Essays on Financial Market

Pushpak S Sarkar

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Yu-Chin Chen, Chair

Eric W Zivot, Chair

Chang Jin Kim

Program Authorized to Offer Degree:
Economics

University of Washington

Abstract

Essays on Financial Market

Pushpak S Sarkar

Co-Chairs of the Supervisory Committee:
Associate Professor Yu-Chin Chen
Department of Economics

Professor Eric W Zivot
Department of Economics

In this dissertation I explore three independent questions related to the financial market in three separate chapters. Below I give a brief outline of each of the chapters.

In the first chapter I explore the "tail dependence" of commodities with other financial assets such as equities, foreign exchange and bonds and use copula to model the time-varying "tail dependence" between commodities and those assets. I show a clear evidence that using copula-based approach we can produce better forecasts of tail-based risk measures such as Value-at-Risk (VaR) and Expected Shortfall (ES) for portfolios consisting of commodities and other financial assets. The results of the statistical tests for evaluating the VaR and ES show clear evidence in favor of time-varying Student's t copula model when compared against the benchmark Dynamic Conditional Correlation (Engel [2002]) and RiskMetrics™ (Morgan et al. [1996]) models. I also look at the effect of tail dependence on optimal portfolio construction by minimizing portfolio expected shortfall and then compare the performance of expected shortfall strategy vs. global minimum variance strategy. For dynamic optimization, the expected shortfall strategy generates a higher portfolio cumulative return than that generated by the global minimum variance strategy.

The second chapter describes a study which is a joint work with Professor R. Douglas Martin.[1] In this study we explore the question of portfolio optimization based on downside risk estimates. It is well known that the mean-variance portfolio optimization does not take into account the skewness and kurtosis of returns distribution. But asset returns data exhibit non-normality in terms of skewness and fat tails. Also, variance is a symmetric measure as it penalizes both negative and positive returns equally. To penalize only the negative returns, investors may profit by using portfolio optimization based on downside risk measures. One of the most popular downside risk measures is Expected Shortfall (ES) or Conditional Value at-Risk (CVaR). ES is defined as the average of loss beyond the Value at-Risk (VaR) and captures the tail characteristics. Expected shortfall gained popularity as the objective risk measure in portfolio optimization with the publication of the seminal paper of Rockafellar et al. [2000] which proposed an optimization algorithm which can be solved by standard linear programming. Krokhmal [2007] further extended the idea and proposed Higher Moment Coherent Risk (HMCR) measures and showed how the HMCR measures can be implemented by reducing it to a p-order conic programming and approximating via linear programming. Krokhmal [2007] discussed the special case where $p = 2$ defined it as the Second Moment Coherent Risk Measure (SMCR). The SMCR has similar properties as CVaR but it measures risk in terms of the second moments of loss distributions. In this paper, we call the Second Moment Coherent Risk as Expected Quadratic Shortfall (EQS), which is a natural variant of ES. We construct the following three types of global minimum risk optimal portfolio using daily returns of 30 small cap stocks (which has the largest market capitalization within the small cap category) - a) Global Minimum Variance (GMV), b) Global Minimum Expected

---

[1]Professor Emeritus, Professor of Statistics, Adjunct Professor of Finance, Former Chair of the Department of Statistics, Founder Director of the Computational Finance and Risk Management Program (Applied Mathematics), University of Washington, Seattle, WA 98195-3330. Email: doug@amath.washington.edu

Shortfall (GMES), and c) Global Minimum Expected Quadratic Shortfall (GMEQS). We conduct both the static as well as dynamic (i.e. with rebalancing) portfolio optimization and analyze the portfolio performance metrics such as Sharpe Ratio (SR), Downside Sharpe Ratio (DSR), Expected Shortfall Ratio (ES Ratio), cumulative return, cumulative return relative to a benchmark and drawdown.

The third chapter describes a study which is a joint work with Professor Yu-Chin Chen [2] and Zihao Chen.[3] It uses machine learning techniques to re-examine the long-standing difficulty in predicting currency returns with macroeconomic indicators by focusing on three possible causes: the general lack of information in the macro predictors, mis-specifications in the forecasting equations, and inherent instabilities in the relationship between the exchange rate and its macro determinants. Using a large international dataset that captures current macroeconomic conditions as well as forward-looking market expectations and perceived uncertainties, we forecast monthly returns from 1995 onward of four major currencies (AUS, CAD, GBP, and JPY) against the USD. In in-sample regressions, we see that while market expectations embedded in derivatives markets may help predict subsequent exchange rate returns, there is little evidence that they contain predictive content above and beyond what is in the macro indicators themselves. Moreover, both types of predictors perform better in non-linear specifications than under the linear specifications which often deliver adjusted R^2 around zero. We take these findings as indicative that the exchange rate is not disconnected to indicators of the macroeconomy - be their current values or expectations, though their functional relation may be more nuanced than simple linear specifications can capture. Moving the analyses to pseudo out-of-sample (OOS) forecasts, we find that a Multilayer Per-

---

[2]Associate Professor, Department of Economics, University of Washington, Seattle, WA 98195-3330. Email: yuchin@uw.edu

[3]Graduate Student, Department of Economics, University of Washington, Seattle, WA 98195-3330. Email: zchen05@uw.edu

ceptron Neural Network can generate improvements over the long-standing Random Walk benchmark, some of which are over 10% and statistically significant. More prominently, we see that the majority of the ML methods considered do not outperform a RW forecast given our small sample context. In fact, unlike results for other asset returns, ML does not appear to help resolve the FX forecasting puzzle. Nevertheless, our ML explorations unveil significant empirical instabilities, especially around the GFC period. These findings support the views that pseudo-OOS exchange rate forecasting in finite samples can be overwhelmed by inherent statistical issues such as parameter and model instabilities, and that the exchange rate dynamics are inherently difficult to distinguish from a RW process statistically (e.g. Engel and West [2005]).

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to thank my committee members Professor Yu-Chin Chen, Professor Eric Zivot, Professor Chang-Jin Kim and Professor Nathalie Williams for their continued guidance and support throughout this journey. I have received very valuable and insightful feedback from each of them. Professor Zivot's graduate level Financial Econometrics course piqued my interest about Copula and its application in the context of financial assets. Professor Zivot has always encouraged me to try different things in my academic pursuit. He was generous enough to offer me a research assistantship which gave me the much needed financial support. It was Professor Chen who has constantly encouraged and guided me throughout my graduate school journey by giving me the best advice based on my strengths and weaknesses. The feedback I received from her during our numerous meetings has shaped this dissertation. I express my deep gratitude to Professor Chen and fellow graduate student Zihao Chen with whom I got an opportunity to work on our joint project. To work under Professor Chen's guidance and mentorship in this project has been an incredible learning opportunity. Zihao has contributed enormously in the project and also helped me finish the manuscript in time. I am also deeply grateful to Professor R. Douglas Martin for his guidance and support in developing the second chapter of this dissertation. This is an ongoing project and I am looking forward to learn more as we continue our research in this topic. I would also like to thank all of my friends who stood by me through thick and thin. I am in debt to them for their generosity.

Finally, this would not have been possible without the love and support of my parents and my brother Pallab. At moments of self-doubt, it was their trust which kept me going.

DEDICATION

To my parents and my brother,

for their love,

and their trust in my abilities.

Chapter 1

MODELING DEPENDENCY OF COMMODITIES WITH OTHER FINANCIAL ASSETS: A COPULA APPROACH

1.1 Introduction

From the early part of the decade of 2000's, commodities started to emerge as an alternative asset class apart from the typical financial assets available on the market. This is corroborated by the fact that from around 2004, institutional investors started investing in the commodity futures in order to reap the benefits of diversification. As noted in Basak and Pavlova [2016] and reported in the CFTC report (2008), the institutional investment surged from $15 billion in 2003 to more than $200 billion in 2008. As a result the commodity prices have undergone dramatic changes in the last decade. Including commodities in long-term asset allocation plan is beneficial for an investor as it offers diversification benefits and it also acts as a hedge against inflation. In the literature, a number of papers discussed this emergence of commodities as an alternative asset class such as Buyuksahin and Robe [2014], Singleton [2014b], Basak and Pavlova [2016] to name a few. The commodity futures index has turned out to be a very popular investment vehicle for the investors. Tang and Xiong [2012] showed that after 2004, the commodities included in these type of indices showed more correlation with the equity market than non-index commodities.

As investment in commodities becomes increasingly popular among the investors, it is a natural question to ask that how commodities are connected with other financial assets available to the investors in general. Because this connection or dependence is a crucial factor in deciding how events originating in one asset market affect the other asset markets. The main objective of this paper is to model the *tail dependence* of commodities with other financial

assets such as stocks, foreign exchange, and bonds with the help of *Copula*. One important stylized fact of the returns of these financial assets is that they are non-gaussian i.e. their distributions are asymmetric or skewed and have fat tails. These returns also have conditional heteroskedasticity (i.e. changing or time-varying volatility) which means that the conditional distribution is also non-normal. If we consider the co-movement, another important feature is that the returns of these assets may show asymmetric dependence i.e. they may show a higher correlation during market downturns than market upturns as reported by Longin F. Solnik [2001], and Ang and Chen [2002]. Risk averse investors are more interested to know the co-movement of the asset returns in an environment of extreme downside event. So it might be more useful to look at the tail behavior of the asset returns rather than focusing solely on the center of the returns distribution. Naturally, multivariate normal distribution is not a good technique to model the joint distribution of these financial returns.

As explained in Patton [2013] copula gives us an alternative way to model the joint distributions of random variables with greater flexibility both in terms of marginal distributions and the dependence structure. With the help of copula, we can specify the marginal distributions of individual random variables separately from the dependence structure that links these distributions to form the joint distribution. To be precise, this dependence structure is the *copula*. Copulas have been used in financial literature for quite some time, for example, see Embrechts et al. [2002], Cherubini et al. [2004], Patton [2006a], Fernandez [2008], Patton [2013].

To understand how commodities evolved as an asset class and what is its impact on portfolio diversification, it calls for exploring the dependence of commodities with other financial assets such as equities, foreign exchange, and bonds. From the risk management perspective it is important to check whether the tails of commodities and other financial asset classes move together. In order to do so, copula approach seems to be a natural choice as it gives us the flexible way of modeling univariate marginal distributions and the joint distribution

of these asset returns in a separate manner, while it also allows us pay special attention to the tails of these assets.

In this paper I model the dependency of commodities (CP) with equities (EQ), foreign exchange (FX) and bonds (BOND) focusing on the U.S. financial market. To model the dependency of commodities with other financial assets, I use both static copula and time-varying dynamic copula models. To model the dependency between CP & EQ, CP & FX, and CP & BOND, bivariate copula models are used. To capture any tail dependence, the copulas which have been used in this paper are Normal copula, Clayton copula, Rotated Clayton copula, and Student's t copula. To model the time varying dependency between CP & EQ and CP & FX, the condtional copula proposed by Patton [2013] are estimated based on the Student's t Copula. Once I estimate the copula models, I show the effectiveness of these models by showing the risk management application based on portfolios comprising of two assets (CP & EQ, CP & FX and CP & BOND), three assets (CP, EQ & FX) and four assets (CP, EQ, FX & BOND). Using the estimated time varying Student's t copula I compute the Value-at-Risk (VaR) and Expected Shortfall (ES) for these portfolios. Through backtesting of the estimated VaR and ES, I show that Student's t copula model produces better forecasts of VaR and ES in comparison to Normal Dynamic Conditional Correlation (Engel [2002]) and RiskMetrics™ (Morgan et al. [1996]) models. This points to the fact that capturing the time varying tail dependency of commodities with other financial assets is important from risk management perspective and copula is an appropriate tool in this context.

While a lot of papers in the literature looked into the correlations between equity returns or foreign exchange returns or bond returns, much less attention has been paid to the commodities market. An investor in not only interested in the upside potential of commodities but also cares about the downside protection. In that context both upper and lower tail dependence between commodities and other financial assets are important. It has not come to my attention whether any study has explored the dependence between commodities and

other financial assets from that point of view. This paper tries to fill in that gap. The main contribution of this paper is that it empirically studies the time varying dependency of commodities with other asset classes such as equities, foreign exchange and bonds by adopting the copula approach and then show applications in risk management and portfolio optimization. While most of the papers in the related literature explore the bivariate copula, I extend this to three dimension and four dimension Student's t copula. I also show that using the time varying Student's t copula model, we can produce better forecasts of tail-based risk measures such as Value-at-Risk (VaR) and Expected Shortfall (ES) for portfolios consisting of commodities, equities, foreign exchange and bonds. I also show that using expected shortfall as the objective risk measure in optimal portfolio construction, we can benefit in terms of higher cumulative portfolio return over time.

The paper is organized in the following way. Section 2 reviews the literature. Section 3 gives a brief review of the copula. Section 4 describes the data used in the exercise and reports the summary statistics. In section 5 estimation of univariate marginal distribution is discussed. In section 6, I discuss about dependence measures such as quantile dependence, tail dependence and exceedance correlation. In section 7, I turn to parametric estimation of constant copulas for full sample and sub-samples. In section 8, I test for structural break in the time varying rank correlation. In section 9, I estimate the time varying Student's t copula. Section 10 discusses about the risk measures such as Value-at-Risk (VaR) and Expected Shortfall (ES) of portfolios. In section 11 and 12, I conduct the backtesting of the estimated VaR measures and in section 13 and 14, I do backtesting of ES. In section 15, I show the portfolio optimization based on ES. Section 16 concludes.

1.2 Review of Literature

Many institutions started to consider commodities as a new asset class when the equity market collapsed in 2000. This was also influenced by the widely publicized discovery of a negative correlation between commodity returns and stock returns reported by Greer [2000],

Gorton and Rouwenhorst [2006], and Erb and Harvey [2006]. Greer [2000] argued that passive unleveraged long-only commodity indices enable an investor to access the commodity returns and commodities also act as a very good hedge hedge against inflation. Because the prices of goods increase, the prices of commodities used to produce those goods also increase. Some commodities such as precious metals and energy products act as a good hedge in this context. This is relevant when we consider that the equity returns may be affected by unexpected inflation. Greer [2000] also points out that commodity returns are negatively correlated with equity and bonds returns. So including commodities in a portfolio offers the diversification benefits i.e. inclusion of commodities in a portfolio will reduce portfolio volatility. These observations find support in Gorton and Rouwenhorst [2006] as they show that, while commodity futures returns are negatively correlated with equity and bond returns, they are positively correlated with inflation, unexpected inflation and changes in expected inflation. Another important fact which plays an important role in the portfolio diversification is that the commodity returns distribution is positively skewed while the equity returns distribution is negatively skewed. It means commodities have less downside risk. Erb and Harvey [2006] argued that annualized excess return of a portfolio of commodity futures can offer returns equivalent to equity portfolios. Tang and Xiong [2012] explored the effects of *financialization* process of commodities which started after the collapse of the equity market in 2000. They found that after 2004 commodity futures index emerged as a popular investment vehicle for the investors. Through the *financialization* process and index investment, prices of non-energy commodities became more correlated with oil prices. They also found that commodity prices became more correlated with the prices of other traditional financial assets and prices of different commodities also showed higher correlation amongst themselves. Singleton [2014a] studied and explored and showed empirical evidence of the impact of investor flows how the financial market conditions caused speculative activity in crude oil futures market. This paper clearly explained the 2008 boom and bust in oil prices. Another important study which explored whether commodity index-investing had any significant influence on commodity prices was HAM [2014]. Buyuksahin and Robe [2014] explored the

relation between returns on investible commodities and the U.S. equity indices using a novel dataset of trader positions in the futures market. They found evidence that the correlation between equity and commodity indices went up due to increasing participation of speculators such as hedge funds. Basak and Pavlova [2016] showed that the price and volatility of commodity futures went up with the help of theoretical model where both institutional investors and traditional market participants play an important role in the commodity futures market.

Given the emergence of commodities as an alternative asset class, it is a natural question to ask how commodities are connected with other financial assets such as equities, foreign exchange and bonds. Because this connection or dependence has an important implication in the portfolio context and in risk management. In the literature attention has been paid to explore the correlation between the returns of two financial assets. In the early literature Li [2000] used copula in analyzing credit risks studying default correlation between two credit risks. Ang and Chen [2002] has pointed out that it is more important to pay attention to the tail events by showing correlation between U.S. stocks and aggregate U.S. is higher in downside market than in upside market i.e. the correlation is asymmetric. In a previous study Longin F. Solnik [2001] employed extreme value theory and found extreme negative returns have higher correlation than positive returns. Forbes and Rigobon [1999] is another important study where they investigated the correlation conditional on volatility. Erb et al. [1994] found cross-country equity correlations are higher during recessions than during expansion. Ang and Bekaert [2015] used regime switching model to explore the benefits of portfolio diversification considering international equities under high correlation and high volatility.

Financial return data has negative skewness and fat tail i.e. it is non-normal. Due to the presence of these asymmetries, the assumption of elliptically distributed returns is not appropriate. As a result normal distribution is not adequate to model these features. In general, correlation suffices as a measure of linear dependence in case of normality. But it gets af-

fected by marginal distribution of data. Embrechts et al. [2002] discussed the properties and disadvantages of correlation and dependence in risk management. To capture the tail dependence effectively, we need an appropriate tool. Copula provides a flexible methodology by which we can model marginal distributions separately from the dependence structure that links these distributions to form the joint distribution. The application of copula in economics and finance has been growing over the years. Patton [2004] showed evidence that copula models can yield better out-of-sample portfolio performance relative to bivariate normal models. Patton [2006a] proposed conditional copula models and explored the change in dependence structure of Deutsche Mark and the Yen during appreciation and depreciation of those exchange rates against the U.S. dollar. There was evidence of higher correlation between those exchange rates during depreciation. To model time-varying conditional copula, Patton [2006a] allowed the copula parameter to vary through time similar to a GARCH model (Engle [1982] and Bollerslev [1986]) of conditional variance. In one of the early papers Embrechts et al. [2002] showed the importance of copula to model risks. Cherubini et al. [2004] discusses application of copulas in finance such as derivative pricing, credit risk analysis, risk management issues etc. Jondeau and Rockinger [2006] used GARCH framework to model time variation in conditional copula model. As an alternative, regime switching model (Hamilton [1989]) also has been used for conditional copula. Rodriguez [2007] used regime switching conditional copula to study contagion and found evidence of higher dependence during downturn. Garcia and Tsafack [2011] used regime switching copula model on the international equity and bond markets to explore the effects of asymmetric dependence. Okimoto [2008] used Markov switching copula to study asymmetric dependence in international equity market. Two excellent references for the detailed theoretical introductions to copula covering its statistical and mathematical properties are Joe [1997] and Nelsen [2006]. Fan and Patton [2014] provides a brief review of the copula theory. Patton [2013] gives one of the most comprehensive reviews of the use of copula in economics for forecasting multivariate time series. This current exercise has closely followed the approach, notations, explanations in Patton [2013] and showed applications in risk management and portfolio optimization.

1.3 A Brief Overview of Copula

Before I go into the details of copula theory, let me clarify why do we need to use copula in the current exercise. Financial return data has negative skewness and fat tail. Normal distribution is not adequate to model these features. And correlation suffices as a measure of linear dependence in case of normality. But it gets affected by marginal distribution of data. Embrechts et al. [2002] discussed the properties and pitfalls of correlation and dependency in risk management. Copula provides a flexible methodology by allowing us to model marginal distributions separately from the dependence structure that links these distributions to form the joint distribution. A copula couples the marginal distributions together to form a joint distribution. It contains all of the information in the joint distribution which can not captured in the marginal distributions. In a word, copula has the entire dependence information.

Now let me briefly go over the theory of copula following the discussions (notations and explanations) in Patton [2004], Patton [2006a], Patton [2009a], Patton [2012] and Patton [2013]. First, I discuss the bivariate case and then extend the analysis to the multivariate scenarios [1]

Suppose there are two random variables X and Y such that,

$$X \sim F$$
$$Y \sim G$$
$$(X, Y) \sim H$$

Let $U = F(X)$ and $V = G(Y)$ such that

$$U \sim Unif(0,1) \qquad \text{and} \qquad V \sim Unif(0,1)$$

where U and V are probability integral transforms of X and Y. Let $(U, V) \sim C$. By standard

[1] This mathematical exposition is taken from the lecture slides of Professor Andrew Patton, Duke University, http://public.econ.duke.edu/~ap172/

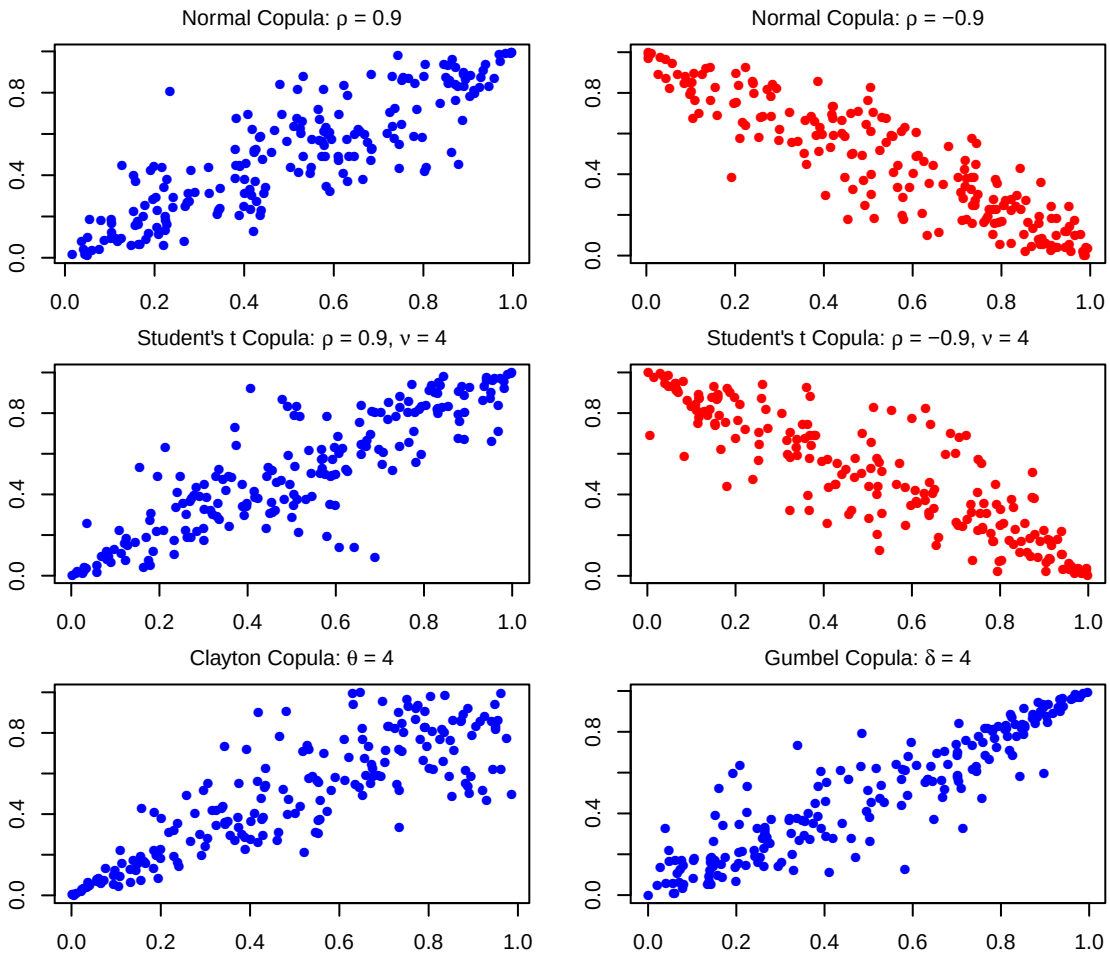

Figure 1.1: Simulation from different Copulas

theory on the distribution of transformations of random variables we obtain:

$$\begin{aligned} c(F(X), G(Y)) &= h(X,Y).\begin{vmatrix} dX/dU & dX/dV \\ dY/dU & dY/dV \end{vmatrix} \\ &= h(X,Y).\begin{vmatrix} f(X)^{-1} & 0 \\ 0 & f(Y)^{-1} \end{vmatrix} \\ &= \frac{h(X,Y)}{f(X).g(Y)} \end{aligned} \tag{1.1}$$

So we can express $h(X, Y)$ as:

$$h(X, Y) = f(X).g(Y).c(F(X), G(Y)) \tag{1.2}$$

where $h(X, Y)$ is joint density, $f(X)$ and $g(Y)$ are marginal densities and $c(F(X), G(Y))$ is the copula density. Sklar [1959] showed that we may decompose the distribution of (X, Y) into three parts:

$$H(x, y) \Leftrightarrow c(F(x), G(y)) \qquad \forall x, y \in \bar{\mathbb{R}} \tag{1.3}$$

This can easily be extended to the multivariate context. Following the notations of Patton [2013], suppose $\boldsymbol{Y} \equiv [Y_1, \ldots, Y_n]' \sim \boldsymbol{F}$, with $Y_i \sim F_i$, then due to Sklar [1959], an n-dimensional joint distribution can be decomposed into n univariate marginals and an n-dimensional copula i.e.

$$\boldsymbol{C} : [\,0, 1]^{\,n} \rightarrow [\,0, 1]$$

such that

$$\boldsymbol{F(y)} = \boldsymbol{C}(F_1(y_1), \ldots, F_n(y_n)) \quad \forall \quad \boldsymbol{y} \in \boldsymbol{R}^n \tag{1.4}$$

So the copula $\boldsymbol{C}$ of the variable $\boldsymbol{Y}$ is a function that maps the univariate marginal distributions F_i to the joint distribution $\boldsymbol{F}$. An alternative way to look at it is *probability integral transform* $U_i \equiv F_i(Y_i)$. According to Casella and Berger [1990], if F_i is continuous, then U_i will have uniform distribution $Unif(0, 1)$ regardless of the original distribution F_i. The copula $\boldsymbol{C}$ can then be thought of as the joint distribution of the vector of probability integral transforms, $\boldsymbol{U} \equiv [\,U_1, \ldots, U_n]\prime$, i.e. a joint distribution of the uniform $Unif(0, 1)$ margins. From the above expression of joint *cdf*, the joint *pdf* can be written as

$$\boldsymbol{f}(y_1, \ldots, y_n) = \boldsymbol{c}(F_1(y_1), \ldots, F_n(y_n)) \times \prod_{i=1}^{n} f_i(y_i) \tag{1.5}$$

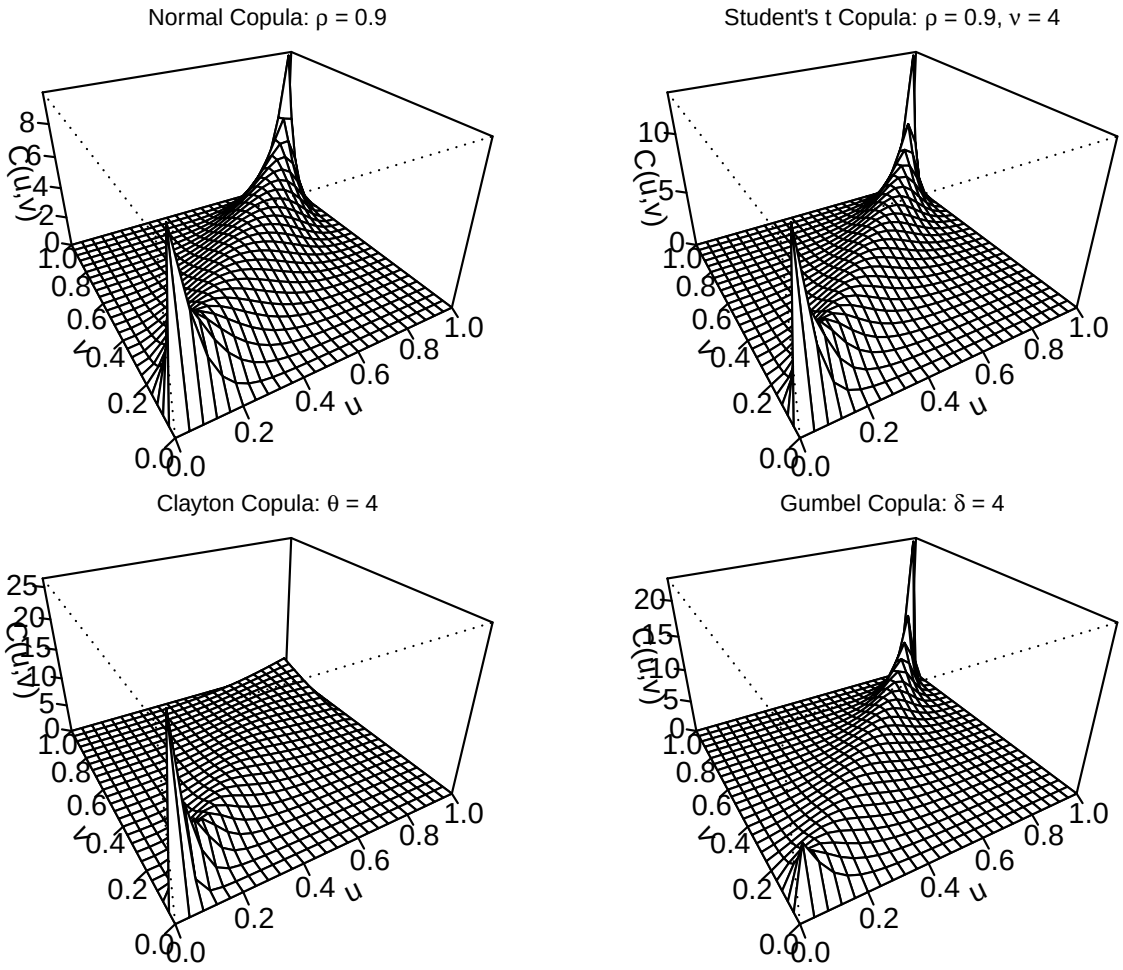

Figure 1.2: Probability Density of different Copulas

where $\boldsymbol{c}(u_1, \ldots, u_n) = \frac{\partial^n \boldsymbol{C}(u_1,\ldots,u_n)}{\partial u_1 \ldots \partial u_n}$. For a detailed exposition the reader is referred to Patton [2013].

To introduce time variation in the copula context, Patton [2006b] introduced the idea of conditional copula i.e. a conditional joint distribution by considering the information set $\mathcal{F}_{t-1}$ and decomposing the conditional distribution of $\boldsymbol{Y_t}$ given $\mathcal{F}_{t-1}$ into its conditional marginal distributions and conditional copula:

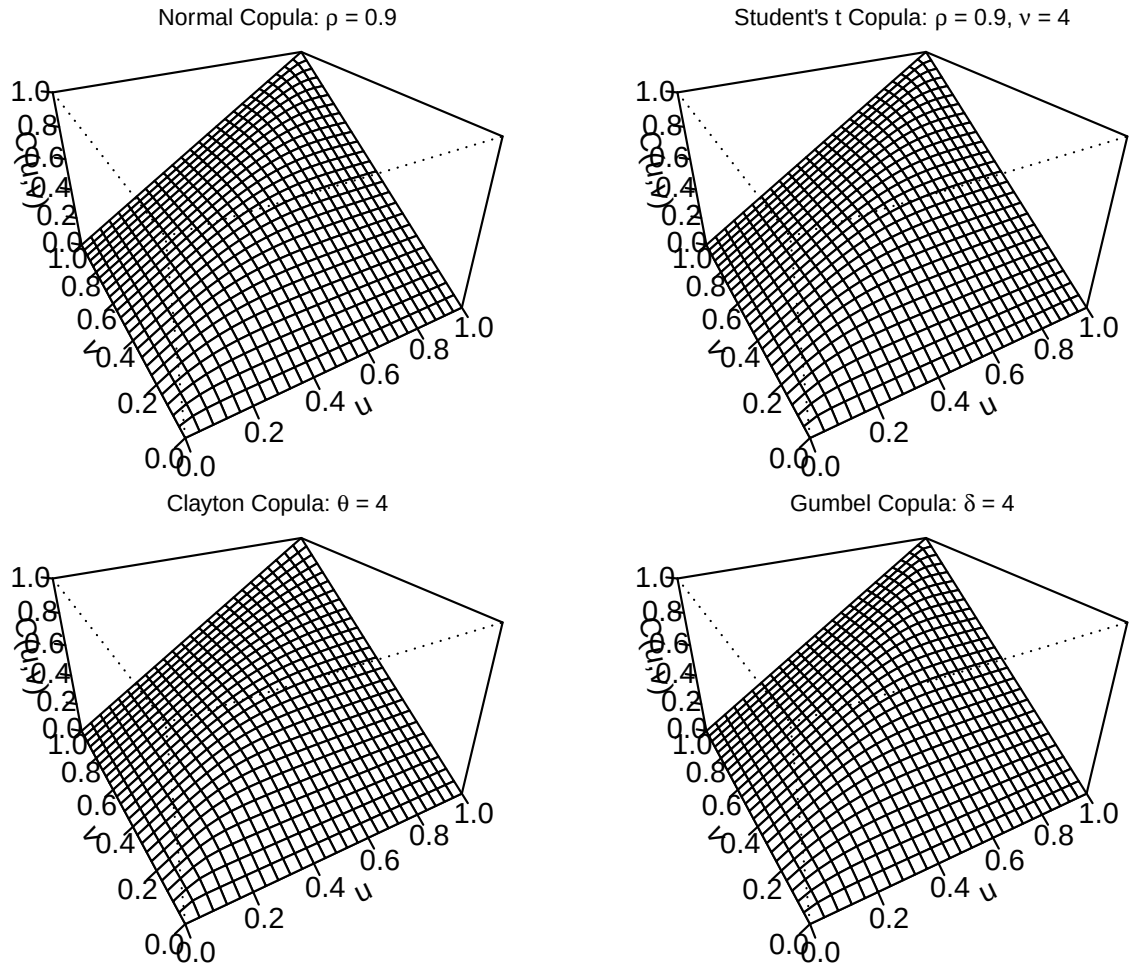

Figure 1.3: Cumulative Distribution Function of different Copulas

Let

$$\boldsymbol{Y_t}|\boldsymbol{\mathcal{F}_{t-1}} \sim \boldsymbol{F}(.|\mathcal{F}_{t-1})$$

with

$$Y_{it}|\mathcal{F}_{t-1} \sim F_i(.|\mathcal{F}_{t-1}), i = 1, 2, \ldots, n$$

then

$$\boldsymbol{F}(\boldsymbol{y}|\mathcal{F}_{t-1}) = \boldsymbol{C}(F_1(y_1|\mathcal{F}_{t-1}), \ldots, F_n(y_1|\mathcal{F}_{t-1}))$$

.

Defining the conditional probability integral transform variables, $U_{it} = F_i(Y_{it}|\mathcal{F}_{t-1})$, the conditional copula of $\boldsymbol{Y_t}|\mathcal{F}_{t-1}$ is the conditional distribution of $\boldsymbol{U_t}|\mathcal{F}_{t-1}$:

$$\boldsymbol{U_t}|\mathcal{F}_{t-1} \sim \boldsymbol{C}(\cdot|\mathcal{F}_{t-1})$$

In this way, the conditional marginal distributions $F_i|\mathcal{F}_{t-1}$ can be estimated in the first step and probability integral transform variables are constructed, and in the second step their joint distribution can be determined. This reduces the computational burden to a great extent. For a detailed exposition, the reader is referred to Patton [2013]. In this paper, both the marginal distribution and the copula models are parametric. Apart from Patton [2013], there are a number of references on copula in the literature such as Joe [1997], Frees and Valdez [1998], Cherubini et al. [2004], McNeil et al. [2005], Nelsen [2006], Genest and Favre [2007], Patton [2009b], Choros et al. [2010], Manner and Segers [2011], Patton [2012].

1.4 Data

Daily data of commodities, equities, foreign exchange and bonds of the U.S. financial market have been used for this exercise. For commodities (CP), the S&P GSCI Commodity Total Return Index is used. It is a leading measure of the commodity price movements and consists of principal physical commodities futures contracts. The source of this data is

	CP	EQ	FX	BOND
Arithmetic Mean	0.004	0.030	0.001	0.024
Minimum	-18.454	-22.900	-3.546	-1.640
Maximum	7.532	10.957	2.995	1.326
Stdev	1.239	1.105	0.536	0.240
Skewness	-0.507	-1.151	-0.101	-0.205
Kurtosis	8.612	26.670	1.967	2.112
Jarque-Bera	0.000	0.000	0.000	0.000
ARCH-LM	0.000	0.000	0.000	0.000

Table 1.1: Summary Statistics

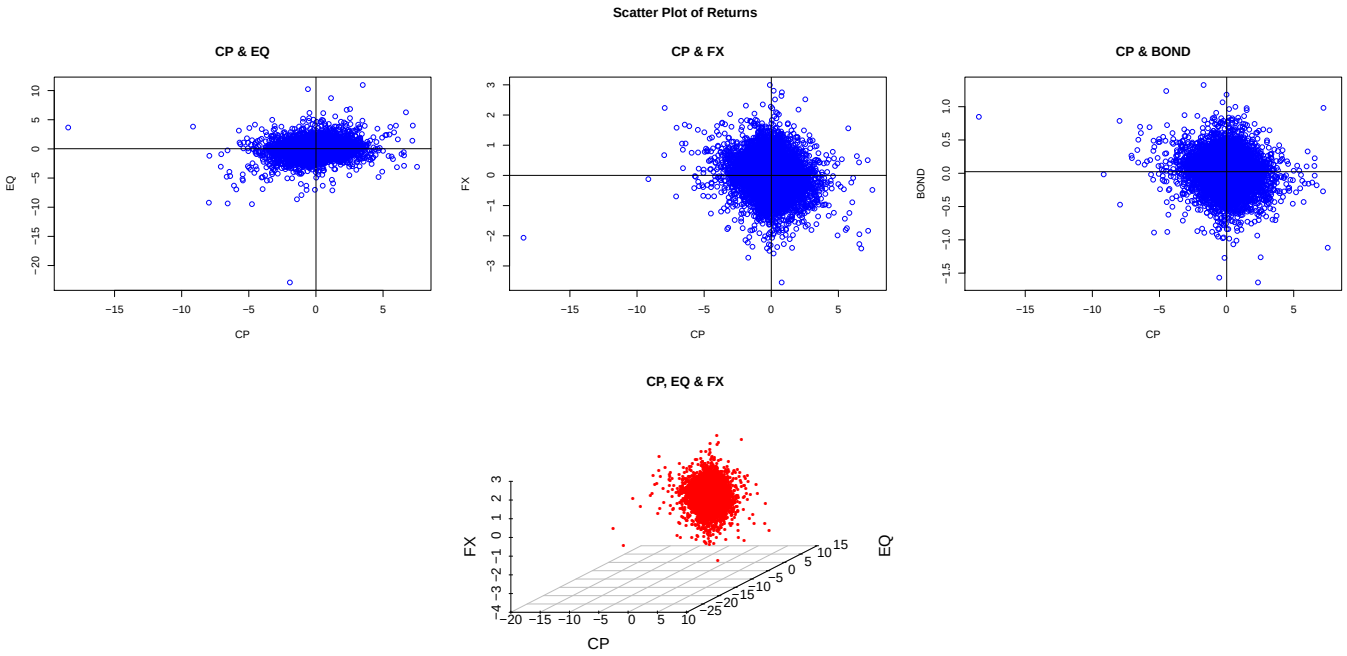

Figure 1.4: Scatter Plot of Returns

Global Financial Database. For equities (EQ), the S&P 500 Total Return Index has been used. For foreign exchange (FX), the U.S. Dollar Index which is traded at Intercontinental Exchange (ICE) is used. For bonds (BOND), the Barclays Aggregate Bond Index data has been used. The source of EQ, FX and BOND data is Bloomberg. The data for commodities, equities and foreign exchange span from January 1980 to December 2018 (9784 daily observations per assets), while the sample for bonds is from January 1989 to December 2018 (7465 daily observations). The table 1.1 shows the summary statistics of the data. The returns of each data are 100 times the first difference of the logarithm of daily data. Each return series shows non-normality as they have negative skewness and excess kurtosis (fat tails). The normality is also rejected by the Jarque-Bera test statistic. The quantile-quantile (QQ) plots in figure 1.5 shows fat-tails of all the series, which is another indication that all of these series are non-normally distributed with fat tails. The p-values of the ARCH-LM test (Engle [1982]) confirm that the return series have conditional heteroscedasticity.

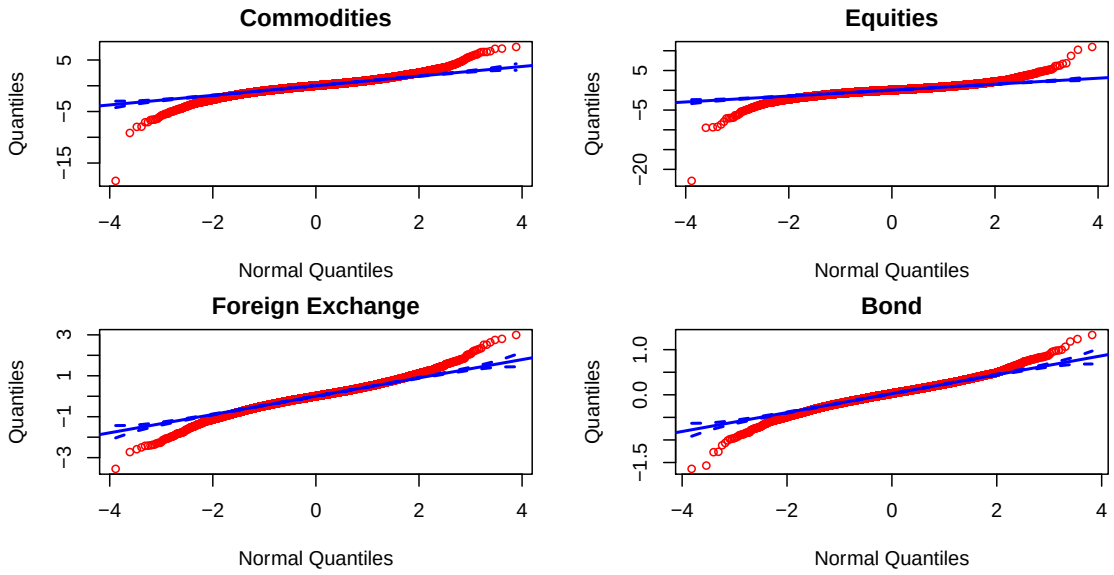

Figure 1.5: QQ Plot

For the computational purpose in this exercise, I have exhaustively used the *MATLAB* (MATLAB [2018]) code provided with Patton [2013] and also different packages (cite later in the chapter) of the statistical programming language *R* (R Core Team [2021]).

