HIGH-FREQUENCY KELLY CRITERION AND FAT-TAILS: GAMBLING WITH AN EDGE

by

Austin Pollok

A Dissertation Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(APPLIED MATHEMATICS)

May 2022

Abstract

This thesis develops a unified theoretical and empirical framework for optimal capital allocation under repeated betting or trading when returns exhibit heavy tails and are realized at high frequency. Building on the Kelly criterion, which maximizes long-run logarithmic wealth growth, we extend classical discrete-time results to continuous-time settings and general stochastic environments driven by Lévy processes. This allows for realistic modeling of skewness, jumps, and fat-tailed return distributions commonly observed in financial markets.

The first part of the thesis provides a rigorous characterization of the Kelly-optimal strategy under independent and identically distributed increments in both discrete and continuous time. New results establish conditions under which optimal growth rates exist and are finite in the presence of heavy-tailed risks, clarifying the trade-offs between growth, exposure, and ruin in non-Gaussian environments. These results yield a high-frequency Kelly criterion that depends explicitly on the Lévy triplet of returns, offering a principled framework for position sizing when jump risk and tail risk are material.

The second part of the thesis applies these theoretical insights to an empirical study of equity options markets. We develop and evaluate high-dimensional and factor-based models for forecasting firm-level realized variance at the daily frequency and demonstrate economically meaningful improvements over benchmark volatility models when applied to option return prediction. Focusing on a variance risk premium–based straddle strategy, we show that the resulting return distributions are strongly non-normal, exhibiting significant skewness and excess kurtosis. The heavy-tailed nature of these returns motivates the use of

the high-frequency Kelly framework developed in the first part to optimize long-run growth.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 History of the Kelly Criterion

Consider the unlikely good fortune of coming across a person who offers you some initial amount of capital and thirty minutes to gamble on a biased coin toss with probability of heads being 60% and probability of tails being 40%. The gamble pays one-to-one odds and you can walk away with whatever amount you end up with at the end of the thirty minutes, you can wager any amount of capital in one cent increments up to your total amount of available capital. Should you play? If you decide you should play, how should you play to have the best chance of walking away a wealthy gambler? This scenario was in fact offered to a small group of quantitative finance students and professionals from Victor Haghani, a former principal of the notorious quantitative hedge fund Long Term Capital Management, with the hopes on discovering if qualified people in quantitative finance were able to answer this seemingly simple decision making under uncertainty [47]. It turns out this simple proposition has a simple solution known as the *Kelly Criterion*.

The Kelly Criterion was first formalized by John Kelly, a physicist from Bell Labs, in 1956 [58] and later popularized by the successful gambler, mathematician, and quantitative hedge fund manager, Ed Thorp, when he used it to determine bet sizes in his newly discovered

card counting system [81]. While working with Claude Shannon at Bell labs, Kelly had a novel interpretation of Shannon's then newly developed theory of communication. Kelly envisioned a horse race speculator who was fed the true odds of the race's outcome through a hypothetical *private wire*, that is, some noisy communication channel, so the speculator had a better estimate of the probabilities than those implied by the public odds. He then showed that a speculator who wishes to take advantage of this private information and maximize their wealth over the long-run could do so by wagering an optimal bet size while compounding their winnings. It turned out the way to calculate this optimal bet size was to maximize the Shannon entropy of the noisy communication channel, which was interpreted as the long-term growth rate of wealth. This provided a framework to prescribe optimal bet sizes to the speculator who wanted to do the best over the long-run and had some side information that allowed him to better estimate probabilities. There was no thought of utility maximization nor trade-offs between expected return and volatility. It was developed solely from an information theoretic interpretation of a gambler wish to exploit an edge.

A few years later, on the opposite side of the country, a newly minted mathematics Ph.D. encountered an article claiming there was a way to get a positive expected return, *edge*, in the game of blackjack [10]. This young mathematician was Ed Thorp. Inspired by the article claiming to yield an edge, Thorp further developed a way to systematically beat blackjack over the long-run by keeping track of what left the deck, what we now know today as card counting [82]. While a post-doc at MIT, Thorp befriended Claude Shannon, who suggested he apply the methods of Kelly to help determine bet sizes while having a better estimate of the edge [81]. Thorp continued his quest to analyze games of chance such as roulette, which he developed a winning system with the help of Shannon, baccarat, where he first coined the term *Kelly Criterion* [88], and many others [85]. Guided by his simple template of finding an edge, then exploiting that edge with proper Kelly bet sizing, Thorp went on to analyze option markets and start the first quantitative market-neutral hedge fund, Princeton-Newport Partners. Aided with a heuristically derived formula for pricing options,later known as the

Black-Scholes option pricing formula, and an implementable version of a dynamic hedging trading strategy [79], Thorp extended the Kelly Criterion for use in securities markets to exploit his edge [83].

Though the system of determining bet sizes with the Kelly Criterion failed to gain much academic attention, likely due to the negative press from the Nobel Prize winning economist Paul Samuelson, it has been reported to have been used by some legendary investors and practitioners such as Warren Buffet, Bill Gross, and James Simons [87, Ch 45, 52, 53, 54]. [1]

1.2 Risk versus Reward

In classical Markowitz portfolio theory, there is a formalization of the vague notions of *risk* and *reward*. Risk is formalized as the standard deviation of a portfolio's return, while reward is formalized as the mean of a portfolio's return. Further, through the construction of Markowitz's set of *efficient portfolios*, one can see a trade-off between risk and reward, namely, more risk implies more reward. This is the standard story that permeates our intuition and is the backbone behind most financial theories. Though this is widely accepted as an explanatory tool, most *hedge* funds, that is funds that aim to be market-neutral, would ideally strive for something that has reward, relative to risk-free assets, and essentially zero risk, if possible. This means, practically, most market-neutral hedge fund managers, speculators, or gamblers with edges, can not accept the story of more risk implies more reward as an operational procedure. The Kelly Criterion combats this narrative. In fact, if allowed to reinterpret risk as exposure, and reward as long-term growth rate, the Kelly Criterion concludes that there is an optimal amount of risk to maximize reward. More importantly, if you take more risk then you'll experience less reward, and most importantly, if you take too much risk then you will eventually be ruined. This is quite contrary to what is traditionally believed from the Markowitz theory of risk versus reward trade-off. It is

[1] For a much more adventurous and colorful history of the Kelly Criterion, see Willam Poundstone's popular science book, Fortune's Formula [72].

important to note that this appears to be the way many successful investors, speculators, or gamblers think about risk, whether it be formally or informally, including Ed Thorp, Ray Dalio, Nassim Taleb, Paul Tudor Jones, and Warren Buffet.

In a 2018 speech, Myron Scholes advocated for a financial theory that pays more attention to compound return, which is growth rate, and noted how prior to the work of Markowitz, most people regarded risk as risk of *loss*, which is exposure, rather than risk of standard deviation. Scholes calls the traditional Markowitz formalization of risk versus reward a *static* theory, as it only applies to a single bet with no consideration of compounding, and further recommends a *dynamic* theory to the investor concerned with compound return and loss. This is something that even Markowtiz agrees with [63]. If investors are concerned with compounding return over long horizons, or if they place many bets over short horizons, and are taking into account the trade-off with exposure, then the Kelly Criterion system is precisely for them [83]. [2]

1.3 Outline of Thesis

The contributions of this thesis begin in part I, which is joint work with Sergey Lototsky [61], with a unified characterization of the Kelly criterion under processes with independent and identically distributed increments, both in discrete and continuous time. In chapter **??** we describe known, though not necessarily well-known, properties of discrete and continuous time Lévy processes which are used for establishing our new results in chapter 4. While there are new results in chapter 2, as well as some simple numerical experiments in section 2.4, the main mathematical contribution comes in chapter 4 where we show a characterization of the Kelly criterion in continuous time in different stochastic environments. This part tells us how to optimally bet in a strategy with an *edge* when the strategy exhibits heavy tailed properties. This concludes the theoretical component of this thesis.

[2] For a further informal comparison between Markotitz and Kelly-type betting see [22, Ch 5].

In part II, which is joint work with Christopher Jones [56], we shift focus by empirically investigating a particular strategy in the options market where we have an *edge*. That is, the variance risk premium, which captures the well-known discrepancy between realized and implied volatility, applied to daily straddle returns. In chapter 6, we build on the work of [46] by investigating option returns, at a daily frequency, which are primarily driven by volatility. We contribute to this strategy by considering non-standard predictive models for daily realized variance in chapter 5. In section 5.4 we consider multiple models for realized variance, including high-dimensional penalized regressions, low-dimensional statistical factor models, as well as ensemble models.

We find, in section 5.6, the benchmark is hard to beat unanimously across different measures of forecast error, yet we are able to improve forecast error using different classes of non-standard models. More importantly, the significance of the improved volatility forecasts is highlighted when applying them to predict option returns, as shown in section 6.4.

Additionally, we find the distribution of the strategy's returns is heavy tailed, thus the tools of part I can help us optimize our strategy for long-run or high-frequency growth.

Part I

Mathematical Formalization of Kelly Criterion

Chapter 2

Discrete Time Kelly Criterion

Consider repeatedly placing bets on a sequence of random gambles where we have identified an edge. We can think of these gambles as the odds or payoffs of a posted wager, or as the rate of return of some financial asset. If we decide our goal is to compound our wealth over time, we can ask how we should bet in order to maximize our long term wealth, while simultaneously avoiding ruin. Let's assume we wish to bet a fixed fraction f of our wealth throughout our repeated gambles. We further assume we cannot short and we cannot use leverage or gamble on margin, which corresponds to assuming $f \in [0, 1]$.

We model our sequence of random gambles as a sequence of random variables that are independent and identically distributed,

$$\forall k \geq 1, \qquad r_k := \frac{P_k - P_{k-1}}{P_{k-1}} = \frac{P_k}{P_{k-1}} - 1 \stackrel{d}{=} r, \tag{2.1}$$

for some price process $P = \{P_k\}_{k \geq 1}$. We interpret the sequence of random gambles as a sequence of random rates of return, or net-returns, and we see $r \geq -1$ as you cannot lose more than 100% of your investment when you avoid short-selling and leveraging.

In a financial context it is clear what a price process is. In a gambling context we can think of the odds as the price, namely for $b : 1$ odds, the initial *price* to play this game or take this bet is $\$1$, and the final *price* will be either $\$(b + 1)$ or $\$0$, yielding a return $r \in \{b, -1\}$.

Most importantly, we assume we are gambling with an **edge**,

$$\mathbb{E}[r] > 0, \tag{2.2}$$

which is an intuitively necessary condition for any successful speculator.

2.1 Discrete Time Wealth Dynamics

A gambler who gambles with their winnings, that is, a gambler who is compounding their returns, will have an evolving wealth process

$$\begin{aligned}
W_0 &= \text{initial wealth/bankroll/portfolio size} \\
W_1^f &= W_0 + r_1(fW_0) = W_0(1 + fr_1) \\
W_2^f &= W_1^f + r_2(fW_1^f) = W_1^f(1 + fr_2) = W_0(1 + fr_1)(1 + fr_2) \\
&\vdots \\
W_n^f &= W_{n-1}^f + r_n(fW_{n-1}^f) = W_{n-1}^f(1 + fr_n).
\end{aligned}$$

More succinctly, the wealth process achieved from betting the fixed fraction f of current wealth in the sequence of random gambles is

$$W_n^f = W_0 \prod_{k=1}^{n} (1 + fr_k). \tag{2.3}$$

In other words, the wealth process is modeling the process of compounding returns with a constant exposure f to those returns.

See code **??** for a generic simulation of a wealth process given you can simulate the returns.

2.2 Kelly Criterion

For the investor, trader, or gambler whose goal is to maximize their long-term wealth, consider

$$\frac{W_n^f}{W_0} = \exp\left(n \ln\left(\frac{W_n^f}{W_0}\right)^{1/n}\right) = \exp\left(n \frac{1}{n} \sum_{k=1}^{n} \ln(1 + f r_k)\right) \xrightarrow[n\to\infty]{a.s.} \exp\left(\infty \mathbb{E}[\ln(1 + fr)]\right), \tag{2.4}$$

by the Law of Large Numbers for an iid sequence of random returns. Given the goal of maximizing long-term wealth, we see we can equivalently maximize

$$g_r(f) := \mathbb{E}[\ln(1 + fr)], \tag{2.5}$$

the long-term growth rate. This leads to the so-called Kelly Criterion:

$$\sup_{f\in[0,1]} g_r(f) = \sup_{f\in[0,1]} \mathbf{E}[\ln(1 + fr)] \tag{2.6}$$

Thus, the investor who compounds his wealth repeatedly over many *similar* bets with the sole purpose of having the largest wealth at the end of the long string of bets should be trying to maximize his long-term growth rate. Note (2.5) is really the long-term *exponential* rate of growth as it is the almost sure limit of

$$G_n(f) := \frac{1}{n} \ln\left(\frac{W_n^f}{W_0}\right), \tag{2.7}$$

the observed path-wise exponential rate of increase over $n-$many bets, that is, the log geometric average of the first n returns with exposure f.

In code **??** we simulate the wealth processes from betting on a coin toss and consider path-wise growth rate and the long-term theoretical growth rate.

2.3 Properties of Kelly Criterion in Discrete Time

In order for (2.3) to be a non-trivial object of study, we need the random variable r to be have the following properties:

$$\mathbb{P}(r \geq -1) = 1; \tag{2.8}$$

$$\mathbb{P}(r > 0),\ \mathbb{P}(r < 0) > 0; \tag{2.9}$$

$$\mathbb{E}[|\ln(1+r)|] < \infty. \tag{2.10}$$

Condition (2.8) quantifies the notion that a loss in a random gamble cannot be more than 100%. Condition (2.9) ensures that both gains and losses are possible on the random gamble. Condition (2.10) is the minimal requirement to define the long-run growth rate of the wealth process, which is the main object of study in this section.

The following properties of the function g_r are immediate consequences of the definition and the assumptions (2.8)–(2.10):

Proposition 2.3.1. *The function $f \mapsto g_r(f)$ is continuous on the closed interval $[0,1]$ and infinitely differentiable in $(0,1)$. In particular,*

$$\frac{dg_r}{df}(f) = \mathbb{E}\left[\frac{r}{1+fr}\right], \quad \frac{d^2 g_r}{df^2}(f) = -\mathbb{E}\left[\frac{r^2}{(1+fr)^2}\right] < 0. \tag{2.11}$$

Corollary 2.3.1. *The function g_r achieves its maximal value on $[0,1]$ at a point $f^* \in [0,1]$ and $g_r(f^*) \geq 0$. If $g_r(f^*) > 0$, then f^* is unique.*

Proof. Note that $g_r(0) = 0$ and, by (2.11), the function g_r is strictly concave on $[0,1]$. □

While concavity of g_r implies that g_r achieves a unique global maximal value at a point f^{**}, it is possible that the domain of the function g_r is bigger than the interval $[0,1]$ and $f^{**} \notin [0,1]$. A simple way to exclude the possibility $f^{**} < 0$ is to consider returns r that are not bounded from above: $\mathbb{P}(r > c) > 0$ for all $c > 0$: in this case, the function

$g_r(f) = \mathbb{E}\ln(1 + fr)$ is not defined for $f < 0$. Similarly, if $\mathbb{P}(r < -1 + \delta) > 0$ for all $\delta > 0$, then the function g_r is not defined for $f > 1$, excluding the possibility $f^{**} > 1$.

Below are more general sufficient conditions to ensure that the point $f^* \in [0, 1]$ from Corollary 2.3.1 is the point of global maximum of g_r: $f^* = f^{**}$.

Proposition 2.3.2. *If*

$$\lim_{f \to 0+} \mathbb{E}\left[\frac{r}{1 + fr}\right] > 0 \qquad \text{and} \tag{2.12}$$

$$\lim_{f \to 1-} \mathbb{E}\left[\frac{r}{1 + fr}\right] < 0, \tag{2.13}$$

then there is a unique $f^* \in (0, 1)$ *such that*

$$g_r(f) < g_r(f^*)$$

for all f *in the domain of* g_r.

Proof. Together with the intermediate value theorem, conditions (2.12) and (2.13) imply that there is a unique $f^* \in (0, 1)$ such that

$$\frac{dg_r}{df}(f^*) = 0.$$

It remains to use strong concavity of g_r. □

Because $r \geq -1$, the expected value $\mathbb{E}[r]$ is always defined, although $\mathbb{E}[r] = +\infty$ is a possibility. Thus, by (2.11), condition (2.12) is equivalent to the intuitive idea of an edge (2.2):

$$\mathbb{E}[r] > 0,$$

which guarantees that $g_r(f) > 0$ for some $f \in (0, 1)$. In other words, if you have an edge, then you will have a positive expected long-run growth rate and you will profit, though the

long-run could be too long. This explains why it is important to be a gambler with an edge. Condition (2.13) can be written as

$$\mathbb{E}\left[\frac{r}{1+r}\right] < 0,$$

with the convention that the left-hand side can be $-\infty$. This condition is necessary in general, to ensure that the edge is not so large to imply leveraged gambling ($f^* > 1$) will lead to an optimal strategy.

Assuming we have an edge, the long-run growth rate (2.5) is well-defined whenever $\mathbb{P}(fr > -1) = 1$. This leads to a larger set of admissible exposures,

$$f \in \left[0,\ \frac{1}{|\text{essinf } r|}\right],$$

where essinf r is the *worst-case scenario* of the return distribution. If essinf $r > -1$, then there is no chance of losing 100% which in some cases could lead to a theoretically optimal exposure which is levered. In these cases where leveraged gambling is theoretically optimal, it may not be practically optimal. One operating in the real world must account for many unknowable uncertainties such as parameter uncertainty, limitations of the theoretical long-run, the true value of essinf r, and perhaps most importantly the increased chance of ruin when gambling with leverage in an uncertain environment. With these practical obstacles, we find it best to eliminate the use of leverage and keep our set of admissible exposures restricted to the interval $[0, 1]$.

2.4 Thin-Tailed Examples

2.4.1 Bernoulli Bet

As an illustrative example, consider making repeated bets in game of chance with only two possible outcomes, that is, a binary bet. The payoff of this gamble can be expressed

with the odds of $b:1$. These odds can be interpreted as the rate of return of a binary bet with possible returns of either $100 \times b\%$ or -100%. For simplicity, we only consider the case of $b:1$ odds as any other odds, say $b:a$, can be reduced to this case by dividing the winning payoff by the losing payoff, $(b/a):1$, and then adapting the following formula as needed, see the end of this section.

Assume we know the distribution of this Bernoulli bet,

$$\mathbb{P}(r = b) = p, \quad \mathbb{P}(r = -1) = 1 - p, \quad 0 < p < 1. \tag{2.14}$$

Further assume this is a favorable game as we have identified an edge,

$$\mathbb{E}[r] = bp - (1 - p) > 0 \iff p > \frac{1}{b+1}.$$

The long-run growth rate function

$$g_r(f) = p \ln(1 + fb) + (1 - p) \ln(1 - f)$$

is defined on $(-1/b, 1)$, and achieves the global maximum at

$$f^* = \frac{bp - (1 - p)}{b} = \frac{\text{edge}}{\text{odds}}, \tag{2.15}$$

and

$$g_r(f^*) = p \ln(p) + (1 - p) \ln(p) + ln(1 + b) + (1 - p) \ln(b),$$

which we know is positive by (2.3.2).

In code **??**, we use Monte-Carlo simulation and a numerical solver to search for f^*.

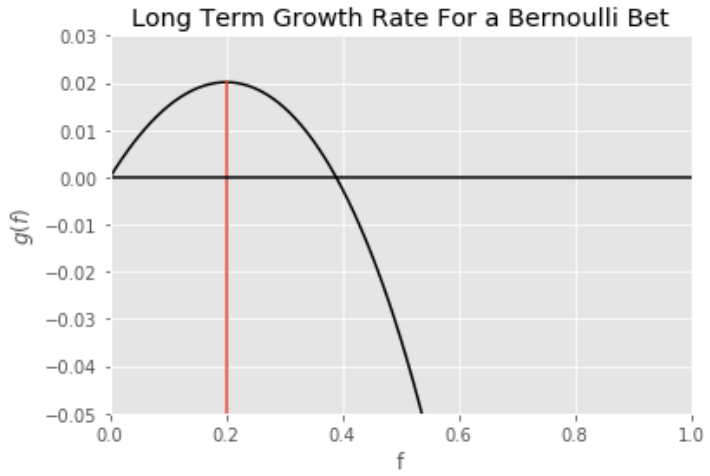

Figure 2.1: Long Term Growth Rate for Binary Bet with b = 1, and p = 0.6. This shows that more *risk* doesn't necessarily mean more *reward*. In fact, too much *risk* decreases *reward* and can lead to ruin.

Above, we only consider the case of $b:1$ odds for illustrative purposes. Any other odds, say $b:a$, can be viewed similarly by dividing the winning payoff by the losing payoff, $(b/a):1$. This changes the formulas as follows:

$$\mathbb{P}(r=-a)=1-p,\quad \mathbb{P}(r=b)=p,\quad 0<a\leq 1,\ b>0,\ 0<p<1. \tag{2.16}$$

The function

$$g_r(f)=p\ln(1+fb)+(1-p)\ln(1-fa)$$

is defined on $(-1/b, 1/a)$, achieves the global maximum at

$$f^*=\frac{p}{a}-\frac{1-p}{b}=\frac{\mathbb{E}[r]}{ab}=\frac{\frac{b}{a}p-(1-p)}{b},$$

and

$$g_r(f^*)=p\ln p+(1-p)\ln(1-p)+\ln\frac{a+b}{a}-(p-1)\ln\frac{b}{a};$$

we know that $g_r(f^*)\geq 0$, even though it is not at all obvious from the above expression.

The no shorting and no leverage constraint $f^* \in [0, 1]$ becomes

$$\frac{a}{a+b} \leq p \leq \min\left(\frac{ab}{a+b}\left(1+\frac{1}{b}\right), 1\right).$$

2.4.2 Uniform Returns

A natural way to extend the Bernoulli bet to a continuous distribution is to assume returns are distributed uniformly over $[-1, b]$,

$$r \sim \text{Uniform}[-1, b].$$

To be a disciplined gambler or investor with an edge, assume

$$\mathbf{E}[r] = \frac{b-1}{2} > 0 \iff b > 1$$

To compute the optimal exposure, consider the long term growth rate function,

$$g_r(f) = \mathbf{E}[\ln(1+fr)]$$

$$= \frac{1}{b+1}\int_{-1}^{b} \ln(1+fx)dx$$

$$= \frac{(1+fb)\ln(1+fb) - (1-f)\ln(1-f)}{f(b+1)} - 1$$

The first order condition is

$$\frac{d}{df}g_r(f) = \frac{fb - \ln(1+fb) + \ln(1-f)}{(b+1)f^2} = 0.$$

We can numerically find the root of this equation. In code **??**, we carry this out.

In the case of $b = 2$, we see $f^* \approx .716...$

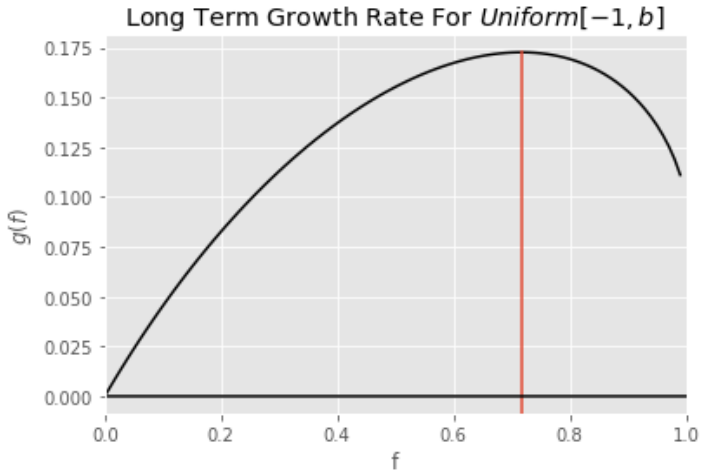

Figure 2.2: Long Term Growth Rate for Uniform Returns with b = 2

2.4.3 Log Normal Returns

A general way to to ensure conditions (2.8)–(2.10) is to consider

$$r := e^R - 1 \tag{2.17}$$

for some random variable R such that $\mathbb{P}(R > 0) > 0$, $\mathbb{P}(R < 0) > 0$, and $\mathbb{E}[|R|] < \infty$. Then (2.12) and (2.13) become, respectively,

$$\mathbb{E}[e^R] > 1 \qquad \text{and} \tag{2.18}$$

$$\mathbb{E}[e^{-R}] > 1. \tag{2.19}$$

For a model such as (2.17) to occur, one could envision a price process $P = \{P_k\}$ of a random gamble we're betting on. Then the net-returns or rate of return is $r_k = \frac{P_k - P_{k-1}}{P_{k-1}}$, as in (2.1). If we assume the log-(gross)returns are independent and identically distributed as R, then

$$\ln(1 + r_k) = \ln\left(\frac{P_k}{P_{k-1}}\right) \sim R, \tag{2.20}$$

which is equivalent to

$$r_k \overset{d}{=} r = e^R - 1,$$

which satisfies (2.17).

If R is normally distributed with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$, then

$$\mathbb{E}[e^R] = e^{\mu+(\sigma^2/2)}, \quad \mathbb{E}[e^{-R}] = e^{-\mu+(\sigma^2/2)},$$

and (2.18), (2.19) are equivalent to

$$-\frac{\sigma^2}{2} < \mu < \frac{\sigma^2}{2}.$$

So to have an edge, we must have $\mu > -\frac{\sigma^2}{2}$, and to guarantee no leverage, we must have $\mu < \frac{\sigma^2}{2}$. For this model, the corresponding f^* is not available in closed form, but can be evaluated numerically.

In code **??**, we carry this out.

