_{t \in [0, T]}$ , compute  $\frac{1}{T} \int_0^T V_t dt$  with a trapezoidal rule, and average values over 50 000 paths (as for the calibrations). The statistics in Table 4.5.4 are based on 100 samples of  $\hat{K}_{\text{vs}}^{\text{mc}}$ . We use a fine discretisation with 2500 equidistant steps for the trapezoidal rules and to simulate  $(V_t)_{t \in [0, T]}$ . We can then, in particular, expect  $\hat{K}_{\text{vs}}^{\text{det}}$  computed with *all* the deterministic methods to closely approximate the true values  $K_{\text{vs}}$ ; for the models nested in (4.15)-(4.17) we of course have equality regardless. Moreover, we can for the Monte Carlo estimates  $\hat{K}_{\text{vs}}^{\text{mc}}$  expect a small distributional error and therefore that the Monte Carlo sampling error should dominate.

In terms of the results, note that the median estimates are all lower than the values computed with the deterministic methods. We believe this is a consequence of  $V_t$  being right-skewed which is natural given that it lies close to but is bounded below by zero. However, for some of the models with log-normal mixture terms, the differences are substantial. For example, S-M-2F-RB calibrated to the joint problem on September 4, 2012, has  $\hat{K}_{\text{vs}}^{\text{det}}$ , which *equals*  $K_{\text{vs}}$  for this example, at 0.1385 whereas the median of  $\hat{K}_{\text{vs}}^{\text{mc}}$  is 0.0950. The 5–95 percentile interval is 0.0833–0.1506 which shows that  $\hat{K}_{\text{vs}}^{\text{mc}}$  also has a very wide distribution. We believe the problems are much worse under the models with

**Table 4.5.2:** Calibration to SPX options. Bar-notation denotes average values till 1 month. Errors are reported in basis points, though, some are left out where price estimation is deemed too difficult; we refer to the main text.

<b>RB-F</b>	$\eta$	$\rho$	$\alpha$	$\xi_0$	Error									
2011-10-14	2.00	-1.00	-0.43	0.0845	35									
2012-09-04	1.90	-0.80	-0.41	0.0347	41									
2017-01-13	2.63	-1.00	-0.45	0.0145	86									
<b>RB</b>	$\eta$	$\rho$	$\lambda$	$\alpha$	$\bar{\xi}_0$	Error								
2011-10-14	1.85	-0.88	0.88	-0.36	0.0848	32								
2012-09-04	1.84	-0.75	1.00	-0.33	0.0354	36								
2017-01-13	2.26	-0.65	14.79	-0.12	0.0127	87								
<b>2F-B</b>	$\theta$	$\eta$	$\rho_{12}$	$\rho_{13}$	$\rho_{23}$	$\lambda_1$	$\lambda_2$	$\bar{\xi}_0$	Error					
2011-10-14	0.90	2.03	-0.50	-0.96	0.27	71.73	1.17	0.0803	7					
2012-09-04	0.90	2.03	-0.35	-0.93	-0.02	92.81	1.46	0.0330	11					
2017-01-13	0.88	2.34	-0.29	-0.97	0.05	74.62	2.89	0.0133	39					
<b>2F-RB</b>	$\theta$	$\eta$	$\rho_{12}$	$\rho_{13}$	$\rho_{23}$	$\lambda_1$	$\lambda_2$	$\alpha_1$	$\alpha_2$	$\xi_0$	Error			
2011-10-14	0.77	2.04	-0.48	-0.97	0.25	50.50	0.96	-0.17	-0.06	0.0807	7			
2012-09-04	0.67	2.00	-0.27	-1.00	0.29	45.67	0.90	-0.27	-0.17	0.0337	11			
2017-01-13	0.91	2.38	-0.22	-0.97	-0.03	90.56	2.41	-0.01	-0.11	0.0134	38			
<b>M-1F-RB</b>	$\mu$	$\eta_1$	$\eta_2$	$\rho$	$\lambda$	$\alpha$	$\bar{\xi}_0$	Error						
2011-10-14	0.82	1.61	7.17	-0.84	1.18	-0.24	0.0897	*						
2012-09-04	0.65	1.58	9.98	-0.75	2.08	-0.05	0.0430	*						
2017-01-13	0.22	1.99	7.58	-1.00	0.46	-0.43	0.0496	*						
<b>M-2F-RB</b>	$\mu$	$\theta_1$	$\theta_2$	$\eta_1$	$\eta_2$	$\rho_{12}$	$\rho_{13}$	$\rho_{23}$	$\lambda_1$	$\lambda_2$	$\alpha_1$	$\alpha_2$	$\xi_0$	Error
2011-10-14	0.22	0.27	0.80	0.87	2.37	-0.21	-0.99	0.23	88.78	0.79	-0.21	-0.27	0.0817	3
2012-09-04	0.69	0.80	0.90	1.37	4.94	-0.27	-0.90	0.00	82.03	1.42	-0.06	-0.12	0.0402	*
2017-01-13	0.18	0.00	0.08	1.55	8.11	-0.99	-0.81	0.85	181.10	3.26	-0.48	-0.09	0.0573	*
<b>QRH</b>	$a$	$b$	$c$	$\rho$	$\lambda$	$\alpha$	$\bar{\xi}_0$	Error						
2011-10-14	0.09	0.78	0.0120	1.00	0.23	-0.43	0.1458	29						
2012-09-04	0.11	0.42	0.0115	1.00	0.83	-0.42	0.1668	30						
2017-01-13	0.21	0.08	0.0024	1.00	8.82	-0.42	-0.0029	70						

Table 4.5.3: Calibration to SPX options (more results).

<b>M-2F-RHyp</b>	$\mu$	$\theta_1$	$\theta_2$	$\eta_1$	$\eta_2$	$\rho_{12}$	$\rho_{13}$	$\rho_{23}$	$\lambda_1$	$\lambda_2$	$\alpha_1$	$\alpha_2$	$\bar{\zeta}_{1,0}$	$\bar{\zeta}_{2,0}$	Error		
2011-10-14	0.38	0.50	0.00	4.87	25.45	-1.00	-0.97	0.96	0.00	6.96	-0.44	-0.04	-4.07	-0.75	27		
2012-09-04	0.17	0.74	0.05	10.54	9.64	-0.99	-1.00	1.00	0.87	0.95	-0.44	0.00	-4.61	-2.34	15		
2017-01-13	0.94	0.01	0.71	47.80	17.66	-0.86	-0.67	0.19	2.07	0.54	-0.40	-0.01	-2.34	-3.09	65		
<b>M-2F-QHyp</b>	$\mu$	$\theta_1$	$\theta_2$	$\eta_1$	$\eta_2$	$d_1$	$d_2$	$\rho_{12}$	$\rho_{13}$	$\rho_{23}$	$\lambda_1$	$\lambda_2$	$\bar{\zeta}_{1,0}$	$\bar{\zeta}_{2,0}$	Error		
2011-10-14	0.42	0.15	0.07	0.43	1.03	1.68	15.00	1.00	0.51	0.50	1.35	91.46	-0.70	-0.09	4		
2012-09-04	0.64	0.12	0.05	0.62	0.89	3.48	8.19	0.90	0.32	-0.01	1.99	232.77	-0.45	-0.02	12		
2017-01-13	0.02	0.09	0.06	1.30	0.78	8.13	6.78	0.99	0.72	0.65	1.45	108.85	-0.71	0.02	20		
<b>M-2F-QRHyp</b>	$\mu$	$\theta_1$	$\theta_2$	$\eta_1$	$\eta_2$	$d_1$	$d_2$	$\rho_{12}$	$\rho_{13}$	$\rho_{23}$	$\lambda_1$	$\lambda_2$	$\alpha_1$	$\alpha_2$	$\bar{\zeta}_{1,0}$	$\bar{\zeta}_{2,0}$	Error
2011-10-14	0.97	0.08	0.18	1.21	0.80	21.95	0.80	1.00	0.52	0.45	1.58	129.11	0.00	-0.03	-0.53	-0.09	9
2012-09-04	0.19	0.27	0.49	1.00	0.48	8.43	4.12	0.91	0.26	-0.07	0.69	45.36	-0.16	-0.26	-1.42	-0.11	5
2017-01-13	0.93	0.02	0.17	0.98	0.96	11.79	5.07	0.99	0.59	0.47	2.09	133.36	0.00	-0.04	-0.50	0.00	22

#### 4.5.2. CALIBRATION RESULTS

log-normal mixture terms because the tails are heavier which should amplify the effects.

