OULUN YLIOPISTO
UNIVERSITY of OULU

OULU BUSINESS SCHOOL

Markku Kurtti

HOW MANY STOCKS MAKE A DIVERSIFIED PORTFOLIO IN A CONTINUOUS-TIME WORLD?

Master's Thesis

Department of Economics, Accounting and Finance

November 2020

UNIVERSITY OF OULU
Oulu Business School

ABSTRACT OF THE MASTER'S THESIS

Unit			
Department of Economics, Accounting and Finance			
Author Markku Kurtti		Supervisor Petri Sahlström	
Title How many stocks make a diversified portfolio in a continuous-time world?			
Subject Finance	Type of the degree Master of Science	Time of publication Nov 2020	Number of pages 233

Abstract

This thesis aims to answer how many stocks make a diversified portfolio in a continuous-time world. The study investigates what are the factors determining diversification effects in a real, continuous-time, world as opposed to thoroughly studied theoretical single period world. Continuous-time world investors care about geometric, instead of arithmetic, rate of return.

We show how methodology based on information theory can be utilized in investing context. Geometric risk premium is explained by the Shannon limit and its derivative, fractional Kelly criterion. Investing world counterpart for the Shannon limit, compounding process capacity, is derived. Geometric risk premium is decomposed to single stock risk premium and diversification premium. Method for estimating diversification premium is provided. Concept of realizable risk premium is derived and used in risk averse investor diversification metrics. Diversification effect is measured as a (realizable) risk premium ratio and as a (realizable) gross compound excess wealth ratio. Both ratios are between a randomly selected portfolio of selected size and fully diversified benchmark.

We show, both analytically and empirically, that diversification in a continuous-time world is a negative price lunch as opposed to free lunch in a single period world. Investor is paid a diversification premium, implying higher geometric risk premium, for consuming a lunch. The magnitude of diversification premium difference to benchmark, the opportunity cost of foregone diversification, is shown to be equal to one half of portfolio's idiosyncratic variance scaled by squared investment fraction. To maintain a constant wealth ratio, required level of diversification for a long-term risk neutral investor is approximately directly proportional to investment time horizon length.

The factors determining required level of diversification in a continuous-time world are number of stocks in the benchmark, Sharpe ratio and variance of the benchmark, idiosyncratic variance of an average stock, investment fraction and time. At investment fraction 1.0, risk averse investor requires more than 100, 200 or 1000 stocks to achieve 90%, 95% or 99% of the maximum diversification benefit, respectively. For short-term risk neutral investor, the corresponding numbers are about 20, 40 or 200 stocks and yet significantly more for long-term risk neutral investor. The numbers increase and decrease as investment fraction increase and decrease, respectively. We find that small firms require substantially more diversification compared to large firms and that there are substantial and consistent differences in diversification premiums between investing styles.

Keywords
diversification, risk premium, geometric return, continuous-time, information theory, Shannon limit, Kelly criterion, ergodicity problem

Additional information

ACKNOWLEDGEMENTS

I would like to thank Professor Petri Sahlström for supervising the thesis. Completing the thesis required strong focus and long hours. I would like to thank my family for bearing with my potential absent-mindedness during the process. Special thanks belong to Jukka Vikstedt, whose numerous and valuable comments and suggestions helped to improve the structure, focus and overall quality of this thesis.

CONTENTS

FIGURES

TABLES

1 INTRODUCTION

1.1 Background, motivation and aims of the thesis

“Diversification is the only free lunch in investing”, “leverage is dangerous” and “stocks for the long-run”. These are familiar commentary to most investors, but are they true and how can we quantify them? In this thesis we develop analytical methodology and empirical tests to assess and quantify these statements from diversification point of view. We discover the price for diversification in a continuous-time world and quantify how the need for diversification is affected by leverage, or more generally the fraction of portfolio allocated to stocks, and by the length of investment time horizon.

Peters (2019) describes using expectation value of wealth interchangeably with long-term time average of wealth as the ergodicity problem in economics. Peters explains that for ergodic observables their expectation value equals their infinitely long-run time average. From the point of view of an individual interested in the growth rate of his wealth, Peters argues, wealth is not ergodic meaning accumulated wealth over time averaged over all possible realizations (the expectation value of wealth) is different from long-run time average of accumulated wealth realized to a typical individual. Peters further argues that, in the context of expected utility theory, assuming ergodicity for wealth requires treating wealth as an aggregate over parallel ourselves.

Similarly, we consider expected arithmetic returns as not a viable option to project individual’s financial wealth in the future. The fact that Bill Gates is a billionaire will increase the arithmetic mean wealth of a population tremendously, but does not directly affect the financial position of a typical individual. Mr. Gates, like all of us, lives his life and consumes his private assets, which are not accessible to typical individual. We do not experience our lives as an aggregate over parallel ourselves (we only have one life) or as an aggregate over all people (we live our individual lives). Therefore, it is the expected geometric, instead of arithmetic, returns that describe how a typical investor experience the growth of his wealth over time. In practice, expected geometric return project median wealth instead of mean wealth.

The ergodicity problem and the distinction between arithmetic and geometric mean returns are at the heart of this thesis. In agreement with Peters (2019), we consider the life of an individual as undiversifiable and irreversible, rendering the use of the expectation value (arithmetic mean) for wealth as not feasible. Unlike the life of an individual, stock holdings are easily diversifiable. Diversification is one of the most fundamental and important concepts in finance. Consequently, there is a fair amount of prior research on the topic and some of the results from the early research are deeply rooted both in the financial literature and in the knowledge base of the practitioners. Especially this is the case with the conventional wisdom that *it takes no more than about ten stocks to make a sufficiently diversified portfolio*.

Importantly, the early literature and the results on diversification are mainly based on single period model comprising of expected arithmetic returns and their volatilities. Single period model by construction ignores the effect of time. Assumption of single period is a major simplification and, as we will show, has led to results which are not the most relevant to an investor who lives in a continuous-time world and aims to accumulate his investment wealth over time. Such investor, as much of the whole investment practitioner community to our knowledge, is more interested in expected geometric returns and volatilities. We will show that expected geometric returns, unlike expected arithmetic returns, are a function of level of diversification. This means diversification in a continuous-time world both reduce risk and increase expected (geometric) return, which in turn implies diversification is beneficial not only for risk averse but also for risk neutral investor, who only cares about the expected reward. Throughout the thesis, growth rates are expressed as instantaneous meaning continuously compounded rates unless otherwise stated.

The aim of the thesis is to explore the diversification effects from the real investor point of view who lives in a continuous-time world with an aim to accumulate his wealth through time in contrast to academic diversification effect important for an imaginary investor who is interested in maximizing his expected utility in a mean-variance single period world. Our practical investor reinvests her returns to fully benefit from the long-term effects of compounding.

The main research question is: *How many stocks make a diversified portfolio in a continuous-time world?* In a single period world this question is thoroughly researched using mean-variance optimization. More recent studies (see e.g. Domian, Louton & Racine, 2007; Bessembinder, 2018; Tidmore, Kinniry, Renzi-Ricci & Cilla, 2019), however, have considered the effect of diversification to long-term realized returns, which implicitly is about incorporating the effects of time and geometric returns into the analysis. In our study, we assess the diversification effect in a continuous-time world systematically by first building a theoretical framework and then by testing the predictions from that framework against empirical data. In addition to the main research question, these specific research questions rise in continuous-time framework: *How diversification effect is affected by fraction of capital allocated to stocks (asset allocation)? How diversification effect is affected by time (investment horizon)? How diversification effect is affected by rebalancing frequency? How diversification effect is affected by investing style?*

1.2 Methodology and data

The methodology used in our study is based on information theory. The work of Claude Shannon, John Kelly and Edward O. Thorp forms the basis on which we build our theoretical framework to assess diversification effect in a continuous-time world. Shannon is the inventor and the father of information theory, Kelly extended the applicability of Shannon's theory to gambling and Thorp is a mathematician, known as the father of quantitative investing, who further extended and applied Kelly's results to investing.

We demonstrate different methods for deriving the formula for geometric risk premium and show how the maximum of the geometric risk premium relates to Shannon's (1948) famous concept: information channel capacity, the Shannon limit. We associate the compounding process with information channel and call the investing world Shannon limit as compounding process capacity. Furthermore, we show, by drawing an analogy with digital communications, that utilizing a single period model implicitly corresponds to assessing a data transfer based on transmitted, instead of received, data rate. The square of the Sharpe ratio is shown to correspond to signal to noise ratio (SNR) in Shannon's channel capacity equation. SNR can be thought to be the single most important measure in digital communications, which explains the central role of Sharpe ratio in the investing word.

Fractional Kelly criterion and its applications to investing context, as defined by Thorp (2006), form the grounds from where we derive our methodology and equations to assess diversification effects in the continuous-time world. Kelly criterion is a derivative of the Shannon limit and states the investment fraction (fraction of investment capital allocated to stocks) at which the Shannon limit (maximum geometric risk premium) is achieved. Furthermore, by using fractional Kelly criterion, we can determine the geometric risk premium at any given investment fraction. The fact that geometric risk premium is a function of level of diversification makes fractional Kelly criterion a perfect tool to assess diversification effect.

Risk premium (throughout the thesis, risk premium refers to geometric risk premium unless otherwise stated) is decomposed to sum of a single stock risk premium and diversification premium. Diversification premium is the risk premium difference between a portfolio of selected size and a single stock portfolio. Linear regression-based method is provided for estimating an average idiosyncratic variance for single stock portfolio from data comprising of exhaustive monthly risk premiums for individual stocks. Very large data set together with its exhaustiveness imply very high precision estimate. Diversification premium for a benchmark portfolio is derived utilizing the estimated average idiosyncratic variance for a single stock portfolio. We further provide methodology to calculate diversification premium for any portfolio size and define diversification premium difference to benchmark as the difference in diversification premiums between a portfolio of selected size and fully diversified benchmark portfolio.

Risk premium is an expectation for infinitely long investment horizon. Human life and realistic investment horizons, on the contrary, are measured in years and decades. To account for the risks associated with limited life expectancies and investment time horizons, we derive a concept called realizable risk premium. Realizable risk premium is the proportion of the risk premium explained by the (expected) risk premium. The remaining proportion is explained by noise (risk). The key property of realizable risk premium metric is that it accounts for both dimensions (risk and reward), but expresses its value in a single (reward) dimension. At infinity, realizable risk premium equals risk premium, but in the meantime, realizable risk premium is an increasing function of investment time horizon length. In the short-term, we show that realizable risk premium is approximated by SNR weighted risk premium. Replacing risk premium with realizable risk premium in our diversification metrics change the investor type from risk neutral to risk averse.

Risk premium describe the growth rate as an average over time and is the average direction of the change in wealth at any given moment. We therefore consider risk premium and realizable risk premium as short-term metrics. For a long-term investor, it is the wealth at the end of investment horizon that matters the most. We therefore consider the gross compound excess wealth (wealth including initial investment in excess of what investment in riskless rate would have produced) as the relevant long-term metric. Gross compound excess wealth or realizable gross compound excess wealth is the wealth accumulated by compounding with risk premium or realizable risk premium as growth rate for risk neutral and risk averse investors, respectively. Using realizable in parenthesis (realizable) before the diversification metric means that we use the word “realizable” in case of risk averse investor metrics, but leave it out from risk neutral investor metrics.

Three diversification metrics for the continuous-time world are defined: 1) number of stocks required to achieve a positive risk premium, 2) number of stocks required to achieve a proportion (e.g. 90%) of the (fully diversified) benchmark risk premium and 3) number of stocks required to achieve a proportion of the benchmark gross compound excess wealth over time. The first metric is the absolute minimum level of required diversification, the second is a short-term diversification metric and the third is a function of investment time horizon and is used as a long-term diversification metric. Second and third metric are further divided to risk neutral and risk averse investor metrics. Risk premium and realizable risk premium are utilized in the diversification metric equations for risk neutral and risk averse investor, respectively.

Our theoretical framework predicts risk premium, diversification premium difference to benchmark and the three defined diversification metrics for any portfolio size. What is remarkable about the predictions is that after deciding the investment fraction and time horizon, only four empirical parameters are required: average monthly number of stocks in the benchmark portfolio, instantaneous Sharpe ratio of the benchmark portfolio, variance of the benchmark portfolio and the average idiosyncratic variance of a single stock portfolio. The developed linear regression-based method is used to estimate the last parameter and the three former parameters are calculated from the benchmark portfolio's time series data.

We test the predictions from the developed theoretical framework against empirical data from the U.S. stock market. Empirical data consists of monthly CRSP stock returns for all common stocks combined with one-month Treasury bill data from July 1926 to June 2018. In addition, we combine annual accounting-based data for U.S. stock market from Compustat with CRSP and Treasury bill data for the period from August 1962 to June 2018. The primary period used in empirical tests is the 45.5-year time span from January 1973 to June 2018, which includes Nasdaq and contains all of the stocks listed in major U.S. stock exchanges. Monthly bootstrapping without replacement is used to create equally weighted portfolios of selected size randomly.

We use in-sample test to investigate how well the theoretical predictions explain the empirical diversification effect. However, as we want to project the results into the future, we show both analytically and using log-normal simulator that forward-looking, out-of-sample, diversification premium (and therefore the required level of diversification in general) is expected to be higher than the historical value. For risk premium, all else equal, the forward-looking risk premium is expected to be lower compared to historical realized risk premium. The reason is that not only the risk (standard deviation), but also the uncertainty about risk (the standard deviation of the standard deviation estimate) affect geometric metrics. Furthermore, we find that our in-sample predictions systematically slightly underestimate empirical diversification premium. We hypothesize the underestimation is attributable to fat-tailedness of the return distribution combined with less than theoretically optimal (less than infinite) monthly rebalancing frequency. Accounting for both the underestimation in the in-sample tests and the uncertainty entailed in the future risk parameters, we consider the required level of diversification estimated by our theoretical framework as the lower bound.

1.3 Results

We test the in-sample predicted diversification metrics against bootstrapped empirical results. Based on the empirical tests the theoretical framework is able to predict the selected diversification effects and metrics very accurately. The slight systematic underestimation of the level of required diversification, which is present with all the metrics, disappears when we cut one percent from both tails of each month's single stock risk premium data. With the fat tails cut, the predictions become extremely accurate. This supports our hypothesis that the underestimation of diversification premium is caused by fat-tailed risk premium distribution combined with lower than infinite rebalancing frequency and further supports the view that our predictions for required level of diversification should be taken as lower bounds.

The most important implication from using geometric, instead of arithmetic, risk premium as the basis for assessing continuous-time world diversification effects is reflected by the price for diversification. We show both analytically and empirically, with very high statistical significance, that diversification in a continuous-time world is a negative price lunch as opposed to free lunch in a single period world. More diversification imply higher risk premium meaning investor is paid a diversification premium for consuming a lunch.

How diversification effect is affected by fraction of capital allocated to stocks (asset allocation)? The cost of foregone diversification is the magnitude of diversification premium difference to benchmark, which is shown to be equal to one half of portfolio's idiosyncratic variance scaled by squared investment fraction. The fact that investment fraction is squared means that for leveraged portfolios diversification is extremely important while for portfolios with low stock allocation the effect from diversification is smaller. At investment fraction one, the risk premium for a randomly picked equally weighted 10-stock portfolio in the 45.5-year period from January 1973 to June 2018 was about 1.6 percentage points less than for a fully diversified benchmark portfolio. The corresponding number was 0.6 and 3.6 percentage points for investment fractions 0.6 and 1.5, respectively. 1.6 percentage points corresponds to a cost of an expensive active mutual fund. In this light, ten stocks are hardly enough to make a diversified portfolio.

The concept of diversification premium difference to benchmark explains the difference in empirical expected geometric returns between a portfolio of selected size and a fully diversified benchmark in the Tidmore et al. (2019) study.

How diversification effect is affected by time (investment horizon)? The aspect of investment time horizon length is reflected in our third, the long-term, diversification metric, the (realizable) gross compound excess wealth ratio. Remarkably, the required number of stocks to keep the gross excess wealth ratio constant for a risk neutral investor (e.g. at 0.9 implying 90% of the benchmark gross compound excess wealth) over time is approximately directly proportional to time horizon length. This means that the required number of stocks in the portfolio doubles when targeted investment time horizon length doubles. This is exactly opposite to risk premium based metrics (the conventional way to assess the effect of time) as the uncertainty related to risk premium decrease (and consequently realizable risk premium increase) as a function of time. However, wealth-based metrics, unlike risk premium-based metrics, account for the compounding effect. The diversification premium difference to benchmark compounds over time, hence the requirement for increased diversification as a function of investment time horizon length.

How diversification effect is affected by rebalancing frequency? We find that while rebalancing is required to maintain the targeted equal weighting in the portfolio, which is theoretically the weighting that leads to highest risk premium, our theoretical predictions on the required level of diversification are not very sensitive to rebalancing frequency. Consequently, even though our theoretical framework assumes infinite rebalancing frequency, the predictions work well against monthly rebalanced empirical portfolios. However, when portfolios are leveraged aggressively, say beyond investment fraction 1.5, we start to see signs that monthly rebalancing frequency is too low as the risk premium declines more rapidly as a function of investment fraction than our theoretical framework predicts.

