FACULTY OF ECONOMICS AND BUSINESS ADMINISTRATION

# THE ART OF ROUGHNESS

## A fractal and neural network combination approach to market risk measurement

Word count: 33,953

Emiel Lemahieu

Student number: 01402798

Supervisor: Prof. Dr. Koen Inghelbrecht

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Master of Science in Business Engineering, main subject: Finance

Academic year: 2018 - 2019

GHENT UNIVERSITY

FACULTY OF ECONOMICS AND BUSINESS ADMINISTRATION

GHENT UNIVERSITY

FACULTY OF ECONOMICS AND BUSINESS ADMINISTRATION

# THE ART OF ROUGHNESS

## A fractal and neural network combination approach to market risk measurement

Word count: 33,953

Emiel Lemahieu

Student number: 01402798

Supervisor: Prof. Dr. Koen Inghelbrecht

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Master of Science in Business Engineering, main subject: Finance

Academic year: 2018 - 2019

GHENT UNIVERSITY

*To Antoon Viaene*

*"With every hypothetical solution of a scientific problem both the number of unsolved problems and the degree of their difficulty increase; they increase much faster than do the solutions. And it would be correct to say that whilst our conjectural knowledge is finite, our ignorance is infinite."*

Karl Popper in '*The World of Parmenides*'

# Table of Contents

# Abstract

This master dissertation revolves around the measurement of market risk. The main methodologies in the field are discussed, focusing on their main limitations. Since these standard models are known to be biased, i.e. under- or overestimating the risk out-of-sample, the thesis proposes a combination approach as to reduce the overall bias. Fractal properties of stock market returns are used to gauge the complexity or roughness of the market. The notion of roughness is intertwined with the assumptions made in standard models through the concept of fractional Brownian motion (fBM). The thesis checks whether measures of roughness contribute to a more effective combination of market risk models. Standard models, together with measures of complexity, are fed into a neural network regression model that is used to recognize and memorize the complex non-linear relationships between the measured complexity, volatility and the eventual risk measure as to minimize the number of unexplained exceedances in loss. The model was trained on Google's Cloud TPU infrastructure for approximately eight hundred traded assets from eleven countries and ten different industries. The findings imply significant improvements in the combination model when adding roughness as a feature. However, the discussion emphasizes model mindfulness because of the limited convergence of in-sample results due to a set of recurrent data issues. The dissertation expands on the implications for risk managers at financial institutions, as well as for asset managers and traders who use the described methodologies for optimization purposes. The thesis concludes by stressing the crucial point of model mindfulness, since black box risk measurement should always be accompanied by a critical mindset.

**Key words:** market risk, combination model, roughness, fractional dimension, Hurst exponent, fractional Brownian motion, machine learning, model mindfulness

# List of Acronyms

| | |
|---|---|
| **AdaDelta** | Variant of AdaGrad that uses only first-order information (Delta) |
| **AdaGrad** | Adaptive Gradient Descent |
| **Adam** | Adaptive Momentum Estimator |
| **AdaMax** | Variant of Adam based on the infinity norm |
| **ADL** | Autoregressive Distributed Lag |
| **API** | Application Programming Interface |
| **ARFIMA** | Autoregressive Fractionally Integrated Moving Averages |
| **ARIMA** | Autoregressive Integrated Moving Averages |
| **BCBS** | Basel Committee on Banking Supervision |
| **BEL20** | Belgian index of 20 biggest listed companies according to market cap |
| **BM** | Brownian Motion |
| **CAPM** | Capital Asset Pricing Model |
| **CDI** | Christian Dior Industries (ticker) |
| **CDO** | Collateralized Debt Obligation |
| **CIO** | Chief Investment Officer |
| **CLT** | Central Limit Theorem |
| **CR** | Conditional Ratio |
| **CRM** | Coherent Risk Measures |
| **CVaR** | Component VaR (not Conditional VaR, which is referred to as Expected Shortfall or ES) |
| **D** | Fractional Dimension |
| **DNN** | Deep Neural Network |
| **EGARCH** | Exponential GARCH |
| **Elu** | Exponential Linear Unit |
| **EMH** | Efficient market hypothesis |
| **ES** | Expected Shortfall |
| **EWMA** | Exponentially Weighted Moving Averages |
| **fBM** | Fractional Brownian Motion |
| **FHS** | Filtered Historical Simulation |
| **fOU** | Fractional Ornstein-Uhlenbeck process |
| **FRTB** | Fundamental Review of the Trading Book |
| **FTGT** | Fisher-Tippet-Gnedenko Theorem |
| **FVaR** | Fréchet VaR |
| **G-SIB** | Global Systemically Important Bank |
| **GA** | Genetic Algorithm |
| **GARCH** | Generalized Autoregressive Conditional Heteroskedasticity |
| **GBM** | Geometric Brownian Motion |
| **GCP** | Google Cloud Platform |
| **GDP** | Gross Domestic Product |
| **GEVT** | Generalized Extreme Value Theory |
| **GFNN** | Genetic Fuzzy Neural Networks |
| **GVaR** | Gumbel VaR |
| **H** | Hurst Exponent |
| **HARCH** | Heterogenous Autoregressive Conditional Heteroskedasticity |
| **HS** | Historical Simulation |
| **i.i.d.** | independent and identically distributed |
| **IDE** | Interactive Development Environment |
| **IS** | In-Sample |
| **IVaR** | Incremental VaR |
| **logVaR** | lognormal VaR |
| **LPHI** | Low Probability, High Impact |
| **LR** | Loglikelihood Ratio |
| **MCS** | Monte Carlo Simulations |
| **ML** | Machine Learning |
| **MPT** | Modern Portfolio Theory |
| **MVaR** | Marginal VaR |

| | |
|---|---|
| **NaN** | Not a Number |
| **Nadam** | Adam with Nesterov Momentum |
| **NVaR** | Normal VaR |
| **OHLCV** | Open, High, Low, Close and Volume |
| **OLS** | Ordinary Least-Squares |
| **OOS** | Out-Of-Sample |
| **PnL** | Profit and loss function |
| **POF** | Points-Of-Failure test or Kupiec test |
| **R/S [analysis]** | Rescaled-Range Analysis |
| **Relu** | Rectified Linear Units |
| **RFSV** | Rough Fractional Stochastic Volatility |
| **RMSE** | Root-mean-squared error or L2 loss |
| **RMSprop** | Root-Mean-Squared propagation |
| **RWA** | Risk-Weighted Assets |
| **SABR** | Stochastic Alpha, Beta & Rho (vol model) |
| **SDE** | Stochastic Differential Equation |
| **SGD** | Stochastic Gradient Descent |
| **SRM** | Spectral Risk Measures |
| **Tanh** | Tangent Hyperbolic |
| **TPU** | Tensor Processing Unit |
| **TVaR** | Student t VaR |
| **UR** | Unconditional Ratio |
| **VaR** | Value-at-risk |
| **VAR** | Vector Autoregression |
| **YCD** | Refinitiv mnemonic for CDI (Christian Dior Industries) |

# List of Tables



# List of Figures

# Introduction

*"Bottomless wonders spring from simple rules…*
*which are repeated without end."*

Benoît B. Mandelbrot

Market risk measurement is a central topic in quantitative finance that gained special attention after the 2007-2009 financial crisis. The quantification of market risk has major consequences for any financial institution's risk management and its profitability through the risk-based allocation of assets and the determination of capital requirements. From an allocation perspective, the proportion of market risk contributed by individual positions to the overall portfolio can give the asset manager valuable insights in the composition of his portfolio in terms of risk. Market risk measurement is also the basis for the quantification of capital buffers. Consequently, the link with recent developments in the Basel framework for market risk will be made.

This dissertation starts off in Chapter 1 by defining this specific type of risk and gives a brief overview of the main parametric, non-parametric and Monte Carlo approaches to its quantification. The goal of these sections is not to delve into the mathematical details behind the models, but to discuss their main assumptions, the intuition behind them and, most importantly, their main limitations. A main point that will be stressed is that alternative approaches often make similar assumptions about the underlying stochastic process. For example, parametric distributions can be mapped on a corresponding stochastic process that would generate identical results in a Monte Carlo. Consequently, some of these models suffer from the same weaknesses. Similarly, it does not matter whether we use one extreme quantile (VaR), a mean of tail VaRs (ES) or a weight function based on risk-aversion (SRM), if we use the exact same methods to come up with the estimations of these concepts. In such a setting, different concepts with similar underlying assumptions will typically have similar shortcomings. Additionally, these standard methods have a consistent bias, i.e. they tend to be

overaggressive or overconservative out-of-sample. This insight tells us that a combination of methods can lower the overall bias.

From an epistemological point of view, one can make the distinction between a theory and a model[1]. A theory has some truthfulness from an objective point of view: it is simply how we assume the world works until the theory is falsified. A model always uses some sort of analogy. We model the problem universe, i.e. what we don't understand, by using building blocks from the world we think we understand. For example, the Black-Scholes model does not describe the world of options in an absolute sense. It does not tell us where an option should trade. The model only comes up with a reasonable price if the modeler tells the model what the future volatility will be and if he makes assumptions about the underlying stochastic process. Models are never correct, but only to the extent that the analogy is justified. The analogy used throughout the dissertation is that one can measure the roughness of a price process, using scaling properties of time and variance, and that subsequently roughness can be used to assess the appropriateness of the yardstick one uses to measure risk. The latter is the research hypothesis that will be tested in the dissertation. Do complexity measures really add something to the models other than '*spurious precision*'? Alternatively: ***"Can measures of roughness contribute to a more effective combination of market risk measures?"*** Because of the fundamental relationship with the assumptions in the classical models, fractal properties might add something to the combination algorithms in a more parsimonious way than including a lot of lagged values for the risk measures. Moreover, measures for the smoothness of the volatility process have recently entered the equations in the so-called rough volatility literature (cf. section 2.6), showing a regained interest of fractals in finance.

To sum up, fractal geometry is the theoretical angle of this dissertation, but I acknowledge that the practical approach is atheoretical in nature, and the outcome is simply a model whose validity is only justified to the extent that the analogy holds.

[1] For great reflections on this topic, read Derman's '*Models. Behaving. Badly.*' (2011).

In the second chapter, the thesis discusses the basics of fractal geometry from an intuitive point of view. The mathematics behind it can be of extraordinary complexity, but the goal is again not to lose ourselves in mathematical detail. The focus of these sections will be on explaining the link between roughness and finance. That is why, without pursuing mathematical rigor, Chapter 2 zooms in on fractional dimensions, Hurst exponents and the link with fractional Brownian motions. These sections further explain how some of the critical assumptions made in standard models are linked with these fractal properties of the stock price time series and how generalizing these assumptions by measuring the actual roughness might be useful in market risk modeling. Throughout the thesis we check whether rough markets are a useful hypothesis for combining risk measures. This main hypothesis thus boils down to roughness somehow measuring the uncontrollable element in financial markets, in another way than the traditional stochastic volatility models.

The practical approach to the combination model is machine learning. As will be explained in Chapter 3, deep neural network regression models are mathematical models that are able to combine different inputs into one output measure with the ability to recognize and memorize complex relationships between the input data in so-called hidden layers. Hence, these models try to go beyond the typical linearities that sneak into the standard equations that link the volatility of an asset to its risk measure. Moreover, through the concept of loss functions, ML models can better cope with the exception notion of loss considered in the backtesting of market risk models, leapfrogging mainstay econometric approaches like least-squares or maximum-likelihood estimators. In addition, these sections point out how the model incorporates the Basel traffic-light approach to backtesting into the model. A custom loss function, based on the Kupiec-Christoffersen framework, was designed to test the significance of the number of exceedances. Furthermore, the implemented code uses genetic algorithms that are able to determine the best values for the model's hyperparameters. These algorithms use a population of different network models that breed, i.e. mutate in newer generations with different hyperparameters, in order to improve the model's

accuracy over time. After a number of generations, the children of these models have superior performances in terms of unexplained exceedances.

In Chapter 4, the set-up of the model is discussed. It explains how the outcomes of the models described in Chapter 1 are combined in the input layer together with our measures of roughness. The transferability of the results in terms of VaR to ES and more general spectral measures will be explained in due detail. All the meaningful technical choices that were made will be explained to ensure the reproducibility of the model[2]. Furthermore, the scope of the dataset will be discussed in detail. It comprised the 11 countries of the G10 with for each country representative stocks from 10 sectors: Finance, Technology, Utilities, Telecommunications, Consumer Services, Health Care, Consumer Goods, Industrials, Basic Materials and Oil & Gas. Next, a visual overview of the main results that were delivered by the model are given. Predicted risk measures, as well as their relationship with returns, are displayed and discussed accordingly.

Chapter 5 discusses the main conclusions that can be drawn from chapter 4. It will focus on the main implications of this thesis for risk managers at financial institutions, asset managers making risk-based allocation decisions and traders that seek for other quantitative ways to assess the efficiency of assets in their search for alpha.
The end of this dissertation will be a plea for model mindfulness. The model has some attractions but also many limitations, with the limited understanding of why results converge (and why not) as the one outstanding shortcoming. The "*I don't know how it works, but it looks like it simply works*" way of thinking had devastating repercussions for the financial system about a decade ago. Therefore, I will underscore the importance of a critical appraisal of algorithms and appropriate governance supporting algorithmic trading and risk management.

[2] Please note that the code is made available on GitHub [emiellemahieu/AOR]. The reader is warmly invited to take a look at the script and share his/her findings with me.

# Chapter 1

# What is market risk and market risk measurement?

## 1.1 Definition

**Market risk:** the prospect of a financial loss due to unforeseen changes in the underlying risk factors, i.e. market prices (e.g. stock prices) or market rates (e.g. interest rates).

The above definition, based on Dowd (2007) , explains that *market risk* boils down to unexpected losses that might occur on assets or portfolios of assets whose value is dependent on the movement of a market price or rate. We are essentially talking about financial contracts that are (contingent) claims on other assets. Stocks are the most well-known claims on financial markets, i.e. claims on the equity of companies and/or their income or dividends. Futures and forwards are examples of linear products that are a claim on some underlying asset in the future, that are therefore prone to market fluctuations. The value that one risks with these products will be a linear function of the measured variability in the underlying. Options are typically non-linear claims on an asset. Their pay-off can be compared to the rectified linear units we will discuss in the machine learning part. E.g. a European call has a zero payoff when the spot rate is below the strike price at maturity and shows linear increments in payoff if the strike increases further. Changes in the volatility, the common perception of riskiness, will have non-linear effects on its price, as the Black-Scholes and other option valuation models imply. Bonds are claims on the periodical payment of coupons and the notional at maturity. They are called fixed income because these cashflows are known in advance, if we neglect default risk or optionality. We could argue that the underlying market rate driving the fluctuations in the bond's price, however, are interest rates. Bonds thus have a linear response in price for small changes in the underlying rate (cf. the bond's duration) and require convexity corrections for larger swings.

In any of these examples, the quantification of the risk in the underlying, or the distributional properties of its returns, will determine the value one risks with the contract itself. In this dissertation, we will try to quantify this risk most effectively and will refer to this as the *measured market risk*. Consequently, quantifying the worst quantiles of some *hypothetical profit and loss* (PnL) function of these assets will serve as a basis for the revaluation of these contracts in unfortunate but probable scenarios. For linear products like futures and forwards, the risk will be a multiple of this quantile, while for more complex products, a full revaluation based on these projected losses needs to be done.

## 1.2 Importance

Financial practitioners will attest to the importance of market risk measurement for effective risk management. As was implied in the previous paragraph, market risk quantification is a first step in order to appropriately hedge the risk that is taken. This explains the inextricable link between the mathematics of risk measurement and risk management. In general, managing risk properly is one of the most important concerns a financial institution faces.

Firstly, it vastly determines the financial institution's *profitability* through *strategic* and *tactical capital allocation*. Risk-return trade-offs and problems of optimization need to be aligned with the bank's risk appetite and need to be based on fully risk-adjusted returns: *how much return do I expect for the risk that I take, and how is the risk of the portfolio distributed across its constituents?* We will reintroduce these questions in section 1.11 and try to answer them in 4.3.

Secondly, risk modeling is the centerpiece in the *calculation of the required capital* of a bank. Naturally, the maximum likely losses that are implied in a financial institution's PnL determine how much money they should set aside for a bad day. For example, the regulatory capital for credit risk is based on the structure of the credit portfolio of the bank. The loans' probabilities of default, the product of a so-called PD model, together

with BCBS[3] defined correlations are fed into the infamous one-factor copula model that determines the regulatory capital. Focusing on market risk, however, Pillar I of the Basel accords (see Table 1 below) describes how the extreme quantiles of a hypothetical PnL should be translated in risk measures. It consists of the quantitative guidelines stipulated by the BCBS that determine how the RWA (risk-weighted assets), the denominator in the leverage ratios, are calculated. The most important elements in the new framework are the shift from Value-at-Risk (VaR) to Expected Shortfall (ES) as the reported measure of risk under stress, and the debate about the PnL attribution test. The latter tests how the hypothetical PnL, used to derive the extreme quantiles in the risk management department, reflects the actual PnL that is realized in the front-office. These revisions of the treatment of market risk are part of the so-called FRTB or fundamental review of the trading book (see Farag, 2017, for an overview).

Table 1: The Basel framework

| **The Basel Framework** | | |
|---|---|---|
| **Pillar I:**<br>**Capital Requirements** | **Pillar II:**<br>**Governance & Supervision** | **Pillar III:**<br>**Reporting** |
| • Credit Risk<br>• Market Risk<br>• Liquidity Risk<br>• Operational Risk | • Supervisory review (SREP)<br>• Internal Capital Adequacy Assessment Process (ICAAP) | • Communication of scope<br>• Risk appetite & exposures<br>• Assessment process<br>• Overall adequacy |

The relevance of market risk measures for the calculation of the RWA can be brought back to the fact that the total market risk of a portfolio (as measured by VaR, see next section), is multiplied by a factor of approximately 12.5 to come up with the RWA for that portfolio. Typically, this number is corrected with multipliers (depending on the model's backtesting results) and/or penalties for G-SIBs[4]. Hence, the required capital in terms of RWA is approximately 8%, without capital surcharges and other corrections[5].

[3] Basel Committee on Banking Supervision
[4] Global Systemically Important Bank
[5] See Allen, Boudoukh & Saunders, 2009, p. 200-232

It goes without saying that the step from risk measure to capital requirement comes with a lot of regulatory complications and refinements in addition to the above representation. However, due to space constraints, I am not going to give an extensive overview of the Basel market risk framework. For that purpose, there is a lot of literature available (Decamps et al., 2004; Dierick et al., 2005; Hannoun, 2010), including the original regulatory texts (BCBS III, 2017). However, the reader should understand the consequences of the models that follow, both on the profitability and robustness of the financial system. There has been a lot of (valid) criticism on quantitative models in finance and the limitations of their use, but the reader should understand that the basis for the leverage of financial institutions and therefore the stability of the system still hinges on these concepts and these types of models.

## 1.3 How market risk is measured

*"If you give a pilot an altimeter that is sometimes defective, he will crash the plane. Give him nothing and he will look out the window."*

Nassim N. Taleb

For many years the mainstay in market risk measurement was VaR (*Value-at-Risk*). It answers the fundamental question: "*How much money are we maximally expecting to lose over a certain time period, given a certain confidence level (cl)?*"

VaR can be calculated on an individual asset's level, portfolio level or VaR can be aggregated over portfolios into an institution's total VaR. VaR essentially boils down to calculating the quantile of the PnL distribution corresponding to a certain *cl* (Liu, 2005):

$$p = \Pr[\Delta V(l) < VaR] = F(VaR) \tag{1.1}$$

This means that the probability p, that the change in market value V of the asset/portfolio is worse than the VaR over the $l$-day horizon is 1-*cl*. Hence, this is one quantile of the PnL distribution that we briefly discussed before.

In literature, one will broadly find three approaches to predict these quantiles. The first set of approaches are called parametric approaches. One simply imposes a statistical distribution on the PnL as to calculate the quantiles from the corresponding quantile formula. Another set of approaches are the non-parametric approaches. There, we make no assumptions about statistical properties of the PnL. We do not impose a best fit on the PnL for the mean, variance, skew and excess kurtosis of returns (or more generally some location, shape and scale parameters). However, we try to extract as much information as possible from the past n observations. Lastly, Monte Carlo simulations are widely used numerical methods where one first assumes a stochastic process driving the market fluctuations. Subsequently, we can simulate thousands of price paths using a computer or random number generator that can mimic the statistical properties of the stochastic process. The worst x% observed paths are then used to revalue the portfolio and calculate the 1-x% VaR.

## 1.4 Parametric models

Parametric models generate quantiles by fitting a statistical distribution on the PnL. Once an appropriate distribution is found, one has closed-form solutions for every quantile. There is a large number of distributions in the statistician's toolbox one can choose from:

- *Normal (Gaussian)* bell curves for their mathematical simplicity.
- *Student t* distributions capture excess kurtosis through empirically determined degrees of freedom.
- *Lognormal* distributions are implied by geometric Brownian motion (cf. infra).
- *Gamma* distribution can accommodate realistic skew and kurtosis and is also quite common.
- *Stable Paretian* distributions accommodate fat tails and specify normal and Cauchy as special cases.
- *Gumbel, Weibull* or *Fréchet* distributions are implied by Generalized Extreme Value theory (cf. infra).
- …

For an overview of frequently used distributions, see Dowd (2007) and Allen et al., (2009). The main advantage of this approach is that we can simply write down VaR as a function of the distribution's parameters and the *cl*, as is done in Table 2 below.

Table 2: Some common parametric specifications of VaR

| Name | Formula | Comment |
|---|---|---|
| NVaR | $-\mu_t + z_{cl}\sigma_t$ | (1.2) with $\mu_t, \sigma_t$ the average daily return and volatility at t respectively. Z refers to the standard z-scores for confidence *cl*. |
| TVaR | $-\mu_t + \sqrt{\frac{\nu-2}{\nu}}t_{cl}\sigma_t$ | (1.3) with $\nu$ degrees of freedom, derived from observed excess kurtosis[6] and t the standard t-score. |
| LogVaR | $1 - e^{\mu_t - z_{cl}\sigma_t}$ | (1.4) |
| Gumbel VaR | $-\mu_t + \sigma_t \ln(-\ln(p))$ | (1.5) when ξ, the shape parameter, equals 0 and p = 1-*cl* (cf. infra) |
| Fréchet VaR | $-\mu_t + \frac{\sigma_t}{\xi}[1-(-\ln(p))^{-\xi}]$ | (1.6) with $\xi > 0$ (cf. infra) |
| … | … | … |

The point of this section is not to drop formulas, nor to be exhaustive or confuse the reader. It just illustrates that parametric VaR is fairly simple to implement and interpret. However, one has a very static or backward-looking view on risk if one picks a certain distribution. It assumes that this distribution does not change, at least for the period of the calculation's purpose. It is clear that this assumption of stationarity does not make any sense in highly dynamic markets and the frequency of recalibration plays a crucial role in the model's usefulness. Moreover, parametric approaches are prone to other commonly encountered statistical assumptions (see 1.4.1). Maybe the most important flaw is that it is very sensitive to its most crucial input: *volatility*.

[6] Again, the point of this section is not to delve into the implementation and calibration issues. However, what is meant by deriving $\nu$ from excess kurtosis K is that $\nu$ is related to K through $K=3(\nu-2)/(\nu-4)$. $\nu$ can be approximated by the closest integer to (4K-6)/(K-3).

The dogma in financial theory is that expected returns are mean returns and the risk of the return is its dispersion around that mean. The dispersion is typically modeled by its variance or sigma. In most applications this is done by *stochastic heteroskedasticity* or *stochastic volatility* ($\sigma_t$). The modeling of $\sigma_t$ is the most crucial part of the risk modeling exercise. Even in non-parametric (for example filtered historical simulations) and MCS approaches (the uncertainty in the stochastic equations), sigma is crucial input. Hence, a good model requires a sophisticated, conditional treatment of volatility that takes into account volatility-altering regimes or volatility clusters like we observe them in reality (Engle, 1982; Engle & Patton, 2007; Mandelbrot, 1972).

This conditional perception of volatility is often implemented by *generalized autoregressive conditional heteroscedasticity* (GARCH) models (Bollerslev, 1986; Engle, 1982):

$$\text{GARCH(p,q):} \qquad \begin{cases} \mu_t = \alpha_0 + \sum_{i=1}^{u} \alpha_{1,i}.\mu_{t-i} + \sum_{i=1}^{v} \alpha_{2,i}.a_{t-i} + a_t \\ {\sigma^2}_t = \beta_0 + \sum_{i=1}^{p} \beta_{1,i}.{a^2}_{t-i} + \sum_{i=1}^{q} \beta_{2,i}.{\sigma^2}_{t-i} \end{cases} \qquad (1.7)$$
$$a_t = \sigma_t \varepsilon_t, \quad \varepsilon_t \sim N(0,1)$$

Typically, a GARCH(1,1) simplification performs well:

$$\text{GARCH(1,1):} \qquad \begin{cases} \mu_t = \alpha_0 + \alpha_1.\mu_{t-1} + u_t \\ {\sigma^2}_t = \beta_0 + \beta_1.{u^2}_{t-1} + \beta_2.{\sigma^2}_{t-1} \end{cases} \qquad (1.8)$$
$$u_t = \sigma_t \varepsilon_t, \quad \varepsilon_t \sim N(0,1)$$

GARCH(1,1) is both parsimonious and powerful to capture most of the PnL's distributional properties. This is demonstrated by Robert Engle in a nice example of calibrating VaR using a GARCH process (Engle, 2001). The example also aptly illustrates that although GARCH is based on a conditional normal return distribution, its unconditional distribution is skewed and has excess kurtosis. However, sometimes there is a correction needed for asymmetries in the volatility process or heavy-tailed error terms (e.g. t-distributed $a_t$ and/or $u_t$). Hence, EGARCH and HARCH with

different distributions for $\varepsilon_t$ are considered in the code[7]. There are many stochastic volatility models such as Heston, SABR, Hull-White and so and so forth. The models in the code are estimated from a GARCH approach since the model is highly tractable and a good starting point for further research[8].

It should be noted that the biggest errors of GARCH models are located in the tails since it is essentially no more than normal distributions stretching out and shrinking again over time (Mandelbrot & Hudson, 2010). That is why GARCH typically performs well to explain variance in an available sample, but has limited use in forecasting out-of-sample volatility for e.g. FHS purposes (Ding & Meade, 2010).

### 1.4.1 Some preliminary reflections on the dangers of variance myopia

> *"Anything that relies on correlation is charlatanism."*
>
> Nassim N. Taleb

However sophisticated the treatment of volatility, the mean-variance framework embedded in portfolio theory, capital asset pricing theory and inherited in some risk management applications brings substantial model risk with it. Focusing too much on mean returns and their variances is dangerous, whether we introduce a stochastic volatility model or not. This has been documented extensively in literature and it urges us to think beyond volatility to model the uncontrollable element in markets.

First of all, conclusions drawn from the framework often support on *'elliptical assumptions'*. These often come down to extreme assumptions of normality or i.i.d. returns invoked by standard methodologies. For instance, normal noise is the rule rather than the exception in a lot of econometric applications. Some other 'well-behaving' distributions are then derived from this standard normal and tend to underestimate risk in the sense that no matter how 'fat' the tails become due to

[7] For a concise overview of these models, see Frömmel (2013), p. 48-58.

[8] It should be mentioned that strictly speaking GARCH is a conditional model of volatility but not a stochastic model, since at time t the volatility is completely pre-determined (deterministic) given previous values (Brooks, 2019).

increasing conditional sigma or unconditional excess kurtosis, they typically decline exponentially[9]. Large swings are considered outliers and outliers tend to be ignored, while from a risk management perspective they are the most important observations in the sample. In risk management, despite the fact that we know that these methods are too aggressive, practical applications (cf. delta-normal VaR) often support on these distributions where the moments are time-invariant (recall the previous section).

Secondly, models related to the mean-variance framework make - often implicit - assumptions about *linearities* between risk and return (or volatility and risk measure in related risk applications). For instance, think about CAPM (Merton, 1973) of which the outcome is a linear relationship between return, as a reward for risk, and the sigma-based[10] risk measure. Another example is Markowitz' portfolio selection (Markowitz, 1952; 1991) where investment decisions are based on variance-covariance matrices that measure the linear comovements of returns. Many of these limitations simply come from how we define correlation and covariance (i.e. *Pearson rho*[11] and models of linear dynamics).

Hidden and unhidden '*Gaussian assumptions'* are omnipresent in these models. While this mainstream finance can be helpful in ordinary portfolio management during '*normal trading time'*, it falls short for extraordinary low-prob, high impact (LPHI) events as considered from a risk perspective (Mandelbrot and Hudson, 2010; Taleb, 2007). The problem of mean-variance myopia is epitomized by the tendency to use elliptical distributions when one is fixated by the first two moments of the return sample only. This empirical fact has been extensively documented and presented in popular media. Think about Nassim Nicholas Taleb's bestseller The Black Swan (Taleb, 2007). This fact is mainly due to the mathematical convenience of these models, as well as assumptions being based on fallacies like the CLT and EMH (cf. 1.4.1). Another

[9] Not all standard models lead to exponential tails. One could roughly distinguish between exponential tails (normal, lognormal,…) and the ones resulting from a power law (Pareto, Lévy,…).

[10] A stock's beta essentially quantifies the covariance between a stock and the overall market over the latter's volatility.

[11] The criticism on Pearson rho and the hazard of static correlations, similarly to stochastic sigma, gave rise to *stochastic rho modeling*. However, many theorists are still very critical about linear correlation models. Meissner calls it the 'work of the devil' (Meissner, 2015), Taleb calls it charlatanism and Wilmott argued: "*Instruments whose pricing requires the input of correlation … are accidents waiting to happen*".

example is the *Black-Scholes formula*, a Nobel prize winning application of physics[12] on financial returns. A typical fallacy that explains some of the contradictions in the model, like the famous volatility smiles and smirks, is that only the drift (risk-free rate) and the volatility (the variance) are considered, but no higher moments of the returns or the volatility process are required. It is therefore not surprising that models that include (stochastic) return kurtosis or return skew, as well as stochastic volatility models that take the volatility of volatility into account, do a better job in explaining option prices.

When JP Morgan published the RiskMetrics' tech document on VaR in October 1994 presenting Value-at-risk as a new paradigm for risk measurement, their models supported on Gaussian distributions. A 95% daily VaR expects around 12,5 exceptions on a yearly basis, assuming the typical 255 trading days. When JPM noticed that the actual number was 'way beyond' this number, they quickly refined their models. Hence, the irony is that models that support on Gaussian assumptions work perfectly when they are not needed, during times when benign markets are just moving smoothly. These models exactly fail where they are needed in risk management: in the tails.
A more fundamental reason why people still rely on these distributions in mainstream finance is maybe best explained by behavioral finance. *Disaster myopia* or the tendency to underestimate shock probabilities was first described in banking by Guttentag and Herring (Guttentag and Herring, 1997) and finds its roots in behavioral economics (Kahneman & Tversky, 1982). Two features of human nature aptly explain the omnipresence of the aforementioned models:

(1) *Availability heuristic* or *Ease of recall*: "*The modeler estimates probability by how easily he can recall a similar event*." Frequent events are usually easier to recall than infrequent events, which is why outliers are neglected.
(2) *Threshold heuristic* or *Too small to matter*: "*When a probability reaches some critically low level, the modeler treats it as if it were zero*." This explains why 'multiple-sigma' events are not explained by models, while they are observed once in a while.

[12] Namely geometric Brownian motion and the Feynman-Kac equation.

The result of all the above is that this blindness for LPHI events is often (implicitly) modeled into the risk models, while they matter most. An important theoretical basis for the inclusion of elliptical distributions is the Central Limit Theorem (CLT) and its link with the celebrated Efficient Market Hypothesis (Fama, 1998). That is why we will briefly zoom in on the CLT and elaborate on the differences with Generalized Extreme Value Theory (GEVT). The latter theorem yields another set of distributions that are included in the code.

In summary, these preliminary reflections on the limitations of focusing too much on volatility imply that these classical models will often understate risk which will lead to anomalous results. As is often the case in science, whenever there appears to be a contradiction, this is because of a (hidden) assumption (Witten, 1995). If you probe more deeply, you can reveal these assumptions and realize there is no contradiction anymore. This is exactly the purpose of this chapter, i.e. continuous awareness of the limitations and the biases of the models that were introduced will help us to understand the purpose of combining different models into a more comprehensive one.

