FINANCIAL MATHEMATICS TEAM CHALLENGE

A collection of the three reports from the 2022 Financial Mathematics Team Challenge.

AIFMRM
AFRICAN INSTITUTE OF FINANCIAL MARKETS AND RISK MANAGEMENT

University of Cape Town • iYunivesithi yaseKapa • Universiteit van Kaapstad

SPES BONA

Preamble

One of the key aims of the FMTC is for South African postgraduate students in Financial and Insurance Mathematics to have the opportunity to focus on a topical, industry-relevant research project, while simultaneously developing links with international students and academics in the field. An allied purpose is to bring a variety of international researchers to South Africa to give them a glimpse of the dynamic environment at UCT in the African Institute of Financial Markets and Risk Management. The primary goal, however, remains for students to learn to work in diverse teams and to be exposed to a healthy dose of fair competition.

The Seventh Financial Mathematics Team Challenge was held from the 27 th of June to the 6 th of July 2022. The challenge this year was unusual because of the lingering effects of the pandemic. All twelve participants were drawn from the 2022 MPhil in Mathematical Finance class at UCT. There were no PhD students involved, so one MPhil student in each team took over the demanding role of the Team Leader. To compensate for the absence of PhD students, each team had two mentors who guided the research on a distinct problem over the ten days. Professional and academic experts from Germany, Australia, South Africa, and the UK mentored the teams, fostering teamwork and providing guidance. As they have in the past, the students applied themselves with remarkable commitment and energy.

This years research included topical projects on (a) South African interest rate parity arbitrage, (b) volatility surface updating, and (c) on using Gaussian mean mixtures for pricing American options. These were chosen from areas of current relevance to the finance and insurance industry. To prepare the teams, guidance and preliminary reading was given to them a month before the meeting in Cape Town. During the final day of the challenge, the teams presented their conclusions and solutions in extended seminar talks. The team whose research findings were adjudged to be the best was awarded a floating trophy. Each team wrote a report containing a critical analysis of their research problem and the results that they obtained. This volume contains these reports, and it will be available to future FMTC participants. It may also be of use and inspiration to Masters and PhD students in Financial and Insurance Mathematics.

FMTC VII was a great success, so 2023 and FMTC VIII is already in the pipeline!

David Taylor, *University of Cape Town*
Andrea Macrina, *University College London & University of Cape Town*

Contents

[1]Winning team of the seventh Financial Mathematics Team Challenge

Covered Interest Parity Arbitrage

TEAM 1

JUSTIN FOUCHÉ, University of Cape Town
SANDISWA DLAMINI, University of Cape Town
TEBOGO MAMPURU, University of Cape Town
DANAE PAVLOU, University of Cape Town

Supervisors:
ALEX BACKWELL, University of Cape Town
ANDREA MACRINA, UCL & UCT

African Collaboration for Quantitative Finance and Risk Research

Contents

1 Introduction

The concept of Covered Interest Parity (CIP) is built on the "no-arbitrage" argument which is fundamental in financial markets. It shows that the implicit interest rates in the foreign exchange swap market coincide with corresponding interest rates in the domestic market. It should not be possible for one to earn a risk-free profit from borrowing in South African Rand (ZAR), and lending in another currency, e.g., United States Dollar (USD), while simultaneously covering the foreign exchange risk, using a foreign exchange swap. In the case where this relationship breaks an arbitrage opportunity arises. This opportunity can be exploited by borrowing at the relatively lower interest rate between the two markets and lending at the higher interest rate with the foreign exchange risk fully hedged by the foreign exchange swap.

We would expect that when there is an arbitrage trade, professional arbitrageurs would trade that strategy and eventually the arbitrage opportunity would evaporate. However, many experts state that there is persistent deviations from CIP. In the G10 countries, the CIP has been consistently violated ever since the 2008 Global Financial Crisis, and it has led to significant arbitrage opportunities in currency markets Du et al. [2018]. Furthermore, it was shown that the spread emerging from violating CIP remained large and persisted, even against the most liquid currencies Du et al. [2018].

There are several explanations as to why this arbitrage is not exploited and thus persists. A possible reason for the persistent CIP deviation can be explained by the regulation that market entities face. Du et al. [2018], found that regulations lead to tightened balance sheets for banks which discourage trading the CIP basis. For example, it was seen that when companies are close to filing their reports, a widening of the CIP deviation was observed. Furthermore, regulators require these firms to maintain a certain amount capital reserves that is commensurate with their balance sheets. Another possible explanation as to why the CIP violation persists is market illiquidity. In illiquid markets the bid/offer spreads tend to be large and financial entities cannot take advantage of the violation as the spread erodes the arbitrage. During 1984 to 1987, De J. Correia and Knight [1987] attributed the CIP violation in South Africa to transaction cost.

Andersen et al. [2019] state that the firm's default risk could deter the firm from exploiting the opportunity created by the CIP violation. If a risky firm needs borrowing to set up a financial strategy that exploits the arbitrage, the firm will have to pay a spread arising from credit risk. This report shall focus on default risk of the firm as a potential explanation for why arbitrage persists in a CIP strategy. Du et al. [2018] found that as USD appreciates against other currencies, the CIP violation increases. This could be attributed to the increased transaction costs involved in trading USD, as it appreciates, which deters professional arbitrageurs from exploiting the strategy.

In looking at why the CIP violations persist, Andersen et al. [2019] was consulted. In their work, the focus is on how funding value adjustments (FVA) influence investment decisions. They found that debt overhang costs to shareholders are one of the key reasons why the arbitrage opportunity persists. When the default risk of the firm is accounted for, the arbitrage is eliminated in most cases. The only time the arbitrage persists is when the firm has a very high survival probability.

Andersen et al. [2019] look at the impact that the FVA has on investment decisions. They develop a model that demonstrates the gain achieved by the firm's shareholders for a specific investment strategy. Their model shows there is an associated cost with implementing an investment strategy and a value adjustment needs to be made to the gains to account for the strategy funding. Their proposed model is a single-period model, where the impact of regulatory requirements, transaction costs and liquidity of the market are ignored when calculating the marginal gain achieved by the strategy.

In this report, our goal is to derive the valuation adjustment arising from the funding costs associated with a foreign exchange (FX) strategy. While Andersen et al. [2019] adopt a general view that could apply to any instrument payoff involved in the arbitrage strategy, our focus is on a specific CIP trading strategy. Despite our setup being a special case, we nevertheless are able to derive, from first principles, the equation to calculate the marginal value to shareholders of debt financing. We consider borrowing directly from the South African market, synthetically lending rand and simultaneously entering a foreign exchange contract (FEC) to fix the exchange rate.

XVA is an umbrella term referring to a number of different "valuation adjustments" that banks make when assessing the value of derivative contracts. Since the 2008 Financial Crisis, credit risk has become a fundamental part of financial trading. The no-default value of a derivative transaction relies on both parties living up to their obligations. When a derivative's credit risk exposure is not collateralised, a credit-valuation adjustment (CVA) is needed to allow for the possibility that the counterparty defaults. Debit-valuation adjustment (DVA), which is analogous to CVA, is the adjustment needed to account for the effects of an arbitrageur defaulting.

FVA is the adjustment to the no-arbitrage value of an uncollateralised derivative that is designed to ensure that a arbitrageur, trading and hedging derivatives, recovers the average funding costs. According to Hull and White [2014], if this adjustment is not made, a loss will be shown for trades that require funding. Intuitively, for trades that create funding, for example the sale of a derivative, an FVA is a benefit as such trades reduce the external funding requirements of a bank.

Within the context of Andersen et al. [2019], the FVA is the cost that arises from the default risk of the firm. There is no other source of risk. Under the main result in Andersen et al. [2019], the FVA is defined as the present value to shareholders of their share of the net financing cost uS, where u is the marginal purchase price of an asset, and S is the firm's credit spread. It is further stated that shareholders pay these financing costs if and only if the firm survives. The main result in Andersen et al. [2019] reflects that even an investment whose up-front cost is strictly below the market value of an asset may be rejected by a firm once the FVA is incorporated. This can be used to explain why the CIP violation seems to persist.

The CIP bases calculated in Section 4.1 have an absolute value of 19 basis points on average across the sample period. Section 4.2 considers longer maturities and results in CIP bases with an absolute value of 85 basis points on average across the period. The CIP bases are positive and negative over the period under consideration, with longer maturities having a tendency to be positive. These averages are lower when considering bid/ask spreads. The net CIP basis remaining after bid/ask spreads are accounted for is the realistic CIP basis that could potentially be exploited as an arbitrage opportunity. The survival probabilities calculated in Section 4 that result in a break-even net position are consistent with realistic levels of bank default. This suggests that potential profits from CIP violations are eroded by the FVA which is driven by these probabilities.

2 Theory

2.1 Funding valuation adjustment

Andersen et al. [2019] consider the impact of a firm setting up a new financial position from the perspective of the firm's shareholders. They let Y denote the position's payoff at maturity and S represent the credit spread of the firm, which would apply to the financing of the new position. Letting G denote the value to the firm's shareholders of adding the new position, we state the main result in Andersen et al. [2019], p.155:

Proposition (The Marginal Value to Shareholders of Debt Financing): The marginal shareholder gain G in equity value is given by,

$$G = p^*\pi - \delta\mathrm{Cov}(\mathbb{I}_D, Y) - \Phi, \tag{1}$$

where

- p^* is the risk-neutral survival probability of the firm,
- δ is the discount factor for the period corresponding to the maturity of the new position,
- $\mathbb{I}_D$ is an indicator for the firm's default,
- $\pi = \delta\mathbb{E}(Y) - u$ is the marginal profit on the trade for a hypothetical risk-free arbitrageur and,
- $\Phi = p^*\delta uS$ is defined to be the FVA.

The FVA is the present value of the shareholders' share of the net financing cost uS, which is only paid if the firm does not default. The covariance term in Eq. (1) is an additional adjustment, applicable if there is correlation between the firm's potential default and the new financial instrument with payoff Y.

2.2 CIP arbitrage strategy

In the CIP strategy the domestic entity is South Africa, with ZAR as the currency, and the foreign entity is the US, with USD as the foreign currency. If the direct and synthetic ZAR positions have the same credit qualities, then the return on the direct and synthetic positions should, in theory, be the same if trade frictions are ignored Keynes [1923]. This equivalence is referred to as CIP. If the equivalence underpinning CIP is broken, then the resulting spread is referred to as CIP basis.

We assume that the initial time is $t = 0$ and maturity time is $t = \tau + 2\beta$, where settlement is at $t = 0 + 2\beta$, where β is a business day. Going long would result in the following strategy:

Strategy 2.1.

1. *Borrow ZAR at the interest rate r_d, directly in South African market.*

2. *Sell ZAR against USD FX Spot (which is denoted by X) to obtain u/S_t USD.*

3. *Invest the USD at the currently available USD interest rate r_f, in USD money market and simultaneously enter an FEC contract, with a forward spot exchange rate X^f, reversing the currency exchange at a predetermined price in the future.*

4. *At maturity, collect the ZAR from the FEC contract and repay the ZAR debt of $(1 + r_d\tau)$.*

Fig. (1) is a visual demonstration of the Strategy 2.1.

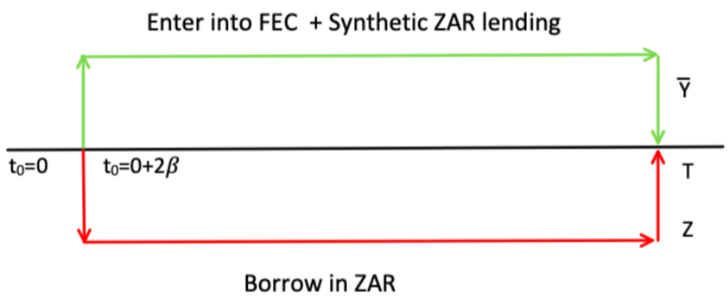


Figure 1: Simple FX swap strategy.

At maturity, the strategy will create a negative cashflow Z corresponding to the debt repayment, denoted as $Z = (1 + r_d\tau) = u(1 + (r + S)\tau)$ and a positive cashflow $\overline{Y}$ corresponding to the investment, denoted as return of $\overline{Y} = (1+r_f\tau)X^f/X = u(1+(r+S+b)\tau)$. Here b is the basis, which represents the size of the violation, r is risk-free rate at which the cashflows grow and S is the credit spread, which accounts for the credit riskiness of the firm. Here we assume that the firm and counterparty have the same credit spread.

However, these cashflows assume that it is impossible for the firm or its counterparty to default. When the realistic possibility of default is incorporated, the cashflows become random. Let X be a random variable representing the gains the arbitrageur will receive from the strategy,

$$X = \begin{cases} \overline{Y} - Z, & \text{if firm and counterparty survive,} \\ \psi\overline{Y} - Z, & \text{if counterparty defaults,} \\ 0, & \text{if firm defaults.} \end{cases}$$

Here ψ is the recovery rate in the case where the firm's counterparty defaults. It is important to note that,

$$\mathbb{E}\Big[\delta\big(\psi(1+\tau r+\tau S)+(1-\psi)(1+\tau r+\tau S)\mathbb{I}_s^{cp}\big)\Big]=1. \tag{2}$$

Now, the marginal gain G is defined as the expected discounted gains in the case where the firm survives, that is

$$G=\mathbb{E}[\delta\mathbb{I}_s^f(Y-u(1+\tau r+\tau S))]$$

where δ is the discount factor, $\mathbb{I}_s^f$ is the firm's survival indicator, and

$$Y=u\psi(1+\tau r+\tau S+\tau b)+u(1-\psi)(1+\tau r+\tau S+\tau b)\mathbb{I}_s^{cp}, \tag{3}$$

where $\mathbb{I}_s^{cp}$ is the counterparty survival indicator.

2.3 Derivation of Proposition 1 under the Strategy 2.1

In the case where the firm does not default, the expected discounted gains are given by

$$G=\mathbb{E}[\delta\mathbb{I}_s^f(Y-u(1+\tau r+\tau S))],$$

where $\delta=1/(1+\tau r)$ is the discount factor and $Y=u\psi(1+\tau r+\tau S+\tau b)+u(1-\psi)(1+\tau r+\tau S+\tau b)\mathbb{I}_s^{cp}$. Thus,

$$G=\mathbb{E}[\delta\mathbb{I}_s^fY]-\mathbb{E}[\delta\mathbb{I}_s^fu(1+\tau r+\tau S)].$$

Because $\mathbb{E}[\mathbb{I}_s^f]=p^*$, the above equation simplifies to

$$G=\mathbb{E}[\delta\mathbb{I}_s^fY]-p^*\delta u(1+\tau r+\tau S).$$

Using the definition of the covariance function, $\text{Cov}(X,Y)=\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]$, we obtain

$$G=\delta(\mathbb{E}[\mathbb{I}_s^f]\mathbb{E}[Y]+\text{Cov}(\mathbb{I}_s^f,Y))-p^*\delta u(1+\tau r+\tau S).$$

Then, it follows that

$$G=\delta p^*\mathbb{E}[Y]+\delta\text{Cov}(\mathbb{I}_s^f,Y)-p^*\delta u(1+\tau r+\tau S).$$

This simplifies to

$$G=\delta p^*\mathbb{E}[Y]-\delta\text{Cov}(\mathbb{I}_D^f,Y)-p^*\delta u(1+\tau r+\tau S).$$

Since the strategy operates in the interest rate market, Eq. (2) applies. This allows for the simplification of $\mathbb{E}[Y] = u(1/\delta + \psi\tau b + (1-\psi)\tau b\hat{p})$, where $\hat{p}$ is the survival probability of the counterparty. Then,

$$G = p^* u\big[1 + \delta\tau b\psi + \delta\tau b(1-\psi)\hat{p}\big] - \delta\mathrm{Cov}(\mathbb{I}_D^f, Y) - p^*\delta u(1 + \tau r + \tau S).$$

Note that $1 + \tau r + \tau S = \delta^{-1} + \tau S$. It then follows that

$$G = p^* u\delta\big[\tau b\psi + \tau b(1-\psi)\hat{p}\big] - \delta\mathrm{Cov}(\mathbb{I}_D^f, Y) - p^*\delta u\tau S. \tag{4}$$

So,

$$G = p^* u\big[1 + \delta\tau b\psi + \delta\tau b(1-\psi)\hat{p} - 1\big] - \delta\mathrm{Cov}(\mathbb{I}_D^f, Y) - p^*\delta u\tau S,$$

and we obtain

$$G = p^*\big[u(1 + \delta\tau b\psi + \delta\tau b(1-\psi)\hat{p}) - u\big] - \delta\mathrm{Cov}(\mathbb{I}_D^f, Y) - p^*\delta u\tau S.$$

From the simplification of $\mathbb{E}[Y]$, we get $\delta\mathbb{E}[Y] = u(1 + \delta\psi\tau b + \delta(1-\psi)\tau b\hat{p})$. So,

$$G = p^*\big[\delta\mathbb{E}[Y] - u\big] - \delta\mathrm{Cov}(\mathbb{I}_D^f, Y) - p^*\delta u\tau S.$$

Define $\pi = \delta\mathbb{E}^*[Y] - u$ as the value of the trade for a hypothetical risk-free agent. Then,

$$G = p^*\pi - \delta\mathrm{Cov}(\mathbb{I}_D^f, Y) - p^*\delta u\tau S. \tag{5}$$

The term π can be thought of as the objective value of the strategy, without accounting for how the strategy is funded. The term $p^*\delta u\tau S$ is the FVA, which adjusts the objective value given how the strategy is funded. It should be noted that the risk of the counterparty defaulting has no impact on the FVA term. This follows naturally as G is the marginal value from the perspective of the shareholders, not the firm. Thus, from the perspective of the shareholders, the firm surviving is all that is relevant within the FVA.

Furthermore, the counterparty defaulting, and thus the CVA, is not of relevance within this strategy. This is not because the counterparty does not have an impact on the valuation, but rather because it will not change the cost involved in implementing the strategy. This becomes clear when looking at the term $p^*\delta u\tau S$, which is the valuation adjustment made when implementing the strategy. This derivation does not consider what happens in the case that the firm defaults, as it is assumed that in the event of default, shareholders give up all assets to creditors. This leads to no DVA term. Given that this valuation adjustment term, $p^*\delta u\tau S$, is neither the DVA nor the CVA, it is plausible to conclude it that represents the FVA. The FVA is the cost to the shareholders of setting up the strategy.

The firm will need to borrow in order to implement this strategy and since the firm can default, there is a possibility that the firm will be unable to repay the lender. This represents a cost to the firm because, if $p^* \neq 1$ then $S \neq 0$. This shows that if borrowing is required in order to implement the strategy, the cost to the firm will increase as the firm's default risk increases. This has nothing to do with the firm or the counterparty defaulting, thus cementing the fact that the CVA and the DVA are not relevant here. Thus, the cost arises due to the firm needing funding to implement the strategy, so it is plausible to deduce that this represents the FVA, to the strategy valuation.

This leads to the derivation of $G = p^*\pi - \delta \mathrm{Cov}(\mathbb{I}_D^f, Y) - p^*\delta u\tau S$, which is exactly the equation derived by Andersen et al. [2019] for a general strategy payoff and assuming $\tau = 1$. This is not a coincidence, the hypothesis is that this is due to the simple payoff that is created by the Strategy 2.1, which allows us to use the Eq. (2). The Strategy 2.1 involves a foreign exchange contract and involves the interest rate market. This creates a linear cashflow even after considering the default scenarios of both the parties involved, which allows for the simplification of the expected value of the cashflow with relative ease. Eq. (2) is vital for the derivation, given that $Y = u\psi(1 + \tau r + \tau S + \tau b) + u(1 - \psi)(1 + \tau r + \tau S + \tau b)\mathbb{I}_s^{cp}$ is the payoff, and since Eq. (2) is a typical relationship encountered in the pricing of interest rate instruments. If, however, a product other than a swap is being considered, it would be unlikely that the derivation of $G = p^*\pi - \delta \mathrm{Cov}(\mathbb{I}_D^f, Y) - p^*\delta u\tau S$ could be relied upon. Andersen et al. [2019] make no assumption about what the payoff function Y is, which is why they rely on a much more involved proof. However, we deal with a simpler setup, since the question of Covered Interest Parity violation is typical in other interest rate markets. A similar payoff structure would be expected when considering any interest rate swap contract.

Example 2.2. *Next we consider a numerical example which can be found in Andersen et al. [2019]*

It is assumed that the two parties are uncorrelated and have the same credit quality. Assume that $p^* = 0.993$, $s = 0.0035$, $b = 0.0025$, $r = 0$, $u = 1$, $\tau = 1$ *and* $\delta = R = 1$. *R100 is lent synthetically in the US market and R100 is borrowed in the South African market to fund this loan. At maturity R100.60 is received and R100.35 is payed which creates an apparent profit of R0.25. But,* $G = p^*\pi - \delta\, Cov(\mathbb{I}_D^{us}, Y) - FVA$ *where* $\delta\, Cov(\mathbb{I}_D^{us}, Y) = 0$ *(since there is no correlation) and* $FVA = p^*\delta uS = 1 \times 0.993 \times 0.35 \times 100 = 0.35$. *Therefore*

$G = 0.25 - 0.35 = -0.10$. *Even though there is a positive CIP basis, once FVA is accounted for there is no profit to be made.*

3 Methodology and data

3.1 Data

Data is obtained from Bloomberg for 3 January 2012 to 31 December 2021, for the US dollar (USD) and rand (ZAR). In order to apply the strategy 2.1, quotes pertaining to the following instruments are gathered:

- USDZAR spot exchange rate (e.g., 15 ZAR buys 1 USD)
- USDZAR forward exchange contracts (FECs) for maturities of 1, 3, 6, 12, 24, 36, 48 and 60 months
- JIBAR for 1, 3, 6 and 12 months
  - The bid/ask spread for 1-year South African negotiable certificates of deposit (NCDs)
- US LIBOR for 1, 3, 6 and 12 months
- SABOR (overnight rates in the South African market)
- Overnight LIBOR
- JIBAR-linked forward rate agreements (FRAs) for 3x6, 6x9, 9x12, 12x15, 15x18 and 18x21 maturity structures
- LIBOR-linked FRAs for 3x6, 6x9, 9x12, 12x15 and 15x18 maturity structures
- JIBAR-linked interest rate swaps (IRSs) for maturities of 2, 3, 4 and 5 years
- LIBOR-linked IRSs for maturities of 2, 3, 4 and 5 years
  - Basis swaps that swap 3-month LIBOR for 6-month LIBOR for maturities of 2, 3, 4 and 5 years

The JIBAR and LIBOR are the standard interbank interest rate benchmarks, giving lending and borrowing rates for typical major banks. Day-count conventions are ignored. A fifteen basis point bid/ask spread was applied to interest

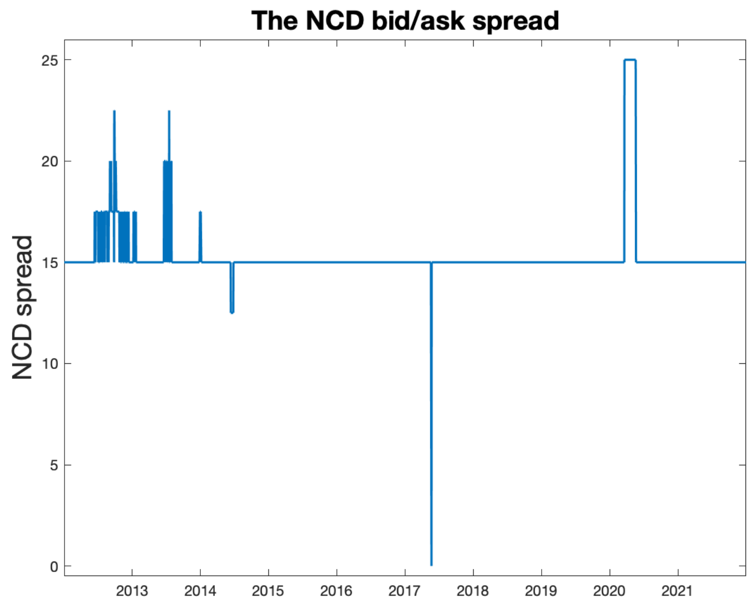


Figure 2: The NCD bid/ask spread.

rates across all tenors. This is inferred from the spread on the South African 1-year NCD which can be seen in Fig. 2 (JIBAR is not quoted with bid/ask spreads but JIBAR is compiled from prevailing NCDs; this therefore gives us a reasonable approximation of bid/ask spreads in interbank markets. Spreads in the US interbank market may be smaller so this seems to be a conservative assumption.)

