8-24-2023 3:00 PM

Modelling long-term security returns

XINGHAN ZHU, *Western University*

Supervisor: Sendova, Kristina, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Statistics and Actuarial Sciences

Abstract

This research focuses on the concerns of Canadian investors regarding portfolio diversification and preparedness for unexpected risks in retirement planning. It models market crashes and two main financial instruments as independent components to simulate clients' portfolios. Initially exploring single distributions on mutual funds such as Laplace and t distributions, the research finds limited success. Instead, a normal-Weibull spliced distribution is introduced to model log returns.

The Geometric Brownian Motion (GBM) model is employed to predict and evaluate returns on common stocks using the Maximum Likelihood Estimator (MLE), assuming that daily log returns follow a normal distribution. Additionally, the Merton Jump Diffusion (MJD) model is considered to account for jumps in stock trajectories with an independent Poisson process term based on the GBM model. Market crashes, defined as a decline of at least 10% in the S&P500 over a maximum of 252 trading days, are modelled using a homogeneous Poisson process. The combined simulation results show that the model is effective in most portfolio predictions.

Keywords— Spliced Distribution, Weibull Distribution, Geometric Brownian Motion, Merton Jump Diffusion, homogeneous Poisson process, market crashes

Summary for Lay Audience

This research is about helping everyday Canadians with their investments and retirement savings. Many people in Canada invest their money in different ways to make sure they have enough for retirement. To make things easier for investors, we have come up with a way to simulate how different types of investments, like stocks and mutual funds, might perform, especially during tough times like market crashes. We used some math and statistics to create a new way to understand these investments and their risks.

We looked at different patterns and found that some of the usual ways of predicting how investments will do did not work so well. So, we came up with a new way to understand these patterns and make better predictions.

We also studied sudden jumps in security prices. To understand this better, we used a model that includes a "jump diffusion" to see how these big jumps happen and how they affect investments.

Lastly, we know that big market crashes can happen unexpectedly and really shake up the investment world. We studied what defines a market crash and how often it happens. We used this information to create a way to understand and prepare for these market crashes when planning our investments.

Acknowledgment

I am grateful to my supervisors, Dr. Kristina Sendova and Dr. Yang Miao. Their expertise, attentive reading, and generous feedback were essential to this dissertation.

Thank you to the members of my examination committee, Dr. Marcos Escobar-Anel, Dr. Hao Yu, Dr. Mehmet A. Begen, for their thoughtful questions and critical engagement with my work.

Table of Content

1 Introduction

This research focuses on the concerns of Canadian investors regarding portfolio diversification and preparedness for unexpected changes in the prices of financial vehicles. Most Canadians invest in some financial instruments to prepare themselves for retirement. Investors from different backgrounds also exhibit diverse investment and risk preferences.

To address these concerns, we introduce a modelling approach to simulate clients' portfolios, focusing on market crashes and financial instruments (stocks and mutual funds) as independent components.

We explore the fit of certain distributions (Laplace and t distribution, in particular) and find that they yield unsatisfactory results. Consequently, we introduce a Normal-Weibull spliced distribution to model log returns of mutual funds. We employ the model on low-to-medium risk as it has the smallest sample size and then extend the analysis to other risk ratings. Splicing is a method for creating new distributions. The advantage of a spliced distribution is that it combines different distributions in non-overlapping intervals to model more closely the behaviour of the underlying random variable.

Geometric Brownian Motion (GBM) is often employed in the literature to predict and evaluate the returns on common stocks. We calculate the maximum likelihood estimator (m.l.e.) of the fitted GBM assuming that daily log returns follow a normal distribution. After that, we consider the Merton Jump Diffusion (MJD) model to model stocks with jumps in their trajectories. More precisely, if we add to the stochastic differential equation (s.d.e.) of the GBM an independent Poisson process that models these jumps, we then obtain the MJD model. The highest level of the stock market circuit breaker is set as a threshold to identify jumps and analyze jump size distribution.

In addition, market crashes are unexpected but may significantly impact our model analysis. We thus adopt the definition of a market crash in Lleo and Ziemba (2017) as a decline of at least 10% in the level of the S&P500 over a time period of at most a year (252 trading days) and then test this definition using historical data on financial crises. We introduce a

Poisson process to model market crashes and combine it with portfolio simulation in stocks and mutual funds.

This thesis is structured as follows: We begin by data description in Section 2. We then have the modelling methodology and results for each component in Section 3, and concluding remarks in Section 4.

2 Data Description

The dataset provided by the Financial Wellness Lab consists of data collected from 31,664 clients with 80,139 accounts who invest in different financial securities. The clients' demographic information contains details on age, gender, residency, annual income and risk tolerance. It is combined with security information to analyze investment products and risk preferences.

Figure 1 shows the types of security that our clients are using:

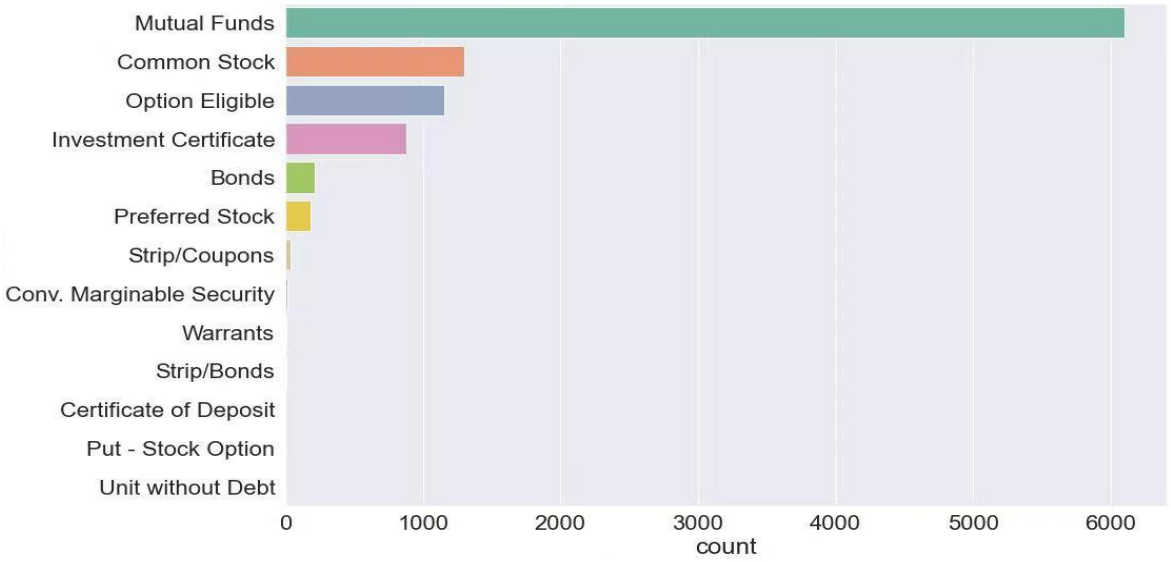


Figure 1: Count plot of clients' investments

Considering investment preferences and missing information about underlying assets of Option Eligible and components of Investment Certificate, we mainly focus on building models for common stocks and mutual funds.

Figure 2 shows different risk preferences of clients of both genders. The pie chart demonstrates that female clients are more conservative in their risk preferences, investing more in low-risk or low-to-medium-risk products (68%), while male clients would choose to invest more in medium-risk or higher-risk products (53%).

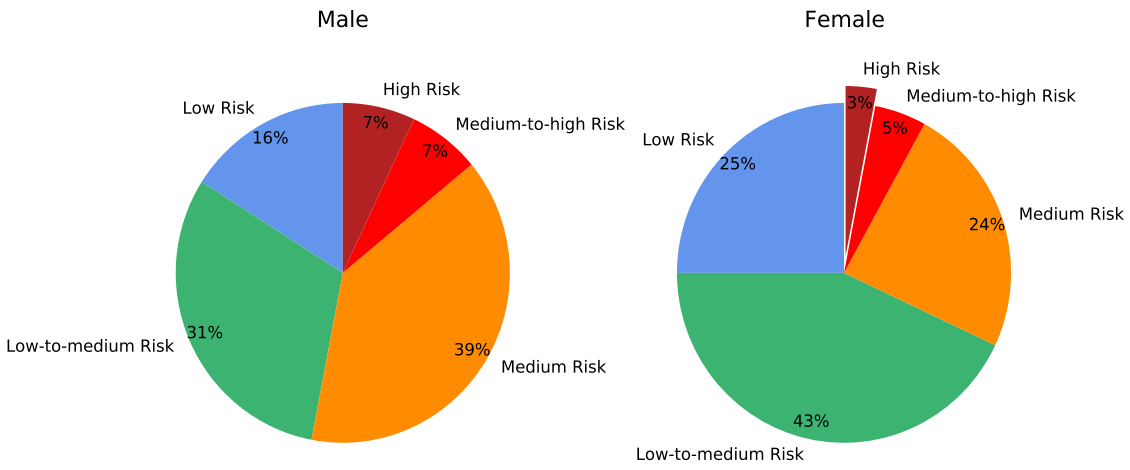


Figure 2: Proportion of investments in different risks for clients of both genders

Figure 3 exhibits the proportions of the population that state various levels of knowledge about investing. The vast majority of clients have some knowledge about investing, while only a small percentage of them have excellent knowledge.

Figure 4 shows the gender distribution of clients. We have close numbers of clients by gender.

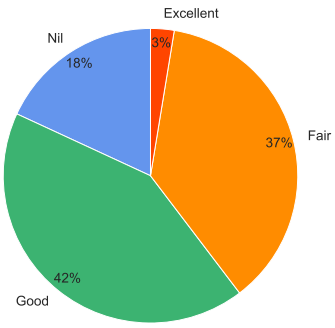


Figure 3: Clients from different knowledge background

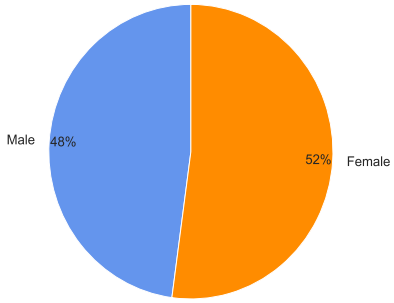


Figure 4: Clients with different gender

Figure 5 illustrates the difference in investment preferences for clients with different investment knowledge. There is not much difference between the first three categories of clients (nill, fair and good), while clients with excellent knowledge of investing will prefer to invest more in common stocks and options.