**Table 4.5.4:** Estimation of the 1-year fair variance swap strike  $K_{vs}$  with the Monte Carlo estimator  $\hat{K}_{vs}^{mc}$ . In the columns with numbers, we show first values of  $K_{vs}$  computed with deterministic methods (denoted  $\hat{K}_{vs}^{det}$ ), then differences between the medians of  $\hat{K}_{vs}^{mc}$  and the values of  $\hat{K}_{vs}^{det}$ , lastly the percentiles 5 and 95 of  $\hat{K}_{vs}^{mc}$ . Statistics for  $\hat{K}_{vs}^{mc}$  are based on 100 samples. The cases where we have deemed estimation to difficult are highlighted in grey.

Date and model	Calibration to SPX options				Joint SPX-VIX calibration			
	$\hat{K}_{vs}^{det}$	Median minus $\hat{K}_{vs}^{det}$	5th perc.	95th perc.	$\hat{K}_{vs}^{det}$	Median minus $\hat{K}_{vs}^{det}$	5th perc.	95th perc.
<b>Oct. 14, 2011:</b>								
RB-F	0.1086	-0.0011	0.1064	0.1085				
RB	0.1120	-0.0008	0.1099	0.1125				
2F-B	0.1167	-0.0018	0.1137	0.1162				
(S-)2F-RB	0.1152	-0.0018	0.1122	0.1150	0.1235	-0.0035	0.1186	0.1216
(S-)M-1F-RB	0.1345	-0.0209	0.1113	0.1263	0.1087	-0.0004	0.1071	0.1101
(S-)M-2F-RB	0.1164	-0.0026	0.1124	0.1150	0.1241	-0.0035	0.1188	0.1244
QRH	0.1219	-0.0010	0.1181	0.1241	0.1173	-0.0008	0.1149	0.1186
(S-)M-2F-RHyp	0.1121	-0.0008	0.1099	0.1124	0.1115	-0.0006	0.1095	0.1123
(S-)M-2F-QHyp	0.1170	-0.0021	0.1140	0.1158	0.1188	-0.0029	0.1147	0.1168
(S-)M-2F-QRHyp	0.1164	-0.0031	0.1124	0.1142	0.1214	-0.0032	0.1169	0.1195
<b>Sept. 4, 2012:</b>								
RB-F	0.0739	-0.0007	0.0723	0.0739				
RB	0.0764	-0.0004	0.0750	0.0770				
2F-B	0.0781	-0.0016	0.0758	0.0774				
(S-)2F-RB	0.0746	-0.0011	0.0726	0.0742	0.0842	-0.0012	0.0813	0.0841
(S-)M-1F-RB	0.1172	-0.0400	0.0762	0.0805	0.0846	-0.0082	0.0755	0.0773
(S-)M-2F-RB	0.0965	-0.0091	0.0821	0.1062	0.1385	-0.0435	0.0833	0.1506
QRH	0.0775	-0.0006	0.0757	0.0785	0.0741	-0.0007	0.0724	0.0749
(S-)M-2F-RHyp	0.0739	-0.0008	0.0722	0.0738	0.0713	-0.0009	0.0697	0.0711
(S-)M-2F-QHyp	0.0815	-0.0031	0.0776	0.0792	0.0773	-0.0022	0.0746	0.0758
(S-)M-2F-QRHyp	0.0765	-0.0013	0.0746	0.0758	0.0766	-0.0011	0.0748	0.0763
<b>Jan. 13, 2017:</b>								
RB-F	0.0398	-0.0006	0.0386	0.0398				
RB	0.0386	-0.0004	0.0378	0.0387				
2F-B	0.0417	-0.0010	0.0399	0.0411				
(S-)2F-RB	0.0423	-0.0013	0.0404	0.0417	0.0456	-0.0015	0.0430	0.0459
(S-)M-1F-RB	0.1703	-0.1246	0.0403	0.0863	0.0954	-0.0542	0.0390	0.0627
(S-)M-2F-RB	0.2083	-0.1664	0.0386	0.1199	0.1693	-0.1195	0.0487	0.0734
QRH	0.0470	-0.0035	0.0414	0.0594	0.0436	-0.0016	0.0409	0.0446
(S-)M-2F-RHyp	0.0376	-0.0006	0.0365	0.0377	0.0390	-0.0005	0.0374	0.0397
(S-)M-2F-QHyp	0.0412	-0.0013	0.0396	0.0403	0.0408	-0.0010	0.0394	0.0401
(S-)M-2F-QRHyp	0.0411	-0.0013	0.0394	0.0402	0.0417	-0.0015	0.0399	0.0405

For evidence of the last point, we refer to Table 4.5.2 and 4.5.5, where we note that the values of either  $\eta_1$  or  $\eta_2$  typically are relatively large whenever we observe an estimation problem in Table 4.5.4. For example, for the problematic model S-M-2F-RB calibrated to the joint problem on September 4, 2012, we have  $\eta_1 = 2.00$  and  $\eta_2 = 5.84$  whereas the maximum value of  $\eta$  for any of the non-mixture log-normal models, and for which no such estimation problems are observed, is  $\eta = 3.03$ .<sup>23</sup> We believe the reason that we do not observe the same estimation problems for the hyperbolic models (despite also being mixtures) is that the tails decay faster under the hyperbolic transformation. This, of course, was precisely our motivation for introducing the hyperbolic transformation.

We turn our attention to the SPX calibrations. We start with a discussion of the one-factor models. Note first that there is a very limited improvement, if any, when we extend the rough Bergomi model (RB-F) to that with the more general gamma kernel (RB); the calibration errors change from 35, 41 and 86 bps to 32, 36 and 87 bps. Results are only slightly better under the quadratic rough Heston model (QRH) which has errors of 29, 30 and 70 bps. We show the fits for RB-F and QRH in figures 4.5.2 and 4.5.3. In line with the results of Section 4.4.3 both fail to capture the term structure of smile (a)symmetry. The problem is barely noticeable on October 14, 2011, where volatility is high, but becomes increasingly visible on the other dates. We can report that the smiles for RB are almost indistinguishable from those of RB-F on October 14, 2011, and September 4, 2012, and that RB too fails to reproduce the term structure of smile (a)symmetry on January 13, 2017. That RB-F, RB, and QRH, calibrate comparably suggests that the main issue with typical one-factor rough volatility models in calibrating to SPX options is to recreate the term structure of smile (a)symmetry. It is then meaningful that the fits only change little when we add more flexibility to the kernel (fractional vs. gamma) or compare different distributions (log-normal vs. quadratic rough Heston).

Much better results are obtained with the two-factor Bergomi model (2F-B) which has errors of 7, 11 and 39 bps.<sup>24</sup> As we can observe in figures 4.5.2-3, the improvement is especially noticeable on the low volatility date. For the first volatility factor, we calibrate  $\rho_{12}$  in the range from -0.50 to -0.29 and  $\lambda_1$  in the range from 71.73 to 92.81. For the second factor, we calibrate  $\rho_{13}$  in the range from -0.97 to -0.93 and  $\lambda_2$  in the range from 1.17 to 2.89. Since  $\theta \approx 0.90$  on all dates, the first factor is much more volatile. The parameters are then in line with the hypothesis that we gave in Section 4.4.3 where we suggested the need for two volatility factors, one of which should mean-revert faster, be more volatile, and less negatively correlated with the index price. The version with the (rough) gamma kernel (2F-RB) results in essentially the same errors (7, 11 and 38 bps).