3.1.1 Plotting the data

Fig. 3 shows the JIBAR and LIBOR for 1-, 3-, 6-, 9- and 12-months tenors. Between 2016 and 2017, the JIBAR reached a peak. They were also at their lowest between 2020 and 2021, with a sharp decline in the first half of the year. The LIBOR are at their highest between 2018 and the end of 2019. However, in the early half of the 2020 year they rapidly decline to the lowest rate for the period under consideration. The rapid decline in 2020, in both sets of rates, could be explained by the COVID pandemic. During this time reserve banks injected large amounts of cash, and borrowing became less important, which resulted in a decrease in interest rates. Ultimately, at shorter tenors both sets of rates are lower than at longer tenors. For example, the overnight rate, over the period under consideration, is lower than the 12-month rate. Both sets of

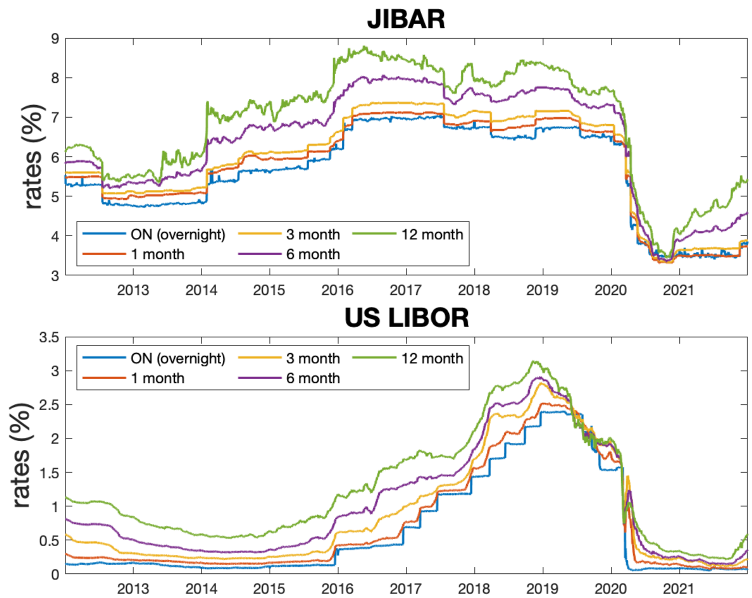


Figure 3: JIBAR and LIBOR from 3 January 2012 to 31 December 2021.

rates look plausible.

Fig.4 shows the FEC rates and the spot exchange rates, in both the US and South African market. Bloomberg quotes the forward points ,X^{pt}, rather than directly quoting the rates. The forward points are converted to rates using the formula $X^f = X^{pt}/10000 + X$.

From Fig.4 one sees that the FEC rates are similar to the USDZAR spot rates, with the 1-month FEC rate being closest to the spot rate. Between 2020 and 2021, there is a large depreciation in the ZAR against the USD. The USDZAR spot rate generally increases over time. This can be interpreted as a depreciation of ZAR against USD.

3.2 Methodology

3.2.1 Bootstrapping

The SA nominal swap curve is bootstrapped out to five years using South African interbank rates, FRAs and IRSs. Similarly, the US nominal swap curve is bootstrapped out to five years using US interbank rates, FRAs and IRSs. The US IRSs are linked to 6-month LIBOR whereas the South African IRSs are linked to 3-month JIBAR. Therefore, the US IRS rates are adjusted using

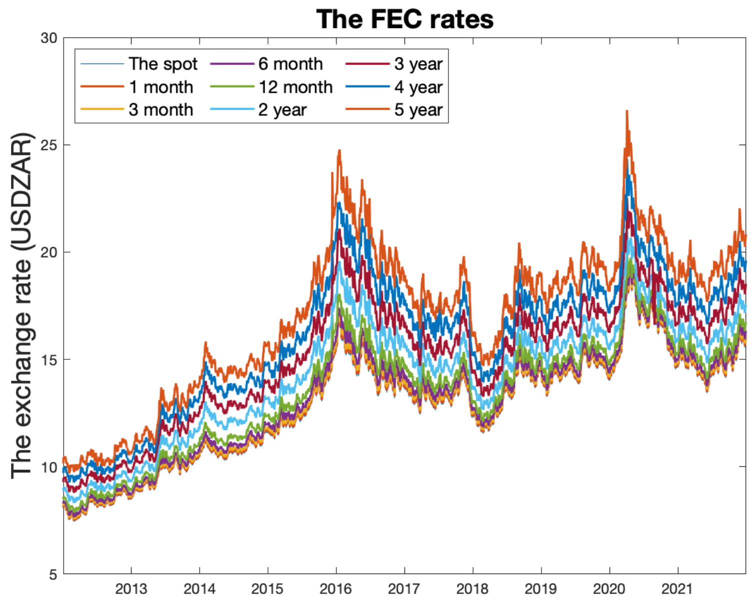


Figure 4: The spot rate and FEC rates over different tenors from 3 January 2012 to 31 December 2021.

market basis-swap spreads. Here the relevant basis swap swaps 3-month LIBOR payments for 6-month LIBOR payments and this therefore converts the US IRSs to 3-month swaps which allows comparison to South African IRSs. Thus, a US and SA bootstrapped curve is obtained for every cross-section of the data, i.e., a curve for each day that there is relevant data for. The discount factors corresponding to the overnight, 1-month and 3-month market rates are derived using

$$Z(t_0, t_i) = \frac{1}{1 + R(t_i - t_0)}.$$

Then, the next set of discount factors are determined with the standard FRA equation:

$$Z(t_0, t_m) = \frac{Z(t_0, t_n)}{1 + f(t_m - t_n)},$$

where f is the t_n-by-t_m FRA rate. The discount factors that correspond to the maturities of the IRSs are obtained from

$$Z(t_0, t_n) = \frac{1 - R\sum_{i=1}^{n-1} Z(t_0, t_i)}{1 + R(t_i - t_{i-1})}.$$

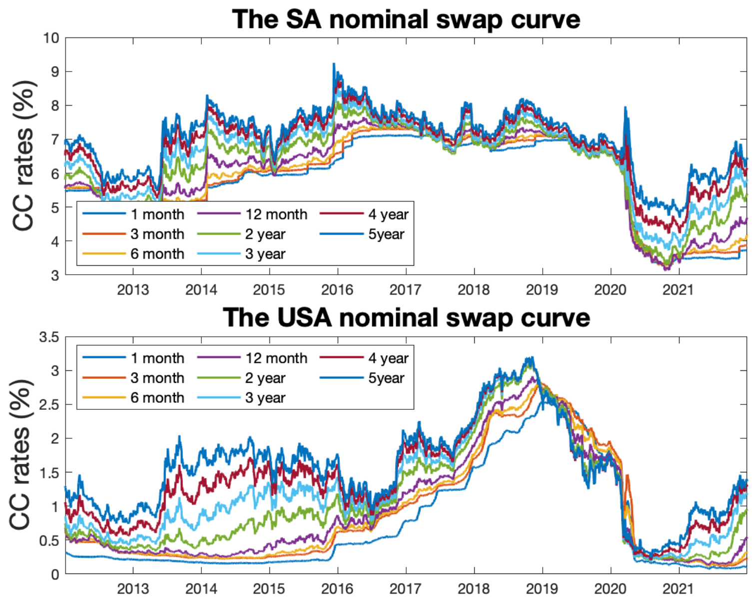


Figure 5: The bootstrapped SA and US nominal swap curves for different tenors from 3 January 2012 to 31 December 2021.

We define x as the quarterly forward discount factor within a particular year, i.e.,

$$\begin{aligned} Z(t_0, t_{n-3}) &= Z(t_0, t_{n-4})x, \\ Z(t_0, t_{n-2}) &= Z(t_0, t_{n-4})x^2, \\ Z(t_0, t_{n-1}) &= Z(t_0, t_{n-4})x^3, \\ Z(t_0, t_n) &= Z(t_0, t_{n-4})x^4. \end{aligned}$$

This amounts to assuming that the quarterly forward discount factors, within a particular year, are equal. This allows us to solve for four unknown discount factors given one swap rate, and this ensures a smooth bootstrapped curve.

3.3 Historical implementation of the Strategy 2.1

The CIP arbitrage strategy is tested using two different approaches. In the first approach, JIBAR and LIBOR are used. These have a maximum maturity of one year. In the second approach, bootstrapped rates are used. These extend up to five years (the longest FEC available). While JIBAR and LIBOR are direct measures of interbank lending and borrowing rates, the bootstrapped rates are *estimates* of longer term lending and borrowing rates available to major banks.

To implement Strategy 2.1, the interest rates need to be adjusted to account for the two-day settlement delay of the foreign-exchange instruments (both spot and FEC). The forward rates for the relevant period of Strategy 2.1 are obtained using,

$$(1 + f(t_0; t_0 + 2\beta, t_0 + t\beta + \tau)\tau) = \frac{Z(t_0, t_0 + 2\beta)}{Z(t_0, t_0 + 2\beta + \tau)}, \tag{6}$$

where the discount factors on the right-hand side are obtained using a spline interpolation scheme on the bootstrapped discount factors. Then, the rate $f(t_0; t_0 + 2\beta, t_0 + t\beta + \tau)$ is a two-day forward JIBAR or LIBOR applicable to the two-day-delayed foreign-exchange transactions.

JIBAR and LIBOR are used to implement the strategy for a tenor of 1, 3, 6, and 12 months (our first approach). Similarly, the bootstrapped rates were used to implement the strategy for a tenor of 1, 3, 6 and 12 months and 2, 3, 4 and 5 years (our second approach). The USDZAR spot rate and a corresponding FEC rate is used to make a synthetic loan in USD which is funded by borrowing in ZAR. This will be referred to as *going the strategy long.* Alternatively, borrowing in USD and lending in ZAR will *short the strategy.* The strategy was implemented daily, for each tenor across the period over which the data extends. Market frictions are accounted for by considering bid and ask rates on the spot and forward exchange rates. The profit from going long the strategy is

$$u\left(\frac{1}{X_a}\underbrace{(1 + (\text{LIBOR} - 0.00075)\tau)X_b^f}_{(1+\tau r+\tau S+\tau b)} - \underbrace{(1 + (\text{JIBAR} + 0.00075)\tau)}_{(1+\tau r+\tau S)}\right), \tag{7}$$

where u is the nominal, which scales the strategy. The first part of the strategy (convert to USD, lend USD at LIBOR and convert back to ZAR with an FEC) can be thought of as a synthetic ZAR-denominated loan that may offer a CIP basis relative to the JIBAR borrowing (which is the second part of the strategy and includes a credit spread). The profit from shorting the strategy, where the bid/ask spread applies in the opposite direction, is

$$u\left(\underbrace{(1 + (\text{JIBAR} - 0.00075)\tau)}_{(1+\tau r+\tau S+\tau b)} - \frac{1}{X_b}\underbrace{(1 + (\text{LIBOR} + 0.00075)\tau)}_{(1+\tau r+\tau S)}X_a^f\right). \tag{8}$$

The accumulation factor of the borrowing leg in either Eq. (7) or Eq. (8) is used to calculate r and S. The profit from the strategy can therefore be used to derive the basis b.

3.4 Calibrating the survival probability

Once the profit for the strategy has been determined using Eq. (7) or Eq. (8), depending on which direction of the strategy is profitable, it is then possible to calibrate the survival probability p^* in Eq. (5) such that the profit is diminished to zero. By assuming that the default risk of the firm and the default risk of the counterparty are correlated, an analytical solution for p^* can be derived using Eq. (12). This represents the firm's minimum survival probability necessary to make a positive net profit from Strategy 2.1.

We now explain how we specifically solve for the value of p^*. Take Eq. (4), and assume that $p = \hat{p} = p^*$. Then,

$$G = \delta pu(\tau b\psi + \tau b(1-\psi)p) - \delta u\tau Sp - \delta \mathrm{Cov}(\mathbb{I}_D^f, Y).$$

Using Eqs (2) and (11) we have

$$G = \delta pu(\tau b\psi + \tau b(1-\psi)p) - \delta u\tau Sp - \delta(-u(1-\psi)(1+\tau r+\tau S+\tau b)\mathrm{Cov}(\mathbb{I}_D^f, \mathbb{I}_D^{cp})). \tag{9}$$

Simplifying the $\mathrm{Cov}(\mathbb{I}_D^f, Y)$ term and noting that $\mathbb{E}[\mathbb{I}_D^f] = 1-p$, leads to

$$\begin{aligned}\sigma_f^2 &= V[\mathbb{I}_D^f] = \mathbb{E}[\mathbb{I}_D^{f\,2}] - \mathbb{E}[\mathbb{I}_D^f]^2 \\ &= (1-p) - (1-p)^2 \\ &= p - p^2.\end{aligned}$$

However, it is assumed that $p = \hat{p} = p^*$ which sets the standard deviations equal. The covariance can be expressed in terms of a correlation ρ as follows:

$$\rho = \frac{\mathrm{Cov}(\mathbb{I}_D^f, \mathbb{I}_D^{cp})}{\sigma_f \sigma_{cp}}. \tag{10}$$

So, we have

$$\mathrm{Cov}(\mathbb{I}_D^f, \mathbb{I}_D^{cp}) = \rho(p - p^2). \tag{11}$$

Using Eqs (2) and (11) we get

$$G = \delta pu(\tau b\psi + \tau b(1-\psi)p) - \tau\delta uSp - \delta(-u(1-\psi)(1+\tau r+\tau S+\tau b)\mathrm{Cov}(\mathbb{I}_D^f, \mathbb{I}_D^{cp})).$$

To find the break-even survival probability, we set $G = 0$. Then, the above equation simplifies to

$$\begin{aligned}0 &= p(\tau b\psi + \tau b(1-\psi)p) - \tau Sp + (1-\psi)(1+\tau r+\tau S+\tau b)\mathrm{Cov}(\mathbb{I}_D^f, \mathbb{I}_D^{cp}) \\ &= p(\tau b\psi + \tau b(1-\psi)p) - \tau Sp + (1-\psi)(1+\tau r+\tau S+\tau b)\rho(p-p^2) \\ &= p(\tau b\psi + \tau b(1-\psi)p) - \tau Sp + p(1-\psi)(1+\tau r+\tau S+\tau b)\rho \\ &\quad - p^2(1-\psi)(1+\tau r+\tau S+\tau b)\rho,\end{aligned}$$

where Eq. (11) is used. After factorising, it follows that,

$$0 = p^2\big[\tau b(1-\psi)-(1-\psi)(1+\tau r+\tau s+\tau b)\rho\big]+p\big[\tau b\psi-\tau S+(1-\psi)(1+\tau r+\tau S+\tau b)\rho\big].$$

Therefore,

$$p = \frac{\tau S - \tau b\psi - (1-\psi)(1+\tau r+\tau S+\tau b)\rho}{\tau b(1-\psi) - (1-\psi)(1+\tau r+\tau S+\tau b)\rho}.$$

We have that $S = (1-p)(1-\psi)/\tau$ and, for simplicity, we set

$$\begin{aligned} y &= -\tau b\psi - (1-\psi)(1+\tau r+\tau S+\tau b)\rho, \\ x &= \tau b(1-\psi) - (1-\psi)(1+\tau r+\tau S+\tau b)\rho \end{aligned}$$

such that

$$p = \frac{\tau S + y}{x}.$$

It then follows that

$$p = \frac{(1-\psi) - p(1-\psi)}{x} + \frac{y}{x}$$

and

$$p + \frac{p(1-\psi)}{x} = \frac{1-\psi}{x} + \frac{y}{x}.$$

So,

$$p\Big[\frac{x+(1-\psi)}{x}\Big] = \frac{(1-\psi)+y}{x}.$$

Therefore, the minimum survival probability p can be expressed by

$$p = \frac{(1-\psi)+y}{x+(1-\psi)}, \tag{12}$$

where the recovery rate ψ is set to 40%.

4 Results

In Section 4.1, we give our first approach based on interbank rates. Then, in Section 4.2, we give our second approach based on the bootstrapped rates which extend to longer maturities.

4.1 JIBAR and LIBOR

4.1.1 The CIP basis

Fig. (6) shows the size of the CIP basis at maturity as calculated on the initiation date of the strategy for the period 03 January 2012 to 31 December 2021 and assuming a nominal of R1, i.e., this figure plots τb in Eq. (9). The

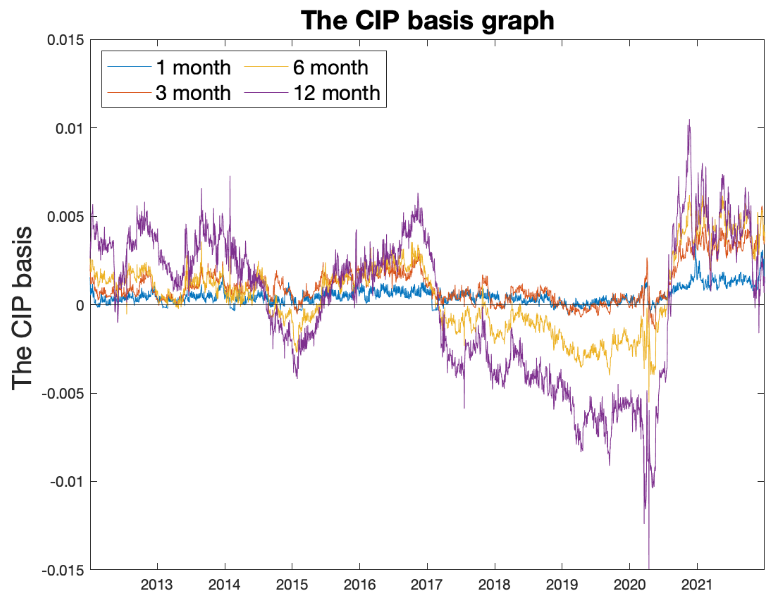


Figure 6: The CIP basis based on JIBAR and LIBOR from 3 January 2012 to 31 December 2021.

calculation of the CIP basis here does not account for the firm's default risk, nor does it incorporate any bid/ask spreads. Only the long position in Strategy 2.1 has been considered, which therefore leads to the CIP basis falling below zero in some cases. Shorting the strategy in these cases would generate a profit. Shorter tenors tend to generate a smaller CIP basis compared to longer tenors. The CIP basis in Fig. 6 has not been annualised which could explain this observation. The strategy for the longer tenors imply a greater period over which the synthetic loan can accumulate and generate a profit at maturity. In contrast, a 1-month tenor, for example, would generally not be long enough to accumulate as large a profit. The CIP basis over the 1-month tenor remains positive over almost the entire period under consideration. This may be due to more stable interest rates and USDZAR FEC rates over a 1-month period. At this tenor, an arbitrageur will almost always go long Strategy 2.1. A similar result is seen for the 3-month tenor, however the CIP basis at this tenor is slightly more volatile. In contrast to the 1- and 3-month tenor, the CIP basis for the 6- and 12-month tenor fluctuates significantly over the period and is negative for a large period extending from February 2017 until July 2020. This may be due to greater discrepancies between the two economies' interest rates offered over these tenors. Less stable FEC rates for these tenors could also explain the volatility in the

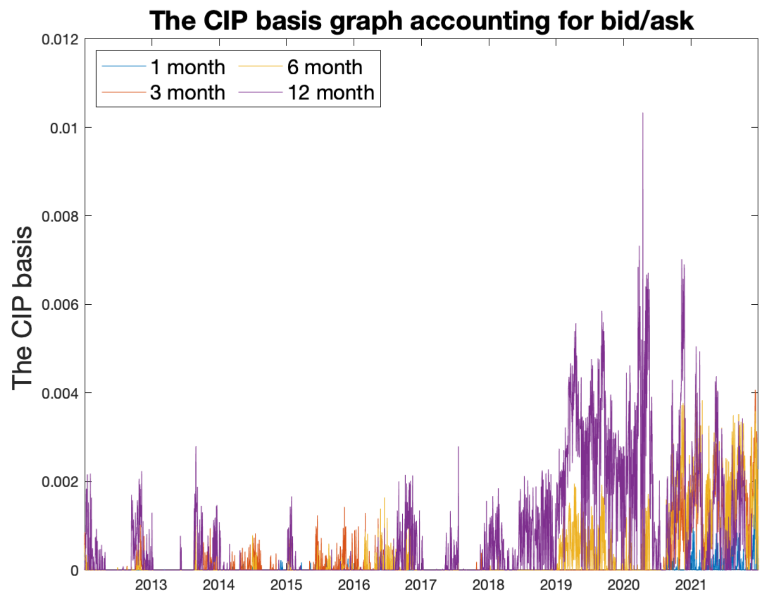


Figure 7: The CIP basis based on JIBAR and LIBOR from 3 January 2012 to 31 December 2021 using bid/ask spread.

CIP basis. Therefore, at longer tenors, it would be possible to long or short Strategy 2.1, depending on whether a positive or negative basis was achieved.

The CIP bases become positive across all tenors for the period extending from August 2020 until December 2021. During this period, the 1-, 3- and 6-month CIP bases are relatively larger than before, while the 12-month CIP basis is relatively smaller. The impact of the COVID-19 pandemic on financial markets is likely to have caused this pronounced change in the CIP basis across these tenors over this period.