### 1.4.2 Central Limit Theorem (CLT)

If you sample n observations from independent populations with identical distributions (i.i.d) than, no matter what the underlying distribution is, the mean of the sample will be normally distributed, with the population mean as expected value and the sample variance divided by n as variance, for n increasing indefinitely. This key finding in statistics is well-known and widely used in financial applications in all kinds of contexts from entry-level regression settings to the test distributions in the most advanced econometric models. In risk measurement, the theorem also sneaks into the equations simply because it is such a fundamental result. For instance, the Hull-White transformation-into-normality approach (see 1.4.4) claims that if you standardize returns by dividing them by their conditional variance, the CLT can be used to determine the PnL distribution (Hull & White, 1998). However, before applying this finding haphazardly to returns we need to critically reflect about its definition. Again, probing more deeply into the assumptions can give us intuition as to where potential improvements lie.

- Only for n going to infinity the mean is normally distributed. Otherwise, it is an approximation that worsens as n gets smaller.

- The returns need to be independent. If returns show some trend or predictability of any kind, this assumption is flawed, and mean returns cannot be modeled by a Gaussian bell curve. However, researchers refer to the Efficient Market Hypothesis (EMH, Fama, 1998) to use the CLT. Indeed, a benign data set of logreturns often looks Gaussian. The EMH basically states that under weak efficiency, trying to predict markets based on past data will yield no excess returns. Moreover, using standard time series analysis techniques, one can even prove that stock markets are random walks (Dickey and Fuller, 1979; Phillips and Perron, 1988). Nevertheless, the econometrist often forgets that this sort of hypothesis testing is prone to a joint hypothesis problem. The Achilles heel of the random walk hypothesis is its close relationship with elliptical distributions which, as we just argued, are unable to explain real-world price fluctuations. Testing the EMH is often done by assuming the EMH as a null hypothesis. Hence, the reliance on the testing distribution is a similar type of issue as the initial problem, namely one relies on the same (erroneous) assumptions about the asymptotic convergence of the distribution of sampling results. Common sense would suggest that the only way to test this hypothesis properly is by looking at the probability of certain events happening under these assumptions. The latter is exactly what we will do in the backtesting of different models in Chapter 4, where we will also conclude that elliptical distributions give too low probabilities to extreme events.

- The returns need to be identically distributed. The question arises whether returns over long time horizons are drawn from the same hypothetical population of potential returns. This is also questionable considering the earlier point on a static versus dynamic view on risk measures. Ideally, a risk measure should not be based on static assumptions but should be a combination of different perspectives that are combined dynamically according to the market circumstances. There are so many

economic, geopolitical,… factors that might impact this distribution over time that this assumption is clearly flawed.

- CLT gives a distribution for the mean, which means that probability statements around the mean of the distribution will be more accurate than probability statements in the tail. In other words, the reliability of defining quantiles drops with lower significance/higher *cl* statements if you support on CLT. If you make rough statements about where the majority of the price movements will be according to measured volatility, you will be approximately right. If you make precise (high confidence level) statements about extreme quantiles, you will be exactly wrong. This is the very opposite of what we need in market risk measurement.

### 1.4.3 Fisher-Tippett-Gnedenko Theorem: Generalized Extreme Value Theory (GEVT)

The *Fisher-Tippett-Gnedenko theorem* (FTGT) samples the maxima from different subsamples and defines the asymptotic distribution for the maximum of the population as the number of subsamples increases indefinitely (De Haan & Ferreira, 2007; Kotz & Nadarajah, 2000). FTGT gives a theoretically correct distribution that is needed for the risk measurement problem, i.e. maxima of the LnP function has a Generalized Extreme Value (GEV) distribution (Smith, 1990):

*Consider the sample X1, … , $X_n$ of n i.i.d. random variables with common cumulative distribution function (cdf) F. We define the ordered sample by $X_{1,n} \leq X_{2,n} \leq \ldots \leq X_{n,n} = M_n$, and we are interested in the asymptotic distribution of the maxima $M_n$ as $n \to \infty$. The distribution of $M_n$ is easy to write down, since $P(M_n \leq x) = P(X_1 \leq x,\ldots,X_n \leq x) = F_n(x)$ and has a GEV distribution which corresponds to either a Weibull, Gumbel or Fréchet distribution depending on the parameters.*

Therefore, as its name suggests, GEV focuses on extreme values instead of coming up with a distribution for the mean. Instead of being blind for LPHI events, we now explicitly use these events to determine the shape of the distribution used for VaR. This shape parameter $\xi$, in the context of the GEV distribution also called the tail index, determines whether this distribution belongs to the Weibull, Gumbel or Fréchet family. In financial applications $\xi$ is typically higher than zero but less than 0.35, therefore belonging to the Fréchet family ($\xi > 0$) which has tails obtained from power laws. Lévy, Pareto, t-distributions all belong to this family. E.g. sampling from a Student t distribution would yield a $\xi > 0$ for the maximum if n gets large. The special case is where $\xi = 0$ corresponds to a Gumbel distribution. Gumbel typically has exponential tails, which we find for normal and lognormal distributions.

The problem with GEV theory, however, is that there are only a handful of meaningful maxima available, while we need a very large number of subsamples with their respective maxima to estimate the parameters robustly. Hence, the theoretical problem is solved but GEV theory leaves us with a practical data issue. As Dowd (2007) concludes, there is no evidence that closed-form solutions delivered by GEV theory give superior risk measures, because of this practical issue. The reader can find the expressions for the Gumbel and Fréchet VaRs that were used in the code in the first section of this chapter. Weibull is not considered in this dissertation, since tails corresponding to a shape parameter $\xi < 0$ are even lighter than normal distribution and have very limited resemblance with real-life PnLs (McNeil, 1999).

### 1.4.4 Other parametric approaches

Some other parametric approaches that were introduced in academics, but which are less frequently used by practitioners are briefly discussed below. They are not included in the code for now, but their relative use in a combination model might be of interest for further research. These are inter alia (based on Dowd, 2007 ):

Table 3: An overview of other parametric approaches

| Name | Description |
|---|---|
| Lévy approaches… (Mandelbrot, Fama, 1965) | … accommodate fat tails but lack conventional closed-form solutions and have infinite variance leading to all sorts of practical issues. They are self-similar and stable. |
| Other elliptical and hyperbolic approaches… (Bauer, 2000) | … yield very normal-like VaR formulas with generalizations of the elliptical distributions to the hyperbolic family. However, they do not always yield closed-form solutions for VaR. |
| Normal Mixture approaches… (Venkataraman, 1997; Wang, 2001) | …where the return process is drawn from a mixture of normal distributions with different variances (cf. link with 1.6). |
| Jump Diffusion Models… (Merton, 1976) | …are similar to mixture approach but the extra component is not another normal with different variance but a jump variable reflecting a large market swing (cf. link with 1.6). For an introduction, see Matsuda, 2004. |
| Cornish-Fisher approximations… (Jaschke, 2001) | …use the Cornish-Fisher expansion to determine the percentiles of distributions which are near normal. |
| The Hull-White Transformation-into-normality approach… (Hull & White, 1998) | …standardize the return data by dividing by e.g. GARCH forecasts so that returns become near i.i.d. and the var-covar framework can be applied. |
| Copula approaches… (Malevergne and Sornette, 2004) | …for multi-variate VaR calculations, copula functions enable to capture the dependence structure between the variables and estimate VaR based on any underlying distribution (Gaussian copula, Student t, etc.) |

## 1.5 Non-parametric models

*"History doesn't repeat itself, but it often rhymes."*

Mark Twain

Non-parametric models do not impose strong distributional assumptions on the PnL data. Instead, we look at the past n observations and make inferences that are purely

based on the data (HS) or manipulated data, though without any necessary parametric restrictions (FHS). Three very widely used approaches are:

I. *Historical Simulation* (HS)

Historical simulation (Allen et al., 2009; Dowd, 2007; Pritsker, 2006) is the most intuitive and easiest approach to market risk quantification. We simply use empirical quantiles of the past n observations. For example, the *cl*% VaR is the 1-*cl* percentile of the past n observations. If we take 500 past observations into consideration, the 95% VaR is the 26$^{th}$ observation when the returns are ordered from smaller to larger. Alternatively:

$$cl\%\ VaR = [r_t]_{1-cl} \tag{1.9}$$

Where $r_t$ are the ordered return observations and [.] takes the 1-*cl* percentile.

II. *Filtered Historical Simulation* (FHS)

A slightly more sophisticated and widely used approach is applying the HS method above on filtered data. Filtering (Vosper et al. 2002; Brandolini & Colucci, 2012) is typically done by standardizing for volatility or by using matrix decompositions in multi-variate contexts. For instance, we estimate a GARCH process on the past n observations. Then, we subtract zero as an approximate mean return over the sample and divide by the estimated sigma. Hence, we have normalized or standardized data. Subsequently, we take the 1-*cl* percentile and multiply it with the best guess for future volatility to come up with an appropriate VaR number. The vol forecasting can be done by moving average forecasting techniques or by observing the implied volatility. By filtering the data, we somehow capture the volatility altering regimes without making strong assumptions about the distribution of the PnL:

$$cl\%\ VaR = \sigma_T[\frac{r_t}{\sigma_t}]_{1-cl} \tag{1.10}$$

where we first standardize the returns $r_t$ by dividing by the volatility $\sigma_t$, then take the (1-*cl*)N$^{th}$ observation and multiply by the volatility forecast $\sigma_T$. This forecast can be a one- or multiple-step-ahead GARCH or RiskMetrics' EWMA forecast (see Ding & Meade, 2010 for a comparison) or by back-solving the implied volatility (IV) from options. Equation 1.10 is sometimes referred to as volatility-weighted historical simulation, which is just one simple example of a filtering mechanism.

*III. Kernel functions*

Kernel functions try to fit a continuous distribution on the discrete empirical PnL without imposing a statistical form, i.e. by linking the points with linear or quadratic mathematical functions (see Butler & Schachter, 1997 for an application). Cubic splines are well-known in financial applications, e.g. for estimating a continuous yield curve out of discrete combinations of maturities and zero-coupon yields. This technique is widely used and deserves mentioning as a non-parametric approach. However, it is not implemented in the code as we will focus on (filtered) historical simulations.

As with the parametric approaches, there are many other techniques that can be qualified as non-parametric (for an overview, again see Allen et al., 2009 and Dowd, 2007). However, we will focus on HS and FHS. In line with what we will discuss in the next section, it is hard to ensure that the scope of the input measures is not too narrow nor too broad. From a feature engineering perspective (also see chapter 3), features should be informative enough, i.e. have enough variation. In our case this means that some measures have a positive bias and others a negative bias, and the more angles we include the better the model gets as long as the approaches are not perfectly correlated. However, we should be wary of not violating the principle of *parsimony* and therefore reduce the number of features to a '*sufficiently informative selection*'. That is why we pick the most tractable models that are available and leave ample room for improvements in follow-on research.

The important caveat attached to this set of approaches is that although there seems like no distribution needs to be assumed, the implicit assumption is that the future PnL distribution is sufficiently similar to the distribution of the past N observations. Therefore, if the model is applied on an economically very benign period, the model will typically lower the VaR as the probability of a new shock increases. Moreover, without manipulation, the biggest loss that the model can predict is the biggest loss in the sample, which is also a major shortcoming. That is why historical simulations are empirically found to be a more aggressive method, i.e. it tends to understate the risk (Pritsker, 2006)[13].

## 1.6 Monte Carlo Simulations

*Monte Carlo Simulations* (MCS) use the vast computing power that is available nowadays to simulate thousands of price paths and revalues the portfolio given these paths. The quantiles needed for our risk measure can then be deduced from the x% worst valuations. This description seems to imply that there is no statistical distribution that needs to be assumed. However, the mechanism driving the variation behind the sample paths assumes a certain distribution. In fact, the *randomness* of this process maps to a certain distribution (Mandelbrot & Hudson, 2010). The infinitesimal differences in prices are modeled by differential equations. The changes are thus linked to an assumed data generating process, which might have mean-reverting or trending properties. The next paragraphs zoom in on these stochastic differential equations (SDEs) and the link with these concepts. Without trying to pursue mathematical rigor, a basic understanding of the modeling choices one has with these SDEs will help the reader to better understand the link with fractals and roughness in the next chapter.

In essence, stochastic differential equations are the mathematics behind the process that drives the creation of sample paths in a Monte Carlo. The mainstay in financial applications is *Brownian motion* (BM). In 1900, it was introduced in a doctoral thesis

[13] The author further explains the dangers of under-responsiveness and so-called *ghost effects* that are typical for historical simulations. The former issue has to do with the critique given in the last paragraph, while the latter means that large changes in VaR are observed when LPHI events drop out of the simulation window.

with the promising title '*Theory of Speculation*' by Louis Bachelier (Bachelier, 1900). BM was borrowed from statistical physics and describes the random movement of a gas molecule through space. It is also referred to as the *diffusion model*. Smoke diffuses randomly from the top of a cigarette according to the same statistics as a stock price moves in a Brownian model. It is closely related to the *random walk theory* and the EMH. When a process shows no mean-reversion nor some deterministic trend, it is said to be a random walk. A random walk for infinitesimal time steps is then called a Brownian motion. A next observation is just a random deviation added to the previous observation. Moreover, it is a so-called *martingale*: the expected value for tomorrow's price is the price of today. Therefore, if $S_t$ is the stock price at time t, this is equal to $S_{t-1}$ plus some random variable. This random variable has a normal distribution with mean zero and volatility sigma. Therefore, the PnL implied by $dS_t$ has a Gaussian distribution.

A very popular model for option valuation is the Nobel prize winning Black-Scholes model (Black and Scholes, 1973). Black-Scholes assumes *geometric Brownian motion* (GBM):

$$\underbrace{dS_t}_{\text{Change in stock price S}} = \underbrace{\mu S_t dt}_{\substack{\text{Drift term:} \\ \text{proportional with} \\ \text{mean over time}}} + \underbrace{\sigma S_t dz}_{\substack{\text{Uncertainty term:} \\ \text{proportional with volatility}}} \tag{1.11}$$

In the GBM case, the logarithm of the returns (logreturns) follows a Brownian motion. In this case, our PnL is lognormally distributed. GBM is a very widely used assumption in finance to model risk and to price assets. For instance, more complex exotic options and other derivatives have no closed-form pricing formula and need to be priced through MCSs that often impose a GBM on the price process. A typical fallacy is that modelers equal unpredictability to BM. There are different levels of randomness, i.e. there are many cases other than pure mean-reversion, trending or BM as we will discuss in the next chapter.

A special property of the BM model is *that increments are proportional with the square root of time* (Velasquez, 2010):

$$dS_t \sim \sqrt{dt} * \varepsilon \tag{1.12}$$

where $\varepsilon$ has a standard normal distribution. We say that for this stochastic process the distance traveled is proportional with the ½th power of the time elapsed (Mandelbrot, 2013; Velasquez, 2010). This property is of extraordinary importance for the next chapter. It is essentially a consequence of BM having no memory. For processes without a memory, we can say that increments scale with the square root of time. This property only holds if the autocovariance between increments is zero. That is why fractional Brownian motion, which will have a generalized autocovariance function to model long memory, does not use the square root rule (see 2.6). This will be explained in detail in the next chapter. For now, it might also be interesting to look at some other often recurring SDEs in finance. We can generalize the GBM expression for a general drift and uncertainty term as a function of $S_t$:

$$dS_t = a(S_t)dt + b(S_t)dW \tag{1.13}$$

The functions a and b can be altered to fulfill the needs of the modeler. Some of the most common SDEs in finance are:

Table 4: Some common SDEs

| Name | Formula | |
|---|---|---|
| Ornstein-Uhlenbeck process | $dS_t = \kappa(\theta - S_t)dt + \gamma dW$ | (1.14) |
| Correlated Brownian motions | $dS_t = \rho dW_1 + \sqrt{1-\rho^2}dW_2$ | (1.15) |
| Merton jump diffusion | $dS_t = (r - v)S_t dt + \sigma S_t dW + (e^{Jt} - 1)S_t dN$ | (1.16) |
| Heston model | $dS_t = \mu S_t dt + \sqrt{v_t}S_t dW_2$<br>$dv_t = \kappa(\theta - v_t)dt + \xi\sqrt{v_t}dW_2$ | (1.17) |

*Ornstein-Uhlenbeck* processes are found in applications where the modeled quantity is pulled towards some long-term mean or equilibrium θ. For an introduction, see Schöbel and Zhu, 1999. The coefficient $\kappa$ is called the gravity and measures the degree to which the series is pulled towards this equilibrium. It is used for modeling bilateral correlations between assets (Meissner, 2013), interest rates (Vasicek, 1977) and many other mean-reverting processes. *Correlated Brownian motions* (Zhou, 2001) are an extension of BM where BMs with different variances are combined using their correlation $\rho$. *Merton models* (Matsuda, 2004; Merton, 1976), of which the *jump diffusion* SDE is probably the most famous one, also combines multiple normal distributions. In this case, however, the other normal is used to model discrete jumps in the process. The Ornstein-Uhlenbeck and Heston processes will provide good insights in the link with the next chapter.

The *Heston model* (Heston, 1993) replaces the general $\sigma$ volatility by the root of the variance function of time $v_t$, which has its own SDE as a model of stochastic volatility. The changes in volatility are modeled similarly to an Ornstein-Uhlenbeck process, i.e. the change is proportional to the deviation of the instantaneous variance $v_t$ and the long-run variance $\theta$. The stochastic part of this equation is proportional with the volatility itself $\sqrt{v_t}$ and the volatility of the volatility $\xi$. This volatility of volatility is introduced mainly to be able to make second order corrections in the time-varying volatility process, for example to accommodate the so-called volatility smile in the time structure of options.

As a conclusion, there are a plethora of SDEs and corresponding distributions to use. We must not forget that, although computers can start from these SDEs to simulate paths without an explicit need for the PnL distribution, for every SDE there exists a corresponding distribution. Therefore, we should remember that essentially the same fitting needs to be done and thus the same assumptions need to be made as considered under the parametric approaches. The advantage of using MCSs is that it is applicable to revalue more complex assets like exotic derivatives, when no clear formula for its value is clear, in which case numeric simulation methods are needed to determine the price. The main advantage of parametric approaches is the availability of closed-form solutions.

## 1.7 Limitations of the VaR approach

One might wonder why the previous sections are explained from the VaR perspective when the limitations of the VaR technique have been extensively reviewed in literature, e.g.:

- It is only *one quantile*; we do not know what happens *if the loss exceeds VaR*.
- It is prone to several *critical assumptions*, like we just illustrated.
- Its estimation often adopts a *static view* on risk: a *rear-mirror view* instead of forward-looking.
- VaR is *not a coherent risk measure*, since it is *not sub-additive* (cf. 1.9).

First empirical evidence on the performance of VaR models at big investment banks goes back to Berkowitz and O'Brien (2002) who show that the reported numbers were highly unsatisfactory. Their simple model, combining ARMA[14] returns with GARCH volatility of the bank's PnL, outperformed the banks' internal models. Biases and approximations at all kinds of structural levels of the bank are typically aggregated and errors worsened, leading to violation ratios which by no means can be reconciled with model integrity.

Moreover, empirical studies show that some techniques systematically underestimate risk (e.g. NVaR) and others overestimate risk (e.g. GEVT distributions) consistently out-of-sample. Inui, Kijima, & Kitano (2005) show that for most distributions with a GARCH approach to estimating volatility, an overestimating bias is present while for HS an underestimating bias is mostly the case (Inui et al., 2005). Moreover, Beder (1995) aptly illustrates how different approaches yield substantially different VaR estimates. Liu (2005) explains that the bias of different techniques often has consistently different signs, which implies that the combination of techniques might decrease the overall bias

[14] Autoregressive Moving Averages (ARMA) models are simple univariate time series models that use the autocorrelation present in a time series to explain a single quantity by looking at its own past (autoregressive part) and the disturbance terms in the past (moving average part). In this case the modeled quantity is the PnL of a bank, where the disturbances have a variance modeled by a GARCH process.

for VaR estimation. In this train of thought, a combination of VaR measures can give a more adequate risk measure.

A crucial nuance to the VaR critique is the transferability from VaR to other risk measures by, for instance, defining ES as an equally weighted average of the tail VaRs that were initially calculated with a VaR model. Therefore, let us not throw VaR in the bin but elaborate on it and define ES in the next section.

## 1.8 Expected Shortfall

As an answer to the critique that VaR only considers one quantile, Expected Shortfall (ES) – also average tail VaR or loss, Conditional VaR, CVaR,… – is the mean of the tail quantiles (Allen et al., 2009; Dowd, 2007). It describes how bad the loss is, given that things turn sour.

$$ES(cl) = E(L|L > VaR(cl)) = \frac{1}{1-cl}\int_0^{1-cl} q_p dp \tag{1.18}$$

Consequently, we can find a closed-form ES formula by integrating a chosen distribution over its tail, as has been done in literature (Andreev et al., 2005; Broda and Paolella, 2011). It is clear that, in contrast to VaR, ES does not only consider one quantile q, but calculates the mean of the VaRs in the tail. As such, a proper VaR model can be used to come up with estimates for ES. However, this alternative way of looking at bad quantiles of a hypothetical PnL does not provide us with an alternative method to calculate those quantiles practically. Therefore, we will need to resort to the same techniques as described above to calculate ES. Again, note that ES was included in the market risk reforms under the finalizations of Basel III - referred to as Basel IV - and is now the new standard for reporting market risk (BCBS III, 2017; Farag, 2017).

## 1.9 Coherent Risk Measures

Instead of building and estimating risk measures and assessing if they are doing a good job afterwards, one should first define the desired properties of such a measure. Artzner & Delbaen (1999), defined 4 such theoretical properties a risk measure should have and defined a *coherent risk measure*:

- *Sub-additivity*: the sum of the risk of positions in a portfolio may not be larger than the risk of those positions held individually.
- *Positive homogeneity*: a multiple of a certain exposure, in terms of the invested amount, should result in a multiple of the risk measure.
- *Translational invariance*: certain values in a portfolio (e.g. cash or certain future cashflows), are not at risk and should be deductible from the risk measure of a portfolio without such a certain future cashflow. E.g. if we maximally expect to lose $2m on a portfolio of bonds, but we do get $1m in cash at T without any risk, our effective risk is only $1m.
- *Monotonicity:* the risk of a position with a higher level of future value should be lower than that of a position with the same notional but a lower future value. Hence, the common link between valuation and discount rate should be implied by the risk measure.

Now that we have defined these desired properties, we can conclude that ES is coherent and VaR not. The main reason for this is that VaR is not sub-additive. If one writes deep out-of-the-money options, for example, quantiles in the tail will have a value of zero until one goes to some very extreme quantile. The result is that the VaR of a portfolio of those options will be higher than the risk of an individual option. This is because the joint probability of one such a low-probability, high-impact event happening given multiple options is higher than its individual probability. This will shift non-zero losses to lower *cls* on the PnL distribution. This is against the common principle of diversification, simply because VaR does not look further than one quantile. This also

explains how traders can game the VaR: as long as some big loss is a little bit less probable than the determined significance level, it will remain unmeasured by the VaR.

## 1.10 Spectral risk measures

*Spectral risk measures* (SRM) are a generalization of the weighting of the PnL quantiles (Dowd, 2007):

$$\text{SRM} = \int_0^1 w(p) q_p dp \tag{1.19}$$

As denoted above, one weighs the quantiles (q) of the PnL for different significance levels p according to a weight w($p$) which is a function of p. Notice that VaR is a special case where w($p$) is 0 for every value, except p = 1-$cl$ where w($p$) = 1. ES is a special case where w($p$) is 1/1-$cl$ for every value up to p=1-$cl$ and w($p$)=0 after. We say a spectral risk measure is coherent if 3 conditions are satisfied concerning the weights $w(p)$ (Acerbi, 2002):

I. $w(p)$ is positive
II. $w(p)$ is decreasing
III. $w(p)$ sums up to one over every p

Or in the words of Dowd (2007): "*The key to coherence is that a risk measure must give higher losses at least the same weight as lower losses*." In theory, these weights $w(p)$ should be derived from a subjective risk-aversion function. Thus, an optimal risk measure for each user depends on the user's risk aversion. In contrast to sigma or VaR, which is supposed to be the same for any market participant, risk according to SRM depends on the user's risk-aversion function. This fits the belief that risk is subjective and not two-dimensional, unlike the way it was presented in classical quantitative models. For instance, an exponential risk-aversion function borrowed from micro-economics (Pratt and Zeckhauser,1987),

$$w_\gamma(p) = \frac{e^{-(1-p)/\gamma}}{\gamma(1-e^{-\frac{1}{\gamma}})} \tag{1.20}$$

where $\gamma$ reflects the user's degree of risk-aversion, can be used to come up with the risk measure:

$$M_\gamma = \int_0^1 \frac{e^{-(1-p)/\gamma}}{\gamma(1-e^{-\frac{1}{\gamma}})} . q_p . dp \tag{1.21}$$

This is just some theoretical example introduced by Dowd (2007) that should familiarize the reader with the idea behind spectral risk measures. The goal of this subchapter, however, is to understand that *any risk measure can be seen as a weighted average of (extreme) quantiles of the PnL*. It is key to understand that if VaR is some extreme quantile measure, other more comprehensive risk measures can be derived from VaR. This point is very critical for the rest of the thesis. Although only parametric and non-parametric approaches to VaR are initially considered and the eventual measure is a combination of VaRs, other *more general risk measures can be calculated by reiterating the model over different values of significance*. Indeed, the outcomes of this reiteration are different losses (VaRs) with corresponding probabilities (1-*cl*) and can be seen as scenarios. These losses predicted by our final model can be seen as drawings from the real tail distribution of our PnL. Their expected value is nothing else than the ES predicted by our final model, and therefore we can say that this reiteration yields coherent risk measures. From the view of SRM, we can iterate our model over any *cl* (not only over the tail) and define a risk-aversion-based weighting function that fulfills the three above requirements. This insight is key for this dissertation, and the link between our final model and coherence will be further explained in chapter 4.

## 1.11 Risk management and allocation decisions: Portfolio VaR

How can an improved risk measure contribute to optimal capital allocation? In the long-run much criticism on modern portfolio theory (cf. 1.4.1) is not grounded, because tail events and deviations from the random walk model do not really make sense when we aggregate returns to multi-year horizons. With temporal aggregation of data, one could argue that strategic allocation decisions can be made correctly using a mean-variance framework. This means that the proportion of the different asset classes in our

portfolio, i.e. the percentage of stocks, bonds and alternative investments, can be derived from some long-run desired *Sharpe ratio*[15] (Markowitz, 1991; Sharpe, 1994). Pricing in tail events for long-term horizons might influence the fund manager's decisions badly. It will move the investor away from asset classes where Black Swan events are probable, which is suboptimal. The biggest loss on financial markets is the opportunity cost of never taking the risk that yields an excess return. Or like Buffet formulates it: in the long-run everyone wins on the financial markets, only those who do not participate lose. Or in the words of Virgil: "*Fortune sides with him who dares*."

Nevertheless, the specific stocks, bonds or alternative investments within these categories and their market timing (i.e. security selection and tactical allocation) need to be chosen more carefully than by just imposing a variance constraint and maximizing the expected return for the combination of assets within this constraint. For tactical capital allocations, i.e. the weights for specific assets within these categories, mean-variance optimization has serious bias. Other methods to determine an optimal risky portfolio like minimum variance (Clarke et al., 2011), naïve Talmudic rules (Duchin and Levy, 2009) and combination methods (Kan and Zhou, 2007; Tu and Zhou, 2011) still hinge on sigma as the predominant measure of uncertainty around the expected returns (see Frömmel, 2013 for an overview). Most of these alternative portfolio methods inherit the MPT mindset and/or do not come up with more comprehensive constraints with regard to risk, or even outright resort to rules of thumb.

As other more comprehensive measures of risk can accommodate more aspects of the underlying market risk than mere sigma, it would be interesting to use VaR or its related concepts as a basis for risk-adjusting returns, i.e. to come up with a better sense of the *reward for variability*. In other words, *what is the Sharpe ratio of an asset where we generalize sigma to any measure of risk like VaR, ES and SRM*? This framework should be consistent with our notion of diversification. The VaR of a portfolio should be lower than the VaR of the constituent assets if held individually. Furthermore, we are

[15] The Sharpe ratio compares the excess return of an asset over the risk-free rate with the variability in the asset's return, as measured by its volatility ($\sigma$).

interested in what the contribution is of one position in the total risk of the portfolio. To answer these questions, we can resort to the portfolio VaR literature[16] (Alexander and Baptista, 2002; Allen et al., 2009; Campbell et al., 2001; Hallerbach, 1999; Stoyanov et al., 2013).

A first important concept is the *individual VaR* or *standalone VaR* of a position. It is simply the VaR calculated using any of the above methods for one individual position as if held separately. We therefore neglect any correlations or comovement with the other positions in the portfolio. Secondly, the *portfolio VaR* or *diversified VaR* is the total VaR of the portfolio that fully takes into account the covariances between all the positions. One could simply calculate the portfolio VaR by applying the above VaR methods on the PnL of the portfolio, instead of using individual PnLs of the positions and taking the covariances into account.

The following three concepts try to best answer the crucial question: "*Which position should I alter to modify my portfolio VaR most effectively?*": (1) Marginal VaR, (2) Incremental VaR and (3) Component VaR.

(1) The *marginal VaR* (MVaR) of a position i is the change in portfolio VaR ($VaR_p$) due to taking an additional unit of exposure of that position $dX_i$. In theory that additional amount is infinitesimal (hence $d$), so that it corresponds with the first derivative of the VaR with respect to the position.

$$MVaR_i = \frac{dVaR_p}{dX_i} \tag{1.22}$$

(2) *Incremental VaR* (IVaR) of a position i is similar to marginal VaR but the additional exposure can now be large. It considers how an actual change a in a given position influences the portfolio VaR p.

$$IVaR_i = \Delta VaR_i = VaR_{p+a} - VaR_p \tag{1.23}$$

[16] Probably the most concise and best introduction is given in chapter 7 of Jorion, 2000.

It is computationally more burdensome than MVaR, since MVaR is some first-order approximation, while incremental VaR requires full revaluation for its accuracy. However, we could write IVaR as the following expansion (if $a$ is small):

$$VaR_{p+a} = VaR_p + a\,\frac{dVaR_p}{dX_i} + \frac{1}{2}a^2\,\frac{d^2VaR_p}{dX_i^2} + \ldots \qquad (1.24)$$

$$IVaR_i \approx a\,MVaR_i \qquad (1.25)$$

This first-order approximation approach is clearly less accurate but way faster and thus less costly.

(3) Lastly, the concept that looks at the total position i at once is called the *component VaR* (CVaR). How much does each position i, in its entirety, contribute to the total risk of the portfolio (with weights w and a total invested amount of P)?

$$CVaR_i = MVaR_i\,w_i\,P \qquad (1.26)$$

$$VaR_p = \sum_{i=1}^{N} CVaR_i \qquad (1.27)$$

Now the link with portfolio management is rather straightforward. The excess returns can easily be compared with the obtained MVaRs. When a position contributes a lot of risk to the portfolio and does not deliver appropriate excess returns ER, the fund manager can start to sell off that position so that the portfolio VaR drops significantly while the expected returns do not. Now the asset manager can use the obtained funds to reinvest in assets with lower VaRs and comparable or higher returns. In an optimal scenario, ER/MVaR ratios are somewhat smoothened out over the portfolio as to increase its efficiency. De facto, this corresponds to a mean-VaR constraint over the portfolio (Alexander & Baptista, 2002). The same authors show that these algorithms yield significantly different conclusions than mean-variance constraints. Moreover, Stoyanov et al., 2013, argue that the return characteristics are not necessarily sensitive to the MVaR values, in contrast to the mean-variance trade-off proposed by MPT. This

sort of analysis can also shine a new light on the risk of the constituent assets of the portfolio by breaking down portfolio VaR (Hallerbach, 1999).
In brief, this section first provided some intuition behind how risk measurement can trickle down to portfolio management decisions. This link will be explained in further detail in 4.2.

## 1.12 Conclusion

To end this first chapter, some concluding remarks need to be made. Hopefully, the text was able to separate the math on the one hand – as quantitative methods for risk measurement is a vast field of research – and the intuition and main takeaways that are relevant for the research question on the other hand. They could be summarized by the following propositions:

- *No matter how promising new risk measures like ES and SRM appear compared to VaR, they support on similar assumptions in the way that they are estimated*. In other words, these practical approaches support on the same essential assumptions, leading to similar shortcomings. The recently adopted ES captures a more comprehensive sense of risk, i.e. it takes into account the tail events of the PnL. Nevertheless, it does not describe a better way to estimate them practically.