1.5 Estimation of Univariate Marginal Distribution

First, the marginal distribution of each series needs to be estimated. To do so, for each series, the conditional mean and conditional variance models are estimated. Generalized Autoregressive Conditional Heteroscedasticity (GARCH) family of models (Engle [1982] and Bollerslev [1986]) are well equipped to handle the conditional heteroskedasticity, asymmetry, and fat tails of financial returns. I consider Auto Regressive Moving Average(ARMA) for conditional mean and GARCH model for modeling conditional variance. For the innovation term of each series, I consider skewed Student's t distribution. I tried with different models under GARCH family and found the following models suitable for different assets. We find the Asymmetric Power GARCH (apARCH) model (see Ding et al. [1993]) suitable for modeling conditional variance of EQ, while GARCH(1,1) suffices for other assets. For model

specifications see table 1.2. The conditional mean ARMA$(1,1)$ model is given by:

$$y_t = \mu_y + \phi_1 y_{t-1} + \phi_2 \epsilon_{t-1} + \epsilon_t \tag{1.6}$$

and the conditional variance GARCH$(1,1)$ is given by:

$$\sigma^2_{y,t} = \omega + \alpha_1 \epsilon^2_{t-1} + \beta_1 \sigma^2_{y,t-1} \tag{1.7}$$

Ding et al. [1993] proposes the Asymmetric Power GARCH (apARCH) model by modeling σ^δ_t for $\delta > 0$

$$\sigma^\delta_t = \omega + \alpha_1 (|\epsilon_{t-1}| - \gamma_1 \epsilon_{t-1})^\delta + \beta_1 \sigma^\delta_{t-1}$$

The estimation method is maximum likelihood estimation (MLE). The table 1.2 shows the models considered for conditional mean, conditional variance and the distribution of the innovations terms. Once these models are estimated, I compute the standardized residuals for each of the series. The fitted density in figure 1.6 appears to be a reasonable fit to the density of the estimated residuals.

	Conditional Mean	Conditional Variance	Innovation
CP	ARMA(1, 0)	GARCH(1, 1)	Skewed t
EQ	ARMA(0, 0)	apARCH(1, 1)	Skewed t
FX	ARMA(1, 0)	GARCH(1, 1)	Skewed t
BOND	ARMA(1, 0)	GARCH(1, 1)	Skewed t

Table 1.2: Models for Conditional Mean and Conditional Variance

	Estimate	Std. Error	t value	Pr(>|t|)
ar1	-0.0022	0.0101	-0.2218	0.8244
omega	0.0048	0.0011	4.1704	0.0000
alpha1	0.0500	0.0042	11.9744	0.0000
beta1	0.9479	0.0041	230.1261	0.0000
skew	0.9434	0.0127	74.3105	0.0000
shape	8.2600	0.6463	12.7800	0.0000

Table 1.3: Commodities: ARMA(1,0) - GARCH(1,1)

	Estimate	Std. Error	t value	Pr(>|t|)
mu	0.0272	0.0080	3.4250	0.0000
omega	0.0167	0.0022	7.6399	0.0000
alpha1	0.0749	0.0056	13.4768	0.0000
gamma1	0.7484	0.0639	11.7204	0.0000
beta1	0.9252	0.0053	173.8590	0.0000
delta	1.0587	0.0826	12.8138	0.0000
skew	0.9287	0.0127	73.1509	0.0000
shape	6.7539	0.4527	14.9203	0.0000

Table 1.4: Equities: ARMA(0,0) - apARCH(1,1)

	Estimate	Std. Error	t value	Pr(>|t|)
mu	0.0028	0.0049	0.5798	0.5621
ar1	-0.0111	0.0099	-1.1213	0.2622
omega	0.0015	0.0004	3.7626	0.0002
alpha1	0.0388	0.0036	10.8701	0.0000
beta1	0.9571	0.0039	245.5614	0.0000
skew	0.9789	0.0137	71.6951	0.0000
shape	7.2393	0.5227	13.8496	0.0000

Table 1.5: Foreign Exchange: ARMA(1,0) - GARCH(1,1)

	Estimate	Std. Error	t value	Pr(>|t|)
mu	0.0223	0.0026	8.7367	0.0000
ar1	0.0107	0.0115	0.9319	0.3514
omega	0.0005	0.0001	3.9340	0.0001
alpha1	0.0339	0.0038	9.0235	0.0000
beta1	0.9583	0.0047	204.6354	0.0000
skew	0.9338	0.0154	60.7098	0.0000
shape	7.9224	0.6736	11.7615	0.0000

Table 1.6: Bond: ARMA(1,0) - GARCH(1,1)

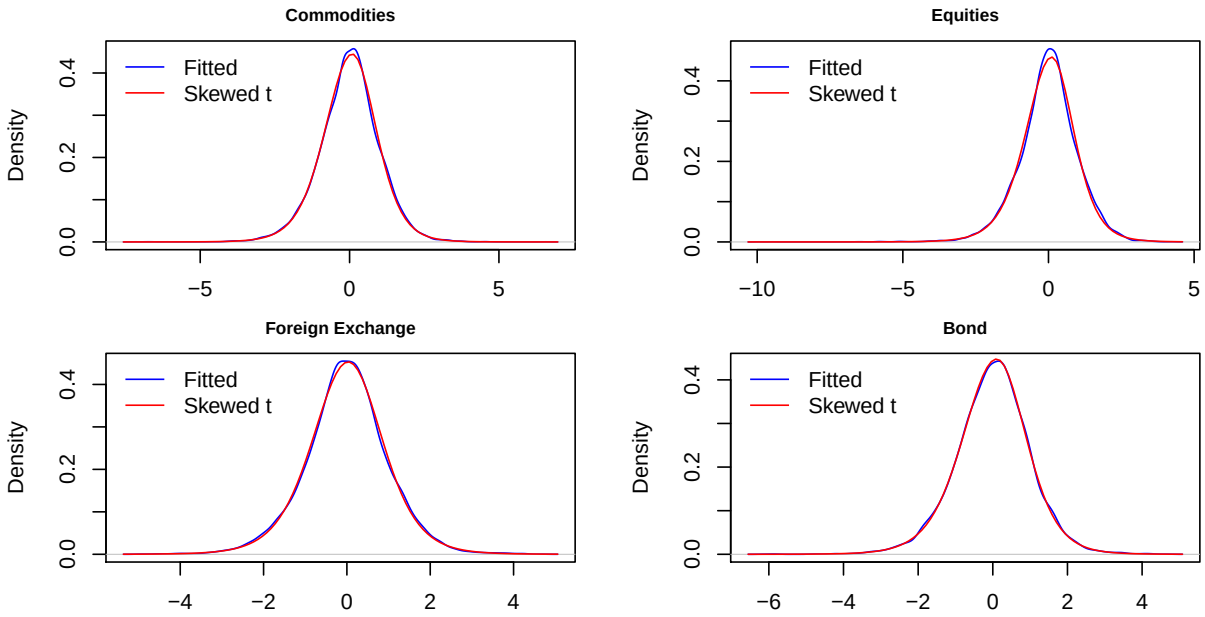

Figure 1.6: Fitted Density Plot

The parameter estimates and standard errors for marginal distribution models are shown in the tables 1.3, 1.4, 1.5, 1.6. Almost all the parameter estimates are highly significant. Once I estimate all the distributional parameters, I proceed to estimate the copula parameters by maximum likelihood in the second stage.

I test the goodness-of-fit of the univariate marginal distribution models with the Cramer-von Mises test which deals with the modelling of a probability distribution of a random vector $\boldsymbol{X} = (X^1, \ldots, X^n)$. It helps in verifying the compatibility between a sample of data $\{x_1, x_2, \ldots, x_N\}$ and a candidate probability distribution. This test is based on the distance between the cumulative distribution function $\widehat{F}_N$ of the sample $\{x_1, x_2, \ldots, x_N\}$ and that of the candidate distribution F. The table 1.7 shows the simulation-based p-values from Cramer-von Mises goodness-of-fit tests for the models of the conditional marginal distributions simulations. Based on the p-values, I can not reject the null hypothesis of uniform distribution. I perform this test using *R* package *goftest*, see Faraway et al. [2021].

	CvM p-value
Commodities	0.42
Equities	0.12
Foreign Exchange	0.48
Bonds	0.99

Table 1.7: Cramer-von Mises Test of Goodness-of-Fit

1.6 Dependence Measures

In case of normality, correlation coefficient is adequate as a measure of dependence. But any pure measure of dependence should be "scale invariant" i.e. it should not be affected by the increasing transformation of the data, as noted in Patton [2013]. This means the measure should be a function of the ranks (or probability integral transform) of the data. Unfortunately, the linear correlation coefficient is not scale invariant and it gets influenced by the marginal distribution of the data. As an alternative, the Spearman rank correlation coefficient measures the degree of association between the rankings of two variables. It gives an idea about the sign of the dependence of two variables. Here I reproduce the definition of population and sample rank correlation from Patton [2013]. I also compute the dependence measures customizing the *MATLAB* (MATLAB [2018]) code provided with Patton [2013].

The population rank correlation ρ is given by

$$\begin{aligned}\rho &= Corr[U_{1t}, U_{2t}] \\ &= 12E[U_{1t}U_{2t}] - 3 \\ &= 12\int_0^1 \int_0^1 uv \; dC(u,v) - 3\end{aligned} \tag{1.8}$$

and the sample rank correlation $\hat{\rho}$ is given by

$$\hat{\rho} = \frac{12}{T}\sum_{t=1}^{T} U_{1t}U_{2t} - 3 \tag{1.9}$$

In this context, another measure of dependence is quantile dependence which captures the degree of dependence between joint lower or joint upper tails of the two variables. It is an appropriate tool to study the joint movement of two markets in times of turmoil. It also enables us to capture the asymmetric dependence structure as dependence might vary based on the center or left or right tail of the distribution. Formally, quantile dependence can be defined as

$$\lambda_q = \begin{cases} Pr[U_{1t} \leq q | U_{2t} \leq q], & 0 < q \leq 1/2 \\ Pr[U_{1t} > q | U_{2t} > q], & 1/2 < q < 1 \end{cases} \tag{1.10}$$

As per Patton [2013], the above can be expressed in terms of a copula as follows -

$$\lambda_q = \begin{cases} \dfrac{C(q,q)}{q}, & 0 < q \leq 1/2 \\ \dfrac{1 - 2q + C(q,q)}{1-q}, & 1/2 < q < 1 \end{cases} \tag{1.11}$$

The figure below shows the estimated quantile dependence for $q \in [0.025, 0.975]$ with 90% *iid* bootstrap confidence intervals.

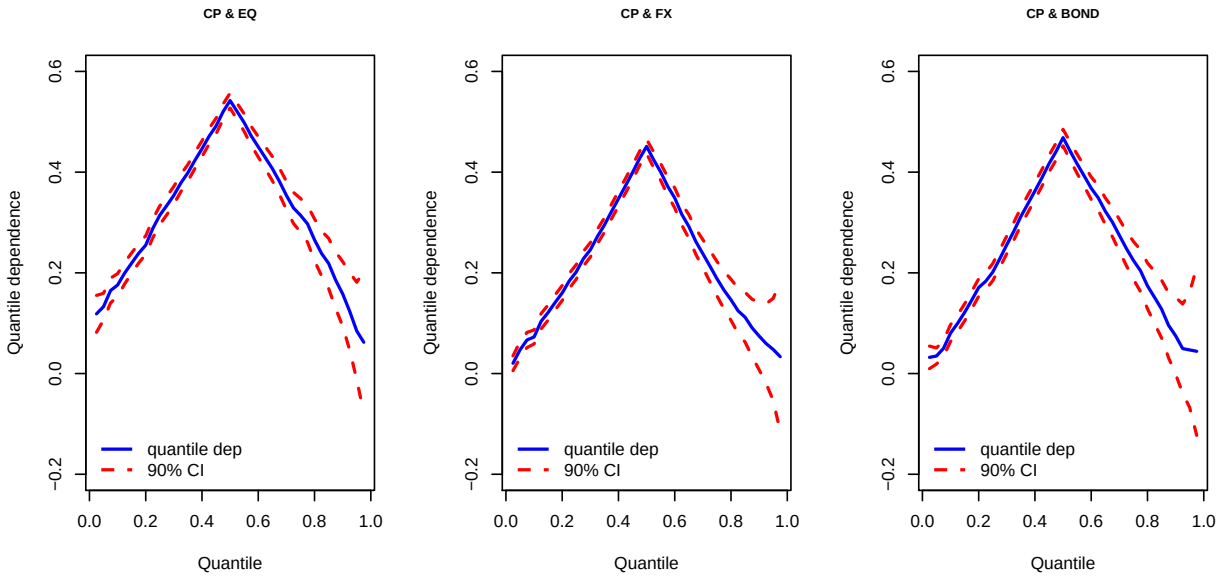

Figure 1.7: Quantile Dependence Plot

For commodities and equities , the lower tail has higher quantile dependence than the upper tail but for commodities & foreign exchange and commodities & bonds, the quantile dependence between two tails is almost the same.

In general, during the time of stress in the financial market, correlation among different assets will increase. This can be quantitatively measured by the coefficient of tail dependence. Bivariate tail dependence measures the degree of dependence in the upper and lower quadrant tail of a bivariate distribution. This is important from the risk management perspective as an investor will try to avoid bad events happening at the same time in both markets. The coefficient of tail dependence was first introduced in the financial context by Embrechts et al. [2002]. Tail dependence is defined as follows. Let $X \sim F_X$ and $Y \sim F_Y$ be two random variables. Then the upper tail dependence is defined as

$$\lambda_u(X,Y) = \lim_{\alpha \to 1} P(Y > F_Y^{-1}(\alpha) | X > F_X^{-1}(\alpha))$$

This computes the probability to observe a large Y, assuming that X is also large. Similarly, the coefficient of lower tail dependence is

$$\lambda_l(X,Y) = \lim_{\alpha \to 1} P(Y \leq F_Y^{-1}(\alpha) | X \leq F_X^{-1}(\alpha))$$

These measures are independent of the marginal distributions of the asset returns and they are invariant under strictly increasing transformations of X and Y.

Tail dependence for Gaussian copula is zero. This means, that regardless of high correlation ρ we choose, if we go far enough into the tail, extreme events appear to occur independently in X and Y. For the Student's t-copula, the coefficients of lower and upper tail dependence are

$$\lambda_l(X,Y) = \lambda_u(X,Y) = 2t_{\nu+1}\left(-\sqrt{\nu+1}\sqrt{\frac{1-\rho}{1+\rho}}\right)$$

where $t_{\nu+1}$ is the distribution function of a univariate Student's t-distribution with $\nu + 1$ degrees of freedom. The stronger the linear correlation ρ and the lower the degrees of freedom ν, the stronger is the tail dependence. The Clayton copula is lower tail dependent i.e. the coefficient of the upper tail dependence $\lambda_u(X,Y) = 0$, and the coefficient of the lower tail dependence is

$$\lambda_l(X,Y) = 2^{-\frac{1}{\delta}}$$

The Gumbel copula is upper tail dependent i.e. the coefficient of the lower tail dependence $\lambda_l(X,Y) = 0$ and the coefficient of the upper tail dependence is

$$\lambda_u(X,Y) = 2 - 2^{\frac{1}{\delta}}$$

Ang and Chen [2002] shows correlation between U.S. stocks and the aggregate U.S. market are much greater for downside moves, particularly for extreme downside moves.

1.7 Parametric Estimation of Constant Copulas

In order to do parametric estimation of static copulas, I have estimated the univariate marginal distribution in the first step. After the marginal distribution parameters are estimated by maximum likelihood (MLE) in the first stage, the copula parameters are estimated again by MLE in the second stage.

$$\sum_{t=1}^{T} log f_t(\boldsymbol{Y_t}; \theta) = \sum_{t=1}^{T} \sum_{i=1}^{n} log f_{it}(Y_{it}; \alpha_i) + \sum_{t=1}^{T} log c_t(F_{1t}(Y_{1t}; \alpha_1), \ldots, F_{nt}(Y_{nt}; \alpha_n); \gamma)$$

This method of estimating parameters for the margin and copula separately in two stages by maximizing the log-likelihood is referred to as *inference function for margin* (Joe and Xu [1996], Joe [1997]). This method is generally known as multi-stage maximum likelihood (MSML) estimation. In this exercise, for the constant parametric copula models, following specifications are considered - Normal copula, Student's t copula, Clayton copula, and Rotated Clayton copula.

The d-dimensional t-copula takes the form

$$C^t_{\nu,P}(u) = t_{\nu,P}(t^{-1}_\nu(u_1), t^{-1}_\nu(u_2), \ldots, t^{-1}_\nu(u_d))$$

where t_ν is the distribution function of a standard univariate t distribution, $t_{\nu,P}$ is the joint distribution function of the vector $X \sim t_d(\nu, 0, P)$ and P is a correlation matrix. Some copulas have simple closed forms like Gumbel or Clayton copula. An example of bivariate Clayton copula is

$$C^{Cl}_\theta(u_1, u_2) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}, 0 < \theta < \infty$$

where the parameter θ represents the strength of dependence. The Gumbel copula exhibits greater dependence in the positive tail than in the negative tail. This copula is given by

$$C^{Gum}_\theta(u_1, u_2) = exp(-[(-logu_1)^\theta + (-logu_1)^\theta]^{1/\theta}, 0 < \theta \leq 1$$

where the parameter θ represents the strength of dependence,(see Patton [2013]).
Next I show the static copula estimation results for different bivariate copulas (CP & EQ, CP & FX, CP & BOND). The tables display the estimated copula parameters, the log-likelihood value and the standard errors of estimated parameters. First, I estimate the copulas based on full-sample, then I estimate splitting the sample into two sub-samples.

1.7.1 Full-sample Copula Estimation

The table 1.8 shows the estimated copula parameters between commodities and equities for the full sample. Among these Normal copula do not have tail dependence. Clayton copula

	Parameter 1	Parameter 2	Kendal's tau	Spearman's rho	LL
Normal	0.12		0.08	0.12	72.75
	(0.01)				
Clayton	0.14		0.07	0.10	82.72
	(0.01)				
Rotated Clayton		0.12	0.06	0.08	50.46
		(0.01)			
Student's t	0.12	9.82	0.08	0.12	121.69
	(0.01)	(1.10)			

Table 1.8: CP-EQ Copula Estimation

	Parameter 1	Parameter 2	Kendal's tau	Spearman's rho	LL
Normal	−0.15		−0.09	−0.14	104.92
	(0.01)				
Clayton	−0.06		−0.03	−0.05	35.43
	(0.01)				
Rotated Clayton		−0.08	−0.04	−0.06	52.24
		(0.01)			
Student's t	−0.15	14.78	−0.09	−0.14	127.61
	(0.01)	(2.39)			

Table 1.9: CP-FX Copula Estimation

is used for modeling lower/left tail dependency while rotated Clayton copula is used for modeling the right tail dependency. The student's t copula assumes symmetric tail dependency for both lower and higher tails. Based on the log-likelihood values, we see that Student's t copula best fits the data for commodities and equities. The table 1.9 shows the estimated copula parameters between commodities and foreign exchange for the full sample. Based on the log-likelihood values, we see that Student's t copula best fits the data for commodities and foreign exchange. The dependence between commodities and foreign exchange implied by the estimated copulas is negative. But the estimated degree of freedom parameter (an indicator of the tail thickness of the joint distribution of the two variables) of the Student's t copula is higher than that of the case between commodities and equities, pointing to lower tail dependency between commodities and foreign exchange.

	Parameter 1	Parameter 2	Kendal's tau	Spearman's rho	LL
Normal	−0.10		−0.06	−0.10	38.36
	(0.01)				
Clayton	−0.07		−0.03	−0.05	22.19
	(0.01)				
Rotated Clayton		−0.05	−0.03	−0.04	14.25
		(0.01)			
Student's t	−0.10	19.51	−0.06	−0.10	48.65
	(0.01)	(4.62)			

Table 1.10: CP-BOND Copula Estimation

The table 1.10 shows the estimated copula parameters between commodities and bonds for the full sample. Based on the log-likelihood values, we see that Student's t copula best fits the data for commodities and bonds. The dependence between commodities and bonds implied by the estimated copulas is negative. But the estimated degree of freedom parameter (an indicator of the tail thickness of the joint distribution of the two variables) of the Student's t copula is higher than that of the case between commodities & equities and commodities & foreign exchange, pointing to lower tail dependency between commodities & bonds.

Given the fact that investment into commodities futures started to go up since 2004, it is worth exploring the dependence between commodities and other financial assets in sub-samples i.e. between 1980-2004 and between 2005-2018. While static copula estimation based on these two sub-samples are not sufficient, it nonetheless gives an indication whether the dependency between commodities and other financial assets changed over time.

1.7.2 Sub-sample Copula Estimation

First, I look into the linear correlation and rank correlation between commodities and equities corresponding to sub-samples 1980-2004 and 2005-2018. The linear and rank correlation in the period 1980-2004 was almost zero, while it rose to 0.3 in the period 2005-2018. The

	Parameter 1	Parameter 2	Kendal's tau	Spearman's rho	LL
Normal	0.03		0.02	0.03	2.36
	(0.01)				
Student's t	0.03	12.86	0.02	0.03	21.51
	(0.01)	(2.28)			

Table 1.11: Estimation of CP-EQ Copula for Sub-sample upto 2004

	Parameter 1	Parameter 2	Kendal's tau	Spearman's rho	LL
Normal	0.29		0.19	0.28	153.68
	(0.01)				
Student's t	0.29	9.38	0.19	0.28	173.26
	(0.02)	(1.70)			

Table 1.12: Estimation of CP-EQ Copula for Sub-sample after 2004

dependence based on static Gaussian and Student's t copula for the period 1980-2004 is almost zero. The dependence went up in the second period 2005-2018. We also see the

degree of freedom parameter from the Student's t copula decreases in the second period, pointing to the fact that tail thickness goes up in the second period. This result is consistent with the fact that commodities started emerging as an alternative asset class after 2004 (Tang and Xiong [2012]) and the commodities market was more integrated with the equities market in general. We observe similar patterns in the relationship between commodities and foreign exchange. The linear and rank correlation in the period 1980-2004 was almost zero, while it rose to -0.3 in the period 2005-2018. The dependence based on static Gaussian and Student's t copula for the period 1980-2004 is almost zero. The dependence went up in the second period 2005-2018. The degree of freedom parameter from the Student's t copula decreases substantially in the second period, pointing to the fact that tail thickness goes up in the second period. While commodities and foreign exchange market showed more dependence in the second period, they showed substantial joint tail movement as was the case between equities and commodities.

	Parameter 1	Parameter 2	Kendal's tau	Spearman's rho	LL
Normal	-0.07		-0.04	-0.06	14.14
	(0.01)				
Student's t	-0.07	22.69	-0.04	-0.07	20.62
	(0.01)	(6.72)			

Table 1.13: Estimation of CP-FX Copula for Sub-sample upto 2004

	Parameter 1	Parameter 2	Kendal's tau	Spearman's rho	LL
Normal	-0.28		-0.18	-0.27	148.86
	(0.01)				
Student's t	-0.29	11.20	-0.18	-0.27	162.22
	(0.02)	(2.39)			

Table 1.14: Estimation of CP-FX Copula for Sub-sample after 2004

Similar patterns were observed in the relationship between commodities and bonds. The linear and rank correlation in the period 1980-2004 was almost zero, while it rose to -0.15 in the period 2005-2018. The dependence based on static Gaussian and Student's t copula for the period 1980-2004 is almost zero. The dependence went up in the second period 2005-2018. The degree of freedom parameter from the Student's t copula decreases substantially in the second period, pointing to the fact that tail thickness goes up in the second period. While commodities and bonds market showed more dependence in the second period, they showed substantial joint tail movement as were in the previous two cases.

	Parameter 1	Parameter 2	Kendal's tau	Spearman's rho	LL
Normal	-0.06		-0.04	-0.05	6.52
	(0.02)				
Student's t	-0.05	23.83	-0.03	-0.05	11.01
	(0.02)	(8.59)			

Table 1.15: Estimation of CP-BOND Copula for Sub-sample upto 2004

	Parameter 1	Parameter 2	Kendal's tau	Spearman's rho	LL
Normal	-0.15		-0.10	-0.15	40.74
	(0.02)				
Student's t	-0.15	15.78	-0.10	-0.15	46.79
	(0.02)	(4.85)			

Table 1.16: Estimation of CP-BOND Copula for Sub-sample after 2004

1.8 Time Varying Dependence

Based on the above result, I argue that the dependence between commodities and other financial assets changes through time. To explore that possibility in detail, first we would like to look into the time varying rank correlation between these assets. The following plots show the 100 day rolling rank correlation between the pairs commodities-equities and

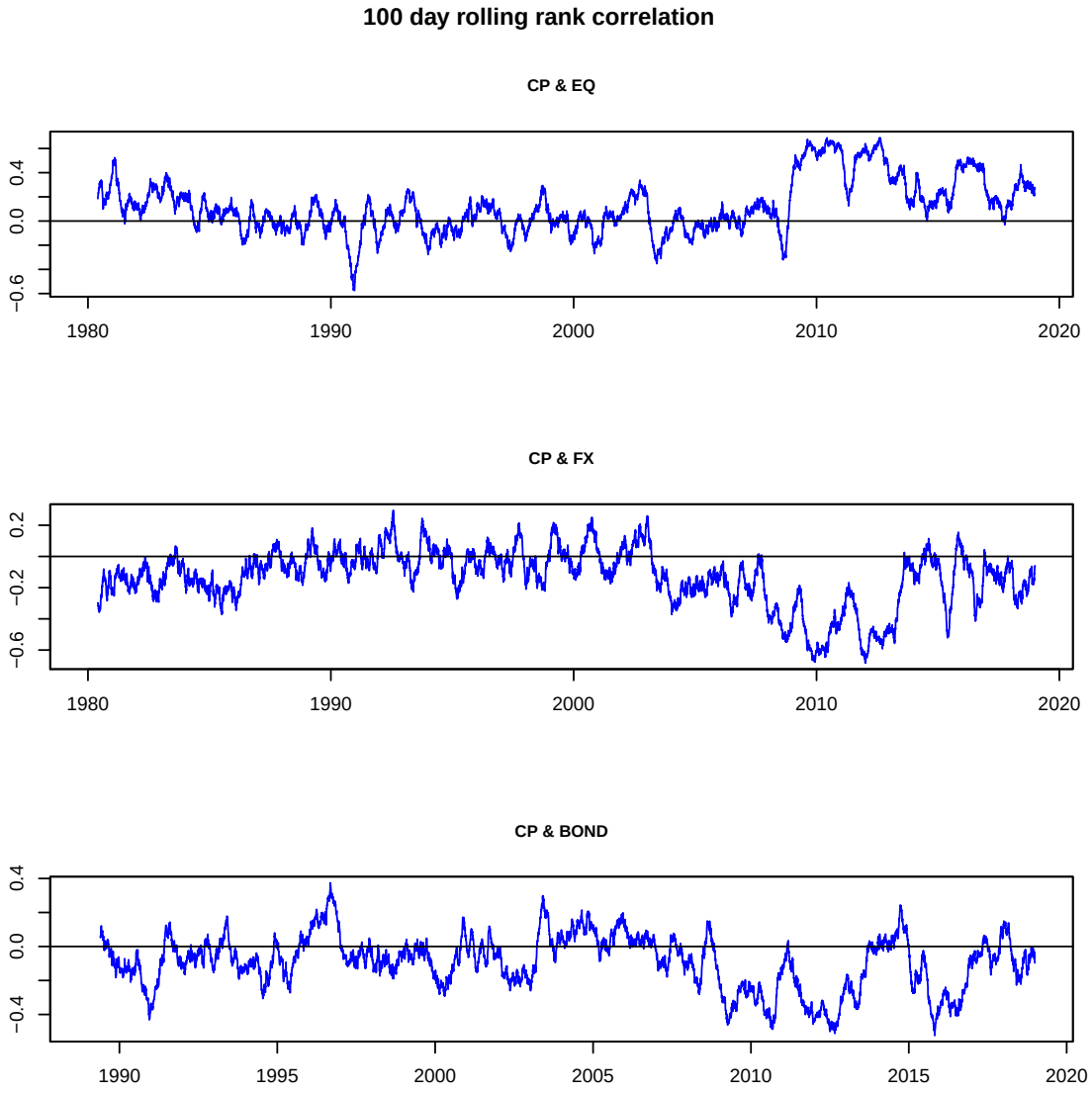

Figure 1.8: Rolling Rank Correlation

commodities-foreign exchange. The rolling rank correlation in both cases do indeed seem to be time varying. Following Patton [2013] I conduct three types of tests for time varying dependence. In the first test, I test for a break in rank correlation at some specified point in the sample, t^*. While under the hypothesis, the dependence measure before and after this particular date will be equal, under the alternative hypothesis they will be different.

$$H_0 : \rho_1 = \rho_2$$
$$H_a : \rho_1 \neq \rho_2$$

where

$$\rho_t = \begin{cases} \rho_1, & t \leq t^* \\ \rho_2, & t > t^* \end{cases} \tag{1.12}$$

Following the *iid* bootstrap procedure suggested by Patton [2013], the critical value of $(\hat{\rho_1} - \hat{\rho_2})$ can be computed. Although this test is simple to implement, it requires us to know a priori when a break in the dependence structure may have occurred. In most of the cases, the date of the break will not be known to us.

As stated in Patton [2013], in the second test, a break in the time varying rank correlation is assumed at an unknown date. To make sure that we have sufficient observations to estimate the pre-break and post-break parameter, we need to assume that the break did not occur "too close" to the start or end of the sample period. The common choice is to search for breaks in an interval $[t_L^*, t_U^*]$ where $t_L^* = [0.15T]$ and $t_U^* = [0.85T]$. For these types of tests, the "sup" test statistic is a popular choice.

$$\hat{B_{sup}} = \max_{t^* \in [t_L^*, t_U^*]} |\hat{\rho_{1,t^*}} - \hat{\rho_{2,t^*}}|$$

where

$$\hat{\rho_{1,t^*}} = \frac{12}{t^*} \sum_{t=1}^{t^*} U_{1t} U_{2t} - 3$$

$$\hat{\rho_{2,t^*}} = \frac{12}{T - t^*} \sum_{t=t^*+1}^{T} U_{1t} U_{2t} - 3$$

The table 1.17 shows the p values of the tests conducted. I consider a one time break at three points in the sample, at $t^*/T \in 0.15, 0.50, 0.85$. For the dependence between commodities-equities, we could detect multiple breaks occurring at different sample points considered.

	Sample split at half	Sample split at 2004
CP-EQ	0.000	0.00
CP-FX	0.001	0.00
CP-BOND	0.001	0.00

Table 1.17: Tests of Time Varying Dependence: p-values

For dependence between commodities and foreign exchange, we could detect a break occurring approximately at the middle of the sample. So broadly, these tests give us enough indication that the dependence between commodities and other financial assets is time varying in nature. So it will be logical to consider the time varying copula models to explore the structure and implications of the dependence.

1.9 Generalised Autoregressive Score (GAS) Estimation of Time Varying Copula

Following the method described in Patton [2013], I now turn to model the conditional time varying copula. Patton [2013] used the Generalized Autoregressive Score (GAS) model proposed by Creal et al. [2013]. The GAS, also called the Dynamic Conditional Score (DCS), model acts as a general framework for introducing time variation in parametric models. In this model the likelihood is available in a closed form. So it enables us to compute the score of parametric conditional observation density with respect to the time varying parameter. Using the notation of Creal et al. [2013], let $p(y_t|f_t)$ denote the conditional observation density for observations y_t and the time varying parameter f_t, where the f_t satisfies the following recursive equation,

$$f_{t+1} = \omega + \beta f_t + \alpha S(f_t)\Big[\frac{\partial \, log \, p(y_t|f_t)}{\partial f_t}\Big]$$

where $S(f_t)$ is a scaling function for the score of the log observation density. Here the scaled score is used to drive the time variation in the parameter f_t. As a result the shape of the conditional observation density is directly linked to the dynamics of f_t.

	ω	α	β	ν	LL
CP-EQ	0.0042	0.0126	0.9990	7.8231	149.1605
	(0.0000)	(0.0000)	(0.0000)	(0.0024)	
CP-FX	−0.0431	0.0163	0.8795	13.9495	123.4796
	(0.0025)	(0.0003)	(0.0000)	(0.0042)	
CP-BOND	−0.0371	0.0308	0.8795	17.1330	50.6517
	(0.0007)	(0.0002)	(0.0000)	(0.0009)	

Table 1.18: Estimation of Student's t GAS Copula

In this method the time varying copula parameter δ_t is expressed as a function of the lagged copula parameter and a "forcing variable" which is related to the standardized or scaled score of the copula log-likelihood. In some cases, copula parameters are constrained to lie in a certain range, e.g., correlation parameter in the Student's t copula should lie between -1 and $+1$. This can be done by applying a strictly increasing function $h(\cdot)$ to δ_t to get f_t. Following the notations used in Patton [2013], let

$$f_t = h(\delta_t) \Leftrightarrow \delta_t = h^{-1}(f_t)$$

where

$$f_{t+1} = \omega + \beta f_t + \alpha I_t^{-1/2} s_t$$
$$s_t \equiv \frac{\partial}{\partial \delta} log\boldsymbol{c}(U_{1t}, U_{2t}; \delta_t)$$
$$I_t \equiv E_{t-1}[s_t s_t'] = I(\delta_t)$$

So the future value of the copula parameter is a function of a constant, current value of the copula parameter and the scaled score of the copula log-likelihood $I_t^{-1/2} s_t$. For the Student's t copula, degrees of freedom parameter is kept constant and the correlation pa-

rameter is time varying. To ensure that this parameter lies between $(-1, 1)$, the function $\delta_t = (1 - exp\{-f_t\})/(1 + exp\{-f_t\})$ is used. The table 1.18 shows the Student's t GAS copula estimation results for the commodities-equity, commodities-foreign exchange and commodities-bond pairs.

1.10 Risk Measures: Value-at-Risk and Expected Shortfall

Modeling dependencies with the help of copula can be useful in the context of risk management. In this paper, I focus on two key tail-based risk measures - Value-at-Risk (VaR) and Expected Shortfall (ES) to show how effectively the copula can capture the tail dependence among different assets. I compute the VaR and ES of equally weighted portfolios consisting of two, three and four assets respectively. The table 1.19 shows the composition of the equally weighted portfolios.

Number of assets	Assets	Weight
2	CP-EQ	1/2
2	CP-FX	1/2
2	CP-BOND	1/2
3	CP-EQ-FX	1/3
4	CP-EQ-FX-BOND	1/4

Table 1.19: Composition of Portfolios

Using the same notations of Patton [2013], let us define VaR and ES. For any portfolio return Y_t with conditional distribution F_t, these risk measures are defined as

$$VaR_t^q \equiv F_t^{-1}(q) \text{ for } q \in (0, 1)$$

$$ES_t^q \equiv E[Y_t | \mathcal{F}_{t-1}, Y_t \leq VaR_t^q] \text{ for } q \in (0, 1)$$

,

So in other words, $q\%$ VaR is q^{th} percentile of the conditional distribution. For a given confidence level the VaR measures the maximum loss of a portfolio over a predetermined

time period. The Expected Shortfall is the expected value of Y_t conditional on it lying below the VaR. Following Patton [2013], I use the method of simulation to compute VaR and ES of the portfolios.

1.11 Evaluation of Risk Measures

While the time varying copula models help us compute the time varying VaR and ES, we need to assess how this model performs. This is more relevant from the perspective of risk management. A good model should be able to predict future VaR accurately. This evaluation of the model is done through a process called *backtesting*. It is a statistical process where the actual profits and loss of a portfolio are compared to the corresponding VaR and ES estimates.

1.11.1 Comparison against Benchmark Models

Once I forecast VaR and ES using time varying Student's t copula, I compare the backtesting preformance with that of two other benchmark models such as RiskMetrics™ (Morgan et al. [1996]) and Normal-Dynamic Conditional Correlation (DCC) (Engel [2002]) model. Both these models are quite popular in the risk management practice.

RiskMetrics™ Model

This model (Morgan et al. [1996]) uses an exponential smoothing mechanism as it uses decreasing weights based on a parameter λ. Let there be N securities. The asset return is modeled as $r_{i,t} = \sigma_{i,t}\epsilon_{i,t}$ where $\epsilon_{i,t} \sim N(0,1)$, $\epsilon_t \sim MVN(0, R_t)$, where $\epsilon_t = [\epsilon_{1t}, \epsilon_{2t}, \ldots, \epsilon_{Nt}]$. R_t is an $N \times N$ conditional or time-dependent correlation matrix. Variance of each return, $\sigma^2_{i,t}$ and the correlations between returns $\rho_{ij,t}$ are functions of time. This model uses exponentially weighted moving average model (EWMA) to forecast variance and covariance of the multivariate normal distribution. Assuming sample mean is zero, the period $t+1$ variance

forecast, given data available at time t, is

$$\sigma^2_{i,t+1|t} = \lambda\sigma^2_{i,t|t-1} + (1-\lambda)r^2_{i,t}$$

Normal-DCC Model

Engel [2002] extended Bollerslev's Constant Conditional Correlation (CCC) model (Bollerslev [1990]) to allow the conditional correlation to be time varying. Let the returns be described as

$$\underset{k\times 1}{r_t} = \mu + \varepsilon_t,\ \varepsilon_t|I_{t-1} \sim i.i.d.\ N(0,\Sigma_t)$$

and R_t be the correlation matrix containing conditional correlations

$$R_t = \begin{bmatrix} 1 & \rho_{12,t} & \cdots & \rho_{1k,t} \\ \rho_{12,t} & 1 & \cdots & \rho_{2k,t} \\ \vdots & & \ddots & \vdots \\ \rho_{1k,t} & \rho_{2k,t} & \cdots & 1 \end{bmatrix}$$

Let D_t be the diagonal matrix containing the volatility of returns along the diagonal

$$D_t = \begin{bmatrix} \sigma_{1t} & 0 & \cdots & 0 \\ 0 & \sigma_{2t} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{kt} \end{bmatrix}$$

For each univariate series, $(i = 1, 2, \ldots, k)$, let $\varepsilon_{it} = \sigma_{it}z_{it}$, $z_{it} \sim N(0,1)$ and $\text{var}(\epsilon_{it}|I_{t-1}) = \sigma^2_{it}$. If the vector of standardized errors is $z_t = (z_{1t}, z_{2t}, \ldots, z_{kt})'$. Then

$$E[z_t z_t'|I_{t-1}] = R_t \neq I_k$$

Once I estimate the univariate GARCH models, I can get estimated standardized residuals. Next I model the pairwise conditional covariances between the standardized residuals $\hat{\rho}_{ij,t} = \widehat{cov}(\hat{z}_{it}, \hat{z}_{jt})$.

1.12 Backtesting of Value-at-Risk

We need to see how good the model is in predicting VaR? To do so , we need to evaluate the model through *backtesting*. Actual profits and loss of a portfolio are compared to the corresponding VaR estimates. If the confidence level used for calculating daily VaR is 99%, expect an exception to occur once in every 100 days on average. There are two types of test as described below.

a) Unconditional coverage test as proposed in Kupiec [1995] statistically tests whether the frequency of violations over some given time interval is in line with the selected confidence level, and

b) Conditional coverage test as proposed by Christoffersen [1998] considers *correct* amount of violations as well as whether those exceptions are independent of each other.

Let I_t be an indicator variable such that $I_t = 0$ if VaR violation occurs and $I_t = 1$ if no VaR violation occurs. Proportion of Failures (POF) test suggested by Kupiec [1995] is described as follows. Let the number of VaR exceptions be N_0 and the total number of observations be T. Let $N_1 = T - N_0$. So the failure rate is given by $\pi = N_0/T$. in an ideal situation, this failure rate should be consistent with the VaR confidence level α. For a confidence level of 99%, the null hypothesis is that the frequency of tail losses should be equal to $p = (1 - \alpha) = 1 - 0.99 = 1\%$. Please note that based on whether VaR violation occurs or not, the sequence of success or failures can be seen as Bernoulli trials. So the number of failures N_0 follows the Binomial distribution.

$$f(N_0) = \binom{T}{N_0} p^{N_0}(1-p)^{N_1} \tag{1.13}$$

The null hypothesis for the POF test can be written as

$$H_0 : p = \pi = \frac{N_0}{T} \tag{1.14}$$

i.e. to check whether the observed failure rate π is significantly different from p.

This test is conducted as a likelihood-ratio (LR) test with the following test statistic

$$LR_{uc} = -2\ln\frac{(1-p)^{N_0}p^{N_1}}{(1-\pi)^{N_0}\pi^{N_1}} \overset{asy}{\sim} \chi^2(1) \tag{1.15}$$

But unconditional coverage omits one important fact i.e. zeros and ones may come in a cluster in a time-dependent manner. In order to rectify it, Christoffersen [1998] (see details) interval forecast test considers both frequency of VaR violations and the time when they occur. As part of the The LR test of independence, the independence part of conditional coverage hypothesis is tested against a first-order Markov alternative.

$$\Pi_1 = \begin{bmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{bmatrix}$$

where $\pi_{ij} = Pr(I_t = j | I_{t-1} = i)$.
The likelihood function for this test is given by $L(\Pi_1; I_1, I_2, \ldots, I_T) = (1-\pi_{01})^{N_{00}}\pi_{01}^{N_{01}}(1-\pi_{11})^{N_{10}}\pi_{11}^{N_{11}}$ where N_{ij} is the number of observations with value i followed by value j.

Next I describe the results of VaR backtesting on the VaR of different bivariate portfolios. For VaR backtesting I used *BacktestVaR* function in the *R* package *GAS* (Ardia et al. [2019]).

1.12.1 Backtesting of VaR for Commodities and Equities Portfolio

The tables 1.20, 1.21, 1.22 contain the results of the unconditional and conditional coverage tests for the correct number of exceedances on the VaR of the equally weighted portfolio consisting of commodities and equities for tail probabilities 1%, 5% and 10%. The tables show the actual number of exceedances, expected number of exceedances, the unconditional coverage test p-value, the decision of the unconditional coverage test based on the p-value, the conditional coverage test p-value, the decision of the conditional coverage test based on the p-value and the actual over expected exceedance ratio.