We now account for bid/ask spreads in the calculation of the CIP basis. This results in a reduction in the absolute value of the CIP basis as shown in Fig. 6. The bid/ask spread reduces the proceeds from the loan (a lower interest rate is earned and conversion between ZAR and USD is done at less attractive terms) and increases the liability due in the form of a higher interest rate in the borrowing arm. The bid/ask spread has this effect for both directions of the strategy. Fig. 7 shows the CIP basis when accounting for these spreads and is expressed under the profitable direction of the strategy. The CIP basis appears to increase on average over the period under consideration and the apparent arbitrage persists. Therefore, the bid/ask spread is not sufficient to rule out the possibility to profit under Strategy 2.1.

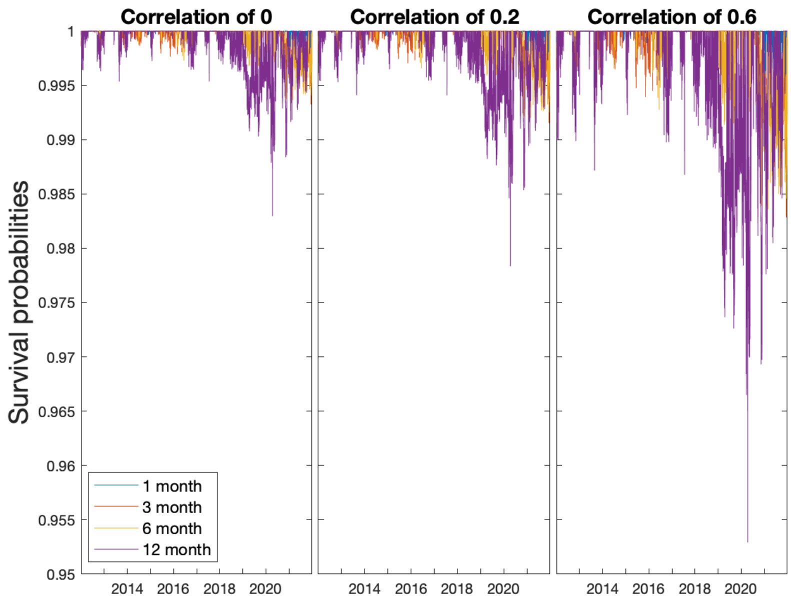


Figure 8: The break-even survival probabilities based on JIBAR and LIBOR from 3 January 2012 to 31 December 2021.

4.1.2 The survival probability

By considering Eq. (5) as derived under Strategy 2.1, it is possible to find a survival probability for the firm that removes the positive CIP basis that remains after accounting for bid/ask spreads. Recall that the derivation for this survival probability is given in Section 3.4. These are the minimum survival probabilities that a firm must have at each date in order for the apparent arbitrage in the Strategy 2.1 to remain after the funding cost of the strategy is accounted for. See Fig. 8 for survival probabilities we obtain over the sample for each maturity. The left-hand panel assumes zero correlation between the arbitrageur's and the counterparty's default indicators, which makes the covariance in Eq. (5) vanish. The middle- and right-hand panel assume correlation values of 0.2 and 0.6, respectively—recall Eq. (10). The left-hand panel of Fig. 8 shows reasonable survival probabilities, in other words, non-excessive default probabilities. For example, with the exception of 2020 and 2021, 1-year survival probabilities are almost always above 99.5%. While there are many possible explanations for a non-zero CIP basis (as discussed in Section 1), we tentatively conclude that the FVA-based explanation of Andersen et al. [2019] is sufficient. In other words, realistic levels of bank default do rule out profiting from a CIP basis. During

2020 and 2021 a larger CIP basis was available. Fig. 8 shows that a lower survival probability would still allow the firm to exploit the potential arbitrage when implementing the Strategy 2.1. The survival probabilities are lower for the longer maturities as the probabilities have not been annualised.

As the correlation between the default indicators of the firm and the counterparty increases, Fig. 8 shows that the minimum survival probability that renders Strategy 2.1 profitable decreases. This is the case since a higher correlation is a benefit to the arbitrageur (in particular to the shareholders, the covariance term in Eq. (5) makes a positive contribution to shareholder value G, as detailed in Section 3.4). When counterparty default tends to coincide with the firm's default, the harm caused by counterparty default becomes less relevant (as the shareholders receive nothing in the case of the firm's own default).

4.2 Bootstrapped rates

4.2.1 The CIP basis

We consider longer tenors for Strategy 2.1 by using bootstrapped rates. Similarly to Fig. 6, we can see in Fig. 9 that these longer tenors tend to generate a larger CIP basis. This is consistent with the view that there is a longer period over which the profit of the strategy can accumulate and is therefore to be expected. The CIP bases for tenors up to 12 months, calculated here under approach 2, are similar in magnitude to those derived using the first approach in Section 4.1. However, these CIP bases are almost always positive over the sample whereas it is negative for a large part of the sample under approach 1. Fig. 9 indicates that the volatility in the CIP basis increases as the tenor increases. This is consistent with the relationship between the volatility and the tenor under approach 1.

The CIP bases for tenors extending beyond 12 months become negative over various periods in the sample, but is positive for the majority of the sample period. There is a pronounced, negative spike in the 4- and 5-year tenor CIP bases on 10 and 11 December 2015. During the period 9 to 11 December 2015, the South African finance minister was replaced twice. This may explain the observed spike. The CIP basis is more volatile for longer tenors which is similar to what is observed in Fig. 6. The CIP bases across all tenors over the period from August 2020 to December 2021 display a similar effect to that observed in Fig. 6 for this period. In contrast to the shorter tenors, Fig. 9 indicates that the magnitude of the CIP bases at longer tenors do not seem to increase on average over the sample period. Instead, Fig. 10 suggests that, on average, CIP bases for longer tenors decrease slightly over the period of investigation.

As in Fig. 7, the CIP basis which accounts for bid/ask spreads is plotted in Fig. 10. A reduction is observed in terms of the absolute value of the CIP bases. This reduction is attributed to bid/ask spreads. A brief period extending from October 2019 to February 2020 seems to achieve CIP. The parity over this period is not observed under approach 1. This suggests that interbank rates prevailing in the markets at this time differ from the rates obtained from bootstrapping the nominal swap curve. However, the CIP bases outside of this period remain positive and this suggests that bid/ask spreads are not sufficient to rule out the profitability of Strategy 2.1. This is true for the tenors extending beyond 12 months as well.

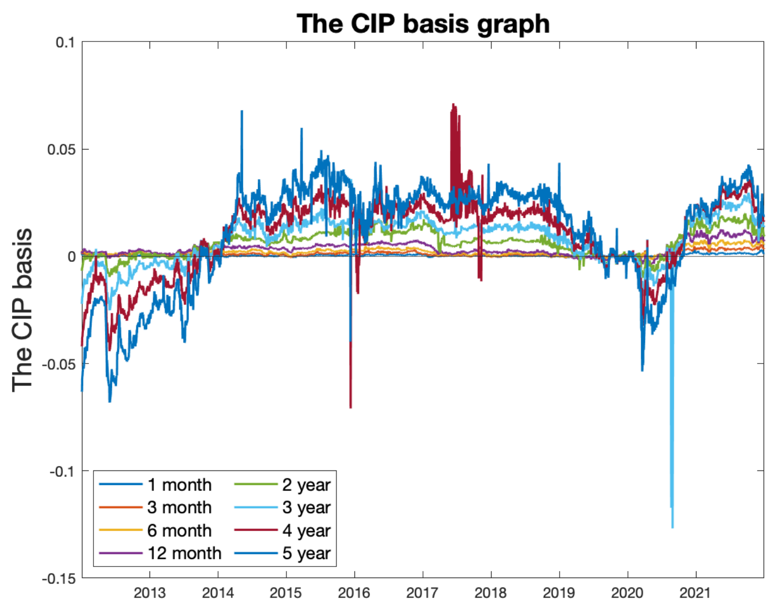


Figure 9: The CIP basis based on bootstrapped rates from 3 January 2012 to 31 December 2021.

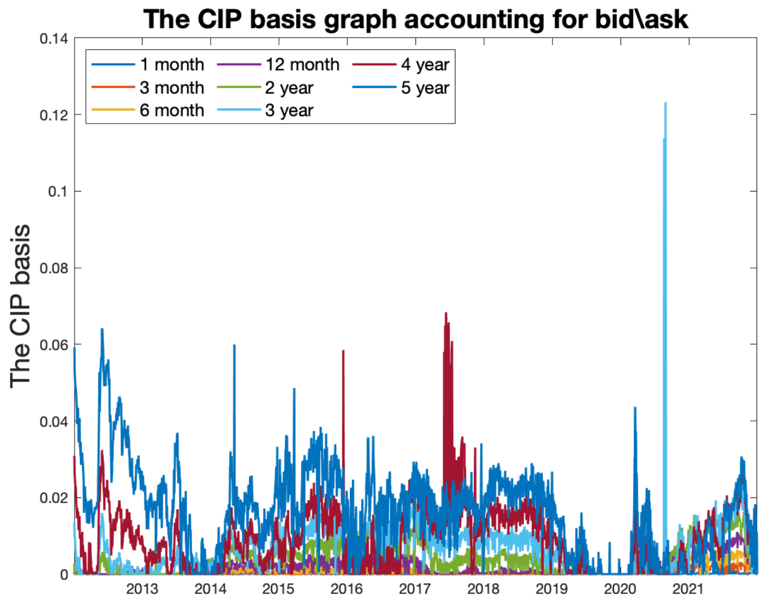


Figure 10: The CIP basis based on bootstrapped rates using bid/ask spread from 3 January 2012 to 31 December 2021.

4.2.2 The survival probability

Fig. 11 plots the minimum survival probability required for Strategy 2.1 to break-even. The probabilities show a similar tendency to decrease as the correlation increases, which is consistent with the view in Section 4.1.2. This holds for all the tenors under consideration in approach 2. The survival probabilities corresponding to the longer tenors tend to increase slightly over the period under consideration. This agrees with the view that CIP bases over these tenors decrease on average over the period. In other words, over time, a greater survival probability would be required to generate the available CIP basis which is declining.

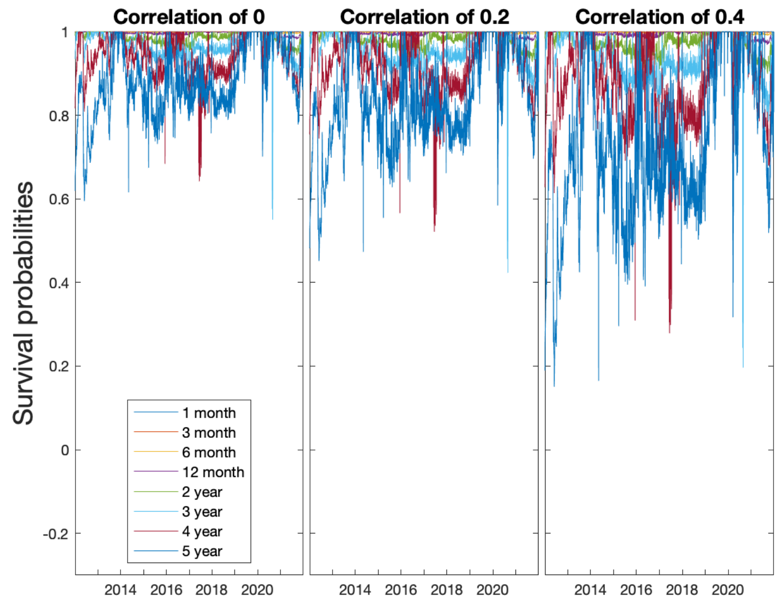


Figure 11: The break-even survival probabilities based on bootstrapped rates from 3 January 2012 to 31 December 2021.

Similarly to what is observed in Fig. 8, the survival probabilities decrease as correlation increases. This can be seen in Fig. 11 for each tenor under consideration. As in Fig. 8, the survival probabilities are generally lower for longer tenors since they have not been annualised. Finally, the observed survival probabilities in the left-hand panel of Fig. 11 are reasonable, as in Fig. 8. This further suggests that realistic levels of bank default erode any potential profit arising from the CIP basis.

5 Conclusions

The CIP basis is calculated for tenors up to 5 years using market data for the 10-year sample period. Adjusting the calculation to account for bid/ask spreads does not reduce the CIP basis to zero. We present a simplified derivation for the main proposition presented in Andersen et al. [2019] for the particular payoff structure created by Strategy 2.1. This allows us to derive the equation needed to calculate the gain achieved by the firm's shareholders for the Strategy 2.1. Thus, we derive the funding valuation adjustment, which adjusts the gain in order to account for the cost of implementing Strategy 2.1. The FVA is seen to arise solely due to the risk of the arbitrageur defaulting. The FVA cost resulting in a break-even strategy suggests survival probabilities that are consistent with default levels among major banks. This is a sufficient reason for non-zero CIP basis to persist.

Possible further research could consider using institution-specific data as opposed to general interbank rates among major banks. Alternatively, this research can be extended by considering how survival probabilities derived from credit default swaps compare to survival probabilities calculated here. Other currency pairs could also be considered to assess whether this FVA-based explanation for non-zero CIP basis holds.

References

L. Andersen, D. Duffie, and Y. Song. Funding value adjustments. *The Journal of Finance*, 74(1):145–192, 2019.

C. De J. Correia and R. Knight. Covered interest arbitrage opportunities in the south african foreign exchange market. *South African Journal of Business Management*, 18(4):209–214, 1987.

W. Du, A. Tepper, and A. Verdelhan. Deviations from covered interest rate parity. *The Journal of Finance*, 73(3):915–957, 2018.

J. Hull and A. White. Valuing derivatives: Funding value adjustments and fair value. *Financial Analysts Journal*, 70(3):46–56, 2014.

J. M. Keynes. A reply to sir william beveridge. *The Economic Journal*, 33(132): 476–486, 1923.

Updating Volatility Surfaces

TEAM 2

JUSTIN WAGNER, UCT
NICKY MALULEKE, UCT
STEHANN VAN JAARSVELT, UCT
TIM ALFELD, UCT

Supervisor:
ERIK SCHLOGL, University of Sydney
JACQUES VAN APPEL, University of Johannesburg

African Collaboration for Quantitative Finance and Risk Research

Contents

1 Introduction

The Black Scholes-Merton model, although serving as a key foundation for future research in asset pricing models, was a simplified version of observed market prices. In particular, it is clear that share price volatility requires a term structure rather than being assumed to remain constant. This is shown by the fact that Black-Scholes implied volatility (the volatility calculated by applying the inverse Black-Scholes function under constant volatility to market European option prices) varies with time. Similarly, when considering implied volatilities for options with the same time to maturity, a so-called 'volatility smile' is observed - options that are nearer to the money trade at lower volatilities than those further out of the money.

Various methods to make the model more consistent with the market have been considered - including non-traded stock jump risks, fees and stochastic volatility, however, the local volatility model, where volatility is a function depending on time to maturity and the current stock price, has become widespread. The local volatility model has greater consistency with market data, as it can account for the volatility smile without introducing any additional non-traded market risks. This is key, as it retains the ability to replicate payoffs within the model, and hence the assumption of arbitrage-free risk neutral pricing of options.

Market data for European options is not available for all combinations of strike price and time to maturity. Additionally, there may be low levels of volume (traded rarely or not at all) and hence low liquidity for certain stocks or indices. This results in greater uncertainty (and likely higher bid-offer spreads) about 'true' prices for options based on market expectations of future asset movements [1]. Discrete, and often sparse market option price data, does not fully specify the local volatility function.

The technique used in this paper is to extract the implied volatilities from the market option price data. These volatilities are then used to calibrate a total variance surface using a choice of parameterisation. Choosing to calibrate the volatility surface has two main benefits. Firstly, a good choice of volatility surface parameterisation can prevent static arbitrage with relative ease compared to the parameterisations of other surfaces. Secondly, extracting implied volatilities and fitting the total variance surface in terms of moneyness allows for easier comparability between different securities [1]. This occurs because while a call price surface is also a function of discount factors and stock price, the volatility surface is information purely relating to European option price data.

Once a full, sufficiently smooth surface of European option prices is attained, the stock price diffusion process becomes fully specified. The Breeden and Litzenberger [4] result can be used to obtain the risk neutral distributions of a stock price by the second derivative of the call surface with respect to strike price for any time to maturity, conditional on the current time. Normally, having specified conditional distributions does not fully specify the underlying diffusion process - because more than one dynamic diffusion process might capture the specified conditional distributions. But when we restrict ourselves to the risk neutral assumption, the diffusion process is fully specified, and the Dupire Formula can be used to derive the local volatility function from the smooth call price surface [2].

The local volatility model has a volatility that depends only on time to maturity and the value of the underlying. This means that once the total variance surface has been fitted to data on a given date, the local volatility for the stock is fully specified up to the last maturity date for which sufficient-volume option prices are available (assuming no extrapolation is performed). There will be some error introduced based on interpolation method or in this specific case, an error based on the parametric form chosen for the fitting of the total variance surface.

In practise the total implied variance surface's parameters are typically re-optimised daily (or on an even more frequent basis.) This would be based on the most relevant option price data - meaning that the previous iteration of the model/ surface is discarded. This presents a clear inconsistency with the model, which has an already fully specified volatility function - that should not change daily.

This paper attempts to utilise a modifying algorithm to update the local volatility surface by fitting a new volatility surface over the strike-time space as new data becomes available. In order to make sure that the surface is as unchanged and smooth as possible whilst fitting new market data, we aim to minimise the Wasserstein distance between the implied risk neutral densities of both the old and new parameterisations. The goal of doing this is to reduce the above mentioned inconsistency whilst still having a model calibrated to the newly available information.

Theoretically, by minimising the Wasserstein distance between risk-neutral densities of stock prices (at pre-determined time points) - our model will be optimally calibrated. These densities are implied by the previous-day volatility optimisation and the optimisation that has been adjusted for new option price data.

This paper further aims to find a minimal number of option price data points, which when fed into the modifying algorithm, results in a volatility surface and resulting option prices that are a reasonably close fit to the new option price data for that day that have not yet been added to the surface calibration. Further description of the modification process and how fit is measured will follow.

The benefit of being able to calibrate a new and accurate surface using only a few extra data points would be two-fold. Firstly, a new 'daily' surface would be able to be proposed without knowledge of most of that day's data points. Secondly, if one could reasonably attribute all, or most, of the change in shape of a Local Volatility Surface (and therefore the change in shape of an option price surface) to only a select few data points - then it is possible that to hedge this movement one need only hold some combination of these few points.

As a check, the calibration of the total variance surface is tested by extracting the local volatility using Dupires formula, and then using a Monte-Carlo scheme to price European options. The resulting prices should match the prices used to calibrate the model. If there is sufficient time, the project will then be concluded by performing a Longstaff-Schwartz Monte-Carlo pricing routine to price American options to examine the existence of an early exercise premium in observed market prices.

2 Volatility Surface Construction

2.1 Data

The test data for this model was daily European puts and calls on the S&P500, and daily American puts and Calls on Moderna and Tesla stock, all of which are at a broad range of strike prices and time to maturities. For each strike and time to maturity combination there is a bid and ask price and volume traded for both the puts and calls. The daily data was collected in batches and is all available to the model at once. This is at odds with how real-world data would arrive. In reality new data points would arrive asynchronously as the options are traded.

If an option at some strike and time to maturity combination had no puts and no calls traded, then it was discarded from the data set before the start of the project as their prices cannot be considered market data. All options with maturities in less than a week were also discarded. This was done in order to minimise the errors that would accompany the inclusion of these near-maturity options. The reason these errors arise is that market pricing becomes unreliable at this stage - as market idiosyncrasies and liquidity issues become prevalent. Further, near-maturity options are more affected by intra-day volatility.

2.2 Discount Factors

In order to calculate the implied volatilities required to create the total variance surface we require the term structures of interest and dividend rates for each stock/index observed. It is sufficient to extract risk-free rate discount factors, $B(t,T)$ and dividend discount factors $Q(t,T)$ from the market option prices using the put-call parity, under the assumption of there being no arbitrage in this regard. The put-call parity relation is as follows:

$$C_t - P_t = S_t Q(t,T) - KB(t,T) \tag{1}$$

where S_t is the the share price, C_t and P_t - the prices of the market call and put options. These prices are measured as the mid point between the bid and the offer price. The risk neutral discount factor is given by $B(t,T)$ and $Q(t,T)$ is the dividend discount factor.

These discount factors were calculated for each maturity in the data. This was done using a Python minimisation function to minimise the objective function below for each specific time to maturity and over all price data for that time to maturity (options of varying strikes are included and the best fit is found.)

$$\sum (S_t Q(t,T) - KB(t,T) - C_t + P_t)^2 \tag{2}$$

The output of this minimisation is an implied risk free discount factor (B_t) and dividend discount factor (Q_t) that make the put call parity hold most true for each collection of options inside an individual time to maturity 'bin'. These time-denominated interest and dividend discount factors imply a term structure of rates which is used throughout this paper.

One issue we may run into when computing the discount factors by optimisation is that some discount factors may be greater than one. This is not a severe issue for the risk-free discount factors - as real-world interest rates need not necessarily be positive, but we require the dividend discount factor factors to be less than or equal to one, because negative dividends are not realistic, nor legally allowed. A solution for this is to set all the divided discount factors that are greater than one, to one, and recalculate the corresponding risk-free discount factor. Given that it is easier to extract these discount factors from the S&P500 data, we may store these and reuse them when working with the Moderna and Tesla data. The Moderna and Tesla stocks do not pay any dividends, so we can set all the the dividend discount factors to one for them. The time to maturities may not be the same, in that case we can just interpolate for the corresponding risk-free discount factors using the following equation (log linear interpolation.)

$$B(t,T) = B(t,T_1)^{\frac{T_2 - T}{T_2 - T_1}} B(t,T_2)^{\frac{T - T_1}{T_2 - T_1}} \tag{3}$$

where T is the time to maturity with a missing risk-free discount factor, T_1 and T_2 are time to maturities with corresponding discount factors $B(t,T_1)$ and $B(t,T_2)$ respectively, and $T_1 \leq T \leq T_2$.

2.3 Implied Volatility

Implied volatility is the volatility which, when inputted into the constant-volatility Black-Scholes pricing formula, recovers market European option prices. These volatilities are found numerically rather than analytically, using a Python root-finding function to set the Black-Scholes equation equal to observed market prices by varying the volatility parameter.

The test data contains prices C_t and P_t, strike prices (K) , time to maturities (τ) and the underlying prices (S_t). The risk-free discount factor (B_t) and dividend discount factor (Q_t) were implied from the data. There is therefore sufficient information to calculate implied volatilities.

In the Black Scholes Merton model, the option price is given by

$$\text{Option price } = \alpha(S_t Q_t \Phi(\alpha d_1) - K B_t \Phi(\alpha d_2)), \tag{4}$$

With

$$d_1 = \frac{\ln \frac{S_t Q_t}{B_t K} + \frac{1}{2}\sigma^2 \tau}{\sigma\sqrt{\tau}} \tag{5}$$

and

$$d_2 = d_1 - \sigma\sqrt{\tau} \tag{6}$$

where $\alpha = 1$ for a call option C_t and $\alpha = -1$ for a put option P_t[1]. Here is σ is the implied volatility we a trying to compute.

