ABSTRACT

Title of Dissertation: ESSAYS IN MATHEMATICAL FINANCE AND MACHINE LEARNING

ZHANG ZHANG
Doctor of Philosophy, 2021

Dissertation Directed by: Professor Dilip B. Madan
Department of Finance

This dissertation consists of three independent essays. Chapter 1, “Exploring Machine Learning in Fixed Income Market” designs a decision support framework that can be used to provide suggested indications of future U.S. on-the-run 10Y Treasury market direction along with the associated probability of making that move. My primary innovation is proposing a framework for applying machine learning methods to U.S. fixed income market. The framework includes a newly proposed performance metric that combines profitability and randomness to select proper outperform models and a sliding window cross-validation method for streaming data learning. I find the Random Forest method provides a decent Sharpe ratio for trading U.S. 10Y Treasury in a “quarantined” testing set but underperforms on Spread trading (10Y Treasury and an asset swap) and Volatility trading (1M10Y Swaption Straddle). Chapter 2, “A Robust Trend Following Framework: Theory and Application” constructs a trend-following signal based on statistical theory and analytically analyzes its properties. I manage to reconcile our model’s theoretical results with stylized facts about trend-following investing – the presence of a "CTA smile". Leveraging on

the theoretical results, we proposed a prototype trend-following framework that is diversified across time-frames and assets. I also discuss the portfolio and risk management of the trend-following strategy. I illustrate the risk-budgeting approach can be used to enhance the trend-following framework. Different approaches to control the costs have also been discussed. Chapter 3, "Markov Modulated Bilateral Gamm Mean Reversion Model" proposed a Markov modulated Bilateral gamma mean-reversion model. Market practitioners argue the market has high volatility regimes and low volatility regimes. I argue the model can capture the mean reversion, asymmetries of returns of up moves and down moves, and other empirical regularities. I derived the characteristic function and provide preliminary parameter estimates by calibrating the model to VIX Index upon the assumption of stationary distribution to avoid using filter methodologies.

ESSAYS IN MATHEMATICAL FINANCE AND MACHINE LEARNING

by

ZHANG ZHANG

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:
Professor Dillip B. Madan, Chair/Advisor
Professor Michael Fu
Professor Leonid Koralov
Professor Mark Loewenstein
Professor Ilya O. Ryzhov

To my family

Acknowledgments

First of all, I would like to give my utmost gratitude to my advisor, Professor Dilip B. Madan. His perpetual energy and enthusiasm in research had motivated all his advisees, including me. Pursing a Ph.D. is challenging, especially when working full-time in the industry. His advice, support, and encouragement drove me forward throughout these years. Without these, this dissertation would not have been possible.

I want to thank Professor Leonid Koralov, Professor Michael Fu, Professor Ilya O. Ryzhov, and Professor Mark Loewenstein to agree to serve on my thesis committee and sparing their invaluable time reviewing the manuscript.

I want to thank participants in the math finance RIT for helpful discussions.

I would like to say thank you to our math finance group members, Sahil Chopra, Yiran Zhang, Yoshihiro Shirai, Khalid Shahnawaz from whom I have learned a lot.

I want to thank Professor Konstantina Trivisa and Professor C. Elman for directing the fantastic program of AMSC. Thanks to all the professors in the classes I took in the first two years of graduate school. They built the foundation for my development in math and engineering. I would also like to acknowledge support from the Math Department staff members, the AMSC Program: Alverda McCoy and Jessica Sadler. They are always helpful.

I am also grateful to my friends Chen Qian, Luyu Sun, Yiran Li, Yiran Zhang, Jinhang Xue for the wonderful graduate life in College Park. I thank my manager Couro Janus and other colleagues at IFC for their support on pursuing my Ph.D.

I owe my deepest gratitude to my family - my mother and father, who have always stood by me and always understood my life decisions. Words cannot express the gratitude I owe them. My grandparents raised me and I owe my grandparents for all the years I cannot spend by their side and not accompanying them in their last days.

I would also like to express my deepest thanks to Ning Li, for her love and unwavering support.

Table of Contents

CHAPTER 1

Exploring Machine Learning in Fixed Income Market

1.1. Introduction

Although predicting stock price direction is something individuals and financial firms have researched for years, and there is plenty of literature written on this subject. However, there is rarely empirical research focusing on the fixed income market direction prediction, especially using the machine learning methodologies, and these kinds of literature are barely repeatable. This paper explores the fixed income market direction prediction using supervised machine learning classification methodologies. These methods would be of considerable interest to quantitative traders who produce mathematical models which account for a decent portion in fixed income already. Our objective is to build a decision support framework that provides the suggested indication of future fixed income market direction and the probability of making that move. We noticed the lack of published working models. We may argue that there is little incentive to publish such models in academic literature.

When predicting the U.S. fixed income products direction, similar to stocks, practitioners typically use one of three approaches. The first is the fundamental analysis which analyzed the economic factors that drive the price. The second approach uses traditional technical analysis to anticipate what others are thinking based on the price and volume. Studies show 80% to 90% of polled professionals, and individual investors rely on at least some form of technical analysis [15]. With advancements in technology and the growing amounts of available data, technical analysis evolves into a more quantitative and statistical approach. This is what we call the quantitative analysis approach, and it is the third approach to predicting market direction.

In this paper, our primary contributions are two folds. First, we proposed a metric to measure the performance of the classifiers in the context of trading. Traditional measurements don't consider randomness, as we presented in Section 1.2. Our metric combines profitability and randomness to measure the performance of the classifiers. Second, and more importantly, we proposed the framework to apply machine learning in the fixed income market. While there is plenty of academic literature on stock prediction, we demonstrate significant economic gains to investors using machine learning forecasts in the U.S. rates market whose trading is not yet fully electronic. A portfolio strategy that times the 10 year U.S. Treasury with random forest tree classifier enjoys an annualized out-of-sample Sharp ratio of 1.35 versus the 0.5 Sharp ratio of a buy-and-hold investor. And sizing daily trades by the machine learning classifier's level of conviction, vis Kelly Criterion, substantially enhanced Sharpe Ratios, across timeframes, asset classes, hold periods, and machine learning methods. Further, the classifiers' perceived conviction in its decision correlated well with its realized hit rate. This suggests a promising application of machine learning in fixed income is in aiding investors in optimal execution.

In economics and finance fields, the application of machine learning methodologies is a promising direction. [126] use neural nets to price options. [127] and [128] also apply a multilayer feedforward neural network to price S&P 500 index calls, and options of Australian All Ordinaries Share Price Index on futures respectively. [129] outline some financial applications of deep learning in portfolios theory. [131] use sequential learning to do return predictability when forming optimal portfolios. [130] introduces the application of deep learning in financial prediction problems.

We begin Section 1.2 with an introduction of different plain vanilla machine learning methods. This includes the description of traditional commonly used supervised learning methods and different ways of evaluating classifiers' performance used for analysis later in the paper. The high dimension nature of these methods improves the flexibility relative to more traditional prediction techniques. This flexibility could help better approximating the unknown and likely complex data generating process. However, due to the enhanced flexibility, overfitting the data becomes the problem. We analyzed different testing methods, including k-fold cross-validation, sliding window, and presequential testing. k-fold cross-validation, the most commonly used technique, is not suitable for our framework. Instead, we combine the sliding window and other methods to avoid information leaks in the cross-validation stage. Then, we briefly discussed the so-called "concept drift." It is an essential concept for streaming data machine learning. In finance, we can consider them as "regimes." When the market condition changes, we need to recalibrate our model using a specific training data set. The timing of the recalibration and the selection of the training data is well discussed in different kinds of literature. Our framework is built based on the assumption that the concept drift occurs and doesn't measure it.

In Section 1.3, we proposed our framework of applying machine learning methods in the fixed income market. We investigate trading liquid duration, spread, and volatility products from 2000 to 2017 using the machine learning classifiers. Our feature space includes around 1000 characteristics of different products in the fixed income market. These products include Treasuries, swaps, swaptions, OIS, TIPS, and international interest rates. Beyond rates products, we also have levels from the corporate bond, FX, commodities, equities markets, and domestic and global economic data. Since the price volatility is high around specific dates, we also include binary 'date flags' for FOMC meetings, payroll releases, and month-ends. Some of our methods expand this feature set much further by having nonlinear transformations and baseline signal interactions.

Our main finding is that machine learning can improve our empirical understanding of fixed income returns at the broadest level. The machine learning classifiers could take the high-dimensional feature set, which is massive from the existing literature's perspective, into a direction classification model. The immediate implication is that machine learning could provide help in solving practical investment problems, such as market timing, portfolio choice, and risk management.

1.2. Methodologies

This section discussed what it meant by machine learning. In Section 1.2.1, we present the collection of most commonly used supervised machine learning methods for regression and classification. Some of them are applied in our analysis. In each subsection, we introduce a new technique and describe the method's general functional form and its objective function for estimating the model parameters. We aim to provide a sufficiently in-depth description of each method so that a reader without machine learning background can understand the basic model structure. In Section 1.2.2, we explores the different commonly used performance metrics for comparing classifiers, such as confusion matrix and ROC. Also, we propose a new performance metric, the Sharpe Ratio Envolope, which is used in our framework for selecting proper model for trading fixed income products. Lastly, we examines different methods of testing in machine learning, such as k-fold cross-validation.

1.2.1. Supervised learning methods. We describe an asset's excess return:

$$r_{i,t+1} = E_t(r_{i,t+1}) + \varepsilon_{i,t+1} \tag{1.2.1}$$

where

$$E_t(r_{i,t+1}) = g^*(z_{i,t}) \tag{1.2.2}$$

Assets are indexed as $i = 1,...,N_t$ and days by $t = 1,...,T$. In prediction, our objective is to find a representation of $E_t(r_{i,t+1})$ as a function of predictor variables (features) that maximizes the out-of-sample explanatory power for realized $r_{i,t+1}$. We can denote those features as vector $z_{i,t}$and assume the conditional expected return $g^*(\cdot)$ is a flexible function of these features. For classification, the set of classess needs to be defined for supervised learning. For example, we could define 3 classes for $r_{i,t+1}$: up move, no move and down move. If we are interested in bigger up moves, we may define more classes. In our analysis, we only define two classes and defer the study of more classes to the later study.

1.2.1.1. *Simple Linear Regression.* We begin with the least complex method – the linear predictive model estimated via ordinary least squares (OLS). We didn't use this model in our analysis and the linear regression is not suitable for the classification job, but we think it is worth to go through predictive models from simplest linear models to more complicated nonlinear models.

The simple linear model imposes $g^*(\cdot)$ can be approximated by a linear function of the raw predictor variables and the parameter vector, θ,

$$g(z_{i,t};\theta) = z_{i,t}^{'}\theta \tag{1.2.3}$$

The model imposes a simple regression specification and does not allow for nonlinear effects or interactions between predictors.

The simple linear model uses a standard least squares, or "l_2" objective function:

$$L(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(r_{i,t+1} - g(z_{i,t};\theta))^2 \tag{1.2.4}$$

Minimizing $L(\theta)$ yields the OLS estimator.

1.2.1.2. *Extension: weighted least squares objective function.* In some cases, replacing 1.2.4 with a weighted least squares objective, such as

$$L_w(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}w_{i,t}(r_{i,t+1} - g(z_{i,t};\theta))^2 \tag{1.2.5}$$

can possibly improve the predictive performance. The weighted least squares objective function allows the users to tilt estimates toward more statistically or economically informative observations. For example, we could set $w_{i,t}$ proportional to the equity market value of stock i at time t. [16] argues the smallest 20% of stocks compose only 3% of aggregate market capitalization. An example of a statistically motivated weighting scheme uses $w_{i,t}$ inversely proportional to an observation's estimated error variance, a choice that potentially improves prediction efficiency in the spirit of generalized least squares.

Heavy tails are a well-known attribute of financial returns. The least squares objective function places extreme emphasis on large errors so that outliers can undermine the stability of OLS. The modified least squares objective functions have been developed to produce more stable forecasts than OLS in the presence of extreme observations. In the machine learning literature, a common choice for counteracting the heavy-tailed observations is the Huber robust objective function, which is defined as

$$L_H(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} H(r_{i,t+1} - g(z_{i,t};\theta), \xi) \tag{1.2.6}$$

where

$$H(x;\xi) = \begin{cases} x^2, & |x| \leqslant \xi \\ \\ 2\xi|x| - \xi^2 & |x| \geqslant \xi \end{cases}$$

The Huber loss, $H(\cdot)$, is a hybrid of squared loss for relatively small errors and absolute loss for relatively large errors, where the combination is controlled by a tuning parameter, ξ , that can be optimized adaptively from the data.[1]

Constructing more robust objective functions are easily applicable in almost all the machine learning methods we study.

1.2.1.3. *Penalized linear.* The simple linear model begins to overfit noise when the number of predictors approaches the number of observations. In return prediciton, the signal-to-noise ratio is notoriously low. It is troublesome to use the simple linear model in this case.

To avoid overfit, we would like to reduce the number of estimated parameters. The most commonly used technique is to append a penalty to the objective function in order to favor more parsimonious specifications. This regularization mechanically deteriorates a model's in-sample performance and hopes it can improve the model's stability out of sample. Penalized methods differ by appending a penalty to the original loss function:

$$L(\theta;\cdot) = L(\theta) + \phi(\theta;\cdot) \tag{1.2.7}$$

[1] OLS is a special case with $\xi = \infty$.

There are several choices for the penalty function. The "elastic net" penalty is a popular one:

$$\phi(\theta;\lambda,\rho)=\lambda(1-\rho)\sum_{j=1}^{P}|\theta_j|+\frac{1}{2}\lambda\rho\sum_{j=1}^{P}\theta_j^2 \tag{1.2.8}$$

The elastic net has two nonnegative hyperparameters, λ and ρ. The $\rho=0$ case corresponds to the lasso and uses l_1 penalization. The lasso can be thought of as a variable selection method. The $\rho=1$ case corresponds to ridge regression, which uses an l_2 penalization. It makes estimates closer to zero but does not impose exact zeros anywhere. In this sense, ridge is a shrinkage method that helps prevent coefficients from becoming large in magnitude. For intermediate values of ρ, the elastic net involves both shrinkage and selection.

1.2.1.4. *Generalized Linear.* Linear models are popular in practice, in part because they can be thought of as a first-order approximation to the data generating process. However, when the "true" model is complex and nonlinear, using linear models introduce approximation error due to model misspecification. Let $g^*(z_{i,t})$ denote the true model and let $g(z_{i,t};\hat{\theta})$ and $r_{i,\hat{t}+1}$ denote the fitted model and predicted return. We can decompose a model's forecast error as:

$$r_{i,t+1}-r_{i,\hat{t}+1}=g^*(z_{i,t})-g(z_{i,t};\theta)+g(z_{i,t};\theta)-g(z_{i,t};\hat{\theta})+\varepsilon_{i,t+1} \tag{1.2.9}$$

The first term represents the approximation error - model misspecification; the second term represents the estimation error; the third term represents the intrinsic error. Intrinsic error is irreducible due to sources of randomness in financial markets. Estimation error is determined by the sample data and can be potentially reduced by adding new observations. Approximation error can be potentially reduced by using more flexible specifications to improve the model's ability to approximate the true model while additional flexibility raises the risk of overfitting and destabilizing the model's out of sample performance.

The generalized linear model introduces nonlinear transformations of the original predictors as new additive terms in an otherwise linear model:

$$g(z;\theta,p(\cdot)) = \sum_{j=1}^{P} p(z_j)'\theta_j \tag{1.2.10}$$

where $p(\cdot) = (p_1(\cdot),...,p_K(\cdot))'$ is a vector of basis functions, and the parameters are now a $K \times N$ matrix. The squared loss function is replaced by the unit deviance d of a distribution in the exponential family. The objective function becomes

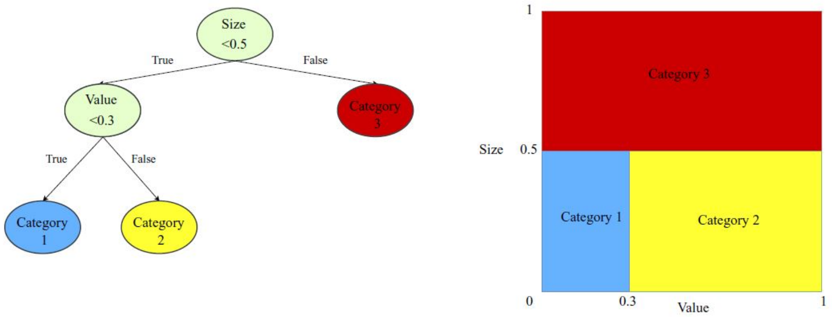
$$\frac{1}{2N}\sum d(r_{i,t},\hat{r}_{i,t}) + \frac{\alpha}{2}||\theta||_2$$

1.2.1.5. *Decision Tree.* The generalized linear model captures individual predictors' nonlinear impacts, but it does not capture the interactions among predictors. One way to add interactions is to expand the generalized model to include multivariate functions of predictors. However, without a priori assumptions for which interactions to include, the generalized linear model becomes computationally infeasible.

As an alternative, decision trees have become a popular machine learning approach for classification and regression by incorporating predictor interactions in practice because the algorithm creates rules which are easy to understand and interpret. At a basic level, trees are designed to find groups of observations that behave similarly to each. A tree "grows" in a sequence of steps. At each step, a new "branch" sorts the data leftover from the preceding step into bins based on one of the predictor variables. This sequential branching slices the space of predictors such that the samples with the same labels or similar target values are grouped togehther, and approximates the unknown function $g^*(\cdot)$ with the average value of the outcome variable within each partition.

Figure 1.3.4-1shows an example with two predictors, "size" and "b/m.". More formally, let the data at node m be represented by Q_m with N_m samples. For each candidate split $\theta = (j, t_m)$ consisting of a feature j and threshold t_m, partition the data into $Q_m^{left}(\theta)$ and $Q_m^{right}(\theta)$ subsets

FIGURE 1.2.1. Tree Example



This figure presents the diagrams of a regression tree (left) and its equivalent representation (right) in the space of two characteristics (size and value). The terminal nodes of the tree are colored in blue, yellow, and red.[a]

[a] Shihao Gu, Bryan Kelly, Dacheng Xiu "Empirical Asset Pricing via Machine Learning"

$$Q_m^{left}(\theta) = \{(x,y\}|z_j \leq t_m\}$$

$$Q_m^{right}(\theta) = Q_m \backslash Q_m^{left}(\theta)$$

The quality of a candidate split of node is then computed using an impurity function or loss function $H(\cdot)$, the choice of which depends on the task being solved (classification or regression)

$$G(Q_m,\theta) = \frac{N_m^{left}}{N_m} H(Q_m^{left}(\theta)) + \frac{N_m^{right}}{N_m} H(Q_m^{right}(\theta)) \tag{1.2.11}$$

Select the parameters that minimise the impurity

$$\theta^* = argmin_\theta G(Q_m,\theta)$$

Recurse for subsets $Q_m^{left}(\theta^*)$ and $Q_m^{right}(\theta^*)$ until the maximum allowable depth is reached, $N_m < min_{samples}$ or $N_m = 1$.

If a target is a classification outcome taking on values $0, 1, ...$ for node m, let

$$p_{mk} = 1/N_m \sum_{y \in Q_m} I(y = k)$$

be the proportion of class k observations in node m. If m is a terminal node, the prediction probability for this region is set to p_{mk}. Common measure of impurity are the following.

(1) Misclassification:

$$H(Q_m) = 1 - max(p_{mk})$$

(2) Gini:

$$H(Q_m) = \sum_k p_{mk}(1 - p_{mk})$$

(3) Entropy

$$H(Q_m) = -\sum_k p_{mk} log(p_{mk})$$

The version we use in this paper is also one of the most popular forms, the C4.5 [2], which extends the ID3 [3] algorithm. The improvements are: 1) it is more robust to noise, 2) it allows for the use of continuous attribute, and 3) it works with missing data. The C4.5 begins as a recursive divide-and-conquer algorithm, first by selecting an attribute from the training set to place at the root node. Each value of the attribute creates a new branch, with this process repeating recursively using all the instances. An ideal node contains all (or nearly all) of one class. To determine the best attribute to choose for a particular node in the tree, the gain in information entropy for the decision is calculated.

Advantages of the tree model are that it is invariant to monotonic transformations of predictors, that it naturally accommodates categorical and numerical data in the same model, that it can approximate potentially severe nonlinearities, and that a tree of depth L can capture (L−1)-way interactions. Their flexibility is also their limitation. Trees are among the prediction methods most prone to overfit, and therefore must be heavily regularized.

1.2.1.6. *Ensemble.* [17] summarizes that an ensemble is a collection of multiple base classifiers that take a new example, pass it to each of its base classifiers, and then combine those predictions according to some method (such as through voting). The motivation is that by combining the predictions, the ensemble is less likely to misclassify. The idea of classifier independence may be unreasonable, given that the classifiers may predict similarly due to the training set. Obtaining a base classifier that generates errors as uncorrelated as possible is ideal. Creating a diverse set of classifiers within the ensemble is considered an important property since the likelihood that a majority of the base classifiers misclassify the instance is decreased. Two of the more popular methods used within ensemble learning are bagging and boosting. These methods promote diversity by building base classifiers on different subsets of the training data or classifiers' weights.

As we discussied in the previous section, decision trees must be heavily regularized.

The first regularization is "boosting". In boosting, instances being classified are assigned a weight; instances that were previously incorrectly classified receive larger weights, with the hope that subsequent models correct the mistake of the previous model. For example, in the AdaBoost [4] algorithm the original training set D has a weight w assigned to each of its N instances $\{(x_1,y_1),....,(x_n,y_n)\}$, where x_i is a vector of inputs and y_i is the class label of that instance. The AdaBoost algorithm then builds k base classifiers with an initial weight $w_i = \frac{1}{N}$. In each step, the weight gets updated according to the error ε_i of each classifier. The reweighting will help the classifiers to correctly classify the instances that were misclassified. The final class is determined by a weighted vote of the classifiers.

The second regularization is "Random Forest". Random forest is a method of tree boosting by bagging. The bagging works by generating k bootstrapped training sets and building a classifier on each (where k is determined by the user). Random forests use a variation on bagging designed to reduce the correlation among trees in different bootstrap samples. If, for example, firm book value is the dominant factor, then most of the bagged trees will have low-level splits on book value resulting in substantial correlation among their ultimate predictions. The forest method decorrelates trees using a method known as “dropout,” which considers only a randomly drawn subset of predictors for splitting at each potential branch. Doing so ensures that, in the example, early branches for at least a few trees will split on characteristics other than book value. This lowers the average correlation among predictions to further improve the variance reduction relative to standard bagging. Depth L of the trees, number of classifiers k in each split and number of bootstrap samples B are the tuning parameters optimized via validation.

While ensembles have shown success in a variety of problems, there are some associated drawbacks. This includes added memory and computation cost in keeping multiple classifiers stored and ready to process. Also the loss of interpretability may be a cause for concern depending on the needs of the problem. For example, a single decision tree can be easily interpreted, while an ensemble of 100 decision trees could be difficult. [5]

1.2.1.7. *Support Vector Machine.* Support vector machines are commonly applied to classification problems. The basic idea of SVM is to find a slice through the predictor (feature) space that best separates disparate outcomes. For a binomial predictor in two dimensions, this can be visualized as a line drawn to separate two prediction classes as cleanly as possible, pick in particular a line with the largest gap (margin) between itself and the data points; in many dimensions, the line becomes a “hyperplane.” Intuitively, a good separation is achieved by the hyper-plane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general, the larger the margin, the lower the generalization error of the classifier. This approach is dubbed “linear” SVM. In general, when the problem isn’t linearly separable, the support vectors are the samples within the margin boundaries.[17]

More formally, given training vectors $x_i \in \mathbb{R}^p$,i=1,..,n in two classes, and a vector $y \in \{1,-1\}^n$, the goal is to find $\omega \in \mathbb{R}^p$ and $b \in \mathbb{R}$ such that the prediction given by $sign(\omega^T \phi(x)+b)$ is correct for most samples.

SVC solves the following primal problem:

$$min_{\omega,b,\zeta}\frac{1}{2}\omega^T\omega + C\sum_{i=1}^{n}\zeta_i$$

$$subject to y_i(\omega^T\phi(x_i)+b) \geq 1-\zeta_i$$

$$\zeta_i \geq 0, i = 1, ...n$$

Intuitively, we’re trying to maximize the margin (by minimizing), while incurring a penalty when a sample is misclassified or within the margin boundary. Ideally, the value would be for all samples, which indicates a perfect prediction. But problems are usually not always perfectly separable with a hyperplane, so we allow some samples to be at a distance from their correct margin boundary. The penalty term C controls the strengh of this penalty, and as a result, acts as an inverse regularization parameter.

The dual problem to the primal is

$$min_{\alpha}\frac{1}{2}\alpha^{T}Q\alpha - e^{T}\alpha$$

$$subject to y^{T}\alpha = 0$$

$$0 \leq \alpha_i \leq C, i = 1, ..., n$$

where e is the vector of all ones, and Q is an n by n positive semidefinite matrix, $Q_{ij} = y_i y_j K(x_i, x_j)$, where $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is the kernel. The terms α_i are called the dual coefficients, and they are upper bounded by C. This dual representation highlights the fact that training vectors are implicitly mapped into a higher (maybe infinite) dimensional space by the function ϕ.[6]

Once the optimization problem is solved, the output for a given x becomes:

$$\sum_{i \in SV} y_i \alpha_i K(x_i, x) + b$$

and the predicted class corresponds to its sign. We only need to sum over the support vectors because the dual coefficients $\alpha_i$are zeo for the other samples.

In summary, the linear support vector classier can be represented as

$$g(x) = \theta_0 + \sum_{i=1}^{n} \alpha_i < x_i, x > \tag{1.2.12}$$

where $\alpha_i \neq 0$ only for all support vectors. Moreover, α_i can also be computed based $< x_i, x_j >$. Only the inner product of the feature space is relevant in computing the linaer support vector classer. The above support vector classier has a linear boundary, which may not be the "ground truth" in practice. To cope with more general cases, one can consider to enlarge the feature space. A straightfoward method is to include the power functions of the inputs. A better approach is the use of the kernel trick, which gives rise to the suppot vector machines. The support vector machine actually enlarges the original feature space to a space of kernel functions:

$$x_i \longrightarrow K(\cdot, x_i)$$

The kernel functions are bivariate functions satisfying the property of nonnegative deniteness:$\sum_{i,j} \omega_i y_i K(x_i, x_j) \geq 0$. The original feature space is the p-dimensional input space. The enlarged feature space is the space of kernel functions, which is in fact of infinite dimension. In actual fitting of the support vector machine, we only need to compute the $K(x_i, x_j)$ for all x_i, x_j in training data.

The commonly used kernel functions are

- linear kernel $K(x_i, x_j) =< x_i, x_j >= x_i^T x_j$

- polynomial kernel of degree d: $K(x_i,x_j) = (1+ < x_i,x_j >)^d$
- Gaussian radial basis function (RBF) kernel: $K(x_i,x_j) = exp(-\gamma \| x_j - x_i \|^2),\ \gamma > 0$
- Sigmoid kernel (Hyperbolic Tangent Kernel): $K(x_i,x_j) = tanh(\gamma < x_i,x_j > +r)$

More information can be found in [7].

1.2.1.8. *Nearest Neighbors.* Nearest neighbor is one of the simplest methods. It takes the most frequent class measured by the weighted euclidean distance (or some other distance measure) among the k closest training examples in the feature space. In specific problems such as text classification, NN has been shown to work as well as more complicated models [8]. A downside of using this model is the slow classification times. However, we can increase speed by using dimensionality reduction algorithms. This chapter used the kNN, k nearest neighbors of each query point, where k is an integer value specified by the user.

The basic nearest neighbors classification uses uniform weights: the value assigned to a query point is computed from a simple majority vote of the nearest neighbors. It is better to weigh the neighbors under some circumstances such that nearer neighbors contribute more to the fit. We use the uniform weights in our case and the brute force computation of distances between all pairs of points in the data.

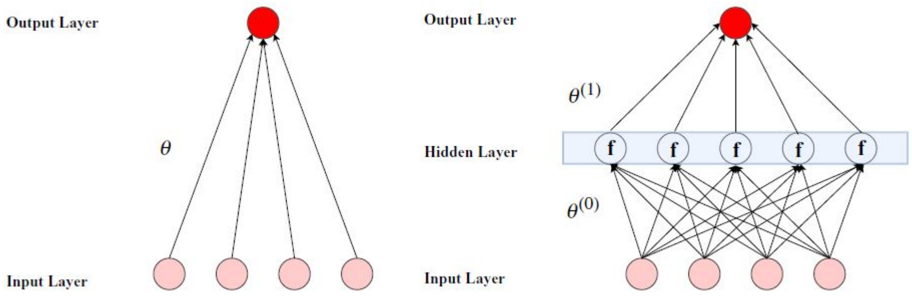
1.2.1.9. *Artificial Neural Net.* Arguably the most powerful model in machine learning, neural networks have theoretical underpinnings as "universal approximators" for any smooth predictive association.[18] They are the currently preferred approach for complex machine learning problems, such as computer vision, natural language processing, and automated game-playing. Their flexibility draws from the ability to connect many telescoping layers of nonlinear predictor interactions. At the same time, their complexity ranks neural networks among the least transparent, least interpretable, and most highly parameterized machine learning methods. We present only a short overview of their structure in this section.

We focuses on traditional "feed-forward" networks. These consist of an "input layer" of raw predictors, one or more "hidden layers" that interact and nonlinearly transform the predictors, and an "output layer" that aggregates hidden layers into an ultimate outcome prediction. 1.2.2 shows two illustrative examples.

The left panel shows the simplest possible network that has no hidden layers. Each of the predictor signals is amplified or attenuated according to a 5-dimensional parameter vector, θ, that includes an intercept and one weight parameter per predictor. The output layer aggregates the weighted signals into the forecast $\theta_0 + \sum_{k=1}^{4} z_k \theta_k$; that is, the simplest neural network is a linear model.

The model incorporates more flexible predictive associations by adding hidden layers between the inputs and output. The right panel of 1.2.2 shows

FIGURE 1.2.2. Neural Networks



This figure provides diagrams of two simple neural networks with (right) or without (left) a hidden layer. Pink circles denote the input layer, and dark red circles denote the output layer. Each arrow is associated with a weight parameter. In the network with a hidden layer, a nonlinear activation function f transforms the inputs before passing them on to the output.[a]

[a] Shihao Gu, Bryan Kelly, Dacheng Xiu "Empirical Asset Pricing via machine learning"

an example with one hidden layer that contains five neurons. Each neuron draws information linearly from all of the input units, just as in the simple network on the left. Then, each neuron applies a nonlinear "activation function" f to its aggregated signal before sending its output to the next layer. In this example, there are a total of $31 = (4+1) \times 5 + 6$ parameters (five parameters to reach each neuron and six weights to aggregate the neurons into a single output).

Formally, given a set of training example (x_i, y_i) where $x_i \in \mathbf{R}^n$ and $y_i \in \{0,1\}$, a one hidden layer neuron learns the function $f(x) = W_2 g(W_1^T x + b_1) + b_2$ where $W_1 \in \mathbf{R}^m$ and $W_2, b_1, b_2 \in \mathbf{R}$ are model parameters. W_1, W_2 represent the weights of the input layer and hidden layer, respectively; and b_1, b_2 represent the bias added to the hidden layer and the output layer, respectively. $g(\cdot) : R \to R$ is the activation function.

[9] state that with foreign exchange rate forecasting, which is similar to stocks because of the high degree of noise, volatility and complexity, it is advisable to use the sigmoidal type-transfer function (i.e. logistic or hyperbolic tangent).

We used the hyperbolic tan given as

$$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{1.2.13}$$

For binary classification, $f(x)$ passes through the logistic function $g(z) = 1/(1+e^{-z})$ to obtain output values between zero and one. A threshold, set to 0.5, would assign samples of outputs larger or equal 0.5 to the positive class, and the rest to the negative class.

If there are more than two classes, $f(x)$ itself would be a vector of size (n classes). Instead of passing through logistic function, it passes through the softmax function, which is written as,

$$softmax(z)_i = \frac{exp(z_i)}{\sum_{l=1}^{k} exp(z_l)} \tag{1.2.14}$$

where $z_i$represents the ith element of the input to softmax, which corresponds to class i, and K is the number of classes. The result is a vector containing the probabilities that sample x belong to each class. The output is the class with the highest probability.

We applied the Cross-Entropy loss function for classification, which in binary case is given as,

$$L(\hat{y}, y, W) = -yln\hat{y} - (1-y)ln(1-\hat{y}) + \alpha \parallel W \parallel_2^2 \tag{1.2.15}$$

where $\alpha \parallel W \parallel_2^2$is a regularizaiton term that penalizes complex models; and $\alpha > 0$ is a non-negative hyperparameter that controls the magnitude of the penalty.

Starting from initial random weights, our neural network minimizes the loss function by repeatedly updating these weights. After computing the loss, a backward pass propagates it from the output layer to the previous layers, providing each weight parameter with an update value meant to decrease the loss.

In gradient descent, the gradient $\bigtriangledown L_W$ of the loss with respect to the weights is computed and deducted from W. More formally, this is expressed as,

$$W^{i+1} = W^i - \varepsilon \bigtriangledown L_W^i \tag{1.2.16}$$

where i is the iteration step, and ε is the learning rate with a value larger than 0.

The algorithm stops when it reaches a preset maximum number of iterations; or when the improvement in loss is below a certain, small number.

In our analysis, we applied the Stochastic Gradient Descent. Also, we consider architectures with up to five hidden layers. We choose the number of neurons in each layer according to the geometric pyramid rule. We used the gridsearch for the regularization parameter α.

1.2.2. Performance Metrics.

1.2.2.1. *Confusion matrix and accurary.* A confusion matrix is a visualization of the performance of a supervised learning method. In a confusion matrix, TP (true positive) is the number of positives correctly identified, TN (true negative) is the number of negatives correctly identified, FP (false positive) is the number of negatives incorrectly identified as positive, and FN (false negative) is the number of positives incorrectly identified as negatives. From the confusion matrix, we could define commonly used measures e.g. Accuracy

$$\begin{aligned} Accurary &= \frac{TP+TN}{TP+TN+FP+FN} \\ Error\, rate &= \frac{FP+FN}{TP+TN+FP+FN} \end{aligned}$$

However, when the dataset is imbalanced, the model may have high accuracy but may not provide us with high-level accuracy in classifying the class we are interested in. Also, the accuracy metric does not take into account pure randomness.

There are other approaches to comparing models with imbalanced datasets. We will discuss these methods in the next few sections. These metrics are precision and recall, harmonic mean, and the F-measure. The harmonic mean considers the class's randomness using Cohen's kappa statistic, while ROC is based on the TP and FP rates.

1.2.2.2. *Precision and Recall.* Precision and recall are both popular metrics for evaluating classifier performance. These metrics are defined

$$\begin{aligned} Precision &= \frac{TP}{TP+FP} \\ Sensitivity\,(Recall) &= \frac{TP}{TP+FN} \\ Specificity &= \frac{TN}{TN+FP} \\ F-measure &= \frac{2(precision)(recall)}{precision+recall} \end{aligned}$$

The F-measure is the harmonic measure of precision and recall in a single measurement. The F-measure ranges from 0 to 1, with a measure of 1 being a classifier perfectly capturing precision and recall. We can use these metrics to measure the performance of the class we are interested even in a imbalanced data.

1.2.2.3. *Cohen's kappa statistic.* Cohen's kappa statistic takes into account the randomness of the class and provides an intuitive result. [10] defined these metrics

$$\kappa = \frac{P_0 - P_c}{1 - P_c}$$

$$P_0 = \sum_{i=1}^{I} P(x_{ii})$$

$$P_c = \sum_{i=1}^{I} P(x_{i.})P(x_{.i})$$

where P_0 is called the total agreement probability P_0 (i.e. the classifier's accuracy), $P(x_{i.})$ is the row marginal probability and $P(x_{.i})$ is the column probability computed from the confusion matrix, and P_c is the agreement probability due to the chance. The kappa statistics is on $[-1,1]$. $\kappa = 0$ means the agreement is equal to random chance, and $\kappa = 1$ and -1 means perfect agreement and perfect disagreement.