How diversification effect is affected by investing style? Interestingly from a stock picker's point of view, we find substantial differences in the level of required diversification as a function of investment style. Our results are in line with the well-known fact that returns for small stocks are more volatile compared to big stocks. We find that big stocks require less diversification than small stocks and the difference is especially large compared to microcap stocks. What is less well known, is that there are substantial differences in required level of diversification outside the size factor. We find that investing in high ROE, high momentum, value, or especially high earnings yield or high earnings yield combined with high momentum style requires substantially lower level of diversification compared to equally weighted market portfolio and especially compared to opposite investing styles. As an example, diversification premium for a benchmark portfolio for high earnings yield style is about 8.1 percentage points while it is about 25.6 percentage points for low earnings yield style. This implies that a 10-stock high earnings yield portfolio lose approximately 0.81 percentage points in risk premium to its benchmark, fully diversified high earning yield portfolio, while corresponding number for low earnings yield portfolio is approximately 2.56 percentage points. Furthermore, the styles (except for the size factor) that have historically required less diversification have also provided higher risk premium. Noteworthy, diversification premium differences between opposite investing styles, unlike risk premium differences, have historically been very consistent over time.

We often hear that “diversification fails when it is the most needed” referring to bear markets. We find that the opposite is true: diversification premium is on average higher at 19.4 percentage points during bear markets compared to 13.3 percentage points during bull markets. The key is to measure relative diversification benefit by comparing diversification premiums between bear and bull markets instead of comparing absolute returns between bull and bear markets. If diversification were able to protect our portfolios from declining returns in bear markets, there would be no such thing as systematic risk and consequently no risk premium.

Comparing continuous-time world diversification effects to single period world, the most striking differences are that in a continuous-time world diversification benefit is an increasing function of investment fraction and time horizon length and that not only risk averse, but also risk neutral investor enjoys from the benefits of diversification. For long-term risk neutral investor, the effect of investment fraction to required level of diversification is squared and the effect of time horizon length is approximately directly proportional to required number of stocks. For long-term risk averse investor, the effect of investment fraction is increasing but milder than squared and increasing the time horizon length first decreases and then (typically after 10 to 20 years) increases the required level of diversification. For short-term risk neutral investor, the effect of investment fraction to required level of diversification is increasing but milder than squared. For short-term risk averse investor, the effect of investment fraction is increasing, but effect is the mildest among all metrics. For the absolute minimum level diversification metric, which ensures positive risk premium, the effect of investment fraction to required level of diversification is increasing but milder than squared. Changing target ratio for risk premium ratio and gross excess wealth ratio metrics approximately double and tenfold the required level of diversification as target ratio change from 90% of maximum diversification benefit to 95% and 99%, respectively.

Even if an investor targets a long-term investment horizon, the long-term is a series of consecutive short-term periods. Therefore, risk averse long-term investor is affected by short-term risks and cannot ignore the short-term required level of diversification. For risk averse investor, the short-term metric is typically dominant and calls for high diversification. For a long-term investor with very high tolerance for short-term risks targeting a liability matching portfolio (e.g., an investor saving for retirement who is not risk neutral in the long-term as he has liabilities to cover), we consider the long-term risk averse metric as the most relevant. For risk neutral investor, we consider the required level of diversification as the maximum between short-term and long-term risk neutral metrics.

How many stocks make a diversified portfolio in a continuous-time world? The answer to this question is multifaceted. The required number of stocks given in following paragraphs are given by our theoretical framework based on empirical parameters from the period from January 1973 to June 2018. Average monthly number of stocks for the period is 5472. The numbers predicted by the theoretical framework are lower bounds and (historical) empirical values are typically about 15% higher. Additionally, when projecting the numbers far into the future, the uncertainty about the future level of average idiosyncratic variance implies the numbers should be slightly increased.

The absolute minimum required level of diversification ensuring positive risk premium is 2, 2, or 4 stocks (the exact average numbers are 1.03, 1.91 and 3.31) for investment fractions 0.6, 1.0 and 1.5, respectively. Below these levels of diversification, investor is expected to earn a higher geometric mean return by investing in riskless rate.

At investment fraction 1.0, short-term risk neutral investor requires about 19, 38 or 184 stocks to achieve 90%, 95% or 99% of the maximum diversification benefit, respectively. Corresponding numbers for short-term risk averse investor are 118, 238 or 1033 stocks. At investment fraction 0.6, the corresponding numbers are 10, 21 or 101 & 93, 190 or 854 stocks, while at investment fraction 1.5, the requirement is 33, 65 or 312 & 156, 312 or 1293 stocks.

At investment fraction 1.0, long-term risk neutral investor (at 20-year horizon) requires about 28, 57 or 279 stocks to achieve 90%, 95% or 99% of the maximum diversification benefit, respectively. Corresponding numbers for long-term risk averse investor are 55, 111 or 520 stocks. At investment fraction 0.6, the corresponding numbers are 10, 21 or 104 & 30, 60 or 291 stocks, while at investment fraction 1.5, the requirement is 62, 127 or 590 & 104, 208 or 912 stocks. At 40- and 60-year horizons, long-term risk neutral investor figures double and triple, respectively. Long-term risk neutral investor figures asymptotically approach the risk averse investor figures.

Furthermore, firm size makes a difference. For short-term risk neutral investor at investment fraction 1.0, about 8, 13 or 23 stocks are required to achieve 90% of the maximum diversification benefit for big, small and microcap stocks, respectively. The corresponding numbers for short-term risk averse investor are 54, 70 and 132 stocks. For long-term investor (at 20-year horizon), the corresponding numbers are 10, 17 or 36 & 24, 37 or 66 stocks.

Investment style affects the required level of diversification. For example, for short-term risk neutral investor at investment fraction 1.0, about 6 or 58 stocks are required to achieve 90% of the maximum diversification benefit for high earnings yield and low earnings yield styles, respectively. The corresponding numbers for short-term risk averse investor are 62 and 204 stocks. For long-term investor (at 20-year horizon), the corresponding numbers are 15 or 47 & 20 or 154 stocks.

1.4 Roadmap to thesis

This thesis proceeds as follows. In chapter 2 we give an overview of the diversification literature. In chapter 3 we derive necessary equations and build the theoretical framework to assess diversification effects in a continuous-time world. More specifically, in section 3.1 we further discuss the difference between single period and continuous-time worlds and the definition of risk premium. In section 3.2 we derive the formula for instantaneous geometric risk premium. In section 3.3, based on the risk premium formula derived in section 3.2, we define and derive the important concept of diversification premium and further develop the ideas originating from information theory to serve as the theoretical basis for this thesis. Diversification metrics for continuous-time world are determined in section 3.4. Section 3.5 defines the hypotheses about empirical diversification effects. In chapter 4 we describe the empirical data and the used methodology including the use of log-normal simulator. In chapter 5, we employ our theoretical framework to produce predictions for selected diversification metrics and compare predicted results to results bootstrapped from empirical data. In section 5.1 we show that fractional Kelly criterion accurately predicts the empirical risk premium. In section 5.2 we confirm statistically that diversification is a negative price lunch and in section 5.3 we show that the empirical opportunity cost of foregone diversification corresponds to our prediction from the theoretical framework. In sections from 5.4 to 5.6 we further illustrate the idea behind using (realizable) risk premium and (realizable) gross compound excess wealth as diversification measures. Furthermore, we show that empirical risk premium ratio and gross compound excess wealth ratio behave as expected based on our theoretical framework for both risk neutral and risk averse investors. In section 5.7 we present the number of stocks required to make a diversified portfolio in a continuous-time world both using figures and numerically in tables. In section 5.8 the empirical effects of firm size and investing style to required level of diversification are tested and shown. The consistency of the historical diversification premium over time is demonstrated in section 5.9. Finally, in chapter 6 we discuss the findings and summarize the thesis.

2 OVERVIEW OF DIVERSIFICATION LITERATURE

2.1 Foundations for diversification

Foundations for diversification from theoretical point of view were set up by Harry Markowitz, who in his seminal paper (Markowitz, 1952) argued that expected return alone is not sufficient for deciding the asset weights in a portfolio. Instead, Markowitz defined a framework where the variables of interest for diversification were expected return of each asset, expected variance of each asset and expected correlation between risky assets in the portfolio. The key of the theory is that while the portfolio expected return is simply the weighted average of the expected returns of the assets, portfolio expected variance is dependent not only on the expected variance of the assets, but also on the correlations between these assets. This implies that there is a benefit from diversification as the expected variance can be less than the weighted sum of the expected asset variances depending on correlation matrix of the portfolio. Markowitz showed that mean-variance optimal portfolio of risky assets lies in the efficient frontier where each point maximizes expected return for given expected variance or alternatively minimizes expected variance for given expected return.

Tobin (1958) introduced the idea that risk/return profile of a portfolio can be selected by combining a portfolio of risky assets with single risk-free asset. Sharpe (1964) defined capital market line (CML) which combines a portfolio of all available risky assets with risk-free interest rate. All efficient portfolios lie on CML and desired risk/return profile can be selected along the line.

From theoretical point of view the question on how much diversification is needed is trivial. As efficient portfolio is constructed by combining a portfolio of all risky assets and one risk-free asset, it follows that the right amount of stocks in a diversified portfolio of stocks is all stocks available. However, if we, e.g., consider the cost of diversification or possible desire to deviate from market return, including all stocks may not be feasible or desirable. This immediately justifies the consideration on the amount of stocks that makes a sufficiently diversified portfolio in a sense that, e.g., cost is sensible or that it is possible to differentiate from market return with tolerable risk.

It is assumed that expected return is not affected by the amount of diversification in the portfolio. Risk on the other hand is the subject of minimization in diversification no matter how risk is defined exactly. The exact definition of risk may however affect the required level of diversification and is therefore central in the literature discussing diversification. Additionally, number of studies consider the cost of diversification together with the benefits to find the equilibrium for diversification.

2.2 Empirical evidence for sufficient diversification

Evans and Archer (1968) pointed out that the optimal portfolio selection where unsystematic variance is completely eliminated as defined by Sharpe (1964) is not reasonable if the cost of diversification is a function of number of stocks of different companies added into portfolio. They put the comparison of marginal cost to marginal benefit from increased diversification into focus. Evans and Archer used 470 securities from Standards and Poor's index from the beginning of 1958 to mid 1967 and observed the data semiannually. They formed 40 portfolios using random selection among the 470 securities with different sizes ranging from 1 to 40 securities. 60 portfolios of each size were formed. From this they calculated the average reduction in standard deviation and the reduction in the mean standard deviation dispersion resulting from adding one security into portfolio for each examined portfolio size. t-tests were used to examine the impact to mean standard deviation and F-tests to test the impact to dispersion in the mean standard deviation when portfolio size is increased.

The main finding of the Evans and Archer (1968) study is that increasing the portfolio size above 8 securities require significant increase in the number of securities to yield a statistically significant reduction in portfolio mean standard deviation or in the mean standard deviation dispersion. They observe the same visually by plotting the expected standard deviation as a function of number of securities in the portfolio. Evans and Archer conclude by stating that considering the costs of diversification, the benefits of increasing the portfolio size beyond ten or so securities are questionable.

The data of Fisher and Lorie (1970) included also companies which were not listed throughout the whole period and thus they avoided the survivorship bias present in earlier studies like Evans and Archer (1968). They used 40 years of data from 1926 to

1965 with different non-overlapping time periods from one year to 40 years in length. In addition to random sampling they also used random sampling controlling for industry groups so that equal number of stocks were selected from each of the 34 industry groups. Fisher and Lorie used three measurements for relative return dispersion: coefficient of variation, relative mean deviation and Gini's coefficient of concentration. They used the universe of stocks available in NYSE and formed random portfolios consisting of 1, 2, 8, 16, 32, 128 and all stocks.

The main result from the Fisher and Lorie (1970) study is that the diversification effect is greatly diminished after 8 stocks in the portfolio. Roughly 80% of the diversification potential is exhausted at portfolio size of 8 stocks, 90% at size 16 stocks, 95 at size 32 stocks and 99% at size 128 stocks. Also, they note that industry diversification does not change the results significantly.

Elton and Gruber (1977) developed an analytical parameter-based model for estimating mean variance. They used parameters obtained from earlier empirical studies. Elton and Gruber defined a total risk measure which consists of portfolio variance plus portfolio mean return difference to market mean return. They recognize that risk is not only the variation of the portfolio, but also the risk of deviating from market return as this deviation is a measure of portfolio mean return uncertainty. Like earlier studies, Elton and Gruber find that most of the diversification gain is realized with a low number of stocks in the portfolio. However, they also find that there still is significant diversification opportunity. For example, portfolio of 15 stocks has 32% more risk than a portfolio of 100 stocks when measured as total risk.

Statman (1987) used Elton and Gruber (1977) model to estimate the effect of number of stocks in the portfolio to diversification gain. Statman determines diversification worthwhile as long as marginal benefit exceeds marginal cost. His innovation is to use return as a common measure for diversification benefit and cost. Statman defines a model where market return (SP500 index fund used as a proxy) is levered until volatility match the volatility of the portfolio size of interest. As the expected return of the market proxy and randomly selected portfolio is the same, the return difference between levered market portfolio and portfolio of interest is dependent on volatility difference between the portfolios, the excess of the borrowing rate over the risk-free rate and estimated equity premium. Statman used 2% estimate for the borrowing rate excess over the risk-free rate and historical value 8.2% for equity premium.

Statman (1987) finds that the return difference, e.g., between a stock portfolio of 10 stocks and levered SP500 index fund is 1.502% (1.986% for lending investor who can operate without leverage). He assumes no cost for stock portfolio and takes the 0.49% cost for the SP500 index fund as the cost for further diversification. Statman finds that for borrowing investor the return benefit is greater than the diversification cost still at the size of 30 stocks and for lending investor at the size of 40 stocks. Based on these results, at least 30 stocks are required to make a diversified portfolio.

Newbould and Poon (1993) found that typical text book recommended 8 to 20 stocks to be enough to eliminate diversifiable risk. They address origins of this recommendation to studies where average standard deviation for a portfolio of different sizes are considered. Newbould and Poon argue that for a risk averse investor who holds one portfolio, not average of many portfolios, it is important not only to assess the average standard deviation, but also the standard deviation in the standard deviation. By using this method and S&P500 stock universe from three-year period 1988-1990, they find confidence intervals for average standard deviations. For an 8-stock portfolio the 99% and 95% upper confidence limits for standardized average standard deviation of 100 are 141 and 131 respectively. For a 20-stock portfolio the corresponding confidence limits are 130 and 123. They find that the confidence limits for standard deviation converge much slower than the marginal diversification benefit diminishes when the number of stocks in the portfolio is increased. Newbould and Poon conclude that the minimum number of stocks needed to eliminate the diversifiable risk significantly exceeds 20.

Average portfolio return standard deviation or variance is the risk in the theoretical models, but also other risk measures have been studied in the literature concerning diversification. According to Surz and Price (2000), reduction in average standard deviation is not the best measure for the magnitude of diversification. Instead, they argue that increase in r-squared (proportion of portfolio return variance explained by market returns) and reduction in tracking error are better measures. Surz and Price argue that reduction in standard deviation measures reduction in total risk (systematic plus idiosyncratic risk) whereas increase in r-squared or reduction in tracking error imply reduction in idiosyncratic (uncompensated) risk. They used the Compustat stock universe from 1986 to 1999 to replicate Fisher and Lorie (1970) study. They added r-squared and tracking error alongside standard deviation as metrics to measure the diversification effect as a function of number of stocks in the portfolio.

Surz and Price (2000) find that r-squared and tracking error require significantly more stocks in the portfolio to exhaust the marginal diversification potential. When a 15-stock portfolio accounts for 93% of the achievable standard deviation reduction, it only accounts for 79% of the achievable diversification benefit when measured as an average between r-squared and tracking error metrics. For a 60-stock portfolio, the numbers are 98% and 88% respectively. Surz and Price also point out that if we accept reduction in standard deviation as the measure of diversification gain, then we must accept that it is possible to diversify the portfolio risk below the level of market risk with very low number of stocks in the portfolio as already some of the simulated 15-stock portfolios have lower standard deviation than market portfolio. This is not possible with any portfolio size if we use r-squared or tracing error to measure diversification gain. Utilizing the diversification metrics recommended by Surz and Price, portfolio size greater than 60 stocks is required to exhaust about 90% of the diversification potential that is exhausted with less than 15 stocks when measured by reduction in average standard deviation.

Many of the seminal studies (Elton & Gruber, 1977; Evans & Archer, 1968; Fisher & Lorie, 1970) on the effect of diversification and number of stocks in the portfolio are from early days in 1960s and 70s after the theoretical framework for diversification was set up. Parameters impacting diversification effect and used in these studies are not time invariant. Campbell, Lettau, Malkiel and Xu (2001) address this in their study. Campbell et al. investigate firm, industry and market level volatility as a function of historical time periods using daily data from 1962 to 1997. In addition to volatilities, they also estimate average correlation between individual stock returns as a function of time periods.

Campbell et al. (2001) find that there has been no trend in market and industry level volatilities between 1961 and 1997, but there is clear upwards trend in firm level volatility. The reason why market or industry volatilities have not increased when firm level volatility has risen is that they also find that average correlation between individual stock returns has decreased. Average monthly correlation (measured from five years of data) between stocks has decreased from 0.28 in early 1960s to 0.08 in late 90s. Campbell et al. find that the same excess standard deviation for a portfolio compared to market portfolio that in the early 60s required 20 stocks in the portfolio, in the late 90s require 50 stocks. Due to increased firm level volatility and decreased average correlation between stocks, the required number of stocks in the portfolio went 2.5-fold in less than 40 years.