Performing an optimisation routine on the Black-Scholes equation for the call option, we get a pair of implied volatilities - for the bid and ask prices. The same is true for the put option. To set an interval of acceptance for calibrating the total variance surface we set the lower acceptance bound to be the minimum of the bid implied volatilities and the upper bound to be the maximum of the ask volatilities over both the put and call implied volatilities. This slightly-relaxed acceptance region creates some leeway in the variance surface parametrisation to improve fit and prevent over-fitting in the case where either put or call price data has low volume and accuracy.

We have cases where the optimisation routine fails to find an optimal implied volatility, possibly as a result of bad data. These cases are very few, so we choose to remove those points from the data and continue with rest of the data. To check if these implied volatilities are correct, we can input them back into the Black-Scholes option pricing formula and check the absolute difference between the Black-Scholes price and the market price in the data. The absolute difference of our data was less than 1×10^{-9} for all data points, as expected.

2.4 Surface Parameterisation

Using the implied volatilities calculated from market data, as described above, we now aim to parameterise a total implied volatility surface, as a smooth function over the Strike (K) and Time to Maturity (τ) space. This surface can then be transformed into a call or put price surface over the Strike (K) and Time to Maturity (τ) space, hence it is a way of interpolating between the observed market prices. The resulting call surface function can be used to obtain conditional risk neutral probability densities for the stock price at future times, as described by Bhupinder Bahra [1]. These risk neutral density functions will be discussed in more depth in the following section on updating the Local Volatility Surface.

Practitioners require there to be no arbitrage in prices produced by their model. If the calibrated surface contains points where a risk free expected profit can be made, this could be used against the 'user' of the model.

Calendar Spread arbitrage occurs when traders are able to use a combination of a long and short call option on the same underlying asset, with the same strike but different maturities to make a potential risk-free profit. Butterfly arbitrage is a risk-free profit made by holding long calls with a high and low strike, and selling two call options with strikes in the middle of the high and low strikes - all with the same time to maturity. This strategy has a positive payoff for terminal stock prices between the low and high strikes, and zero payoff otherwise. If the net premium from buying and selling the calls is equal or greater than zero, an arbitrage strategy exists.

The following parametrisation of the total variance surface, suggested by Gatheral and Jacquier [3], fits market volatilities well using a limited number of parameters. This parametrisation has the desired qualities of being a smooth function, and precluding both calendar and butterfly arbitrage. This is given by

$$w(t,k) = \frac{\theta_T}{2}\left(1 + \rho\psi(\theta_T)k + \sqrt{(\psi(\theta_T)k+\rho)^2 + 1 - \rho^2}\right),$$

$$\text{Where } \theta_T, T > 0 \,,$$

$$\psi(\theta) = \frac{\eta}{\theta^{\gamma_1}(1+\beta_1\theta)^{\gamma_2}(1+\beta_2\theta)^{1-\gamma_1\gamma_2}}$$

k is referred to as log-moneyness, and satisfies the following equation in terms of strike price: $E(S_T|\mathcal{F}_0)e^k$ = K ,
where $\gamma_1 = 0.238, \gamma_2 = 0.253, \beta_1 = e^{5.18}, \beta_2, \eta = 2.016048e^{\epsilon}$, and $\epsilon \in (-1,1)$ and ρ are parameters to be fitted [8].

θ_T = w(T,0) is the total implied variance for options with strike equal to the forward price. θ_T is parameterised by finding the total implied variance (average of bid and ask) for the at the money options for every time to maturity in the dataset. Linear interpolation is used to find values in-between data points.
Using Dupire's Formula, this parameterisation can be used to extract a parametric form for the local volatility surface of the stock/index price, which can later be used to price various options using Monte Carlo simulation.

The parameters ρ and ϵ were calibrated by using a Python minimisation algorithm where the objective function was set as the squared relative distance by which the fitted total variance was outside of the bid-offer spread for implied total variance:

$$\sum_{\text{Market Data}} \left(\max(w_{\text{fitted}}(t,k) - \sigma^2_{\text{implied, ask}}t; \sigma^2_{\text{implied, bid}}t - w_{\text{fitted}}(t,k); 0\right)^2 \tag{7}$$

The following shows the fit achieved on the S&P500 European Call and Put option market data on 08 March 2022:

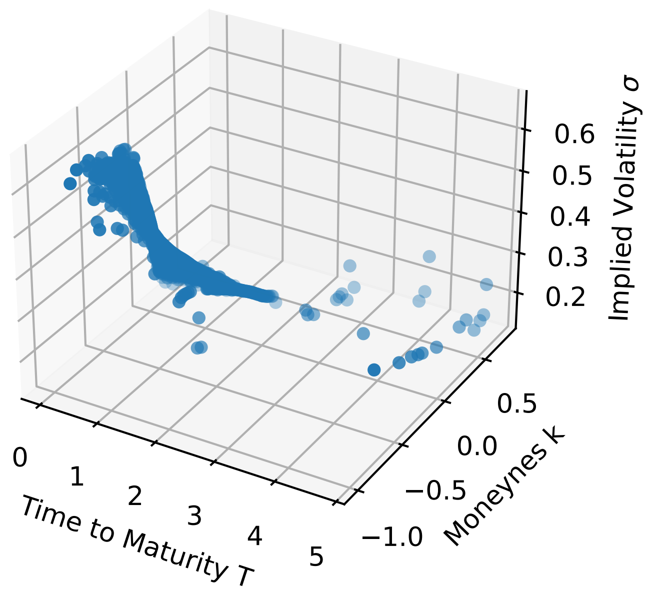


Figure 2.1: Implied Volatilities for S&P500 on the 8th of March 2022

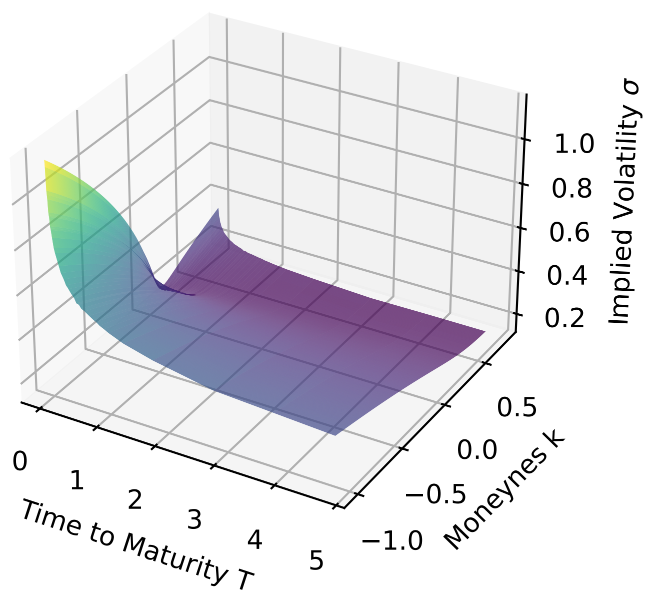
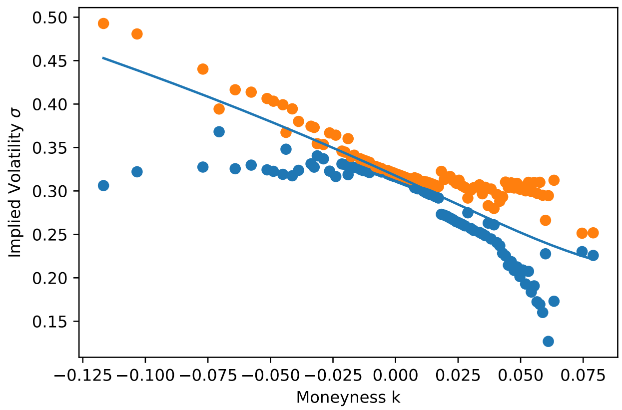
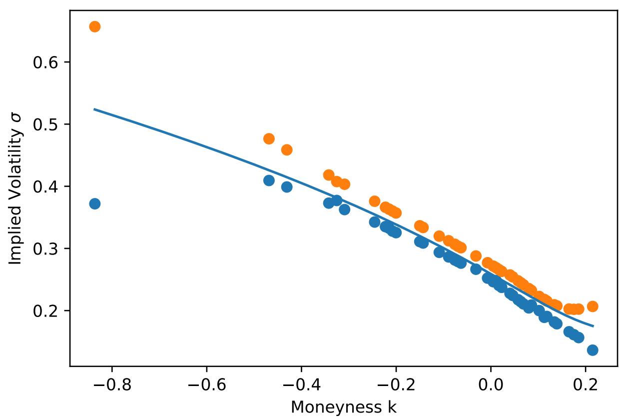


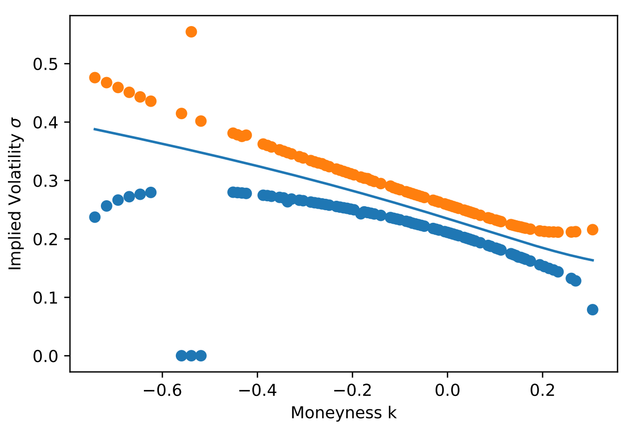
Figure 2.2: Parameterised Total Variance Surface for the S&P500 on the 8th of March 2022



(a) 15 days



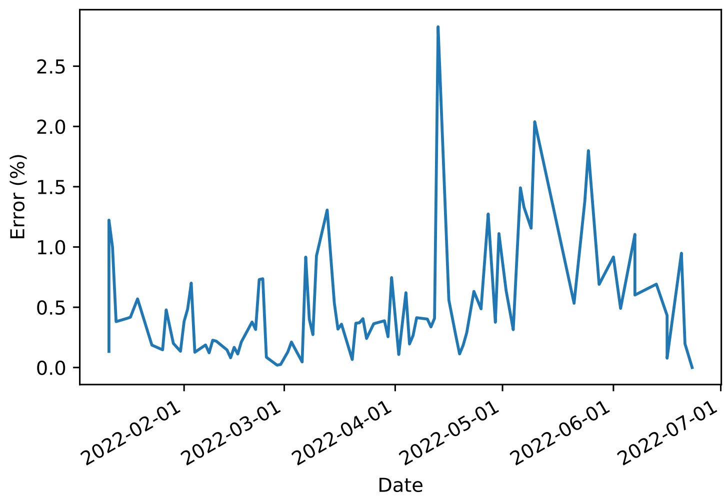
(b) 206 days



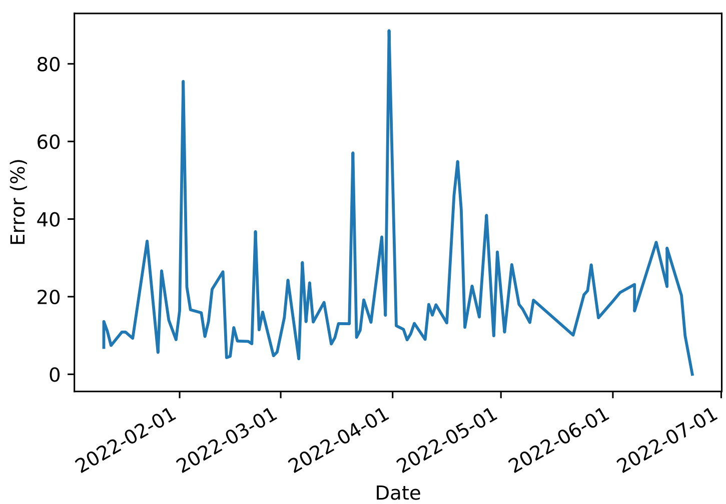
(c) 647 days

Figure 2.3: Time slices showing fit of total variance surface over a range of log-moneyness values

The graphs above show the total variance curve for fixed times and a varying log-moneyness range. The parameterised surface passes in between the bid (blue dots) and ask (orange dots) total implied variances for most data points, indicating an acceptable fit has been achieved. $\epsilon = 0.4844985584878775$ and $\rho = -0.7851212547854564$ were found to be the optimal values for the surface.



(a) Average errors



(b) Maximum errors

Figure 2.4

The above figures show the average and maximum relative error of fitted values for total variance, compared to the market implied values for approximately a six month period. The average error on a particular day is the total error of the calibrated total variance surface divided by the number of data points that day. The average error is acceptably low.

The points that spike on the maximum error graph indicate data points that may need to be removed. If there was more time available for this project, the data would have been cleaned more rigorously to eliminate the outliers clearly visible above in both the error plots (Figure 2.4) and the 2D surface fit plots (Figure 2.3).

3 Updating the Volatility Surface

3.1 Wasserstein Distance

The one dimensional Wasserstein Distance metric is defined as follows[5]:

$$W_1(\mu_1, \mu_2) = \int_{\mathbb{R}} |F_1(x) - F_2(x)| \, dx$$

(where F_1 and F_2 are cumulative density functions)
The Wasserstein metric, commonly referred to as the earth mover's distance, is applied in statistics to compare two different but related random variables and gives the optimal/lowest cost of changing from one distribution (F_1) to another (F_2).

This paper utilises the Wasserstein Distance metric to compare Risk Neutral Density functions as part of the modifying algorithm. Intuitively, after modifying a previous-day total variance surface to capture implied total variance for newly-added current-day data, we will aim to minimise the Wasserstein Distance metric between implied risk neutral distributions for the old and the new adjusted surface, thereby changing the local volatility surface as "little as possible".

The reasoning is that by modifying the surface in this way, it minimises the inconsistency described in the introduction. Additionally, by retaining as much data as possible from the previous day it should be possible to achieve a good fit using few data points which, as mentioned, is desirable because of hedging implication.

3.2 Risk-Neutral Density Functions as Function of Surface Parameters

Using the Breeden-Litzenberger equation - sourced from [4], one can obtain conditional risk-neutral stock distributions for a specific time-to-maturity slice. These will be functions of the parametrisation variables ρ, ϵ and the function θ_T.
The Breeden-Litzenberger equation is given by:

$$q_T(x) = \frac{\partial^2 C}{\partial K^2}_x B_T^{-1}$$

where $q_t(k)$ is the risk neutral price density function at time T, C is the call price surface and B_t is the interest discount factor.

The risk neutral distribution is utilised in the Wasserstein metric. The metric uses the cumulative distribution function of the price. Integrating, on obtains the following result:

$$F_T(x) = \int_{-\infty}^{x} [q_T(k)]\, dk$$

Note, integration is over the positive real line since the stock price cannot be negative.

$$= \int_{0}^{x} \left[\frac{\partial^2 C}{\partial K^2}_k B_T^{-1} \right] dk$$

$$= \left[\frac{\partial C}{\partial K} \right]_0^x B_T^{-1}$$

$$\text{Now} \quad \lim_{x\to\infty} F_T(x) = \left(\lim_{x\to\infty} \frac{\partial C}{\partial K}(x) - \frac{\partial C}{\partial K}(0) \right) B_T^{-1}$$

$$= (0 - \frac{\partial C}{\partial K}(0))$$

$$= 1$$

$$\text{So} \ \frac{\partial C}{\partial K}(0) = -B_T$$

$$\text{Finally, } F_T(x) = \left(\frac{\partial C}{\partial K}(x) + B_T \right) B_T^{-1}$$

We therefore need to solve for the analytical solution of $\frac{\partial C}{\partial K}(x)$, where C_t is the Black-Scholes Call Price:

$$C = (S_t Q_t \Phi(d_1) - K B_t \Phi(d_2))$$

$$d_1 = \frac{\ln \frac{S_t Q_t}{B_t K} + \frac{1}{2}\sigma^2 \tau}{\sigma\sqrt{\tau}} \quad \text{and}$$

$$d_2 = d_1 - \sigma\sqrt{\tau}$$

$$\sigma = \sqrt{\frac{w}{\tau}}$$

$$\frac{\partial C}{\partial K} = S_t Q_t \phi(d_1) \times \frac{\partial d_1}{\partial K} - K B_t \phi(d_2) \times \frac{\partial d_2}{\partial K} - B_t \Phi(d_2)$$

$$\frac{\partial d_1}{\partial K} = \frac{-\frac{1}{K}\sigma\sqrt{\tau} - \ln\frac{S_t Q_t}{K B_t}\sqrt{\tau \frac{\partial \sigma}{\partial K}}}{\sigma^2 \tau} - \frac{1}{2}\frac{\partial \sigma}{\partial K}\sqrt{\tau}$$

$$\frac{\partial d_2}{\partial K} = \frac{\partial d_1}{\partial K} - \frac{\partial \sigma}{\partial K}\sqrt{\tau}$$

$$\frac{\partial \sigma}{\partial K} = \frac{1}{2}\sqrt{\frac{1}{\tau w}}\frac{\partial w}{\partial K}$$

and where $w(T,k)$ is the parameterised total variance surface[8]

$$w(T,k) = \frac{\theta_T}{2}(1 + \rho\psi(\theta_T)k + \sqrt{(\psi(\theta_T)k + \rho)^2 + 1 - \rho^2}$$

$$\frac{dw}{dK} = \frac{\theta}{2}\frac{\psi(\theta_T)k + \rho)\psi(\theta_T}{\sqrt{\rho\psi(\theta_T) + (\psi(\theta_T)k + \rho)^2 + (1 - \rho^2))}}\frac{\partial k}{\partial K}$$

Furthermore, we obtain F_1 and F_2, the input functions to the Wasserstein Metric, at set time to maturities. F_1 is derived from the previous-day data, hence is a function of fixed parameters ρ and ϵ. F_2 is extracted from the modified variance surface (and the resulting call surface), and hence is a function of ρ and ϵ that will be varied in order adjust to optimise the fit to new data points.

As expected, the numerical and analytical CDFs are a close fit.
A key point to note is that F_2 is also of function of θ_T. θ_T = w(0,T)[8], the exact at the future money total variance. Because it is not a function of ρ or ϵ, θ_T cannot be interpolated from surrounding data points and must be specified before running the optimisation routine. This is minor, as data points are unlikely to be from options that are exactly at the forward price, but a bigger issue is how to obtain θ_T. One

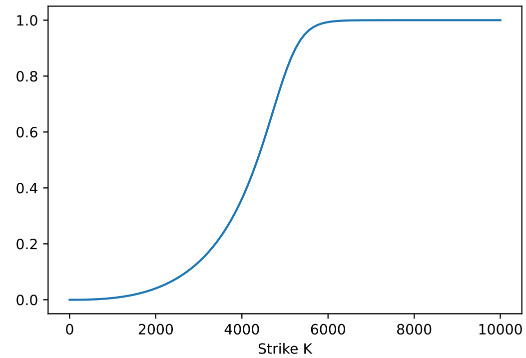


Figure 3.1: CDF of Risk-Neutral Density Function for SP500 share prices, calculated analytically

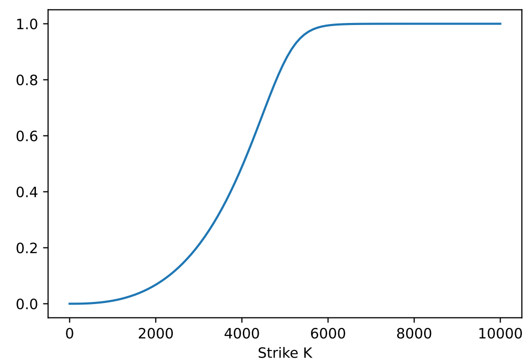


Figure 3.2: CDF of Risk-Neutral Density Function for SP500 share prices, calculated numerically

way to address this is to ensure that all at-the-money data points from the new day are included in the added data points so that θ_T can be defined before calibrating the new ρ or ϵ, in the same way as the one-day parameterisation earlier.

Having all our required inputs, we now proceed to construct an objective function which will be minimised in order to calibrate the adjusted values for ρ and ϵ.

3.3 Objective Function for Modifying Algorithm

The required objective function must consider two sources of error: firstly, if the adjusted total variance surface does not lie within the bid-ask spread for implied total variance from the new market data. Secondly, if the Wasserstein Distance metric is large. Weighting can be used to shift the importance of these two sources of error.

An appropriate objective function is as follows:

$$w_1 \sum_{\text{New Market Data}} \left(\max\left(w_{\text{fitted}}(t,k) - \sigma^2_{\text{implied, ask}} t; \sigma^2_{\text{implied, bid}} t - w_{\text{fitted}}(t,k); 0 \right)^2 \right)$$

$$+ w_2 \sum_{T_1, T_2, \cdots, T_n} \int_{-\infty}^{\infty} |F_1(x) - F_2(x)| dx$$

The time-to-maturity slices (T_i) were chosen to be the times for which option price data was available, so that θ_{T_i} values are available.

The interest rate and dividend discount factors, B_T and Q_T, which are used to find implied volatilities from market data with Black-Scholes, were solved from the full new day's data as described in Section 2. It is assumed that these values can be obtained from separate sources.

Using a minimisation algorithm, estimates for the adjusted ρ and ϵ values, and hence the entire adjusted variance surface (and resulting call surface) is obtained, for a given set of data additional data points.

The fit can be checked by considering the maximum and average errors of non-added market call and put prices versus prices derived from the fitted surface. This will indicate how well the adjustment method has performed based on the limited new data added. It will also assist in determining the necessary or optimal number of data points that need to be added in the modifying algorithm, which points to add and how good a fit can be expected after modification. How to add points and different uses of weights, w_1 and w_2 is discussed in the following section.

3.4 Performing the Update Step

The update step is tested in three test phases, gradually adding more data and altering the weights of the objective function:

Case 1:
The ϵ and ρ values from the previous day's calibration are used in the total variance function, along with the θ_T values calculated for the current day from at the money forward option prices.

Case 2:
The same data as in case one are used, and ϵ and ρ are calibrated. w_2 is set to 1. w_1 is set to a large constant, to ensure that the optimised first fits the surface to new data points then minimises the Wasserstein Distance.

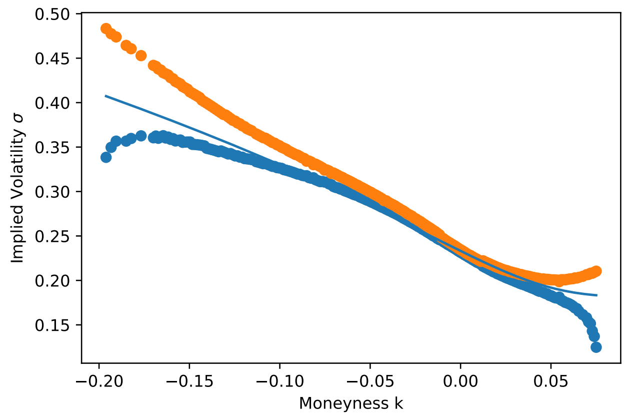
Case 3:

This builds on Case 2 by using the same weights and adding additional sparse data points. Data points are added selectively in an attempt to get the best fit with as few additional points as possible.