- *However sophisticated the treatment of vol or the statistical process assumed, it is hard to ensure a priori that the eventual risk measure will not be overconservative or overaggressive.* The first problem concerns underutilization of capital, or a misallocation of means because value-creating investments are not made from a risk perspective. Industry adoption of these models would result in slower economic growth. Again, the worst loss on financial markets is the opportunity cost of not taking risk. The second problem concerns taking exposures too aggressively, which inevitably leads to increased and unsustainable levels of leverage. These aggressive models are very vulnerable to exogenous shocks (referred to as Black Swan events, correlation breakdowns, volatility breakouts,… also see 3.3).

Because of these remarks, this dissertation starts from the point of view that *only a combination of techniques can give satisfying results*. One could for example calculate risk measures under different assumptions, report them in a table and discuss their meaning and limitations according to the prevalent economic 'context'. Alternatively, one could calculate VaR under different assumptions and choose the model that fits the purpose (regulatory reporting, portfolio optimization etc.) best. However, this dissertation proposes that *one can also combine these different outputs quantitatively into a more comprehensive risk measure.* The goal is to get a result of which the overall bias is lower, i.e. by combining alternative assumptions dynamically according to the 'context'. This contextual parameter could be modeled by any informative quantity linked to the models' underlying assumptions like fundamentals of the underlying asset ('micro-signals'), 'macro-signals' on the business cycle, Twitter sentiment data on the underlying and so and so forth. The practical way of combining these methods that is proposed in this research is based on *roughness* and is discussed in the following chapters on fractal geometry and neural networks.

# Chapter 2

# What is roughness? On fractional dimensions, Hurst exponents and fractional Brownian motions

## 2.1 Predicting predictability, a coastline analogy

**Fractal**: any of various extremely irregular curves or shapes for which any suitably chosen part is similar in shape to a given larger or smaller part when magnified or reduced to the same size. – Merriam-Webster Dictionary

**Fractal**: a geometrical or physical structure having an irregular or fragmented shape at all scales of measurement between a greatest and smallest scale such that certain mathematical or physical properties of the structure behave as if the [fractal] dimensions of the structure are greater than the spatial dimensions. – Dictionary.com

Benoît B. Mandelbrot (1924-2010) was the father of fractal mathematics. He was born in Poland, grew up in France and worked mainly in the USA at IBM, Harvard and Yale. He was an extraordinary mathematician and a textbook example of a '*polymath*' who had broad interests in the practical sciences. He especially contributed in those fields where what he coined as '*the art of roughness*' and '*the uncontrolled element in life*' had practical implications (Nathan, 2015).

Fractal geometry is essentially the *art of roughness*. It is, in contrast to the perfect shapes of Euclidean geometry, the study of the irregularities in nature. Rough clouds, ragged surfaces and other common shapes in nature can hardly be described by perfect triangles, squares and other Euclidean building blocks. Mandelbrot therefore preferred the term roughness before irregular, since nature is not smooth, and roughness is very regular. Recall the very first quote in this dissertation: "*Bottomless wonders spring from simple rules… which are repeated without end.*" In one of his last speeches, Mandelbrot concluded with this concise but very powerful statement. Extremely complex things are often a product of very simple rules, which are repeated to such an extent that the

system looks extremely complex from the outside. Once you start to iterate very simple mathematical expressions, for instance, one can find sets of numbers of infinite complexity. The Mandelbrot set is probably the most well-known set of this kind, whose almost psychedelic images have become famous[17]. Similarly, the SDEs of the previous chapter try to reproduce such a complex, chaotic system like a stock market, where dynamics are captured by 2-inch equations. These are essentially very simple rules, albeit stochastic instead of deterministic. Important to understand is that fractals are way more than mathematical constructions. Fractals are everywhere in nature: from the bronchi in our lungs to the branches of trees, from the roughness of clouds to the irregular shapes of coast lines.

In his 1967 paper '*How long is the coast of Britain? Statistical self-similarity and fractional dimension*' (Mandelbrot, 1967), Mandelbrot introduced the concept of fractional dimensions as a good measure of roughness. By a fluke, he discovered that particular cases of power laws, once discovered by a mathematician named Felix Hausdorff, are applicable to measure the roughness of surfaces that are no perfect Euclidean shapes (Mandelbrot, 2010). The paradox in the paper concerned the measurement of Britain's coast line. Huge disagreements existed between different researchers on how long it was, known as the *coastline paradox*.


Figure 1: How long is the coast of Britain? Answer: vastly measure-dependent. Source: simulations based on Wolfram code

[17] Psychedelic in the sense that it is really a sensory overload. I highly recommend to Google search '*Mandelbrot Set*' or '*Mandelbrot Zoom*' and experience it yourself.

Building on the work of L.F. Richardson, Mandelbrot showed that the unit of measurement greatly influenced the eventual length, since the smaller the unit, the more irregularities are taken into account and the longer the eventual measured distance. Figure 1 aptly illustrates this effect, which is referred to as *the Richardson effect*. Consequently, there is no such thing as a 'real' coast length. Imagine that one would be able to scale down to the size of a grain of sand, then one could still find more irregularities, driving the length of the coast line to an enormous distance. The total length will not converge to some 'real' length, as one would intuitively assume. Indeed, the very intuitive notion of the length of something can be a complete fallacy, when that something does not adhere to the laws of Euclid. For these real-world shapes, Mandelbrot showed that a power law exists between the scale and the length, which leads to a fixed exponent: *the roughness*.

Consider a scale factor, l, a number N as a function of l, which will denote the number of smaller parts needed to replace the initial larger part N(l) and the dimension D. We scale down by dividing the initial line by l in the one-dimensional case. For the two- and three-dimensional case, we consider squares and cubes with a basis l (see Figure 2 below).

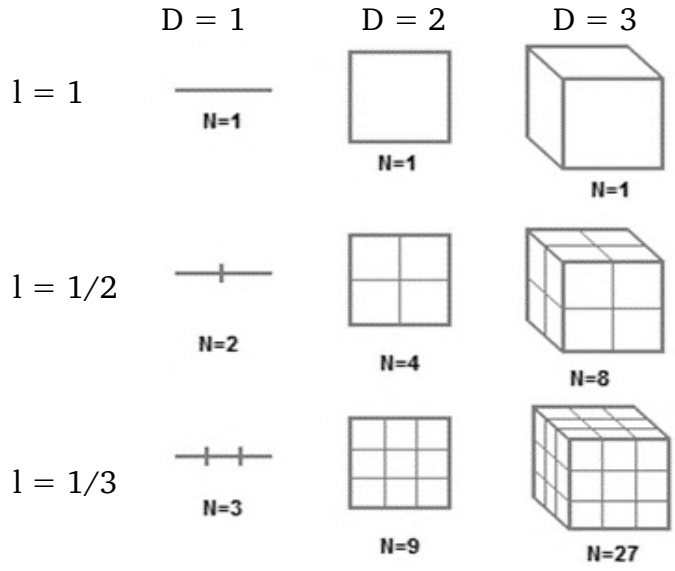


Figure 2: Fractional dimensions. Source: adapted from Ryan (2007), based on the work of Felix Hausdorff (Hausdorff, 1919)

We need 2 (= N) lines of length 1/2 (=l) to replace the initial line. We need 4 squares and 8 cubes to do the same in the 2- and 3-dimensional cases. Similarly, we need 3 lines of length 1/3 to replace the initial 1-dimensional shape. We require 9 squares and 27 cubes to extend to the higher dimensions. In general, we need $(1/l)^D$ shapes. We clearly notice a power law between the number of smaller scaled D-dimensional structures and 1/l, with the dimension D as the fixed exponent:

$$N(l) = (\frac{1}{l})^D \tag{2.1}$$

Using logarithms, we can isolate the dimension D:

$$\log\big(N(l)\big) = D \log\left(\frac{1}{l}\right) \tag{2.2}$$

Furthermore, if we generalize D to both integer and fractional values, we can define the *fractional dimension* as:

$$D = \frac{\log\big(N(l)\big)}{\log\left(\frac{1}{l}\right)} \tag{2.3}$$

where D is called the *Hausdorff dimension*[18]. For 'pure' fractals, shapes that have scale-invariant complexity, this fractional dimension is an exact fit in the log-log plot and referred to as *fractal dimension*. For real-world shapes, a similar power law rationale can be applied, but we generally find an approximating fit in the log-log plot (cf. infra). Fractional dimensions are thus a generalization of our common perception of dimensions to non-integer values, where we always focus on the number of smaller pieces required to reconfigure a larger piece. Intuitively, we could say that fractional dimensions measure how detail or complexity behaves at different scales. Practically, as we will soon demonstrate, this means roughness.

[18] The Hausdorff dimension should not be confused with the Minkowski–Bouligand dimension, often called the box-counting dimension, which takes a limit of a similar expression and is equivalent for most fractals. For illustration purposes, we pick the simplest version here as to give some intuition behind the ideas, not to be mathematically rigorous.

When we apply a similar power law rationale to coast lines, we find what Mandelbrot discovered in 1967. The size of the ruler used to measure the coast length can be compared with l, while N(l) now corresponds to the total coast length. Mandelbrot pointed out that we cannot simply quote one measured length because it is context dependent, i.e. depending on the unit of measurement. This means that things we measure – thus any dimension - of which our common sense would suggest that it is an absolute dimension - like coast lines or, indeed, risk - are in fact a relative dimension, i.e. dependent on context. Questions like '*how do I measure it*?' and '*is this yardstick appropriate?*' are often more important than the eventual outcome. First, we need to assess the roughness of the coast line, or how detail changes with scale, and then we can draw conclusions on the appropriateness of the yardstick. For instance, small scales and rough borders will inflate distances compared to other measurements of larger scale, smoother coast lines. Thus, before comparing apples with oranges, one should consider both *how you measure* something and *how rough* that something is you measure, i.e. how much you deviate from a perfect Euclidean shape.

These last statements allow us to make the link with financial risk management. In market risk measurement, one should also consider both what yardstick you use to measure the risk and to what extent the assumptions behind that yardstick are consistent with the roughness of the real process. From the previous chapter we know that it is very hazardous to unleash Gaussian models on very rough markets. Now we can say that this would be tantamount to trying to measure the circumference of a cloud using straight rulers only. Nature cannot be captured by smooth shapes only, nor can markets. You are doomed to neglect irregularities, which are – ironically - more important for our problem than the regularities.

Since risk is a latent variable, there is no such thing as the 'real' VaR. Similar to the length of coast lines, defining it as some absolute number is a complete fallacy. It is therefore very hard to claim more precision by mathematically enhancing new VaR models – recall the comparison of GEV with CLT or the extensions on common SDEs - as it is often perceived 'overengineered' or too technical to use practically. An alternative, however, is to challenge existing models with respect to their assumed

roughness. We could measure to what extent the assumed roughness is consistent with the measured roughness and give these different yardsticks weights accordingly in a combination model. In other words, we are not trying to reinvent the wheel but to combine different existing methods according to their appropriateness.

Obviously, the analogy that thus runs through this dissertation as a common thread is about measuring roughness and linking it to yardsticks. It is not about predicting what the markets are going to do, since that would imply having a crystal ball and that would make this dissertation not worth the paper it is written on. The purpose of the model should not be confused with *technical analysis* (in its strict sense, i.e. predicting future prices based on past prices). However, the goal of applying fractal dimensions and Hurst exponents in this thesis, is to assess the different roughnesses of stocks and indices to predict the adequacy of the models used, based on the link between roughness and the assumptions made in the standard risk models. Providing intuition towards this link and elaborating on its implications is thus the main goal of this chapter.

In summary, measuring roughness is about *predicting predictability*. In line with the analogy, we are not trying to come up with spuriously precise coast lengths, but with a sense of how accurate estimates are likely to be, given the measured roughness and the link with the unit of measurement that was used. We try to predict how appropriate a risk measure (‘*predictability of risk*’) will be under a certain *context* (‘*the roughness price process*’). If this analogy holds, roughness can be used to *more effectively combine risk measure models*. As will be pointed out in 2.6, fractional Brownian motion as a generalization of BM can serve as a link between the previously discussed methods and thus qualifies as a simple and logical but powerful connector in a combination model.

As a final remark, similar to the conclusion in Mandelbrot and Hudson (2010), we realize that since we cannot better predict markets using fractal theoretical properties, this will not bring us fortune, but it can save us a lot of money if it gives us insight in the *model risk* we are taking. The research hypothesis in this dissertation borrows from this intuition and states that the deviation from the standard assumptions, as measured by the roughness, can be used to give weights to models with different biases.

### 2.1.1 Fractional dimension: Algorithms

For the calculation of D for financial time series, different theoretically equivalent algorithms exist: Higuchi, Katz, box-counting methods and so and so forth. For our purpose, all these algorithms basically try to measure to what extent one-dimensional data starts to fill the two-dimensional plane.

We could say that a perfect Euclidean line does not fill the plane at all. From the moment more detail appears on smaller timescales, it is just as if the line starts to fill the plane. 'Just as if', since this only really applies for fractal structures with infinite complexity. In the words of Mandelbrot: "*Mathematicians thought a curve was a curve, a plane was a plane and the two don't mix. Well, they do mix.*"

However, this intuition is also applicable to shapes that are no pure fractals, like financial time series that show resemblance with statistical self-similarity, albeit not perfect. Consequently, very smooth stock charts will have a fractional dimension higher than but close to 1 (a straight line between the first and the last trading day). Very rough charts that contain a lot of detail, on the other hand, have a fractional dimension lower than but close to 2 (a plane), because they look like they cover a lot of the plane. Figure 3 below shows Python simulations for extreme values of D (1.99, 1.5 and 1.01) respectively.

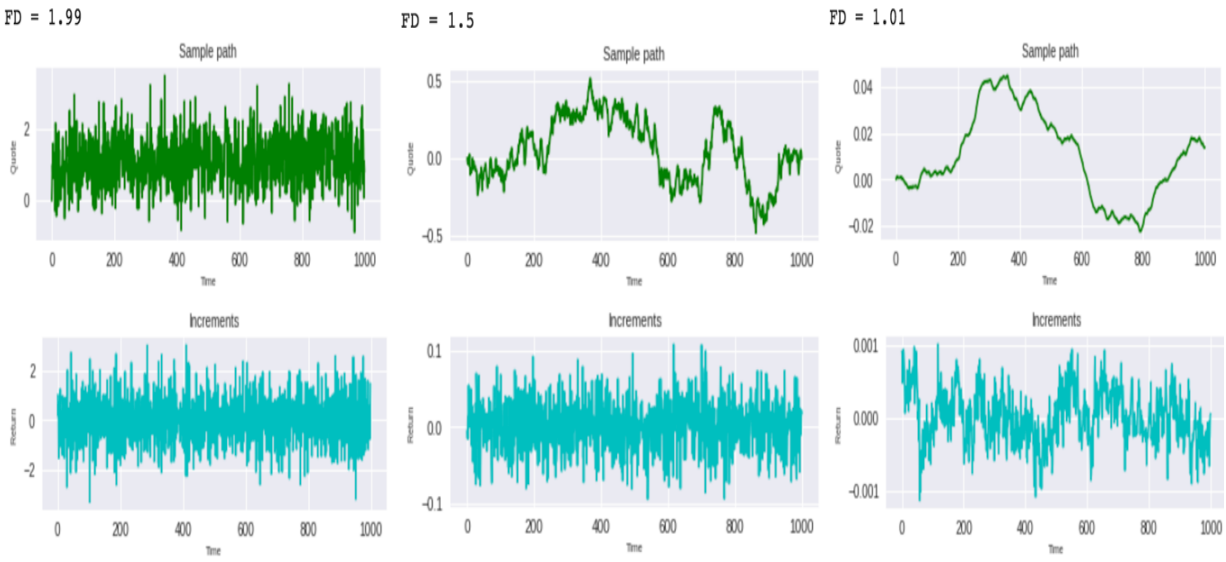


Figure 3: A Python simulation of sample quote paths for extreme values of D

We consider two of the most widely used algorithms for time series data in literature: Higuchi's and Katz' algorithm. Higuchi's algorithm is used, amongst other applications, in medical research to assess EEG diagrams, for example to distinguish between different modes of sleep. Research can be found where Katz' algo is used for ECG diagrams to distinguish between healthy heart sequences versus disorders. There are many other applications in geology, biology, et cetera.

#### 2.1.1.1 Higuchi's algorithm

*Higuchi's algorithm* basically measures Euclidean distances between consecutive observations for different timescales and compares the average distance with the scale. The description is borrowed from Cervantes-De la Torre in the Journal of Physics (Cervantes-De la Torre et al., 2013):

*In order to obtain the fractal dimension D, Higuchi considered a finite set of observations, taken at a regular interval:*

$$X(1),\ X(2),\ X(3),\ \dots, X(N) \qquad (2.4)$$

*From this series, a new one* $X_k^m$ *, must be constructed, which is defined as follows:*

$$X_k^m;\ X(m),\ X(m+k),\ X(m+2k),\ \dots, X\left(m+\left[\frac{N-m}{k}\right]k\right) \qquad (2.5)$$

*with (m = 1, 2,. . . , k); and where [·] denotes the Gauss notation, that is the bigger integer, and both k and m are integers. m and k indicate the initial time and the interval time, respectively.*

*For a time interval equal to k, one gets k sets of new time series. For example, for k = 4 and N = 100, four new time series are obtained:*

$$\begin{aligned} X_4^1&:\ X(1),\ X(5),\ X(9),\ \dots, X(97) \\ X_4^2&:\ X(2),\ X(6),\ X(10),\ \dots, X(98) \\ X_4^3&:\ X(3),\ X(7),\ X(11),\ \dots, X(99) \\ X_4^4&:\ X(4),\ X(8),\ X(12),\ \dots, X(100) \end{aligned} \qquad (2.6)$$

*Higuchi defines the length of the curve associated to each time series, $X_k^m$, as follows:*

$$L_m(k) = \frac{1}{k}\left(\sum_{i=1}^{\left[\frac{N-m}{k}\right]} \left(X(m+ik) - X(m+(i-1)k)\right)\right)\left(\frac{N-1}{\left[\frac{N-m}{k}\right]k}\right)$$

*where the term $\frac{N-1}{\left[\frac{N-m}{k}\right]k}$ is a normalization factor. Higuchi takes the average value $\langle L(k)\rangle$ of the k lengths associated to the time series given by the previous formula for $L_m(k)$. If the average value follows a power law:*

$$\langle L(k)\rangle \propto k^{-D}$$

*then the curve has a fractal dimension D.*

#### 2.1.1.2 Katz' algorithm

*Katz* provides a slightly different approach to measuring the distances. This description was borrowed from '*A Comparison of Waveform Fractal Dimension Algorithms*' (Esteller et al., 2001):

*Katz' D can be defined as:*

$$D = \frac{log(L)}{log(d)} \qquad (2.9)$$

*where L is the total length of the curve or sum of distances between successive points, and d is the diameter estimated as the distance between the first point of the sequence and the point of the sequence that provides the farthest distance.*

*Mathematically, d can be expressed as:*

$$d = max(distance(1, i)) \tag{2.10}$$

*Considering the distance between each point of the sequence and the first, point i is the one that maximizes the distance with respect to the first point.*

*The D compares the actual number of units that compose a curve with the minimum number of units required to reproduce a pattern of the same spatial extent. Ds computed in this fashion depend upon the measurement units used. If the units are different, then so are the Ds. Katz's approach solves this problem by creating a general unit or yardstick: the average step or average distance between successive points* $\underline{a}$.

*Normalizing distances by this average results in:*

$$D = \frac{log_{10}\left(\frac{L}{\underline{a}}\right)}{log_{10}\left(\frac{d}{\underline{a}}\right)} \tag{2.11}$$

*Defining n as the number of steps in the curve, then n = L/*$\underline{a}$*, and the above equation can be written as:*

$$D = \frac{log(n)}{log\left(\frac{d}{l}\right) + log(n)} \tag{2.12}$$

*This expression summarizes Katz's approach to calculate the fractional dimension.*

These algorithms are fairly easy to implement in Python[19] since there is already a lot of sample code available on GitHub, so that it is just a matter of finetuning the code to the requirements of the application.

[19] If you might have any questions on how these are implemented, again, please take a look at emiellemahieu/AOR on GitHub. The code is pretty self-explanatory and well-documented.

## 2.2 An introduction to (anti-)persistence in econometrics: mean reversion, trends and random walks

In chapter 1, we briefly introduced the concepts of mean reversion, trends and random walks in section 1.6. In order to bridge the gap between the concepts of roughness and the methods of the previous chapter, we need a sidetrack to mainstream econometrics. In simple applications[20], *mean reversion* is often quantified with a degree of mean reversion using a process that goes back to the Ornstein-Uhlenbeck SDE mentioned earlier (Meissner, 2013):

$$S_t - S_{t-1} = a(\mu_S - S_{t-1})\Delta t + \sigma_S \varepsilon \sqrt{\Delta t} \tag{2.13}$$

where the change in price is proportional with the difference of the previous price and the long-term mean $\mu_S$. This expression is no different from a discrete version of the SDE in 1.6. The $a$ coefficient denotes the *degree of mean reversion*, the mean reversion rate or *gravity*. $\sigma_S \varepsilon \sqrt{\Delta t}$ is the stochasticity part where $\sigma_S$ denotes the volatility in stock price S and $\varepsilon$ a random drawing, typically (but not necessarily) drawn from a standard normal, i.e. $\varepsilon(t) \sim N(0,1)$. A simple and widely used model for mean reversion, for instance to predict interest rates based on some long-term mean, is one without the stochasticity term:

$$S_t - S_{t-1} = a(\mu_S - S_{t-1})\Delta t \tag{2.14}$$

If we look at daily returns with the unit of t in days, then $\Delta t = 1$

$$S_t - S_{t-1} = a\mu_S - aS_{t-1} \tag{2.15}$$

Which can be easily estimated with a simple linear regression $Y = \hat{b} + \hat{a}X$

$$\underbrace{S_t - S_{t-1}}_{Y} = \underbrace{a\mu_S}_{\hat{b}} - \underbrace{a}_{\hat{a}}\underbrace{S_{t-1}}_{X} \tag{2.16}$$

[20] Meissner (2013) uses this process to measure the anti-persistence of correlation between an individual stock and an index, i.e. the mean reversion of this $\rho$ around its long-term mean. Of course, one could take a different angle to measure the same thing, but the description he provides is very intuitive.

After obtaining estimates for $Y = \hat{b} + \hat{a}X$, we can obtain the degree of mean reversion - $\hat{a}$, and the long-term mean $\hat{b}/\hat{a}$ . For instance, if $\hat{a}$ = -0,50 we close the gap between an observation and the long-term mean with, on average, 50% of that distance towards that LT mean over the next time period. For example, if the long-term mean of inflation is 2% and we are currently at 1%, a rational estimate of the quantity at the next time step (e.g. next year) would be 1,50% or 50% percent of the difference closer to the long-run equilibrium. Thus, significant gravity or the anti-persistent tendency to mean-revert causes some degree of predictability in market prices or rates.

The opposite property of mean reversion is the *autocorrelation* ($\rho$), which can be measured as:

$$\rho = \frac{cov(S_t, S_{t-1})}{\sigma(S_t).\sigma(S_{t-1})} \tag{2.17}$$

Autocorrelation is the reverse property to mean reversion. It is the most common measure of persistence, since it is simply the Pearson correlation between subsequent observations. Note that $\rho$ sums up to one with the degree of mean reversion: $\rho$ + a = 1[21]. Hence, quantities with high autocorrelation will have the tendency to 'trend' and low tendency to mean revert, i.e. a next observation will be some increment added or subtracted to the previous observation in line with its past, rather than be pulled towards the long-term mean.

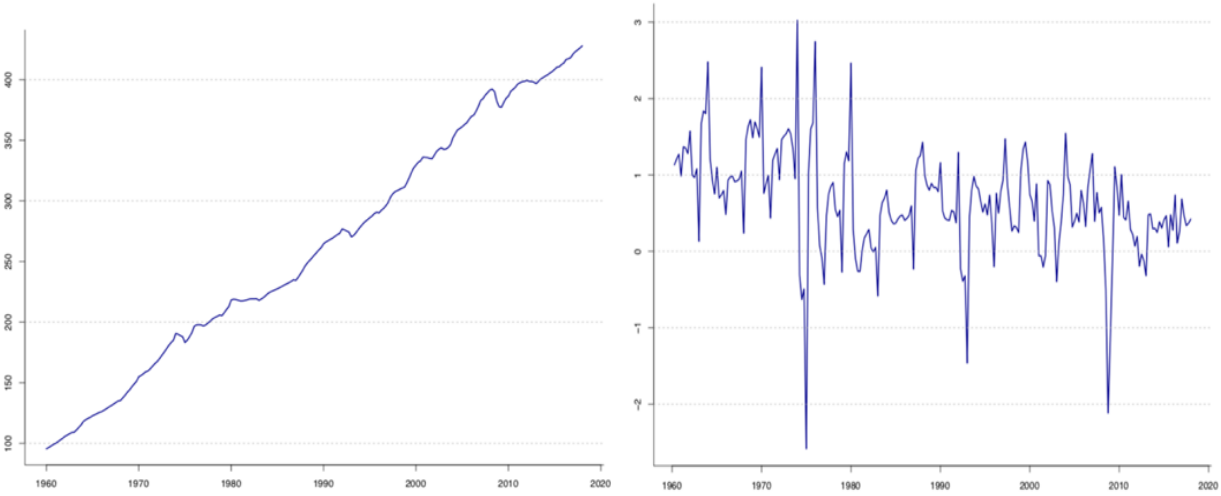


Figure 4: Three types of predictability (1): Trending Belgian GDP (left), mean reverting Belgian Output Growth (right)

[21] Remark that we estimate $\rho$ and a from the levels of S and not the returns. The three types of predictability then would correspond to positive, negative or no autocorrelation respectively.

Roughly speaking, we can distinguish between three types of predictability: trending series, mean reverting series and random walks. Real-world examples of these types of predictability are depicted below in Figure 4 and 5.


Figure 5: Three types of predictability (2): Random Walk BEL20 Index
Source: Econometrics, Time Series Analysis, Everaert G. (2019).

These sample time series all have some sort of predictability. If you put your hand on the right-hand-side of a graph of your choice and have to predict what will happen you will probably (1) continue going up a little every year for the Belgian GDP, (2) apply the above mean-reversion principle intuitively and say that if the growth in GDP is higher this year than usually, it is most likely to drop and vice versa and (3) that you do not really know what the BEL20 is going to do. The BEL20 is sloping upwards since 2014 but one can see that at previous times of higher growth (e.g. in the beginning of 2015) this was not persistent in the second half of 2015.

We could state that a random walk is a special case of a trend with perfect autocorrelation[22] and no memory. Does this mean that the degree of mean reversion is zero? What does this imply for the long-term mean in Equation 2.13? It is clear that this is a special case. Although beyond the scope of this section, the random walk corresponds to a *unit root* case in financial time series analysis. In such a setting, the mean is continuously changing and consequently the long-term mean of the BEL20 has

[22] Again, note there is perfect correlation in the prices but no autocorrelation in the returns. No memory thus implies that there is perfect correlation between $S_t$ and the one-period lag $S_{t-1}$, since all the information is reflected in the last price and all the prices before $S_{t-1}$ do not matter.

no meaning. The only relevant piece of information is the price today, since the mean of past prices is changing every day. This is called *non-stationarity*.

Although very intuitive and widely used[23], measuring this level of mean reversion or autocorrelation like described above is a simple heuristic for the predictability of time series. It is a linear approach of analysis, which is too strong an assumption to work fine in finance. It makes assumptions on linear correlations (e.g. the $\rho$ calculation and the initial linear regression) between consecutive observations. Additionally, this process is not able to capture the *long memory*[24] of the time series in contrast to Hurst exponents. That is why we move away from these linear concepts and delve into R/S analysis in the next section.

As a concluding remark, it has to be said that the above description is a very reductive view on modern time series econometrics. Of course, things are more complicated than depicted above, but it nicely represents the intuition needed for the next section. In a proper analysis, autocorrelation would give rise to an autoregressive model (ARMA) in a univariate case and ADL or VAR models for multivariate analyses. The formal tests for unit roots are called (Augmented) Dickey-Fuller tests (Dickey and Fuller, 1979; Phillips and Perron, 1988) and enable us to distinguish deterministic trends and random walks with or without drift from stationary series. Additionally, important relationships between different quantities can complicate the analysis. For instance, cointegration relationships and the multi-dimensionality of the data (e.g. a panel structure) have to be tested for and taken into account. In brief, this section did not introduce common time series models but looked at some of the basic concepts under their bonnet to make the link with R/S analysis in the next section.

---

[23] A substantial part of the research w.r.t. *technical analysis* is essentially based on autocorrelation of stock prices - e.g. momentum analysis, price reversals, moving averages (ARMA) models and patterns, etc. The research favoring the efficient market hypothesis (cf. 1.4.2) then mostly comes down to providing evidence against significant autocorrelation in return series (i.e. a random walk) using the tests mentioned above.

[24] Remark that researchers have tried to integrate this long memory in these autocorrelation-based models (ARIMA models) using the concept of fractional orders of integration in ARFIMA models (Granger and Joyeux, 1980; Hosking, 1984). These are, in fact, the discrete versions of their continuous counterpart fBM (cf. infra). Although beyond the scope of this dissertation, it can be shown that this fractional order of integration is closely linked to H.

## 2.3 Hurst exponents: rescaled-range analysis and long memory

In Egypt, Britain's former colony, hydrologist Harold Edwin Hurst (1880-1978) was tasked to come up with an answer to the following critical question: "*What are the ideal dimensions of a dam in the Nile?"* Hurst studied numerous time series on water levels (Hurst, 1952) in order to strike a good balance for the following trade-off. If you build it too high, you waste huge amounts of resources; whereas if you build it too low, flooding can lead to huge human and economic disasters. For an anecdotic account of the story, see Mandelbrot and Hudson (2010).

Whilst doing extensive quantitative analysis, Hurst recognized a lot of variability in the levels, where years of large changes of either sign - extreme droughts or flooding - tended to follow each other. The problem boiled down to how one could model this variability and come up with measures to describe the unpredictability. What is the level of persistence or anti-persistence in the water level data? This was very relevant for the dam issue since the level of anti-persistence or unpredictability in the water levels directly relates to the need for 'overdimensioning' or safety buffers in terms of the dam's dimensionality (Hurst, 1956). The reader will probably recognize that the previous problem is a similar trade-off to market risk and capital requirements at financial institutions, with the water levels as PnL and the dam's dimensions as a capital buffer. Borrowing from Hurst's work, how can we model the persistence versus anti-persistence of financial time series?

The answer is *rescaled-range analysis* (or R/S analysis). It was initially developed by Hurst and later rediscovered by Mandelbrot (Mandelbrot, 2002). It tries to come up with a measure for the persistence of a time series in another way then just regressing the water level on lagged values ($\rho$). R/S analysis uses a power law instead: *how does the rescaled range behave for different scales*?

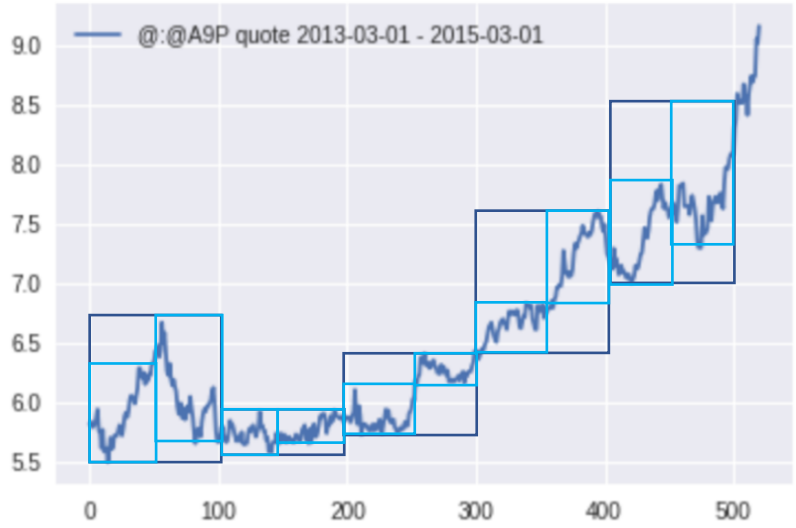


Figure 6: Boxes try to capture the behavior of the ranges for different timescales of the A9P stock

Figure 6 illustrates the idea. Say N is the size of the sample, and n is the size of the intervals at smaller timescales. We first calculate the mean for every interval:

$$mean_i(n) = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{for all k} = \lfloor N/n \rfloor \text{ intervals} \quad (2.18)$$

Then, we calculate the deviations $Y_i$ as a mean-adjusted time series:

$$Y_i(n) = X_i - mean_i(n) \quad (2.19)$$

Next, we sum the deviations as to get a series of cumulative deviation:

$$y_t(n) = \sum_{i=1}^{t} Y_i(n) \qquad \text{for j} = 1,\ldots,n \quad (2.20)$$

The range of the interval is then defined as the widest difference in the series of deviations:

$$R_i(n) = \max\left(y_1(n), y_2(n), \ldots, y_n(n)\right) - \min\left(y_1(n), y_2(n), \ldots, y_n(n)\right) \quad (2.21)$$

The scale of each of the intervals is then defined as their standard deviations:

$$S_i(n) = \sqrt{\frac{1}{t}\sum_{t=1}^{T}(X_t - mean_i(n))^2} \quad (2.22)$$

for every interval of size n.