Statman (1987) had used the model and parameters obtained by Elton and Gruber (1977) in modeling the diversification effect and concluded that at least 30 stocks with the parameters available at that time were required to form a sufficiently diversified portfolio. However, given the study by Campbell et al. (2001) the parameters had changed and Statman (2004) reassessed the degree of diversification required based on this new information. In addition to increased firm level idiosyncratic volatility and decreased average correlation between stocks Statman updated other parameters too for the new study. New proxy for the market and full diversification was Vanguard total market index fund with annual cost of 0.20% compared to Vanguard SP500 index fund with 0.49% annual cost in the original study. New forward-looking estimate for equity premium was 3.44% compared to historical average or 8.2% in the original study. Additionally, Statman now assumed 0% excess cost for borrowing over the risk-free rate whereas he had used 2% in the 1987 study. Finally, Statman now assumed 0.14% annual cost for buying and holding a portfolio of individual stocks, compared to zero cost assumed in 1987. As a result, the net cost for further diversification became 0.20% - 0.14% = 0.06% compared to 0.49% - 0.0% = 0.49% in the original study. The same methodology as in the original study was used in the 2004 study. The main finding of the study is that, with the new parameters and assumptions, more than 300 stocks are needed for the marginal cost to exceed the marginal benefit of additional diversification.

All of the studies reviewed above, except Surz and Price (2000), have used mean-variance optimization as the framework in their assessment of diversification effect. Domian et al. (2007) however, take an alternative viewpoint by measuring shortfall risk and deciding required level of diversification based on this measure. Domian et al. define shortfall risk as the possibility of ending wealth being below a target ending wealth at the end of the investment period. They argue that shortfall risk is a useful risk measure especially for investors with long investment horizons. Domian et al. use a 20-year investment period from 1985 to 2004 and a universe of 1000 stocks with largest market capitalization in their simulations to form randomly picked stock portfolios of various sizes. In addition, they use another 20-year period from 1965 to 1984 as a robustness check. They measure the compound return for the portfolios over the 20-year period and compare this to risk-free 20-year treasury bond return from the same period. Shortfall is measured as a probability of a stock portfolio compound return being below the risk-free return from the period. In addition to random sampling, they also use an alternative sampling technique to ensure that each of the 10 selected industries get equal amount of stocks in the portfolio. Various sizes of portfolios from 10 to 200 stocks are simulated.

Domian et al. (2007) find that the number of stocks required to achieve significant diversification effect is substantially greater when assessing shortfall risk compared to previous studies focusing on mean-variance optimization. To reduce the probability of realizing a stock portfolio return below risk-free return over a 20-year period to 1%, 168 stocks are required in the portfolio. To achieve 5% and 10% shortfall risk, 93 and 63 stocks are required respectively. If only 10 stocks are included in the portfolio, the shortfall risk is 40.2%. Furthermore Domian et al. find that the magnitude of industry diversification effect, when present, can be matched by only slightly increasing the number of randomly selected stocks in the portfolio. E.g. for portfolio sizes 10 to 30 stocks, 10% increase in the number of randomly selected stocks delivers the same diversification effect as diversifying across industry sectors. For larger portfolio sizes, there is no diversification effect from industry diversification. Domian et al. conclude that more than 100 stocks are needed to make a diversified portfolio.

Bessembinder (2018), similarly as Domian et al. (2007), assess long horizon investing and shortfall risk. Bessembinder emphasize the fact that the cross-sectional return distribution of a portfolio of stocks resembles more lognormal than normal distribution. It follows that the cross section of individual stock returns is positively skewed which implies that median return is lower than mean return of the distribution.

The greater the skewness, the greater the difference between median and mean returns. This is especially true with longer time horizons as skewness in the lognormal distribution arises not only from the possible lognormal shape of the original one period return distribution, but from the effect of compounding consecutive time periods. Compounding induces skewness as skewness is a function of volatility and volatility increases as a function of square root of time. Bessembinder used monthly CRSP U.S. data from the period from 1926 to 2016 and calculated hypothetical realized portfolio returns assuming reinvested dividends. He compared the returns to monthly treasury bills (risk-free) return, value-weighted market return and nominal 0% return threshold from the corresponding time period. Bessembinder used a bootstrap simulation where he randomly selected a value weighted portfolio of stocks for each month and linked the monthly returns for time periods of one, ten and ninety years. He formed portfolios with sizes of 1, 5, 25, 50 and 100 stocks. By this method he simulated the effect of randomly selecting a portfolio of stocks for a selected investment horizon and measured the shortfall metrics after the investment period.

Bessembinder (2018) finds that the skewness risk is extreme at 90-year horizon, significant at 10-year horizon and exists already at one-year horizon. Skewness manifests itself in the shortfall metrics. At 90-year horizon, one stock strategy loses to risk-free return in 73% of simulations and to value weighted market benchmark in 96% of simulations. 49% of the simulations have negative return. Had Bessembinder used real returns, the result would be even more striking indicating more than half of the single stock strategy returns being negative. The results are similar but not as extreme when either number of stocks in the portfolio is increased or investment horizon is shortened. The effect of both is the same, i.e., volatility of the portfolio (which causes skewness) is decreased. With 100 stock portfolio and 10 years investment horizon 7% of simulations have lower return than risk-free return and 52.5% of simulated portfolios have lower return than market return. This is the effect of skewness risk as even with this high number of stocks in the portfolio, the portfolio return is likely to lose to market return while the expected value for the portfolio is equal to the expected value of the benchmark portfolio. 90-year period result gives some insight to diversification effect in the very long term. Comparison to market return reveals that the probability to have lower return than market increases monotonically when starting with 100 stock portfolio and ending with one stock portfolio. Even with 100 stock portfolio, market return is higher in 57% of simulations. On the other hand, already with 25 stock portfolios, and with all simulated portfolios with greater number of stocks, 100% of simulated portfolios have return higher than risk-free return. Bessembinder does not provide numerical value for sufficient number of stocks in the portfolio, but gives a theoretical explanation and comprehensive simulation results which indicate that, even with 100 stock portfolio, an investor is likely to have a realized return lower than market return and the probability of losing to market return increases as investment horizon increases. Empirical results are similar, if not even more dramatic, for global stocks in the period from January 1990 to December 2018 (Bessembinder, Chen, Choi & Wei, 2019).

Tidmore et al. (2019) utilize the universe of Russel 3000 Index from January 1987 to December 2017 to assess the effect of diversification. Their method is to create randomly selected, quarterly rebalanced, equally weighted portfolios of different size and compare the average realized geometric rate of return to fully diversified benchmark portfolio. This return difference is called expected excess return. The most interesting finding of Tidmore et al. is that the empirical (geometric) expected excess return for a portfolio is an increasing function of number of stocks in the portfolio and that the expected excess return is always negative (or zero for very widely diversified portfolios). A single stock portfolio is expected to lose 9.9 percentage points while for a ten-stock portfolio the difference to benchmark is diminished to 1.0 percentage points. In other words, increasing diversification on average increases portfolio's (geometric) expected excess return compared to fully diversified benchmark portfolio.

2.3 Empirical evidence for the average investor level of diversification

Literature debates about the number of stocks needed to make a sufficiently diversified portfolio and the evidence ranges from about ten to more than three hundred stocks. But what is the average or median level of diversification among retail investors?

Goetzmann and Kumar (2008) study equity portfolio diversification and provide a distribution of number of stocks held by over 60000 individual investors in the time span from 1991 to 1996. Data was acquired from a large U.S. discount brokerage house. The data does not distinguish between direct stock holdings and other assets possibly held by the investors meaning that on average the investors are better diversified than can be interpreted from the data used in the study. Nevertheless, the data gives us an idea how diversified the direct stock holding were during the period.

The striking finding of the Goetzmann and Kumar (2008) study is that during the six-year time span, about 28% of the investors had only one stock in their portfolio, about 75% of investor had five or fewer stocks and less than 10% had more than ten stocks in their portfolio. Empirical evidence supports investing in about ten stocks at minimum, but less than 10% of the investors in the study exceeded this minimum recommendation.

Findings of Polkovnichenko (2005) study are largely similar. Polkovnichenko utilize data from Survey of Consumer Finances carried out six times between 1983 and 2001. The survey provides data on U.S. households assets including direct and indirect equity holdings. Indirect holdings include diversified portfolios such as mutual funds. Our main interest is the results from direct equity holdings.

Polkovnichenko (2005) summarize the direct holdings providing a median number of stocks held. Over the time period covered by the surveys, of the direct investors about 80% held in maximum of five stocks and about 90% held less than ten stocks. About 40% held only one stock in their direct equity portfolio.

2.4 Summary of the literature

Conventional wisdom has been and still is that a low number of stocks – say ten – is sufficient to make a diversified portfolio. Early studies support this view, but a large number of studies over the years argue that substantially larger number – say more like a hundred stocks – is needed. The disagreement originates from the definition of diversification effect. Early studies focus on the fact that the marginal benefit of diversification measured as a decrease in average portfolio standard deviation when adding one stock to portfolio diminishes quickly after few stocks are in the portfolio. There is no disagreement on this. Also, there is an agreement among the studies that industry diversification does not provide significant diversification effect.

Several studies suggest different measures for diversification effect instead of just focusing on average portfolio standard deviation or variance. Measures like uncertainty of the expected portfolio return, standard deviation in standard deviation, i.e., the uncertainty of the expected portfolio standard deviation, marginal diversification benefit measured as a difference between levered market return and portfolio return, focusing on reduction in uncompensated risk, time varying nature of the parameters determining diversification effect, shortfall probabilities for long horizon investors and increasing skewness risk as a function of investment horizon all call for substantially higher degree of diversification than mere assessment of mean standard deviation of the portfolio.

Perhaps contrary to common belief, diversification can be deemed the more important the longer the investment horizon is due to skewness risk. If investment horizon is short and expected volatility of the portfolio is the risk that is to be diversified away, then the conventional wisdom of not holding more than ten stocks may be justified. On the other hand, if liability matching ending wealth is a concern, if investor is not comfortable bearing uncompensated risk or if time horizon is long, then one may be better advised by many more recent studies recommending rather 100 than 10 stocks to make a sufficiently diversified portfolio.

Especially we note that when long-run realized return or metrics based on long-run realized return are used to assess the need for diversification, such as in studies by Domian et al. (2007), Bessembinder (2018) and Tidmore et al. (2019), then we are seeing much higher number of stocks required compared to studies focusing on annualized mean standard deviation mitigation. In our view, this is the case because, in the long-run, realized rate of return converges towards growth rate (geometric rate of return) and not towards arithmetic rate of return. The focus in our study is on the diversification effect metrics based on growth rate.

While the number of stocks recommended by the academic studies ranges from about ten to hundreds, the bulk of number of stocks held directly by individual investors seem to range from one to ten. Even though direct holdings don’t represent the whole picture of the diversification, we have a reason to believe that diversification effect at the low range of number of stocks bears significance for individual investors.

3 DIVERSIFICATION IN A CONTINUOUS-TIME WORLD

3.1 The essence of continuous-time world and risk premium

3.1.1 Single period versus continuous-time world

Practical investor lives in a continuous-time world where the word "return" refers to geometric rate of return. The concept and word "return" in people's minds and in common parlance is associated to geometric returns so strongly that reporting arithmetic returns can be considered to be misleading or even cheating. Consequently, according to Hull (2015, p. 348), use of geometric returns in mutual fund reporting is ensured by regulation in some jurisdictions.

As stocks are priced continuously in the markets, in the strictest interpretation, we can consider the single period world to exist only instantaneously at an infinitely short time period. Immediately when the time period is longer than infinitely short the compounding process takes place and we enter, a multiperiod, continuous-time world. Admittedly, the effects of compounding become more tangible at longer time horizons and require reinvesting dividends to fully realize, but the underlying state of affairs of an equity investor is always a continuous-time world.

Markowitz (1959, pp. 116–125) and Thorp (1971) show that a portfolio selected based on optimizing the mean variance properties of a geometric growth rate and, the conventional way, optimizing mean variance of an arithmetic return, typically approximately lead to the same risky asset portfolio construction. Practical portfolio selection, including the central role of Sharpe ratio, therefore is not different between the single period and continuous-time worlds. The difference between these two worlds arises from the non-linearity of the continuous-time world. The fact that portfolio's mean growth rate (mean annualized log return) as a function of portfolio volatility is non-linear, as opposed to linearly behaving portfolio arithmetic mean return, makes a difference in many aspects including diversification effect.

According to Kelly (1956), the use of log return has nothing to do with the investor utility but the mathematical fact that it is the logarithm of period return which is additive over periods and to which the law of large numbers applies. The fact that the logarithm of one plus the arithmetic return, the instantaneous geometric rate of return, is governed by the law of large numbers is the reason why realized rate of return tends toward geometric, not arithmetic, rate of return. As it is the wealth compounding at the rate of realized returns that the investors accumulate and eat, it becomes apparent why it is the geometric, not arithmetic, rate of return that they care about.

The theoretical academic dispute over which return, geometric or arithmetic, is appropriate or what are the properties of the related utility functions may continue forever. Paul Samuelson (see e.g. Samuelson (1971) and Thorp (2008)) has perhaps been the most notable opponent of the adoption of geometric return metrics and the related Kelly criterion (Thorp, 2008). The practitioners, however, seem to have made their choice by almost exclusively measuring, reporting and debating about their performance using geometric rate of return. It really does not need to be more complicated than that. You measure what you care about.

Practitioners, living in a continuous-time world, care about geometric rate of return. Our logical inference is that when assessing the effect of diversification, we should assess the effect on what we care about: the effect on expected geometric rate of return and related risk. This is the main aspect where our study differs from the majority of the existing literature which is focused on assessing the effect of diversification on expected arithmetic rate of return and related risk in a one period world.

3.1.2 Definition of risk premium

Equity risk premium is the mean return that investor earns in excess of a riskless investment alternative such as short-term government bills or government bonds. The logic is that investor is compensated for bearing the systematic market risk associated with equity investments. As put by Dimson, March and Staunton (2003), equity risk premium is by many, with a good reason, considered as “the most important number in finance”. One could argue that without equity risk premium there would be no private risky investments and therefore no system like capitalism as we know it.

The definition of equity risk premium, however, is not as unambiguous as one could assume from "the most important number in finance". The same potential confusion that is practically present with all return related metrics in finance, arithmetic versus geometric, is there with risk premium as well. Typically, when we hear risk premium, we cannot be exactly sure if it is about the arithmetic or geometric mean excess returns.

Dimson, March and Staunton (2003; 2011), however, make the distinction clear and separate the concept to two distinct numbers: arithmetic equity risk premium and geometric equity risk premium. Importantly, Dimson et al. appear to consider the geometric equity risk premium as the determinant of the arithmetic risk premium and not the other way around. Dimson et al. (2011), e.g., state that the historical geometric equity premiums are the sum of investors' *ex ante* expectations and the random component of luck.

Similarly, in our consideration the geometric equity risk premium is the primary risk premium over the arithmetic alternative as the realized return converge towards geometric mean, not arithmetic mean. At the end of the day if an investor is going to eat one of the returns, it will be the realized return which gravitates towards the geometric expectation.

Historical risk premiums are typically measured and refer to large, practically fully diversified, value weighted indices such as country indices. In our study, equity risk premium or put shortly, risk premium, will refer to geometric risk premium. Additionally, our measure of risk premium is for equally weighted portfolio of different levels of diversification. Our risk premium of interest in this study is therefore a geometric risk premium for equally weighted portfolio which we denote as $RP_{EW,G}$, where subscript EW denote equally weighted and G geometric. Corresponding arithmetic risk premium is denoted as $RP_{EW,A}$, where subscript A denote arithmetic. Risk premium in this study will refer to geometric risk premium unless otherwise stated.

3.2 Derivation of the instantaneous geometric risk premium

The expected excess portfolio growth rate is in the heart of our analysis. Excess growth rate means growth rate in excess of riskless rate of return corresponding to the definition of the "most important number in finance": geometric risk premium of an equally weighted portfolio $RP_{EW,G}$. Next, we derive the instantaneous expected excess growth rate formula, the instantaneous geometric risk premium, and show how it can be arrived at from different angles.

3.2.1 Itô's lemma for geometric Brownian motion approach

First, we use Itô's lemma to derive the instantaneous expected excess growth rate formula. Itô's lemma is widely used in financial literature and, e.g., the famous Black and Scholes option pricing formula (Black & Scholes, 1973) is derived utilizing it.

Hull (2015) shows how expected growth rate, i.e., constant drift rate of a logarithmic price process is derived utilizing Itô's lemma. Hull (2015, p. 335) utilizes Itô process:

$$dx = a(x,t)\,dt + b(x,t)\,dz, \tag{1}$$

where a and b are functions of x and t while dz is a Wiener process. Variable x has a drift rate of a and variance rate of b^2.

By using Itô's lemma Hull (2015, p. 335) shows function G, which is a function of x and t, follows the process:

$$dG = \left(\frac{\partial G}{\partial x}a + \frac{\partial G}{\partial t} + \frac{1}{2}\frac{\partial^2 G}{\partial x^2}b^2\right)dt + \frac{\partial G}{\partial x}b\,dz, \tag{2}$$

where process G has a drift rate of

$$\frac{\partial G}{\partial x}a + \frac{\partial G}{\partial t} + \frac{1}{2}\frac{\partial^2 G}{\partial x^2}b^2$$

and a variance rate of

$$\left(\frac{\partial G}{\partial x}\right)^2 b^2.$$

Hull use the Itô's lemma (2015, pp. 336–337) to derive a process followed by logarithmic stock price ln *P* when stock price *P* follows a process:

$$dP = \; mP\,dt + \; sP\,dz, \tag{3}$$

where *m* is the expected (arithmetic) return per year and *s* is the standard deviation of the stock price per year. Note that our notation differs from Hull's notation. Hull use price *S*, expected return μ and standard deviation σ. Our notation is compliant, where applicable, with Thorp (2006) notation which will be followed throughout the thesis.