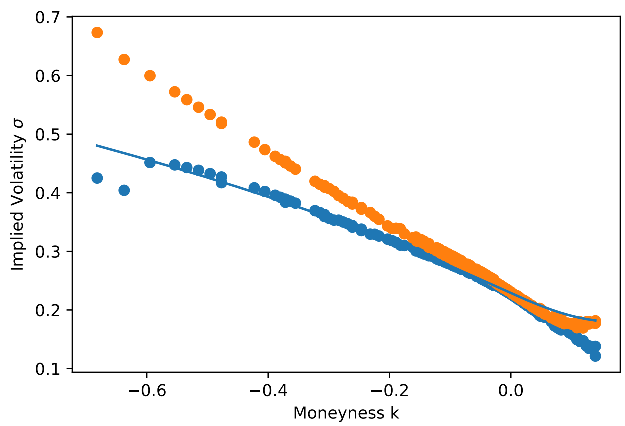
The cases above will be performed in sequence, stopping when a good-enough fit is found. If, for example, a close enough fit is found under Case 1 it is best not to increase the complexity unnecessarily. If case 3 is required to obtain a close fit, then an additional objective will be to add as few points as possible.
The fit after an update is performed will be examined by computing maximum and average errors for predicted total variance versus implied total variance from data that has not been added to the calibration. This should be checked over multiple days as, for example, more data may be required on highly volatile trading days.

3.5 Results from the Modification Algorithm

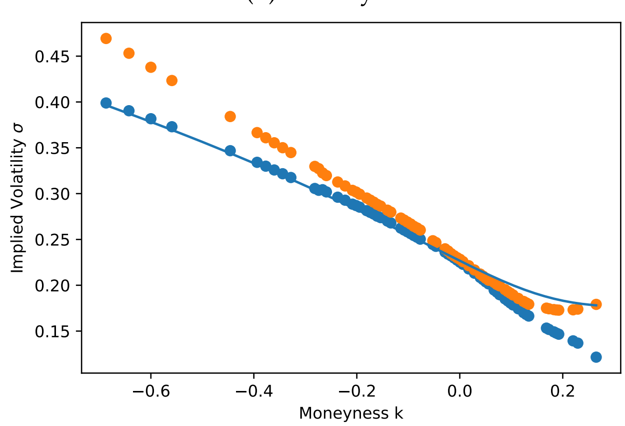
For all of the one-day intervals tested, Case 1 as described above, was sufficient for obtaining a good fit. This indicates that the one-day change in the total variance surface was minor enough that retaining the previous day values for ρ and ϵ with the new θ_T obtained from implied volatilities on current day at the money option prices.



(a) 15 days



(b) 121 days



(c) 366 days

Figure 3.3: Time slices showing fit of total variance surface over a range of log-moneyness ″k″.

A check of the time series of ρ and ϵ values was done to identify a day where a substantial change occurred, but the one-day change was still minor. Next, intervals longer than one day were considered, where there is a more drastic change in the total variance surface. On the 8th of March 2022 we found ϵ and ρ, the next day with a significant change was the 16th of March. Performing the Modified algorithm (Case 3) using the ϵ and ρ from the 8th of March (reference day) we obtained a new ϵ and ρ which was used to produce a total variance surface for the 16th of March. The new values are $\epsilon = 0.5529650569606183$ and $\rho = -0.6246179076807962$.

The fit achieved was satisfactory. This is a good indication that the Wasserstein metric may have merits in parameterising future volatility surfaces on days which appear to have volatility that is more than just marginally different to their reference date. We were unable to check this Wasserstein adjusted algorithm broadly enough to comment on how many future points we should include for optimal performance, or how it would fair compared to other metrics that retained information from the reference day. We can, however, conclude that it does appear to be a feasible way to include the information that comes with future data points while still retaining some information about the shape of the volatility surface (or more accurately the implied risk neutral densities) on the reference date.

4 Pricing Options Using Monte Carlo

4.1 Dupire Equation for Local Volatility

Using the Dupire equation, the total variance surface parameterised above can be transformed into the local volatility surface[6].

$$\sigma(T,K) = \sqrt{2\frac{\dfrac{\partial C}{\partial T} + rK\dfrac{\partial C}{\partial K}}{K^2\dfrac{\partial^2 C}{\partial K^2}}}$$

It can then be helpful to express this of the total variance surface $w(k,T)$, as opposed to its expression in terms of the Call Surface as stated above[7]:

$$\sigma(T,k) = \sqrt{\frac{\dfrac{\partial w}{\partial T}}{1 - \dfrac{k}{w}\dfrac{\partial w}{\partial k} + \dfrac{1}{2}\dfrac{\partial^2 w}{\partial k^2} + \dfrac{1}{4}(-\dfrac{1}{4} - \dfrac{1}{w} + \dfrac{k^2}{w^2})(\dfrac{\partial w}{\partial k})^2}}$$

For this investigation, the choice of parameterisation w is as follows, to eliminate calendar spread and butterfly arbitrage:

$$w(T,K) = \frac{\theta_T}{2}\left(1 + \rho\psi(\theta_T)k + \sqrt{(\psi(\theta_T)k + \rho)^2 + 1 - \rho^2}\right)$$

In order to use our results obtained with this parametrisation to price these American put options, we must obtain the volatility surface's time derivative, and first and second derivative with respect to moneyness. The following are the results from the derivations.

$$\frac{dw}{dt} = \frac{1}{2}\frac{\partial \theta_T}{\partial t}\left(1 + \rho\psi(\theta_T)k + \sqrt{(\psi(\theta_T)k + \rho)^2 + 1 - \rho^2}\right)$$
$$+\frac{\theta_T}{2}\left(\rho k\frac{\partial \psi(\theta_T)}{\partial t} + \frac{\psi(\theta_T)k + \rho)k\dfrac{\partial \psi(\theta_T)}{\partial t}}{\sqrt{(\psi(\theta_T)k + \rho)^2 + 1 - \rho^2}}\right)$$

$$\frac{d\psi(\theta_T)}{dt} = -\frac{\eta}{(\theta^{\gamma_1}(1+\beta_1\theta)^{\gamma_2}(1+\beta_2\theta)^{1-\gamma_1-\gamma_2})^2}\Big(\gamma_1\theta^{\gamma_1-1}(1+\beta_1\theta)^{\gamma_2}(1+\beta_2\theta)^{1-\gamma_1-\gamma_2}$$

$$+\gamma_2\theta^{\gamma_1}(1+\beta_1\theta)^{\gamma_2-1}(1+\beta_2\theta)^{1-\gamma_1-\gamma_2}\beta_1\frac{\partial\theta}{\partial t} + (1-\gamma_1-\gamma_2)\theta^{\gamma_1}(1+\beta_1\theta)^{\gamma_2}(1+$$

$$\beta_2\theta)^{-\gamma_1-\gamma_2}\beta_2\frac{\partial\theta}{\partial t}\Big)$$

$$\frac{d\theta}{dt} = \frac{\theta_{T2}-\theta_{T1}}{T_2-T_1}$$

$$\frac{dw}{dk} = \frac{\theta}{2}(\rho\psi(\theta_T) + \frac{(\psi(\theta_T)k+\rho)\psi(\theta_T)}{\sqrt{\rho\psi(\theta_T)+(\psi(\theta_T)k+\rho)^2+(1-\rho^2))}})$$

$$\frac{d^2w}{dk^2} = \frac{\theta\psi(\theta_T)}{2}(-\frac{1}{2}((\psi(\theta_T)k+\rho)^2+1-\rho^2)^{-\frac{3}{2}} \times 2\psi(\theta_T)(\psi(\theta_T)k+\rho)^2$$

$$+\frac{\psi(\theta_T)}{\sqrt{(\psi(\theta_T)k+\rho)^2+1-\rho^2}}$$

The derivatives above, when fed into the Dupire equation, specify a Local volatility Curve. The curve has the general correct shape[2] and is smooth in the strike direction. However, it has a sub-optimal shape in time dimension due to the exponential interpolation between times for which at the money option prices were available in the data (interpolation between θ_t).

Using this Local Volatility Curve one can read off a value for volatility, σ, at any point in strike and time to maturity space. This value for volatility could then be used and recalculated in each step in a Monte-Carlo pricing routine.

4.2 Longstaff Schwartz

To further test that the volatility surface calibrated in Section 3 generates prices consistent with the market, a method to price American options is required. This is done by using the Longstaff-Schwartz method adjusted to intake Monte-Carlo sample price paths generated using the local volatility surface.

The Longstaff-Schwartz method allows the pricing of American puts, whose values can differ from European puts due to their early exercise premium. The method achieves this by iterating backwards through times to maturity - from the maturity date to the start date. At each date where the option can be exercised the algorithm estimates the continuation value of the option if it is not exercised and compares it to the exercise value at that point.

There was insufficient time to reach this stretch goal of the project. The option prices retrieved using the Local Volatility Surface were reasonable, however, we were unable to complete the checking of the Local Volatility Surface using numerical techniques. We could not continue with the attempt of pricing the American Options using Longstaff-Schwartz. This means that we cannot comment on whether or not the market has accurately priced in an early exercise premium.

Bibliography

[1] Bahra, B., 1997. Implied risk-neutral probability density functions from option prices: theory and application .

[2] Dupire, B., 1994. Pricing with a smile .

[3] Jacquier, J.G..A., 2013. Arbitrage-free svi volatility surfaces .

[4] Litzenberger, D.B..R., 1978. Prices of state-contingent claims implicit in option prices .

[5] Panaretos, V.M., 2019. Statistical aspects of wasserstein distances .

[6] Schlögl, E., 2022a. Local volatility models .

[7] Schlögl, E., 2022b. Obtaining local volatility from implied total variance.

[8] Schlögl, E., 2022c. Specifying a volatility surface free of static arbitrage.

Using Gaussian Mean Mixtures for Pricing American Options

TEAM ALPHA

KEFENTSE DIPUDI, University of Cape Town
STEFRO MILLARD, University of Cape Town
ALEXIO PHYTIDES, University of Cape Town
NICHUME ZAMXAKA, University of Cape Town

Supervisor:
JÖRG KIENITZ, University of Cape Town / University of Wuppertal
THOMAS MCWALTER, University of Cape Town

African Collaboration for Quantitative Finance and Risk Research

Contents

1 Introduction

The field of numerical option pricing is vast, as the need for fast and accurate valuation and hedging of option prices is crucial in the realm of quantitative finance. Black and Scholes (1973) made a breakthrough by proposing a model with the underlying price following a geometric Brownian motion and deriving an analytical formula for European option prices. The Black-Scholes model is widely recognised as one of the most important contributions in modern financial theory, allowing analytical simplicity and tractability for European options. However, the overwhelming majority of traded options are of American type. The valuation of such financial instruments remains a topic of active research as there is a lack of closed-form solutions for evaluating such problems. This is due to the requirement of sophisticated approaches to tackle the free-boundary and optimal stopping problems inherent in the formulation of pricing American options. Many techniques, such as Least Squares Monte Carlo (Carriere, 1996; Longstaff and Schwartz, 2001), Finite Difference methods (Duffy, 2013), Fourier Transform methods (Carr and Madan, 1999) and Quantization (Pagès and Sagna, 2015), have been used to find approximate solutions of various efficiency and accuracy. Recently there has been a surge of interest in statistical machine learning approaches to pricing American type options, including the use of Gaussian Mean Mixture Models (Kienitz, 2021).

On 19 October 1987 (Black Monday), the Dow Jones Industrial Average fell by more than 22%, which marked the start of a global stock market decline (Bernhardt and Eckblad, 2013). Since then, volatility smiles and smirks have been prevalent in option markets, making the assumption of constant volatility controversial (Jones, 2003). The Heston model was subsequently established to overcome this limitation, as it models stochastic volatility which accounts for variations in the asset price and volatility and therefore, provides more realistic dynamics for the underlying asset prices.

This report investigates the use of the Gaussian Mean Mixture with Dynamically Controlled Kernel Estimation (GMM-DCKE) approach for pricing American options, under the Heston model. GMM-DCKE is a data driven and model free approach. It is model-free in the sense that it is not specific to a given model or class of models. We rely only on observed realizations at given time points. The benefit of GMM is that it can fit discontinuities better, which is beneficial in pricing certain exotic options that can have discontinuities in option payoff. To investigate the performance of the GMM-DCKE approach for pricing American options, we shall implement two traditional methods of pricing American options, under the Heston model, which will be used as benchmark approaches. These two benchmark approaches are the Least squares Monte Carlo and Finite Difference approaches. Through the implementation of these two traditional methods, we shall perform a rigorous comparison between these approaches and various GMM-DCKE approaches. The Least Squares Monte Carlo (LSMC) method, although widely used, has its shortcomings. The prices calculated using LSMC become less accurate for out-the-money options because less data is used to calculate the fitted values. The prices are also sensitive to the choice of polynomials used in the regression procedure. This led to the exploration of the Gaussian Mean Mixtures Model as it makes use of sophisticated machine learning techniques rather than regression to calculate the American option prices.

This report will focus on the valuation of options on a non-dividend paying underlying asset and therefore, will only investigate the valuation of American put options. We only consider the valuation of American put options, as under equivalent conditions, both American and European call options on a non-dividend paying underlying asset have the same valuation. The price of an American call option can be directly evaluated through semi-analytical option pricing formula, which uses Fourier inversion.

In addition to investigating the performance of the GMM-DCKE approach for pricing American put options, we also price exotic options under the same model. We shall specifically investigate the evaluation of prices for both American-type basket and rainbow options. These exotic options are dependent on multiple underlying assets and therefore, pricing these options will allow us to investigate the use of the GMM-DCKE approach in a multi-dimensional setting.

In this report, we give a brief overview of the Heston stochastic volatility model and the Quadratic Exponential (QE) Scheme, as these will be used to compute the stock price paths required in the American option pricing valuations. We then give a background on the various techniques that we have implemented to price American options, with the GMM-DCKE being the crucial method. Thereafter, we discuss the implementation of Least Squares Monte Algorithm, PDE schemes and GMM-DCKE method. The GMM-DCKE method, as mentioned previously, will be the focus of this report. Finally, we present the results and demonstrate comparisons of the prices computed by the GMM-DCKE method, with the two benchmark methods.

2 Background

2.1 Options

A call option is an agreement between a buyer and a seller that lets the buyer, at their discretion, purchase a financial instrument (in this case a stock) at a future date for a specified price. A put option is a similar contract that allows the holder to sell a financial instrument at a future date for a specified price. Two of the most common types of options are the European and American options. A European option can only be exercised at the expiration date, i.e. the maturity of the option. An American option allows the buyer exercise the option at any time up until the expiration date.

Since American options, unlike their European counterparts, allow exercise at any time between their issue and maturity, they are worth at least the same as, but usually more than, their European equivalent, due to the additional optionally offered by American options. The difference in price between American and European options is called the Early Exercise Premium. It can be shown that if a stock does not pay any dividends, the price of an American and European call option is the same, as it is never optimal to exercise an American call option before maturity (provided the underlying asset pays no dividends). This is because there is a time premium associated with the remaining life of an option, which makes early exercise suboptimal. Therefore, in this instance, one may treat American call options as European call options (in other words, there is no early exercise premium for American call options). However, this is not the case for an American put option (regardless of whether the underlying asset pays dividends or not), as there are conditions in which an American put option may be rationally early-exercised. This is because the holder can benefit from interest by exercising early and investing the payoff at the risk-free rate.

2.2 Heston Stochastic Volatility Model

The Heston Model is a stochastic volatility model that seeks to overcome the shortcomings presented by the assumption of constant volatility of the underlying asset, assumed in the Black-Scholes model (Black and Scholes, 1973). It allows the asset's variance to be modelled as a stochastic process in order to produce a more realistic model for the dynamics of asset prices. The Heston model assumes that the underlying asset, S_t, follows a Black-Scholes type stochastic process, but with a stochastic variance V_t (Rouah, 2013). The variance is mean reverting and follows a square-root Cox, Ingersoll, and Ross (CIR) process. Hence, the Heston Model is defined by the following bivariate system of stochastic differential equations (SDEs), which represent the instantaneous asset price and change in variance under a risk neutral probability measure,

$$
\begin{aligned}
dS_t &= rS_t dt + \sqrt{V_t} S_t dW_t^s & (1)\\
dV_t &= \kappa(\theta - V_t)dt + \sigma\sqrt{V_t} dW_t^v, & (2)
\end{aligned}
$$

where $(W_t^s)_{t\geq 0}$ and $(W_t^v)_{t\geq 0}$ are Brownian motions with $\mathbb{E}[dW_t^s dW_t^v] = \rho dt$, for the correlation constant $\rho \in$ [-1,1] (Rouah, 2013).

The parameters of the Heston model are defined in Table 1.

Parameter	Definition
r	The risk-free interest rate
θ	The long-term price variance
κ	The rate of mean reversion to the long-term price variance
v_0	The initial level of the variance
σ	The volitility of the variance

Table 1: Heston model parameters.

2.3 Quadratic Exponential Scheme

The calculation of American option prices using Monte Carlo methods will require simulation of the stock price paths. Under the Heston model, this will mean simulating a set of time-dependent stock prices and variances for each path, from which option prices can be evaluated. There are a range of methods available to generate these paths such as the popular Euler, Milstein or implicit Milstein schemes. While these methods have well-known convergence properties, they occassionally generate negative values for the variances, even when the Feller condition, $2\kappa\theta > \sigma^2$, is met. The negative variances can be overridden by either adopting a full truncation scheme or a reflection scheme. In the full truncation scheme, the negative variances v_t are sent to zero. In the reflection scheme, the negative variances v_t are reflected with $-v_t$. Although this ensures positive variances, the full truncation scheme creates zero variances and the reflection scheme can create very large variances. Other techniques include IJK scheme, quadratic-exponential scheme, transformed volatility scheme and the scheme of Alfonsi (Rouah, 2013). In Andersen (2007), the Quadratic Exponential (QE) scheme was proposed for simulating the Heston variance process. This will be explored further.

Solving the stochastic differential equation of the variance process results in realizations of v_{t+dt} conditional on v_t that follow the non-central chi-square distribution. The QE scheme relies heavily on this. The v_t is estimated by sampling from an approximation of the non-central chi-square distribution. The method proposes two algorithms that approximate the non-central chi-square distribution with the choice of each algorithm depending on the magnitude of v_t. If the value of v_t is moderate to high, then v_{t+dt} can be approximated by applying a power function to a standard normal random variable Z_V,

$$v_{t+dt} = a(b + Z_V)^2, \tag{3}$$

where a and b are determined by moment matching methods using the mean and the variance of the Cox, Ingersoll and Ross (CIR) process. The mean and the variance of the CIR process are:

$$\begin{aligned} m &= \theta + (v_t - \theta)e^{-\kappa dt} = \mathbb{E}\left[v_{t+dt} \mid v_t\right] \\ s^2 &= \frac{v_t \sigma^2 e^{-\kappa dt}}{\kappa}(1 - e^{-\kappa dt}) + \frac{\theta\sigma^2}{2\kappa}(1 - e^{-\kappa dt})^2 = \text{Var}\left[v_{t+dt} \mid v_t\right]. \end{aligned}$$

Given that v_{t+dt} can be expressed as in (3), this can be further simplified to

$$\begin{aligned} v_{t+dt} &= a(b + Z_V)^2, \\ &= a(b^2 + 2bZ_V + Z_V^2), \end{aligned}$$

where $\mathbb{E}\left[v_{t+dt}\right] = a(b^2+1)$ and $\text{Var}[v_{t+dt}] = 2a^2(1+2b^2)$ are deduced accordingly. By equating these two equations with m and s^2 and solving for a and b simultaneously, the solutions to a and b are:

$$\begin{aligned} b &= \left(\frac{2}{\psi} - 1 + \sqrt{\frac{2}{\psi}\left(\frac{2}{\psi} - 1\right)}\right)^{\frac{1}{2}} \\ a &= \frac{m}{1 + b^2}, \end{aligned}$$

where $\psi = \frac{s^2}{m^2}$. The value of b is only defined for $\psi \leq 2$ and this agrees with the density of the variance process v_t being far from zero. For values of v_t that are small, v_{t+dt} will be approximated differently. For small values of v_{t+dt}, the cumulative distribution function of v_{t+dt} can be approximated using

$$\mathbb{P}\left[v_{t+dt} \leq x + dx\right] = p + (1 - p)(1 - e^{-dx}), \quad x \geq 0,$$

with the corresponding probability distribution function being the weighted average of a term including the Dirac delta function δ and a term including e^{-dx} given by the following expression:

$$\mathbb{P}\left(v_{t+dt} \in [x + dx]\right) = p\delta(0) + \beta(1 - p)(1 - e^{-dx}), \quad x \geq 0.$$

The expressions for p and β are also found by moment matching as in the first algorithm with $p \in [0, 1]$ and $\beta \in \mathbb{R}$. From the PDF of v_{t+dt}, we have

$$\mathbb{E}\left[v_{t+dt}\right] = \frac{1 - p}{\beta} \quad \text{and} \quad \text{Var}\left[v_{t+dt}\right] = \frac{1 - p^2}{\beta^2}.$$

Equating the above with m and s^2 and solving for p and β gives

$$p = \frac{\psi - 1}{\psi + 1} \quad \text{and} \quad \beta = \frac{1-p}{m}.$$

The value of p must be greater than 0 and this will happen if and only if $\psi \leq 1$. The first approximation of v_{t+dt} required that $\psi \leq 2$. Combining these two conditions means that the critical value ψ_c that is chosen should be such that $\psi_c \in [1,2]$. To approximate the values of v_{t+dt}, the inverse distribution function for the above must be computed. Inverting the cumulative distribution function produces the inverse distribution function, which can be written as:

$$\mathbf{\Psi}^{-1}(u) = \begin{cases} 0 & \text{for} \quad 0 \leq u \leq p \\ \frac{1}{\beta}\ln\frac{1-p}{1-u} & \text{for} \quad p \leq u \leq 1. \end{cases}$$

Finally, $v_{t+dt} = \mathbf{\Psi}^{-1}(U_V)$ where U_V is a uniform random number. The QE scheme can be summarized as a rule that switches between the two sampling algorithms. The first algorithm being executed for $\psi \leq \psi_c$ and the second for $\psi > \psi_c$. We now look at generating the stock prices using the variance process generated above.