The rescaled range of an interval is defined as:

$$R_i(n)/\ S_i(n) \quad (2.23)$$

The total rescaled range corresponding to an interval size of n is:

$$R/S\ (n) = \frac{1}{k}\sum_{l=1}^{k}\frac{R_i(n)}{S_i(n)} \quad with\ k = \lfloor \mathrm{N/n} \rfloor \quad (2.24)$$

Now, the *Hurst exponent* (H) is derived from the following power law:

$$H = \frac{\log\left(R/S(n)\right)}{\log\left(n\right)} \quad (2.25)$$

Practically, H is therefore estimated from a linear regression fit in the log-log plot[25] of R/S for different values for n.

Notice the similarities with fractional dimensions: the timescale relates to the rescaled range through a power law with a fixed exponent H. Instead of comparing the covered distance with the time elapsed, we compare the changing variance with the smaller and smaller time steps. Hence, we expect an intimate relationship for self-affine time series where fractional dimensions need not to be approximated. Notice that even if the underlying process is not a pure fractal, like stock prices, there is no need for an algorithm that approximates H, since H can be calculated directly using its definition. However, the fit in the log-log plot will then not be perfect, which suggest that stock markets are never perfectly self-similar. Again, this power law rationale is relatively

[25] Remark the similarities with the previous sections. To quote Jim Gatheral: "*It's the one thing we always get in econophysics papers: straight lines on a log-log plot*." (Gatheral, 2017).

easy to implement in Python and some sample code is available on GitHub. An example of an estimation of H for the US Finance performance index (H=0.4295) is given in Figure 7.

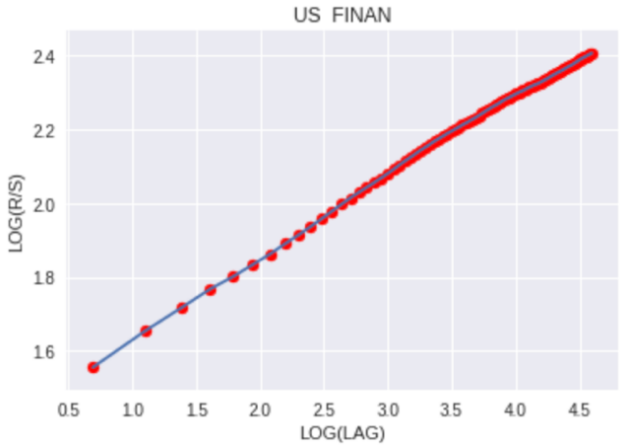


Figure 7: An R/S analysis for the US Finance performance index (H=0.4295)

Mandelbrot proved that for self-affine processes, the local properties are reflected in the global ones, resulting in the celebrated relationship D+H=n+1 between fractal dimension D and Hurst exponent H for a self-affine surface in n-dimensional space (Mandelbrot, 1985). Therefore, for one-dimensional financial time series (e.g. the BEL20 index over time) with self-affine statistical properties we can say H ≈ 2 – D.

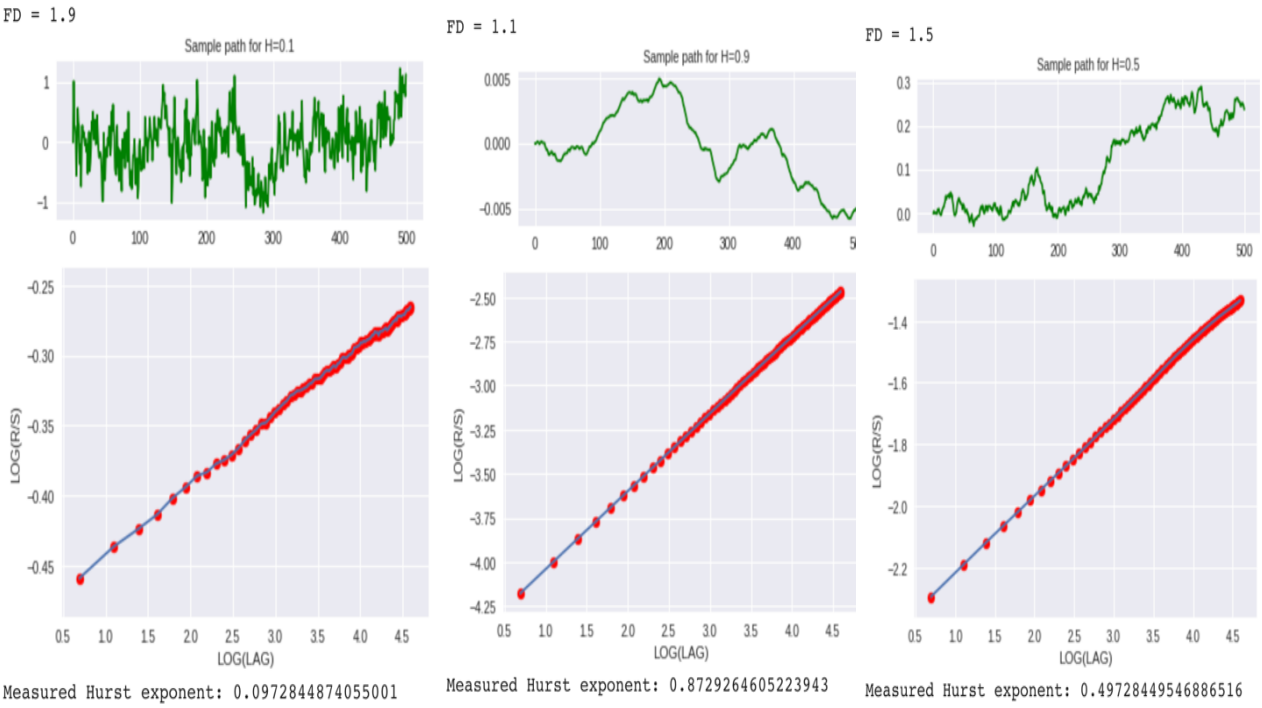


Figure 8: The estimation of H for processes with D equal to 1.9, 1.1 and 1.5 (BM) respectively

The above simulations (see Fig. 8) of D=1.9, D=1.1 and D=1.5 sample data in Python yields some pretty robust estimates of H in terms of the H = 2-D relationship.
So far, we have not really delved into the meaning of H. The box below explains the most important implications of this exponent. Since one-dimensional data can have a fractional dimension between 1 and 2, a Hurst exponent is a number ranging from 0 to 1, with the following characteristics:

***R/S and persistence***

| | |
|---|---|
| ***0 < H < 0,5*** | ***Anti-persistent series**:* *A rough, 'wildly random' series with mean-reversion features and negative autocorrelation in returns* |
| ***H = 0,5*** | ***Brownian motion**:* *Archetypical assumption in finance, no correlation with past returns (independence) and a martingale (expected deviation is zero with normal noise)* |
| ***0,5 < H < 1*** | ***Persistent series**:* *Trending behavior, smoother series or 'milder randomness' and positive autocorrelation in returns* |

As we discussed before and is displayed in the box above, Brownian motion is the case where the level of mean reversion and autocorrelation level each other out. There is no predictability in terms of a tendency to move towards a long-term equilibrium mean, nor is there a tendency to trend[26]. This corresponds with the H=0.5 case. This case corresponds to the panel on the right-hand-side in Figure 8[27].

The careful reader now probably wonders why we are in need of H if, according to the box, H exactly models these mean reverting (negatively correlated returns) or trending

[26] However, there might be a stochastic trend or drift, but this is by definition stochastic and not deterministic. This means that, similar to the BEL20 example in Fig. 5, this tendency might break down at any point. There is no *expected* autocorrelation in returns, but by pure chance (hence *stochastic*) there might be periods with continuation.

[27] Again, notice the similarities with the plots for D in 2.1.1.

(positively correlated returns) properties of our time series. The answer lies in the concept of *long memory*. If our R/S analysis leads to H>0.5 (<0.5), we are more likely to find positive (negative) autocorrelation following the methodologies in 2.2. However, the opposite is not necessarily true. A series can be both mean reverting and persistent at the same time according to its long memory parameter. The difference lies in the fact that although first lags might not be positively or negatively correlated in the frameworks of 2.2, a more complex non-linear relationship between $S_t$ and S at multiple lags can be present. This is what H measures: *Given the persistence in the past n prices, what is the likelihood that the trend visible in these n prices will persist in the next n observations?* This is completely different in nature from autocorrelation that can be found in the most recent p lags[28]. This type of persistence can thus only be measured by non-linear methods[29]. In other words, long memory implies that relationships exist between any two observations, depending on this measure H. Both observations can be many lags away from each other, and recent lags could nevertheless be insignificantly correlated.

To put it briefly, long memory implies *long range relationships* between prices and is closely related to the previously described measure of roughness, the fractional dimension.

Does this analysis imply that Brownian motion is a fractal? Indeed, just imagine we go from the SDE notation to an implementation in code where we model actual price moves for actual time intervals. If we rewrite the equations from 1.6 into a discrete form, it does not matter for what size of time interval we define the stochastic process. We can do this straightforwardly if we dilate the measures of mean (typically close to zero) and volatility (with $\sqrt{T}$) with time. Therefore, zooming in on this Brownian motion, the statistical properties for a process simulated per month, per day or per

[28] An ARMA process will never use autoregressive terms of e.g. 50 observations away and skip more recent intermediate terms to capture long memory. It will only include terms up till a certain lag that is still significant. This is in clear contrast with long memory. However, remember the ARFIMA models of footnote 24, which try to leapfrog this issue by using a fractional order of integration.

[29] Id est R/S analysis, based on a power law, in contrast to covariance analysis, based on Pearson $\rho$.

minute would be the same. The self-similarity that we find will not be as visually attractive like other famous fractals, but the graph would be statistically self-similar. This finding is consistent with the fact that even seasoned chartists cannot tell what the timescale of a chart is if you do not tell them what the scale of the price is. Obviously, the changes will be bigger in magnitude for larger time steps as is implied in the increasing annualized volatility[30]. However, the overall 'look' and raggedness of the chart is completely the same. This further implies that the roughness or complexity of the price process is equal at every timescale. Remark that Mandelbrot's most concise definition of a fractal focuses exactly on this property of scale-invariant complexity: "*a fractal is a shape of which the complexity is constant at every scale*". The aim of this chapter is exactly to do this: measure complexity, see whether this scaling of time and variance is significantly different from ½ and whether this deviation is robust over sectors and geographies. We will provide first estimations for fractional dimensions and Hurst exponents of the data set in 2.5 and use those insights for our risk measure in chapter 4.

## 2.5 Fractal dimensions and Hurst exponents in financial markets: some empirical results

Before making the link with fractional Brownian motion (fBM), let us first consider some empirical results for the fractional dimensions and Hurst exponents of the price series in the data set. Is Brownian motion a realistic assumption for the data? If not, what are more realistic estimates for D and H? We find good fits for the log-log plots of the aggregate indices (Finance, Technology, Utilities, Telecommunications, Consumer Services, Health Care, Consumer Goods, Industrials, Basic Materials and Oil & Gas) of the 11 countries (Belgium, Canada, France, Germany, Italy, Japan, Netherlands, Sweden, Switzerland, United Kingdom and the United States). Let us first zoom in on these estimates and discuss them, before calculating the roughness on individual tickers in chapter 4.

[30] This corresponds to the 'accepted wisdom' that volatility scales for larger time steps with a scaling factor of ½ (under the assumption of independence, as we remember from the discussion at the end of page 19).

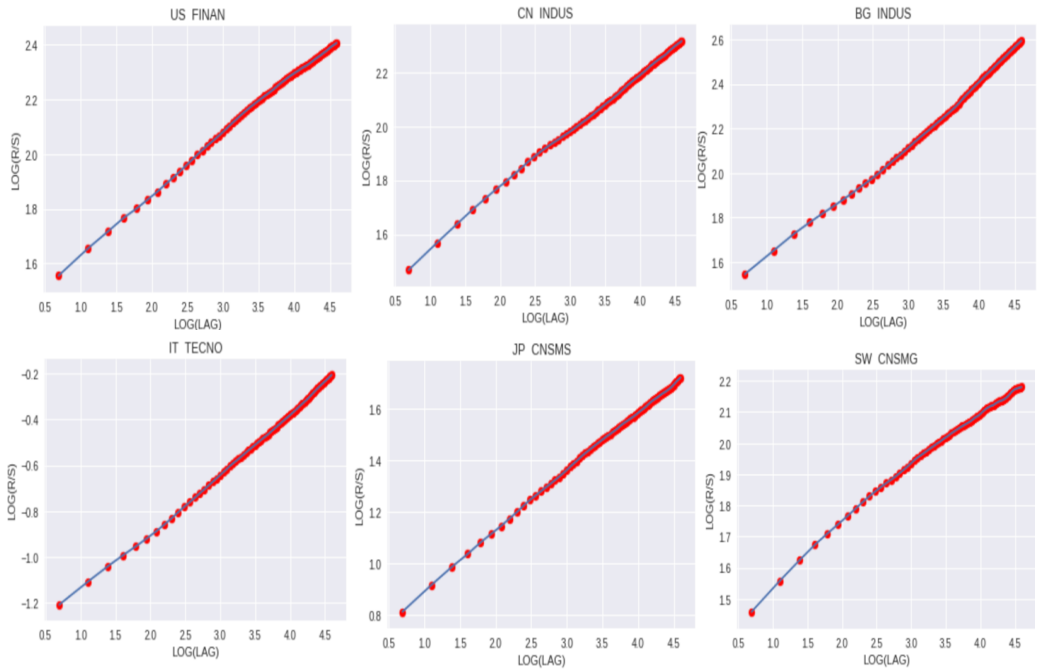


Figure 9: Six sample log-log plots from the data set

From the above results for 6 sample indices in Figure 9, we can conclude that the scaling of rescaled range and interval size indeed appears to hold quite nicely for the indices, with H equal to 0.4295, 0.4201, 0.5574, 0.5270, 0.4529 and 0.3417 for US Finance, Canadian Industrials, Belgian Industrials, Italian Technology, Japanese Consumer Services and Swedish Consumer Goods respectively. Below, the distribution of H for the 110 indices is depicted in Figure 10:

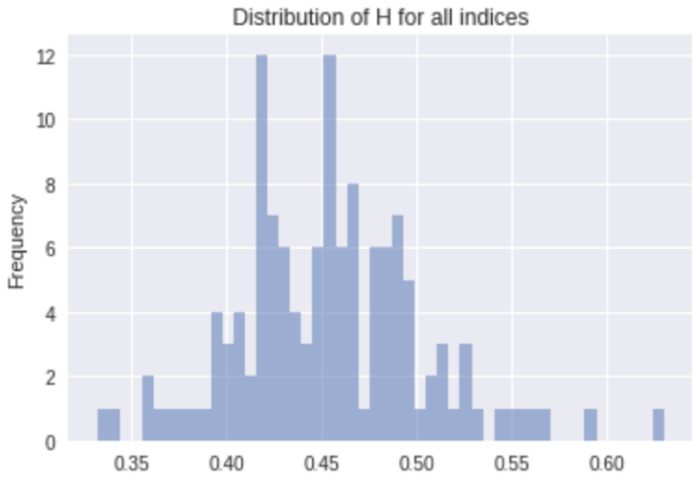


Figure 10: The distribution of H for all 110 indices

These are some pretty important results. First and foremost, we find that the Brownian motion case of H=0.50 is not consistent with our data. More precisely, ½ is just a special case and, naturally, H has a distribution rather than a value. Overall, *the roughness of the indices is higher than assumed in a Gaussian model*, resulting in lower Hurst exponents. Furthermore, the observed range in H (also see Tables 5 and 6 below) is *too large to say that these deviations are due to noise in the estimators*. We conclude that the roughness of a real-world price series is somewhere between 0.35 and 0.60, with most of the probability lying between 0.40 and 0.50.

Table 5: The 5 indices with the highest measured roughness (Low H)

| Country | Industry | Ticker | Hurst exponent |
|---|---|---|---|
| Japan | Telecom | TELCMJP | 0.331955 |
| Switzerland | Consumer Goods | CNSMGSW | 0.341743 |
| Netherlands | Utilities | UTILSNL | 0.357753 |
| USA | Consumer Goods | CNSMGUS | 0.357834 |
| USA | Consumer Services | CNSMSUS | 0.363426 |

Table 6: The 5 indices with the lowest measured roughness (High H)

| Country | Industry | Ticker | Hurst exponent |
|---|---|---|---|
| Canada | Healthcare | HLTHCCN | 0.630145 |
| Switzerland | Utilities | UTILSSW | 0.591553 |
| Netherlands | Healthcare | HLTHCNL | 0.567236 |
| Netherlands | Telecom | TELCMNL | 0.563048 |
| Belgium | Industrials | INDUSBG | 0.557427 |

Some of the most remarkable deviations are displayed above. For instance, American and Swiss volatile indices like Consumer Goods and Consumer Services result in more ragged charts with a roughness of approximately 0.35. The most persistent indices can be found in Canada, with Healthcare leading with a Hurst exponent of 0.63. A more

visual representation of these first calculations of H for all the countries and aggregate indices are depicted in the heatmaps below (Figure 11). In addition, the index performance (as a return multiple over the total time window) is shown.


Figure 11: Performance (upper panel) and H (lower panel) heatmaps

These heatmaps convey a lot of significant information. Firstly, some *extreme deviations from H=0.5 can be found*. For instance, in Anglo-Saxon countries (US/UK) and Sweden, an overall brighter orange is observed compared to the other geographies. These first markets are thus found to be more rough than the latter markets. In terms of industries, Consumer Goods, Consumer Services and Telecom seem to be brighter than average. This might be due to more nervousness on these fast-paced markets. Healthcare is a special case with both very smooth (Canada, The Netherlands) and very rough

(Germany, Sweden, Switzerland & UK) indices. Given recent trends, these idiosyncratic country effects can be explained by individual pharmaceutical companies undertaking M&A activities that do not affect the overall industry but, because of their disproportional weights in the local index, induce sudden changes (and thus additional roughness or continuity) in the country index. The total country average of the H exponent is shown below in Figure 12.

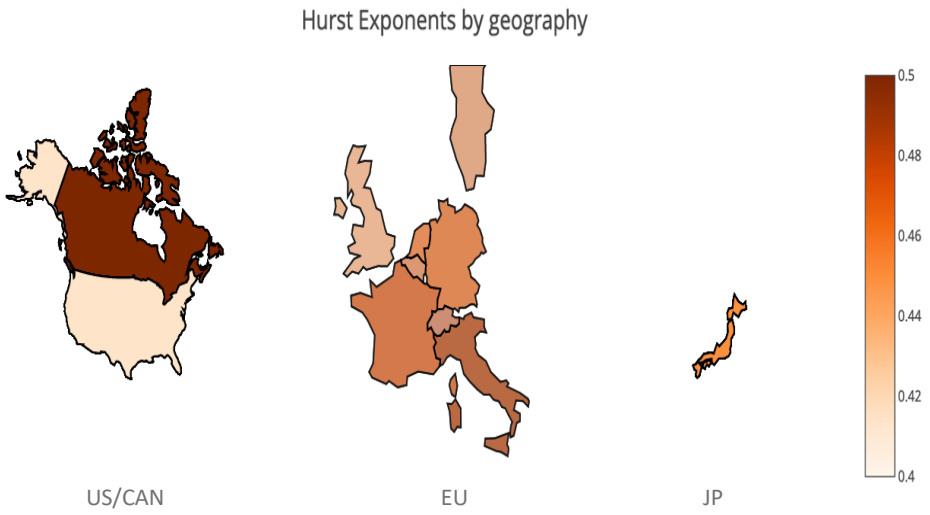


Figure 12: H index and clear country differences

Upon critical inspection of both panels in Figure 11, we find no clear-cut relationship between roughness and returns. This is exemplified by the UK and Dutch Technologies indices, two of the indices with the biggest returns, whose measured roughnesses are not remarkable. Rough markets first seem to have higher expected returns (UK, US & Sweden), but these conjectures are hard to corroborate based on the above evidence[31]. Moreover, we see that abysmal returns like the Swiss Utilities have a very smooth chart. All things considered, markets can go up in a rough fashion and go down smoothly, or the other way around.

We thus conclude that the *overall returns are not linked to the roughness of the index.* However, the scatteredness of the return time series is linked with roughness, as roughness essentially measures the dispersion of prices for smaller and smaller time scales. We will delve into the link between roughness and volatility in the next section.

[31] Neither did the numbers imply any significant correlations, but I preferred to let the pictures talk instead of crosstables.

Other useful visualizations of the distribution of H are boxplots. These plots below seem to confirm our view that was shaped by the heatmaps. We clearly notice that H, as well as its dispersion, differs across countries and industries.

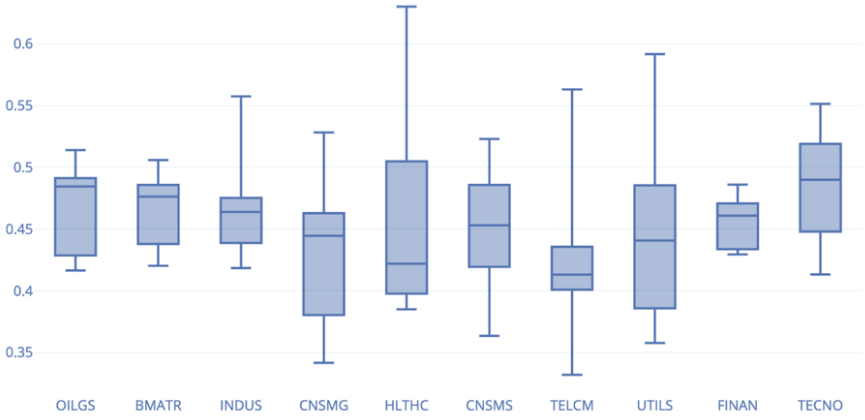


Figure 14: Boxplots confirm clear industry differences for H and its spread

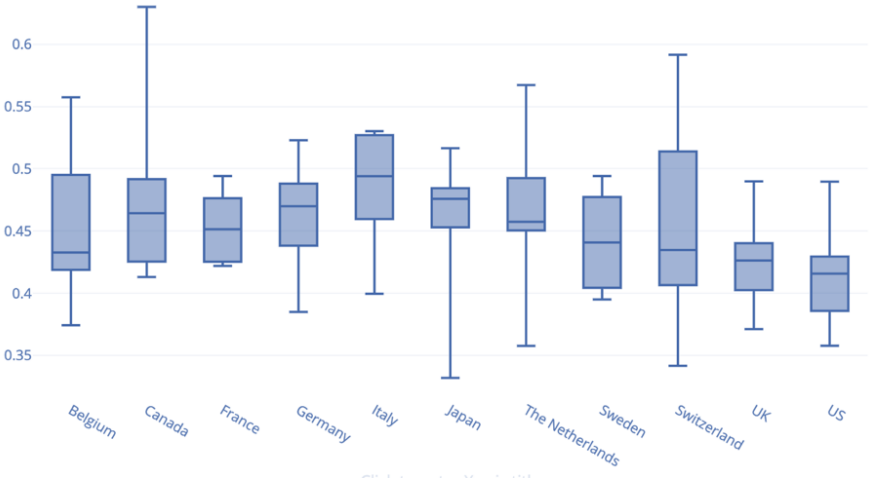


Figure 13: Boxplots also confirm our country view on H

We already provided evidence that Consumer Goods, Consumer Services and Telecom have the most occurrences of a low H and Figure 13 reaffirms this. We notice that Utilities also has some remarkably low outliers (e.g. Dutch Utilities, H = 0.36). Moreover, we can see that indeed Healthcare is a rather idiosyncratic industry with a high spread of values and a low mean for H. Consequently, Healthcare has together with Telecom the lowest expected H. Interestingly, the financial stocks have an unusual low spread for H, possibly meaning that they trade with the same roughness.

In terms of geographies, Figure 14 reasserts our view that Anglo-Saxon countries (US & UK), Sweden and Switzerland have the roughest markets. However, country effects should be nuanced from the point of view that the high spread of H for most countries implies that large deviations from BM can occur anywhere.

To conclude this rather pictorial part of the dissertation, let us also consider a similar heatmap for the Higuchi fractional dimension. Although it is closely linked to H and should therefore lead to similar conclusions, Higuchi's D is calculated with another approximating algorithm (cf. 2.1.1.1) and is therefore expected to be slightly different.

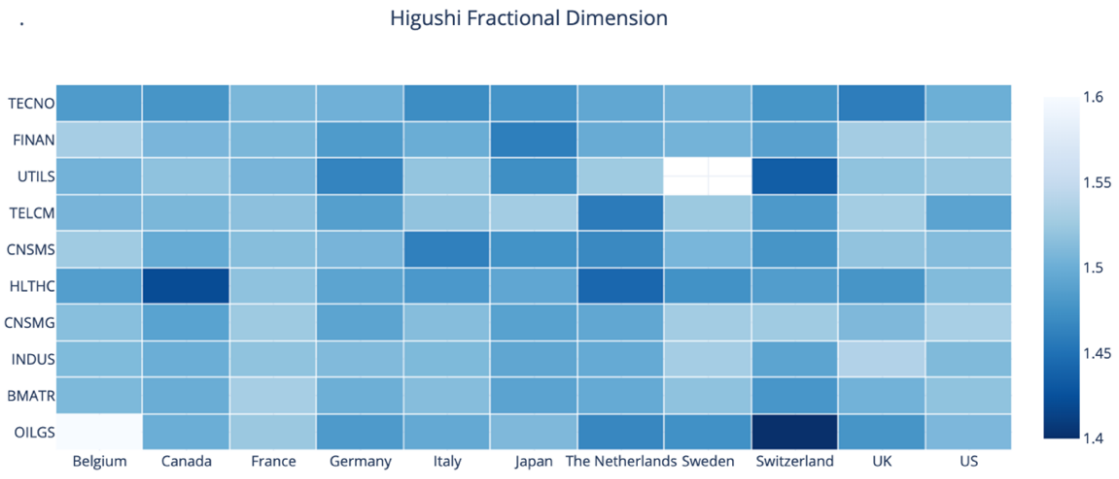


Figure 15: Higuchi D – Convergent conclusions, though not identical

Recall that a low H should result in a high D, since H ≈ 2 – D. We could argue that the typical grouping of the US, UK and Sweden still holds for D. However, Switzerland shows some completely different results. From the heatmap and boxplot of Switzerland we can see that the country has both outliers with a very high (≈0.60) and very low (≈0.35) Hurst exponent. From the fractional dimension we would expect values between 1.4 and 1.65. Rather, we see that the special cases have smoother Ds than expected, i.e. way smaller than 1.6. This could potentially mean that the used algorithms for fractional dimension have a bias towards picking a D that is too smooth, compared to the algorithm for H. However, France is now showing rougher figures than expected so this potential bias cannot be accounted for straightforwardly.

Of course, it has to be said that H is also a best fit in the log-log plot. It is therefore not an absolute number either, but an approximation. In addition, the formula theoretically only holds for self-affine processes with infinite complexity, while our data is at most an approximation of statistical self-similarity with finite complexity (cf. supra).
In summary and as a last conclusion of our empirical analysis, *the 'D approach' towards quantifying roughness yields convergent conclusions with the 'H approach', but they are not identical*. Although this is probably due to the different approximating algorithms and the fact that we do not have perfect fractals, the real reason is still ambiguous.

## 2.6 Fractional Brownian motions

In the previous section, we concluded that the roughness of an arbitrary stock in our horizon is not consistent with the Brownian case. To recap briefly, we said that Brownian motion, the random walk with infinitesimal time steps, is the case where we have no predictive power from either a certain level of autocorrelation, nor some gravity or degree of mean reversion. We explained Hurst modeled exactly this property in a more general way than linear autocorrelations, i.e. we look at long memory which is based on a power law instead of rho-based autoregressions. Hurst ranged from zero to unity, with ½ corresponding to the Brownian case. We said fractional dimensions of the considered time series were between 1 and 2 depending on the raggedness of the curve, so that 1.5 corresponded to BM. Additionally, from chapter 1 we know that the Brownian case is inextricably linked with a Gaussian view of the world. In this regard, consider the following random walk for illustration purposes:

$$S_{t+1} = S_t + \varepsilon_t \qquad \varepsilon \sim N(0, \sigma^2) \tag{2.26}$$

where the price of tomorrow is the price of today where we add a random deviation with mean zero. In other words, $dS_t$ is just Gaussian noise with zero expectation and volatility $\sigma^2$. The returns are thus serially uncorrelated Gaussian increments. This standard definition of a random walk therefore has the *Markov property*, i.e. it has no memory. This corresponds to no expected autocorrelation between increments $\varepsilon_t$ and

$\varepsilon_s$, t>s. However, prices have an observed autocorrelation between observations at t and s, t>s [32]:

$$\mathrm{E}[S(t)S(s)] = \sigma^2 \mathrm{s} \tag{2.27}$$

Deliberately, we avoided formal proof or a mathematical rigorous explanation for D and H of this process, since it requires overly sophisticated mathematics for the aim of this section[33]. The only aim is to provide intuition with regard to the link between roughness and the assumptions we heavily criticized in chapter 1. One elegant angle to see this, and which is of importance for this dissertation, is to regard these figures from the fractional Brownian motion generalization of BM. *Fractional Brownian motion* (Mandelbrot and Van Ness, 1968) has a covariance function, t>s:

$$\mathrm{E}[S_H(t)\ S_H(s)] = \frac{\sigma^2}{2}(\ |t|^{2H} + |\mathrm{s}|^{2H} - |t-s|^{2H}) \tag{2.28}$$

where we generalize our Brownian motion for a covariance ('*memory*') function that is a function of H. In a fBM model, the autocorrelation between observations for an arbitrary shift $t - s$ is the expression above (t>s). Indeed, one can easily check that if we plug in H=0.5 into the covariance function, we find ordinary BM, such that BM is just a special case of fBM with no long memory.

Formal proof requires solving a complex integral[34] (Mandelbrot, 2002) and is beyond the scope of this dissertation, but it should be intuitively clear that the persistence in the series goes up with the Hurst exponent H. Indeed, rescaled range analysis enables us to model the memory feature of financial time series by looking at the roughness of

[32] Or generally $\mathrm{E}[S(t)S(s)] = \sigma^2 \min(\mathrm{t}, \mathrm{s})$. This can easily be seen from equation 2.26. Backward iterating yields $S_t = S_0 + \sum_{i=0}^{t-1} \varepsilon_{t-i}$. The last term is the *stochastic trend*, such that $\mathrm{cov}(S(t)S(s))=\mathrm{E}(\varepsilon_t+\ldots+\varepsilon_1)(\varepsilon_s+\ldots+\varepsilon_1))=\mathrm{E}({\varepsilon_s}^2+\ldots+{\varepsilon_1}^2) = \mathrm{s}\sigma^2$. Thus, returns are uncorrelated and prices are correlated because of the stochastic trend only. Therefore, in line with the discussion in footnote 26, BM has no memory.

[33] As this chapter only provides an introduction to fractals and their link with finance, we only included sample plots to prove this (Figure 3 and 8). The interested reader with a more mathematical background is referred to Mandelbrot (2013).

[34] Mandelbrot replaced the Riemann–Liouville fractional integral proposed by Lévy by a Weyl integral. White noise, the Gaussian increments $dS_t$ (the $\varepsilon_t$ in 2.26), are fractionally integrated using the factor $(t-s)^{H-1/2}$ in which we recognize the autocovariance function of BM. Notice the link with ARFIMA models in footnote 24.

the data. We leapfrog the econometric approaches of 1.6 that try to find linear comovements in an autoregressive scheme. Recall that these methods often provided evidence for the EMH and the application of BM for risk management purposes (1.4.2 and 2.2). Now we are able to go beyond this linear approach starting from R/S analysis. The worthwhileness of this approach stems from the observation that the roughness of BM is not consistent with real stock data (cf. the last section). This thus implies that *there is long memory present in our stock price data*.