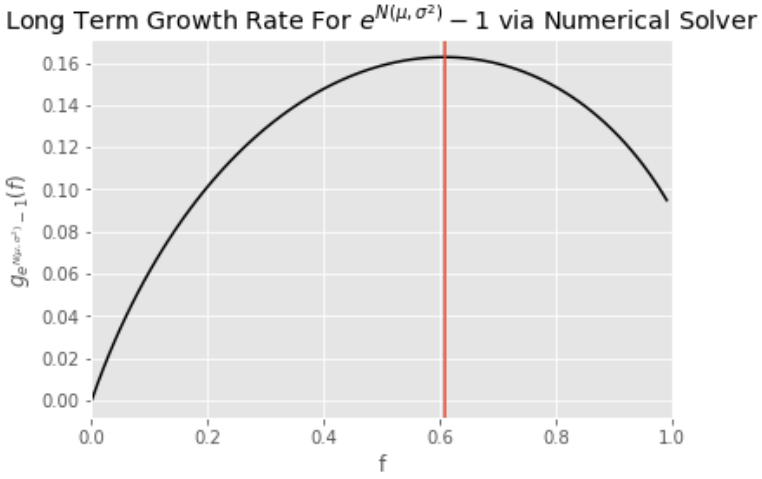

Figure 2.3: Long Term Growth Rate for with Log Normal Returns with $\mu = 0.09$ and $\sigma = 1$

Chapter 3

From Random Walks to Lévy Processes

In this background section, we will use notation consistent with the rest of the thesis, which means most of the objects of study will be modeling rates of returns, r, or log-gross returns, $R = \ln(1 + r)$ of various financial gambles or assets.

There are many approaches to embark upon when studying Lévy processes and these different paths are dependent upon what you assume. We start by trying to understand Lévy processes as *high-frequency* limits of random walks using the tools of weak convergence and stochastic calculus. The theorems as well as their proofs can be found in [52], [75], and [73]. For examples using this approach in the context of volatility modeling, see [1].

For the remainder of this chapter, we fix a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ satisfying the usual conditions [52, Definition I.1.3]. We write $\xi \overset{d}{=} \eta$ to indicate equality in distribution of two random variables, and $X \overset{\mathcal{L}}{=} Y$ to indicate equality in law (as function valued random elements).

The simplest object of study comes from summing independent and identically distributed random variables.

Definition 3.0.1 (Random Walk)**.** *A stochastic process* $S = \{S_n\}_{n \geq 0}$ *is called a* random

walk, *if* $S_n = \sum_{k=1}^{n} \xi_k$, *with* $\{\xi_k\}$ *being a sequence of independent and identically distributed random variables. Note the increments are independent and stationary. That is, it satisfies the following properties:*

$$S_0 = 0, \tag{3.1}$$

$$S_{k+h} - S_k \perp\!\!\!\perp (S_k, S_{k-1}, \cdots, S_1, S_0), \quad \forall h, k \geq 1, \tag{3.2}$$

$$S_{k+h} - S_k \overset{d}{=} S_h, \quad \forall h, k \geq 1. \tag{3.3}$$

For the random walk, there exists limit theorems such as Law of Large numbers and the Central Limit Theorem.

Theorem 3.0.1 (LLN)**.** *If* $\mathbb{E}[|\xi_1|] < \infty$, *then* $\frac{S_n}{n} \xrightarrow[n\to\infty]{a.s.} \mathbb{E}[\xi_1]$.

Theorem 3.0.2 (CLT)**.** *[52, Remark VII.5.3]*

$$\textit{If } Var(\xi_1) < \infty, \textit{ then } \frac{S_n - \mathbb{E}[S_n]}{\sqrt{n}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, Var(\xi_1)).$$

But what if we took a more and more remote look at our random walk, that is, zoomed out further and further? More specifically, imagine these discrete observations $\{\xi_k\}_{k\leq n}$ occurring during a continuous time interval, for simplicity say $[0, 1]$. Thus we have $n-$many observations per time period. *Zooming out* is equivalent to speeding up the frequency of our observations per time period. A theorem by Donsker describes what happens to our random walk as the frequency per time period gets larger and larger.

Set the centered and scaled pre-limiting process to be

$$Y_t^n := \frac{S_{\lfloor nt \rfloor} - \mathbb{E}[S_{\lfloor nt \rfloor}]}{\sqrt{n}}$$

and the limiting process to be

$$Y_t := \sigma B_t \overset{d}{=} \mathcal{N}(0, \sigma^2 t), \quad \sigma^2 := \mathrm{Var}(\xi_1)$$

where B is a Brownian motion process.

The following limit theorem is often referenced as Donsker's Invariance Principle, [52, Corollary VII.3.11], which involves sequences of stochastic processes converging as function-valued random elements. Convergence in law and the more general concept of weak convergence is further explained in [52, Section VI.3].

Theorem 3.0.3 (Functional CLT). *If $\sigma^2 < \infty$, then $Y^n \xrightarrow[n\to\infty]{\mathcal{L}} Y$.*

It can be shown that Y has continuous sample paths, as well as independent and stationary increments. This gives us a way of formally showing how a continuous time process can be approximated, in the appropriate sense, by a random walk.

What happens if the observations have fat enough tails that the variance is infinite,

$$\mathrm{Var}(\xi_1) = \infty?$$

If we assume the tails are not so fat as to have infinite expected value, then the Law of Large number still applies, however the Central Limit Theorem must be appropriately adjusted.

Theorem 3.0.4 (Infinite Variance CLT). *[20, Proposition 9.25] Assume there exists constants $\{a_n\}$ and $\{b_n\}$ such that $a_n \to \infty$ and $\frac{b_n}{n} \to 0$, as $n \to \infty$, then the normalized sum converges in distribution*

$$\frac{S_n}{a_n} - b_n \xrightarrow[n\to\infty]{d} Y,$$

where Y is a stable distribution, meaning

$$\sum_{i=1}^{n} Y_i \overset{d}{=} a_n Y + b_n,$$

where Y_i is an independent and identically distributed copy of Y, for all $1 \le i \le n$.

To make the jump from sums of random variables to Lévy processes, we have go slightly beyond the random walk.

Definition 3.0.2 (Infinitely Divisible Distribution)**.** *A random variable Y is said to have an* infinitely divisible distribution *if for every n there exists a sequence of independent and identically distributed random variables $\{\xi_{n,k}\}$ with the property*

$$S_n^n := \sum_{k=1}^{n} \xi_{n,k} \overset{d}{=} Y.$$

Theorem 3.0.5 (Triangular Array CLT)**.** *[52, Theorem VII.2.35] If we have a row-wise independent sum of random variables $S_n^n := \sum_{k=1}^{n} \xi_{n,k}$ which satisfies the infinitesimal property:*

$$\forall \varepsilon, \ \sup_{1 \le k \le n} \mathbb{P}(|\xi_k^n| > \varepsilon) \underset{n \to \infty}{\longrightarrow} 0,$$

and we assume

$$S_n^n \underset{n \to \infty}{\overset{d}{\longrightarrow}} Y,$$

then Y must be an infinitely divisible distribution.

There are hosts of theorems that allow one to derive necessary and sufficient conditions for the sequence $\{S^n\}$ to converge towards any given infinitely divisible distribution. To make this result hold for continuous time stochastic processes, we define the pre-limiting process

$$Y_t^n := S_{\lfloor nt \rfloor}^n = \sum_{k=1}^{\lfloor nt \rfloor} \xi_{n,k} \tag{3.4}$$

and let

$$Y = \{Y_t\} \text{ be a stochastic process satisfying } Y_1 \overset{d}{=} Y, \tag{3.5}$$

where Y is defined in Theorem 3.0.5.

Theorem 3.0.6 (Functional Triangular Array CLT)**.** *[52, Corollary VII.3.6]*

Assume $\{Y^n\}$ *and* Y *are defined as in* (3.4) *and* (3.5)*, and assume*

$$Y_1^n \xrightarrow[n\to\infty]{d} Y_1$$

as in Theorem 3.0.5 where the convergence is in distribution as random variables. Then the stochastic processes converge in law as function valued random elements

$$Y^n \xrightarrow[n\to\infty]{\mathcal{L}} Y.$$

The above theorems illustrate how we can start with random walks and rigorously converge to continuous time stochastic processes of different forms, namely that satisfy the property of infinitely divisible distributions at time one. Continuous time stochastic processes with the property that their time one random variable has an infinitely divisible distribution can be show to be Lévy processes.

Definition 3.0.3 (Lévy Process)**.** *A stochastic process* $R = \{R_t\}_{t\geq 0}$ *is called a* Lévy process *if it satisfies the following properties:*

$$R_0 = 0, \tag{3.6}$$

$$R_t - R_s \perp\!\!\!\perp R_{[0,t]} := R_{\cdot\wedge t}, \tag{3.7}$$

$$R_t - R_s \overset{d}{=} R_{t-s}, \tag{3.8}$$

$$\mathbb{P} - \lim_{s\to t} R_s = R_t. \tag{3.9}$$

Properties (3.6), (3.7), (3.8) are the continuous time analogues of properties (3.1), (3.2), (3.3) of a random walk. Property (3.9) is called *stochastic continuity* of the process R, which means, more precisely, $\forall \varepsilon > 0, \exists \delta_\varepsilon$ such that whenever $|t-s| < \delta$, $\mathbb{P}\left(|R_t - R_s| > \varepsilon\right) < \varepsilon$. Equivalently, for any $\varepsilon > 0$,

$$\lim_{s\to t} \mathbb{P}\left(|R_t - R_s| > \varepsilon\right) = 0.$$

One can show conditions (3.6)-(3.9) are equivalent to the random variable R_1 having an infinitely divisible distribution, [75, Theorem 7.10], which shows Lévy processes are in one-to-one correspondence with the class of infinitely divisible distributions and, in particular, (3.5) is a Lévy process which can be interpreted as the limit of (3.4) by Theorem 3.0.6.

It turns out we can get an additional limit theorem that isn't a central limit theorem for Lévy processes.

Theorem 3.0.7 (LLN for Lévy Processes)**.** *[75, Theorem 36.5]*

$$\textit{If } \mathbb{E}[|R_1|] < \infty, \textit{ then } \frac{R_t}{t} \xrightarrow[t\to\infty]{a.s.} \mathbb{E}[R_1].$$

Thus far we have understood Lévy processes as limits of sums of random variables and thought about them from a distributional point of view. Next, we change our point of view from the distributional laws of Lévy processes to a pathwise view.

A very pragmatic reason for studying Lévy processes comes from the fact that they are allowed to *jump* with a relatively flexible class of jump distributions. The associated jump process is defined as

$$\Delta R_t := R_t - R_{t-},$$

where $R_s \xrightarrow[s\to t]{} R_{t-}$, for $s < t$.

In order to adequately study the jumps of a Lévy process, we need a way to measure the jumps, that is, a jump measure formalized by the notion of a random measure of a stochastic process [52, Section II.1].

For a measurable subset of the real line A, we define

$$\mu^R(\omega; t, A) := \sum_{s\le t} \delta_{(s,\Delta R_s(\omega))}([0,t],\ A) \tag{3.10}$$

to be the jump measure that counts the number of jumps of the process R of size in A up to time t. We can equivalently write (3.10) in a set notation

$$\mu^R(\omega; t, A) = \#\left\{0 \le s \le t |\ \Delta R_s(\omega) \in A\right\},$$

or in infinitesimal notation

$$\mu^R(\omega; ds, dx) := \sum_{s>0} \delta_{(s,\Delta R_s(\omega))}(ds, dx).$$

This allows one to decompose a Lévy process into a sum of three processes: a white noise process with drift, a well-behaved small jump process, and a well-understood large jump process.

Theorem 3.0.8 (Lévy-Itô Pathwise Decomposition)**.** *[75, Theorem 19.2] Given a Lévy process R and its associated jump measure* (3.10)*, there exists a triple (μ, σ, F^R) such that R can be decomposed as follows*

$$R_t = \underbrace{\mu t + \sigma W_t}_{\text{Brownian motion with drift}} + \underbrace{\int_0^t \int_{\mathbb{R}} x 1_{\{|x|\leq 1\}} (\mu^R - mF^R)(dx, ds)}_{\mathbb{L}^2\text{-martingale with countably many jumps}} + \underbrace{\int_0^t \int_{\mathbb{R}} x 1_{\{|x|>1\}} \mu^R(dx, ds)}_{\text{compound Poisson process}},$$

where m is the ordinary Lebesgue measure and mF^R is the compensator of the jump measure (3.10)*.*

For more formal and general treatments of the above results, as well as much more results regarding Lévy processes and semimartingales, see [52] and [75].

Chapter 4

High Frequency Kelly Criterion

The following chapter is adapted from work with Sergey Lototsky published in the SIAM Journal of Financial Mathematics.

We extend the discrete time Kelly Criterion to a continuous time version. That is we investigate how to optimally bet when you are continuously compounding returns. This begs the question: why go to continuous time? In mathematics, one goes to continuous time to use the powerful tools of stochastic calculus to approximate discrete processes by continuous processes, much like in ordinary calculus. In finance, one goes to continuous time also to use the tools of stochastic calculus, however, one needs practical justification for the approximation of discrete financial processes with continuous time processes. Such an approximation is well motivated if one is doing anything frequently over short horizons, or less frequently but over long horizons.

Consider an investor, gambler, or trader who identifies more short-term edges and then frequently trades or re-balances their portfolio so as to maximize their compound wealth in the long-run.

4.1 A Random Walk Model

Following the methodology in [84, Section 7.1] we assume we're compounding a sufficiently large number n of bets in a time period $[0, T]$, where T could represent days, months, or years depending on the application. The returns $r_{n,1}, r_{n,2}, \ldots$ of the bets are modeled as

$$r_{n,k} = \frac{\mu}{n} + \frac{\sigma}{\sqrt{n}}\xi_{n,k} \tag{4.1}$$

for some $\mu > 0$, $\sigma > 0$ and independent and identically distributed random variables $\{\xi_{n,k}\}$ with mean 0 and variance 1. The classical simple random walk corresponds to $\mathbb{P}(\xi_{n,k} = +1) = \mathbb{P}(\xi_{n,k} = -1) = 1/2$ and can be considered a *high frequency* version of the Bernoulli bet (2.14) with $b = 1$ and $p = 1/2$. To satisfy condition (2.8), we need $r_{n,k} \geq -1$, which, in general, can only be achieved with *uniform boundedness* of $\xi_{n,k}$:

$$|\xi_{n,k}| \leq C_0, \tag{4.2}$$

and then, we can assume without loss of generality, that n is large enough so that

$$|r_{n,k}| \leq \frac{1}{2}. \tag{4.3}$$

Assume we have an edge on every bet as in (2.2),

$$\mathbb{E}[r_{n,k}] = \frac{\mu}{n} > 0. \tag{4.4}$$

The wealth process, similar to (2.3), corresponding to n-many bets per unit time period with exposure $f \in [0, 1]$ at time $t \in (0, T]$ is given by:

$$W_t^{n,f} = W_0 \prod_{k=1}^{\lfloor nt \rfloor} \left(1 + f r_{n,k}\right); \tag{4.5}$$

where $\lfloor nt \rfloor$ denotes the largest integer less than nt. From now on, we assume without loss of generality, $W_0 = 1$.

Let $B = \{B_t\}_{t\geq 0}$ be a standard Brownian motion on a stochastic basis $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, that is, a filtered probability space, satisfying the usual conditions, [52, Definition 1.3], and define the process

$$W_t^f := \exp\left(\left(f\mu - \frac{f^2\sigma^2}{2}\right)t + f\sigma B_t\right). \tag{4.6}$$

We interpret this process as our continuous time or *high frequency* wealth process corresponding to continuous re-balancing with exposure f.

Theorem 4.1.1. *For every $T > 0$ and for every $f \in [0,1]$, the sequence of processes $\left\{\left(W_t^{n,f}\right)_{t\in[0,T]}\right\}_{n\geq 1}$ converges in law to the process $W^f = \left(W_t^f\right)_{t\in[0,T]}$.*

Proof. The objective is to show weak convergence of

$$Y_t^{n,f} := \ln(W_t^{n,f}),$$

as $n \to \infty$ to the process

$$Y_t^f := \left(f\mu - \frac{f^2\sigma^2}{2}\right)t + f\sigma B_t,\ t \in [0,T].$$

The proof relies on the method of predictable characteristics for semimartingales from [52]. In particular, we make suitable changes in the proof of Corollary VII.3.11.

By (4.5)

$$Y_t^{n,f} = \sum_{k=1}^{\lfloor nt \rfloor} \ln(1 + f r_{n,k}).$$

Then (4.1) and (4.2) imply

$$\mathbb{E}\left(Y_t^{n,f} - \mathbb{E}Y_t^{n,f}\right)^4 \leq \frac{C_0^4\sigma^4}{n^2}(nT + 3nT(nT-1)) \leq 3C_0^4\sigma^4T^2,$$

which further implies the sequence of processes $\left\{\left(Y_t^{n,f}\right)_{t\in[0,T]}\right\}_{n\geq 1}$ is uniformly integrable.

Next, by [52, Theorem VII.3.7], to establish weak convergence in law it suffices to verify the following:

$$\lim_{n\to\infty}\sup_{t\le T}\left|\lfloor nt\rfloor\mathbb{E}[\ln(1+fr_{n,1})]-\left(f\mu-\frac{f^2\sigma^2}{2}\right)t\right|=0, \tag{4.7}$$

$$\lim_{n\to\infty}\lfloor nt\rfloor\left(\mathbb{E}\left(\ln(1+fr_{n,1})\right)^2-\left(\mathbb{E}[\ln(1+fr_{n,1})]\right)^2\right)=f^2\sigma^2 t,\ \ \forall t\in[0,T], \tag{4.8}$$

$$\lim_{n\to\infty}\lfloor nt\rfloor\mathbb{E}[\phi(\ln(1+fr_{n,1}))]=0,\ \ \forall t\in[0,T]. \tag{4.9}$$

Where (4.9) must hold for all functions $\phi=\phi(x),\ x\in\mathbb{R}$ that are continuous and bounded on $\mathbb{R}$ and satisfy $\phi(x)=o(x^2),\ x\to 0$, that is,

$$\lim_{x\to 0}\frac{\phi(x)}{x^2}=0. \tag{4.10}$$

Equalities (4.7) and (4.8) follow from

$$r_{n,1}^2=\frac{\sigma^2}{n}\xi_{n,1}^2+\frac{2\mu\sigma}{n^{3/2}\xi_{n,1}}+\frac{\mu^2}{n^2},$$

and (4.3) along with an inequality

$$\left|\ln(1+x)-x-\frac{x^2}{2}\right|\le|x|^3,\ \ \forall|x|\le\frac{1}{2}.$$

More specifically,

$$\mathbb{E}\left[\left(\ln(1+fr_{n,1})\right)^2\right]=\frac{f^2\sigma^2}{n}+o(1/n),\ \ n\to\infty.$$

Finally, to establish (4.9), use the fact that (4.1) and (4.10) imply

$$\phi(\ln(1+fr_{n,1}))=o(1/n),\ \ n\to\infty.$$

□

With the same spirit as (2.5), we define the long-term continuous time growth rate

$$g(f) := \lim_{t\to\infty} \frac{1}{t} \ln(W_t^f).$$

We can see

$$g(f) = f\mu - \frac{f^2\sigma^2}{2}, \tag{4.11}$$

which implies

$$f^* = \frac{\mu}{\sigma^2} \tag{4.12}$$

for which the maximal long-term continuous time growth rate becomes

$$g(f^*) = \frac{\mu^2}{2\sigma^2}. \tag{4.13}$$

Our constraint of no shorting and no leverage, $f^* \in [0, 1]$, holds if $0 \le \mu \le \sigma^2$, which is consistent, to the order $1/n$, with (2.12) and (2.13) when applied to (4.1):

$$\mathbb{E}[r_{n,k}] = \frac{\mu}{n}, \quad \mathbb{E}\left[\frac{r_{n,k}}{1 + r_{n,k}}\right] = \frac{\mu - \sigma^2}{n} + o(1/n).$$

We remark that the wealth process (4.6) is for someone who is *continuously* placing bets, that is, adjusts their exposures or positions instantaneously, and, for large n, is a good approximation for someone that is discretely betting with a high frequency as in (4.5). In general, when the returns are given by (4.1), a direct optimization of (4.5) with respect to f will not lead to a closed-form formula for the corresponding optimal strategy f_n^*, yet Theorem 4.1.1 suggests that for sufficiently large n, (4.12) can be an approximation of f_n^* which suggests (4.13) can be an approximation to the corresponding long-term growth rate. As an illustration of this approximation, of a discrete process with a continuous process,

consider an adaptation of a Bernoulli bet (2.16) to a high frequency model:

$$\mathbb{P}\left(r_{n,k} = \frac{\mu}{n} \pm \frac{\sigma}{\sqrt{n}}\right) = \frac{1}{2}, \tag{4.14}$$

which, for fixed n, is is a particular case of the general Bernoulli model (2.16) with $p = q = 1/2$,

$$a = \frac{\sigma}{\sqrt{n}} - \frac{\mu}{n}, \ b = \frac{\sigma}{\sqrt{n}} + \frac{\mu}{n}.$$

Then, by direct computation,

$$f_n^* = \frac{\mu}{\sigma^2 - (\mu^2/n)} \to \frac{\mu}{\sigma^2}, \ n \to \infty,$$

and

$$\lim_{n\to\infty} g_{r_n}(f_n^*) = \frac{\mu^2}{2\sigma^2}.$$

4.2 A Random Walk Model in a Random Environment

The objective of this section is to analyze high frequency limits for betting *in business time*. In other words, the number of bets is not known a priori, so that a natural model of the corresponding wealth process is

$$W_t^{n,f} = \prod_{k=1}^{\lfloor \Lambda_{n,t} \rfloor} (1 + f r_{n,k}) \tag{4.15}$$

where, for each n, the process $t \mapsto \Lambda_{n,t}$ is a subordinator, that is, a non-decreasing Lévy process, independent of all $r_{n,k}$.

To study (4.15), we will follow the methodology in [59], where convergence of processes is derived after *assuming* a suitable convergence of the random variables. The main result in this connection is as follows.

Theorem 4.2.1. *Consider the following objects:*

- *random variables* $X_{n,k},\ n,k \geq 1$ *such that* $\{X_{n,k},\ k \geq 1\}$ *are iid for each* n*, with mean zero and, for some* $\beta \in [0,1]$*,* $m_n := \left(\mathbb{E}|X_{n,1}|^\beta\right)^{1/\beta} < \infty$*;*
- *random processes* $\Lambda_n = \Lambda_{n,t},\ n \geq 1,\ t \geq 0$*, such that, for each* n*,* Λ_n *is a subordinator independent of* $\{X_{n,k},\ k \geq 1\}$ *with the properties* $\Lambda_{n,0} = 0$*, and for some numbers* $0 < \delta, \delta_1 \leq 1$ *and* $C_n > 0$*,* $\left(\mathbb{E}\Lambda_{n,t}^\delta\right)^{1/\delta} \leq C_n t^{\delta_1/\delta}$*.*

Assume that there exist infinitely divisible random variables Y *and* U *such that*

$$\lim_{n\to\infty} \sum_{k=1}^{n} X_{n,k} \stackrel{d}{=} \bar{Y},\ \lim_{n\to\infty} \frac{\Lambda_{n,1}}{n} \stackrel{d}{=} \bar{U}.$$

If

$$\sup_n \left(C_n m_n^\beta\right) < \infty, \tag{4.16}$$

then, as $n \to \infty$*, the sequence of processes*

$$t \mapsto \sum_{k=1}^{\lfloor \Lambda_{n,t} \rfloor} X_{n,k},\ t \in [0,T],$$

converges, in the Skorokhod topology, to the process $Z = Z_t$ *such that* $Z_t = Y_{U_t}$*, where* Y *and* U *are independent Lévy processes satisfying* $Y_1 \stackrel{d}{=} \bar{Y}$ *and* $U_1 \stackrel{d}{=} \bar{U}$*.*

The proof is a word-for-word repetition of the arguments leading to [59, Theorem 1]: the result of [45], together with the assumptions of the theorem, implies

$$\lim_{n\to\infty} \sum_{k=1}^{\lfloor \Lambda_{n,1} \rfloor} X_{n,k} \stackrel{d}{=} Z_1,$$

and therefore the convergence of finite-dimensional distributions for the corresponding processes; together with condition (4.16), this implies the convergence in the Skorokhod space. Because we deal exclusively with Lévy processes, it is possible to avoid the heavy machinery from [52].

We now consider the wealth process (4.15) and apply Theorem 4.2.1 with

$$X_{n,k} = \ln(1 + fr_{n,k}) - \mathbb{E}\ln(1 + fr_{n,k}).$$

Consider an example, assume that the returns $r_{n,k}$ are as in (4.1), and let $\Lambda_{n,t} = S_{n^\alpha t}$, where $\alpha \in (0, 1]$ and $S = \{S_t\}_{t\geq 0}$ is the Lévy process such that S_1 has an α-stable distribution with both scale and skewness parameters equal to 1. Recall that an α-stable Lévy process $L^\alpha = \{L^\alpha_t\}_{t\geq 0}$ satisfies the following equality in distribution (as processes):

$$L^\alpha_{\gamma t} \overset{\mathcal{L}}{=} \gamma^{1/\alpha} L^\alpha_t, \ \gamma > 0. \tag{4.17}$$

Then

$$\Lambda_{n,t} \overset{\mathcal{L}}{=} nS_t$$

and, in the notations of Theorem 4.2.1, $\bar{Y}$ is normal with mean zero and variance σ^2. Keeping in mind that

$$\mathbb{E}\ln(1 + fr_{n,k}) = \mathbb{E}\ln(1 + fr_{n,1}) = \left(f\mu - \frac{f^2\sigma^2}{2}\right)n^{-1} + o(n^{-1}),$$

we repeat the arguments from [59, Example 1] to conclude that

$$\lim_{n\to\infty} \ln W^{n,f}_t \overset{\mathcal{L}}{=} \left(f\mu - \frac{f^2\sigma^2}{2}\right)S_t + Z_t,$$

where Z_1 has symmetric 2α-stable distribution. By (4.17),

$$S_t \overset{d}{=} t^{1/\alpha} S_1, \ \lim_{t\to+\infty} t^{-1/\alpha} Z_t \overset{d}{=} \lim_{t\to+\infty} t^{-1/(2\alpha)} Z_1 \overset{d}{=} 0,$$

and the "natural" long term growth rate becomes

$$\lim_{t\to\infty} t^{-1/\alpha}\Big(\lim_{n\to\infty} \ln W_t^{n,f}\Big) \stackrel{d}{=} \Big(f\mu - \frac{f^2\sigma^2}{2}\Big)S_1,$$

which is random, but, for each realization of S, is still maximized by f^* from (4.12). Therefore, if the time with which we compound our wealth is random, then the growth rate is also random as we don't know when we will stop compounding, yet it is still maximized by a deterministic fraction. Note that, for the purpose of this computation, the (stochastic) dependence between the processes S and Z is not important.