Andersen (2007) abandons the Euler discretization for $\ln S_t$ to avoid the problem of "leaking correlation". Using the Euler discretization results in the correlation between $\ln S_{t+dt}$ and v_{t+dt} being closer to zero than ρ. He shows through the use of a characteristic function that the correlation between $\ln S_{t+dt}$ and v_{t+dt} is in fact closer to ρ. We use the discretization in the paper that corrects for this correlation. The integral form of the process for v_t is

$$v_{t+dt} = v_t + \kappa\theta dt - \kappa \int_t^{t+dt} v_u du + \sigma \int_t^{t+dt} \sqrt{v_u} dB_{2,u},$$

which can also be written as

$$\int_t^{t+dt} \sqrt{v_u} dB_{2,u} = \frac{1}{\sigma}\left(v_{t+dt} - v_t - \kappa\theta dt + \kappa \int_t^{t+dt} v_u du\right). \tag{4}$$

Using Ito's lemma to solve for $\ln S_t$ together with (1) and applying the Cholesky decomposition thereafter produces the following integral form for the increment of $\ln S_t$

$$\begin{aligned} \ln S_{t+dt} = \ln S_t + rdt - \frac{1}{2}\int_t^{t+dt} v_u du \\ + \int_t^{t+dt} \sqrt{v_u}\left(\rho dB_{2,u} + \sqrt{1-\rho^2} dB_{1,u}\right), \end{aligned} \tag{5}$$

where $B_{1,u}$ and $B_{2,u}$ are independent Brownian motions. Substituting (4) into (5) gives

$$\begin{aligned}\ln S_{t+dt} &= \ln S_t + rdt + \frac{\rho}{\sigma}\left(v_{t+dt} - v_t - \kappa\theta dt\right) \\ &\quad + \left(\frac{\kappa\rho}{\sigma} - \frac{1}{2}\right)\int_t^{t+dt} v_u du + \sqrt{1-\rho^2}\int_t^{t+dt}\sqrt{v_u}dB_{1,u}.\end{aligned} \tag{6}$$

The following approximations for the integrals above were made in the paper:

$$\int_t^{t+dt} v_u du \approx dt(\gamma_1 v_t + \gamma_2 v_{t+dt}) \tag{7}$$

$$\int_t^{t+dt}\sqrt{v_u}dB_{1,u} \approx Z_V\sqrt{dt}\sqrt{\gamma_1 v_t + \gamma_2 v_{t+dt}}. \tag{8}$$

Finally, substituting the approximations into (6) produces the discretization for $\ln S_t$

$$\begin{aligned}&\ln S_{t+dt} = \ln S_t + rdt + K_0 + K_1 v_t + K_2 v_{t+dt} + \sqrt{K_3 v_t + K_4 v_{t+dt}}Z_V \\ \Rightarrow \quad &S_{t+dt} = S_t e^{rdt + K_0 + K_1 v_t + K_2 v_{t+dt} + \sqrt{K_3 v_t + K_4 v_{t+dt}}Z_V},\end{aligned}$$

where

$$\begin{aligned}K_0 &= \frac{-\kappa\rho\theta}{\sigma}dt, \\ K_1 &= \left(\frac{\kappa\rho}{\sigma} - \frac{1}{2}\right)\gamma_1 dt - \frac{\rho}{\sigma}, \\ K_2 &= \left(\frac{\kappa\rho}{\sigma} - \frac{1}{2}\right)\gamma_2 dt + \frac{\rho}{\sigma}, \\ K_3 &= \left(1-\rho^2\right)\gamma_1 dt, \\ K_4 &= \left(1-\rho^2\right)\gamma_2 dt.\end{aligned}$$

Andersen (2007) suggests that there are multiple ways for setting the constants γ_1 and γ_2. The common ways include setting $\gamma_1 = \gamma_2$, giving a central discretisation, or setting $\gamma_1 = 1, \gamma_2 = 0$, giving an Euler-type discretization.

2.4 American Option Pricing

Since the introduction of the Black-Scholes model, the problem of pricing American options has become extremely important, because the majority of financial derivatives traded in the markets are American-style derivatives. While the Black-Scholes model provides an analytical solution to the valuation of European options, the possibility of early exercise makes pricing an American option a more complicated task compared with its European counterpart. Therefore, to price an American option one would need to generalise the method to allow for early exercise.

The theory of American option valuation presents some mathematically challenging obstacles such the relationship between optimal stopping and free-boundary problems. These problems stem from the fact that pricing American options involves finding an optimal exercise strategy for the option. Therefore, due to the complexity of the additional optionality offered by an American option, there is no closed-from solution to the valuation of this option.

In general, the price of an option can be calculated by evaluating the expectation of the discounted option payoff under a risk-neutral measure. Since the holder of an American option can choose to exercise the option at any time until maturity, the pricing of an American option involves an optimal stopping problem. The value of an American option is the supremum over a range of possible stopping times of the risk-neutral expectation of the discounted payoff of the option. Therefore, the value of an American put option is defined as

$$\sup_{\tau \in \mathcal{T}} \mathbb{E}^{\mathbb{Q}} \left[e^{-r\tau} \left(K - S_\tau \right)^+ \right].$$

Here, $\mathbb{Q}$ is a risk neutral probability measure and $(K - S_\tau)^+$ represents the payoff of the put option at time τ. Therefore, valuing an American option involves finding the optimal time to exercise the option.

There are several methods for pricing an American option which value the option by approximating it as a Bermudan option. A Bermudan option is a restricted form of the American option that allows for early exercise at set dates.

To accurately value an American option, one needs to make use of a numerical approach. Some of the most commonly used numerical methods for pricing American options include the Binomial Lattice model, Finite Difference methods, Quadratic approximation and Least Squares Monte Carlo methods (Kienitz and Wetterau, 2013). In this report we shall implement two traditional techniques for American put option pricing, namely the Least squares Monte Carlo and Finite Difference method techniques, and use them as benchmark approaches to compare to our evaluation of the GMM-DCKE approach for pricing American options.

2.5 Least-Square Monte Carlo Method

The Least-Squares Monte Carlo (LSMC) first proposed by Carriere (1996) and popularized by Longstaff and Schwartz (2001) provides an efficient algorithm for pricing American options. The algorithm can be applied to any stock price process that lends itself to simulation (Rouah, 2013). The purpose of this method is to determine the optimal time to exercise the American option with each simulation done. This is because the American option should only be exercised once during the term of the option. Each path simulation has only one optimal time to exercise.

The method implicitly determines the optimal time by finding the optimal stock prices to exercise the option. The option prices are evaluated using the backward dynamic programming formulation.
The method relies on assuming that the underlying variable processes are Markovian. The key idea of this approach is to use least squares to estimate the conditional expected continuation value at each time step. The continuation value is the value of the option if the holder does not exercise immediately at that point in time. The LSMC method relies on discretizing exercise times to a finite set $\mathcal{T}_i = \{t_i < t_{i+1} < ... < t_N = T\}$ with $t_0 = 0$. The LSMC method thus values Bermudan-style options, with the values of these options approaching American-style options as the discretization time steps become larger (Palupi et al., 2015).

The Least-Squares method calculates the early-exercise option through a backward inductive process. The value of the option at each time step for a single (risk-neutral) path is given recursively by the following equations:

$$\begin{aligned} &\mathcal{V}_N(S_N) = H_N(S_N) \\ &\mathcal{V}_{i-1}(S_{i-1}) = \begin{cases} H_{i-1}(S_{i-1}) & \text{if} \quad H_{i-1}(S_{i-1}) > \mathbb{E}\left[e^{-r\Delta t_i}\mathcal{V}_i(S_i) \mid S_{i-i}\right] \\ e^{-r\Delta t_i}\mathcal{V}_i(S_i) & \text{otherwise,} \end{cases} \end{aligned}$$

for $N \geq i \geq 1$ and where S_i represents the asset price at time t_i. The $H_{i-1}(S_{i-1})$ stands for the immediate exercise payoff at time t_{i-1}. At each time step, the value of the option is calculated as the maximum of the immediate exercise payoff and the continuation value at that time. Calculating the American option price is then given as $V_0 = \mathbb{E}\left[\mathcal{V}_1(S_1)\right]$. The difficulty with valuing American options is the evaluation of the conditional expectations at each time.
Longstaff and Schwartz (2001) propose a way of calculating these conditional expectations as the fitted values of a least-square regression of a set of basis functions. They assume that the conditional expectation can be written as

$$e^{-r\Delta t_i}\mathbb{E}\left[\mathcal{V}_i(S_i) \mid S_{i-1} = x\right] = f(\hat{\beta}_{i-1}, x),$$

where $f(\hat{\beta}_{i-1}, x) = \sum_{r=0}^{R} \beta_r \phi_r(x)$ in terms of the basis functions $\phi_r(x), 0 \leq r \leq R$ and $\beta = [\beta_0, \beta_1, ..., \beta_R]$. Longstaff and Schwartz select the basis functions to be the Laguerre polynomials. Other basis functions including Hermite, Legendre, Chebyshev, Gegenbauer and Jacobi polynomials are also possible options. The β parameters are estimated using a regression procedure and simultaneously used to calculate the conditional expectations. The algorithm for the Least Squares Monte Carlo is presented in algorithm 1.
At maturity, the vector $\mathcal{V}_N$ of the terminal payoffs is computed. At each t_k for $1 \leq k \leq N-1$, the value of the option is set to be the maximum of the early

exercise payoff and the realized continuation value. This process is performed recursively until time t_1 is reached. The price of the option is then the average of the continuation values assuming that there is no early exercise is not possible at t_0.

2.6 Heston Model PDE

The price (U) of a contingent claim, on a non-dividend paying stock, under in the Heston model, (1) and 2, is given as the solution to the following PDE :

$$\frac{\partial U}{\partial \tau} = \frac{1}{2}\nu S^2 \frac{\partial^2 U}{\partial S^2} + rs\frac{\partial U}{\partial S} - rU + \frac{1}{2}\sigma^2\nu\frac{\partial^2 U}{\partial \nu^2} + rS\frac{\partial U}{\partial S} + \kappa(\theta - \nu)\frac{\partial U}{\partial \nu}, \tag{9}$$

where $\tau = T - t$, ν is the spot volatility and the rest of the parameters are defined as in Table (1). As with any PDE, a boundary and initial conditions are required. We give these for an American put option.

$$\begin{aligned}
U(S,\nu,T) &= (K-S)^+ \\
\frac{\partial U}{\partial S} &\xrightarrow[S]{\infty} 0 \\
U(0,\nu,\tau) &= 0 \\
\frac{\partial U(S,0,\tau)}{\partial \tau} &= rS\frac{\partial U(S,0,\tau)}{\partial S} - rU(S,0,\tau) + rS\frac{\partial U(S,0,\tau)}{\partial S} + \kappa\theta\frac{\partial U(S,0,\tau)}{\partial \nu} \\
U(S,\infty,\tau) &= (K-S)^+.
\end{aligned} \tag{10}$$

The boundary condition in (10) can also be stated as $U(\infty,\nu,\tau) = K$. The finite difference method is used to find numerical approximations of this PDE, and is described in detail in the implementation section.

2.7 Gaussian Mean Mixture Models

A Gaussian Mixture Mean model is a weighted sum of Gaussian component densities. It is a parametric probability density function, where parameters are estimated using training data and an iterative Expectation-Maximization(EM) algorithm. More formally it is specified as follows:

2.7.1 Definition

Definition 2.1. let $\mathbf{x} \in \mathbb{R}^d$ be a continuous-valued data vector, A Gaussian mixture model with K components, denoted as GMM(K) is defined as

$$p(\mathbf{x}|\theta) = \sum_{k=1}^{K} \omega_k f(\mathbf{x}|\mu_{\mathbf{k}}, \mathbf{\Sigma_k}), \tag{11}$$

where ω_k and $f(\mathbf{x}|\mu_{\mathbf{k}}, \mathbf{\Sigma_k})$, $k = 1, ..., K$, are the mixture weights and component Gaussian densities respectively. The mixture weights are strictly positive and sum to one. Each density is a d-dimensional Gaussian function of the form,

$$f(\mathbf{x}|\mu_{\mathbf{k}}, \mathbf{\Sigma_k}) = \frac{1}{(2\pi)^{(\frac{d}{2})}|\Sigma_k|^{\frac{1}{2}}} \exp \frac{1}{2}(\mathbf{x} - \mu_{\mathbf{k}})^T \Sigma_k^{-1}(\mathbf{x} - \mu_{\mathbf{k}})$$

with covariance matrix Σ_k and mean vector $\mu_{\mathbf{k}}$. Gaussian mixture models are parameterized by the mean vectors, covariance matrices and mixture weights from all the component densities, denoted as $\theta = \{\omega_k, \mu_{\mathbf{k}}, \Sigma_k\}$, $k = 1, ..., K$.

GMM comes in a variety of forms. The covariance matrices, may be restricted to be diagonal or full rank. The Gaussian components can also share or be connected to parameters, such as having a single covariance matrix for all components, the quantity of data available for estimating the GMM parameters frequently dictates the model architecture (number of components, complete or diagonal covariance matrices, and parameter) Reynolds (2009).

Gaussian mixture models are often used for data clustering. Clustering is defined as the unsupervised classification of data into homogeneous groups (Najar et al., 2017). We can use GMMs to perform either hard clustering or soft clustering on data. Hard clustering assigns data point to the multivariate normal components that maximize the component posterior probability given data, it assigns a data point to one cluster. Soft clustering assigns a score to a data point for each cluster, the value of the score indicates the associated strength of the data point to the cluster. They can assign a data point to more than one cluster. GMM clustering accommodates clusters that have different sizes and correlation structures within them, however it requires you to specify the number of clusters before fitting the model.

2.7.2 Parameter Estimation

Estimating the parameters for the GMM model is done using maximum likelihood estimation (MLE). The likelihood function is defined as

$$\mathcal{L}(\theta|\mathbf{x}) = \prod_{i}^{N} p(\mathbf{x_i}|\theta). \tag{12}$$

In other words the likelihood function is the joint probability distribution of the random vectors $\mathbf{x_i}$, $i = 1, \cdots, N$. Each $\mathbf{x_i}$ has density function as in (11). The parameters of the GMM model are estimated by finding the parameters θ that maximise the joint probability. In order to avoid the product in (12), the log-likelihood is rather maximised since it will have an extremum at the same parameter values. The log-likelihood function for the GMM, (11), is then,

$$l(\theta|\mathbf{x}) = \sum_{i}^{N} \log\left(\sum_{k}^{K} \omega_i f(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)\right). \tag{13}$$

This is a constrained maximisation problem since $\sum_i^K \omega_i = 1$ and $\omega_k > 0$. The log-likelihood function can be maximised with the Expectation Maximisation (EM) scheme. First we need to introduce the idea of a latent variable. A latent variable, is a variable that can not be directly observed. In the GMM framework the label of component k, is the latent variable z, where

$$\begin{aligned} p(z) &= \omega_k \\ p(\mathbf{x}|z) &= f(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k) \\ p(\mathbf{x}, z) &= \omega_k f(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k). \end{aligned}$$

The EM scheme has 2 steps, namely an E-step and a M-step. In the E-step, calculate

$$\begin{aligned} p(z|\mathbf{x}_i) &= \frac{p(\mathbf{x}_i, z)}{p(\mathbf{x}_i)} \\ &= \frac{\omega_k f(\mathbf{x}_i|\mu_k \mathbf{\Sigma_k})}{\sum_l^K \omega_l f(\mathbf{x}_i|\mu_k \mathbf{\Sigma}_k)} =: \gamma_{ik}. \end{aligned}$$

In the M-step, maximise

$$\begin{aligned} & \sum_{i}^{N} \sum_{k}^{K} \gamma_{ik} \log(p(\mathbf{x}_i, z)) \\ = & \sum_{i}^{N} \sum_{k}^{K} \gamma_{ik} \log(f(\mathbf{x}_i|\mu_k, \mathbf{\Sigma_k})) + \sum_{i}^{N} \sum_{k}^{K} \gamma_{ik} \log(\omega_k). \end{aligned} \tag{14}$$

Maximising this with respect to ω_k can be done using the Lagrange multiplier method, where

$$L = \sum_i^N \sum_k^K \gamma_{ik} \log(\omega_k) - \lambda \left(\sum_k^K \omega_k - 1 \right).$$

Conditioning on $\omega_k > 0$ is not necessary since the $\log(\omega_k) \xrightarrow[\omega_k]{0} \infty$. solving $\frac{dL}{d\omega_k} = 0$, yields

$$\begin{aligned} \hat{\omega_k} &= \frac{\sum_i^N \gamma_{ik}}{\sum_i^N \sum_k^K \gamma_{ik}} \\ &= \frac{\sum_i^N \gamma_{ik}}{N}. \end{aligned} \quad (15)$$

Maximising (14) with respect to μ_i and $\mathbf{\Sigma}_i$ leads to the following estimates

$$\hat{\mu_k} = \frac{\sum_i^N \gamma_{ik} \mathbf{x}_i}{\gamma_{ik}} \quad (16)$$

$$\hat{\mathbf{\Sigma}}_k = \frac{\sum_i^N \gamma_{ik} (\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^T}{\sum_i^N \gamma_{ik}}. \quad (17)$$

From (15), (16) and (18), it might seem as if the estimates can be calculated directly, but the estimates are dependent on $p(z|\mathbf{x}_i)$, which itself is dependent on $\{\omega_k, \mu_k, \mathbf{\Sigma}_k\}$. To solve this problem, we implement the EM algorithm. The EM algorithm, (3), is a recursive scheme that takes as input an initial guess of parameters. It then uses these initial parameters to recursively calculate accurate estimates of the parameters.

3 Implementation of Pricing Techniques

Throughout this report, the parameters as defined in Table 1, will take on the values in Table 2. These parameters will be used to generate the stock paths needed to implement the different approaches for evaluating the option prices.

Parameter	Value
r	0.1
θ	0.16
κ	5
v_0	0.0625

Table 2: Heston model parameter values.

3.1 Least-Squares Monte Carlo Method

A summary of the Least Squares Monte Carlo method is as follows:

A The first step is to simulate the stock price paths. The quadratic exponential scheme using the parameters above was used to generate the stock paths. A sample size/number of paths of 50000 and weekly exercise dates were used.

B The next step is to calculate the terminal payoff for each path. Since an American put option is being considered, the payoff is in this form:

$$H_N(S_N) = \text{max}(K - S_N, 0).$$

The payoffs are then discounted to the previous time step t_{N-1} using the following:

$$\mathcal{V}_{N-1}(S_{N-1}) = e^{-r\Delta(t_N - t_{N-1})}\text{max}(K - S_N, 0).$$

The stock prices for which $H_{N-1}(S_{N-1})$ is greater than zero are found. These will be used as regressors in the least squares regression procedure.

C Calculate $\mathcal{V}(S_{N-1})$ for only the stock prices identified under B. If we set this vector of values to Y and the regressors as X, the formulation is as follows:

$$\hat{\beta} = (FF^T)^{-1}FY \quad \text{and} \quad f(\hat{\beta}, X) = F^T\hat{\beta}$$

where F is the matrix of the Laguerre polynomials presented below.

D The next step is to compare $H_{N-1}(X)$ to $f(\hat{\beta}, X)$. For values of X where $H_{N-1}(X) < f(\hat{\beta}, X)$ this means that early exercise is optimal for those stock prices. Hence, $\mathcal{V}_{N-1}(X^{'})$ must be set equal to the early exercise payoff, where $X^{'}$ is the stock prices for which the above condition is met.

E The above steps are repeated recursively backward through the time steps until t_1 is reached. It is assumed that early exercise is not applicable at t_0 as mentioned above. Hence, the price of the American put option is computed as $\mathcal{V}_0 = \mathbb{E}\left[\mathcal{V}_1\right]$. A summary of the algorithm is presented below.

Algorithm 1 Algorithm for the Least Squares Monte Carlo.

1. Use the Quadratic Exponential Scheme to generate the stock price paths with values at each of the exercise dates.
2. Set $\mathcal{V}_N$ to be the vector of terminal payoffs for each of the paths.
3. At each iteration evaluate the realised continuation values as:
$$\mathcal{V}_{i-1} = e^{-r\Delta t_i}\mathcal{V}_i.$$
4. Identify paths for which $H_{i-1}(S_{i-1}) > 0$ where H represents the payoff function.
5. Set the vector X to be the stock prices for the corresponding paths (S_{i-1}) and Y to be the realized continuation values for these paths.
6. Perform least squares regression on Y and $f(\hat{\beta}, X)$ to produce an estimate of $\hat{\beta}$.
7. Calculate $f(\hat{\beta}, X)$ as the conditional expected continuation values. For the stock prices in X where the early exercise payoff is greater than the corresponding continuation value, set $\mathcal{V}_{i-1}$ to be the early exercise values($H_{i-1}(S_{i-1})$).
8. Repeat steps 3-7 for $i = N, N-1, ..., 2$.
9. Compute the value of the option as $\mathcal{V}_0 = \mathbb{E}\left[e^{-r\Delta t_1}\ \mathcal{V}_1\right]$.

In this report, we explored changing the number of Laguerre polynomials to see whether this improved the accuracy of the option prices calculated. The comparison section provides more detail on this. Presented below are the Laguerre polynomials that were used in the regression procedure:

$$\begin{aligned}
\phi_0(x) &= 1 \\
\phi_1(x) &= 1 - x \\
\phi_{k+1} &= \frac{(2k+1-x)\phi_k(x) - k\phi_{k-1}(x)}{k+1}
\end{aligned} \tag{18}$$

3.2 Finite Difference Method

In order to use the finite difference method to solve, (9), it is necessary to discretise the state variables S, ν and τ. This is done by creating a grid on which the variables lie. This grid may have uniform spacings or non-uniform spacings. A non-uniform grid will give more accurate results than a uniform grid. A uniform grid is formed by discretising the variables as follows,

$$\begin{aligned} Si &= i \times ds, & i &= 0, ..., Ns \\ \nu_j &= j \times d\nu, & j &= 0, ..., N\nu \\ \tau_n &= n \times d\tau, & n &= 0, ..., N\tau \end{aligned}$$

where $ds = \frac{S_{max} - S_{min}}{Ns}$, $d\nu = \frac{\nu_{max} - \nu_{min}}{N\nu}$ and $d\tau = \frac{\tau_{max} - \tau_{min}}{N\tau}$.
A non-uniform grid can be created by discretising the state variables so that the grids are finer around the strike price and $\nu_0 = 0$ (In't Hout and Foulon, 2010). This grid has the following form

$$S_i = K + c \sinh \xi_i, \qquad i = 0, ..., Ns$$

where

$$\begin{aligned} \xi_i &= \sinh^{-1}\left(\frac{K}{C}\right) \\ \Delta\xi_i &= \frac{1}{Ns}[\sinh^{-1}\left(\frac{S_{max} - K}{C}\right) - \sinh^{-1}\left(\frac{-K}{C}\right)]. \\ \nu_j &= d \sinh(j\Delta\eta), \qquad j = 0, ..., N\nu, \end{aligned} \tag{19}$$

with

$$\Delta\eta = \frac{1}{N\nu} \sinh^{-1}\left(\frac{V_{max}}{d}\right).$$

In't Hout and Foulon (2010) , use $C = \frac{K}{5}$ and $d = \frac{\nu_{max}}{500}$. The grid for τ remains uniform. Figure 1 illustrates this non-uniform grid.