The previous statement can be best understood from the discussion we had about the scaling property of Brownian increments with the square root of time in 1.6:

$$dS \sim \sqrt{dt}\,\varepsilon \tag{2.29}$$

We said that this expression is only valid if there is *no memory* (the Markov property), i.e. H = 0.5. Since we now have memory in the fBM case, we expect a different relationship. We can start from the same perspective as BM: *the distance traveled is proportional to the power H of the time elapsed*. However, with fBM we say that the power is not necessarily ½, but can be any H (Mandelbrot, 2002; Velasquez, 2010):

$$dS_H \sim dt^H\,\varepsilon \tag{2.30}$$

This is yet another example of a scaling law: the quantity S (a stock price, interest rate, volatility, etc.) has increments that are proportional with the $H^{th}$ power of the size of the time step (the scale). For instance, this means that a 10-daily VaR can be approximated by $10^H$ times the daily VaR, where H is the roughness of the time series of the daily VaR. This is in clear contrast to the Basel-compliant $\sqrt{T}$-rule. BIS regulation basically implies that the time structure of VaR is the root of time. BIS capital adequacy rules stipulate that banks should operate with a holding period of two weeks (10 days), which implies the application of $\sqrt{10}$ times daily VaR[35]. Again, this means that BIS

[35] Basel refers to daily VaR as DEAR (Daily Earnings At Risk). The internal models framework of the BIS capital regulations (cf. 1.1) sets the minimum market risk capital requirement to be the larger of (1) $\sqrt{10}$ * Previous DEAR or (2) Multiplier * $\sqrt{10}$ ∗Average DEAR (Allen et al., 2009).

assumes that there is no memory in consecutive losses. This is in clear contrast with what we observe in bad trading periods where losses are accumulated. Then, the H of the daily VaR time series becomes larger and the losses over one day should be dilated by a larger factor if we try to use them to predict what we are losing over multiple days. We will come back to this key insight of roughness for VaR in 5.2.

Another more common example is when the quantity is volatility, where H measures the roughness of the volatility process. Gaussian models imply a term structure $\sqrt{T}$ for volatility (recall the discussion on annualized vol in chapter 1), yielding results that are not consistent with realistic volatility patterns.

The main literature using fBM for financial applications is the so-called *rough volatility* literature. Rough volatility uses fBM for the volatility process rather than the stock process. Hence, it tries to measure the long memory in the volatility process. Gatheral, Jaisson and Rosenbaum (Gatheral et al., 2014) showed that *logvolatility* behaves like a fractional Brownian Motion with H of order 0.1. Recall the Heston model in 1.6:

$$dS_t = \mu S_t dt + \sqrt{v_t} S_t dW_2 \tag{2.31}$$

$$dv_t = \kappa(\theta - v_t)dt + \xi\sqrt{v_t} dW_2 \tag{2.32}$$

Gatheral et al. found a remarkably robust *monofractal scaling property* for the time series of volatility:

$$\log(\sigma_{t+\delta}) - \log(\sigma_t) = \vartheta(B^H_{t+\delta} - B^H_t) \tag{2.33}$$

In other words, the logincrements in volatility over a shift $\delta$ is proportional with the increments of a fBM process, with a proportionality constant $\vartheta$. Gatheral and Rosenbaum showed that this relationship holds for all 21 equity indices in the Oxford-Man database, Bund futures, Crude Oil futures and Gold futures, with an H in the order of 0.1 (Gatheral et al., 2018). This allows us to write down the following stochastic volatility model:

$$dS_t = \sqrt{v_t} S_t dW \tag{2.34}$$

$$\log\left(\sqrt{v_t}\right) = X_t \tag{2.35}$$

$$dX_t = \vartheta dB_H - \alpha(X_t - m)dt \tag{2.36}$$

Lo and behold, the *rough fractional stochastic volatility* (RFSV) model, which looks like a *rough* version of the Heston model in Eq. 2.31 and 2.32. As its name suggests RFSV replaces '*Brownian, Markovian stochastic vol*', by '*rough, fractional stochastic vol*'. In the first equation (Eq. 2.34), we recognize the stochasticity term for the change in S, which we have seen in all the previous SDEs. The second and third equation just rewrite the monofractal property (Eq. 2.33) above. Note that the $\alpha(X_t - m)$ term is a correction term that drives the logvolatility to its long-term level, much like the Ornstein-Uhlenbeck process discussed earlier. That is why $X_t$ in its complete form ($\alpha \neq 0$) is called a *fractional Ornstein-Uhlenbeck process* (fOU). This model is applied for enhancing option pricing in, inter alia, the *rough Bergomi* (rBergomi) model. This is a slightly simplified version of the RFSV model with special merit in fitting the observed volatility surface. "*The rBergomi model fits the SPX volatility markedly better than conventional Markovian stochastic volatility models, and with fewer parameters*" according to Bayer et al. (2016). Heston models, much like SABR, Hull-White and other stochastic volatility models, were already able to model the curvature of the surface by looking at the volatility of volatility (a second order model of vol), but RFSV has special merit in deriving even more realistic smiles (Bayer et al., 2016; Gatheral et al., 2018; Jacquier et al., 2018). The core difference is thus to replace Markovian models by long memory models, using the measured roughness of volatility.

To reiterate my earlier point, the purpose of this section is to give the reader insight in how roughness is entering financial modeling at vast pace, not to confuse him with equations. My only aim is to provide intuition for the link between roughness and the assumptions in chapter 1. In that regard, the key take-away from the work of Gatheral and Rosenbaum, is that it proves that fractal-inspired scaling properties of financial time series are really measuring some underlying dynamics and fractals are not just '*a cult applied to finance*'. Moreover, the interest in rough volatility has invited researches to do empirical work on the implications of the changing roughness of volatility. This empirical work is particularly useful for this dissertation. Figure 16, for instance, shows

the roughness of the volatility process and its link with real-world events (Gatheral et al., 2018).

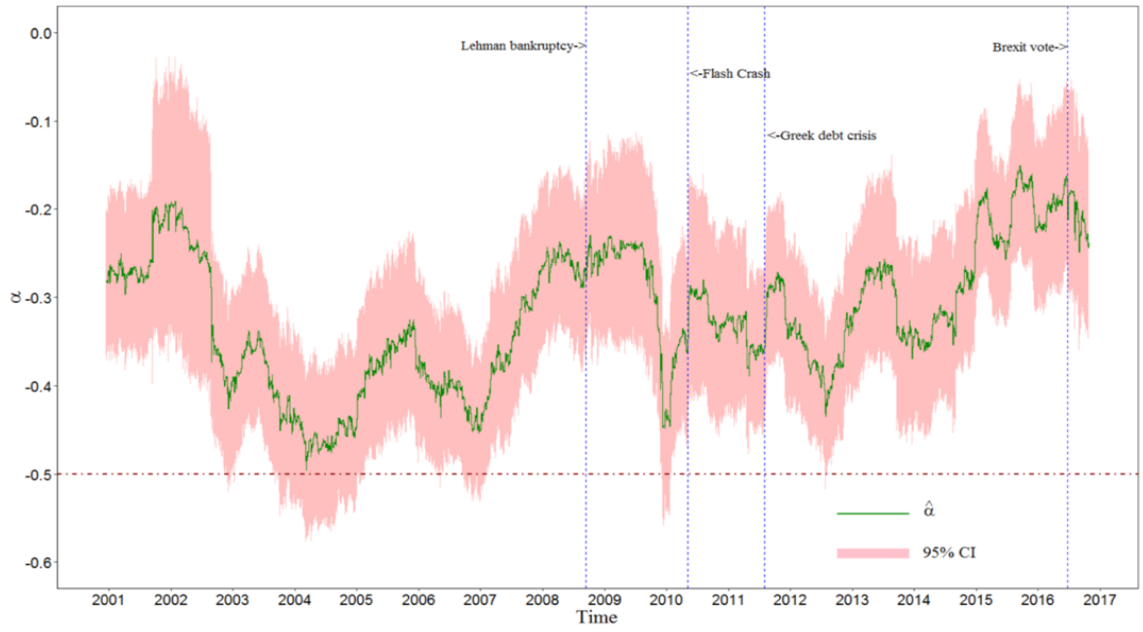


Figure 16: Gatheral et al. (2018) show that the roughness of the volatility process ($\alpha = H - \frac{1}{2}$) is highly correlated with real-life events. When the persistence goes up, it is clearly linked to market turbulence like the 2008 Lehman bankruptcy, the 2010 Flash Crash, the 2011 Greek debt crisis and the 2016 Brexit vote.

This shows that (1) *markets are rougher than most models* (cf. 1.6) *imply*, (2) *roughness changes over time* and (3) *roughness is highly correlated with real-world events* (Figure 16). That is why we will tap from their intuition and borrow some of these scaling properties for the combination of risk models in the next chapter. An obvious but crucial simplification is that this thesis considers the roughness of the price process and not the underlying volatility process. Though the scaling property (Eq. 2.33) and the observations in 2.5 are related (time-varying persistence in volatility may induce time-varying anti-persistence in stock prices), they are fundamentally different processes. The intuition is therefore transversal, but the concepts cannot be used interchangeably.

All things considered, there is a mathematical foundation (or at least intuition) that links the assumptions in the standard models to Hurst exponents and fractional dimensions. Some of the relationships that were found are remarkably robust. A first intuitive consequence of the previous paragraphs is that an assumed data generating process based on roughness (e.g. fBM) would yield more extreme VaR numbers if

volatility is higher or *the smoothness of the volatility process is higher/the roughness of the price process is higher*. Note that there is an '*inverse property'* between the *roughness of the price process and the roughness of the volatility process*. Persistence in volatility occurs when we are in a so-called volatility cluster, when returns are more scattered and prices look more rough. Just keep the sample plots in 2.3 in mind. When H of the volatility process is decreasing, the volatility estimates look very rough. This is typically outside of the cluster, when vol is very anti-persistent. Consequently, this corresponds with more benign periods in the time series of the stock price, i.e. when the D and H of the price process will be indicative of a smoother period.

The reader will by now understand that all these remarks imply that randomness is more than sigma, and this adds another dimension to the '*uncontrollable element*' that is put into the model by the modeler. The methods in chapter 1 introduced volatility as only element of uncertainty, i.e. a Brownian noise term was multiplied with a (stochastic) sigma. Although these stochastic volatility models were designed explicitly to include persistence in the volatility process, we have concluded that they are not quite able to do so because of their dependence on linear autocorrelation. Indeed, GARCH is an *autoregressiv*e scheme that suffers from the same essential weaknesses as our common Gaussian models, since it is a memoryless volatility model. Although Heston, SABR, Hull-White, etc. all provide different approaches, they typically belong to the same 'family' of Markovian models[36].

Therefore, we can conclude that *an additional element of uncertainty needs to be included: the roughness of the process*. As we just extensively discussed, *roughness is linked to volatility*, the common perception of riskiness, *but it is not quite the same*.

While realizing that the approach of this dissertation is 'atheoretical' in comparison with the work on RFSV, the intuition behind a roughness-based combination model for risk measures seemed to find support in this work. This is mainly because it shows that time-varying roughness is a variable that is both related to the underlying data generating process assumed in our models and a 'contextual variable' that correlates

[36] Again, this is only partly true since long memory features returns have, similarly to ARFIMA models, led to a set of FIGARCH models (fractionally integrated GARCH). However, the results are less convincing than rough volatility (Baillie et al., 1996; Gatheral et al., 2018)

with the dynamics in the underlying market. Although ML models (cf. the next chapter) might be seen by quant purists as brute force models, they are not necessarily doomed to be lesser mathematical models than explicit closed-form equations, not in the least in terms of performance. Admittedly, the set-up or internal structure of the model will be less transparent, but the nature of the model is not necessarily different. This corresponds to more recent trends in numerical and computational finance where the need for a solution is identified, as well as the means to get there (by using heavy computational power), but there is no need for a closed form solution. Therefore, I cannot mathematically prove that the combination proposed by the model is optimal, but I can show you the backtesting results.

With all the previous remarks in mind, the goal of this thesis and what we will continue to do in the next chapters can now be summarized as follows:

This dissertation investigates the use of standard parametric and non-parametric approaches to estimate Value-at-risk (VaR) and combine them in a neural net with the purpose to:

1. Reduce the overall bias of the methods by combining them (Inui, Kijama, Itano, (2003), Liu (2005)).
2. Explicitly tell the machine - in the loss function - to learn a combination that minimizes the exceptions in financial loss (based on Kupiec, Christoffersen, 1998).
3. Investigate the use of fractal-inspired complexity measures for this combination, given the intimate relationship between fractional Brownian motion (fBM, Mandelbrot & Van Ness, 1968) and the generalization of standard assumptions in the classical models.

The backtesting framework of Kupiec and Christoffersen (Christoffersen, 2008; Kupiec, 1995, 1999) will be explained in more detail in chapter 3. Introducing the backtesting measure into the model explains why the model is not necessarily going to underperform closed-form solutions based on roughness. This is because now we can do (2), i.e. define exceptions as explicit loss. However, this increases the perception of our risk measure as a black box. This pitfall will be discussed in due detail in the next chapters.

## 2.7 Conclusion

*"The art of asking questions is more fruitful than the art of finding solutions."*

Georg Cantor

We can conclude this chapter by saying that the intuitive concepts of the *art of roughness* and *rough markets* created a vast interest by quants and financial theorists to study the subject in recent years. I hope I was able to cherry-pick the most exciting ideas behind this art without being too mathematical[37] nor giving the basic math, that was used in the code, short shrift.

Recently, roughness has entered the equations through fractional Brownian motion. fBM is a generalization of BM, like fractional dimensions are a generalization of ordinary dimensions, and can possibly link the methods from chapter 1 to each other. All of these standard methods rely on a hypothetical data generating process as they are formalized by SDEs. In this chapter, we started very intuitively by reflecting about different types of predictability and eventually showed that this boiled down to all of these SDEs having an *assumed roughness*. The question is to what extent this roughness is consistent with the specific stock or market in question. That is why this dissertation starts from measuring this roughness before haphazardly applying any of the above methods without thinking about its appropriateness (recall the coast line analogy for predicting predictability). First estimates for these complexity measures for our data set were given in this chapter, where it seemed obvious that the diffusion model does not make a lot of sense for real-world markets.

Because of all these remarks, roughness might serve as a simple and logical but powerful connector in a combination model. This realization, combined with the rapid

[37] It is very hard for a subject like this to focus on the main implications and not on the math, without giving the math short shrift. I know this chapter will probably sound technical for finance audiences and very heuristic for math audiences. However, the main ideas do not change when we refine the mathematics gradually, i.e. in further research.

development of machine learning techniques (see the next chapter for a primer on ML) that enable the modeler to make sense of huge realms of PnL data in a cheap and fast way, suggests that there might be a *power in combination*. Both in combining different methods for VaR, as well as combining fractal properties with a machine learning framework. The research question stated in the introduction now comes down to *whether roughness can be an informative feature in a combination model, or whether this approach only contributes to spurious precision*. To answer this question, we will focus on critically investigating the backtesting results of our different final models in chapter 4.

# Chapter 3

# What are Deep Neural Networks?

## 3.1 A brief introduction to machine learning

*"Solving the right problem numerically beats solving the wrong problem analytically every time."*

Richard Martin

Because of (1) the explosion of data, (2) algorithmic advancements, (3) the availability of vast storage space and (4) the increase in computing power (Manyika, 2017), we have seen a proliferation of *machine learning* applications in finance and our daily lives over the last years. In order to better explain the concepts behind the used machine learning algorithms, I think this dissertation deserves a primer on ML. How does a machine learn? A few basic concepts can help explain the design freedom one has in developing an ML model and can therefore help explain the motivation behind the modeling choices that were made.

Machines learn by minimizing what the modeler defines as *loss*. Imagine a classic OLS regression. There we tell the machine to come up with weights for our feature variables that minimize the loss, i.e. the squared deviation of the predicted output with the real output. In general, this loss could be anything: least-squares, mean absolute errors, mean absolute percentage errors, mean squared logarithmic errors and more technical ones like binary cross-entropy, sparse categorical cross-entropy and so and so forth (for an overview see Vapnik, 1999). The machine translates the input features into an output as to minimize this measure of loss over time. This first design parameter gives us the opportunity to explicitly define an *exception* - a financial loss higher than predicted by the model - as loss. More exactly, we can now tell the machine this number

of exceptions should be consistent with the confidence level and the loss should increase as the number of exceptions diverges further from the theoretical number (see 3.3).

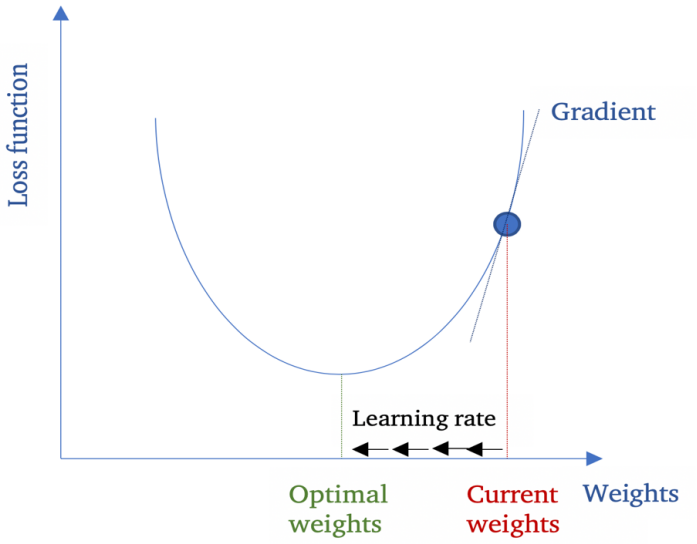


Figure 17: Minimizing loss using gradients

The machine minimizes the loss over different iterations or steps called *epochs*, using an *optimizer* algorithm. These algorithms are based on *gradient descent* (see Fig. 17), a mathematical technique using the gradient, i.e. the first partial derivatives vector, which shows in which direction the loss function decreases the fastest. This explains why the loss function should be differentiable. In plain English, the machine alters the weights of the model a little in each direction and looks at the response in loss. Subsequently, it chooses to change the weights in the next step in the direction where the impact on the loss function was the most significant decrease. It is the sort of hot-and-cold game children like to play. There is no artificial intelligence behind machine learning, the machine only 'knows' in which direction it should move towards achieving a minimized loss function. The specific algorithms I considered were AdaGrad (Adaptive gradient algorithm), RMSprop (Root mean square propagation), Adam (Adaptive moment estimation), SGD (stochastic gradient descent), AdaDelta, AdaMax and Nadam (for an overview see Ruder, 2016). These are all variants of gradient descent with their own merits and limitations. The power of genetic algorithms, a set of bio-inspired algorithms we will discuss in 3.4, is that they determine throughout the

generations which algorithms work best for the data and loss function you are working with.

The so-called *hyperparameters* the designer has to come up with are the learning rate and the number of epochs. The *learning rate* boils down to the size of the steps the machine takes. Large steps can bring you to a solution fast, but that solution might be suboptimal since the algorithm will jump from one side of the minimum to the other when it is close to it, not being precise enough to find the 'real' minimum. The number of steps or *epochs* is just the number of iterations you use to optimize the loss function. A large number might bring you closest to the optimum, but it may take an inconvenient amount of time to do this, or the machine may even time out before getting there. Other hyperparameters are, in the case of a DNN, the *number of layers* and the *number of neurons* per layer (cf. 3.2).

## 3.2 What is a Deep Neural Network?

**Neural network** - *a mathematical model or computational model based on biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In more practical terms, neural networks are nonlinear statistical data modeling tools used to characterize highly complex and convoluted relationships between inputs and outputs or to find correlation patterns in financial data* (Sun et al., 2008).

A *neural network* is a mathematical function $f$, linking a vector X of input variables $x_i$ to an output variable $Y$:

$$f: X \rightarrow Y \tag{3.1}$$

The power of this set of models is that neural nets use so-called *hidden layers* where the input is connected by the weights $w_i$ and transformed by a special function called the activation function $G$. In these layers, complex sequences of essentially simple non-linear transformations are performed.

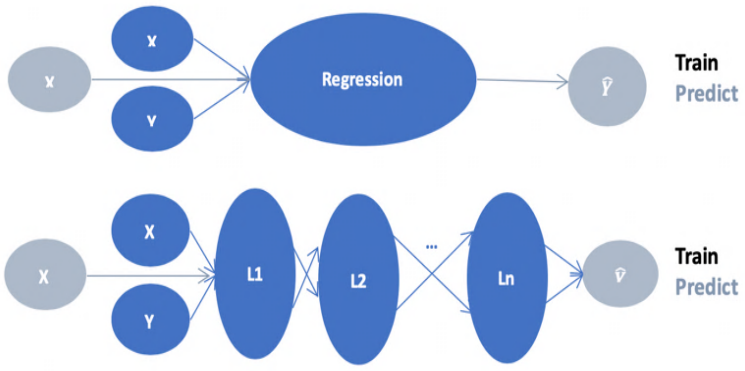


Figure 18: Comparing the monolayered regression model with the multilayered architecture of a DNN

Again, consider a standard regression model as an example. We use the input variables $X$ and initialize the model by giving them arbitrary weights, as to explain $Y$. Then we impose that the weights should minimize the sum of squares of deviations of the predicted Y with the real Y. Using a sample to do this, one could call this training the model. As a result, we have a model that gives weights to the input variables, as a linear transformation of the input, to explain the output variable. Naturally in a linear model, the coefficients are fully based on the Pearson correlation between $X$ and $Y$ and their individual variances. In a typical regression framework, you only measure the linear comovement between the variables as to come up with the coefficients. Moreover, there is only one transformation of the input. We can now evolve this model into a neural network model.

In the multilayered architecture of a DNN model (see Fig. 18), we have different types of layers: the *input layer*, the *hidden layers* and the *output layer* (Giudici, 2005) denoted by the three equations (3.2), (3.3) and (3.4) below. As is also clear in Fig. 19, input is transformed through sequences of transformations. Similar to the human brain, the most elementary computational unit is called a *neuron* (Bolland et al., 1998). These neurons are connected by the weights $w_i$ and activated by a function $G$, so that in essence (Sun et al., 2008):

$$n_k = w_{k,0} + \sum_{i=1}^{i*} w_{k,i} x_i \quad (3.2)$$

$$N_k = G(n_k) \quad (3.3)$$

$$y = \gamma_0 + \sum_{k=1}^{k*} \gamma_k N_k \quad (3.4)$$

In words, the elementary mathematical functions transforming the input $x_i$ are the k neurons $n_k$ that give weights to the input features in the first layer. These neurons are then activated by an activation function $G$.

*Activation functions* are the very essence of DNNs. They are the non-linear transformations that enable the model to learn non-linear relationships between the input and the output. They thus introduce more complexity into the model, but also increase their performance dramatically. Figure 19 shows that these functions can be seen as an additional layer of mathematical functions that activate neurons before they are given a weight, i.e. connected with the next layer. Many breakthroughs in voice and image recognition can be explained by the fact that researchers solved the problem to separate non-linear sets of data by introducing activation functions in the model. Examples of activation functions are sigmoid functions, rectified linear units (relus), tanh, etc[38]. These functions are seemingly simple transformations of the data. A *relu* takes any number as input and returns the same number if it is larger than zero and zero otherwise, much like the payoff of a call option. *Sigmoid functions* take a number between $-\infty$ and $+\infty$ and transform it into a number between -1 and 1. Alternatively, sigmoids are used to transform any number into one between 0 and 1, as to come up with a probability in logistic regressions. This is used for classification algorithms like voice and image recognition, and many more.

Recall the opening quote: "*Bottomless wonders spring from simple rules… which are repeated without end*"[39]. Again, the nesting of seemingly simple transformations can

[38] For a comparison see DasGupta and Schnitger (1993)

[39] Although a lot remains unknown about the brain, chances are it is a fractal too. This might explain the existence of some applications in engineering where fractals meet neural networks (Castillo and Melin, 2002; Ryeu et al., 2001; Steeb, 1999).

deliver very interesting results since we are now able to model complex relationship by introducing non-linearities.

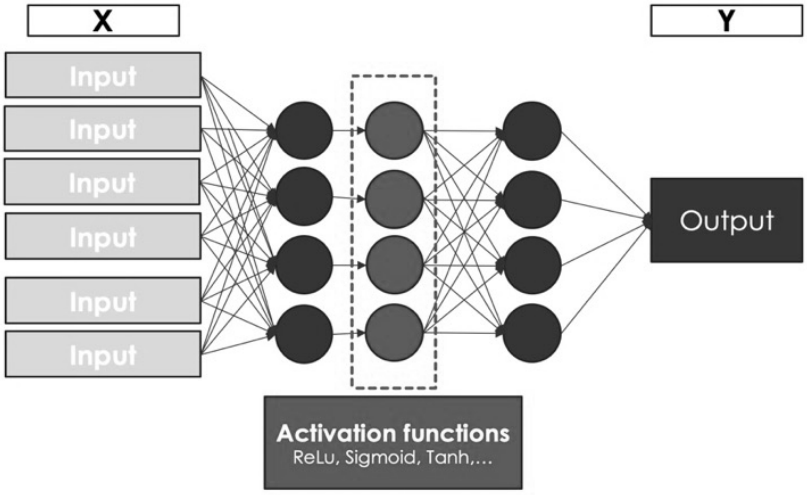


Figure 19: Activation functions - Architecture

In a fascinating online article with the title '*Neural Networks, Manifolds, and Topology*' Christopher Olah from Google Brain explains the power of introducing non-linearities in a model using an original perspective: *topology*. Suppose you want to separate a dataset into two groups according to a label, say male or female, based on two known features that you will tell the machine, say height and weight.

You could plot the two known features on the X and Y axis and see how these two groups behave (see Figure 20). In the example, we clearly see that the blue and the red dataset are not linearly separable if we only use X and Y. So how can activation functions help us out? The full story is rather involved in terms of mathematical and geometrical concepts and would take us too far[40]. The intuition, however, is that an ML model can learn an alternative *representation* of the dataset so that it is linearly separable. The model starts to transform the two datasets differently according to the feature we want to distill. Activation functions learn the best non-linear representation that enables the last layer to make a linear combination of the transformed data as to distinguish the two labels. In the words of Olah's blog "*Each layer stretches and squishes*

[40] However, the paper itself is a must-read for anyone interested in how neural networks work and who has a basic understanding of math and topology. Olah (2014) is included in the references.

*space, but it never cuts, breaks or folds it. Intuitively, we can see that it preserves the topological properties.*" One's mind boggles when you think about how involved these models get for multiple dimensions when using simple transformations that are repeated to a large extent. But why do we need these non-linear transformations?

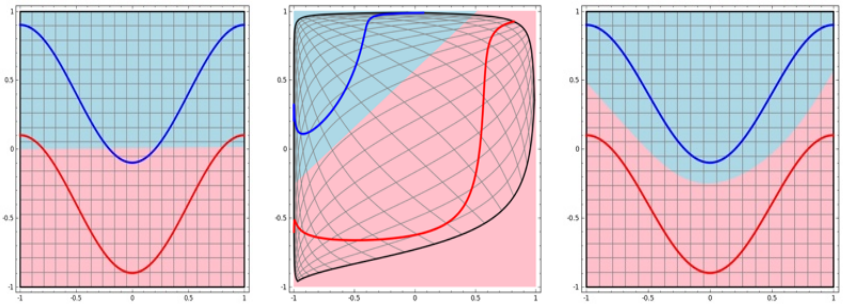

Figure 20: "*The hidden layer learns a representation so that the data is linearly separable*" (*Neural Networks, Manifolds, and Topology*, Colah's blog 2014)

## 3.3 Why a DNN for our model?

*"All models are wrong, but some are useful."*

George E.P. Box

One could argue that the traditional methods (individually) have a very linear view on risk (also see 1.4.1). We typically decide on a static confidence level, which is dependent on the purpose of the calculation. This is done by the regulator (see BCBS III, 2017) for e.g. reporting or the calculation of capital requirements. On the other hand, hedge funds and other financial institutions can calibrate these statistical models on any confidence level they like for optimization purposes. The chosen (set of) distribution(s) is recalibrated over time but their nature stays the same. Examples are the standard z-scores, t-scores, etc. based on the predetermined confidence level. Furthermore, we notice that the recalibrated parameters (e.g. other distributional parameters like excess kurtosis for the degrees of freedom of a Student t, the shape of the GEVT,…) do not move a lot over time. Consequently, our main input for the

eventual risk measure is volatility. Once we have decided on the previous elements, our eventual risk measure is some linear transformation of volatility. All these methods (except for a simple HS) essentially stretch out volatility in a linear fashion (cf. Figure 22).

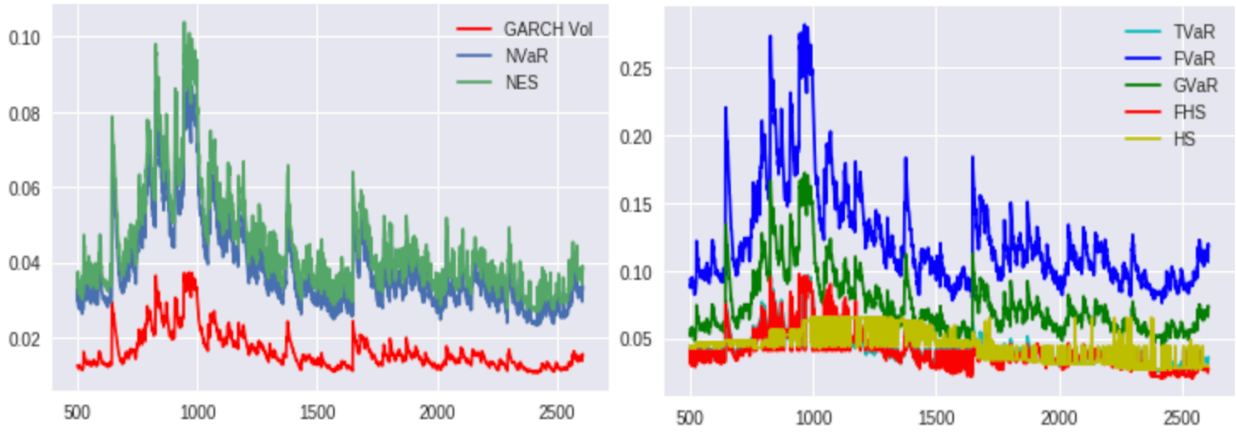


Figure 21: Some standard model estimations of bad quantiles moving over time – The estimated stochastic volatility is linearly stretched out

If we had critically reflected about the selection of formulas that was given in Table 2, we could already have figured out that volatility is the most important variable in the equation, bar none. We already extensively elaborated on the dangers of variance myopia and a linear transformation of vol into risk measure. These linearities result in abrupt responses if the vol changes, which casts doubt on the reliability of the initial model. For instance, in the beginning of 2018 G-SIBs were announcing their trading risk tripling at their equity unit in less than one month because of surging volatility. Terms like volatility breakouts and correlation breakdowns to denote sudden pernicious impacts on financial models often boil down to high sigma-dependence. We should not be silly and stop using statistical sigma as our main definition of volatility or abandon our quests to improve stochastic volatility models (cf. 2.6). The point of repeating this critique in this section is, again, to probe more deeply into why it does not work. The reasons are plenty, but the blatant one is again that is too focused on a limited set of (static) assumptions.

In this regard, a combination model could potentially give more robust estimates, not only by (1) combining different approaches and altering their weights over time, but also simply by (2) introducing non-linearities.

(1) refers to the fact that we are combining different VaRs into one risk measure, which is in essence no different from a SRM. Any combination of VaRs that adheres to the three conditions summed up in 1.10 can be seen as a SRM and will be coherent as long as we take enough projected VaRs into account. This means that these projected losses, or scenarios with corresponding probabilities, need to be informative enough to satisfy the three conditions. An additional motivation, next to coherence, is that the weights adapt over time based on a learning process. Static assumptions are implied in the features, but the combination approach with dynamic weights makes the model more adaptive and responsive. One could argue that in light of the discussion on the efficiency of markets, and the distributional properties this implies for VaR, we now evolve towards an adaptive market view (see 5.2).