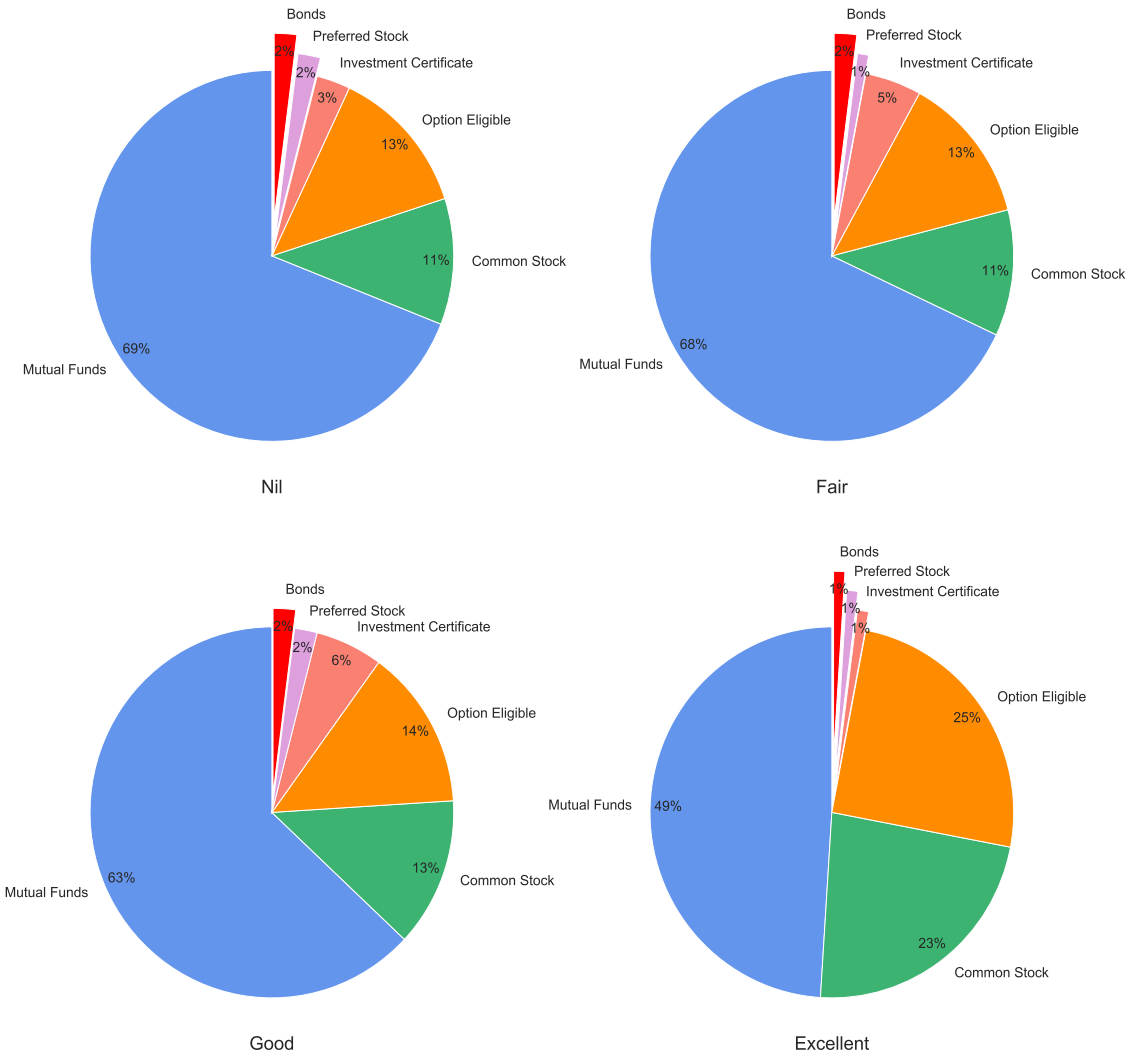


Figure 5: Proportion of investments in different securities for clients with different knowledge of investing

We used the information about security tickers and downloaded historical prices to model mutual funds. Stocks were modelled based on 3-year close prices provided by the Financial Wellness Lab and extended to the longer term through detailed tickers. Due to the limitation on stock information, we only had data for 21.3% of common stocks. In total, 2,171 stocks and 25,500 mutual funds with tickers are available to us for modelling. Furthermore, mutual funds are rated with four risk levels by the Financial Wellness Lab – low, low-to-medium, medium and high. The distribution of risk ratings is shown in the following table:

Low risk	Low-to-medium Risk	Medium risk	High risk
937	195	21406	2962

Table 1: Mutual funds with different risk rating

In addition, the average number of years in existence for mutual funds is 11.6 years, and the tractories of mutual funds with the same risk rating show a similar pattern. We randomly select 20 securities from each risk rating in the following figures to avoid overlapping:

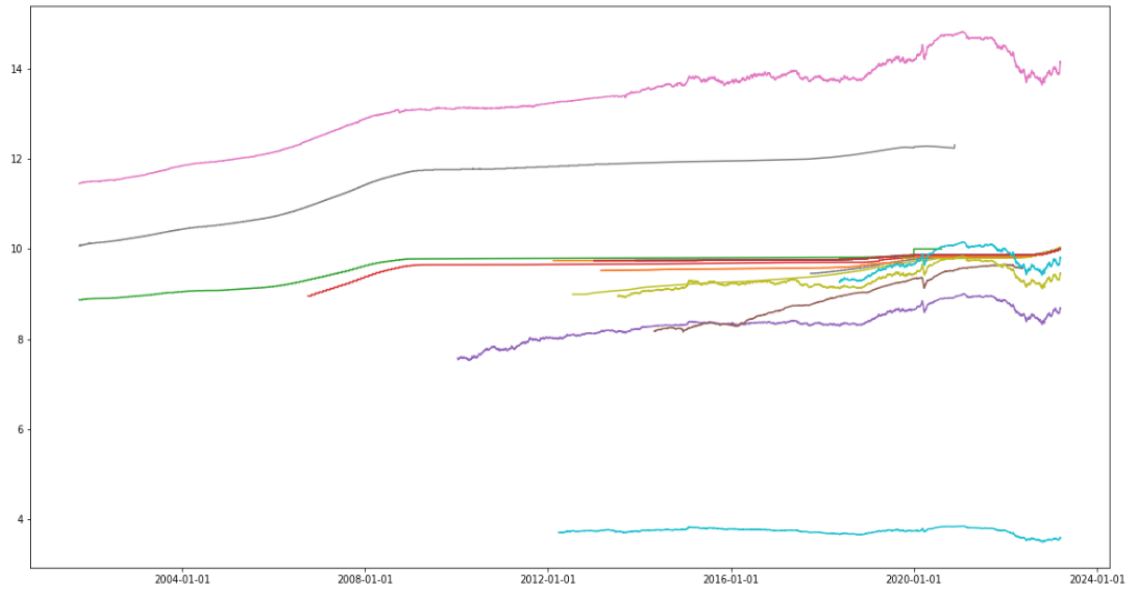


Figure 6: Low risk

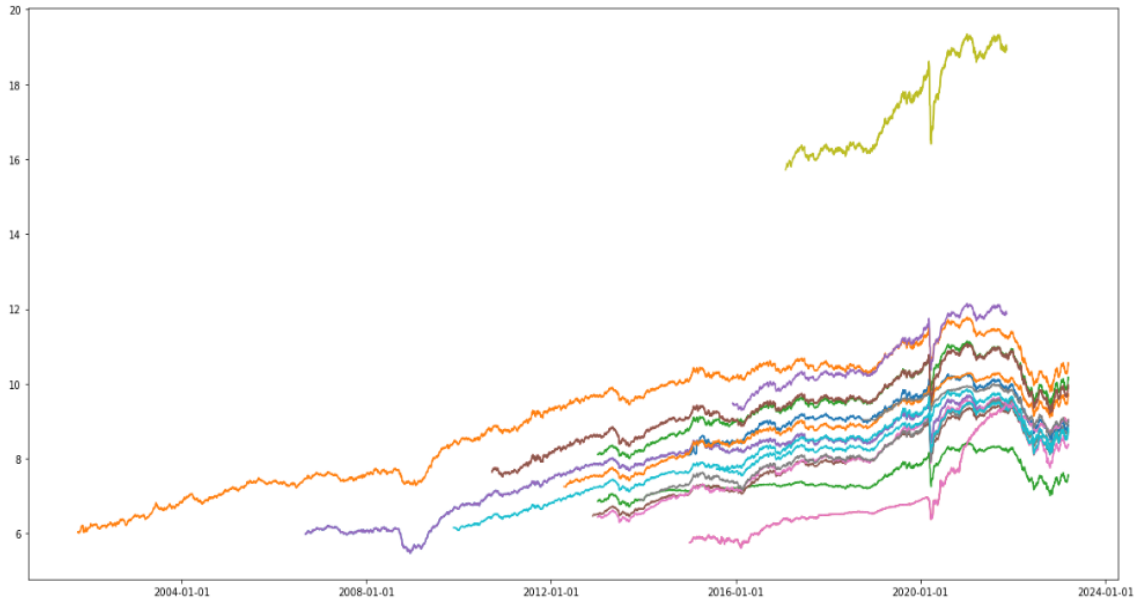


Figure 7: Low-to-medium risk

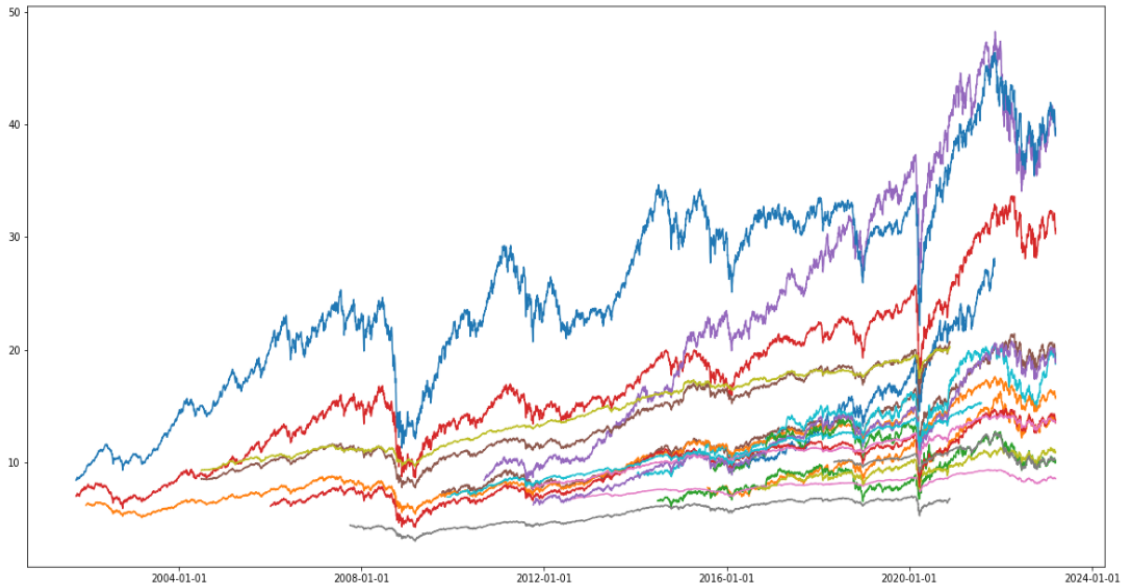


Figure 8: Medium risk

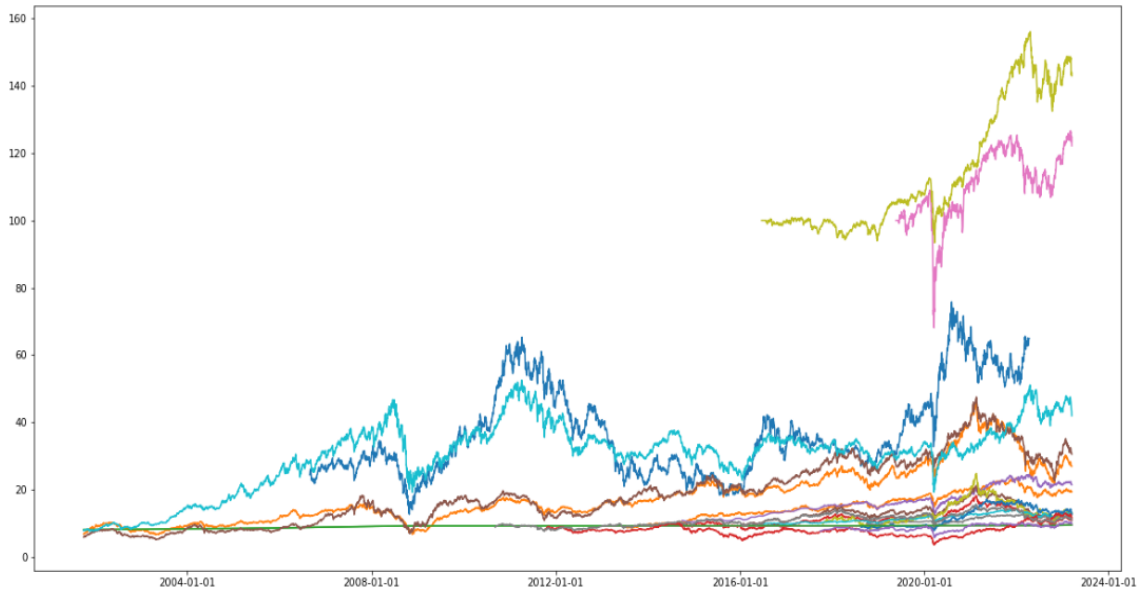


Figure 9: High risk

These observations explain the way mutual funds are grouped.