We replicate Hull's derivation of the process followed by ln *P* with the exception, following Thorp (2006), that we derive a process for logarithmic portfolio value ln *V* and we add riskless rate *r* and investment fraction *f* (fraction of capital allocated to stocks) into process of portfolio value:

$$dV = \; [r + f(m - r)]V\,dt + \; fsV\,dz. \tag{4}$$

The portfolio consisting of fraction f allocated to stocks and $1 - f$ allocated to riskless investments (such as one-month government bonds) is assumed to be rebalanced continuously and dividends are assumed to be reinvested immediately. $\big(r + f(m - r)\big)$ is the drift rate and $f^2s^2V^2$ is the variance rate of portfolio value V.

Applying Itô's lemma as in equation (2), we get:

$$dG = \left(\frac{\partial G}{\partial V}a + \frac{\partial G}{\partial t} + \frac{1}{2}\frac{\partial^2 G}{\partial V^2}b^2\right)dt + \frac{\partial G}{\partial V}b\,dz. \tag{5}$$

Next, we calculate the derivates needed in constructing the process followed by G = ln V:

$$\frac{\partial G}{\partial V} = \frac{1}{V}, \frac{\partial^2 G}{\partial V^2} = -\frac{1}{V^2}, \frac{\partial G}{\partial t} = 0, \tag{6}$$

and substitute the calculated derivates from equation (6) and portfolio value drift rate and variance rate from equation (4) into equation (5):

$$\begin{aligned} dG &= \left(\frac{1}{V}[r + f(m-r)]V + 0 - \frac{1}{2}\frac{1}{V^2}f^2s^2V^2\right)dt + \frac{1}{V}fsV\,dz \\ &= \left[r + f(m-r) - \frac{f^2s^2}{2}\right]dt + fs\,dz, \end{aligned} \tag{7}$$

implying G = ln V has a constant drift rate, i.e., growth rate $g_\infty = r + f(m-r) - \frac{f^2s^2}{2}$ and a constant variance rate of f^2s^2. Following Thorp (2006) notation, subscript ∞ denotes instantaneous drift rate, i.e., continuous compounding. We will from now on denote continuous compounding with subscript ∞.

We are interested in expected portfolio excess growth rate (excess of riskless rate) so we denote expected instantaneous excess return as $m_e = m - r$ and expected instantaneous excess growth rate as $g_\infty^e = g_\infty - r$. Note that now $s_e = Sdev(G_\infty^e)$ in the equation (8) compared to $s = Sdev(G_\infty)$ in equation (7). This means that s_e now represents standard deviation of continuously compounding excess growth (excess of riskless rate growth) instead of standard deviation of continuously compounding growth s. Note that expected instantaneous excess growth rate is equal to instantaneous geometric risk premium for equally weighted portfolio RP_{EW,G_∞}. We therefore have:

$$RP_{EW,G_\infty} = g_\infty^e = fm_e - \frac{f^2 s_e^2}{2}. \tag{8}$$

With f = 1 and r = 0, equation (7) is identical to derivation of lognormal property of stock prices given by Hull (2015, pp. 336–337). It follows that expected portfolio excess value (value in excess of what riskless rate earns) is lognormally distributed and expected logarithmic excess value of a portfolio, i.e., expected instantaneous excess growth rate of a portfolio is normally distributed between time 0 and some future time T with mean $g_\infty^e T$ and variance $f^2 s_e^2 T$.

3.2.2 Power series expansion approach

Thorp (2006) derives the expected growth rate formula differently compared to Itô's lemma approach. While Thorp arrives to the same formula, the important difference is that Thorp, as opposed to Itô's lemma approach, does not assume normal distribution of logarithmic returns or the log-normal distribution of returns. This is important as empirical logarithmic stock returns are well known to not be normally distributed but to entail fat tails (Gabaix, Gopikrishnan, Plerou & Stanley, 2003). Skewness in the distribution of empirical logarithmic returns is not precluded either.

Thorp (2006, p. 406) starts by definitions: X is a random variable with $P(X = m + s) = P(X = m - s) = 0.5$ and $E(X) = m$, $Var(X) = s^2$. Thorp then defines V_0 as the initial capital and, variables introduced in equations (3) and (4), m, r and f. This leads to formula for capital as a function of investment fraction:

$$V(f) = V_0[1 + (1 - f)r + fX] = V_0[1 + r + f(X - r)], \tag{9}$$

leading to expected portfolio growth rate:

$$g(f) = E[G(f)] = E\left(ln\frac{V(f)}{V_0}\right)$$
$$= E(ln[1 + r + f(X - r)])$$
$$= 0.5ln[1 + r + f(m - r + s)] +$$
$$0.5ln[1 + r + f(m - r - s)]. \quad (10)$$

Next, Thorp (2006, p. 406), keeping the same drift and total variance, subdivide the time interval into *n* equal independent steps. This requires dividing *m*, s^2 and *r* by *n*. We now have $X_i, i = 1, \ldots, n$ with probabilities:

$$P\left(X_i = \frac{m}{n} + sn^{-1/2}\right) = P\left(X_i = \frac{m}{n} - sn^{-1/2}\right) = 0.5, \quad (11)$$

leading to compound return:

$$\frac{V_n(f)}{V_0} = \prod_{i=1}^{n}\left[1 + (1 - f)\frac{r_i}{n} + fX_i\right]. \quad (12)$$

According to Thorp (2006, p. 406), by taking $E[\ln(\cdot)]$ both sides of equation (12) and expanding the result in power series, gives the formula for expected capital growth rate:

$$g(f) = r + f(m - r) - \frac{s^2 f^2}{2} + O\left(n^{-1/2}\right), \quad (13)$$

which, as $n \to \infty$ in equation (13), gives the expected instantaneous capital growth rate:

$$g_\infty(f) = r + f(m - r) - \frac{s^2 f^2}{2}. \quad (14)$$

According to Thorp (2006, p. 407), standard deviation s, in the limit when $n \to \infty$, is equal for return and growth: $s = Sdev(X) = Sdev(G_\infty)$.

We note that equation (14) is identical to drift rate in equation (7) which was derived using Itô's lemma. By subtracting r from equation (14) we get expected instantaneous excess growth rate $g_\infty^e(f)$ which is identical to equation (8) derived using Itô's lemma.

The basic principle of the instantaneous compounding process leading to instantaneous geometric risk premium RP_{EW,G_∞}, i.e., the expected instantaneous excess growth rate $g_\infty^e(f)$, is shown in Figure 1. We associate the instantaneous expected excess return m_e to signal and the standard deviation of the excess return, equaling to standard deviation of continuously compounding excess growth, s_e to noise. This association will be explained later.

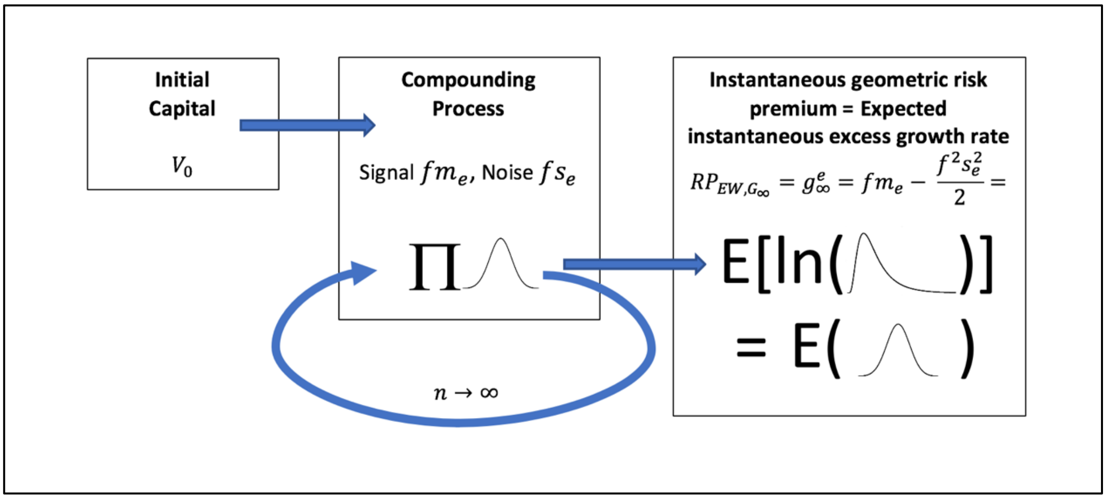

Figure 1. The principle of the instantaneous compounding process.

Importantly Thorp (2006, p. 407) notes that there are no requirements regarding the distribution of random variable X except that X must be bounded with mean m and variance s^2. This means, as opposed to Itô's lemma derivation, there is no assumption of lognormality of returns in Thorp's derivation. For equation (8) and equation (14) to hold, normal distribution of logarithmic returns, the expected instantaneous capital growth rate, is therefore not required.

3.2.3 Euler's number identity approach

Next, we show an alternative method for deriving equation (8) utilizing the definition of Euler's number:

$$e = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n. \tag{15}$$

Thorp (2006 p. 409) notes that by using expected excess return, excess growth rate or excess growth instead of expected return, growth rate or growth respectively, the equations simplify and excess metric equations have complexity equal to non-excess metric equations where r is set to zero. Our interest is in excess growth rate and by substituting $V^e(f) = V(f) - V_0 r$ and $m_e = m - r$ we enter the domain of excess metrics as desired and get the simpler math as a bonus.

First, we derive excess (in excess of riskless capital) capital $V^e(f)$. $V(f)$ is obtained from equation (9):

$$\begin{aligned} V^e(f) = V(f) - V_0 r &= V_0[1 + r + f(X - r)] - V_0 r \\ &= V_0[1 + f(X - r)]. \end{aligned} \tag{16}$$

Then we derive excess growth rate $g^e(f)$. With R denoting the random variable which expected value is the riskless rate r, s_e now denotes the standard deviation of the excess return $Sdev(X - R)$, which is equal to $s = Sdev(X)$ if $Sdev(R) = 0$.

$$g^e(f) = E[G^e(f)] = E\left(ln\frac{V^e(f)}{V_0}\right) = E(ln[1 + f(X - r)])$$
$$= 0.5ln[1 + f(m - r + s_e)] + 0.5ln[1 + f(m - r - s_e)], \quad (17)$$

Next, we substitute $m_e = m - r$ into equation (17) and simplify:

$$g^e(f) = 0.5(ln[1 + f(m_e + s_e)] + ln[1 + f(m_e - s_e)]). \quad (18)$$

Following Thorp (2006), we subdivide the time to *n* equal individual steps by dividing m_e, s_e^2 and *r* by *n*. This gives us equation (18) divided to *n* time steps:

$$g^e(f) = 0.5\begin{pmatrix} \sum_{i=1}^{n} \ln\left[1 + f\left(\frac{m_{e,i}}{n} + \frac{s_{e,i}}{\sqrt{n}}\right)\right] + \\ \sum_{i=1}^{n} \ln\left[1 + f\left(\frac{m_{e,i}}{n} - \frac{s_{e,i}}{\sqrt{n}}\right)\right] \end{pmatrix}. \quad (19)$$

We simplify and rearrange equation (19) to contain auxiliary function $a(f)$ and equation (15) form (for intermediate steps, see Appendix 1):

$$a(f) = 2fm_e + f^2\left(\frac{m_e^2}{n} - s_e^2\right), \quad (20)$$

$$g^e(f) = \frac{a(f)}{2} ln\left[\left(1 + \frac{a(f)}{n}\right)^{\frac{n}{a(f)}}\right]. \quad (21)$$

Next, we use equation (15) identity. When $n \to \infty$ in equations (20) and (21), we have:

$$a_\infty(f) = 2fm_e - f^2 s_e^2, \quad (22)$$

$$g_{\infty}^{e}(f) = \frac{a_{\infty}(f)}{2} \ln e = \frac{a_{\infty}(f)}{2} = f m_e - \frac{f^2 s_e^2}{2}. \tag{23}$$

Again, in equation (23), we have expected instantaneous excess growth rate $g_{\infty}^{e}(f)$ which is identical to equation (8). As our assumptions are based on Thorp (2006) power series expansion derivation, our derivation does not rely on excess growth rate being normally distributed.

3.2.4 Information theory approach

To derive expected instantaneous excess growth rate based on Shannon's (1948) information theory, we look into concepts of entropy and mutual information (MI).

Entropy $H(X)$ is a measure of average uncertainty in a single random variable *X*. The higher the entropy, the more information on average is required to describe the random variable. Conditional entropy $H(X|Y)$ is the entropy, i.e., the average (remaining) uncertainty of random variable *X* conditional to knowledge of random variable *Y*. Mutual information $I(X;Y)$ is a measure of dependence between two random variables *X* and *Y*. Mutual information is the average amount of information obtainable on random variable *X* by observing random variable *Y* and vice versa. (Cover & Thomas, 2005).

Following Cover and Thomas (2005), we denote the entropy of a continuous variable with lower case h as opposed to capital H which is used for discrete random variables. Then, e.g., the entropy $h(X)$ is a measure of uncertainty in a single continuous random variable *X*.

According to Cover and Thomas (2005), mutual information is defined as a function of input entropy $h(X)$ and conditional entropy $h(X|Y)$ or alternatively as a function of output entropy $h(Y)$ and conditional entropy $h(Y|X)$:

$$I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X). \tag{24}$$

Shannon (1948) calls $h(Y)$ as the entropy of the received signal and $h(Y|X)$ as the entropy of the noise.

According to Shannon (1948), entropy $H(\phi)$ of a normally distributed random variable, whose standard deviation is σ, is given as:

$$h(\phi) = \ln(\sqrt{2\pi e}\sigma). \tag{25}$$

Entropy $h(\ln \phi)$ of a log-normally distributed random variable is given by Cover & Thomas (2005, p. 662). Denoting excess return as μ, it can be written as:

$$h(\ln \phi) = \ln(\sigma e^{(\mu - \sigma^2/2 + 1/2)}\sqrt{2\pi}). \tag{26}$$

To apply mutual information formula into investing context, we assume a model where excess return (output or received signal of our investment process) is log-normally distributed. We assume a channel noise distribution equal to normally distributed single period excess return (input or the transmitted signal of our investment process). We can think this in terms of Figure 1 where m_e is the expected excess input (to multiplicative compounding process) return (signal) and s_e is the standard deviation of the input (noise). We therefore define the entropy of a log-normally distributed random variable as the output entropy $h(Y)$, i.e., the entropy of the received signal, and the entropy of a normally distributed random input variable $h(X)$ as being equal to the conditional entropy $h(Y|X)$, i.e., the entropy of the noise. To be consistent with our previous equations, we add investment fraction *f* in the equations of $h(Y)$ and $h(Y|X)$ to enable leveraging of excess returns. Substituting equations (25) and (26) with investment fraction *f* into equation (24) gives mutual information:

$$
\begin{aligned}
I(X;Y) &= h(Y) - h(Y|X) \\
&= \ln\left(f\sigma e^{(f\mu - f^2\sigma^2/2 + 1/2)}\sqrt{2\pi}\right) - \ln\left(\sqrt{2\pi e}f\sigma\right) \\
&= \ln\left(\sqrt{2\pi}f\sigma e^{(f\mu - f^2\sigma^2/2 + 1/2)}\right) - \ln\left(\sqrt{2\pi}f\sigma e^{1/2}\right) \\
&= \ln\left(\frac{\sqrt{2\pi}f\sigma e^{1/2} e^{(f\mu - f^2\sigma^2/2)}}{\sqrt{2\pi}f\sigma e^{1/2}}\right) && (27) \\
&= f\mu - \frac{f^2\sigma^2}{2}. && (28)
\end{aligned}
$$

We can write equation (27) as a function of entropies $h[X(f)]$, $h[Y(f)]$ and $h[(Y(f)|X(f))]$ and arrive to the result of equation (28):

$$
\begin{aligned}
I(X;Y) &= \ln\left(\frac{e^{h[Y(f)]}}{e^{h[(Y(f)|X(f))]}}\right) \\
&= \ln\left(\frac{e^{h[X(f)]} e^{(f\mu - f^2\sigma^2/2)}}{e^{h[(Y(f)|X(f))]}}\right) \\
&= \ln\left(\frac{e^{h[X(f)]} e^{(f\mu - f^2\sigma^2/2)}}{e^{h[X(f)]}}\right) \\
&= f\mu - \frac{f^2\sigma^2}{2}. && (29)
\end{aligned}
$$

We can see from equation (29) that simplification reduces the term $e^{h[X(f)]}$ from the equation. Simultaneously the entropy of the normally distributed random input variable, i.e., the input entropy $h[X(f)]$ is reduced away. This happens as the noise entropy $h[(Y(f)|X(f))]$ is equal to input entropy $h[X(f)]$. In other words, the inherent uncertainty in the input signal is equal to the uncertainty remaining about the input signal after knowing the output signal. Intuitively, the reason why input entropy would be equal to noise entropy is that we can think of transmitting the expected return and receiving a realized return. The received signal, a realized return, is the expected return plus the noise which are the mean and the standard deviation from the same input distribution, respectively. The fact that the input entropy $h[X(f)]$ is reduced from the equation suggests that the distribution of the input signal is not significant to resulting mutual information $I(X;Y)$. This would imply the resulting mutual information is not dependent on input signal being normally distributed.