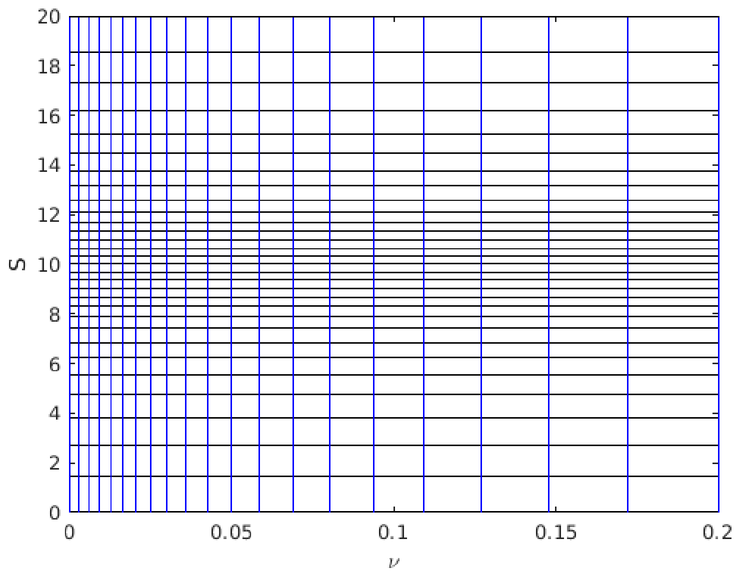


Figure 1: Non-uniform grid as described in In't Hout and Foulon (2010).

The finite difference method uses approximations for the derivatives in (9). The interior points on the grid use central differences formulae as follows,

$$
\begin{aligned}
\frac{\partial U(S_i,\nu_j)}{\partial S} &= \frac{U^n_{i+1,j}-U^n_{i-1,j}}{S_{i+1}-S_{i-1}} && (20)\\
\frac{\partial U(S_i,\nu_j)}{\partial \nu} &= \frac{U^n_{i,j+1}-U^n_{i,j-1}}{\nu_{j+1}-\nu_{j-1}} && (21)\\
\frac{\partial^2 U(S_i,\nu_j)}{\partial S^2} &= \frac{U^n_{i-1,j}}{(S_i-S_{i-1})(S_{i+1}-S_{i-1})} - \frac{2U^n_{i,j}}{(S_i-S_{i-1})(S_{i+1}-S_i)} && (22)\\
&\quad + \frac{U^n_{i+1,j}}{(S_{i+1}-S_i)(S_{i+1}-S_{i-1})}\\
\frac{\partial^2 U(S_i,\nu_j)}{\partial \nu^2} &= \frac{U^n_{i,j-1}}{(\nu_j-\nu_{j-1})(\nu_{j+1}-\nu_{j-1})} - \frac{2U^n_{i,j}}{(\nu_j-\nu_{j-1})(\nu_{j+1}-\nu_j)} && (23)\\
&\quad + \frac{U^n_{i,j+1}}{(\nu_{j+1}-\nu_j)(\nu_{j+1}-\nu_{j-1})}\\
\frac{\partial^2 U(S_i,\nu_i)}{\partial \nu \partial S} &= \frac{U^n_{i+1,j+1}+U^n_{i-1,j-1}-U^n_{i+1,j-1}-U^n_{i-1,j+1}}{(S_{i+1}-S_{i-1})(\nu_{j+1}-\nu_{j-1})}. && (24)
\end{aligned}
$$

On the boundary of the grid forward and backward difference formulae for the derivatives are used. For example the forward difference formula for the first derivative is

$$\frac{\partial U(S_i,\nu_i)}{\partial S} = \frac{U^n_{i+1,j} - U^n_{i,j}}{(S_{i+1} - S_i)}. \tag{25}$$

3.2.1 Explicit Method

The Explicit method is the simplest method to numerically solve the Heston PDE, (9). The explicit method solves the following,

$$\begin{aligned}\frac{U^{n+1}_{i,j} - U^n_{i,j}}{dt} = \frac{1}{2}\nu S^2 \frac{\partial^2}{\partial S^2}U^n_{i,j} + rS\frac{\partial}{\partial S}U^n_{i,j} - rU^n_{i,j} + \frac{1}{2}\sigma^2\nu\frac{\partial^2}{\partial \nu^2}U^n_{i,j} \\ +\kappa(\theta-\nu)U^n_{i,j}\frac{\partial}{\partial\nu}U^n_{i,j} + \rho\sigma\nu\frac{\partial^2}{\partial s\partial\nu}U^n_{i,j}\end{aligned} \tag{26}$$

Since the Solution (U) on the boundary can easily be calculated, (25). The derivatives are calculated in the interior using (20) – (24) . This method is explicit in the sense that the derivatives are calculated at time point n instead of $n-1$. There are numerous other techniques that can also be used to numerically approximate the solution of the PDE. We also consider ADI methods.

3.2.2 Alternating Direction Implicit (ADI) Method

When using any other method that is not the explicit method, it is difficult to approximate the solution since derivatives w.r.t two variables together with a cross derivative term are required. ADI methods implements this by treating each state variable separately. In order to do that first notice that the PDE can be specified in terms of an operator

$$\frac{\partial U}{\partial \tau} = LU,$$

where

$$L = \frac{1}{2}\nu S^2\frac{\partial^2}{\partial S^2} + rS\frac{\partial}{\partial S} - r + \frac{1}{2}\sigma^2\nu\frac{\partial^2}{\partial\nu^2} + \kappa(\theta-\nu)\frac{\partial}{\partial\nu} + \rho\sigma\nu\frac{\partial^2 U}{\partial S\partial\nu}.$$

This operator is then split up into parts only containing derivatives of S, ν and it cross derivative terms. That is,

$$\begin{aligned} A_0 &= \rho\sigma\nu\frac{\partial^2 U}{\partial S\partial\nu}. \\ A_1 &= \frac{1}{2}\nu S^2\frac{\partial^2}{\partial S^2} + rS\frac{\partial}{\partial S} - \frac{1}{2}r \\ A_2 &= \frac{1}{2}\sigma^2\nu\frac{\partial^2}{\partial\nu^2} + \kappa(\theta-\nu)\frac{\partial}{\partial\nu} - \frac{1}{2}r. \end{aligned}$$

ADI schemes usually handle each one of these operators implicitly, hence the name Alternative direction implicit method. It can be solved using a variety of different ADI schemes, some of which we list below.

- Douglas Scheme,
- Craig-Sneyed scheme and
- Modified Craig-Sneyed scheme.

The algorithms for the different schemes are shown in, algorithm 2. The operators

Algorithm 2 ADI schemes

- Douglass scheme
  1. $\mathbf{Y}_0 = [\,\mathbf{I} + dt\mathbf{L}]\mathbf{U}_{t-1}$
  2. $\mathbf{Y}_k = [\,\mathbf{I} - \theta dt\mathbf{A}_k]^{\,-1}[\,\mathbf{Y}_{k-1} - \theta dt\mathbf{A}_k\mathbf{U}_{t-1}\,]$ $\quad k$=1, 2
  3. $\mathbf{U}_t = \mathbf{Y}_2$
- Craig-Sneyed scheme
  1. $\mathbf{Y}_0 = [\,\mathbf{I} + dt\mathbf{L}]\,\mathbf{U}_{t-1}$
  2. $\mathbf{Y}_k = [\,\mathbf{I} - \theta dt\mathbf{A}_k]^{\,-1}[\,\mathbf{Y}_{k-1} - \theta dt\mathbf{A}_k\mathbf{U}_{t-1}\,]$ $\quad k$=1, 2
  3. $\tilde{\mathbf{Y}}_0 = \mathbf{Y}_0 + \frac{1}{2}dt[\,\mathbf{A}_0\mathbf{Y}_2 - \mathbf{A}_0\mathbf{U}_{t-1}\,]$
  4. $\mathbf{Y}_k = [\,\mathbf{I} - \theta dt\mathbf{A}_k]^{\,-1}[\,\tilde{\mathbf{Y}}_{k-1} - \theta dt\mathbf{A}_k U_{t-1}\,]$ $\quad k$=1, 2
  5. $\mathbf{U}_t = \tilde{\mathbf{Y}}_2$
- Modified Craig-Sneyed scheme
  1. $\mathbf{Y}_0 = [\,\mathbf{I} + dt\mathbf{L}]\,U_{t-1}$
  2. $\mathbf{Y}_k = [\,\mathbf{I} - \theta dt\mathbf{A}_k]^{\,-1}[\,\mathbf{Y}_{k-1} - \theta dt\mathbf{A}_k\mathbf{U}_{t-1}\,]$ $\quad k$=1, 2
  3. $\hat{\mathbf{Y}}_0 = \mathbf{Y}_0 - \theta[\,\mathbf{A}_0\mathbf{Y}_2 - \mathbf{A}_0\mathbf{U}_{t-1}\,]$
  4. $\tilde{\mathbf{Y}}_0 = \hat{\mathbf{Y}}_0 + (\frac{1}{2} - \theta)[\,\mathbf{L}\mathbf{Y}_2 - \mathbf{L}\mathbf{U}_{t-1}\,]$
  5. $\tilde{\mathbf{Y}}_k = [\,\mathbf{I} - \theta dt\mathbf{A}_k]^{\,-1}[\,\tilde{\mathbf{Y}}_{k-1} - \theta dt\mathbf{A}_k\mathbf{U}_{t-1}\,]$ $\quad k$=1, 2
  6. $\mathbf{U}_t = \tilde{\mathbf{Y}}_2$

must be matrices and $\mathbf{U}$ must be a vector instead of a grid. Refer to Rouah (2013) to see how to build these operator matrices, and solution vectors.

3.3 Gaussian Mean Mixture with Dynamically Controlled Kernel Estimation

The GMM-DCKE builds upon the Dynamically Controlled Kernel Estimation (DCKE) proposed by Kienitz et al. (2021). DCKE uses kernel density estimation which includes local bandwidth selection and applies Gaussian process regression for inter-/extrapolation and smoothing. GMM-DCKE replaces the numerical methods using analytic expressions for conditional expectations after numerically fitting a GMM (Kienitz, 2021).

3.3.1 GMM-DCKE Algorithm

The algorithm is as follows,

- **Input**
  Training set $\mathbb{X} = \{x_1, ..., x_N\}$ with d-dimensional elements and $\mathbb{Y} = \{y_1, ..., y_N\}$, $y_n \in \mathbb{R}$, (x_n, y_n) represents the joint realisations of some random variables (X, Y), where X is the random vector of underlying risk factors for some $t < T$ and Y is a function of X. In our setting Y represents the payoff of the Bermudan derivative at time T. We let $\mathbb{X}^* = \{x_1^*, ..., x_M^*\}$ be the test set, however we assume $\mathbb{X} = \mathbb{X}^*$.

- **Output**
  Predictions $y^* = (y_1^*, ..., y_m^*)$ for $y_i^* \in \mathbb{R}$, such that $y_i^* \approx \mathbb{E}[Y|X = x_i^*]$, i.e., conditional expected values representing the value of a Bermudan derivative.

- **Calculation**
  We fit a GMM model to approximate the joint distribution of the risk factors X and payoff Y, using $\mathbb{X}$ and $\mathbb{Y}$ by applying the EM. However, we need to specify number of components K for the GMM model. Each component of the GMM model is a multivariate normal distribution given by,

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix}\right). \tag{27}$$

  For the fitted GMM(K) we have the conditional mean and variance resp. given as

$$\mu_{Y|X} = \mu_Y + \Sigma_{YX}\Sigma_{YX}^{-1}(X - \mu_X) \tag{28}$$

$$\Sigma_{Y|X} = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} \tag{29}$$

  for each single component $k = 1, ..., K$ of GMM(K). Furthermore

$$p_{Y|X}(\mathbf{y}|\theta_k) = \frac{f_k(\mathbf{x}, \mathbf{y}|\mu_\mathbf{k}, \mathbf{\Sigma_k})}{\int f_k(\mathbf{x}, \mathbf{y}|\mu_\mathbf{k}, \mathbf{\Sigma_k})dy} \sim \mathcal{N}(\mu_{k,X|Y}, \Sigma_{k,X|Y})$$

and for the conditional distribution of GMM(K),

$$p_{Y|X}(\mathbf{y}) = \sum_{k=1}^{K} \tilde{\omega_k} p_{Y|X}(\mathbf{y}|\theta_k) \tag{30}$$

$$\tilde{\omega_k} = \frac{\omega_k \phi(\mathbf{x}|\mu_{k,X}, \Sigma_{k,XX})}{\sum_l \omega_k \phi(\mathbf{x}|\mu_{k,X}, \Sigma_{k,XX})},$$

where $\mathcal{N}(\mu, \Sigma)$, $\phi(\cdot)$ denote the multivariate normal density with parameters μ, Σ and corresponding PDF respectively.

- **Control Variate**
  We can improve the accuracy of our estimator by applying the control-variate variance reduction technique. Suppose we have a random variable Z with known conditional expectation given by $\mu_{Z|X} = \mathbb{E}[Z|X = x]$. Then for any $x \in \mathbb{R}^d$ and any $\beta_x \in \mathbb{R}$, the random variable,

  $$Y^* := Y|X + \beta_X(Z|X - \mu_{Z|X})$$

  has the same conditional expectation. Choosing

  $$\beta_{X=x} := \frac{-\mathbb{C}\mathrm{ov}[Y, Z|X = x]}{\mathbb{V}\mathrm{ar}[Z|X = x]}$$

  minimizes the variance since

  $$\mathbb{V}\mathrm{ar}[Y^*] = (1 - \mathbb{C}\mathrm{orr}[Y, Z|X = x]^2)\mathbb{V}\mathrm{ar}[Y|X = x]$$

  for $x \in \mathbb{X}$. The higher the control variate Z is conditionally correlated to Y, the higher the variance reduction. The realisations of the control variate $\mathcal{Z} = \{Z_1, ..., Z_N\}$ are included into the fitting of GMM(K). Hence in this case

  $$y_i^* = \widehat{\mathbb{E}}[Y|X = x_i^*] + \widehat{\beta}_{X=x_i^*}(\widehat{\mathbb{E}}[Z|X = x_i^*] - \mu_{Z|X})$$

  with

  $$\widehat{\beta}_{X=x_i^*} = \frac{-\widehat{\mathbb{C}\mathrm{orr}}[Y, Z|X = x_i^*]}{\widehat{\mathbb{V}\mathrm{ar}}[Z|X = x_i^*]}.$$

  where $\widehat{\mathbb{E}}[\cdot]$, $\widehat{\mathbb{V}\mathrm{ar}}$ and $\widehat{\mathbb{C}\mathrm{orr}}$ are the estimates which can be calculated from the fitted data using (28) and (29). When we use the underlying as the control variate, $\beta_{X=x}$ represents the minimum variance delta.
  We have to choose the optimal number of components K, choosing a larger K over-fits the model and a smaller K under-fits the model. The number of components K for the GMM model are determined using empirical results or statistical methods, e.g. Silhouette scores or minimizing information criteria. Kienitz (2021) proposes all values of K between 3 and 6.

3.3.2 Parameter Estimation and GMM Regression

The Parameters $\theta = \{\omega_k, \mu_k, \mathbf{\Sigma}_k\}$ for the GMM are estimated using the EM algorithm. The EM algorithm is presented below in Algorithm3. Once the parameters are estimated, the conditional mean for each component $\mu_{Y|X}$ can be calculated using 28. The conditional distribution for the GMM(K) model is then calculated with (30). This is effectively a regression problem since the expected value of the conditional distribution is then needed for pricing American options. We illustrate the EM algorithm used to calculate the conditional distribution of the GMM(K) model below. Figure 2, shows the EM algorithm in action. It is clear that the estimated mean values converge to the true mean values.

Figure 2 shows the GMM regression together with non-linear regression using Laguerre polynomials as basis functions. When regressing using Laguerre polynomials, the number of basis functions needs to be set. This can easily lead to under or over fitting, where as the GMM model is less prone to overfitting. Figure 3 shows the conditional distribution of the GMM(K) model for 3D data. No Laguerre polynomials were fitted for the 3D data. It is clear from these two Figures that the GMM(K) model is adequate for use in non-linear regression, with the advantage that the GMM(K) handles discontinuities better, Figure (2), whereas the Laguerre polynomials are not accurate for data with discontinuities. This can be beneficial when pricing certain exotic options where discontinuities can arise in the payoff of the option. The EM algorithm suffers from the curse of dimensionality since it has to calculate a $d \times d$ covariance matrix. This only marginally effects the success of the GMM model as can be seen from the ease of calculating the conditional expectation in multiple dimensions.

Algorithm 3 EM algorithm for multivariate Gaussian mixtures.

1. Choose a set of initial values $\boldsymbol{\theta}$, that is $\pi_k^{old}, \mu_k^{old}, \boldsymbol{\Sigma}_k^{old}$.
2. E-Step: Calculate the $p(z|\mathbf{x_i})$

$$\gamma_{ik}^{new} = \frac{\omega_k^{old} f(\boldsymbol{x}_i|\mu_k^{old}, \boldsymbol{\Sigma}_k^{old})}{\sum_{j=1}^{K} \omega_j^{old} f(\boldsymbol{x_i}|\mu_j^{old}, \boldsymbol{\Sigma}_j^{old})}.$$

3. M-Step: Update the unknown parameters, that is

$$\omega_k^{new} = \frac{N_k}{N}.$$

   where $N_k = \sum_{i=1}^{N} \gamma_{ik}^{new}$, and

$$\mu_k^{new} = \frac{\sum_{i=1}^{n} \gamma_{ik}^{new} \boldsymbol{x}_i}{N_k},$$

   and

$$\boldsymbol{\Sigma}_k^{new} = \frac{1}{N_k} \sum_{i=1}^{n} \gamma_{ik}^{new} (\boldsymbol{x}_i - \mu_k^{new})(\boldsymbol{x}_i - \mu_k^{new})^T.$$

4. Set $\boldsymbol{\theta}^{old} = \boldsymbol{\theta}^{new}$.
5. Repeat steps (2) to (4) until convergence.

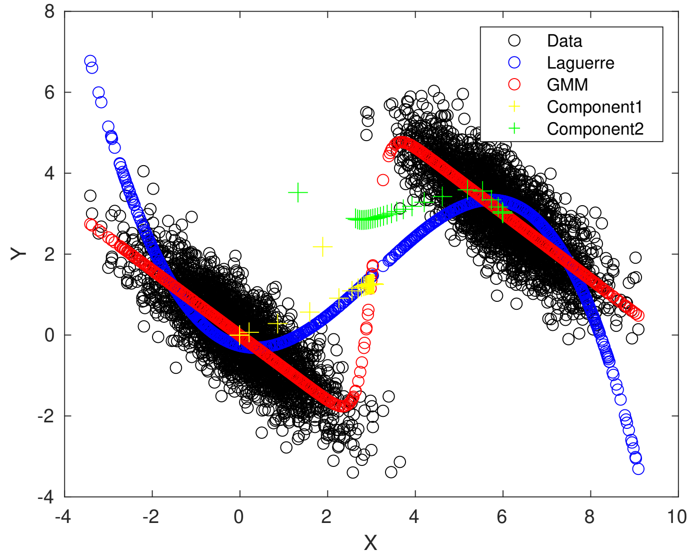


Figure 2: Plot of conditional distribution using a) GMM(K), b) Laguerre polynomials. The crosses indicate the the estimate of the mean of each component at each step in the EM algorithm. The GMM(K) model above used two components to calculate the conditional distribution, where as five polynomials were used for fitting with Laguerre polynomials as the basis function.

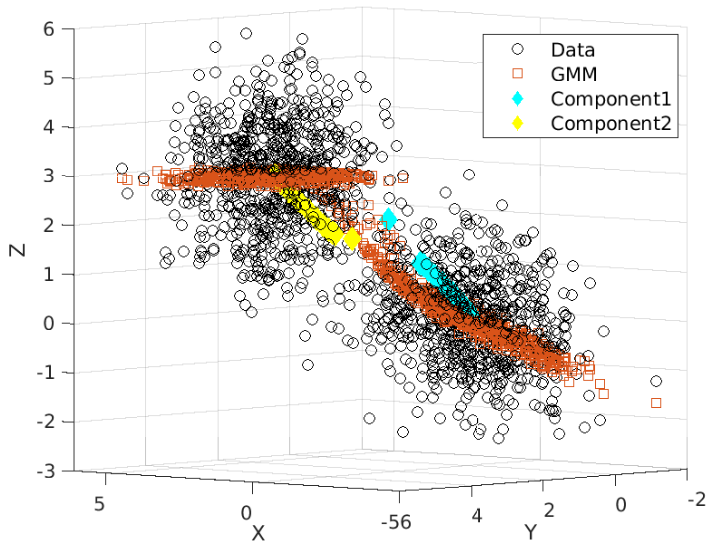


Figure 3: Plot of the conditional distribution of the GMM(K) model on 3D data. The diamonds indicate the mean of each component at each step of the EM algorithm. The GMM(K) model used two components to calculate the conditional distribution.

3.3.3 Pricing Using the GMM-DCKE Algorithm

Analogous with the Least-Squares Monte Carlo Algorithm, pricing using the GMM-DCKE algorithm follows a Dynamic Programming Formulation. However instead of regressing realised payoffs at each exercise time from continuation values on the function of state variables (prices), we fit the GMM model at each exercise time and compute the conditional expectation using (28). Essentially we are replacing step 5, 6 and 7 from the Least Squares Monte Carlo algorithm, with the following steps, using two different methods,

- **GMM-DCKE without Control Variate**
  - 5′. Set the vector X to be the stock prices for the corresponding paths (S_{i-1}), Y to be the realized continuation values for these paths and the vector Z to be corresponding variance (v_{i-1}).
  - 6′. Fit the data (Y, X, Z) to a GMM-model with (X, Z) representing our risk factors.
  - 7′. Calculate a conditional expected continuation values using (28). For the paths in X where early exercise is greater than the corresponding continuation value, set $\mathcal{V}_{i-1}$ to be the early exercise values $(H_{i-1}(S_{i-1}))$.

– **GMM-DCKE with Control Variate**

We use the underlying at each exercise time as a control variate, since we can calculate its conditional expectation using the Martingale property, that is:

$$\mathbb{E}[S_{t_i}|S_{t_{i-1}}] = \mathbb{E}[S_i|S_{i-1}] = S_{i-1}e^{r\Delta_{t_i}}.$$

Now we have the following steps:

* 5′. Set the vector X to be the stock prices for the corresponding paths (S_{i-1}), Y to be the realized continuation values for these paths and the vector Z to be control variate (S_i).
* 6′. Fit the data (Y, X, Z) to a GMM-model.
* 7′. Calculate a conditional expected continuation values using (31). For the paths in X where early exercise is greater than the corresponding continuation value, set $\mathcal{V}_{i-1}$ to be the early exercise values $(H_{i-1}(S_{i-1}))$.

The GMM-DCKE with control variate, rewards us with the minimum variance delta $\hat{\beta}_{X=x_i^*}$ at each time step, which is our conditional trading strategy with respect to the underlying. Pricing using the GMM-DCKE algorithm can be extended to multi-dimensions, whereby our risk factor is actually a vector of underlying stocks, and therefore can be used to price exotic options which are dependent on multiple underlying stocks.