3 Methodology and Results

3.1 Modelling Mutual Funds

A mutual fund is a type of investment fund. An investment fund is a collection of investments, such as stocks, bonds or other financial instruments. Unlike most other types of investment funds, mutual funds are open-ended, which means as more people invest, the fund issues new units or shares.

A mutual fund typically focuses on specific types of investments. For example, a fund may invest mainly in government bonds, stocks from large companies or stocks from certain countries. Some funds may invest in a mix of stocks and bonds, or other mutual funds.

3.1.1 Risk Rating

As we explained in the dataset description, mutual funds with low-to-medium risk have the smallest sample size, which implies that it is easier to obtain the data and analyze it. We built a model based on this risk level and fitted the model to other risk levels while calibrating the parameters accordingly.

3.1.2 Log Return

The daily return

$$\frac{S_t - S_{t-1}}{S_{t-1}}$$

is easy to calculate and is used to simulate prices. However, it has a limitation in its range: the maximum loss for a security is capped by its total market value, which means it has a lower bound of -1. This problem is solved by calculating the logarithm of the daily returns, i.e.,

$$\log \frac{S_t}{S_{t-1}}$$

and modelled mutual funds based on this variable.

In our dataset, time intervals of two successive market close prices are mostly 1-day and 3-days. We verified that all 3-day time intervals are weekends. Besides, Kolmogorov-Smirnov tests on 1-day and 3-day log-returns of mutual funds showed that 81% of mutual funds have these two sets of samples drawn from the same probability distribution. Therefore, we treated all log returns as one sample for each mutual fund in modelling.

We also only considered the time periods without market crashes and will discuss modelling market crashes in detail in Section 3.3.

Log returns of each mutual fund showed heavy tails in Q-Q plots with normal distribution as the theoretical distribution(see Figure 10), which led us to introduce the following Sections 3.1.3 and 3.1.4.

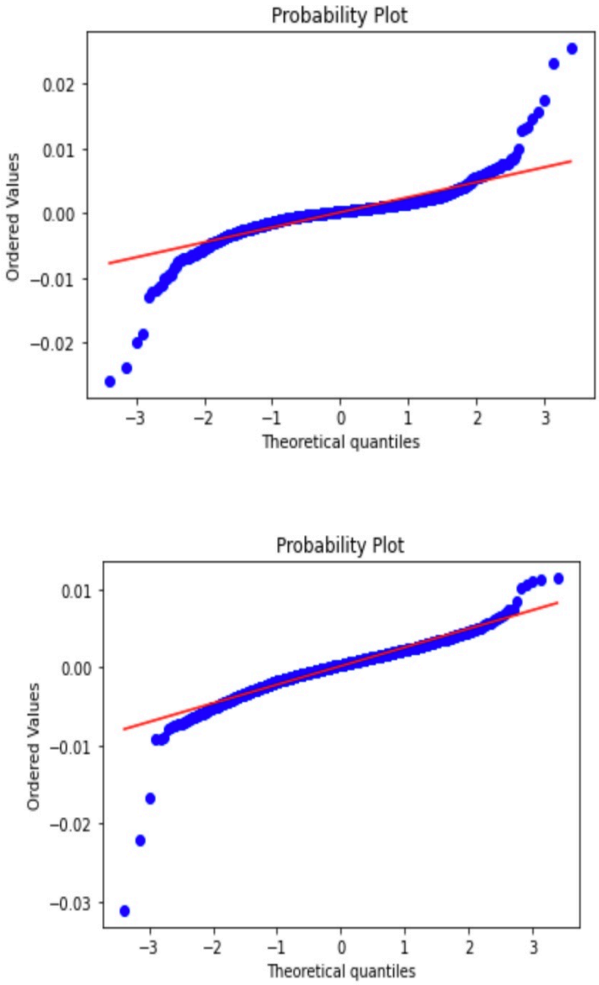


Figure 10: QQ plots of log return

3.1.3 Single Distribution

In observation of all density plots and Q-Q plots, almost all mutual funds have approximately symmetrical density and have heavier tail than normal distribution. We started with simple models to test if a particular distribution fits the data well. In this case, we considered the Laplace distribution and the Student's t-distribution. We also tested normal distribution as benchmark to see whether other distributions provide better outcomes. The results of one-sample Kolmogorov-Smirnov tests on 194 mutual funds with low-to-medium risk (test whether a sample comes from a specific distribution) are shown in Table 2:

Test result	Normal	Laplace	Student's t
Reject	172	128	167
Fail to reject	22	66	27

Table 2: Kolmogorov-Smirnov tests of three distributions on 194 mutual funds

We tested each mutual fund with reference probability distributions and made the decision based on a significant level of 0.1. The table shows that all three distribution does not fit the data well.

Laplace Distribution: The Laplace distribution, one of the earliest known probability distributions, is a continuous probability distribution named after the French mathematician Pierre-Simon Laplace. Like the normal distribution, this distribution is unimodal (one peak) and it is also a symmetrical distribution. However, it has a sharper peak than the normal distribution. The Laplace distribution is the distribution of the difference of two independent random variables with identical exponential distributions. It is often used to model phenomena with heavy tails or when data has a higher peak than the normal distribution.

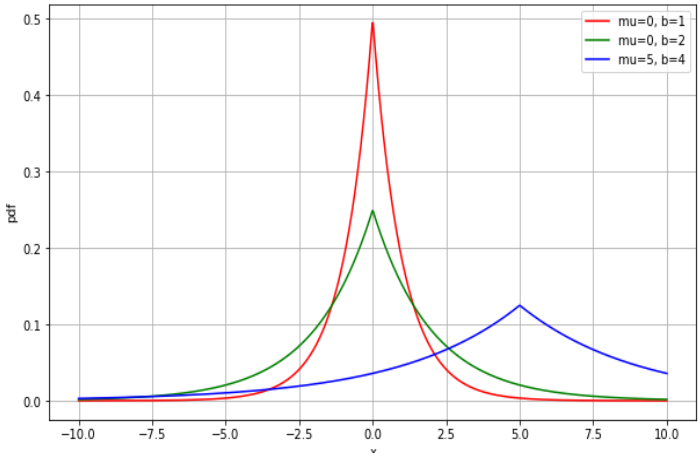


Figure 11: Probability density function of Laplace distribution

Student's t Distribution: Student's t distribution (or simply the t-distribution) t_ν is a continuous probability distribution that generalizes the standard normal distribution. Like the latter, it is symmetric around zero and bell-shaped. However, t_ν has heavier tails and the amount of probability mass in the tails is controlled by the parameter ν. For $\nu = 1$ the Student's t distribution t_ν becomes the standard Cauchy distribution, whereas for $\nu \to \infty$, it becomes the standard normal distribution $\mathcal{N}(0, 1)$.

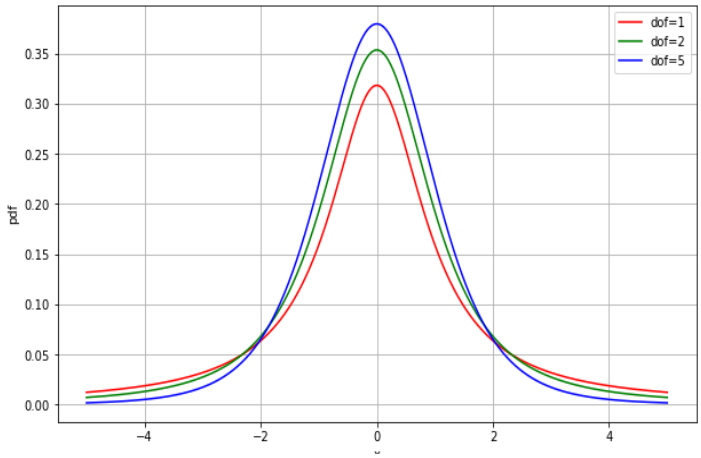


Figure 12: Probability density function of Student's t distribution

3.1.4 Normal-Weibull Spliced Distribution

The density function of an n-component spliced distribution is defined as follows (Klugman et al., 2012):

$$f(x) = \begin{cases} a_1 f_1(x) & \text{if } x \in C_1 \\ a_2 f_2(x) & \text{if } x \in C_2 \\ \vdots & \\ a_n f_n(x) & \text{if } x \in C_n \end{cases} \tag{1}$$

Here $a_1, a_2, \cdots, a_n$ are positive weights that add up to one:

$$\sum_{i=1}^{n} a_i = 1.$$

For $i = 1, 2, \cdots, n$, $f_i(x)$ is a proper density function with all probability on the interval C_i:

$$\int_{C_i} f_i(x)\, dx = 1.$$

The intervals $C_1, C_2, \cdots, C_n$ are mutually exclusive:

$$C_i \cap C_j = \emptyset, \forall i \neq j.$$

The intervals are also sequentially ordered: $x < y$ if $x \in C_i$ and $y \in C_j$ for all $i < j$. For example, the intervals can be formed by $C_1 = [c_0, c_1], C_1 = (c_1, c_2], \cdots, C_n = (c_{n-1}, \infty)$, where $c_0, c_1, \cdots, c_{n-1}$ are break points or thresholds. An advantage of spliced distributions is that they allow us to model different parts of a response variable with different distributions (Gan and Valdez, 2018). In our case, we considered a three-part spliced distribution as our model with two tail parts and one middle part.

Threshold Selection: To find the proper distribution of each component of the spliced distribution, we used the Median Absolute Deviation (MAD) method to identify the middle part and the tail parts. Absolute deviation from the median was (re-)discovered and popularized by Hampel (1974), who attributes the idea to Carl Friedrich Gauss (1777-1855). It is a robust way to identify outliers and replaces standard deviation or variance with the median

deviation and the mean with the median. Like the mean, the median is a measure of central tendency but offers the advantage of being very insensitive to the presence of outliers (Leys et al., 2013). The result is a method that is not as affected by outliers as using the mean and standard deviation would be.