The results reveal the advantage of Student's t copula. The VaR model estimated using the

time varying Student's t copula performs much better better in estimating the portfolio risk.

	Student's t Copula	RiskMetrics™	DCC
Expected exceedances	97	97	97
Actual exceedances	95	206	183
UC LR p values	0.773	0.000	0.000
UC Decision	VaR consistent	VaR inconsistent	VaR inconsistent
CC LR p value	0.213	0.000	0.000
CC Decision	VaR consistent	VaR inconsistent	VaR inconsistent
AE	0.971	2.106	1.871

Table 1.20: 1% VaR Backtesting for CP and EQ Portfolio

	Student's t Copula	RiskMetrics™	DCC
Expected exceedances	489	489	489
Actual exceedances	459	566	517
UC LR p values	0.158	0.000	0.200
UC Decision	VaR consistent	VaR inconsistent	VaR inconsistent
CC LR p value	0.000	0.000	0.000
CC Decision	VaR inconsistent	VaR inconsistent	VaR inconsistent
AE	0.938	1.157	1.057

Table 1.21: 5% VaR Backtesting for CP and EQ Portfolio

	Student's t Copula	RiskMetrics™	DCC
Expected exceedances	978	978	978
Actual exceedances	968	1000	944
UC LR p values	0.728	0.466	0.245
UC Decision	VaR consistent	VaR consistent	VaR consistent
CC LR p value	0.000	0.000	0.000
CC Decision	VaR inconsistent	VaR inconsistent	VaR inconsistent
AE	0.989	1.022	0.965

Table 1.22: 10% VaR Backtesting for CP and EQ Portfolio

The VaR measures estimated using the time varying normal copula model picks up too many VaR violations, thus greatly overestimating the portfolio risk. Also, the average loss of the

VaR violations is much smaller in case of the Student's t copula. The better performance of Student's t copula is due to the fact that it captures tail dependence of commodities with equities to some degree of accuracy.

1.12.2 Backtesting of VaR for Commodities and Foreign Exchange Portfolio

The tables 1.23, 1.24, 1.25 (corresponding to tail probabilities 1%, 5% and 10%) showing the results of unconditional and conditional coverage tests on the VaR of the equally weighted

	Student's t Copula	RiskMetrics™	DCC
Expected exceedances	97	97	97
Actual exceedances	102	135	116
UC LR p values	0.674	0.000	0.073
UC Decision	VaR consistent	VaR inconsistent	VaR consistent
CC LR p value	0.003	0.001	0.002
CC Decision	VaR inconsistent	VaR inconsistent	VaR inconsistent
AE	1.043	1.380	1.186

Table 1.23: 1% VaR Backtesting CP and FX Portfolio

	Student's t Copula	RiskMetrics™	DCC
Expected exceedances	489	489	489
Actual exceedances	521	483	445
UC LR p values	0.144	0.775	0.038
UC Decision	VaR consistent	VaR consistent	VaR inconsistent
CC LR p value	0.001	0.048	0.001
CC Decision	VaR inconsistent	VaR inconsistent	VaR inconsistent
AE	1.065	0.987	0.910

Table 1.24: 5% VaR Backtesting CP and FX Portfolio

	Student's t Copula	RiskMetrics™	DCC
Expected exceedances	978	978	978
Actual exceedances	1022	882	839
UC LR p values	0.143	0.001	0.000
UC Decision	VaR consistent	VaR inconsistent	VaR inconsistent
CC LR p value	0.010	0.000	0.000
CC Decision	VaR consistent	VaR inconsistent	VaR inconsistent
AE	1.045	0.902	0.858

Table 1.25: 10% VaR Backtesting CP and FX Portfolio

portfolio consisting of commodities and foreign exchange clearly reveals the advantage of Student's t copula. The VaR model estimated using the time varying Student's t copula performs better in estimating the portfolio risk. The VaR measures estimated using the time varying normal copula model picks up only a few VaR violations, thus substantially under estimating the portfolio risk. Also, the average loss of the VaR violations is smaller in case of the Student's t copula. The better performance of Student's t copula is due to the fact that it captures tail dependence of commodities with equities and foreign exchange to some degree of accuracy.

1.12.3 Backtesting of VaR for Commodities and Bonds Portfolio

The tables 1.26, 1.27, 1.28 (corresponding to tail probabilities 1%, 5% and 10%) showing the results of unconditional and conditional coverage tests on the VaR of the equally weighted portfolio consisting of commodities and bonds reveals the advantage of Student's t copula. The VaR model estimated using the time varying Student's t copula performs better in estimating the portfolio risk. The VaR measures estimated using the RiskMetrics™ model picks up only a few VaR violations, thus substantially under estimating the portfolio risk. The better performance of Student's t copula is due to the fact that it captures tail dependence of commodities with bonds to some degree of accuracy.

	Student's t Copula	RiskMetrics™	DCC
Expected exceedances	74	74	74
Actual exceedances	75	101	91
UC LR p values	0.967	0.004	0.066
UC Decision	VaR consistent	VaR inconsistent	VaR consistent
CC LR p value	0.028	0.007	0.018
CC Decision	VaR consistent	VaR inconsistent	VaR consistent
AE	1.005	1.353	1.219

Table 1.26: 1% VaR Backtesting CP and BOND Portfolio

	Student's t Copula	RiskMetrics™	DCC
Expected exceedances	373	373	373
Actual exceedances	371	374	327
UC LR p values	0.907	0.966	0.012
UC Decision	VaR consistent	VaR consistent	VaR inconsistent
CC LR p value	0.090	0.997	0.010
CC Decision	VaR consistent	VaR consistent	VaR inconsistent
AE	0.994	1.002	0.876

Table 1.27: 5% VaR Backtesting CP and BOND Portfolio

	Student's t Copula	RiskMetrics™	DCC
Expected exceedances	746	746	746
Actual exceedances	371	674	633
UC LR p values	0.000	0.005	0.000
UC Decision	VaR inconsistent	VaR inconsistent	VaR inconsistent
CC LR p value	0.000	0.012	0.000
CC Decision	VaR inconsistent	VaR consistent	VaR inconsistent
AE	0.497	0.903	0.848

Table 1.28: 10% VaR Backtesting CP and BOND Portfolio

1.13 Backtesting of Expected Shortfall

VaR as a measure of risk has some drawbacks as Artzner et al. [1997] and Artzner et al. [1999] pointed out some of the deficiencies of VaR. It gives only an upper bound of the losses

that occur with a given frequency but tells nothing about the size of the loss. VaR is not a *"coherent"* measure of risk as VaR is also not *"subadditive"*, i.e., VaR of a portfolio can be higher than the sum of VaRs of the individual assets in the portfolio. But Expected Shortfall (ES) overcomes this deficiency and is a *"coherent"* measure of risk. McNeil and Frey [2000] proposed a test for expected shortfall. The test is described as follows. ES_α^t is the expected shortfall of the conditional loss distribution $F_{L_{t+1}}$ and define

$$S_{t+1} = (L_{t+1} - ES_\alpha^t)I_{t+1}.((Z_{t+1} - ES_\alpha(Z)) \,|\, I(Z_{t+1} > q_\alpha(Z))$$

is a zero-mean i.i.d. sequence of innovation variables. The null hypothesis is that the excess conditional shortfall (excess of the actual series when VaR is violated), is i.i.d. and has zero mean. The test is a one sided t-test against the alternative that the excess shortfall has mean greater than zero and thus that the conditional shortfall is systematically underestimated. I have used *ESTest* function in the *R* package *rugarch* (Ghalanos [2020]).

1.13.1 Backtesting of ES for Commodities and Equities Portfolio

The tables 1.29, 1.30, 1.31 show the results of the expected shortfall test of CP-EQ portfolio corresponding to tail probabilities 1%, 5% and 10%. Based on the confidence level and p-value, we conclude whether the expected shortfall is systematically underestimated. The results in the tables clearly shows that Student's t copula model produces better forecasts of ES when compared against RiskMetrics™ and Normal-DCC model.

	p value	Decision
Student's t Copula	0.3082	ES correctly estimated
RiskMetrics™	0.0000	ES underestimated
DCC	0.0000	ES underestimated

Table 1.29: CP and EQ: 1% Expected Shortfall Backtesting

	p value	Decision
Student's t Copula	0.3235	ES correctly estimated
RiskMetrics™	0.0000	ES underestimated
DCC	0.0000	ES underestimated

Table 1.30: CP and EQ: 5% Expected Shortfall Backtesting

	p value	Decision
Student's t Copula	0.4670	ES correctly estimated
RiskMetrics™	0.0000	ES underestimated
DCC	0.0000	ES underestimated

Table 1.31: CP and EQ: 10% Expected Shortfall Backtesting

1.13.2 Backtesting of ES for Commodities and Foreign Exchange Portfolio

The results in the tables 1.32, 1.33, 1.34 clearly show that Student's t copula model produces better forecasts of ES (for CP-FX portfolio corresponding to tail probabilities 1%, 5% and 10%) when compared against RiskMetrics™ and Normal-DCC model.

	p value	Decision
Student's t Copula	0.2875	ES correctly estimated
RiskMetrics™	0.0000	ES underestimated
DCC	0.0000	ES underestimated

Table 1.32: CP and FX: 1% Expected Shortfall Backtesting

	p value	Decision
Student's t Copula	0.4584	ES correctly estimated
RiskMetrics™	0.0000	ES underestimated
DCC	0.0000	ES underestimated

Table 1.33: CP and FX: 5% Expected Shortfall Backtesting

	p value	Decision
Student's t Copula	0.4564	ES correctly estimated
RiskMetrics™	0.0000	ES underestimated
DCC	0.0000	ES underestimated

Table 1.34: CP and FX: 10% Expected Shortfall Backtesting

1.13.3 Backtesting of ES for Commodities and Bonds Portfolio

The results in the tables tables 1.35, 1.36, 1.37 clearly show that Student's t copula model produces better forecasts of ES (for CP-BOND portfolio corresponding to tail probabilities 1%, 5% and 10%) when compared against RiskMetrics™ and Normal-DCC model.

	p value	Decision
Student's t Copula	0.2551	ES correctly estimated
RiskMetrics™	0.0000	ES underestimated
DCC	0.0000	ES underestimated

Table 1.35: CP and BOND: 1% Expected Shortfall Backtesting

	p value	Decision
Student's t Copula	0.4482	ES correctly estimated
RiskMetrics™	0.0000	ES underestimated
DCC	0.0000	ES underestimated

Table 1.36: CP and BOND: 5% Expected Shortfall Backtesting

	p value	Decision
Student's t Copula	0	ES underestimated
RiskMetrics™	0	ES underestimated
DCC	0	Reject H0

Table 1.37: CP and BOND: 10% Expected Shortfall Backtesting

1.14 Backtesting of VaR and ES on Three Assets Portfolio

Next I consider the three assets portfolio consisting of CP, EQ and FX and compute the VaR and ES corresponding to tail probabilities 1%, 5% and 10%. The tables 1.38, 1.39 and

	Student's t Copula	RiskMetrics™	DCC
Expected exceedances	97	97	97
Actual exceedances	102	181	164
UC LR p values	0.674	0.00	0.000
UC Decision	VaR consistent	VaR inconsistent	VaR inconsistent
CC LR p value	0.655	0.00	0.000
cc Decision	VaR consistent	VaR inconsistent	VaR inconsistent
AE	1.043	1.85	1.676

Table 1.38: 1% VaR Backtesting for CP, EQ, FX Portfolio

	Student's t Copula	RiskMetrics™	DCC
Expected exceedances	489	489	489
Actual exceedances	510	530	490
UC LR p values	0.337	0.061	0.969
UC Decision	VaR consistent	VaR consistent	VaR consistent
CC LR p value	0.001	0.005	0.003
CC Decision	VaR inconsistent	VaR inconsistent	VaR inconsistent
AE	1.043	1.084	1.002

Table 1.39: 5% VaR Backtesting for CP, EQ, FX Portfolio

	Student's t Copula	RiskMetrics™	DCC
Expected exceedances	978	978	978
Actual exceedances	1057	971	904
UC LR p values	0.009	0.805	0.011
UC Decision	VaR inconsistent	VaR consistent	VaR inconsistent
CC LR p value	0.000	0.000	0.000
CC Decision	VaR inconsistent	VaR inconsistent	VaR inconsistent
AE	1.080	0.993	0.924

Table 1.40: 10% VaR Backtesting for CP, EQ, FX Portfolio

1.40 show the VaR backtesting results. Student's t copula produces relatively better forecasts of VaR when compared against the other two alternative models. The tables 1.41, 1.42, and 1.43 show the ES backtesting results. The results in the tables clearly show that Student's t copula model produces better forecasts of ES when compared against RiskMetrics™ and Normal-DCC model.

	p value	Decision
Student's t Copula	0.1819	ES correctly estimated
RiskMetrics™	0.0000	ES underestimated
DCC	0.0000	ES underestimated

Table 1.41: CP, EQ, FX: 1% Expected Shortfall Backtesting

	p value	Decision
Student's t Copula	0.2869	ES correctly estimated
RiskMetrics™	0.0000	ES underestimated
DCC	0.0000	ES underestimated

Table 1.42: CP, EQ, FX: 5% Expected Shortfall Backtesting

	p value	Decision
Student's t Copula	0.4268	ES correctly estimated
RiskMetrics™	0.0000	ES underestimated
DCC	0.0000	ES underestimated

Table 1.43: CP, EQ, FX: 10% Expected Shortfall Backtesting

1.15 Backtesting of VaR and ES on Four Assets Portfolio

Next I consider the four assets portfolio consisting of CP, EQ, FX and BOND and compute the VaR and ES corresponding to tail probabilities 1%, 5% and 10%. The tables 1.44, 1.45, and 1.46 show the VaR backtesting results. Student's t copula produces relatively

better forecasts of VaR when compared against the other two alternative models when tail probability is very small.

	Student's t Copula	RiskMetrics™	DCC
Expected exceedances	74	74	74
Actual exceedances	75	185	157
UC LR p values	0.966	0.000	0.000
UC Decision	VaR consistent	VaR inconsistent	VaR inconsistent
CC LR p value	0.483	0.000	0.000
CC Decision	VaR consistent	VaR inconsistent	VaR inconsistent
AE	1.005	2.479	2.104

Table 1.44: 1% VaR Backtesting for CP, EQ, FX and BOND Portfolio

	Student's t Copula	RiskMetrics™	DCC
Expected exceedances	373	373	373
Actual exceedances	466	479	433
UC LR p values	0.000	0.000	0.002
UC Decision	VaR inconsistent	VaR inconsistent	VaR inconsistent
CC LR p value	0.000	0.000	0.000
CC Decision	VaR inconsistent	VaR inconsistent	VaR inconsistent
AE	1.249	1.284	1.160

Table 1.45: 5% VaR Backtesting for CP, EQ, FX and BOND Portfolio

	Student's t Copula	RiskMetrics™	DCC
Expected exceedances	746	746	746
Actual exceedances	940	808	736
UC LR p values	0.00	0.019	0.690
UC Decision	VaR inconsistent	VaR inconsistent	VaR consistent
CC LR p value	0.00	0.000	0.000
CC Decision	VaR inconsistent	VaR inconsistent	VaR inconsistent
AE	1.26	1.083	0.986

Table 1.46: 10% VaR Backtesting for CP, EQ, FX and BOND Portfolio

The tables 1.47, 1.48, and 1.49 show the ES backtesting results. The results in the tables clearly show that Student's t copula model produces better forecasts of ES when compared against RiskMetrics™ and Normal-DCC model.

	p value	Decision
Student's t Copula	0.9099	ES correctly estimated
RiskMetrics™	0.0000	ES underestimated
DCC	0.0000	ES underestimated

Table 1.47: CP, EQ, FX, BOND: 1% Expected Shortfall Backtesting

	p value	Decision
Student's t Copula	0.9998	ES correctly estimated
RiskMetrics™	0.0000	ES underestimated
DCC	0.0000	ES underestimated

Table 1.48: CP, EQ, FX, BOND: 5% Expected Shortfall Backtesting

	p value	Decision
Student's t Copula	0.9961	ES correctly estimated
RiskMetrics™	0.0000	ES underestimated
DCC	0.0000	ES underestimated

Table 1.49: CP, EQ, FX, BOND: 10% Expected Shortfall Backtesting

1.16 Portfolio Optimization

To find out the effect of tail dependence between commodities and other financial assets, I also explore the question of how to construct an optimal portfolio by minimizing expected shortfall or expected tail loss. Then I compare the performance of Global Minimum Variance (GMV) portfolio to that of Global Minimum Expected Shortfall (GMES) portfolio. A brief overview of these methods have been given in sections 2.2 and 2.3 in Chapter 2. The discussion and

notations in this section closely follow an unpublished manuscript of Martin, Philips, Scherer and Li [2021][2]. I dynamic portfolio optimization using the *R* package *PortfolioAnalytics* (Peterson and Carl [2018]). To evaluate the performance of the portfolios, I look at the cumulative return (for dynamic optimization using rolling window).

1.16.1 Global Minimum Variance Portfolio Optimization

The portfolio Mean-Variance optimization theory was proposed by Harry Markowitz in his seminal paper Markowitz [1952a]. The basic principle of this theory states that an investor, in order to choose an optimal portfolio of assets, minimizes the variance of the portfolio's returns at any given level of the portfolio's mean return. As the returns mean vector $\boldsymbol{\mu}$ and the returns covariance matrix $\boldsymbol{C}$ are unknown, in practice they are estimated from the historical data $\boldsymbol{r}_t = (r_{t1}, \ldots, r_{tN})', t = 1, 2, \ldots, N$. The estimator for $\boldsymbol{\mu}$ is the $N \times 1$ vector of sample means and the estimator for the covariance matrix $\boldsymbol{C}$ is the $N \times N$ sample covariance matrix estimator $\hat{\boldsymbol{C}}$. When the returns follow multivariate normal distribution, the sample covariance matrix is the maximum-likelihood-estimator (MLE) of the true but unknown covariance matrix $\boldsymbol{C}$. At time t, let us assume that $\boldsymbol{w}_t = (w_{1t}, w_{2t}, \ldots, w_{Nt})$ denote the vector of portfolio weights for N assets. For simplicity, we can exclude the time subscript and define the portfolio return as $r_P = \boldsymbol{w'r}$. These weights are determined at the beginning of the time interval over which the returns are computed at the end of the interval.

We can express the portfolio mean return vector in terms of asset mean return vector and portfolio weight vector as $\mu_P = \boldsymbol{w'\mu}$ and portfolio variance $\sigma_P^2 = var(r_P)$ in terms of covariance matrix $\boldsymbol{C}$ as $\sigma_P^2 = \boldsymbol{w'Cw}$. This is in quadratic form and we assume that the covariance matrix is positive definite. Then for any non-zero weight vector $\boldsymbol{w}$ portfolio variance is positive. The *global minimum variance* (GMV) portfolio is defined as the portfolio with minimum possible variance that can be achieved with any fully invested portfolio. The

[2]To prevent copyright violation, please do not cite or use materials from this section without explicit consent from the authors of the unpublished manuscript. Professor R. Douglas Martin (Professor Emeritus, University of Washington, Seattle) can be reached at doug@amath.washington.edu

GMV portfolio weight vector $\boldsymbol{w}_{gmv}$ is the solution of the following minimization problem

$$\boldsymbol{w}_{gmv} = \arg\min_{\boldsymbol{w}} \sigma^2_{gmv}(\boldsymbol{w}) = \arg\min_{\boldsymbol{w}} \boldsymbol{w'C w}$$

subject to the full investment constraint $\boldsymbol{w'1} = 1$. The optimal weight vector is computed as

$$\boldsymbol{w} = \frac{\boldsymbol{C}^{-1}\mathbf{1}}{(\mathbf{1}\boldsymbol{C}^{-1}\mathbf{1})^{-1}}$$

To evaluate the performance of the portfolio performance we look at Sharpe Ratio (SR) Sharpe [1964]. This is the most commonly used risk-adjusted performance measure. The Sharpe Ratio is defined as the portfolio mean return μ_P less a risk-free rate r_f divided by the portfolio volatility

$$\sigma_P : SR = \frac{\mu_P - r_f}{\sigma_P}$$

1.16.2 Global Minimum Expected Shortfall Optimization

Now I turn to construct an optimal portfolio based on the minimum expected shortfall. Rockafellar et al. [2000] proposed portfolio optimization based on *conditional Value-at-Risk* (CVaR) or expected shortfall (ES). Expected shortfall is defined as

$$ES_\gamma(F) = -E(R|R \leq q_\gamma(F))$$

where R is the return following the distribution function F and $E(R|R \leq q_\gamma(F))$ is the conditional expectation of return R, conditioned on the fact that R is less than or equal to its γ-quantile $q_\gamma(F)$. In the context of a portfolio, R is the portfolio return $r_P = r_P(\boldsymbol{w}) = \boldsymbol{w'r}$ where $\boldsymbol{r}$ is the return vector of assets included in the portfolio and $\boldsymbol{w}$ is the portfolio weight vector. In case of portfolio optimization, the expected shortfall is dependent on the weight vector $\boldsymbol{w}$. The portfolio optimization problem based on minimum expected shortfall can be written as

$$\boldsymbol{w}_{minES} = \arg\min_{\boldsymbol{w}} ES_\gamma(\boldsymbol{w})$$

where the weight vector $\boldsymbol{w}$ is subject to a convex set of constraints i.e. $\boldsymbol{w} \in \boldsymbol{C}$ where $\boldsymbol{C}$

contains full-investment constraint $\boldsymbol{w'1} = 1$ and the portfolio mean return constraint

$$E(\boldsymbol{w'r}) \geq \mu_P, -\infty < \mu_P < +\infty$$

Rockafellar et al. [2000] proposed an algorithm which solves the above optimization problem using linear programming. For more detailed overview see section **??** of Chapter 2. To evaluate the portfolio the ES Ratio is defined as

$$ES\ Ratio(\boldsymbol{w}) = \frac{\mu_P(\boldsymbol{w}) - r_f}{ES(r_P(\boldsymbol{w}))} = \frac{\boldsymbol{w'\mu} - r_f}{ES(\boldsymbol{w'r})}$$

Next I outline the results of the dynamic portfolio optimization using the daily data of commodities, equities and foreign exchange. To evaluate the performance of dynamic portfolio optimization, we look at the cumulative portfolio return from both strategies. Figure 1.9 clearly shows that the optimal strategy of minimizing portfolio expected shortfall generates a higher cumulative portfolio return starting from early 2002 than the cumulative return generated by global GMV strategy.

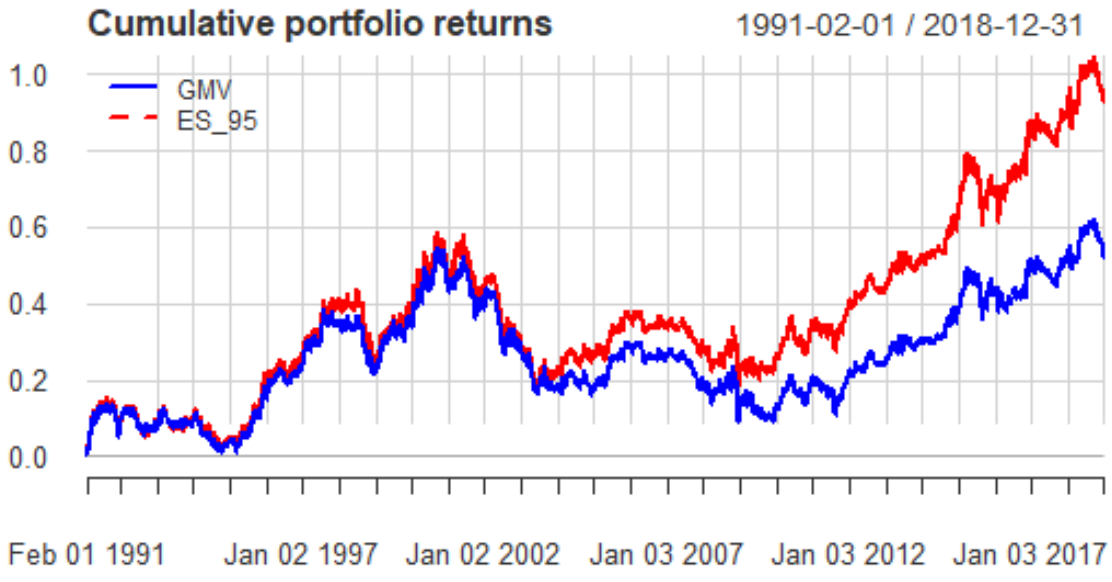

Figure 1.9: Portfolio cumulative return

Figure 1.10: Portfolio drawdown

Figure 1.10 shows the portfolio drawdown from both GMV and GMES strategies. Starting from late 2002 the GMES portfolio showed a lower drawdown relative to that GMV showing that GMES strategy results in lower loss starting from late 2002.

1.17 Conclusion

In this paper, I model the time-varying *"tail dependence"* of commodities with other financial assets - equities, foreign exchange and bonds with the help of Student's t copula. I extend the two-dimensional time varying Student's t copula model to higher dimensions such as three and four dimension copula. Based on the static copula models, I find that the tail dependence of commodities with other financial assets increased after 2004 as *"financialization"* of commodities started. I also show that using time varying Student's t copula model, we can produce better forecasts of tail-based risk measures such as Value-at-Risk (VaR) and Expected Shortfall (ES) for portfolios consisting of commodities and other financial assets. The results of the statistical tests for evaluating the VaR and ES show clear evidence

in favor of time-varying Student's t copula model when compared against the benchmark RiskMetrics™ or Normal-DCC models as these models are not equipped to capture the tail dependence between different financial assets. I think paying attention to tail dependence modeling can help us gain new insights in the portfolio risk management. I also looked at the effect of tail dependence on portfolio optimization by forming optimal portfolio minimizing portfolio expected shortfall. I compare the performance of expected shortfall strategy vs. global minimum variance strategy. For dynamic optimization, the expected shortfall strategy generates a higher portfolio cumulative return and lower drawdown than those generated by the GMV strategy. As a conclusion we can say using copula in modeling tail dependence between commodities and other financial assets can help us in risk management and portfolio optimization.

Chapter 2

PORTFOLIO OPTIMIZATION BASED ON DOWNSIDE RISK ESTIMATES

This second chapter describes a study which is a joint work with Professor R. Douglas Martin.[1] In this chapter we construct the following three types of global minimum risk optimal portfolio using daily returns of 30 small cap stocks (which has the largest market capitalization within the small cap category) - a) Global Minimum Variance (GMV), b) Global Minimum Expected Shortfall (GMES), and c) Global Minimum Expected Quadratic Shortfall (GMEQS). We conduct both the static as well as dynamic (i.e. with rebalancing) portfolio optimization and analyze the portfolio performance metrics such as Sharpe Ratio (SR), Downside Sharpe Ratio (DSR), Expected Shortfall Ratio (ES Ratio), cumulative return, cumulative return relative to a benchmark and drawdown.

2.1 Introduction and Review of Literature

Portfolio mean-variance optimization (MVO), which is usually implemented as an equivalent quadratic utility (QU) optimization, has been the workhorse model since the publication of Harry Markowitz's seminal paper Markowitz [1952b], the Monograph Markowitz [1959], the book Markowitz (1987), and the Journal of Finance paper Markowitz [1991]. Due to its mathematical foundation, intuitive appeal and simplicity, MVO has been quite popular and widely applied in practice in asset management, and the QU version is the foundation of a number of commercial portfolio optimization and risk management software products[2].

[1]Professor Emeritus, Professor of Statistics, Adjunct Professor of Finance, Former Chair of the Department of Statistics, Founder Director of the Computational Finance and Risk Management Program (Applied Mathematics), University of Washington, Seattle, WA 98195-3330. Email: `doug@amath.washington.edu`

[2]These include Qontigo Axioma, MSCI Barra, and Northfield, among other.

The basic principle of mean-variance portfolio optimization is that an investor should select a portfolio of assets such that the variance of the portfolio's returns is minimized at any specified level of the portfolio's mean return. If we plot the portfolio volatility (i.e. standard deviation) on the horizontal axis and portfolio mean return on the vertical axis, the shape of the plot of all possible portfolio mean-volatility combinations appear to have a smooth curved left boundary and this curve is called the portfolio frontier. Each portfolio on the portfolio frontier has the minimum possible variance for that level of portfolio mean return. The set of all such portfolios is called MinVar portfolios. Among them the portfolio that achieves the minimum possible variance that can be achieved by any fully invested portfolio is referred to as a *global minimum variance* (GMV) portfolio. The GMV portfolio is positioned at the leftmost tip of the portfolio frontier. The portion of the portfolio frontier where the MinVar portfolios have higher portfolio mean return and higher volatility is called the *efficient frontier*. It can be shown that the points along the efficient frontier can be obtained via maximizing mean-variance quadratic utility.

While both MVO and QU are intuitively appealing and easy to implement, they do not result in portfolios that whose weights take into account the skewness and kurtosis of returns distribution. But asset returns data (especially of higher frequency such as weekly and daily) exhibit non-normality in terms of skewness and fat tails. In addition to that, variance is a symmetric measure as it penalizes both negative and positive returns equally. In reality investors will only care about downside when their returns fall below the average or some threshold return and will not mind an upside. This threshold level of return which distinguishes a desirable outcome from an undesirable one, is sometimes called Minimum Acceptable Return (MAR). Downside risk measures the risk below this minimum acceptable return. Therefore, in order to penalize only the negative returns, investors may profit by using portfolio optimization based on downside risk measures. In Chapter 9 of his 1959 monograph, Markowitz focused on the fact that symmetric nature of variance does not make it an accurate measure of risk, and proposed semi-variance (SV) as an alternative risk measure.

Using semi-variance as the risk measure, an investor should invest in an optimal portfolio that maximizes a mean-semivariance utility function. But it failed to gain popularity as the mean-semivariance "utility" function maximization problem is not a quadratic programming problem. So it never caught on in practice. Among other alternatives for downside risk measures, lower partial moments of order two or higher have been used by Fishburn [1977] and Price et al. [1982]. Eventually, MVO became the most well known workhorse portfolio construction model in both academia and industry.

One of the most popular downside risk measures is Expected Shortfall (ES), also known as Expected Tail Loss (ETL) and Conditional Value-at-Risk (CVaR). ES is defined as the average of loss beyond the Value-at-Risk (VaR) and captures the tail characteristics. The VaR is a limited risk measure as it is only a quantile of the returns distribution (giving no information about the size of losses beyond that quantile) and it does not satisfy the "coherence" axioms as suggested in Artzner et al. [1997] and Artzner et al. [1999] and is thoroughly discussed in McNeil et al. [2015] book. While the ES not only is more informative, it also satisfies the "coherence axioms" mentioned earlier. The fact that expected shortfall is a coherent risk measure and variance is not a coherent risk measure makes ES based portfolio optimization particularly attractive. As a matter of fact, by 2013 and 2014, Bank of International Settlements (BIS) began recommending the use of ES instead of VaR. Expected shortfall gained popularity as the objective risk measure in portfolio optimization with the publication of the seminal paper of Rockafellar et al. [2000] on Conditional Value-at-Risk (CVaR) portfolio optimization. In order to construct an optimal portfolio by minimizing the ES, Rockafellar et al. [2000] proposed an optimization algorithm which can be solved by standard linear programming with linear inequality constraints. As a result, very large dimension problem involving hundreds or thousands of assets can be rapidly solved using open source and commercial linear program solvers. This not only has overcome the technical problem of minimizing the VaR as objective risk measure in the portfolio optimization context as VaR is non-convex, but the more important aspect is that VaR does not quantify the losses

beyond the VaR. Using the ES instead of VaR has shown promising results. Some important references related to this discussion are Acerbi and Tasche [2002], [Gaivoronski and Pflug, 2005], Hellmich and Kassberger [2011], Brandtner [2013], Alexander and Baptista [2004], Benati [2003], Bertsimas et al. [2004].

These methods of portfolio optimization follow the risk-reward approach. But the classical approach to decision making under uncertainty is built on the von Neumann and Morgenstern (vNM) utility theory as suggested in von Neumann and Morgenstern [1947]. The idea of making the risk-reward optimization model and risk measures consistent with expected utility maximization has gained much attention in research. As discussed in Krokhmal [2007] and other papers such as Rothschild and Stiglitz [1970], Pflug [2000], it is possible to construct coherent risk measures consistent with the vNM utiltiy theory. Krokhmal [2007] showed ways how risk measures can be expressed as solutions to stochastic programming problems and proposed Higher Moment Coherent Risk (HMCR) measures. It showed how the HMCR measures can be implemented by reducing it to a p-order conic programming and approximating via linear programming. Krokhmal [2007] also discussed the special case where $p = 2$ and defined it as the Second Moment Coherent Risk Measure (SMCR). The SMCR has similar properties as CVaR but it measures risk in terms of the second moments of loss distributions. To formulate the SMCR measure into a mathematical programming problem, second-order cone constraints should be used. The necessary tool to solve convex optimization problems involving second-order cone constraints is the Second-order Cone Programming (SOCP). Krokhmal [2007] outlines an algorithm how a portfolio optimization problem of minimizing HMCR can be transformed into a linear programming problem with a single p-order cone constraint. In this paper, we call the Second Moment Coherent Risk as Expected Quadratic Shortfall (EQS), which is a natural variant of ES.

2.2 Review of Mean-Variance Portfolio Theory

In this section we briefly outline the Mean-Variance portfolio theory following the discussion in Martin, Philips, Scherer and Li [2021][3]. The portfolio Mean-Variance optimization the-

ory was introduced by Harry Markowitz in his seminal papers Markowitz [1952b], Markowitz [1959]. The basic principle of mean-variance portfolio optimization is that an investor should select a portfolio of assets such that the variance of the portfolio's returns is minimized at any specified level of the portfolio's mean return. Let a portfolio have N assets and $\boldsymbol{r}_t$ be an $N \times 1$ vector of asset returns at time $t = 1, 2, \ldots, T$. For simplicity we assume that asset returns $r_{i,t}$ have constant mean $\mu_i = E(r_i)$, and the covariances between returns r_i and r_j are $\boldsymbol{C}_{ij}$ where $i, j = 1, \ldots, N$ and $t = 1, 2, \ldots, T$. Let $\boldsymbol{\mu}$ be the $N \times 1$ vector of asset mean returns and $\boldsymbol{C}$ be the $N \times N$ covariance matrix of asset returns. At time t, let us assume that $\boldsymbol{w}_t = (w_{1t}, w_{2t}, \ldots, w_{Nt})$ denote the vector of portfolio weights for N assets. For simplicity, we can exclude the time subscript and define the portfolio return as $r_P = \boldsymbol{w'r}$. These weights are determined at the beginning of the time interval over which the returns are computed at the end of the interval. We can express the portfolio mean return vector in terms of asset mean return vector and portfolio weight vector $\mu_P = E(r_P) = \boldsymbol{w'\mu}$. The portfolio variance can be written in terms of covariance matrix $\boldsymbol{C}$ as $\sigma_P^2 = var(r_P) = var(\boldsymbol{w'r}) = \boldsymbol{w'Cw}$. As this is in quadratic form and assuming that the covariance matrix is positive definite, we can say for any non-zero weight vector $\boldsymbol{w}$ portfolio variance is positive. By taking square root of σ_P^2, we get the portfolio volatility $\sigma_P = (\boldsymbol{w'Cw})^{1/2}$.

The condition that a portfolio is fully invested in risky assets is represented by the condition that the weight vector satisfies the *full–investment* constraint:

$$\boldsymbol{w'1} = \sum_{i=1}^{N} w_i = 1 \tag{2.1}$$

The weights do not have to be non–negative, as negative weights that correspond to borrowing through short–selling are generally possible. When the full-investment constraint is

[3]This is an unpublished manuscript. To prevent copyright violation, please do not cite or use materials from this section without explicit consent from the authors of the unpublished manuscript. Professor R. Douglas Martin (Professor Emeritus, University of Washington, Seattle) can be reached at doug@amath.washington.edu

satisfied, the portfolio weight vector $\boldsymbol{w}$ solves one of the following types of portfolio optimization problems.

1. *MinVar Portfolios*: The fully invested *minimum variance* (MinVar) portfolio minimizes the portfolio variance $\sigma_P^2(\boldsymbol{w}) = \boldsymbol{w'Cw}$ with respect to $\boldsymbol{w}$, subject to full-investment constraint $\boldsymbol{w'1} = 1$ and the portfolio mean return constraint $\boldsymbol{w'\mu} = \mu_P$, where μ_P is the portfolio manager's specified portfolio mean return.

2. *QU Portfolios*: The *maximum quadratic utility* (MaxQU) portfolios maximize the quadratic utility function $QU(\boldsymbol{w}) = \mu_P(\boldsymbol{w}) - \frac{1}{2}\lambda\sigma_P^2(\boldsymbol{w})$ subject to the full-investment constraint, where $\lambda > 0$ is the portfolio manager's risk aversion parameter.

3. *MaxMean Portfolios* The MaxMean portfolios maximize the portfolio mean return $\mu_P(\boldsymbol{w}) = \boldsymbol{w'\mu}$, subject to the full-investment constraint and the portfolio variance constraint $\boldsymbol{w'Cw} = \sigma_P^2$, where σ_P is the portfolio manager's specified portfolio volatility.

The weight vector $\boldsymbol{w}_{mv}$ of a fully invested minimum variance portfolio solves the following constrained optimization problem is the solution of the following minimization problem

$$\boldsymbol{w}_{mv} = \arg\min_{\boldsymbol{w}} \boldsymbol{w'Cw} \tag{2.2}$$

subject to the full-investment constraint

$$\boldsymbol{w'1} = 1 \tag{2.3}$$

and mean return constraint

$$\boldsymbol{w'\mu} = \mu_P \tag{2.4}$$

The MinVar problem has a solution for all possible positive and negative values of μ_P.

The *global minimum variance* (GMV) portfolio is defined as the portfolio with minimum possible variance that can be achieved with any fully invested with no portfolio mean return constraint. The GMV portfolio weight vector $\boldsymbol{w}_{gmv}$ is the solution of the following minimization problem

$$\boldsymbol{w}_{gmv} = \arg\min_{\boldsymbol{w}} \boldsymbol{w}'\boldsymbol{C}\boldsymbol{w} \tag{2.5}$$

subject to the full-investment constraint

$$\boldsymbol{w}'\mathbf{1} = 1 \tag{2.6}$$

As the above minimization problem involves a linear equality constraint, it can be solved by the method of Lagrange multipliers. The Lagrangian for the GMV problem is given by:

$$L(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}'\boldsymbol{C}\boldsymbol{w} + \gamma(1 - \boldsymbol{w}'\mathbf{1}) \tag{2.7}$$

Solving for GMV portfolio, we get

$$\boldsymbol{w}_{gmv} = \frac{\boldsymbol{C}^{-1}\mathbf{1}}{(\mathbf{1}\boldsymbol{C}^{-1}\mathbf{1})^{-1}} \tag{2.8}$$

The portfolio mean return is given as

$$\mu_{gmv} = \boldsymbol{w}'_{gmv}\boldsymbol{\mu} = \frac{\boldsymbol{C}^{-1}\mathbf{1}\boldsymbol{\mu}}{(\mathbf{1}\boldsymbol{C}^{-1}\mathbf{1})^{-1}} \tag{2.9}$$

and the variance of the *global minimum variance* portfolio is given by

$$\sigma^2_{gmv} = \boldsymbol{w}'_{gmv}\boldsymbol{C}\boldsymbol{w}_{gmv} = (\mathbf{1}\boldsymbol{C}^{-1}\mathbf{1})^{-1} \tag{2.10}$$

The MinVar problem has a solution for all possible positive and negative values of μ_P, and in particular it has solutions for values of μ_P less than, as well as greater than, the mean return μ_{gmv} of the global minimum variance portfolio. But, as we shall see, for every MinVar portfolio with $\mu_P = \mu_{gmv} - c$ with $c < 0$, there exists a MinVar portfolio with $\mu_P = \mu_{gmv} + c$.

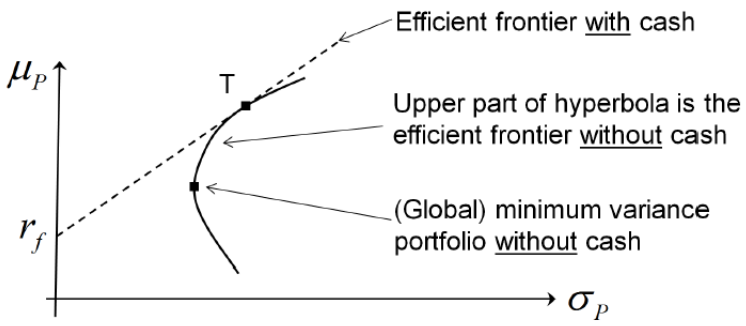

Figure 2.1: Efficient frontier

Source: Martin [2015]

The latter portfolios are called efficient frontier MinVar portfolios, and the former are inefficient MinVar portfolios. As $\boldsymbol{\mu}$ and $\boldsymbol{C}$ are unknown, in practice they are estimated from the historical returns data $\boldsymbol{r}_t$. Let $\hat{\boldsymbol{\mu}}$ be the $N \times 1$ vector of sample means estimator and let the sample covariance matrix estimator be $\hat{\boldsymbol{C}}$. The MaxQU portfolio problem has solutions for all values of $\lambda > 0$, and in the limit when the portfolio manager's risk aversion becomes arbitrarily large, i.e., when $\lambda \to \infty$, the portfolio becomes the global minimum variance portfolio. The MaxMean problem has a solution only for values of the portfolio variance that are at least as large as that of the global minimum variance portfolio. MaxQU and MaxMean portfolios are equivalent in the sense that the set of all solutions to these two problems are the same. Also MaxQU portfolios and efficient MinVaR portfolios are equivalent in the sense that the set of all efficient MinVar portfolios is the same as the set of all MaxQU portfolios. Mean-Variance optimal portfolio theory has been extensively covered in graduate level text

books such as Cochrane [2009], Huang and Litzenberger [1988], Pennacchi [2008].