Now when we change to our notation where $m_e \equiv \mu$ and $s_e \equiv \sigma$, equation (28) is identical to equation (8). As m_e and s_e are expressed as yearly instantaneous metrics, it means we can interpret the mutual information as growth per unit of time: nats per year. We will later show that compounding (C) rate f_C is reduced from the growth rate formula in the limit when it approaches infinity (which will happen here as we have instantaneous metrics). We therefore can write the expected instantaneous excess growth rate as equal to mutual information:

$$g_{\infty}^{e} = I(X;Y), \tag{30}$$

which means, considering our assumed lognormal excess return model, expected instantaneous excess growth rate is equal to the mutual information between normally distributed input (or transmitted) and log-normally distributed output (or received) random variables. We can think this in conjunction with Figure 1. We input (the feedback arrow in the figure) the instantaneous single period normally distributed excess return *X* and, after *n* multiplications in the compounding channel/process, get an output in the form of log-normally distributed excess return *Y*. The output is transformed to expected instantaneous excess growth rate equaling mutual information by taking an expectation of the natural logarithm of the *Y*. Furthermore, as discussed above, it appears that the equality between expected instantaneous excess growth rate and mutual information is not tied to lognormal model, but is more general similarly as we have shown that for the expected instantaneous excess growth rate equation (8) to hold, normally distributed returns are not required.

Furthermore, as we are in the investing context, the expected instantaneous excess growth rate is better described as mutual excess capital growth (MG) than mutual information. MG is the expected amount of excess capital growth realizable from the average excess arithmetic single period return (the input) after (at output) the compounding process (the channel). By comparison, as defined by Cover and Thomas (2005), MI is the average amount of information obtainable on the input by observing the output after the channel.

3.3 Derivatives of the instantaneous geometric risk premium formula

3.3.1 Definition and derivation of diversification premium

Next, based on the instantaneous geometric risk premium formula, we will define and derive the central concept of our study which we label as diversification premium. Diversification premium is the average difference in instantaneous excess growth rate between a portfolio of selected size and a single stock portfolio. We will further define diversification premium difference to benchmark as portfolio diversification premium difference between a portfolio of selected size and a fully diversified benchmark portfolio. We will show that there is a premium, in terms of higher expected growth rate, associated with accepting more diversification like there is a premium for accepting more risk.

Diversification premium for a portfolio of n_P stocks (where n is number of stocks and P denotes portfolio) is defined as the instantaneous geometric risk premium (the instantaneous expected excess growth rate) difference between portfolio of n_p stocks and single stock $(n_P = 1)$ portfolio:

$$DP^P = RP^P_{EW,G_\infty} - RP^{n=1}_{EW,G_\infty} = g^{e,P}_\infty - g^{e,n=1}_\infty = g^P_\infty - g^{n=1}_\infty, \tag{31}$$

where RP^P_{EW,G_∞} is the instantaneous geometric risk premium of a portfolio of n_P stocks and $RP^{n=1}_{EW,G_\infty}$ is the instantaneous geometric risk premium of a single stock portfolio. $g^{e,P}_\infty$ is the expected instantaneous growth rate of a portfolio of n_P stocks and $g^{e,n=1}_\infty$ is the expected instantaneous growth rate of a single stock portfolio. Diversification premium can be calculated similarly using expected instantaneous growth rates g^P_∞ and $g^{n=1}_\infty$ instead of expected instantaneous excess growth rates.

It is obvious from equation (8) that, given a constant expected instantaneous excess return m_e, expected instantaneous excess growth rate difference between two portfolios is determined solely by the variance of the excess return s_e^2. This is because in case of a population of n_{BM} (where *BM* denotes benchmark) stocks we have a constant expected instantaneous excess return m_e leaving s_e^2 as the only differentiator between portfolios of different sizes. We can substitute equation (8) into equation (31) and show that:

$$
\begin{aligned}
DP^P &= g_\infty^{e,P} - g_\infty^{e,n=1} \\
&= fm_e^P - \frac{f^2(s_e^P)^2}{2} - \left(fm_e^{n=1} - \frac{f^2\left(s_e^{n=f=1}\right)^2}{2}\right) \\
&= fm_e - \frac{f^2(s_e^P)^2}{2} - \left(fm_e - \frac{f^2\left(s_e^{n=f=1}\right)^2}{2}\right) \\
&= \frac{f^2}{2}\left[\left(s_e^{n=f=1}\right)^2 - (s_e^P)^2\right]. \qquad (32)
\end{aligned}
$$

We denote average variance of a single stock portfolio $\left(s_e^{n=f=1}\right)^2$ as $Var_{n=f=1}$ and average variance of a n_P–stock portfolio $(s_e^P)^2$ as Var_P. Furthermore, we denote average idiosyncratic (firm specific) variance for a single stock portfolio and n_P stock portfolio as $Ivar_{n=f=1}$ and $Ivar_P$ respectively. Subscript *n=f=1* is to emphasize that the variable is for single stock portfolio with 100% allocation to stocks. Fully diversified benchmark portfolio (consisting of all stocks in the population) variance is denoted as Var_{BM}. With this notation, utilizing the orthogonality between systematic variance Var_{BM} and idiosyncratic variance ($Ivar_{n=f=1}$ or $Ivar_P$), we decompose variance terms and enter the domain of idiosyncratic variance with mathematically convenient zero average correlation between individual stocks.

Average variance of a n_P–stock portfolio Var_P is a sum of systematic benchmark variance and portfolio's idiosyncratic variance. The idiosyncratic variance term can be further decomposed as a function of both, number of stocks in the portfolio n_P and number of stocks in the benchmark n_{BM}. Involving n_{BM} in the equation allows us to derive exact formulas instead of settling for approximations. We can write average variance of a n_P–stock portfolio as:

$$\begin{aligned} Var_P &= Var_{BM} + Ivar_P = Var_{BM} + \frac{Ivar_{n=f=1}}{n_P} - \frac{Ivar_{n=f=1}}{n_{BM}} \\ &= Var_{BM} + \left(\frac{1}{n_P} - \frac{1}{n_{BM}}\right) Ivar_{n=f=1}, \end{aligned} \tag{33}$$

which reduces to systematic benchmark variance exactly when $n_P = n_{BM}$.

We then can rewrite equation (32) and have diversification premium for a portfolio of n_P–stocks:

$$\begin{aligned} DP^P_{n>1} &= \frac{f^2}{2}\left(Var_{n=f=1} - Var_P\right) \\ &= \frac{f^2}{2}\Big[\left(Var_{BM} + Ivar_{n=f=1}\right) \\ &\qquad - \left(Var_{BM} + \frac{Ivar_{n=f=1}}{n_P} - \frac{Ivar_{n=f=1}}{n_{BM}}\right)\Big] \\ &= \frac{f^2}{2}\left(Ivar_{n=f=1} - \frac{Ivar_{n=f=1}}{n_P} + \frac{Ivar_{n=f=1}}{n_{BM}}\right) \\ &= \left(1 - \frac{1}{n_P} + \frac{1}{n_{BM}}\right)\frac{Ivar_{n=f=1}}{2} f^2. \end{aligned} \tag{34}$$

Note that above equation holds for portfolios larger than one stock. When $n_P = 1$, then the diversification premium is zero. In the limit, when $n_{BM} \to \infty$, diversification premium simplifies to:

$$DP^P_{n_{BM}\to\infty} = \left(1 - \frac{1}{n_P}\right)\frac{Ivar_{n=f=1}}{2} f^2. \tag{35}$$

When n_{BM} is large compared to n_P, equation (35) can be used to approximate the diversification premium of a n_P–stock portfolio.

It turns out that the approximate diversification premium equation (35) is basically the same equation that Erb and Harvey (2006) provide for diversification return:

$$Diversification\ return = \frac{1}{2}\left(1 - \frac{1}{K}\right)\bar{\sigma}^2(1 - \bar{\rho}), \tag{36}$$

where *K* is the number of securities in the portfolio, $\bar{\sigma}^2$ is average variance of a security and $\bar{\rho}$ is the average correlation of a security. In equation (35) we have entered the domain of idiosyncratic variance implying average variance of a security ($\bar{\sigma}^2$) corresponds to idiosyncratic variance of a security ($Ivar_{n=f=1}$) and average correlation ($\bar{\rho}$) between securities is zero. Finally, *K* corresponds to n_P implying equations (35) and (36) are inherently identical with the exception that our equation (35) includes additional parameter investment fraction *f*, which is assumed to be one in equation (36).

Equation (35) is also consistent with the diversification bonus and rebalancing bonus formulas derived by Bernstein and Wilkinson (1997) who give an approximate (for practical rebalancing frequencies instead the theoretical infinite frequency, hence the "≈") formula for rebalancing bonus as:

$$G - G' \approx \left[\sum_i X_i(1 + G_i)\right] - \left[\sum_i X_i(1 + G_i)^N\right]^{1/N} + \sum_{i<j} X_i X_j \left(\frac{V_{ii}}{2} + \frac{V_{jj}}{2} - V_{ij}\right). \tag{37}$$

where G is geometric return of rebalanced portfolio and G' is the geometric return of the corresponding portfolio that is not rebalanced. X_i and X_j are asset weights (sum of X_i is constrained to one) of the i-th and j-th asset respectively while G_i is the geometric return of the i-th asset. N is the number of compounding periods and V_{ii} and V_{jj} are variances while V_{ij} is covariance. The term in the second line is called diversification bonus and the sum $i < j$ means the sum over all asset pairs i and j for which $i < j$.

Comparing equation (37) to our equation (35) we first note that the expected geometric return between each of the randomly picked stocks in our portfolio is equal. In accordance with how Bernstein and Wilkinson (1997) interpret their formula, equal expected asset returns imply the right-hand side of the first row of their formula vanishes leaving the second row, the diversification bonus, equal to rebalancing bonus. Now, as our equation is in the idiosyncratic variance domain, also the covariance term V_{ij} vanishes from the second row. Furthermore, investment fraction in Bernstein and Wilkinson equation is assumed to be one and asset weights in our equation are assumed to be equal. After substituting investment fraction one into our equation and equal weighting into Bernstein and Wilkinson equation, equations (35) and (37) are the same. For example, for a two-stock portfolio, both equations imply diversification premium (diversification bonus in Bernstein and Wilkinson terms) equal to one fourth of a single stock idiosyncratic variance.

When $n_P < n_{BM}$, randomly picked portfolio of n_P stocks from the benchmark population of stocks will always have an average excess return variance greater than the average excess return variance of a fully diversified benchmark portfolio of n_{BM} stocks. This implies, when $n_P < n_{BM}$, that expected instantaneous excess growth rate of a randomly picked portfolio of n_P stocks from this population of stocks will always be lower in comparison to fully diversified benchmark portfolio of n_{BM} stocks. Hence, there is a premium, i.e., greater expected growth rate paid for accepting more diversification.

We can calculate diversification premium for benchmark portfolio, i.e., expected growth rate difference between fully diversified portfolio and single stock portfolio. In this case $n_P = n_{BM}$ and substituting to equation (34) gives:

$$DP^{BM} = \frac{Ivar_{n=f=1}}{2} f^2 = DP^{BM}_{f=1} f^2, \tag{38}$$

where we can see that diversification premium for benchmark portfolio with $f = 1$ is:

$$DP^{BM}_{f=1} = \frac{Ivar_{n=f=1}}{2}. \tag{39}$$

We can see the diversification premium, i.e., the expected portfolio growth benefit from diversification, for a benchmark portfolio is directly proportional to average idiosyncratic variance of a single stock and proportional to the square of investment fraction.

We will further define diversification premium difference to benchmark as portfolio diversification premium difference between a n_P–stock portfolio and fully diversified benchmark portfolio. This equals the difference in the expected instantaneous excess growth rates. New symbol $g^{e,BM}_{\infty}$ in the equation is the expected instantaneous excess growth rate of a benchmark portfolio:

$$\begin{aligned}\Delta DP^{BM}_{n>1} &= g^{e,P}_{\infty} - g^{e,BM}_{\infty} = DP^{P} - DP^{BM} \\ &= \left(1 - \frac{1}{n_P} + \frac{1}{n_{BM}}\right)\frac{Ivar_{n=f=1}}{2} f^2 - \frac{Ivar_{n=f=1}}{2} f^2 \\ &= -\frac{Ivar_{n=f=1}}{2}\left(\frac{1}{n_P} - \frac{1}{n_{BM}}\right) f^2 \\ &= \left(\frac{1}{n_{BM}} - \frac{1}{n_P}\right) DP^{BM}_{f=1} f^2. \end{aligned} \tag{40}$$

Note that above equation holds for portfolios larger than one stock. When $n_P = 1$, portfolio's diversification premium is zero and diversification premium difference to benchmark is equal to the opposite of diversification premium of a benchmark portfolio:

$$\Delta DP_{n=1}^{BM} = -DP^{BM} = -DP_{f=1}^{BM} f^2 = -\frac{Ivar_{n=f=1}}{2} f^2. \quad (41)$$

Diversification premium difference to benchmark gives us the n_P–stock portfolio expected instantaneous excess growth rate difference to benchmark. Diversification premium difference to benchmark is always negative when $n_P < n_{BM}$.

In the limit, when $n_{BM} \to \infty$ in equation (40), diversification premium difference to benchmark simplifies to:

$$\Delta DP_{n_{BM}\to\infty}^{BM} = -\frac{Ivar_{n=f=1}}{2n_P} f^2 = -\frac{Ivar_P}{2} f^2 = -\frac{DP_{f=1}^{BM}}{n_P} f^2 = -\frac{DP^{BM}}{n_P}. \quad (42)$$

When n_{BM} is large compared to n_P, equation (42) can be used to approximate diversification premium difference to benchmark for a n_P–stock portfolio as a function of idiosyncratic variance of a single stock portfolio with 100% stock allocation. Alternatively, we can think of equation (42) as a function of idiosyncratic variance of a n_P–stock portfolio with 100% stock allocation $Ivar_P$. This is because when n_{BM} is large compared to n_P, idiosyncratic variance of a n_P–stock portfolio is approximated by the idiosyncratic variance of a single stock portfolio with 100% stock allocation $Ivar_{n=f=1}$ divided by n_P.

Furthermore, when n_{BM} is large compared to n_P, we can see from equation (42) that the diversification premium difference to benchmark for an investment portfolio is approximated by the diversification premium of a benchmark portfolio divided by the number of stocks in the investment portfolio.

3.3.2 Estimation of diversification premium

We have established that portfolio diversification premium is a function of average idiosyncratic variance of a single stock portfolio $Ivar_{n=f=1}$. Next, we will show that $Ivar_{n=f=1}$ can be estimated from exhaustive historical stock return data by using a linear ordinary-least-squares (OLS) regression. In addition, we will show how diversification premium can be arrived to from several different angles by utilizing different regression outputs leading to three different formulas expressing diversification premium difference to benchmark. Important property of our approach is that it utilizes exhaustive data consisting of the whole selected benchmark population of individual stock returns from selected time period. The approach is enabled by deriving equations describing the generalization of the parameters estimated from individual stock data to any n_P–stock size portfolio.

As we are interested in geometric rate of return, we use logarithmic returns instead of arithmetic (simple) returns in our regression. Our regression model is the following:

$$StockLnERet_{t,i} = \alpha + \beta BmLnERet_t + \epsilon_{t,i}. \tag{43}$$

In equation (43), $i = 1, \dots, n_t$ denotes the i-th individual stock occurrence in sub-period t, while n_t is the number of individual stock occurrences in sub-period t. $t = 1, \dots, T$ denotes the t-th sub-period, while T is the number of sub-periods in whole period of interest. Our empirical sub-period length is one month. $StockLnERet_{t,i} = \ln(1 + StockERet_{t,i})$ and $StockERet_{t,i}$ is the i-th simple excess return of an individual stock in month t. $BmLnERet_t = \ln(1 + BmERet_t)$ and $BmERet_t$ is the simple excess return of an equally weighted benchmark portfolio (including the whole population of benchmark's stocks) in month t. α is the intercept, i.e., the expected unexplained difference between $StockLnERet_{t,i}$ and $BmLnERet_t$. β is the sensitivity of $StockLnERet_{t,i}$ to changes in $BmLnERet_t$ corresponding to the concept of market beta. $\epsilon_{t,i}$ is the residual or the error term of i-th individual stock data point in month t. $StockERet_{t,i}$ and $BmERet_t$ have 100% allocation to underlying stocks implying investment fraction *f* is one.

We are interested in portfolio's expected excess growth rate which is used in the compounding process. Each sub-period (e.g. one month) has equal weight in the compounding process. The regression model described by equation (43) therefore is directly usable only if each sub-period in the data has equal number of individual stock returns as the number of returns per sub-period equals the weight of a sub-period in the compounding process over the total period. Equal number of returns per sub-period is typically the case in simple simulations, but is not the case with empirical data where, e.g., the number of stocks per month varies and may differ significantly especially when the data spans a longer period of time like several decades. We therefore introduce two alternative regression model specifications which account for the differences in the number of data points per time sub-period.