4.3 A Heavy-Tailed Model

With the results of the previous sections in mind, we consider the following *high frequency* generalization of (2.1) and (2.17):

$$r_{n,k} = \frac{P_{k/n} - P_{(k-1)/n}}{P_{(k-1)/n}},\ k = 1, 2, \ldots, \tag{4.18}$$

where the process $P = \{P_t\}_{t\geq 0}$, has the form $P_t = e^{R_t}$, and $R = \{R_t\}_{t\geq 0}$ is a Lévy process. In other words,

$$r_{n,k} = e^{R_{k/n} - R_{(k-1)/n}} - 1. \tag{4.19}$$

By the Lévy Itô decomposition (3.0.8), the process $R = \{R_t\}_{t\geq 0}$ can be decomposed into a drift, diffusion/small jump, and large jump components:

$$R_t = \mu t + \sigma B_t + \int_0^t \!\!\!\!\!\!- \int_{-1}^{1} x\big(\mu^R(dx, ds) - F^R(dx)ds\big) + \int_0^t \int_{|x|>1} x\mu^R(dx, ds); \tag{4.20}$$

we use the notation $\int\!\!\!\!\!\!-_a^b$ first introduced in (**??**).

The function F^R in (4.20) is a non-random non-negative measure on $(-\infty, 0)\bigcup(0, +\infty)$

such that

$$\int_{-\infty}^{+\infty} \min(x^2, 1) F^R(dx) < \infty.$$

Equality (4.20) has a natural interpretation in terms of financial risks [80]: the drift represents the edge ("guaranteed" return), diffusion and small jumps represent small fluctuations of returns, and the large jump component represents (sudden) large changes in returns. Similar to (4.5), the corresponding wealth process is

$$W_t^{n,f} = \prod_{k=1}^{\lfloor nt \rfloor} \left(1 + f r_{n,k}\right). \tag{4.21}$$

Denote by $\mathbb{D}((0,T))$ the Skorohod space on $(0,T)$ [52, Chapter VI]. We have the following generalization of Theorem 4.1.1.

Theorem 4.3.1. *Consider the family of processes* $\left\{\left(W_t^{n,f}\right)_{t\in[0,T]}\right\}_{n\geq 1}$, $f \in [0,1]$, *defined by* (4.21). *If* $r_{n,k}$ *is given by* (4.19), *with* $P_t = e^{R_t}$, *and* $R = R_t$ *is a Lévy process with representation* (4.20) *and* $\mathbb{E}|R_1| < \infty$, *then, for every* $f \in [0,1]$ *and* $T > 0$,

$$\lim_{n\to\infty} W^{n,f} \overset{\mathcal{L}}{=} W^f$$

in $\mathbb{D}((0,T))$, *where*

$$\begin{aligned} W_t^f = \exp\Bigg(& f R_t + \frac{f(1-f)\sigma^2}{2} t \\ & + \int_0^t \int_{-\infty}^{+\infty} \Big[\ln\left(1 + f(e^x - 1)\right) - fx \Big] \mu^R(dx, ds) \Bigg). \end{aligned} \tag{4.22}$$

Proof. By (4.19) and (4.21),

$$\ln W_t^{n,f} = \sum_{k=1}^{\lfloor nt \rfloor} \ln\left(1 + f\left(e^{R_{k/n} - R_{(k-1)/n}} - 1\right)\right).$$

Step 1: For $s \in \left(\frac{k-1}{n}, \frac{k}{n}\right]$, let

$$r_s^{n,k} = e^{R_s - R_{(k-1)/n}} - 1, \tag{4.23}$$

and apply the Itô's formula [73, Theorem II.32] to the process

$$s \mapsto \ln\left(1 + f r_s^{n,k}\right), \quad s \in \left(\frac{k-1}{n}, \frac{k}{n}\right].$$

The result is

$$\begin{aligned}\ln\left(1 + f r_s^{n,k}\right) &= \int_{\frac{k-1}{n}}^{s} \frac{f(1 + r_{u-}^{n,k})}{1 + f r_{u-}^{n,k}}\, dR_u + \frac{\sigma^2}{2} \int_{\frac{k-1}{n}}^{s} \frac{f(1-f)(1 + r_{u-}^{n,k})}{\left(1 + f r_{u-}^{n,k}\right)^2}\, du \\ &\quad + \int_{\frac{k-1}{n}}^{s} \int_{-\infty}^{+\infty} \Bigg[\ln\left(1 - f + f e^x (r_{u-}^{n,k} + 1)\right) \\ &\quad - \ln(1 + f r_{u-}^{n,k}) - x \frac{f(1 + r_{u-}^{n,k})}{1 + f r_{u-}^{n,k}} \Bigg] \mu^R(dx, du).\end{aligned}$$

Step 2: Putting $s = \frac{k}{n}$ in the above equality and summing over k, we derive the following expression for $\ln W_t^{n,f}$:

$$\begin{aligned}\ln W_t^{n,f} &= \sum_{k=1}^{\lfloor nt \rfloor} \left(\int_{\frac{k-1}{n}}^{\frac{k}{n}} h_{n,k}^{(1)}(s)\, dR_s + \int_{\frac{k-1}{n}}^{\frac{k}{n}} h_{n,k}^{(2)}(s)\, ds + \int_{\frac{k-1}{n}}^{\frac{k}{n}} \int_{-\infty}^{+\infty} h_{n,k}^{(3)}(s,x) \mu^R(dx, du) \right) \\ &= \int_0^t H_{n,t}^{(1)}(s)\, dR_s + \int_0^t H_{n,t}^{(2)}(s)\, ds + \int_0^t \int_{-\infty}^{+\infty} H_{n,t}^{(3)}(s,x) \mu^R(dx, ds)\,, \end{aligned} \tag{4.24}$$

where

$$\begin{aligned} h_{n,k}^{(1)}(s) &= \frac{f(1 + r_{s-}^{n,k})}{1 + f r_{s-}^{n,k}}\,, \quad h_{n,k}^{(2)}(s) = \frac{\sigma f(1-f)}{2} \frac{1 + r_{s-}^{n,k}}{(1 + f r_{s-}^{n,k})^2}\,, \\ h_{n,k}^{(3)}(s,x) &= \ln\left(1 - f + f e^x (r_{s-}^{n,k} + 1)\right) - \ln(1 + f r_{s-}^{n,k}) - f x \frac{1 + r_{s-}^{n,k}}{1 + f r_{s-}^{n,k}}\,; \\ H_{n,t}^{(i)}(s) &= \sum_{k=1}^{\lfloor nt \rfloor} h_{n,k}^{(i)}(s) \mathbf{1}_{(\frac{k-1}{n}, \frac{k}{n}]}(s), \ i = 1, 2; \quad H_{n,t}^{(3)}(s,x) = \sum_{k=1}^{\lfloor nt \rfloor} h_{n,k}^{(3)}(s,x) \mathbf{1}_{(\frac{k-1}{n}, \frac{k}{n}]}(s). \end{aligned}$$

Step 3: Because

$$\lim_{n\to\infty,\, k/n\to s} R_{(k-1)/n} = R_{s-},$$

equality (4.23) implies

$$\lim_{n\to+\infty,\, k/n\to s} r_{s-}^{n,k} = 0$$

for all s. Consequently, we have the following convergence in probability:

$$\lim_{n\to+\infty} H_{n,t}^{(1)}(s) = f,\ \lim_{n\to+\infty} H_{n,t}^{(2)}(s) = \frac{\sigma^2 f(1-f)}{2},$$
$$\lim_{n\to+\infty} H_{n,t}^{(2)}(s,x) = \ln\big(1 + f(e^x - 1)\big) - fx.$$

To pass to the corresponding limits in (4.24), we need suitable bounds on the functions $H^{(i)}$, $i = 1, 2, 3$.

Using the inequalities

$$0 < \frac{1+y}{1+ay} \le \frac{1}{a},\quad 0 < \frac{1+y}{(1+ay)^2} \le \frac{1}{4a(1-a)},\quad y > -1,\ a \in (0,1),$$

we conclude that

$$0 < h_{n,k}^{(1)}(s) \le 1,\ 0 < h_{n,k}^{(2)}(s) \le \sigma^2,$$

and therefore

$$0 < H_{n,t}^{(1)}(s) \le 1,\ 0 < H_{n,t}^{(2)}(s) \le \sigma^2. \tag{4.25}$$

Similarly, for $f \in (0,1)$ and $y > -1$,

$$\left|\ln\frac{1 - f + fe^x(y+1)}{1+fy} - fx\frac{1+y}{1+fy}\right| \le 2\big(|x| \wedge |x|^2\big), \tag{4.26}$$

so that

$$|h_{n,k}^{(3)}(s,x)| \le 2\big(|x| \wedge |x|^2\big)$$

and

$$\left|H_{n,t}^{(3)}(s)\right| \le 2\left(|x| \wedge |x|^2\right). \tag{4.27}$$

To verify (4.26), fix $f \in (0,1)$ and $y > -1$, and define the function

$$z(x) = \ln \frac{1 - f + fe^x(y+1)}{1 + fy}, \ x \in \mathbb{R}.$$

By direct computation,

$$\begin{aligned} z(0) &= 0, \\ z'(x) &= \frac{fe^x(y+1)}{1 - f + fe^x(y+1)} = 1 - \frac{1-f}{1 - f + fe^x(y+1)}, \\ z'(0) &= \frac{f(y+1)}{1+fy}, \end{aligned}$$

so that, using the Taylor formula,

$$\ln \frac{1 - f + fe^x(y+1)}{1+fy} - fx\frac{1+y}{1+fy} = z(x) - z(0) - xz'(0) = \int_0^x (x-u)z''(u)du. \tag{4.28}$$

It remains to notice that

$$0 \le z'(x) \le 1, \quad 0 \le z''(x) \le 1,$$

and then (4.26) follows from (4.28).

With (4.25) and (4.27) in mind, the dominated convergence theorem [73, Theorem IV.32] makes it possible to pass to the limit in probability in (4.24); the convergence in the space $\mathbb{D}$ then follows from the general results of [52, Section IX.5.12]. □

Now that the process R_t is exponentiated,the analog of (2.10) becomes $\mathbb{E}[|R_1|] < \infty$, which, by Theorem **??**, is equivalent to

$$\int_{|x|>1} |x| F^R(dx) < \infty.$$

The following is a representation of the long-term growth rate of the limiting wealth process W^f which is a generalization of both (2.5) and (4.11).

Theorem 4.3.2. *Let $R = \{R_t\}_{t\geq 0}$ be a Lévy process with representation* (4.20).

If $\mathbb{E}[|R_1|] < \infty$, then the process $W^f = \{W_t^f\}_{t\geq 0}$ defined in (4.22) *satisfies*

$$\begin{aligned}\lim_{t\to+\infty} \frac{\ln W_t^f}{t} = f\left(\mu + \int_{|x|>1} xF^R(dx)\right) + \frac{f(1-f)\sigma^2}{2} \\ + \int_{-\infty}^{+\infty} \left[\ln\left(1 + f(e^x - 1)\right) - fx\right] F^R(dx), \quad a.s.\end{aligned} \tag{4.29}$$

Proof. By (4.22),

$$\frac{\ln W_t^f}{t} = f\frac{R_t}{t} + \frac{f(1-f)\sigma^2}{2} + \frac{1}{t}\int_0^t \int_{-\infty}^{+\infty} \left[\ln\left(1 + f(e^x - 1)\right) - fx\right] \mu^R(dx, ds).$$

It remains to apply the law of large numbers for Lévy processes [75, Theorem 36.5]. □

If, in addition, we assume that

$$\int_{-1}^{1} |x| F^R(dx) < \infty,$$

that is, the small-jump component of R has bounded variation, then, after a change of variables and re-arrangement of terms, (4.29) becomes

$$g_R(f) = f\bar{\mu} - \frac{f^2\sigma^2}{2} + \int_{-1}^{\infty} \ln(1 + fx) F^{\tilde{R}}(dx), \tag{4.30}$$

where

$$\bar{\mu} = \mu - \int_{-1}^{1} xF^R(dx) + \frac{\sigma^2}{2},$$

and

$$F^{\tilde{R}}(dx) = F^R(d(\ln(1 + x))).$$

Equality (4.30) serves as a generalization of both (2.5) and (4.11) when we're in continuous

time and have processes not only have drift and diffusion, but can jump as well.

Note $\tilde{R}$ is the process such that $\mathcal{E}(\tilde{R}) = e^R$ as in [35, Section 8.4], which corresponds to the stochastic exponential process.

Similar to Proposition 2.3.2, we also have the following sufficient conditions to ensure the constraints of no shorting and no leverage for the optimal exposure.

Theorem 4.3.3. *In the setting of Theorem 4.3.2, denote the right-hand side of* (4.29) *by* $g_R(f)$ *and assume that*

$$\lim_{f\to 0+}\int_{-\infty}^{+\infty}\left(\frac{e^x-1}{1+f(e^x-1)}-x\right)F^R(dx) > -\left(\mu+\frac{\sigma^2}{2}+\int_{|x|>1}xF^R(dx)\right),$$
$$\lim_{f\to 1-}\int_{-\infty}^{+\infty}\left(\frac{e^x-1}{1+f(e^x-1)}-x\right)F^R(dx) < -\left(\mu+\int_{|x|>1}xF^R(dx)\right),$$

Then there exists a unique $f^* \in (0,1)$ *such that*

$$g_R(f) < g_R(f^*)$$

for all f *in the domain of* g_R.

4.4 Future Work

When trying to build a practical theory of investment based on the Kelly Criterion, there are some glaring issues that need to be addressed. The following questions were posed by Ed Thorp in [86] and they aim to bring light to some issues not addressed in traditional Kelly Criterion literature that could make the theory more practical by investigating these issues. Not only are questions helpful from a practical point of view, but it appears there are many possible paths for academic research to aid in the development of these useful adaptations to the Kelly Criterion. I attempt to summarize the issues and give my preliminary thoughts about research paths.

Opportunity Costs

This is a problem of a gambler placing simultaneous bets at a given instance in time, such as holding a portfolio with assets that have some dependence structure between them. In particular, when making our decision, to avoid overbetting, we need to know all the bets that we currently have and are currently considering adding, as well as their joint properties. We could attempt to solve this issue by finding, either explicitly or numerically, the optimal $\vec{f}^*$ when we have a log-return vector $\vec{R}_t = (R_t^1, \cdots, R_t^d)^\top \in \mathbf{R}^d$ that has a particular dependence structure, for example, $\gamma_t(i,j) = Cov(R_t^i, R_t^j)$. I think the mathematics of heavy tailed copulas could help answer this question, see [35, Chapter 5]. Also, for intuition on how this can be used and why it would be important, see [42, Section 10.2]. At the end of this section, Ethier has an example where he has two betting scenarios, one which is more diversified than the other, with an edge, but you would choose the one with lower edge per trial since the lower edge per trial strategy leads to more diversification and hence a higher long term growth rate with lower volatility.

The opportunity cost refers to over-diversifying in the presence of large edges as this can lead to missed opportunities for compounding. Also see [83].

Risk of Volatility and Risk of Ruin Constraints

As discussed in [84] and [87], betting your optimal fraction f^* can lead to large swings in your wealth in the short run, and there are drawdowns which are too large for the comfort of many investors. So a heuristic that Thorp implemented was to use a percentage of the optimal fraction, called "fractional Kelly," which tamed the large drawdowns while also reducing the long term growth rate. This is very important as drawdowns destroy long-run compounding! One way this can be formalized is by adding a constraint to the long term grown rate

optimization problem. Note that a large drawdown can lead to effective ruin if someone decides to pull all of their money from a given investment strategy. Thus drawdown risk is really just the risk of effective ruin. We are also interested in the most important type of risk which is risk of literal ruin for a given initial wealth and betting fraction,

$$\mathbb{P}\left(\inf_{t\geq 0} W_t^f \leq 0 \middle| W_0 = w\right) =: \Psi(w, f).$$

It appears that one way these problems could be investigated is with the tools of fluctuation theory for Lévy processes, see [8, Chapter 11]. Another possible approach to investigating these problems is with the tools of stochastic control. There are many additional approaches in [87, Part III].

Model Uncertainty

Another reason for using a "fractional Kelly" strategy, apart from reducing short term drawdown risk, is when we are unsure about the distribution of our returns, which in finance is an understatement. If we get the distribution wrong or make many estimation errors, then our optimal fraction could still lead to over-betting. A useful investigation is in [23], and could lead to further investigations.

This section is particularly essential for connecting the Kelly Criterion with forecasting models which is what all quantitative investors or traders use to attempt to find edges. That is, when returns are being generated by a parametric statistical model,

$$r \approx \hat{r}(\vec{X}; \vec{\theta}).$$

One could further try to model optimal exposures directly as a statistical parametric model as in [19], that is,

$$f \approx \hat{f}(\vec{X}; \vec{\theta}).$$

Black Swans

A main goal of this thesis is attempting to resolve this issue. It involves considering the Kelly Criterion under more general, including fat-tailed, Lévy processes.

The "Long Run"

Here one would be concerned with convergence rates of the Strong Law of Large Numbers and its variants. In other words, how long do I need to keep compounding my wealth, or about how many bets do I need to place, to achieve my long term growth rate. Note that not every gambler can make enough bets for the long run to kick in, at least with a high probability. These asymptotic properties that are the main staple of the Kelly Criterion will not necessarily be true if the gambler or investor doesn't have enough opportunities to make it to the long run.

Numerical Investigations

It is the subject of future research to give an detailed numerical and empirical study of the optimal exposure $f^*(\mu, \sigma, F^R)$ in terms of the Lévy characteristics of the log-returns R for various well known heavy tailed Lévy processes, as well as processes that we can fit to return data. One such study in this direction is by Steve Schulist of PIMCO [76].

Part II

Finding the Edge

Chapter 5

Volatility

Volatility as a vague concept refers to both the frequency and magnitude of the ups and downs of a some observable phenomenon over time. Despite the phenomenon being observable, its variation over time is, in fact, not directly observable. Many efforts have been made to estimate, model, and forecast this latent process due to the wide-ranging applications of volatility. Particularly in finance, the range of uses for predicting volatility include risk management, choosing portfolio weights for optimal portfolio construction[1], pricing and hedging of derivative contracts, and trading on mispricings found in the options market and, more generally, the volatility market.

Historically, one of the most common variables financial economists have focused on studying is the volatility of financial assets' return series, which is an inherently unobserved quantity. We do not sample volatility, we sample prices, which can, themselves, be tricky. We then aggregate those sampled prices into a measure which summarizes past price fluctuations, ex-post, and possibly predicts future price fluctuations, ex-ante. To further complicate the study of volatility, the frequency and magnitude of ups and downs of prices appears to change throughout time.

Those with a need to understand volatility of returns are faced with answering the fol-

[1] As in the Kelly criterion in part I.

lowing questions:

(i) How to *measure* volatility of returns, ex-post, over a given time period? That is, volatility estimation.

(ii) How to *predict* volatility of returns, ex-ante, over a given time period? That is, volatility forecasting.

Consider a random variable R modeling the log-gross return[2] of some financial asset, which we'll often reference as simply the *return*. We can define the volatility as the standard deviation of this return. As a first approximation, we might consider estimating this quantity with sample moments. Given we have observations $\{R_t\}_{t=1}^T$, we define the *sample volatility* as

$$\hat{\sigma} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(R_t - \hat{\mu})^2}, \tag{5.1}$$

where $\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} R_t$. At short horizons, say daily, the expected return is often very small implying its square is effectively zero. Thus it is fair to replace (5.1) with

$$\hat{\sigma} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} R_t^2}.$$

This could easily be improved by using a moving average to discard stale data and take advantage of the observed phenomenon of volatility clustering. Of course, this method is subject to balancing the trade-off between "staleness" and noise of the estimator. This trade-off is commonly referred to as the bias-variance trade-off where "staleness" refers to bias and variance refers to noise. If you estimate with a shorter sample, your estimator will have more noise despite being more in-tune with the current volatility environment; if you estimate with a longer sample, your estimator will have less noise but will also be less reflective of the current volatility environment. The traditional approach to optimally balance these

[2]The log-gross return is also known as the continuously compounded return.

trade-offs has been to use the (G)ARCH-type models as first introduced by Engle [40] and Bollerslev [14]. Additionally, in a world of stochastic observations, particularly stochastic volatility, where the distribution of volatility is changing over time, taking an average over many days as in (5.1) doesn't make much sense due to the lack of independent and identical distributions of squared returns.

Another simpler, yet effective, approach for the estimation of volatility, ex-post, has been to use range-based estimators, [69], [6]. In any given trading period, the most commonly recorded prices are the open, close, high, and low price. If we're considering daily trading periods, then the daily range is the difference between the high and low price

$$P_t^H - P_t^L; \tag{5.2}$$

and if we want to reduce effects related to price levels, we would need to normalize (5.2), typically by the closing price

$$\hat{\sigma} = \frac{P_t^H - P_t^L}{P_t^C}.$$

Different flavors and improvements of range-based estimators for volatility have been created and studied by Parkinson [69], Garman and Klass [44], Rogers and Satchell [74], and Yang and Zhang [90].

A more modern approach to estimating volatility, ex-post, is to use *better* data, where *better* often refers to higher frequency data, which gives rise to the notion of *realized volatility*.

5.1 Realized Variance

Due to the unobservable nature of variance, it is common practice to rely on an explicit model of the (G)ARCH-type, for measurement. However, using higher frequency data, there exists a model-free methodology to measure variance which, in turn, provides a natural benchmark for forecast evaluation purposes. Given this model-free methodology for meas-

uring realized variance, one can begin to treat volatility as an observed process and hope to use more traditional forecasting techniques, which treat the variables as observables.

5.1.1 Measuring Realized Variance

Intuitively, a world with availability of tick-by-tick transaction prices and quotes should yield opportunities to estimate volatility with a high degree of precision. If we assume prices evolve in continuous time, then this idea can be made precise [1]. Consider working on a daily time horizon with access to a rich data set including minute-by-minute price data, that is, we are no longer restricted to working solely with open, close, high, and low daily price data [3]. A technique, which has become the de-facto standard, to exploit this high frequency data is to use a quadratic variation estimator for an ex-post volatility measurement, as was initially studied in [5]. Such estimators are called *realized volatility* or, for the more mathematically inclined, *realized variation* estimators.

For a given price process P, define *returns* over a given window of time $[t_j, t_{j+1}]$ as the continuously compounded return

$$R_{t_j,t_{j+1}} := \ln\left(\frac{P_{t_{j+1}}}{P_{t_j}}\right).$$

Using returns sampled at the Δ intraday frequency on day t, we construct the realized variance estimator by summing the squared returns computed every Δ-period of time

$$RV_t^{\Delta} := \sum_{j=1}^{\lfloor t/\Delta \rfloor} R^2_{t-1+(j-1)\Delta,\ t-1+j\Delta} \tag{5.3}$$

where $1/\Delta$ could be, as it is for our application, 78 for 5-min returns over a 6.5 hour trading day, or 144 for 10-min returns in 24 hour markets.

In the theoretical world of continuous time price evolution, sending $\Delta \to 0$, which cor-

[3] Such a dataset can be constructed from the NYSE TAQ (Trades and Quotes) database.

responds to continuous sampling, will lead the realized variance estimator to approach the true integrated variance of the price process

$$RV_t^{\Delta} \underset{\Delta \to 0}{\longrightarrow} IV_t, \tag{5.4}$$

in some appropriate sense, where the integrated variance[4] is defined as the quadratic variation of returns over the $[t-1, t]$ period

$$IV_t = [R, R]_t - [R, R]_{t-1} \ . \tag{5.5}$$

Under the rather weak assumption of returns evolving as a continuous time semimartingale, we have a method, with theoretical justification, for constructing daily estimates of realized return variation, to any desired degree of precision, using directly observable data. Quadratic variation of a continuous time semimartingale is best interpreted as the actual return variation that transpired over a given period of time, and as such it is the clear target for realized variance measurement. Given that the quadratic variation of a process is the realization of a random quantity, that is difficult to forecast, it serves as a reference for which forecasts should be compared against. Using realized variance as an ex-post variance proxy is completely model-free, requiring nothing more than intraday prices. This allows one to harness the information inherent in high-frequency returns for assessment of lower frequency return variation.

Some issues with high frequency data include the lack of being able to actually achieve the limit. If you take finer and finer samples, you will run into missing data and liquidity effects that will unfavorably bias the estimates. We do not see prices over night which could lead to huge swings in prices between the open and the close and hence jumps in volatility. Further microstructure issues include discreteness of the price grid, as well as bid-ask spreads. This implies high-frequency realized volatility estimators are still noisy, due to a non-negligible

[4]Here, we use IV to mean integrated variance, as opposed to implied volatility.

error term in the approximation (5.4), and likely biased particularly if the sampling frequency is too high.