For the log-normal mixtures, we report only a single error in Table 4.5.2 and that for

<sup>23</sup> We find that it generally is for values  $\eta > 3$  that it becomes difficult to estimate the mean of  $Y := \exp(\eta X - \frac{1}{2}\eta^2)$  where  $X \sim \mathcal{N}(0, 1)$ . To illustrate the problem for the example of the main text note this: Let  $\eta = 5.84$  and  $N \in \mathbb{N}$ ,  $\hat{\mu} := \frac{1}{N} \sum_{i=1}^N Y_i$ , where  $Y_i, i = 1, \dots, N$ , are i.i.d. and distributed as  $Y$ . Sampling 10 000 values of  $\hat{\mu}$  where  $N = 50\,000$ , we get a median of 0.0812 and a 5–95 percentile interval of 0.0257–1.0627. This should be compared to the true mean which is 1. Convergence is also very slow in  $N$ : Sampling 10 000 values of  $\hat{\mu}$  where  $N = 10^6$ , we get a median of 0.2202 and a 5–95 percentile interval of 0.0985–1.4136.

<sup>24</sup> In [38] there is also an example of the (market) at-the-money term structure of skew that is more consistent with a two-factor Bergomi model than with a power-law.

## 4.5.2. CALIBRATION RESULTS

---

M-2F-RB calibrated on October 14, 2011. As it nests the model 2F-B, the error is unsurprisingly low at 3 bps. The models M-2F-QHyp and M-2F-QRHyp perform well on all three dates with errors of 4, 12 and 20 bps, respectively, 9, 5 and 22 bps. The model M-2F-RHyp calibrates worst among the two-factor models; its errors are 27, 15 and 65 bps. Given that M-2F-QRHyp attains much lower errors than M-2F-RHyp but uses the same kernel parameterisation, the poorer results for the latter model likely reflects a distributional problem and not a lack of flexibility for the volatility autocorrelations.

Our results show that substantial gains can be made with an additional factor—especially when volatility is low. If we compare 2F-B and M-2F-QHyp in Figure 4.5.2, we see that they both achieve essentially perfect fits on October 14, 2011, and September 4, 2012. We can report that 2F-RB and M-2F-QRHyp also do, although, we, as noted, leave out the fits to improve visibility. However, on January 13, 2017, differences emerge. Even if 2F-B reproduces some of the decoupled term structures of skew and curvature on this date, there is yet a very visible residual error; see Figure 4.5.3 in particular. In contrast, M-2F-QHyp attains an almost perfect fit on this date also. While not shown, we can report that also on January 13, 2017, are the volatility smiles for M-2F-QRHyp, respectively, 2F-RB, almost indistinguishable from those of M-2F-QHyp, respectively, 2F-B.

Since the model M-2F-QHyp, respectively, M-2F-QRHyp, uses the same kernel parameterisation as the model 2F-B, respectively, 2F-RB, but calibrates almost perfectly on the low volatility date, we believe it is unlikely that the poorer fits for the models 2F-B and 2F-RB result from a lack of flexibility for the volatility autocorrelations. Rather we believe the problem is distributional, and, possibly is related to the observation that the log-normal distribution is too light-tailed; c.f. the commentary on page 148. Another potential explanation is that the poorer fits result from a misspecified elasticity of variance-of-variance which part one of our analysis suggested for the one-factor rough Bergomi model. We can expect this to be more problematic when the term structure of volatility is more pronounced as it then, presumably, will be more difficult to get the effective volatility-of-volatility right at both short and long time horizons. Indeed, we find that the term structure of volatility is more marked on January 13, 2017: the values of  $\frac{1}{T} \int_0^T \xi_0(t) dt$  with  $T = 1$  (1 year) divided by those with  $T = \frac{5}{252}$  (1 week) are 2.14, 2.49 and 5.82 on respectively October 14, 2011, September 4, 2012, and January 13, 2017; these numbers are based on  $\xi_0$  as calibrated for the two-factor Bergomi model (2F-B).

It is worth highlighting again that 2F-RB and M-2F-QRHyp that both use the gamma kernel, calibrate about equally well to each of their non-rough counterparts 2F-B and M-2F-QHyp. Since M-2F-QHyp calibrates almost perfectly on all three dates, the conclusion then appears to be that while we do need two volatility factors to calibrate to SPX options, their autocorrelation structure need not be very complicated as even exponential kernels seemingly suffice. We believe this is a non-trivial observation given all the evidence for roughness, mostly notably in [33], and also the generally optimistic view of rough volatility that one is inclined to take on based on existing literature. The apparent success and popularity of one-factor rough volatility is likely a consequence of the fact that the rough fractional kernel generates mean-reversion at multiple time-scales even if driven by a single Brownian motion and which is in contrast to classical one-factor models whose autocorrelation structure likely is much too simple. It is then somewhat

meaningful that the differences disappear, or at least significantly diminish, when we consider two-factor models where even non-rough ones allow mean-reversion at both short and long time-scales. However, as two factors are needed in any case to capture the decoupled nature of the volatility surface, and because computational costs are worse under rough volatility, we find it hard to justify modelling volatility rough based on our results so far. As the reader will see, the remaining results are in line with this conclusion.

Let us stress that we cannot from our results conclude whether or not volatility *truly* is rough as that can only be finally answered in the asymptotic limit of very small time-scales. Rather the point to be made is that, when viewed in a two-factor framework, roughness does not appear to matter at the time-scales relevant to the pricing of SPX options and, as the reader will see, VIX options. The debate of roughness in many ways resembles that of long-memory in volatility which deals with the other asymptotic limit and for which the authors of [33] make a comparable observation noting that while their Rough Fractional Stochastic Volatility (RFSV) model does not truly display long-memory as is found in many works (see e.g. [10]) common statistical tests applied to sample paths under the model may result in a false positive detections of long memory.

In what follows, we discuss the joint calibrations. We focus on the models with log-normal terms first and start with October 14, 2011, where errors are reported for them all. With S-M-1F-RB we get an error of 197 bps which is significantly worst. As is evident from Figure 4.5.4, although the VIX smiles are captured reasonably well, too little skew is produced for short term SPX options. We take it as evidence that the model does not sufficiently decouple the variability of the spot variance from that the remainder of the forward variance curve on which the VIX index depends. We note that the model struggles to reconcile both markets despite that it calibrates a correlation parameter of<sup>25</sup>  $\rho = -1$  and uses the gamma kernel which nests the rough fractional kernel and allows mean-reversion at multiple time-scales. We find it unlikely that the problem should arise from the inability of the model to create a term structure of (a)symmetry of which there is little on this date or that the problem is distributional as the VIX smiles are fitted fairly well. In line with the results of Section 4.4.3, a more plausible explanation is that the gamma kernel, the fractional kernel in particular, as used in this model, lacks flexibility in decoupling the volatility autocorrelations at short and long lags. The same problem is reflected in [37] where a similar 'skewed' rough Bergomi model is calibrated.

The models S-2F-RB and S-M-2F-RB perform much better with roughly half the errors at 84 and 88 bps. In Figure 4.5.4, we see that the fit for S-2F-RB also visually is very decent. We do not show S-M-2F-RB in the plot but can report that the fit is comparable to that of S-2F-RB. For September 4, 2012, we only report errors for S-2F-RB among the log-normal type models. It nevertheless again performs well with an error of 70 bps and a very decent fit (Figure 4.5.5). It is worth noting that S-2F-RB calibrated to both of these dates, and S-M-2F-RB calibrated to October 14, 2011, all have  $\alpha_1 = \alpha_2 = 0$  so they in fact are two-factor classical models. This is more evidence that only two exponential terms are needed to capture the relevant volatility autocorrelations.

---

<sup>25</sup> Note that  $\rho$  allows us to influence the SPX skews independent of the VIX smiles.