Alternative regression model specifications are such that we first calculate a normalization weight for each time sub-period, assign the weight to each stock return belonging to the sub-period, and then multiply both $StockLnERet_{t,i}$ and $BmLnERet_t$ in equation (43) with this weight to neutralize potential differences in the number of stock return data points and therefore equalize the weights among sub-periods. The lower the number of stock returns per sub-period is, the greater the weight for each individual stock return belonging to the sub-period will be and vice versa. The first of the alternative specifications has a normalization weight for stock returns in the t-th sub-period as follows:

$$w_t^{norm\_\alpha} = \overline{n_t}/n_t, \tag{44}$$

where $\overline{n_t}$ is the average number of stock returns per sub-period over the whole period of interest and n_t is the number of stock returns in the t-th time sub-period.

The normalization weight for stock returns in the t-th sub-period for the second alternative specification is:

$$w_t^{norm\_Var} = \sqrt{\overline{n_t}/n_t}. \tag{45}$$

The first alternative specification is then derived from regression equation (43) by normalizing logarithmic returns by $w_t^{norm\_\alpha}$:

$$w_t^{norm\_\alpha} StockLnERet_{t,i} = \alpha + \beta w_t^{norm\_\alpha} BmLnERet_t + \epsilon_{t,i}, \quad (46)$$

and similarly, the second alternative specification becomes:

$$w_t^{norm\_Var} StockLnERet_{t,i} = \alpha + \beta w_t^{norm\_Var} BmLnERet_t + \epsilon_{t,i}. \quad (47)$$

Note that when each sub-period in the data has equal number of individual stock returns, normalization weight becomes one and equations (46) and (47) simplify to equation (43). However, as the empirical number of individual stock returns typically differ between sub-periods, we choose to exclusively use either regression equation (46) or (47). As the notation in the normalization weight equations (44) and (45) suggests, former is used with alpha-based equations and latter with variance-based equations. The difference of the two normalization weights stems from the fact that alpha scales linearly while variance scales with a squared scale factor. We therefore take a square root when obtaining the normalized number of stocks per time sub-period in the variance-based case. We will refer to appropriate regression equation when the alpha- and variance-based equations are introduced.

We start with a variance-based metric. Average idiosyncratic variance of a single stock portfolio $Ivar_{n=f=1}$ can now be obtained from the residual of the regression equation (47):

$$Ivar_{n=f=1} = Var(\epsilon_{t,i}). \quad (48)$$

Alternatively, we can arrive to diversification premium difference to benchmark formula by simply using alpha from the regression equation (46). As we have defined, the diversification premium difference to benchmark for a single stock portfolio is the expected growth rate difference between single stock portfolio and fully diversified benchmark portfolio. This is the exact definition of α in regression equation (43) and therefore also in (46). To be consistent with the notation $Ivar_{n=f=1}$, we denote alpha as $\alpha_{n=f=1}$ emphasizing that alpha is for single stock portfolio with 100% stock allocation. We can write $\alpha_{n=f=1}$ equal to $n_P = 1, f = 1$ and equation (39) substituted into equation (41) and have alpha for single stock portfolio with 100% stock allocation:

$$\alpha_{n=f=1} = \Delta DP^{BM}_{n=f=1} = -\frac{Ivar_{n=f=1}}{2}. \tag{49}$$

We can solve $Ivar_{n=f=1}$ from equation (49):

$$Ivar_{n=f=1} = -2\alpha_{n=f=1}, \tag{50}$$

which by substituting equation (39) gives:

$$DP^{BM}_{f=1} = -\alpha_{n=f=1}, \tag{51}$$

and by substituting equations (50) and (51) into equation (40) we have diversification premium difference to benchmark as function of alpha:

$$\begin{aligned} \Delta DP^{BM}_{n>1} &= \alpha_{n=f=1}\left(\frac{1}{n_P} - \frac{1}{n_{BM}}\right) f^2 \\ &= DP^{BM}_{f=1}\left(\frac{1}{n_{BM}} - \frac{1}{n_P}\right) f^2. \end{aligned} \tag{52}$$

Note that above equation holds for portfolios larger than one stock. When $n_P = 1$, portfolio's diversification premium is zero and diversification premium difference to benchmark is equal to the opposite of diversification premium of a benchmark portfolio which now equals the alpha scaled by the squared of investment fraction:

$$\Delta DP^{BM}_{n=1} = -DP^{BM} = -DP^{BM}_{f=1}f^2 = \alpha_{n=f=1}f^2. \tag{53}$$

In the limit, when $n_{BM} \to \infty$, equation (52) simplifies to:

$$\Delta DP^{BM}_{n_{BE}\to\infty} = \frac{\alpha_{n=f=1}}{n_P}f^2 = -\frac{DP^{BM}_{f=1}}{n_P}f^2. \tag{54}$$

When n_{BM} is large compared to n_P, equation (54) approximates instantaneous expected excess growth rate difference between n_P–stock portfolio and fully diversified benchmark portfolio as a function of alpha of a single stock portfolio with 100% stock allocation.

Yet another alternative to present diversification premium difference to benchmark formula is by utilizing R-squared metric from regression (47) (we can see from equation (55) that R-squared is a variance-based metric) and benchmark variance.

R-squared can be expressed as one minus the ratio of the residual sum of squares (RSS) to the total sum of squares (TSS) (Brooks, 2014, p. 153). We can consider the RSS as the idiosyncratic variance of the portfolio $Ivar_P$ and the TSS as the total variance of the portfolio which equals benchmark variance Var_{BM} plus the idiosyncratic variance of the portfolio $Ivar_P$. Now, when $n_P = 1$ and $f = 1$, we denote R-squared as $R^2_{n=f=1}$ emphasizing that R-squared from regression is for single stock portfolio with 100% stock allocation. Similarly, $Ivar_P = Ivar_{n=f=1}$. We therefore can write the R-squared as a function of benchmark variance and idiosyncratic variance of a single stock portfolio with 100% stock allocation.

$$\begin{aligned} R^2_{n=f=1} &= 1 - \frac{Ivar_{n=f=1}}{Var_{BM} + Ivar_{n=f=1}} \\ &= \frac{1}{1 + Ivar_{n=f=1}/Var_{BM}}. \end{aligned} \tag{55}$$

Next, we solve $Ivar_{n=f=1}$ from equation (55):

$$Ivar_{n=f=1} = \left(\frac{1}{R^2_{n=f=1}} - 1\right) Var_{BM}, \tag{56}$$

which by substituting (39) gives:

$$DP^{BM}_{f=1} = \left(\frac{1}{R^2_{n=f=1}} - 1\right) \frac{Var_{BM}}{2}, \tag{57}$$

and substituting equations (56) and (57) into equation (40) gives diversification premium difference to benchmark as function of R-squared and benchmark variance:

$$\begin{aligned} \Delta DP^{BM}_{n>1} &= \left(1 - \frac{1}{R^2_{n=f=1}}\right) \frac{Var_{BM}}{2} \left(\frac{1}{n_P} - \frac{1}{n_{BM}}\right) f^2 \\ &= \left(\frac{1}{n_{BM}} - \frac{1}{n_P}\right) DP^{BM}_{f=1} f^2. \end{aligned} \tag{58}$$

Note that above equation holds for portfolios larger than one stock. When $n_P = 1$, portfolio's diversification premium is zero and diversification premium difference to benchmark is equal to the opposite of diversification premium of a benchmark portfolio:

$$\Delta DP^{BM}_{n=1} = -DP^{BM} = -DP^{BM}_{f=1} f^2 = \left(1 - \frac{1}{R^2_{n=f=1}}\right) \frac{Var_{BM}}{2} f^2. \tag{59}$$

In the limit, when $n_{BM} \to \infty$, equation (58) simplifies to:

$$\Delta DP^{BM}_{n_{BM}\to\infty} = \left(1 - \frac{1}{R^2_{n=f=1}}\right)\frac{Var_{BM}}{2n_P}f^2 = -\frac{DP^{BM}_{f=1}}{n_P}f^2. \quad (60)$$

When n_{BM} is large compared to n_P, equation (60) approximates instantaneous expected excess growth rate difference between n_P–stock portfolio and fully diversified benchmark portfolio as a function of R-squared of a single stock portfolio with 100% stock allocation and benchmark variance.

Equation (56) decompose idiosyncratic variance into two components: First component $\left(\frac{1}{R^2_{n=f=1}} - 1\right)$ which is a function of R-squared and second component Var_{BM} which is the systematic risk, i.e., the variance of the fully diversified benchmark portfolio. This decomposition gives us good intuition what drives the diversification premium difference between n_P–stock portfolio and fully diversified benchmark portfolio in equation (58). R-squared basically tells us how representative the benchmark is to individual stock risks, i.e., what proportion of the risks, and therefore expected growth rate, is explained by the benchmark. The first component therefore is the size, in relation to systematic risk component size, of a single stock portfolio risks not explained by the benchmark, i.e., relative size of idiosyncratic, costly risk. The second component tells us how risky the benchmark is by itself. Multiplying the first and the second component together gives us the absolute amount of idiosyncratic risk for a single stock portfolio.

From above decomposition of idiosyncratic variance, it seems like a reasonable hypothesis that investment style may affect the diversification premium difference between n_P–stock portfolio and fully diversified benchmark portfolio. Investment style loads on additional risk factors, not just market factor. For example, small cap value style loads on market, size and value factors which should increase the $R^2_{n=f=1}$ metric which now accounts for exposure to three factors instead of one[1]. Additionally,

[1] Since we have rebalanced equally weighted benchmarks instead of buy and hold value weighted, we acknowledge that there is some base exposure to size, value and momentum factors because of the equal weighting per se. Regardless the base factor exposures, deliberate exposure to distinct investment style is expected to be reflected in the factor exposures associated with the style.

it is obvious that the benchmark variance affects the diversification premium difference to benchmark. If we think of focusing on different firm sizes as investment style and compare two benchmarks, small cap stocks and large cap stocks, we expect small cap stocks benchmark to have higher variance. This supports the hypothesis that investment style affects diversification premium difference to benchmark.

3.3.3 Diversification is a negative price lunch

As opposed to one period world where diversification is famously “a free lunch”, in continuous-time, multi-period world, diversification is “a negative price lunch”. This is based on the fact that in the one period world idiosyncratic risk is “uncompensated risk” whereas in the continuous-time world idiosyncratic risk is “costly risk”. The costly nature of the idiosyncratic risk in the continuous-time world is obvious from the approximate diversification premium difference to benchmark equation (42) where we can see that expected instantaneous growth rate of a n_P–stock portfolio is the expected instantaneous growth rate of fully diversified benchmark portfolio minus half the portfolio’s idiosyncratic variance. To be precise, in continuous-time world also the systematic risk, while compensated, is simultaneously costly as can be seen from equation (8) where any variance, including systematic variance, decreases the growth rate.

More specifically, one half of the portfolio’s idiosyncratic variance approximates the magnitude of diversification premium difference to benchmark, the opportunity cost of foregone diversification. In one period world arithmetic rate of return is constant regardless the level of diversification, i.e., idiosyncratic risk is unnecessary and uncompensated, but not costly.

The opportunity cost above holds for portfolio’s with exactly 100% stock allocation implying investment fraction one. Importantly, as shown by equation (42), the effect of asset allocation between risky and riskless assets dramatically affects the opportunity cost of foregone diversification as the impact from idiosyncratic variance is multiplied by the square of the investment fraction. Accounting for asset allocation, we can more precisely determine one half of the portfolio’s idiosyncratic variance scaled by the squared investment fraction as approximating the magnitude of diversification premium difference to benchmark, the opportunity cost of foregone diversification.

What “a negative price lunch” means is that an investor is paid for consuming the lunch of diversification. An investor earns the higher expected portfolio growth rate the more diversification he accepts. This description bears a similarity to description of the term “risk premium”: An investor earns the higher expected portfolio growth rate the more (systematic) risk he accepts. Hence the term “diversification premium”.

In continuous-time world risk is polarized. All risk is costly, but systematic risk is compensated while idiosyncratic risk is not. Risk therefore is either potentially beneficial (compensated systematic risk) or plain costly (idiosyncratic risk), there is no middle ground. The only way to earn a potentially higher geometric risk premium is to bear more systematic risk(s) while any exposure to idiosyncratic risk will only decrease the geometric risk premium.

Even the compensated systematic risk entails the underlying costly nature common to all continuous-time world risks. Therefore, even the systematic risk, as we will learn in the coming discussion about the Kelly criterion, when overdosed, may decrease the expected growth rate of a portfolio.

3.3.4 Kelly criterion and the magic of Sharpe ratio

Kelly (1956) utilized information theory, introduced by Shannon (1948), in the context of gambling. Kelly introduced a new interpretation of Shannon’s information rate. This interpretation is known as the Kelly criterion. Kelly criterion states that gambler, playing positive expectation games, shall not target maximizing the expected value of his bankroll, but instead targets maximizing the expected growth rate (the expected value of the logarithm) of his bankroll in consecutive bets. The gambler will, by following the policy of betting a constant fraction (as given by Kelly criterion) of his bankroll, achieve the maximum expected growth rate for his bankroll while simultaneously (given indefinitely divisible bets) ensure that the probability of ruin approaches zero. Alternative strategy of maximizing the expected value of his bankroll would imply betting the whole bankroll at each trial and, in the long term, would lead to almost certain ruin. The key is to maximize expected growth rate, not expected value. (Kelly, 1956; Thorp, 2006).

Thorp (2006), derives a continuous variable, continuous-time version of the discrete variable, discrete time Kelly criterion. Thorp provides us with tools to input the standard metrics in finance, expected return and volatility of the returns, into Kelly criterion formula and shows how the instantaneous drift rate, i.e., log return or expected growth rate of a portfolio is a non-linear function of risk (volatility of the growth rate) and fraction of capital allocated to risky assets. Thorp shows how the implications and conclusions from Kelly criterion as determined by Kelly (1956) in discrete gambling domain can be transferred into continuous stock investment domain.

Thorp (2008) cites an interview where Warren Buffett is asked about diversification and his view on position sizing. Buffett says he has two views on diversification: Those who are not professional and don't have stock picking skill (edge) should diversify maximally, while those who are professional and have an edge use Kelly and concentrated portfolios.

Kelly criterion is applicable in two aspects. The first aspect is the portfolio selection which maximizes the reward to risk trade-off. Stock picking skill combined with concentrated, properly weighted bets can increase the expected return and hence improve the portfolio selection outcome. The second aspect is the portfolio risk level selection or minimization for the overall portfolio. Buffett can actively utilize both aspects by selecting and weighting the stocks while keeping an eye on the risk level of his potentially leveraged portfolio. Diversifier takes a passive stance towards the first aspect but can maximally utilize the second. Importantly, Kelly criterion makes evident that passive diversifier has one significant advantage over the active concentrated bettor. Namely a higher expected portfolio growth rate before accounting for the potential edge of the stock picker. This is the edge of the diversifier arising from lower risk. The stock picker therefore needs a skill level high enough to overcome the opportunity cost of foregone diversification edge to make his concentrated stock picking efforts worthwhile.

Our application of Kelly criterion will assume market efficiency which corresponds to investors without stock picking skill and emphasizes the role of diversification. We will assume randomly picked equally weighted stock portfolio and seek to explain the characteristics of that portfolio using the Kelly criterion. We emphasize the importance of rebalancing and assume a continuously rebalanced and continuously compounding portfolio. We will show how Sharpe ratio plays a key role as the most important parameter determining the opportunity set for a continuous-time world investor. We touch the subjects of rational weight on risky assets and the efficient frontier and how they differ between a single-period and continuous-time world investor. We also show how Sharpe ratio based on geometric returns is very different metric compared to Sharpe ratio based on arithmetic returns. Importantly from typical individual investor point of view, who as shown by Goetzmann and Kumar (2008) and Polkovnichenko (2005) is poorly diversified, we show how Kelly criterion explains the poor expected geometric risk premium at the low end of the level of diversification.

Thorp (2006, p. 407) shows that maximum expected instantaneous growth rate, applicable to an investment portfolio, is achieved at point f^*:

$$f^* = \frac{m-r}{s^2}, \tag{61}$$

where m is the expected instantaneous (arithmetic) return per year, r is the instantaneous riskless rate per year and s is the standard deviation of the continuously compounding growth per year, i.e., the standard deviation of the logarithmic return per year. f^* is obtained by differentiating equation (14) with respect to f and finding the maximum. f^* can be referred to as “full investment fraction” or “full Kelly”. According to Thorp (2006, p. 407), in the infinitely short compounding interval the standard deviation of the continuously compounding growth of a continuously rebalanced portfolio, i.e., the standard deviation of logarithm of one plus arithmetic return is equal to standard deviation of the arithmetic returns.

Similarly, we can find the investment fraction *f* delivering the maximum expected instantaneous excess growth rate by differentiating equation (8) with respect to *f*. "Full Kelly" then is:

$$f^* = \frac{m_e}{s_e^2}, \tag{62}$$

Where expected instantaneous (arithmetic) excess return $m_e = m - r$ per year and s_e is standard deviation of the continuously compounding excess growth (excess of riskless rate growth) per year.

Following Thorp (2006, p. 407), we substitute equation (62) into equation (8) and have maximum expected instantaneous excess growth rate:

$$g_\infty^e(f^*) = \frac{m_e}{s_e^2} m_e - \frac{\left(m_e/s_e^2\right)^2 s_e^2}{2} = \frac{m_e^2}{2s_e^2}. \tag{63}$$

Sharpe ratio, the corner stone of modern portfolio theory, is central also in the continuous-time world where investors care about the growth rate, i.e., the geometric rate of return. Thorp (2006, p. 407) shows that for fixed riskless rate, maximum expected instantaneous growth rate depends only on Sharpe ratio. This means that maximum expected instantaneous excess growth rate for a portfolio depends on Sharpe ratio and nothing else.