5.1.2 Forecasting Realized Variance

Given realized variance is a measurable proxy for the latent quadratic variation, and the associated measurement errors over time are uncorrelated, one could imagine using standard time series models to capture the temporal features of return variance. The most common class of time series models is the autoregressive fractionally integrated moving averaging (ARFIMA(p,d,q)). The class of ARFIMA models does a good job of capturing the statistical properties of logarithmic realized variance [7]. Such statistical properties include a long memory dependence structure in variance, as well as logarithmic realized variance being much closer to homoskedastic and approximately unconditionally Gaussian. Such a model, however, leads to a number of practical modeling issues. These issues include the choice of the sampling frequency at which realized variance measures are constructed, the difficulty in disentangling the jumps and diffusive variance components of the realized variance process, and finally, the approach used to best accommodate the indications of *long memory*. If, in good fortune, one can adequately respond to these practical modeling issues, then forecasting is straightforward once the realized variance has been cast in a traditional time series modeling framework, and the model parameters have been estimated.

Additionally, one must pay attention to how they compute realized variance for a calendar period when the trading day has an official closing. For example, the over-the-counter foreign exchange market has a 24-hour trading period, so this is a minor problem, but this is typically not the case for equity or fixed-income markets. If one uses intraday returns to estimate realized variance, they must be aware the actual object being estimated is the *intraday* realized variance and not the full day's variance. This estimator is not incorporating overnight variance, despite the price process still moving overnight. It is common to see substantial price changes between a market's close and subsequent open, in which case the

daily realized variance estimator is missing out on the overnight variance.

Finally, the preferred sampling frequency can become quite a challenge when the underlying asset is relatively illiquid. If updated price observations are only available intermittently throughout a trading day, then the effective sampling frequency required to generate the *high frequency* returns is lower than the frequency intended to compute the realized variance. To further enhance the problem of an illiquid price series, their bid-ask spreads are typically larger and are more sensitive to random fluctuations in order flow. This implies the associated return series, constructed from the noisy bid-ask spreads, will contain a relatively large amount of noise as well. A simple solution to help reduce the problems of large bid-ask spreads and intermittent trading would be to use a lower sampling frequency, but this too comes at a cost of increasing the measurement error in realized variance. Many of these issues have been previously studied and an even more comprehensive coverage of the literature is presented in [7].

5.2 Forecasting Firm Volatility

Financial return series presents stylized facts that induce challenges to classical econometric modeling techniques. These include autocorrelations of the *squared* returns that exhibit very strong persistence which can last for long periods of time, as well as return distributions being heavy tailed and highly peaked while displaying a very slow convergence to the normal distribution as the horizon increases. Additional stylized facts are summarized in [34]. In principle, one hopes volatility models capture the most important stylized facts of stock return volatility; this includes time series clustering, negative correlation with returns known as the leverage effect, log-normality, and long-memory. As quoted in [62], "large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes," which is an informal way of describing time series clustering.

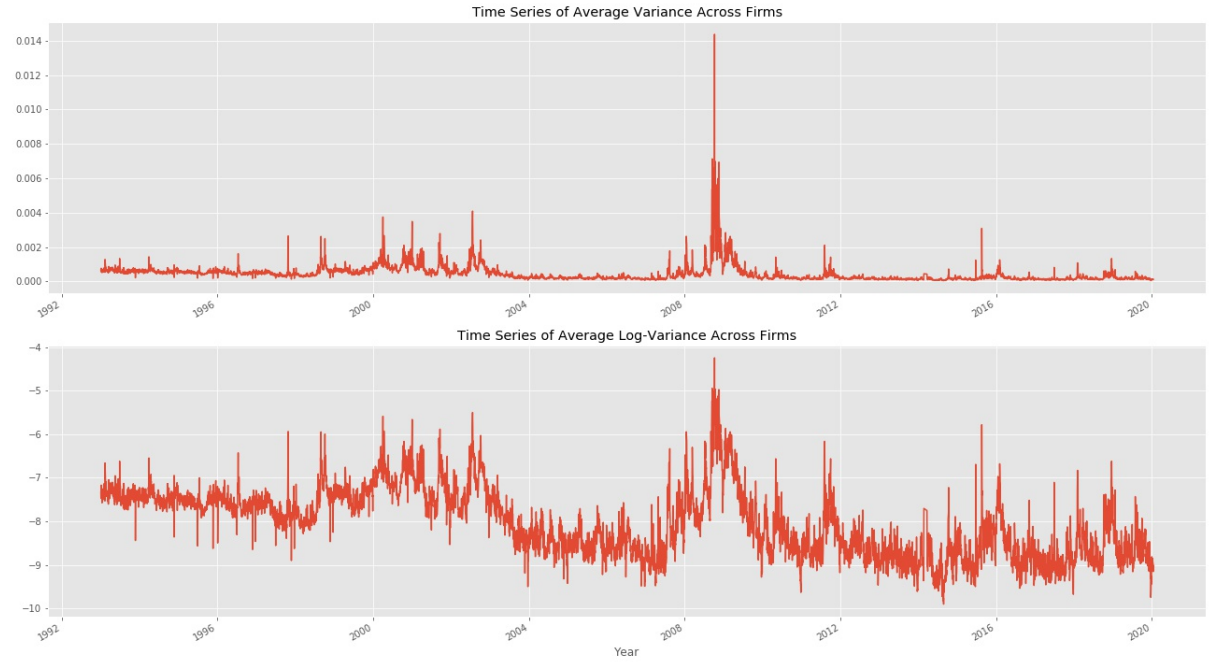

Figure 5.1: We plot the time-series of firm-level realized variance, we note that there appears to be lots of common movement (co-movement) amongst the firms. This suggests the existence of a common factor structure that explains the common movement.

We see time variation in the firm-level average of realized variances, leading us to believe there exists a common factor structure in variances of individual firms. The time-varying average firm-level volatility points to predictability, in the sense of it not following a random walk. Despite this observation, improving the forecasting ability of the level of volatility of some financial asset's return series turns out to be a highly non-trivial pursuit. As described in [18], forecasting the future level of volatility is difficult for multiple reasons. Volatility forecasts are sensitive to the specification of the model; in the case of linear regression this is simply the predictor variable specification. It is important to strike the right balance between capturing the signal in the data and overfitting the noise. Depending on the particular model specification, the volatility forecast can be subject to overfitting the data, rather than extracting the salient features of the data. Given the task of correct model specification, the problem of volatility forecasting is further enhanced by the need to correctly estimate the model's parameters, which can be difficult due to the latent nature of volatility. The further

the estimated parameters are from the true parameters, the worse the volatility forecasts will be. Most volatility forecasts depend on current and past levels of volatility which are proxied, or estimated, by noisy variables, which leads, even in the presence of perfectly specified and estimated models, to inherited or possibly amplified uncertainty in the forecast. In our set of experiments, we model the conditional realized variance of returns, and we focus on different model classes, model specification, and out-of-sample forecasting.

5.3 Methodology

Contrary to most conditional volatility forecasting studies, which test their models on a relatively small number of firms or a small number of distinct asset classes, we investigate our volatility forecasts on the entire cross section of stocks in the S&P 500 index at the daily frequency starting in 1/1993 and ending in 6/2019.

To achieve the most reliable estimate of generalization error while using a static data set, we use a walk-forward out-of-sample forecasting methodology, which to practitioners is known as a back-test, and in the machine learning community as leave-one-out cross-validation (LOOCV).

Given our out-of-sample forecasts, we measure the deviation from the realized variance measurements according to loss functions commonly used in the volatility forecasting literature and theoretically studied in Patton [71]. We consider mean-squared error, mean-absolute deviation, QLIKE, and Mincer-Zarnowitz regressions. Additionally, we consider, empirically, the differences between aggregating forecast errors across a time-series, a cross-section, and the pooled panel of forecasts.

5.3.1 Out-of-Sample Analysis

Though often hailed as a virtue in machine learning, "data mining" is typically viewed as a demerit in empirical asset pricing, and empirical finance more broadly. Data mining

refers to running many repeated experiments to find the one specification that best fits the data, but does not generalize well on unseen data, that is, doesn't fit well on data that was used to estimate the model. One way to account for data mining is to adjust the statistical significance threshold according to the number of parameter searches undertaken, this is known as the problem of multiple-testing in machine learning [50, Section 18.7] and was studied in the finance literature by Harvey, et al. [48]. Another reasonable response to the problem of data mining is to use an out-of-sample (OOS) walk-forward validation approach, which financial practitioners call a *back-test*. For more on the issues of data snooping, see [43, Section 32.1.3.2].

True out-of-sample tests are often not feasible because we're working with a static data set, so we must rely on pseudo-OOS tests, which imitate real-time forecasting on historical data, ex-post. We must concede that pseudo-OOS tests can still be subjected to p-hacking, multiple testing, and data mining as described in [33]. We perform a rolling out-of-sample analysis on our panel of data (5.6). The vague idea of this error estimation technique is to forecast exactly as we would use the forecast in real-time on live, unseen data. More specifically, on every day, and for every firm, we split our sample into a set of data that is used to estimate our forecasting model and a set of data that is used to evaluate the forecast against the measured response variable. Next we roll the sets forward by one period and repeat. Such sets have varying names depending on the academic circle. The set used to estimate the model can be called the estimation set, training set, or learning set. The set used to evaluate the the model's forecast against the measured variable can be called the out-of-sample set, test set, or validation set.

An additional reason for using a walk-forward out-of-sample analysis is to account for structural change in the data. Structural changes can come from two sources, either the ambient environment changing or environment participants learning from past data and adapting their behavior going forward. A particularly simple technique for addressing the problem of structural change is to use a rolling window estimation set where data prior to

a certain time period is completely discarded, or used with a weight of zero. Another data weighting scheme is to use exponential weighting that gradually down-weights old data. It is not exactly clear which schemes make the most sense in financial applications. With these structural change concerns in mind, we construct our estimation set by employing a rolling window approach which gives a simple technique for estimating time varying conditional volatilities for a given estimation window $[t - W, t]$. The window length, W, directly determines the bias-variance trade-off of the forecast, with larger values of W increasing the forecast bias, due to nonstationarity of financial return data, while reducing the forecast's variance.

More specifically, we estimate our model for conditional variance on every day t, and for every stock i in the sample[5], using $W = 250$ days of observations of the predictor variables. Every time t we estimate our forecast using data from $t - W$ to $t - 1$, producing a data set of the form

$$\left\{ \underbrace{(\mathbf{x}_{t-W}, y_{i,t-W+1}), \ldots, (\mathbf{x}_{t-2}, y_{i,t-1})}_{\text{estimation (training) set}}, (\mathbf{x}_{t-1}, y_{i,t}), \underbrace{(\mathbf{x}_t, y_{i,t+1})}_{\text{OOS (test) set}} \right\}^{6}, \tag{5.6}$$

which gives a forecast $\hat{y}_{i,t+1} = f(\mathbf{x}_t; \hat{\beta}_t)$ for the next period $t+1$, which we will compare against the next period's observed value $y_{i,t+1}$.

To preserve an out-of-sample analysis, it is important to recognize the information we have available at time t when we estimate our forecast. Another way of saying this is we must preserve adaptedness of our forecasts to the available filtration $\mathbb{F} = \{\mathcal{F}_t\}$. We're trying to forecast the $t + 1$ daily realized variance at time t using information up to $3:55$pm ET. We use data before $3:55$pm ET in order to allow for a tradeable strategy employing the variance forecasts. Using data before the close of the trading day, 4pm ET, allows us to place trades on day t using our estimated forecast for $t + 1$.

The order of operations is as follows:

[5] There are 1078 unique firms in our sample starting from 1993 and ending in 2019.

[6] We do not have a "standard" machine learning setup for every stock and for every day. The observations $(\mathbf{x}, y)$ are not independent and identically distributed, but rather have dependence and a time-varying distribution between the observations.

1. Estimate a forecast model which is trained between the close in period $t-1$ and before $3:55$pm ET in period t.

2. Measure the predictor variables at time $3:55$pm ET, $\mathbf{x}_t \in \mathcal{F}_t^{3:55}$, and plug them into our estimated model with produces a forecast $\hat{y}_{i,t+1} = f(\mathbf{x}_t; \hat{\beta}_t) \in \mathcal{F}_t^{3:55}$.

3. Place trades before the close of the trading day at 4pm ET in period t to be realized in the next period $t+1$.

4. Measure the quality of the forecast in period $t+1$ based on $\hat{y}_{i,t+1}$ and the measured $y_{i,t+1}$.

Because we estimate our model between the close of period $t-1$ and before $3:55$pm ET in period t, we cannot include the observation $(\mathbf{x}_{t-1}, y_{i,t})$ in our estimation set. We do not have access to $y_{i,t}$, which we wouldn't have access to until after the close of period t, even though we do have access to $\mathbf{x}_{i,t-1}$. Including $(\mathbf{x}_{t-1}, y_{i,t})$ in our estimation set would destroy the out-of-sample analysis by introducing look-ahead bias to a trading strategy making use of our forecast.

Another subtle issue with our sample is not all stocks remain in the sample the entire estimation or out-of-sample set. As firms can enter and exit the S&P 500 index based off market valuations, the panel of data becomes unbalanced. Because we are working with an unbalanced panel at each point in the sample (5.6), we choose to apply a filter which drops any firms that enter or exit the estimation set. This allows us to avoid having to impute values for firms that enter or exit the estimation set, and does not use introduce any look-ahead bias.

5.3.2 Out-of-Sample Performance Measurement

Using the walk-forward out-of-sample methodology to forecast the daily realized variance yields a sequence

$$\left\{(y_{i,t+1}, \hat{y}_{i,t+1}); 1 \leq t \leq T, 1 \leq i \leq N\right\} \tag{5.7}$$

of daily measured realized variance and the associated forecast for that day, for every firm. Let N be the number of firms and T be the number of trading days in our sample[7]. There exists familiar techniques for measuring volatility forecast error, which often vary depending on the empirical properties of the data, though most of these techniques do not address how to aggregate forecast error measurements in a large panel of forecasts, as we have.

In the volatility forecasting literature, numerous authors have expressed concern that a few extreme observations may have an excessively large impact on the outcomes of forecast evaluation and comparison tests. The common response to such concerns is to use loss functions that are less sensitive to large observations, such as mean absolute deviation or proportional error, instead of the traditional mean squared error loss function. Patton [71] theoretically shows this approach can still lead to incorrect inferences and selection of inferior forecasts.

The forecaster must then consider how different loss functions penalize deviations differently. The mean squared error is more sensitive to outliers relative to other commonly used loss functions. QLIKE[8], however, is an asymmetric loss function, viewed as a function of forecast value. The far left tail of the QLIKE loss function is far more sensitive, yet less sensitive to the far right tail of the loss function relative to MSE. This property of QLIKE partially solves the problem of volatility loss functions being dominated by a few very large observations by being robust to extreme observations in the right tail while sacrificing robustness in the left tail. Being a non-symmetric loss function, QLIKE penalizes positive and negative loss values differently which leaves the ranking of forecasts by QLIKE to incorrectly favor positively biased forecasts. Further problems arise through distortions in the rankings of competing forecasts when using a noisy volatility proxy. Our realized variance volatility proxy still contains noise, despite being less noisy than traditional proxies such as squared returns or ranged-based estimators. Patton [71] demonstrates the MSE and QLIKE loss functions are, theoretically, robust to noise in the volatility proxy for the lat-

[7] In our sample, $T = 6350$, $N = 1078$.

[8] The QLIKE loss function is defined as $\mathcal{L}(y, \hat{y}) := \log(\hat{y}) + \frac{y}{\hat{y}}$.

ent volatility process. Meddahi [64] shows the ranking of forecasts on the basis of R^2 from Mincer-Zarnowitz (MZ) regressions[9], where one regresses a noisy variance proxy on the variance forecast, is robust to noise in the true variance proxy, but still might not handle the few extreme observations problem we see in volatility forecasting. Corsi, [36, Section 3.3], also uses MZ-regressions for evaluating forecast performance. Given the data (5.7), we estimate the out-of-sample prediction error using a squared loss function, absolute loss function, and the QLIKE loss function analyzed in [71] and applied in [70]. We also use the R^2 from MZ regressions [66] where we regress the ex-post measurement onto the ex-ante forecast to measure performance of the forecasts, as well as the bias. For these MZ-regression based forecast evaluations, unbiasedness of the forecasts requires an intercept of zero, and a slope of one. MZ-regressions rank volatility forecast models by their ability to *explain* subsequent realized volatility measurements, as measured by R^2. When using the R^2 of MZ-regressions to rank forecasts, recall

$$R^2 = 1 - \frac{Var(\hat{\varepsilon})}{Var(y)} = 1 - \frac{\sum_i \hat{\varepsilon}_i^2}{\sum_i (y_i - \bar{y})^2},$$

where $\hat{\varepsilon} = y_i - \hat{y}_i$ which leads to the ratio $\frac{\sum_i \hat{\varepsilon}_i^2}{\sum_i (y_i - \bar{y})^2}$ being interpreted as the fraction of unexplained variance, that is, variance not explained by a model. For R^2 to increase, variance of the residuals needs to decrease. In the extreme case of perfect over-fitting where $\hat{y}_i = y_i$, for all i, we have $R^2 = 1$. If we consider a constant baseline model $\hat{y}_i = \bar{y}$, then $R^2 = 0$, with models forecasting worse than the mean leading to $R^2 < 0$.

Regardless of the choice of forecast error measurement, it has been known, despite the obvious increase in validity, that a walk-forward out-of-sample evaluation of forecast errors provides evidence that is hard to interpret. For example, Campbell and Thompson [24] find, when predicting the first moment of returns, a variable can have economic predictive power, yet still fail to outperform a naive benchmark when predicting using a walk-forward validation test. All of the different methods for forecast evaluation have imperfections, which is why in section 6.4 we rank volatility forecasts by evaluating returns of option portfolios

[9] MZ regressions take the form $y_t = a + b\hat{y}_t + \varepsilon_t$.

formed on the basis of the forecast.

5.4 Models

Historically, volatility modeling has evolved along two distinct paths corresponding to the statistical, $\mathbb{P}$, and risk-neutral, $\mathbb{Q}$, probability measures. The statistical $\mathbb{P}$ probability measure path has typically followed discrete-time (G)ARCH-type models as in Engle [40] and Bollerslev [14], while the risk-neutral $\mathbb{Q}$ probability measure path has typically followed continuous-time dynamics modeling as in [2]. The (G)ARCH-style literature has typically focused on explicitly modeling the stylized facts of volatility, such as autocorrelation and persistence.

Without explicitly focusing on modeling stylized facts of volatility in-sample we, instead, focus on out-of-sample forecasting of realized variance utilizing both high and low-dimensional machine learning models.

5.4.1 Heterogeneous Autoregressive Model of Realized Variance

A volatility forecasting model that has been demonstrated to have good forecasting performance, on a small sample of assets, is the heterogeneous autoregressive (HAR) model of Corsi 2009 [36]. This model is a simple alternative to the seminal (G)ARCH-type models of [40] and [14]; it uses ordinary least squares to regress a firm's daily realized volatility on its own lagged realized volatility at various horizons. Corsi finds by exploiting the information content in high frequency data to construct variance measures, and by removing the need for numerical optimization, the HAR model out-performs classical (G)ARCH-type models which explicitly model stylized facts of return volatility.

Using high-frequency intraday data to construct realized variance estimators, Corsi [36] makes a simple improvement to this realized variance estimator by modeling realized variance using an autoregressive model and lagged realized variance estimates over daily, weekly, and

monthly horizons,

$$RV_{t+1}^{(d)} = c + \beta^{(d)} RV_t^{(d)} + \beta^{(w)} RV_t^{(w)} + \beta^{(m)} RV_t^{(m)} + \varepsilon_{t+1}, \tag{5.8}$$

where, for t in days, $RV_t^{(d)}$ is the usual ex-post daily realized variance estimator (5.3), and $RV_t^{(w)}$, $RV_t^{(m)}$ are equal-weighted averages of daily realized variance over the past week and month, respectively,

$$RV_t^{(w)} = \frac{1}{5}\left(RV_t^{(d)} + RV_{t-1}^{(d)} + \cdots + RV_{t-4}^{(d)}\right), \tag{5.9}$$

$$RV_t^{(m)} = \frac{1}{22}\left(RV_t^{(d)} + RV_{t-1}^{(d)} + \cdots + RV_{t-21}^{(d)}\right). \tag{5.10}$$

We can loosely view (5.8) as a time series counterpart to a three factor stochastic volatility model where the factors are the past realized variances viewed at different frequencies.

Equation (5.8) has a simple autoregressive structure in realized volatility with the additional feature of considering volatilities over different horizons, this observations leads Corsi to name this model heterogeneous autoregressive of realized volatility, HAR(3)-RV model, which we simply refer to as HAR. Another nice feature of (5.8) is the simplicity of the implementation method. A standard linear regression, fit with the method of ordinary least squares, will suffice to estimate a volatility forecasting model.

If we assume log-prices move as a continuous time semimartingale, then the theoretical volatility over the course of the day is the quadratic variation of the semimartingale (5.5), coined integrated variance, which is well approximated by (5.3), as seen by a corresponding limit theorem (5.4). Corsi empirically showed we can use an autoregressive model depending on different horizons of realized variance (5.8) to approximate (5.3), both as an ex-post estimator and an ex-ante forecast. Due to the simplicity and forecasting success of the HAR model, we choose to use (5.8) as our baseline for forecasts of daily realized variances [10].

[10] We consider other specifications of the HAR model, such as using returns squared as the dependent variable, different frequencies of model re-estimation, different horizons for the predicted realized variances, as well as some nonlinear functions of the Corsi regressors, with little to worsening differences in performance.

Viewing (5.8) more generally as a factor model, one can speculate as to whether it would be of further use to include additional factors such as various measures of jumps, overnight or holiday volatility indicators, factors for news announcements, implied volatility estimates from option prices or volatility swap contracts, etc. In addition to adding more regressors, we could also try to incorporate some nonlinear effects by changing the underlying linear model to a nonlinear one. We begin to tackle some of these questions in the following sections.

5.4.2 High-Dimensional Regularized Models of Realized Variance

We know when measuring volatility all estimators are noisy. The realized variance of a stock's daily return, $RV_{i,t}$, as demonstrated by Corsi [36], is affected by its own previous realized variance $RV_{i,t-1}$ at different horizons. Additionally, as first described by Black [12], volatility is related to returns, $R_{i,t-1}$. Black documents the tendency of variances to go up after the market return goes down; specifically, volatility responds to a negative return much greater than a positive return of the same magnitude. Black postulated this effect was caused from a drop in a stock's price, which changes the capital structure of the firm by increasing the debt-to-equity ratio since a falling price decreases equity value. This explanation was coined as the leverage effect. This has been modeled in continuous time by Carr and Wu [27], and empirically by Lo and Hasanhodzic [49], where they showed the leverage effect is not actually due to firms' leverage. Nevertheless, the leverage effect has come to be known as the asymmetry of magnitudes in upward and downward movements of returns, which is commonly observed in equity markets. We hypothesize firm-level realized variance is also related to the previous realized variance and returns of other related stocks in the cross-section. How these stocks in the cross-section become related could be through a wide variety of mechanisms, or common factors, some more macroeconomic, others more micro-structure related. We hope to capture some of the related stocks whose past volatility and returns have some influence over $RV_{i,t}$, that is we hope to capture *sparse-signals* in the cross section of realized volatility. This is motivated by Chinco et al., [32] who study the sparse-signals in

the cross-section of stock returns. Additionally by adding more variables and information we hope to reduce noise and increase bias in a favorable direction. By exploiting the trade-off between bias and variance we hope to reduce forecast error. With the above observations in mind, we model realized variance using the cross-section of returns and realized variances at different horizons, which forces the forecasting problem into a high-dimensional framework.

In high-dimensional settings, where the number of predictor variables isn't small relative to the number of observations, or when there are more predictors relative to the number of observations, predictions based on ordinary least squares estimates are often unreliable. When the number of covariates is large relative to the number of observations, ordinary least squares overfits the data by tipping the parameters $\hat{\beta}^{OLS}$ to fit noise rather than true signal. In the case where the number of input variables is strictly larger than the number of observations, there are an infinite number of of solutions to $\hat{\beta}^{OLS}$, as ordinary least squares estimates are not unique. In particular, there are an infinite number of solutions that fit the training data perfectly, but much of this fit is from fitting the $\varepsilon_{i,t}$ in the observation $y_{i,t}$ which we're modeling as $y_{i,t} = f(\mathbf{X}_{i,t-1}; \beta_{t-1}) + \varepsilon_{i,t}$. One approach to help reduce the problem of a model fitting the noise in an observation, especially in a high-dimensional setting, is to *regularize* a model. For our application, regularization loosely means constraining the model's parameters in a particular way. Regularization can reduce variance, but at the cost of biasing the forecast. That is, a forecast which fits the training data without much regularization can lead to a forecast with high estimation error and low bias. Increasing the *amount* of regularization, as controlled by a hyperparameter, or the *type* of regularization, as controlled by the regularization function, can lead to a forecast with lower estimation error and a higher bias.

To model realized variance in this high-dimensional setting, we use the least absolute shrinkage and selection operator (LASSO) of Tibshirani [89] to simultaneously select and shrink regression coefficient estimates, which is used to prevent overfitting. The LASSO

estimates parameters by minimizing the objective function

$$\hat{\beta}_{\lambda}^{LASSO} = \underset{\beta}{\operatorname{argmin}} \Big((\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda ||\beta||_1 \Big). \tag{5.11}$$

The LASSO parameter estimates $\hat{\beta}_{\lambda}^{LASSO}$ are non-linear in the dependent variable, unlike the ordinary least squares parameter estimates, and there is no closed-form solution. To compute $\hat{\beta}_{\lambda}^{LASSO}$ requires solving a quadratic programming problem, for which there are a number of efficient algorithms to solve 5.11, as described in [50, Sections 3.4.4 and 3.8]. The L^1 penalty function in 5.11 induces shrinkage of coefficients towards zero, as well as *sparse* coefficient estimates. That is, only a small number of the coefficients in $\hat{\beta}_{\lambda}^{LASSO}$ will be nonzero[11].