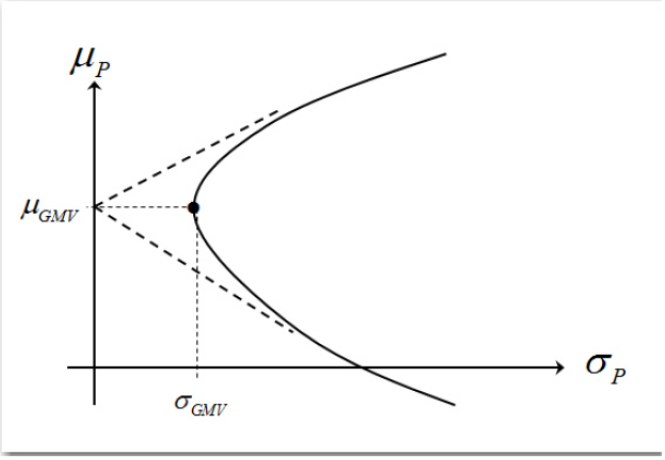

Figure 2.2: No-cash MinVar efficient frontier

Source: Martin [2015]

It can be shown that the following relationship exists between MVO portfolio mean and variance (for details see the cited references):

$$\mu_{mv} = \mu_{gmv} \pm \sqrt{d}\sigma_{gmv}(\sigma_{mv}^2 - \sigma_{gmv}^2)^{1/2}, \quad \sigma_{mv} \geq \sigma_{gmv} \tag{2.11}$$

where d is given by

$$d = \mu' \mathbf{C}^{-1} \mu \mathbf{1}' \mathbf{C}^{-1} \mathbf{1} - (\mathbf{1}' \mathbf{C}^{-1} \mu)^2 > 0$$

The above equation 2.11 i.e. the μ_{mv} vs. σ_{mv} curve describes what is known as *portfolio frontier*.

To evaluate the performance of a portfolio, the Sharpe Ratio (SR) (Sharpe [1964]) is the most commonly used risk-adjusted performance measure. It is defined as the portfolio mean return μ_P less a risk-free rate r_f divided by the portfolio volatility σ_P

$$SR = \frac{\mu_P - r_f}{\sigma_P} \tag{2.12}$$

where the risk–free rate in the U.S. is usually taken to be the return on a three–month Treasury bill.

2.3 Expected Shortfall and Expected Quadratic Shortfall Portfolio Theory

While both MVO and QU are intuitively appealing and easy to implement, they do not result in portfolios that whose weights take into account downside portfolio risk due to the skewness and kurtosis of returns distribution. But asset returns data (especially of higher frequency such as weekly and daily) exhibit non-normality in terms of skewness and fat tails. In addition to that, variance is a symmetric measure as it penalizes both negative and positive returns equally. But a large positive return is not undesirable to an investor. Therefore, in order to penalize only the negative returns, investors may profit by using portfolio optimization based on downside risk measures. In the 1990's, J.P. Morgan introduced and popularized Value–at–Risk (VaR) for the purpose of measuring extreme downside returns risks, as is fully described in their classic RiskMetrics™ Technical Document (1996). Value–at–Risk is defined in terms of a quantile $q_\gamma(F)$ of a returns distribution $F(r)$, where γ is a small tail probability and the γ-quantile satisfies the equation $P(R \leq q_\gamma(F)) = F(q_\gamma(F)) = \gamma$.[4]

Intuitively, VaR is a very limited tail risk measure since, being a quantile of the distribution of change in wealth, it gives no indication of the size of losses beyond that quantile. Furthermore, VaR suffers greatly from lacking coherence (see Artzner et al. [1997] and Artzner et al. [1999]). Many possible risk measures have been proposed in the literature. But how should

[4]Here we follow the discussion and notations from 2021 draft chapters of the book Portfolio Optimization and Risk Management being developed by Martin, Philips, Scherer and Li.

the risk manager choose one? The axioms of coherent risk measures serve as a guidance. We describe these axioms.

L : end of period loss random variable (loss taken as positive)

ρ : a "function" of L(through distribution function of L)

Axiom 1 ***Positive Homogeneity (POSHOM)*** $\rho(\lambda L) = \lambda\rho(L), \quad \lambda \geq 0$ *i.e. when loss is multiplied by a positive factor, risk must increase by the same factor.*

Axiom 2 ***Monotonicity (MONO)*** $L_1 > L_2 \implies \rho(L_1) > \rho(L_2)$ *i.e. greater loss implies greater risk. (inequality with probability one).*

Axiom 3 ***Sub-Additivity (SUBADD)*** $\rho(L_1 + L_2) \leq \rho(L_1) + \rho(L_2)$ *i.e. the risk of aggregated losses is not greater than the sum of the individual risks.*

Axiom 4 ***Cash Reduction of Risk (CRR)*** $\rho(L - \alpha) = \rho(L) - \alpha$ *i.e. the addition of cash (risk-free position) decreases the risk.*

Axiom 1 and **Axiom 2** imply that ρ is convex:

$$\rho(\lambda L_1 + (1-\lambda)L_2) \leq \lambda\rho(L_1) + (1-\lambda)\rho(L_2) \quad \forall\, L_1, L_2,\, 0 < \lambda < 1$$

Around the turn of the century the academic literature began to focus on a risk measure called expected shortfall (ES), which is defined loosely as the average of the losses beyond VaR. If we assume that the random return R has a continuous and strictly increasing distribution function F, the expected shortfall, with loss defined as a positive quantity, is defined as:

$$ES_\gamma(F) = -E(R|R \leq q_\gamma(F)) \tag{2.13}$$

where $E(R|R \leq q_\gamma(F))$ is the conditional expectation of return R, conditioned on the fact that R is less than or equal to its γ-quantile $q_\gamma(F)$. The VaR is the negative of the γ of

the quantile. In the context of a portfolio, R is the portfolio return $r_P = r_P(\boldsymbol{w}) = \boldsymbol{w'r}$ where $\boldsymbol{r}$ is the return vector of assets included in the portfolio and $\boldsymbol{w}$ is the portfolio weight vector. Substituting portfolio return r_P in place of general asset return R in (2.13), portfolio expected shortfall has the form:

$$ES_\gamma(F_{\boldsymbol{r}}) = -E(\boldsymbol{w'r}|\boldsymbol{w'r} \leq q_\gamma(\boldsymbol{w}, F(\boldsymbol{r}))) \tag{2.14}$$

where $F(\boldsymbol{r})$ is the multivariate distribution of asset return vector $\boldsymbol{r}$. In case of portfolio optimization, the expected shortfall is dependent on the weight vector $\boldsymbol{w}$. So the above expression can also be written as

$$ES_\gamma(\boldsymbol{w}) = -E(\boldsymbol{w'r}|\boldsymbol{w'r} \leq q_\gamma(\boldsymbol{w})) \tag{2.15}$$

It is straightforward to show that ES can be written in the integral form

$$ES_\gamma(\boldsymbol{w}) = -\frac{1}{\gamma}\int_{-\infty}^{q_\gamma(\boldsymbol{w})} \boldsymbol{w'r}\, dF(\boldsymbol{r}) \tag{2.16}$$

Although VaR is a much simpler risk measure than ES, an increased interest in ES was driven by two important factors. The first is that ES quantifies the size of the losses beyond VaR. The second reason for increased interest is ES has been shown to satisfy the above coherence axioms whereas this is not the case for VaR. Consequently, by 2013 and 2014 the Bank of International Settlements began recommending the use of ES instead of VaR. Expected shortfall gained popularity as the objective risk measure in portfolio optimization with the publication of the seminal paper of Rockafellar et al. [2000] on Conditional Value-at-Risk (CVaR) portfolio optimization. In order to construct an optimal portfolio by minimizing the ES, Rockafellar et al. [2000] proposed an optimization algorithm which can be solved by standard linear programming with linear inequality constraints. The portfolio optimization problem based on minimum expected shortfall can be written as

$$\boldsymbol{w}_{minES} = \arg\min_{\boldsymbol{w}} ES_\gamma(\boldsymbol{w}) \tag{2.17}$$

where the weight vector $\boldsymbol{w}$ is subject to a convex set of constraints i.e. $\boldsymbol{w} \in \boldsymbol{C}$ where $\boldsymbol{C}$ is any convex set in Rockafellar et al. [2000]. The general problem analogous the MVO is to choose $\boldsymbol{C}$ to be linear inequality constraints, the most basic of which is the full-investment and target mean return constraints, but other important constraints include commonly used long-only and box constraints.

The full-investment constraint is given by

$$\boldsymbol{w'1} = 1 \tag{2.18}$$

and the portfolio mean return constraint is given by

$$E(\boldsymbol{w'r}) \geq \mu_P, -\infty < \mu_P < +\infty \tag{2.19}$$

and the long-only weight constraint can be expressed as

$$0 \leq w_i \leq 1, \quad i = 1, \ldots, N$$

.

To formulate portfolio optimization problem based on minimum expected shortfall as per Rockafellar et al. [2000], we add and subtract $q_\gamma(\boldsymbol{w})$ inside the integral of (2.16), thereby obtaining the two term ES form:

$$\begin{aligned} ES_\gamma(\boldsymbol{w}) &= -\frac{1}{\gamma} \int_{-\infty}^{q_\gamma(\boldsymbol{w})} \left[\boldsymbol{w'r} + q_\gamma(\boldsymbol{w}) - q_\gamma(\boldsymbol{w})\right] dF(\boldsymbol{r}) \\ &= -q_\gamma(\boldsymbol{w}) + \frac{1}{\gamma} \int_{-\infty}^{q_\gamma(\boldsymbol{w})} \left[q_\gamma(\boldsymbol{w}) - \boldsymbol{w'r}\right] dF(\boldsymbol{r}) \\ &= -q_\gamma(\boldsymbol{w}) + \frac{1}{\gamma} \int_{-\infty}^{q_\gamma(\boldsymbol{w})} \left[q_\gamma(\boldsymbol{w}) - \boldsymbol{w'r}\right]^+ dF(\boldsymbol{r}) \end{aligned} \tag{2.20}$$

where $\left[x\right]^+$ denotes the positive part of x. Replacing $q_\gamma(\boldsymbol{w})$ by a free variable t, the objective function can be defined as

$$F_\gamma(\boldsymbol{w}, t) = -t + \frac{1}{\gamma} \int_{-\infty}^{q_\gamma(\boldsymbol{w})} \left[t - \boldsymbol{w}'\boldsymbol{r}\right]^+ dF(\boldsymbol{r}) \tag{2.21}$$

With the above expression we can now state the important main result of Rockafellar et al. [2000].

Theorem 1 (Rockafellar and Uryasev, 2000) *The function $F_\gamma(\boldsymbol{w}, t)$ is convex in $\boldsymbol{w}$ and t, continuously differentiable in t and:*

$$ES_\gamma(\boldsymbol{w}) = \arg\min_t F_\gamma(\boldsymbol{w}, t)$$

$$\arg\min_{\boldsymbol{w}} minES_\gamma(\boldsymbol{w}) = \arg\min_{\boldsymbol{w},t} F_\gamma(\boldsymbol{w}, t), \quad \boldsymbol{w} \in \boldsymbol{C} \text{ convex}$$

When there are N observed return vectors $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N$, each consisting of returns on N assets, we can obtain an empirical version $\hat{\boldsymbol{F}}_\gamma(\boldsymbol{w}, t)$ of the objective function $\boldsymbol{F}_\gamma(\boldsymbol{w}, t)$ by using the plug-in rule of substituting the empirical distribution function F_n for the unknown true distribution F of the asset returns as

$$\hat{\boldsymbol{F}}_\gamma(\boldsymbol{w}, t) = -t + \frac{1}{\lceil n\gamma \rceil} \sum_{i=1}^{N} \left[t - \boldsymbol{w}'\boldsymbol{r}_i\right]^+ \tag{2.22}$$

and compute the estimated ES optimal portfolio weights $\boldsymbol{w}$ by solving:

$$\arg\min_{\boldsymbol{w},\boldsymbol{t}} \hat{\boldsymbol{F}}_\gamma(\boldsymbol{w}, t)$$

.

Now we show the linear programming formulation of the above optimization problem. It is clear that minimizing $\hat{\boldsymbol{F}}_\gamma(\boldsymbol{w}, t)$ is the same as maximizing $-\hat{\boldsymbol{F}}_\gamma(\boldsymbol{w}, t)$, and it is shown that maximizing,

$$-\hat{\boldsymbol{F}}_\gamma(\boldsymbol{w}, t) = t - \frac{1}{n\gamma} \sum_{i=1}^{N} \left[t - \boldsymbol{w}'\boldsymbol{r}_i\right]^+ \tag{2.23}$$

where $t \in R$, and $\boldsymbol{w} \in \boldsymbol{C} \subset \boldsymbol{R}^p$.

And it is equivalent to solving the following linear programming problem (see the proof[5] in Appendix B.1):

$$\underset{\boldsymbol{w_i},\{\boldsymbol{e_i}\},\boldsymbol{t}}{\arg\max}\left(t-\frac{1}{n\gamma}\sum_{i=1}^{N}e_i\right) \tag{2.24}$$

such that

$$e_i \geq t-\sum_{j=1}^{p}w_j r_{ij}, \quad i=1,2,...,n \tag{2.25}$$

$$e_i \geq 0 \quad i=1,2,...,n \tag{2.26}$$

$$\sum_{j=1}^{p}w_j\mu_j \geq \mu_0 \tag{2.27}$$

subject to $t \in R, \boldsymbol{w} \in \boldsymbol{C} \subset \boldsymbol{R}^p$, e.g., long-only, box, group constraint.

We know that a standard linear programming problem can be formulated as follows:

Find $\mathbf{x}$ vector of order $m \times 1$ that solves:

$$\underset{\boldsymbol{x}}{\arg\max}\,\mathbf{c}'\mathbf{x}$$

subject to:

$$\mathbf{Ax} \leq \mathbf{b}$$

$$\mathbf{x} \geq \mathbf{0}$$

[5] The proof is due to Professor R. Douglas Martin, see Martin [2015].

where

$$\begin{array}{lllll} \mathbf{c} & \text{is} & m \times 1 & \text{vector} \\ \mathbf{A} & \text{is} & k \times m & \text{matrix} \\ \mathbf{b} & \text{is} & k \times 1 & \text{vector} \end{array}$$

If we want the constraint $\mathbf{Ax} \geq \mathbf{b}$ then we can just write it in the above form by changing the signs of both sides:

$$-\mathbf{Ax} \geq -\mathbf{b}$$

So the linear programming formulation of ES minimization can also be expressed using the matrix notation as follows:

$$\begin{aligned} \mathbf{x} =& (w_1, \ldots, w_p, e_1, e_2, \ldots e_n, t)' \quad \text{is} \quad (p+n+1) \times 1 \\ \mathbf{c}' =& \left(\mathbf{0}_{1\times p}, -\frac{1}{\gamma}\mathbf{1}_{1\times n}, 1 \right) \quad \text{is} \quad 1 \times (p+n+1) \end{aligned}$$

$$\mathbf{R}_{n\times p} = \begin{pmatrix} r_{11} & \ldots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{n1} & \ldots & r_{np} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\mu}_{1\times p} = (\mu_1, \ldots, \mu_p)$$

$$\mathbf{A} = \begin{pmatrix} \mathbf{R}_{n\times p} & \mathbf{I}_{n\times n} & -\mathbf{1}_{n\times 1} \\ \mathbf{0}_{n\times p} & \mathbf{I}_{n\times n} & \mathbf{0}_{n\times 1} \\ \boldsymbol{\mu}_{1\times p} & \mathbf{0}_{1\times n} & 0 \end{pmatrix} \quad \text{is} \quad (2n+1) \times (p+n+1)$$

$$\mathbf{b}_{2n+1} = (\mathbf{0}_{1\times n}, \mathbf{0}_{1\times n}, \mu_0)'$$

In the risk management context, the emphasis on ES has been for small tail probabilities i.e. $\gamma = 0.025$ or 2.5%. But in case of ES portfolio optimization small tail probabilities, e.g., $1\%, 2.5\%, 5\%$ may cause considerable uncertainty in data based estimate of ES. As a consequence, the portfolio performance may be affected. See for example Scherer [2009]. For our empirical study in Section 2.5, we use not only a traditional 5% tail probability, but also

10%, 25% and 50%[6]. The choice of a 50% tail probability in ES optimal portfolio is very close in spirit to semi–variance optimal portfolios that Markowitz (see Markowitz [1959]) had initially been interested in. See also this recent paper Markowitz et al. [2020].

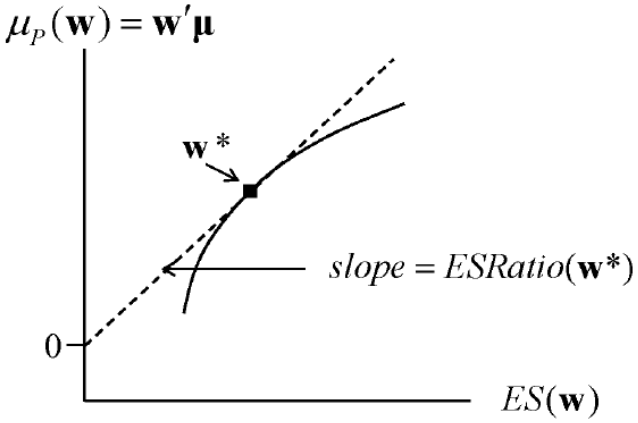

Figure 2.3: ES ratio

Source: Martin [2015]

To evaluate the portfolio the *ES Ratio* is defined as

$$ES\ Ratio(\boldsymbol{w}) = \frac{\mu_P(\boldsymbol{w}) - r_f}{ES(r_P(\boldsymbol{w}))} = \frac{\boldsymbol{w'\mu} - r_f}{ES(\boldsymbol{w'r})} \tag{2.28}$$

Minimum expected shortfall optimization lacks the quadratic downside penalty of minimum variance optimization. So it is natural to consider an expected quadratic shortfall risk measure to minimize. Krokhmal [2007] showed ways how risk measures can be expressed as solutions to stochastic programming problems and proposed Higher Moment Coherent Risk (HMCR) measures. It showed how the HMCR measures can be implemented by reducing it

[6]Basel II recommends 2.5%, but we regard that as having too much ES estimation risk.

to a p-order conic programming and approximating via linear programming. Here we briefly describe HMCR using the notations following Krokhmal [2007]. Let $\Re(X)$ be a risk measure of a random outcome X defined as $\chi \mapsto \boldsymbol{R}$ is some linear space of $\mathcal{F}$-measurable functions. Here X can be thought of as a loss function implying small values are good while large values are bad.

Theorem 2 *Let the function $\phi : X \mapsto \boldsymbol{R}$ satisfy axioms (A1)–(A3), and be a lower semi-continuous (lsc) function such that $\phi(\eta) > \eta$ for all real $\eta \neq 0$. Then the optimal value of the stochastic programming problem*

$$\rho(X) = \inf_{\eta} \ \eta + \phi(X - \eta) \tag{2.29}$$

is a proper coherent risk measure, and the infimum is attained for all X, so inf in (2.29) may be replaced by $\min_{\eta \in \boldsymbol{R}}$.

For some $0 < \alpha < 1$ and $\phi(X) = (1-\alpha)^{-1}\|(X)^+\|_p$, where $\|(X)\|_p = (E|X|^p)^{1/p}$, Krokhmal [2007] proposed a Higher Moment Risk Measure (HMCR) which is defined as

$$HMCR_{p,\alpha}(X) = \arg\min_{\eta \in \boldsymbol{R}} \eta + (1-\alpha)^{-1} \left\|(X-\eta)^+\right\|_p \tag{2.30}$$

where $p \geq 1, \quad \alpha \in (0,1)$.

The optimal $\eta_{p,\alpha}(X)$ is the tail cut-off point and we can adjust it by choosing the parameter α. The implementation of HMCR measures reduces to a p-order conic programming and can be approximated via linear programming. The case $p = 2$ defines the Second Moment Risk Measure (SMCR)

$$SMCR_{\alpha}(X) = \arg\min_{\eta \in \boldsymbol{R}} \eta + (1-\alpha)^{-1} \left\|(X-\eta)^+\right\|_2, \quad 0 < \alpha < 1 \tag{2.31}$$

The $SMCR_{\alpha}(X)$ has similar properties as $CVaR_{\alpha}(X)$ but it measures risk in terms of the second moments of loss distributions. To formulate the SMCR measure into a mathematical

programming problem second-order cone constraints should be used. The necessary tool to solve convex optimization problems involving second-order cone constraints is the second-order cone programming (SOCP).

The development of interior-point methods in the 1980s made it possible to solve new variants of convex optimization problems such as semi-definite problems and second-order cone problems etc. As stated in Alizadeh and Goldfarb [2003], the second-order cone programming (SOCP) problems are a special kind of convex optimization problems in which a linear objective function is minimized subject to a second-order cone constraint. As per Boyd et al. [2004], a set C is called a *cone,* if for every $x \in C$ and $\theta \geq 0$ we have $\theta x \in C$. And C is a convex cone, if for any $x_1, x_2 \in C$ and $\theta_1, \theta_2 \geq 0$, we have $\theta_1 x_1 + \theta_2 x_2 \in C$. A convex cone is depicted in Figure 2.4 where all the points of the form $\theta_1 x_1 + \theta_2 x_2$ are shown by the pie slice. If $\| \, . \, \|$is any norm on $\boldsymbol{R}^n$, then the norm cone can be defined as the set

$$C = \{(x, t) \mid \| \, x \, \| \leq \, t\} \subseteq \boldsymbol{R}^{n+1} \tag{2.32}$$

In case of the Euclidean norm, the norm cone is the *second-order cone* given by

$$\begin{aligned} C =& \{(x, t) \in \boldsymbol{R}^{n+1} \mid \| \, x \, \|_2 \leq \, t\} \\ =& \left\{ \begin{bmatrix} x \\ t \end{bmatrix} \middle| \begin{bmatrix} x \\ t \end{bmatrix}^T \begin{bmatrix} \boldsymbol{I} & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x \\ t \end{bmatrix} \leq 0, \, t \geq 0 \right\} \end{aligned} \tag{2.33}$$

It is also called a *quadratic cone* or the *Lorentz cone* or the *ice-cream cone.* Figure 2.5 displays the second-order cone in $\boldsymbol{R}^3$.

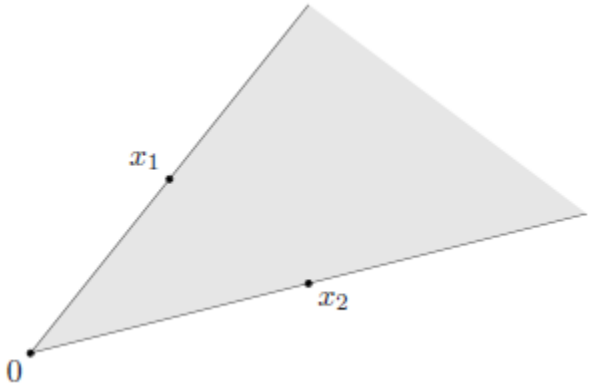

Figure 2.4: A convex cone

Source: Boyd et al. [2004]

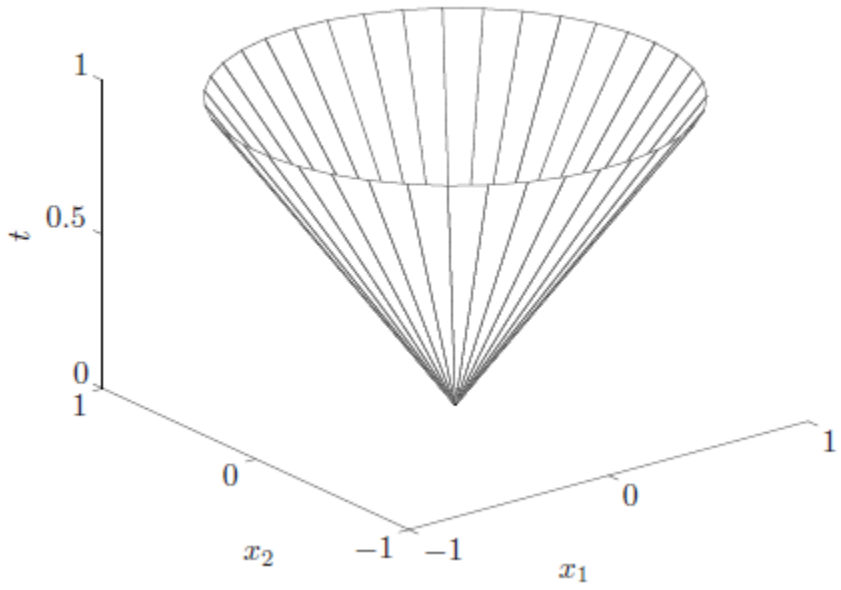

Figure 2.5: A second-order cone in $\boldsymbol{R}^3$, $\{(x_1, x_2, t) \quad (x_1^2 + x_2^2)^{1/2} \leq t\}$

Source: Boyd et al. [2004]

Krokhmal [2007] outlines an algorithm how a portfolio optimization problem of minimizing

HMCR can be transformed into a linear programming problem with a single p-order cone constraint. The problem can be formulated as follows

$$\begin{aligned} \min_{\boldsymbol{x}} \Re(-\boldsymbol{r}^T\boldsymbol{x}) \\ \text{s.t. } \boldsymbol{e}^T\boldsymbol{x} = 1 \\ E(\boldsymbol{r}^T\boldsymbol{x}) \geq r_0 \\ \boldsymbol{x} \geq 0 \end{aligned} \tag{2.34}$$

where $\boldsymbol{x} = (x_1, x_2, ..., x_n)^T$ the vector of portfolio weights and the asset returns vector is given by $\boldsymbol{r} = (r_1, r_2, ..., r_n)^T$ and $\boldsymbol{e} = (1, 1, ..., 1)^T$. The three constraints are budget constraint, minimum required portfolio return as r_0 and no short-selling constraint. The budget constraint along with no short-selling constraint imply all funds are invested. In stochastic programming the random asset return r_i can be modeled using J discrete equiprobable scenarios $\{r_{i1}, r_{i2}, ..., r_{iJ}\}$.Now we can write the term $\|(X-\eta)\|_2^+$ in Equation 2.31 in the following way:

$$\begin{aligned} \left\|(X-\eta)^+\right\|_2 =& (E|(X-\eta)^+|^2)^{1/2} \\ =& \left(\frac{1}{J}\left[\left((X_1-\eta)^+\right)^2 + \left((X_2-\eta)^+\right)^2 + \ldots + \left((X_J-\eta)^+\right)^2\right]\right)^{1/2} \\ =& \frac{1}{J^{1/2}}\left[w_1^2 + w_2^2 + \ldots + w_J^2\right]^{1/2} \end{aligned} \tag{2.35}$$

where in the minimization problem $(X_1-\eta)^+, (X_2-\eta)^+, \ldots, (X_J-\eta)^+$ are replaced by auxiliary variables $w_1 \geq 0, w_2 \geq 0, \ldots, w_J \geq 0$ respectively.

Following the algorithm in Krokhmal [2007], the problem of minimizing second moment coherent risk as defined in Equation 2.31 can be transformed into a linear programming problem with a second-order cone constraint as follows:

$$\min \ \ \eta + \frac{1}{1-\alpha} \left\| (X-\eta)^+ \right\|_2 = \min \ \ \eta + \frac{1}{1-\alpha} \frac{1}{J^{1/2}} t \tag{2.36}$$

$$\text{s.t.} \ \sum_{i=1}^{n} x_i = 1$$

$$w_j \geq -\sum_{i=1}^{n} r_{ij} x_i - \eta, \ \ j = 1, 2, ..., J$$

$$t \geq (w_1^2 + w_2^2 + ... + w_J^2)^{1/2}$$

$$x_i \geq 0, \ \ i = 1, 2, ..., n$$

$$w_j \geq 0, \ \ j = 1, 2, ..., J$$

In the above formulation of the minimization problem $w_1, w_2, \ldots, w_J$ and t are introduced as auxiliary variables. It is clear that the constraint

$$t \geq (w_1^2 + w_2^2 + ... + w_J^2)^{1/2} \quad \text{i.e.} \quad \| \boldsymbol{w} \|_2 \leq t$$

is a second order cone constraint as $C = \{(\boldsymbol{w}, t) \mid \| \boldsymbol{w} \|_2 \leq t\} \subseteq \boldsymbol{R}^{J+1}$ is a second order cone. All other constraints in the above problem are linear constraints. So the above problem boils down to Second Order Cone Programming (SOCP) problem. To solve this SOCP problem we have used an R package *CVXR* developed by Fu et al. [2020].

2.4 Data and Methodology

We have returns data of daily frequency for 300 stocks from Center for Research in Security Prices (CRSP). These stocks are divided into four groups based on market capitalization - a) Large Cap, b) Mid Cap, c) Small Cap and d) Micro Cap. For our analysis we have selected the top 30 stocks based on the market capitalization within the Small Cap sector. In this exercise we consider log returns data as for small returns both log returns and arithmetic returns are close and the multi-period log returns have a simple additive structure. We constructed an optimal portfolio based on three strategies - GMV, GMES and GMEQS.

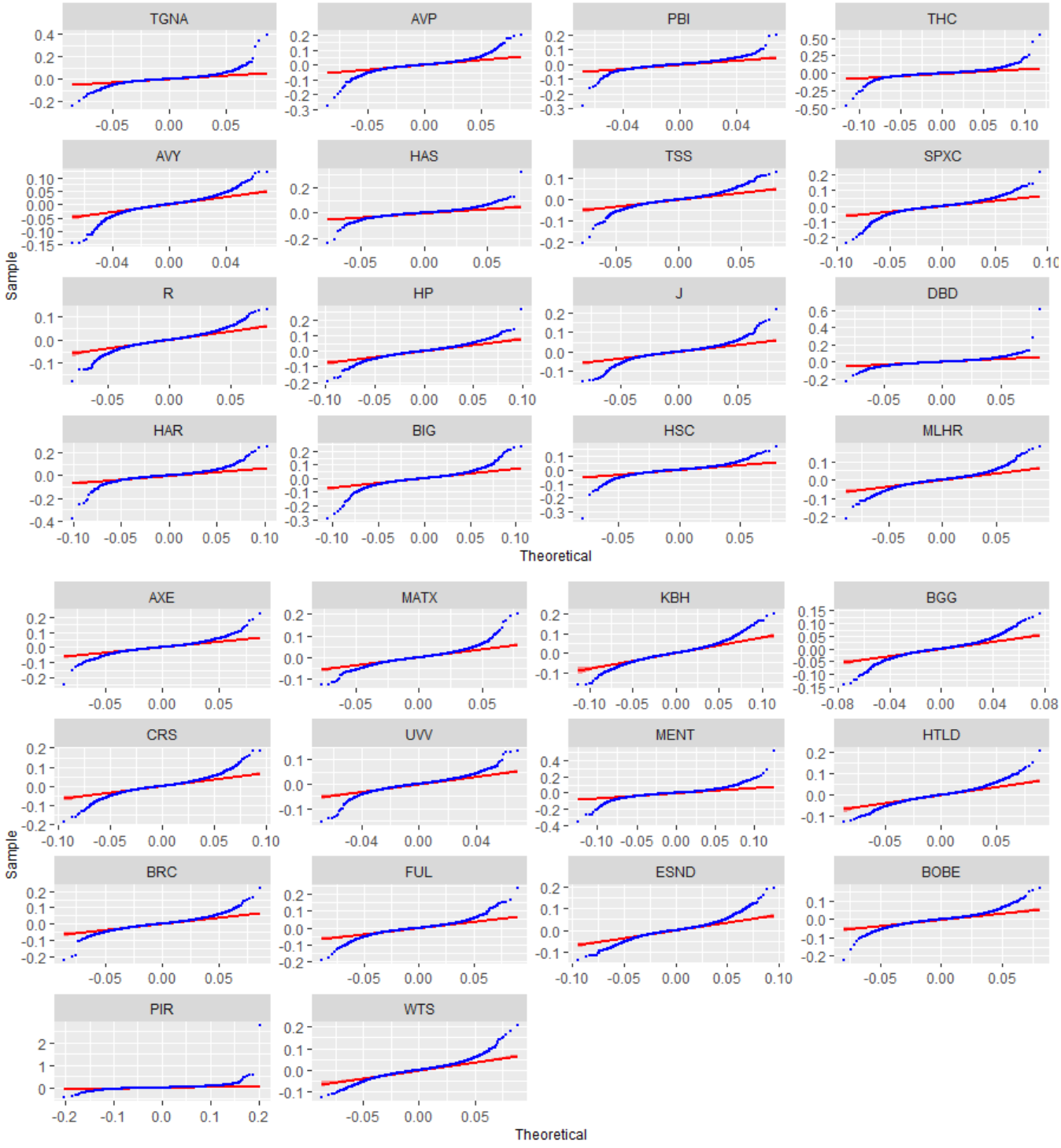

Figure 2.6: QQ Plot of CRSP Small Cap Stock Returns

The sample spans from January 1993 to December 2015 giving us 5793 daily observations for each stock. For the static portfolio optimization, we use full sample and for dynamic optimization we optimize the portfolio rebalancing on "months". It is already established in empirical asset pricing literature that asset returns of higher frequency exhibit non-normality i.e. negative skewness and fat tails. From the QQ (quantile-quantile) plot as shown in figure 2.6, we see that most of the 30 stock returns have fat tails and negative skewness. This is also confirmed from the skewness and kurtosis displayed in table B.1 in Appendix B. From the daily returns data of 30 small cap stocks, we constructed optimal portfolio with monthly rebalancing i.e. we computed optimal portfolio weights at the end of each month and used a rolling window of 500 days. For each of the three strategies GMV, GMES and GMEQS, we first constructed optimal portfolio subject to long-only full-investment constraint. But rebalancing a portfolio involves selling some assets in the portfolio and buying others. Such buying and selling does not come for free, and the associated transaction costs results in reduced profit. Indeed, uncontrolled buying and selling can result in transaction costs that can wipe out a very high percentage if not all of the profit in a portfolio strategy. There are several levels of sophistication in controlling transaction costs. The simplest is to assume that each purchase and sale of assets costs the same amount per unit of asset, and in our analysis we focus on one way *turnover* (TO) defined at time t as:

$$TO(\boldsymbol{w}_t, \boldsymbol{w}_{t-1}) = \sum_{i=1}^{N} |w_{t,i} - w_{t-1,i}| \tag{2.37}$$

With $w_{t-1,i} = w_i^{init}$ we can express $w_{t,i} = w_i^{init} + w_i^+ - w_i^-$ where $w_i \geq 0$ represents a buy and $w_i^- \geq 0$ represents a sale, only one of which will occur. Then we have

$$\begin{aligned} TO(\boldsymbol{w}, \boldsymbol{w}^{init}) &= TO(\boldsymbol{w}^+, \boldsymbol{w}^-) \\ &= \sum_{i=1}^{N} |w_i^+ - w_i^-| \\ &= \sum_{i=1}^{N} \left(w_i^+ - w_i^-\right) \end{aligned} \tag{2.38}$$

A basic long-only turnover constrained MinVar optimization problem with positive definite covariance matrix $\boldsymbol{\Sigma}$ and turnover upper bound constraint toc is defined by (see Martin [2015]):

$$\min_{\boldsymbol{w}} \ \boldsymbol{w'\Sigma w}$$

$$\text{subject to } \boldsymbol{w'1} = 1$$

$$\boldsymbol{w'\mu} = \mu_P$$

$$\boldsymbol{w} - \boldsymbol{w}^+ + \boldsymbol{w}^- = \boldsymbol{w}^{init}$$

$$1'\boldsymbol{w}^+ + 1'\boldsymbol{w}^- \leq toc$$

$$\boldsymbol{w}^+ \geq 0$$

$$\boldsymbol{w}^- \geq 0$$

$$\boldsymbol{w} \geq 0$$

This set of constraints presents a problem in that the there is now an effective weights vector $\tilde{\boldsymbol{w}} = (\mathbf{w},\mathbf{w}^+,\mathbf{w}^-)$ of dimension $3N$ instead of the original dimension N of $\mathbf{w}$. Consequently we now need to minimize the quadratic form

$$\min_{\tilde{\mathbf{w}}} \tilde{\mathbf{w}}'\tilde{\boldsymbol{\Sigma}}\tilde{\mathbf{w}}$$

where

$$\tilde{\boldsymbol{\Sigma}} = \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0}_N & \mathbf{0}_N \\ \mathbf{0}_N & \mathbf{0}_N & \mathbf{0}_N \\ \mathbf{0}_N & \mathbf{0}_N & \mathbf{0}_N \end{pmatrix}$$

is an $3N \times 3N$ matrix and $\mathbf{0}_N$ is an $N \times N$ matrix of zeros. However, a slightly modified formulation of the above problem is given as follows:

$$\min_{\mathbf{w}} \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}$$

$$\text{subject to } \boldsymbol{w}'\mathbf{1} = 1$$

$$\boldsymbol{w}'\boldsymbol{\mu} = \mu_P$$

$$\boldsymbol{w} - \boldsymbol{w}^+ + \boldsymbol{w}^- = \boldsymbol{w}^{init}$$

$$1'\boldsymbol{w}^+ + 1'\boldsymbol{w}^- = y$$

$$y \geq 0$$

$$\boldsymbol{w}^+ \geq 0$$

$$\boldsymbol{w}^- \geq 0$$

$$\boldsymbol{w} \geq 0$$

In the process of optimal portfolio construction, sometimes too much weight may be concentrated on a few assets of the portfolio. An investor would like to minimize the concentration risk by diversifying the portfolio. Generally, the sum of the squared portfolio weights i.e. $\sum_{i=1}^{N} w_i^2$ can serve as a weights concentration measure. If all the weight is concentrated in one asset, then we have$\sum_{i=1}^{N} w_i^2 = 1$. On the other hand, if the portfolio is equally weighted, i.e., if $w_i = 1/N$, then $\sum_{i=1}^{N} w_i^2 = 1/N$. So to measure the diversification of a portfolio, we compute the following metric $DIV(\boldsymbol{w}) = 1 - \sum_{i=1}^{N} w_i^2$. The concentration risk can be reduced either by using box constraints or by including the concentration risk term $\sum_{i=1}^{N} w_i^2$ as a penalty in the objective function of the portfolio optimization.

To implement the portfolio optimization strategies in this exercise we have used R statistical programming language, R Core Team [2021] and R packages such as *PerformanceAnalytics* (Peterson and Carl [2020]), *PortfolioAnalytics* (Peterson and Carl [2018]) and *CVXR* (Fu et al. [2020]).

2.5 Results

We now describe the results of our analysis. We use daily returns data of 30 small cap stocks and form optimal portfolio at the end of every month using three strategies - GMV, GMES and GMEQS and long-only constraints. We use tail probabilities 5%, 10%, 25% and 50% when forming portfolios based on ES and EQS and we also do the analysis both with and without turnover constraints. Below we describe the results of portfolio optimization without turnover constraints and with turnover constraints.

2.5.1 Without Turnover Constraints

The figures 2.7, 2.8, 2.9 and 2.10 show the performance of the portfolios formed by GMV, GMES and GMES strategies for 5%, 10%, 25% and 50% tail probabilities. We show the cumulative geometric return over time and the relative performance of the portfolios against the S&P 500 index return. The function *chart.RelativePerformance* plots the ratio of cumulative geometric return of these portfolios comparing against the benchmark S&P 500 index. The plot in the top panel of figure 2.7 shows the portfolio growth i.e. the cumulative geometric return (reflecting compounding) of the portfolio over time and the plot in the bottom panel shows the active performance of the portfolios against the benchmark S&P 500 index return. Cumulative return of the portfolio formed by the GMEQS strategy is greater than that of two other strategies for almost entire duration of the time span we are considering. Even during the Global Financial Crisis (GFC) of 2007-08 while all the portfolios show a steep decline in returns, the GMEQS strategy perform slightly better than the other two.