(2) refers to the fact that the world does not work in a linear way. Figures 22 and 23 compare the linear thinking of the standard models with an ML framework. Much like Colah's example, the 'real VaR' will probably not lie somewhere in the space of all linear combinations of the VaRs that go into the model. If that would be true, a simple regression framework could do the job from this linear perspective (i.e. apart from the bespoke loss function, etc.).

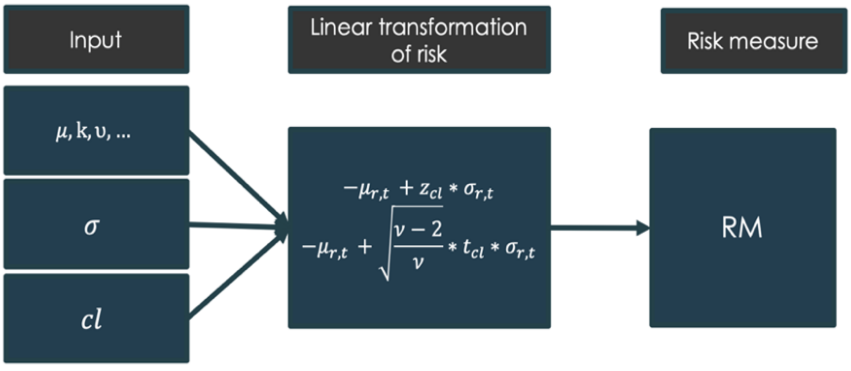


Figure 22: Traditional models

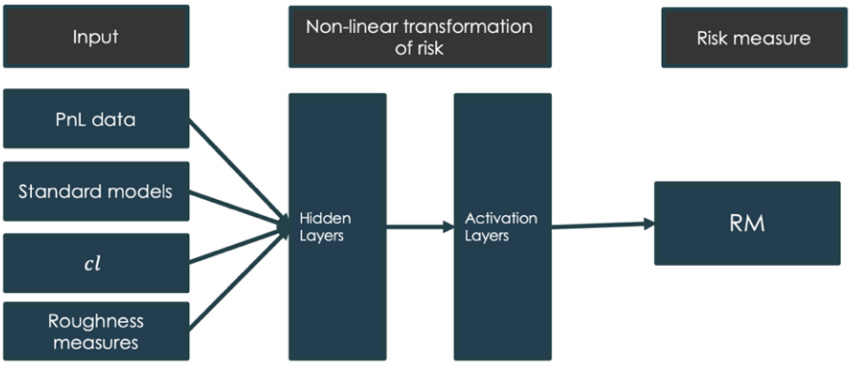


Figure 23: ML models – Introducing non-linearities is no quick-fix, but it is a step in the right direction

Once we agree on the fact that non-linearities have a lot to offer for our problem, we can choose to use either raw PnL data (Sun et al., 2008) or a combination of standard VaR approaches (Liu, 2005) as input for our model. It is clear by now that we will combine different 'instantaneous' VaR estimates in a DNN to predict the magnitude of a bad quantile of the PnL consistent with the proposed confidence level cl. One important motivation for this, is that the alternative approaches require a lot of lagged PnL or VaR data as features. This means that the ML model is basically looking for patterns in the past n observations of the PnL/VaR, where n stands for the number of included lags. Sun et al. (2008) uses a lag of 10. This could be seen as a VAR[41] model of x VaRs with lag n=10, where the estimation is done through a DNN instead of classical methods. Anyone familiar with these models will know that a lag of 10 very quickly erodes the power of any of the tests that is done on the parameters of the model[42]. The latter is caused by the detrimental impact on the individual weights of having a large lag, i.e. for practical lag lengths it conflicts with *the principle of parsimony*.

[41] A Vector Auto-Regression (VAR) model is another standard econometric time series model for multi-variate analysis. It looks for autoregressive patterns in the history of the variables and their *common* past. Again emphasizing *autoregressive,* it uses linear correlation for its estimation. The point of mentioning this kind of model is that the approach of Sun et al. (2008) is similar, as they use a lag of 10 for their different approaches, though using a DNN.
[42] As the degrees of freedom are consumed by the estimation of n times x coefficients.

Therefore, we start from the *instantaneous VaR* at any time *t* and include the measures of roughness as explicit features to the model (see Chapter 4 for specific details). The rationale behind this choice should be clear by now, but Figure 24 below can make it more intuitive. The research question formulated in the introduction now comes down to testing whether roughness measures are informative. Roughness could be some sort of a missing link in a VaR combination model given the fallacies of many of the underlying assumptions of the standard models that were discussed in Chapter 1, and their link with the fBM generalization of the underlying stochastic data generating process. For instance, we could argue that Gaussian methods should get lower weights as the roughness increases, i.e. the Hurst exponent drops way below the Brownian 0.5 case. More aggressive versus conservative distributions might get a weight according to the measured deviation from the standard assumption of H = 0.5, as is shown in the figures below. Of course, this approach is heuristic and atheoretical in nature, but there are good arguments for the analogy on which the model hinges.

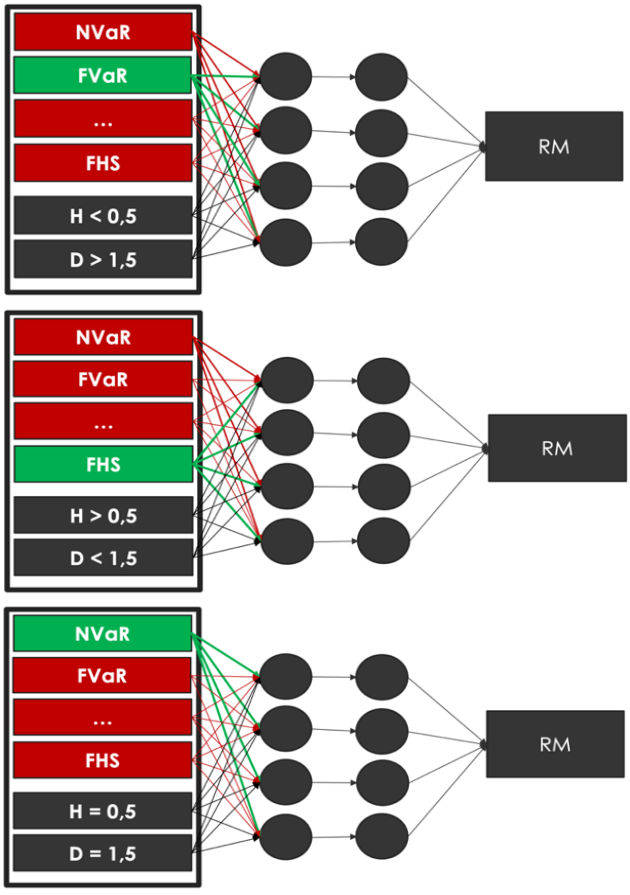


Figure 24: ML combination models – Roughness as a missing link?

So why might an ML model outperform our classical statistical model we started this chapter with? In the words of Sun, Rachev, Chen and Fabozzi:

> *"Why do the traditional time series models dramatically underperform the NN model? Standard time series models require the residuals to follow a normal distribution. Second, time series models do not have a memory while the NN model can memorize the [nonlinear] dynamics between the features and save it with hidden layers. The normal distribution for a random variable cannot capture the memory effect and the accuracy is reduced. Third, the NN model allows us to use very accurate computing methods, i.e. a recursive method referred to as backpropagation, while the typical maximum likelihood estimation for time series models is less accurate in its applicable algorithm."*

## 3.4 Genetic algorithms

*"To breed or not to breed, that is the question."*

It may not come as a surprise that building an ML model requires both science and art. One should be able to program the model, define the right loss function, etc. from some scientific motivation. However, there is no one single prescription on how to come up with the hyperparameters. Like it was argued in 3.1, the best hyperparameters depend on the data and the loss function one is working with. The only certainty is what we want to obtain: low loss, high accuracy without overfitting the sample data, and in terms of this dissertation: exceptions that are consistent with the confidence level. To obtain this goal by tuning the hyperparameters we have two options: brute force trial and error or genetic algorithms (GA) (see Carr, 2014). With *brute force* methods, you try every (sensible) combination of hyperparameters and wait a lifetime until the program spits out the best model. *Genetic algorithms* provide a much better and faster way. Say we start our brute force with a set of models with random hyperparameters. Why would we, in a next step, try models that have slightly different hyperparameters

from models that were least accurate? Why would we not only slightly alter the hyperparameters of the models that are performing best instead? This, in a brief, summarizes the reasoning behind GA.

**Genetic algorithms** - *commonly used to generate high-quality solutions to optimization and search problems by relying on bio-inspired operators such as mutation, crossover and selection.*—Wikipedia

*At its core, a genetic algorithm* (Carr, 2014)*:*

1. *Creates a population of (randomly generated) members*
2. *Scores each member of the population based on some goal. This score is called a fitness function.*
3. *Selects and breeds the best members of the population to produce more like them*
4. *Mutates some members randomly to attempt to find even better candidates*
5. *Kills off the rest - Survival of the Fittest - and*
6. *Repeats from step 2. Each iteration through these steps is called a generation.*

These steps are implemented in the code, where random neural nets are made from combinations of:

- 2, 4, 8, 16, 32, 64, 128, 256 or 512 neurons per layer
- 1, 2, 3 or 4 layers
- Relu, elu, tanh or sigmoid activation
- RMSprop, Adam, SGD, AdaGrad, AdaDelta, AdaMax or Nadam optimization

The algorithms are designed such that a bad parameter quickly disappears from the decision tree (or 'family tree') that the code goes through. Of course, the code always uses the same bespoke loss function, features and labels. The number of generations and number of models in the initial population is a choice that primarily hinges on the available computing power[43].

[43] This can be adapted in the form sheet the code starts from. The estimations shown in this document are based on 10 generations and an initial population of 10 networks (again, see emiellemahieu/AOR on GitHub).

## 3.5 Backtesting the model: a bespoke Kupiec-based loss function

A key point that was stressed in the previous sections, was that ML models can take any kind of loss function as long as it is differentiable. This is extremely useful in risk measurement, where we cannot simply compare a 'real' risk measure with a predicted risk measure and e.g. take the squared deviation. For instance, the closest thing to a real VaR at some point in time is the empirical distribution of the PnL in the future and can therefore only be assessed afterwards. It is not like one tries to predict house prices based on location and living space, where at any point in time one has both data on the features X and the label Y (= sample house prices). Due to its different nature, the integrity of risk models is tested differently than common regression frameworks (= variance). The most well-known framework for backtesting is the *Kupiec-Christoffersen framework* (Christoffersen, 2008; Kupiec, 1995).

The framework is mathematically very similar to binomial backtesting of exceptions. With a *cl*% confidence level, the probability of an exception is 1-*cl*%. Therefore, the occurrence of x exceptions on N observations can be tested binomially:

$$P(X = x) = \binom{N}{x}(1 - cl)^x cl^{N-x} \tag{3.5}$$

The ratio $x/N$ is also referred to as the *violation ratio*. The binomial distribution, however, is discrete. This means that we need shortcuts for the calculation of a p-value to test the hypothesis whether x is significant. One such a shortcut is the normal approximation of the binomial distribution. Luckily, the Kupiec test provides us with an alternative.

The *Kupiec test* (Kupiec, 1995), also called the *points-of-failure test*, uses a log-likelihood ratio (LR) based on the percentage of exceptions or violation ratio (x/N) and the *cl* (p = 1-*cl*), which is $\chi^2$ distributed with 1 degree of freedom. Therefore, we can now

calculate a p-score to assess whether the number of exceptions is statistically significant, i.e. the model likely to be incorrect:

$$LR_{POF} = -2\log\left(\frac{(1-p)^{N-x}p^{x}}{\left(1-\frac{x}{N}\right)^{N-x}\left(\frac{x}{N}\right)^{x}}\right) \sim \chi^{2}{}_{1} \tag{3.6}$$

The loglikelihood ratio under the points-of-failure test ($LR_{POF}$) is also referred to as the unconditional loglikelihood ratio ($LR_{uc}$). This test forms the theoretical background behind the Basel traffic-light assessment of internal models. Taken a 99% *cl* and 255 trading days, we expect 2,55 exceptions a year. The traffic-light refers to three zones a model can end up after the backtest: a green zone with up to 4 exceptions, an orange zone with 5-9 exceptions and a red zone with 10 exceptions or more. In the first case, the model is assumed to have no integrity problems. In the second case, the multiplier (that was shortly mentioned in 1.2) is increased as a penalty. Thus, the risk that was measured is increased before it is translated into RWA. In the red zone, intervention is needed. It indicates that there are certain quality and accuracy problems in the model.

The framework was extended by Christoffersen in 1998 with the concept of conditional coverage. For an overview, see Christoffersen, 2008. *Conditional coverage*[44], in contrast to the unconditional ratio, boils down to including the autocorrelation of exceptions in the framework. Like any accident, an exception does not come alone. Christoffersen therefore calculates an independent LR ($LR_{IND}$) which measures the serial independence of the data. Then he sums up the $LR_{POF}$ with the $LR_{IND}$ and gets the conditional LR ($LR_{cc}$).

$$LR_{cc} = LR_{IND} + LR_{uc} \tag{3.7}$$

Since these expressions are commonly used to assess the model after construction, it makes sense to include these expressions in the loss function so that the machine learns

[44] In the words of Christoffersen: "*An accurate VaR measure should satisfy both the unconditional coverage property and the independent property. The unconditional coverage property means that the probability of realizations of losses in excess of the estimated VaR must be exactly (1-cl)%. The independent property means that previous VaR violations do not presage future VaR violations.*" (Christoffersen, 2008)

to build a model that is optimal from this exception point of view instead of e.g. only looking at L2 loss (Liu, 2005). To achieve this end, a custom loss function was developed that compares the tensor of predicted VaRs with the actual returns. Exceptions in our loss function are thus defined as the number of times that the actual losses exceed the predicted VaR. Based on the violation ratio, the LR is calculated from the conditional and unconditional perspective. Since these ratios should ideally be insignificant, these ratios cannot be too high and they would thus qualify as a measurable quantity of loss. If we use this as loss[45], however, we find that this definition is not differentiable and leads to NaN loss. Therefore, we start from classical L2 loss where the predicted VaR is compared with the 'real' lookback empirical VaR and then add a penalty for an excessive violation ratio[46]. The latter approach is always differentiable and does not lead to NaN loss.

In summary, the concepts of this section are at the core of how our machine learns, i.e. the feedback it gets from giving our combination model weights depends on its current backtest results. The model is therefore expected to generate superior results – at least from a backtest perspective - if it is properly trained in-sample[47]. Whether this can be generalized out-of-sample, will be the main focus of the next chapter.

## 3.5 Conclusion

In this chapter we looked under the bonnet of machine learning models, focusing on the main design parameters we have in building neural network algorithms. We elaborated on the modelling opportunities ML models give to solve our combination problem. We particularly emphasized that the solution of our combination does not lie in the set of linear combinations of the input risk models. Building on the critique in chapter 1 and the insights of chapter 2, we provided the rationale behind a DNN VaR combination model with roughness as 'contextual' variable. The set-up and model implementation, as well as the results, will be explained in the next chapter.

[45] Or we could take the corresponding p-value by using a $\chi^2$-distribution.

[46] Note that the relative weight of the L2 loss compared to the Kupiec-based penalty for excessive violation can be tuned in the form (again, see emiellemahieu/AOR on GitHub.

[47] Unfortunately, this is not a trivial assumption, as the next chapter will soon illustrate.

# Chapter 4

# The model & results

## 3.1 Data & model set-up

*"There are no facts, only interpretations."*

Friedrich Nietzsche

The main design ideas behind the model have been discussed in chapter 2 and 3, culminating in the figures on page 81. This chapter describes the actual set-up and implementation of the model, as to ensure maximal reproducibility for further research. Again, note that the code is made available through GitHub [emiellemahieu/AOR][48]. After going through the model for one ticker, we will discuss the model for all 780 tickers and elaborate on the implications for the efficiency of risk and reward of the assets in our horizon. Essentially, the code follows the next steps:

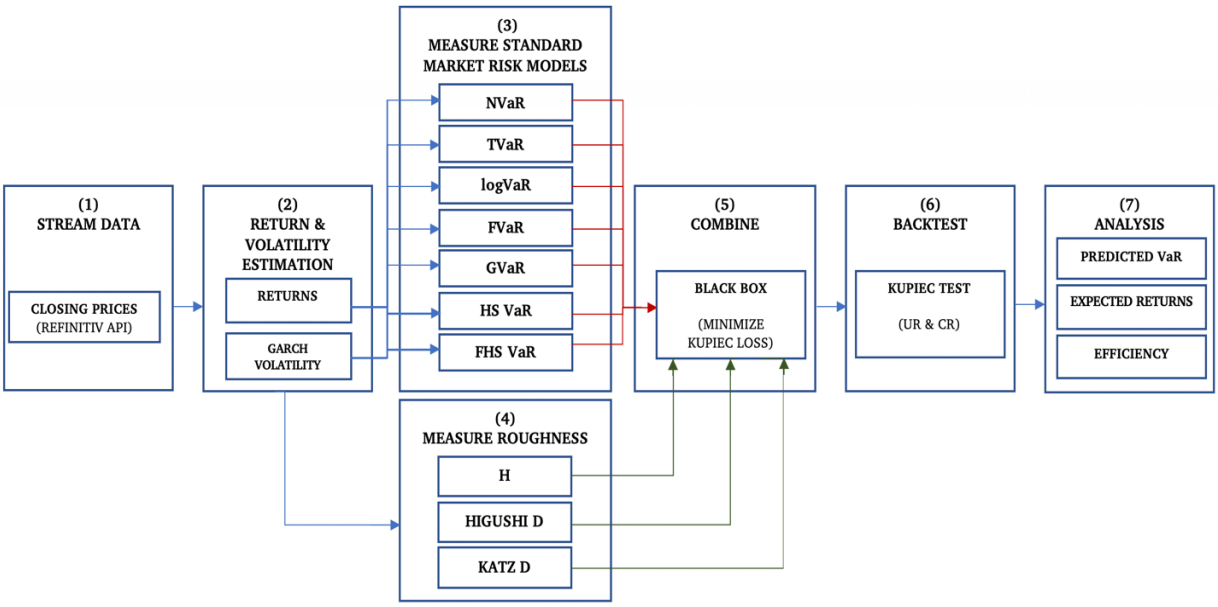


Figure 25: Model workflow

[48] The python script was developed on Google Colaboratory, a platform that was initially developed for internal use by Google's ML researchers but is now available for students and developers. The code is presented in the form of an IPython Jupyter notebook and the platform gives the possibility to run the code local or on the cloud (GCP) leveraging Google's GPUs (Graphics Processing Units) and TPUs (TensorFlow Processing Units).

(1) It takes a cross-section of 11 country codes (Belgium, Canada, France, Germany, Italy, Japan, Netherlands, Sweden, Switzerland, United Kingdom and the United States) and 10 sector codes (Finance, Technology, Utilities, Telecommunications, Consumer Services, Health Care, Consumer Goods, Industrials, Basic Materials and Oil & Gas) and fetches the constituents of these performance indices, i.e. the 10 biggest tickers according to market cap. However, not all countries have 10 large-cap stocks in all industries. E.g. Belgium only has 2 relevant stocks in the Oil & Gas industry. Therefore, we eventually end up with only 780 stocks. The code downloads OHLCV (Open-High-Low-Close-Volume) using the Refinitiv Eikon API for 15 years of data (1/3/2004 – 1/3/2019) on a daily basis. Below, an example is given for the YCD quote:

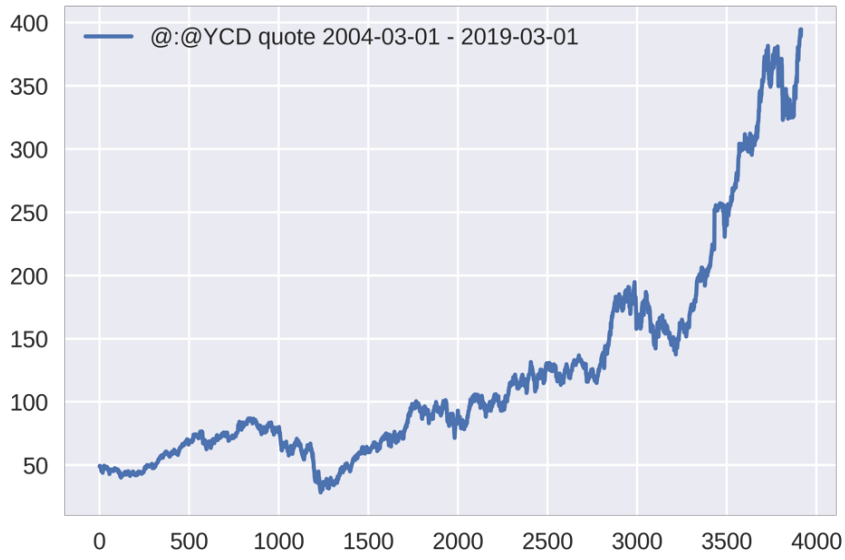


Figure 26: YCD (Refinitiv mnemonic for CDI or Christian Dior Industries) stock quote

(2) It then determines the daily geometric returns, $\log(P_t/P_{t-1})$, and estimates different GARCH processes for its vol. Below the returns are shown in Fig. 27, as well as the in-sample forecasts of a GARCH(1,1) process (cf. Equation 1.8).

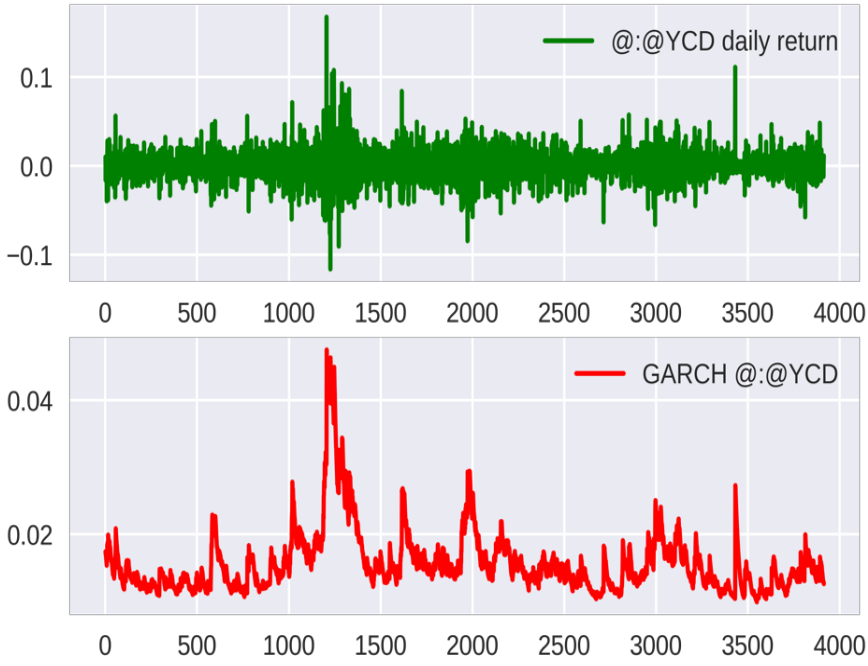


Figure 27: The corresponding return series (geometric returns) and estimated GARCH(1,1) process

Since we have a large time window of 15 years, the obvious feature of the time series of volatility is the financial crisis of 2007-2009. In the volatility cluster between observations 1000 and 1500, we notice the biggest daily returns, ranging from 15% gains to 12% losses. During all volatility spikes, we see that returns are more scattered and typically correspond with a downturn in the market. Indeed, when the persistence of volatility goes up, i.e. periods of continued higher sigma, we notice that the roughness of the price chart increases. These periods will be of particular interest for our machine learning model, since including this return, volatility and roughness data will enable the network to learn patterns that occur during financially distressed periods. Please find Table 7 below for the main price, return and volatility data and quantiles of the CDI stock.

Table 7: CDI price, return and volatility information

| | Close | High | Low | Open | Volume | Daily return | Daily volatility (GARCH) |
|---|---|---|---|---|---|---|---|
| Count | 3915 | 3915 | 3915 | 3915 | 3915 | 3915 | 3915 |
| Mean | 123.14 | 124.14 | 121.80 | 122.98 | 149.23 | 0.00053 | 0.0156 |
| Min | 28.44 | 29.00 | 26.89 | 28.31 | 1.90 | -0.12330 | 0.0098 |
| 25% | 65.93 | 66.95 | 65.39 | 66.25 | 71.85 | -0.00073 | 0.0127 |
| 50% | 93.59 | 95.82 | 93.54 | 94.79 | 119.00 | 0.00081 | 0.0144 |
| 75% | 154.40 | 153.90 | 150.95 | 152.22 | 186.00 | 0.00879 | 0.0168 |
| Max | 384.70 | 398.00 | 390.90 | 393.50 | 3283.00 | 0.1546 | 0.0475 |

(3) The code uses these return and volatility estimates - and, if required, other distributional parameters - to calculate the standard VaRs discussed in chapter 1 on a daily horizon with a 99% confidence level. Time-varying mean returns, kurtosis, skew etc. are calibrated over a lookback window of the past n observations[49].

(4) Next, the code estimates three measures of roughness for the past n daily prices. The Hurst exponent (H) of the daily price process, Higuchi and Katz' fractional dimensions (D) of the price charts are calculated based on the past n observations, where n corresponds to the lookback window defined for the recalibration of the standard VaRs. Therefore, we obtain time series of 'rolling' measures of roughness which are shown in Table 8.

Table 8: CDI rolling roughness exponents and standard risk measures

| | Hurst Exp | Higuchi D | NVaR | TVaR | logVaR | HS | FHS | FVaR | GVaR |
|---|---|---|---|---|---|---|---|---|---|
| Count | 3915 | 3915 | 3915 | 3915 | 3915 | 3915 | 3915 | 3915 | 3915 |
| Mean | 0.3678 | 1.5061 | 0.0360 | 0.0404 | 0.0352 | 0.0433 | 0.0404 | 0.1187 | 0.0726 |
| Min | 0.3156 | 1.3836 | 0.0214 | 0.0249 | 0.0212 | 0 | 0 | 0.0751 | 0.0461 |
| 25% | 0.3274 | 1.4844 | 0.0287 | 0.0331 | 0.0284 | 0.0335 | 0.0327 | 0.0972 | 0.0597 |
| 50% | 0.4108 | 1.5024 | 0.0328 | 0.0377 | 0.0322 | 0.0396 | 0.0394 | 0.1100 | 0.0674 |
| 75% | 0.4737 | 1.5325 | 0.0391 | 0.0447 | 0.0383 | 0.0456 | 0.0423 | 0.1272 | 0.0777 |
| Max | 0.5888 | 1.6125 | 0.1145 | 0.1307 | 0.1083 | 0.0806 | 0.1222 | 0.3544 | 0.2145 |

[49] This can be modified in the form fields the code starts from. For these results, it was set at 500 days.

(5) As was seen in the previous chapters, some of these VaR estimates are critically understating risk, while very conservative models drastically overstate projected losses. Looking at the estimates in Table 8, for instance, we see huge differences between the light- and fat-tailed approaches. This is because we work with a 99% VaR and the quantile we look for lies deep in the tail.

Therefore, the code combines the results of (3) and (4) in a neural network regression model that supports on elements from the TensorFlow and Keras frameworks[50]. The sample is split up, since cross-validation will be done for our model assessment. Hence, we split the data into a training set (in-sample or IS, which we will use to train our weights), a validation set (which we will use to tune our hyperparameters) and a testing or forecasting set (out-of-sample or OOS, for comparing the adequacy of our different models on unseen data). Given that we have almost 4000 observations (15 years with approximately 255 days a year), we train our model on approximately 2600 observations, which leaves us 700 observations for validation and testing each. The hyperparameters are determined by genetic algorithms (see 3.4) and the loss function is a custom loss function that combines the standard L2 loss with a penalizing factor for an excessive violation ratio (see 3.4). The result is a combined VaR model based on roughness, or a '*rough VaR model*'.

(6) Next, it backtests the trained rough VaR model and compares the results with the standalone models in terms of violation ratio, unconditional and conditional p-value (also see 3.4). In Table 9, we clearly see that in-sample (IS) our model uses the 2631 training features of VaRs and roughness to fit the 1% significance level nicely (31 exceptions compared to a theoretical 27). This is because the loss function penalizes any exception over the 27 threshold severely. However, in-sample this could be due to *overfitting*, a typical issue in machine learning, so we focus on out-of-sample (OOS) results. This is what risk management essentially is about - controlling *future* losses based on *historical* data - so we should not be too confident

[50] Again, for details, I warmly invite you to take a look at emiellemahieu/AOR on GitHub.

about in-sample results and be wary about overfitting. That is why we focus on forecasting power (OOS results) here and in the next section to compare models.

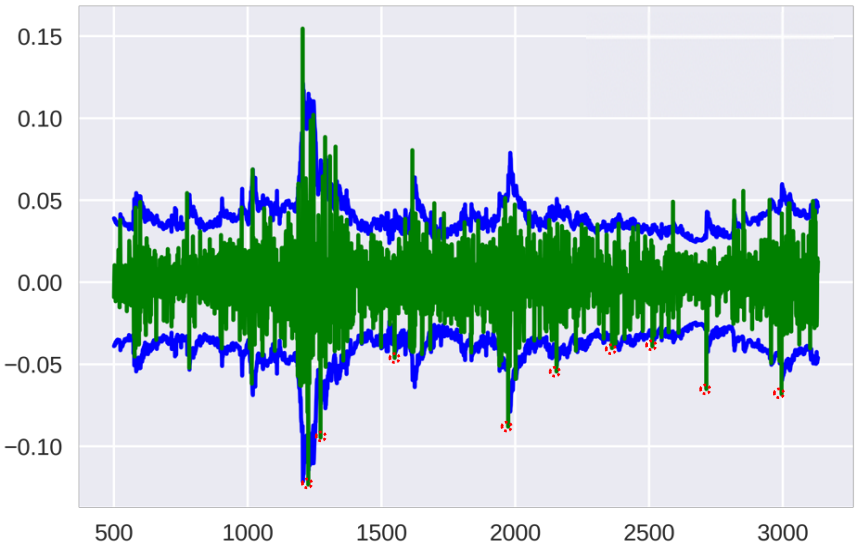


Figure 28: Backtesting of in-sample (IS) predicted 99% cl returns with indicated exceptions

In Table 9, we compare our *rough* model with a more aggressive (Gaussian) model and a more conservative (Gumbel) model. We find that our rough model yields 9 exceptions over 782 OOS returns, which corresponds to a violation ratio of 1,15% which is in line with the in-sample results. When we look at the unconditional loglikelihood ratio, we find that this number of exceptions is not significantly different from 8 (p-value = 0.6787).

Table 9: Backtesting the results using the Kupiec-Christoffersen loglikelihood framework

| | Rough model (IS) | | Rough model (OOS) | | Gaussian (OOS) | | Gumbel (OOS) | |
|---|---|---|---|---|---|---|---|---|
| Controlled | 2600 | | 773 | | 770 | | 781 | |
| Exception | 31 | | 9 | | 12 | | 1 | |
| total | 2631 | | 782 | | 782 | | 782 | |
| Violation Ratio | 0.0117 | | 0.0115 | | 0.0153 | | 0.0012 | |
| | Statistic | p-value | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| Unconditional | 0.7988 | 0.3714 | 0.1715 | 0.6787 | 1.9399 | 0.1637 | 9.5865 | 0.0019 |
| Independence | 0.7714 | 0.3797 | 0.2096 | 0.6471 | 1.8686 | 0.1716 | 0.0026 | 0.9596 |
| Conditional | 1.5703 | 0.4561 | 0.3811 | 0.8265 | 3.8085 | 0.1489 | 9.5891 | 0.0082 |

The autocorrelation of these 9 exceptions is not significant, so that we do not need to look at the conditional ratio. This implies that the 9 occurred exceptions were not consecutive losses in the same period of distress, which suggests that our algorithm is adaptive enough to recognize these types of patterns. When we compare with a Gaussian model, we find 12 exceptions, which is more but not problematic from the Kupiec-Christoffersen point of view (p-value = 0.1637). Gumbel, on the other hand, only yields one exception (p-value = 0.0019) and is therefore way to conservative to be a correct model.