1.2.2.4. *ROC.* Receiver operating characteristic curve (ROC) is a plot of the true positive rate which is also called recall against the false positive rate, which is 1-specificity.

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{TN + FP}$$

The top left corner on the ROC graph means the best performance with the highest TP and TN. The Area Under the ROC Curve (AUC) is calculated by integrating the ROC curve. AUC is a single number comparison. Random classifier would therefore have an AUC of 0.50, and a classifier better and worse than random would have an AUC greater than and less than 0.50, respectively. It is most commonly used with two-class problems.

1.2.2.5. *Costs.* The cost-based method is based on the "cost" associated with making incorrect decisions[11]. The performance metrics we discussed so far do not consider the possibility that not all classification errors are equal. For example, an opportunity cost can be associated with missing a significant move in a stock.

One of the advantages of the cost-based evaluation metric for trading is the cost associated with making incorrect decisions is known. For example, the cost of the wrong prediction of the "no change" state will only cost us the opportunity, while the cost of the wrong prediction of "up move" will cost us real money. Hence, different errors of the classifier's results would have a different associated cost. The table shows a hypothetical cost matrix of the trading problem by classifying the "down move", "no move", "up move" of the stock market.

TABLE 1. Hypothetical Cost Matrix

		Classifier's results		
		down move	no move	up move
Actual Class	down move	0	1.0	4.0
	no move	2.0	0	2.0
	up move	4.0	1.0	0

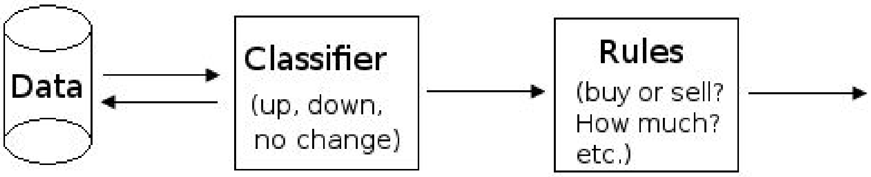
1.2.2.6. *Profitability - Sharpe Ratio Envelope.* While the objective of correctly predicting the directional movement of the US Treasury is to achieve better profitability, the performance metrics we discussed so far are only based on the ability to correctly classify but not on the trading system's overall profitability. For example, a classifier may have very high accuracy, kappa, AUC, but this may not necessarily produce a profitable trading strategy. The profitability of individual trades may be more critical. For example, making $1 on one hundred trades is not as profitable as losing $0.5 95 times and making $30 on each of the five trades. On the contrary, we also can argue that a less volatile approach is more ideal (i.e., making small sums consistently). This depends on the overall objective of the trader.

As shown in the figure, a classifier is trained on the historical data and make classification decision "up", "hold", "down" on the next one week or one month. The classification results are passed to the rule system, which sets the trading rules based on the classification results. For example, a "up" leads to buy a 10Y US Treasury, and a "down" leads to short a US 10Y Treasury. The rules shall also address the amount of US 10Y Treasury to be purchased, how much risk to take, etc.

In this chapter, our trading system's rules are to follow the classification results to open a position daily and close the position after one week or one month. The position size is determined based on the classifiers' confidence.

We proposed a profitability based method to measure the performance of the classifiers. The method takes into consideration the randomness

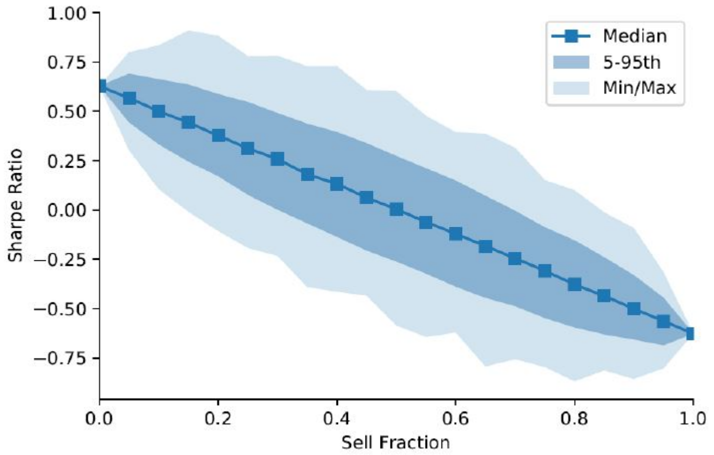
FIGURE 1.2.3. Trading system



which we is commonly called “luck” on the trading floor. We call it Sharpe Ratio Envelope in this paper.

1.2.4shows the SRE (Sharpe Ratio Envelop) we constructed. We take our classifier’s daily position decisions, randomly shuffle those decisions (permute them), and calculate the new Sharpe. We then repeat this exercise many thousands of times, recording the 95th and 99th percentile Sharpe Ratio of these randomly permuted returns along with the maximal Sharpe Ratio from all the trials. The median, 95th, and maximal randomly discovered Sharpe ratios from 3,000 random trials are shown as a function of sell fraction, i.e., the percentage of days you were short instead of long. The median Sharpe threshold drops from roughly 0.6 for an all-long strategy to 0.0 for a 50/50 split, to -0.6 for all-short over this period. The maximal Sharpe from 3000 random trials peaks near a sell fraction of 20%, meaning if you were required to go short 1 out of 5 trades, you have a roughly 1:3,000 chance of achieving those risk-adjusted returns by luck. It is the “residual Sharpe Ratio” in excess of these percentiles of randomly permuted decisions that we consider when declaring an improvement over luck for all our classifiers. Using this metric, it is much more impressive if a machine learning strategy with a 50% sell fraction produces a Sharpe of 0.75 than if an 80%-long strategy does likewise.

FIGURE 1.2.4. Sharpe Ratio Envelope



1.2.3. Methods of testing. Once a model's parameters are calibrated on the training set, its performance needs to be evaluated on a testing set. The testing set is used since the model is biased toward the training set and may over-fit the data. Overfitting leads to an artificially high-performance measure. The following subsection reviews some of the methods used to evaluate the performance of classifiers.

1.2.3.1. *Holdout.* The holdout method split dataset D into two disjoint sets, $D_{training}$ and D_{test}. The split varies from a 50-50 to a two-thirds for training and a one-third split for testing.

There are several problems associated with the holdout methods. First, splitting the data into disjointed sets reduces the amount of data available for learning. This can be mostly solved by using random subsampling. However, it doesn't fit for our time series data analysis.

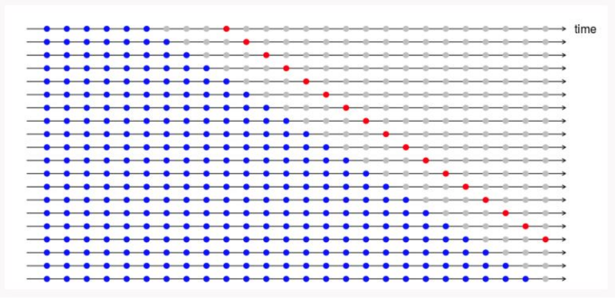
The holdout method's second potential problem is the performance metric can be high if the classes are imbalanced between the training and testing set. It is pertinent, especially when evaluating the streaming US Treasury data, where the underlying structure of the data may change over time due to changing market dynamics—for example, training a model on an upward moving market, where the class "large up move" outweighs "large down move". Then, we test the model on a downward moving market where "large down move" dominates "large up move".

1.2.3.2. *Sliding window.* We propose to use the sliding window approach for our analysis. It can prevent information leaks assuming history could be repeated in the future in financial market.

As shown in the figure, The model is trained on the blue data sets and then tested on the data sets after the red dot. There is a gap between the training and testing set because the gap is the holding period of the security. The method is the most intuitive and it is the main method we will use in our trading system. In our analysis, the size of the sliding window are predefined by us. However, such priori could be determined by an adaptive algorithm that takes account of the level of concept drift in the data. We defer the adaptive approach to the later study.

1.2.3.3. *Prequential.* In data streaming, prequential [12] becomes popular because it monitors a model's error over time by predicting unseen

FIGURE 1.2.5. Sliding Window in our Framework



instances one-by-one. It will add those instances into the training set after the observed value is known. The error is computed as the sum of a loss function between the observed values y_i and predicted values $\hat{y}_i$. In the beginning, few instances are included in the model, and it causes high error rates. The high error rates then lead to the inclusion of a forgetting factor in the model, giving less weight to previously seen examples. The forgetting factor can include either a sliding window of size n or a decay factor. More information can be found in [13].

1.2.3.4. *k-fold Cross Validation.* k-fold Cross-validation split the dataset into k equal subsets. K − 1 subsets are used as the training set, and the remaining subset is used as the testing set. This is repeated k times so that each subset is used as a testing set once. The errors obtained during all k runs are added together, and then the performance metric is computed as the average across runs.

Although k-fold cross-validation is the most popular approach to test the model, we think it doesn't fit our purpose because the future market conditions are being leaked. It may lead to an overly optimistic and biased classifier. Models should be tested only on data that was not available at the time when the model was created.

1.2.4. Concept Drift.

1.2.4.1. *Definition and Causes.* The changing of the underlying target variable's statistical properties, or called concept drift, makes learning from streaming data difficult. It also makes the task of keeping models relevant difficult. As the concept drift changes, model performance may decrease and require a change or update in the training data. Ideally, if the concept drift is known, the traders could use different models for each specific market condition.

However, the assumption is that the concept generating function is unknown. We know the market periodically displays reoccurring behavior such as economic cycles or certain market moods but specific market conditions are rarely consistently known as a priori. Also, the idea of using the most recent training data may not be optimal for all problems.

[14] may give the most common definition of concept drift explaining it in three forms that concept drift can occur: (1) the class priors, $P(c_i)$ may change over time, (2) the distribution of the classes may change, $P(X|c_i)$, where X is a vector of labeled instances, and (3) the posterior distribution of the class member $P(c_i|X)$ may change.

In traditional offline machine learning, we usually assume the training and testing data are from a stationary distribution with the same concept generating function. However, this assumption is often violated in the financial market. Many adaptive algorithms are developed for streaming data, but they still have difficulty maintaining high performance dealing with concept drift. We may overreact to the noise and update the training set so that the model loses past knowledge that may be helpful in the future. Meanwhile, not updating training data frequently enough leads to a model with poor performance. There are some reasons that concept drift occurs in the market. For example, traders preference changes; the political environment changes; crowding trades eliminates the predictability of a predictor.

1.2.4.2. *Approaches to learning with concept drift.* Different literature proposes different approaches to build algorithms to learn with drifting concepts, but they usually take two forms. The first is to detect concept drift, such as the novelty detection algorithms. Once detected, we update the classifiers. The second approach assumes the concept drift occurs and considers this assumption when building the model. In the second approach, the actual level of concept drift or its occurrence may not be measured. For example, we could train the model using a fixed size of the training set, such as the sliding windows approach. Another solution is to use ensembles that use a pool of trained classifiers updated according to the heuristic.

There are several algorithms developed to detect the concept drift and adjust the model training accordingly. We will not review these methods in this chapter. Building an adaptive online learning framework for trading is deferred to the later study.

Our framework applied the second approach, which assumes the concept drift occurs and doesn't measure it. There are several advantages of this approach. First, although our approach requires longer training times than adaptive classifiers, it is offset by only training models periodically. Second, our approach also allows us to use the traditional machine learning models that could be trained parallelly on a multiple cores machine. Third, using the up-to-date data for the model may not be ideal since markets may stabilize and old knowledge may become useful again. Fourth, to handle the US rates market data problems, such as class imbalance and dimensionality reduction, using the full subset of data is much more easier.

1.3. An Empirical Study of machine learning methods trading Fixed Income products

1.3.1. Data and Data Preprocessing.

1.3.1.1. *Data Universe.* We obtain the data from 2000-2016 (1) the daily close levels across all benchmark tenors USD interest rates and beyond such as Treasuries, swaps, swaptions, OIS, TIPS, and international interest rates; (2) the daily close levels across the corporate bond, agency mortgage, FX, commodities, equities markets, and domestic economic data; (3) the binary date flag for FOMC meetings, payrolls releases, and month-ends; (4) the 1-week, 1-month and 3-month changes in levels and 3-month, 6-month, 1-year and 2-year trailing z-scores for most of the products with levels.Table 1 summarizes this broad feature set. Altogether this consists of roughly 1,000 raw input features.

The broad feature sets reflect our "agnostic" preliminary approach to explore machine learning in the fixed income market. We could have used our domain knowledge to create a more focused list of drivers. Also, we could have created signals without strong collinearities (e.g., using the level, slope, and curvature of the Treasuries curve, rather than all the benchmark yields themselves). This "agnostic" approach comes at the cost of a lack of transparency.

We separate the data into two timespans: a post-crisis period from mid-2008 to 2016 and a more extended millennial period from 2000-2016. The first time-period has a richer feature set: many of our input features have only been reliably and frequently tracked since the financial crisis. The second period covers a broader set of market conditions, including the complete arc of the global financial crisis, albeit with a sparser set of the feature set. We elected to quarantine all 2017 data, which was not used for training or testing purposes under any circumstances, until we had arrived at our final, tuned, and productized machine learning strategies.

This limited time frame and the slow frequency of our data (daily closes) give us a limited sample size of roughly 1500 testing days for the post-crisis set and approximately 3500 testing days for the millennial period. The sample points will also exhibit a strong degree of autocorrelation, even if the features and asset performance have no autocorrelation. Hence, the total number of plausibly independent observations is much lower.

TABLE 2. Input Features Set

Treasuries	yield, carry, repo rates
Swaps	realized volatility, carry
Swaptions	Implied Volatility Surface, skew
OIS	Rates
TIPS	Breakevens
MBS	Mortage basis, convexity
Cross-Asset	HG and HY bond indices, equity index levels and volatility, FX indices and volatility, commodity indices
Economic	Global and regional economic indices, various date flags
Dates	Flags for FOMC meetings, payrolls and month-ends

1.3.1.2. *Data normalization.* Normalization, refers to the process of transforming the data for use in a training model. We showed the most common techniques below

$$x_t = \frac{x_t - x_{min}}{x_{max} - x_{min}}$$

$$x_t = \frac{x_t}{x_{max}}$$

$$x_t = \frac{x_t - \mu}{\sigma}$$

$$x_t = log(x_t)$$

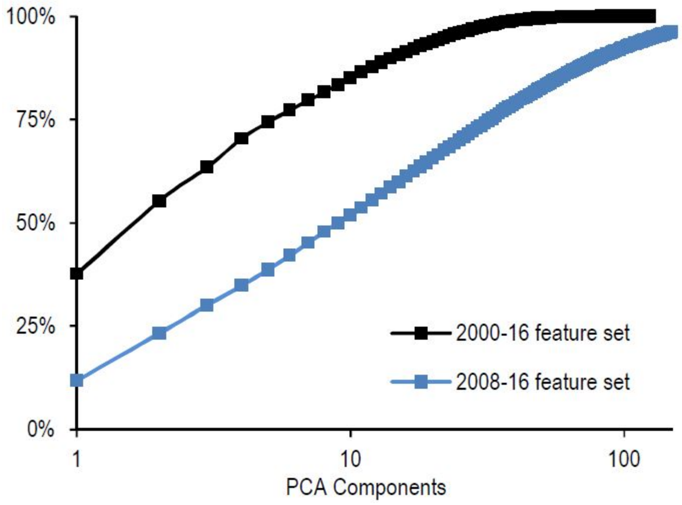
There are two reasons for normalization. First, some models such as ANN are prone to outliers. The normalization can help to eliminate the problem. Second, trends may be present in the time series, which is called non-stationarity. The presence of trends often degrades the classifiers' performance. The transformation is used to stabilize the variability of the series. In online learning, normalization is more difficult because we cannot assume that the data's future distribution will remain the same as the past, and we will not know the min and max values until all the information is available. In our framework, there is no such problem. However, minimum and maximum values are necessarily stored to make similar comparisons with future data.

1.3.1.3. *Dimension Reduction.* The benefits of dimension reduction are commonly well recognized. It can help the model make a better prediction and reduce the computational and memory/storage burden. There are multiple approaches to reduce dimensionality. The first one is by creating new features that combine existing features, such as combine stock price and earing using the P/E ratio. The second way to reduce dimensionality is by selecting a subset of the features. Feature subset selection is the process of removing as much irrelevant information as possible. There are three commonly used methods: filter, wrapper feature selection, and embedded methods.

Recall that our approach is the "agnostic" preliminary approach. Hence, we didn't apply feature selection techiniques. We applied one of the feature project techinique - PCA to reduce the dimensionality of our dataset. We defer discussion of the feature selection in our framework to the later study.

We apply the transformation to all our continuous variables, excluding binary "date flags" from the process. It results in a set of orthogonal (and thus uncorrelated) features, the first few of which – the "principal components" – explain a majority of the variance in the original raw feature set. As illustrated in 1.3.1, ten components capture roughly half the variation in the richer 2008-16 dataset while capturing approximately 80% of variation on the longer but sparser 2000-16 dataset. Approaching 90% of the variation requires ~100 components for the 08-16 set, our primary sample for this piece.

FIGURE 1.3.1. PAC reduces the dimentionality of the raw feature set



1.3.2. Framework For Machine Learning in Fixed Income Trades.

1.3.2.1. *Trading Strategy and Rules.* Recall we illustrate the basic framework in section 1.2. The classifiers' prediction results are passed to the rule box for executing based on a particular set of predefined rules. We developed the strategies (rules) in this chapter to execute trades daily based on the prediction results and close the position after one week or one month. Thus the only decision to make is positioning: whether to buy or sell, and in what size. In some cases, the trade size is uniform for each day, and the predictor simply needs to decide to buy or sell. In other cases, the algorithm is allowed to arrive at an optimal trade size. A slightly more advanced approach would allow the classifiers to select the proper structure to trade daily and/or arrive at the decision of when to close the position on its own. We looked primarily at three different trade structures: 10-year Treasuries, 10-year matched maturity swap spreads, and 1Mx10Y ATMF swaption straddles.

Thus our problem boils down to predicting the optimal daily trade positioning, given our set of input features. Supervised machine learning can work in two broad tasks: classification and regression. In this chapter, the machine learning models can either predict the magnitude and direction of returns (regression) or simply whether the asset will rally, sell-off, or move sideways (classification). We limited our investigations to the latter approach, exploring classification schemes. For most of our work, we trained the classifiers to classify just two outcomes: rally or sell-off, and thus whether to buy or sell. We defer the multinomial classification study, with three outcomes (rally, sell-off, or roughly unchanged) and four outcomes (rally and sell-off broken into two different size classes) to the future study.

We found the simple binary classifier (buy or sell) is substantially improved when we size the trade based on the classifier's conviction level. We size the trade using the Kelly Criterion, a simple two-outcome bet-sizing strategy. Assuming expected gains and losses are approximately equal magnitude (roughly accurate for a simple duration trade), the optimal size to maximize expected returns is $S = 2P - 1$ where P is the probability you made the right choice (S will be between 0 and 1, since P is necessarily above 50%).

We also explore other features of the strategies' behavior (volatility, average returns, and the fraction of days you buy or sell). We find certain ML techniques can converge to multiple, qualitatively distinct solutions of comparable performance, depending on the choice of the technique's so-called 'hyperparameters'.

1.3.2.2. *Testing Method.* The most commonly used method for testing is the k-fold cross-validation. To the extent the markets tomorrow look and perform similar to the markets today, randomly placing adjacent days in the training and testing samples will cause the information to leak between the two sets. Even if the markets exhibit no intrinsic autocorrelation, our P/L series has strong autocorrelation. The entire holding period returns on adjacent days are built from many of the same daily returns. Recall we discussed the testing method in 1.2.3. We applied the sliding window in our framework. Therefore, our testing set always consists of a contiguous block of days that occurs entirely after an expanding or rolling set of training days. We also remove the first n days from the testing set, where n is our trade's holding period since the training space's final points are built from daily returns on those days.

1.3.2.3. *Performance Measurement.* Our metric for success is the risk-adjusted returns, typically captured with the Sharpe ratio for linear instruments and the non-parametric Sharpe ratio for nonlinear instruments. We know returns across all financial products exhibit a high degree of kurtosis so that the sample means and standard deviations are noisy and heavily biased by extreme moves. Said another way, the positioning decision on one or two days where the market moved aggressively can exert an outsize influence on these performance metrics. Also, many of these classifiers tend to produce an extensive range of performance levels on high dimensional data with only modest sample size (the case in our framework). In this case, simply cherry-picking a solution with the highest Sharpe ratio from the sliding window cross-validation will lead to a strategy that veers into massive losses on novel data points. This issue is especially acute for financial strategies, given the aforementioned high kurtosis of daily returns.

Therefore, recall that we proposed the SRE performance metric (Sharpe Ratio Envelope) in section 1.2.2.6. The method takes into consideration the randomness. We create a threshold from the SRE to compare our classifiers' success - 95th and 99th percentile Sharpe: a threshold Sharpe above which the good signal is implausible to have arisen by random chance.

Besides, we “quarantine” a final contiguous set of days which we set aside from use throughout the cross-validation and hyperparameter selection process. In our case, this was the entirety of 2017 thus far, a 10-month period. After selecting the optimal strategy for each classifier through cross-validation, we test their performance on this quarantined set.

Nevertheless, the interpretation is quite different when we test a couple of hand-selected ‘best’ classifiers from the cross-validation on the ‘quarantined’ 2017 data. We are essentially running a hypothesis test on every single trial. The solutions need only clear the 95% threshold, and this procedure is a generalized approach to the classic t-score hypothesis test.

Beyond, we also eliminate solutions that exhibited red flags. In particular, we remove classifiers with exorbitantly high in sample accuracy and the classifiers that isolated outperformers that did not exist within a cluster of nearby hyperparameters that performed similarly. We are most confident when a classifier is not especially sensitive to hyperparameter’s exact choice or when a classifier showed a clear, intuitive trend in terms of in-/out-of-sample accuracy out-of-sample volatility and Sharpe after dialing up/down the hyperparameters.

1.3.3. Results.

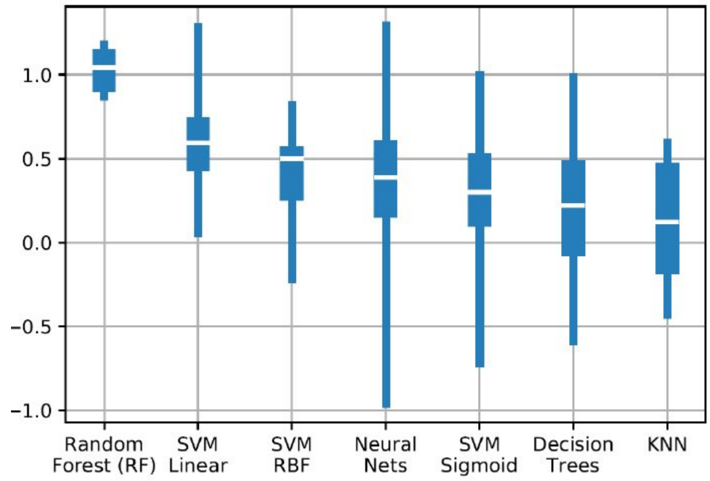
1.3.3.1. *Initial Cross Validation Results.* We performed the sliding window cross-validation procedure beginning with daily trades of 10-year on-the-run Treasury notes held for one week. We trained each classifier using the broad set of input features available for the 2008-16 period, holding off on using all 2017 data we placed in "quarantine." We will discuss the results of other asset classes, hold periods, and training epochs.

Across machine learning models and regardless of parameter choices, we found a substantial improvement in outcomes when we sized trades based on the classifier's conviction level by using the Kelly Criterion. Going forward, we present only results using this sizing approach.

We applied the grid search procedure for each model for hyperparameters' selection. We calibrate the classifiers on a training sub-sample and test its performance on a later, distinct test sample. We then computed summary statistics for that particular run, such as training and test sample hit rates, daily buy/sell fractions, and ultimately Sharpe ratios. As discussed in 1.3.2.3, we removed classifiers with high overfitting likelihood. All surviving classifiers are summarized in 1.3.2, which shows the distribution of test-sample Sharpe ratios broken out by the machine learning models. From this chart, we found several ML techniques, for a subset of hyperparameter choices, produced Sharpe ratios meaningfully higher than uniform buying.

After the cross-validation, it is tempting to simply select a few high performers with a certain set of hyperparameters (with Sharpes above 1.2) in

FIGURE 1.3.2. Sharpe Ratio Distribution

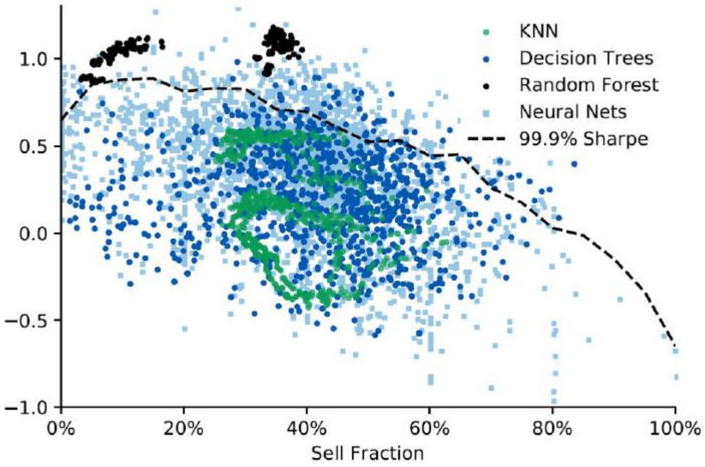


The thin lines show the min/max range, thick lines the inner-quartile range, and white strip is the median outcome; unitless
Classifiers were trained on data beginning in mid-2008 and tested out-of-sample beginning in early 2010. The first 5 days were removed from the testing period, and Sharpe ratios and sell fractions were then computed on the remaining out-of-sample period of roughly 1.5 years. The training window was then expanded four times, until all dates up until 12/30/2016 were tested.

most machine learning applications. This is, after all, out-of-sample performance. However, given our sparse input feature set, the kurtosis of returns, and the relative opaqueness of many of these methods, cherrypicking these predictors is not a responsible approach.

Instead, we looked for structure and consistency in these results. As 1.3.3 illustrates, the various ML classifiers arrived upon an extensive range of trading strategies, measured by the fraction of days they short. While random forests showed a high degree of performance and stability (two tight clumps, both more than unity Sharpe), ANN classifiers' performance is all over the map. The best ANN classifier is shown to be an isolated point on the map, and we had trouble explaining exactly why it out-performed. Neighboring choices of hyperparameters produced a drastically different performance. We argue the performance is highly sensitive to a random seed used to initiate the optimization suggesting the algorithm often did not converge to the globally optimal network. These issues may be caused by our sparse dataset and the complexity of the financial markets. ANN is known to have such issues when applied towards prediction tasks instead of object recognition, where it has enjoyed much success.

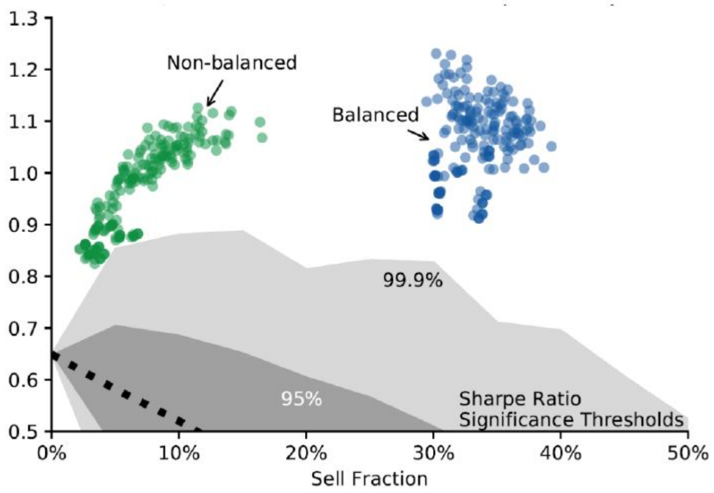
FIGURE 1.3.3. Performance of Classifiers with Different Hyperparameter sets



While random forest performs well almost regardless of hyperparameters' choice, and ANNs' performance seemingly is irrelevant to their selection, the other classical methods' performance sat somewhere in between. Decision trees, KNN, and SVM produced a broad spectrum of Sharpe ratios and sell fractions (1.3.3). In general, low-depth decision trees, also pruned for nodes with too few samples points, generalized more accurately. KNN with a broad sample of nearby neighbors (k~80 as opposed to, say, 10) performs better out of sample. While these properly tuned classifiers showed some promise, the clear "winner" is the random forest.

1.3.3.2. *Deep Dive into Random Forest Trees.* By every performance metric we calculated, the most consistent results came via the random forest tree, the only “ensemble” method we explored in this preliminary work. As we discussed in the methodology section, a random forest is built from an ensemble of basic decision trees. Each decision tree is trained on a randomly selected sub-sample of the training data and a randomly selected subset of the input features. The final result of the RF is the aggregated votes from all the decision trees in the ensemble. Through all these efforts, the resulting ensemble of trees is typically less susceptible to overfitting.

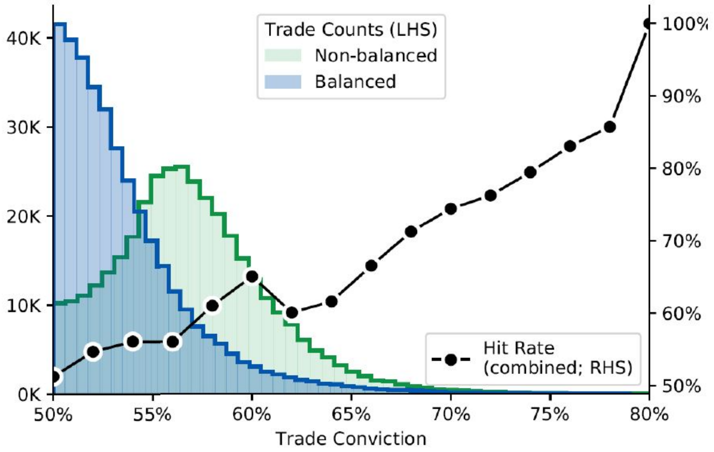
FIGURE 1.3.4. Performance of Random Forest Tree



Since RF performs the best among all the models, we take some time to present its performance and general behavior from the sliding window cross-validated sample data. We present the strategy to trade 10-year Treasuries daily and hold for one-week periods, from 2008-16. After removing the classifiers with obvious overfitting signs, we were left with classifiers consistently producing Sharpe Ratios above unity (1.3.4). The classifiers belong to two general clumps by the sell fraction based on one parameter setting that dictates how important it is for correct classification to be "balanced." In the context of our strategy, where buying the 10Y note every day can generate a decent hit rate, this parameter effectively requires the decision trees to find opportunities to short with as much accuracy as opportunities to buy. We find both classifiers outperform all long by a comfortable margin, despite markedly different aggregate behavior. While for the sake of simplicity, none of the results include transaction costs in 1.3.4. The standard bid/ask on 10-year Treasuries reduces all Sharpe ratios by slightly less than 0.1 (~10%).

As discussed previously, across different models, we found sizing trades by the classifiers' conviction level can substantially increase the Sharpe ratios, despite the long/short positions having the same "hit rate." For RF, we found the balanced and non-balanced clumps had disparate levels of conviction. The balanced classifier behaves in a more 'timid' fashion, with most days having a conviction ~50% (very low conviction) than the non-balanced ones where the conviction rate is ~56% (1.3.5). These two different behaviors do not prevent either from performing well in aggregate. In addition, all the RF classifiers enjoy the realized test-sample hit rates move much higher

FIGURE 1.3.5. Trade Conviction and Hit Ratio



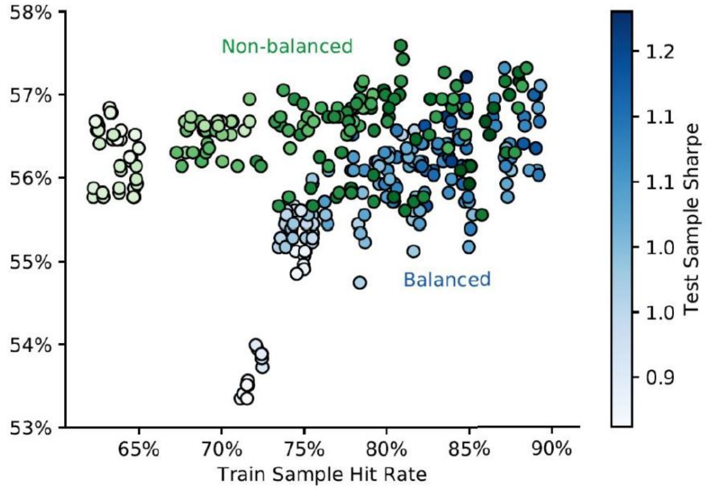
Distribution of days on which the RF classifiers had X% confidence ('conviction') in their decision to go long or short (LHS, count) broken out by weighting strategy, also shown is realized hit-rate vs conviction (RHS; %). Balanced and non-balanced denote whether or not the decision tree seeks to classify buy and short with equal accuracy—balanced classifiers cared more about spotting short opportunity.

when they have high conviction. It suggests using machine learning models for timing execution is a promising application.

We also found that the out-of-sample hit rate of our RF classifiers is not well correlated with the Sharpe ratio (1.3.6). For non-balanced classifiers, the test sample hit rate is at ~57% for all classifiers, regardless of the training set's performance. However, as the hit rates increase in the training set, so does the Sharpe, thanks to the sizing strategy based on how confident the machine learning prediction. The balanced RF also enjoys a similar pattern, though to a lesser extent.

1.3.7 shows that the RF classifiers outperform the all-long strategy when RF models managed to avoid the large draw-downs such as the taper tantrum of 2013 and the widening in rates in late 2015. Within the 2010-16 testing

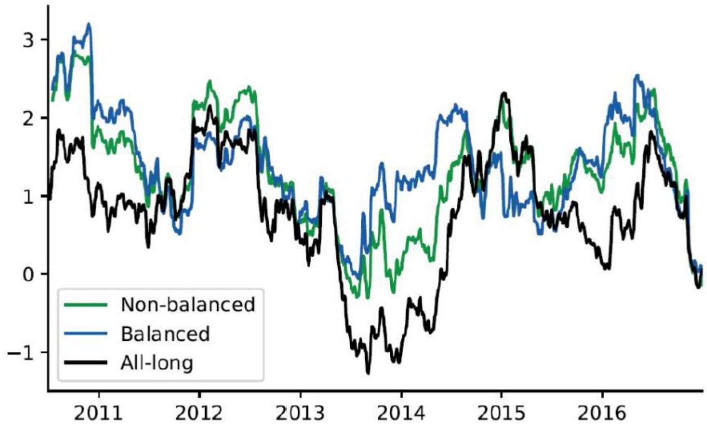
FIGURE 1.3.6. Hit rates in Training Sets and Test Sets



Test (out-of-)sample hit-rate vs train (in-)sample hit rate, broken out by weighting strategy (LHS; %), with shading connoting test sample Sharpe ratio (colors; unitless)
Balanced and non-balanced denote whether or not the algorithm sought to predict rallies and selloffs with equal accuracy—balanced predictors cared more about spotting selloffs

time and across all our RF classifiers, the performance rarely fell below the simple all-long strategy.

FIGURE 1.3.7. RF Classifiers Outperform All-long Strategy



Rolling 1-year Sharpe ratio for daily trades of 10-year Treasuries held for one week from 2010-16, for our RF predictors and uniform, all-long (unitless)

1.3.8 shows the RF classifiers generalizes well to the strategy to trade 10-year Treasury with holding period of one month within longer 2000-16. For both weekly and monthly hold bond trades tested from 2003-16, RF classifiers manage to avoid all the significant drawdowns associated with all-long strategies though they do dip into the red, on a rolling 1-year basis, on occasion.