To use the median absolute deviation method, we followed these steps:

- Calculate the median of the sample ($\tilde{X}$).
- Calculate the deviation from the median for each sample, using the absolute value of the deviations, $|X_i - \tilde{X}|$ = deviation.
- Find the median of the absolute deviations, median(|deviations|),
  MAD = median($|X_i - \tilde{X}|$).
- Values within range $(\tilde{X} - 3 \cdot MAD, \tilde{X} + 3 \cdot MAD)$ are identified as "middle part."
- The rest of the data are identified as "tail part."

The rejection criterion of a value is the unavoidable subjective aspect of the decision. Depending on the stringency of the researcher's criteria, which should be defined and justified by the researcher, Miller (1991) proposes the values of 3 (very conservative), 2.5 (moderately conservative) or even 2 (poorly conservative). We tested numbers with an interval of 0.1 between 2 and 3, and 3 yielded the best result.

We applied same one-sample Kolmogorov-Smirnov tests on middle parts of 194 mutual funds with low-to-medium risk and the outcome is shown in Table 3:

Test result	Normal	Laplace	Student's t
Reject	67	191	143
Fail to reject	127	3	51

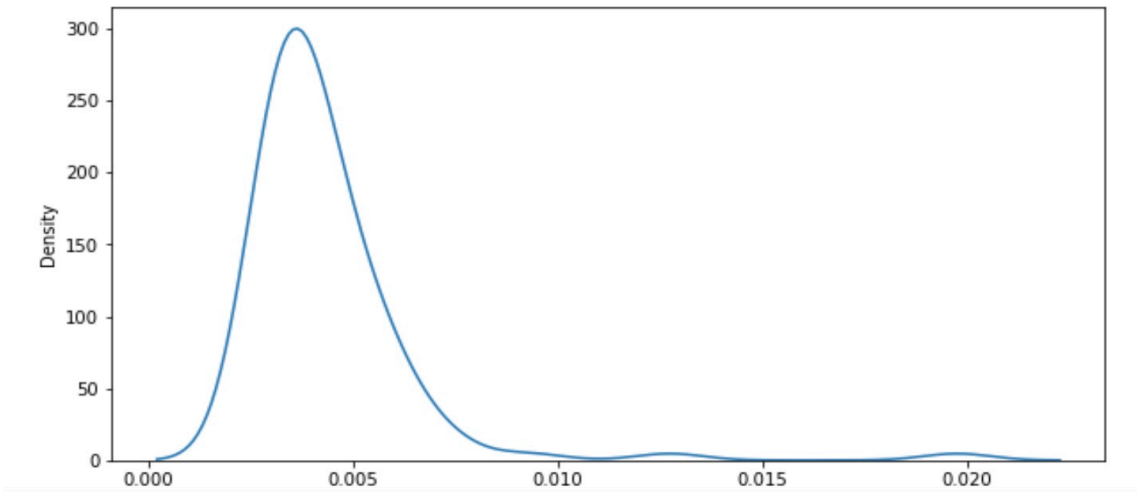
Table 3: Kolmogorov-Smirnov tests of three distributions on middle parts of 194 mutual funds

From Table 3, we noticed that the normality test of middle parts over 194 mutual funds returned a better result.

Tail Distribution: To find a distribution that fits well the tails, we used a more involved approach: we shifted the sample distribution in a way to make it symmetric with respect to the y-axis and then took the absolute value of all sample values in the tails. This method is based on the observational symmetry of the sample density, which is also a main assumption of our model. Furthermore, there are two main reasons for us to choose this method:

- It is hard to find a single distribution that could generate both negative and positive extreme values.
- If we only focus on the right tail or on the left tail, the sample might not be large enough to find the proper distribution and estimate its parameters.

The following figures are examples of density plots of tail parts:



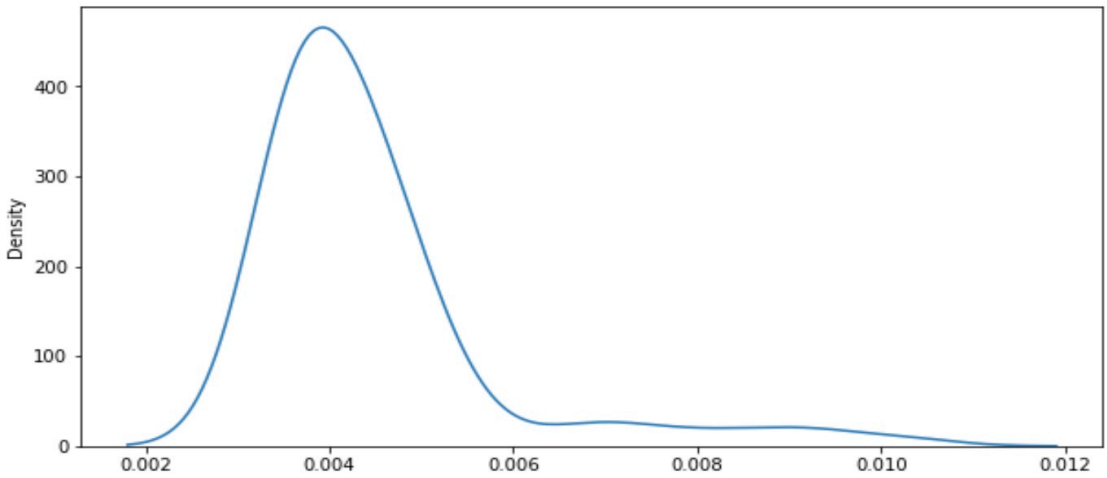


Figure 13: Examples of densities with a heavy right tail

We considered Gamma distribution and Weibull distribution on the Kolmogorov-Smirnov test, and the result showed that 77.9% of the tail data passed the test on Weibull distribution with a significant level of 0.1 which meant they fit the Weibull distribution well.

The Weibull distribution is a continuous probability distribution used to analyze life data, model failure times and access product reliability. It can also fit a large range of data from many other fields like economics, hydrology, biology, and engineering. It is a family of extreme value distributions which is frequently used to model reliability, survival, wind speeds and other data.

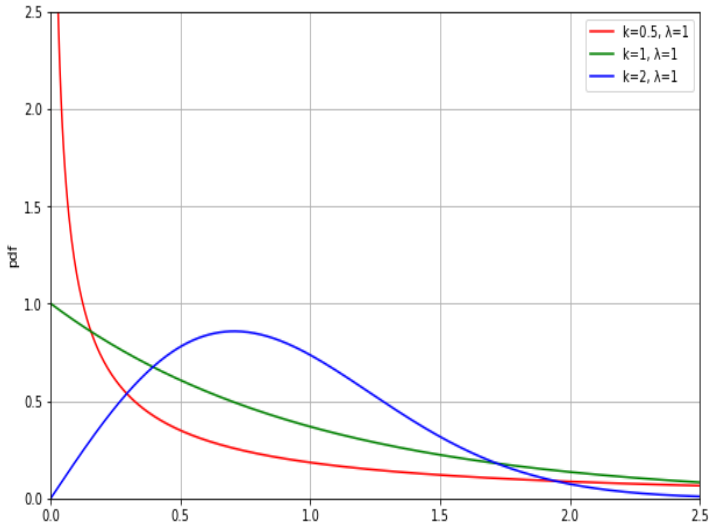


Figure 14: Probability density function of Weibull distribution

The result of Kolmogorov-Smirnov tests on mutual funds (pass rate) with other risk levels is shown in Table 4:

Test Result	Low Risk	Medium Risk	High Risk
Normal(Middle)	78.6%	80.5%	89.7%
Weibull(Tail)	55.5%	67.8%	71.7%

Table 4: Kolmogorov-Smirnov tests of mutual funds with different risk levels

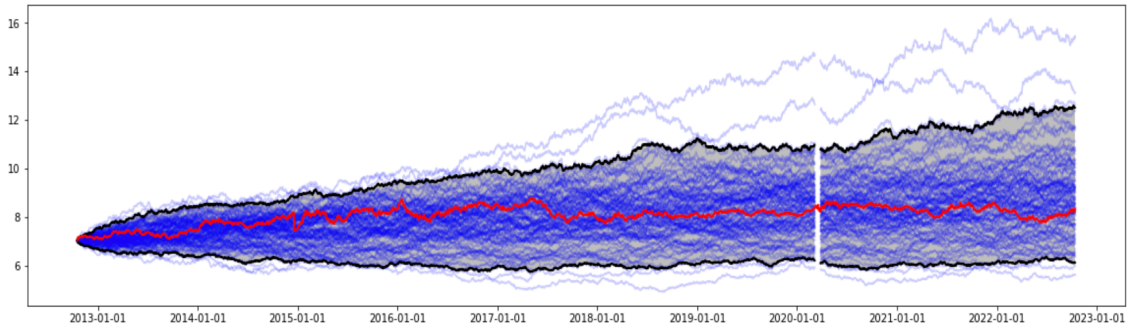
We notice that the higher the risk, the higher the proportion of mutual funds with middle parts normally distributed. A possible reason is that mutual funds with higher risk invest more stocks, and the log returns of stock usually follow a normal distribution without considering extreme values (McDonald and Lynch, 2006).

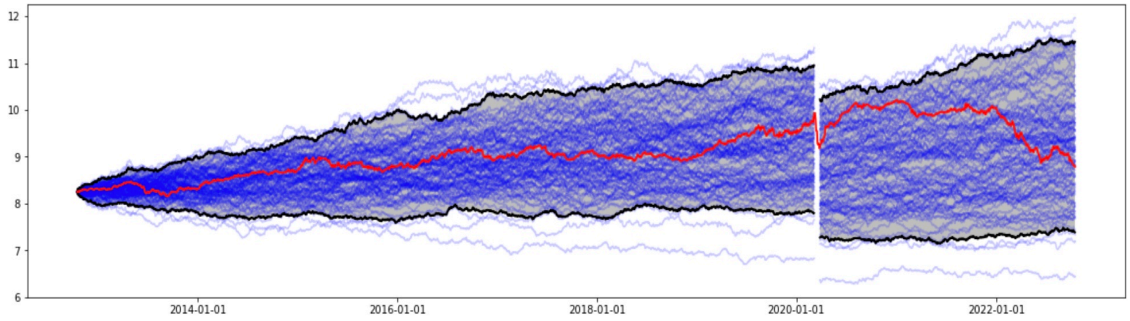
Simulation: To simulate mutual fund performance, we need to generate random variables from our model. Since there is no common method in Python to generate random values from a spliced distribution directly, we introduced an approximation method using step function with the following steps:

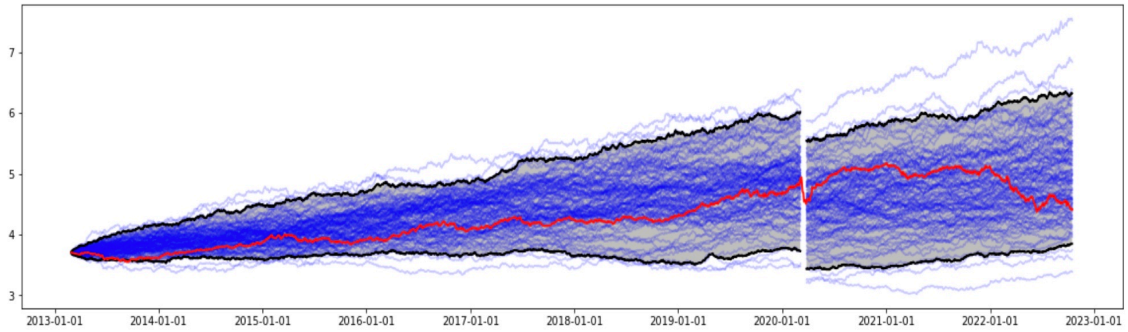
- Choose a very small value h ($h = 0.000001$ in our simulation).
- Calculate approximate values of the cumulative density function (c.d.f.) by setting $F(h) = F(0) + h \cdot f(0)$, where $f(\cdot)$ is the probability density function (p.d.f.).
- Repeat the step by $F(nh) = F((n-1)h) + h \cdot f((n-1)h)$.
- Stop once $F(\cdot) = 1$, then build a corresponding 'lookup table.'
- Generate a uniform random variable between 0.5 and 1.
- Use the 'lookup table' to find the value X of the inverse c.d.f.
- "Flip a coin", if the result is positive, $X_i = X+$sample mean; if the result is negative, $X_i = -X+$ sample mean.