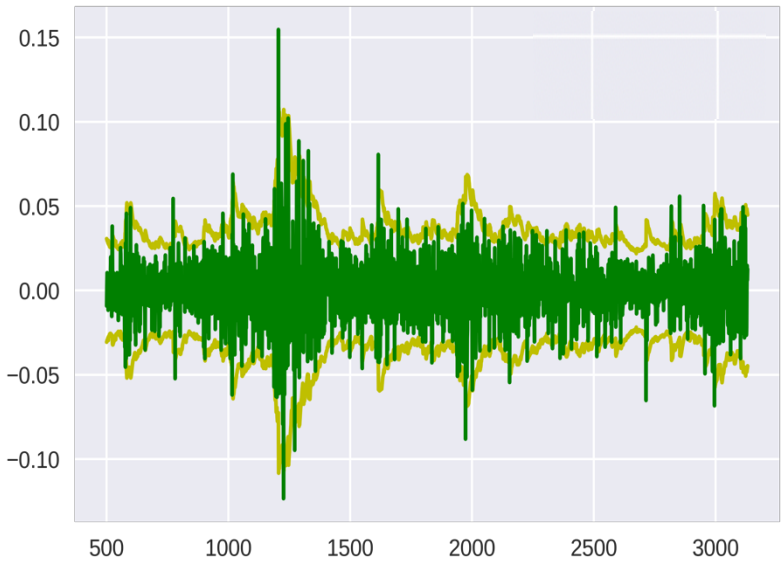


Figure 30: For comparison, backtesting results (IS) of another more aggressive model (logVaR)


Figure 29: For comparison, backtesting results (IS) of another more conservative model (TVaR with varying kurtosis)

(7) The final and seventh implemented step consists of the analysis of the risk-return trade-off implied by the model. Given the different nature of our model compared to standard portfolio theory, it is worthwhile to take a look at the implications of the model in terms of the efficiency of the assets in our horizon. These steps thus make the link with our model and section 1.11. Consequently, the code reports the VaRs and the expected returns on the stocks. The performance ratio compares the return with the risk measure to describe the stock's efficiency, i.e. what percentage return do we expect compared to the percentage loss we risk on the same horizon according to the model? The next section will delve into the meaning of these ratios, since the individual estimates do not say much compared to the aggregate results.

According to these performance ratios, stocks in a hypothetical portfolio could get a buy, hold or sell signal based on the quantile of their efficiency score. This essentially means that the return of the portfolio would relatively increase more than the portfolio VaR when buying, or the portfolio VaR relatively decreases more than the return when selling, thus increasing its efficiency (cf. 1.11). However, these signals are crude and therefore need to be more refined[51].

## 3.2 The link with ES and SRM

A short but crucial remark needs to be made on the link between our VaR prediction model and the coherent models discussed in chapter 1, i.e. Expected Shortfall (ES) and Spectral Risk Measures (SRM). Figure 31 illustrates that we can augment our rough VaR combination model by simply introducing a for-loop in the code that iterates the model over all the needed (tail) probabilities. As such, by adding an additional step 7 (a and b)[52], we see that the calculation of ES and/or SRM becomes possible. For ES, we will iterate the model over all tail confidence levels. The results are scenarios (7a), where we have projected losses with their probability (1-*cl*). Then we take a simple

[51] To question the practical relevance of these efficiency ratios, the only option is to test the profitability of cross-sectional long-short strategies based on these quantities. This will be briefly discussed in section 5.3.
[52] Note that 'Analysis' now becomes step 8.

average to know the average tail VaR or ES. For SRM, we do the same thing but typically over all (or at least more) probabilities. The weighting function then has to be defined based on the level of risk aversion by the user, but this will mostly depend on the application's purpose.

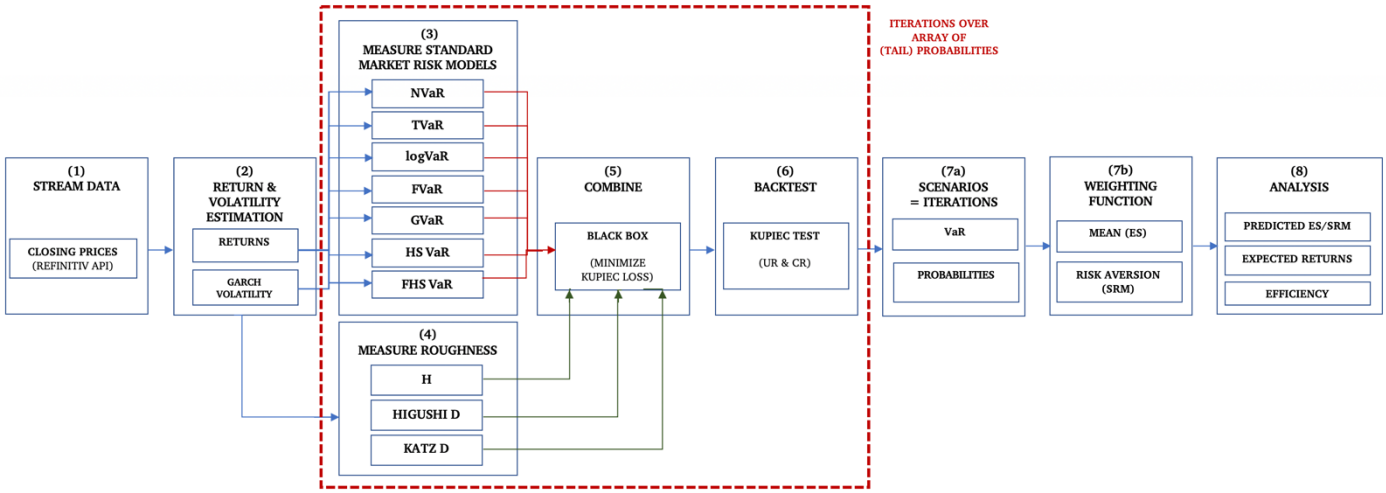


Figure 31: From VaR to ES and SRM

## 4.3 Results and discussion

### 4.3.1 Model performance

The first section of this chapter discussed only one example (CDI) of how the code combines different models into one combined measure using roughness. In order to be able to provide an answer to the research question, we need to reflect on the model's generalization and therefore need to look at the results for more than one ticker. In this section, we will compare the results of our model (referred to as *rough model*) to a model without roughness (referred to as *combination model*) to see whether roughness is a significant feature. We could also critically investigate the weights of our model directly but, given the complexity of the multiple layers, it is very hard, if not impossible, to draw any conclusions from these arrays. The motivations for opting for a black-box model were discussed in the previous chapter, but it is clear that the best counterargument against our approach is that there is opaqueness around how our

model learns as weights cannot be assessed straightforwardly. That is why, instead of focusing on these weights, we will be focusing on comparing backtesting results to see what our model has learned.

Table 10: Out-of-sample performance (1)

| **Total number of assets** | | 780 |
|---|---|---|
| **Proper training possible*** | 49% | 381 |
| **Proportion of significant models**** | **UR** | **CR** |
| Rough model | 8% | 8% |
| Combination model | 10% | 12% |
| Normal model | 20% | 25% |
| Gumbel model | 67% | 67% |

* Models are properly trainable if in-sample loss is not excessive (UR/CR not significant in-sample).

** Significant percentage of models of the 'properly trained models', from the unconditional (UR) and conditional (CR) perspective respectively.

Similar to Table 9, Table 10 summarizes the out-of-sample performances of our model compared to a more conservative (Gumbel) model and a more aggressive (normal) model. Moreover, a simple combination model without roughness is added to investigate the importance of roughness in our model. At first glance, the most remarkable feature of this table is that not all models were considered well-trained. From the 780 assets in our horizon, a meager 381 models were considered properly trained (49%). The other 51% had a significant in-sample loss, i.e. UR or CR is significant on the fitted training sample. We observed that when a network does not even manage to fit the training set properly, this is most likely due to a set of recurring data issues. As our model combines different individual models, it should always be at least as performant as the strongest model in the input features. In other words, even if our model would not be of any added value, it should then give 100% weight to the best individual model. However, in roughly half of the cases the model gave nonsense results in-sample while in the other half, where in-sample loss was not significant, the model outperformed the other models out-of-sample. Moreover, most of these issues occurred when there was no 15 years of data available. When a sample includes substantially less then 4000 observations, the number of available training days quickly erodes to impractical levels, as we also need a validation and testing sample. If you

only investigate a very limited set of tickers or indices, it is easy to guarantee that there is enough data available and that there are no severe data issues. This is what all ML papers with a similar set-up (i.e. financial time series) did. Given that roughness is both country, industry and asset specific (cf. 2.5), it would make no sense to make inferences after having trained the model on only a handful of indices. We need a broad universe of assets from different markets to question the model's generalization, but that inevitably leads to data being streamed in a more mechanical fashion and more data anomalies[53] occur. These data anomalies can lead to *butterfly effects* (cf. 5.5), such that our model does not converge to desirable results.

From the tickers that were properly trained, we find that only 8% still had a significant amount of exceptions according to both the conditional and unconditional loglikelihood ratio. This is a substantial improvement compared to the more aggressive normal method (which yields 20% and 25% erroneous models respectively) and the more conservative Gumbel method (which yields 67% for both UR and CR). The higher percentage of erroneous models is (1) due to too many exceptions in the normal case and (2) due to too few exceptions in the Gumbel case. This is shown in Table 11 where the average violation ratio of our model is compared with the simple combination

[53] These typically include (1) NaN traps, (2) (reverse) stock splits and/or (3) improper scaling.

(1) A *NaN trap* means that if some observations are unavailable and replaced by a NaN (Not a Number) value, a machine learning model can show anomalous results. Even if these NaN values are rare, the loss function can diverge away from an optimum or can result in so-called NaN loss.

(2) (*Reverse) stock splits* lead to (+)-100% returns, which will confuse the model severely because there is no pattern behind this sudden excess volatility. These are accounted for by replacing the closing prices in the model by a performance index (PI) that takes into account dividends and stock splits. However, most of the tickers that did not converge using closing prices also do not converge using performance indices. This brings us to the conclusion that these stocks splits are not that important.

(3) *Improper scaling* means that if the features are not correctly preprocessed, different scales for our input variables may lead to spurious variation. E.g. if daily returns vary with $\pm 0.02$, but Higuchi D varies with $\pm 0.20$ this leads the model to believe that there is ten times more variability in the latter variable. Spuriously, the model will attach more importance to Higuchi. That is why we first standardize the data in the preprocessing phase. However, some erroneous values seem to slip through the cracks and disturb the model severely.

model and the individual standard methods. We should not attach too much importance to an average and more to the likelihood ratios, but they give an indication of the models' overall biases.

Table 11: Out-of-sample performance (2)

| | **Significant UR*** | **Significant CR*** | **Average Violation ratio** |
|---|---|---|---|
| Rough | 8% | 8% | 0.7757% |
| Combination | 10% | 12% | 1.4728% |
| | | | |
| Normal | 20% | 25% | 1.2956% |
| Lognormal | 22% | 29% | 1.3741% |
| Gumbel | 67% | 67% | 0.1114% |
| Fréchet | 96% | 96% | 0.0154% |
| Historical simulation | 17% | 19% | 0.9475% |
| Filtered historical simulation | 17% | 19% | 1.0398% |

* Of the 381 models that were properly trainable (in-sample loss is not excessive, UR/CR not significant in-sample).

From these numbers, we are inclined to believe that an ML model nicely outperforms the individual models. Moreover, the difference between the rough model and the simple combination model is quite significant (8% compared to 10/12%). Averages might not be the best way to summarize the violation ratios of all models, but it is clear that our model fits the 1% violation ratio rather nicely.

A simple combination model clearly outperforms individual methods in terms of the significance of its violation ratios, but given the average ratio it tends to underestimate the risk slightly. We further see that normal and lognormal methods are quite similar in terms of performance. Gumbel and Fréchet methods yield dramatic results. Although they are based on the theoretically correct answer to our VaR problem (cf. 1.4.2), they more often than not generate quantiles that are never exceeded by the real losses. A case of no exceptions inevitably leads to a significant loglikelihood ratio, such that most of these models are useless. The simple historical simulation performs remarkably well. More sophistication clearly does not always imply superior performance. HS is both very easy to understand, implement and apparently works quite nicely too. There is,

however, no significant difference with the results of a filtered historical simulation. This leads us to suspect that volatility-weighting returns is not as useful out-of-sample as we would think beforehand.

All in all, these numbers illustrate that *roughness appears to be an informative feature for combining different risk measures*, as our rough model consistently outperforms the simple combination model. Table 11 sums up what we need to know to answer the research question, as it shows that over 381 models our model was consistently better at controlling the projected losses. However, one blatant limitation of our model quickly became apparent and ruined the party: machine learning models only outperform traditional models if they are properly trained, and that percentage was surprisingly low. We could argue that *if the model is well-trained, then our ML model will outperform standard models*. If we take a random stock, we might as well throw a coin and if it is head the ML model would train properly and outperform all other models. If it is tail, it would give nonsense results and be completely unreliable. This is not a very comforting conclusion for people that would mechanically use black-box models for certain applications. A point that will be emphasized in the next chapter, is that we should therefore be more interested in why our model converges in some cases and not in other cases, instead of being interested in mere backtesting results.

We should not be overly critical neither. The model clearly has its usefulness since we know that at any time, if the model is able to train properly on the available data up to that point, it is likely to be better at predicting future losses than other methods. Another insight is that, because of this issue of convergence, we could move away from training a model per ticker. We could train a model on a very long, hypothetical time series that is simply a concatenation of time series of many different stocks, collected from a pool of similar stocks according to some features. Then, we could use one trained 'supermodel' to make predictions for many other stocks. However, this was not the approach of this set-up and also poses new questions[54].

[54] Such as *'What stocks are similar?'*, *'What data can be concatenated without introducing new inconsistencies or data anomalies?'*, et cetera.

Whether the claimed contribution is *substantial* compared to a simple combination is up to debate, since there is no formal test to compare the percentages of Table 11 with each other. One way to look at these figures is to break down the percentage of significant rough models in the ones that were also significant for a simple combination model and the ones that were not, and vice versa.

Table 12: Breaking down the % of significant models

| **Model** | **UR** | **%** | **CR** | **%** |
|---|---|---|---|---|
| **Rough model** | **8%** | | **8%** | |
| Combination model | | 38% | | 35% |
| Normal model | | 26% | | 20% |
| Gumbel model | | 32% | | 34% |
| **Combination model** | **10%** | | **12%** | |
| Rough model | | 26% | | 27% |
| Normal model | | 30% | | 20% |
| Gumbel model | | 32% | | 32% |

Table 12 breaks down the 8% significant rough models in what percentage of these tickers were also problematic for the other models. These percentages are clearly quite low. Of the significant models in our model, only 38% (UR) was also significant in the simple combination model and 26 and 32 percent in the normal and Gumbel models respectively. This leads us to believe that the errors in the models are not perfectly correlated, and the improvements by our model will lead to *different conclusions*. We therefore argue that the additional precision is *not merely spurious precision*[55].

Finally, no clear-cut patterns were found between significant models and countries or industries.

[55] As we define *spurious precision* as a lack of making different decisions after the improvement of the model. For instance, a VaR of 12.3456% will not make any difference compared to a VaR of 12.3% for e.g. the decisions in section 4.3.2 and 5.3. Given that we find low correlations in errors and completely different numbers of exceptions (which reflects the overall level of VaR), we assume decisions would be different (e.g. allocation decisions, capital requirement, etc.).

### 4.3.2 Implications for risk and return

Let us now see what the model implies about risk-return trade-offs. Figure 32 plots the risk as measured by the daily VaR delivered by our rough model, compared to the average daily return over the same period. Alternatively: *'how much money do we get on average over a one-day period, and how much money can we expect to maximally lose over the same period in 99% of the cases?'*

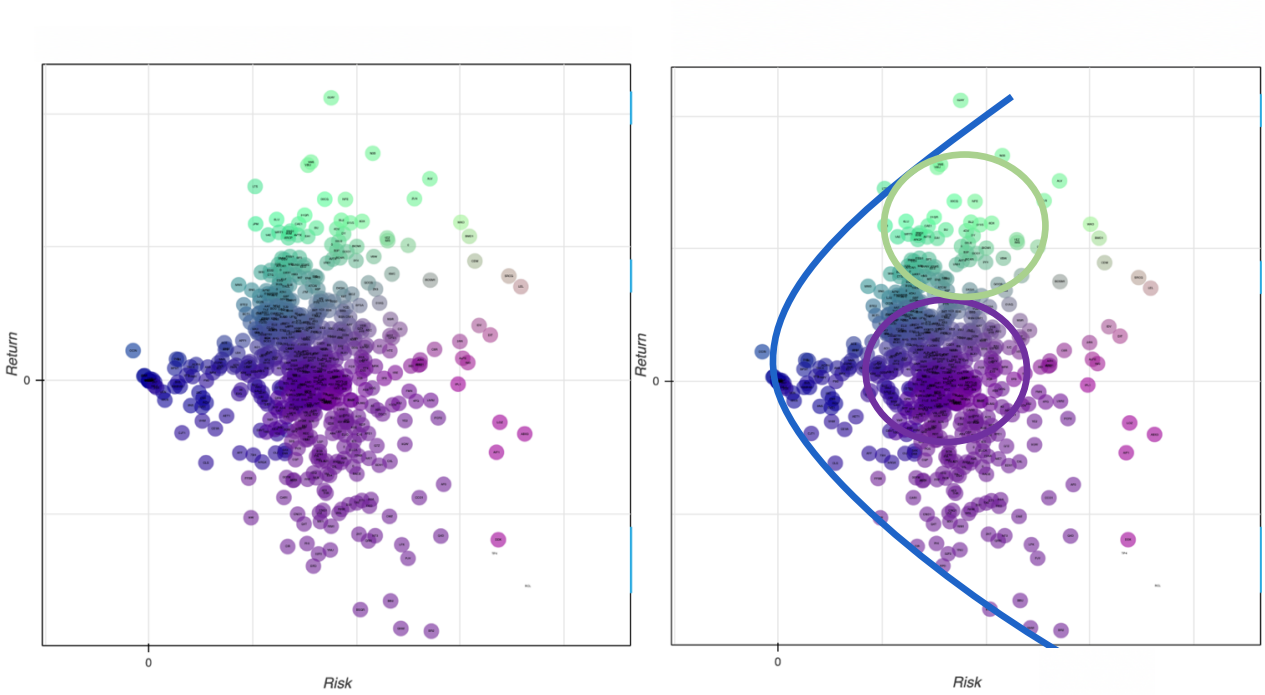


Figure 32: Risk (MVaR) and reward according to the model

With a bit of imagination, one can see the 'banana' shape we expect from an efficient frontier. More risky stocks deliver more extreme high or low returns, but the vast majority of returns is found in the region with close-to-zero returns and average riskiness. The color is derived from the performance ratio. The least efficient tickers are the purple ones, having moderate returns with high risk or outright abysmal returns. The blue cluster are stocks that have moderate returns but very low risk. Given the short time horizon, even low returns can compound to nice returns if the anticipated losses on that asset are very low. The most efficient region is the green region, there

our model predicts either nice returns with low risk or very nice returns with moderate risk.

In the panel on the right-hand-side, Figure 32 shows an alternative efficient frontier. The mean-variance trade-off as suggested by Modern Portfolio Theory is replaced by an expected return versus (marginal) VaR trade-off. Recall our discussion in 1.11, the Sharpe ratio of efficient portfolios in MPT is now replaced by ER/MVaR ratios.

The Sharpe ratio is often used as an ex post performance measure. However, the Sharpe ratio can also be seen as an a priori expected measure of performance, if we assume past return efficiency is indicative of its future efficiency. From the latter perspective, we can assess the future expected return efficiency of stocks from the perspective of our model using the ER/MVaR estimates delivered by the model.

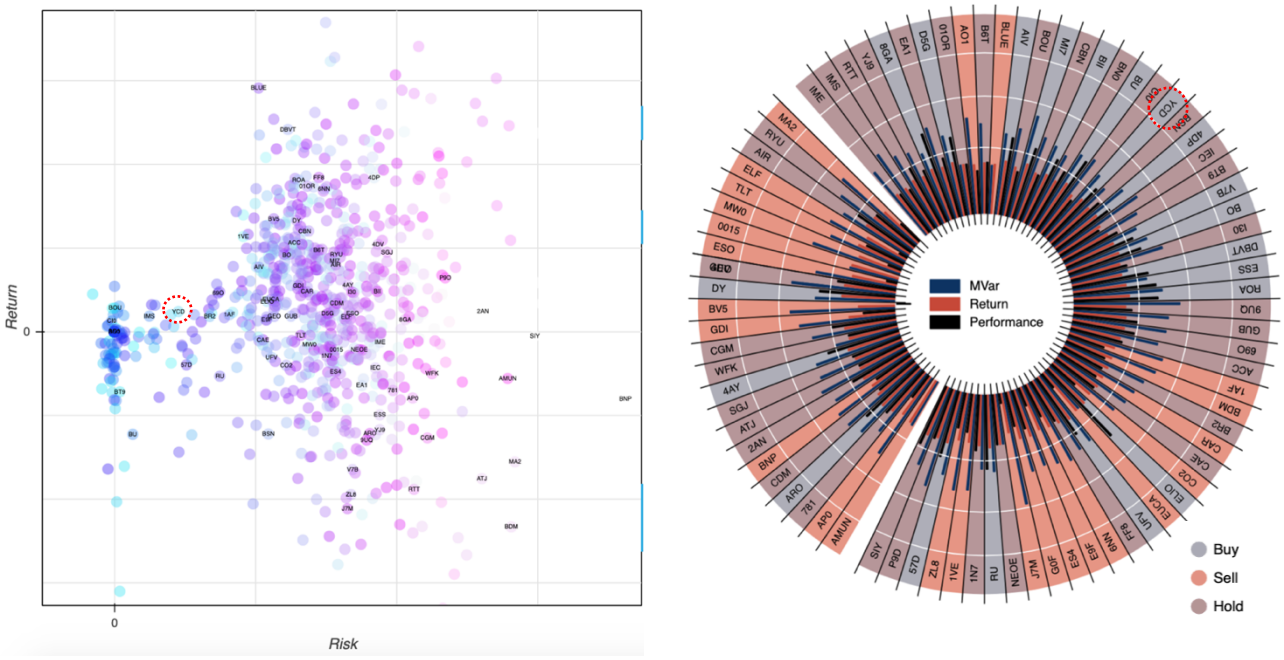


Figure 33: ER/MVaR efficiency, an example (France)

How can we increase the efficiency of a portfolio using the concepts in 1.11 and our model? This corresponds to migrating from the purple-circled region in Figure 32, i.e. the region where we most likely end up if we pick stocks at random or with gut feeling, to the green region. Since it is hard to visualize all assets of our universe at once, let us consider the efficiency of the French stocks in the sample. Figure 33 shows the average returns, MVaR and performance ratio characteristics of French stocks. On the right-

hand-side, we find a performance map that compares the French stocks. For illustration purposes, an additional dimension is added to the plot, namely a buy, sell and hold color. This is based upon three buckets[56] of performance: the top decile of performance ratios, the bottom decile and the middle majority of stocks respectively. On the left-hand-side, we plotted the risk-return characteristics where French stocks are indicated by their ticker. Our CDI stock from 4.1, for instance, performs quite nicely. On the left panel of Figure 33, we see that YCD offers low returns, but with a very low risk. Given the short time horizon (daily), even small returns can compound to substantial returns when the underlying risk is very low. This sort of efficiency clearly is the case for CDI, as can be seen from the right panel: it has lower than average returns with a measured risk way below average. That is why the performance ratio is one of the highest, as it crosses the first concentric circle.

The migration strategy proposed in the previous paragraph can be achieved by a cross-sectional long-short strategy, as we will discuss in 5.3. Given the completely different nature of this strategy compared to mean-variance optimization, it is expected to generate different Sharpe ratios and/or alpha than traditional mean-variance optimization. Whether this is a fruitful approach is still to be tested, as we will discuss in the next chapter.

## 4.4 Conclusion

In this penultimate chapter, we described the workflow behind the code of our roughness-based combination model. We first discussed the data set and the collection of methods and algorithms that were implemented in the code. Next, we elaborated on the link between our VaR model and the coherent risk models from chapter 1. We then delved into the results of all the 780 trained models. We compared the percentages of significant models of our rough model to a simple combination model and the other individual models. We concluded that our model works well, given that it is trained properly in-sample. We therefore expanded on some recurrent data issues behind this

[56] Of course, this division into buckets is crude and needs to be more refined in order for the cross-sectional long-short strategy to make sense (cf. 5.3), but it illustrates the idea.

observation. Moreover, we were realistic enough to acknowledge that the main reason behind this issue is the black-box principle underlying our model. We therefore have to conclude that the added value of this analysis is not only in the model but - maybe to a larger extent - in *the critical appraisal of machine learning models* and their limited convergence to desirable results. This thesis was therefore not only an exercise in modeling and programming, but even more an exercise in *model mindfulness and relativism*.

Apart from this skepticism, we can say that if we look at models that have learned something in-sample, our model almost guarantees outperformance. Our model's usefulness thus stems from the fact that*, if at any given time our model converges based on past information, it is likely to be better at controlling the PnL in the future than any other method*. We therefore came to the conclusion that *roughness is a significant feature*, which was the core aspect of our research question. Moreover, we find that erroneous ML models are not perfectly correlated with the other methodologies. This leads us to believe that *the improvements that we found for half of the tickers are not spurious*, since they imply that we would make different decisions. In addition, given the convergence of results when we have sufficient information, it would make sense to construct one single model that is trained on a concatenation of many time series. This introduces extra challenges, however, and is therefore subject to further research.

All things considered, it cannot be overemphasized that there is still a lot of ambiguity in the model. This is not surprising from a black box, but it has clear implications for model mindfulness. This point will be stressed again in the next chapter.

# Chapter 5
# Conclusions and recommended further research

## 5.1 Implications for risk managers

*"There is enough math in finance already. What is missing is imagination."*

Emanuel Derman

The implications for risk managers who model market risk and translate these measures in capital requirements (cf. 1.2) is epitomized by the following caveat: "*Markets are rougher than most people think.*"

A first obvious implication of the empirical work that was done in this thesis, is that the roughness implied by standard risk models is not consistent with real-life markets. Whether this means that risk managers should include Hurst exponents and fractional dimensions in their equations is another debate. The essential take-away is that this observation forces the risk modeler to be mindful about his models and their assumptions.

Secondly, a dynamic combination of models with varying degrees of aggressiveness into a combined measure is a useful approach, whether we include roughness or not. This quickly became apparent when we compared a simple combination model with individual standard methods.

Thirdly, this dissertation suggests that increasing roughness of the underlying stock process, or increasing persistence in the volatility process, is indeed highly correlated with market turbulence when more conservative models should get higher weights. That working hypothesis was confirmed by our backtesting results. This dissertation thus argues that *roughness contributes to a more effective combination*. Whether

roughness is superior to other 'contextual variables', however, is a conclusion that cannot be drawn from this work and calls for further research. There is a plethora of other modeling opportunities to determine those weights (e.g. including fundamentals of the underlying in question, including order flow data in high-frequency contexts, etc.). In summary, we are both optimistic and skeptical about the obtained results and their implications for practical risk management.

Fourthly, whether machine learning is the best way to combine those measures is also an interesting debate to which this dissertation contributed. We would be inclined to favor ML, because of (1) its adaptive nature, (2) the way it learns compared to standard statistical techniques (backpropagation versus least-squares or maximum likelihood), (3) the bespoke loss function which penalizes exceptions and therefore forces the model to be consistent with the confidence level and (4) simply because of its superior backtesting results. However, we are well-aware of its disadvantages (cf. 4.4) such that it is wishful thinking that a regulator would accept internal models that are based on black boxes (cf. 5.4). Black box risk measurement is therefore something that only financial institutions that are not bound by stringent capital requirement models are able to use in the short run.

Lastly, a rule of thumb that could be useful for risk managers, which was briefly discussed in the section on the link between roughness and finance (2.6), is to dilate projected losses by the $H^{th}$ power[57] of time instead of ½. This is a simple rule that nevertheless captures the underlying roughness of the process in an elegant way. This would mean that the *VaR surfac*e (see Dowd, 2007, p. 31, for an example) which shows all combinations of VaR, holding period (l-days) and confidence level (*cl*) would look different depending on the underlying roughness of the market.

Figure 34 illustrates these findings. It shows the normal VaR of a process with 0% expected return and 1% volatility on the Z axis, with the holding period (l days) and confidence level (cl) on the x- and y-axis respectively. Panel A, B and C show this

[57] Note that here we mean the H of the predicted VaR numbers and not of the price or volatility process. In other words, we apply R/S analysis to the time series of daily VaR to dilate VaR to an l-day period using l to the power of H. H would then capture the persistence in the series of losses. E.g. an accumulation of losses would increase H and the losses dilated by a bigger factor. This assumes that the roughness of the market is indirectly measured by the roughness of the risk measure.

surface for 3 values of H ( ½ , 0.30 and 0.60 respectively) from two angles (cl and l change axes for illustration purposes). Looking at the difference in order of magnitude of VaR between the different panels, it is obvious why H is important to take into account. If the persistence in losses increases, the error one makes in applying the Basel-compliant compliant $\sqrt{T}$-rule becomes really worrisome. This conclusion thus suggests that monitoring the roughness of the price process, the volatility process and the time series of VaR all have a different but complementary use in risk management applications.

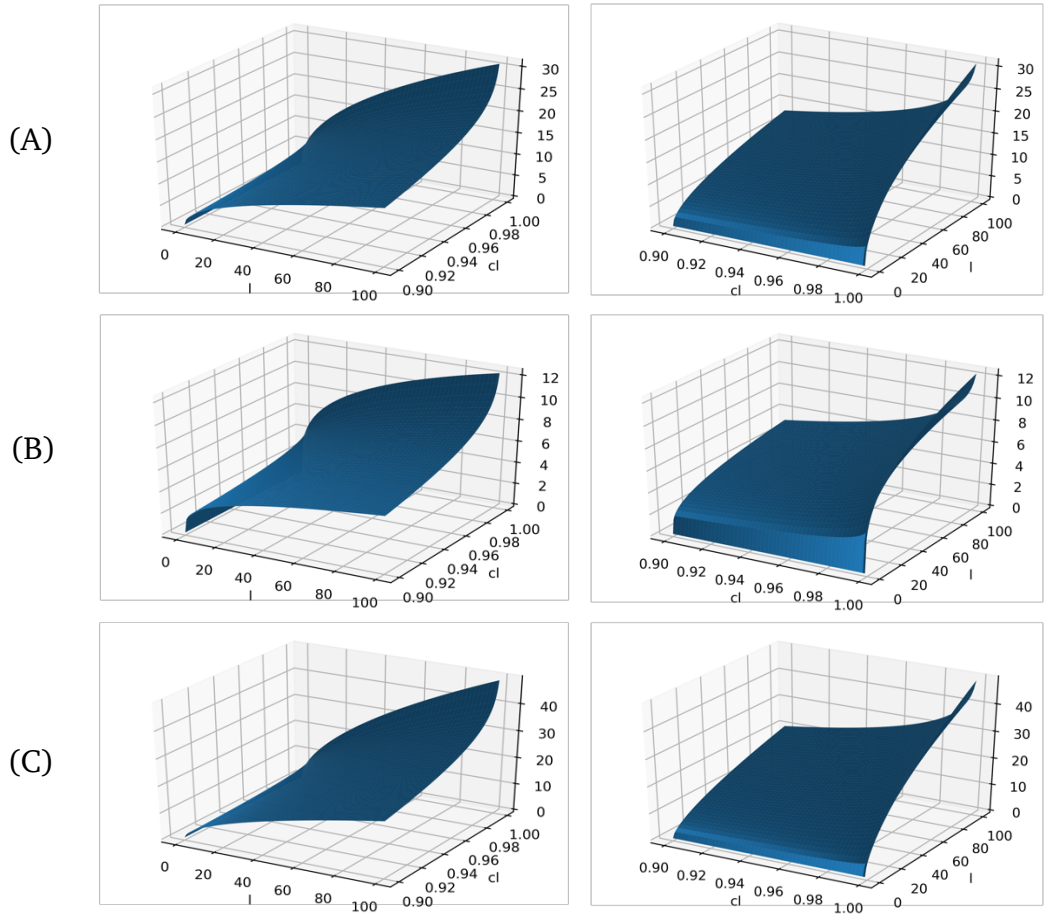


Figure 34: VaR surfaces and H

## 5.2 Implications for asset managers

*"Contrary to popular opinion, mathematics is about simplifying life, not complicating it."*

Benoît B. Mandelbrot

*"Make things as simple as they are, but not simpler*." is a truism that is commonly attributed to Albert Einstein and unfortunately applies here. Asset managers, even more than risk managers, are not interested in the statistical implications of research but in the practical ones. If a simple model performs well (e.g. HS), the marginal improvement of complicated statistical models is not worth the cost of trying to grasp that extra complexity. Although I am fully aware of this, there is a reason for risk managers to listen.