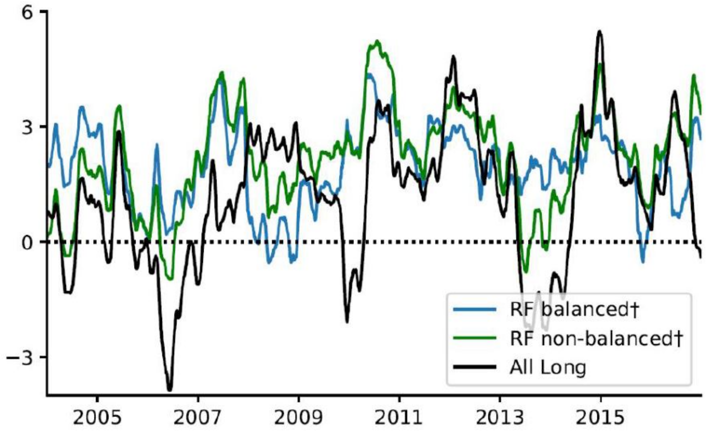
Emboldened by our success in trading duration, we explored how well the machine learning classifiers perform on spreads and volatility. For spreads, we trade the matched-maturity swap spreads to on-the-run 10Y Treasuries. For volatility, we trade 1Mx10Y swaption straddles (no delta hedging) with a holding period of either one week or one month (to maturity). Once more, we found RF classifiers provided the best performance.

However, while the RF classifiers could produce the Sharpe well above both our 95th and 99.9th percentile significance thresholds, 1.3.9 illustrate that the story is more mixed for spreads and volatility.

In the case of spreads, it is difficult to determine the proper ‘benchmark’ strategy. Within the post-crisis time frame, narrowers have been the winning spread trade on average, while the “default” all-long positive carry strategy is to buy spreads (wideners). We use selling spreads as the benchmark strategy for this exercise, while for the final test on the quarantined 2017 data, we use the all-long strategy as the benchmark (buying spreads).

This inconsistent benchmark seems unfair to our classifiers, and we are interested in whether the ML classifiers can beat both benchmarks over the relevant timeframes. 1.3.9 shows that the entirety of Random Forest classifiers mostly fall above 95% of the Sharpe Ratio Envelop, but they are within the maximum Sharpe, which is the 99.9% threshold. This is not the very clean outperformance we observe in the duration trading.

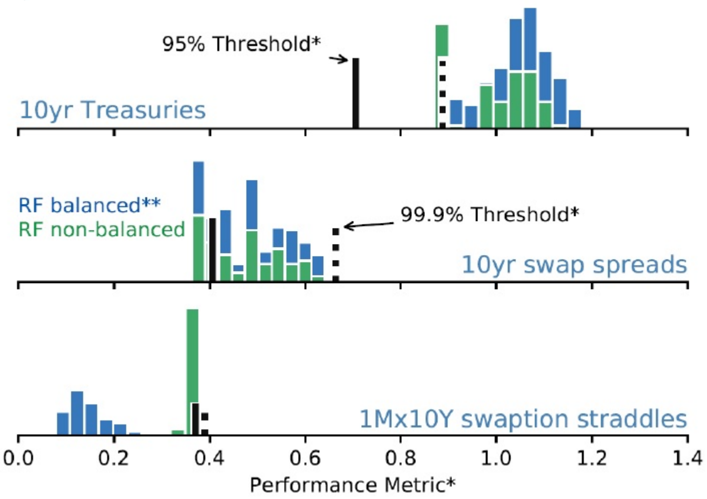
FIGURE 1.3.8. RF Classifiers Outperform All-long Strategy (2000-16)



For volatility, we do not measure the performance by the Sharpe ratio. Instead, we use a non-parametric ratio better suited to the non-linear products with skewed and long-tailed returns. We found our non-balanced classifiers simply converged to the popular short-gamma strategy on trading the swaption straddle, while the balanced classifiers underperform the systematic short-gamma strategy. We note that this underperformance may be due to our categorization/sizing scheme, which was designed intentionaly for a symmetric P/L distribution.

1.3.3.3. *Performance on Quarantined 2017 Data.* The results presented thus far came from an iterative, experimental fit-then-test approach. While the success of the most robust predictors proved insensitive to the details

FIGURE 1.3.9. RF Classifiers Performance on Spread and Volatility



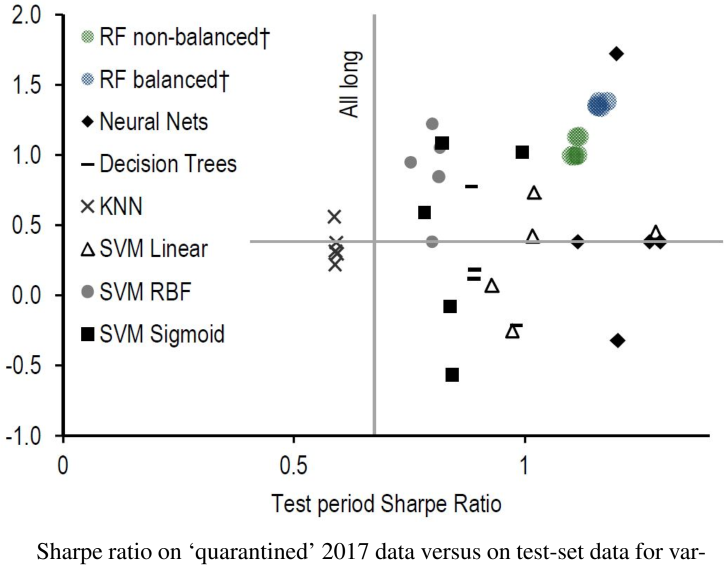
*The nonparameteric ratio for volatility is the average of 1) Non-parametric Sharpe: the average of median versus inter-quartile range; 2) Sterling ratio: median returns versus median losses; and 3) draw-down ratio: returns versus 5th percentile as expected returns versus downside risk.

of implementation, we nonetheless fully admit that during the cross validation process parameters were tuned, design decisions were tweaked, and techniques and strategies were proposed and then abandoned, until out-of-sample performance delivered a respectable risk adjusted return. Along the way, high level information from our test set has inevitably seeped into the fitting and selection process.

Thus from the start we elected to "quarantine" data from 2017, removing it from all analysis until we had selected the top candidate predictors from each ML technique. This final sample serves to validate the predictors. We have a testable hypothesis: that the best predictors, particularly the random forest predictors, will continue to out-perform a daily buying strategy. And we have a framework for judging the statistical significance of the out performance (outlined in the Framework Implementation Section).

Beyond this hypothesis-testing, testing the performance on a quarantined data is particularly important since the financial time series are notoriously nonstationary. That said, the 2017 time frame is not exactly a brand new era of financial markets, and performance therein may not be able to tell how these classifiers will behave through a large and exogenous shock to the rates markets.

FIGURE 1.3.10. RF Classifiers vs All-long Treasury Strategies Quarantined 2017 dataset



Sharpe ratio on 'quarantined' 2017 data versus on test-set data for various ML predictors trained* to trade 10-year Treasury notes daily for 5-day hold periods (unitless)

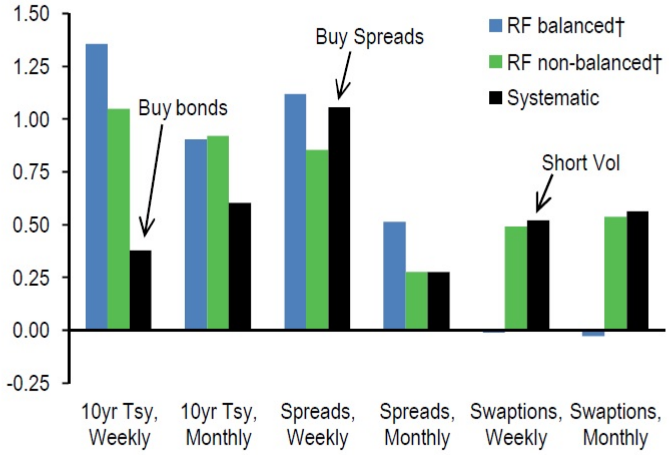
*All predictors cross-validated on 10-year Treasury performance (daily trades, 1-week holding) from mid-2008-16; trades sized with the Kelly Criterion. For each predictor we pre-selected the 5 top candidates from each technique before setting it loose on the 'quarantined' 2017 data. Absolutely no information from 2017 was used while training and vetting these predictors.

† Balanced and non-balanced denote whether or not the algorithm sought to predict rallies and selloffs with equal accuracy – balanced predictors cared more about spotting selloffs.

For trading 10-year Treasuries with a holding period of one week, 1.3.10 shows the Sharpe ratios for the top 5 classifiers we selected from the sliding window cross-validation. Consistent with the results from the cross-validation, RF is still the most consistent outperformer, with balanced RF classifiers producing a Sharpe of 1.35 over 2017, compared to 0.38 for all long. The ensemble method outperforms all classical models, where the top 5 classifiers proved very inconsistent. Only SVM RBF consistently performs at-or-above all-long strategy. The single best classifier is a lone neural network that produced a Sharpe above 1.72, while the four other ANN classifiers do not reproduce such a good Sharpe. Three of them essentially match the all-long strategy, while the fourth one produces a negative Sharpe.

We found the machine learning classifiers on trading spreads and volatility do not perform as well as on trading duration given the cross-validation results. This trend persisted into the quarantine period. 1.3.11 For spreads, it is worth mentioning the RF classifiers outperformed the narrower benchmark strategy in cross-validation; meanwhile, they also almost match the wideners benchmark strategy in the quarantine 2017 time frame. For spreads, it is worth mentioning the RF classifiers outperformed the narrower benchmark strategy in cross-validation; meanwhile, they also almost match the wideners benchmark strategy in the quarantine 2017 time frame.

FIGURE 1.3.11. RF Classifers' Performance on 2017 data Across Different Strategies

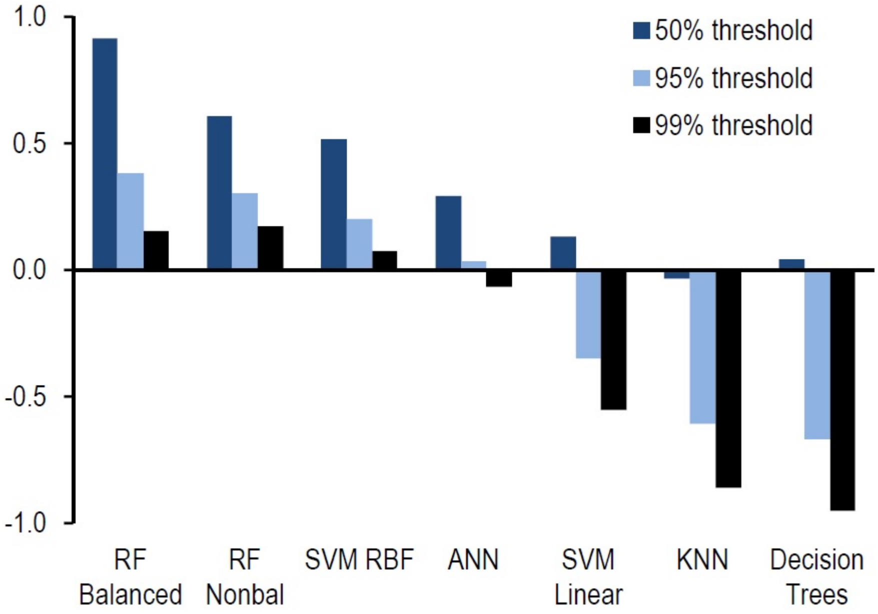


Performance* of RF classifiers** on 'quarantined' 2017 data compared to systematic strategies for 10-year Treasuries, matched-maturity swap spreads, and 1Mx10Y swaptions (unitless)
*For Treasuries and swap spreads, "performance" refers to annualized Sharpe Ratio, for 1Mx10Y swaptions, we instead use the average of non-parametric Sharpe, Sterling and drawdown ratios.
†Balanced and non-balanced denote whether or not the RF seeks to predict long and short with equal accuracy – balanced predictors cared more about spotting short opportunity
**All classifiers cross-validated on daily trades from mid-2008-16; trades sized with the Kelly Criterion. For each classifier we pre-selected the 5 top candidates from each model before setting it loose on the 'quarantined' 2017 data.

For volatility, we discussed that implementation (particularly how we size trades based on probabilities) is designed within the mind of a symmetric distribution of risk, which is more appropriate for linear products. Hence, the underperform is anticipated in this case.

Finally, we would like to discuss how significant the out-performance is. For this, we produce the Sharpe Ratio Envelope on the 2017 data and compute the significant threshold. Each classifier's performance is compared against this threshold. 1.3.12 shows this residual Sharpe across different ML models. We find RF once more outperforms, at both the 95th and 99th

FIGURE 1.3.12. Residual Sharpe Ratio on 2017 Data



Residual Sharpe* of various ML predictors on 'quarantined' 2017 data for strategies trading 10-year Treasuries; positive values denote statistically significant out performance (unitless)
*Residual Sharpe: for each predictor, we take its daily trade decisions (trade size and direction) and randomly permute (shuffle) them across all days in the quarantine period, recomputing Sharpe for this randomized set. We repeat the exercise thousands of times, taking the 50th, 95th and 99th percentile outcomes. These Sharpes are then subtracted from the predictor's Sharpe. If this residual Sharpe is positive, we deem it statistically significant to that percentile level. In the context of our quarantine set, this is somewhat analogous to a one-sample hypothesis test.

percentile level, as does SVM RBF. The other methods failed to meet this benchmark.

1.4. Conclusion

We propose a framework to apply machine learning models for trading liquid fixed income products. Based on the prediction from the machine learning classifiers, the trade is executed on a daily basis for liquid, benchmark rates products held for one-week to one-month. Specifically, we explore the k-nearest neighbors (KNN), decision trees, and support vector machines (SVM); the 'ensemble' technique random forest; and artificial neural networks (ANN). We propose a sliding window cross-validation procedure instead of using the most commonly used k-fold cross-validation for our application. Also, we propose a Sharpe Ratio Envelope approach to measure the performance of different models with different hyperparameter sets. Among the various models, the only ensemble method, random forest (RF), is the most successful. It consistently produces Sharpe ratios greater than unity while trading 10 year Treasuries, both in the cross-validation and 'quarantined' 2017 data. This performance comes for both 1-week and 1-month hold periods, using a broad set of input features available from 2008 and a more limited set available since 2000. We found sizing daily trades by the ML classifiers' conviction level via the Kelly Criterion can substantially enhance the Sharpe Ratios across timeframes, asset classes, hold periods, and ML classifiers. Further, the classifiers' perceived conviction is correlated well with its realized hit rate. This suggests a promising application of ML models in the fixed income for timing the market. We found trading 10-year matched-maturity swap spreads, RF classifiers incrementally outperform the narrowers benchmark on the cross-validation dataset, and then the same classifiers roughly match the wideners benchmark on 2017 data. Although the consistent performance is promising, we are not that confident on trading spread using ML models. For trading 1Mx10Y ATMF swaption straddles, RF classifiers failed to outperform the popular systematically short-gamma strategy.

1.5. Future Research

Firstly, we use the ML classifiers to predict a binary classification in either one week or one month, which are up move and down move. A multinomial classification setup could be explored in the future. For example, a three classes classification, which consists of move, no move, and down move, could potentially enhance the strategy's performance. Investors are usually also interested in capturing the big market moves, so big up moves, such as 10% up move, could be appropriately defined. Hence, the five classification setups could be explored in future study too. Secondly, we briefly discussed the concept drift used in online learning literature in section 1.2. Our framework is an offline learning framework. The online learning methodologies could be explored since these algorithms are designed to conduct the learning in a streaming data environment. Especially as the market regime changes, training data selection and parameter calibration frequency are necessary setups for online learning. Thirdly, we found the random forest tree, an ensemble method, is the most promising one among all the classifiers. Different ensemble methods could be explored in the future. Primarily, we could explore setting up an ensemble, which consists of different classifiers, such as decision trees, ANNs, and SVMs, to conduct the prediction job. Fourthly, we found the sizing strategy can substantially enhance the performance. However, we designed the sizing strategy with symmetric return distribution in mind. It may be why the volatility trading using ML classifiers does not perform well in our framework. We could explore a different sizing strategy for the nonlinear products, and we may find that with a proper sizing strategy, the ML classifiers can also outperform the traditional sell-gamma strategy for trading volatility. Finally, we could explore intraday trading using our framework since the intraday return predictability is well documented in the literature.

CHAPTER 2

A Trend Following Framework: Theory and Application

2.1. Introduction

A price momentum effect means that the price of an asset has trends instead of being randomly distributed. A trending price means that an asset that recently appreciated is more likely to continue moving higher and vice versa. The existence of Momentum effects is an asset pricing anomaly and would violate the efficient markets hypothesis and enable Momentum traders to consistently outperform the broad market. Trend-following (also referred to in academic circles as time-series momentum[1]) has actively been on investors' radar for the last few decades. Managed Futures Hedge Funds, also known as 'CTAs' (short for Commodity Trading Advisors), often trade futures contracts based on Trend-Following techniques.[2]The longevity of the strategy and the appealing performance in the midst of the crisis of 2008 have helped to propel the assets managed by CTAs to more than $348bln[3].

A host of academic literature provides potential explanations for the Momentum effect. They include inefficiencies in investor behavior (see

[1] See Tobias, M., Ooi, Y. and Pedersen, L. "Time Series Momentum". Journal of Financial Economics 104 (2012): 228–250

[2] Earlier CTAs mainly traded Currencies and Commodity futures. Nowadays, CTAs trade a broad range of financial instruments (cash, forwards, futures, options, etc) across different asset classes and geographies.

[3] Estimate by BarclayHedge in the 3rd Quarter of 2017

[133][134]), macroeconomic supply and demand frictions, positive feedback loops between risk assets and economic growth, and even in the market microstructure. Momentum in equities is a well documented in academic literature. [1] is one of the early literature that documents equity Momentum. [73][74] provides a detailed review of commodity Momentum strategies and the CTA practitioners applied momentum strategies over the past 30 years. [72] document the momentum in fixed income. The Momentum effect in Currency Markets was demonstrated in the research of [71], [70] and [69]. [68] documented Momentum effects in global equity index, currency, commodity and bond futures markets since the 1970s. Based on extended datasets, [67] validated significant Momentum effects across assets since 1903, while [66] did a similar exercise for Equity index and commodity markets since 1800.

This chapter focuses on a concrete trend-following solution and analyzes its analytical properties alongside its practical implementation. We find the majority of the research on trend-following has been empirical. In our opinion, there has been a relative lack of theoretical research linking the empirically observed characteristics of the strategy to theoretical results with a model framework. To some extent, we try to fill this void with the current paper.

The contribution of the chapter is threefold. First, we propose a trend-following signal based on statistical theory and analytically analyze its properties. We manage to reconcile the theoretical results with the stylized facts, a ‘CTA smile’ (see [67]) and prove the signals based on longerterm horizons to outperform (see [64]). Second, we propose a prototype trend-following framework that is diversified across time-frames and assets and uses a unified approach across assets. Third, we discuss the portfolio and risk management of the trend-following strategy. We illustrate how the risk-budgeting approaches can be applied to the trend-following framework.

We start by presenting a signal that is based on statistical hypothesis testing. We show that under certain assumption and specification, the trend-following signal is also the delta of a straddle. Therefore, we can explicitly link the trend-following fund’s performance and the long straddle positions’ performance [65].

Next, we analyze the profit drivers for the trend-following strategy using our proposed signal. We prove the strategy is profitable whenever the underlying assets have trends in either direction. Hence we demonstrate that the so-called “CTA smile” [67] can be justified within our theoretical model as well. We found that the absolute value of the Sharpe ratio of the underlying asset is important for the profitability of the strategy. Furthermore, we found the signal based on longer bookback periods have better profitability than the signal based on shorter lookback periods.

We explicitly take into account the time series properties and assume the underlying asset return follows a AR(1) process. We analytically demonstrate that the autocorrelation is important for the profitability of signals based on short lookback periods. It is intuitive that positive autocorrelation leads to profits while negative autocorrelation leads to substantial losses. On the other hand, the signals' profitability based on longer lookback periods is unaffected by the time-series properties of the underlying. The impact of transaction costs is also explicitly modeled. Results show that transaction costs increase with the bid-ask spread but decrease with the volatility and the lookback period.

Besides, the correlation between the P&L of the signal based on different lookback periods is derived. We analytically show that the correlation depends on the ratio of the lookback periods, and the correlations' theoretical values closely match the empirical observations. It is demonstrated that the average of the signals across various lookback periods is optimal if an appropriate correlation structure between P&Ls of signals is present. While averaging the signals among different lookback windows has been common, certain conditions have to be present for its optimality.

We propose a prototype trend-following framework that uses a unified methodology across asset and asset classes based on the theoretical results. The solution is diversified across various time-frameworks. The performance prototype trend-following framework is compared to benchmark indices under various fee structure scenarios. The diversification and hedging properties of trend-following with respect to long-only portfolios are also demonstrated in simulations.

We apply a risk budgeting approach to the proposed trend-following framework and compare it with the inverse volatility approach. Last but not least, we discussed different cost control approaches. We tempt to incorporate short-signals that provide quicker reaction at inflection points in a cost-efficient way. We discuss the impact of ‘carry’ and show how our framework allows for incorporating the carrier component in the strategy design.

2.2. Methodology

2.2.1. Trend-Following Signal and Options. A simple and intuitive measure of a trend is the average asset's return over a certain period. If it is positive, we can conclude the asset is trending upwards and vice versa. The greater the average return in absolute value, the higher our conviction for the presence of a trend.

We denote the average return over period T at time t as $R_{\bar{t},T}$ and the estimated volatility as $\hat{\sigma}_t$. Under the assumption that R_t is i.i.d. $N(0,\sigma^2)$, it is well-known that the t-statistic $t_{t,T} = \frac{\sqrt{T}R_{\bar{t},T}}{\hat{\sigma}_t}$ has a Student's t-distribution with $T-1$ degrees of freedom. Later, we relax the assumption and assume R_t follows a AR(1) process. Then, the t-statistics becomes $\frac{\sqrt{T}R_{\bar{t},T}}{\hat{\sigma}_t} \cdot \sqrt{\frac{1+\rho}{1-\rho}}$, where ρ is the autocorrelation coefficient. In the case of daily return data, the absolute value of autocorrelation is small and hence the value of the t-statistics is not significantly impacted. When the sample size increases, the t-distribution converts to the standard normal distribution (it occurs when $T > 30$ as will be the case in most of our subsequent work).

We can easily construct the statistical tests to test whether the average return is greater than zero when the estimate $\bar{R_{t,T}}$ turns out positive:

$$H0 : \mu = 0 \;\; H1 : \mu > 0$$

The decision whether to accept or reject $H0$ at a certain confidence level is based on comparison of the calculated t-value to a critical value depending on the chosen confidence level. Hence, we will reject $H0$ when $1 - N(t_{t,T})$ is below the required confidence level, where N stands for the standard normal cumulative density function. In general, the smaller $1 - N(t_{t,T})$ the higher is our confidence that $\mu > 0$. As $t_{t,T} > 0$, $(1 - N(t_{t,T})) \in [0, \frac{1}{2}]$. In case we want to construct a trend-following signal ranging from 0 to 1, we can show that the linear combination $2 \cdot N(t_{t,T}) - 1$ achieves that goal.

Similarly we consider the case when the estimated average return is negative:

$$H0 : \mu = 0 \;\; H1 : \mu < 0$$

In this case, the smaller is $N(t_{t,T})$, the greater the confidence with which we can reject $H0$. We want to map $N(t_{t,T}) \in [0, \frac{1}{2}]$ to a signal ranging from $[-1, 0]$. Again, the linear transformation that achieves this goal is $2 \cdot N(t_{t,T}) - 1$.

In the end, we can construct our trend-following signal: $2 \cdot N(t_{t,T}) - 1$.

The stylized facts of the trend-following fund's performance profile is the so-called "CTA" smile which has always been thought to resemble the P&L of a straddle. Besides, the trend-following also tends to exhibit positive convexity. For example, [65] used lookback straddles to replicate the track record of actual trend-followers. Below we make an explicit link between our trend-following signal and the typical option strategies.

In the Black-Scholes world, the delta of a straddle is given by $2N(d1_t)-1$. Let's assume that the strike of the option is set to the price T days ago and the maturity of the option is T. Under the assumption of zero interest rate,

$$d1_t = \frac{1}{\sigma\sqrt{T-t}}[ln(\frac{S_t}{K})+(r+\frac{\sigma^2}{2})(T-t)] = \frac{1}{\sigma\sqrt{T}}(ln(\frac{S_t}{S_{t-T}})+\frac{\sigma^2 T}{2})$$

Using the assumptions of the Geometric Brownian Motion and introducing $\varepsilon_t \sim N(0,1)$, we can write

(2.2.1)

$$\begin{aligned} d1_t &= \frac{1}{\sigma\sqrt{T}}[\sum_{s=t-T+1}^{t} ln(\frac{S_s}{S_{s-1}})+\frac{\sigma^2 T}{2}] = \frac{1}{\sigma\sqrt{T}}[\sum_{s=t-T+1}^{t}(\mu+\sigma\varepsilon_s)] \\ &= \frac{1}{\sigma\sqrt{T}}(\sum_{s=t-T+1}^{t} R_s) = \frac{\sqrt{T}\bar{R}_{t,T}}{\sigma} \end{aligned}$$

If we plug in an estimate of the volatility σ, we arrive at $d1_t = t_{t,T}$. Hence, the delta of a straddle with appropriately chosen strike and maturity can also be viewed as a trend-following signal $2N(t_{t,T}) - 1$.

2.2.2. Profit drivers of "delta-straddle' trend-following signals. The profit generation mechanism of trend-following has not been well comprehended beyond the general statement that "trend-following is profitable when there are strong trends." The sections below analyze the interactions between the Sharpe ratio of the asset and the Sharpe ratio of the trend-following and demonstrate that trend-following exhibits a straddle-like P&L profile. We also derive expressions for the expected transaction costs and elaborate on the trade-off implications between having a reactive trend-following system and keeping a lid on the costs.

2.2.2.1. *Gross P&L.* In the Appendix, we derived the relationship between the gross P&L of the trend-following strategy and its lookback window, underlying assets' Sharpe ratio and the time series' autocorrelation. We deviate from the assumptions of the Black-Scholes world and assume the underlying assets follows AR(1) process.

PROPOSITION 1. *Assume that underlying asset returns follow an AR(1) model: $R_t = a + \rho R_{t-1} + \varepsilon_t$ where $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ and $|\rho| < 1$. It follows that $R_t \sim N(\frac{a}{1-\rho}, \frac{\sigma_\varepsilon}{1-\rho^2}) \sim N(\mu, \sigma^2)$. The expected gross P&L for a "straddle" signal based on a lookback T is:*

$$(2.2.2)\quad E(PL_{t,T}) = 2\frac{\mu}{\sigma}\Phi(\frac{\mu_{d_1}}{\sqrt{1+\sigma_{d_1}^2}}) - \frac{\mu}{\sigma} + 2\phi\frac{\sigma_{d_1}}{\sqrt{1+\sigma_{d_1}}}f(\frac{\mu_{d_1}}{\sqrt{1+\sigma_{d_1}^2}})$$

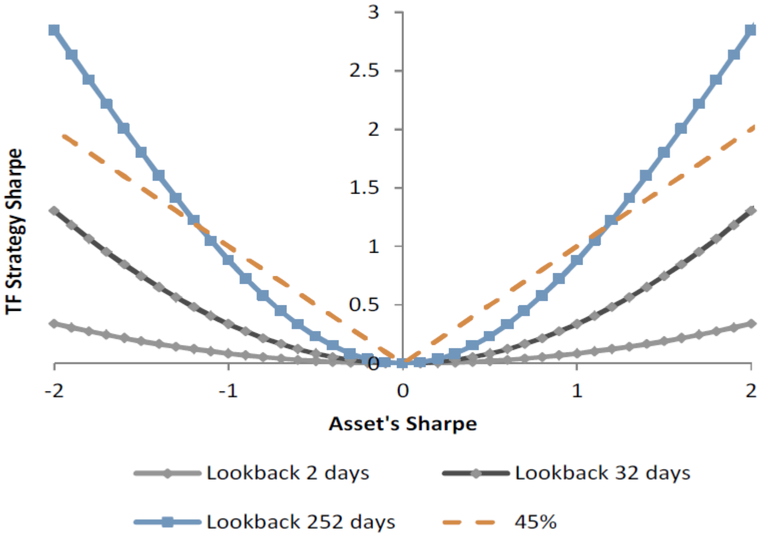
where μ_{d_1},σ_{d_1}*and* ϕ *are functions of* μ,σ^2,ρ *and T,* Φ *stands for the standard normal c.d.f and f for the standard normal p.d.f.*

In case $\rho = 0$ *(B-S assumption), it follows that* $E(PL_{t,T}) = \frac{\mu}{\sigma}(2\Phi(\frac{\mu}{\sigma}\frac{\sqrt{T}}{\sqrt{2}}) - 1)$.

Similarly, if $\mu = 0$,*we obtain that* $E(PL_{t,T}) = \frac{2\rho(1-\rho^T)}{\sqrt{2\pi}\sqrt{2T(1-\rho)-2\rho(1-\rho^T)}}$

In 2.2.1, we show the profile of the gross P&L for various lookbacks and Sharpe ratios of the underlying asset for the case $\rho = 0$. It is obvious the P&L of the trend-following strategy based on our proposed signal has the typical straddle P&L payoff. Both positive and negative drift (positive and negative values of μ) can produce positive profit. Besides, we shows analytically the profitability of the strategy is linked to the Sharpe ratio of the underlying asset (μ/σ). We can also notice the convexity exists in the strategy. Especially, when the lookback period is relatively large and the Sharpe ratio of the asset is sizable, the Sharpe ratio of the trend-following strategy exceeds the underlying's. This is desirable when the Sharpe ratio of the underlying is sizably negative. A subtle implication of this result is that if the Sharpe ratio of an asset is stable and below 1, an investor might be better off holding the asset rather pursuing a trend-following strategy.

FIGURE 2.2.1. Sharpe ratio of the trend-following strategy v.s. the Sharpe ratio of the underlying ($\rho = 0$)



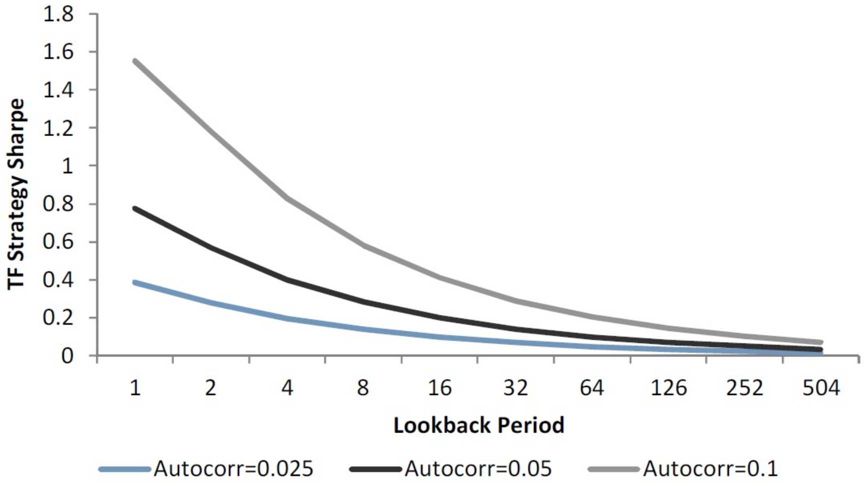
In 2.2.2 and 2.2.3, we plot the Sharpe ratio of the trend-following strategy for various positive and negative values of the autocorrelation when there is no drift ($\mu = 0$). As expected, positive autocorrelation leads to profits for the trendfollowing strategy and vice versa. Besides, there are two important conclusions from the results.

First, the impact of autocorrelation is more pronounced for the profits produced by the signals based on short-term lookback periods. The P&L of the signals based on longer-term lookback periods is expected to be immune to the impact of autocorrelation. Hence, the signals based on the longer-term lookback periods will tend to be pure trend-following play for reasonable values of the autoregressive coefficients. For sufficently large T and realistic values of ρ it follows

$$E(PL_{t+1,T}) = \frac{2\rho(1-\rho^T)}{\sqrt{2\pi}\sqrt{2T(1-\rho)-2\rho(1-\rho^T)}} \sim \frac{2\rho}{\sqrt{2\pi}\sqrt{2T(1-\rho)-2\rho}} \sim 0$$

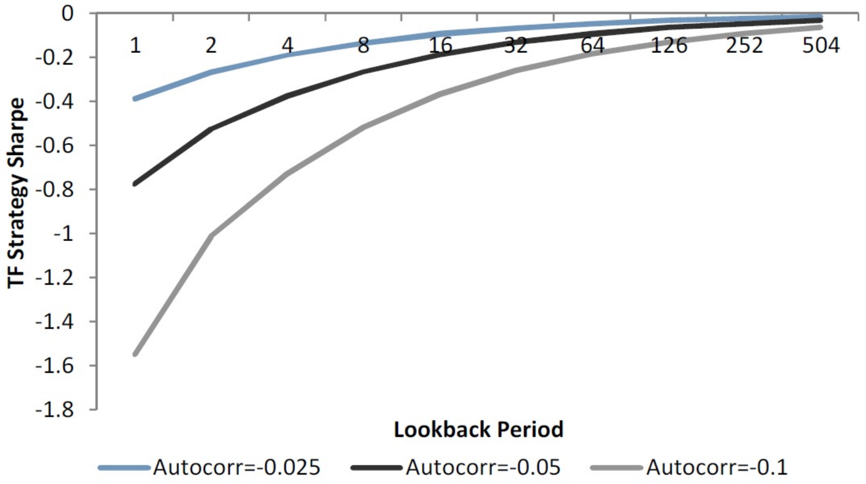
Second, even small values of autocorrelation can lead to a substantial positive or negative P&L when signals are based on short-term lookback periods. For example, when the autocorrelation coefficient is 0.1 a trend-following strategy based on a lookback period of 4 days is expected to produce a Sharpe ratio above 0.8.

FIGURE 2.2.2. TF Sharpe Ratio v.s. Positive Autocorrelation for the underlying AR(1)



2.2.2.2. *Transaction Costs.* To implement every systematic strategy, it is important to have a good understanding of the transaction costs of the strategy. There are two types of transaction costs: running and execution. The running costs are the cost to maintain the position which is linked to the size of the position. The execution costs are the bid-ask spread cost in our case which is linked to the change in the position.

FIGURE 2.2.3. TF Sharpe Ratio v.s. Negative Autocorrelation for the underlying AR(1)



PROPOSITION 2. *Under the assumption that the underlying assets returns follows an AR(1) and denote the unit running cost as RC, the expected running costs of the strategy based on a lookback of T are:*

$$E(RU_{t,T}) = (2\Phi(\mu_{d_1}/\sqrt{\sigma_{d_1}^2+1}) + 2\Phi(-\mu_{d_1}/\sigma_{d_1}) - 4BvN(\mu_{d_1}/\sqrt{\sigma_{d_1}^2+1}, -\mu_{d_1}/\sigma_{d_1}; corr = -\sigma_{d_1}/\sqrt{\sigma_{d_1}^2+1}))RC/\sigma \tag{2.2.3}$$

where BvN(U,W;ρ) stands for the c.d.f of the standard bivariate normal distribution with correlation ρ evaluated at U and W. μ_{d1},$\sigma_{d_1}$and ϕ are functions of μ,σ^2,ρ and T. Φ stands for the standard normal c.d.f.