4 Results and Discussion

4.1 Pricing Vanilla American Options

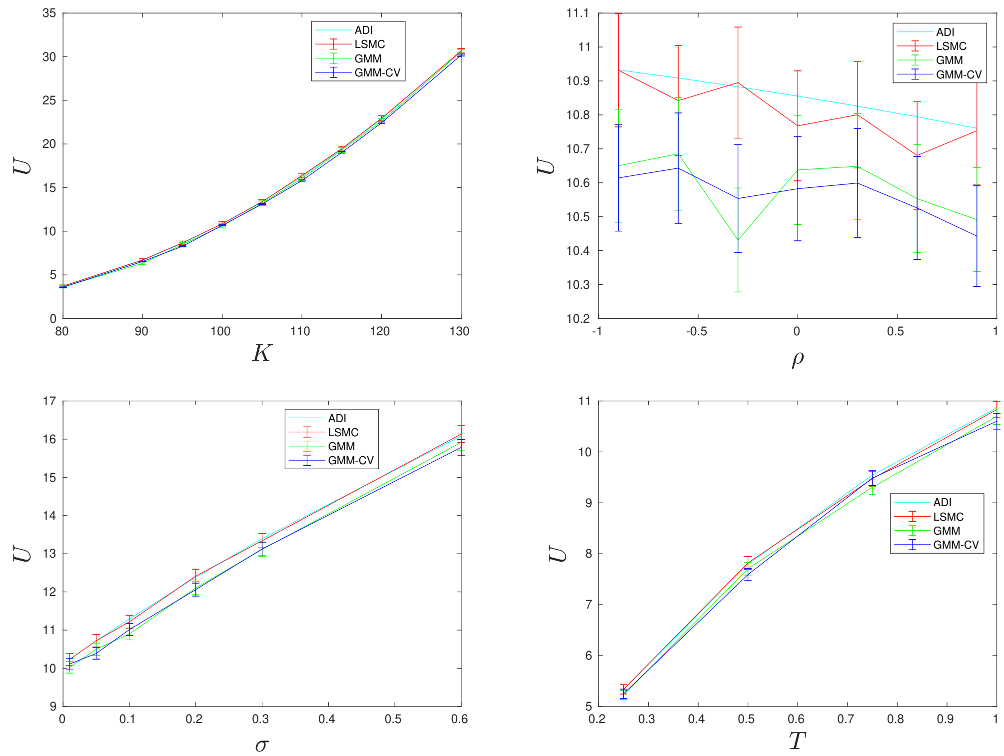


Figure 4: Plot showing the price of a vanilla American option as a function of strike price, spot volatility, volatility of variance and time-to-maturity. The standard parameters used were $\kappa = 5.0$, $\theta = 0.16$, $\sigma = 0.2$, $\rho = -0.1$, $K = 120$, $r = 0.1$, $S_0 = 100$, and $V_0 = 0.0625$. Note that these parameters were kept constant unless used as the independent variable for each graph.

In this section we present our results on pricing vanilla American options using PDE, LSMC and GMM Monte-carlo methods. The GMM Monte-carlo methods are then compared to the PDE and LSMC method. Figure 4 shows the prices of a vanilla American option using these methods. The price values are all within the Monte-carlo error bounds at a 3σ confidence level. The GMM-DCKE method uses the EM algorithm to estimate the parameters of the GMM model. These parameters are needed to calculate the expected continuation value from the cross-sectional data at each time-step, as is standard for these Monte-Carlo methods of pricing American options. Similarly the LSMC uses Laguerre polynomials to fit the cross sectional data. Under the Heston model, the expected continuation value are a function of both the underlying and the spot volatility. These methods must thus be used in three dimensions (For exotic options the number of dimensions can be considerably more, for example basket and rainbow options). The GMM-DCKE model has the advantage that mixtures of Gaussian are relatively easy to generalise to multiple dimension, whereas using Laguerre polynomials for example are not straight-forward to generalise to multiple dimensions.

Figure 5 illustrates the cross sectional data fitted using the GMM-DCKE model. Some exotic options can have a payoff with discontinuities. When that is the case the GMM-DCKE method is superior. Another benefit of using the GMM-DCKE method is that together with control variates, it produces the Greeks as well. It is considerably harder to get stable estimates of the Greeks using the LSMC model. PDE methods are also able to calculate the Greeks effiently but only if the PDE is a few dimensions. Once the option is more complex, e.g basket or rainbow options, the PDE method suffers from the curse of dimensionality.

The next section illustrates the sensitivity of price as a function of the number of Laguerre polynomials used.

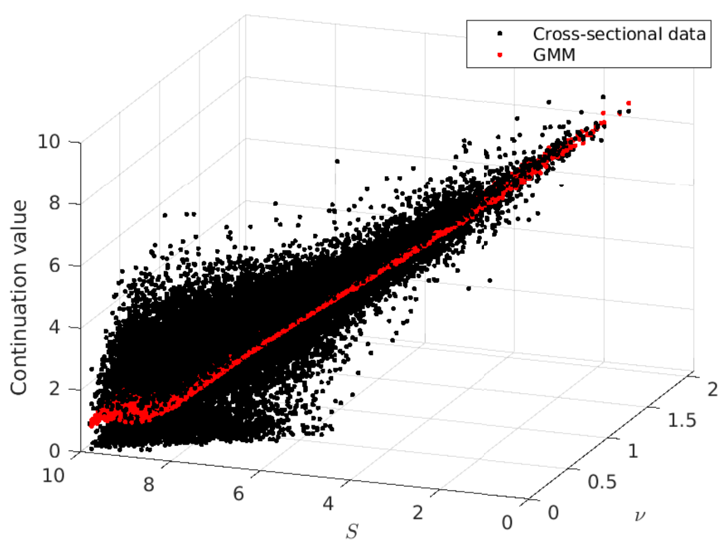


Figure 5: Plot of the continuation value as a function of spot stock price and volatility. The GMM(K) model is used to calculate the conditional expectation.

4.2 Price Sensitivity as a Function of Number of Components

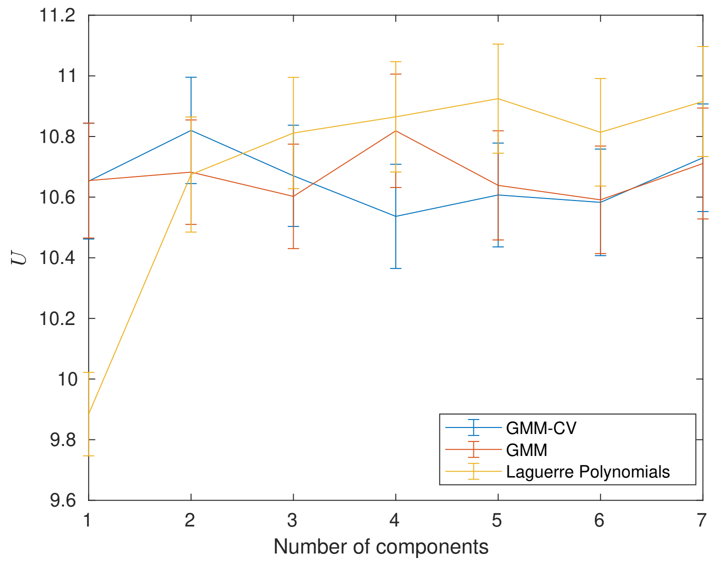


Figure 6: Comparison of GMM(K) method and least squares approach. The y-axis indicates the price of the option. The x-axis is the number of components used for calculating the conditional expectation. The standard parameters used were $\kappa = 5.0$, $\theta = 0.16$, $\sigma = 0.2$, $\rho = -0.1$, $K = 120$, $r = 0.1$, $S_0 = 100$, and $V_0 = 0.0625$. The number component for least squares approach is the number of Laguerre polynomials used for calculating the conditional expectation.

We have mentioned above that one of the shortcomings of the Least Squares Monte Carlo is that the prices calculated will be sensitive to the number of polynomials used to calculated the expected conditional continuation values. Figure 6 shows the effect on the option prices by varying the degree of the Laguerre polynomials used. All the other model parameters were kept constant. The values of the parameters that were used in computing these prices are: $S_0 = 100, K = 100, v_0 = 0.0625$, $r = 0.1, \rho = -0.1, \theta = 0.16, \kappa = 5, \sigma = 0.8$. Although, the price sensitivity does not seem too large, the sensitivity may be large for more exotic options. The GMM(K) model substantially improves on this. The sensitivity on option prices is minuscule for this vanilla American option but for exotic options that will not be the case.

4.3 Multi-dimensional GMM-DCKE

To illustrate the performance of the GMM-DCKE method in a multi-dimensional setting, we consider options dependant on multiple underlying assets. For this report we consider both American-type basket and rainbow options. Both options will be considered to be dependent on five underlying assets and therefore, we use a five dimensional Heston model to simulate five sets of time-dependent stock prices and variances for each path, as referred to in (2.3), from which the option prices can then be evaluated. While we consider only five underlying assets, both options can be easily adjusted to any number of underlying assets, within reason.

We price both the American-type basket option and American-type rainbow option with the fixed parameters described in Table 3.

Parameter	**Value**
Sample Size	50 000
Number of Components	5
Number of Underling Assets	5
Initial Stock Prices	[100 100 100 100 100]
Initial Variances	[0.0625 0.0625 0.0625 0.0625 0.0625]
Volatility's of Variance	[0.15 0.4 0.15 0.2 0.3]
Correlation between Stock and Variance	[0.1 -0.7 -0.4 0.15 -0.9]

Table 3: Option pricing parameters.

The parameters required in the five-dimensional Heston model used to generate the stock paths, will be the same as in Table 2. We show the effect of a change in strike price and maturity time by varying these parameters. Additionally, we value these options by considering them as Bermudan options with weekly exercise dates to approximate an American option.

4.3.1 Pricing American-type Basket Options

An American-type basket option is an exotic option which is a contract dependant on multiple underlying assets, and its payoff is determined by the weighted average price of these assets on or before the expiration date. This type of option entitles a holder to the right, but not the obligation, to trade at the strike price within a specified date. Therefore, there is a certain price to be paid for acquiring this right, which produces the problem of pricing such an option. A lot of literature shows that the price of a basket option is usually cheaper than that of option portfolios on the same individual underlying assets (Hanbali and Linders, 2019). Based on this advantage, basket options are popular among investors. The payoff for an American-type basket option at time $\tau \in \mathcal{T}$ (where $\mathcal{T}$ is the set of all possible exercise times) is of the form

$$\left(\eta\left(\sum_{i=1}^{d}\omega_i S_{i\tau} - K\right)\right)^+, \tag{31}$$

where d is the number of underlying assets the option is dependent on, ω_i are the weights and $\eta = 1$ for an American-type call basket option and $\eta = -1$ for an American-type put basket option.

A basket call option can be viewed similarly to that of a vanilla call option where the single underlying is a basket of assets. Therefore, as with an American call option on a non-dividend paying underlying asset where it is never optimal to exercise the option before maturity, an American-type call basket option is equivalent to a European-type call basket option under the same conditions and on non-dividend paying underlying assets. As a result of this, we only consider pricing an American-type put basket option, on five underlying assets (i.e. d = 5), which are all equally weighted (i.e. $\omega_i = \frac{1}{5}$ for all i). The asset correlation matrix (positive definite matrix) used to parameterize the Heston model is

$$C = \begin{pmatrix} 1.0 & 0.2 & 0.0 & 0.5 & 0.7 \\ 0.2 & 1.0 & 0.4 & 0.0 & 0.1 \\ 0.0 & 0.4 & 1.0 & 0.3 & 0.2 \\ 0.5 & 0.0 & 0.3 & 1.0 & 0.25 \\ 0.7 & 1 & 0.2 & 0.25 & 1.0 \end{pmatrix}. \tag{32}$$

We now show the effect of a change in both the strike price and maturity time by evaluating an American-type put basket option with the parameters shown in Table 3.

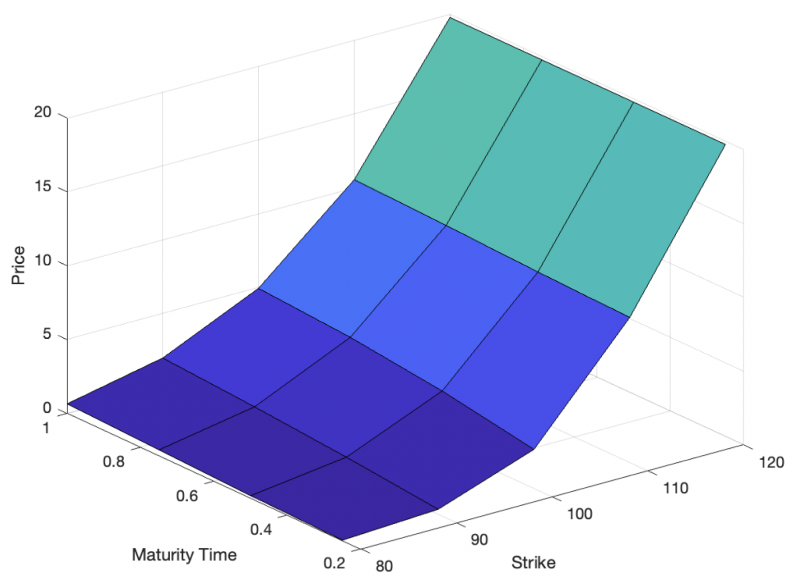


Figure 7: Surface plot of American-type put basket option price.

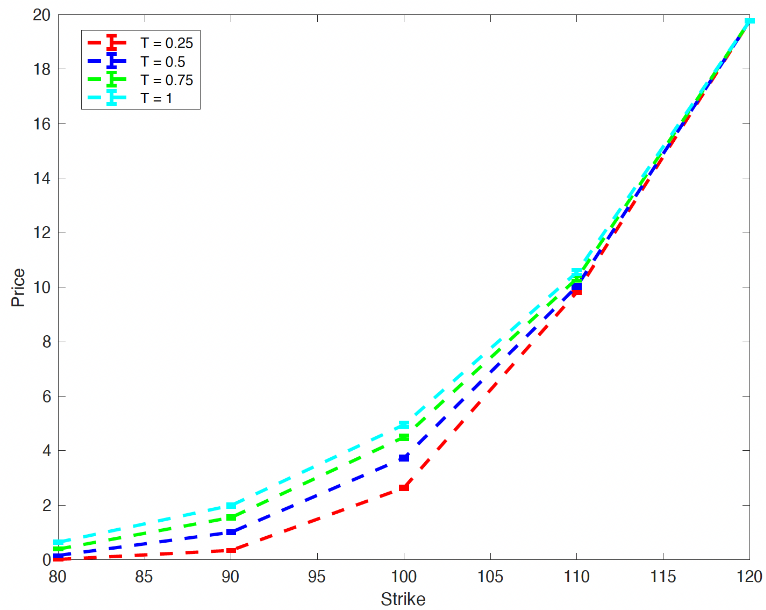


Figure 8: Plot of American-type put basket option price.

Figure 7 depicts the effect on the American-type put basket option price with changes in both the strike price and maturity time. The option has been priced at the strike prices [80 90 100 110 120] and maturity times [0.25 0.5 0.75 1.0]. Figure 8 clearly indicates how the option price increases with an increase in strike price, which is a result of the option moving further in the money with an increase in the strike price. Additionally, the option price increases with an increase in maturity time, as a longer maturity time results in greater optionality for the option holder which will result in a higher early exercise premium and hence a greater price.

4.3.2 Pricing American-type Rainbow Options

Similarly to an American-type basket option, an American-type rainbow option is an exotic option which is a contract dependant on multiple underlying assets. However, unlike an American-type basket option, rainbow options are instead structured as calls and/or puts on the best or worst performer as it relates to the underlying assets involved, where each underlying asset is referred to as a colour of the rainbow (Ouwehand and West, 2006). Therefore, the payoff of an American-type rainbow option can take on many different forms. Some of the forms that American-type rainbow options take on, include the following payoffs at time $\tau \in \mathcal{T}$ (where $\mathcal{T}$ is the set of all possible exercise times),

$$\max\left(S_{1\tau}, S_{2\tau}, ..., S_{n\tau}, K\right) \tag{33}$$
$$\left(\max\left(S_{1\tau}, S_{2\tau}, ..., S_{n\tau}\right) - K\right)^{+} \tag{34}$$
$$\left(\min\left(S_{1\tau}, S_{2\tau}, ..., S_{n\tau}\right) - K\right)^{+} \tag{35}$$
$$\left(K - \max\left(S_{1\tau}, S_{2\tau}, ..., S_{n\tau}\right)\right)^{+} \tag{36}$$
$$\left(K - \min\left(S_{1\tau}, S_{2\tau}, ..., S_{n\tau}\right)\right)^{+}, \tag{37}$$

where (33) is referred to as a best of assets or cash option, (34) is referred to as a call on max option, (35) is referred to as a call on min option, (36) is referred to as a put on max option and (37) is referred to as a put on min option (Ouwehand and West, 2006). In this report we only consider pricing a put on min American-type rainbow option (37), however the methodology used to price such an option can be easily adjusted to account for any of the American-type rainbow options mentioned in payoffs (33) – (37). Similarly to Section 4.3.1, we now show the effect of a change in both the strike price and maturity time by evaluating a put on min American-type rainbow option with the parameters shown in Table 3 and asset correlation matrix (32).

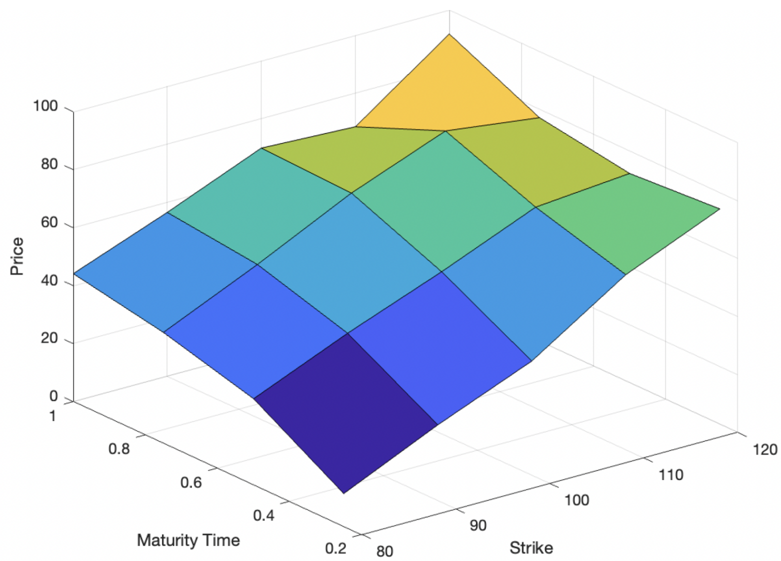


Figure 9: Surface plot of put on min American-type rainbow option.

Figure 9 depicts the effect on the put on min American-type rainbow option from a change in both the strike price and maturity time. Similarly to Section 4.3.1, the option has been priced at the strike prices [80 90 100 110 120] and maturity times [0.25 0.5 0.75 1.0]. Figure 9 clearly indicates how the option price increases with an increase in strike price, which is a result of the option moving further in the money with an increase in the strike price.

5 Conclusion

We have demonstrated that GMM-DCKE is purely data driven and model-free, in the sense that we do not fix a particular model or class of models, and no other information such as the underlying stochastic differential equation is required. Hence, it is easy to incorporate transaction costs and other market frictions into the GMM-DCKE model. We illustrated the use of the GMM-DCKE approach to pricing American options in a multi-dimensional setting by pricing an American-type basket option with five underlying stocks, this problem can be extended to price baskets with more than five underlying stocks. It is impossible to price such options using the PDE scheme due to the curse of dimensionality, and using the LSMC will be inefficient. The GMM-DCKE model overcomes the curse of dimensionality by fitting the underlying stocks and payoff process, and analytically calculating the price. We also priced a rainbow option which has a discontinuous payoff. Because of the high dimensionality of the problem, it is difficult to price such derivatives using PDE or LSMC approaches. When we incorporate the control variate to the GMM-DCKE approach, the variance of our price estimates is reduced and we obtain the Greeks for free. This is important for both trading and risk management.

Bibliography

Andersen, L.B., 2007. Efficient simulation of the Heston stochastic volatility model. Available at SSRN 946405 , 3–31.

Bernhardt, D., Eckblad, M., 2013. Stock market crash of 1987. Federal Reserve Bank of Chicago .

Black, Scholes, Myron, F., 1973. The pricing of options and corporate liabilities. Journal of political Economy 81, 637–654.

Carr, P., Madan, D., 1999. Option valuation using the fast fourier transform. Journal of computational finance 2, 61–73.

Carriere, J.F., 1996. Valuation of the early-exercise price for options using simulations and nonparametric regression. Insurance: mathematics and Economics 19, 19–30.

Duffy, D.J., 2013. Finite Difference methods in financial engineering: a Partial Differential Equation approach. John Wiley & Sons.

Hanbali, H., Linders, D., 2019. American-type basket option pricing: a simple two-dimensional partial differential equation. Quantitative Finance 19, 1689–1704.

In't Hout, K., Foulon, S., 2010. ADI finite difference schemes for option pricing in the heston model with correlation. International Journal of Numerical Analysis & Modeling 7, 306–307.

Jones, C.S., 2003. The dynamics of stochastic volatility: evidence from underlying and options markets. Journal of Econometrics 116, 181–224.

Kienitz, J., 2021. GMM DCKE-semi-analytic conditional expectations. Available at SSRN 3902490 , 1–15.

Kienitz, J., Nowaczyk, N., Geng, N.Q., 2021. Dynamically controlled kernel estimation. Available at SSRN 3829701 , 1–17.

Kienitz, J., Wetterau, D., 2013. Financial modelling: Theory, implementation and practice with MATLAB source. John Wiley & Sons.

Longstaff, F.A., Schwartz, E.S., 2001. Valuing American options by simulation: a simple least-squares approach. The Review of Financial Studies 14, 113–147.

Najar, F., Bourouis, S., Bouguila, N., Belghith, S., 2017. A comparison between different Gaussian-based mixture models, in: 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), IEEE. pp. 704–708.

Ouwehand, P., West, G., 2006. Pricing rainbow options. Wilmott magazine 5, 74–80.

Pagès, G., Sagna, A., 2015. Recursive marginal quantization of the Euler scheme of a diffusion process. Applied Mathematical Finance 22, 463–498.

Palupi, I., Sitorus, I.U., Umbara, R.F., 2015. Using least-square Monte Carlo simulation to price American multi underlying stock options, in: 2015 3rd International Conference on Information and Communication Technology (ICoICT), IEEE. pp. 504–509.

Reynolds, D.A., 2009. Gaussian mixture models. Encyclopedia of Biometrics 741.

Rouah, F.D., 2013. The Heston model and its extensions in Matlab and C#. John Wiley & Sons.