First and foremost, in line with the conclusion for risk managers, the empirical work done in this dissertation urges asset managers to be aware that markets are rough and that the uncontrollable element in risk models is typically understating or overstating the real risk. This should incentivize asset managers to (1) always be mindful about the models they use and (2) grasp the usefulness of combining different methods. Whether this leads them to use machine learning for this combination problem or simple rules of thumb, is fully up to them. The most important conclusion they should draw is to combine methods sensibly in the first place.

Secondly, the rule of thumb that was proposed in the previous section can be useful for asset managers too. For instance, it can give them insight in how big a loss can get on a bad position, given the period needed to get rid of that position, their current daily loss on that position and the historical H of the VaR time series of comparable assets in crisis periods. It is clear that this is just one example and a similar type of rationale can be applied on all kinds of similar estimation problems.

Thirdly, another more qualitative way to consider roughness and that would appeal to asset managers is to look at H and D as measures of market efficiency that can help for market and/or security selection. Given that H is directly linked to stock market

predictability, it is linked to weak-form market efficiency. Depending on the strategy (value investing, statistical arbitrage, etc.), different markets in terms of efficiency are typically sought after by asset managers. For example, more high-frequency applications like technical analysis and order flow analysis require less efficient markets like crypto markets. The empirical work that was done in this thesis provides a heuristic tool to distinguish between different levels of market efficiency across sectors and countries. E.g. Canadian Health-Care proved to be an inefficient market with extreme levels of persistence. There are already multiple studies that start from this perspective (Alvarez-Ramirez et al., 2008; Cajueiro and Tabak, 2004), but this dissertation clearly contributed to this perspective given our really broad scope of assets.

Fourthly, in line with the previous point, as roughness does not only differs across markets and assets but also in the time dimension, this has major consequences for the discussion on market efficiency. It is clear that markets are, on average, not efficient. However, market efficiency varies considerably over time, given that the dispersion in H and D for individual tickers is huge, and thus not due to noise in the estimators. This would provide evidence for an adaptive view on markets like the Adaptive Market Hypothesis (Lo, 2004). This is nothing new under the sun, as this link was already stressed by other authors. However, this dissertation provided more empirical evidence for time-varying roughness for a broad cross-section of stocks. The practical relevance for asset managers is therefore not only market/security selection based on the link between their strategy and market efficiency, but also for market timing[58].

Finally, we provided a rationale for assessing stock return efficiency that goes beyond mean-variance analysis and classical Sharpe ratios. This should trigger asset managers to reflect on how much they still rely on classical portfolio models and to what extent the vulnerabilities that were emphasized in this thesis apply to them. Although further research is required to make statements on the usefulness of the new frontier introduced in this dissertation, the different angle introduced in this thesis could instigate this thinking exercise.

[58] For instance, we will look for inefficient markets when we implement a simple moving average strategy. There the trick is to get in the market when inefficiency increases (large deviations for H/D). E.g. when central banks intervene and cause momentum, persistence will go up (H>0.5/D<1.5).

## 5.3 Implications for trading: using efficiency as alpha factors

*"All of us are drawn, like a moth to a flame, to high returns with low risk: a high Sharpe ratio."*

Andrew L. Lo

*Systematic* or *algorithmic trading*, also called *robo-trading*, regained a lot of attention in the ML era. In its early days, systematic trading focused on trade execution, i.e. machines optimizing execution in limit order markets from order data. Only later, machines became independent agents that trade autonomously using hard-coded technical analysis rules, typically in a high-frequency context. In this new era of artificial intelligence, these independent agents trade using reinforcement learning or adaptive algorithms instead of hard-coded rules. In this section, we will discuss a typical 'quant workflow' for constructing these algorithms based on the shared experience of the former CIO of the crowd-sourced hedge fund Quantopian. This platform provides an interactive development environment (IDE) for quants that develop trading algorithms using predefined APIs for constructing data pipelines, trade execution,

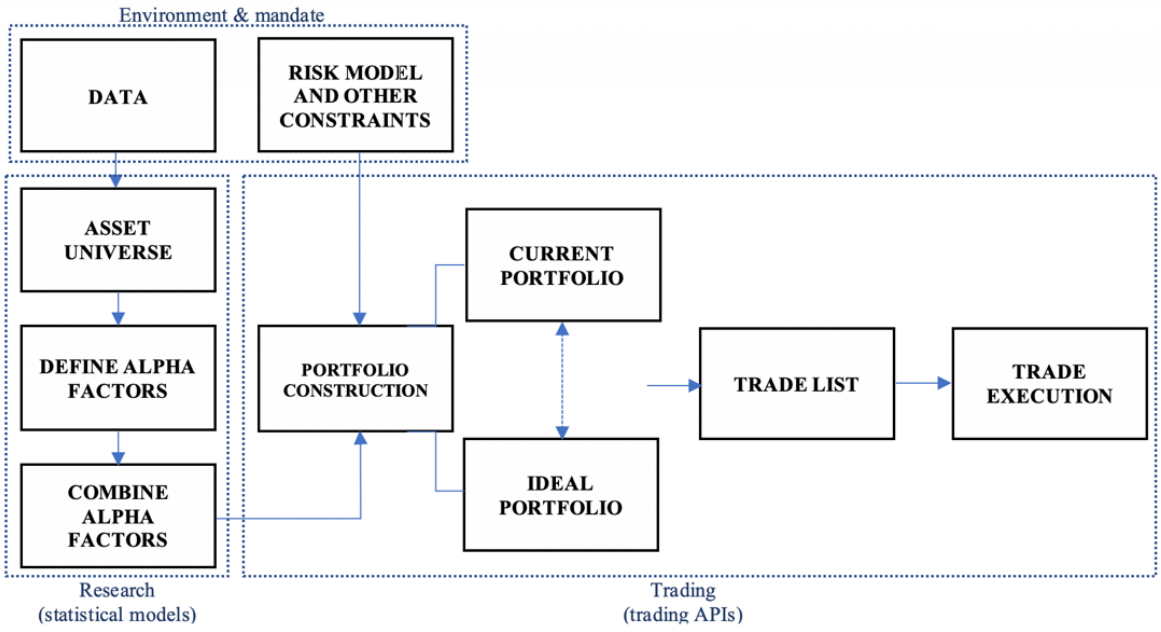


Figure 35: The Quant Workflow (based on Larkin, 2016)

backtesting strategies and so and so forth. While going through these steps, we will make the link with the efficiency ratios discussed in the previous chapter and reflect on how these can be implemented in an algorithmic trading strategy. Unfortunately, implementation and backtesting of such a strategy is beyond the scope of this dissertation, but this might be subject to further research.

The development process always starts from a certain *Environment*. For instance, the algorithm will trade under a certain *Mandate* (e.g. only long-short in equities and fixed income, limited to certain geographies, certain constraints on leverage and volatility, certain limitations on the concentration of certain buckets of risk and amount allocated to individual assets, etc.). Due to the nature of our earlier analysis, a cross-sectional long-short strategy makes most sense. *Cross-sectional long-short* means that we score assets according to some factor and go long the ones that perform best, and short the ones that score worst, thereby being dollar neutral. Furthermore, the environment also consists of the data sets one starts from. For instance, data can range from the public financial markets data used in this dissertation, to proprietary sets on collected Twitter and Google Search sentiment data.

We further distinguish between the *Research* and *Trading* environment. In the research environment we try to find interesting, novel ideas that translate into statistical relationships between stock returns, risk and/or their efficiency. This is therefore much like what this thesis implemented. We start by figuring out for what asset universes these relationships might apply best. For example, technical analysis strategies like moving averages models (the principles we discussed in 2.2), which try to exploit momentum, work better in FX markets than equity markets. Therefore, it is key to research for which level of market efficiency and at what frequency these strategies might be most lucrative. These relationships finally result into so-called alpha factors.

*Alpha factors* can be summarized as quantitative factors that are indicative of an asset's future return (or the return's efficiency). Indeed, the term alpha stems from classical portfolio analysis where returns are explained by market risk or other factors (betas) and an additional portion through managerial skill (alpha). These alpha factors are then gathered in a combined alpha factor, i.e. a score for every asset in the universe

denoting how attractive it is from the perspective of the statistical relationship that was found. Next, portfolio construction is done by dividing assets in quantiles according to their alpha factors. For our long-short strategy, these quantiles are then used to go long the x% best stocks in terms of efficiency ratio and to go short the worst quantiles. The difference between the desired portfolio and the current portfolio then results in a trade list that is executed automatically using the *Trading* APIs, which rebalance the portfolio with the frequency that is defined by the programmer.

Given these building blocks for developing algorithmic trading code, we can conclude that our approach to streaming equity price data and transforming it to efficiency ratios (expected excess returns over MVaR) already covers a large part of the Environment and Research building blocks. Therefore, in order to implement these ideas into a genuine strategy, we still need to integrate the existing code and ML frameworks with the Trading APIs. Given that these are typically written in Python, the integration can be done straightforwardly. However, we still need to define sensible constraints (in term of leverage, concentration, etc.) that make sense given our broad stock universe. These steps could be subject to further research (see 5.5).

## 5.4 Implications of model mindfulness

*"If there is any risk related to the role of humans being overwhelmingly replaced by AI, that would be when humans stop thinking independently and autonomously."*

Haruhiko Kuroda

For an innumerable amount of times, this dissertation was skeptical about the fact that ML models are black boxes which focus on results and lose transparency in terms of the why and the how. We quickly realized that the most blatant issue with the model was that sometimes it did not converge in-sample, given that a combination model should at least be as strong as the strongest individual model in the features. Upon critical investigation, these nonsense results appeared to stem from a set of recurrent

data issues. The disappointingly low percentage of converging models underscored the importance of being skeptical about the results. However, in contrast to a standard econometric regression framework, we cannot investigate the weights (i.e. the coefficients) and their significance using predefined tests. We resorted to mere backtest results to see whether roughness was a significant feature, but we never really understood why, apart from some theoretical conjectures on the link between roughness and risk models. We even went as far as making the link with algorithmic trading, such that trading decisions could be made based on this black box, i.e. black box trading.

As a result, it makes common sense that this dissertation ends with the caveat it began with in the abstract and the introduction. The presence of appropriate governance around the proposed algorithms is more important than the integrity of the models or code itself. Governance relates to clear roles and procedures around the development of these algorithms, i.e. people taking up responsibility when these algorithms start to 'misbehave'. Data can always be abused, no matter how 'objective' the code claims to be. It always comes down to reducing perverse incentives, or the incentives that stem from people using algorithms they know will only work well in the short run by fudging the risks they take. In this train of thought, Emanuel Derman's 'Models. Behaving. Badly.' (Derman, 2011) pretty much sums it up: "*The syntax of finance and physics have become very similar, but the semantics are very different*." Financial data is peculiar in the sense that it is not deterministic, such that one could show almost anything by using the models and tools of the exact sciences on noisy financial data. This corresponds to the disclaimer used in the introduction concerning models being analogies. Andrew Lo said: "*The difference between data and information is narrative, the story we tell by using the data*." and this is exactly what all ML models do (Lo, 2018). Financial data, unfortunately, do not always speak for themselves. Data do not have a voice, so the modeler has the freedom to impose almost any kind of narrative he wants. We used this freedom to test whether different angles to calculate VaR could be combined based on roughness. However, this is only one way to look at the world. It enables us to understand only part of reality. It should therefore be used accordingly. Hence, instead

of emphasizing good backtests it is more important to try to explain why some backtests were considerably worse. In the eloquent words of Emanuel Derman:

"*A model is just a toy. A good toy doesn't reproduce every feature of the real object but illustrates for its intended audience the qualities of the original object most important to them. Similarly, good models should aim to do only a few important things well.*"

In our case, we combined different models to reduce their individual bias and improve the overall backtests. That is a very clear objective for a machine learning model, such that there is no reason to assume that the algorithm understands context beyond roughness, let alone be good at anything else. Therefore, *Black Swan* events unknown to our system of input data, features and equations are likely to be handled as outliers, no matter how many crises we include in the training data and how well our code will adapt to new information. The most important limitation of our model is that the model is only as good as the quality and quantity of the available data. Data-heavy models naturally have a very data-dependent convergence of results. On this point, I would like to add a metaphor of my own, about a machine learning model and a genius kid:

"*An ML model is like a genius kid. It absorbs and retains information like no other kid. You could send her to a school where they teach quantum mechanics at the age of 6, she would play with it. However, we send our whizz-kid to a strange school; a school with no books or teachers, merely desks and walls. What we would find is that the kid would not seem to be the genius she really is."*

What sometimes went wrong with our model, to continue the analogy, is that, even worse than these strange schools, we send our model to a terrible school:

*"In this school, instead of no books, there are some books with wrong information. They tell the kid an orange is an apple, and an apple is a pear. Now the outside world will think of the kid as retarded."*

That is the problem with questionable data quality and machine learning models. The first school correspond to an environment where a powerful model has insufficient data, while the second school is a data set of dubious quality where a few anomalies ruin the whole thing. All things considered, we should see any model as a toy that represents only part of reality and emphasize its weaknesses (the qualities that were not modeled) more than its strengths.

## 5.5 Limitations of the set-up and recommended further research

*"There is a saying that bad traders divorce their spouse sooner than abandon their positions. Loyalty to ideas is not a good thing for traders, scientists - or anyone."*

Nassim N. Taleb

It is important to emphasize the limitations encountered during the thesis. I am not a mathematician, nor a computer scientist. Much like a fractal, one could zoom in on any section of this thesis and refine it, and again and again. *"Get the fundamentals down, and you will improve the quality of everything you do."* is a popular saying. Mathematicians could easily improve and add to chapters 1 & 2, by including more models and by calibrating them more professionally. Computer scientists can undoubtedly make the code more correct and efficient, leveraging parallel computing or even quantum computing.

Moreover, the code itself has some clear limitations. Very data-heavy models like our ML model mean very data-dependent convergence of results. This can be seen as a *butterfly effect*. The old analogy of a butterfly swinging its wings, which causes a tornado many miles away, really applies here. Strange singletons in the data, like reverse stock splits causing a sudden onetime +100% return can cause the model to show anomalous results. The irony that our model works best for benign datasets is that this was our initial critique on Gaussian and historical simulation models. The difference, however, is that when we include many crisis periods, our model learns how to better cope with it, in contrast to these static models. In the case of our model, we are not talking about LPHI events in the *real* return data, but anomalies like data quality issues (e.g. unexpected NaN values instead of prices).

The different angle of this thesis allows for further research. First of all, the real value of the obtained efficiency scores (compared to the traditional Sharpe ratio) can only be discovered through the backtesting of strategies based on the concept (cf. 5.3). Secondly, marginal VaR is linked to our common perceptions of an asset's beta (Jorion,

2000). If our approach towards calculating MVaR indeed appears to be useful, one could also wonder what this means for the *implied beta*. Is it possible to reverse engineer more 'sophisticated' betas from our predicted MVaR? If that would be possible, how effective would a dynamic beta strategy[59] be for tactical asset allocation or security selection?

Another set of models that were very briefly discussed are copula models. Copula models allow to model the dependence structure between variables in an elegant way. They were highly discredited after the CDO crisis, since the Gaussian version was an essential ingredient in the pricing of CDOs. This was mainly due to the fact that they used the simplest (Gaussian) version with only one static correlation measure. In general, however, copula models stay a very powerful tool to get a sense of the comovement structure between variables. We could wonder how we can implement copulas in combination with neural networks so that the comovement structure between our VaR features, which is implicitly understood by the machine, can be better understood by the modeler.

Apart from these suggestions, I am quite keen to know how the model could be more refined from an overall perspective. How can one improve the efficiency of the model as it is today? How can it be speeded up to use in more real-time application? Furthermore, I used quite generalized hyperparameters for a simple DNN regressor. Maybe a completely different type of network would work better? There is a vast field of research devoted to the taxonomy of neural nets, where more appropriate architectures for this problem probably already exist. One of the most promising types are so-called GANs, or Generative Adversarial Networks, where different neural nets compete against each other, based on the principles of game theory, as to increase the overall model's performance. Another interesting field are fuzzy logic models. Normally, computers (and thus neural nets) save data in ones and zeros. Fuzzy logic can be used to temporarily save data in a similar way to probabilities, i.e. a number between 0 and 1. So-called Genetic Fuzzy Neural Networks (GFNN) use fuzzy

[59] I.e. measuring the time-varying beta using this model, and adjusting the beta of the portfolio based on the strength of some market signal.

reasoning and could offer extra modeling opportunities that might be relevant for our problem. Lastly, reinforcement learning is doing a great job in improving AI models. For instance, think about AlphaGo's victory against the world's best Go player. In finance, reinforcement learning is, inter alia, used to optimize utility functions for agents based on risk and returns. Algorithmic trading already uses these concepts for autonomous trading agents. In these settings, so-called Q-learning defines rewards and losses that reinforce the algorithms, thus enabling a learning process. Very similar to our bespoke loss function, exceptions could be defined as losses and controlled losses can get rewards in a so-called Q-table. Therefore, integration of the concepts in this dissertation with Q-learning where agents have a more comprehensive sense of risk, might be a very interesting research path to pursue in the future.

# References


Acerbi, C., 2002. Spectral measures of risk: A coherent representation of subjective risk aversion. J. Bank. Finance 26, 1505–1518.

Alexander, G.J., Baptista, A.M., 2002. Economic implications of using a mean-VaR model for portfolio selection: A comparison with mean-variance analysis. J. Econ. Dyn. Control 26, 1159–1193.

Allen, L., Boudoukh, J., Saunders, A., 2009. Understanding market, credit, and operational risk: the value at risk approach. John Wiley & Sons.

Alvarez-Ramirez, J., Alvarez, J., Rodriguez, E., Fernandez-Anaya, G., 2008. Time-varying Hurst exponent for US stock markets. Phys. Stat. Mech. Its Appl. 387, 6159–6169. https://doi.org/10.1016/j.physa.2008.06.056

Andreev, A., Kanto, A., Malo, P., 2005. On closed-form calculation of CVaR.

Artzner, P., Delbaen, F., Eber, J.-M., Heath, D., 1999. Coherent measures of risk. Math. Finance 9, 203–228.

Bachelier, L., 1900. Theory of speculation. Dimson E M Mussavian 1998 Brief Hist. Mark. Effic. Eur. Financ. Manag. 4, 91–193.

Baillie, R.T., Bollerslev, T., Mikkelsen, H.O., 1996. Fractionally integrated generalized autoregressive conditional heteroskedasticity. J. Econom. 74, 3–30.

Barone-Adesi, G., Giannopoulos, K., Vosper, L., 2002. Backtesting derivative portfolios with filtered historical simulation (FHS). Eur. Financ. Manag. 8, 31–58.

Bauer, C., 2000. Value at risk using hyperbolic distributions. J. Econ. Bus. 52, 455–467.

Bayer, C., Friz, P., Gatheral, J., 2016. Pricing under rough volatility. Quant. Finance 16, 887–904.

BCBS III, B., 2017. Finalizing post-crisis reforms. December.

Beder, T.S., 1995. VAR: Seductive but dangerous. Financ. Anal. J. 51, 12–24.

Berkowitz, J., O'Brien, J., 2002. How accurate are value-at-risk models at commercial banks? J. Finance 57, 1093–1111.

Black, F., Scholes, M., 1973. The pricing of options and corporate liabilities. J. Polit. Econ. 81, 637–654.

Bolland, P.J., Connor, J.T., Refenes, A.P., 1998. Application of neural networks to forecast high frequency data: foreign exchange. Nonlinear Model. High Freq. Financ. Time Ser. Wiley.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. J. Econom. 31, 307–327.

Brandolini, D., Colucci, S., 2012. Backtesting value-at-risk: a comparison between filtered bootstrap and historical simulation. J. Risk Model Valid. 6, 3.

Broda, S.A., Paolella, M.S., 2011. Expected shortfall for distributions in finance, in: Statistical Tools for Finance and Insurance. Springer, pp. 57–99.

Brooks, C., 2019. Introductory econometrics for finance. Cambridge university press.

Butler, J.S., Schachter, B., 1997. Estimating value-at-risk with a precision measure by combining kernel estimation with historical simulation. Rev. Deriv. Res. 1, 371–390.

Cajueiro, D.O., Tabak, B.M., 2004. The Hurst exponent over time: testing the assertion that emerging markets are becoming more efficient. Phys. Stat. Mech. Its Appl. 336, 521–537.

Campbell, R., Huisman, R., Koedijk, K., 2001. Optimal portfolio selection in a Value-at-Risk framework. J. Bank. Finance 25, 1789–1804.

Carr, J., 2014. An introduction to genetic algorithms. Sr. Proj. 1, 40.

Castillo, O., Melin, P., 2002. Hybrid intelligent systems for time series prediction using neural networks, fuzzy logic, and fractal theory. IEEE Trans. Neural Netw. 13, 1395–1408.

Cervantes-De la Torre, F., González-Trejo, J.I., Real-Ramirez, C.A., Hoyos-Reyes, L.F., 2013. Fractal dimension algorithms and their application to time series associated with natural phenomena, in: Journal of Physics: Conference Series. IOP Publishing, p. 012002.

Christoffersen, P., 2008. Backtesting.

Clarke, R., De Silva, H., Thorley, S., 2011. Minimum-variance portfolio composition. J. Portf. Manag. 37, 31.

Cont, R., 2007. Volatility clustering in financial markets: empirical facts and agent-based models, in: Long Memory in Economics. Springer, pp. 289–309.

DasGupta, B., Schnitger, G., 1993. The power of approximating: a comparison of activation functions, in: Advances in Neural Information Processing Systems. pp. 615–622.

De Haan, L., Ferreira, A., 2007. Extreme value theory: an introduction. Springer Science & Business Media.

Decamps, J.-P., Rochet, J.-C., Roger, B., 2004. The three pillars of Basel II: optimizing the mix. J. Financ. Intermediation 13, 132–155.

Derman, E., 2011. Models. Behaving. Badly.: Why Confusing Illusion with Reality Can Lead to Disaster, on Wall Street and in Life. Simon and Schuster.

Dickey, D.A., Fuller, W.A., 1979. Distribution of the estimators for autoregressive time series with a unit root. J. Am. Stat. Assoc. 74, 427–431.

Dierick, F., Pires, F., Scheicher, M., Spitzer, K.G., 2005. The New Basel Capital framework and its implementation in the European Union.

Ding, J., Meade, N., 2010. Forecasting accuracy of stochastic volatility, GARCH and EWMA models under different volatility scenarios. Appl. Financ. Econ. 20, 771–783.

Dowd, K., 2007. Measuring market risk. John Wiley & Sons.

Duchin, R., Levy, H., 2009. Markowitz versus the Talmudic portfolio diversification strategies. J. Portf. Manag. 35, 71.

Engle, R., 2001. GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics. J. Econ. Perspect. 15, 157–168. https://doi.org/10.1257/jep.15.4.157

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econom. J. Econom. Soc. 987–1007.

Engle, R.F., Patton, A.J., 2007. What good is a volatility model?, in: Forecasting Volatility in the Financial Markets. Elsevier, pp. 47–63.

Esteller, R., Vachtsevanos, G., Echauz, J., Litt, B., 2001. A comparison of waveform fractal dimension algorithms. IEEE Trans. Circuits Syst. Fundam. Theory Appl. 48, 177–183.

Fama, E.F., 1998. Market efficiency, long-term returns, and behavioral finance. J. Financ. Econ. 49, 283–306.
Fama, E.F., 1965. Portfolio analysis in a stable Paretian market. Manag. Sci. 11, 404–419.
Farag, H.M., 2017. Bracing for the FRTB: Capital, business and operational impact. J. Secur. Oper. Custody 9, 160–177.
Frömmel, M., 2013. Portfolios and investments. BoD–Books on Demand.
Gatheral, J., 2017. Rough volatility: An overview by Jim Gatheral - YouTube [WWW Document]. URL https://www.youtube.com/watch?v=gW073Tnx7CE&t=732s (accessed 4.2.19).
Gatheral, J., Jaisson, T., Rosenbaum, M., 2018. Volatility is rough. Quant. Finance 18, 933–949.
Gatheral, J., Jaisson, T., Rosenbaum, M., n.d. Volatility is rough. Preprint, 2014. ArXiv Prepr. ArXiv14103394.
Giudici, P., 2005. Applied data mining: statistical methods for business and industry. John Wiley & Sons.
Granger, C.W., Joyeux, R., 1980. An introduction to long-memory time series models and fractional differencing. J. Time Ser. Anal. 1, 15–29.
Guttentag, J.M., Herring, R.J., 1997. Disaster myopia in international banking. J Repr. Antitrust Econ 27, 37.
Hallerbach, W.G., 1999. Decomposing portfolio value-at-risk: A general analysis. Tinbergen Institute Discussion Paper.
Hannoun, H., 2010. The Basel III capital framework: a decisive breakthrough. BIS Hong Kong.
Heston, S.L., 1993. A closed-form solution for options with stochastic volatility with applications to bond and currency options. Rev. Financ. Stud. 6, 327–343.
Hosking, J.R., 1984. Modeling persistence in hydrological time series using fractional differencing. Water Resour. Res. 20, 1898–1908.
Hull, J., White, A., 1998. Value at risk when daily changes in market variables are not normally distributed. J. Deriv. 5, 9–19.
Hurst, H.E., 1956. Methods of using long-term storage in reservoirs. Proc. Inst. Civ. Eng. 5, 519–543.
Hurst, H.E., 1952. The Nile: a general account of the river and the utilization of its waters.
Inui, K., Kijima, M., Kitano, A., 2005. VaR is subject to a significant positive bias. Stat. Probab. Lett. 72, 299–311.
Jacquier, A., Pakkanen, M.S., Stone, H., 2018. Pathwise large deviations for the Rough Bergomi model. J. Appl. Probab. 55, 1078–1092.
Jaschke, S.R., 2001. The Cornish-Fisher-expansion in the context of Delta-Gamma-Normal approximations. SFB 373 Discussion Paper.
Jorion, P., 2000. Value at risk.
Kahneman, D., Slovic, S.P., Slovic, P., Tversky, A., 1982. Judgment under uncertainty: Heuristics and biases. Cambridge university press.
Kan, R., Zhou, G., 2007. Optimal portfolio choice with parameter uncertainty. J. Financ. Quant. Anal. 42, 621–656.
Kotz, S., Nadarajah, S., 2000. Extreme value distributions: theory and applications. World Scientific.
Kupiec, P., 1995. Techniques for verifying the accuracy of risk measurement models. FEDS Pap.

Kupiec, P.H., 1999. Risk capital and VaR. J. Deriv. 7, 41–52.
Liu, Y., 2005. Value-at-risk model combination using artificial neural networks. Emory Univ. Work. Pap. Ser.
Lo, A., 2018. Quantcast 29/08/2018 Andrew Lo – Risk.net, Quantcast by Risk.net.
Lo, A.W., 2004. The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. J. Portf. Manag. Forthcom.
Malevergne, Y., Sornette, D., 2004. Value-at-Risk-efficient portfolios for a class of super- and sub-exponentially decaying assets return distributions. Quant. Finance 4, 17–36.
Mandelbrot, B., 2010. Benoit Mandelbrot: Fractals and the art of roughness - YouTube [WWW Document]. URL https://www.youtube.com/watch?v=ay8OMOsf6AQ&t=1048s (accessed 4.1.19).
Mandelbrot, B., 2002. Gaussian self-affinity and fractals: Globality, the earth, 1/f noise, and R/S. Springer Science & Business Media.
Mandelbrot, B., 1972. Certain Speculative Prices"(1963). J. Bus. 45, 542–543.
Mandelbrot, B., 1967. How long is the coast of Britain? Statistical self-similarity and fractional dimension. science 156, 636–638.
Mandelbrot, B.B., 2013. Fractals and scaling in finance: Discontinuity, concentration, risk. Selecta volume E. Springer Science & Business Media.
Mandelbrot, B.B., Hudson, R.L., 2010. The (mis) behaviour of markets: a fractal view of risk, ruin and reward. Profile books.
Mandelbrot, B.B., Van Ness, J.W., 1968. Fractional Brownian motions, fractional noises and applications. SIAM Rev. 10, 422–437.
Manyika, J., 2017. A future that works: AI, automation, employment, and productivity. McKinsey Glob. Inst. Res. Tech Rep.
Markowitz, H., 1952. The utility of wealth. J. Polit. Econ. 60, 151–158.
Markowitz, H.M., 1991. Foundations of portfolio theory. J. Finance 46, 469–477.
Matsuda, K., 2004. Introduction to Merton jump diffusion model. Dep. Econ. Grad. Cent. City Univ. N. Y.
McNeil, A.J., 1999. Extreme value theory for risk managers. Departement Math. ETH Zent.
Meissner, G., 2015. The Pearson Correlation Model–Work of the Devil? Retrieved 4, 2018.
Meissner, G., 2013. Correlation Risk Modeling and Management: An Applied Guide including the Basel III Correlation Framework-With Interactive Models in Excel/VBA. John Wiley & Sons.
Merton, R.C., 1976. Option pricing when underlying stock returns are discontinuous. J. Financ. Econ. 3, 125–144.
Merton, R.C., 1973. An intertemporal capital asset pricing model. Econometrica 41, 867–887.
Nathan, C., 2015. Benoit Mandelbrot: A Life in Many Dimensions. World Scientific.
Olah, C., n.d. Neural networks, manifolds, and topology, 2014. URL Httpcolah Github Ioposts2014-03-NN-Manifolds-Topol.
Phillips, P.C., Perron, P., 1988. Testing for a unit root in time series regression. Biometrika 75, 335–346.
Pratt, J.W., Zeckhauser, R.J., 1987. Proper risk aversion. Econom. J. Econom. Soc. 143–154.
Pritsker, M., 2006. The hidden dangers of historical simulation. J. Bank. Finance 30, 561–582.
Ruder, S., 2016. An overview of gradient descent optimization algorithms. ArXiv Prepr. ArXiv160904747.

Ryeu, J.K., Aihara, K., Tsuda, I., 2001. Fractal encoding in a chaotic neural network. Phys. Rev. E 64, 046202.

Schöbel, R., Zhu, J., 1999. Stochastic volatility with an Ornstein–Uhlenbeck process: an extension. Rev. Finance 3, 23–46.

Sharpe, W.F., 1994. The sharpe ratio. J. Portf. Manag. 21, 49–58.

Smith, R.L., 1990. Extreme value theory. Handb. Appl. Math. 7, 437–471.

Steeb, W.-H., 1999. The nonlinear workbook: chaos, fractals, cellular automata, neural networks, genetic algorithms, fuzzy logic with C++, Java, SymbolicC++ and reduce programs. World Scientific Publishing Company.

Stoyanov, S.V., Rachev, S.T., Fabozzi, F.J., 2013. Sensitivity of portfolio VaR and CVaR to portfolio return characteristics. Ann. Oper. Res. 205, 169–187.

Sun, W., Rachev, S., Chen, Y., Fabozzi, F.J., 2008. Measuring intra-daily market risk: A neural network approach. Technical report, Karlsruhe Institute of Technology (KIT).

Taleb, N.N., 2007. The black swan: The impact of the highly improbable. Random house.

Tu, J., Zhou, G., 2011. Markowitz meets Talmud: A combination of sophisticated and naive diversification strategies. J. Financ. Econ. 99, 204–215.

Vapnik, V.N., 1999. An overview of statistical learning theory. IEEE Trans. Neural Netw. 10, 988–999.

Vasicek, O., 1977. An equilibrium characterization of the term structure. J. Financ. Econ. 5, 177–188.

Velasquez, T., 2010. Chaos Theory and the Science of Fractals in Finance.

Venkataraman, S., 1997. Value at risk for a mixture of normal distributions: the use of quasi-Bayesian estimation techniques. Econ. Perspect. 21, 2–14.

Wang, J., 2001. Generating daily changes in market variables using a multivariate mixture of normal distributions, in: Proceedings of the 33nd Conference on Winter Simulation. IEEE Computer Society, pp. 283–289.

Wikipedia contributors, 2019. Benoit Mandelbrot — Wikipedia, The Free Encyclopedia.

Witten, E., 1995. Of Beauty and Consolation Episode 9 Edward Witten - YouTube [WWW Document]. URL https://www.youtube.com/watch?v=RfwsvSjXkJU&t=267s (accessed 3.25.19).

Zhou, C., 2001. An analysis of default correlations and multiple defaults. Rev. Financ. Stud. 14, 555–576.