## 4.5.2. CALIBRATION RESULTS

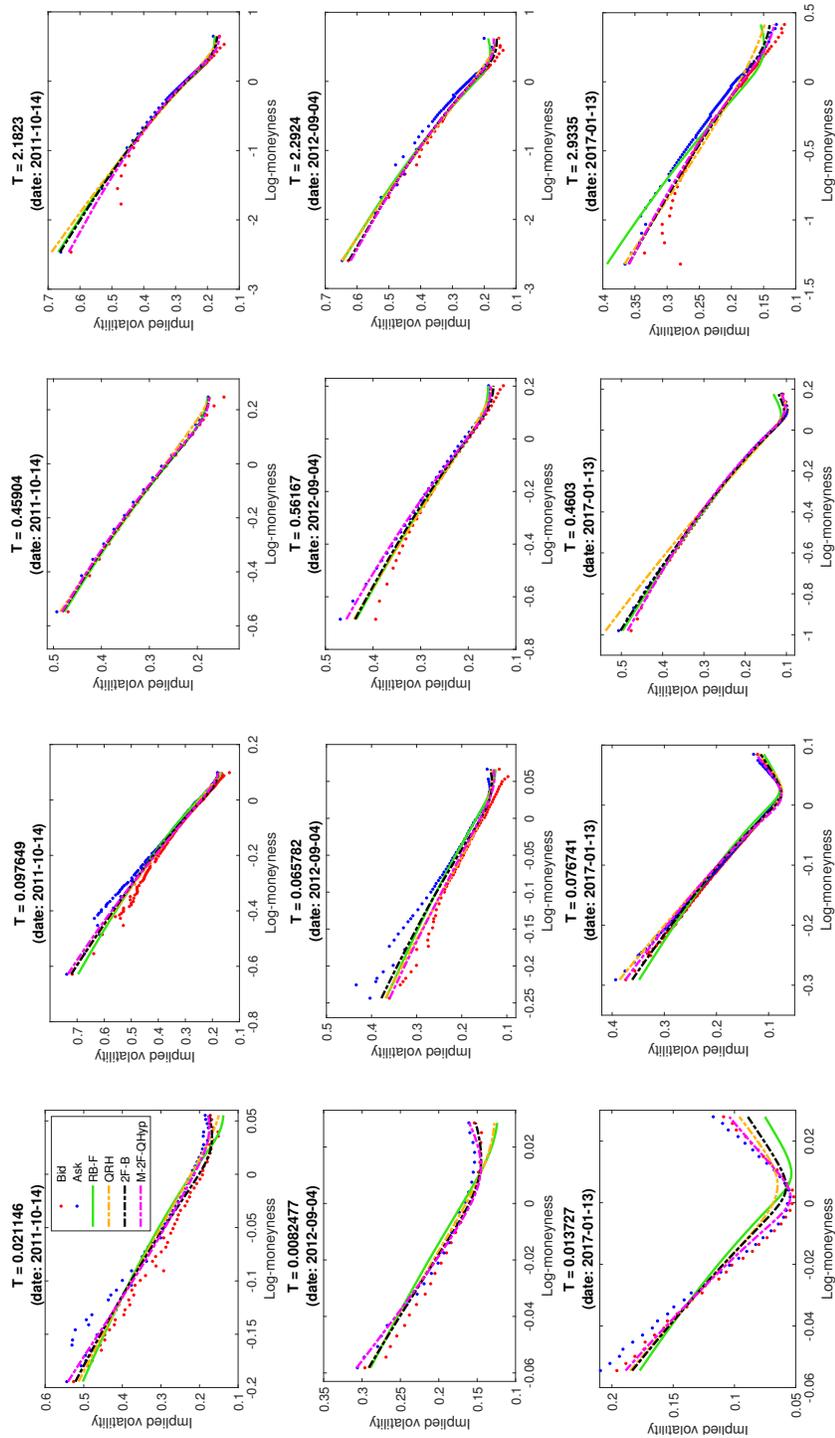


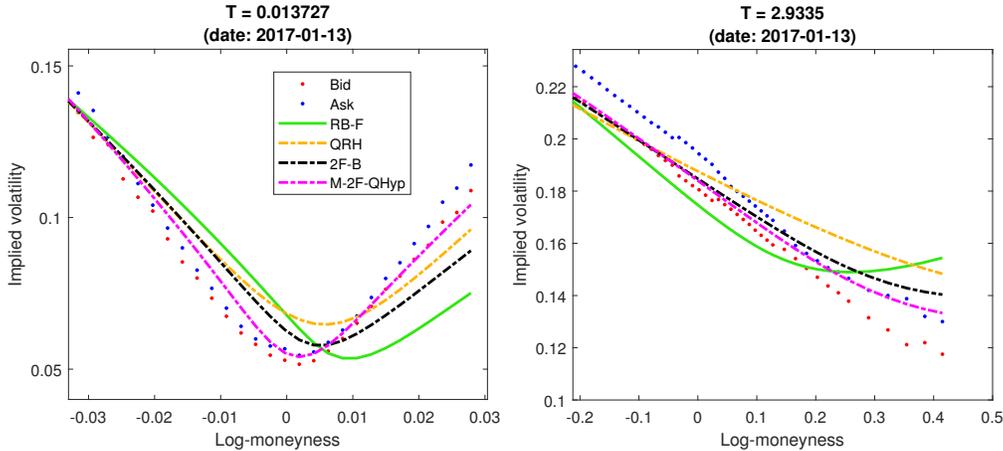
Figure 4.5.2: Calibration to SPX options (selected models and expiries).

CHAPTER 4. EMPIRICAL ANALYSIS OF ROUGH AND CLASSICAL STOCHASTIC VOLATILITY MODELS TO THE SPX AND VIX MARKETS

**Table 4.5.5:** Calibration to SPX and VIX options. Bar-notation denotes average values till 1 month. Errors are reported in basis points, though, some are left out where price estimation is deemed too difficult; we refer to the main text.