For every stock in S&P 500, from 1993 to 2019, we use all cross-sectional stocks' lagged daily, weekly, and monthly realized variance and return estimates[12], measured at at $3:55$pm ET, as predictor variables to forecast the next day's firm-specific realized variance

$$\begin{aligned} \widehat{RV}_{i,t+1} = f\Big(& \{RV_{j,t}^{3:55,(d)}\}_{j=1}^{500}, \ \{R_{j,t}^{3:55,(d)}\}_{j=1}^{500}, \\ & \{RV_{j,t}^{3:55,(w)}\}_{j=1}^{500}, \ \{R_{j,t}^{3:55,(w)}\}_{j=1}^{500}, \\ & \{RV_{j,t}^{3:55,(m)}\}_{j=1}^{500}, \ \{R_{j,t}^{3:55,(m)}\}_{j=1}^{500}; \quad \hat{\beta}_t \Big). \end{aligned} \tag{5.12}$$

As discussed in section 5.3, the training sample is based on W-many observations from a window of time, $[t-W, t-1]$. If our forecasting model, $f(\mathbf{x}; \beta)$, is a penalized linear model such as LASSO, Ridge, or Elastic-Net regressions, also called *shrinkage* models, we need a methodology for how to best choose the hyperparameters in the context of a time series forecast. In our setting, λ in (5.11) and the window parameter W are considered hyperparameters.

One data-driven methodology to choose hyperparameters is *cross-validation*. In any

[11] We should note the LASSO can run into problems with correlated variables.

[12] Computed in the same fashion as (5.9) and (5.10).

prediction or forecasting problem, one fits a model to a particular data set, that is, estimates a model's parameters to make the output of the model optimally close to the data set. Classical statistics often emphasizes the *in-sample error* which is a theoretical quantity commonly formulated as $MSE(\hat{y})$, $RMSE(\hat{y})$, or $MAD(\hat{y})$, which one can estimate empirically. Another theoretical quantity which measures a model's performance on unseen data, is know as the *out-of-sample (generalization) error*. Out-of-sample error can be estimated with the technique of cross-validation

$$\varepsilon_{OOS} := \mathbb{E}\Big[error(\hat{Y}, Y)\Big] \approx \hat{\varepsilon}_{OOS}^{CV}. \tag{5.13}$$

There exists many variants to cross-validation that can reduce the variation in the out-of-sample error estimate (5.13), including a *resampling* methodology coined $K-$fold cross-validation. The method of cross-validation can also be used to estimate the theoretically optimal set of model hyperparameters. We use the method of cross-validation to simultaneously select the best set of hyperparameters as well as estimate the out-of-sample error; this double cross-validation is referred to as *nested cross-validation*. Specifically, a particular out-of-sample error estimate depends on the choice of model hyperparameters,

$$\hat{\varepsilon}_{OOS}^{CV}(\lambda) \approx \mathbb{E}\Big[error(\hat{Y}^{\lambda}, Y)\Big], \tag{5.14}$$

and one wants to choose the set of hyperparameters that that minimizes the out-of-sample error estimate

$$\hat{\lambda}^{*,CV} = \underset{\lambda}{\operatorname{argmin}}\, \hat{\varepsilon}_{OOS}^{CV}(\lambda) \approx \lambda^{*}, \tag{5.15}$$

computed with, say, $K-$fold cross-validation, where

$$\lambda^{*} := \arg\min_{\lambda} \mathbb{E}\Big[error(\hat{Y}^{\lambda}, Y)\Big]. \tag{5.16}$$

For every day t, and for every stock i in the sample[13], using $W = 250$ days of observations of the predictor variables, we estimate our forecast using data from $t - W$ to $t - 1$, using a data set of the form

$$\begin{Bmatrix} \underbrace{(\mathbf{x}_{t-W}, y_{i,t-W+1}), \ldots, (\mathbf{x}_{t-W+j}, y_{i,t-W+1+j})}_{\text{estimation (training) set}}, \\ \underbrace{(\mathbf{x}_{t-W+j+1}, y_{i,t-W+1+j+1}), \ldots, (\mathbf{x}_{t-2}, y_{i,t-1})}_{\text{validation set}}, (\mathbf{x}_{t-1}, y_{i,t}), \\ \underbrace{(\mathbf{x}_t, y_{i,t+1})}_{\text{OOS (test) set}} \end{Bmatrix} \tag{5.17}$$

for some $j \leq t - 2$. The out-of-sample set is used for evaluating forecasts for the estimated model, and the validation set is for evaluating forecasts for different model configurations, that is, different hyperparameters; finally the estimation set is used to estimate the model parameters for a given hyperparameter. For more on nested cross-validation, see [50, Section 7.10], and for a general outline of a machine learning methodology see appendix A.

It is worth noting that in regularized models, there is an issue of how to track structural changes in the parameters of the forecast model, and also how to estimate and adapt the hyperparameters over time. There isn't a good reason to fix the hyperparameters over time which necessitates a methodology for a data-driven method to estimate how the hyperparameters changes over time. Simply re-estimating the penalty hyperparameters every period in a rolling windows walk-forward out-of-sample validation may be computationally too expensive. Concerns of nonstationary data, that is, structural change, raise questions about the suitability of cross-validation methods for both model parameter and hyperparmeter estimation. Implementations of $K-$fold cross-validation assume the temporal position of the validation fold relative to the training folds is irrelevant. Validation folds can be drawn from the data that is older than some of the training data, which isn't a problem with stationary data. When data is nonstationary, it isn't clear that $K-$fold cross-validation is the best

[13] There are 1078 unique firms in our sample starting from 1993 and ending in 2019.

methodology for estimating parameters, because the direction of time can matter. So far, there has been little research in tackling this question and is an important open question, which we leave for future work.

We should mention Carr, Wu, and Zheng [30] use ridge regression, random forests, and feed-forward neural networks to predict realized variance of the S&P 500 index (SPX). They find that machine learning methods have marginal improvement when predicting the realized variance of SPX using option prices as predictor variables. Instead, they find greater improvements when predicting a risk premium, that is, the difference between the realized variance of SPX and a VIX-styled volatility index, which they synthesize in the paper.

There are a few other variants of the high-dimensional regularized models of realized variance that we investigate, but nothing seems to deviate much from the LASSO specification[14].

Finally, as mentioned in [68, Section 2.4], regularization can be interpreted from a Bayesian perspective. An open research agenda, as laid out in Nagel [68], is to look for links between regularization and economic restrictions, which requires the Bayesian interpretation of regularization.

5.4.3 Low-Dimensional Statistical Factor Models of Realized Variance

In general, consider observing the time-series of a measurable quantity on a collection of entitites, such as test assets.If the test assets' time-series tend to jointly move together, then this suggests there may be a common factor structure that is driving this common movement or *synchronization* of the time-series. By a common factor structure, we mean a set of variables, independent of the test assets, of which the measurable quantity is a function. A big question is how to extract the factors from a sample. The difference in the

[14]We don't change the left or right-hand side regression variables, but we do change the functional form, f, of the model to ridge regression and the elastic-net. To reduce computational time, we also only retrain our model every twenty trading days, so approximately monthly.

size of the common movements across the test assets is due to sensitivity, also known as asset exposure or loadings, of the test assets to the common factors. Typically, we assume these exposures are constant over time, but they needen't be[15].

Figure 5.1 is suggestive of a common factor structure in firm-level realized variances. This requires us to consider how to best extract factors that explain the common variation in firm-level realized varainces. It also begs the economic question of what causes a firm to have a high level of return volatility on a given day, and is this reason something common to all firms? Recall, return volatility for an average firm appears to go through regimes where large changes are common, and other regimes where small changes are common. Furthermore, when changes in returns are large, they tend to stay large, with similar behavior for for small changes in returns. Such properties are called time-varying volatility and volatility clustering, respectively, and are displayed in 5.1. There exist many models, typically of the (G)ARCH-type flavor, which try to explain these stylized facts through explicit time-series modeling, but a more fundamental question is to explain these stylized facts using economic rationale. Schwert [77] looks at possible sources causing time-varying volatility in the S&P 500 index. He finds little evidence, at the monthly level, that volatility in economic fundamentals such as bond returns, inflation, short-term interest rates, growth in industrial production, and monetary growth, has any observable influence on stock market return volatility. In Schwert's study, much of the movement in stock market return volatility is not explained by the economic variables he examined.

In a first-effort attempt to see if there exists factors that can explain the common time-variation in firm-level volatilities, we use the linear method of principal components to extract common factors. In many situations, we have a large number of predictor variables that tend to be highly correlated. Commonly, we want to find small numbers of linear combinations of the original predictor variables which are used in place of the predictor variables. Therefore,

[15]Often we would like to determine if the test assets' loadings are time-varying, and then how to model this time-variation. A simple estimation scheme is to sequentially estimate the loadings so we now have a time-series of loadings and repeat the process for identifying factor structures in commonly varying time-series. If there exists a factor structure in the loadings, we can model the loadings as $\beta_{i,j,t} \approx a_{i,j} + b_{i,j} Z_t$.

we seek to find variables that are compressed summaries of the data which captures its *essential* properties. In principal components analysis, a set of variables is approximated with a small number of *underlying factors* that capture a large amount of the common variation among the variables. To construct the principal components used in our regressions, we use the eigen-decomposition of the firm-level realized variance matrix,

$$V = Q\Lambda Q^T,$$

where $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_N)$ is the diagonal matrix of eigenvalues, ordered in decreasing magnitude, and Q is the matrix of eigenvectors of the firm-level realized variance matrix V. More precisely, again, due to the large degree of common movement of firms' total variance over time, which is suggestive of a common factor structure in firm specific total variance, we believe

$$\sigma^2_{i,t} = a_i + \sum_{j=1}^{K} b_{i,j} F_{j,t} + \varepsilon_{i,t}, \quad \forall t, \forall i, \tag{5.18}$$

for some set of common factors $\{F_j\}_{1 \le j \le K}$. Believing firm-level realized variance is a linear function of contemporaneous factors at each point in time, allows us to express (5.18) in cross-sectional form as

$$\mathbf{V}_{N\times 1}(t) = \mathbf{a}_{N\times 1} + \mathbf{B}_{N\times K}\mathbf{F}_{K\times 1}(t) + \varepsilon_{N\times 1}(t), \quad \forall t, \tag{5.19}$$

where the factors are the variables that best explain the historical variability of the cross-section of realized volatilities, $\mathbf{V}$. We can then use ordinary least squares to estimate $\mathbf{a}$, $\mathbf{B}$, and ε. We could consider stacking the above vector equation in time, so we have a $T \times N$ matrix of realized variances; it is useful to stack and think in terms of matrices and linear algebra when considering projections and rotations of data, but this isn't always the best way to think of these models in a rolling time-series regression.

Our goal is to predict $\mathbb{E}_t[\mathbf{V(t+1)}]$. If we assume firm-level realized variances are linear

combinations of common factors as in (5.19) and take conditional expectations, we get

$$\widehat{\mathbf{V}}(t+1) := \underbrace{\mathbb{E}_t\left[\mathbf{V}(t+1)\right]}_{\substack{\text{forecast for}\\ \text{firm-level}\\ \text{realized variances}}} = \mathbf{a} + \mathbf{B}\underbrace{\mathbb{E}_t\left[\mathbf{F}(t+1)\right]}_{\substack{\text{forecast for}\\ \text{factor values}}} + \underbrace{\mathbb{E}_t\left[\varepsilon(t+1)\right]}_{\substack{\text{forecast for}\\ \text{residual, which isn't}\\ \text{necessarily zero}}}, \tag{5.20}$$

which necessitates the need to forecast the conditional expectation of future factor levels and residuals if we hope to forecast the firm-level realized variances. We forecast the conditional expected future factor levels as the linear combination of average lagged factor levels over daily, weekly, and monthly horizons

$$\hat{\mathbf{F}}(t+1) := \mathbb{E}_t\left[\mathbf{F}(t+1)\right] = \tilde{\mathbf{a}} + \tilde{\mathbf{b}}_1\mathbf{F}(t) + \tilde{\mathbf{b}}_2\bar{\mathbf{F}}(t-5,t) + \tilde{\mathbf{b}}_3\bar{\mathbf{F}}(t-22,t)^{16} \tag{5.21}$$

Similarly, we can model the noise as

$$\varepsilon(t+1) = \phi_0 + \phi_1\varepsilon(t) + \phi_2\bar{\varepsilon}(t-5,t) + \phi_3\bar{\varepsilon}(t-22,t) + \eta(t+1),$$

which implies a forecast of the conditional expected future residual of

$$\hat{\varepsilon}(t+1) = \mathbb{E}_t\left[\varepsilon(t+1)\right] = \phi_0 + \phi_1\varepsilon(t) + \phi_2\bar{\varepsilon}(t-5,t) + \phi_3\bar{\varepsilon}(t-22,t). \tag{5.22}$$

Plugging in (5.21) and (5.22) into (5.20) gives us a forecast for firm-level realized variances.

We do this factor extraction and forecasting process on a rolling daily basis. Every day, we extract the factors from the cross-section of firm-level realized variances over the previous year. These factors are driving the variability of firm-level realized variance over that year; we use these factors to predict the next-day realized variance of all firms in the cross section.

[16] If we believed factors are particularly persistent, we could forecast the factors as the lagged values $\hat{\mathbf{F}}(t+1) = \mathbf{F}(t)$.

Our main model of interest uses this factor extraction methodology and adds the the firm-specific predictive variables from the HAR model in section 5.4.1,

$$\hat{\sigma}^2_{i,t+1} = a_i + \sum_{j=1}^{K} b_{i,j}\hat{F}_{j,t+1} + \beta_i^{(d)} RV_{i,t}^{(d)} + \beta_i^{(w)} RV_{i,t}^{(w)} + \beta_i^{(m)} RV_{i,t}^{(m)} + \hat{\varepsilon}_{i,t+1}, \quad \forall t, \forall i. \tag{5.23}$$

This variable combination of common factors and firm-specific factors[17] turns out to yield positive forecasting results, as we document in table 5.2.

There are a few other variants of this methodology that we try, and again, nothing seems to differ too much from this simpler specification [18].

5.4.4 Ensemble Models of Realized Variance

The quality of volatility forecasts depend crucially on a well specified model and the use of *informative* data. An example of a combining models is [18] where they find a two-factor range-based EGARCH model dominate the extensive set of models and data combinations that they consider, both in-sample and out-of-sample. They combine EGARCH and multi-factor volatility models, while using range-based volatility estimators. They attribute their results to a less misspecified volatility model and a more informative volatility proxy. This indicates volatility forecasts might be improved with other ensembling techniques.

Diebold and Shin [39] study regularization methods for forecast ensembling. Their methodology, which we replicate in our study, is coined the *egalitarian LASSO* and the *partial-egalitarian LASSO* for forecast combination. The *egalitarian LASSO* is used for selection and shrinkage towards equal weights. By changing the LASSO penalized estimation problem, we can change the shrinking of weights to zero towards a shrinking of the deviations from equal weights towards zero. eLASSO shrinks in the right direction, however eLASSO selects in the

[17] This is referred to as model nesting. Specifically the model we consider in (5.23) nests the HAR model and also nests a principal components factor model.

[18] We consider different numbers of factors, we consider a different factor extraction methodology that accounts for heteroskedasticity in residuals as outlined in Jones [54], and models that don't include a forecast of residuals, as well.

wrong direction. This means that eLASSO tends to not discard forecasters, because it selects and shrinks towards equal weights, rather than zero weights. eLASSO implicitly presumes that all forecasters belong in the set to be combined [19]. The *partially-egalitarian LASSO* is a modified version of the eLASSO such that some forecasts are potentially discarded, and the survivors are selected and shrunken toward equality. The two step implementation, advocated for in [39, Section 2.6.2], is first to apply the ordinary LASSO to the set of forecasters to potentially discard some, and second apply eLASSO to the remaining set of forecasters to shrink their weights towards equality.

These procedures can be computationally expensive due to the hyperparamater estimation using cross-validation, and as pointed out in [39], most of these effectively act as equal weighted averages of all forecasts, or a subset of top-performing forecasts. This insight leads us to consider simple, equal-weighted averages of forecasts from all the previous sections, which has striking success as documented in table 5.2.

5.5 Data Description

Herein we describe the construction of our data set for forecasting firm-level realized variances[20]. The main variables we construct are daily first and second moments of returns for all firms in the S&P 500 index from January 1993 to June 2019. To compute the variance of returns, we use the realized variance estimator[21], which is computed as the sum of all squared 5-minute returns within a single day. To compute the intraday returns on individual firms, we use the NYSE Trades and Quotes (TAQ) database. For every 5-minute

[19] eLASSO can be implemented with a straightforward adaptation of standard LASSO methodology. In particular in Appendix A of Diebold and Shin, we can see to get the eLASSO coefficients, all one needs to do is run the standard LASSO regression on

$$\tilde{y}_t := y_t - \frac{1}{K}\sum_{i=1}^{K} \hat{y}_{i,t} \ .$$

[20] All credit for the data set construction goes to Christopher Jones.

[21] The implementation of the realized variance estimator was motivated by the findings of [60].

interval between 9 : 30am and 4pm ET, we use the last recorded transaction price to be the closing price for that interval. For every firm in the data set, and for every day, there are a total of 84 price observations.

Given these 84 price observations for a single firm on a particular day, we use them to construct 5-minute returns, but before we do that, we have to filter the TAQ observations. In particular, we exclude observations with a price or size of zero, corrected orders, and trades with condition codes B, G, J, K, L, O, T, W, or Z. We include trades on all exchanges, and we also eliminate observations that result in transaction-to-transaction return reversals of 25% or more, as well as observations that are outside the CRSP daily high/low range. Lastly, we compute size-weighted median prices for all transactions with an identical time stamp. For more on the construction and computing of realized variance using high-frequency data, see [6].

5.6 Results

We conduct a large-scale empirical investigation of realized variance forecasting on a large panel of firms, which to the best of our knowledge, hasn't been done with this magnitude before. We compare the different forecasting models, from section 5.4, with some standard benchmark models such as rolling standard deviation of returns. We find it is hard to unambiguously and unanimously beat the HAR benchmark model when analyzed through the lens of forecast errors.

To start, we compute the first four moments of daily stock returns and realized variances. Using our large panel of data, we compute the moments for an average firm, and for an average cross-section of firms. The moments for an average firm were estimated by first computing the moments for a single firm using its time-series, and then averaging those firm-specific moments across all stocks in the sample. On the other hand, moments for an average cross-section were estimated by first computing the moments for a cross-section of

stocks, and then averaging those cross-section moments over all time periods in the sample[22].

We can think about results corresponding to an average firm and an average cross-section of firms as follows. When one seeks to understand a phenomenon for a general entity, such as individual stocks, bonds, etc., one should use a cross-section average of time-series statistics which will summarize the average entity-specific distribution of the relevant variables,

$$\frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_i(X_{i,1},\ldots,X_{i,T_i}). \tag{5.24}$$

Similarly, when one seeks to understand a cross-sectional phenomenon of entities, one should use a time series average of cross-section statistics that summarize the average cross-section distribution of the relevant variables,

$$\frac{1}{T}\sum_{t=1}^{T}\hat{\theta}_t(X_{1,t},\ldots,X_{N_t,t}). \tag{5.25}$$

In a particular cross-section at time t, there are N_t-many firms. One may be concerned with the cross-sectional distribution, for reasons such as forming a portfolio of cross-section stocks.

In this study, the relevant variables are firm-level realized variance and returns of all S&P 500 stocks from 1/1993 to 6/2019. We see in table 5.1 the kurtosis of the realized variance of the firms in both the average cross-section and the average firm is very large. This tells us we should expect, on an average day, to see firms with very large realized variances relative to other firms in the cross-section. That is, there are firms that dominate the cross-section in terms of their realized variance. The kurtosis for the average firm is also very large, which tells us we should expect an average firm in the S&P 500 index to have large realized variance outliers over the course of its time-series.

Table 5.1 says there are days where particular firms display very large spikes in realized

[22]We should note the firms in the cross-section are changing over time, as our sample is an unbalanced panel.

Table 5.1: Daily summary statistics of the first and second moments of stock returns. That is, we look at the first four moments of stock returns and realized variances. Panel A gives summary statistics for an average firm in our sample. Panel B gives summary statistics for an average cross-section of firms. We see realized variance is prone to large outliers both cross-sectionally and firm-specifically.

Variable	Mean	SD	Skew	Kurtosis
Panel A: Average Firm				
Return	0.00028	0.01605	0.136	1.64
Realized Variance	0.00045	0.00072	0.312	71.2
Panel B: Average Cross-Section of Firms				
Return	0.00009	0.02054	0.155	4.41
Realized Variance	0.00057	0.00096	9.68	258

variances relative a firm's own history, as well as relative to other firms in a cross-section.

We move to forecasting realized variance using the variables summarized in table 5.1. To start, we first look at a measure of historical realized variance, which is a moving average of squared full-day returns, which we call rolling standard deviation squared. Next, we test Corsi's HAR model of realized variance, as described in 5.4.1. Such a model has a self-imposed sparsity structure on the predictor variables and is very simple to implement on a rolling basis with three predictors and 250 observations. We take the rolling standard deviation squared and the HAR model to be our benchmarks. Subsequently, we investigate the ability of the class of *naive* penalized high-dimensional regression models to forecast daily firm-level realized variance. We consider these to be naive because we add all lagged realized variances and lagged returns over daily, weekly, and monthly horizons of all firms in the cross-section for the entire estimation sample. Every penalized regression forecast has approximately 6,000 regressors with 250 observations in the estimation sample.

The common time-varying nature of average firm-level realized variances, as shown in 5.1, is suggestive of a common factor structure in daily firm-level realized variances. Accordingly, we use principal components methodology to extract sources of common variation of realized variance across firms. We then use the top three principal components as well as the predictor

variables for the HAR baseline model to form a *nested* model which forecasts next-day realized variance. Finally, motivated by Diebold and Shin [39], we consider simple equal-weighted averages of our forecasting models. This simple version of forecast ensembling is known to perform very well out-of-sample, though not necessarily optimally, as documented in [39, Section 4], yet have the benefit of not having to train additional models or search for optimal hyperparameter configurations.

Table 5.2: Model scoring statistics using four common scoring functions: root-mean-squared-error, mean-absolute-error, QLIKE, and R^2 from MZ-regressions. Panel A reports errors for an average firm, Panel B reports error for an average cross-section, and Panel C reports the pooled error.

Forecast	RMSE	MAE	QLIKE	$R^2_{\text{MZ Reg}}$	$\alpha_{\text{MZ Reg}}$	$\beta_{\text{MZ Reg}}$
Panel A: Average Firm Error						
PCA + Lagged RVs	0.001702	0.001202	3.186	0.4575	-0.000016	0.2686
Average Forecasts	0.001051	0.000688	1.625	0.4866	-0.000015	0.4000
HAR	0.000519	0.000200	0.2591	0.4692	0.000044	0.8763
Rolling SD Squared	0.000633	0.000306	0.6449	0.3229	-0.000050	1.230
LASSO	0.000549	0.000217	0.3448	0.3906	0.000051	0.8763
Panel B: Average Cross-Section of Firms Error						
PCA + Lagged RVs	0.002071	0.001392	3.437	0.2598	0.000112	0.2023
Average Forecasts	0.001276	0.000810	1.835	0.3135	0.000015	0.3752
HAR	0.000728	0.000254	0.3309	0.3146	0.000081	0.8042
Rolling SD Squared	0.000841	0.000371	0.8299	0.0949	0.000212	0.6835
LASSO	0.000730	0.000275	0.4676	0.2804	0.000121	0.7942
Panel C: Average Pooled Error						
PCA + Lagged RVs	0.002563	0.001192	3.179	0.2318	0.000007	0.2555
Average Forecasts	0.001843	0.000682	1.621	0.2378	0.000040	0.3519
HAR	0.001593	0.000198	0.2581	0.1223	0.000264	0.3418
Rollling SD Squared	0.001335	0.000304	0.6449	0.0768	0.000107	0.7175
LASSO	0.001348	0.000214	0.3374	0.1611	0.000184	0.5424

We see in table 5.2 the results are not unanimous across panels and across error measurements. In panel A, the benchmark HAR has the smallest loss function error, for RMSE,

MAE, and QLIKE[23] loss functions, with the LASSO high-dimensional penalized regression coming in a close second. However, the R^2 of the MZ-regression, as described in 5.3.2, is largest for the simple ensemble model, with HAR coming in a close second. In panel B, we see a similar story, namely, the benchmark HAR model minimizes the loss function error with the LASSO model coming in an even tighter second place. Again, as in panel A, the R^2 of the MZ-regression ranks HAR and forecast averaging in the top two, but this time the ranking is reversed with HAR beating forecast averaging. Finally, in panel C, we see the most diverse set of rankings. According to RMSE, the rolling standard deviation squared beats HAR and the LASSO, where this time the LASSO loss is lower than the HAR loss. We see, however, according to MAE and QLIKE, panel C agrees with panel A and B with HAR minimizing loss and LASSO coming in a close second. The nested PCA-HAR model as well as forecast averaging perform the best according the the R^2 of the MZ-regression. As previously noted, the presence of large outliers and noise in the proxy for realized variance likely contributed to the lack of consensus of forecast rankings across all panels and error measurements.