Some details need to be explained in our method: we used again the assumption of symmetry. We considered a shifted sample, which is symmetric with respect to the y-axis and only the positive part. In this case, the cumulative density function (c.d.f.) started from 0.5 ($F(0) = 0.5$), which is why we generated uniform random variables between 0.5 and 1.

Figures 15 show some examples of simulation results:







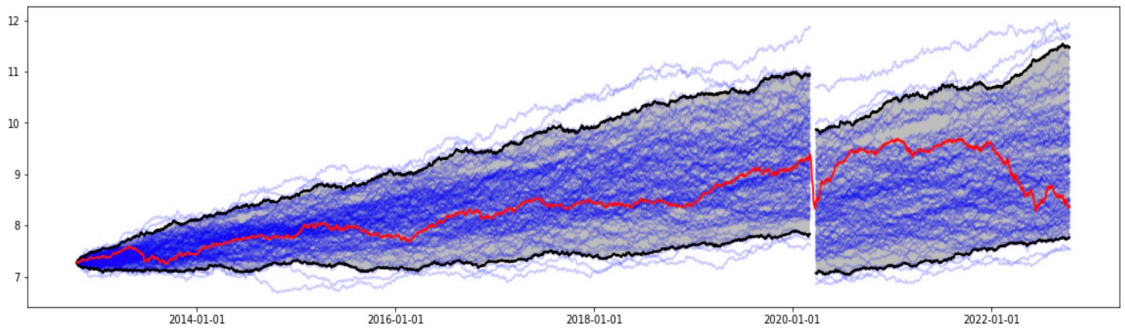


Figure 15: Examples of mutual funds simulation

The historical prices are plotted with red lines, and blue lines are our simulated trajectories. Two black lines are 95% confidence interval of simulated prices. The severity of market crashes in simulation is the actual percentage of drop during that period, and we will discuss more details in Section 3.3.

3.1.5 Value at Risk

Value at risk (VaR) is a statistic that quantifies the extent of possible financial losses within a firm, portfolio, or position over a specific time frame. This metric is most commonly used by investment and commercial banks to determine the extent and probabilities of potential losses in their institutional portfolios. Thus, if the VaR on an asset is $100 million at a one-week, 95% confidence level, there is only a 5% chance that the value of the asset will drop more than $100 million over any given week.

We randomly selected three mutual funds in each risk level and simulated 100 trajectories with a 3-year time frame for each one. The following table shows tickers and corresponding values with 95% confidence that the return will not fall below:

In our simulation, mutual funds with higher risk rating overall have a higher 95% Value at Risk. Generally, investments that are perceived to be riskier tend to have higher potential returns, but they also come with a higher likelihood of significant losses. This relationship is a fundamental tenet of modern portfolio theory and risk management.

Low risk			Low-to-medium risk		
IGI1015	MAX1450	FID5198	MFC13052	CIG16651	FID5821
-1.061%	2.221%	3.005%	-0.041%	-3.331%	-6.011%

Medium risk			High risk		
MMF1028	MFC1235	TML5603	DYN2978	JHN6416	CIG55203
-16.890%	-8.553%	-21.201%	-31.556%	-35.034%	-20.218%

Table 5: 95% VaR of Random Mutual Funds

3.1.6 Mixture Distribution

A mixture distribution is a mixture of two or more probability distributions. Random variables are assumed to come from a number of different populations to create a new distribution. These populations can be univariate or multivariate, although the mixed distribution should have the same dimensionality. In addition, they should either be all discrete probability distributions or all continuous probability distributions. The distributions can be made up of different distributions (e.g., a normal distribution and a t-distribution) or the same distribution with different parameters.

We considered a mixture distribution of normal distribution and t-distribution and used the 'minimize' function in Python to find the best weights (probability) of two distributions by minimizing the negative likelihood function. Then we applied the Kolmogorov-Smirnov test on the empirical c.d.f. with the theoretical c.d.f.. However, the test result was highly unsatisfactory. Only about 30% of mutual funds could fit the mixture distribution well.

Thus, we employ the spliced distribution that we fitted above as the only reliable statistical model for mutual funds.

3.2 Modelling Stocks

3.2.1 Geometric Brownian Motion Model

A geometric Brownian motion (also known as exponential Brownian motion) is a continuous-time stochastic process in which the logarithm of the randomly varying quantity follows a

Brownian motion (also called a Wiener process) with drift. It is an important example of stochastic processes satisfying a stochastic differential equation. We began with a basic introduction to the GBM model and then continue with the estimation of its parameters and simulations.

Assumption: A stochastic process S_t is said to follow a GBM if it satisfies the following stochastic differential equation:

$$dS_t = \mu S_t dt + \sigma S_t dW_t,$$

where W_t is a Wiener process or Brownian motion, and μ and σ are constants.

The interpretation of the variables is as follows:

- S_t: spot price of the asset at time t;
- μ: the percentage drift;
- σ: the percentage volatility.

Parameter Estimation: To introduce the Geometric Brownian Motion model as the model of all common stocks in the research, we plotted the density of daily log-returns to confirm it is approximately normal distributed in Figure 16.

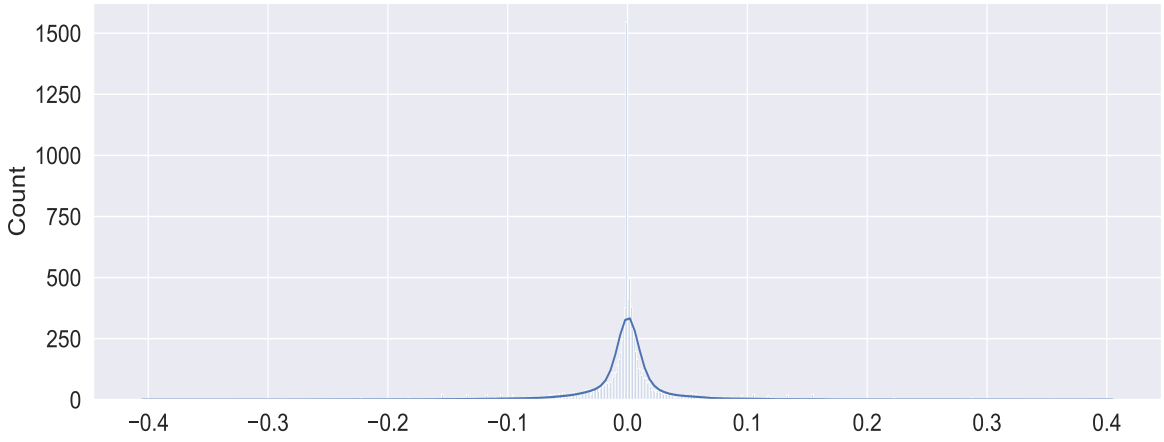


Figure 16: Density plot of log-returns

We derived maximum likelihood estimators for the parameters with the historical market close price. We used N as the sample size and t_n as the date of the n^{th} observation. Let $\log(S_t) = X_t$, $\delta X = X_{t_N}$ - X_{t_0} and $\delta t = t_N$ - t_0 for brevity. The m.l.e. of the drift $\hat{\mu}$ and the volatility $\hat{\sigma}$ are:

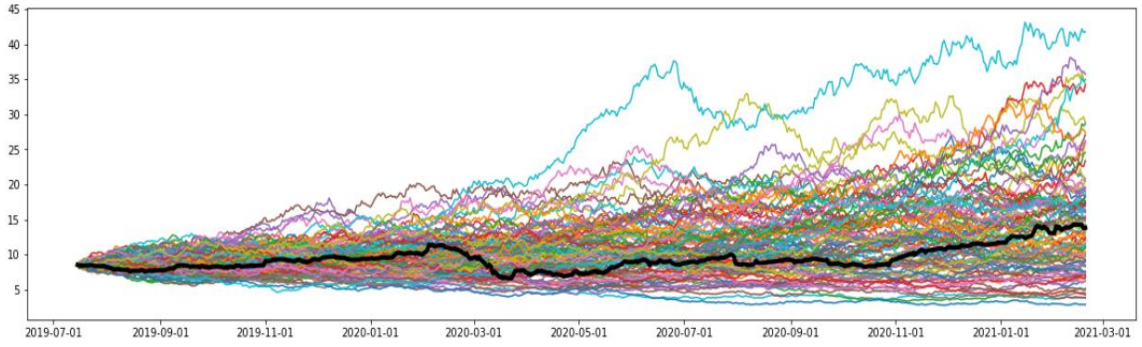
$$\hat{\mu} = \frac{\delta X}{\delta t} + \frac{1}{2}\hat{\sigma}^2,$$
$$\hat{\sigma}^2 = -\frac{1}{N}\frac{(\delta X)^2}{\delta t} + \frac{1}{N}\sum_{n=1}^{N}\frac{\Delta X_n^2}{\Delta t_n},$$

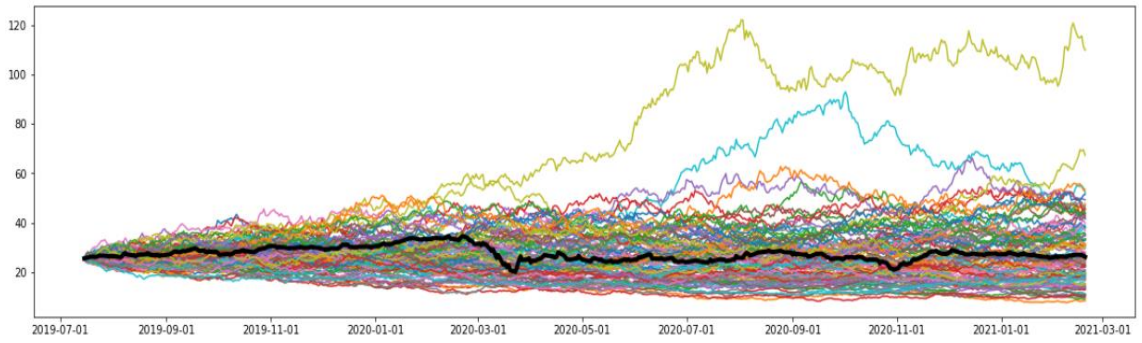
where $\Delta X_n = X_{t_n}$ - $X_{t_{n-1}}$, $\Delta t_n = t_n$ - t_{n-1} and $\Delta W_n = W_n$ - W_{n-1}.