<b>S-2F-RB</b>	$\theta$	$\eta$	$\rho_{12}$	$\rho_{13}$	$\rho_{23}$	$\lambda_1$	$\lambda_2$	$\alpha_1$	$\alpha_2$	$c$	$\xi_0$	Error						
2011-10-14	0.93	2.50	-0.70	-0.94	0.41	106.66	0.83	0.00	0.00	0.0197	0.0940	84						
2012-09-04	0.86	2.57	-0.59	-0.94	0.83	59.92	3.65	0.00	0.00	0.0134	0.0305	70						
2017-01-13	0.93	3.03	-0.39	-0.97	0.15	56.55	1.94	0.00	-0.19	0.0083	0.0157	118						
<b>S-M-1F-RB</b>	$\mu$	$\eta_1$	$\eta_2$	$\rho$	$\lambda$	$\alpha$	$c$	$\xi_0$	Error									
2011-10-14	0.65	0.73	2.59	-1.00	0.03	-0.30	0.0000	0.1038	197									
2012-09-04	0.90	2.13	9.23	-0.97	1.41	-0.43	0.0134	0.0322	*									
2017-01-13	0.69	1.99	7.68	-0.96	0.46	-0.43	0.0129	0.0469	*									
<b>S-M-2F-RB</b>	$\mu$	$\theta_1$	$\theta_2$	$\eta_1$	$\eta_2$	$\rho_{12}$	$\rho_{13}$	$\rho_{23}$	$\lambda_1$	$\lambda_2$	$\alpha_1$	$\alpha_2$	$c$	$\xi_0$	Error			
2011-10-14	0.33	0.79	0.93	0.79	2.80	-0.89	-0.99	0.83	141.87	0.90	0.00	0.00	0.0120	0.0957	88			
2012-09-04	0.44	0.75	0.97	2.00	5.84	-0.51	-0.99	0.59	68.56	1.97	-0.05	-0.34	0.0150	0.0439	*			
2017-01-13	0.24	0.28	0.34	1.30	8.48	-1.00	-0.85	0.86	174.34	2.88	-0.35	-0.05	0.0102	0.0777	*			
<b>QRH</b>	$a$	$b$	$c$	$\rho$	$\lambda$	$\alpha$	$\xi_0$	Error										
2011-10-14	0.09	0.93	0.0075	1.00	0.16	-0.43	0.1550	54										
2012-09-04	0.07	0.74	0.0032	0.82	0.33	-0.44	0.3541	81										
2017-01-13	0.07	0.23	0.0009	0.94	2.21	-0.46	0.0417	89										
<b>S-M-2F-RHYP</b>	$\mu$	$\theta_1$	$\theta_2$	$\eta_1$	$\eta_2$	$\rho_{12}$	$\rho_{13}$	$\rho_{23}$	$\lambda_1$	$\lambda_2$	$\alpha_1$	$\alpha_2$	$c$	$\bar{\xi}_{1,0}$	$\bar{\xi}_{2,0}$	Error		
2011-10-14	0.13	1.00	0.10	5.10	7.14	-0.98	-0.56	0.36	0.00	1.05	-0.43	0.00	0.0000	-3.58	-0.99	157		
2012-09-04	0.26	0.77	0.15	18.00	7.35	-1.00	-1.00	1.00	2.28	0.03	-0.43	-0.04	0.0023	-4.85	-3.69	78		
2017-01-13	0.93	0.00	0.71	48.36	19.98	-0.98	-0.58	0.42	2.64	1.42	-0.40	-0.01	0.0076	-3.06	-2.35	120		
<b>S-M-2F-QHYP</b>	$\mu$	$\theta_1$	$\theta_2$	$\eta_1$	$\eta_2$	$d_1$	$d_2$	$\rho_{12}$	$\rho_{13}$	$\rho_{23}$	$\lambda_1$	$\lambda_2$	$c$	$\bar{\xi}_{1,0}$	$\bar{\xi}_{2,0}$	Error		
2011-10-14	0.31	0.06	0.11	0.54	0.98	3.47	14.80	0.96	0.70	0.46	1.46	118.55	0.0112	-0.71	-0.10	73		
2012-09-04	0.56	0.17	0.06	0.72	0.84	5.37	7.19	1.00	0.48	0.48	4.78	159.85	0.0081	-0.42	0.01	75		
2017-01-13	0.02	0.31	0.04	1.41	0.73	22.47	6.24	0.88	0.72	0.32	3.66	98.43	0.0000	-0.73	0.01	40		
<b>S-M-2F-QRHYP</b>	$\mu$	$\theta_1$	$\theta_2$	$\eta_1$	$\eta_2$	$d_1$	$d_2$	$\rho_{12}$	$\rho_{13}$	$\rho_{23}$	$\lambda_1$	$\lambda_2$	$\alpha_1$	$\alpha_2$	$c$	$\bar{\xi}_{1,0}$	$\bar{\xi}_{2,0}$	Error
2011-10-14	0.95	0.10	0.07	1.11	0.71	20.96	3.38	0.96	0.73	0.53	1.29	115.49	0.00	-0.02	0.0137	-0.58	-0.14	79
2012-09-04	0.19	0.29	0.23	1.04	0.74	9.90	6.53	1.00	0.46	0.45	1.03	22.87	-0.16	-0.25	0.0095	-1.37	0.01	61
2017-01-13	0.95	0.04	0.43	1.01	0.86	11.05	5.64	0.80	0.68	0.10	3.46	104.31	-0.01	-0.04	0.0009	-0.49	0.00	46

## 4.5.2. CALIBRATION RESULTS



**Figure 4.5.3:** Calibration to SPX options (zoomed examples).

On January 13, 2017, we, among the log-normal type models, again only report an error for S-2F-RB. However, on this date, it calibrates much worse with an error of 118 bps and a notably poor fit (Figure 4.5.6). While the VIX smiles are fitted fairly well, the short term SPX smiles are too flat. We believe the problem is distributional because S-M-2F-QRHyp, which is driven by the same two Gaussian processes, attains an error of only 46 bps; we discuss the hyperbolic models more momentarily. The problem is reflected in the shift parameter that is calibrated to  $c = 0.0083$  (9.11% in volatility terms) and which is high relative to the volatility level (the VIX index is at 11.23). This indicates that S-2F-RB overrelies on the volatility shift to skew the VIX smiles. As shown in Figure 4.5.1, that can precisely flatten the SPX skews. Since S-2F-RB calibrates poorly on the low volatility date and the log-normal mixture type models have estimation problems, we conclude that the log-normal distribution, all in all, is ill-suited for the joint problem.

Next, we consider the quadratic rough Heston model which, with errors of 54, 81, and 89 bps, performs better overall compared to S-2F-RB and S-M-2F-RB. Figures 4.5.4-6 show that it captures both the SPX *skews* and the VIX smiles; that is, besides the very short term SPX skew on January 13, 2017. This is impressive given that the model only has six parameters besides the curve  $\zeta_0$  and only one Brownian motion driving volatility. It is somewhat surprising that it performs so much better than S-M-1F-RB on October 14, 2011, (the only date we can compare on) which is also one-factor with a gamma kernel. We suspect that QRH calibrates fairly well to the joint problem because of the squaring operation. More precisely, we believe it might increase mean-reversion by amplifying relative differences between small and large values thereby helping to decouple the SPX skews from the level of the VIX implied volatilities. Nevertheless, being one-factor, the model is yet unable to reproduce the term structure of smile (a)symmetry that is very visible on January 13, 2017. We conclude that the model, perhaps, is a bit *too* simple.

We turn our attention to the hyperbolic models: We note that S-M-2F-RHyp, overall, calibrates poorly with errors of 157, 78 and 120 bps. This is unsurprising given the

also somewhat unsatisfactory calibrations of M-2F-RHyp to SPX options. Much better results are achieved with S-M-2F-QHyp which has errors of 73, 75 and 40 bps. Figures 4.5.4-6 show that the fits themselves visually are also very decent on all three dates. Compared to quadratic rough Heston, we note that the fit is just slightly worse on the high volatility date, though still fairly decent, but notably better on the low volatility date. The rough version S-M-2F-QRHyp performs similarly with errors 79, 61 and 46 bps.

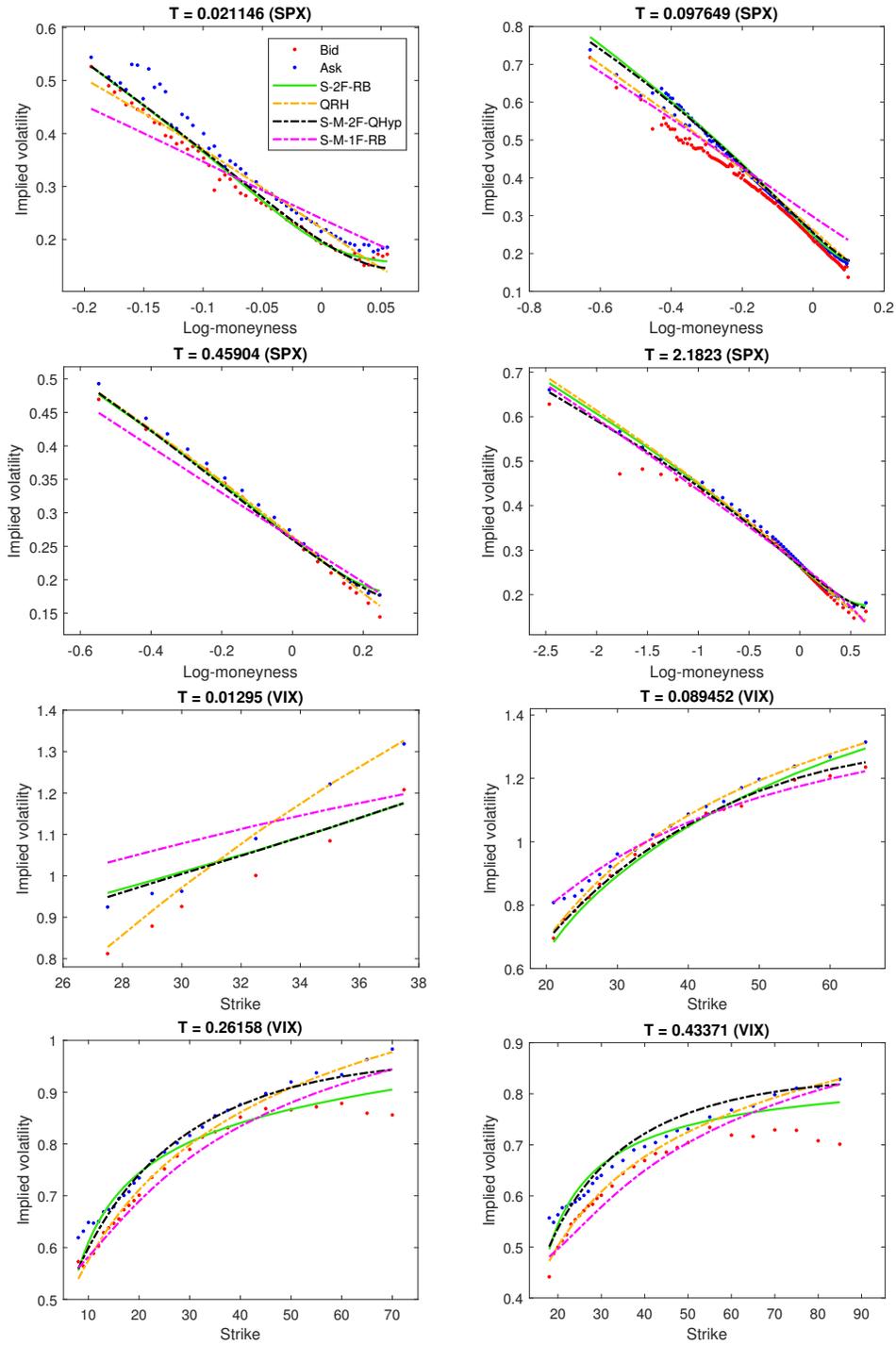
The very decent and robust calibrations of S-M-2F-QHyp suggests that the joint calibration problem is largely solvable with classical two-factor volatility. The observation goes against the idea that jumps or roughness is needed to calibrate jointly. As the analysis is based on a small sample size, the conclusion should though be taken with some moderation. We leave a more extensive historical calibration of the model for future work.