Under assumption that $\mu = 0$ and $\rho = 0$ (i.e. Gaussian noise), it follows that

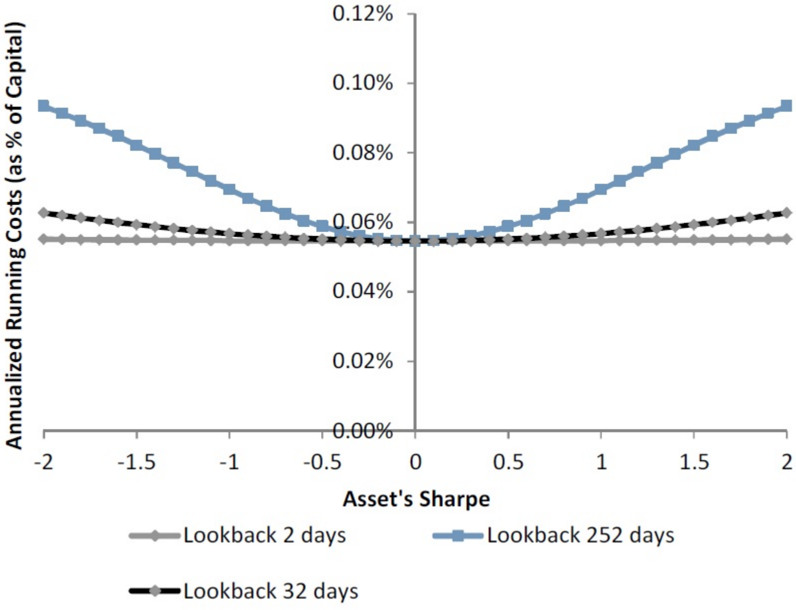
$$(2.2.4) \qquad E(RU_{t,T}) = -2\frac{asin(-\frac{1}{\sqrt{2}})}{\pi}\frac{RC}{\sigma} = \frac{1}{2}\cdot\frac{RC}{\sigma}$$

All other things equal, the running costs are an increasing function of the ratio between the unit running cost and the volatility. In the case when returns are a Gaussian white noise, the running costs are equal to half of that ratio. Furthermore, the graph below show expected annualized running costs as a percentage of employed capital[4]. The running costs are increasing with the lookback period and the absolute value of the Sharpe Ratio of the underlying. Such results are intuitive as higher in magnitude Sharpe ratios generated more significant signals and positions. For the same Sharpe ratio, the signals based on the longer-term periods are more significant than the signals based on the shorter-term periods. In general, the magnitude of the running costs is small and rarely exceeds the underlying asset's running cost.

In the case of pure autoregressive (no drift), the running costs have a flat structure across various lookback periods and autocorrelation coefficients. Note that the coeffient in the bivariate normal distribution $corr = -\sigma_{d_1}/\sqrt{\sigma_{d_1}^2+1}$ is decreasing in ρ and T when $\rho > 0$ and increasing in ρ and T when $\rho < 0$. Therefore, the running costs are increasing in ρ and T when $\rho > 0$ and decreasing in ρ and T when $\rho < 0$. Given the P&L

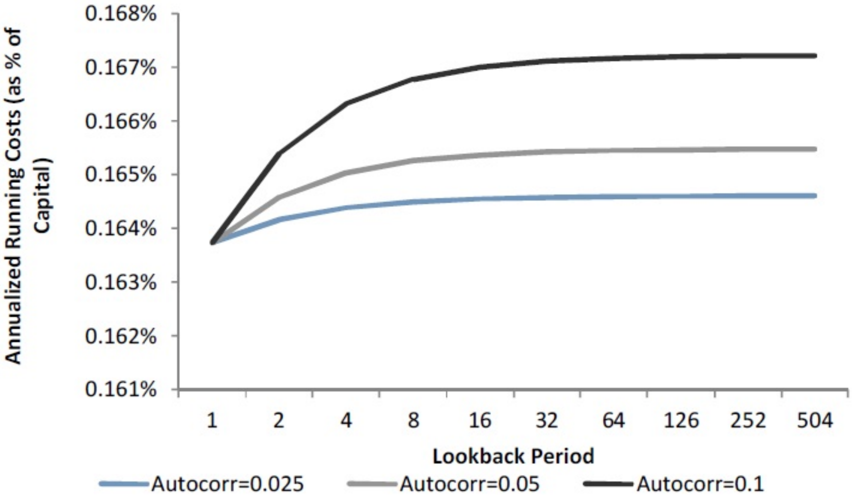
[4] We assume that we target 10% annualized volatility and hence the employed capital is 10×annualized volatility

FIGURE 2.2.4. Annualized Expected Running Costs



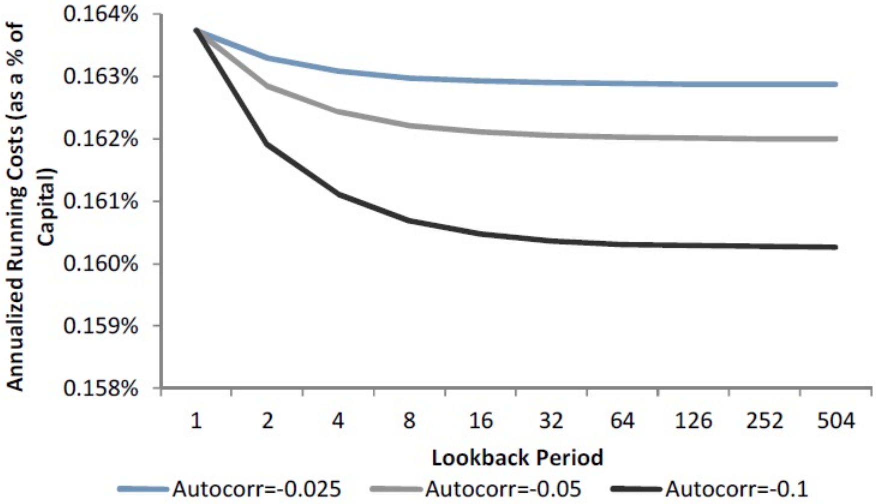
arguments discussed before, the estimated running cost structure supports focusing on shorter term lookback periods when the strategies are designed to benefit from the autocorrelation properties.

FIGURE 2.2.5. Annualized running costs as a % of capital for positive autocorrelation



2.2.2.3. *Execution Cost.*

FIGURE 2.2.6. Annualized running costs as a % of capital for negative autocorrelation



PROPOSITION 3. *Under the assumption that the underlying asset returns follow an AR(1) and denote the unit execution cost as EC, the expected execution costs for a signal based on a lookback of T are:*

$$E(XC_{t,T}) = 4\cdot(\Phi(\frac{-\mu_{d_1}}{\sqrt{\sigma_{d_1}^2+1}}) - BvN(\frac{-\mu_{d_1}}{\sqrt{\sigma_{d_1}^2+1}}, \frac{-\mu_{d_1}}{\sqrt{\sigma_{d_1}^2+1}}; corr = 1-\frac{(\frac{1-\rho^T}{T})}{1+\sigma_{d_1}^2}))\frac{EC}{\sigma} \tag{2.2.5}$$

where BvN(U,W;ρ) stands for the c.d.f. of the standard bivariate normal distribution with correlation ρ evaluated at U and W. $\mu_{d_1}, \sigma_{d_1}$and ϕ are functions of μ, σ^2, ρ and T. Φ stands for the standard normal c.d.f.

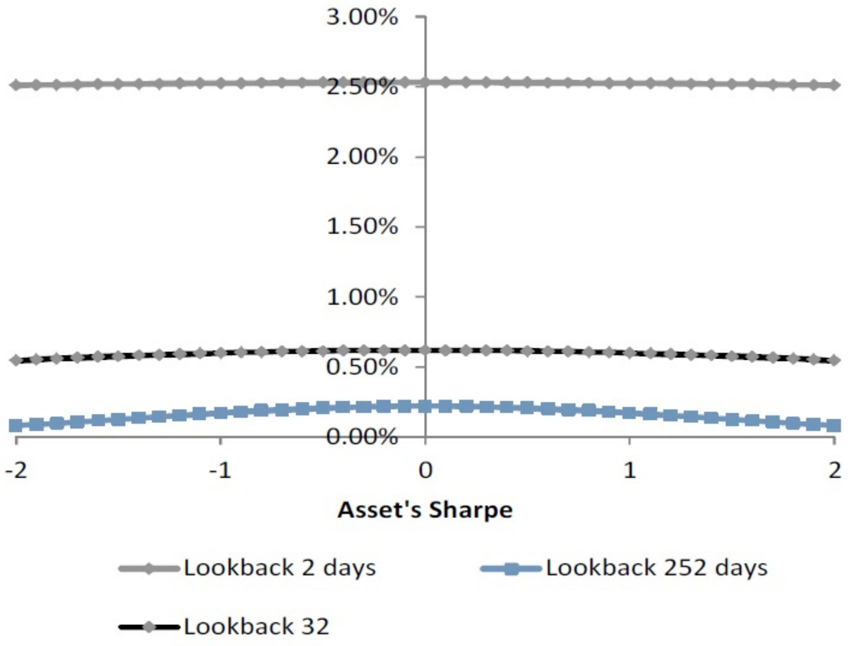
Under the simplified assumptions that $\mu = 0$ and $\rho = 0$, it follows that

(2.2.6)
$$E(XC_{t,T}) = \frac{2EC}{\pi\sigma} acos(1 - \frac{1}{2T})$$

Similarly to running costs, the ratio between the unit execution cost and volatility is important. Under the assumption that returns are a Gaussian noise, the execution costs are a decreasing function of the lookback period.

When there is no autocorrelation, the execution costs are also decreasing with the lookback period. The impact of the Sharpe ratio of the underlying is more pronounced for longer-term lookback periods, and the execution costs decrease with the absolute value of the Sharpe ratio.

FIGURE 2.2.7. Annualized Expected Execution Costs



In the case of pure autoregressive (no drift), the execution costs are dependent on the lookback period. The longer lookback periods produce smaller execution fees. The execution costs are decreasing in ρ when $\rho > 0$ and increasing in ρ when $\rho < 0$. The impact of the autocorrelation is much more muted in comparison to the period.

FIGURE 2.2.8. Annualized Expected Execution Costs

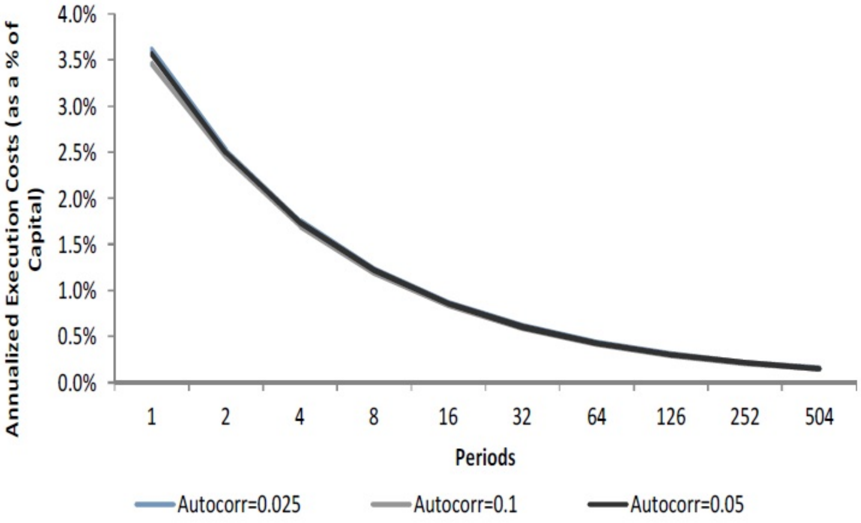
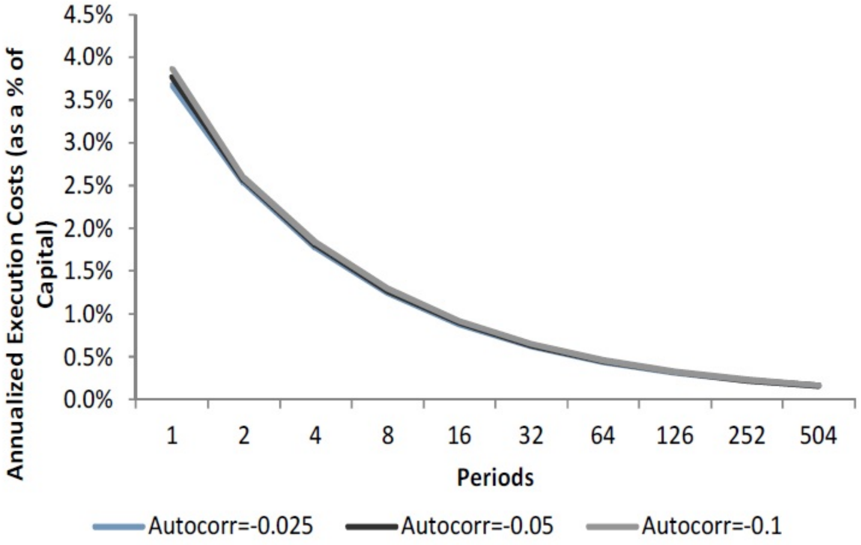
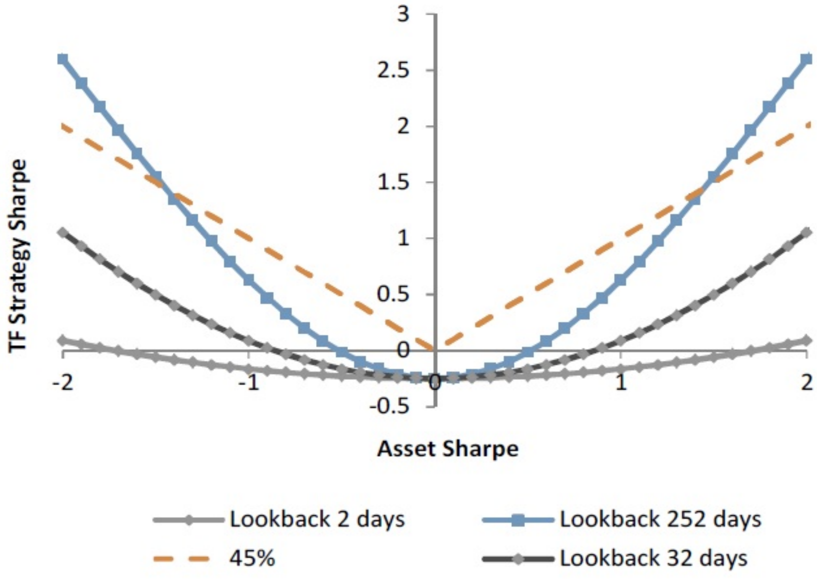


FIGURE 2.2.9. Annualized Expected Execution Costs



2.2.2.4. *Net P&L.* Knowing the the expected P&L and transaction costs, we can derive the net P&L. We start by assuming the autocorrelation coefficient is zero.

FIGURE 2.2.10. TF Sharpe ratio v.s. Sharpe ratio of the underlying ($\rho = 0$)



In 2.2.10 we plot the Sharpe ratio based on the net P&L of the trend-following strategy for various lookback periods. We use the transaction cost structure of S&P and assume daily volatility of 1%. It is evident that signals based on short-term lookbacks can only be profitable if the asset's Sharpe ratio is quite sizable in either direction. For example, for a signal based on 2 days we need a Sharpe ratio above 2 and below -2 to assure the strategy's profitability. For a signal based on 32 days, the Sharpe ratio should be above 1 or below -1. Even a signal based on a 1 year lookback period requires the Sharpe ratio's absolute value to be bigger than 0.5 so that profitability is assured.

While such threshold seems quite high, at first sight, the table below shows that empirically such absolute values of the Sharpe ratios are normal. It is evident that, empirically, the Sharpe ratios' absolute values have sufficient magnitude to render the trend-following strategy profitable. Hence, the persistence of the underlying assets' Sharpe ratio is very important for the profitability of the trend-following strategy using the proposed signal. Besides, the trends should last a sufficiently long time so that the signals can capture them.

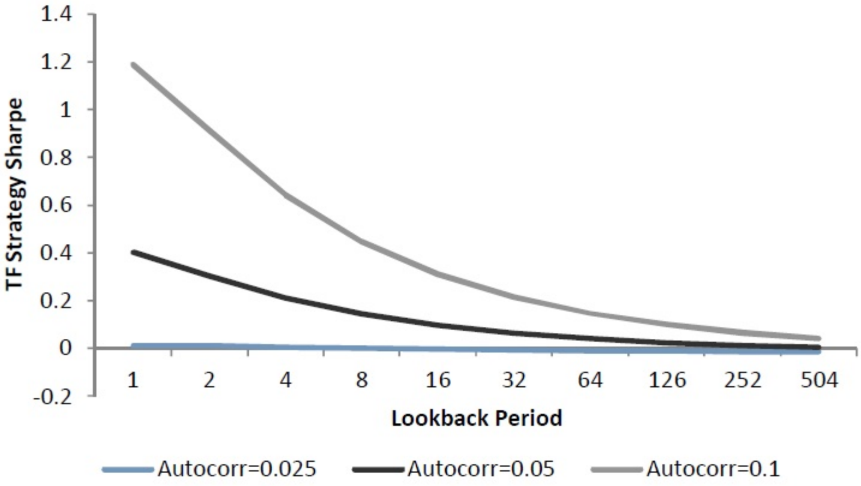
TABLE 1. Average absolute value of the Sharpe ratio over various timeframes

Asset Class	Data Size (in Days)							
	4	**8**	**16**	**32**	**64**	**126**	**252**	**504**
Equities	8.4	5.3	3.2	2.2	1.7	1.3	0.8	0.7
FX	8.6	5.1	3.1	2.4	1.6	1.1	0.7	0.7
Commodities	8.3	5.5	3.1	2.2	1.7	1.2	0.6	0.5
Rates	8.2	5.4	3.3	2.1	1.4	1.5	1.1	0.8

Furthermore, we expect the Sharpe ratio of the trend-following strategy to be below the asset's Sharpe ratio's absolute value. A bigger positive or negative Sharpe ratio of the underlying and long-term lookback period are both necessary for the TF Sharpe ratio to exceed the absolute value of the Sharpe of the underlying asset. For example, we need the Sharpe ratio of the underlying to be bigger in absolute value than 1.5 so that the trend-following is more profitable than either holding or shorting the asset.

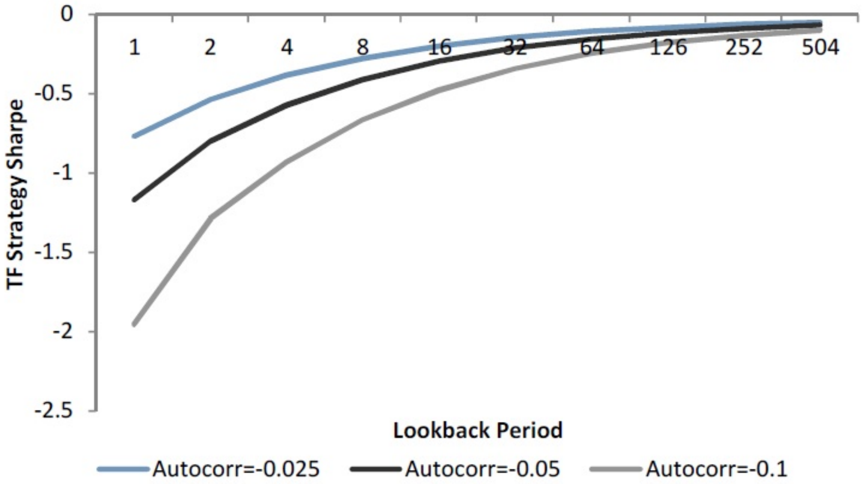
If the asset's drift is stable, it is much more profitable and cost-efficient to use signals based on longer lookback periods. For example, if we expect equities to have a positive drift due to the equity risk premia, it is preferable to use signals with longer lookback periods. The shorter term lookback periods becomes attractive in two scenarios. Firstly, the duration of the trend might be smaller than a long lookback period. For example, if the trend changes direction every 3 months, using a signal based on a 6 months lookback periods will be useless. Secondly, during market sell-offs, signals based on shorter lookback periods are more reactive.

FIGURE 2.2.11. TF Sharpe ratio after accounting for transaction costs v.s. positive autocorrelation



2.2.12 and 2.2.11 shows that when we apply signals based on the short-term lookback, positive autocorrelations leads profitability even after accounting for costs. But negative autocorrelation of the same magnitude can lead to two times higher losses when quicker signals are employed. The profitability of signals based on long term periods remains relatively immune to the autocorrelation. The Sharpe ratio of the P&L generated by a signal based on a 1 year lookback is 0.06 when the autocorrelation is 0.1 and -0.13 when autocorrelation is -0.1.

FIGURE 2.2.12. TF Sharpe ratio after accounting for transaction costs v.s. negative autocorrelation



Below we have shown the average autocorrelation coefficients through time for various asset classes and the average value across all asset classes. While most of the autocorrelation values were positive at the beginning of our sample, they have gradually turned negative. The autocorrelation dynamics is important for the profitability of the strategy using the short-term lookback period.

2.2.2.5. *Lookback period selection.* In empirical work, the selection of the size of the lookback window is always tricky. On the one hand, a long lookback window is needed to produce reliable estimates. On the other hand, too long a window may not be reactive to recent market development.

The selection of the lookback window size is documented in a host of literature. The selection is often based on backtested performance. On one hand, a single window size is selected. [133] document that 12-month horizon generates the largest Sharpe ratio for trend-following strategies across

FIGURE 2.2.13. Average autocorrelation coefficient per asset class

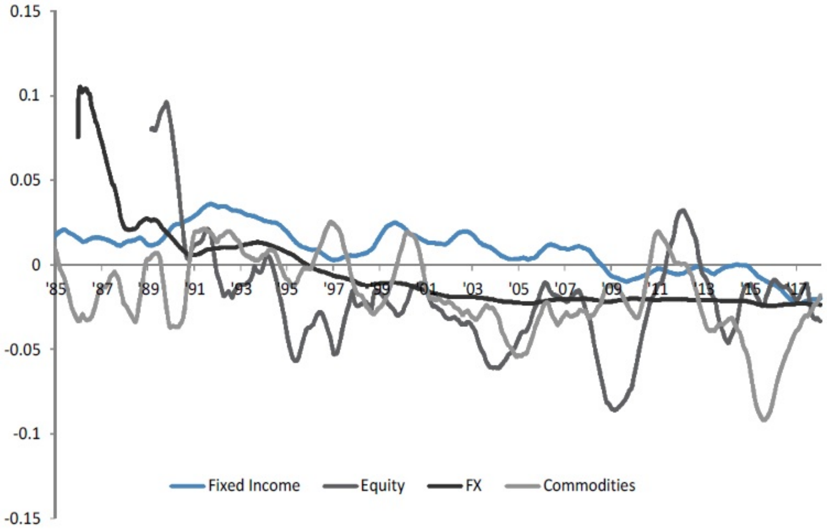
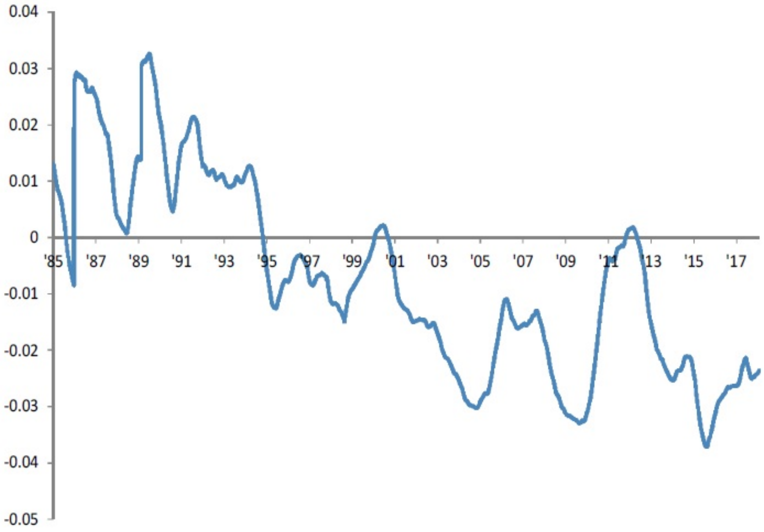


FIGURE 2.2.14. Average autocorrelation coefficient across all asset classes



each asset class. [72] uses the past 12-month cumulative raw return on the asset skipping the most recent month's return. [68] also focused on the 12 month time-series momentum strategy with a 1 month holding period. [66] make use of a 5 month lookback period.

On the other hand, several lookback periods are combined to achieve diversification. [67] use an equally weighted combination of 1-month, 3-month and 12-month time series momentum strategies. [64] construct an aggregated signal based on 3 EWMA Crossovers.

In the following, we discuss the optimal way to select the lookback periods by explicitly considering the properties of the signal. The correlation between the P&L generated by signals based on various lookback periods is implemented in our analysis. The relevant derivations can be found in the Appendix.

PROPOSITION 4. *The correlation between the P&L produced by our proposed signals based on lookback periods T_1 and T_2 $(T_1 < T_2)$ is given by the formula below:*

$$\rho = 6 \cdot asin(0.5 \cdot \sqrt{\frac{T_1}{T_2}})/\pi \tag{2.2.7}$$

The main caveat of the result is that it is the ratio between the lookback periods $\frac{T_1}{T_2}$ is important rather than their difference $(T_1 - T_2)$. For example, if we plug in $\frac{T_2}{T_1}$ = 2 then $\rho = 0.69$.

In **??** and **??** we show theoretical and empirical correlations for selected lookback periods. Though the theoretical correlation is derived under the simplified assumption, the difference between the theoretical and the empirical values are small enough to neglect. The biggest average absolute difference is at 0.04.

We can estimate Equal Risk Contribution (ERC) weights based on the correlation matrix. The optimal weights are pretty close to equal, meaning that averaging the signals is close to an optimal solution.

TABLE 2. ERC weights

Period	1	2	4	8	16	32	64	126	252	504
Weight	0.120	0.103	0.095	0.092	0.090	0.090	0.092	0.095	0.103	0.120

2.3. Application: A Robust Trend-Following Prototype

2.3.1. Data and Transaction Costs. We collect liquid futures across equities, currencies, commodities, and fixed income to backtest our trend-following prototype. More details are documented in Appendix B. We allocate equally across the asset classes at the beginning and replace such allocation with risk budgeting allocation in the next section. Furthermore, every asset receives a risk weight that reflects its liquidity relative to the rest of the assets in the same asset class. For futures, we use the daily volumes from the relevant exchanges to measure the liquidity, while for currencies, we use the Bank for International Settlements transactions volume data.

As a robustness check, we also analyzed the performance of alternative datasets. We did not find substantial performance differences, so we do not present the alternative dataset results. The first alternative dataset shares the same asset universe, but the weights are equally distributed among the same asset class assets. The second alternative dataset has a larger data universe, and the weights are also based on liquidity within the same asset class. More details can be found in the subsequent sections.

We assume a conservative cost structure. The table in Appendix outlines the average execution and running costs per asset class in various periods. We assume the transaction costs were on average four times higher than the current levels and 1.5 times higher between 1993-2002. These adjustments are in accordance to [62]. [5].

2.3.2. Trend-Following Prototype. We apply the derived theoretical correlation of the signals with different lookback periods in the previous section and design our benchmark prototype to use a signal that averages our proposed signals with lookback periods of 32 days, 64 days, 126 days, 252 days and 504 days. The implications of the earlier results justify the equally weighting scheme among signals with different lookback periods is optimal.

We employ standard portfolio and risk management techniques in our benchmark prototype. Every asset's position is proportional to the signal and the risk weight and inversely proportional to its volatility. [6]. The portfolio is dynamically risk-managed on an expanding window basis. We also

[5] The transaction costs are 6 times higher at the beginning and gradually decrease to 2 times higher at the end of 1992. From 1993 to 2003 the transaction costs gradually decrease from 2 times higher to the current levels.

[6] The smoothing parameter used is 0.94 (half-life of approx. 11 days).

target the annualized volatility of 10% for our portfolio for leveraging purposes.

We also apply a floor on the adjustment of the position. If the absolute value of the position's adjustment is below the floor, the position will not be adjusted. The floor corresponds to a change in the signal of 0.25. Such a transaction cost mitigating approach has become a standard cost management technique. Later in the chapter, we demonstrate the results for various values of the floor parameter.

2.3.3. Backtest Performance. 3 shows the cumulative performance of the benchmark prototype in various asset classes and the combined portfolio. Commodities have historically had the most appealing trend-following track-record (which the CTA industry originated). Equities have historically been the most challenging for the trend-following.

TABLE 3. Performance Statistics by Asset Class

	Commodities	**Equities**	**Rates**	**FX**	**Combination**
Annualized Return	6.82%	3.39%	6.23%	5.74%	9.27%
Annualized Volatility	9.47%	9.73%	9.39%	8.77%	9.04%
Sharpe	0.72	0.35	0.66	0.65	1.03
Max Drawdown	-20.93%	-23.45%	-21.38%	-16.37%	-13.60%

4 shows there exist substantial diversification benefits due to the very low average correlation between the trend-following strategies in different asset classes. 3 demonstrates that the Sharpe ratio of the combined portfolio is more than 40% greater than the Sharpe ratio of the best performing asset class – commodities. The combined portfolio's drawdown is also well-controlled and stands at less than 1.5 times the annualized volatility.

TABLE 4. Correlation Matrix among the P&L in various asset classes

	Equities	FX	Commodities	Rates
Equities	1.00			
FX	0.06	1.00		
Commodities	0.03	0.15	1.00	
Rates	0.09	0.11	0.04	1.00

We have already discussed the diversification benefits among the various lookback periods within our theoretical framework. 5 shows the backtest results are in line with our theoretical framework. Although the combined portfolio does not substantially improve the Sharpe ratio compared with the best performing lookback period (1 year), the drawdown is improved by more than 5%. The empirical results also follow theoretical results suggesting that longer-term lookback periods can do better as the Sharpe ratio's threshold value ensures profitability is lower and the overall expected transaction costs are lower. The additional pre-requisite for appealing performance is to have sufficient stability in the trends to be captured by the signals based on longer-term lookback windows.

TABLE 5. Performance Statistics for Signlas Based on various Lookback periods

	32 days	64 days	126 days	252 days	504 days	Combined
Annualized Return	4.67%	5.55%	6.70%	8.87%	7.28%	9.27%
Annualized Volatility	9.23%	9.23%	9.28%	9.18%	9.15%	9.04%
Sharpe	0.51	0.60	0.72	0.97	0.80	1.03
Max Drawdown	-26.67%	-23.80%	-17.77%	-19.05%	-22.10%	-13.60%

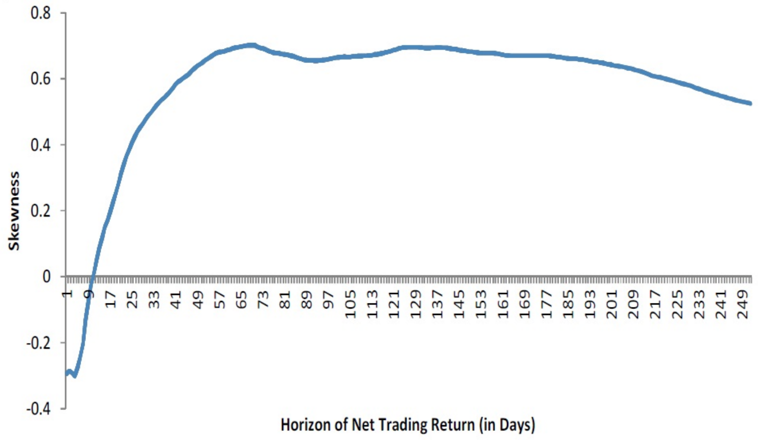
6 shows the strategy would have returned 85bp per month on average. It is interesting to note the positive skewness of the strategy. While the

worst monthly return has been -6.38%, the maximum return would have been 9.63%. Literature often discuss one of the main merits of the trend-following strategies is the positive skewness. It is known that many typical risk-premia strategies have negative skewnes such as the short volatility strategies and carry strategies. [66] documents a strong relationship between the negative skewness of the risk-premia strategies and the expected return. Since trend-following exhibits both positive expected return and positive skewness, Leperiere consider the trend-following strategy is a market anomaly rather than a risk-premia strategy. [60] analyzed the skewness of various non-linear strategies. They argue that even when there are no-trends, the trend-following's cumulated return over a certain period also exhibits skewness. They also show that the skewness peaks at a certain period. 2.3.1 shows our strategy's skewness over various periods and we can indeed notice such a result. However, the transaction costs and negative autocorrelation can lead to negative skewness over very short periods in reality.

TABLE 6. Performance statistics for the monthly strategy returns

Average Monthly Return	Volatility of the Monthly Return	Min	Max	Skewness	Kurtosis
0.85%	2.88%	-6.38%	9.63%	0.33	3.34

FIGURE 2.3.1. Skewness of the Net Return



2.3.2 shows our benchmark strategy also manages to capture most parts of the returns of the big trend-following funds as represented by the NEIXCTAT Index[7]. We simulated the performance of our strategy under various fee assumptions.

The first assumption is a fee structure of 1.5% management fee and 15% performance fee on average. We impose the typical 2 and 20 fee structure at the beginning of the period and adjust it to 1 and 10 fee structure following industry trend on fees. The second assumption is a fee structure of 50bp flat management fee and no performance fee. The correlation between our benchmark strategy with different fee structures and the NEIXCTAT Index is very high, over 0.8. 7 shows our benchmark strategy has a more significant Sharpe ratio even in the aggressive fee scenario and better-controlled drawdown.

[7] The NEIXCTAT Index is designed to track the 10 largest (by AUM) trend following CTAs. The index is equally weighted, and rebalanced and reconstituted annually.

FIGURE 2.3.2. Compounded Performance of our Benchmark Strategy v.s. NEIXCTAT Index

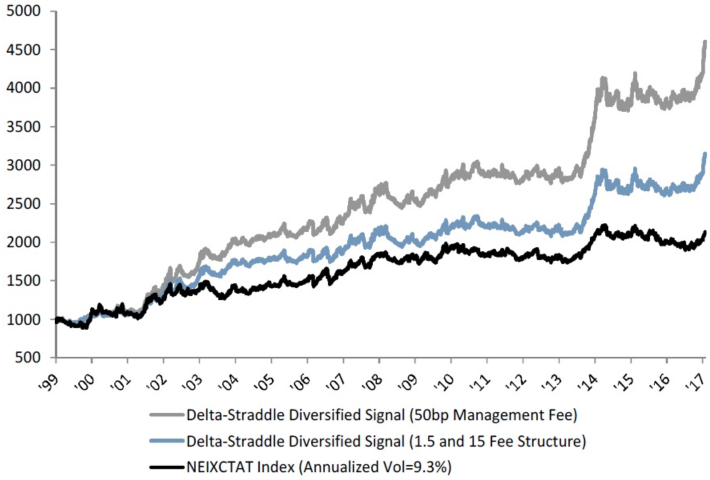


TABLE 7. Comparative Performance Statistics

	Benchmark 1.5 and 15 Fee Structure	Benchmark 50bp Management Fee	NEIXCTAT Index
Annualized Return	6.56%	8.61%	4.48%
Annualized Volatility	9.32%	9.43%	9.32%
Sharpe	0.70	0.91	0.48
Max Drawdown	-12.26%	-11.86%	-15.64%

As a robustness check, we have also checked the Z-score signal and the binary signal. 8 shows our signal outperforms, but all the signals share similar characteristics. The average correlation between the three approaches is 0.97. [59] argues many commonly used trend-following signals are equivalent once the lookback periods as appropriately adjusted. As we discussed earlier, having a robust signal from a theoretical point of view is important, but the choice (or combined) of the lookback windows is not less important.

FIGURE 2.3.3. Cumulative Performance of Various trend-following Signals

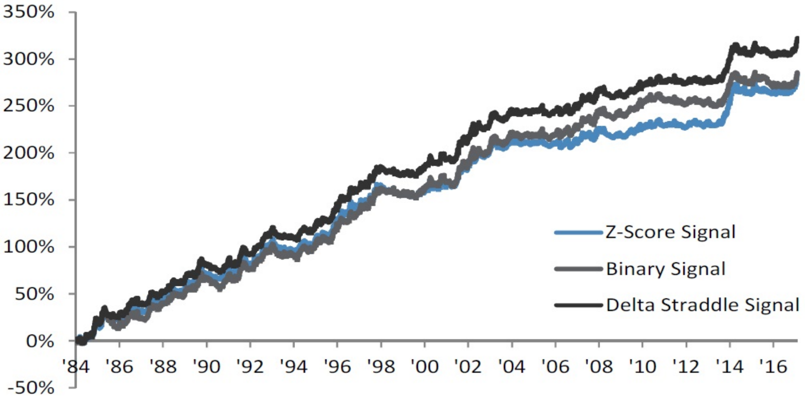


TABLE 8. Performance statistics for various signals

	Z-Score Signal	Binary Signal	Delta Straddle Signal
Annualized Return	8.16%	8.25%	9.27%
Annualized Volatility	8.98%	9.12%	9.04%
Sharpe	0.91	0.9	1.03
Max Drawdown	-14.00%	-17.70%	-13.60%

2.3.4. Trend-following strategies Diversify Long-only Portfolio. In addition to the attractive feature of positive skewness, trend-following strategies can also add substantial diversification benefits for the long-only portfolios. As we discussed, the trend-following strategies exhibit convexity so that its return can more than compensate the loss of the underlying when the sell-off is sizable enough. We observe CTAs substantially outperformed in the GFC. It is well-known that the magnitude of the sell-offs is typically sizable. Hence, the trend-following strategies become appealing to hedge the long-only portfolio.