Now, remember that the definition of the general (*ex ante*) Sharpe ratio (SR) as given by Sharpe (1994) is:

$$SR = \frac{\bar{d}}{\sigma_d}, \tag{64}$$

where $\bar{d}$ is the expected differential return between fund arithmetic return R_F and benchmark arithmetic return R_B and σ_d is the predicted standard deviation of the differential return $d = R_F - R_B$.

In our case, portfolio's instantaneous expected excess return m_e corresponds to expected differential return $\bar{d}$ and predicted standard deviation of the continuously compounding excess growth s_e corresponds to predicted standard deviation of the differential return σ_d. Note that there is a subtle difference both between returns and standard deviations. $\bar{d}$ is periodically compounded rate while m_e is an instantaneous (continuously compounded) rate. σ_d is the predicted standard deviation for the periodically compounded arithmetic differential return whereas s_e is the predicted standard deviation for the continuously compounded excess growth, i.e., the predicted standard deviation for the logarithm of one plus the arithmetic differential return. We follow Baz and Guo (2017) and call our Sharpe ratio as the instantaneous Shape Ratio and use subscript ∞ to denote the instantaneousness:

$$SR_\infty = \frac{m_e}{s_e}. \tag{65}$$

It is the instantaneous Sharpe ratio that Thorp (2006, p. 407) shows is the determinant of maximum expected instantaneous growth rate together with riskless rate. When compounding and balancing periods are infinitely short, the standard deviation of the arithmetic differential return equals the standard deviation of the logarithm of one plus the arithmetic differential return (Thorp, 2006, p. 407). Also, periodically compounded excess return becomes instantaneous excess return when period becomes infinitely short. Therefore, at infinite compounding and rebalancing frequency the conventional and instantaneous Sharpe ratios are equal.

When compounding and rebalancing frequency is lower than infinite, we can think of the difference between the standard deviations between instantaneous and periodically compounding returns as the difference between the standard deviation of a normally distributed instantaneous return and the standard deviation of an associated log-normal distribution, which is the result from compounding the normally distributed returns over time. When the arithmetic expected return is greater than zero, the standard deviation for the log-normal distribution is always greater than the standard deviation for the normal distribution and the difference increases as a function of time (length of rebalancing period). On the other hand, the excess return component of the Sharpe ratio is slightly greater for periodically compounded excess return and the difference increases as a function of time, which help to compensate the Sharpe ratio difference arising from the difference between standard deviations. When utilizing monthly return data, we find the difference between conventional and instantaneous Sharpe ratio with typical stock market return parameters is negligible.

Rebalancing is essential. In addition to assumed infinite compounding frequency, a Kelly investor, who cares about the portfolio growth rate which (as will be shown) is a function of Sharpe ratio, will rebalance his portfolio theoretically at infinite frequency (Thorp, 2006, p. 408). Rebalancing now entails both rebalancing between riskless rate and stock portfolio (maintaining desired investment fraction f) and rebalancing among stocks to maintain equal weighting. As can be seen from the equation (37) derived by Bernstein and Wilkinson (1997), the rebalancing bonus, and therefore the expected growth rate of a portfolio, for assets with equal expected growth rate is the greatest when asset weights are equal. Rebalancing bonus is the expected growth rate difference between rebalanced and non-rebalanced portfolio. The results for rebalancing bonus for assets with equal expected growth rate apply to our case as we treat each randomly picked individual stock as sharing the characteristics of an average stock in the benchmark population. Weights in a non-rebalanced, buy and hold, portfolio increasingly deviate from original weights as time passes on. This implies the greater the rebalancing frequency the more equal the weighting among stocks and theoretically, not accounting for costs or any empirical phenomena like momentum, the greater the expected excess growth rate for our randomly picked stock portfolio.

The literature about rebalancing does not support very frequent rebalancing for practitioners. Typically, monthly, quarterly, and annual rebalancing frequencies don’t produce significantly differing results when measured as average portfolio growth rate. If anything, after costs, too frequent rebalancing may cause a growth rate disadvantage. According to empirical tests by Jaconetti, Kinniry and Zilbering (2015), rebalancing broadly diversified stock and bond asset allocation using monthly, quarterly or annual frequency does not yield a significant difference in reward or risk characteristics between rebalanced portfolios. Jaconetti et al., however, do find a significant difference between the rebalanced portfolios and never rebalanced portfolio and conclude that, at reasonable frequencies, rebalancing is foremost a method to maintain desired reward/risk trade-off rather than to target enhancing portfolio growth rate.

Kuhn and Luenberger (2010) specifically study the effect of rebalancing frequency to growth rate for log-optimal portfolio. They show that continuous rebalancing mathematically yields the highest possible growth rate. This implies that infinite rebalancing frequency theoretically maximizes instantaneous Sharpe ratio. However, Kuhn and Luenberger also show theoretically that the difference in achieved growth rate before costs is negligible between continuous rebalancing (infinite rebalancing frequency) and annual rebalancing.

Based on the literature it seems safe to assume that no significant portfolio growth rate difference is expected when deviating from the theoretical continuous rebalancing as long as the rebalancing takes place at a reasonable frequency which we consider to be once per year at lowest. Therefore, even though our equations assume continuous rebalancing, we expect the equations to very closely approximate the results obtainable at practical and reasonable rebalancing frequencies. There is one caveat however. Contrary to most studies, our analysis includes investment fraction f which means the portfolio can be leveraged. If portfolio is aggressively leveraged, it will deviate from typical parameters used in studies. Use of leverage therefore could potentially change the rebalancing frequency requirements for a portfolio. Intuitively, if anything, increased f should call for higher rebalancing frequency.

Now, building on the instantaneous Sharpe ratio, substituting equation (65) into equation (63) gives us maximum expected instantaneous excess growth rate as a function of instantaneous Sharpe ratio:

$$g_{\infty}^{e}(f^{*}) = \frac{SR_{\infty}^{2}}{2}. \tag{66}$$

What is remarkable is that the maximum achievable expected instantaneous excess growth rate of a portfolio is a function of instantaneous Sharpe ratio and nothing else. This is the expected instantaneous excess growth rate at “full Kelly”, i.e., at investment fraction f^*. “Fractional Kelly”, as determined by Thorp (2006, p. 408), is the Kelly fraction c multiplied by “full Kelly” investment fraction f^* leading to investment fraction $f = cf^*$. For example, “half Kelly” corresponds to investment fraction $f = \frac{1}{2}f^*$. When Kelly fraction c in a long-only portfolio is increased from zero to one, expected growth rate and risk (standard deviation of the growth) monotonically increase until the maximum growth rate and risk at “full Kelly” reached. As the rational Kelly fraction opportunity set spans from zero to one, the associated expected instantaneous excess growth rate determines the rational frontier. At rational frontier, expected instantaneous excess growth rate is always the highest achievable given the level of risk or alternatively the level of risk is the lowest achievable given the level or expected instantaneous excess growth rate. What all this means is that Kelly fraction c and instantaneous Sharpe ratio SR_∞ together determine the whole rational expected instantaneous excess growth rate opportunity set (rational frontier) available to a continuous-time world investor.

Thorp (2006, p. 409) derives investment fraction $f = cf^*$ as a function of Kelly fraction c and “full Kelly” investment fraction f^*:

$$f = cf^* = \frac{c(m-r)}{s^2}, \qquad (67)$$

which in our notation corresponds to:

$$f = cf^* = \frac{cm_e}{s_e^2}. \qquad (68)$$

Following Thorp (2006, p. 409), we substitute equations (68) and (65) into equation (8) and find the expected instantaneous excess growth rate as a function of Kelly fraction and instantaneous Sharpe ratio:

$$g_{\infty}^{e}(cf^{*}) = \frac{cm_e}{s_e^2}m_e - \frac{(cm_e/s_e^2)^2 s_e^2}{2} = \left(c - \frac{c^2}{2}\right)\left(\frac{m_e}{s_e}\right)^2$$
$$= c\left(1 - \tfrac{c}{2}\right)SR_{\infty}^{2}. \quad (69)$$

It is easy to interpret from equation (69), and is visualized in Figure 2, that expected instantaneous excess growth rate is at maximum when Kelly fraction $c = 1$, zero when $c = 0$ or $c = 2$ and negative when $c < 0$ (implying short position on stocks) or $c > 2$. The shape of the curve is parabolic and therefore symmetric around the maximum at $c = 1$. The implication is that the rational opportunity set, i.e., the rational frontier of a continuous-time world investor spans from Kelly fraction zero to one. Risk neutral investor, who only cares about maximizing growth rate and is indifferent to risk, in the absence of leverage constraints will always choose full Kelly $(c = 1)$ allocation. For risk averse investor, Kelly fraction $1 < c \leq 2$ is irrational (on irrational frontier) as investor can always obtain the same expected instantaneous excess growth rate from the rational frontier $(0 \leq c \leq 1)$ with lower risk. Naturally Kelly fraction leading to negative expected instantaneous excess growth rate is irrational as well. Negative expected instantaneous excess growth rate implies negative instantaneous geometric risk premium meaning that investor is expected to earn a portfolio growth lower than expected growth of riskless rate.

The above considers the Kelly fraction choice set $0 \leq c \leq 1$ as being the rational frontier for a portfolio. As is evident from Thorp (2008), if the portfolio itself is the instantaneous Sharpe optimal portfolio (portfolio with the greatest possible instantaneous Sharpe ratio), then the rational frontier is also the geometric efficient frontier of a continuous-time world investor who cares about expected growth rate (geometric rate of return), not the rate of (arithmetic) expected return. Thorp (2008) describes the relationship between fractional Kelly strategies and the geometric efficient frontier. Maximum diversification will lead to maximum instantaneous Sharpe ratio implying any portfolio with less than perfect diversification will lie below the parabola of the instantaneous Sharpe optimal portfolio and will be off the geometric efficient frontier. At any targeted level of reward, a portfolio off the geometric efficient frontier will offer a lower reward to risk ratio (expected instantaneous excess growth rate relative to predicted standard deviation of the excess growth rate) compared to what is offered by a portfolio on the geometric efficient frontier.

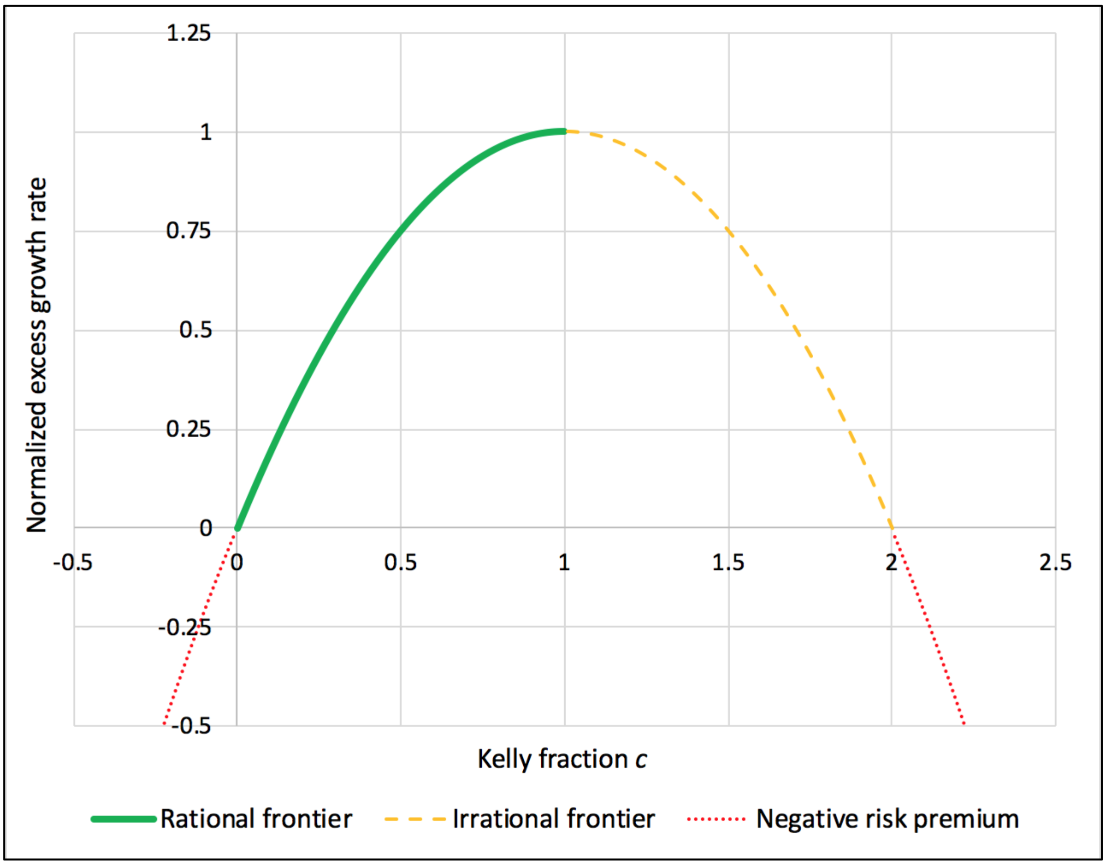

Figure 2. Expected instantaneous excess growth rate normalized by its maximum.

Equation (69) crystallizes that Kelly fraction c and instantaneous Sharpe ratio SR_∞ together determine the whole expected instantaneous excess growth rate opportunity set available to an investor. Rational opportunity set is the part of the whole opportunity set that lies on the rational frontier. Differentiating equation (69) with respect to c leads to maximum expected instantaneous excess growth rate of $\frac{1}{2}SR_\infty^2$ when $c = 1$. This is in line with the result from equation (66). As shown by Thorp (2006, p. 409), we can further calculate the expected instantaneous excess growth rate of fractional Kelly allocation relative to full Kelly allocation:

$$\frac{g_\infty^e(cf^*)}{g_\infty^e(f^*)} = c\left(1-\frac{c}{2}\right)SR_\infty^2/\left(\frac{SR_\infty^2}{2}\right) = c(2-c). \tag{70}$$

Equation (70) implies, and Figure 2 demonstrates, that while risk (predicted standard deviation for the continuously compounding excess growth rate) decreases linearly as Kelly fraction is decreases, the expected instantaneous excess growth rate in relation to full Kelly allocation decreases more slowly and non-linearly. For example, as can be seen from Figure 2, half Kelly allocation entails half of the risk but retains three quarters of the expected instantaneous excess growth rate compared to full Kelly allocation. Because of this diminishing marginal benefit from increasing the risk, the riskier part of the rational frontier, as Kelly fraction approaches one, may not be very attractive to risk averse investor. It is not only pure risk aversion, but the fact that the uncertainty about the future reflected in the parameter estimates, such as expected return, rationalize a conservative Kelly fraction in its own right (Thorp, 2006, pp. 411–412).

By substituting equation (65) into equation (62) and solving SR_∞ we find a connection between the instantaneous Shape Ratio and the standard deviation at full Kelly as given also by Baz and Guo (2017):

$$SR_\infty = f^* s_e = Sdev[G_\infty^e(f^*)]. \tag{71}$$

In other words, as visualized by "the instantaneous Sharpe triangle" in Figure 3, instantaneous Sharpe ratio equals the risk at full Kelly, i.e., the risk at investment fraction delivering the maximum expected instantaneous excess growth rate. That is, instantaneous Sharpe ratio equals the standard deviation of the continuously compounded excess growth of the portfolio G_∞^e levered to full Kelly f^*. Instantaneous Sharpe ratio is equal to the standard deviation at the maximum point on the rational frontier of a continuous-time world investor. Instantaneous Sharpe ratio therefore is equal to the maximum risk rational continuous-time world investor with a given portfolio shall ever take. Finally, instantaneous Sharpe ratio of a maximally diversified Sharpe optimal portfolio is equal to the maximum risk on the geometric efficient frontier which is the maximum risk any rational continuous-time world investor will ever take.

The instantaneous Sharpe triangle in Figure 3 describes how everything is linked and determined by instantaneous Sharpe ratio (0.5 in the figure) when portfolio is levered to the rational maximum risk $s_e f^*$ equaling full Kelly allocation $c = 1$. It is well known from modern portfolio theory that the slope of the arithmetic efficient frontier is equal to Sharpe ratio. This is the case in the instantaneous Sharpe triangle (the slope of the hypotenuse) as well as shown by the arcus tangent function in the figure. In addition to the slope, at full Kelly allocation, rational maximum risk (the length of the horizontal cathetus) is equal to instantaneous Sharpe ratio, expected excess return and the variance of the portfolio (the length of the vertical cathetus) are equal to the square of instantaneous Sharpe ratio and finally the expected excess growth rate equals half of the square of instantaneous Sharpe ratio. This gives us an intuitive understanding what the instantaneous Sharpe ratio means in the continuous-time world and how it completely characterizes the investment opportunity available to a risk neutral investor operating at maximum rational risk and maximum expected reward.

The figure shows how the benefit (the expected excess return of the portfolio) equals the cost (the variance of the portfolio) exactly at full Kelly allocation. The expected excess growth rate, i.e., the geometric risk premium is maximized as marginal cost equals marginal benefit. Exceeding the full Kelly allocation implies that marginal cost, which is a squared function of investment fraction, exceeds marginal benefit, which is a linear function of investment fraction, causing the geometric risk premium to decline.