The proof can be done by the fact that (McDonald and Lynch, 2006):

$$\Delta X_n = log\left(\frac{S_t}{S_{t-1}}\right) \sim \mathcal{N}\left(\left(\mu - \frac{1}{2}\sigma^2\right)\Delta t,\ \sigma^2\Delta t\right).$$

The following figures are five examples of simulations under the GBM model and the real stock price trajectories are showed in bold black line.





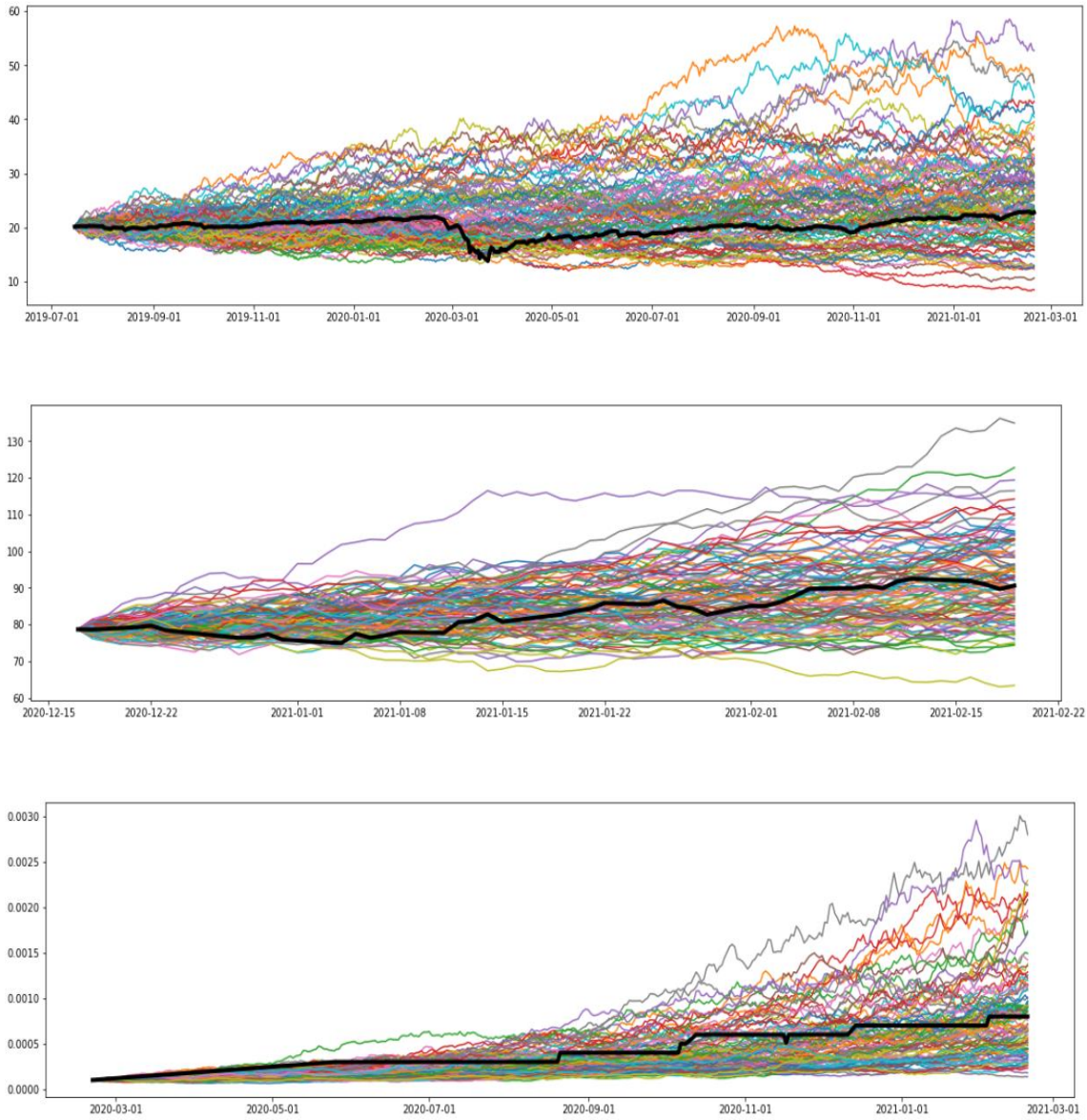


Figure 17: Examples of GBM Simulation

The trajectories are smooth since our simulation results are based on the Geometric Brownian Motion model. However, we noticed some jumps in the historical data, and we tried to apply another model with jumps in the following section.

3.2.2 Merton Jump Diffusion Model

As the previous examples demonstrate, there were some stock trajectories that have jumps, so we introduce the Merton Jump Diffusion model in this section (Bates, 1996).

Assumption: In a Merton Jump Diffusion model, the s.d.e. for the stock price is given as:

$$dS_t = \mu S_t dt + \sigma S_t dW_t + S_t dJ_t,$$

where W_t is a Wiener process or Brownian motion and

$$J_t = \sum_{i=1}^{M_t} Y_i$$

is a compound Poisson process where the jump sizes Y_i are independent and identically distributed with distribution F and the number of jumps M_t is a Poisson process with a constant jump intensity λ.

The compound Poisson process is a model for a series of discrete events (e.g., jumps) where the average inter-arrival time is known, but the exact timing of events that occurred is random. A Poisson process is a counting process (see detail in Section 3.3.2) with following properties (Last and Penrose, 2017):

- $M_0 = 0$;
- M_t has independent and stationary increments;
- The number of arrivals in any interval with length T has Poisson distribution with parameter λT.

Parameter Estimation: To estimate the parameters in the Merton Jump Diffusion model, we extracted jumps from historical prices (computed as log returns) and used the same m.l.e. method as for the GBM model. It is worth noting that we had piece-wise continuous trajectories and estimated parameters based on each part.

Since jumps are also daily log returns which are approximately normally distributed as mentioned before, we applied a threshold to treat relatively larger daily log returns as jumps. Then jumps will follow truncated normal distribution as in Figure 18, and we will discuss how to choose the threshold in Section 3.2.2.

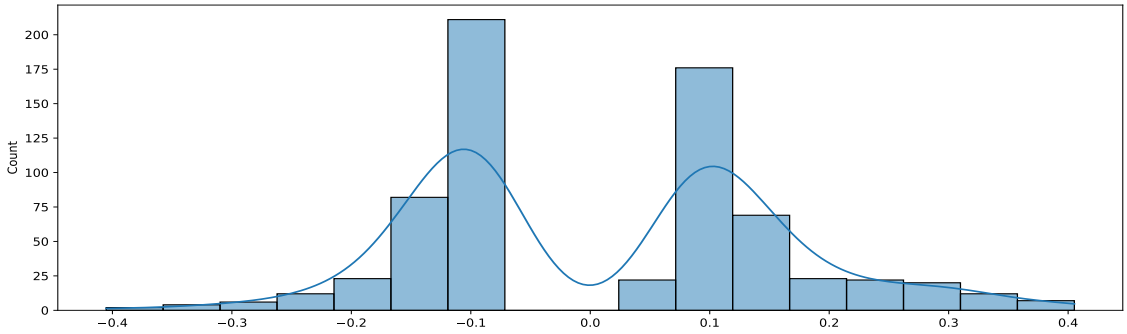


Figure 18: Density plot of Jumps

The mean μ of the truncated normal distribution can be regarded as a perturbation of the mean μ of the parent normal distribution. Its value can be determined by referencing the normal p.d.f. ϕ and c.d.f. Φ, as presented in Johnson et al. (1995). First, define:

$$\alpha = \frac{a - \bar{\mu}}{\bar{\sigma}}; \quad \beta = \frac{b - \bar{\mu}}{\bar{\sigma}},$$

where a and b are limits of truncation range (a, b), and $\bar{\mu}$ and $\bar{\sigma}$ are parameters of parent normal distribution. Then we have:

$$\mu = \bar{\mu} - \bar{\sigma} \cdot \frac{\phi(0,1;\beta) - \phi(0,1;\alpha)}{\Phi(0,1;\beta) - \Phi(0,1;\alpha)};$$
$$\sigma^2 = \bar{\sigma}^2 \cdot \left[1 - \frac{\beta\phi(0,1;\beta) - \alpha\phi(0,1;\alpha)}{\phi(0,1;\beta) - \phi(0,1;\alpha)} - \left(\frac{\phi(0,1;\beta) - \phi(0,1;\alpha)}{\Phi(0,1;\beta) - \Phi(0,1;\alpha)}\right)^2\right].$$

We assumed all jumps to be following a single model to estimate λ, which is the total number of jumps divided by the total number of the trading days. We also estimated λ based on each stock's historical data and simulate stock prices, which is discussed in Section 3.2.3.

Threshold Selection: To choose the proper threshold hat will help us identify jumps reliably, we use levels of the circuit breaker in the stock market. If a stock moves up or down too quickly within a 5-minunte period, it can cause an automatic circuit breaker halt that will pause trading for 5 minuntes.

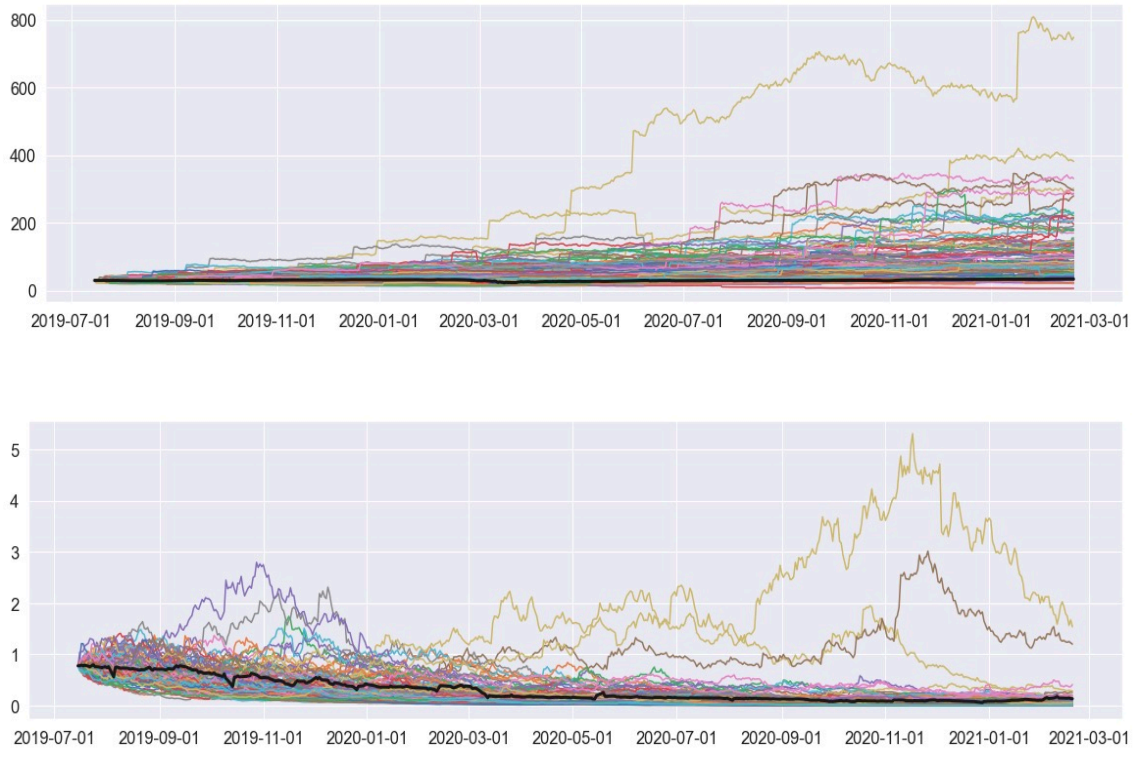
U.S. regulations have three circuit breaker levels, which are set to halt trading when the

S&P 500 Index drops 7%, 13%, and 20%. A market volatility that triggers a Level 1 (7%) or Level 2 (13%) circuit breaker will halt market-wide trading for 15 minutes. A market volatility that triggers a Level 3 (20%) circuit breaker at any time during the trading day will halt market-wide trading for the remainder of the trading day (Santoni and Liu, 1993).