To verify the hypothesis empirically, we construct portfolios that long positions in the underlying from our asset universe. The portfolios all have an annualized volatility of 10% target and use the same weights scheme as our benchmark strategy. We also construct combined portfolios that invest 50% in the long-only portfolio and 50% in our benchmark strategy. 9 and 10 show the diversification benefits are evident in all asset classes except for fixed income. The market's directionality has led to much overlap between our trend-following positions and those of the long-only portfolio in fixed income.

TABLE 9. Performance Statistics for Fixed income and Equities

	Fixed Income			Equities		
	Trend-Following	Long-Only	Combo	Trend-Following	Long-Only	Combo
Return	6.23%	7.23%	6.78%	3.39%	4.20%	3.86%
Vol	9.39%	9.60%	8.47%	9.73%	9.82%	7.56%
Sharpe	0.66	0.75	0.80	0.35	0.43	0.51
Max DD	-21.38%	-31.34%	-22.52%	-23.45%	-42.90%	-14.44%

TABLE 10. Performance Statistics for Commodities and FX

	Commodities			FX		
	Trend-Following	Long-Only	Combo	Trend-Following	Long-Only	Combo
Return	6.82%	3.47%	5.14%	5.74%	1.66%	3.70%
Vol	9.47%	10.08%	7.71%	8.77%	9.32%	6.92%
Sharpe	0.72	0.34	0.67	0.65	0.18	0.53
Max DD	-20.93%	-40.62%	-17.55%	-16.37%	-53.88%	-22.29%

2.3.4 and 11 show the diversification benefits stand out too when we have the multi-asset portfolios. The Sharpe ratio of the combined portfolio is greater than the trend-following and the drawdown is decreased by more than 60%.

FIGURE 2.3.4. Cumulative Returns of Long-only, Trend-following, and Combo

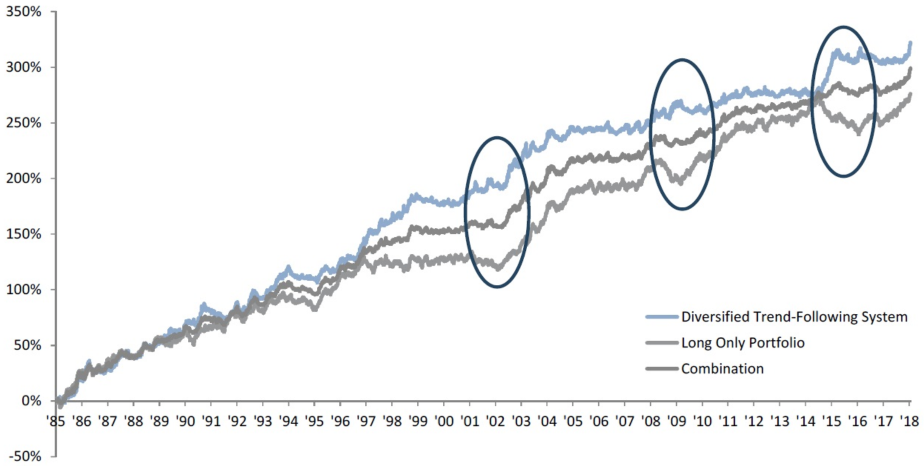


TABLE 11. Performance Statistics of All Asset Classes Portfolio

	All Asset Classes		
	Trend-Following	Long-Only	Combo
Annualized Return	9.27%	7.86%	8.61%
Annualized Volatility	9.04%	9.22%	7.31%
Sharpe	1.03	0.85	1.18
Max DD	-13.60%	-29.15%	-11.20%

2.4. Portfolio Management of The Trend-Following Portfolio

2.4.1. Risk Budgeting. So far, we have been constructing portfolios without considering the correlations among assets. The allocation scheme in our benchmark portfolio – allocating proportionally to the signal-to-vol ratio would have been optimal if the underlying assets were perfectly correlated [58]. Next, we aim additionally to incorporate the correlation structure in our trend-following portfolio.

An asset's risk contribution is set to be proportional to the absolute value of the signal and the risk weight. In the optimization procedure, we also constrain that we have long positions in the assets with positive signals and short positions in the assets with the negative signals. [57] shows such an optimization problem is well-defined and has a unique solution. A similar approach has been adopted by [56] but the paper does not consider signals of different scales, and asset-wise risk weights.

Similarly to the benchmark case, the portfolio volatility is also targeted to be on average 10%. Note that the portfolio's volatility can be below or above the targeted value depending on the signals' strengths at a certain point in time. [8]We estimate the covariance matrix with the last 252 days and the forecasted volatilities.

One of the potential dangers with the risk-budgeting allocation scheme is when short-positions are allowed the scheme may take too much leverage. It is well known that there is much uncertainty in the estimation of the covariance matrix. It is often possible that the two assets' signals are in opposite directions when the assets' returns are positively correlated or vice versa. A potential estimation error in the covariance matrix can lead to excessive leverage due to an ill-envisaged offset in such situations.

To mitigate the potential danger, we introduce an additional parameter that limits the size of the offset allowed by the risk budgeting approach. When the signals' direction does not match what the correlation coefficient

[8] The annualized portfolio volatility =

$$(10\% \times \sum_{t=1}^{n} abs(S_{it}) \times RiskWeight_i)$$

$$\times t/(\sum_{s=1}^{t}\sum_{i=1}^{n} abs(S_{is}) \times RiskWeight_i)$$

, where S_{is} stands for signal for asset i and time s.

implies, we compare the absolute value of the correlation coefficient to a cap parameter. If the absolute value is above the cap, we set the correlation to the cap parameter and adjust for the original sign of the correlation.

TABLE 12. Performance Statistics of Various Correlation Floor Parameter

	RB, No Corr Cap	RB, Corr Cap 0.5	RB, Corr Cap 0.25	RB, Corr Cap 0.001	Benchmark Vol Targeting
Annualized Return	7.66%	9.30%	10.03%	9.76%	9.27%
Annualized Volatility	10.24%	8.07%	9.28%	9.83%	9.04%
Sharpe	0.75	1.15	1.08	0.99	1.03
Maximum Drawdown	-29.2%	-14.3%	-21.6%	-23.3%	-13.60%

12 shows that the optimal choice lies between the extreme values. No cap setup leads to excessive leverage when the signal predictions turn wrong, and extreme capping might reduce the underlying assets diversification benefits.

We note that the performance of the risk budgeting approach depends on a mixture of factors. From a portfolio management point of view, the risk could be more efficiently distributed if the correlation of the assets differ significantly. The assets that have lower correlations to others will receive higher allocations in comparison to our benchmark case. Therefore, to some extent, the performance will depend on the assets' trend-following performance with low correlations.

2.4.2. Costs Control Exploration.

Reduce Turnover Costs. As discussed, our benchmark prototype implements a simple mechanism to control the turnover costs. In simple terms, a trade is executed only when its size is sufficiently large. At every point in time, the change of the position, which is proportional to the ratio of the signal and volatility, should be greater than the ratio of cap parameter relative volatility.

13 shows the results for various values of the floor parameter (up to a maximum value of 0.25). We observe that even such a simple rule can improve performance with higher floor parameter values typically produces a better Sharpe. We select the value of 0.25 as it improves the Sharpe ratio by around 10% and argues that values beyond 0.25 will affect the system's agility.

TABLE 13. Performance Statistics for Various Cap Parameter

	Cap=0.25	Cap=0.2	Floor=0.15	Floor=0.1	Floor=0.05	No Floor
Annualized Return	9.27%	9.11%	8.92%	9.12%	8.56%	8.32%
Annualized Volatility	9.04%	9.07%	9.06%	9.07%	9.05%	9.04%
Sharpe	1.03	1.00	0.99	1.01	0.95	0.92
Maximum Draw-down	-13.60%	-13.45%	-14.17%	-13.54%	-14.00%	-14.22%

Limiting Costs When No Trend. From the expected transaction costs we derived in the previous section, we conclude that costs play a much more significant role when the trend-following strategies are based on short-term lookback periods. 14 empirical illustrates those results.

Assuming there is no cost, the performance of our benchmark strategy has just been marginally improved. In contrast, the Sharpe ratio of a trend-following strategy with signals based 4, 8, and 16 days of lookbacks moves from positive to negative after the costs are considered. Our objective in this subsection is to control the impact of costs while the signals based on short-term lookback periods are implemented when volatility spikes.

TABLE 14. Performance Statistics for the Trend-following Strategy With Costs

	Trend-Following No Costs (32-504d)	Trend-Following With Costs (32-504d)	Trend-Following No Costs (4,8,16d)	Trend-Following With Costs (4,8,16d)
Annualized Return	11.00%	9.27%	7.85%	-3.26%
Annualized Volatility	9.10%	9.04%	9.08%	9.06%
Sharpe	1.21	1.03	0.86	-0.36
Maximum Drawdown	-11.60%	-13.60%	-28.53%	-89.17%

We already discussed in the theoretical part that the most challenging environment for the trend-following strategy is the lack of trends. The problem becomes even more acute if volatility is low and the signals are based on the short-term lookback periods. Below we have plotted the expected transaction costs per year for various asset classes using the volatility for the whole sample and the volatility in 2017.[9]. We observe that the signals based on the short-term periods can lead to excessive losses in the absence of trends. The losses have been accentuated with the drop in volatility in

[9] Targeting annualized volatility on the invested period is 10%.

2017. In terms of costs, FX and Equities are observed to be the most expensive asset classes due to the low volatility in 2017.

FIGURE 2.4.1. Expected Transaction Costs for Equities in the Absence of Trends

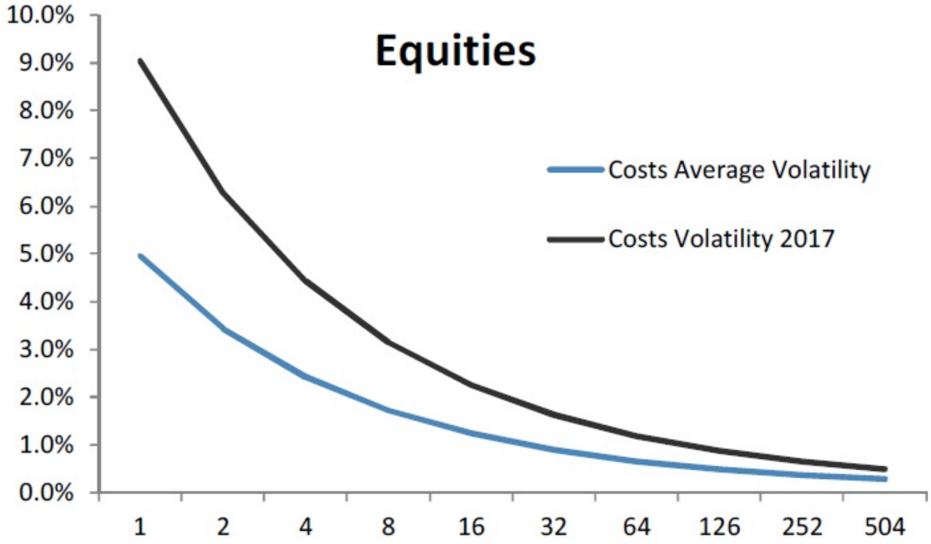


FIGURE 2.4.2. Expected Transaction Costs for Fixed Income in the Absence of Trends

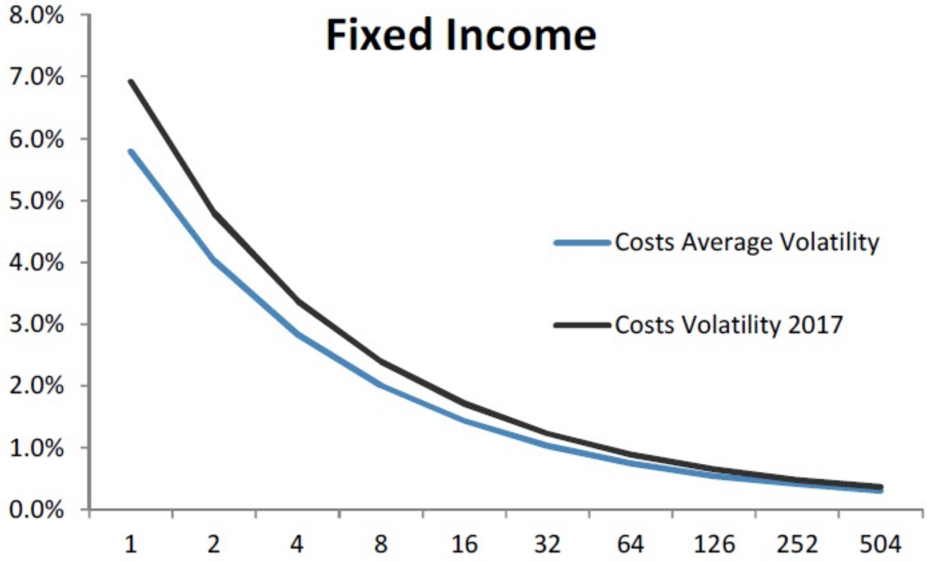


FIGURE 2.4.3. Expected Transaction Costs for FX in the Absence of Trends

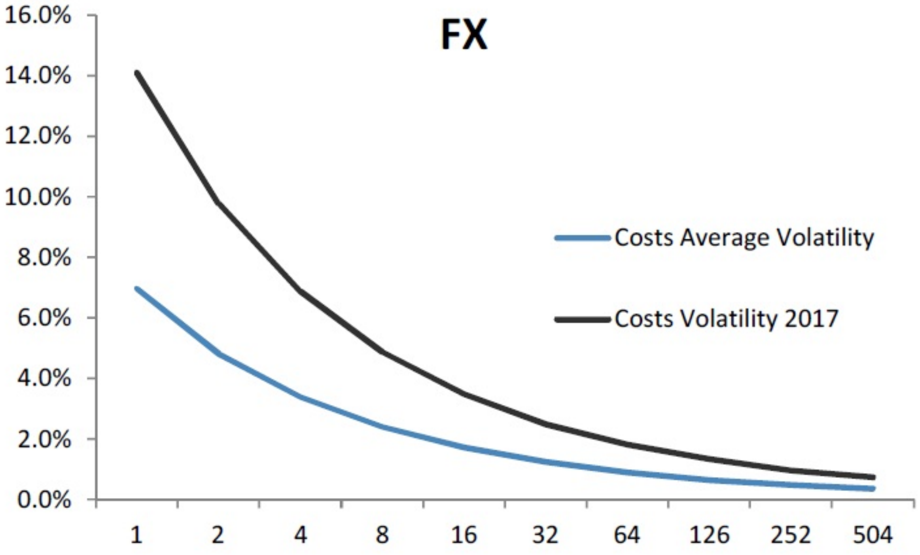
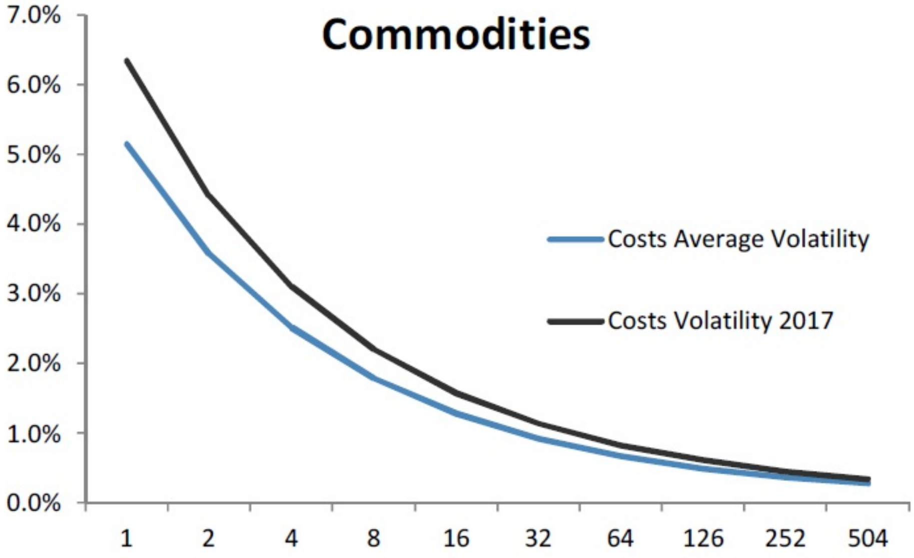


FIGURE 2.4.4. Expected Transaction Costs for Commodities in the Absence of Trends

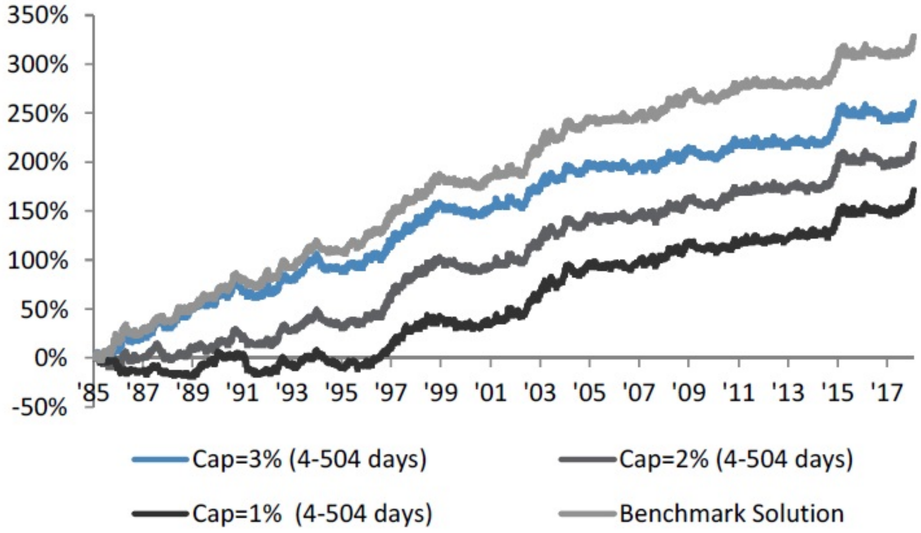


Based on our theoretical framework, we aim to limit the downside in the case of trendless. We add the signals based on lookback periods of 4, 8, and 16 days to our benchmark strategy's signal. Subsequently, we impose a cap on the costs under the assumption that the market is trendless. In such a way, a signal will be excluded if the costs calculated for this signal exceed the cap. The rest of the signals are aggregated for positioning decisions.

We empirically investigate the cap from 1% to 3%, and we have considered the performance from 1985 as well as from 2003 due to different transaction costs in a different era.

2.4.5 shows that the trend-following strategies that include short-term lookback periods with caps underperform the benchmark strategy since 1985. As we have substantially increased the transaction costs, the cap can often switch off some signals. In general, the cap approach comes at a cost. It limits the losses in trendless markets, but potential profits will not materialize if the markets turn out to be strongly trending.

FIGURE 2.4.5. Cumulative Returns since 1985



2.4.6 shows that the trend-following strategies that employ short-term lookback periods with the cap approach can produce a performance similar

FIGURE 2.4.6. Cumulative Returns since 2003

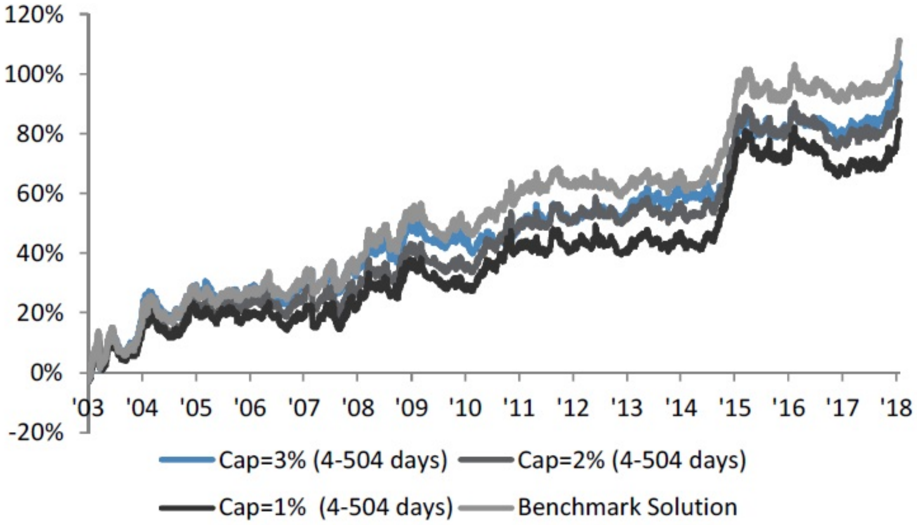


TABLE 15. Performance Statistics for Various Caps with short-term Signal

	Since 1985				Since 2003			
	Cap=1% (4-504d)	Cap=2% (4-504d)	Cap=3% (4-504d)	Prototype	Cap=1% (4-504d)	Cap=2% (4-504d)	Cap=3% (4-504d)	Prototype
Annualized Return	5.00%	6.27%	7.50%	9.27%	6.64%	6.23%	5.40%	7.12%
Annualized Volatility	8.37%	8.58%	8.74%	9.04%	9.04%	8.99%	8.96%	9.34%
Sharpe	0.60	0.73	0.86	1.03	0.73	0.69	0.60	0.76
Maximum Drawdown	-21.04%	-17.64%	-17.33%	-13.60%	-10.71%	-14.08%	-15.44%	-11.85%

to the benchmark strategy. The 1% cap strategy even has a smaller drawdown (even after accounting for costs).15

Such a result is quite appealing as employing signals based on a wide range of lookback periods leads to better diversification and a swifter reaction by the trend-following strategy at inflection points.

Carry Consideration for Short Position. Often the trend-following strategies benefit from the carry present in the underlying futures or FX forwards. For example, the fixed income trend-following strategies have benefited substantially from keeping a long futures position and profiting on slide in the yield curve. Similarly in FX many of the high-yielding currencies have tended to appreciate. While trendfollowing in high carry assets on the long side is to some extent straightforward, having a short position in high carry assets might is challenging as the trend-following gain might not offset the loss due to negative carry. For example, the potential reversal in the bullish fixed income trend might pose challenges in front the trend-following systems.

Our framework is well-suited to take into account carry within trend-following. It can consider the carry as an additional component in the expected P&L calculation and hence link the strength of the signal, the current trend and the size of the carry within a unified framework.

2.4.7 compares the strategy with the carry to the benchmark strategy. We observe that the average correlation is 0.13, which needs to be investigated.

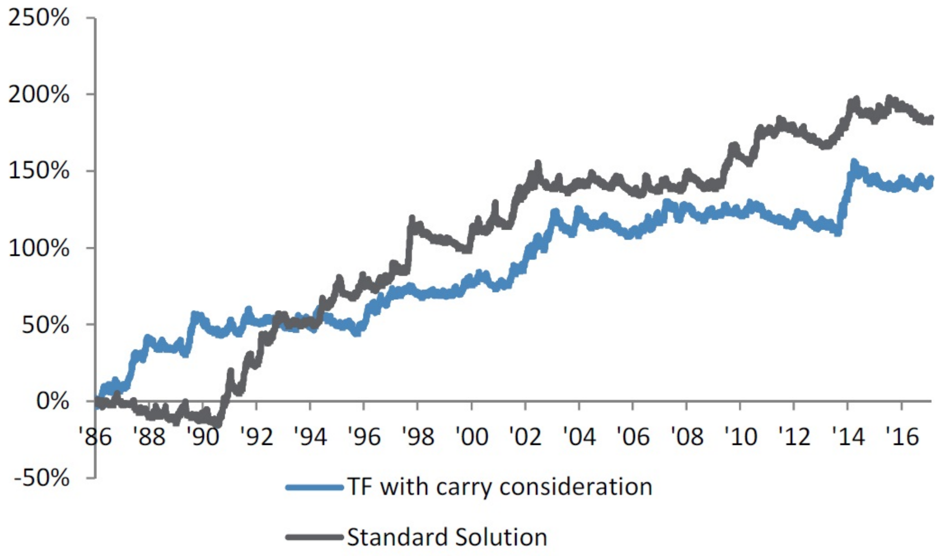
TABLE 16. Performance Characteristics in FX with and without carry considerations

	TF with carry consideration	Standard Solution
Return	4.47%	5.64%
Vol	8.76%	9.31%
Sharpe	0.51	0.61
Max DD	-20.25%	-20.28%

2.5. Conclusion

In this chapter, we proposed a trend-following signal based on statistical theory and analytically analyzed factors that have an impact on the

FIGURE 2.4.7. Cumulative Returns for FX TF strategies with and without carry



performance of the trend-following strategy. Our theoretical model is able to justify the empirically observed performance characteristics of the trend-following funds, such as the so-called "CTA smile" and the convexity. We argue that the underlying assets' Sharpe ratio is critical for the profitability of the trend-following strategy. Higher underlying assets' Sharp ratio leads to better return due to the bigger convexity. Furthermore, we take into account the time series properties of the underlying assets and show the autocorrelation is important for the profitability of the strategy with signals based on shorter lookback periods. The transaction costs' theoretical results imply that the costs increase with the bid-ask spread and decrease with the volatility and the lookback periods. Besides, we derived the correlation between the performance of the signals with different lookback periods and demonstrated the conditions that guarantee the optimality of the approach that averaging the signals among different lookback periods. Leveraging the theoretical results, we proposed a benchmark trend-following framework and backtested its performance. The benchmark strategy is compared to the trend-following index, NEIXCTA Index, under various fee structure assumptions, and correlation is above 80%. We also discussed the diversification and hedging benefits of our trend-following benchmark strategy with respect to the long-only portfolio. The benchmark framework allocates capital proportional to the strength of the signal relative to its volatility. We consider the correlation between different assets by applying the risk budget portfolio allocation scheme to our framework. The backtests results show the risk budget approach could enhance our benchmark strategy's performance. To make our signal more reactive at the inflection points in a cost-efficient way, we discussed methods to control the costs using a signal based on short-term lookback periods. Last but not least, we showed that our model could incorporate the carry effect, and the impact of the negative carry is discussed.

2.6. Future Research

Firstly, our model assumes the underlying assets' return follows an AR(1) process. It is interesting to explore the model results if we assume the underlying assets' return follows a more general ARMA(p,q) process. Secondly, the momentum effect in academics often refers to cross-sectional momentum. The well-known Famma French Carhart four-factor model adds the cross-sectional momentum to Fama French three-factor model. Since our benchmark strategy, which allows short, is diversified across assets and the lookback periods. It is interesting to understand the relationship between the time series momentum and cross-sectional momentum and their contribution to our benchmark strategy. For example, we could decompose our benchmark strategy's excess return to time series momentum effect and cross-sectional momentum effect. Furthermore, we could also further decompose the time series momentum into predictability of the futures' price and the rolling yield from the shape of the future curve. Finally, it is interesting to explore the links between the excess return of a trend-following strategy and different types of investors, such as speculators and hedgers. CFTC records the positions of commercial and non-commercial traders. It could be used to explore such a link.

CHAPTER 3

Markov Modulated Bilateral Gamma Mean Reversion Process

3.1. Introduction

A widely cited cliché of the financial market is that market rides an escalator up and takes an elevator down. Such references may be found on the web, and we cite two examples [31]. Such remarks suggests the assets' prices rise and fall assymmetrically. They invite non-diffusive modeling of asset returns.

There is extensive literature on such non-diffusive models, and we may cite [20, 21, 22, 23, 24, 25, 26, 27] as examples. Many of these models incorporate an exponential tilt to the arrival rate of the jumps that allowing for higher arrival rates for negative relative to positive moves of the same size. Such approach is used to differentiate up and down moves. However, even when there is infinite jumps, the aggregate arrival on the two sides is comparably modeled. An exception to such a formulation is the bilateral gamma model of [27].

There are a series of properties making the Bilateral gamma processes very interesting. Bilateral Gamma distributions are self-decomposable, stable under convolution, and have simple cumulant generating function and characteristic function. The associated Lévy processes are finite-variation processes making infinitely many jumps at each interval with positive length, and all their increments are Bilateral gamma distributed. In particular, one can easily simulate the trajectories of the Bilateral gamma processes. An extension of this model was studied risk neutrally in [28]. It was observed that the risk-neutral markets rose with a greater frequency of smaller jumps, but on the downside, the jumps were fewer and larger.

Besides, there exists other empirical regularities such as volatility clustering and aggregational Gaussianity [1]. These regularities imply that a suitable model for the asset returns must be able to capture the time variation in volatility as well as in other higher moments. In particular, the term structure of volatility and other higher moments imply the model must also allow for time-inhomogeneity. This is important for the derivative pricing and the risk management such as the measurement of value-at-risk.

Bilateral gamma processes and other Lévy processes are time homogeneous. [34] shows that the theoretical behavior of the term structure of their moments does not match empirical observations. For example, the variance theoretically increases with a factor t, skewness decreases with a factor $\sqrt{t}$, and kurtosis decreases with a factor t. However, the empirically observed moments do not exhibit patterns that are even close to these

[1] Volatility clustering is the observed tendency of high volatility periods to be followed by more high volatility periods, while aggregational Gaussianity is the observation that at short time scales, the distribution of returns is highly non-normal with fat tails, while at longer time scales the distribution tends to look more and more normal.

theoretical results. In order to allow for time-inhomogeneity, there has developed an interest in markov modulated processes. [29]look at a two states switching process where the underlying processes are geometric Brownian motions. [34] consider a two-state Markov chain where the underlying state processes are VG processes. [34] suggest that more than two states are supposed to be considered. However, the mathematical approach used in their paper cannot be easily leveraged for more than two states.

In this chapter we extend the work of [34] to more than two states. We propose the N-state Markov modulated Mean Reversion Bilateral gamma process. The statistics switch between drift and compensator pairs. We generalize the underlying process to any pure jump process with triplet $(\mu, 0, \nu)$. There are several advantages of the model we proposed. The first is that our model allows for any number of switching states without adding complexity to the model. The underlying Bilateral gamma processes allow asymmetries on the up and down moves, and the closed-form representation is more concise and transparent.

The outline of the rest of the chapter is as follows. Section 3.2 introduces the definition and properties of the Lévy processes. Section 3.3 introduces the Bilateral distribution, Bilateral gamma processes and their properties. Section 3.4 proposed the Markov modulated Mean Reversion Bilateral gamma model. The characteristic function is derived. Section 3.5 provides the estimation of the parameters using a single time series data. We defer more rigorous econometric analysis of the estimation to later studies. Section 3.6 proposes several potential applications using the model. We defer more rigorous analysis of the trading implications to later studies.

3.2. Lévy Process and Properties

3.2.1. Definition. Lévy processes play an important role in many fields of science, such as in engineering, for study of networks, queues and dams; in economics, for continuous time-series models; and in mathematical finance to price the different sorts of derivative securities. The most famous continuous time model is the Black-Scholes model, which assumes the log return of the underlying asset is normally distributed. Empirically, practioners and academia all found the log returns of most financial assets do not follow a Normal law. They are skewed and have an actual kurtosis higher than that of the Normal distibution. So we need more flexible distributions which generalized Brownian motion.

DEFINITION 5. (Brownian Motion). A stochastic process $B = \{B(t)\}$ is a standard Brownian motion on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ if:

(i): $B(0) = 0$ a.s.

(ii): B has independent increments, i.e. $B(t) - B(s)$ is independent of $\mathscr{F}_s$, for any $0 \leq s < t \leq \mathscr{T}$

(iii): B has stationary increments, i.e. for any $s,t \geq 0$ the distribution of $B(t+s) - B(t)$ doesn't dependent on t

(iiii): For every $t > 0$, B has a Normal(0,t) distribution

Looking at the Brownian motion, we would like to have a similar process, based on a more general distribution than the Normal. However, to define such a stochastic process with independent and stationary increments, the distribution has to be infinitely divisible. Such processes are called Lévy processes, in honour of Paul Lévy, the pioneer of the theory.

DEFINITION 6. (Lévy Process). A càdlàg, adapted, real valued stochastic process $L = \{L(t)_{t \geq 0}\}$ with $L(0) = 0$, a.s. is called a Lévy process if the following conditions are satisfied:

(i): L has independent increments, i.e. $L(t) - L(s)$ is independent of $\mathscr{F}_s$

(ii): L has stationary increments, i.e. for any $s, t \geq 0$ the distribution of $L(t+s) - L(t)$ does not depend on t

(iii): L is stochastically continuous, i.e. for every $t \geq 0$ and $\varepsilon > 0$: $lim_{s \to t} P(|L(t) - L(s)| > \varepsilon) = 0$

The simplest Lévy process is the linear drift, a deterministic process. Brownian motion is the only (non-deterministic) Lévy process with continuous sample paths. Other examples of Lévy processes are the Poisson and compound Poisson processes. Notice that the sum of a linear drift, a Brownian motion and a compound Poisson process is again a Lévy process; it is often called a jump-diffusion process.

3.2.2. Infinitely Divisible Distribution and the Lévy-Khintchine formula. Let X be a real valued random variable, denotes its characteristic function by φ_X and its law by P_X , hence $\varphi_X(u) = \int_{\mathbb{R}} e^{iux} P_X(dx)$.

DEFINITION 7. The law P_X of a random variable X is infinitely divisible, if for all $n \in \mathbb{N}$there exists i.i.d. random variable $X_1^{(1/n)},...,X_n^{(1/n)}$ such that

$$X = X_1^{(1/n)} + ... + X_n^{(1/n)} \tag{3.2.1}$$

Alternatively, we can characterize an infinitely divisible random variable using its characteristic function. The law of a random variable X is infinitely divisible, if for all $n \in \mathbb{N}$,there exisits a random variable $X^{1/n}$, such that

$$\varphi_X(u) = (\varphi_{X^{1/n}}(u))^n \tag{3.2.2}$$

The next theorem provides a complete characterization of random variables with infinitely divisible distributions via their characteristic functions; this is the celebrated Lévy-Khintchine formula.

THEOREM 8. *The law $P_X$of a random variable X is infinitely divisible if and only if there exists a triplet (μ, c, ν), with $\mu \in \mathbb{R}$,$c \in \mathbb{R}_+$and a measure satisfying $\nu(\{0\}) = 0$ and $\int_{\mathbb{R}}(1 \wedge |x|^2)\nu(dx) < \infty$, such taht*

$$\mathbb{E}[e^{iuX}] = exp\left[i\mu u - \frac{u^2 c}{2} + \int_{\mathbb{R}}(e^{iux} - 1 - iux1_{|x|<1})\nu(dx)\right] \tag{3.2.3}$$

Every Lévy process can be associated with the law of an infinitely divisible distribution.

THEOREM 9. *For every Lévy process $L = (L_t)_{0\leq t\leq T}$, we have that*

$$(3.2.4) \quad \mathbb{E}[e^{iuL_t}] = e^{t\psi(u)}$$
$$= exp\left[t(i\mu u - \frac{u^2c}{2} + \int_{\mathbb{R}}(e^{iux} - 1 - iux1_{|x|<1})\nu(dx))\right]$$

where $\psi(u)$ is the characteristic exponent of L_1, a random variable with an infinitely divisible distribution.

3.2.3. Analysis of jumps and Poisson random measures. The jump process $\Delta L = (\Delta L_t)$ associated to the Lévy process L is defined, for each $0 \leq t \leq T$, via

$$\Delta L_t = L_t - L_{t-}$$

where $L_{t-} = lim_{s\uparrow t} L_s$.

A convenient tool for analyzing the jumps of a Lévy process is the random measure of jumps of the process. Consider a set $A \in \mathscr{B}(\mathbb{R}\backslash\{0\})$. Define the random measure of the jumps of the process L by

$$\mu^L(\omega;t,A) = \#\{s \in [0,t] : \Delta L_s(\omega) \in A\}] \tag{3.2.5}$$

$$= \sum_{s \leq t} 1_A(\Delta L_s(\omega))$$

hence, the measure $\mu^L(\omega;t,A)$ counts the jumps of the process L of size in A up to time t. The measure has stationary and independent increments, so $\mu^L(\cdot,A)$ is a Poisson process and μ^L is a Poisson random measure. The intensity of this Poisson process is $\nu(A) = E[\mu^L(1,A)]$.