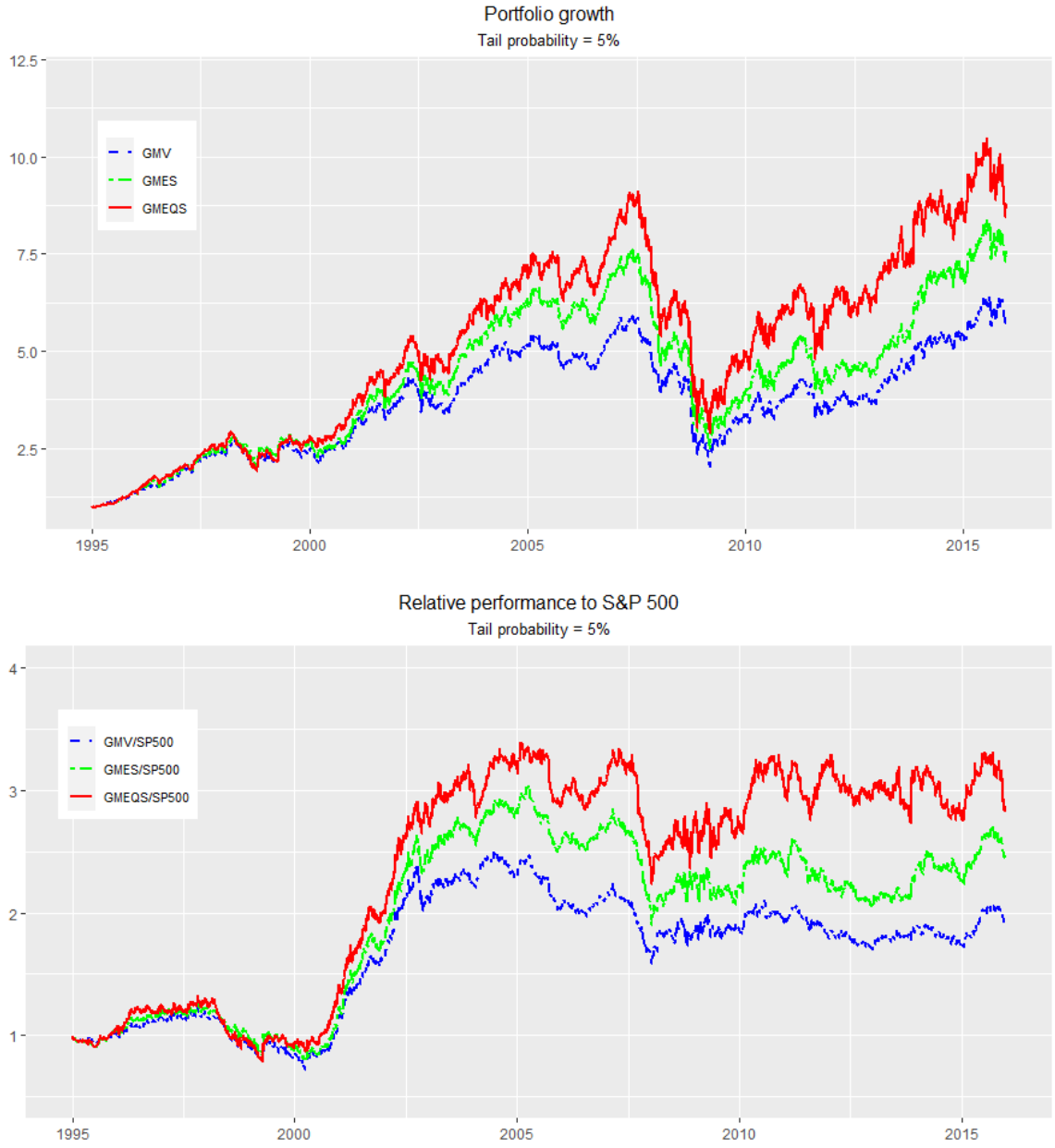

Figure 2.7: Cumulative Return of Portfolio: without TOC (tail prob. 5%)

This is also evident from plot in the lower panel where we see GMEQS shows the best relative performance. Among the three strategies GMV performs the worst indicating that we need to pay attention to the non-normal features of the asset returns data and focus on tail based

Figure 2.8: Cumulative Return of Portfolio: without TOC (tail prob. 10%)

risk measures while optimizing portfolios. When tail probability is 10% GMEQS performance dominates that of the GMV and GMES portfolios in terms of cumulative return and relative performance as evident in figure 2.8. Interestingly, GMES and GMV do not show much

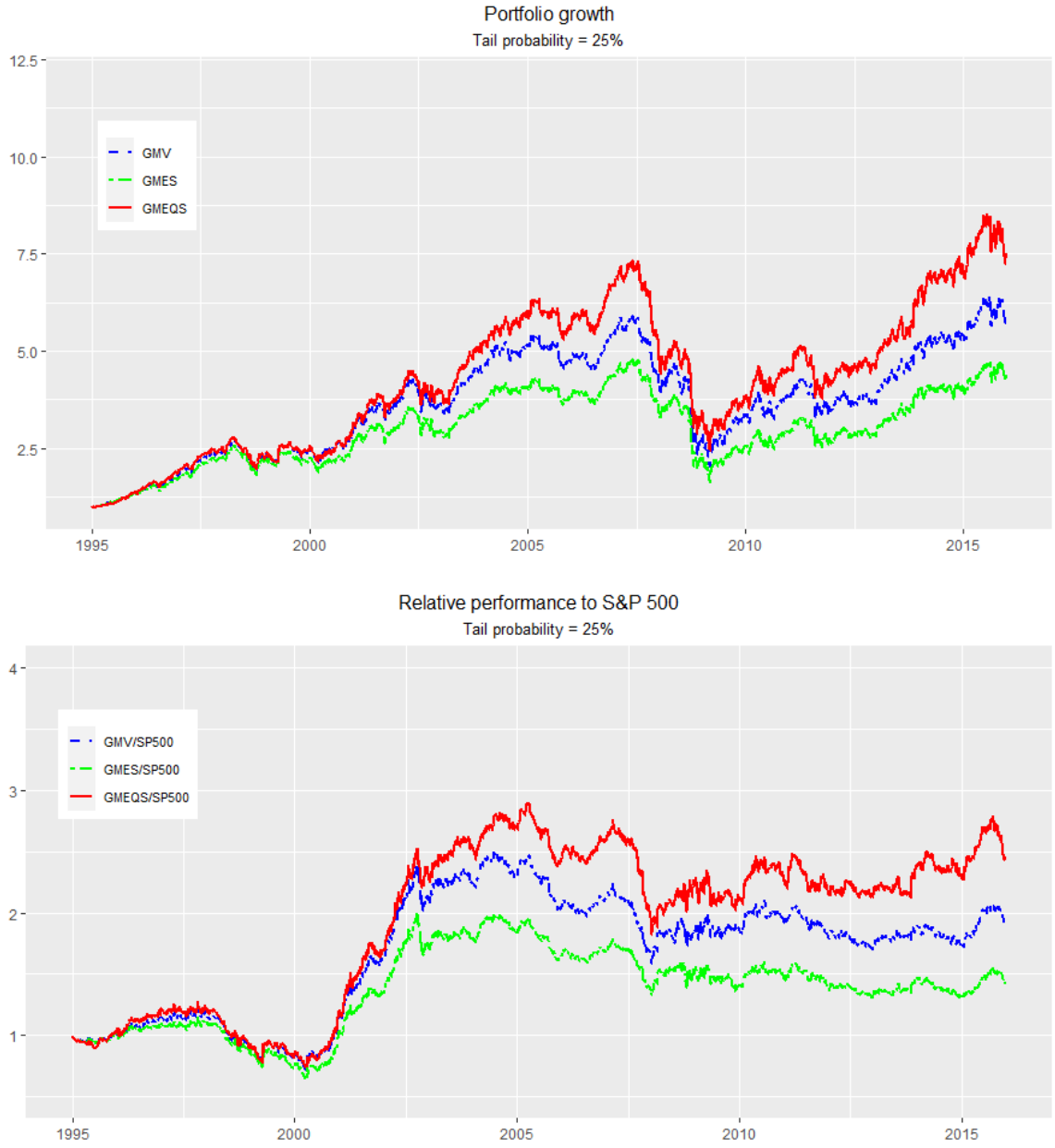

Figure 2.9: Cumulative Return of Portfolio: without TOC (tail prob. 25%)

difference in performance in this case. When tail probability is 25%, figure 2.9 shows that GMEQS is still the best performer of all three strategies. But GMV now shows a better performance than GMES. This is because as we move towards the center of the distribution

Figure 2.10: Cumulative Return of Portfolio: without TOC (tail prob. 50%)

(with increasing tail probability) the returns distribution seems closer to normal distribution and the skewness and kurtosis play a less important role. So minimizing expected shortfall does give more advantage than minimizing variance. The same logic applies to the case

when tail probability is 50%. As it covers the left-half of the distribution, the skewness and kurtosis do not play an important role. In this case, as shown in figure 2.10, both GMEQS and GMV show similar performance (slightly better than GMES).

We also compute the diversification and turnover of each of the strategies are shown in figures 2.11, 2.12, 2.13 and 2.14. The average diversification and turnover of each of the strategies are shown in table 2.1. To evaluate the performance of these strategies we compute different return to risk ratios. From table 2.2 we see that the sharpe ratio of GMEQS is slightly higher than that of GMV and GMES. As we move from lower tail probability to higher tail probability these return to risk ratios show a declining trend which is expected.

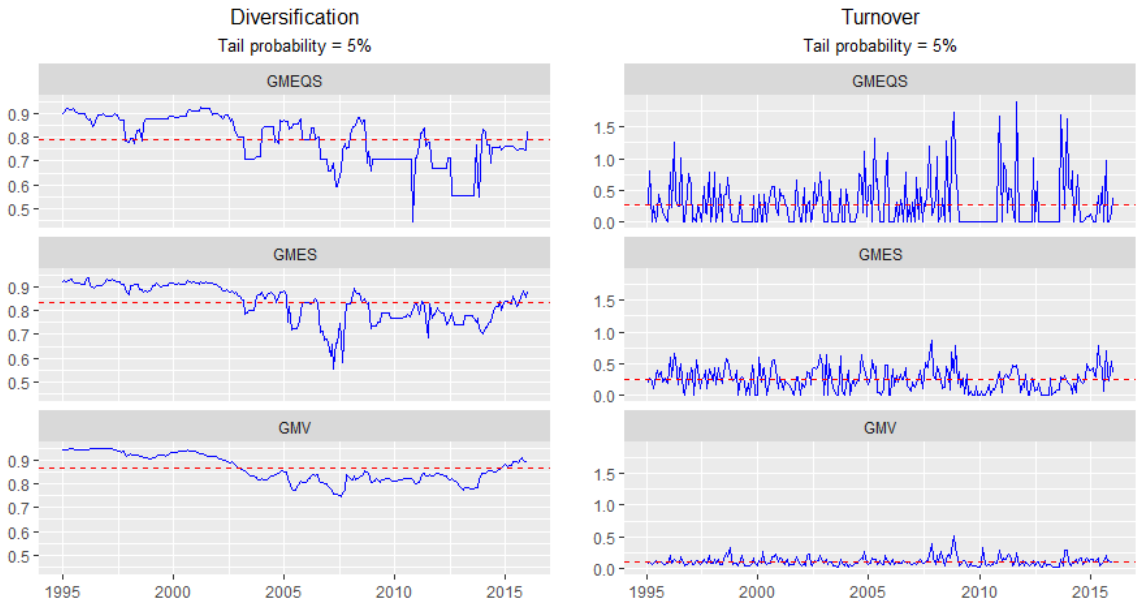

Figure 2.11: Diversification and Turnover: without TOC (tail prob. 5%)

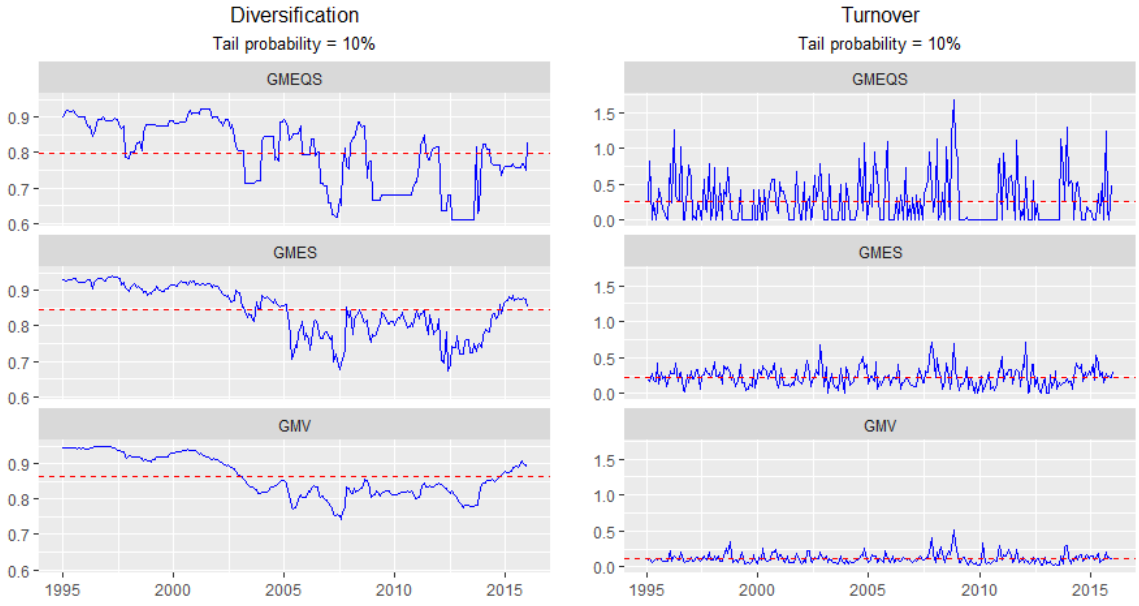

Figure 2.12: Diversification and Turnover: without TOC (tail prob. 10%)

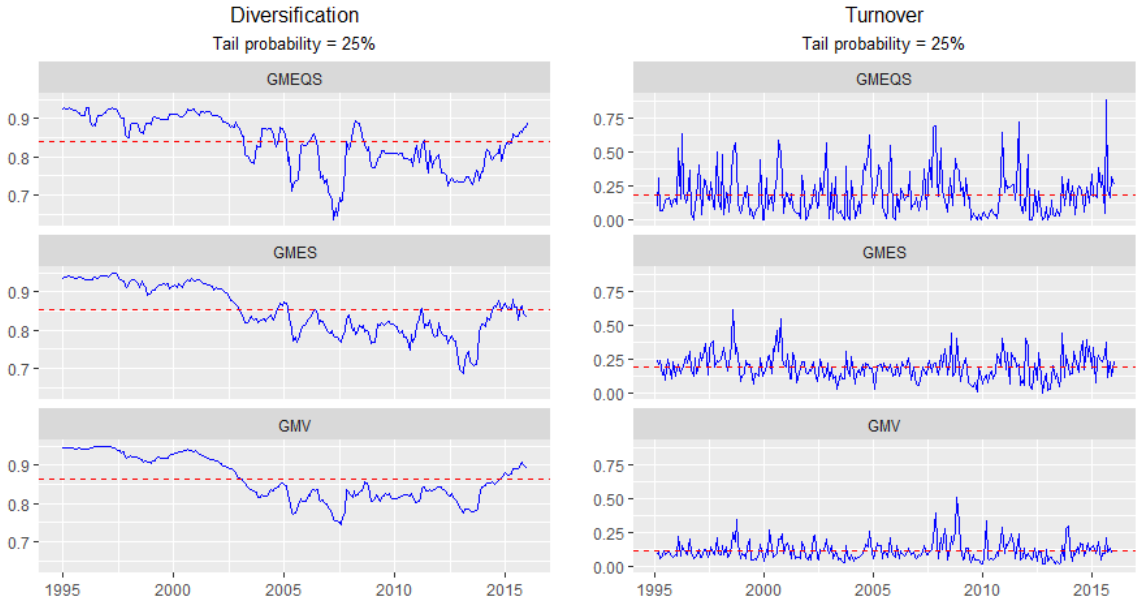

Figure 2.13: Diversification and Turnover: without TOC (tail prob. 25%)

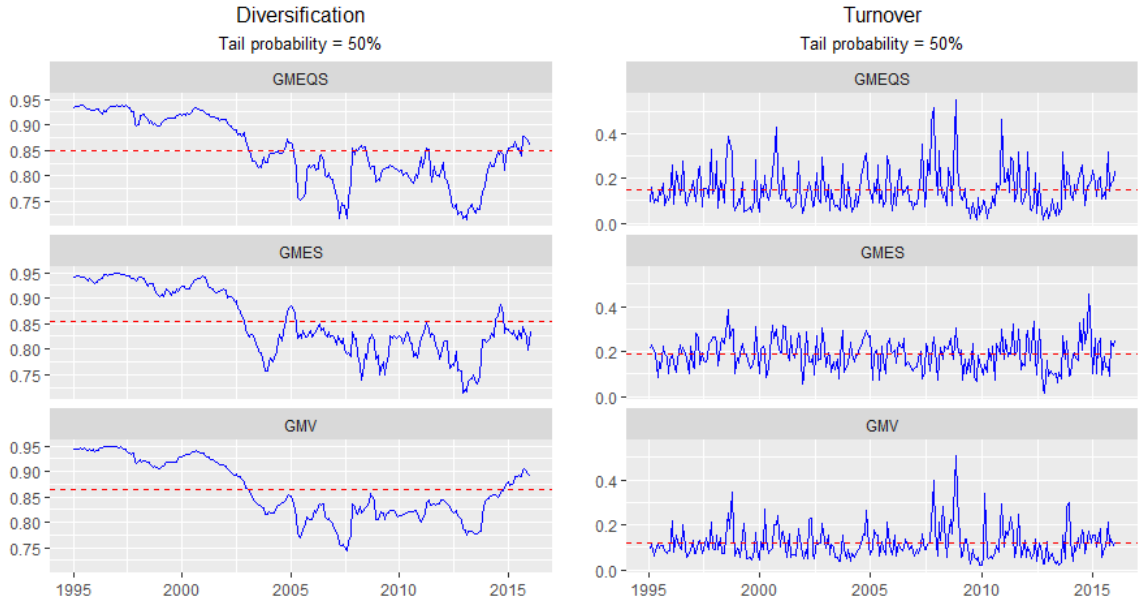

Figure 2.14: Diversification and Turnover: without TOC (tail prob. 50%)

	Tail probability = 5%		Tail probability = 10%		Tail probability = 25%		Tail probability = 50%	
	Diversification	Turnover	Diversification	Turnover	Diversification	Turnover	Diversification	Turnover
GMV	0.8636	0.1167	0.8636	0.1167	0.8636	0.1167	0.8636	0.1167
GMES	0.8349	0.2581	0.8446	0.2299	0.8517	0.1960	0.8533	0.1885
GMEQS	0.7912	0.2738	0.7977	0.2590	0.8387	0.1898	0.8506	0.1505

Table 2.1: Diversification and Turnover: without TOC

	Tail probability = 5%	Tail probability = 10%	Tail probability = 25%	Tail probability = 50%
GMV	0.0350	0.0350	0.0350	0.0350
GMES	0.0385	0.0359	0.0299	0.0316
GMEQS	0.0385	0.0393	0.0384	0.0351

Table 2.2: Sharpe Ratio: without TOC

	Tail probability = 5%	Tail probability = 10%	Tail probability = 25%	Tail probability = 50%
GMV	0.0345	0.0345	0.0345	0.0345
GMES	0.0382	0.0356	0.0294	0.0309
GMEQS	0.0389	0.0395	0.0382	0.0348

Table 2.3: Downside Sharpe Ratio: without TOC

We also compute the downside sharpe ratio and ES ratio for these portfolios. We see from table 2.3 that the downside sharpe ratio of GMEQS portfolio is slightly higher than that of GMV and GMES portfolios. As we move from lower to higher tail probability both GMEQS and GMES generally show a declining trend (except for GMEQS it slightly goes up at 10% tail probability) in downside sharpe ratio. From table 2.4 too we see that the GMEQS portfolio has slightly higher ES ratio than GMV and GMES portfolios. And similar to sharpe ratio and downside sharpe ratio, ES ratio too generally declines as we move away from the tail of the distribution.

	Tail probability = 5%	Tail probability = 10%	Tail probability = 25%	Tail probability = 50%
GMV	0.0195	0.0195	0.0195	0.0195
GMES	0.0216	0.0202	0.0166	0.0175
GMEQS	0.0221	0.0225	0.0215	0.0197

Table 2.4: ES Ratio: without TOC

2.5.2 With Turnover Constraints

Here we show the results of portfolio optimization with turnover constraint $toc = 0.4$. The figures 2.15, 2.16, 2.17 and 2.18 show the performance of the portfolios formed by GMV, GMES and GMES strategies for 5%, 10%, 25% and 50% tail probabilities. Figure 2.15 shows that the cumulative return of the portfolio formed by the GMEQS strategy is higher than that of two other strategies. This is also evident from plot in the lower panel where we see GMEQS shows the best relative performance. Among the three strategies GMV performs the worst indicating that we need to pay attention to the non-normal features of the asset

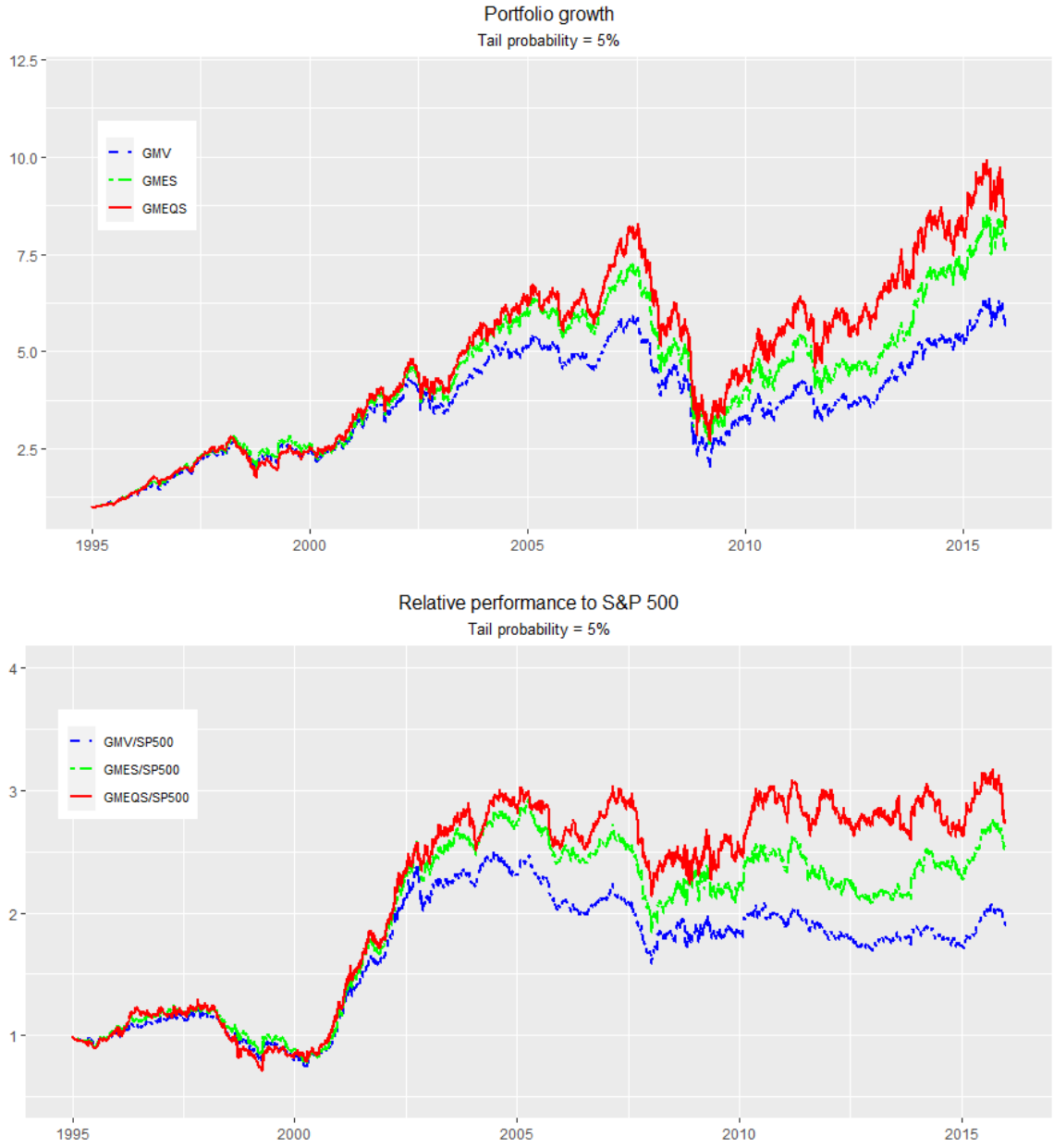

Figure 2.15: Cumulative Return of Portfolio: with TOC (tail prob. 5%)

returns data and focus on tail based risk measures while optimizing portfolios. For tail probability 10% also GMEQS beats GMV and GMES in terms of cumulative return and relative performance as evident in figure 2.16. Interestingly, GMES and GMV do not show

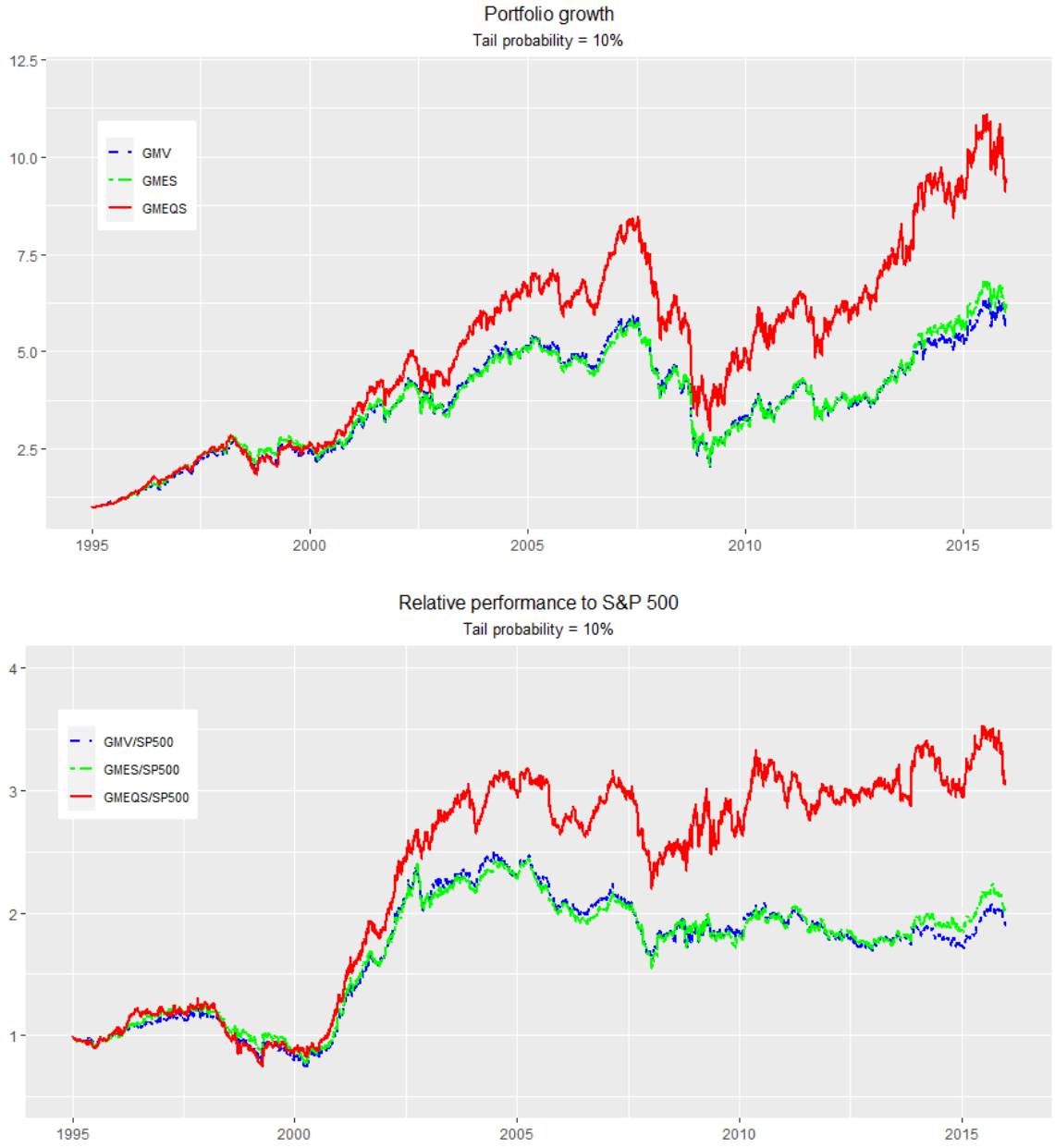

Figure 2.16: Cumulative Return of Portfolio: with TOC (tail prob. 10%)

much difference in performance in this case. When tail probability is 25%, figure 2.17 shows that GMEQS is still the best performer of all three strategies. But GMV now shows a better performance than GMES.

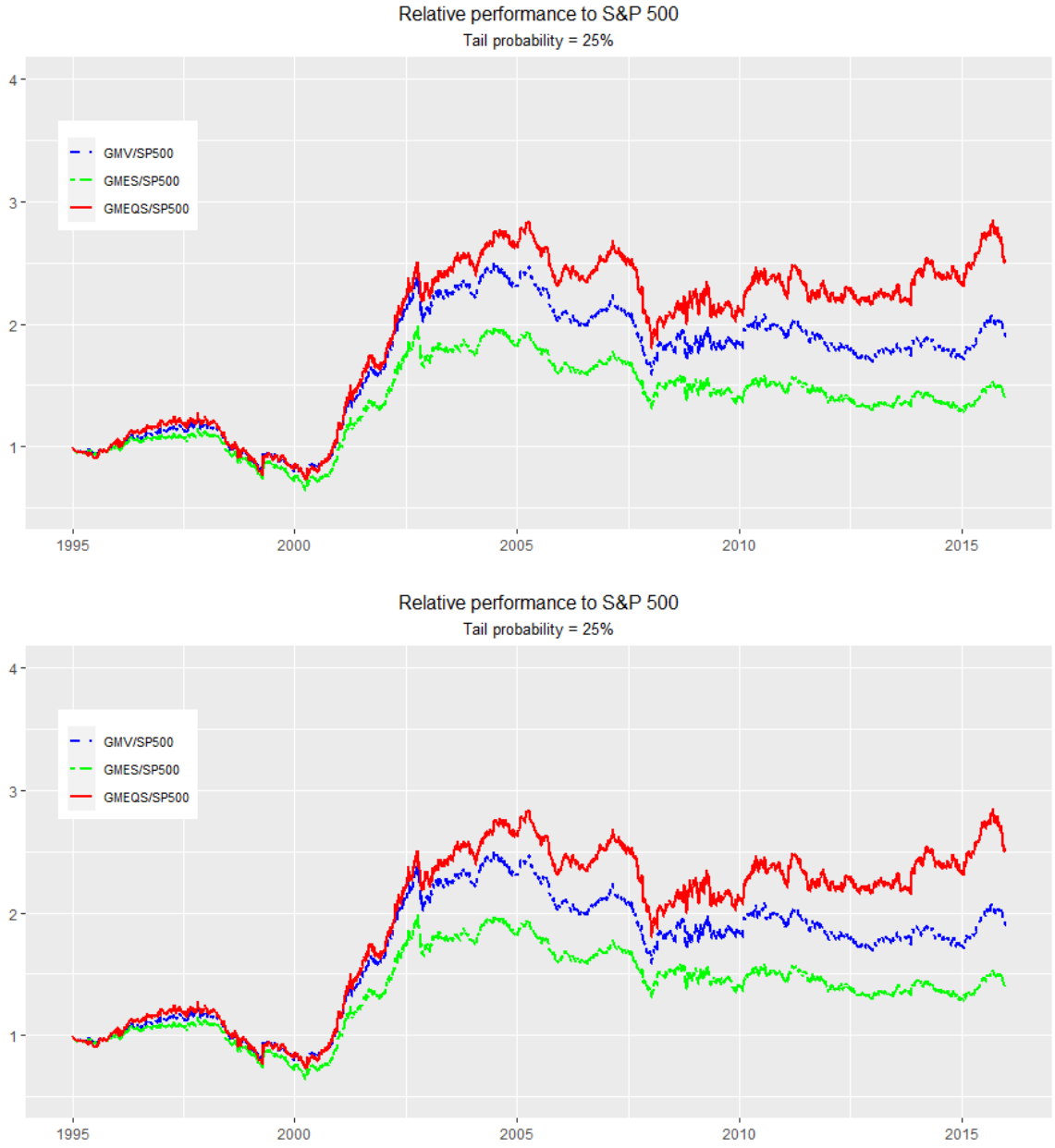

Figure 2.17: Cumulative Return of Portfolio: with TOC (tail prob. 25%)

Figure 2.18: Cumulative Return of Portfolio: with TOC (tail prob. 50%)

For tail probability 50%, as shown in figure 2.18, both GMEQS and GMV show similar performance (slightly better than GMES).

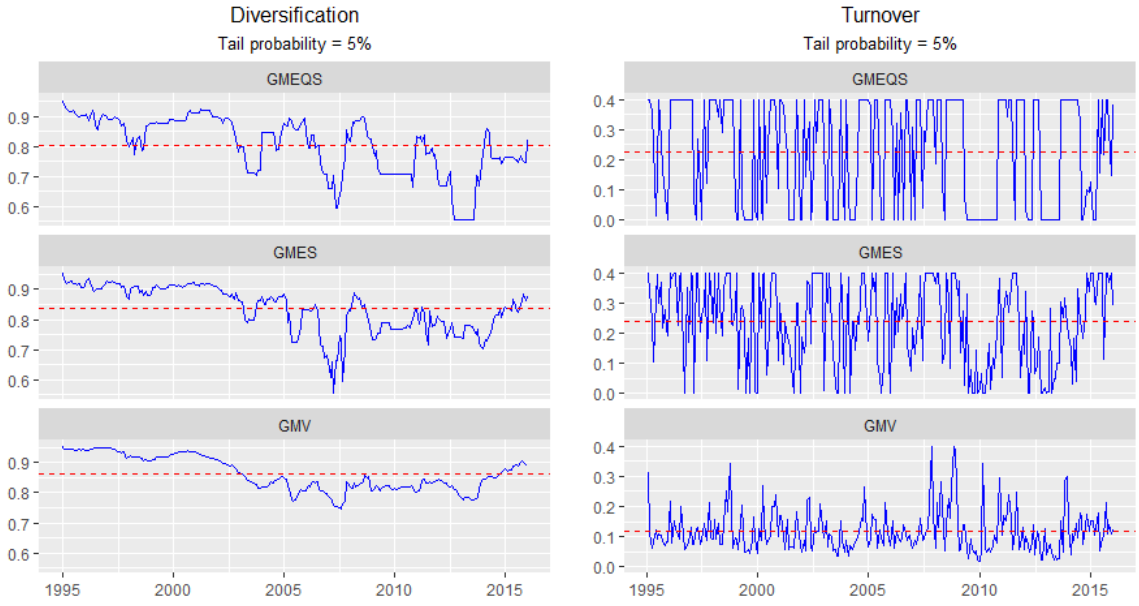

Figure 2.19: Diversification and Turnover: with TOC

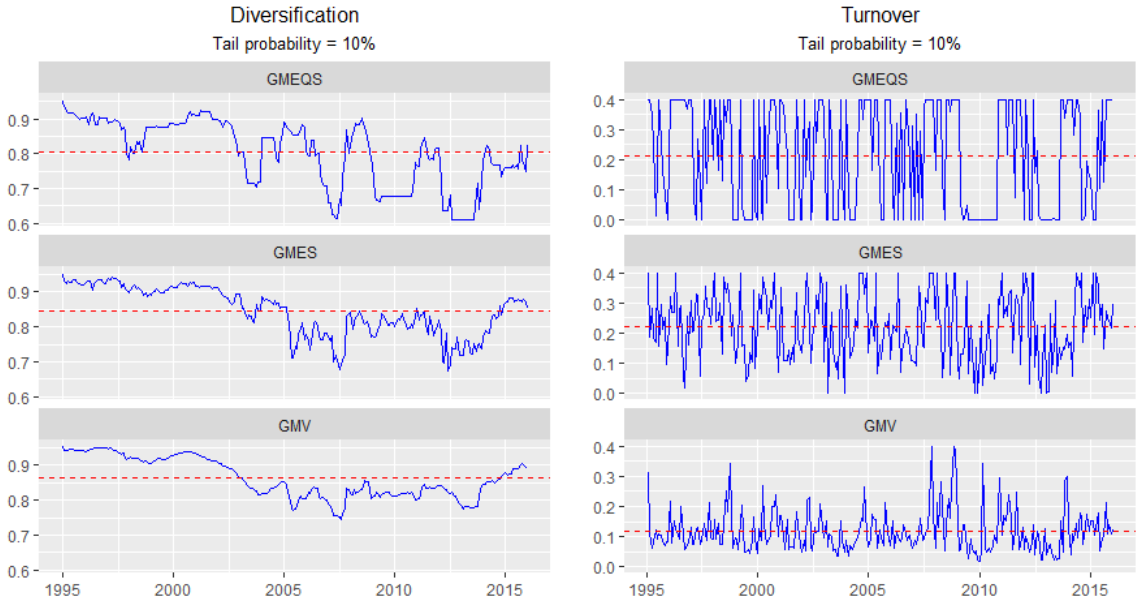

Figure 2.20: Diversification and Turnover: with TOC

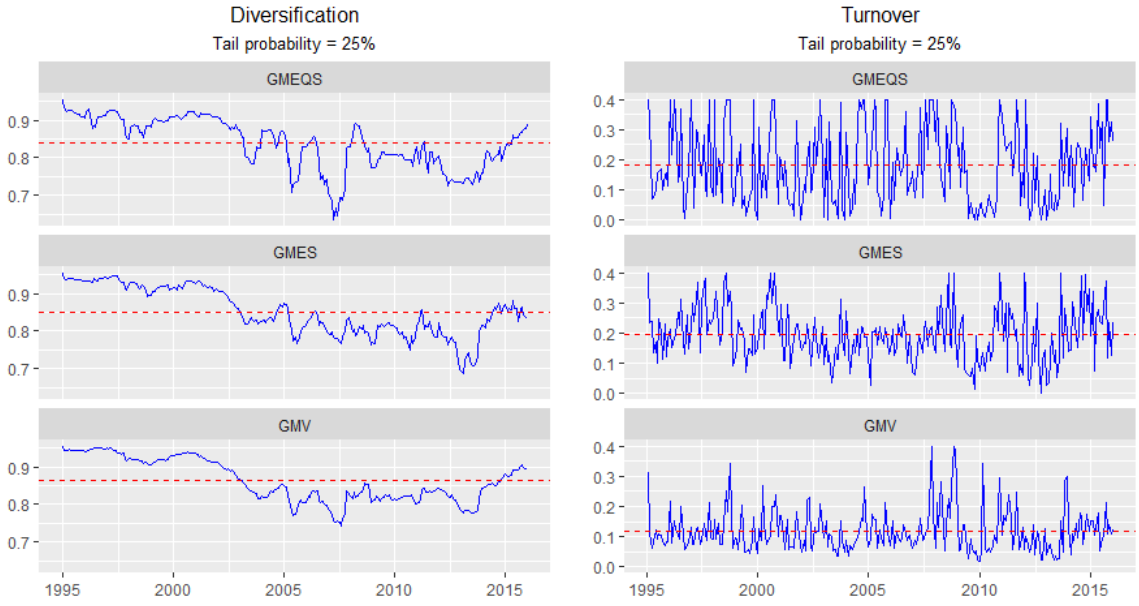

Figure 2.21: Diversification and Turnover: with TOC

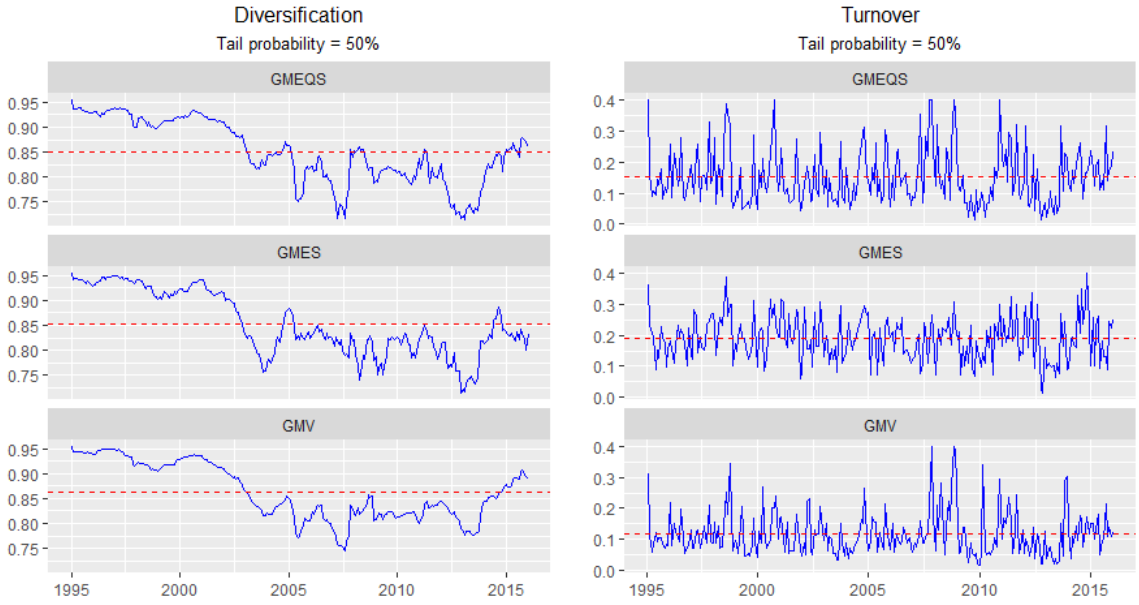

Figure 2.22: Diversification and Turnover: with TOC

	Tail probability = 5%		Tail probability = 10%		Tail probability = 25%		Tail probability = 50%	
	Diversification	Turnover	Diversification	Turnover	Diversification	Turnover	Diversification	Turnover
GMV	0.8636	0.1174	0.8636	0.1174	0.8636	0.1174	0.8636	0.1174
GMES	0.8358	0.2386	0.8454	0.2230	0.8518	0.1949	0.8534	0.1888
GMEQS	0.8032	0.2284	0.8031	0.2154	0.8388	0.1813	0.8507	0.1506

Table 2.5: Diversification and Turnover: with TOC

	Tail probability = 5%	Tail probability = 10%	Tail probability = 25%	Tail probability = 50%
GMV	0.0348	0.0348	0.0348	0.0348
GMES	0.0390	0.0355	0.0297	0.0315
GMEQS	0.0379	0.0402	0.0389	0.0352

Table 2.6: Sharpe Ratio: with TOC

The diversification and turnover of each of the strategies are shown in figures 2.19, 2.20, 2.21 and 2.22. The average diversification and turnover of each of the strategies are shown in table 2.5. From table 2.6 we see that the sharpe ratio of GMEQS is slightly higher than that of GMV and GMES.

	Tail probability = 5%	Tail probability = 10%	Tail probability = 25%	Tail probability = 50%
GMV	0.0344	0.0344	0.0344	0.0344
GMES	0.0386	0.0352	0.0292	0.0308
GMEQS	0.0374	0.0402	0.0385	0.0348

Table 2.7: Downside Sharpe Ratio: with TOC

	Tail probability = 5%	Tail probability = 10%	Tail probability = 25%	Tail probability = 50%
GMV	0.0195	0.0195	0.0195	0.0195
GMES	0.0219	0.0200	0.0165	0.0174
GMEQS	0.0215	0.0229	0.0218	0.0197

Table 2.8: ES Ratio: with TOC

We also compute the downside sharpe ratio and ES ratio for these portfolios. We see from table 2.7 that the downside sharpe ratio of GMEQS portfolio is slightly higher than that of GMV and GMES portfolios. As we move from lower to higher tail probability both GMEQS and GMES generally show a declining trend (except for GMEQS it slightly goes up at 10% tail probability) in downside sharpe ratio. From table 2.8 too we see that the GMEQS portfolio has slightly higher ES ratio than GMV and GMES portfolios. And similar to sharpe ratio and downside sharpe ratio, ES ratio too generally declines as we move away from the tail of the distribution.