We chose 20% as the threshold to identify jumps, which means daily log returns that are greater than log(1.2) or smaller than log(0.8) will be considered as jumps.

3.2.3 Model Comparison

Figures 19 are examples of Merton Jump Diffusion model simulations when we consider that all jumps are from the same Poisson process.



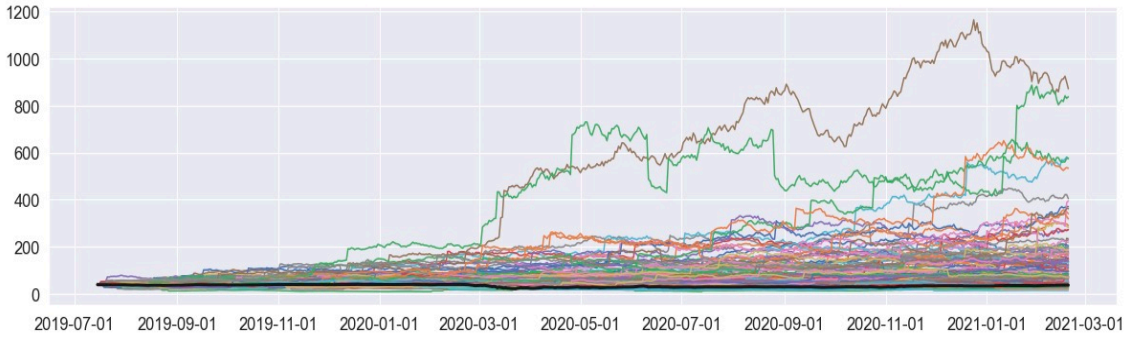
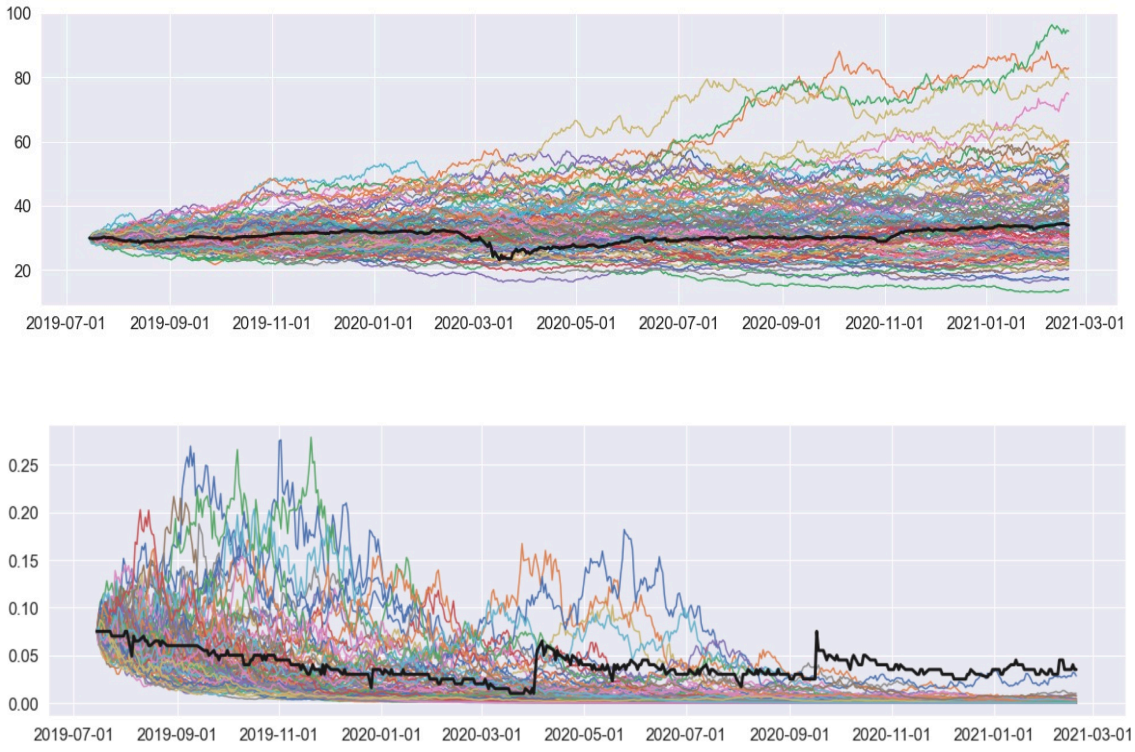


Figure 19: Examples of Merton Jump Diffusion Model Simulation (All)

However, we can see that it might cause some stocks with smooth trajectories to have some jumps in the simulation, resulting in a significant difference in the stock price prediction.

When we estimate the value of λ based on the historical data of each stock, as shown in Figure 20, we can better simulate the trajectories.



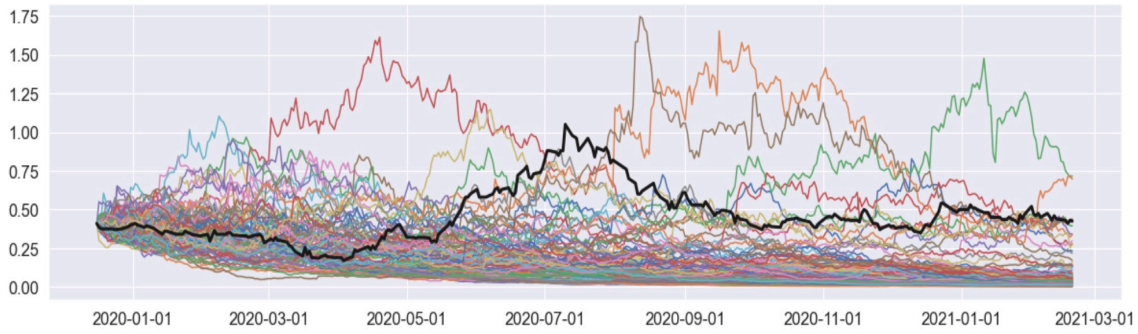


Figure 20: Examples of Merton Jump Diffusion Model Simulation(Single)

In this case, if a stock has no jump in its trajectory, the estimation of λ will become 0, and the model will be exactly the same as the Geometric Brownian Motion model.

3.2.4 Normal-Weibull Spliced Distribution

We also wanted to replicate the success of the spliced distribution on common stocks. However, when we tested the model on 10-year data, only about 46% of stocks showed normality in their central parts. It suggested that the spliced distribution may not be a good choice in modelling stocks.

3.3 Modelling Market Crashes

3.3.1 Definition

A stock market crash is a rapid and often unanticipated drop in stock prices. A stock market crash can be a side effect of a major catastrophic event, economic crisis, or the collapse of a long-term speculative bubble. Reactionary public panic about a stock market crash can also be a major contributor to it, inducing panic selling that depresses prices even further.

To detect historical market crashes from the market index, we followed the definition in Lleo and Ziemba (2017) as an equity market crash as a decline of at least 10% in the level of the S&P500 over a time period of at most a year (252 trading days). We extended the results in Lleo and Ziemba (2017) to 2022 and verified historical financial crises with our detection

(see the Appendix). The average inter-arrival time of a crash was 728 days, and the average decline was 22.6%.

3.3.2 Homogeneous Poisson Process

Due to the limited number of market crashes (only 21 observations available), we wanted to use a simple model in this section. So we considered the homogeneous Poisson process as it is the simplest point process.

Counting Process: A counting process, $N(t), t \geq 0$, is any integer valued process with the following properties:

- $N(0) = 0$;
- $N(t+s) \geq N(t), \forall s \geq 0$.

Homogeneous Poisson Process: A homogeneous Poisson process, $N(t)$, with rate λ is defined as a counting process with independent and stationary increments with the property that the number of points counted in an interval $(t, t+s]$ is given by a Poisson distribution with parameter λs, i.e.,

$$N(t, t+s] \sim Po(\lambda s).$$

Here $N(t, t+s] \stackrel{\Delta}{=} N(t+s) - N(t)$.

The property of independence and stationarity of the increments implies that the number of points counted in any two disjoint intervals is given by two independent Poisson random variable whose parameters are proportional to the size of the corresponding intervals by the proportional constant λ.

Data Fitting: To verify if the homogeneous Poisson process provides a proper fit for our data, we tested data with the following property: Let $\{N(t), t \geq 0, \}$ be a homogeneous Poisson process with rate λ, then the inter-arrival times $\{X_n, n \geq 1, \}$ are independent random variables, each of them distributed as an exponential distribution with parameter λ. Due to the limited samples, we compared the moments (first moment ($E(X)$), second

moment ($E(X^2)$) and third moment ($E(X^3)$)) and concluded that the inter-arrival times are approximately exponentially distributed, and severity roughly follows a Pareto distribution.

3.3.3 Mutual Funds

Unlike stocks, mutual funds experienced different market crashes. Since most mutual funds show one market crash in their trajectories, we could only analyze the average severity in each risk level and noticed that the greater the risk lever of the fund, the greater the severity. The result is shown in the following table:

	Low risk	Low-to-medium risk	Medium risk	High risk
Average severity	11.5%	15.2%	28.0%	39.7%

Table 6: Average serverity of market crashes in each risk level

3.3.4 Markov Switching Dynamic Regression

The Markov Switching Dynamic Regression model is a type of Hidden Markov Model that can be used to represent phenomena in which some portion of the phenomenon is directly observed while the rest of it is 'hidden'. The hidden part is modeled using a Markov model, while the visible portion is modeled using a suitable time series regression model in such a way that, the mean and variance of the time series regression model change depending on which state the hidden Markov model is in.

We only considered the Markov Switching Dynamic Regression as a complement to our method and will not introduce the methodology of this method in detail.

The following graph shows the Markov Switching Dynamic Regression model's regression result in Volatility Index (VIX) historical data, in which the yellow dots are market crashes period.