Though rarely beating the HAR benchmark, it is important to notice the LASSO forecast does a very good job for being so naive. This is impressive due to the large number of highly correlated predictors relative to the small number of observations, as well as the naive "throw-everything-in" approach to modeling firm-level volatility. This points to the possibility of greatly improving a volatility forecasting model using more tailored penalized regression techniques. Issues needing to be more deeply considered when using out-of-the-box high-dimensional penalized regressions are how to handle collinearity and scaling of the predictor variables. Moreover, it would be a worthwhile pursuit to consider customized penalty functions which encode economically motivated priors that can help steer the model towards capturing the stylized facts in volatility[24]. We see the deviations in forecast ranking

[23]QLIKE is less sensitive to large observations, relative to RMSE, by more heavily penalizing under-predictions than over-predictions.

[24]Including clustering, leptokurtic distributions, asymmetry and the leverage effect, as well as response to external shock events [41].

when using loss functions, which tend to favor HAR and LASSO, as well as using the R^2 of MZ-regressions, which tend to favor the nested PCA-HAR model and the equal-weighted forecast averaging model.

Though the results are the least unanimous in panel C, we believe panel C's methodology is the most well-suited for aggregating the large panel of forecast errors[25] because to perform our out-of-sample walk-forward validation, as described in section 5.3.1, we re-estimate our forecasts every day, for every firm. In the full panel of data, we model each entry (day and time) individually, and panel C's measurement of errors averages each entry of errors (day and time) equally. Specifically, we do not model a single firm for all time, nor do we model an entire cross-section at each point in time, which corresponds to panel A and B's measurement of errors.

There remains open questions as to how to properly interpret these results. This is no surprise. The ambiguity of out-of-sample volatility forecast error measurement was documented in [71], who favored mean-squared error and QLIKE, as opposed to Meddahi [64] who favored the R^2 of MZ-regressions. To further add to the ambiguity, marginal improvements to out-of-sample statistical measures of forecast error can lead to economically large improvements in portfolio performance, as observed for the first moment of returns in Campbell and Thompson [24], and is observed in our sample as well, as documented in 6.4. We believe these models collectively provide favorable evidence of utilizing techniques such as model ensembling, factor models of firm volatility, as well as high-dimensional penalized regressions to forecast volatility, which breaks the $\mathbb{P}-$measure tradition of classical (G)ARCH-type volatility forecasting which is well-studied in the literature. We have chosen rather simple and off-the-shelf implementations of ensembling, factor, and penalized regression models. This begs the question of investigating these classes of models with implementations more specific-

[25] Pooled forecast error corresponds to

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T_i}\widehat{\varepsilon}_{i,t}$$

where N and T are the total number of firms and days in the entire sample, respectively, and T_i is the number of days in the i^{th} firm's time-series.

ally tailored towards stylized facts of volatility. Further work should be done to enhance the different model classes to exploit their full potential and see how much improvement they can have over traditional methods of volatility modeling and forecasting. One particular question that also warrants further attention is how to make sense of the outliers? That is, as documented in table 5.1, the kurtosis is extremely large and these outliers are important for predicting variance[26]. It remains to be seen if outliers aren't necessarily as important for predicting option returns, as in section 6.4, as they are for predicting volatility. This leads us to a conjecture: an outlier in volatility forecasts isn't an outlier in the spread between realized voaltility and implied volatility, which is the variable used for predicting option returns. That is, when realized volatility is large, it is likely implied volatility is also large, so the sorting variable doesn't have as much kurtosis as volatility itself, which we observe in panel C of table 6.1.

[26]If we miss the outliers, then we have large forecast errors; also note that large R^2s are frequently caused by a single outlier.

Chapter 6

Options

The breadth of volatility research and innovation in the context of mathematical and econometric volatility modeling, as well as the numerous empirical studies, have lead to several stylized facts in volatility, including time-variation and high-persistence[1] in asset returns, which in turn has lead to the creation of volatility as an asset class. There are currently many vehicles available for taking implicit or explicit positions in volatility. Traditionally, volatility-based products had a major use case for hedging volatility exposure in portfolios ($\mathbb{Q}$ use cases), however, important new uses show up in asset allocation where portfolio managers take active positions in volatility to capture mispricing or risk premiums ($\mathbb{P}$ use cases). Treating volatility as an asset class can be done because it is easily tradable through various portfolios of exchange-listed options, [9] and [46], VIX futures and volatility based exchange-traded products (ETPs), or by using other more specialized derivatives [26] including over-the-counter variance or volatility swaps [28], and even correlation swaps [2].

Herein, we take a view on option markets as markets where we can express our view on volatility, where our view is crafted through our forecasts in chapter 5. These markets can express more than a view on volatility, but also a view on quantities related to volatility such

[1] Also known as stochastic volatility and volatility clustering.

[2] For a review of equity volatility markets, see Van Tassel [78].

as volatility of volatility, or co-volatility as shown in [29][3]. As a particular example, one can directly view delta-hedged options returns as being proportional to a volatility risk premium, meaning the returns of someone who trades delta-hedged option positions are rewarded as compensation for bearing volatility risk, as in [9].

We should mention that the terms used in part I, specifically *edge*, are consistent with the terminology used in this chapter. There are many terms to describe the ability to make a profitable trading strategy and they are all associated with different explanations. In particular some common terms that suggest the ability to create a profitable trading strategy are *edge*, *excess expected return*, and *non-zero risk premium*, where a premium refers to having a higher return stream relative to some other return stream. Both edge and excess return imply there is an expectation for profit but don't necessarily provide an explanation as to why such profitability has existed in the past, or why it might hope to persist in the future. This is where the notion of risk premium comes in, it seeks to explain the ability to create a profitable trading strategy by showing that the profits are related to a particular type of risk, which investors dislike. Investors' disdain for this type of risk means they won't hold this risky portfolio unless they are compensated by some required rate-of-return in excess of the risk-free rate, known as the risk premium. This implies the strategy is not an arbitrage, or near-arbitrage, strategy implying the markets are still efficient. Regardless of the terminology, we investigate a hypothetical options trading strategy, and demonstrate, at least as a first approximation, that we can find an edge which is economically significant, and is impacted by our previous methods for volatility forecasting.

[3]Many studies implicitly ask how to trade the implied volatility surface, how to trade statistics associated with realized volatility, and more generally how to trade discrepancies between risk-neutral and statistical distributions, $(f^{\mathbb{Q}}, f^{\mathbb{P}})$, as in [3].

6.1 Option Returns

6.1.1 Previous Research on Option Returns

Traditionally, financial research on options has been in support of the "sell-side" where it is essential to price options and find hedging portfolios that synthesize the option's payoff structure. This follows the vein of Black, Scholes, and Merton in [13] and [65], respectively. Option returns, as concerned from the "buy-side," have, historically, received considerably less attention.

A first theoretical study of option returns, outside the context of pricing and hedging, was in Cox and Rubinstein [38]. In [38, Section 5.5] they show how the equilibrium risk and expected return of an option are related to the risk and expected return of the underlying stock in the context of a binomial pricing model. Additionally, they derive relationships between option returns and the CAPM, and, in particular, they show how to calculate the alpha and beta of an option. Cox and Rubinstein also do this in continuous time [38, Section 5.6]. In [38, Section 7.8], they briefly mention analyzing option risk premium in terms of a factor model. They point to the fact that market participants who are interested in valuing an option in terms of the underlying, or in pursuing riskless arbitrage profits contingent on market prices differing from this computed value, needen't study risks and returns of options. That is, risk-neutral market participants ($\mathbb{Q}-$investors) weren't particularly worried about if an option was a good investment as part of a larger portfolio; such a task was reserved for risk-averse market participants ($\mathbb{P}-$investors). Only later did it become common to include options as part of an investment portfolio, where one needs to consider the risk and return trade-off[4]. Very little empirical research examining these risks, returns, or premiums seem to have been done until Coval and Shumway [37].

As noted in [21, Appendix A], prior to the development of markets on index options,there were a few studies on individual security option returns where they assumed the Black-

[4]We should mention an additional distinction between returns and prices. Returns represent realized gains or losses on trades, as opposed to prices which are not realized.

Scholes model was correct and simulated returns from the model. Once the market for index options was developed in the latter-half of the 1980's, many studies, apart from pricing and hedging of index options, tried to better understand the discrepancies between the risk-neutral and statistical probability measures. In particular, [3] used option prices to characterize the risk-neutral density and compare with the underlying's statistical density, and investigate difference between risk-neutral and statistical moments. Additionally, Jackwerth [51], tried to better understand the implied volatility smile of S&P 500 index options which implies out-of-the-money puts are expensive relative to at-the-money puts, and determine if these puts are actually mispriced. Accordingly, many studies analyzed returns to put selling strategies, most at the monthly frequency.

Coval and Schumway [37] was the earliest paper studying empirical option risks and returns from an asset pricing point-of-view. They analyzed weekly beta-neutral call and put option returns, as well as straddle returns, and found significantly negative returns, which they claimed is evidence toward additional priced risk factors, rather than a mispricing. Though most of the literature on empirical option returns calculates returns at a monthly horizon, Jones [53] considers daily option returns while using a nonlinear factor model and finds deep out-of-the-money put options have statistically significant alphas and simple put-selling strategies have attractive Sharpe ratios.

More recently, there have been several studies focused on individual equity option returns starting with [46] and [25]. They focus on monthly returns from delta-hedged calls and puts, as well as at-the-money straddles, and find these monthly option portfolio returns are predictable due to various measures of volatility such as realized minus implied volatility spreads and idiosyncratic stock volatility, respectively. They postulate their significant option returns come from mispricing or limits to arbitrage.

Most recently, Jones et al. [57] study the time-series and cross-sectional properties of option returns by computing monthly returns on at-the-money straddles on individual equities. They find option returns exhibit momentum, short-run reversal, and seasonality which fur-

ther justifies thinking of the option and volatility markets as its own asset class which exhibit a factor structure with similarities to the equities market.

In this study, we focus on returns of at-the-money equity straddles, at the daily frequency, and evaluate the strength of the realized minus implied volatility spread, as in [46], when we have better forecasts for daily equity volatility.

6.1.2 Premiums for Volatility Risk

From an investor point-of-view, under the $\mathbb{P}-$measure, it is of great importance to understand if volatility risk is good or bad for portfolio returns, and to know how portfolios with exposure to volatility risk should expect to be compensated, or if they should be expected to forfeit returns. One belief is that investors do not like risk, and volatility, as defined by the standard deviation of returns, was one of the first formalized measures of risk. This implies investors do not like volatility which in turn, implies investors *should* be compensated in excess of the risk-free return for bearing such undesired risk. Another belief as to why volatility risk *should* be priced in assets' returns is due to the counter-cyclical dynamics of volatility. It is natural to conjecture that risk-averse investors require compensation for being exposed to volatility shocks and to ask how highly variance risk is priced.

In response to such needs, a large literature spawned on the analysis of volatility, or variance, risk and how it is priced in a number of different markets. Carr and Wu [28] document variance risk premiums in both the over-the-counter swaps market, as well as the equity market. Similar abnormal returns attributed to variance or volatility risk have been documented in the Treasury market, options market, commodities market, foreign-exchange market, interest-rate swaps market, VIX markets, as well as credit markets as summarized in [4].

For some time now, traders and academics have observed a discrepancy between the volatility implied by theoretical option pricing models and statistical measures of historical volatility. This discrepancy can be interpreted as risk-averse investors having a different

view of volatility than risk-neutral investors. Such a disagreement is one way to define a variance risk premium, and many authors have interpreted the spread between the realized volatility and implied volatility as a measure of volatility risk, and any abnormal returns associated with this measure are capturing a premium associated with volatility risk. There exists many studies that document a negative volatility risk premium. An example of such a study is Carr and Wu [28] who find strong evidence of a negative variance risk premium in expected future stock returns, where they measure variance risk by comparing variance swap rates implied from option prices to realized variances. They also find average at-the-money straddle returns, as well as delta-hedged call and put returns, are large and unconditionally negative, which shows a negative price of volatility risk. It is important to mention Carr and Wu [28] show the variance risk premium is itself a return premium to a class of over-the-counter variance and volatility swaps, which they try to explain with commonly used equity risk factors, such as Fama-French factors; the variance risk premium can also be interpreted as a priced risk factor in equity or option markets[5]

Different studies and authors can define variance or volatility risk in different ways, and they can use their defined measure to explain priced risks in various markets, or they can use their defined measure as a return that needs explaining in terms of other risk factors. For the purpose of this study, we consider the volatility risk premium to be the spread between a forecast of future realized volatility and option implied volatility, which provides another forward-looking volatility measure. Implied volatilities are based on the market's forecast of future realized volatilities, extracted from the prices of options on the asset whose volatility we're trying to forecast. One problem with using implied volatility as a forward looking volatility measure is that option prices typically reflect a volatility risk premium due to the fact that options aren't redundant securities, because of limits to arbitrage affecting perfect hedging of volatility risk. This leads to some volatility risk premium being reflected in the computed implied volatilities. It has been documented that the relation between the realized

[5]That is to say, the variance risk premium can be used in the left-hand side and right-hand side of return regressions, which requires it to be explained and to be used for explaining, respectively.

minus implied volatility spread

$$\left(\widehat{RV}_{t+1} - IV_t,\ \mathbb{E}_t\left[R^{\text{stocks}}_{t+1}\right]\right)$$

and future expected stock returns is negative. Additionally, the discrepancy in the realized and implied volatility forecasts has been documented in the cross-section of equity option returns by Goyal and Saretto [46]. They show that the realized minus implied volatility spread

$$\left(\widehat{RV}_{t+1} - IV_t,\ \mathbb{E}_t\left[R^{\text{options}}_{t+1}\right]\right).$$

has a very strong positive cross-sectional relation with future expected returns of at-the-money straddles, as well as delta-hedged calls and puts. Goyal and Saretto [46] attribute this pheomenon above to mispricing in the options market. They claim high implied volatility compared to realized volatility are indications of overpriced options, similarly low implied volatility compared to realized volatility indicate underpriced options.

For a more comprehensive summary of volatility risk in the equity markets see [43, Chapter 34] and volatility risk in options and stock markets, see [11, Chapter 17]. For papers on equilibrium models explaining variance and volatility risk premiums is the S&P 500 index, see works by Bollerslev et al. [17], [15], and [16].

6.1.3 Calculating Option Portfolio Returns

In this section, we digress for the interests of the mathematically inclined reader to add some formalism. We consider options, calls or puts, and denote the price as

$$\mathcal{O}\left(t, S_t, I_t; K, T\right), \tag{6.1}$$

where at time t, the underlying spot price is S_t, with an implied volatility of I_t, a strike of K and an expiration at time T [6].

We will be selecting from this investable universe of options, i.e. choosing options from the (K, T)-plane, based on properties such as time-to-maturity, $\tau = T - t$, moneyness, $\frac{S_t}{K}$, value of implied volatility, I_t [7], as well as possibly other risk measures such as the delta and other Greeks. Upon the appropriate selection of options from the investable (K, T) universe, we will weight the options and add them together to form an option portfolio.

A portfolio is a linear combination of securities. The change in dollar value of a portfolio at time t is defined as

$$V_t^p - V_{t-1}^p := \sum_{i=1}^{N} n_{i,t-1}(S_{i,t} - S_{i,t-1}), \tag{6.2}$$

where $n_{i,t-1}$ is the number of shares of the i^{th} security, set at time $t-1$, and $S_{i,t}$ is the price of the i^{th} security observed at time t, for all N-many securities in the portfolio. We can also call (6.2) the PnL (profit-and-loss) of the portfolio, and it can be interpreted as the return per dollar value invested[8]. The change in portfolio value from time $t-1$ to t comes from the change in security prices since the number of shares remains fixed.

We can convert this to a portfolio rate of return

$$R_t^p = \frac{V_t^p - V_{t-1}^p}{V_{t-1}^p} = \sum_{i=1}^{N} \left(\frac{n_{i,t-1} S_{i,t-1}}{V_{t-1}^p} \right) \left(\frac{S_{i,t} - S_{i,t-1}}{S_{i,t-1}} \right) = \sum_{i=1}^{N} w_{i,t-1} R_{i,t}, \tag{6.3}$$

where we define the portfolio weights and security rate of returns as

$$w_{i,t-1} := \frac{n_{i,t-1} S_{i,t-1}}{V_{t-1}^p},$$
$$R_{i,t} := \frac{S_{i,t} - S_{i,t-1}}{S_{i,t-1}}.$$

[6] We can think of the (K, T)-plane as the space of all options written, at time t, on the underlying S_t. In other words, the (K, T)-plane is the investable universe of options on an underlying at a fixed point in time.

[7] Specifically the spread between implied volatility and the underlying's realized volatility.

[8] A *return* is, most generally, a payoff divided by a unit of some price, where this price needn't be the initial price or value of the portfolio. That is, a *return* can be thought of as the dollar payoff (PnL) of a portfolio per unit of some other variable, typically a price.

Finally, we can define the portfolio returns in excess of the risk free rate

$$R_t^{e,p} := R_t^p - R_t^{rf} = \sum_{i=1}^{N} w_{i,t-1}\left(R_{i,t} - R_t^f\right). \tag{6.4}$$

We now consider how (6.2), (6.3), and (6.4) change in the context of specific portfolios of options. For simplicity, consider a vanilla option, either a call or put, written on an underlying security, say a stock. We denote the price of the option and its underlying at a particular time by $\mathcal{O}_{i,t}^{(K,T)}$, as short-hand for (6.1), and $S_{i,t}$, respectively. Though simple, we highlight the fact that a delta-hedged option is, in fact, a portfolio. This portfolio consists of a long position in an option and a short position in its underlying in the amount delta-many shares. The change in dollar value (6.2) of a delta-hedged portfolio at time t is

$$V_t^{\substack{\text{delta}\\\text{hedged}}} - V_{t-1}^{\substack{\text{delta}\\\text{hedged}}} = \left(\mathcal{O}_{i,t}^{(K,T-1)} - \mathcal{O}_{i,t-1}^{(K,T)}\right) - \Delta_{i,t-1}^{\mathcal{O}}\left(S_{i,t} - S_{i,t-1}\right), \tag{6.5}$$

where the option's delta is defined as

$$\Delta_{i,t-1}^{\mathcal{O}} := \frac{\partial \mathcal{O}_{i,t-1}^{(K,T)}}{\partial S_{i,t-1}}(t-1, S_{i,t-1})$$

and measures the option's price sensitivity to small changes in the underlying spot price. It is important to realize that the although the delta is a function of spot price and time, it is measured at time $t-1$, and thus a constant at time t, not a function. At $t-1$ this portfolio is formed to have zero delta, but at time t, you are effectively holding a different option because the time to maturity has decreased. That is, a $T-$day option at time $t-1$ is now a $(T-1)-$day option at time t with a different delta. This allows us to compute the option portfolio's new delta at time t,

$$\frac{\partial V_t^{\substack{\text{delta}\\\text{hedged}}}}{\partial S_{i,t}} = \frac{\partial\left(\mathcal{O}_{i,t}^{(K,T-1)} - \Delta_{i,t-1}^{\mathcal{O}} S_{i,t}\right)}{\partial S_{i,t}} = \Delta_{i,t}^{\mathcal{O}} - \Delta_{i,t-1}^{\mathcal{O}}.$$

Now that we know the change in value of a delta-hedged portfolio (6.5), we can compute the excess rate of return (6.4) of a delta-hedged option portfolio by

$$R_t^{e,\ \text{delta hedged}} = \frac{\mathcal{O}_{i,t-1}^{(K,T)}}{V_{t-1}^{\text{delta hedged}}} \left(\frac{\mathcal{O}_{i,t}^{(K,T-1)} - \mathcal{O}_{i,t-1}^{(K,T)}}{\mathcal{O}_{i,t-1}^{(K,T)}} - R_{t-1}^{rf} \right) - \frac{\Delta_{i,t-1}^{\mathcal{O}} S_{i,t-1}}{V_{t-1}^{\text{delta hedged}}} \left(\frac{S_{i,t} - S_{i,t-1}}{S_{i,t-1}} - R_{t-1}^{rf} \right),$$

where R_{t-1}^{rf} is the product of the riskless rate of return available at time $t-1$ and the number of calendar days between time $t-1$ and t, since we know interest accrues on a calendar time basis.

Given we now understand how to connect familiar notions of option prices and deltas to portfolios and returns, we can now begin to think about options from an investment point-of-view ($\mathbb{P}$-measure), rather than a pricing and hedging perspective ($\mathbb{Q}$-measure). In this study, we're interested in straddle returns, and in section 6.3 we describe how we compute the straddle returns from our data set. More specifically, we're interested in at-the-money, delta-neutral straddles. For the record, a straddle is a different type of portfolio of options where you go long, buy, a call and put option on the same underlying security with the same strike and expiration date. Such a portfolio has the property, by definition, that exactly one *leg*, i.e. one option in the portfolio, will expire in-the-money as long as the terminal price differs from the spot price at the time of the portfolio formation. Again, this is due to the fact that we're only considering at-the-money straddles, so the strike of the call and put equals the spot price, and if the spot differs from the price at expiry, then the price at expiry differs from the strike, and exactly one of the straddle's legs will be in-the-money.

To construct such straddle portfolios, rather than trading in the underlying to create a delta-neutral portfolio, for each option in the portfolio, we needn't trade in the underlying and we can just rebalance our options portfolio. Just change the weights, rather than buying a call and a put, and then delta-hedging in the underlying, you can just rebalance the portfolio weights. That is, we simply use the weights in the call and put to make the straddle delta-neutral.

Analogous to (6.1), let $C_{i,t}^{(K,T)}$, $P_{i,t}^{(K,T)}$ be the price of a call and put option, respectively, on the i^{th} underlying with spot value of $S_{i,t}$, at time t, with an implied volatility $I_t^C(K,T)$ and $I_t^P(K,T)$, for a fixed strike K and expiry T. Then the value of the straddle portfolio can be written as

$$V_{i,t}^{\text{strad}} = n_{i,t-1}^{C} C_{i,t}^{(K,T)} + n_{i,t-1}^{P} P_{i,t}^{(K,T)}. \tag{6.6}$$

Equivalently, we see the change in portfolio value is due to changes in the assets in the portfolio

$$V_{i,t}^{\text{strad}} - V_{i,t-1}^{\text{strad}} = n_{i,t-1}^{C}\left(C_{i,t}^{(K,T-1)} - C_{i,t-1}^{(K,T)}\right) + n_{i,t-1}^{P}\left(P_{i,t}^{(K,T-1)} - P_{i,t-1}^{(K,T)}\right). \tag{6.7}$$

Dividing by the total value of the portfolio and cleverly multiplying by one, we can express the dollar value of the straddle portfolio in terms of returns

$$R_t^{\text{strad}} = \frac{V_{i,t}^{\text{strad}} - V_{i,t-1}^{\text{strad}}}{V_{i,t-1}^{\text{strad}}} = \left(\frac{n_{i,t-1}^{C} C_{i,t-1}^{(T,K)}}{V_{i,t-1}^{\text{strad}}}\right)\frac{\left(C_{i,t}^{(K,T-1)} - C_{i,t-1}^{(K,T)}\right)}{C_{i,t-1}^{(K,T)}} + \left(\frac{n_{i,t-1}^{P} P_{i,t-1}}{V_{i,t-1}^{\text{strad}}}\right)\frac{\left(P_{i,t}^{(K,T-1)} - P_{i,t-1}^{(K,T)}\right)}{P_{i,t-1}^{(T,K)}}, \tag{6.8}$$

which can be written as

$$R_{i,t}^{\text{strad}} = w_{i,t-1}^{C} R_{i,t}^{C} + w_{i,t-1}^{P} R_{i,t}^{P}. \tag{6.9}$$

The delta of the straddle is the sensitivity of the portfolio with respect to a change in spot

$$\Delta_{i,t}^{\text{strad}} = \frac{\partial V_{i,t}^{\text{strad}}}{\partial S_{i,t}} \approx \frac{V_{i,t}^{\text{strad}} - V_{i,t-1}^{\text{strad}}}{S_{i,t} - S_{i,t-1}}, \tag{6.10}$$

where the numerator is calculated in (6.7). Using (6.6), we can calculate the straddle delta in terms of the delta of the portfolio's call and put, as well as the number of shares in each asset

$$\Delta_{i,t}^{\text{strad}} = n_{i,t-1}^{C}\Delta_{i,t}^{C} + n_{i,t-1}^{P}\Delta_{i,t}^{P}, \tag{6.11}$$

which shows we can control the delta of the straddle by choosing a portfolio of calls and puts

weighted in the right amount. We will now show what the *right amount* is, and because we only care about the weights of the straddle portfolio, we can assume, without loss of generality, we'll always hold one call option which allows us to only have to find the number of puts required to make the straddle portfolio delta-neutral. Assume

$$n^C_{i,t-1} = 1, \text{ then } \Delta^{\text{strad}}_{i,t-1} = \Delta^C_{i,t-1} + n^P_{i,t-1}\Delta^P_{i,t-1} = 0 \tag{6.12}$$

which implies

$$\left(n^C_{i,t-1}, n^P_{i,t-1}\right) = \left(1, -\frac{\Delta^C_{i,t-1}}{\Delta^P_{i,t-1}}\right). \tag{6.13}$$

Given the number of puts and number of calls, we can then determine the weights for a delta-neutral straddle by

$$w^C_{i,t-1} = \left(\frac{n^C_{i,t-1}C_{i,t-1}}{V^{\text{strad}}_{i,t-1}}\right) = \left(\frac{1C_{i,t-1}}{V^{\text{strad}}_{i,t-1}}\right), \tag{6.14}$$

and

$$w^P_{i,t-1} = \left(\frac{n^P_{i,t-1}P_{i,t-1}}{V^{\text{strad}}_{i,t-1}}\right) = \left(\frac{-\frac{\Delta^C_{i,t-1}}{\Delta^P_{i,t-1}}P_{i,t-1}}{V^{\text{strad}}_{i,t-1}}\right). \tag{6.15}$$

Computing the weights is further complicated because there isn't a unique *price* for the calls and puts, but rather a bid-ask spread. Given we know quotes at which people are willing to buy and sell, for particular volumes, we must decide what quote, or function of quotes, we should choose as a proxy for price when computing returns. A common, though not necessarily realistic, estimate of a fill price is to take the midpoint price, that is the transaction price where no bid-ask spread is paid. In other words, one can only observe a price when a transaction occurs; if no transaction has occurred, then we can use the midpoint of the bid and ask quotes as a proxy for a price,

$$\mathcal{O}^{\text{mid}}_{i,t} = \frac{1}{2}\mathcal{O}^{\text{bid}}_{i,t} + \frac{1}{2}\mathcal{O}^{\text{ask}}_{i,t} \tag{6.16}$$

which is an average of the highest bid and the lowest ask price[9].