We end with a simulation example of S-M-2F-QHyp as calibrated jointly to SPX and VIX options on January 13, 2017; we refer to Figure 4.5.7 where we show a trajectory of  $S_t$  and  $VIX_t$  across 15 years. Despite that volatility is driven by two ordinary Ornstein-Uhlenbeck processes, the model is able to generate jump-like behaviour for the S&P 500 index characteristic of rough volatility. The trajectory of the VIX index also qualitatively resembles the historical trajectory shown in Figure 4.4.1.

## 4.6 Conclusion

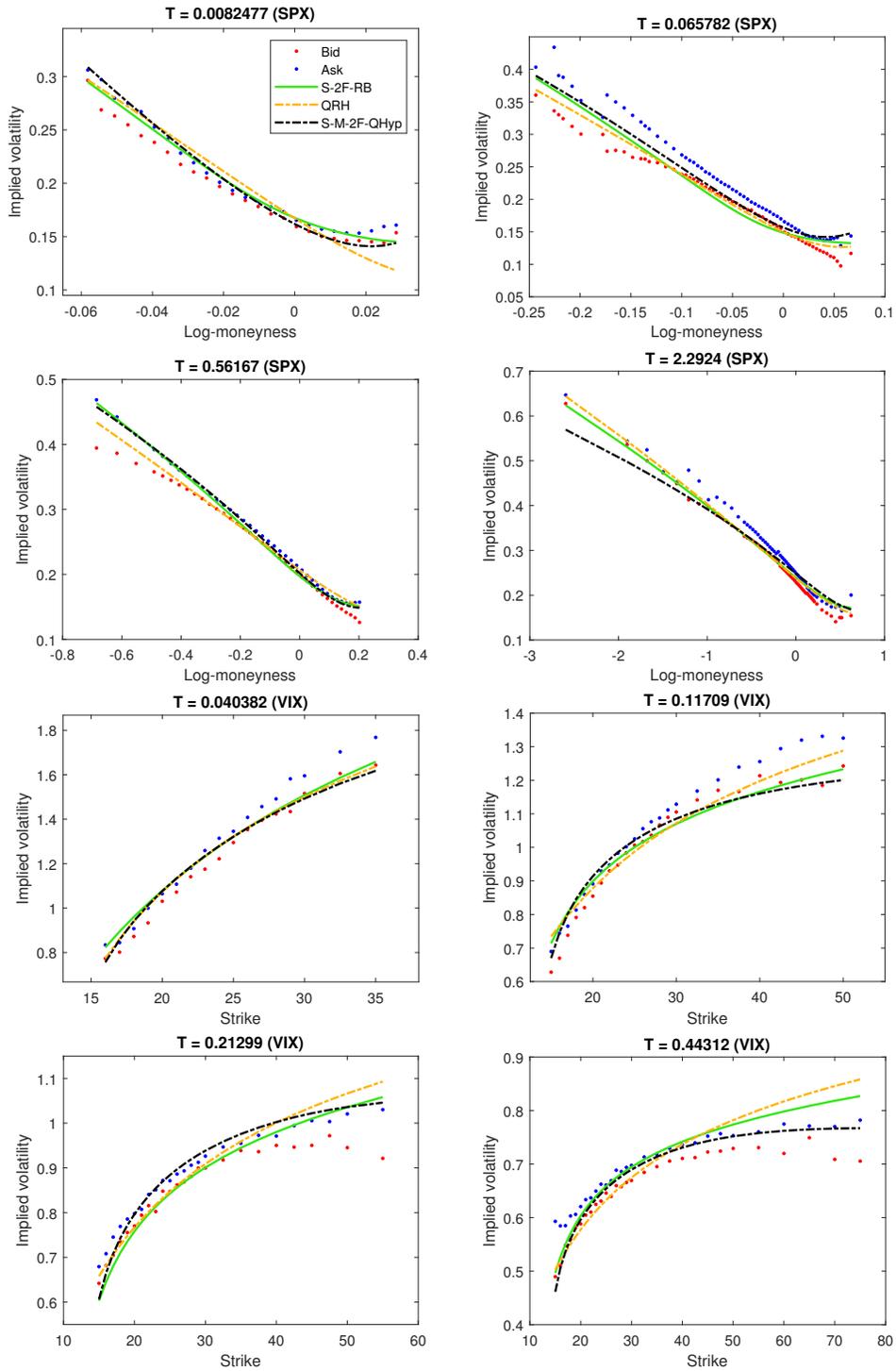
We have in this paper shown that the (one-factor) rough Bergomi model fails to create a pronounced enough term structure of smile effect for SPX options. Moreover, we have discovered that short expiry SPX smiles, generally, are more symmetric than long expiry smiles, a feature we found the rough Bergomi model neither could recreate. To calibrate to SPX options with a continuous-path stochastic volatility model our results suggests that two factors are needed for volatility, three for the S&P 500 index. One volatility factor should be less negatively correlated with the index price, more noisy, and mean-revert faster. Once extended to a two-factor volatility framework, we found no meaningful differences between comparable rough and non-rough models and obtained decent fits with even a (log-normal, non-rough) two-factor Bergomi model. The two-factor Bergomi model was though still not fully aligned with market prices on an example date with low volatility. Furthermore, we were unable to calibrate (across volatility levels) jointly to SPX and VIX options with models based on the log-normal distribution. We considered several alternative models, including a certain two-factor volatility model based on the quadratic and hyperbolic transformations and which is driven by two ordinary Ornstein-Uhlenbeck processes. We demonstrated that this model can calibrate almost perfectly to SPX options and very decently to the joint problem—also across volatility levels. From this we conclude that the SPX and VIX options markets can largely be reconciled with classical two-factor volatility, all without roughness and jumps.