DEFINITION 10. The measure ν defined by

$$\nu(A) = E[\mu^L(1,A)] = E[\sum_{s \leq 1} 1_A(\Delta L_s(\omega))]$$

is the Lévy measure of the Lévy process L.

Now, we can define an integral with respect to the Poisson random measure.

THEOREM 11. *Consider a set $A \in \mathscr{B}(\mathbb{R}\backslash\{0\})$ and a function $f{:}\mathbb{R} \to \mathbb{R}$, Borel measurable and finite on A.*

A. The process $\int_0^t \int_A f(x)\mu^L(ds,dx)_{0\leq t \leq T}$ is a compound Poisson process with characteristic function

$$E[exp(iu\int_0^t \int_A f(x)\mu^L(ds,dx))] = exp(t\int_A (e^{iuf(x)} - 1)\nu(dx)) \tag{3.2.6}$$

B. If $f \in L^1(A)$, then

$$E[\int_0^t \int_A f(x)\mu^L(ds,dx)] = t\int_A f(x)\nu(dx) \tag{3.2.7}$$

3.2.4. The Lévy-Ito Decomposition.

THEOREM 12. *Consider a triplet (μ, c, ν) where $\mu \in \mathbb{R}, c \in \mathbb{R}_+$ and a measure satisfying $\nu(\{0\}) = 0$ and $\int_{\mathbb{R}}(1 \wedge |x|^2)\nu(dx) < \infty$. Then, there exists a probability space $(\Omega, \mathscr{F}, P)$ on which four independent Lévy processes exist, $L^{(1)}, L^{(2)}, L^{(3)}$ and $L^{(4)}$, where $L^{(1)}$ is a constant drift, $L^{(2)}$ is a Brownian motion, $L^{(3)}$ is a compound Poisson process and $L^{(4)}$ is a square integrable (pure jump) martingale with an a.s. countable number of jumps of magnitude less than 1 on each finite time interval. Taking $L = L^{(1)} + L^{(2)} + L^{(3)} + L^{(4)}$, we have that there exists a probability space on which a Lévy process L with characteristic exponent*

$$\psi(u) = i\mu u - \frac{u^2 c}{2} + \int_{\mathbb{R}} (e^{iux} - 1 - iux1_{|x|<1})\nu(dx) \tag{3.2.8}$$

as follows

(3.2.9)

$$L_t = \mu t + \sqrt{c}W_t + \int_0^t \int_{|x|\geq 1} x\mu^L(dx,ds) + \int_0^t \int_{|x|<1} x(\mu^L - \nu^L)(ds,dx)$$

where $\nu^L(dx,ds) = \nu(dx)ds$.

PROPOSITION 13. *Let L be a Lévy process with triplet* (μ, c, ν).

A. If $\nu(\mathbb{R}) < \infty$*, then almost all paths of L have a finite number of jumps on every compact interval. In that case, the Lévy process has finite activity.*

B. If $\nu(\mathbb{R}) = \infty$*, then almost all paths of L have an infinite number of jumps on every compact interval. In that case, the Lévy process has infinite activity.*

Whether a Lévy process has finite variation or not also depends on the Lévy measure (and on the presence or absence of a Brownian part).

PROPOSITION 14. *Let L be a Lévy process with triplet* (μ, c, ν).

A. If $c = 0$ *and* $\int_{|x|\leq 1} |x|\nu(dx) < \infty$*, then almost all paths of L have finite variation.*

B. If $c \neq 0$ *or* $\int_{|x|\leq 1} |x|\nu(dx) = \infty$*, then almost all paths of L have infinite variation.*

If the Lévy process has finite variation, the characteristic exponent is

$$\psi(u) = i\mu u + \int\int_{-\infty}^{+\infty} (exp(iux) - 1)\nu(dx,ds) \tag{3.2.10}$$

Basically, the sum of all jumps smaller than some $\varepsilon > 0$ doesn't converge. However, the sum of jumps compensated by their mean does converge. This peculiarity leads to the necessity of the compensator term $iux1_{|x|<1}$.

3.2.5. Examples of Lévy Process.

EXAMPLE 15. The Compound Poisson Process

Suppose $N = N_t, t \geq 0$ is Poisson process with intensity $\lambda > 0$ and that Z_i is an i.i.d. sequence of random variables independent of N and following a law L, with characteristic function $\phi_X(u)$. Then we could say that

$$X_t = \sum_{k=1}^{N_t} Z_i, \ \ t \geq 0 \tag{3.2.11}$$

is a compound Poisson process. The value of the process at time t, is the sum of N_t random numbers with law L. The ordinary Poisson process is the case where $Z_i$with the law L degenerate at the point 1.

We could write the distribution function of the law L as:

$$P(Z_i \in A) = \frac{\nu(A)}{\lambda} \tag{3.2.12}$$

where $\nu(R) = \lambda < \infty$. Then the characteristic function of X_t is given by

$$E[exp(iuX_t)] = exp(t\int_{-\infty}^{+\infty}(e^{iux}-1)\nu(dx)) \tag{3.2.13}$$

$$= exp(t\lambda(\phi_Z(u)-1))$$

From this we can easily obtain the Lévy triplet:

$$\left[\int_{-1}^{+1} x\nu(dx), 0, \nu(dx)\right] \tag{3.2.14}$$

Next, we look at the Gamma process

EXAMPLE 16. The Gamma process

The density function of the Gamma distribution $\Gamma(\alpha,\lambda)$ with parameters $\alpha > 0$ and $\lambda > 0$ is given by

$$f_{Gamma}(x;\alpha,\lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}exp(-x\lambda),\ x > 0 \tag{3.2.15}$$

The density function clearly has a semi-heavy (right) tail. The characteristic function is given by

$$\phi_{Gamma}(u;\alpha,\lambda) = (1-iu/\lambda)^{-\alpha} \tag{3.2.16}$$

Clearly, this characteristic function is infinitely divisable. The Gamma process

$$X_t^{Gamma} = \{X_t^{Gamma},\, t \geq 0\} \tag{3.2.17}$$

with parameters $\alpha, \lambda > 0$ is defined as the stochastic process which starts at zero and has stationary and i.i.d. Gamma distributed increments. More precisely, time enters in the first parameter: X_t^{Gamma} follows a Gamma$(\alpha t, \lambda)$ distribution.

The Lévy triplet of the Gamma process is given by

$$[\alpha(1-e^{-\lambda})/\lambda, 0, \alpha e^{-\lambda x} x^{-1} 1_{x>0} dx] \tag{3.2.18}$$

The properties of the Gamma(α, λ) could be derived from the characteristic function. $E(X) = \frac{\alpha}{\lambda}$, $Var(X) = \frac{\alpha}{\lambda^2}$, $Skewness(X) = 2\alpha^{-1/2}$, $Kurtosis(X) = 3(1+2\alpha^{-1})$. Note that we also have the following scaling property. If X is Gamma(α, λ), then for $c > 0$,cX is Gamma$(\alpha, \lambda/c)$. In addition, the sum of two independent gamma process is also a gamma process: $X_1^{Gamma}(t; \alpha_1, \lambda) + X_2^{Gamma}(t; \alpha_2, \lambda) \sim X(t; \alpha_1 + \alpha_2, \lambda)$. The gamma process could also be parameterized in terms of the mean $\mu = \alpha/\lambda$ and variance $\nu = \alpha/\lambda^2$ of the increase per unit time, which is equivalent to $\alpha = \mu^2/\nu$ and $\lambda = \mu/\nu$. In this parametrization, cX is Gamma$(c\mu, c^2\nu)$ and $X_1(t;\mu_1,\nu) + X_2(t;\mu_2,\nu) \sim X(t;\mu_1+\mu_2,\nu+\nu)$.

Next, we look at the CGMY Process

EXAMPLE 17. The CGMY Process

The CGMY distribution and the associated Lévy process was introduced by [50] to model to logreturns of financial assets. We introduce the four parameter distribution using the notation in [50].

The CGMY distribution with C, G, M and Y is defined through its cumulant function:

$$\psi_{CGMY}(u) = C\Gamma(-Y)[(M-iu)^Y + (G+iu)^Y - G^Y] \tag{3.2.19}$$

CGMY distribution is infinitely divisible and therefore a Lévy process $L(t)$, where $L(1)$ is CGMY distributed, can be constructed. The Lévy measure for this process is absolutely continuous with respect to the Lebesgue measure

$$\nu_{CGMY}(dj) = \begin{cases} C|j|^{-1-Y} exp(-G|j|)dj, & j<0 \\ C|j|^{-1-Y} exp(-M|j|)dj, & j>0 \end{cases} \tag{3.2.20}$$

Here, $Y < 2$ in order to have a Lévy measure which integrates $|j|^2$ around zero.

From the cumulant function, we could obtain the moment generating function $\phi(u) = \psi(-iu)$:

$$\phi(u) = C\Gamma(-Y)[(M-u)^Y + (G+u)^Y - G^Y] \tag{3.2.21}$$

The characteristic function

(3.2.22)

$$\varphi_{CGMY}(u;C,G,M,Y) = exp(C\Gamma(-Y)((M-iu)^Y - M^Y + (G+iu)^Y - G^Y))$$

The CGMY distribution is infinitely divisible and we can define the CGMY Lévy process $X^{(CGMY)} = X_t^{(CGMY)}, t \geq 0$, as the process which starts at zero and has independent and stationary distributed increments and in which the increment s follows a $CGMY(sC,G,M,Y)$ distribution; in other words, the characteristic function of X_t^{CGMY} is given by

$$\begin{aligned} E[e^{iuX_t^{CGMY}}] &= \phi_{CGMY}(u;tC,G,M,Y) \\ &= (\phi(u;C,G,M,Y))^t \\ &= exp(C\Gamma(-Y)((M-iu)^Y - M^Y + (G+iu)^Y - G^Y)) \qquad (3.2.23) \end{aligned}$$

The first parameter of the Lévy triplet:

$$\mu_{CGMY} = C\left(\int_0^1 e^{-Mx}x^{-Y}dx - \int_{-1}^0 e^{Gx}|x|^{-Y}dx\right) \tag{3.2.24}$$

The range of the parameters are restricted to $C, G, M > 0$ and $Y < 2$. Choosing the Y parameters greater than or equal to 2 does not yield a valide Lévy measure.

3.3. Bilateral Gamma Distribution and Process

3.3.1. Bilateral Gamma Distribution. A Bilateral gamma distribution [49] with parameters $\alpha_p, \lambda_p, \alpha_n, \lambda_n > 0$ is defined as the distribution of $Y - Z$, where Y and Z are independent gamma variables with shape and rate parameters, $Y \sim \Gamma(\alpha_p, \lambda_p)$ and $Z \sim \Gamma(\alpha_n, \lambda_n)$. For $\alpha, \lambda > 0$ we denote by $\Gamma(\alpha, \lambda)$ a Gamma distribution, i.e. the absolutely continuous probability distribution with density

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}1_{\{x>0\}}(x), \;\; x \in \mathbb{R}$$

with the characteristic function

$$\varphi(u) = \left(\frac{\lambda}{\lambda - iu}\right)^\alpha \tag{3.3.1}$$

The characteristic function of a bilateral Gamma distribution $\Gamma(\alpha_p, \lambda_p, \alpha_n, \lambda_n)$ is given by

$$\varphi(u) = (\frac{\lambda_p}{\lambda_p - iu})^{\alpha_p}(\frac{\lambda_n}{\lambda_n + iu})^{\alpha_n}, \; u \in \mathbb{R} \tag{3.3.2}$$

LEMMA 18. *(1) Suppose* $X \sim \Gamma(\alpha_{1,p}, \lambda_p, \alpha_{1,n}, \lambda_n)$ *and* $Y \sim \Gamma(\alpha_{2,p}, \lambda_p, \alpha_{2,n}, \lambda_n)$, *and that* X *and* Y *are independent. Then* $X + Y \sim \Gamma(\alpha_{1,p} + \alpha_{2,p}, \lambda_p, \alpha_{1,n} + \alpha_{2,n}, \lambda_n)$.

(2) For $X \sim \Gamma(\alpha_p, \lambda_p, \alpha_{n,}\lambda_n)$ *and* $c > 0$ *it holds* $cX \sim \Gamma(\alpha_p, \frac{\lambda_p}{c}, \alpha_n, \frac{\lambda_n}{c})$

As it is seen from 3.3.2, bilateral Gamma distribution is infinitely divisible. The Lévy measure is given by

$$\nu(dx) = \Big(\frac{\alpha_p}{x}e^{-\lambda_p x}1_{(0,\infty)}(x) + \frac{\alpha_n}{x}e^{-\lambda_n x}1_{(-\infty,0)}(x)\Big)\,dx \tag{3.3.3}$$

Thus, we can also express the characteristic function φ as

$$\varphi(u) = exp\left(\int_R (e^{iux} - 1)k(x)dx\right) \; u \in \mathbb{R} \tag{3.3.4}$$

where $k : \mathbb{R} \to \mathbb{R}$is the function

$$k(x) = \alpha_p \frac{e^{-\lambda_p x}}{x} 1_{(0,\infty)}(x) + \alpha_n \frac{e^{-\lambda_n |x|}}{|x|} 1_{(-\infty,0)}(x),\ x \in \mathbb{R} \tag{3.3.5}$$

The bilateral Gamma distributions are selfdecomposable. It moreover holds

$$\int_{|x|>1} e^{ux} \nu(dx) < \infty \ forall z \in (-\lambda_n, \lambda_p)$$

Consequently, the cumulant generating function

$$\Psi(u) = ln \mathbb{E}[e^{uX}] \tag{3.3.6}$$

exists on $(-\lambda_n, \lambda_p)$, and Ψ and Ψ' are, with regard to 3.3.2, given by

$$\Psi(u) = \alpha_p ln(\frac{\lambda_p}{\lambda_p - u}) + \alpha_n ln(\frac{\lambda_n}{\lambda_n + u}),\ u \in (-\lambda_n, \lambda_p) \tag{3.3.7}$$

$$\Psi'(u) = \frac{\alpha_p}{\lambda_p - u} - \frac{\alpha_n}{\lambda_n + u},\ u \in (-\lambda_n, \lambda_p) \tag{3.3.8}$$

Hence, the n-th order cumulant is gien by

$$k_n = (n-1)!\left(\frac{\alpha_p}{(\lambda_p)^n} + (-1)^n \frac{\alpha_n}{(\lambda_n)^n}\right),\ n \in \mathbb{N} = \{1, 2, ...\} \tag{3.3.9}$$

Then, we can specify

$$\begin{aligned}
\mathbb{E}[X] &= \kappa_1 = \frac{\alpha_p}{\lambda_p} - \frac{\alpha_n}{\lambda_n}, \\
Var[X] &= \kappa_2 = \frac{\alpha_p}{(\lambda_p)^2} + \frac{\alpha_n}{(\lambda_n)^2}, \\
\gamma_1(X) &= \kappa_3/\kappa_2^{3/2} = 2(\frac{\alpha_p}{(\lambda_p)^3} - \frac{\alpha_n}{(\lambda_n)^3})/(\frac{\alpha_p}{(\lambda_p)^2} + \frac{\alpha_n}{(\lambda_n)^2})^{3/2} \\
\gamma_2(X) &= 3 + \kappa_4/\kappa_2^2 = 3 + 6(\frac{\alpha_p}{(\lambda_p)^4} + \frac{\alpha_n}{(\lambda_n)^4})/(\frac{\alpha_p}{(\lambda_p)^2} + \frac{\alpha_n}{(\lambda_n)^2})^2
\end{aligned}$$

3.3.2. Statistics of Bilateral Gamma Distribution. Let $X_1, ..., X_n$ be an i.i.d. sequence of $\Gamma(\Theta)$-distributed random variables, where $\Theta = (\alpha_p, \alpha_n, \lambda_p, \lambda_n)$, and let $x_1, ..., x_n$ be a realization. We start with the method of moments and estimate the k-th moments $m_k = E[X_1^k]$ for $k = 1, ..., 4$ as

$$\hat{m}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \tag{3.3.10}$$

By [47], the following relations between the moments and the cumulants are valid:

$$
\begin{cases}
k_1 = m_1 \\
k_2 = m_2 - m_1^2 \\
k_3 = m_3 - 3m_1m_2 + 2m_1^3 \\
k_4 = m_4 - 4m_3m_1 - 3m_2^2 + 12m_2m_1^2 - 6m_1^4
\end{cases}
\tag{3.3.11}
$$

We can solve the system of equations explicitly. In general, if we avoid the trivila case, it has finite many, but more than one solution. However, in practice, the restriction $\alpha_p, \alpha_n, \lambda_p, \lambda_n > 0$ ensures uniqueness of the solution. This procedure yields a vector $\hat{\Theta}_0$ as first estimation for the parameters. In order to perform a maximum likelihood estimation, we need adequate representations of their density functions. Since the densities satisfy the symmetry relation

$$
f(x;\alpha_p,\lambda_p,\alpha_n,\lambda_n) = f(-x;\alpha_n,\lambda_n,\alpha_p,\lambda_p) \ x \in \mathbb{R}\backslash\{0\}
\tag{3.3.12}
$$

We only analyze the density function on the positive real line

(3.3.13)

$$
f(x) = \frac{\lambda_p^{\alpha_p}\lambda_n^{\alpha_n}}{(\lambda_p+\lambda_n)^{\alpha_n}\Gamma(\alpha_p)\Gamma(\alpha_n)} e^{-\lambda_p x} \int_0^\infty v^{\alpha_n-1}(x+\frac{v}{\lambda_p+\lambda_n})^{\alpha_p-1} e^{-v} dv
$$

We can express the density f by means of the Whittaker function $W_{\lambda,\mu}(z)$, which is a well-studied mathematical function. According to [46], the Whittaker function has the representation

$$W_{\lambda,\mu}(z) = \frac{z^{\lambda} e^{-z/2}}{\Gamma(\mu-\lambda+0.5)} \int_0^{\infty} t^{\mu-\lambda-0.5} e^{-t} (1+\frac{t}{z})^{\mu+\lambda-0.5} dt \tag{3.3.14}$$

$$for \mu - \lambda > -0.5$$

We obtain for $x > 0$

$$f(x) = \frac{\lambda_p^{\alpha_p} \lambda_n^{\alpha_n}}{(\lambda_p+\lambda_n)^{0.5(\alpha_n+\alpha_p)} \Gamma(\alpha_p)} x^{\frac{1}{2}(\alpha_p+\alpha_n)-1} e^{-\frac{x}{2}(\lambda_p-\lambda_n)} \times W_{\frac{1}{2}(\alpha_p-\alpha_n),\frac{1}{2}(\alpha_p+\alpha_n-1)}(x(\lambda_p+\lambda_n)) \tag{3.3.15}$$

The logarithm of the likelihood function for $\Theta = (\alpha_p, \alpha_n, \lambda_p, \lambda_n)$ is, by the symmetry relation and the representation of the density, given by

$$\begin{aligned} lnL(\Theta) = & -n^{+} ln(\Gamma(\alpha_p)) - n^{-} ln(\Gamma(\alpha_n)) \\ & + n(\alpha_p ln(\lambda_p) + \alpha_n ln(\lambda_n) - \frac{\alpha_p+\alpha_n}{2} ln(\lambda_p+\lambda_n) \\ & + (\frac{\alpha_p+\alpha_n}{2} - 1)(\sum_{i=1}^{n} ln|x_i|) - \frac{\lambda_p-\lambda_n}{2}(\sum_{i=1}^{n} x_i) \\ & + \sum_{i=1}^{n} ln(W_{\frac{1}{2}sgn(x_i)(\alpha_p-\alpha_n),\frac{1}{2}(\alpha_p+\alpha_n-1)}(|x_i|(\lambda_p+\lambda_n)) \end{aligned} \tag{3.3.16}$$

where n^+denotes the number of positive, and n^- the number of negative observations. We could take the vector $\hat{\Theta_0}$, obtained from the method of moments, as starting point of an algorithm which maximize the logarithmic likelihood function numerically.

3.3.3. Bilateral Gamma Processes. Bilateral Gamma distributions are infinitely divisible. The Lévy measure $\nu(\mathbb{R}) = \infty$ and $\int_R |x|\nu(dx) < \infty$. Hence, the Bilateral Gamma processes are finite-variation processes making infinitely many jumps at each interval with positive length, and they are equal to the sum of their jumps

$$X_t = \sum_{s\leq t} \Delta X_s = \int_0^t \int_{\mathbb{R}} x\mu^X(dx,ds),\ t \geq 0 \tag{3.3.17}$$

where μ^X denotes the random measure of jumps of X. Bilateral Gamma processes are special semimartingales with canonical decomposition

$$X_t = x * (\mu^X - v)_t + (\frac{\alpha_p}{\lambda_p} - \frac{\alpha_n}{\lambda_n})t,\ t \geq 0 \tag{3.3.18}$$

where v is the compensator of μ^X, which is given by $v(dt,dx) = dtk(dx)$.

$$k(x) = \left(\frac{\alpha_p}{x} e^{-\lambda_p x} 1_{(0,\infty)}(x) + \frac{\alpha_n}{|x|} e^{-\lambda_n |x|} 1_{(-\infty,0)}(x) \right) \tag{3.3.19}$$

We could see that all increments of X have a bilateral gamma distribution, more precisely

$$X_t - X_s \sim \Gamma(\alpha_p(t-s), \lambda_p; \alpha_n(t-s), \lambda_n) \ \ 0 \leqslant s < t \tag{3.3.20}$$

3.4. Markov Modulated Bilateral Gamma Process with Mean Reversion

3.4.1. Pure Jump. Consider a real-valued pure jump process $L = \{L_t, t \geq 0\}$ defined on a probability space (Ω, σ, P). We can write the process

$$L_t = L_0 + \sum_{0 \leq s \leq t} (L_s - L_{s-})$$

So if $\mu^L$is the random measure which gives the random jump time and random jump size $x = L_s - L_{s-}$, then we can write L_t

$$L_t = L_0 + \int_0^t \int_{\mathbb{R}} x\mu^L(dx, ds)$$

The statistical properties are determined by its compensator. We suppose the Lévy measure $\nu(dx, ds) = k(x)dxds$ where $k(x)$ that accounces the arrival rate of jumps of size x. Then, under P, we have the local martingale M_t with respect to the filteration generated by X.

$$M_t = L_0 + \int_0^t \int_{\mathbb{R}} x(\mu^L(dx, ds) - k(x)dxds)$$

We note that the compensator is unique and must be predictable with respect to the filtration generated by L. Also, for the pure jump processes with finite variation and independent and homogeneous increments, then its characteristic function is uniquely given by the Khintchine theorem in terms of the drift μ and the Lévy density $k(x)$

$$E[e^{iuL_t}] = exp[iu\mu t + t\int_{\mathbb{R}}(e^{iux} - 1)k(x)dx]$$

We could easily extend the process X to include a diffusion component. However, [30] demonstrates that this is unnecessary if we allow for infinite jump activity by specifying that $\int_{\mathbb{R}} k(x)dx = \infty$.

3.4.2. Bilateral Gamma Mean Reversion Model. In this section, motivated by the OU process, we propose a model for asset prices with positive values such as stocks and VIX. The model intends to capture the mean reversion feature using a mean reversion drift while capture other empirical regularities using Bilateral gamma process which is a pure jump process with finite variation. In order to allow for time-inhomogeneity, we introduce an independent Markov process to modulate the drift. We think the model could be a good choice to capture the empirical stylized regularities and also we derived a closed form characteristic function of such a model. Our model for a price process S will be the exponential of a mean reversion drift and a pure jump process. There are two ways to show the dynamics, SDE and Levy exponential Model. We demonstrates the two approaches below:

3.4.2.1. *SDE and the VIX process.* We write the dynamic of the asset prices is the solution of the following SDE:

$$dS_t = S_{t-}(-a(t)dt + dL_t) \tag{3.4.1}$$

The process is then the stochastic exponential of a semimartingale. The relation between the jump x in L_t and those of S_t is given by the function $(e^x - 1)$. 3.4.13shows such a relation. Hence we can define the martingale

$$\begin{aligned} m(t) &= (e^x - 1) * (\mu^L - \nu^L) \\ &= \int_0^t \int_{\mathbb{R}} (e^x - 1)(\mu^L(dx,ds) - \nu^L(dx,ds)) \end{aligned} \tag{3.4.2}$$

We know

$$\nu^L(dx,ds) = k(x,s)dxds \tag{3.4.3}$$

Let the martingale $M(t)$ be the stochastic exponential of m.

$$M(t) = \xi(m)_t \tag{3.4.4}$$

$$= exp(L(t) - \int_0^t \int_{\mathbb{R}} (e^x - 1)k(x,s)dxds)$$

$$= exp(\int_0^t \int_{\mathbb{R}} x\mu^L(dx,ds) - \int_0^t \int_{\mathbb{R}} (e^x - 1)k(x,s)dxds) \tag{3.4.5}$$

Further define

$$\theta(t) = \int_{\mathbb{R}} (e^x - 1)k(x,t)dx \tag{3.4.6}$$

By results of stochastic exponentials of semimartingales we have that

$$S_t = S_0 exp\left((\int_0^t \theta(s) - a(s))ds \right) M(t) \tag{3.4.7}$$

$$= S_0 e^{X_t}$$

where

$$X_t = -a(t) + \int_0^t \int_{\mathbb{R}} x\mu^L(ds,dx) \tag{3.4.8}$$

3.4.2.2. *Jump process exponential and the VIX process.* In this subsection, we could first define the process X_t:

$$X_t = -a(t) + \int_0^t \int_{\mathbb{R}} x\mu^L(ds,dx) \tag{3.4.9}$$

where μ^L is the random measure and the ν^L is the Lévy measure of the pure jump process L_t which is a compensator of the μ^L and $k(x)dx = \nu(dx)$ is the Lévy density. Hence, the first term is the mean reversion drift and the second term is the pure jump process. Also, we know the pure jump process X_t with Lévy triplet $(\mu, 0, \nu)$ can be written

Let S_t be the asset price and we take $S = \{S_t, t \geq 0\}$ to be the form

$$S_t = S_0 e^{X_t} \tag{3.4.10}$$

We state a version of Ito's formula directly to the pure jump semimartingales from [32]

(3.4.11)

$$f(X_t) = f(X_0) + \int_0^t f'(X_{s-})dX_s + \sum_{0<s\leq t}[f(X_s) - f(X_{s-}) - f'(X_{s-})\Delta X_s]$$

(3.4.12)

$$= f(X_0) + \int_0^t f'(X_{s-})dX_s + \int_0^t \int_{\mathbb{R}} (f(X_{s-} + x) - f(X_{s-})$$

$$- f'(X_{s-})x)\mu^L(ds, dx)$$

$$= f(X_0) + \int_0^t (-a_s)f'(X_s)ds$$

$$+ \int_0^t \int_{\mathbb{R}} (f(X_{s-} + x) - f(X_{s-}))\mu^L(dx, ds)$$

Let $f(X_t) = S_0 e^{Y_t}$, it gives

(3.4.13)
$$S_t = S_0 + \int_0^t (-a_s)S_s ds + \int_0^t S_{s-} \int_{\mathbb{R}} (e^x - 1)\mu^L(dx, ds)$$

It shows that the asset price is changed by a factor of e^x when a jump x occurs. It ensures the asset price always stay positive. The $e^x - 1$ term ensures the existence of the integral $\int_{\mathbb{R}}(e^x - 1)k(x)dx$. We decide the rewrite the model into a different parameteric form. Define

$$\theta(t) = \int_{\mathbb{R}} (e^x - 1)\nu^L(dx, ds) \tag{3.4.14}$$
$$= \int_{\mathbb{R}} (e^x - 1)k(x,t)dx$$
$$= \int_{\mathbb{R}} (e^x - 1)k(x,t)dx$$

We can rewrite the S_t as

$$S_t = S_0 + \int_0^t S_s(\theta(s) - a(s))ds + \int_0^t S_{s-} \int_{\mathbb{R}} (e^x - 1)(\mu^L(dx, ds) - k(x)dxds) \tag{3.4.15}$$

Recall that $(e^x - 1) * (\mu^L - \nu^L)$ is a martingale, define

$$M_t = \int_0^t \int_{\mathbb{R}} (e^x - 1)(\mu^L(dx, ds) - k(x)dxds)) \tag{3.4.16}$$

The dynamics of the S_t becomes

$$S_t = S_0 + \int_0^t (\theta(s) - a(s))ds + M_t \tag{3.4.17}$$

Therefore, we can also write X_t as

$$X_t = (\theta_t - a_t)t + \int_0^t \int_{\mathbb{R}} x\mu^L(dx,ds) - \int_0^t \int_{\mathbb{R}} (e^x - 1)k(x)dxds \tag{3.4.18}$$

3.4.3. Characteristic functions. In order to derive the characteristic function of the Markov modulated process, we derive the characteristic function of X_t with Lévy density $k^{BG}(x)$. Apply the differentiation rule we discussed in the last section to the function $f(X) = e^{iuX_t}$. It gives

$$\begin{aligned} e^{iuX_t} &= e^{iuX_0} + iu\int_0^t (-a)e^{iuX_s}ds \\ &+ \int_0^t \int_{\mathbb{R}} e^{iuX_{s-}}(e^{iux} - 1)\mu^L(dx,ds) \\ &= e^{iuX_0} + iu\int_0^t (-a_s)e^{iuX_s}ds \\ &+ \int_0^t \int_{\mathbb{R}} e^{iuX_{s-}}(e^{iux} - 1)(\mu^L(dx,ds) - k(x)dx,ds) \\ &+ \int_0^t \int_{\mathbb{R}} e^{iuX_{s-}}(e^{iux} - 1)k(x)dxds \end{aligned} \tag{3.4.19}$$

The second integral above is a martingale. We take the expected values

(3.4.20)

$$\begin{aligned} \varphi_{X_t} = E[e^{iuX_t}] &= e^{iuX_0} + \int_0^t E(e^{iuX_s})\left(iu(-a)t + \int_{\mathbb{R}} (e^{iux} - 1)k(x)dx\right)ds \\ &= exp\{iuX_0 + iu(-a)t + t\left(\int_{\mathbb{R}} (e^{iux} - 1)k(x)dx\right)\} \end{aligned}$$

The unit characteristic function is the value when $t = 1$ and $X_0 = 0$. It is the Lévy-Khintchine representation of the process X_t. Next, we apply the bilateral gamma density and it gives:

$$\begin{aligned}\varphi_{X_1^{BG}}(u) &= exp(iu(-a))\varphi_{BG}(u) \\ &= exp(iu(-a))[(\frac{\lambda_p}{\lambda_p - iu})^{\alpha_p}(\frac{\lambda_n}{\lambda_n + iu})^{\alpha_n}]\end{aligned} \tag{3.4.21}$$

Since we are also interested in θ, we find

$$\begin{aligned}\varphi_{BG}(-i) &= \int_{\mathbb{R}}(e^x - 1)k_{BG}dx \\ &= (\frac{\lambda_p}{\lambda_p - 1})^{\alpha_p}(\frac{\lambda_n}{\lambda_n + 1})^{\alpha_n} \\ &= \theta\end{aligned} \tag{3.4.22}$$

3.4.4. Markov modulated Mean Reversion Bilateral gamma process. The empirical studies indicate that the dynamics of asset prices are not time-homogenous. In this section, we capture the time-inhomogeneity by allowing the drift and the compensator measure to switch between a finite set of drift/measure pairs.

Suppose $U = \{U_t, t \geq 0\}$ is a Markov chain, independent of the Bilateral gamma process (or the genearalized pure jump process), and with a state space

$$\{e_1, e_2, e_3, ..., e_N\}, \;\; where\, e_i = (0, 0, ..., 1, 0, ..., 0)' \in R^N.$$

Suppose also that the generator, or Q-matrix, of U is $\Pi = \{\Pi_{ji}\}$, $1 \leq i, j \leq N$. Π_{ji} represents the rates at which the process U jumps from state i to state j, that is, the transition rates. Π solves

$$\frac{dp_t}{dt} = \Pi p_t,$$

where $p_t^i = P(U_t = e_i)$ is the historical probability of being in state i at time t, and $p_t = (p_t^1, p_t^2, ..., p_t^N)$. Then, from [41], we know that U can be written as

$$U_t = U_0 + \int_0^t \mathit{\Pi} U_s ds + \Gamma_t, \tag{3.4.23}$$

where $\Gamma = \{\Gamma_t, t > 0\}$ is an R^N-valued martingale under P with respect to the filtration generated by U. In the case of Bilateral gamma process, we suppose for each j there is a drift μ_j and Lévy density $k_j^{BG}(x)$ where

$$k_j^{BG}(x) = \frac{\alpha_p^j}{x} e^{-\lambda_p^j x} 1_{(0,\infty)}(x) + \frac{\alpha_n^j}{|x|} e^{-\lambda_n^j |x|} 1_{(-\infty,0)}(x) \tag{3.4.24}$$

Using $< \cdot, \cdot >$ denote the innter product operator, we can write the $k_{BG}(dx, ds, U)$ as

$$k_{BG}(dx,ds,U)=\sum_{j=1}^{N}<U_{s-},e_j>k_j^{BG}(x)dxds \tag{3.4.25}$$

and with $a=(a_1,a_2,...,a_N)\in R^N$, the drift at time s is $<U_s,a>$. Then,

$$\begin{aligned} S_t &= S_0e^{X_t} \\ &= S_0exp\left\{-\int_0^t <U_s,a>ds+\int_0^t\int_{\mathbb{R}}x\mu^L(dx,ds)\right\} \end{aligned} \tag{3.4.26}$$

Then, we have that

(3.4.27)

$$e^{iuX_t}=1+iu\int -<U_s,a>e^{iuX_s}ds+\int_0^t\int_{\mathbb{R}}e^{iuX_{s-}}(e^{iuX}-1)\mu^L(dx,ds)$$

Note that unlike [34] that we only consider one jump process and it is the drift and the compensator that depends on the state of the Markov chain U.

Write the $\{\mathscr{F}^{\mathscr{U}}\}$ for the filteration generated by U, and for $0\le s\le t$, we define

$$\lambda(u,U,s)=\mathbb{E}\left[e^{iuX_s}|\mathscr{F}_t^U\right] \tag{3.4.28}$$

Now, the U is independent of X, so conditioning on $\mathscr{F}_t^U$ and we have for fixed t that

$$\lambda(u,U,t) = 1 + iu\int_0^t - < U_s, a > \lambda(u,U,s)ds \\ + \sum_{j=1}^{N}\int_0^t\int_{\mathbb{R}} < U_s, e_j > (e^{iux}-1)\lambda(u,U,s)k_j(x)dxds \tag{3.4.29}$$

Write $H_t^j = \int_0^t < U_s, e_j > ds$ for the amount of time the process U has spent in state j up to time t. Also, write

$$\phi_j(u) = iu(-a_j) + \int_{-\infty}^{+\infty}(e^{iux}-1)k_j(x)dx \tag{3.4.30}$$

for the unit log-characteristic function. In the case of Bilateral gamma process, we have

$$\phi_j(u) = iu(-a_j) + \alpha_p^j log(\frac{\lambda_p^j}{\lambda_p^j - iu}) + \alpha_n^j log(\frac{\lambda_n^j}{\lambda_n^j + iu}) \tag{3.4.31}$$

Then, noting that $< U_s, -a >= \sum_{j=1}^N < U_{s-}, e_j > (-a_j)$, we obtain

$$\lambda(u,U,t) = exp\left(H_t^1\phi_1(u) + H_t^2\phi_2(u) + ... + H_t^N\phi_N(u)\right) \tag{3.4.32}$$

The characteristic function of the process X_t is therefore

(3.4.33)

$$\varphi_{X_t}(u) = E\left[e^{iuX_t}\right] = E\left[\lambda(u,U,t)\right] = \Phi_{H(t)}(\phi_1(u),\phi_2(u),...,\phi_N(u))$$

where

(3.4.34)
$$\Phi_{H(t)}(\lambda) = \Phi_{H(t)}(\lambda^1,\lambda^2,...,\lambda^N)$$
$$= E\left[exp(\lambda^1 H_t^1 + \lambda^2 H_t^2 + ... + \lambda^N H_t^N\right]$$

is the Laplace transform of $H_t = (H_t^1, H_t^2, ..., H_t^N)$.