Finally, given we can now compute the weights for a delta-neutral straddle portfolio for every straddle in the investable universe (K, T), we must narrow down all of the straddles to those straddle portfolios which are at-the-money. Thus we have a mathematical sketch of how to create a data set of delta-neutral, at-the-money straddle portfolio returns, further details are given in 6.3.

6.2 Methodology

In order to evaluate the economic significance of our volatility forecasts from chapter 5, we must compute the spread between our realized volatility forecast and the market's implied volatility forecast. We will use this signal to form cross-section portfolios made up of *daily*, at-the-money, delta-neutral straddles. We do not decompose the investable universe based on time-to-maturity of the available straddles. Our formation of these cross-section portfolios uses the *portfolio-sort* methodology, which is a very simple nonparametric technique for estimating the future conditional expected rate-of-return, conditional on a *sorting* variable. All variables are appropriately lagged to avoid any look-ahead bias, data snooping, and leakage into our portfolio return estimate. Finally, we evaluate the portfolio performance by looking at its time-series and computing the average rate-of-return, both absolute and risk-adjusted.

In empirical studies where a trading strategy is being investigated, the importance of removing all possible leakage of data and look-ahead biases cannot be overly emphasized. One source of look-ahead bias is not ensuring variables are appropriately lagged. Specifically, data needs to be shifted backwards to ensure information used to make a forecast for today could not have been known after yesterday. That is to say all variables used for forecasting must be appropriately adapted to the filtration. When we sort portfolios, it is imperative

[9]Note this could be refined by taking a weighted average of all quotes in the limit order book, where the weights are a function of the volume.

that the sorting variable is adapted to the time period in the *denominator* of returns. If

$$R_t = \frac{P_t}{P_{t-1}} - 1,$$

then the sorting variable

$$X_{t-1} \in \mathcal{F}_{t-1}$$

must be $(t-1)$–measurable because we form portfolios in the previous period and then observe how they perform over the current period. For the purpose of our experiments, we have forecasts for firm-level realized and implied volatilities[10]

$$\widehat{RV}_t,\ IV_{t-1} \in \mathcal{F}_{t-1}^{3:55\text{pm}}$$

which gives us enough time to plug our predictor variables into our trained model and get our forecast for tomorrow's realized-variance, calculate a sorting variable, sort options, and subsequently form portfolios whose returns will be realized over the following period. We use a measure of the spread between realized and implied volatility as our portfolio sorting variable,

$$X_{t-1} := \text{VRP}_{t-1} = f\left(\widehat{RV}_t, I_{t-1}\right) \in \mathcal{F}_{t-1}^{3:55\text{pm}},$$

where $f(x,y) = x - y,\ \frac{x}{y}$, or $\ln(\frac{x}{y})$.

Upon forming the properly lagged sorting variable, we form portfolios from the assets in the investable universe as a way to test the predictive power of the sorting variable on the conditional expected future return[11]. This also yields a hypothetical naive trading strategy.

For each period, $[t_{i-1},\ t_i]$, we consider an investable universe, and weights $w_{i,t-1}$ constructed at the beginning of the period, for simplicity we use equal weights. Next, we select, or

[10] We ensure IV and $\sqrt{RV}$ are in the correct units. Because IV is annualized, we need to de-annualize by dividing by $\sqrt{250}$, to make the units daily rather than yearly, to match the units of realized variance.

[11] In Characteristic-Sorted Portfolios: Estimation and Inference, Cattaneo et al. [31] investigate the commonly used nonparametric technique of portfolio sorts from a statistical perspective, and show it is a nonparametric estimator for conditional expected returns, as well as derive many of its statistical properties.

sort, assets from the investable universe based on their rank ordering, i.e. binning, according to some characteristic, X_{t-1}. It is important to realize this particular classification of assets is based on a rank-ordering from quantiles according to a characteristic variable, and can be further generalized by other machine learning methods used to discriminate based on different characteristics. Given we've sorted the investable universe by rank ordering, we form portfolios by multiplying the returns of the selected groups, typically the highest and lowest quintile or decile, by weights $w_{i,t-1} = \frac{1}{N}$, where N is the number of assets we've selected by binning. This yields a portfolio return for every point in time, which gives us a time-series of portfolio returns and allows us to analyze the portfolio's performance.

When the investable universe is options on the universe of S&P 500 stocks, there are lots of ways to select options, and there are common heuristics researchers use with regards to the strike or moneyness and term-structure of all available options.

6.3 Data Description

For the study of equity option returns, we use the OptionMetrics database[12]. OptionMetrics also provides data on interest rates, individual stocks, and equity indices, and more recently borrowing data to analyze short selling. For the study of equity returns, we use the CRSP database, which is the source of stock prices, returns, trading volume, market capitalization, and adjustments for stock splits[13] The set of all optionable stocks, i.e. stocks upon which options trade, is much smaller than full set of U.S.-based common stocks that comprise a typical CRSP sample. The OptionMetrics database provides end-of-day bid-ask quotes on all options traded on U.S. exchanges, price, implied volatility[14], and Greeks for all

[12] All credit for the data set construction goes to Christopher Jones.

[13] OptionMetrics has a database link-table for the WRDS database, used to merge options and equity data into a single table.

[14] OptionMetrics interpolates implied volatilities to calculate implied volatilities for options that aren't actually on the market with the correct strike or expiry. They then use the interpolated IVs to calculate option prices and Greeks using the binomial option pricing model, because it allows for early exercise which is common of exchange-traded options.

U.S.-listed index and equity options. Our options sample starts in January 1996 and ends in June 2019, which restricts our sample period for the volatility forecasts. We apply various filters to our sample in an effort to remove illiquid securities from the investable universe, while retaining a sufficiently large investable universe to deliver statistically sound results. Many liquidity filters, such as requiring open interest to be nonzero, tend to significantly reduce the sample size, which can cause some issues.

Our sample consists of call and put options on stocks that are members of the S&P 500 index. For every stock, we consider the shortest maturity options that have at least ten trading days until expiration and then impose the following filters, where the necessary specification will be subsequently provided:

- The bid-ask spread must be less than or equal to half of the midpoint.
- The price must be at least 0.1.
- Arbitrage bounds must not be violated.
- The price cannot exhibit a major reversal.
- The quotes must be reasonable.

With the above filters applied, we find the call with a delta closest to 0.5, and finding the matching put. Using this call and put, we form a straddle, and compute returns on this straddle using bid-ask midpoints as proxies for the price.

Specifically, we delete observations that violate simple *arbitrage bounds*. We compare the option's bid or ask to the end-of-day bid or ask on the underlying stock, which is obtained from CRSP. The four different bounds we check are:

1. For calls, the closing option bid is less than or equal to the closing stock ask price, which prevents the possibility of locking in a guaranteed profit by buying the stock and selling the call.

2. Also for calls, we require that the closing option ask is bounded below the value of immediate exercise, which is computed based on the stock's closing bid price.

3. For puts, we require that the closing option bid is less than or equal to the strike price, which prevents short put strategies that are guaranteed to have positive instantaneous profits.

4. Also for puts, we require that the closing option ask is bounded below by the value of immediate exercise, which is computed based on the stock's closing ask price.

Reversals are defined as cases in which option returns, hedged or unhedged, exceed 2000% (or is below -95%) and is followed by a similar return below -95% (or greater than 2000%). Such cases are extrememly rare (less than 0.01%) in the sample, and when they do appear, they seem to be the result of data errors, such as put quotes mistakenly being given for the corresponding call. In these cases, we remove the return on the day of and the day after the apparently incorrect price.

We preserve *reasonable quotes* by deleting what appear to be inaccurate or non-competitive quotes at the start and the end of the holding period. Specifically, we remove data when the bid price is greater than the ask, or the bid-ask spread is greater than the minimum of $10 or the stock's closing price. While this step only eliminates 0.22% of the sample, the excluded observations include some highly unrealistic prices, such as an ask price of $9,999 for a call on a $40 stock, that have the potential to affect some results.

For more on the use of option filters to construct our sample of options, see [46] and [67], and for other subtle remarks regarding construction of option return data sets see [21, Appendix C, D]. We should also mention that options can disappear from the investable universe from one period to the next, in-sample, but it is unlikely.

The construction of the call and put options data set from OptionMetrics allows us to create a subset of data containing S&P 500 at-the-money, delta-neutral straddles. Such a sub-data set includes for every day t, a term structure of implied volatility, for a fixed

moneyness, specifically at-the-money. Thus we have the time-series evolution of the implied volatility term structure of at-the-money straddles,

$$\left\{IV_t^{\text{ATM-straddle}}(\tau)\right\}_t,$$

where for a fixed time t, $IV_t^{\text{ATM-straddle}}(\tau)$ is the term structure as a function of time-to-maturity, τ. This subset of data does not contain the implied volatility smile/skew for fixed maturities, because everything is at-the-money, by construction. We need to compute the implied volatility for at-the-money straddles[15]. Rather than solving for the Black-Merton-Scholes model's implied volatility using binomial trees, we use an approximation for the implied volatilities

$$I_{i,t-1}^{\text{strad}} \approx w_{i,t-1}^{C} I_{i,t-1}^{C} + w_{i,t-1}^{P} I_{i,t-1}^{P},$$

which is the weighted average of the call and put implied volatilities.

Finally, we link the above data description with section 6.1.3. Our data set contains firm identifiers, calendar-time, strike, time-to-maturity, option (call and put) price, delta, implied volatility, and excess-return computed from the midpoint price, as well as the weights in the call and put, and finally the excess-return of the straddle portfolio, written symbolically:

$$\begin{aligned}
&i,\ t,\ K,\ \tau,\\
&C_{i,t-1},\ \Delta_{i,t-1}^{C}, I_{i,t-1}^{C},\ R_{i,t}^{C,\text{ mid, }e},\\
&P_{i,t-1},\ \Delta_{i,t-1}^{P}, I_{i,t-1}^{P},\ R_{i,t}^{P,\text{ mid, }e},\\
&w_C,\ w_P,\\
&\text{and}\ \ R_{i,t}^{\text{strad}} = w_{i,t-1}^{C} R_{i,t}^{C} + w_{i,t-1}^{P} R_{i,t}^{P},
\end{aligned}$$

where the weights are chosen to make the straddle delta-neutral as in (6.14) and (6.15). Note,

[15]Note, this is an instance of the general question of how to compute the implied volatility of a portfolio of options.

all of the lags $(t-1)$ are due to weights being computed at the beginning of the holding period.

6.4 Results

In many machine learning and forecasting applications, the objective for estimating a model's parameters, in-sample, is to minimize the sum of squared forecast errors or functions of it, such as R^2. The corresponding sum of squared, out-of-sample, prediction errors is used to measure the out-of-sample predictive performance. This method of minimizing prediction error is also employed in the context of variance or volatility forecasting. As noted in [68, Chapter 3.2] it is not obvious that this is the right approach to evaluate the power of a forecast. In a trading or asset management setting, the ultimate goal is not to forecast moments of returns, but rather to use forecasts to construct tradable portfolios that earn large out-of-sample expected returns, either absolute or risk-adjusted. Methods that lead to better forecasts in terms of standard loss functions do not necessarily lead to better portfolios in terms of standard portfolio metrics. Similarly, methods that lead to better portfolios may not necessarily be related to better forecasts, and certainly not in the same magnitudes [24].

We first consider summary statistics of the unconditional average excess returns of an equally-weighted at-the-money equity straddle portfolio. As documented in panel A of table 6.1, the average unconditional excess straddle returns are slightly negative, at approximately negative half of a basis point, have positive skewness, and have a high degree of excess kurtosis. In panel B of table 6.1, we see the average implied volatility of at-the-money straddle returns is much larger than realized volatility, and far less skewed. We also notice the average realized volatility coming from the forecasts are significantly lower than the measured realized volatility, with a significant increase in skewness and kurtosis, with the LASSO and HAR models showing the most dramatic increase skewness and kurtosis. Panel C in table 6.1 computes summary statistics for the sorting variable, typically named the

volatility risk premium,

$$\ln\left(\frac{\widehat{RV}}{IV}\right)$$

for each forecast. The sorting variable's average value is positive for the PCA-HAR nested model and the forecast averaging model, and negative for HAR, LASSO, and rolling standard deviation, with a skewness much closer to zero, and kurtosis[16] much lower than the realized variance forecasts.

Table 6.1: Panel A shows the average excess returns of straddles in our sample, which are, unconditionally, slightly negative. The computed sample moments are the time-series average for the moments of the cross-sectional distribution of the excess straddle returns. The mean gives the average return of an equal-weighted portfolio of all the straddles in our sample. Panel B gives the average cross-section of firms' summary statistics, and panel C gives the summary statistics of the average cross-section of firms' sorting variables, dependent upon the forecasting model.

Option Variables	Mean	SD	Skew	Kurtosis
Panel A: Unconditional Excess Returns				
ATM Straddle	-0.00006	0.8336	4.23	50.9
Panel B: Volatility Variables				
ATM Straddle Implied Volatility	0.3274	0.1164	1.59	5.64
Realized Volatility	0.0165	0.0071	2.44	12.9
$\widehat{RV}$ - PCA + Lagged RVs	0.0015	0.0015	4.25	29.4
$\widehat{RV}$ - Average Forecasts	0.0010	0.0010	4.21	28.3
$\widehat{RV}$ - HAR	0.0004	0.0005	6.80	78.4
$\widehat{RV}$ - Rolling SD Squared	0.0004	0.0003	2.50	7.12
$\widehat{RV}$ - LASSO	0.0004	0.0004	6.40	73.2
Panel C: Option Return Conditioning Variable - $\ln(RV/IV)$				
VRP - PCA + Lagged RVs	0.5082	0.2008	-0.28	3.62
VRP - Average Forecasts	0.1410	0.1750	-.0018	4.40
VRP - HAR	-0.1664	0.1961	0.0373	4.91
VRP - Rolling SD Squared	-0.0730	0.1777	-0.2087	5.76
VRP - LASSO	-0.1589	0.2179	-0.1653	5.18

[16] We're using excess kurtosis, meaning a standard normal distribution has an excess kurtosis of zero.

In table 6.2, we find economically significant unannualized average returns ranging from 1% to 2% per day, and unannualized daily Sharpe ratios ranging from 0.57 to 0.80. This confirms the so-called volatility risk premium is large and economically significant in daily equity options returns, which adds to the existing literature of Goyal and Saretto [46] who document such returns at the monthly horizon.

Given the problems outlined in section 5.3.2 with using traditional forecast error measurements, we use the conditional straddle portfolio performance measures, in table 6.2, as a different way to rank volatility forecasts without reference to a particular loss function. We are particularly interested in the high minus low portfolio, that is, the portfolio that longs the highest quintile of stocks and shorts the stocks in the lowest quintile[17]. All portfolios are formed by equally weighting stocks in the portfolio.

Table 6.2: We consider the average excess return to equally-weighted portfolios formed on the spread between a forecast of realized volatility and implied volatility. Panel A gives risk adjusted expected excess returns, and panel B gives the raw or absolute expected excess returns to the sorted straddle portfolios. These numbers are all unannualized at the daily frequency.

Forecast	1	2	3	4	5	5 minus 1
Panel A: Sharpe Ratio of Straddle Portfolios						
PCA + Lagged RVs	-0.3249	-0.1299	-0.0251	0.0940	0.2665	0.8044
Average Forecasts	-0.2971	-0.1192	-0.0242	0.0776	0.2613	0.7876
HAR	-0.1967	-0.0876	-0.0221	0.0567	0.2043	0.6317
Rolling SD Squared	-0.2171	-0.0799	-0.0148	0.0572	0.1902	0.5769
LASSO	-0.1706	-0.0675	-0.0138	0.0469	0.1701	0.5744
Panel B: Average Straddle Portfolio Excess Returns						
PCA + Lagged RVs	-0.0097	-0.0043	-0.0009	0.0035	0.0109	0.0206
Average Forecasts	-0.0089	-0.0041	-0.0009	0.0029	0.0105	0.0195
HAR	-0.0063	-0.0031	-0.0008	0.0021	0.0076	0.0139
Rolling SD Squared	-0.0068	-0.0028	-0.0005	0.0021	0.0075	0.0143
LASSO	-0.0056	-0.0023	-0.0005	0.0017	0.0063	0.0118

[17] We can think of these long-short portfolios as "market neutral" because the portfolio weights will sum to zero.

We observe the forecasts distinguish themselves in economically significant average returns on long-short quintile portfolios, on an absolute and risk adjusted basis. These portfolios are formed on the spread between forecasted realized and implied volatility. The PCA-HAR nested model, from section 5.4.3, separates stocks in the cross-section to form long-short straddle portfolios that outperform all other volatility forecasting models. It beats the benchmark by increasing the Sharpe ratio by 20% as well as improving by average daily return by approximately 0.6% (or 60 basis points). The equally-weighted average forecast ensemble has similar improvements over the benchmark. While the portfolios formed from the LASSO model have lower daily average returns, the model also produces a drop in portfolio return volatility. This is evidenced by the LASSO-based long-short portfolio's Sharpe ratio, which performs similar to the rolling standard-deviation benchmark model.

These results strengthen the case of equity option return predictability at the daily frequency, based on the spread between realized and option implied volatility. There is more work to be done here in making these results more pointed both for understanding forecast evaluation, and for crafting a true trading strategy, in particular understanding the effects of transaction costs and margin requirements as documented in [46].

One thing to note is the moments in the conditional and unconditional straddle portfolios, as in tables 6.3 and 6.1, respectively. Particularly, when compared to the unconditional straddle portfolios, the conditional straddle portfolios display an increase in mean, a decrease in standard deviation, a decrease in skewness, and an especially large decrease kurtosis. This suggests the conditional returns are brought closer a positively biased normal distribution. In table 6.3, we particularly note the increase in means from the baseline models to top-performing models is associated with an increase in skewness and kurtosis.

The above results hint at the need for a heavy-tailed, high-frequency, Kelly criterion-based optimization. We can further see this by looking at the time-series and histogram of the top performing strategy, where we clearly see non-normally distributed returns.

Table 6.3: Summary statistics of the time-series of straddle portfolios formed from different forecasts.

Forecast	Mean	SD	Skew	Kurtosis
High-Low Sorted Straddle Portfolio Returns				
PCA + Lagged RVs	0.0206	0.0256	2.11	18.64
Average Forecasts	0.0195	0.0248	1.75	14.47
HAR	0.0139	0.0221	0.603	8.04
Rolling SD Squared	0.0143	0.0248	0.966	7.35
LASSO	0.0118	0.0206	0.501	6.33

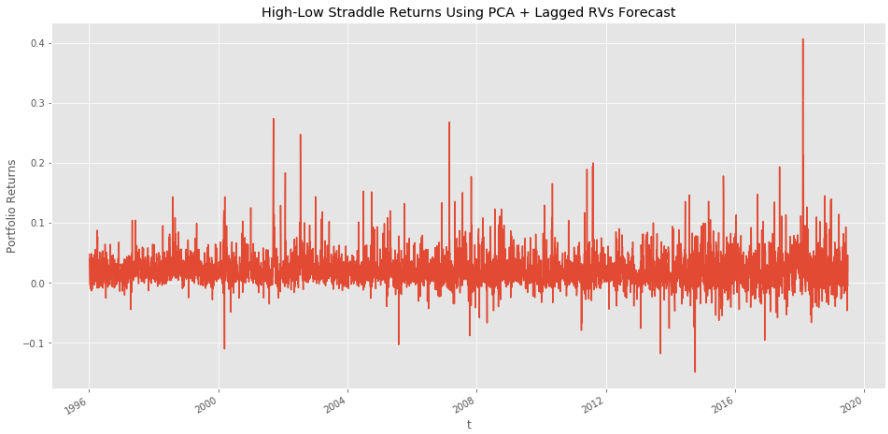

Figure 6.1: Time-series of top performing straddle portfolio.

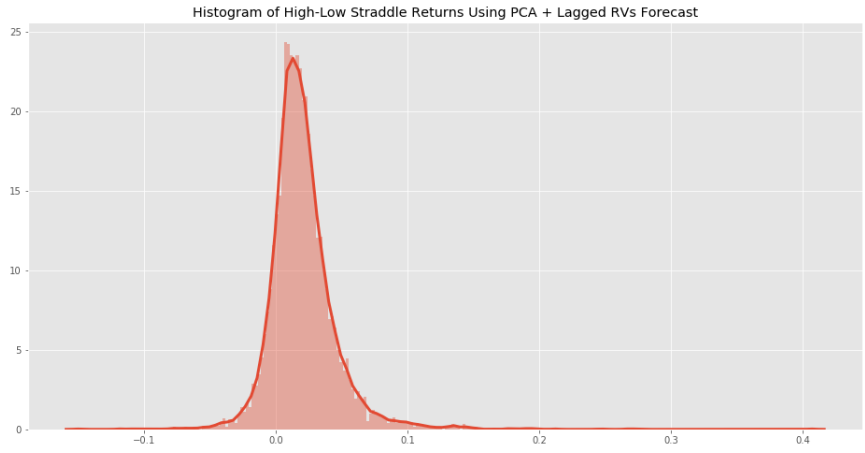

Figure 6.2: Histogram of the time-series of returns of the top performing straddle portfolio.

We then simulate normally distributed returns with the same mean and standard devi-

ation as the top-performing strategy, to see the difference, which is rather stark.

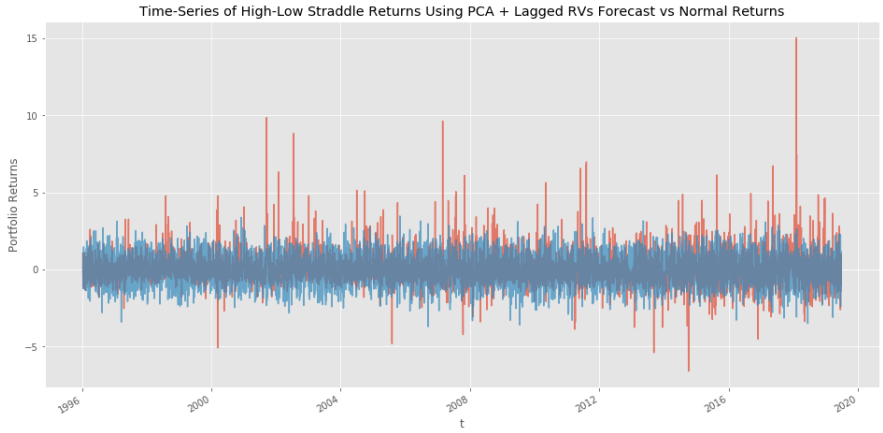

Figure 6.3: Time-series of top performing straddle portfolio returns versus simulated normal returns.

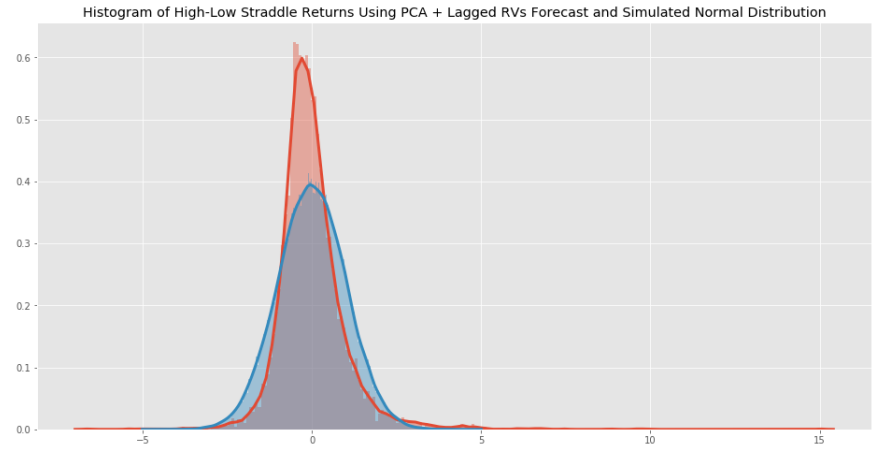

Figure 6.4: Histogram of the time-series of returns of the top performing straddle portfolio versus simulated normal returns.

Additionally, we can plot the quantiles of a theoretical normal distribution versus the quantiles of our top-performing conditional long-short straddle portfolio returns.

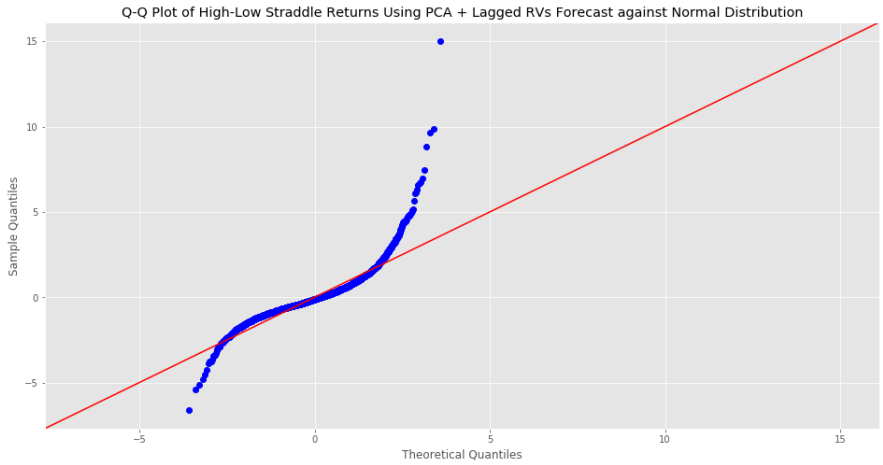

Figure 6.5: Q-Q plot of the quantiles of returns of the top performing straddle portfolio versus theoretical normal quantiles.