## 4.6. CONCLUSION



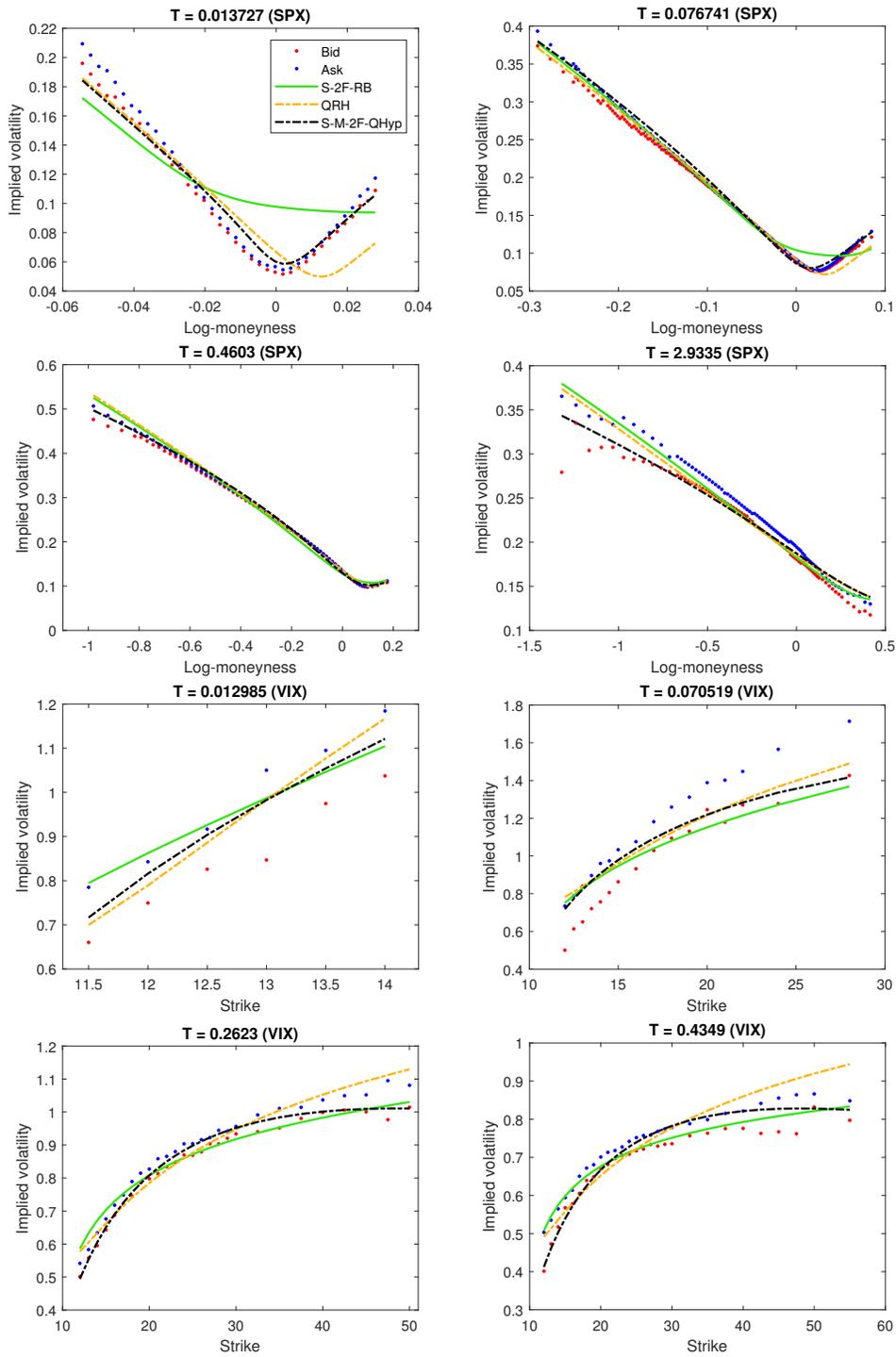
**Figure 4.5.4:** Joint calibration to SPX and VIX options (selected models and expiries). Date: 2011-10-14.

CHAPTER 4. EMPIRICAL ANALYSIS OF ROUGH AND CLASSICAL STOCHASTIC VOLATILITY MODELS TO THE SPX AND VIX MARKETS

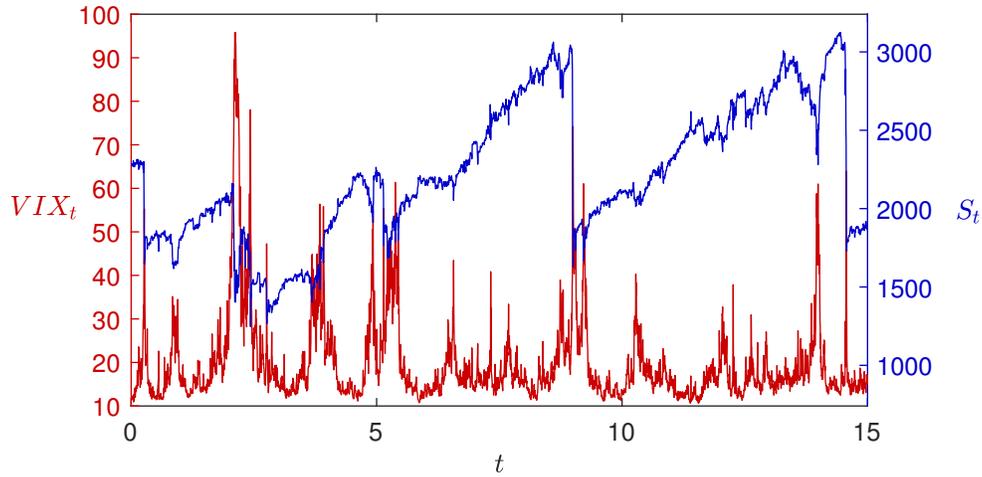


**Figure 4.5.5:** Joint calibration to SPX and VIX options (selected models and expiries). Date: 2012-09-04.

## 4.6. CONCLUSION



**Figure 4.5.6:** Joint calibration to SPX and VIX options (selected models and expiries). Date: 2017-01-13.



**Figure 4.5.7:** Simulation of S-M-2F-QHyp as calibrated to SPX and VIX options on January 13, 2017. The curves  $\zeta_{1,0}$  and  $\zeta_{2,0}$  are extended flat and  $S_t$  is simulated without drift.

## References

- [1] Abi Jaber, E., Lifting the Heston model. *Quantitative Finance*, 2019, **19**(12), 1995–2013.
- [2] Abi Jaber, E., El Euch, O., Markovian structure of the Volterra Heston model. *Statistics and Probability Letters*, 2019, **149**, 63–72.
- [3] Abi Jaber, E., El Euch, O., Multifactor approximation of rough volatility models. *SIAM Journal on Financial Mathematics*, 2019, **10**(2), 309–349.
- [4] Abi Jaber, E., Larsson, M., Pulido, S., Affine Volterra processes. *The Annals of Applied Probability*, 2019, **29**(5), 3155–3200.
- [5] Alòs, E., García-Lorite, D., Muguruza, A., On smile properties of volatility derivatives: Understanding the VIX skew. *SIAM Journal on Financial Mathematics*, 2022, **13**(1), 32–69.
- [6] Bayer, C., Friz, P. K., Gassiat, P., Martin, J., Stemper, B., A regularity structure for rough volatility. *Mathematical Finance*, 2020, **30**(3), 782–832.