We still need to obtain the closed form expression for $\Phi_{H(t)}(\lambda)$. The proposition below gives a closed form solution.

PROPOSITION 19. *For the N-state Markov switching model, the Laplace transform of the occupation time H_t is given by*

$$\Phi_{H(t)}(\lambda) = E\left[exp(<\lambda, H_t>)\right] = < exp\{(\Pi + diag(\lambda))t\}E[U_0], \bar{1} >$$

where $\bar{1} \in R^N$is a vector of ones and Π is the Q-matrix of the Markove chain U.

PROOF. For a vector of transform variabls $\lambda = (\lambda^1, \lambda^2, ..., \lambda^N) \in R^N$, the Laplace transform of H is

$$E\left[exp(<\lambda, H_t>)\right] = E\left[exp(\lambda^1 H_t^1)...exp(\lambda^N H_t^N)\right]$$

Define Z_t as the random vector process

$$Z_t = exp(<\lambda, H_t>)U_t = exp(\int_0^t <\lambda, U_s> ds)U_t$$

Then $Z_t \in R^N$and

$$dZ_t = <\lambda, U_t> Z_t dt + exp(\int_0^t <\lambda, U_s> ds)dU_t$$

Recall that $U_t = U_0 + \int_0^t \Pi U_s ds + \Gamma_t$ and so we can substitute to get

$$dZ_t = <\lambda, U_t> Z_t dt + \Pi Z_t dt + exp(\int_0^t <\lambda, U_s> ds)d\Gamma_t$$

Now $<\lambda, U_t> Z_t = diag(\lambda)Z_t$, so

$$Z_t = U_0 + \int_0^t (\Pi + diag(\lambda))Z_s ds + \int_0^t exp(\int_0^s <\lambda, U_v> dv)d\Gamma_s$$

The last integral is a martingale, and taking expectations gives

$$E[Z_t] = E[U_0] + \int_0^t (\Pi + diag(\lambda))E[Z_s]ds$$

Solving yields

$$E[Z_t] = exp\{(\Pi + diag(\lambda))t\}E[U_0]$$

Observing that $E[exp(<\lambda, H_t>)] = E[<exp(<\lambda, H_t>)U_t, \bar{1}>]$ and denoting $\Phi_{H(t)}(\lambda) = E[exp(<\lambda, H_t>)]$, we obtain that

$$\begin{aligned}\Phi_{H(t)}(\lambda) &= <E[Z_t], \bar{1}> \\ &= <E[exp(<\lambda, H_t>)U_t], \bar{1}> \\ &= <exp\{(\Pi + diag(\lambda))t\}E[U_0], \bar{1}> \\ &= <E[U_0], exp\{(\Pi + diag(\lambda))t\}, \bar{1}>\end{aligned}$$

which is the desired result. □

3.5. Parameter Estimation

This section discusses three potential procedures to estimate the parameters using the characteristic function we derived from the last section. To avoid using the filtering method, we assume the return of VIX is in a stationary state. The parameters obtained this way will be the real-world parameters, which will be different from the risk-neutral parameters. We do not intend to determine the optimal number of states and the optimal estimation methodologies in this section. We defer the risk-neutral model's calibration and comprehensive econometric analysis of the estimation methodologies to a later study.

3.5.1. Stationary Assumption. We assume the return of VIX is in a stationary state. This imposes a certain restrictions on the initial probabilities so that the initial probabilities cannot be estimated freely. Denote the initial probability of being in state i as $p_0^i = P(U_0 = e_i)$ and $p_0 = (p_0^1, ...p_0^N)$. Recall that the transition rate matrix solves

$$\frac{dp_t}{dt} = \Pi p_t$$

For the process to be stationary, we need

$$\frac{dp_t}{dt} = 0$$

This is the case if the initial probability distribution $p_0$satisfies

$$\Pi p_0 = 0$$

3.5.2. Maximu Likelihood estimation. The probability density function of the returns could be derived by inverting the characteristic function $\Phi_{S_t}(z)$ using inverse Fourier transform. Assuming the individual state process is a Bilateral gamma process, a maximum likelihood estimation of the parameters can be calculated. Since the procedure needs to invert the characteristic function at every return point, a well-established fast Fourier transform (FFT) can be implemented on binned sampled returns to reduce the computational intensity.

However, it is well known that the MLE may suffer degeneracy problems, especially for the mixtures of distributions. The singularities on the likelihood surface caused these problems. In such a case, any numerical routine to maximize the likelihood function will result in unreliable results. The EM algorithm or the penalized maximum likelihood has been proposed to deal with such cases.

3.5.3. Generalized method of moments. The generalized method of moments exploits the characteristic fucntion directly using empirical characteristic function. Applications of the method can be found in [39, 40, 38]. The characteristic function is defined as

$$\Phi_{X_t}(u,\theta) = E[e^{iuX_t}]$$

where θ is a set of parameters of the characteristic funciton.

The empirical characteristic function is defined as

$$\hat{\Phi}_N(u) = \frac{1}{N}\sum_{j=1}^{N} e^{iuX_j}$$

Define a partition grid $u^p = (u_1, u_2, ..., u_p)$ of transform variables, which is used to form the basis of the moment conditions, and let

$$h(u_k, X_j, \theta) = e^{iu_kX_j} - \Phi_{X_t}(u_k, \theta)$$

so that $E[h(u_k, X_k, \theta)] = 0$. Then, the sample moment is derived as

$$h_N(u_k, \theta) = \hat{\Phi}_N(u_k) - \Phi_{X_t}(u_k, \theta)$$

which forms a vector of sample moments $h_N(u^q, \theta)$. θ is found by minimizing the quadratic function $h_N(u^q, \theta)' W h_N(u^q, \theta)$, where W is an optimal weighting matrix. The matrix is the inverse of the covariance matrix associated with the moments. Intuitively, the estimator becomes more efficient as the number of moments gets larger (the partition grid gets finer). However, the covariance matrix may become singular and cannot be inverted. To solve the problem, we can formulate the GMM objective problem as a minimum norm problem on a new space of moment conditions, a spaced endowed with the norm

$$\| h_N \|^2 = \int_R h_N(z) \bar{h}_N(x) \pi(z) dz$$

where $\bar{h}_N(z)$ is the complex conjugate of h_N. $\pi(z)$ is a Gaussian probability density function. [36, 37] provides details of the approach which is called the C-GMM approach.

3.5.4. Tail Moment Matching. To extend the generalized method of moments, [35] argued that the use of bounded tail probabilities for moment matching was more reliable than the use of other possibly unbounded moments. The latter was less susceptible to the present of outliers that naturally occur when the data is not known to be coming from the proposed model.

For three state model, the characteristic function may be used to evaluate model probabilities for tail probabilities

$$P(X(t) < a < 0), P(0 < a < X(t))$$

From the data we may construct the empirical observations for $i = 1, ..., 99$

$$i\% = P(X(t) < y_i)$$

The points y_i may be interpolated from the empirical distribution function for the required percentiles. The data may then be split into the two tails by

$$i\% = P(X(t) < y_i < 0)$$

$$1 - i\% = P(0 < y_i < X(t))$$

Model parameters are estimated by least squares matching of observed tail probabilities to model tail probabilities.

3.5.5. Results. Terms such as high and low volatility regimes are frequently used by investors to describe the market environment. Although the two-state classification is intuitively appealing, we believe a three regime approach is more appropriate. For example, there exists effective three regime quantitative strategies applied to VSTOXX in industry. Using the daily data from 1990 to 2018 for the VIX Index, we use the tail moment matching estimation to calibrate the three states model. We name these states low, medium and high volatility states.

As we defined in Section 3.4, there are two drifts in the model. One is the OU mean reversion drift a and the other is the pure jump drift θ. We found that the mean reversion drift is increasing from the low vol states to high vol states. It it intuitive because VIX Index reverse to its mean faster when it is in the high regime. We calculated the total drift of the process $\theta - a$. We found VIX has a positive drift in low vol regime and negative drifts in mid and high vol regimes.

Also, we observed the kurtosis is increasing from low vol states to high vol states and the model well captured the positive skew in different regimes. We observe that the $c_n$exceeds c_p and b_p is greater than b_n. It illustrates that the positive moves are less frequent and larger while the negative moves are more frequent and small. Usually we can obeserve in the market that the VIX index have rare big positive jumps and then the index will drift down. The $c_n > c_p$ and $b_n < b_p$ well capture the asymmetries between up moves and down moves.

These results are only suggestive however, and as mentioned previously, we defer a more detailed empirical analysis to a later study.

REMARK. There are two parameterization which are $\alpha_p = c_p$ and $\lambda_p = \frac{1}{b_p}$. Parameters in the bracket refer to the equivalent parameters.

TABLE 1. Estimates on VIX Index 1990-2018

State	Low Vol	Medium Vol	High Vol
a	1.0065	1.0082	1.0267
c_p	1.0354	1.3820	0.6720
b_p	0.0376	0.0376	0.0702
c_n	1.2146	1.9828	0.7843
b_n	0.0263	0.0266	0.0458
θ	1.0083	1.0009	1.0216
$\theta - a$	0.0016	-0.0073	-0.0128
p	0.3218	0.3514	0.3268
volatility	0.0479	0.0579	0.0703
skewness	0.5958	0.3724	0.8999
kurtosis	5.9922	4.9998	7.8272

3.6. Trading Strategies

3.6.1. Implications for volatility strategies. The identification of volatility regimes also has important implications for the performance of volatility strategies. Short volatility strategies tend to perform well in low and mid volatility regimes, and poorly in high volatility regimes. Therefore, we would deem low or mid vol regimes to be normalized environments for monetizing volatility premium. The sharpe ratio of shorting VIX strategies is demonstrated below.

TABLE 2. Sharpe ratio of short 1M volatility strategy

	Low Vol	Mid Vol	High Vol
Sharpe Ratio	1.5	0.7	-0.3

3.6.2. Forcasting future volatility and generating trading strategies.

While the model helps us with understanding and describing the behavior of VIX Index, it can also be used to forecast future values and lead to directly actionable trading strategies. By simulation, we could obtain a distribution of the possible terminal values of VIX at the target expiry. By doing so, we could forecast the probability distribution rather than a point estimate of futures VIX values, and obtain a better understanding of the risk/reward.

Overlaying the two probability distributions, we can identify the strikes where the Index options are under or overvalued relative to historically realized. For instance, the VIX options could be underestimate the density on the left tail and overestimate on the right tail, i.e. the SPY put options are underpriced and calls are overpriced.

3.7. Conclusion

Popular clichés about markets refer to markets taking escalators up and elevators down. Such characterizations suggest that one models the rise of markets differently from the fall. We present a class of Lévy processes for modeling such market fluctuations: Bilateral gamma processes. Lévy processes are time homogenous. We propose a model for asset prices which is the exponential of the bilateral gamma process whose statistical behavior is allowed to switch between N states. This is achieved by applying Markov switching drift/compensator pairs. We also added the OU mean reversion drift to the model so that the resulting asset price model has the potential to capture any empirically observed behavior such as different up and down moves dynamics, term structures of moments, etc. In addition to the Bilateral gamma processes, we also generalize it to a pure jump process. We derived the closed-form expression for the characteristic function and calibrated the model to the VIX Index, assuming stationary distribution.

3.8. Future Research

Firstly, we derived the characteristic function of the proposed model in a physical world. In order to calibrate the model to option prices, we can explore the measure change of the process and derive the characteristic function in risk neutral world in the future. Secondly, we assume the distribution is stationary in this chapter and calibrate the model to the time series data. In the future, we can explore to use the filter method, such as Unscented Kalman Filter, to estimate the model without the stationary assumption. Finally, we only conceptually discussed the model's potential application in trading, especially in the short volatility strategy. In the future, we could explore to simulate the model and backtest the strategy we proposed using real data.

APPENDIX A

Appendix to Chapter 1

A.1. Simulating P/L

We briefly outline details of our daily trades. In all cases, we ignore transaction costs.

- For 10-year Treasuries, we long the current issue, financing the bond in an overnight repo including specials. If a new security is auctioned during the trade period of either one-week or one-month, we do not roll to the new bond. The performance is given by the total return, which consists of the change in the bond's dirty price, less total repo cost.
- For 10-year matched maturity swap spreads, we long the most recently auctioned 10-year Treasury and long the asset swap to pay the fixed on a weighted notional of the swap with maturity matched to the Treasury. We hold the package (10-year Treasury and the asset swap) for either one week or one month and compute total return including carry from the package.
- For 1Mx10Y ATMF swaption straddles, we long the straddle for either one week or one month without delta hedging. For one week holding period, we compute the change in the premium. For the

one month holding period, we compare initial premium to terminal payoff.

APPENDIX B

Appendix to Chapter 2

B.1. Data Universe

	Bloomberg Ticker	**Name of Asset**	**Risk Weight**	**Asset Class Weight**
Equities	GX1 Index	DAX Index	2.24%	25%
	VG1 Index	DJ Euro Stoxx 50	3.53%	
	Z1 Index	FTSE100 Index	1.00%	
	ES1 Index	S&P 500 Index	12.06%	
	FTJGUSSE Index	Russell 2000 EMini	0.69%	
	NQ1 Index	Nasdaq 100 E-Mini	1.32%	
	NI1 Index	Nikkei 225 Index	1.08%	
	TP1 Index	OSE Japan Topix Index	0.88%	
	KM1 Index	KOSPI 200 Index	0.94%	
	HI1 Index	Hang Seng Index	1.26%	

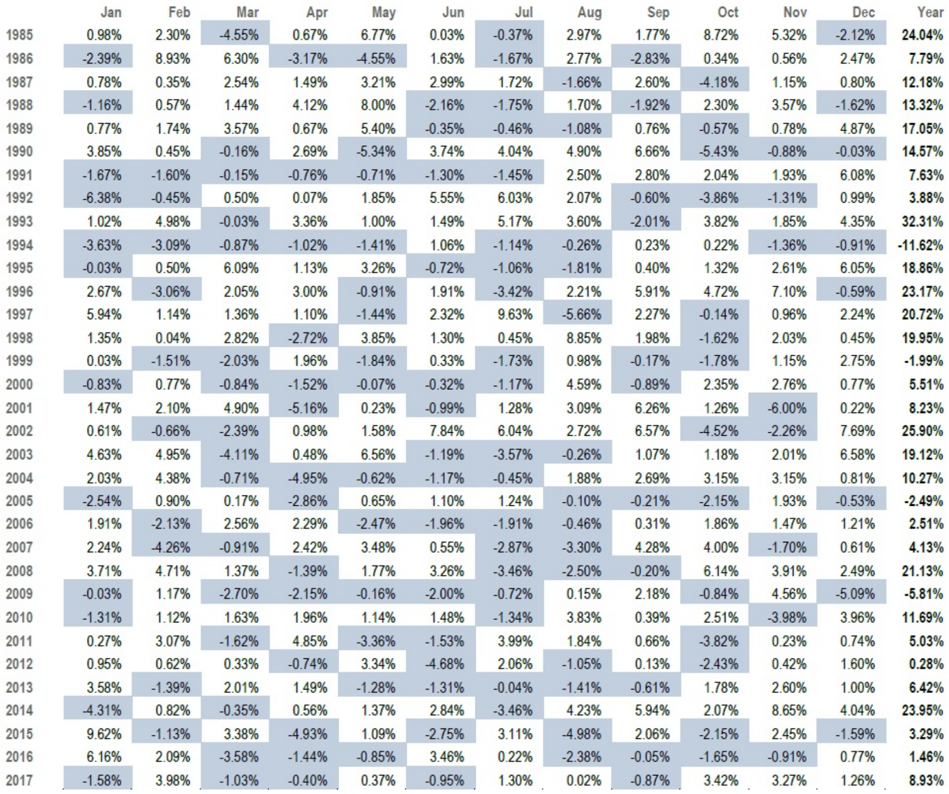
Curr	EURUSDCR Index	EUR Total Return	9.46%	25%
	GBPUSDCR Index	GBP Total Return	3.87%	
	SEKUSDCR Index	SEK Total Return	0.67%	
	CADUSDCR Index	CAD Total Return	1.54%	
	JPYUSDCR Index	JPY Total Return	6.75%	
	AUDUSDCR Index	AUD Total Return	2.07%	
	NZDUSDCR Index	NZD Total Return	0.63%	
Comm	CO1 Comdty	Brent Crude	4.64%	25%
	CL1 Comdty	WTI Crude	9.75%	
	HO1 Comdty	Heating Oil	1.06%	
	XB1 Comdty	Gasoline	0.96%	
	NG1 Comdty	Natural Gas	1.77%	
	GC1 Comdty	Gold	2.89%	
	SI1 Comdty	Silver	0.60%	
	HG1 Comdty	Comex Copper	0.65%	

	C 1 Comdty	Corn	0.60%	
	S 1 Comdty	Soybean	2.09%	
Fixed Income	ED4 Comdty	Eurodollar	2.52%	25%
	TU1 Comdty	US Treasury Note 2Y	1.38%	
	FV1 Comdty	US Treasury Note 5Y	2.06%	
	TY1 Comdty	US Treasury Note 10Y	3.83%	
	US1 Comdty	US Treasury Long Bond	0.85%	
	DU1 Comdty	Euro Schatz (2y)	1.58%	
	OE1 Comdty	Euro Bobl (5y)	2.96%	
	OAT1 Comdty	French Govt. Bonds (10y)	0.82%	
	RX1 Comdty	Euro Bund (10y)	5.04%	
	IK1 Comdty	Italian Govt. Bonds (10y)	0.65%	
	G 1 Comdty	Long Gilt (10y)	0.60%	
	JB1 Comdty	Japanese Gov’t Bond (10y)	1.47%	

	YM1 Comdty	Australian Gov’t Bond (3y)	0.56%	
	XM1 Comdty	Australian Gov’t Bond (10y)	0.66%	

B.2. Monthly Return Series

FIGURE B.2.1. Monthly Return Series



	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year
1985	0.98%	2.30%	-4.55%	0.67%	6.77%	0.03%	-0.37%	2.97%	1.77%	8.72%	5.32%	-2.12%	**24.04%**
1986	-2.39%	8.93%	6.30%	-3.17%	-4.55%	1.63%	-1.67%	2.77%	-2.83%	0.34%	0.56%	2.47%	**7.79%**
1987	0.78%	0.35%	2.54%	1.49%	3.21%	2.99%	1.72%	-1.66%	2.60%	-4.18%	1.15%	0.80%	**12.18%**
1988	-1.16%	0.57%	1.44%	4.12%	8.00%	-2.16%	-1.75%	1.70%	-1.92%	2.30%	3.57%	-1.62%	**13.32%**
1989	0.77%	1.74%	3.57%	0.67%	5.40%	-0.35%	-0.46%	-1.08%	0.76%	-0.57%	0.78%	4.87%	**17.05%**
1990	3.85%	0.45%	-0.16%	2.69%	-5.34%	3.74%	4.04%	4.90%	6.66%	-5.43%	-0.88%	-0.03%	**14.57%**
1991	-1.67%	-1.60%	-0.15%	-0.76%	-0.71%	-1.30%	-1.45%	2.50%	2.80%	2.04%	1.93%	6.08%	**7.63%**
1992	-6.38%	-0.45%	0.50%	0.07%	1.85%	5.55%	6.03%	2.07%	-0.60%	-3.86%	-1.31%	0.99%	**3.88%**
1993	1.02%	4.98%	-0.03%	3.36%	1.00%	1.49%	5.17%	3.60%	-2.01%	3.82%	1.85%	4.35%	**32.31%**
1994	-3.63%	-3.09%	-0.87%	-1.02%	-1.41%	1.06%	-1.14%	-0.26%	0.23%	0.22%	-1.36%	-0.91%	**-11.62%**
1995	-0.03%	0.50%	6.09%	1.13%	3.26%	-0.72%	-1.06%	-1.81%	0.40%	1.32%	2.61%	6.05%	**18.86%**
1996	2.67%	-3.06%	2.05%	3.00%	-0.91%	1.91%	-3.42%	2.21%	5.91%	4.72%	7.10%	-0.59%	**23.17%**
1997	5.94%	1.14%	1.36%	1.10%	-1.44%	2.32%	9.63%	-5.66%	2.27%	-0.14%	0.96%	2.24%	**20.72%**
1998	1.35%	0.04%	2.82%	-2.72%	3.85%	1.30%	0.45%	8.85%	1.98%	-1.62%	2.03%	0.45%	**19.95%**
1999	0.03%	-1.51%	-2.03%	1.96%	-1.84%	0.33%	-1.73%	0.98%	-0.17%	-1.78%	1.15%	2.75%	**-1.99%**
2000	-0.83%	0.77%	-0.84%	-1.52%	-0.07%	-0.32%	-1.17%	4.59%	-0.89%	2.35%	2.76%	0.77%	**5.51%**
2001	1.47%	2.10%	4.90%	-5.16%	0.23%	-0.99%	1.28%	3.09%	6.26%	1.26%	-6.00%	0.22%	**8.23%**
2002	0.61%	-0.66%	-2.39%	0.98%	1.58%	7.84%	6.04%	2.72%	6.57%	-4.52%	-2.26%	7.69%	**25.90%**
2003	4.63%	4.95%	-4.11%	0.48%	6.56%	-1.19%	-3.57%	-0.26%	1.07%	1.18%	2.01%	6.58%	**19.12%**
2004	2.03%	4.38%	-0.71%	-4.95%	-0.62%	-1.17%	-0.45%	1.88%	2.69%	3.15%	3.15%	0.81%	**10.27%**
2005	-2.54%	0.90%	0.17%	-2.86%	0.65%	1.10%	1.24%	-0.10%	-0.21%	-2.15%	1.93%	-0.53%	**-2.49%**
2006	1.91%	-2.13%	2.56%	2.29%	-2.47%	-1.96%	-1.91%	-0.46%	0.31%	1.86%	1.47%	1.21%	**2.51%**
2007	2.24%	-4.26%	-0.91%	2.42%	3.48%	0.55%	-2.87%	-3.30%	4.28%	4.00%	-1.70%	0.61%	**4.13%**
2008	3.71%	4.71%	1.37%	-1.39%	1.77%	3.26%	-3.46%	-2.50%	-0.20%	6.14%	3.91%	2.49%	**21.13%**
2009	-0.03%	1.17%	-2.70%	-2.15%	-0.16%	-2.00%	-0.72%	0.15%	2.18%	-0.84%	4.56%	-5.09%	**-5.81%**
2010	-1.31%	1.12%	1.63%	1.96%	1.14%	1.48%	-1.34%	3.83%	0.39%	2.51%	-3.98%	3.96%	**11.69%**
2011	0.27%	3.07%	-1.62%	4.85%	-3.36%	-1.53%	3.99%	1.84%	0.66%	-3.82%	0.23%	0.74%	**5.03%**
2012	0.95%	0.62%	0.33%	-0.74%	3.34%	-4.68%	2.06%	-1.05%	0.13%	-2.43%	0.42%	1.60%	**0.28%**
2013	3.58%	-1.39%	2.01%	1.49%	-1.28%	-1.31%	-0.04%	-1.41%	-0.61%	1.78%	2.60%	1.00%	**6.42%**
2014	-4.31%	0.82%	-0.35%	0.56%	1.37%	2.84%	-3.46%	4.23%	5.94%	2.07%	8.65%	4.04%	**23.95%**
2015	9.62%	-1.13%	3.38%	-4.93%	1.09%	-2.75%	3.11%	-4.98%	2.06%	-2.15%	2.45%	-1.59%	**3.29%**
2016	6.16%	2.09%	-3.58%	-1.44%	-0.85%	3.46%	0.22%	-2.38%	-0.05%	-1.65%	-0.91%	0.77%	**1.46%**
2017	-1.58%	3.98%	-1.03%	-0.40%	0.37%	-0.95%	1.30%	0.02%	-0.87%	3.42%	3.27%	1.26%	**8.93%**

B.3. Correlation between the P&L of Two Trend-Following Signals

PROOF. Assume that the asset's return R_t is i.i.d. following normal distribution $N(0,\sigma^2)$.[1] Under such assumptions,

[1] At first sight the assumption of a Gaussian white noise might seem restrictive. In practice the returns processes over different timeframes may differ. In extreme cases we can even have trends in opposite direction. Hence, we consider the assumption a good compromise. Later we show that it is also a realistic one as the theoretical and empirical correlation matrices are quite close.

$$\text{(B.3.1)}\quad d1_{t,T} = \frac{ln(\frac{S_t}{S_{t-T}}) + \sigma^2 T/2}{\sigma\sqrt{T}} = \frac{\sum_{s=t-T+1}^{t} ln(\frac{S_s}{S_{s-1}}) + \frac{\sigma^2 T}{2}}{\sigma\sqrt{T}}$$

$$= \frac{\sum_{s=t-T+1}^{t} R_s}{\sigma\sqrt{T}} = \frac{\sqrt{T}\bar{R_{t,T}}}{\sigma} \sim N(0,1)$$

where $d1_{t,T}$ is the Black-Scholes $d1$ statistics calculated at time t, for an option maturity T and strike S_{t-T} and $\bar{R_{t,T}}$is the average asset return from $t-T+1$ to t. If we consider two lookback periods $T_1$and T_2 ($T_1 < T_2$), it follows that the correlation .

$$\text{(B.3.2)}\quad \rho(d1_{t,T_1}, d1_{t,T_2}) = E[d1_{t,T_1} d1_{t,T_2}] = E[\frac{\sum_{t-T_1+1}^{t} R_s}{\sigma\sqrt{T_1}} \cdot \frac{\sum_{t-T_2+1}^{t} R_s}{\sigma\sqrt{T_2}}]$$

$$= \frac{T_1\sigma^2}{\sigma^2\sqrt{T_1 T_2}} = \frac{\sqrt{T_1}}{\sqrt{T_2}}$$

Let $S_{t,T}$ denote the trend-following signal at time t based on lookback period T and $PL_{t+1,T}$ is the P&L at time $t+1$ based on signal $S_{t,T}$. The position is proportional to the signal and inversely proportional to the volatility and hence $PL_{t+1,T} = R_{t+1}S_{t,T}/\sigma$. Furthermore, let Φ denote the c.d.f of the standard normal distribution.

(B.3.3)

$$Var[PL_{t+1,T}] = E[PL^2_{t+1,T}] = E[E_t[PL^2_{t+1,T}]] = E[S^2_{t,T}E[R^2_{t+1}]/\sigma] = E[S^2_{t,T}]$$

$$= E[(2\Phi(d1_{t,T}) - 1)^2] = 4E[(\Phi(d1_{t,T})^2] - 4E[\Phi(d1_{t,T})] + 1$$

If a random variable $X \sim N(0,1)$, we can show that

$$E[\Phi(X)^2] = \int_{-\infty}^{+\infty} (\Phi(X)^2) f(x)dx = 1/3$$

$$Var(PL_{t+1,T}) = \frac{1}{3}$$

It follows that

$$\text{(B.3.4)} \quad \rho(PL_{t+1,T_1}, PL_{t+1,T_2}) = \frac{E[(2\Phi(d1_{t,T_1}) - 1)(2\Phi(d1_{t,T_2}) - 1)]}{\sqrt{Var(PL_{t+1,T_1})Var(PL_{t+1,T_2})}}$$

$$= 12E(\Phi(d1_{t,T_1})\Phi(d1_{t,T_2})) - 3$$

$d1_{t,T_1}$ and $d1_{t,T_2}$ have a bivariate normal distribution with mean vector $\mu_1 = [0,0]$ and covariance matrix $\Sigma_1 = \begin{bmatrix} 1 & \sqrt{T_1/T_2} \\ \sqrt{T_1/T_2} & 1 \end{bmatrix}$. Using Lemma 1 in [55], we can show that $E[\Phi(d1_{t,T_1})\Phi(d1_{t,T_2})] = P(x < 0, y <$

0) where x and y have a bivariate normal distribution with mean vector $\mu_2 = [0,0]$ and covariance matrix $\Sigma_2 = \begin{bmatrix} 2 & \sqrt{T_1/T_2} \\ \sqrt{T_1/T_2} & 2 \end{bmatrix}$.[2]

We can make use of the properties of the bivariate normal distribution and further conclude that

$$E[\Phi(d1_{t,T_1})\Phi(d1_{t,T_2})] = P(x<0,y<0) = P(\frac{x}{\sqrt{2}}<0,\frac{y}{\sqrt{2}}<0) \tag{B.3.5}$$

$$= \frac{1}{4} + \frac{asin(0.5\cdot\sqrt{\frac{T_1}{T_2}})}{2\pi}$$

[3]After simplication, we have

$$\rho = 6asin(0.5\cdot\sqrt{\frac{T_1}{T_2}})/\pi \tag{B.3.6}$$

.

Note that the correlation is soley dependent on the ratio between the period $T_1$and T_2. If we target particular value of ρ, it follows that $T_1/T_2 = 4(sin(\frac{\rho\pi}{6}))^2$. The correlations can be considered relatively high. For example, for $\frac{T_1}{T_2} = 0.5$ it follows that $\rho = 0.69$. □

[2] Alternatively, we can use the conditional distribution of $d1_{t,T_2}$given $d1_{t,T_1}$ integrate out $d1_{t,T_2}$

[3] The result can be found in [53] and [54]

B.4. Expected P&L when the Asset's Return Follows an AR(1) Process

PROOF. Assume that $R_t = a + \rho R_{t-1} + \varepsilon_t$ where $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$. It follows that $R_t \sim N(\frac{a}{1-\rho}, \frac{\sigma_\varepsilon^2}{1-\rho^2}) \sim N(\mu, \sigma^2)$. We know that $PL_{t+1,T} = \frac{R_{t+1}S_{t,T}}{\sigma} = R_{t+1}(2\Phi(d1_{t,T}) - 1)/\sigma$. Note that in the following it is assumed that the volatility $\sigma^2$of the AR(1) process is known quantity. In an AR(1) the estimate of the volatility is asymptotically normal, $\hat{\sigma^2} \sim N(\sigma^2, \frac{2\sigma^4(1+\rho^2)}{T(1-\rho^2)})$ (see [52]). For financial daily data $abs(\rho)$ is sufficiently small and the sample size dominates the error of the estimate. For example, if we assume $\rho = 0$,the standard error of the estimate will be less than 10% of the true value when $T = 252$ days. For the subsequent derivations we need to find the correlation between R_{t+1}/σ and $d1_{t,T}$:

(B.4.1)

$$\begin{aligned} Cov(\frac{R_{t+1}}{\sigma}, d1_{t,T}) = Cov(\frac{R_{t+1}}{\sigma}, \frac{\sum_{s=t-T+1}^{t} R_s}{\sigma\sqrt{T}}) &= \frac{\sum_{s=t-T+1}^{t} Cov(R_{t+1}R_s)}{\sigma^2\sqrt{T}} \\ &= \frac{1}{\sqrt{T}}(\rho + \rho^2 + ... + \rho^T) = \frac{\rho(1-\rho^T)}{\sqrt{T}(1-\rho)} \end{aligned}$$

The derivation of $Var(d1_{t,T})$ requires more algebraic operations:

(B.4.2) $$Var(d1_{t,T}) = Var(\frac{\sum_{s=t-T+1}^{t} R_s}{\sigma\sqrt{T}}) = \frac{1}{T}(T + 2(T-1)\rho$$

$$+ 2(T-2)\rho^2 + ... + 2\rho^{T-1})$$

$$= \frac{T(1-\rho^2) - 2\rho(1-\rho^T)}{T(1-\rho)^2}$$

Hence, the correlation between $\frac{R_{t+1}}{\sigma}$ and $d1_{t,T}$ is

(B.4.3) $$\rho(\frac{R_{t+1}}{\sigma}, d1_{t,T}) = \frac{\rho(1-\rho^T)}{\sqrt{T(1-\rho^2) - 2\rho(1-\rho^T)}}$$

Now let's denote $X = \frac{R_{t+1}}{\sigma} \sim N(\mu/\sigma, 1)$ and $Y = d1_{t,T} \sim N(\mu_{d1_{t,T}}, \sigma^2_{d1_{t,T}})$ with $\mu_{d1_{t,T}} = \frac{\sqrt{T}\mu}{\sigma}$ and $\sigma^2_{d1_{t,T}} = Var(d1_{t,T})$. Note that X and Y are jointly bivariate normal with correlation $\phi = \frac{\rho(1-\rho^T)}{\sqrt{T(1-\rho^2)-2\rho(1-\rho^T)}}$. Hence,

$$X|Y \sim N(\frac{\mu}{\sigma} + \frac{\phi(Y - \mu_{d1_{t,T}})}{\sigma_{d1_{t,T}}}, 1 - \phi^2)$$

$$Y|X \sim N(\mu_{d1_{t,T}} + \sigma_{d1_{t,T}}\phi(X - \frac{\mu}{\sigma}), (1-\phi^2)\sigma^2_{d1_{t,T}})$$

It follows that

$$\text{(B.4.4)}\quad E[PL_{t+1,T}] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x(2\Phi(y)-1)f(x,y)dxdy =$$

$$2\int_{-\infty}^{+\infty}\Phi(y)f(y)\int_{-\infty}^{+\infty} xf(x|y)dxdy - \int_{-\infty}^{+\infty} xf(x)\int_{-\infty}^{+\infty} f(y|x)dydx$$

Let $X^* \sim N(0,1)$ and $Y^* \sim N(0,1)$. Making use of the conditional distribution of $X|Y$ and $Y|X$ and formulas 10010.8 and 10011.3 in [51] and with f denoting the c.d.f. of the standard normal distribution, we obtain:

(B.4.5)

$$E[PL_{t+1,T}] = 2\int_{-\infty}^{+\infty}\Phi(y)f(y)(\frac{\mu}{\sigma}+\frac{\phi(y-\mu_{d1,T})}{\sigma_{d1,T}})dy - \int_{-\infty}^{+\infty} xf(x)dx =$$

$$2(\frac{\mu}{\sigma}-\phi\frac{\mu_{d1,T}}{\sigma_{d1,T}})\int_{-\infty}^{+\infty}\Phi(\sigma_{d1,T}y^*+\mu_{d1,T})f(y^*)dy^*$$

$$+2\frac{\phi}{\sigma_{d1,T}}\int_{-\infty}^{+\infty}\Phi(\sigma_{d1,T}y^*+\mu_{d1,T})f(y^*)(\sigma_{d1,T}y^*+\mu_{d1,T})dy^* - \frac{\mu}{\sigma}$$

$$= 2\frac{\mu}{\sigma}\Phi(\frac{\mu_{d1,T}}{\sqrt{1+\sigma_{d1,T}^2}}) - \frac{\mu}{\sigma} - 2\phi\frac{\sigma_{d1,T}}{\sqrt{1+\sigma_{d1,T}^2}}f(\frac{\mu_{d1,T}}{\sqrt{1+\sigma_{d1,T}^2}})$$

In case $\rho = 0$,it follows that

$$\text{(B.4.6)}\qquad E[PL_{t+1,T}] = \frac{\mu}{\sigma}(2\Phi(\frac{\mu}{\sigma}\frac{\sqrt{T}}{\sqrt{2}})-1)$$

Similarly, if $\mu = 0$,we obtain

$$\text{(B.4.7)} \qquad E[PL_{t+1,T}] = \frac{2\rho(1-\rho^T)}{\sqrt{2\pi}\sqrt{2T(1-\rho)-2\rho(1-\rho^T)}}$$

Given that we will be using estimates of parameters of the AR(1) process, the uncertainty embedded in the estimates based on shorter periods is greater. Below we make use the Delta Theorem to approximate the volatility in our estimate of the expected P&L. For simplicity we will assume that the uncertainty arises only due to estimate $\hat{\mu}$of the mean μ. [4]Let's assume that $f(\mu) = E(PL_{t+1,T})$. In an AR(1) process, $\sqrt{T}(\hat{\mu}-\mu) \sim N(0, \sigma^2(1+\rho)/(1-\rho))$.[5] From the Delta Theorem it follows that $\sqrt{T}(f(\hat{\mu})-f(\mu)) \sim N(0, (f(\mu)')^2\sigma^2(1+\rho)/(1-\rho))$.