The above table 6.3 and figures 6.3, 6.4, 6.5, show the returns to our options portfolio of straddles, formed on the signal of the spread between the realized and implied volatility, has an edge, is positively skewed, and has heavy tails. This shows a Kelly-criterion optimization could be applied to optimize this strategy.

It should be noted that the returns highlighted in table 6.2 are not entirely representative of the actual returns that could be realized by a long-short straddle portfolio. This is due to transaction costs and margin requirements when short selling straddles, as mentioned in [46, Section 5.2]. Further, option returns have a bias, especially at higher frequencies, as shown in Jones, Duarte, and Wang [55]. We conjecture such microstructure biases won't affect our results too much because we're working with at-the-money options, and microstructure biases tend to be larger in options that are away-from-the-money.

A deeper analysis on transaction costs, or how to optimally implement the above edge through transaction cost and order type analysis is certainly warranted. In the literature, transaction costs are typically estimated by estimating the bid-ask spread and using this measure as a proxy for transaction costs, which will be an upward biased estimate of the true cost one would incur while trading. It is important to recognize bid-ask spreads are not constant, and in practice one's goal is to trade when the spread is small, which would require additional modeling of the dynamics of the spread. Also, the use of order types other than a market order, such as a limit order or more complex order types, can help reduce trading costs. And finally, the signal we investigated above, the discrepancy between realized and implied volatility, is not intended to be used as a stand-alone strategy, but as an additional signal meant to be used in a complex trading program combining multiple signals. Though we believe transaction costs must be analyzed, the above reasons mildly critique the literature's method of conducting transaction cost analysis, and suggest a list of questions for further research.

With regards to models, the results in section 5.6 point to the class of high-dimensional penalized regression as having the greatest potential for minimizing forecast errors, and from results in this section 6.4 point to the class of volatility factor models as having the greatest potential to improve option portfolio performance. Further research needs to investigate this discrepancy between the volatility forecast error measurements and the standard portfolio performance metrics. Such discrepancies raise questions regarding the best way to train

machine learning models for forecasting volatility. In particular, should we be using classical error measurements to evaluate the power of an out-of-sample volatility forecast? Also, should we be using the standard maximization of the cross-validated R^2 to tune model hyperparameters? This is left for future research.

Chapter 7

Summary

This thesis is three tiered with the intended goal of performing research at each stage of the investment or trading process.

First, we consider how to optimally bet when we want to maximize the high-frequency short run, or the low-frequency long run, growth rate of a sequence of heavy-tailed bets in the presence of an *edge*. Interpreted financially, we derive the optimal growth rate function for a strategy that holds a risky portfolio, which reflects our edge, and a risk-free asset, which we can assume is cash. Furthermore, we allow for the risky portfolio to be non-normal, that is, allowing for heavy tails and skewness in the return distribution. To make use of this optimization, we require a forecast for the expected return, volatility, and jump intensity (which reflects portfolio jump and skew risk) of the risky portfolio reflecting our edge. Mathematically, we require a forecast of the Lévy triple of the risky portfolio

$$\left(\widehat{\mu}_p^{\mathbb{P}},\ \widehat{\sigma}_p^{\mathbb{P}},\ \widehat{F}_p^{\mathbb{P}}\right)$$

under the statistical $\mathbb{P}$-measure.

Second, we seek to find a variance forecasting methodology which beats the de-facto benchmark of Corsi [36]. We examine high-dimensional penalized regressions, such as LASSO, low-dimensional statistical factor models, as well as simple ensembling methods, and find

we can beat the benchmark, though marginally and not unanimously according to different measures of forecast errors. We find, however, the forecasts, $\widehat{\sigma}^{\mathbb{P}}$, do clearly distinguish themselves in the context of equity option returns, and provide economically significant improvements to the benchmark. One distinction of our investigation is we forecast individual firm-level return variance at the daily frequency for all firms in the S&P 500, over our twenty year sample. To our knowledge, no study of this magnitude has been done in the literature before.

Finally, we investigate a source of *edge* in the equity options market which makes use of our volatility forecasts, $\widehat{\sigma}^{\mathbb{P}}$. Inspired by the work of Goyal and Saretto [46], we consider a high-frequency, daily, at-the-money straddle return strategy which captures discrepancies between our forecasts of volatility and the *market's* forecast of volatility, $\widehat{\sigma}^{\mathbb{Q}}$. Strictly speaking, we consider the log difference of realized volatility and Black-Scholes at-the-money implied volatility as a predictor variable for straddles. This is one specification of the general discrepancy between statistical volatility, $\sigma^{\mathbb{P}}$ and risk-neutral volatility, $\sigma^{\mathbb{Q}}$. To contextualize, we emphasize these portfolios are trying to capture the signal between a measure of the spread between the second moments of the $\mathbb{P}$ and $\mathbb{Q}$ distribution of returns,

$$\ln\left(\frac{\sigma^{\mathbb{P}}}{\sigma^{\mathbb{Q}}}\right).$$

We observe economically significant returns, both absolute and risk-adjusted, that clearly distinguish our above forecasts, favoring the nested factor model and simple ensembling models. Additionally, the time series of the daily straddle returns, conditional on the discrepancy between realized and implied volatility, exhibits not only an *edge*, as characterized by the average returns, but also large skewness and heavy-tails, making the return series amenable to optimization with the above heavy-tailed and high-frequency Kelly criterion.

Part III

Appendix

Appendix A

General Outline for ML Prediction

General Outline for Machine Learning Prediction Model

We are analyzing samples from a population over some time period.

1. Problem formulation and exploration: Must answer what we're trying to predict and with what data, i.e. what are $(\mathbf{X}, Y)$ and what is the relation $Y = g_\gamma(\mathbf{X}; \theta)$. Further, perform some preliminary exploratory data analysis (EDA) with some basic correlations and summary statistics motivating further analysis. That is, analyzing the joint relationship

$$(\mathbf{X}, Y).$$

Data preparation, cleaning, and formatting. This step transforms the data

$$\tilde{\mathbf{X}} = T(\mathbf{X})$$

for some transformation function T which does feature extraction. (which could be a discretization/binning function)

Determine estimation (train) set and out-of-sample (test) set. That is, what is the the

outer-fold cross-validation scheme?

$$\{(X_i, Y_i)\} = \left\{(X_i^{train}, Y_i^{train})\right\} \cup \left\{(X_i^{test}, Y_i^{test})\right\}$$

2. Establish a baseline model to compare against our model.

$$Y = b(X),$$

b for baseline model. This should be simple and intuitive.

Determine an optimization algorithm to do the estimation of the model parameters and hyperparameters, this step requires the specification of a loss function to minimize to find the optimal model parameters as well as the deciding on a method for the inner-fold cross-validation which determines the optimal set of hyperparameters. That is,

(a) We model the outputs as a parametric function of the inputs

$$Y = g_\gamma(\mathbf{X}; \theta),$$

where the model parameters (or coefficients) θ are supposed to capture the effect of the inputs on the outputs, for very complex and nonlinear models these effects can be very difficult to interpret beyond achieving a prediction. The model's hyperparameters γ are meant to capture the structure of the model rather than the effect of the variables, though these are non-trivially linked.

(b) Before we estimate the model parameters, we first have to estimate the model's hyperparameters. To do this we somehow split the training set into a sub-training set and a validation set

$$\left\{(X_i^{train}, Y_i^{train})\right\} = \left\{(X_i^{sub-train}, Y_i^{sub-train})\right\} \cup \left\{(X_i^{validation}, Y_i^{validation})\right\}$$

which makes up the inner-fold of the nested cross-validation scheme.

(c) To estimate the model's hyperparameters we minimze the error over validation set and we estimate the model's parameters for a given hyperparameter configuration over the sub-train data set. First solve

$$\hat{\theta}^*(\gamma) = \arg\min_{\theta} \mathbb{E}_{(\mathbf{X},Y)}\left[\mathcal{L}\left(Y_{sub-train}, g_\gamma(\mathbf{X}_{sub-train};\theta)\right)\right].$$

Next we estimate the error over the validation set and choose the hyperparameter configuration that minimizes the validation error

$$\hat{\gamma}^* = \arg\min_{\gamma} \hat{\varepsilon}^{CV}(\gamma) = \arg\min_{\gamma} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}\left(Y_i^{validation}, g_\gamma(\mathbf{X}_i^{validation};\hat{\theta}^*(\gamma))\right)$$

where we are estimating the expected loss over the validation set using empirical risk minimization and n is the number of observations in the validation set.

3. Now that we have our optimal hyperparameter configuration, we need to estimate the model's parameters by solving

$$\hat{\theta}^*(\gamma^*) = \arg\min_{\gamma} \frac{1}{n_{train}} \sum_{i=1}^{n_{train}} \mathcal{L}\left(Y_i^{train}, g_{\gamma^*}(\mathbf{X}_i^{train};\hat{\theta})\right) \approx \arg\min_{\theta} \mathbb{E}_{(\mathbf{X},Y)}\left[\mathcal{L}(Y_{train}, g_{\gamma^*}(\mathbf{X}_{train};\theta))\right],$$

which we solve for over the training set data.

4. We finally have our estimated/fitted/trained model with an "optimal" set of parameters and hyperparameters

$$\hat{y}(\mathbf{x}) = g_{\hat{\gamma}^*}(\mathbf{x};\hat{\theta}^*),$$

which we use to make predictions/forecasts on the test set

$$\left\{(\mathbf{x}_i^{test}, \hat{y}_i^{test})\right\}.$$

5. Lastly we estimate the out-of-sample (generalization) error which is supposed to tell us how well our model predicts on unseen data

$$\varepsilon_{OOS} \coloneqq \mathbb{E}\Big[error(Y, \hat{Y})\Big] \approx \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} error\left(Y_i^{test}, \hat{y}(\mathbf{X}_i^{test})\right) =: \hat{\varepsilon}_{OOS}^{CV}.$$

Note the statistical methodology of cross-validation can be used to estimate both the out-of-sample (generalization) error, as well as to estimate the optimal hyperparameter configuration.

Also note most classical finance problems deal with panel data, while most classical machine learning problems deal with iid cross-section data. This means we have to be a little more careful when framing these problems.

References

[1] Yacine Aït-Sahalia and Jean Jacod. *High Frequency Financial Econometrics*. First. Princeton University Press, 2014.

[2] Yacine Aït-Sahalia, Chenxu Li, and Chen Xu Li. 'Implied Stochastic Volatility Models'. In: *The Review of Financial Studies* 34.1 (Mar. 2020), pp. 394–450.

[3] Yacine Aït-Sahalia, Yubo Wang, and Francis Yared. 'Do Option Markets Correctly Price the Probabilities of Movement of the Underlying Asset?' In: *Journal of Econometrics* 102.1 (2001), pp. 67–110.

[4] Manuel Ammann and Mathis Moerke. 'Credit Variance Risk Premiums'. In: Working paper (2021).

[5] Torben Andersen and Tim Bollerslev. 'Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts'. In: *International Economic Review* 39.4 (1998), pp. 885–905.

[6] Torben Andersen et al. 'The distribution of realized stock return volatility'. In: *Journal of Financial Economics* 61.1 (2001), pp. 43–76.

[7] Torben Andersen et al. 'Volatility and Correlation Forecasting'. In: *Handbook of Economic Forecasting*. Ed. by G. Elliott, C.W.J. Granger, and Granger Timmermann. Handbook of Economic Forecasting. 2006, pp. 777–878.

[8] Søren Asmussen and Hansjörg Albrecher. *Ruin Probabilities*. Second. Vol. 14. World Scientific Publishing Co., Hackensack, NJ, 2010.

[9] Gurdip Bakshi and Nikunj Kapadia. 'Delta-Hedged Gains and the Negative Market Volatility Risk Premium'. In: *The Review of Financial Studies* 16.2 (June 2015), pp. 527–566.

[10] Roger R. Baldwin et al. 'The Optimum Strategy in Blackjack'. In: *Journal of the American Statistical Association* (1956).

[11] Turan G. Bali, Robert F. Engle, and Scott Murray. *Empirical Asset Pricing: The Cross Section of Stock Returns*. Wiley, 2016.

[12] Fischer Black. 'Studies of Stock Price Volatility Changes'. In: *Proceedings of the Business and Economics Section of the American Statistical Association* (1976), pp. 177–181.

[13] Fischer Black and Myron Scholes. 'The Pricing of Options and Corporate Liabilities'. In: *Journal of Political Economy* 81.3 (1973), pp. 637–654.

[14] Tim Bollerslev. 'Generalized autoregressive conditional heteroskedasticity'. In: *Journal of Econometrics* 31.3 (1986), pp. 307–327.

[15] Tim Bollerslev, Natalia Sizova, and George Tauchen. 'Volatility in Equilibrium: Asymmetries and Dynamic Dependencies'. In: *Review of Finance* 16.1 (Mar. 2011), pp. 31–80.

[16] Tim Bollerslev, George Tauchen, and Hao Zhou. 'Expected Stock Returns and Variance Risk Premia'. In: *The Review of Financial Studies* 22.11 (Feb. 2009), pp. 4463–4492.

[17] Tim Bollerslev et al. 'Risk and return: Long-run relations, fractional cointegration, and return predictability'. In: *Journal of Financial Economics* 108.2 (2013), pp. 409–424.

[18] Michael W. Brandt and Christopher S. Jones. 'Volatility Forecasting With Range-Based EGARCH Models'. In: *Journal of Business & Economic Statistics* 24 (2006), pp. 470–486.

[19] Michael W. Brandt, Pedro Santa-Clara, and Rossen Valkanov. 'Parametric Portfolio Policies: Exploiting Characteristics in the Cross-Section of Equity Returns'. In: *The Review of Financial Studies* (2009).

[20] Leo Breiman. *Probability*. Addison-Wesley Publishing Company, Reading, Mass.-London-Don Mills, Ont., 1968.

[21] Mark Broadie, Mikhail Chernov, and Michael Johannes. 'Understanding Index Option Returns'. In: *The Review of Financial Studies* 22.11 (May 2009), pp. 4493–4529.

[22] Aaron Brown. *Red-Blooded Risk. The Secret History of Wall Street*. Wiley, 2011.

[23] Sid Browne and Ward Whitt. 'Portfolio Choice and the Bayesian Kelly Criterion'. In: *Adv. in Appl. Probab.* 28 (1996).

[24] John Y. Campbell and Samuel B. Thompson. 'Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average?' In: *The Review of Financial Studies* 21.4 (Nov. 2007), pp. 1509–1531.

[25] Jie Cao and Bing Han. 'Cross section of option returns and idiosyncratic stock volatility'. In: *Journal of Financial Economics* 108.1 (2013), pp. 231–249.

[26] Peter Carr and Roger Lee. 'Volatility Derivatives'. In: *Annual Review of Financial Economics* 1.1 (2009), pp. 319–339.

[27] Peter Carr and Liuren Wu. 'Leverage Effect, Volatility Feedback, and Self-Exciting Market Disruptions'. In: *Journal of Financial and Quantitative Analysis* 52.5 (2017), pp. 2119–2156.

[28] Peter Carr and Liuren Wu. 'Variance Risk Premiums'. In: *The Review of Financial Studies* 22.3 (Apr. 2008), pp. 1311–1341.

[29] Peter Carr and Liuren Wu. *Vol, Skew, and Smile Trading*. 2016.

[30] Peter Carr, Liuren Wu, and Zhi-bai Zhang. 'Using Machine Learning to Predict Realized Variance'. In: *Journal of Investment Management* 18 (2020), pp. 1–16.

[31] Matias D. Cattaneo et al. 'Characteristic-Sorted Portfolios: Estimation and Inference'. In: *The Review of Economics and Statistics* 102.3 (July 2020), pp. 531–551.

[32] Alex Chinco, Adam D. Clark-Joseph, and Mao Ye. 'Sparse Signals in the Cross-Section of Returns'. In: *The Journal of Finance* 74.1 (2019), pp. 449–492.

[33] Tarun Chordia, Amit Goyal, and Alessio Saretto. 'Anomalies and False Rejections'. In: *The Review of Financial Studies* 33.5 (Feb. 2020), pp. 2134–2179.

[34] Rama Cont. 'Empirical properties of asset returns: stylized facts and statistical issues'. In: *Quantitative Finance* 1.2 (2001), pp. 223–236.

[35] Rama Cont and Peter Tankov. *Financial Modelling with Jump Processes*. Chapman & Hall/CRC, Boca Raton, FL, 2004.

[36] Fulvio Corsi. 'A Simple Approximate Long-Memory Model of Realized Volatility'. In: *Journal of Financial Econometrics* 7.2 (Feb. 2009), pp. 174–196.

[37] Joshua D. Coval and Tyler Shumway. 'Expected Option Returns'. In: *The Journal of Finance* 56.3 (2001), pp. 983–1009.

[38] John C. Cox and Mark Rubinstein. *Options Markets*. Englewood Cliffs, N.J, Prentice-Hall, 1985.

[39] Francis X. Diebold and Minchul Shin. 'Machine learning for regularized survey forecast combination: Partially-egalitarian LASSO and its derivatives'. In: *International Journal of Forecasting* 35.4 (2019), pp. 1679–1691.

[40] Robert F. Engle. 'Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation'. In: *Econometrica* 50.4 (1982), pp. 987–1007.

[41] Robert F. Engle and Andrew J. Patton. 'What good is a volatility model?' In: *Quantitative Finance* 1 (Feb. 2001).

[42] Stewart N. Ethier. *The Doctrine of Chances*. Springer-Verlag, Berlin, 2010.

[43] Wayne Ferson. *Empirical Asset Pricing: Models and Methods*. MIT Press, 2019.

[44] Mark B. Garman and Michael J. Klass. 'On the Estimation of Security Price Volatilities from Historical Data'. In: *The Journal of Business* 53.1 (1980), pp. 67–78.

[45] B. V. Gnedenko and G. Fahim. 'A certain transfer theorem'. In: *Dokl. Akad. Nauk SSSR* (1969).

[46] Amit Goyal and Alessio Saretto. 'Cross-Section of Option Returns and Volatility'. In: *Journal of Financial Economics* 94.2 (2009), pp. 310–326.

[47] Victor Haghani and Richard Dewey. 'Rational Decision Making Under Uncertainty: Observed Betting Patterns on a Biased Coin'. In: *arxiv.org/abs/1701.01427* (2016).

[48] Campbell R. Harvey, Yan Liu, and Heqing Zhu. '... and the Cross-Section of Expected Returns'. In: *The Review of Financial Studies* 29.1 (Oct. 2015), pp. 5–68.

[49] Jasmina Hasanhodzic and Andrew W. Lo. 'On Black's Leverage Effect in Firms with No Leverage'. In: *The Journal of Portfolio Management* 46.1 (2019), pp. 106–122.

[50] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Second. Springer Series in Statistics. Data mining, inference, and prediction. Springer, New York, 2009.

[51] Jens Carsten Jackwerth. 'Recovering Risk Aversion from Option Prices and Realized Returns'. In: *The Review of Financial Studies* 13.2 (June 2015), pp. 433–451.

[52] Jean Jacod and Albert N. Shiryaev. *Limit Theorems for Stochastic Processes*. Second. Springer-Verlag, Berlin, 2003.

[53] Christopher S. Jones. 'A Nonlinear Factor Analysis of SP 500 Index Option Returns'. In: *Journal of Finance* 61.5 (2006), pp. 2325–2363.

[54] Christopher S Jones. 'Extracting factors from heteroskedastic asset returns'. In: *Journal of Financial Economics* 62.2 (2001), pp. 293–325.

[55] Christopher S. Jones, Jefferson Duarte, and Junbo Wang. 'Very Noisy Option Prices and Inferences Regarding Option Returns'. In: *Working Paper* (2021).

[56] Christopher S. Jones and Austin Pollok. 'Predicting Out-of-Sample Option Returns from Volatility Risk: Can Anything Beat the Benchmark?' In: Working paper (2021).

[57] Christopher S. Jones et al. 'Option Momentum'. In: Working paper (2021).

[58] John L. Kelly. 'A New Interpretation of Information Rate'. In: *The Kelly Capital Growth Investment Criterion*. World Scientific, 1956. Chap. 2, pp. 25–34.

[59] V. Yu. Korolev, L. M. Zaks, and A. I. Zeifman. 'On convergence of random walks generated by compound Cox processes to Lévy processes'. In: *Statist. Probab. Lett.* (2013).

[60] Lily Y. Liu, Andrew J. Patton, and Kevin Sheppard. 'Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes'. In: *Journal of Econometrics* 187.1 (2015), pp. 293–311.

[61] Sergey Lototsky and Austin Pollok. 'Kelly Criterion: From a Simple Random Walk to Lévy Processes'. In: *SIAM Journal on Financial Mathematics* 12.1 (2021), pp. 342–368.

[62] Benoit Mandelbrot. 'The Variation of Certain Speculative Prices'. In: *The Journal of Business* (1963).

[63] Harry M. Markowitz. 'Investment for the Long Run: New Evidence for an Old Rule'. In: *Journal of Finance* (1976).

[64] Nour Meddahi. 'A Theoretical Comparison between Integrated and Realized Volatility'. In: *Journal of Applied Econometrics* 17.5 (2002), pp. 479–508.

[65] Robert C. Merton. 'Theory of Rational Option Pricing'. In: *The Bell Journal of Economics and Management Science* 4.1 (1973), pp. 141–183.

[66] Jacob Mincer and Victor Zarnowitz. 'The Evaluation of Economic Forecasts'. In: *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*. National Bureau of Economic Research, Inc, 1969, pp. 3–46.

[67] Dmitriy Muravyev. 'Order Flow and Expected Option Returns'. In: *The Journal of Finance* 71.2 (2016), pp. 673–708.

[68] Stefan Nagel. *Machine Learning in Asset Pricing*. Vol. 8. Princeton University Press, 2021.

[69] Michael Parkinson. 'The Extreme Value Method for Estimating the Variance of the Rate of Return'. In: *The Journal of Business* 53.1 (1980), pp. 61–65.

[70] Andrew J. Patton. 'Data-Based Ranking of Realised Volatility Estimators'. In: *Journal of Econometrics* 161.2 (2011), pp. 284–303.

[71] Andrew J. Patton. 'Volatility Forecast Comparison Using Imperfect Volatility Proxies'. In: *Journal of Econometrics* 160.1 (2011), pp. 246–256.

[72] William Poundstone. *Fortune's Formula. The Untold Story of the Scientific Betting System that Beat the Casinos and Wall Street*. Hill and Wang, 2006.

[73] Philip E. Protter. *Stochastic integration and differential equations*. Second. Vol. 21. Stochastic Modelling and Applied Probability. Springer, 2005.

[74] L. C. G. Rogers and S. E. Satchell. 'Estimating Variance From High, Low and Closing Prices'. In: *The Annals of Applied Probability* 1.4 (1991), pp. 504–512.

[75] Ken-iti Sato. *Lévy Processes and Infinitely Divisible Distributions*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2013.

[76] Stephen Schulist. 'Fat-Tailed Kelly'. In: *Wilmott* 2021.115 (2021), pp. 64–69.

[77] William G. Schwert. 'Why Does Stock Market Volatility Change Over Time?' In: *The Journal of Finance* 44.5 (1989), pp. 1115–1153.

[78] Peter Van Tassel. 'The Law of One Price in Equity Volatility Markets'. In: *Working Paper* (2020).

[79] Edward O. Thorp. *A Man for All Markets: From Las Vegas to Wall Street, How I Beat the Dealer and the Market*. Random House, New York, 2017.

[80] Edward O. Thorp. 'A perspective on quantitative finance: models for beating the market'. In: *The Best of Wilmott 1: Incorporating the Quantitative Finance Review*. Ed. by P. Wilmott. Willey, 2005.

[81] Edward O. Thorp. *Beat the Dealer. A Winning Strategy for the Game of Twenty-One*. Random House, New York, 1966.

[82] Edward O. Thorp. 'Favorable Strategy for Twenty One'. In: *Proceedings of the National Academy of Sciences of the United States of America* (1961).

[83] Edward O. Thorp. 'Portfolio Choice and the Kelly Criterion'. In: *The Kelly Capital Growth Investment Criterion: Theory and Practice*. Ed. by L. C. MacLean, W. T. Ziemba, and E. O. Thorp. Vol. 3. World Scientific Handbook in Financial Economics Series. World Scientific, 2011, pp. 81–90.

[84] Edward O. Thorp. 'The Kelly Criterion in Blackjack, Sports Betting, and the Stock Market'. In: *The Kelly Capital Growth Investment Criterion*. World Scientific, 2006. Chap. 54, pp. 791–834.

[85] Edward O. Thorp. *The Mathematics of Gambling*. Gambling Times, 1984.

[86] Edward O. Thorp. 'Understanding the Kelly Criterion'. In: *The Kelly Capital Growth Investment Criterion*. World Scientific, 2012. Chap. 36, pp. 511–525.

[87] Edward O. Thorp, Leonard C. MacLean, and William T. Ziemba. *The Kelly Capital Growth Investment Criterion*. World Scientific, 2011.

[88] Edward O. Thorp and William E. Walden. 'A Favorable Side Bet in Nevada Baccarat'. In: *Journal of The American Statistical Association* (1966).

[89] Robert Tibshirani. 'Regression Shrinkage and Selection via the Lasso'. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), pp. 267–288.

[90] Dennis Yang and Qiang Zhang. 'Drift-Independent Volatility Estimation Based on High, Low, Open, and Close Prices'. In: *The Journal of Business* 73.3 (2000), pp. 477–492.