## REFERENCES

---

- [7] Bayer, C., Friz, P., Gatheral, J., Pricing under rough volatility. *Quantitative Finance*, 2016, **16**(6), 887–904.
- [8] Bayer, C., Friz, P. K., Gulisashvili, A., Horvath, B., Stemper, B., Short-time near-the-money skew in rough fractional volatility models. *Quantitative Finance*, 2019, **19**(5), 779–798.
- [9] Bayer, C., Stemper, B., Deep calibration of rough stochastic volatility models (2018). Working paper. Available online at arXiv:1810.03399v1 (accessed July 27, 2021).
- [10] Bennedsen, M., Lunde, A., Pakkanen, M. S., Decoupling the short- and long-term behavior of stochastic volatility. *Journal of Financial Econometrics*, 2021, 1–46.
- [11] Bennedsen, M., Lunde, A., Pakkanen, M. S., Decoupling the short- and long-term behavior of stochastic volatility (2017). Working paper. Available online at arXiv:1610.00332v2 (accessed July 27, 2021).
- [12] Bennedsen, M., Lunde, A., Pakkanen, M. S., Hybrid scheme for Brownian semistationary processes. *Finance and Stochastics*, 2017, **21**, 931–965.
- [13] Bergomi, L., Smile dynamics III. *Risk*, October 2008, pp. 90–96.
- [14] Bergomi, L., Guyon, J., Stochastic volatility’s orderly smiles. *Risk*, May 2012, pp. 60–66.
- [15] Bishop, C. M., *Pattern Recognition and Machine Learning*, Springer, New York, 2006.
- [16] Carr, P., Madan, D. B., Option valuation using the fast Fourier transform. *Journal of Computational Finance*, 1999, **2**(4), 61–73.
- [17] Chung, S.-L., Tsai, W.-C., Wang, Y.-H., Weng, P.-S., The information content of the S&P 500 index and VIX options on the dynamics of the S&P 500 index. *The Journal of Futures Markets*, 2011, **31**(12), 1170–1201.
- [18] Cont, R., Kokholm, T., A consistent pricing model for index options and volatility derivatives. *Mathematical Finance*, 2013, **23**(2), 248–274.
- [19] De Marco, S., Volatility derivatives in (rough) forward variance models. Stochastic Analysis and its Applications, Oaxaca, May 2018.
- [20] El Euch, O., Fukasawa, M., Gatheral, J., Rosenbaum, M., Short-term at-the-money asymptotics under stochastic volatility models. *SIAM Journal on Financial Mathematics*, 2019, **10**(2), 491–511.
- [21] El Euch, O., Fukasawa, M., Rosenbaum, M., The microstructural foundations of leverage effect and rough volatility. *Finance and Stochastics*, 2018, **22**(2), 241–280.
- [22] El Euch, O., Gatheral, J., Rosenbaum, M., Roughening Heston. *Risk*, May 2019b, pp. 84–89.
- [23] Forde, M., Gerhold, S., Smith, B., Small-time, large-time, and  $H \rightarrow 0$  asymptotics for the rough Heston model. *Mathematical Finance*, 2021, **31**(1), 203–241.

- 
- [24] Forde, M., Smith, B., Viitasaari, L., Rough volatility and CGMY jumps with a finite history and the Rough Heston model – small-time asymptotics in the  $k\sqrt{t}$  regime. *Quantitative Finance*, 2021, **21**(4), 541-563.
- [25] Forde, M., Zhang, H., Asymptotics for rough stochastic volatility models. *SIAM Journal on Financial Mathematics*, 2017, **8**(1), 114–145.
- [26] Friz, P. K., Gassiat, P., Pigato, P., Precise asymptotics: Robust stochastic volatility models. *The Annals of Applied Probability*, 2021, **31**(2), 896–940.
- [27] Friz, P. K., Gassiat, P., Pigato, P., Short-dated smile under rough volatility: Asymptotics and numerics. *Quantitative Finance*, 2021, DOI:10.1080/14697688.2021.1999486.
- [28] Fukasawa, M., Short-time at-the-money skew and rough fractional volatility. *Quantitative Finance*, 2017, **17**(2), 189–198.
- [29] Gulisashvili, A., Time-inhomogeneous Gaussian stochastic volatility models: Large deviations and super roughness. *Stochastic Processes and their Applications*, 2021, **139**, 37–79.
- [30] Gatheral, J., Jacquier, A., Arbitrage-free SVI volatility surfaces. *Quantitative Finance*, 2014, **14**(1), 59–71.
- [31] Gatheral, J., Consistent modeling of SPX and VIX options. *The Fifth World Congress of the Bachelier Finance Society*, London, 18 July, 2008.
- [32] Gatheral, J., Jusselin, P., Rosenbaum, M., The quadratic rough Heston model and the joint S&P 500/Vix smile calibration problem. *Risk*, May 2020.
- [33] Gatheral, J., Jaisson, T., Rosenbaum, M., Volatility is rough. *Quantitative Finance*, 2018, **18**(6), 933–949.
- [34] Gatheral, J., *The Volatility Surface: A Practitioner’s Guide*, John Wiley & Sons, New Jersey, 2006.
- [35] Goodfellow, I., Bengio, Y., Courville, A., *Deep Learning*, MIT Press, 2016.
- [36] Guyon, J., Inversion of convex ordering in the VIX market. *Quantitative Finance*, 2020, **20**(10), 1597–1623.
- [37] Guyon, J., On the joint calibration of SPX and VIX options. QuantMinds, Lisbon, 16 May, 2018.
- [38] Guyon, J., The VIX future in Bergomi models: Analytic expansions and joint calibration with S&P 500 skew (2020). Working paper. Available online at SSRN:3720315 (accessed July 27, 2021).
- [39] Hagan, P. S., Kumar, D., Lesniewski, A. S., Woodward, D. E., Managing smile risk. *Wilmott Magazine*, January 2002, pp. 84–108.

## REFERENCES

---

- [40] Heston, S. L., A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 1993, **6**(2), 327–343.
- [41] Hornik, K., Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 1991, **4**(2), 251–257.
- [42] Horvath, B., Jacquier, A., Tankov, P., Volatility options in rough volatility models. *SIAM Journal on Financial Mathematics*, 2020, **11**(2), 437–469.
- [43] Horvath, B., Muguruza, A., Tomas, M., Deep learning volatility: A deep neural network perspective on pricing and calibration in (rough) volatility models. *Quantitative Finance*, 2021, **21**(1), 11–27.
- [44] Jacquier, A., Martini, C., Muguruza, A., On VIX futures in the rough Bergomi model. *Quantitative Finance*, 2018, **18**(1), 45–61.
- [45] Kahl, C., Jäckel, P., Fast strong approximation Monte Carlo schemes for stochastic volatility models. *Quantitative Finance*, **6**(6), 513–536.
- [46] Kidger, P., Lyons, T., Universal approximation with deep narrow networks. *Proceedings of Machine Learning Research*, 2020, **125**, 2306–2327.
- [47] Kingma, D. P., Ba, J., Adam, A method for stochastic optimization (2015). Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego. Available online at arXiv:1412.6980v9 (accessed July 27, 2021).
- [48] Kokholm, T., Stisen, M., Joint pricing of VIX and SPX options with stochastic volatility and jump models. *Journal of Risk Finance*, 2015, **16**(1), 27–48.
- [49] Lacombe, C., Muguruza, A., Stone, H., Asymptotics for volatility derivatives in multi-factor rough volatility models. *Mathematics and Financial Economics*, 2021, **15**, 545–577.
- [50] Lewis, A. L., A simple option formula for general jump-diffusion and other exponential Levy processes (2001). Working paper. Available online at SSRN:282110 (accessed July 27, 2021).
- [51] Lord, R., Kahl, C., Optimal Fourier inversion in semi-analytical option pricing. Tinbergen Institute Discussion Paper No. 2006-066/2. Available online at SSRN:921336 (accessed July 27, 2021).
- [52] McCrickerd, R., Pakkanen, M. S., Turbocharging Monte Carlo pricing for the rough Bergomi model. *Quantitative Finance*, 2018, **18**(11), 1877–1886.
- [53] McGhee, W. A., An artificial neural network representation of the SABR stochastic volatility model (2018). Working paper. Available online at SSRN:3288882 (accessed July 27, 2021).
- [54] Pacati, C., Pompa, G., Renò, R., Smiling twice: The Heston++ model. *Journal of Banking and Finance*, 2018, **96**, 185–206.

- [55] Pacati, C., Renò, R., Santilli, M., Heston model: Shifting on the volatility surface. *Risk*, November 2014.
- [56] Roper, M., Arbitrage free implied volatility surfaces. Preprint, School of Mathematics and Statistics, The University of Sydney, Australia, 2010.
- [57] Rømer, S. E., Poulsen, R., How does the volatility of volatility depend on volatility? *Risks*, 2020, **8**(2), 59.
- [58] Rømer, S. E., Hybrid multifactor scheme for stochastic Volterra equations with completely monotone kernels (2022). Working paper. Available online at SSRN:3706253 (accessed March 10, 2022).
- [59] Scott, L. O., Option pricing when the variance changes randomly: Theory, estimation, and an application. *The Journal of Financial and Quantitative Analysis*, **22**(4), 419–438, 1987.