The derivative of the expected P&L with respect to μ can be derived straightforwardly as:

$$\text{(B.4.8)} \quad (f(\mu)') = \frac{2}{\sigma}\cdot\Phi(\frac{\mu_{d1,T}}{\sqrt{1+\sigma^2_{d1,T}}}) + 2\frac{\mu}{\sigma}f(\frac{\mu_{d1,T}}{\sqrt{1+\sigma^2_{d1,T}}})\frac{\sqrt{T}/\sigma}{\sqrt{1+\sigma^2_{d1,T}}}$$
$$-\frac{1}{\sigma} - 2\phi\frac{\sigma_{d1,T}}{\sqrt{1+\sigma^2_{d1,T}}}f(\frac{\mu_{d1,T}}{\sqrt{1+\sigma^2_{d1,T}}})\frac{\mu_{d1,T}}{1+\sigma^2_{d1,T}}\frac{\sqrt{T}}{\sigma}$$

□

[4] An alternative (at the cost of complexity) is to use the variance-covariance matrix of the estimates of the autoregressive process $(\hat{\mu}, \hat{\rho}, \hat{\sigma}^2)$ that can be obtained from Maximum Likelihood estimation and apply the Delta theorem accordingly.

[5] See [52]

B.5. Expected Transaction Costs when the Asset's Return is an AR(1) Process

B.5.1. Expected Running Costs.

PROOF. The running costs are proportional to the absolute nominal value of the position that we hold every day. If $RU_{t,T}$ denotes the running costs at time t for a signal based on a lookback of T days and RC stands for the per unit running cost then $RU_{t,T} = Abs(S_{t,T}) \cdot RC/\sigma$. Subsequently,

(B.5.1)

$$E(RU_{t,T}) = (P(S_{t,T} > 0)E(S_{t,T}|S_{t,T} > 0) + P(S_{t,T} < 0)E(-S_{t,T}|S_{t,T} < 0)) \cdot \frac{RC}{\sigma}$$

Introducing the standard normal variable $Z \sim N(0,1)$,we obtain

(B.5.2)

$$\begin{aligned} P(S_{t,T} > 0)E(S_{t,T}|S_{t,T} > 0) &= P(d1_{t,T} > 0)E(2\Phi(d1_{t,T}) - 1|d1_{t,T} > 0) \\ &= 2P(d1_{t,T} > 0)E(\Phi(d1_{t,T})|d1_{t,T} > 0) - P(d1_{t,T} > 0) \\ &= 2P(Z < d1_{t,T}, d1_{t,T} > 0) - (1 - \Phi(-\mu_{d1,T}/\sigma_{d1,T})) \end{aligned}$$

We have shown shown in the previous section that if returns follow an AR(1) process:

$$d1_{t,T} \sim N(\mu_{d1,T}, \sigma^2_{d1,T}) = N(\frac{\sqrt{T}\mu}{\sigma}, \frac{T(1-\rho^2)-2\rho(1-\rho^T)}{T(1-\rho)^2})$$

It follows that

(B.5.3) $$X = Z - d1_{t,T} \sim N(-\mu_{d1,T}, \sigma^2_{d1,T}+1)$$

(B.5.4) $$Cov(X, d1_{t,T}) = Cov((Z-d1_{t,T}), d1_{t,T}) = -Var(d1_{t,T})$$

(B.5.5) $$Correlation - \rho(X, d1_{t,T}) = -\sigma_{d1,T}/\sqrt{\sigma^2_{d1,T}+1}$$

It follows that

(B.5.6)

$$P(Z < d1_{t,T}, d1_{t,T} > 0) = P(Z < d1_{t,T}) - P(Z < d1_{t,T}, d1_{t,T} < 0)$$

$$= P(X < 0) - P(X < 0, d1_{t,T} < 0)$$

$$= \Phi(\frac{\mu_{d1,T}}{\sqrt{\sigma^2_{d1,T}+1}}) - BvN\left(\frac{\mu_{d1,T}}{\sqrt{\sigma^2_{d1,T}+1}}, -\frac{\mu_{d1,T}}{\sigma_{d1,T}}; corr = -\frac{\sigma_{d1,T}}{\sqrt{\sigma^2_{d1,T}+1}}\right)$$

where $BvN(U,W;\rho)$ stands for the c.d.f of the standard bivariate normal distribution with correlation ρ evaluated at W.

Similarly,

(B.5.7)

$$P(S_{t,T}<0)E(-S_{t,T}|S_{t,T}<0) = -P(d1_{t,T}<0)E(2\Phi(d1_{t,T})-1|d1_{t,T}<0)$$

$$= -2BvN\left(\frac{\mu_{d1,T}}{\sqrt{\sigma^2_{d1,T}+1}}, -\frac{\mu_{d1,T}}{\sigma_{d1,T}}; corr = -\frac{\sigma_{d1,T}}{\sqrt{\sigma^2_{d1,T}+1}}\right) + \Phi(-\frac{\mu_{d1,T}}{\sigma_{d1,T}})$$

Therefore,

(B.5.8) $E[RU_{t,T}] =$

$$(2\Phi(\frac{\mu_{d1,T}}{\sqrt{\sigma^2_{d1,T}+1}}) + 2\Phi(\frac{-\mu_{d1,T}}{\sigma_{d1,T}}) -$$

$$4\cdot BvN(\frac{\mu_{d1,T}}{\sqrt{\sigma^2_{d1,T}+1}}, -\frac{\mu_{d1,T}}{\sigma_{d1,T}}; corr = -\frac{\sigma_{d1,T}}{\sqrt{\sigma^2_{d1,T}+1}}))$$

$$-1\frac{RC}{\sigma}$$

Under simplified assumptions that $\mu = 0$ and $\rho = 0$ (i.e. returns are a Gaussian noise), it follows that

(B.5.9) $$E(RU_{t,T}) = -2\frac{asin(-\frac{1}{\sqrt{2}})}{\pi}\frac{RC}{\sigma} = \frac{1}{2}\cdot\frac{RC}{\sigma}$$

Note that in this case the expected running costs are independent of the lookback period. For example, if assume 10bp running fee per year and a volatility of 10%, the expected running costs are 0.3% per year. □

B.5.2. Expected Execution Costs.

PROOF. The execution costs are linked to the absolute value of the change in nominal position. If $XC_{t,T}$ denotes the execution costs at time t for a signal based on a lookback period of T and EC is the per unit execution cost then $XC_{t,T} = Abs(S_{t,T} - S_{t-1,T}) \cdot \frac{EC}{\sigma}$.

Let start by analyzing the case when $S_{t,T} > S_{t-1,T}$. We are interested in the expression below:

$$\text{(B.5.10)} \quad E(S_{t,T} - S_{t-1,T} | S_{t,T} > S_{t-1,T}) P(S_{t,T} > S_{t-1,T})$$

$$= 2E(\Phi(d1_{t,T}) - \Phi(d1_{t-1,T}) | d1_{t,T} > d1_{t-1,T}) P(d1_{t,T} > d1_{t-1,T})$$

$$= 2P(d1_{t-1,T} < Z < d1_{t,T}) = 2P(d1_{t-1,T} - Z < 0, d1_{t,T} - Z > 0)$$

where $Z \sim N(0,1)$

Let's assume that returns follow an AR(1) process, i.e. $R_t = a + \rho R_{t-1} + \varepsilon_t$ where $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$. It follows that $R_t = N\left(\frac{a}{1-\rho}, \frac{\sigma_\varepsilon^2}{1-\rho^2}\right) \sim N(\mu, \sigma^2)$. Previously we have shown that

$$d1_{t,T} \sim N(\mu_{d1,T}, \sigma^2_{d1,T}) \sim N\left(\frac{\sqrt{T}\mu}{\sigma}, \frac{T(1-\rho^2)-2\rho(1-\rho^T)}{T(1-\rho)^2}\right)$$

Let denote $X = d1_{t,T} - Z$ and $Y = d1_{t-1,T} - Z$. It follows that $X \sim N(\mu_{d1,T}, \sigma^2_{d1,T}+1)$ and $Y \sim N(\mu_{d1,T}, \sigma^2_{d1,T}+1)$.

Subsequently,

$$\text{(B.5.11)} \quad Cov(d1_{t,T}, d1_{t-1,T}) = Cov\left((\frac{R_t - R_{t-T}}{\sigma\sqrt{T}} + d1_{t-1,T}), d1_{t-1,T}\right)$$
$$= \frac{1}{\sigma\sqrt{T}}E(R_t d1_{t-1,T}) - \frac{1}{\sigma\sqrt{T}}E(R_{t-T} d1_{t-1,T}) + \sigma^2_{d1,T}$$

Proceeding further,

$$\text{(B.5.12)} \quad \frac{1}{\sigma\sqrt{T}}Cov(R_t, d1_{t-1,T}) = \frac{\sum_{s=t-T}^{t-1}(R_t R_s)}{\sigma^2 T}$$
$$= \frac{(\rho+\rho^2+...+\rho^T)\sigma^2 + T\mu^2}{\sigma^2 T} = \frac{\rho(1-\rho^T)}{T(1-\rho)} + \frac{\mu^2}{\sigma^2}$$

Similarly,

$$\text{(B.5.13)} \quad \frac{1}{\sigma\sqrt{T}}Cov(R_{t-T}, d1_{t-1,T}) = \frac{\sum_{s=t-T}^{t-1}(R_{t-T} R_s)}{\sigma^2 T} + \frac{\mu^2}{\sigma^2}$$
$$= \frac{(\rho+\rho^2+...+\rho^{T-1})\sigma^2 + T\mu^2}{\sigma^2 T} = \frac{(1-\rho^T)}{T(1-\rho)} + \frac{\mu^2}{\sigma^2}$$

It follows that

(B.5.14)
$$Cov(d1_{t,T}, d1_{t-1,T}) = -\frac{1-\rho^T}{T} + \sigma^2_{d1,T}$$

(B.5.15)
$$Cov(X,Y) = Cov(d1_{t,T}, d1_{t-1,T}) + 1$$

(B.5.16)
$$Corr(X,Y) = 1 - (\frac{1-\rho^T}{T})/(\sigma^2_{d1,T} + 1)$$

We can proceed similarly for the case when the position is decreasing:

(B.5.17)
$$E(S_{t-1,T} - S_{t,T}|S_{t,T} < S_{t-1,T})P(S_{t,T} < S_{t-1,T}) = 2P(X < 0, Y > 0)$$

Subsequently, making use of the bivariate normal distribution:

(B.5.18) $$E(XC_{t,T}) = 2\cdot(P(Y<0,X>0)+P(X<0,Y>0))\cdot\frac{EC}{\sigma}$$

$$= 4\frac{EC}{\sigma}$$

$$\left(\Phi(\frac{-\mu_{d1,T}}{\sqrt{\sigma^2_{d1,T}+1}}) - BvN\left(\frac{-\mu_{d1,T}}{\sqrt{\sigma^2_{d1,T}+1}}, \frac{-\mu_{d1,T}}{\sqrt{\sigma^2_{d1,T}+1}}; corr = 1 - \frac{(\frac{1-\rho^T}{T})}{1+\sigma^2_{d1,T}}\right)\right)$$

In the special case when returns are a Gaussian white noise, it follows that $Corr(d1_{t,T}, d1_{t-1,T}) = 1 - 1/2T$. Using the properties of the bivariate normal distribution, in this case we obtain:

(B.5.19)

$$E(XC_{t,T}) = 4(0.25 - asin(1-\frac{1}{2T})\cdot\frac{1}{2\pi})\frac{EC}{\sigma} = \frac{2EC}{\pi\sigma}acos(1-\frac{1}{2T})$$

□

B.6. P&L Volatility under AR(1) Return Dynamics

The derivation of the P&L volatility under the general assumption of an AR(1) return process is quite evolved. We prefer to evaluate numerically $E(PL^2_{t,T})$ when needed.

It is straightforward to calculate the P&L volatility when return process is a Gaussian white noise with a drift ($\rho = 0$). Let's again use the notations $X = \frac{R_{t+1}}{\sigma} \sim N(\mu/\sigma, 1)$ and $Y = d1_{t,T} \sim N(\mu_{d1,T}, \sigma^2_{d1,T})$.It follows that

(B.6.1)

$$E(PL_{t+1,T}^2) \;=\; E(x^2(2\Phi(y)\;-\;1)^2) \;=\; E(x^2)E(4\Phi(y))^2\;-\;4\Phi(y)\;+\;1)$$

(B.6.2)
$$E(x^2) = 1 + (\frac{\mu}{\sigma})^2$$

Let $Y^* \sim N(0,1)$ and making use of formulas 10010.8 and 20010.3 in [51]:

(B.6.3)

$$E(\Phi(Y)^2) = \int_{-\infty}^{+\infty} \Phi(y)^2 f(y)dy = \int_{-\infty}^{+\infty} (\Phi(\mu_{d1,T} + \sigma_{d1,T}y^*))^2 f(y^*)dy^*$$

$$= BvN(\frac{\mu_{d1,T}}{\sqrt{1+\sigma_{d1,T}^2}}, \frac{\mu_{d1,T}}{\sqrt{1+\sigma_{d1,T}^2}}; corr = \frac{\sigma_{d1,T}^2}{(1+\sigma_{d1,T}^2)})$$

(B.6.4)
$$E(\Phi(Y)) = \Phi(\frac{\mu_{d1,T}}{\sqrt{1+\sigma_{d1,T}^2}})$$

If returns are a Gaussian white noise ($\mu = 0$ and $\rho = 0$), the P&L volatility is independent of the lookback period.

(B.6.5)

$$Var(PL_{t,T}) = E(PL_{t,T}^2) = 4BvN(0,0;corr = \frac{1}{2}) - 4\Phi(0) + 1 = \frac{2asin(\frac{1}{2})}{\pi} = \frac{1}{3}$$

B.7. Theoretical Correlation Matrix

TABLE 2. Theoretical correlation matrix for various lookback periods

Period	1	2	4	8	16	32	64	126	252	504
1	1.00	0.69	0.48	0.34	0.24	0.17	0.12	0.09	0.06	0.04
2		1.00	0.69	0.48	0.34	0.24	0.17	0.12	0.09	0.06
4			1.00	0.69	0.48	0.34	0.24	0.17	0.12	0.09
8				1.00	0.69	0.48	0.34	0.24	0.17	0.12
16					1.00	0.69	0.48	0.34	0.24	0.17
32						1.00	0.69	0.49	0.34	0.24
64							1.00	0.70	0.49	0.34
126								1.00	0.69	0.48
252									1.00	0.69
504										1.00

B.8. Empirical Correlation Matrix

TABLE 3. Empirical correlation matrix for various lookback periods

Period	1	2	4	8	16	32	64	126	252	504
1	1.00	0.70	0.49	0.35	0.25	0.17	0.12	0.07	0.03	-0.01
2		1.00	0.69	0.50	0.35	0.25	0.18	0.10	0.05	0.01
4			1.00	0.70	0.50	0.36	0.25	0.15	0.09	0.03
8				1.00	0.71	0.51	0.37	0.25	0.17	0.09
16					1.00	0.72	0.53	0.38	0.27	0.16
32						1.00	0.74	0.55	0.41	0.26
64							1.00	0.75	0.56	0.38
126								1.00	0.74	0.53
252									1.00	0.74
504										1.00
△[6]	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.04	0.04

[6] Average absolute deviation from theoretical values

B.9. Average Transaction Costs

TABLE 4. Average Transaction Costs

Periods	**Equities**		**FX**		**Commodities**		**Fixed Income**	
	Exe Cost (bps)	Run Cost (bps)	Exe Cost (bps)	Run Cost (bps)	Exe Cost (bps)	Run Cost (bps)	Exe Cost (bps)	Run Cost (bps)
Before 1993	20.0	44.0	18.8	32.0	17.2	39.6	18.4	33.6
1993-2002	7.5	16.5	7.0	12.0	6.5	14.9	6.9	12.6
Since 2003	5.00	11.00	4.7	8.0	4.3	9.9	4.6	8.4

Bibliography

[1] Jegadeesh, N. and Titman, S. (1993), “Returns to buying winners and selling losers: Implications for stock market efficiency”, Journal of Finance 48(1), 65–91.

[2] C4. 5: programs for machine learning. Morgan Kaufmann, 1993.

[3] Quinlan, J. R. 1986. Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81–106

[4] Yoav Freund and Robert Schapire. A decision-theoretic generalization of online learning and an application to boosting. In Computational learning theory, pages 23–37. Springer, 1995.

[5] Albert Bifet and Richard Kirkby. Data stream mining a practical approach. 2009.

[6] Kernel Methods for Pattern Analysis. Cambridge University Press.

[7] Vladimir Vapnik. The Nature of Statistical Learning Theory. Springer, 1995

[8] Yiming Yang and Xin Liu. A re-examination of text categorization methods. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and development in information retrieval, pages 42–49. ACM, 1999.

[9] Lean Yu, Shouyang Wang, and Kin Keung Lai. Foreign-Exchange-Rate Forecasting with Artificial Neural Networks. International Series in Operations Research & Management Science. Springer, 2007.

[10] Arie Ben-David. About the relationship between ROC curves and Cohen's kappa. Engineering Applications of Artificial Intelligence, 21(6):874–882, 2008.

[11] Charles Elkan. The foundations of cost-sensitive learning. In International Joint Conference on Artificial Intelligence, volume 17, pages 973–978. Citeseer, 2001

[12] A Philip Dawid and Vladimir G Vovk. Prequential probability: Principles and properties. Bernoulli, 5(1):125–162, 1999.

[13] Joao Gama and Pedro Rodrigues. Stream-based electricity load forecast. Knowledge Discovery in Databases: PKDD 2007, pages 446–453, 2007.

[14] Mark G Kelly, David J Hand, and Niall M Adams. The impact of changing populations on classifier performance. In Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 367– 371. ACM, 1999.

[15] Lukas Menkhoff. The use of technical analysis by fund managers: International evidence. Journal of Banking & Finance, 34(11):2573–2586, 2010.

[16] Fama EF, French KR. Dissecting anomalies. The journal of finance. 2008;63(4):1653-1678.

[17] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[18] Hornik, K.,M. Stinchcombe, and H. White. 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2:359–66.

[19] Option pricing and Esscher transform under regime switching

[20] Merton, R. C. Option Pricing When Underlying Stock Returns Are Discontinuous. Journal of Financial Economics 1976, 3 (1), 125–144 DOI: 10.1016/0304-405X(76)90022-2.

[21] Madan, D. B, and E Seneta. 1989. "Chebyshev Polynomial Approximations for Characteristic Function Estimation: Some Theoretical Supplements." Journal of the Royal Statistical Society. Series B (Methodological) 51 (2): 281–85.

[22] Madan, Dilip, Peter Carr, and Eric Chang. 1998. "The Variance Gamma Process and Option Pricing." European Finance Review 2 (1): 79–105.

[23] Barndorff-Nielsen, O., Jensen, J., & Sørensen, M. (1998). Some stationary processes in discrete and continuous time. Advances in Applied Probability, 30(4), 989-1007. doi:10.1239/aap/1035228204

[24] Carr, Peter, and Liuren Wu. 2003. "What Type of Process Underlies Options? A Simple Robust Test." The Journal of Finance 58 (6): 2581–2610.

[25] Kou, S. G, and Hui Wang. 2004. "Option Pricing Under a Double Exponential Jump Diffusion Model." Management Science 50 (9): 1178–92. https://doi.org/10.1287/mnsc.1030.0163.

[26] Carr, Peter, Helyette Geman, Dilip B Madan, and Marc Yor. 2011. "Options on Realized Variance and Convex Orders." Quantitative Finance 11 (11): 1685–94. https://doi.org/10.1080/14697680903397675.

[27] Küchler Uwe, and Stefan Tappe. 2008. "Bilateral Gamma Distributions and Processes in Financial Mathematics." Stochastic Processes and Their Applications 118 (2): 261–83. https://doi.org/10.1016/j.spa.2007.04.006.

[28] Madan, Dilip B, and Marc Yor. 2016. "On Valuing Stochastic Perpetuities Using New Long Horizon Stock Price Models Distinguishing Booms, Busts, and Balanced Markets." Mathematical Finance 26 (2): 296–328. https://doi.org/10.1111/mafi.12056.

[29] American options with regime switching. Int J Theor Appl Finance 5, 497–514 (2002)

[30] The fine structure of asset returns: An empirical investigation. J Bus 75, 305–332 (2002)

[31] http://allstarcharts.com/ escalator-up-and-elevator-down/, http://seekingalpha.com/article/ 3746396-market-stars-ride-escalarot-elevator.

[32] Jacod, Jean, and Shiriev Al´bert Nikolaevich. 2003. Limit Theorems for Stochastic Processes. 2nd ed. Grundlehren Der Mathematischen Wissenschaften, 288. Berlin: Springer.

[33] Pricing and hedging in exponential L´evy models: review of recent results

[34] Konikov, M.,Madan,D.B.(2002) Option pricing using variance-gamma Markov chains.Rev Derivatives Res 5, 81–115

[35] Madan, D. B. (2015), "Estimating Parameteric Models of Probability Distributions," Methodology and Computing in Applied Probability, 17, 823- 831.

[36] Carrasco, M., Chernov, M., Florens, J.P., Ghysels, E (2004).: Efficient estimation of jump diffusions and general dynamic models with a continuum of moment conditions.Working paper, Department of Economics, University of Rochester 2004

[37] Carrasco, M., Florens, J.P. (2000): Generalization of GMM to a continuum of moment conditions. Econ Theory 16, 797–834 (2000)

[38] : Empirical characteristic functions in time series estimation. Econ Theory 18, 691–721 (2002)

[39] Carrasco, M., Florens (2000), J.P.: Generalization of GMM to a continuum of moment conditions. Econ Theory 16, 797–834 (2000)

[40] Tran,K.C. (1998): Estimatingmixtures of normal distributions via the empirical characteristic function. Econ Rev 17, 167–183 (1998)

[41] Elliott, R.J. (1993), New finite dimensional filters and smoothers for noisily observed Markov chains. IEEE Trans Inf Theory 39, 265–271

[42] Madan, D. B. and M. Yor (2016), On valuing stochastic perpetuities us- ing new long horizon stock price models distinguishing Booms, Busts and Balanced Markets,Mathematical Finance, 26, 296-328.

[43] Madan D. B. and E. Seneta (1990), "The variance gamma (VG) model for share market returns," Journal of Business, 63, 511-524.

[44] Madan, D. B. (2015), "Estimating Parameteric Models of Probability Dis- tributions," Methodology and Computing in Applied Probability, 17, 823- 831.

[45] Madan D., P. Carr and E. Chang (1998), "The variance gamma process and option pricing," European Finance Review, 2,79-105.

[46] Gradshteyn, I. S. and Ryzhik, I. M. (2000) Table of integrals, series and products. Academic Press, San Diego.

[47] Muller, P. H. (1991) Lexikon der Stochastik. Akademie Verlag, Berlin.

[48] D. Madan, E. Seneta, The variance gamma (V.G) model for share market returns, J. Bus. 63 (4) (1990) 135155.

[49] U. Kuchler, S. Tappe, Option pricing in bilateral gamma stock model, Statist. Decisions 27 (4) (2009) 281307.

[50] P. Carr, H. Geman, G.B. Madan, and M. Yor. The ne structure of asset returns: an empirical inverstigation. J. Business, 75(2):305

[51] Owen, D. B. (1980), “A Table of Normal Integrals”, Communications in Statistics - Simulation and Computation, 9(4):389– 419.

[52] Crack, T. and Ledoit, O. (2010). "Using Central Limit Theorems for Dependent Data"

[53] Stuart, A. and Ord, J. K. (1998), ”Kendall’s Advanced Theory of Statistics, Vol. 1: Distribution Theory”, 6th ed. New York: Oxford University Press, 1998.

[54] Rose, C. and Smith, M. D. (2002), "The Bivariate Normal”, §6.4 A in “Mathematical Statistics with Mathematica”., New York: Springer-Verlag, 2002.

[55] “Extending Owen’s integral table and a new multivariate Bernoulli distribution”

[56] , “Trend-Following, Risk-Parity and the Influence of Correlations", Chapter 3 in “Risk-Based and Factor Investing", Elsevier & ISTE Press.

[57] , "Least-squares approach to risk parity in portfolio selection”, Quantitative Finance, Volume 16, Issue 3.

[58] Bruder, B. and Roncalli, T. (2012),“Managing Risk Exposures using the Risk Budgeting Approach", Lyxor Asset Management Research Paper (http://www.thierry-roncalli.com/download/risk-budgeting.pdf).

[59] Levine, A. and Pedersen, L.(2016), “Which Trend Is Your Friend?”, Financial Analysts Journal, May/June 2016, Volume 72 Issue 3.

[60] Martin, R. and Bana, A.(2012), “Nonlinear Momentum Strategies”, RISK Magazine, Nov 2012.

[61] Martin, R. and Zou, D. (2012), “Momentum Trading: ‘Skews Me’”, RISK, Aug 2012.

[62] Jones, C. (2002), “A Century of Stock Market Liquidity and Trading Costs”, working paper, Columbia Business School.

[63] , “Building Diversified Portfolios that Outperform Out of Sample”, Journal of Portfolio Management, 2016, 59–69

[64] Baltas, N. (2015), “Trend-Following, Risk-Parity and the Influence of Correlations", Chapter 3 in “Risk-Based and Factor Investing", Elsevier & ISTE Press.

[65] Fung, W. and Hsieh, D. (2001), “The Risk in Hedge Fund Strategies: Theory and Evidence from Trend Followers”, The Review of Financial Studies, Vol. 14, No. 2. (Summer, 2001), pp. 313-341.

[66] Lemperiere, Derenble, Seager, Potters and Bouchaud (2014), “Two centuries of trend following”, Journal of Investment Strategies, 2014 Volume 3, Number 3

[67] , “A Century of Evidence on Trend-Following Investing”, forthcoming

[68] Moskowitz, T., Y.H. Ooi, and L.H. Pedersen (2012), “Time Series Momentum,” Journal of Financial Economics, 104(2), 228-250.

[69] Menkhoff, L., Sarno, L., Scmeling, M. and Schrimpf, A., (2012). “Currency Momentum Strategies”, Journal of Financial Economics

[70] Burnside, C.; Eichenbaum, M.; and Rebelo, S. (2011), “Carry trade and Momentum in currency markets”, Annual Review of Financial Economics 3(1), 511–535.

[71] John Okunev and Derek White (2003) “Do Momentum-Based Strategies Still Work in Foreign Currency Markets?”, Journal of Financial and Quantitative Analysis, 2003, vol. 38, 425–447.

[72] Asness, Clifford S.; Tobias J. Moskowitz; and Lasse Heje Pedersen (2013), “Value and Momentum Everywhere," The Journal of Finance, 68(3), 929-985

[73] Erb, Claude, and Campbell Harvey (2006), “The strategic and tactical value of commodity futures,” Financial Analysts Journal 62, 69–97.

[74] Miffre, Joëlle and Georgios Rallis (2007), “Momentum in Commodity Futures Markets,” Journal of Banking and Finance 31(6), 1863- 1886.

[75] Bishop C.M. (1995) Neural networks for pattern recognition. Oxford, UK: Oxford University Press.

[76] Breiman, L. (2001) Random forests. Machine Learning 45:5–32.

[77] Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen. (1984) Classification and regression trees. Boca, Raton, FL: CRC press.

[78] Dietterich, T. G. (2000) Ensemble methods in machine learning. In International workshop on multiple classier systems, eds. F. Schwenker, F. Roli, and J. Kittler, 1–15. New York: Springer.

[79] Dimopoulos, Y., P. Bourret, and S. Lek. (1995) Use of some sensitivity criteria for choosing networks with good generalization ability. Neural Processing Letters 2:1–4.

[80] Eldan, R., and O. Shamir. (2016) The power of depth for feedforward neural networks. In 29th Annual Conference on Learning Theory, eds.V. Feldman, A. Rakhlin, and O. Shamir, 907–40. Brookline,MA:Microtome Publishing.

[81] Fama, E. F., and K. R. French. (1993) Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33:3–56.

[82] Freund, Y. (1995) Boosting aWeak Learning Algorithm by Majority. Information and Computation 121:256–85.

[83] Freyberger, J., A. Neuhierl, and M. Weber. (2020) Dissecting characteristics nonparametrically. Review of Financial Studies 33: 2326–77.

[84] Friedman, J. (2001) Function approximation: A gradient boosting machine. Annals of Statistics 5:1189– 232.

[85] Glorot, X., A. Bordes, and Y. Bengio. (2011) Deep Sparse Rectifier Neural Networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, vol. 15, 315–323. Brookline, MA: Microtome Publishing.

[86] Goodfellow, I., Y. Bengio, and A. Courville. (2016) Deep learning. Cambridge: MIT Press.

[87] Gu, S., B. T. Kelly, and D. Xiu. (2019) Autoencoder asset pricing models. Working Paper, Yale University.

[88] Gu, S., B. T. Kelly, and D. Xiu. (2020) Empirical Asset Pricing via Machine Learning. The Review of Financial Studies, Yale University.

[89] Hansen, L. K., and P. Salamon. 1990. Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence 12:993–1001.

[90] Hastie, T., R. Tibshirani, and J. Friedman. 2009. The elements of statistical learning. New York: Springer.

[91] Heaton, J. B., N. G. Polson, and J. H. Witte. 2016. Deep learning in finance. Preprint, https:// arxiv.org/abs/1602.06561.

[92] Hinton, G. E., S. Osindero, and Y.-W. Teh. 2006. A fast learning algorithm for deep belief nets. Neural Computation 18:1527–54.

[93] Jaggi, M. 2013. An equivalence between the lasso and support vector machines. In Regularization, optimization, kernels, and support vector machines, eds. J. A. K. Suykens, M. Signoretto, and A. Argyriou, 1–26. Boca Raton, FL: CRC Press.

[94] Kelly, B., and S. Pruitt. 2013. Market expectations in the cross-section of present values. Journal of Finance 68:1721–56.

[95] Kelly, B., S. Pruitt, and Y. Su. 2019. Characteristics are covariances: A unified model of risk and return. Journal of Financial Economics.

[96] Kingma, D., and J. Ba. 2014. Adam: A method for stochastic optimization. Preprint, https:// arxiv.org/abs/1412.6980.

[97] Masters, T. 1993. Practical neural network recipes in C++. New York: Academic Press.

[98] Moritz, B., and T. Zimmermann. 2016. Tree-based conditional portfolio sorts: The relation between past and future stock returns. Working

Paper, Ludwig Maximilian University of Munich.

[99] Rapach, D., and G. Zhou. 2013. Forecasting stock returns. In Handbook of economic forecasting, eds. G. Elliott and A. Timmermann, 328–83. Amsterdam, the Netherlands: Elsevier.

[100] Rosenberg, B. 1974. Extra-Market Components of Covariance in Security Returns. Journal of Financial and Quantitative Analysis 9:263–74.

[101] Schapire, R. E. 1990. The Strength of Weak Learnability. Machine Learning 5:197–227.

[102] Tukey, J. W. 1960. A survey of sampling from contaminated distributions. In Contributions to probability and statistics, eds. I. Olkin, S. G. Ghurye, W. Hoeding, W. G. Madow, and H. B. Mann. Stanford, CA: Stanford University Press.

[103] Wilson, D. R., and T. R. Martinez. 2003. The General Inefficiency of Batch Training for Gradient Descent Learning. Neural Networks 16:1429–51.

[104] Yao, J., Y. Li, and C. L. Tan. 2000. Option price forecasting using neural networks. Omega 28:455–66.

[105] Arthur E Hoerl and Robert W Kennard. “Ridge regression: Biased estimation for nonorthogonal problems.” Technometrics, 12(1):55–67, 1970.

[106] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. “Least angle regression.” The Annals of statistics, 32(2):407–499,

2004.

[107] Robert Tibshirani. “Regression shrinkage and selection via the lasso.” Journal of the Royal Statistical Society. Series B (Methodological), pp. 267–288, 1996.

[108] Hui Zou and Trevor Hastie. “Regularization and variable selection via the elastic net.” Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2):301–320, 2005.

[109] . “The regression analysis of binary sequences.” Journal of the Royal Statistical Society. Series B (Methodological), pp. 215–242, 1958.

[110] David W Hosmer Jr and Stanley Lemeshow. Applied logistic regression. John Wiley & Sons, 2004.

[111] Svante Wold, Arnold Ruhe, Herman Wold, and WJ Dunn, III. “The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses.” SIAM Journal on Scientific and Statistical Computing, 5(3):735–743, 1984.

[112] James H Stock and Mark W Watson. “Macroeconomic forecasting using diffusion indexes.” Journal of Business & Economic Statistics, 20(2):147–162, 2002.

[113] . “A training algorithm for optimal margin classifiers.” In Proceedings of the fifth annual workshop on Computational learning theory, pp. 144–152. ACM, 1992.

[114] Corinna Cortes and Vladimir Vapnik. “Support-vector networks.” Machine learning, 20(3):273–297, 1995.

[115] Alex Smola and Vladimir Vapnik. “Support vector regression machines.” Advances in neural information processing systems, 9:155–161, 1997.

[116] Biau, G. 2012. Analysis of a Random Forests Model. Journal of Machine Learning Research 13:1063 1095.

[117] Scornet, E., G. Biau, and J.-P. Vert. 2015. Consistency of random forests. Annals of Statistics 43:1716{1741. URL https://doi.org/10.1214/15-AOS1321.

[118] . “Random forests.” Machine learning, 45(1):5–32, 2001

[119] B¨uhlmann, P., and B. Yu. 2003. Boosting With the L2 Loss. Journal of the American Statistical Association 98:324–339.

[120] Chen, T., and C. Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16, pp. 785–794. New York, NY, USA: ACM. URL http://doi.acm.org/10.1145/2939672. 2939785.

[121] Fan, J., C. Ma, and Y. Zhong. 2019. A Selective Overview of Deep Learning. Tech. rep., Princeton University.

[122] Eldan, R., and O. Shamir. 2016. The Power of Depth for Feedforward Neural Networks. In V. Feldman, A. Rakhlin, and O. Shamir (eds.), 29th Annual Conference on Learning Theory, vol. 49 of Proceedings of Machine Learning Research, pp. 907–940. Columbia University, New York, New York, USA: PMLR. URL

http://proceedings.mlr.press/v49/eldan16.html.

[123] 2017. Why Does Deep and Cheap Learning Work So Well? Journal of Statistical Physics 168:1223–1247. URL https://doi.org/10.1007/s10955-017-1836-5.

[124] Rolnick, D., and M. Tegmark. 2018. The power of deeper networks for expressing natural functions. In ICLR.

[125] James M Hutchinson, Andrew W Lo, and Tomaso Poggio. "A nonparametric approach to pricing and hedging derivative securities via learning networks." The Journal of Finance, 49(3):851–889, 1994.

[126] Mary Malliaris and Linda Salchenberger. "Using neural networks to forecast the S&P 100 implied volatility." Neurocomputing, 10(2):183–195, 1996.

[127] M Qi and GS Maddala. "Option pricing using artificial neural networks: the case of S&P 500 index call options." In Neural Networks in Financial Engineering: Proceedings of the Third International Conference on Neural Networks in the Capital Markets, pp. 78–91. New York: World Scientific, 1996.

[128] P Lajbcygier, C Boek, M Palaniswami, and A Flitman. "Neural network pricing of all ordinaries SPI options on futures." In 3rd Conference on Neutral Networks in the Capital Markets, 1995.

[129] Matthew Francis Dixon, Diego Klabjan, and Jin Hoon Bang. "Classificationbased Financial Markets Prediction using Deep Neural Networks." Available at SSRN 2756331, 2016.

[130] JB Heaton, NG Polson, and JH Witte. “Deep Learning in Finance.” arXiv preprint arXiv:1602.06561, 2016.

[131] Michael Johannes, Arthur Korteweg, and Nicholas Polson. “Sequential learning, predictability, and optimal portfolio returns.” The Journal of Finance, 69(2):611– 644, 2014.

[132] Liu, Laura Xiaolei and Lu Zhang, (2008), “Momentum Profits, Factor Pricing, and Macroeconomic Risk,” Review of Financial Studies 21 (6), 2417-2448

[133] Baltas, A. N.; and Kosowski, R. (2012), “Improving Time Series Momentum strategies: The role of trading signals and volatility estimators,” Working paper: Available at SSRN

[134] Hong, H. and Stein, J. C. (1999), “A unified theory of underreaction, Momentum trading, and overreaction in asset markets”, Journal of Finance 54(6), 2143–2184