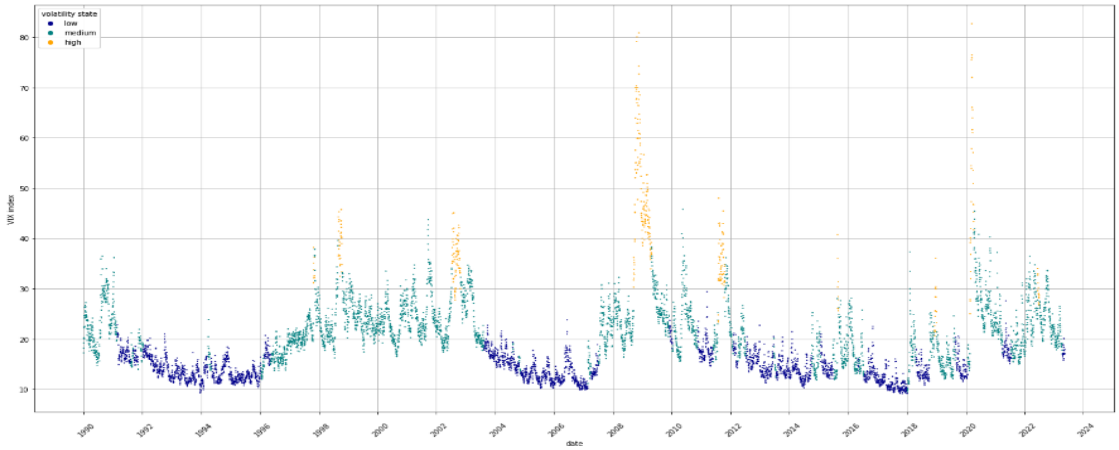


Figure 21: MSDR Results on Volatility Index

Our detection method, by 10% decline, is more sensitive than the Markov Switching Dynamic Regression model. The detection of market crashes, shown in yellow dots, is a subset of our result (see Appendix). It also shows that market crashes always tend to be accompanied by an increase in the volatility index.

4 Conclusions

Modelling long-term portfolio returns is important for our clients since it reflects whether their investments position them well for retirement and can offset unanticipated risks. When modelling the mutual funds, we considered the median absolute deviation method to identify each component of the spliced distribution and estimate the corresponding parameters of each distribution. Although there is no perfect model for the prediction of market securities, the numerical results and simulation plots showed our Normal-Weibull model good fitness and prediction on mutual funds data. It is also worth mentioning that it provided an alternative to the Merton Jump Diffusion model other than the classical GBM model in modelling stocks.

With our modelling approach, we can model individual securities that clients invest in and simulate long-term returns. It can also be combined with more detailed information about investment portfolio proportion and the correlation between each security to show how likely our clients are well-prepared for retirement.

However, our model still has limitations on predicting market crashes as the small sample available. One drawback is that we could only consider a simple model for market crashes since a complex model is hard to build based on 21 observations. Another drawback is that the limited sample led to inaccurate Kolmogorov-Smirnov tests with a specific distribution of our data.

References

Bates, D. S. (1996). Jumps and stochastic volatility: Exchange rate processes implicit in deutsche mark options. The Review of Financial Studies, 9(1):69–107.

Gan, G. and Valdez, E. A. (2018). Fat-tailed regression modeling with spliced distributions. North American Actuarial Journal, 22(4):554–573.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. Journal of the american statistical association, 69(346):383–393.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). Continuous univariate distributions, volume 2, volume 289. John wiley & sons.

Klugman, S. A., Panjer, H. H., and Willmot, G. E. (2012). Loss models: from data to decisions, volume 715. John Wiley & Sons.

Last, G. and Penrose, M. (2017). Lectures on the Poisson process, volume 7. Cambridge University Press.

Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of experimental social psychology, 49(4):764–766.

Lleo, S. and Ziemba, W. T. (2017). Does the bond-stock earnings yield differential model predict equity market corrections better than high p/e models? Financial Markets, Institutions & Instruments, 26(2):61–123.

McDonald, R. L. and Lynch, M. R. (2006). Derivatives markets. Addison-Wesley Boston.

Miller, J. (1991). Reaction time analysis with outlier exclusion: Bias varies with sample size. The quarterly journal of experimental psychology, 43(4):907–912.

Mishkin, F. S. and White, E. N. (2002). Us stock market crashes and their aftermath: implications for monetary policy.

Santoni, G. J. and Liu, T. (1993). Circuit breakers and stock market volatility. Journal of Futures Markets, 13(3):261–277.

5 Appendix

In the following table, we list well-documented market crashes since 1968.

Number	Peak Date	Interarrival Time(days)	Trough Date	Peak-to-Trough decline(%)
1	1968-11-29	784	1970-05-26	36.1%
2	1971-04-28	337	1971-11-23	13.9%
3	1973-01-11	415	1974-04-25	25.5%
4	1975-07-15	446	1975-09-16	14.1%
5	1976-09-21	371	1978-03-06	19.4%
6	1978-09-12	199	1978-11-14	13.6%
7	1980-02-13	456	1980-03-27	17.1%
8	1980-11-28	246	1982-08-12	27.1%
9	1983-10-10	424	1984-07-24	14.4%
10	1987-08-25	1127	1987-12-04	33.5%
11	1990-07-16	955	1990-10-1	19.9%
12	1998-07-17	2836	1998-08-31	19.3%
13	1999-07-16	319	1999-10-18	11.6%
14	2000-03-24	158	2001-09-21	36.8%
15	2007-10-09	2209	2008-11-20	51.9%
16	2010-04-23	519	2010-07-02	16.0%
17	2011-04-29	301	2011-10-03	19.4%
18	2014-09-03	1066	2015-12-14	18.9%
19	2018-07-12	941	2018-12-09	11.1%
20	2020-02-20	438	2020-03-23	37.4%
21	2022-03-29	736	2022-10-12	17.5%

We follow the definition of the market crash in Lleo and Ziemba (2017) as a decline of at

least 10% in the level of the S&P500 over a time period of less than a year (252 trading days). The Peak Date is the date at which the local peak that preceded the crash was reached. The Interarrival Time is the number of calendar days between the through date of the last market crash and the peak date of this market crash. The Trough Date is the date at which the local trough that followed the crash was reached. The Peak-to-Trough Decline is the percentage loss on the S&P 500 from the local peak to the local trough. The Peak-to-Trough duration measures the duration of the decline as the number of calendar days between the local peak and the local trough.

The historical financial crises are listed as follows((Mishkin and White, 2002)):

1968-1970(Recession of 1969-1970):In 1966, the Vietnam War escalated, inflation was rampant, interest rates were surging, and concerns over a global recession pounded stocks. Stocks suffered a bear market in 1966, with the S&P 500 falling about -22% from peak to trough.

1971:The international monetary crisis of 1971 began on August 15, 1971 with the announcement of President Nixon's "new economic policy.

1973-1974: The 1973-1974 stock market crash caused a bear market between January 1973 and December 1974. Affecting all the major stock markets in the world, particularly the United Kingdom, it was one of the worst stock market downturns since the Great Depression.

1975: The 1973-1975 recession or 1970s recession was a period of economic stagnation in much of the Western world during the 1970s, putting an end to the overall post-World War II economic expansion. It differed from many previous recessions by involving stagflation, in which high unemployment and high inflation existed simultaneously

1976-1978: The 1976 sterling crisis was a currency crisis in the United Kingdom. Inflation (at close to 25% in 1975, causing high bond yields and borrowing costs), a balance of payments deficit, a public spending deficit, and the 1973 oil crisis were contributors.

1980: The early 1980s recession was a severe economic recession that affected much of the world between approximately the start of 1980 and 1983. It is widely considered to have

been the most severe recession since World War II. A key event leading to the recession was the 1979 energy crisis, mostly caused by the Iranian Revolution which caused a disruption to the global oil supply, which saw oil prices rising sharply in 1979 and early 1980. The sharp rise in oil prices pushed the already high rates of inflation in several major advanced countries to new double-digit highs, with countries such as the United States, Canada, West Germany, Italy, the United Kingdom and Japan tightening their monetary policies by increasing interest rates in order to control the inflation. These G7 countries each, in fact, had "double-dip" recessions involving short declines in economic output in parts of 1980 followed by a short period of expansion, in turn, followed by a steeper, longer period of economic contraction starting sometime in 1981 and ending in the last half of 1982 or in early 1983. Most of these countries experienced stagflation, a situation of both high inflation rates and high unemployment rates.

1987: Black Monday, Infamous stock market crash that represented the greatest one-day percentage decline in U.S. stock market history, culminating in a bear market after a more than 20% plunge in the S&P 500 and Dow Jones Industrial Average. Among the primary causes of the chaos were program trading and illiquidity, both of which fueled the vicious decline for the day as stocks continued lower even as volume grew lighter. Today, circuit breakers are in place to prevent a repeat of Black Monday. After a 7% drop, trading would be suspended for 15 minutes, with the same 15 minute suspension kicking in after a 13% drop. However, in the event of a 20% drop, trading would be shut down for the remainder of the day.

1990: Early 1990s recession, Iraq invaded Kuwait in August 1990, causing oil prices to increase. The Dow Jones Industrial Average dropped 18% in three months, from 2,911.63 on July 3 to 2,381.99 on October 16, 1990. This recession lasted approximately 8 months.

1998: Global stock market crash that was caused by an economic crisis in Asia. The Russian government devalues the ruble, defaults on domestic debt, and declares a moratorium on payment to foreign creditors.

2000 - 2001: Collapse of a technology bubble. The September 11 attacks caused global stock markets to drop sharply. The attacks themselves caused approximately $40 billion in insurance losses, making it one of the largest insured events ever.

2007-2008: The 2007-2008 financial crisis, or Global Financial Crisis (GFC), was a severe worldwide economic crisis that occurred in the early 21st century. It was the most serious financial crisis since the Great Depression (1929). Predatory lending targeting low-income homebuyers, excessive risk-taking by global financial institutions, and the bursting of the United States housing bubble culminated in a ”perfect storm”.

2010: 2010 flash crash, The Dow Jones Industrial Average suffered its worst intra-day point loss, dropping nearly 1,000 points before partially recovering.

2011: August 2011 stock markets fall, S&P 500 entered a short-lived bear market between 2 May 2011 (intraday high: 1,370.58) and 4 October 2011 (intraday low: 1,074.77), a decline of 21.58%. The stock market rebounded thereafter and ended the year flat.

2015: 2015-2016 stock market selloff,The Dow Jones fell 588 points during a two-day period, 1,300 points from August 18-21. On Monday, August 24, world stock markets were down substantially, wiping out all gains made in 2015, with interlinked drops in commodities such as oil, which hit a six-year price low, copper, and most of Asian currencies, but the Japanese yen, losing value against the United States dollar. With this plunge, an estimated ten trillion dollars had been wiped off the books on global markets since June 3.

2018: 2018 cryptocurrency crash, The S&P 500 index peaked at 2930 on its September 20 close and dropped 19.73% to 2351 by Christmas Eve. Bitcoin price peaked on 17 Dec ’17, then fell 45% on 22nd Dec ’17. The DJIA falls 18.78% during roughly the same period. Shanghai Composite dropped to a four-year low, escalating their economic downturn since the 2015 recession.

2020: 2020 stock market crash, The S&P 500 index dropped 34%, 1145 points, at its peak of 3386 on February 19 to 2237 on March 23. This crash was part of a worldwide recession caused by the COVID-19 lockdowns.

2022: 2022 stock market decline, The S&P 500 index peaked at 4,796 on its January 3 close and dropped 23.55% to 3,666 by June 16, 2022. As part of the global decline in most risk assets, the price of Bitcoin collapsed 59% during the same time period, and 72% from its November 8 all time high. The DJIA fell 18.78% since its January 4 high. Nasdaq Composite fell 33.70% from its November 19 high.

VITA

Name:	Xinghan Zhu
Post-secondary Education and Degrees:	The University of Western Ontario London, Ontario, Canada 2019-2021 B.Sc. The University of Western Ontario London, Ontario, Canada 2021-2023 M.Sc.
Honours and Awards:	Manulife Financial Scholarship in Actuarial Science 2019
Related Work Experience	Teaching Assistant The University of Western Ontario 2021-2023