2.6 Conclusion

In this paper we focus on downside risk estimates in order to construct an optimal portfolio. Based on a sample of small cap CRSP daily stock returns data, we construct optimal portfolio by minimizing different risk measures such as variance, Expected Shortfall (ES) and Expected Quadratic Shortfall. The EQS has similar properties as ES but it measures risk in terms of the second moments of loss distributions. Following the algorithm proposed by Krokhmal [2007], we have constructed an optimal portfolio by minimizing the EQS after formulating it as a Second Order Cone Program (SOCP). For implementation We used the conic solver functionality of the R package *CVXR* developed by Fu et al. [2020]. We form optimal portfolios based on GMV, GMES and GMEQS strategies at the end of each month using the rolling window method and compute cumulative portfolio return and relative return comparing against the benchmark S&P 500. We also look at the return to risk ratios such as sharpe ratio, downside sharpe ratio and ES ratio. Based on cumulative return and return to risk ratios, GMEQS shows better performance over GMV and GMES. This is slightly more pronounced when the tail probability is low. These results seem quite promising for the EQS as an objective measure for constructing optimal portfolios. This calls for further exploration of this area and it should be tested on larger datasets. We plan do more exhaustive study in future.

Chapter 3

PREDICTING EXCHANGE RATES WITH MACHINE LEARNING: EXPECTATIONS, NONLINEARITY, AND PARAMETER INSTABILITY

3.1 Introduction

Explaining exchange rate movements at the short-to medium-horizons (monthly to 2-years) has long been a challenge, giving rise to decades of literature exploring various empirical exchange rate puzzles and disconnects (see, for example, Handbook of International Economics chapters Frankel and Rose [1995] and more recently by Engel [2014]). While the literature has identified various theoretical or structural channels of how macroeconomic forces, market expectations, and investor sentiments contribute to currency movements, their empirical support remain elusive.

This paper aims to re-examine the long-standing empirical disconnect between the exchange rate and its theoretical macroeconomic determinants (or "fundamentals") by exploring a large set of monthly data that capture both current macroeconomic conditions (à la the factor analysis literature of Ludvigson and Ng [2009]; as well as market expectations and perceived uncertainties about them, as embodied in the term structures of relevant asset prices (e.g. Chen and Tsang [2013], Chen et al. [2018]). Instead of the restricting our empirical testing to the standard linear analyses, we employ a variety of linear and non-linear machine learning techniques, to examine their predictive power for subsequent exchange rate movements, both in in-sample regressions and in pseudo-out-of-sample forecasts. Using these more flexible forecasting techniques, our empirical design aims to shed further light on the sources of previous failures as well as evaluate explanations put forth in the literature, in-

cluding parameter instability (e.g. Rossi [2013]) , non-linear dynamics (e.g. Gourinchas and Tornell [2002]), and other methodological issues emphasized in the previous literature (e.g. Engel and West [2005]).

Machine learning methods have been getting increasing attention from the empirical asset pricing literature for the last couple of years. So far the research has been focused on finding the value of machine learning in the context of prediction. Gu et al. [2020] is an exhaustive study exploring the suitability of machine learning methodologies in the context of financial asset return predictability and risk premium measurement. They showed non-linear supervised machine learning technique to be effective in forecasting the expected stock excess returns and measure the equity risk premium. Bianchi et al. [2021] in their study explored the effectiveness of machine learning methods in bond return predictability and found that non-linear methods are useful in out-of-sample prediction of excess bond returns.

Our specific motivation behind this exercise is to provide a closer examination of the elusive to non-existent out-of-sample success over the Random Walk (RW). We used four categories of data combinations: (i) macroeconomic data only, (ii) macroeconomic variables with yield curve data, (iii) macroeconomic, yield curve and option data, and (iv) only forward looking term structure variables, i.e. yield curve and option data.

The results of the linear models confirm the long existing verdict that random walk is hard to beat when it comes to prediction of exchange rate. Ordinary Least Squares fails to produce meaningful forecasts precisely because there are too many features than observations. For the same reason, least squares methods with parameter regularization have a reasonable performance because they are able to force coefficient estimates into a small range, or select a small set of features by forcing other coefficients to zero. However the results are reasonable relative to the OLS results, but still do not show improvements over RW. In most cases the R^2's from regression are slightly negative, with some exceptions. The small absolute values

also indicate that a big proportion of their predictions coincides with those of random walk. As we are using a large number of predictors, there can be high correlation among many variables thus making the prediction exercise less effective. In order to mitigate that effect, we orthogonalize the data and extract principal components from each different groups and re-run all the models. Comparing across different data combinations, there is no clear evidence that one specification is better than another.

On the other hand, using the simplest neural network specification - multilayer perceptrons, we are able to produce statistically significant improvements over random walk at 13% – 14% in the pseudo out-of-sample exercise. Although the improvement is not over all specifications, We take these findings as indicative that the exchange rate is not disconnected to indicators of the macroeconomy - be their current values or expectations, though their functional relation may be more nuanced than simple linear specifications can capture. Further exploration involves using random forest, more of a classification method than a regression method, we find orthogonalized data generally produces the best out-of-sample performance, which hint at the information extraction process that we will explore further. Our results with machine learning methods, albeit positive in some aspects, are not as outstanding as those in Bianchi et al. [2021], who look at bond excess returns. A future extension of the paper include trying to identify the differences in the underlying data generating process for exchange rates and that of bond returns.

This paper is organized as follows: next section reviews the relevant literature. Section 3 introduces our data and methodology. Section 4 reviews our results, followed by a discussion about instability in section 5. We conclude in section 6.

3.2 Review of Literature

Machine learning methods have been getting increasing attention in empirical asset pricing literature. This is mainly due to the greater access of high computing power enabling us

to implement more advanced techniques in statistics and econometrics to solve problems in economics and finance. But predominantly machine learning methods have been used in predictive analysis. Gu et al. [2020] pointed out that machine learning methods are suitable for prediction i.e. in the context of our problem they approximate the conditional expectation $E(\Delta s_{t+1}|I_t)$ where Δs_{t+1} is the exchange rate return and I_t is the unobserved information set at time t. Earlier Mullainathan and Spiess [2017] also made the same observation that machine learning methods are not helpful for structural analysis and in problems of inference. So it is difficult to understand the underlying economic process purely on the basis of machine learning methods. Nevertheless, they can play a supportive role in understanding that economic process.

In this paper we look at number of modeling techniques - simple linear regression i.e. ordinary least squares (OLS), linear regression with regularization (Lasso and Ridge), dimension reduction methods such as principal components regression (PCR), regression tree methods such as random forest (RF), and neural network. Early works in financial economics which used principal components or factors in forecasting problems are Stock and Watson [2002a], Stock and Watson [2002b], Stock and Watson [2006], Bai and Ng [2002], Bai and Ng [2006], Bai and Ng [2008], Boivin and Ng [2006], Forni and Reichlin [1998], Rorni and Reichlin [1996], Some notable works which attempt at combining machine learning and equilibrium asset pricing are Kelly and Pruitt [2013], Feng et al. [2017], Kelly et al. [2017]. Early works which used neural networks in economics and finance are Kuan and White [1994], Lo [1994], Hutchinson et al. [1994], Yao et al. [2000]. Regression trees have been used to predict credit card risk by Khandani et al. [2010] and Butaru et al. [2016]. Sirignano et al. [2016] used deep neural network to analyze the mortgage risk while Messmer [2017] used deep learning to study cross-section of expected returns. Heaton et al. [2017] used deep neural network in the context of smart indexing in finance to automate portfolio selection. Dimension reduction and regularization methods have been successfully used by Kelly and Pruitt [2015], Kozak et al. [2020], Freyberger et al. [2020]. Gu et al. [2020] conducted a comparative analysis

of machine learning methods to predict the expected stock excess return and showed that non-linear methods such as regression trees and neural network were able to produce better out-of-sample forecasts and more accurate equity risk premium measurement. Bianchi et al. [2021] used machine learning methods to predict bond excess return and found that deep neural network outperforms dimension reduction methods principal component regression and partial least squares in forecasting out-of-sample bond excess return thus pointing to the fact that neural networks can capture the complex non-linearities of the data.

3.3 Data and Methodology

We pick four currency pairs: USD per AUD, USD per CAD, USD per GBP, and JPY per USD. Our sample starts from January 1995. We obtain the daily spot exchange rate from Global Financial Data, and convert them into monthly data by using the end-of-month values. We use the FRED-MD data set as in Ludvigson and Ng [2009] and McCracken and Ng [2016], which includes 135 variables about the U.S. economy spanning eight categories: output and income, labor market, housing, consumption and orders, money and credit, interest rate, prices, and stock market. To avoid overlapping with the yield curve data and with our dependent variable, we deleted all related variables from the data set. As for foreign macroeconomic variables, for easy reproducibility, we manually composed the data set for each country by selecting the foreign equivalent variables as in FRED-MD that are available in the FRED database. We then transform each series in the same way as in the FRED-MD appendix. Fortunately having four advanced foreign economies means we are able to find at least a handful of equivalent variables for each category. In the end, after making the trade-off between data lengths and variable counts, we are able to have more than 37 variables for each foreign country, and monthly data available through at least mid-2012.

Our daily yield curve data from Bloomberg contains yields for 13 maturities: 3m, 6m, 12m, 24m, 36m, 72m, 84m, 96m, 108m, 120m and 15y. The option price data consists of daily over-the-counter option prices for the four currency pairs from JP Morgan from December

1999 to May 2012. We use at-the-money, risk reversal, and strangle prices with maturities 7d, 1m, 2m, 3m, 6m, 9m, and 12m in our estimation. All daily series are converted to monthly by selecting their end-of-month values. In the end, our data coverage is limited by foreign macroeconomic data and option price data. The results we show in Section 6 are run on the longest coverage for each data combinations.

The aim of this exercise is to disentangle the exchange rate disconnect puzzle. We are particularly interested to find out a) whether the data do not have any useful information to predict exchange rate (i.e. full exchange rate disconnect), b) whether the relationship between exchange rate return and the predictors is linear or non-linear, c) if the selection mechanism is unstable, d) whether the collinearity of the data is playing any role. Given our systematic approach, we are also able to compare how these machine learning methodologies performs with exchange rate data. In this exercise we have macroeconomic variables, yield curve data and FX options data as the set of predictors to explain the variation in FX return data.

We use both raw data and extract principal components from those raw data to use them separately in our analysis. First, we do an in-sample analysis taking the exchange rate return as the dependent variable and the set of predictors belonging to different categories such as macroeconomic data, yield curve data and FX options data. Ordinary least squares as well as the machine learning methods such as Lasso, Ridge, Partial Least Squares, Random Forest, Neural Network are used. We discuss these methods briefly in the next section. Assuming there are l macroeconomic data, m yield curve data and n FX option data, for regression based analysis we consider the following equation when using raw data

$$\Delta s_{t+1} = \beta_0 + \sum_{i=1}^{l} \boldsymbol{\beta}_{macro,i} \boldsymbol{X}_{macro,it} + \sum_{j=1}^{m} \boldsymbol{\beta}_{yield,j} \boldsymbol{X}_{yield,jt} + \sum_{k=1}^{n} \boldsymbol{\beta}_{option,k} \boldsymbol{X}_{option,kt} \qquad (3.1)$$

where Δs_{t+1} is the log difference of month-end exchange rates; $\boldsymbol{X}_{macro,it}$ is the i^{th} macroeconomic data; $\boldsymbol{X}_{yield,jt}$ is the j^{th} yield curve data; $\boldsymbol{X}_{option,kt}$ is the k^{th} option data; and β's are the coefficients to be estimated.

We then conduct our analysis using the orthogonalized data. We orthogonalize the raw data by extracting the principal components separately from all three categories such as macroeconomic (l factors), yield curve (m factors) and option data (n factors). For regression based analysis we consider the following equation when using the orthogonalized data

$$\Delta s_{t+1} = \beta_0 + \sum_{i=1}^{l} \boldsymbol{\beta}_{macro,i} \boldsymbol{F}_{macro,it} + \sum_{j=1}^{m} \boldsymbol{\beta}_{yield,j} \boldsymbol{F}_{yield,jt} + \sum_{k=1}^{n} \boldsymbol{\beta}_{option,k} \boldsymbol{F}_{option,kt} \tag{3.2}$$

where Δs_{t+1} is the log difference of month-end exchange rates; $\boldsymbol{F}_{macro,it}$ is the i^{th} macroeconomic factor (i.e. principal component); $\boldsymbol{F}_{yield,jt}$ is the j^{th} yield curve factor; and $\boldsymbol{F}_{option,kt}$ is the k^{th} option factor. We also do our analysis selecting only the top factors which summarize 95% of the variation in the corresponding dataset. These will be detailed in the results section.

After in-sample analysis we proceed to produce the out-of-sample forecasts by using a) expanding window and b) rolling window procedure. We also compute the mean squared error (MSE) of the forecasts. To evaluate these forecasts we compare them with the forecasts produced by our benchmark *random walk* model and use the metric out-of-sample R^2 (as suggested in Campbell and Thompson [2008], Bianchi et al. [2021]) defined as

$$R^2_{OOS} = 1 - \frac{MSE_m}{MSE_{rw}} = 1 - \frac{MSE_m}{Var(\Delta s_{t+1})} \tag{3.3}$$

where MSE_m is the mean squared error of model m and MSE_{rw} is the mean squared error of benchmark *random walk* model. In rolling window approach parameters are estimated

over a rolling window of a fixed size through the sample. The data is initially split into an estimation sample and an out of sample. Let, the full sample be $t = 1, t = 2, \ldots t = T, t = T + 1, \ldots, t = T + n$. The initial estimation estimation sample is from $t = 1$ to $t = T$ and out of sample is from $t = T + 1$ to $t = T + n$. The parameters are estimated from the sample $t = 1$ to $t = T$ and a forecast is produced for $t = T + 1$ i.e. $\Delta\hat{s}_{t+1}$ is computed. In the next iteration, we drop one observation from the beginning and add one observation at the end i.e. the estimation sample is from $t = 2$ to $t = T + 1$. After parameters are estimated based on this sample, we produce the forecast for $t = T + 2$ i.e. $\Delta\hat{s}_{t+2}$ is computed. We continue in this fashion until we produce forecasts for all the data points in the out of sample i.e. from $t = T + 1$ to $t = T + n$. Once forecasts are computed, we proceed to evaluate those forecasts. In expanding window approach parameters are estimated over an increasing window through the sample. The data is initially split into an estimation sample and an out of sample. Let, the full sample be $t = 1, t = 2, \ldots t = T, t = T+1, \ldots, t = T+n$. The initial estimation sample is from $t = 1$ to $t = T$ and out of sample is from $t = T+1$ to $t = T+n$. The initial sample using data from $t = 1$ to $t = T$ is used to estimate the model, and 1-step ahead out-of-sample forecast is is produced for $t = T + 1$ i.e. $\Delta\hat{s}_{t+1}$ is computed. Then the sample is increased by one and the model is re-estimated based on sample from $t = 1, t = 2, \ldots t = T, t = T + 1$ to produce a forecast for $t = T + 2$ i.e. $\Delta\hat{s}_{t+2}$ is computed. This is continued until we produce forecasts for all the data points in the out of sample i.e. from $t = T + 1$ to $t = T + n$. Once forecasts are computed, we proceed to evaluate those forecasts.

To evaluate the forecast accuracy we employ Diebold and Mariano [1995] and Clark and West [2007] tests. We have two competing forecasting models - our machine learning model and the benchmark RW model. The forecast errors are computed as $e_{it} = y_{it} - y_t \quad i = 1, 2$. The loss associated with forecast i is given by $g(e_{it}) = e_{it}^2$ i.e. the loss is a function of the squared error. In the Diebold and Mariano [1995] test, the loss differential between two forecasts are computed as $d_t = g(e_{1t}) - g(e_{2t})$. Now the two forecasts will have equal accuracy if and only if the loss differential on an average will be zero. So the null hypothesis of the test is given

as the following

$$H_0 : E(d_t) = 0 \quad \forall t \quad \text{vs.} \quad H_1 : E(d_t) \neq 0$$

Under H_0, the test statistic is asymptotically $N(0; 1)$ distributed. But Diebold and Mariano [1995] is not suitable when competing forecasts are obtained from two nested models as is the case for us. To address that shortcoming Clark and West [2007] proposed a test which is appropriate for comparing a parsimonious null model to a larger model that nests the null model. According to this test, under the null the parsimonious model generates the data and the larger model introduces noise into its forecasts by estimating parameters whose population values are zero. Therefore parsimonious model will produce a smaller mean squared prediction error than that of the larger model. To account for this noise, Clark and West [2007] introduced a special adjustment to the mean squared prediction error. Then the usual t-statistics are computed to test whether the adjusted difference in mean squared errors is zero or not. For conducting Diebold-Mariano test, we used the function *dm.test* in *R* package *forecast*. For details see Hyndman and Khandakar [2008] and Hyndman et al. [2021]. For conducting Clark-West test we have used the *R* code provided by Gray Calhoun[1].

We also explored whether combining multiple forecasts produced by different subset of regressors can give better forecasts. This is motivated by *Complete Subset Regressions* as proposed by Elliott et al. [2013]. We briefly describe this method following the discussion in Elliott and Timmermann [2013]. In this method all possible combinations of models including a fixed number of regressors k out of K regressors are considered for producing forecasts and then an equal weighted combination of those forecasts is computed as the final forecast. The dependent variable y_t is regressed on each subset of regressors and then all

[1]Calhoun, G. 2011, An asymptotically normal out-of-sample test of equal predictive accuracy for nested models. Unpublished manuscript.

Calhoun, G. 2011, Documentation appendix: An asymptotically normal out-of-sample test of equal predictive accuracy for nested models. Unpublished manuscript.
https://github.com/grayclhn/oosanalysis-R-library/blob/master/man/clarkwest.Rd

the forecasts are combined together by taking a simple average. There will be a total of $n_{k,K} = K!/(k!(K-k)!)$ models to consider. If we set $k = 1$ i.e. we take only one regressor at a time, there will be a total of $n_{1,K} = K$ models. The equal weighted combination of forecasts from these K models is given by

$$f^c = \frac{1}{K}\sum_{i=1}^{K} x_i'\hat{\beta}_i \tag{3.4}$$

In general combination of forecasts or ensemble forecasts are likely to perform better as we know individual models are likely to be misspecified, see Elliott and Timmermann [2013] and the book *Economic Forecasting* by Elliott and Timmermann (2016) for details. In this exercise we used the *Complete Subset Regressions* as proposed by Elliott et al. [2013] setting $k = 1$. From the set of all regressors (macroeconomic data, yield curve data and FX options data), we take one variable at a time and regress the FX return on that regressor. After producing the forecasts from all the models, we take a simple average of all forecasts to generate the final forecast of the outcome. Next, we compute the out-of-sample R^2 comparing this forecast against our benchmark model of random walk.

As pointed out in Gu et al. [2020] machine learning methods have a number of advantages - a) they can be used in high dimensional modeling in the prediction context, b) they help in model selection and reduce overfit bias via regularization, c) they are computationally efficient in searching among a large number of potential specifications. All these aid us in approximating an unknown data generating process in the problem context. In our case we have a large number of predictors (macroeconomic, yield curve and option data) and the frequency of our data is monthly. We know the conventional least squares method is not the best candidate for estimation in such a scenario as number of predictors is very close to the number of sample observations. Predictors belonging to different groups can also be highly correlated. So multicollinearity can be an issue. We are also agnostic about the functional form through which dependent variable and the predictors are related i.e. we do

not know whether the association linear or non-linear. Machine learning helps us in tackling these challenges. We have pointed out earlier (and also discussed in Gu et al. [2020]) that machine learning methods are suitable for prediction i.e. they approximate the conditional expectation $E(\Delta s_{t+1}|I_t)$ where Δs_{t+1} is the exchange rate return and I_t is the unobserved information set at time t. But they fall short in the context of a structural problem and in problems of inference. So it is difficult to understand the underlying economic process purely on the basis of machine learning methods. Nevertheless, they can play a supportive role in understanding that economic process.

In the appendix we briefly discuss the machine learning methods used in the prediction of foreign exchange return in this paper. We consider simple linear regression i.e. ordinary least squares (OLS), linear regression with regularization (Lasso and Ridge), dimension reduction methods such as principal components regression (PCR), regression tree methods such as random forest (RF), and neural network.

Here we briefly outline the methods of hyperparameter tuning in case of different machine learning methods. For Lasso we use a time series cross-validator which splits the sample into five segments. Time series cross-validator ensures that the test observations come after the train observations to main temporal ordering. We create a grid of 1000 values of the hyperparameter and for each value we estimate the model and evaluate the performance based on cross-validation mean-squared-error (MSE). We choose the optimal value of the hyperparameter as the one which gives minimum cross-validation MSE. This same strategy is following in hyperparameter tuning for both Ridge and Elastic Net regression. To fit random forest we consider a number of parameters such as number of trees in the forest, maximum depth of the tree, minimum number of samples needed to split an internal node, minimum number of samples needed in a leaf node, number of features required for an optimal split. We create a grid by combining different values of these parameters and fit the random forest to data. The optimal combination of the parameter values are chosen based

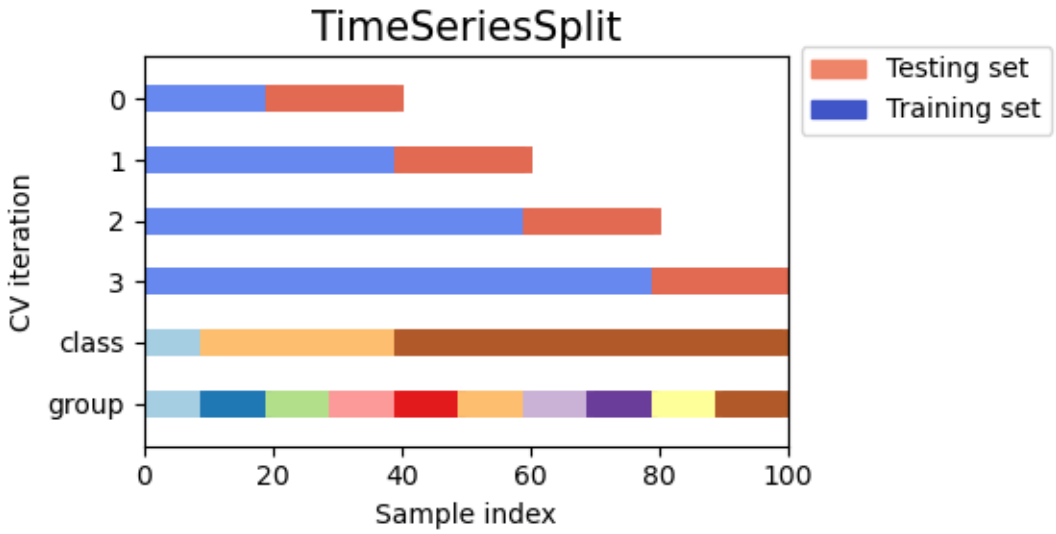

Figure 3.1: Time Series Cross Validation

Source: https://scikit-learn.org/

on the best fit. The computational cost for random forest hyperparameter tuning was much higher than that of linear regularization type models. Exhaustive search of hyperparameter values may be one of the reasons why random forest shows better fit relative to linear models. For fitting neural network in our analysis, we use simple a simple structure where two intermediate (i.e. hidden) layers contain 16 and 8 nodes and output layer consists of one node. The number of nodes in the input input layer is equal to the number of predictors to be used. To reduce computational burden and also due to paucity of data (short time series) we have not done conventional hyperparameter training for fitting neural network. For details on neutral network training the readers are referred to the cited references.

Both random forest and neural network can capture non-linear relationship. Some popular textbook references for for an in depth discussion of these methods are James et al. [2014], Efron and Hastie [2016] and Hastie et al. [2001]. Gu et al. [2020] and Bianchi et al. [2021] are two very useful references which discuss application of machine learning methodologies

in economics and finance. We follow the notation and discussion of these references. To implement the machine learning we have used standard packages of the *Python* programming language, Van Rossum and Drake [2009]. To prepare data and to conduct forecast evaluation tests we have used *R* statistical programming language, R Core Team [2021].

3.4 Results

In this section, we present and discuss out-of-sample and in-sample performance of our predictive model using R^2 values. In-sample R^2 explains the amount of variance explained in exchange rate movements explained by the model, and usually takes a value between 0 and 1. This value occasionally becomes slightly negative, indicating that the model prediction is worse than a constant prediction at the mean of our dependent variable, $\overline{\Delta s_{t+1}}$. Out-of-sample R^2 is defined as the percentage improvement in mean squared errors of our model prediction relative to a benchmark model. Following the literature, we examine whether our model can consistently outperform the random walk in out-of-sample predictions.

$$R^2_{OOS} = 1 - \frac{MSE_m}{MSE_{rw}} = 1 - \frac{MSE_m}{Var(\Delta s_{t+1})} \tag{3.5}$$

The out-of-sample R^2 takes values from $(-\infty, 1]$, where the value of 1 means zero MSE from the model and a negative number means the model does not forecast as well as the benchmark model.

The overall result suggests that Meese-Rogoff puzzle still persists - our models cannot consistently outperform the random walk in out-of-sample forecasts. However, there are individual cases where neural network delivers an sizeable improvement, and some of them are statistically significant. We view this as a positive signal of neural network's ability to extract relevant information to predict future exchange rate fluctuations. Comparing across models, we find that random forest and partial least squares, unlike neural network in some cases, do not present an advantage over their more commonly used linear counterparts, potentially

due to the lesser degree of flexibility comparing to neural network. Among linear models the performance is mediocre at the best, providing only very slightly positive OOS R^2 if not negative. Further analysis shows that penalized regressions pick up random signals from the dataset, either because they cannot efficiently extract information or because the information is too sparse. As robustness checks, we also explore using the top principal components (or "factors") and fully orthogonalized data to avoid multi-collinearity. The factor loadings of the top factors in some countries suggest strong structural shifts in the macroeconomic data set, which may help explain the difficulty in forecast using traditional methods.

In-sample results from the linear methods suggest there are relevant information between our RHS variables and exchange rate movements a month later. However in-sample results do not necessarily translates to out-of-sample performances. Especially in the case of non-linear methods, under the existence of large number of parameters, the degrees of freedom is very low, therefore making it hard to interpret the in-sample R^2s from neural network and random forest.

3.4.1 In-sample Results

The in-sample results are presented in Table 3.1, where instead of running multiple neural network, we ran only one neural network with fully connected layers. The results illustrate

Table 3.1: In-sample adjusted R^2 using raw data

		(1)	(2)	(3)	(4)
Australia	Lasso	0.0000	0.0000	-0.0043	-0.0019
	Ridge	0.0229	0.0369	0.0298	0.0199
	EL Net	0.0000	0.0000	-0.0043	-0.0019
	RF	0.7837	0.7647	0.7791	0.6253
	NN	0.0407	0.0987	0.1088	0.0462
Canada	Lasso	-0.0045	-0.0045	-0.0087	-0.0058
	Ridge	0.0156	0.0191	0.0217	-0.0015
	EL Net	-0.0045	-0.0045	-0.0087	-0.0057
	RF	0.7402	0.3990	0.4164	0.5819
	NN	0.2955	-0.1559	0.3394	0.0206
UK	Lasso	-0.0002	-0.0002	-0.0002	-0.0002
	Ridge	0.0283	0.0304	0.0704	0.0011
	EL Net	-0.0002	-0.0002	-0.0002	-0.0002
	RF	0.6699	0.7672	0.5835	0.4097
	NN	0.1866	0.0217	0.2415	0.0473
Japan	Lasso	-0.0002	-0.0002	-0.0001	0.0284
	Ridge	0.0186	0.0211	0.0283	0.0394
	EL Net	-0.0002	-0.0002	-0.0001	0.0284
	RF	0.6700	0.3343	0.7808	0.3554
	NN	0.1263	-0.1652	0.1761	0.0575

Note: The four model specifications are (1) with macroeconomic variables only, (2) with yield curve and macroeconomic variables, (3) with yield curve, option, and macroeconomic variables, (4) with yield curve and option variables.

the effect of regularization and early stopping callbacks, where the in-sample R^2 never goes above 0.70. In our earlier results without early stopping, the in-sample R^2s are all higher and can be as high as 0.98 while having a worse out-of-sample results. Intuitively, without regularization and a dedicated validation set to evaluate when to stop estimating, the only objective for the MLP is to minimize in-sample MSE to as low as possible. With hundreds of thousands parameters in a MLP, it is easy for the network to 'memorize' the outcomes, making predictions using new samples meaningless.

Table 3.2: In-sample adjusted R^2 using orthogonalized data

		(1)	(2)	(3)	(4)
Australia	Lasso	0.0109	0.0109	-0.0043	-0.0043
	Ridge	0.0597	0.0658	0.0705	0.0119
	EL Net	0.0109	0.0109	-0.0043	-0.0043
	RF	0.7061	0.5456	0.6236	0.6309
	NN	0.9858	0.9847	0.9912	0.9503
Canada	Lasso	-0.0033	-0.0033	-0.0087	-0.0087
	Ridge	0.0582	0.0661	0.0670	0.1364
	EL Net	-0.0033	-0.0033	-0.0087	-0.0087
	RF	0.5453	0.5882	0.4964	0.6332
	NN	0.9888	0.9840	0.9914	0.9533
UK	Lasso	0.0419	-0.0002	-0.0002	0.0687
	Ridge	0.1952	0.0761	0.0871	0.0703
	EL Net	0.0419	-0.0002	-0.0002	0.0685
	RF	0.5582	0.6716	0.5113	0.6367
	NN	0.9834	0.9832	0.9904	0.9461
Japan	Lasso	-0.0002	-0.0002	-0.0039	-0.0039
	Ridge	0.0577	0.0696	0.0779	0.0190
	EL Net	-0.0002	-0.0002	-0.0039	-0.0039
	RF	0.6028	0.7246	0.7469	0.4526
	NN	0.9849	0.9861	0.9891	0.9685

Note: The four model specifications are (1) with macroeconomic variables only, (2) with yield curve and macroeconomic variables, (3) with yield curve, option, and macroeconomic variables, (4) with yield curve and option variables.

Table 3.3: In-sample adjusted R^2 using factors

		(1)	(2)	(3)	(4)
Australia	Lasso	0.0330	0.0170	0.0039	-0.0043
	Ridge	0.0564	0.0639	0.0034	-0.0006
	EL Net	0.0329	0.0169	0.0039	-0.0043
	RF	0.6828	0.3239	0.5054	0.4091
	NN	0.3849	0.4653	0.7748	0.5329
Canada	Lasso	-0.0046	-0.0046	-0.0087	-0.0087
	Ridge	-0.0017	-0.0008	-0.0048	-0.0078
	EL Net	-0.0046	-0.0046	-0.0087	-0.0087
	RF	0.2469	0.4999	0.5297	0.5016
	NN	0.5348	0.6174	0.6442	0.4774
UK	Lasso	-0.0002	0.0774	-0.0002	-0.0002
	Ridge	0.0691	0.0419	0.0114	0.0046
	EL Net	-0.0002	0.0774	-0.0002	-0.0002
	RF	0.3213	0.4966	0.3910	0.3215
	NN	0.3628	0.5885	0.7639	0.4720
Japan	Lasso	-0.0002	-0.0002	-0.0039	-0.0039
	Ridge	0.0029	0.0062	0.0027	0.0014
	EL Net	-0.0002	-0.0002	-0.0039	-0.0039
	RF	0.2548	0.5310	0.5503	0.3289
	NN	0.2017	0.4552	0.5450	0.3659

Note: The four model specifications are (1) with macroeconomic variables only, (2) with yield curve and macroeconomic variables, (3) with yield curve, option, and macroeconomic variables, (4) with yield curve and option variables.

3.4.2 Out-of-sample Results

We describe the out-of-sample forecasting results starting with the neural network. In the plots, the blue color line is the actual data and the orange color line is the forecast.

Neural Network Results

We start our detailed discussion with neural network results. There is no rule-of-thumb in how to construct a MLP. So we follow the model specification and estimation procedure in Bianchi et al. [2021], where the neural network has three dense layers with 32, 16, and 8 neurons respectively, and we estimate 100 neural networks for each point in our forecast sample and use the average prediction of the best 10 neural networks as our point estimate.

Figure 3.2: Prediction of AUS exchange rate movements using orthogonalized yield curve, option, and macroeconomic data

Preliminary MLP results using raw data, top principal components, and fully orthogonalized data sets are presented in Table 3.4. When using top principal components, we extract five factors from each countries' macroeconomic data and three factors from term structure variables. The latter group are often interpreted as the level, slope, and curvature of the corresponding variables, while the first group is an arbitrary cut-off we impose for dimension reduction. In practice, due to the variations among countries and at different times, the

variance captured by the five factors vary and may dip below 80%. However, as we will discuss later in this section, the instability issue with factor extraction is more prominent than not hitting a certain threshold of variance captured. For the orthogonalized data sets, we recursively extract the most amount of PCs from the original data set, thereby producing a data set whose features are all orthogonal to each other.

Table 3.4: Rolling window out-of-sample R^2s of MLP

		(1)	(2)	(3)	(4)
Australia	Raw data	2.21%	6.00%	2.82%	2.18%
	PCs	2.15%	4.93%	5.10%	2.18%
	Orthogonalized data	2.17%	-0.49%	10.07%	2.04%
Canada	Raw data	3.49%	-3.95%	-33.25%	3.38%
	PCs	3.37%	2.48%	-7.08%	3.49%
	Orthogonalized data	3.39%	-2.32%	-0.45%	3.42%
UK	Raw data	1.08%	-13.49%	-4.87%	1.03%
	PCs	0.96%	-0.91%	14.35%	1.09%
	Orthogonalized data	1.12%	1.97%	1.80%	0.97%
Japan	Raw data	2.04%	-0.55%	13.61%	1.83%
	PCs	2.05%	1.60%	6.58%	2.10%
	Orthogonalized data	2.01%	-5.58%	-14.46%	1.68%

Note: The four model specifications are (1) with macroeconomic variables only, (2) with yield curve and macroeconomic variables, (3) with yield curve, option, and macroeconomic variables, (4) with yield curve and option variables.

In the table, we see consistent improvements over random walk prediction when using macroeconomic data (spec (1)) and term structure variables (spec (4)) separately, despite the improvement being small and not statistically significant. We also see some sizeable and significant improvement when combining macro and term structure variables. Fig 3.4 plots the best prediction performance in the table - using PCs extracted from specification (3) of UK data. As with most forecasts, it has trouble replicating the variance in the actual data, but it outperforms random walk in capturing the overall trend of exchange rate movements.

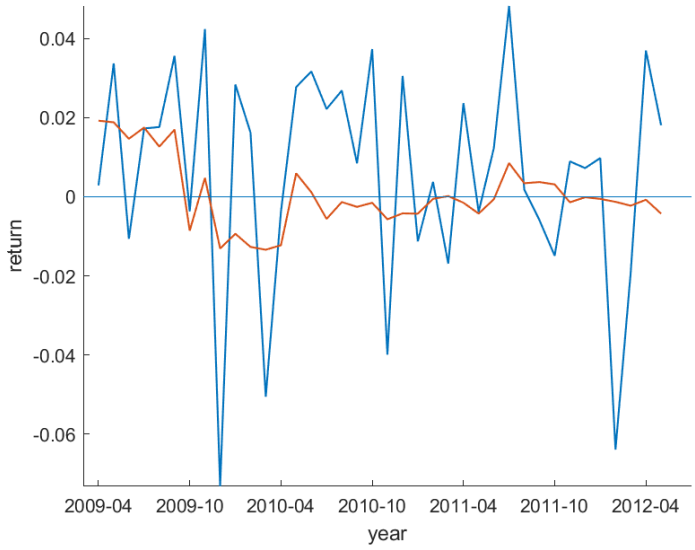

Figure 3.3: Prediction of JPY exchange rate movements using yield curve, option, and macroeconomic (raw) data

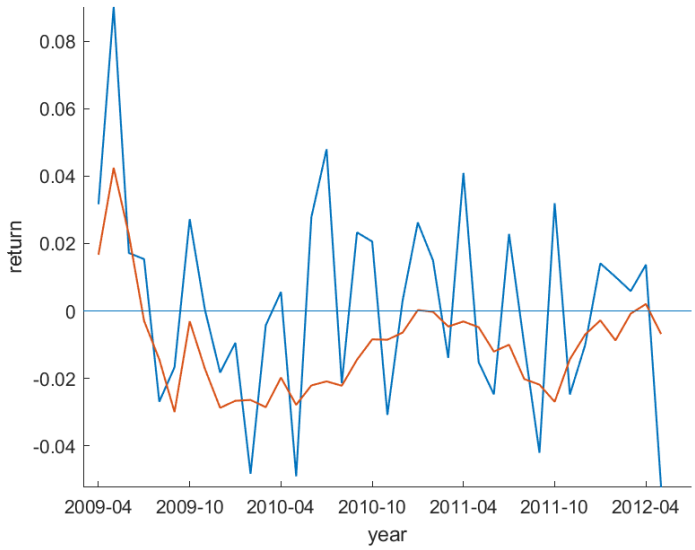

Figure 3.4: Prediction of UK exchange rate movements using PCs extracted from yield curve, option, and macroeconomic variables

This is seen in other well-performing cases as well and confirms that instead of the improvements all come from fitting one extreme data point, neural network with its non-linearity and flexibility is able to extract useful information for exchange rate forecasts to some extent. The inconsistent pattern in column (2) and (3) suggest the information contained by macro data and term structure variables is different and their interaction matters for forecast exercises. In our neural networks, the two groups of data are kept separate until the output layer to prevent them from interacting with each other when passing through the network. Thus the results in the middle two columns where we would combine two different types of data at the end lacks the consistency in the other two columns. In our case, the curse of dimensionality is not a major concern, although traditional wisdom suggest otherwise. Comparing the rows of raw data and orthogonalized data shows that the structure of the data set affects how effective neural network can extract information, but there is no clear indication about which method is better. If we add principal components into the comparison, the discussion becomes even harder because now we are bringing in some cross-country differences in their economic structure, which affects the information contained in the top principal components and are not necessarily related to the exchange rate.

The mixture of uncertain results are expected because our sample size is very small comparing to the application of neural network in other disciplines or in the industry. The stochastic nature of neural network's estimation procedure may have also contributed to these results. Although we followed a robust procedure, in practice we still find slightly different results between multiple runs of the program. The complexity in neural network estimation is also seen when comparing across columns. As we will see later in this section, with traditional methods, it is easy to eyeball the relative contribution (positive or negative) of each additional group of predictors. However, using Canada raw data as an example, adding in term structure variables to macroeconomic variables significantly reduces the forecast performance, while term structure variables alone yield improvements over the benchmark. The advantage of neural network also makes it hard to fully understand the differences between

results and we leave the extended discussion to future work. Despite the limitations with our current study, we are still able to show positive evidence that neural network provides a new way of extracting exchange rate signals from macroeconomic and term structure variables.

We also find dropout layers to be very important while training the neural network. Dropout layers randomly selects a fixed proportion of neurons to deactivate in each training step, artificially introduces noise such that the network would only learn strong signals rather than memorizing both signal and noise from the data, i.e. overfitting the data. Batch normalization is another regularization method commonly used in neural network estimation. It normalizes the input distribution of each neuron to stabilize and aid the SGD process. We use both methods in the results reported above. In our exercise, we did two other runs by turning off both or one of them. Turning off the BN layer only very slightly reduces the OOS R^2, while turning off the dropout layer would cause the forecast performance to severely deteriorate, sometimes even by over 100%.

Random Forest

Similar to neural network, random forest is another very flexible nonlinear methods. But unlike MLP, random forest does not perform as good. As a classification method looking for the closest relatives of a data point, it has a different training objective than all other methods we examine. The result presented in Table 3.5 can be evidence verifying a suspicion that information extracted from random forest may not completely overlap with other methods, thus giving it room for random forest to be incorporated either as a standalone method or to pre-process the data. Indeed random forest has been combined with neural network in some exercises in industrial organization literature to predict price changes (e.g. Bajari et al. [2020]).

Table 3.5: Rolling window out-of-sample R^2s of random forest

		(1)	(2)	(3)	(4)
Australia	Raw data	-9.50%	-13.70%	-6.66%	-34.27%
	PCs	-4.85%	-21.60%	-22.00%	-50.74%
	Orthogonalized data	-8.64%	-13.16%	-11.04%	-21.68%
Canada	Raw data	-6.87%	-11.04%	-16.10%	-42.83%
	PCs	-20.20%	-25.60%	-39.75%	-64.34%
	Orthogonalized data	-0.84%	-0.95%	-6.61%	-7.47%
UK	Raw data	-9.31%	-14.92%	-10.04%	-20.96%
	PCs	-7.21%	-14.35%	-18.27%	-18.74%
	Orthogonalized data	-7.85%	-11.15%	-6.61%	-8.97%
Japan	Raw data	1.51%	-0.29%	0.18%	-15.75%
	PCs	-5.15%	2.01%	-18.97%	-30.00%
	Orthogonalized data	-9.91%	-4.59%	-4.65%	-2.63%

Note: The four model specifications are (1) with macroeconomic variables only, (2) with yield curve and macroeconomic variables, (3) with yield curve, option, and macroeconomic variables, (4) with yield curve and option variables.

Linear Models

Table 3.6 shows the out-of-sample performance of linear methods applied to the three formats of Japanese data. We use this as an example of our analysis in this section. The full results for all countries are available in the appendix. We broadly define linear methods to include not only OLS, but also constrained linear methods, i.e. LASSO, ridge, and elastic net, as well as partial least squares, which involves a non-linear process of selecting factors but the final estimation procedure is still linear. In the exercise using principal components, we excluded PLS, because PLS by itself is a supervised dimension reducing method, serving the same purpose as PCA. In addition to raw data, we apply all methods to top principal components hoping to improve the results by feeding the information along axes with highest variance directly to our methods, then to all principal components (i.e. the orthogonalized data set) to use all the information yet avoid colinearity.

As expected OLS presents serious overfitting issue, resulting in bad forecast performance. Even only using principal components, with 22 features in spec (3), PCR still cannot produce reasonable forecasts using our monthly data set. Least squares methods with parameter regularization have a better performance than OLS because they are able to force coefficient estimates into a small range, and in the case of LASSO, select a small set of features by forcing other coefficients to zero. By construction, elastic net always has an R^2 between that of LASSO and ridge. In most cases the R^2s hover around zero, with some slight improvements over random walk that are neither consistent throughout specifications nor statistically significant. The small absolute values also indicate that a big proportion of their predictions coincides with those of random walk with occasional non-zero forecasts that are in the correct direction.

Partial least squares, albeit it being a supervised dimension reduction method, performs poorly in the out-of-sample category. Our intuition is that the factor extraction process of PLS considers the covariances between dependent and independent variables, but does not take into account out-of-sample predictabilities. As a result, we can see in table **??** the in-sample R^2s for PLS are higher than the above three methods, but out-of-sample performance is significantly worse. Another contributing factor to this difference is that when estimating LASSO, ridge, and elastic net models, we use time series cross validation to select the value of λ, governing regularizing strengths, best for out-of-sample predictions. This is important to keep in mind when examining the in-sample results from table **??** as the low in-sample R^2 in, for example, LASSO with macroeconomic variables is not the 'best' achievable number, but rather one chosen by a cross validation procedure eyeing on out-of-sample performance.

Comparing across different data combinations, there is no clear evidence that one specification is better than another. However unlike neural networks, it does not seem that linear methods are able to pick out relevant information in the term structure variables.

This mediocre performance of linear methods is shared among all countries, without a pattern of one method consistently outperform another. There are three potential reasons: 1) there is significant cross-country variations in terms of what factor affects exchange rates the most, which makes the cross-country comparison difficult structurally; 2) the information extracted by these methods is not very relevant, either due to the linear nature of the models or because there is indeed an empirical disconnect between exchange rate and other variables; 3) there is strong structural shifts in our sample period, making our model unstable. The first two are out of the scope of this paper, but we discuss the last point in the next section.

Table 3.6: Rolling window out-of-sample R^2s using Japanese data

		(1)	(2)	(3)	(4)
Raw	Lasso	0.00%	0.00%	-0.10%	0.00%
	Ridge	-8.36%	-9.21%	-2.58%	-3.60%
	Elastic Net	0.00%	0.00%	-0.10%	0.02%
Factors	PCR	-4.16%	-17.97%	-32.91%	-31.13%
	Lasso	0.00%	-2.51%	-1.32%	-3.13%
	Ridge	-0.16%	-0.32%	-1.90%	-4.68%
	EL Net	0.00%	-2.50%	-1.31%	-3.12%
Orthogonalized	Lasso	0.31%	0.57%	1.70%	0.83%
	Ridge	1.74%	2.11%	-4.91%	-14.16%
	Elastic Net	0.31%	0.56%	1.69%	0.83%

Note: The four model specifications are (1) with macroeconomic variables only, (2) with yield curve and macroeconomic variables, (3) with yield curve, option, and macroeconomic variables, (4) with yield curve and option variables.

3.5 Discussion on the Instability in Data

Our results in the previous section show that in general there exists information in macroeconomic data which is related to exchange rate movements. However as in previous literature, linear framework cannot extract the information and even with fairly complicated neural

networks, we still cannot extract them in a consistent and robust way. In this section, we discuss the instability issue and potential solutions.

Instability may come from the structure of the dataset itself, as shown by a closer look at our PCA procedure. We extract principal components recursively, meaning we re-extract principal components in each iteration of the estimation procedure to prevent information leakage, as if a person observes today's data, extracts the important factors using his past information, estimates the parameters, and produces his forecast for next period exchange rate changes. By doing this, we prevent future observations from having an effect on how the PCs are extracted today. The caveat is that the composition of principal components (i.e. the factor loadings) can be unstable - with the data evolving, geometrically the direction of the first PC may rotate and subsequently affect the direction of other PCs. On the other hand, this gives us an opportunity to examine the underlying structure of the dataset. If the weights from the PCA changes dramatically, it indicates that the variances along different axis (i.e. information) in the dataset also changes dramatically. This instability in the right hand side variables can strongly affect our estimation, not only in methods using principal components but in all other methods as well.

To check for this, we monitor the absolute values of the factor loadings for each PC we use in our estimation. Fig 3.5 plots the factor loadings of the first PC from the Canadian macroeconomic data set on all features. The loading values corresponding to each month, i.e. each vertical slice in the graph, is the value of the unit vector that determines the direction of the principal component. The graph shows that indeed there is significant structural shifts in the composition of that PC around the Great Recession, that would cause the principal component to rotate. This problem is shown in all PCs from Canadian data set. Referring back to Table 3.4, Canada is the only one country where the nerual network didn't produce any sizeable improvements.

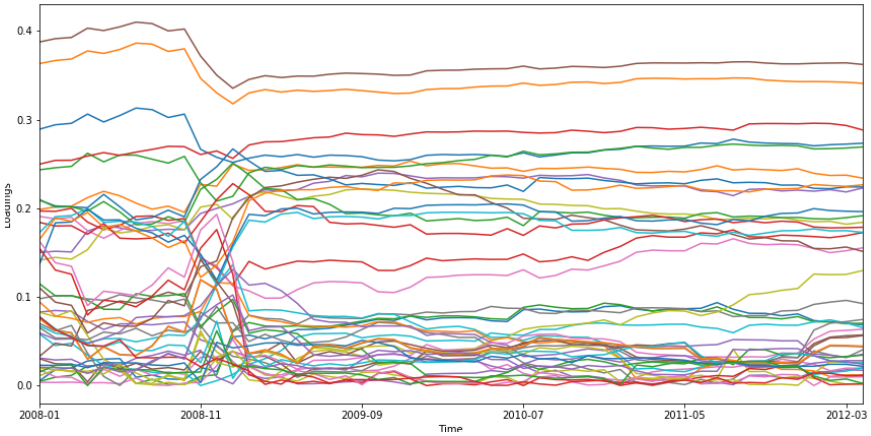

Figure 3.5: Factor loadings of the first PC extracted from Canadian macro data set

Table 3.7: Five Variables with highest loadings for the first PC of each country

	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
US	Baa bond yield - FFR	VXO index	Unemp rate	Unemp age 15+	Aaa bond yield - FFR
Australia	Emp rate age 15+	Car registration	Imports	Unemp rate age 25-54	Bus confidence
Canada	Emp age 15+	Emp rate age 15+	Emp in service industry	Emp rate age 25-54	Emp rate age 15-24
UK	Manuf production	Total industry production w/o construction	Investment goods production	Intermediate goods production	Consumer durable production
Japan	Retail trade	Unemp rate age 25+	Hourly wage	M3	M1

Analyzing PCs can also help understand why there are cross-country differences. We look at the most influential variable on the PCs, i.e. the variables that get the most loadings. We find there to be significant overlapping among the top 5 influencers for each PCs extracted from one country's macroeconomic data set, but across countries, the variables vary to a large degree, usually involving variables from different sectors of the economy. In Table 3.7 we report the top 5 influencer of the first PC to illustrate the difference between countries. We

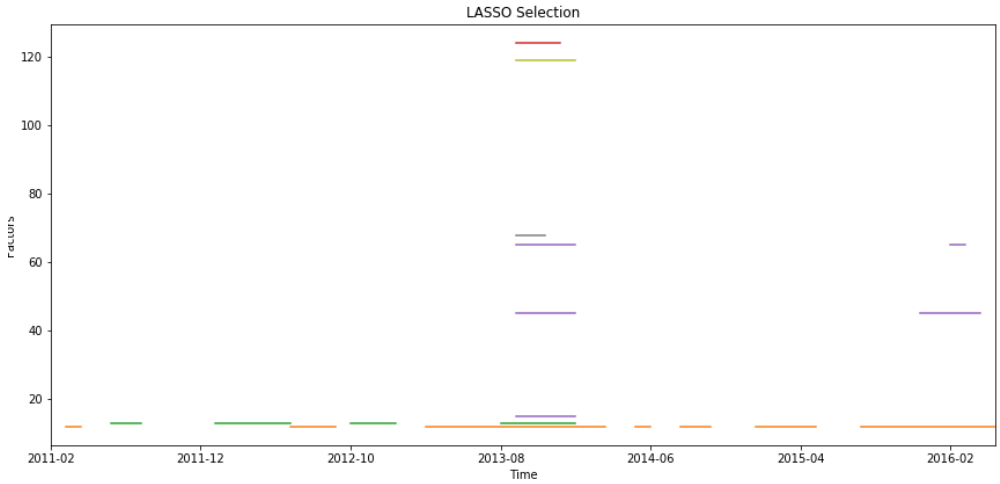

Figure 3.6: LASSO selection in raw Australian data

think these differences between countries and those in the result tables are due to structural differences, since exchange rates being the relative price of two countries' currency can never be said to not relate to a country's economic structure.

Next we look at the features selected by LASSO (Fig 3.6), as well as the selected PCs in the case of orthogonalized dataset (Fig 3.7). With raw data, we see that LASSO usually only selects one or two features if not nothing. With orthogonalized data, the selection does not concentrate at the first several PCs as expected, and similar to the case of raw data, the selection is extremely sparse. The two factors (green and yellow) that are most frequently picked up in the raw data case corresponds to two measures of industry production in the U.S., which is never expected to be the sole driver of exchange rates. When plotting the predictions of LASSO, they are all a very flat line around 0 with very rare non-zero predictions. Given the significant albeit inconsistent success with neural network, we take this as a sign that linear methods have trouble exploiting the relationship between exchange rate and macroeconomic fundamentals.

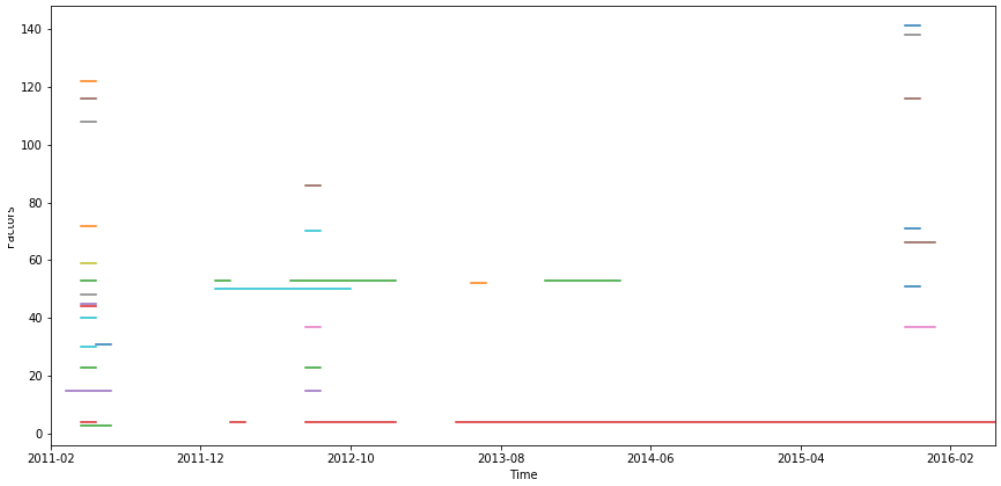

Figure 3.7: LASSO selection in orthogonalized Australian data

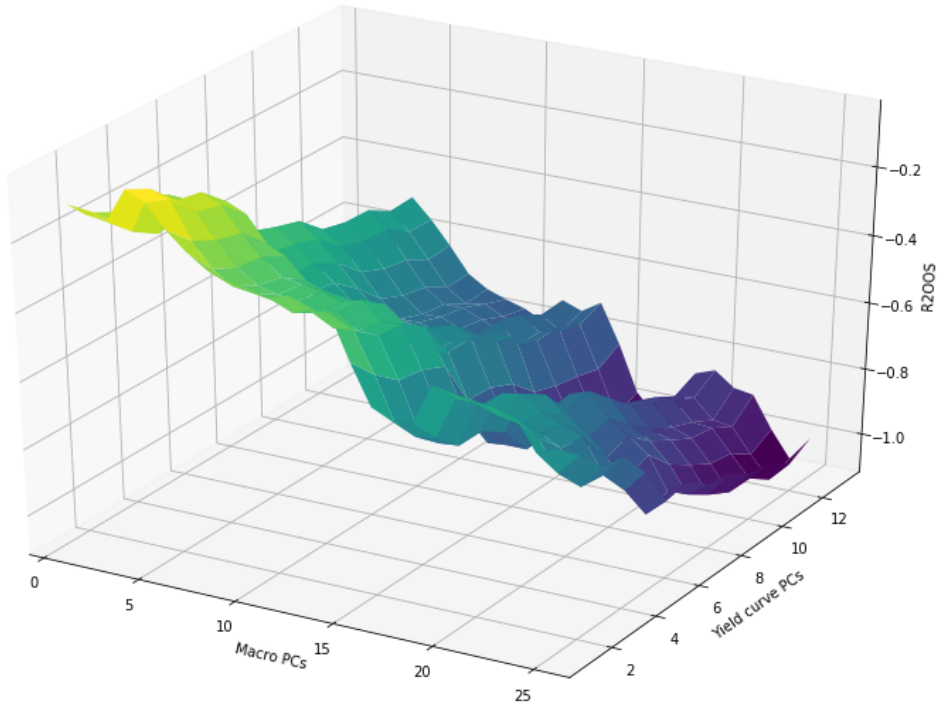

Figure 3.8: R^2_{OOS} with various combinations of PCs using Australian macro and yield curve data

To further strengthen the previous inference and as a robustness check for our choice of PC numbers, we re-run the PCR with all combinations of numbers of PCs between macroeconomic data and term structure variables. None of the PC combinations are able to beat random walk, e.g. all points in Fig 3.8 have a negative z-value. Despite that with more numbers of PCs, the regression suffers from overfitting, the steep decline in performance is an evidence of linear models' struggle.

Lastly we discuss some recent work in neural networks that may offer some insights to our results. Belkin et al. [2020] and Nakkiran et al. [2019] propose a double-descent phenomenon, where the mean squared prediction error (MSPE) would first increase with the complexity of the model, representing the traditional bias-variance trade-off and transition from underfitting to overfitting. Then after the peak, MSPE would decrease as the model becomes more complex. The double descent result can help explain why in forecasting exchange rates, OLS with 150-200 features produces horrible results while neural network with hundreds of thousands of parameters can have reasonable results. Fan et al. [2021] use neural network differently - instead of using neural network as a enclosed estimation 'box', they use neural networks to estimate each component of the traditional structural model. When applied to factor pricing models, their estimation with only structural component and without the idiosyncratic noise yields significantly more predictive power than the traditional application. We are optimistic about applying this structural deep learning approach in forecasting exchange rates.

3.6 Conclusion

In this paper we re-examine the long-standing empirical disconnect between the exchange rate and its theoretical macroeconomic determinants by exploring a large set of monthly data that captures both current macroeconomic conditions as well as market expectations and perceived uncertainties about them, as embodied in the term structures of relevant asset prices. We employ a variety of linear and non-linear machine learning techniques, to

examine their predictive power for subsequent exchange rate movements, both in in-sample regressions and in pseudo-out-of-sample forecasts.

Our in-sample results confirm the correlation between macroeconomic fundamentals and exchange rate movements. However the out-of-sample results of the linear models confirm the long existing verdict that random walk is hard to beat in the prediction of exchange rates, even with principal component regression and penalized least squares. In the non-linear front, the neural network produces some statistically significant improvements over the random walk prediction, with a generally better overall performance. From the PCA factor loadings, we see instability in the macroeconomic datasets and cross-country differences in economic structure, both of which make it hard to extract a consistent and robust relationship for exchange rate prediction. LASSO selection also makes it clear that linear methods struggle to select meaningful variables. On the neural network front, we plan to fine tune the neural network, try to disentangle the reason of the better performance, and incorporate some recent developments in the field in the future.

BIBLIOGRAPHY

Risk premia in crude oil futures prices. *Journal of International Money and Finance*, 42:9 – 37, 2014. Understanding International Commodity Price Fluctuations.

Carlo Acerbi and Dirk Tasche. On the coherence of expected shortfall. *Journal of Banking & Finance*, 26(7):1487–1503, 2002.

Gordon J Alexander and Alexandre M Baptista. A comparison of var and cvar constraints on portfolio selection with the mean-variance model. *Management science*, 50(9):1261–1273, 2004.

Farid Alizadeh and Donald Goldfarb. Second-order cone programming. *Mathematical programming*, 95(1):3–51, 2003.

A. Ang and J. Chen. Asymmetric correlations of equity portfolios. *Journal of Financial Economics*, 63(3):443–494, 2002.

Andrew Ang and Geert Bekaert. International Asset Allocation With Regime Shifts. *The Review of Financial Studies*, 15(4):1137–1187, 06 2015. doi: 10.1093/rfs/15.4.1137.

David Ardia, Kris Boudt, and Leopoldo Catania. Generalized autoregressive score models in R: The GAS package. *Journal of Statistical Software*, 88(6):1–28, 2019. doi: 10.18637/jss.v088.i06.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. A characterization of measures of risk. Technical report, Cornell University Operations Research and Industrial Engineering, 1997.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

Jushan Bai and Serena Ng. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150, 2006.

Jushan Bai and Serena Ng. Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2):304–317, 2008.

Pat Bajari, Z. Cen, V. Chernozhukov, R. Huerta, J. Li, M. Manukonda, and G. Monokrousos. Quality-adjusted price indices powered by ai. Working paper, Amazon Core AI, 2020.

Suleyman Basak and Anna Pavlova. A model of financialization of commodities. *Journal of Finance*, 71:1511–1556, 2016.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

Stefano Benati. The optimal portfolio problem with coherent risk measure constraints. *European Journal of Operational Research*, 150(3):572–584, 2003.

Dimitris Bertsimas, Geoffrey J Lauprete, and Alexander Samarov. Shortfall as a risk measure: properties, optimization and applications. *Journal of Economic Dynamics and control*, 28 (7):1353–1381, 2004.

Daniele Bianchi, Matthias Büchner, and Andrea Tamoni. Bond risk premiums with machine learning. *The Review of Financial Studies*, 34(2):1046–1089, 2021.

Jean Boivin and Serena Ng. Are more data always better for factor analysis? *Journal of Econometrics*, 132(1):169–194, 2006.

T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31:307–327, 1986.

Tim Bollerslev. Modelling the coherence in short-run nominal exchange rates: a multivariate generalized arch model. *The review of economics and statistics*, pages 498–505, 1990.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Mario Brandtner. Conditional value-at-risk, spectral risk measures and (non-) diversification in portfolio selection problems–a comparison with mean–variance analysis. *Journal of Banking & Finance*, 37(12):5526–5537, 2013.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Florentin Butaru, Qingqing Chen, Brian Clark, Sanmay Das, Andrew W Lo, and Akhtar Siddique. Risk and risk management in the credit card industry. *Journal of Banking & Finance*, 72:218–239, 2016.

Bahattin Buyuksahin and Michel A. Robe. Speculators, commodities and cross-market linkages. *Journal of International Money and Finance*, 42:38–70, 2014.

John Y Campbell and Samuel B Thompson. Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, 21(4):1509–1531, 2008.

G. Casella and R. L. Berger. *Statistical Inference, Duxbury Press, U.* S.A, 1990.

Yu-chin Chen and Kwok Ping Tsang. What does the yield curve tell us about exchange rate predictability? *The Review of Economics and Statistics*, 95(1):185–205, 2013. URL https://EconPapers.repec.org/RePEc:tpr:restat:v:95:y:2013:i:1:p:185-205.

Yu-chin Chen, Ranganai Gwati, and Jingyi Ren. Currency returns and the term structure of fx derivatives. Working paper, University of Washington, 2018.

U. Cherubini, E. Luciano, and W. Vecchiato. *Copula Methods in Finance*. John Wiley Sons, England, 2004.

B. Choros, R. Ibragimov, and E. Permiakova. Copula estimation. In Lecture Notes, editor, *Workshop on Copula Theory and its Applications*. in Statistics - Proceedings, F. Durante, W. H?rdle, P. Jaworski, and T. Rychlik, eds., Springer, 2010.

Peter F Christoffersen. Evaluating interval forecasts. *International economic review*, pages 841–862, 1998.

Todd E Clark and Kenneth D West. Approximately normal tests for equal predictive accuracy in nested models. *Journal of econometrics*, 138(1):291–311, 2007.

John H Cochrane. *Asset pricing: Revised edition*. Princeton university press, 2009.

John H Cochrane and Monika Piazzesi. Bond risk premia. *American economic review*, 95 (1):138–160, 2005.

Drew Creal, Siem Jan Koopman, and André Lucas. Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, 28(5):777–795, 2013. doi: 10.1002/jae.1279. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.1279.

Francis X Diebold and Roberto S Mariano. Com paring predictive accu racy. *Journal of Business and Economic Statistics*, 13(3):253–263, 1995.

Zhuanxin Ding, Clive WJ Granger, and Robert F Engle. A long memory property of stock market returns and a new model. *Journal of empirical finance*, 1(1):83–106, 1993.

Bradley Efron and Trevor Hastie. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press, USA, 1st edition, 2016. ISBN 1107149894.

Graham Elliott and Allan Timmermann. *Handbook of economic forecasting*. Elsevier, 2013.

Graham Elliott, Antonio Gargano, and Allan Timmermann. Complete subset regressions. *Journal of Econometrics*, 177(2):357–373, 2013.

P. Embrechts, A. McNeil, and D. Straumann. Correlation and dependence properties in risk management: Properties and pitfalls. In *M.* Value at Risk and Beyond, Cambridge University Press, Risk Management, 2002.

Charles Engel. Chapter 8 - exchange rates and interest parity. In Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff, editors, *Handbook of International Economics*, volume 4 of *Handbook of International Economics*, pages 453–522. Elsevier, 2014. doi: https://doi.org/10.1016/B978-0-444-54314-1.00008-2. URL https://www.sciencedirect.com/science/article/pii/B9780444543141000082.

Charles Engel and Kenneth D. West. Exchange rates and fundamentals. *Journal of Political Economy*, 113(3):485–517, 2005. ISSN 00223808, 1537534X. URL http://www.jstor.org/stable/10.1086/429137.

Robert Engel. Dynamic conditional correlation: a simple class of multivariate garch models. *Journal of Business and Economic Statistics*, 20(3):339–350, 2002.

R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of uk inflation. *Econometrica*, 50:987–1007, 1982.

Claude B. Erb and Campbell R. Harvey. The strategic and tactical value of commodity futures. *Financial Analysts Journal*, 62(2):69–97, 2006.

Claude B. Erb, Campbell R. Harvey, and Tadas E. Viskanta. Forecasting international equity correlations. *Financial Analysts Journal*, 50(6):32–45, 1994. doi: 10.2469/faj.v50.n6.32.

Jianqing Fan, Tracy Ke, Yuan Liao, and Andreas Neuhierl. Structual deep learning in conditional asset pricing. *working paper*, 2021.

Yanqin Fan and Andrew J. Patton. Copulas in econometrics. *Annual Review of Economics*, 6(1):179–200, 2014. doi: 10.1146/annurev-economics-080213-041221.

Julian Faraway, George Marsaglia, John Marsaglia, and Adrian Baddeley. *goftest: Classical Goodness-of-Fit Tests for Univariate Distributions*, 2021. URL https://CRAN.R-project.org/package=goftest. R package version 1.2-3.

Guanhao Feng, Stefano Giglio, and Dacheng Xiu. Taming the factor zoo. *Chicago Booth research paper*, (17-04), 2017.

Viviana Fernandez. Copula-based measures of dependence structure in assets returns. *Physica A: Statistical Mechanics and its Applications*, 387(14):3615–3628, 2008.

Peter C Fishburn. Mean-risk analysis with risk associated with below-target returns. *The American Economic Review*, 67(2):116–126, 1977.

K. Forbes and R. Rigobon. No contagion, only interdependece: Measuring stock market co-movements. *Unpublished working paper. National Bureau of Economic Research, Cambridge, MA.*, 1999.

Mario Forni and Lucrezia Reichlin. Let's get real: a factor analytical approach to disaggregated business cycle dynamics. *The Review of Economic Studies*, 65(3):453–473, 1998.

Jeffrey A. Frankel and Andrew K. Rose. Chapter 33 empirical research on nominal exchange rates. volume 3 of *Handbook of International Economics*, pages 1689–1729. Elsevier, 1995. doi: https://doi.org/10.1016/S1573-4404(05)80013-9. URL https://www.sciencedirect.com/science/article/pii/S1573440405800139.

E. W. Frees and E. A. Valdez. Understanding relationships using copulas. *North American Actuarial Journal*, 2(1):1–25, 1998.

Joachim Freyberger, Andreas Neuhierl, and Michael Weber. Dissecting characteristics nonparametrically. *The Review of Financial Studies*, 33(5):2326–2377, 2020.

Anqi Fu, Balasubramanian Narasimhan, and Stephen Boyd. CVXR: An R package for disciplined convex optimization. *Journal of Statistical Software*, 94(14):1–34, 2020. doi: 10.18637/jss.v094.i14.

Alexei A Gaivoronski and Georg Pflug. Value-at-risk in portfolio optimization: properties and computational approach. *Journal of risk*, 7(2):1–31, 2005.

R. Garcia and G. Tsafack. Dependence structure and extreme comovements in international equity and bond markets. *Journal of Banking Finance*, 35(8):1954–1970, 2011.

C. Genest and A.-C. Favre. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12:347–368, 2007.

Alexios Ghalanos. *rugarch: Univariate GARCH models.*, 2020. R package version 1.4-4.

Gary Gorton and K. Geert Rouwenhorst. Facts and fantasies about commodity futures. *Financial Analysts Journal*, 62(2):47–68, 2006.

Pierre-Olivier Gourinchas and Aaron Tornell. Exchange rate dynamics, learning and misperception. Working Paper 9391, National Bureau of Economic Research, December 2002. URL http://www.nber.org/papers/w9391.

Robert J. Greer. The nature of commodity index returns. *The Journal of Alternative Investments*, 3(1):45–52, 2000.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, 33(5):2223–2273, 02 2020. ISSN 0893-9454. doi: 10.1093/rfs/hhaa009. URL https://doi.org/10.1093/rfs/hhaa009.

J. D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57:357–384, 1989.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

James B Heaton, Nick G Polson, and Jan Hendrik Witte. Deep learning for finance: deep portfolios. *Applied Stochastic Models in Business and Industry*, 33(1):3–12, 2017.

Martin Hellmich and Stefan Kassberger. Efficient and robust portfolio optimization in the multivariate generalized hyperbolic framework. *Quantitative Finance*, 11(10):1503–1516, 2011.

Chi-fu Huang and Robert H Litzenberger. *Foundations for financial economics*. North-Holland, 1988.

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.

Rob Hyndman, George Athanasopoulos, Christoph Bergmeir, Gabriel Caceres, Leanne Chhay, Mitchell O'Hara-Wild, Fotios Petropoulos, Slava Razbash, Earo Wang, and Farah Yasmeen. *forecast: Forecasting functions for time series and linear models*, 2021. URL https://pkg.robjhyndman.com/forecast/. R package version 8.15.

Rob J Hyndman and Yeasmin Khandakar. Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 26(3):1–22, 2008. URL https://www.jstatsoft.org/article/view/v027i03.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014. ISBN 1461471370.

H. Joe. Multivariate models and dependence concepts. *Monographs in Statistics and Probability*, 73, 1997.

H. Joe and J. J. Xu. The estimation method of inference functions for margins for multivariate models. working paper, Department of Statistics, University of British Columbia, 1996.

E. Jondeau and M. Rockinger. The copula-garch model of conditional dependencies: an international stock market application. *Journal of International Money and Finance*, 25 (5):827–853, 2006.

Bryan Kelly and Seth Pruitt. Market expectations in the cross-section of present values. *The Journal of Finance*, 68(5):1721–1756, 2013. doi: https://doi.org/10.1111/jofi.12060. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12060.

Bryan Kelly and Seth Pruitt. The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics*, 186(2):294–316, 2015. URL https://EconPapers.repec.org/RePEc:eee:econom:v:186:y:2015:i:2:p:294-316.

Bryan Kelly, Seth Pruitt, and Yinan Su. Some characteristics are risk exposures, and the rest are irrelevant. *Unpublished Manuscript, University of Chicago*, 2017.

Amir E Khandani, Adlar J Kim, and Andrew W Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.

Serhiy Kozak, Stefan Nagel, and Shrihari Santosh. Shrinking the cross-section. *Journal of Financial Economics*, 135(2):271–292, 2020.

Pavlo A Krokhmal. Higher moment coherent risk measures. 2007.

Chung-Ming Kuan and Halbert White. Artificial neural networks: An econometric perspective. *Econometric reviews*, 13(1):1–91, 1994.

Paul Kupiec. Techniques for verifying the accuracy of risk measurement models. *The J. of Derivatives*, 3(2), 1995.

David X. Li. On default correlation. *The Journal of Fixed Income*, 9(4):43–54, 2000. doi: 10.3905/jfi.2000.319253.

AW Lo. Neural networks and other nonparametric techniques in economics and finance. *Blending Quantitative and traditional Equity Analysis, Association for Investment Management and Research*, 1994.

B. Longin F. Solnik. Correlation structure of international equity markets during extremely volatile periods. *Journal of Finance*, 56:649–676, 2001.

Sydney C. Ludvigson and Serena Ng. Macro Factors in Bond Risk Premia. *The Review of Financial Studies*, 22(12):5027–5067, 10 2009. ISSN 0893-9454. doi: 10.1093/rfs/hhp081. URL https://doi.org/10.1093/rfs/hhp081.

H. Manner and J. Segers. Tails of correlation mixtures of elliptical copulas. *Insurance: Mathematics and Economics*, 48(1):153–160, 2011.

Harry Markowitz. Portfolio selection*. *The Journal of Finance*, 7(1):77–91, 1952a. doi: https://doi.org/10.1111/j.1540-6261.1952.tb01525.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1952.tb01525.x.

Harry Markowitz. The utility of wealth. *Journal of political Economy*, 60(2):151–158, 1952b.

Harry Markowitz. Portfolio selection, 1959.

Harry M Markowitz. Foundations of portfolio theory. *The journal of finance*, 46(2):469–477, 1991.

Harry M Markowitz, David Starer, Harvey Fram, and Sander Gerber. Avoiding the downside: A practical review of the critical line algorithm for mean–semivariance portfolio optimization. *HANDBOOK OF APPLIED INVESTMENT RESEARCH*, pages 369–415, 2020.

Douglas Martin. Lecture slides: Portfolio optimization and asset management, September 2015.

MATLAB. *9.7.0.1190202 (R2019b)*. The MathWorks Inc., Natick, Massachusetts, 2018.

Michael W McCracken and Serena Ng. Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589, 2016.

A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative Risk Management: Concepts*. Techniques and Tools, Princeton University Press, New Jersey, 2005.

Alexander J McNeil and Rüdiger Frey. Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of empirical finance*, 7(3-4):271–300, 2000.

Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management: concepts, techniques and tools-revised edition*. Princeton university press, 2015.

Marcial Messmer. Deep learning and the cross-section of expected returns. *Available at SSRN 3081555*, 2017.

JP Morgan et al. Riskmetrics technical document. 1996.

Sendhil Mullainathan and Jann Spiess. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2):87–106, Spring 2017. URL https://ideas.repec.org/a/aea/jecper/v31y2017i2p87-106.html.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*, 2019.

R. B. Nelsen. *An Introduction to Copulas, Second Edition, Springer, U.* S.A, 2006.

Tatsuyoshi Okimoto. New evidence of asymmetric dependence structures in international equity markets. *Journal of Financial and Quantitative Analysis*, 43(3):787–815, 2008. doi: 10.1017/S0022109000004294.

A. J. Patton. On the out-of-sample importance of skewness and asymmetric dependence for asset allocation. *Journal of Financial Econometrics*, 2(1):130–168, 2004.

A. J. Patton. Modelling asymmetric exchange rate dependence. *International Economic Review*, 47(2):527–556, 2006a.

A. J. Patton. Estimation of multivariate models for time series of possibly different lengths. *Journal of Applied Econometrics*, 21(2):147–173, 2006b.

A. J. Patton. Copula-based models for financial time series. In G. Andersen, J.-P. Kreiss R. A. Davis, and T. Mikosch, editors, *T.* Handbook of Financial Time Series, Springer Verlag, 2009a.

A. J. Patton. Are "market neutral" hedge funds really market neutral? *Review of Financial Studies*, 22(7):2495–2530, 2009b.

A. J. Patton. A review of copula models for economic time series. *Journal of Multivariate Analysis*, 2012.

Andrew Patton. Copula methods for forecasting multivariate time series. *Handbook of Economic Forecasting*, 2:899–960, 2013.

George Gaetano Pennacchi. *Theory of asset pricing*. Pearson/Addison-Wesley Boston, 2008.

Brian G. Peterson and Peter Carl. *PortfolioAnalytics: Portfolio Analysis, Including Numerical Methods for Optimization of Portfolios*, 2018. URL https://CRAN.R-project.org/package=PortfolioAnalytics. R package version 1.1.0.

Brian G. Peterson and Peter Carl. *PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis*, 2020. URL https://CRAN.R-project.org/package=PerformanceAnalytics. R package version 2.0.4.

Georg Ch Pflug. Some remarks on the value-at-risk and the conditional value-at-risk. In *Probabilistic constrained optimization*, pages 272–281. Springer, 2000.

Kelly Price, Barbara Price, and Timothy J Nantell. Variance and lower partial moment measures of systematic risk: some analytical and empirical results. *The Journal of Finance*, 37(3):843–855, 1982.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

J. C. Rodriguez. Measuring financial contagion: a copula approach. *Journal of Empirical Finance*, 14(3):401–423, 2007.

Mario Rorni and Lucrezia Reichlin. Dynamic common factors in large cross-sections. *Empirical economics*, 21(1):27–42, 1996.

Barbara Rossi. Exchange rate predictability. *Journal of Economic Literature*, 51(4):1063–1119, December 2013. doi: 10.1257/jel.51.4.1063. URL https://www.aeaweb.org/articles?id=10.1257/jel.51.4.1063.

Michael Rothschild and Joseph E Stiglitz. Increasing risk: I. a definition. *Journal of Economic theory*, 2(3):225–243, 1970.

Bernd Scherer. More than you ever wanted to know about conditional value at risk optimization. *Satchell, S. Optimizing Optimization: The Next Generation of Optimization Applications and Theory. San Diego: Elsevier*, pages 283–299, 2009.

William F Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, 19(3):425–442, 1964.

Kenneth J. Singleton. Investor flows and the 2008 boom/bust in oil prices. *Management Science*, 60(2):300–318, 2014a.

Kenneth J. Singleton. Investor flows and the 2008 boom/bust in oil prices. *Management Science*, 60:300–318, 2014b.

Justin Sirignano, Apaar Sadhwani, and Kay Giesecke. Deep learning for mortgage risk. *arXiv preprint arXiv:1607.02470*, 2016.

A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut Statistique de l'Universite´ de Paris*, 8:229–231, 1959.

James H Stock and Mark W Watson. Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, 97(460):1167–1179, 2002a.

James H Stock and Mark W Watson. Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162, 2002b.

James H Stock and Mark W Watson. Forecasting with many predictors. *Handbook of economic forecasting*, 1:515–554, 2006.

Ke Tang and Wei Xiong. Index investment and the financialization of commodities. *Financial Analysts Journal*, 68:54–74, 2012.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

J. von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton University Press, 1947.

Jingtao Yao, Yili Li, and Chew Lim Tan. Option price forecasting using neural networks. *Omega*, 28(4):455–466, 2000.

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. doi: https://doi.org/10.1111/j.1467-9868.2005.00503.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00503.x.

Appendix A

MODELING DEPENDENCY OF COMMODITIES WITH OTHER FINANCIAL ASSETS: A COPULA APPROACH

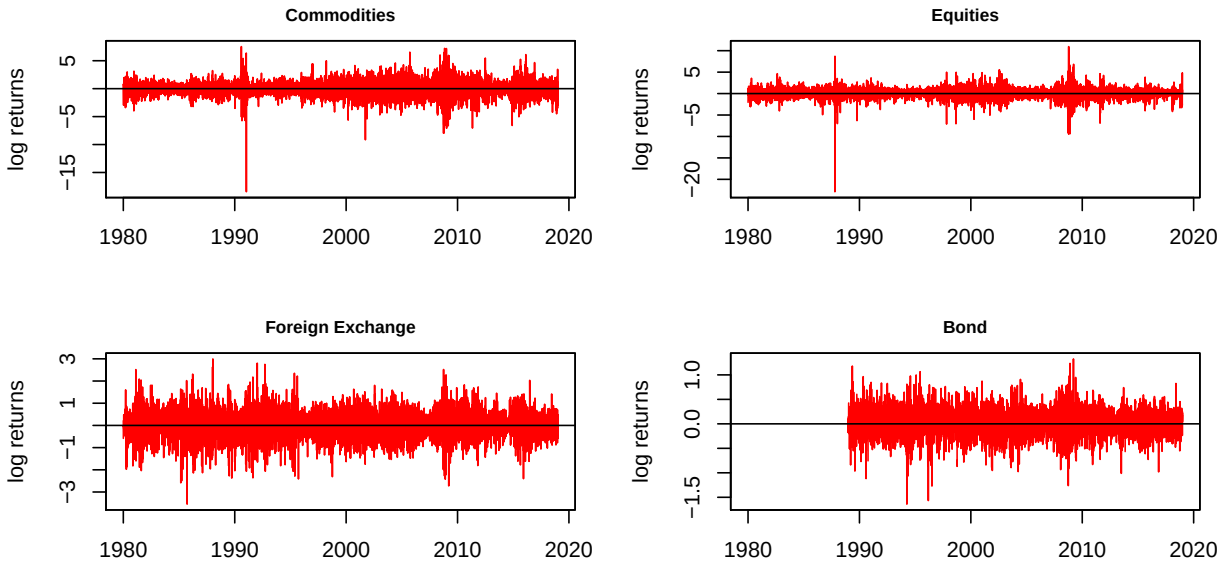

Figure A.1: Daily Log Returns

Appendix B

PORTFOLIO OPTIMIZATION BASED ON DOWNSIDE RISK ESTIMATES

	Minimum	Maximum	Mean	Stdev	Skewness	Kurtosis
TGNA	-0.2397	0.3941	0.0004	0.0229	1.5090	35.9399
AVP	-0.2767	0.2031	0.0003	0.0229	-0.1981	14.9416
PBI	-0.2842	0.2028	0.0003	0.0182	-0.5464	20.5571
THC	-0.4669	0.5501	0.0005	0.0313	0.4433	43.2263
AVY	-0.1467	0.1214	0.0005	0.0182	-0.3454	7.2654
HAS	-0.2431	0.3224	0.0005	0.0204	0.2954	20.1809
TSS	-0.2104	0.1316	0.0008	0.0205	-0.1564	7.5308
SPXC	-0.2403	0.2184	0.0006	0.0246	-0.5278	9.5194
R	-0.1796	0.1329	0.0005	0.0211	-0.1481	5.2634
HP	-0.1923	0.2650	0.0008	0.0262	0.0107	5.1405
J	-0.1506	0.2160	0.0006	0.0222	0.0625	8.4189
DBD	-0.2233	0.6103	0.0005	0.0221	3.8599	109.2075
HAR	-0.3765	0.2473	0.0009	0.0270	-0.1497	17.9787
BIG	-0.2941	0.2297	0.0006	0.0282	-0.1290	12.2267
HSC	-0.3473	0.1625	0.0003	0.0211	-0.8514	19.0424
MLHR	-0.2109	0.1783	0.0007	0.0240	0.1633	6.4375
AXE	-0.2486	0.2254	0.0006	0.0231	0.1756	8.3077
MATX	-0.1269	0.2046	0.0005	0.0207	0.6163	7.7300
KBH	-0.1629	0.2032	0.0006	0.0303	0.3883	3.5629
BGG	-0.1424	0.1403	0.0004	0.0202	0.0789	5.4174
CRS	-0.1890	0.1853	0.0006	0.0251	0.0601	5.9388
UVV	-0.1573	0.1365	0.0004	0.0182	-0.3132	7.3931
MENT	-0.3575	0.5123	0.0007	0.0333	0.5668	19.8750
HTLD	-0.1273	0.2067	0.0006	0.0236	0.3269	4.0207
BRC	-0.2222	0.2190	0.0006	0.0229	0.2027	8.0151
FUL	-0.1888	0.2344	0.0006	0.0239	0.3425	7.9940
ESND	-0.1364	0.1934	0.0008	0.0252	0.3505	4.6587
BOBE	-0.2227	0.1719	0.0004	0.0213	0.1042	7.8078
PIR	-0.4382	2.7500	0.0011	0.0540	23.2468	1164.3567
WTS	-0.1271	0.2056	0.0005	0.0234	0.4737	5.3489

Table B.1: CRSP Small Cap Stock Returns: summary statistics

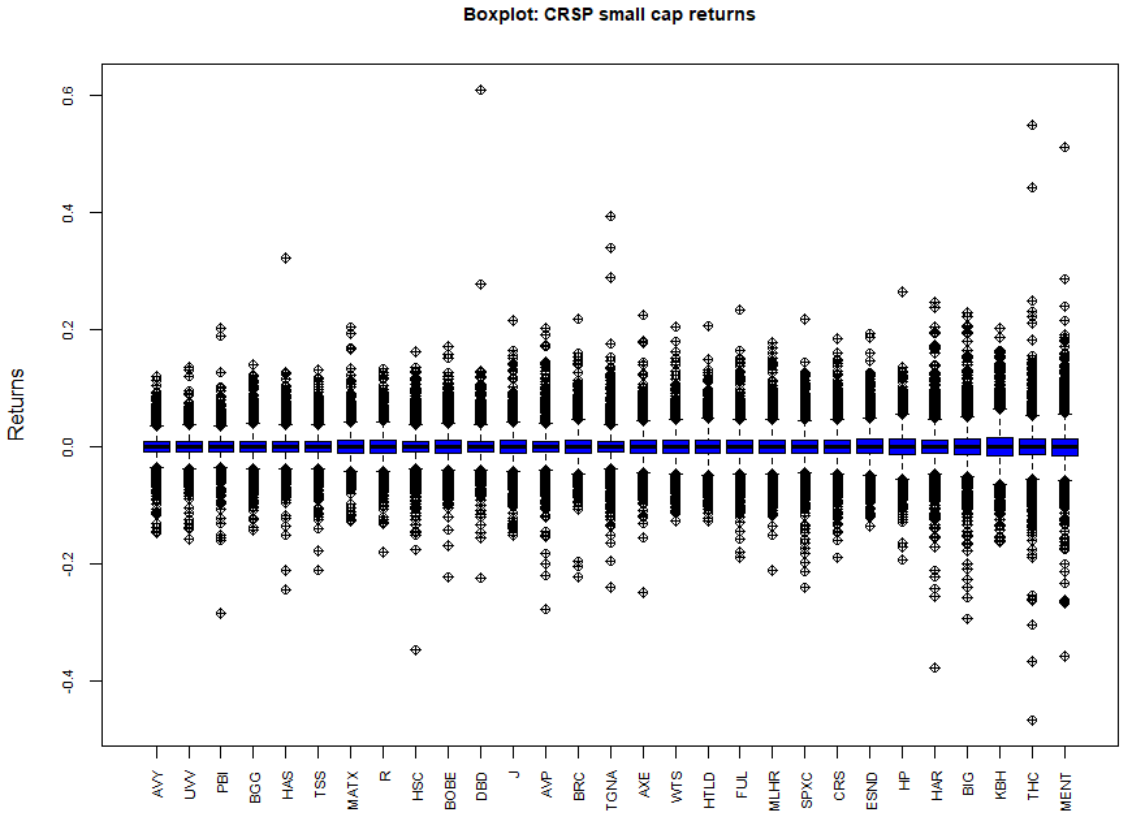

Figure B.1: Distribution of CRSP Small Cap Stock Returns

B.1 Minimum Expected Shortfall Linear Programming Proof

This proof is due to Professor R. Douglas Martin (Martin [2015]).
First, note that $e_i \geq t - \sum_{j=1}^{p} w_i r_{i,j}$ and $e_i \geq 0$ implies that:

$$e_i \geq \left[t - \sum_{j=1}^{p} w_j r_{i,j} \right]^+$$

$$\Rightarrow t - \frac{1}{n\gamma} \sum_{i=1}^{n} e_i \leq t - \frac{1}{n\gamma} \sum_{i=1}^{n} \left[t - \sum_{j=1}^{p} w_j r_{i,j} \right]^+$$

Now consider the maximization over e_i, t, w_j by first maximizing over e_i,t and subsequently maximizing with respect to w_j. From the above inequality we have:

$$\max_{\{e_i\},t} \left(t - \frac{1}{n\gamma} \sum_{i=1}^{n} e_i \right) \leq \max_{\{e_i\},t} \left(t - \frac{1}{n\gamma} \sum_{i=1}^{n} \left[t - \sum_{j=1}^{p} w_j r_{i,j} \right]^+ \right)$$

Note that if $e_i = 0$ then $e_i = \left[t - \sum_{j=1}^{p} w_j r_{i,j} \right]^+$

Suppose that in the LHS of (1) $e_i > \left[t - \sum_{j=1}^{p} w_j r_{i,j} \right]^+$ for some i.

In that case $e_i > 0$ can be decreased until $e_i = \left[t - \sum_{j=1}^{p} w_j r_{i,j} \right]^+$, thereby increasing $t - \frac{1}{n\gamma} \sum_{i=1}^{n} e_i$ and contradicting that the LHS of (1) is maximized.

This shows that the maximizing the LHS of (1) gives a result equal to maximizing the RHS.

Comment: The existence of high-quality large-scale LP software allows to compute ES optimal portfolios for portfolios with very large numbers of asset, e.g., 1,000 to 3,000 assets.

Appendix C

PREDICTING EXCHANGE RATES WITH MACHINE LEARNING: EXPECTATIONS, NONLINEARITY, AND PARAMETER INSTABILITY

C.1 Result tables and graphs

Table C.1: Rolling window out-of-sample R^2s of linear models using raw data

		(1)	(2)	(3)	(4)
Australia	OLS	-446.72%	-1483.95%	-588.18%	-365.52%
	Lasso	-1.47%	-1.38%	-12.23%	-0.30%
	Ridge	-1.35%	-6.53%	-2.30%	-1.37%
	Elastic Net	-1.46%	-1.38%	-12.19%	-0.34%
Canada	OLS	-8611.22%	-3570.88%	-664.21%	-272.75%
	Lasso	-2.62%	-3.20%	0.10%	-0.63%
	Ridge	-6.22%	-6.62%	-18.57%	-6.75%
	Elastic Net	-2.61%	-3.19%	0.10%	-0.63%
UK	OLS	-751.62%	-4832.92%	-543.16%	-138.70%
	Lasso	0.71%	0.71%	-2.84%	-3.66%
	Ridge	-2.85%	-4.08%	-16.61%	-10.93%
	Elastic Net	0.71%	0.71%	-3.00%	-3.66%
Japan	OLS	-599.64%	-2606.61%	-507.77%	-172.34%
	Lasso	0.00%	0.00%	-0.10%	0.00%
	Ridge	-8.36%	-9.21%	-2.58%	-3.60%
	Elastic Net	0.00%	0.00%	-0.10%	0.02%

Note: The four model specifications are (1) with macroeconomic variables only, (2) with yield curve and macroeconomic variables, (3) with yield curve, option, and macroeconomic variables, (4) with yield curve and option variables.

Table C.2: Rolling window out-of-sample R^2s of linear models using principal components

		(1)	(2)	(3)	(4)
Australia	PCR	-0.45%	-15.79%	-64.48%	-40.43%
	Lasso	4.25%	5.04%	0.78%	0.89%
	Ridge	1.05%	-1.19%	1.85%	-0.07%
	EL Net	4.24%	5.01%	0.79%	0.62%
Canada	PCR	-13.02%	-17.84%	-57.06%	-32.47%
	Lasso	-1.19%	-0.80%	-2.15%	-0.42%
	Ridge	-0.58%	-0.36%	-2.78%	-1.90%
	EL Net	-1.16%	-0.80%	-2.13%	-0.42%
UK	PCR	-4.99%	-9.94%	-63.21%	-40.65%
	Lasso	-3.41%	-1.47%	-9.07%	-2.90%
	Ridge	-0.53%	0.13%	-3.01%	-2.99%
	EL Net	-3.34%	-1.63%	-9.25%	-2.89%
Japan	PCR	-4.16%	-17.97%	-32.91%	-31.13%
	Lasso	0.00%	-2.51%	-1.32%	-3.13%
	Ridge	-0.16%	-0.32%	-1.90%	-4.68%
	EL Net	0.00%	-2.50%	-1.31%	-3.12%

Note: The four model specifications are (1) with macroeconomic variables only, (2) with yield curve and macroeconomic variables, (3) with yield curve, option, and macroeconomic variables, (4) with yield curve and option variables.

Table C.3: Rolling window out-of-sample R^2s of linear models using orthogonalized data

		(1)	(2)	(3)	(4)
Australia	OLS	-446.72%	-1483.95%	-1898.55%	-365.52%
	Lasso	-0.11%	-0.04%	2.78%	0.10%
	Ridge	-2.91%	-6.78%	-1.45%	-2.74%
	Elastic Net	-0.11%	-0.04%	2.77%	0.09%
Canada	OLS	-2364.58%	-1282.82%	-2038.73%	-272.75%
	Lasso	0.16%	-0.46%	0.06%	-4.18%
	Ridge	-0.70%	-1.94%	-9.43%	-12.91%
	Elastic Net	0.16%	-0.55%	0.07%	-4.16%
UK	OLS	-751.62%	-4832.92%	-57.10%	-138.70%
	Lasso	0.40%	0.28%	-8.39%	-7.56%
	Ridge	-3.28%	-5.50%	-7.08%	0.25%
	Elastic Net	0.39%	0.28%	-8.37%	-7.75%
Japan	OLS	-599.64%	-2606.61%	-1660.20%	-172.34%
	Lasso	0.31%	0.57%	1.70%	0.83%
	Ridge	1.74%	2.11%	-4.91%	-14.16%
	Elastic Net	0.31%	0.56%	1.69%	0.83%

Note: The four model specifications are (1) with macroeconomic variables only, (2) with yield curve and macroeconomic variables, (3) with yield curve, option, and macroeconomic variables, (4) with yield curve and option variables.

Table C.4: Australia: Evaluation of rolling window forecast using raw data

	Diebold-Mariano				Clark-West			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Lasso	0.910	0.890	0.984	0.627	0.888	0.853	0.966	0.617
Ridge	0.625	0.832	0.673	0.647	0.386	0.631	0.367	0.548
EL Net	0.910	0.890	0.984	0.646	0.889	0.855	0.966	0.636
Random Forest	0.872	0.898	0.693	0.954	0.535	0.606	0.311	0.754
Neural Network	0.248	0.252	0.420	0.251	0.149	0.049	0.102	0.151

Note: The four model specifications are (1) with macroeconomic variables only, (2) with yield curve and macroeconomic variables, (3) with yield curve, option, and macroeconomic variables, (4) with yield curve and option variables.

Table C.5: Canada: Evaluation of rolling window forecast using raw data

	Diebold-Mariano				Clark-West			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Lasso	0.810	0.821	0.251	0.910	0.762	0.790	0.246	0.921
Ridge	0.831	0.855	0.994	0.994	0.627	0.636	0.991	0.994
EL Net	0.810	0.822	0.251	0.910	0.762	0.791	0.246	0.921
Random Forest	0.831	0.918	0.877	0.994	0.560	0.637	0.691	0.932
Neural Network	0.184	0.842	0.969	0.188	0.102	0.693	0.186	0.105

Note: The four model specifications are (1) with macroeconomic variables only, (2) with yield curve and macroeconomic variables, (3) with yield curve, option, and macroeconomic variables, (4) with yield curve and option variables.

Table C.6: Japan: Evaluation of rolling window forecast using raw data

	Diebold-Mariano				Clark-West			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Lasso	0.899	0.926	0.670	0.467	0.901	0.928	0.658	0.458
Ridge	0.912	0.896	0.714	0.944	0.784	0.621	0.462	0.938
EL Net	0.130	0.159	0.670	0.121	0.128	0.157	0.658	0.117
Random Forest	0.406	0.520	0.491	0.918	0.130	0.138	0.153	0.559
Neural Network	0.242	0.515	0.076	0.272	0.156	0.101	0.013	0.178

Note: The four model specifications are (1) with macroeconomic variables only, (2) with yield curve and macroeconomic variables, (3) with yield curve, option, and macroeconomic variables, (4) with yield curve and option variables.

Table C.7: UK: Evaluation of rolling window forecast using raw data

	Diebold-Mariano				Clark-West			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Lasso	0.345	0.345	0.901	0.768	0.227	0.228	0.868	0.697
Ridge	0.682	0.754	0.962	0.967	0.403	0.491	0.898	0.947
EL Net	0.344	0.345	0.898	0.768	0.227	0.228	0.866	0.697
Random Forest	0.916	0.977	0.878	0.881	0.621	0.846	0.606	0.349
Neural Network	0.284	0.830	0.593	0.292	0.232	0.502	0.019	0.240

Note: The four model specifications are (1) with macroeconomic variables only, (2) with yield curve and macroeconomic variables, (3) with yield curve, option, and macroeconomic variables, (4) with yield curve and option variables.

Table C.8: Australia: Evaluation of rolling window forecast using factors

	Diebold-Mariano				Clark-West			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
PCR	0.531	0.916	0.996	0.990	0.167	0.303	0.706	0.819
Lasso	0.049	0.032	0.253	0.142	0.019	0.014	0.199	0.108
Ridge	0.270	0.706	0.194	0.520	0.154	0.482	0.106	0.435
EL Net	0.050	0.032	0.250	0.157	0.019	0.014	0.196	0.111
Random Forest	0.649	0.923	0.920	0.990	0.252	0.617	0.604	0.962
Neural Network	0.258	0.235	0.378	0.251	0.155	0.055	0.007	0.150

Note: The four model specifications are (1) with macroeconomic variables only, (2) with yield curve and macroeconomic variables, (3) with yield curve, option, and macroeconomic variables, (4) with yield curve and option variables.

Table C.9: Canada: Evaluation of rolling window forecast using factors

	Diebold-Mariano				Clark-West			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
PCR	0.856	0.919	0.999	0.993	0.542	0.614	0.989	0.974
Lasso	0.668	0.653	0.887	0.915	0.493	0.534	0.886	0.923
Ridge	0.580	0.608	0.972	0.997	0.468	0.505	0.968	0.998
EL Net	0.665	0.653	0.887	0.915	0.489	0.534	0.886	0.923
Random Forest	0.958	0.992	0.999	0.981	0.548	0.653	0.973	0.906
Neural Network	0.187	0.363	0.629	0.183	0.103	0.245	0.157	0.101

Note: The four model specifications are (1) with macroeconomic variables only, (2) with yield curve and macroeconomic variables, (3) with yield curve, option, and macroeconomic variables, (4) with yield curve and option variables.

Table C.10: Japan: Evaluation of rolling window forecast using factors

	Diebold-Mariano				Clark-West			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
PCR	0.791	0.977	0.963	0.996	0.491	0.566	0.668	0.973
Lasso	0.740	0.877	0.901	0.847	0.741	0.836	0.907	0.817
Ridge	0.570	0.554	0.912	0.928	0.504	0.259	0.900	0.888
EL Net	0.280	0.877	0.901	0.847	0.279	0.836	0.907	0.818
Random Forest	0.680	0.422	0.928	0.987	0.045	0.015	0.674	0.873
Neural Network	0.243	0.441	0.280	0.236	0.155	0.175	0.153	0.152

Note: The four model specifications are (1) with macroeconomic variables only, (2) with yield curve and macroeconomic variables, (3) with yield curve, option, and macroeconomic variables, (4) with yield curve and option variables.

Table C.11: UK: Evaluation of rolling window forecast using factors

	Diebold-Mariano				Clark-West			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
PCR	0.720	0.855	0.986	0.973	0.274	0.432	0.769	0.790
Lasso	0.929	0.816	0.916	0.876	0.873	0.732	0.922	0.875
Ridge	0.603	0.467	0.919	0.879	0.430	0.359	0.897	0.717
EL Net	0.927	0.828	0.920	0.876	0.869	0.751	0.924	0.875
Random Forest	0.769	0.932	0.931	0.838	0.138	0.478	0.733	0.204
Neural Network	0.297	0.541	0.278	0.283	0.245	0.285	0.034	0.231

Note: The four model specifications are (1) with macroeconomic variables only, (2) with yield curve and macroeconomic variables, (3) with yield curve, option, and macroeconomic variables, (4) with yield curve and option variables.

Table C.12: Australia: Evaluation of rolling window forecast using orthogonalized PCs

	Diebold-Mariano				Clark-West			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Lasso	0.838	0.635	0.202	0.222	0.839	0.615	0.173	0.208
Ridge	0.853	0.916	0.706	0.803	0.728	0.833	0.615	0.730
EL Net	0.838	0.635	0.202	0.222	0.839	0.615	0.173	0.208
Random Forest	0.866	0.894	0.838	0.956	0.699	0.740	0.547	0.884
Neural Network	0.253	0.517	0.196	0.265	0.152	0.106	0.003	0.161

Note: The four model specifications are (1) with macroeconomic variables only, (2) with yield curve and macroeconomic variables, (3) with yield curve, option, and macroeconomic variables, (4) with yield curve and option variables.

Table C.13: Canada: Evaluation of rolling window forecast using orthogonalized PCs

	Diebold-Mariano				Clark-West			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Lasso	0.448	0.717	0.496	0.853	0.386	0.635	0.446	0.822
Ridge	0.649	0.793	0.983	0.970	0.603	0.744	0.980	0.964
EL Net	0.448	0.756	0.495	0.853	0.386	0.680	0.445	0.822
Random Forest	0.580	0.590	0.787	0.713	0.150	0.166	0.594	0.431
Neural Network	0.185	0.685	0.515	0.186	0.102	0.231	0.050	0.103

Note: The four model specifications are (1) with macroeconomic variables only, (2) with yield curve and macroeconomic variables, (3) with yield curve, option, and macroeconomic variables, (4) with yield curve and option variables.

Table C.14: Japan: Evaluation of rolling window forecast using orthogonalized PCs

	Diebold-Mariano				Clark-West			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Lasso	0.211	0.161	0.145	0.059	0.208	0.159	0.110	0.054
Ridge	0.254	0.261	0.909	0.937	0.097	0.095	0.852	0.873
EL Net	0.211	0.161	0.145	0.059	0.208	0.159	0.109	0.054
Random Forest	0.977	0.794	0.739	0.608	0.911	0.477	0.533	0.217
Neural Network	0.246	0.676	0.799	0.284	0.160	0.345	0.254	0.190

Note: The four model specifications are (1) with macroeconomic variables only, (2) with yield curve and macroeconomic variables, (3) with yield curve, option, and macroeconomic variables, (4) with yield curve and option variables.

Table C.15: UK: Evaluation of rolling window forecast using orthogonalized PCs

	Diebold-Mariano				Clark-West			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Lasso	0.189	0.157	0.918	0.986	0.184	0.153	0.918	0.978
Ridge	0.829	0.936	0.831	0.461	0.710	0.861	0.588	0.322
EL Net	0.189	0.157	0.918	0.988	0.184	0.153	0.919	0.981
Random Forest	0.949	0.976	0.842	0.845	0.770	0.903	0.628	0.573
Neural Network	0.280	0.457	0.469	0.301	0.229	0.162	0.111	0.247

Note: The four model specifications are (1) with macroeconomic variables only, (2) with yield curve and macroeconomic variables, (3) with yield curve, option, and macroeconomic variables, (4) with yield curve and option variables.

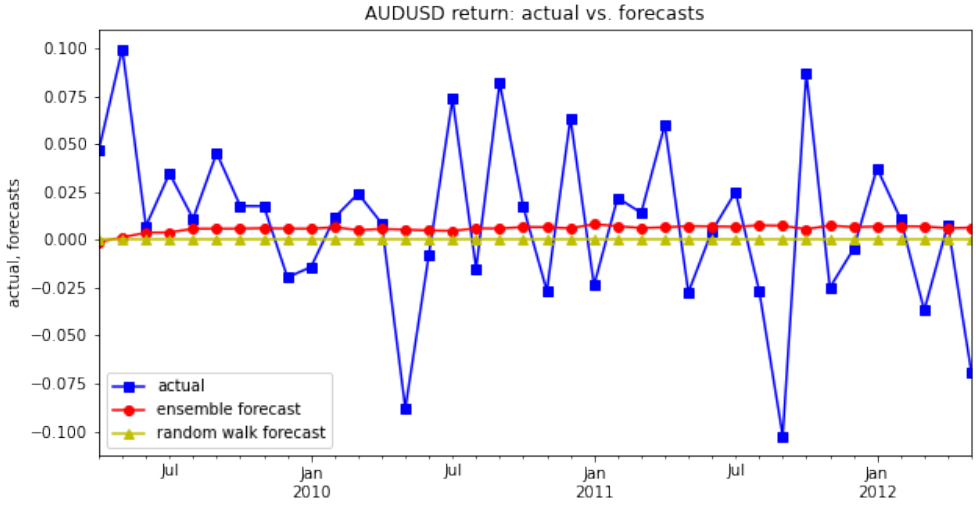

Figure C.1: AUDUSD: rolling out-of-sample ensemble forecast

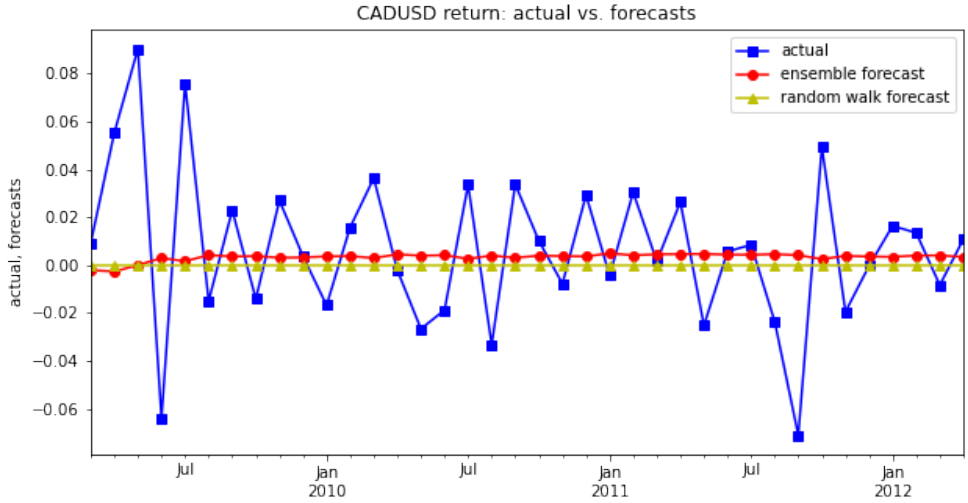

Figure C.2: CADUSD: rolling out-of-sample ensemble forecast

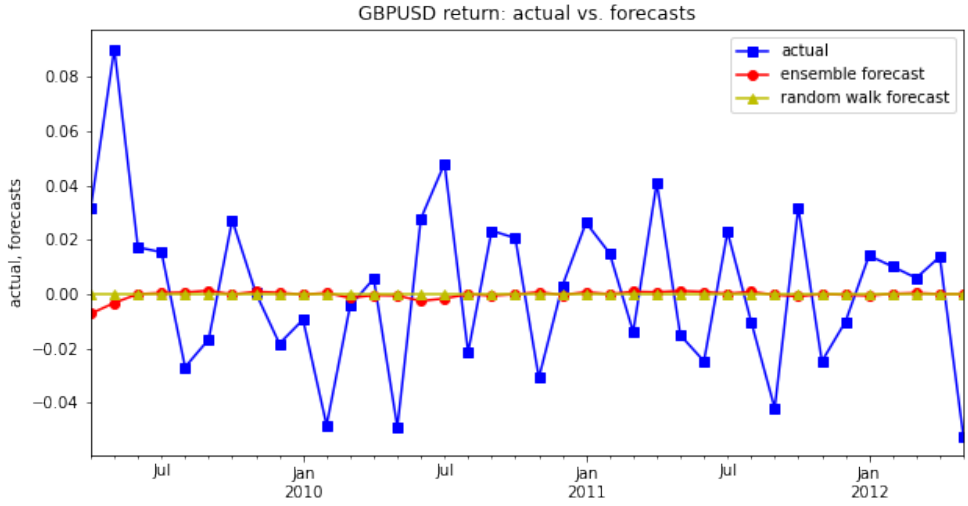

Figure C.3: GBPUSD: rolling out-of-sample ensemble forecast

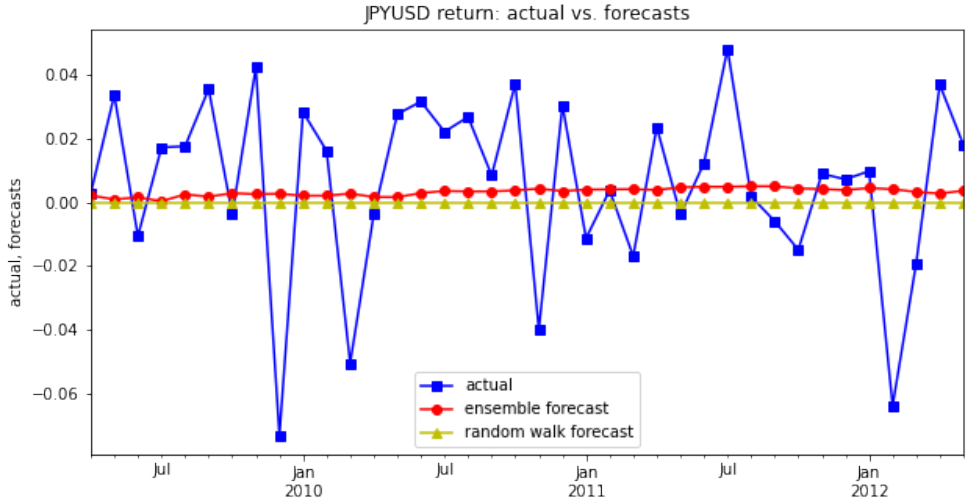

Figure C.4: JPYUSD: rolling out-of-sample ensemble forecast

Table C.16: Rolling window ensemble out-of-sample R^2

	OOS R^2
Australia	0.0023
Canada	-0.0196
UK	-0.0488
Japan	0.0196

C.2 Machine Learning Methods

We give brief overview of the machine learning methods used in this exercise. This section has heavily used materials (exposition, explanation, diagrams etc.) from these references Tibshirani [1996], James et al. [2014], Zou and Hastie [2005], Breiman [2001], Efron and Hastie [2016], Hastie et al. [2001].

C.2.1 Ridge Regression

When N is not very large relative to p, the estimated OLS coefficients can show large variance. One way to handle such a scenario is to *constrain* or *regularize* the coefficients to reduce the variance. This is done by shrinking the coefficients to zero. In penalized regression techniques, the regression coefficients are constrained by using a penalty on their size. The ridge coefficients are obtained by minimizing the penalized residual sum of squares

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{C.1}$$

where $\lambda \geq$ is a *tuning parameter* which controls the amount of the shrinkage. Note that the intercept β_0 is not included in the penalty term.
Making the size constraint on the parameters explicit, an alternative way to write the above problem is

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \tag{C.2}$$

subject to $\sum_{j=1}^{p} \beta_j^2 \leq t$

A greater amount of shrinkage can be achieved with a larger value of λ. When $\lambda = 0$, the penalty term hs no effect and we get the least squares estimates. When $\lambda \to \infty$, the

coefficients estimates will approach zero. Using the matrix notation, we can write

$$RSS(\lambda) = (y - \boldsymbol{X}\beta)'(y - \boldsymbol{X}\beta) + \lambda\beta'\beta \tag{C.3}$$

and the estimates are give by $\hat{\beta}^{ridge} = (\boldsymbol{X'X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X'y}$ where $\boldsymbol{I}$ is the $p \times p$ identity matrix. In case of orthogonal predictors, the ridge coefficients turn out to be a scaled version of the least squares estimates $\hat{\beta}^{ridge} = \frac{\hat{\beta}}{(1+\lambda)}$. As λ increases, greater amount of shrinkage of ridge coefficients causes a reduction in the variance of the predictions at the cost of a slight increase in bias. In situations where least squares estimates show high variance, ridge regression performs well.

C.2.2 LASSO Regression

The ridge regression selects all the p predictors and shrinks the coefficients towards zero but does not set any of them to exactly zero through the penalty term $\lambda \sum_{j=1}^{p} \beta_j^2$. In presence of a large number of predictors, this might pose a problem to interpreting the model. In stead of choosing all the predictors, it will be more sensible to choose only the most relevant variables. The method which enables us to do so is LASSO or Least Absolute Shrinkage and Selection Operator suggested by Tibshirani [1996]. The LASSO estimates are obtained as the solution to the following problem

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{C.4}$$

While the ridge regression uses L_2 penalty $\sum_{j=1}^{p} \beta_j^2$, the LASSO uses the L_1 penalty

$\sum_{j=1}^{p} |\beta_j|$. The L_1 penalty forces some of the coefficients to be exactly equal to zero. In that sense, LASSO performs variable selection. And based on the value of λ, LASSO can choose any number of variables while the the ridge regression retains all the variables. The L_1 penalty makes the LASSO solution nonlinear in y_i and there is no closed form solution.

An alternative way to set up the above problem

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \tag{C.5}$$
$$subject\ to \sum_{j=1}^{p} |\beta_j| \leq t$$

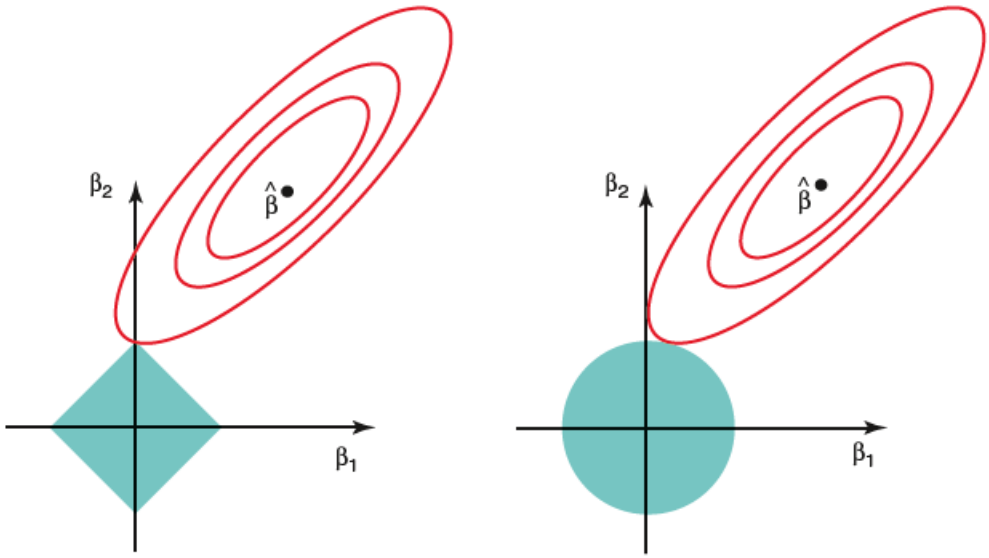

Figure C.5: Contours of the error and constraint functions for the lasso (left) and ridge regression (right).

The solid areas show the constraint regions, $|\beta_1| + |\beta_2| \leq s$ and $\beta_1^2 + \beta_2^2 \leq s$. The ellipses show the contours of the residual sum of squares. Source: James et al. [2014]

While finding the LASSO coefficients by minimizing the residual sum of squares, the constraint budget t determines how large $\sum_{j=1}^{p} |\beta_j|$ can be. If t is small, $\sum_{j=1}^{p} |\beta_j|$ needs to be small so that the constraint is not violated. For a large enough t, the least squares solution can be obtained. But when t is sufficiently small, some of the coefficients will be set to zero. The LASSO regression translates the coefficients by a constant factor *lambda* and this is referred to as *soft thresholding*. LASSO performs well in situations where some of the predictors have significant coefficient estimates but rest of the predictors have very small or almost zero coefficient estimates.

C.2.3 Elastic Net Regression

The elastic net penalty term is a combination of ridge and LASSO and can be written as

$$\lambda \sum_{j=1}^{p} \Big(\alpha \beta_j^2 + (1-\alpha)|\beta_j| \Big) \tag{C.6}$$

This was proposed by Zou and Hastie [2005]. The elastic net has two tuning parameters λ and α. It becomes L_1 penalty term when $\alpha = 0$ and it corresponds to ridge regression. The penalty term becomes L_2 norm when $\alpha = 1$ and it corresponds to LASSO. For any $0 < \alpha < 1$, the elastic net regression retains both the shrinkage and selection property.

C.2.4 Principal Component Regression

Principal Component Regression (PCR) is a dimension reduction technique and is applied when we have a large number of highly correlated predictors. In dimension reduction technique the predictors are first transformed so that they are orthogonal to each other and then a predictive least squares model is fit using those transformed predictors. The transformed predictors are obtained through the *Principal Component Analysis* (PCA). In PCA, the original predictors are linearly combined into a small set of regressors so that the original covariance structure of the predictors are preserved. Following the discussion of James et al.

[2014]. Let $X_1, X_2, \ldots, X_p$ be the p original predictors and $Z_1, Z_2, \ldots, Z_M$ be the the set of $M < p$ linear combination of p predictors such that

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j \tag{C.7}$$

where $\phi_{1m}, \phi_{2m}, \ldots, \phi_{pm}$, $m = 1, 2, \ldots, M$ are constants. In the second step these $Z_1, Z_2, \ldots, Z_M$ are used in the predictive linear regression. So instead of estimating $p + 1$ coefficients, $M + 1$ coefficients need to be estimated. Hence the term *dimension reduction*. Through PCA the dimension of $n \times p$ data matrix is reduced and the p principal components are obtained. The first principal component Z_1 indicates the direction of the data along which the observations vary the most and the second principal component Z_2 has the second largest variance and it is uncorrelated with Z_1. The key intuition behind principal component regression is that a smaller set of principal components can explain the most variance of the data as well as the relationship with the dependent variable. Thus it can overcome the problem of over fitting. When we have a large number of predictors and a small sample size i.e. p is large relative to sample size n, selecting $M << p$ will lead to a substantial reduction of variance of fitted coefficients.

C.2.5 Random Forest

In decision tree methods the predictor space is split into a number of segments based on some splitting rules. To build the regression tree, the predictor space $X_1, X_2, \ldots, X_p$ is divided into J distinct and non-overlapping regions $R_1, R_2, \ldots, R_J$. To make a prediction for any given observation belonging to region R_j, the mean of the dependent variable of all observations falling in region R_j is computed. To find J regions $R_1, R_2, \ldots, R_J$, the following RSS is minimized

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \tag{C.8}$$

where $\hat{y}_{R_j}$ is the mean of the dependent variable of observations belonging to region R_j. A computationally feasible way to divide the predictor space is binary recursive splitting. This is a top-down greedy approach as it starts at the top of the tree and splits the predictor space into two branches moving downward. For recursive binary splitting, the predictor X_j and the split point s divide the predictor space into two regions in such a way that it results in the largest possible reduction in RSS. Formally, for any j and s, $R_1(j, s) = \{X | X_j < s\}$ and $R_2(j, s) = \{X_j \geq s\}$.Then the j and s are determined by minimizing the following

$$\sum_{i: x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i: x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 \tag{C.9}$$

where $\hat{y}_{R_1}$ and $\hat{y}_{R_2}$ are the means of the dependent variable of observations falling within regions R_1 and R_2 respectively. To divide further, this same process is repeated and the most suitable predictor and the split point are identified in each of these two regions by minimizing the RSS. The process can continue until a stopping condition is reached. To improve the predictive performance, many such trees can be aggregated. Bootstrap aggregation or bagging is a statistical method which reduces the high variance of decision trees. In bagging many training sets are chosen from the population and using each training set a different prediction model is built and then the average of all predictions is selected. Considering there are B training sets chosen from the sample, we compute B predictions

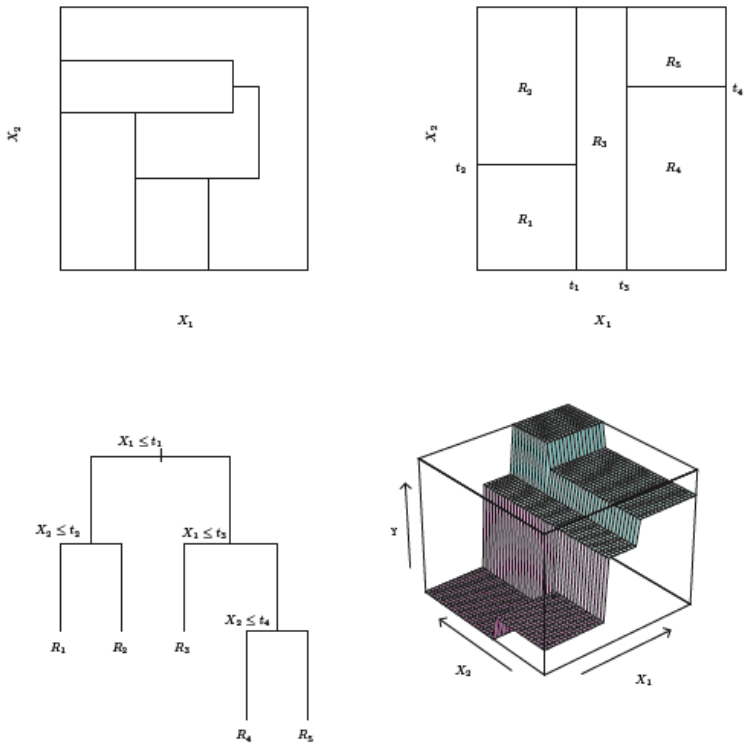

Figure C.6: Regression tree

Source: James et al. [2014]

$\hat{f}^1(x), \hat{f}^2(x), \ldots, \hat{f}^B(x)$ and then average them to get the final prediction

$$f_{avg}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(x) \tag{C.10}$$

Multiple training sets are generated by bootstrapping i.e. taking repeated samples (i.e. B times) from the original training set. Averaging B such noisy but approximately unbiased trees leads to a lower variance. In bagging the variance of B such averages can be shown to be $\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$ where σ^2 is variance of B identically distributed random variables and ρ is the pairwise correlation. For a very large B, the second term vanishes but the first term does not disappear. Reducing the correlation in the first term will lead to more

variance reduction. *Random Forest* proposed by Breiman [2001] improves the bagging method by de-correlating the trees and then takes an average of them. The tree is built on a bootstrapped dataset and before each split $m \leq p$ predictors are randomly selected as candidates for splitting. Typically, for classification problems, the default value for $m = \lfloor \sqrt{p} \rfloor$ and the default node size is one. For regression problems default value for $m = \lfloor p/3 \rfloor$ and the minimum node size is five.

After constructing B trees $\{T(x; \Theta_b)\}_1^B$ in this way, the random forest predictor is given by

$$\hat{f}_{rf}^B(x) = \frac{1}{B} \sum_{b=1}^{B} T(x; \Theta_b) \tag{C.11}$$

Random forest at each split consider only a subset of the predictors. The main difference between bagging and random forest is in the size of predictor subset i.e. m. For bagging $m = p$. Reducing the value of m will reduce the correlation between any pair of trees and will reduce the variance of the average. When there are a large number of predictors but the number of relevant variables is small, random forest does not perform well with small m. This is due to the fact that at any split there is a smaller chance that a relevant variable will be picked as a candidate for splitting.

C.2.6 Neural Network

Neural network is a nonlinear statistical model and is one of the most powerful modeling techniques in machine learning. It has found wide applications in different fields such as artificial intelligence, biology, image processing, natural language processing, computer vision, neuroscience. It is getting increasing attention in the field of economics and finance as a modeling technique for nonlinear models. Neural network is a highly parametrized machine learning tool which tries to approximate an arbitrary function f and is described as a *universal approximator*. For example, it maps an input $\boldsymbol{x}$ to output y i.e. $y = f(\boldsymbol{x})$. With sufficient data it can learn any smooth predictive relationship. Here we briefly outline the

structure of Multilayer Perceptron (MLP) or the traditional feed-forward neural network. A feed-forward neural network has at least three layers of nodes - one input layer which accepts the inputs or predictors, one or more hidden layers which transform the predictors through nonlinear activation function and one output layer which combines the hidden layers to a final output. The motivation behind the name "neural networks" comes from the human brain where each unit is a neuron, and the connections are synapses.

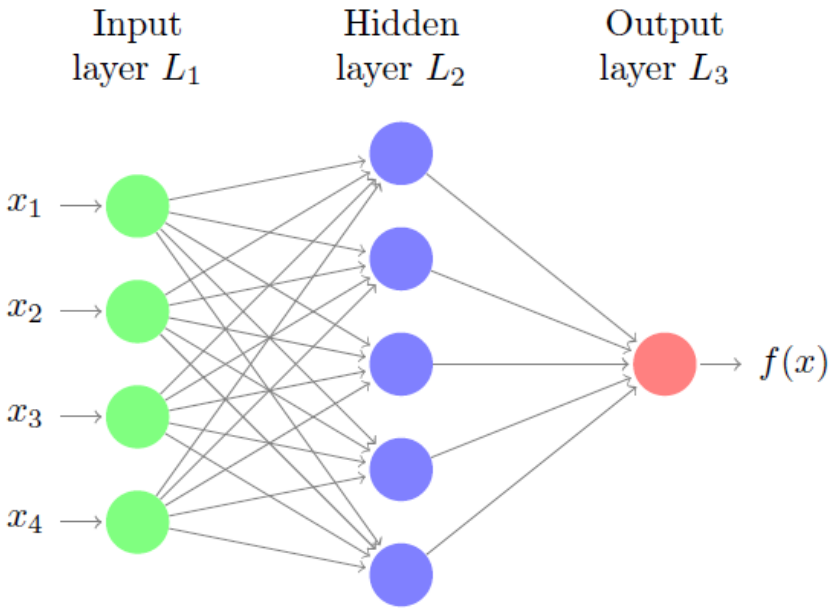

Figure C.7: Feed-forward neural network

Source: Efron and Hastie [2016]

Figure C.7 shows a feed-forward neural network with one input layer, one hidden layer and an output unit. The input layer has five predictors or inputs x_j and the hidden layer has five hidden units $a_l = g\big(w_{l0}^{(1)} + \sum_{j=1}^{4} w_{lj}^{(1)} x_j\big)$. The output unit can be represented as $o = h\big(w_0^2 + \sum_{l=1}^{5} w_l^{(2)} a_l\big)$. Through supervised learning, the nodes or memory units or neurons learn new features from the data. A vector of weights $w_{lj}^{(1)}$ (with (1) referring to the first layer and lj referring to the jth variable and lth unit) connect each node a_l in

the hidden layer to the input layer. There are also intercept terms $w_{l0}^{(1)}$ known as bias parameters. The function g is a nonlinear activation function, for example as the Sigmoid function $g(t) = 1/(1 + e^{-t})$. The output layer also has weights and an output function h. For regression problems, h is typically the identity function. Total number of parameters in this neural network $(4 + 1) \times 5 + 6 = 31$. To estimate these parameters or weights, neural network is trained on the data so that the model fits the data well. As a measure of fit, the sum-of-squared errors is used.

$$R(\theta) = \sum_{i=1}^{N} (y_i - f(x_i))^2 \tag{C.12}$$

where θ is the set of parameters.

To minimize the loss function, stochastic gradient descent (SGD) method is used. It is also called *back propagation* in this context. To prevent an overfitted solution, the loss function is optimized with various regularization methods, including adding penalty terms, using dropout layers, and implementing early stopping. For theoretical details we refer the readers to Efron and Hastie [2016] and Hastie et al. [2001].

In our implementation, for out-of-sample predictions, we borrow the structure and estimation procedure of Bianchi et al. [2021], where we estimate 100 neural networks at the same time, choose the best 10 according to the MSE of the validation sample, and use the average prediction of the 10 neural networks to be our final point forecast. This is aimed at reducing the randomness produced in initiating the neural networks and in the SGD procedure. In terms of the structure, macro data are passed through the hidden layers while term structure variables are added in before the output layer. This corresponds to the "hybrid" framework in Bianchi et al. [2021], which combines the ideas in Ludvigson and Ng [2009] and Cochrane and Piazzesi [2005]. Between layers, there are batch normalization (BN) layers and dropout layers. BN layers normalizes the input from the previous neuron, which can vary drastically

between each step of SGD, to a stable distribution, stabilizing and speeding up the estimation of the neural network. Dropout layers randomly deactivates (drops) neurons during each estimation step, introducing additional randomness and thus preventing the network from overfitting. When we leave out the dropout layers, our neural networks dramatically overfit the data. In some cases OOS R^2 decreases by over 100 percentage points.