The figure also displays the difference between a one period world and continuous-time world. The former is bound by Sharpe ratio which determines the slope of the arithmetic efficient frontier but the risk and reward may otherwise be chosen freely and both may grow up to infinity. The continuous-time world is bound by Sharpe ratio in a much more restrictive manner. The rational frontier (the efficient frontier in case of a fully diversified portfolio), and therefore the whole continuous-time world opportunity set for a rational investor exists only within the boundaries of the instantaneous Sharpe triangle. The risk and reward may be chosen freely inside the triangle implying risk may grow up to instantaneous Sharpe ratio and reward up to one half of the squared instantaneous Sharpe ratio.

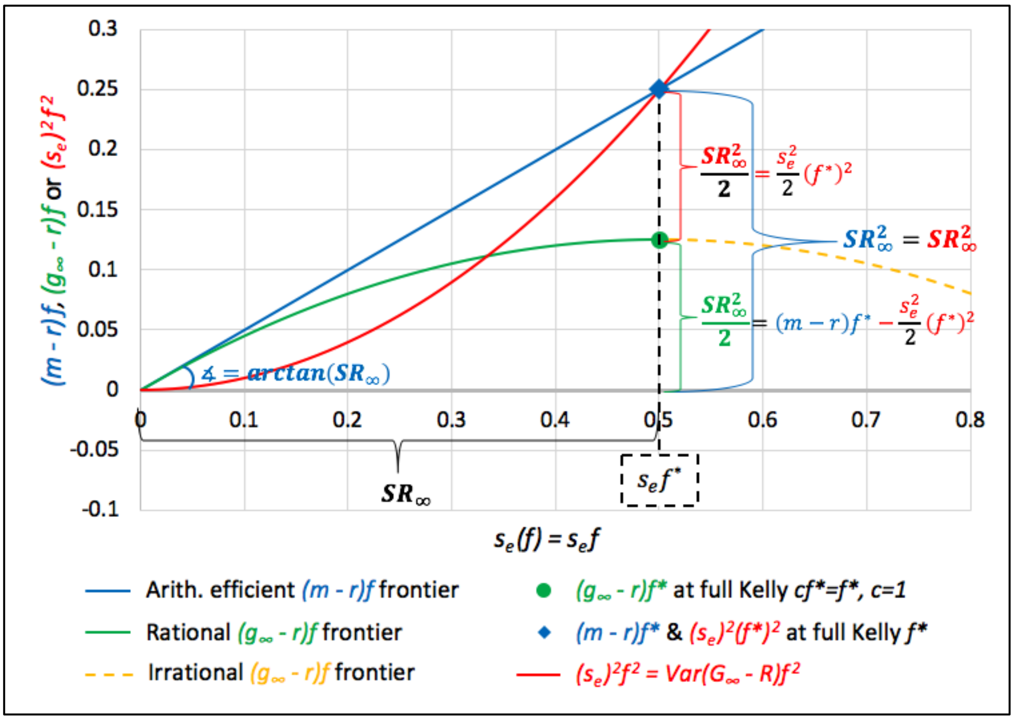

Figure 3. The instantaneous Sharpe triangle.

It is worth noting the surprisingly little-acknowledged fact that Markowitz (1976) himself, the inventor of the assessment of diversification effect in an arithmetic one period model, concluded his paper by cautiously recommending not to present the part of the arithmetic efficient frontier that exceeds the risk level of full Kelly point. His rationale was that a long-term investor should not expose himself to higher short-term variability and lower long-term return than what would be expected below full Kelly point which Markowitz referred to as “Kelly-Latané” point. In Figure 3 Markowitz recommendation would imply ending the blue arithmetic efficient frontier at full Kelly point $s_e f^*$.

We have derived the expected instantaneous excess growth rate as a function relative risk fraction, Kelly fraction c. However, in practical applications we often operate with absolute risk fraction, investment fraction f. To change to absolute risk domain, we first solve c from equation (68):

$$c = \frac{fs_e^2}{m_e} = f\frac{Var(G_\infty^e)}{m_e} = f\frac{s_e}{SR_\infty}, \tag{72}$$

and substitute to equation (69) to obtain the expected instantaneous excess growth rate as a function of investment fraction, standard deviation and instantaneous Sharpe ratio:

$$g_\infty^e(f) = f\left(m_e - \frac{fs_e^2}{2}\right) = fs_e\left(SR_\infty - \frac{fs_e}{2}\right). \tag{73}$$

At full Kelly $(f = f^*)$, substituting equation (71) into equation (73) magically reduces the dependencies down to instantaneous Sharpe ratio and gives equation (66).

Hudson and Gregoriou (2015) show that the arithmetic expected excess returns as the numerator of Sharpe ratio, as originally specified by Sharpe (1966), is often used interchangeably with geometric expected excess returns in the literature. They show that arithmetic and geometric returns are different concepts implying resulting two Sharpe ratios, both theoretically and from practical point of view, are two different metrics and should not be used interchangeably. We agree with Hudson and Gregoriou and derive an instantaneous geometric Sharpe ratio SR_{G_∞} (where subscript G stands for geometric growth and the subscripts subscript ∞ denotes instantaneous geometric growth) with expected instantaneous excess growth rate as the numerator. Instantaneous arithmetic Sharpe ratio is important as it determines the expected instantaneous excess growth rate opportunity set (as shown by equation (69)) for a continuous-time world investor. Instantaneous geometric Sharpe ratio, however, is the metric that better describes the reward to risk trade-off for a continuous-time world investor whose reward is measured as a geometric excess growth rate, not as arithmetic excess return.

Expected instantaneous excess growth rate as a function of Kelly fraction *c* relative to standard deviation of instantaneous excess growth scaled by Kelly fraction gives us the instantaneous geometric Sharpe ratio at a given level of risk. By substituting equations (69) and (71) respectively, we have the instantaneous geometric Sharpe ratio as a function of instantaneous arithmetic Sharpe ratio SR_∞ and relative risk fraction c:

$$SR_{G_\infty}(cf^*) = \frac{g_\infty^e(cf^*)}{cSdev[G_\infty^e(f^*)]} = c\left(1-\frac{c}{2}\right)SR_\infty^2/cSR_\infty$$
$$= \left(1-\frac{c}{2}\right)SR_\infty, \quad (74)$$

which, by substituting equation (72), can be written as a function of instantaneous arithmetic Sharpe ratio SR_∞, standard deviation s_e and absolute risk fraction f:

$$SR_{G_\infty}(f) = SR_\infty - \frac{fs_e}{2}. \quad (75)$$

An important difference between instantaneous arithmetic Sharpe ratio and instantaneous geometric Sharpe ratio is that while the former is constant in relation to risk the latter is a linearly decreasing function of risk. The two Sharpe ratios approach each other when risk approaches zero. Figure 4 illustrates the relation between the two Sharpe ratios as function of relative risk fraction, i.e., Kelly fraction c. In the figure, the Sharpe ratios are normalized by instantaneous arithmetic Sharpe ratio to highlight their relation. Importantly, the greater risk a continuous-time world investor takes, the lower risk adjusted instantaneous excess growth rate he should expect. A one period world investor, on the contrary, is set to receive a constant level of risk adjusted arithmetic excess return no matter how much risk he takes. Clearly excess risk taking and leveraging appears more appealing to a one period world investor or, perhaps more realistically, to an investor who invests in a continuous-time world but assess his strategy in a one period framework.

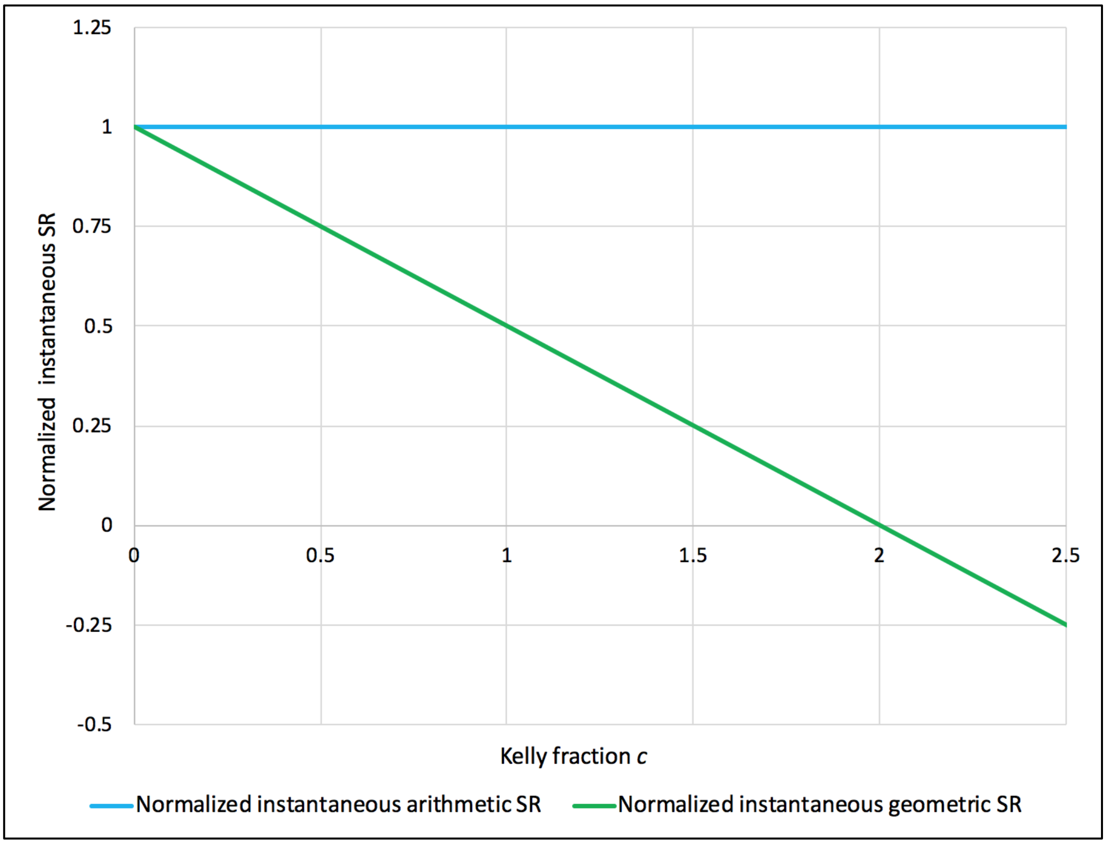

Figure 4. Instantaneous geometric Sharpe ratio in relation to instantaneous arithmetic Sharpe ratio

We showed and visualized in Figure 3 how the instantaneous Sharpe ratio characterizes the investment opportunity available to a risk neutral investor at fully Kelly allocation. However, it is not only the specific full Kelly point, but equations (69) and (73) determine the whole expected instantaneous excess growth rate opportunity set available to a risk averse continuous-time world investor. Rational opportunity set is the rational frontier (the green line in Figure 2 and Figure 3) which is a function of the rational stock allocation opportunity set spanning from Kelly fraction zero to one or, when expressed as an investment fraction, from zero to fraction f^* which, when equation (65) is substituted into equation(62), can be expressed as:

$$f^* = \frac{SR_\infty}{s_e}. \tag{76}$$

The bottom line is, as shown by equation (69), that a rational continuous-time world equity investor only has two parameters to care about: The instantaneous Sharpe ratio SR_∞ and the Kelly fraction c. Maximizing the former involves maximizing diversification and the latter is chosen from the rational stock allocation opportunity set $0 \leq c \leq 1$ based on desired expected reward to predicted risk trade-off.

3.3.5 The Shannon limit as a function of square of Sharpe ratio

Thorp (2006) builds his continuous variable, continuous-time results based on Kelly criterion introduced by Kelly (1956). Kelly in his part, derives Kelly criterion based on information theory provided by Shannon (1948). Therefore, we should study Shannon's work to understand where and how continuous variable and time applications of Kelly criterion originate from.

Shannon's information theory lays the foundations for digital communications. We have a vast literature on information theory and its applications to digital communications and signal processing. We will show how information theory, by applying standard methods from digital signal processing, can be used to explain and analyze the diversification effect in continuous-time world.

In digital communications and signal processing, signal to noise ratio (SNR) is the center of analysis. More specifically, SNR refers to signal power divided by noise power. It is essential to notice that power domain is utilized to assess effect size, not amplitude domain. In finance we often hear that returns are noisy and hear people referring to signal. However, in finance, we don't find such systematic use or accurately quantified definition for SNR as we do in the digital communications engineering. Information theory and the concept of SNR is central in digital communications and signal processing and we see no reason why the same would not be the case in finance.

Sharpe ratio is well known to be a function of diversification. We have shown how Kelly criterion and Sharpe ratio are connected. We have also shown that rebalancing frequency is a determinant of Sharpe ratio. Next, we will show how a square of Sharpe ratio is equivalent to SNR which in turn is a determinant of Shannon's channel capacity, the Shannon limit. In addition, we hypothesize a link between channel bandwidth, the other determinant of channel capacity, and portfolio compounding frequency or compounding bandwidth. We therefore hypothesize a connection between the core concepts in the worlds of finance (Sharpe ratio & compounding) and digital communications (SNR & bandwidth). The common application is information channel capacity which we call compounding process capacity in the world of finance. The compounding process capacity is equivalent to the maximum expected instantaneous excess growth rate, the maximum instantaneous geometric risk premium, achieved at full Kelly allocation. It is the Shannon limit that no rational long-term capital compounder will attempt to exceed. Squared Sharpe ratio implies that the significance of diversification as a determinant of the Shannon limit is raised to second power. It follows that the important concepts of diversification, risk premium, Kelly criterion, Sharpe ratio, compounding frequency, rebalancing frequency, SNR and channel bandwidth appear to be linked and related to the concept of channel capacity, the Shannon limit. Information theory therefore appears applicable to finance, including diversification effect, similarly as it is applicable to digital communications.

Perhaps the most influential result by Shannon (1948) is the determination of information channel capacity, also known as the Shannon limit. Information channel capacity C is the upper theoretical bound for average information transfer rate per channel use (per symbol) between source and target in a given channel and can be written in a general form as show by (Cover & Thomas, 2005, p. 184):

$$C = max\{I(X;Y)\} = I_{MI}^{Max}. \tag{77}$$

The interpretation is that the theoretical maximum average information transfer rate at an arbitrary low error-rate is bound by the maximum mutual information between the source transmitting X and target receiving Y. Exceeding the capacity implies communication errors which in a communications system means lower average information transfer rate than would be achieved at a source rate adjusted to channel capacity. This is necessarily the case regardless how sophisticated the transmitter or the receiver is. (Cover & Thomas, 2005, pp. 6–9; Shannon, 1948).

In equation (**77**) the channel capacity is expressed as information per transmitted symbol. In practice, as shown by Chen (pp. 81–82), the information per symbol is often multiplied by symbol rate f_S expressed as symbols per unit time. The result of this multiplication yields information rate per unit time. In digital communications the maximum $f_S = f_N$ where f_N is Nyquist rate (Proakis, 1995, pp. 13–14). This yields channel capacity expressed as information per unit time (bits per second):

$$C = f_N I_{MI}^{Max}. \tag{78}$$

Shannon (1948) shows that when the noise is white thermal noise (with average power *N*) and we have a limited constant average transmit power (*P*), the channel capacity *C* per unit time for a limited bandwidth *W* is given as:

$$C = W log_2 \left(\frac{P+N}{N}\right). \tag{79}$$

Equation (79) can be written as a function of signal power to noise power ratio *SNR*:

$$C = W log_2 (1 + SNR). \tag{80}$$

Equations (79) and (80) give us the theoretical maximum achievable average information transfer rate as bits per second as the base of the logarithm is two and the unit of the bandwidth in communications systems is hertz which is one per second. We can see that the channel capacity depends only on the communication bandwidth *W* and the *SNR* of the received signal.

Proakis (1995, p. 384) shows that the noise power can be decomposed to components W and power spectral density N_0 for the additive white gaussian noise (AWGN) giving, when substituted to equation (79), the capacity as:

$$C = Wlog_2\left(\frac{P+WN_0}{WN_0}\right) = Wlog_2\left(1+\frac{P}{WN_0}\right). \tag{81}$$

The unit of N_0 is watts per hertz which is equivalent of watt-seconds meaning N_0 is the average noise power per unit of bandwidth. We therefore can define SNR_0 as average signal power to average noise power per unit of bandwidth ratio:

$$SNR_0 = \frac{P}{N_0}, \tag{82}$$

implying we can write equation (81) as:

$$C = Wlog_2\left(1+\frac{SNR_0}{W}\right). \tag{83}$$

Now let's consider channel capacity C_∞ in the theoretical case of infinite communication bandwidth *W*. If *W* is infinite, the only capacity limiting factor in equation (80) is *SNR* and given a constant noise power per unit of bandwidth N_0 the limiting factor becomes the average signal power *P*. Proakis (1995, p. 385) shows that C_∞ is a function of *P* and N_0 and by substituting equation (82) we have:

$$C_\infty = \frac{P}{N_0} log_2\, e = \frac{P}{N_0 \ln 2} = \frac{SNR_0}{\ln 2}, \tag{84}$$

giving the maximum bits/s communication rate when capacity is limited entirely by SNR_0 or, given a constant N_0, more precisely by average signal power *P*.

More generally, as can be seen from equation (83), when SNR_0 is small meaning average signal power *P* is small in relation to N_0, the effect from sub-dividing the power to smaller transferrable chunks by shortening the transmission time interval (*TTI*) and proportionally increasing the communication bandwidth *W* becomes small.

In this case we are said to be in a power-limited regime of channel capacity (Tse & Viswanath, 2005, p. 174). The opposite occurs when SNR_0 is large as a result of average signal power P being large in relation to N_0, in which case we are in a bandwidth-limited regime (Tse & Viswanath, 2005, p. 174).

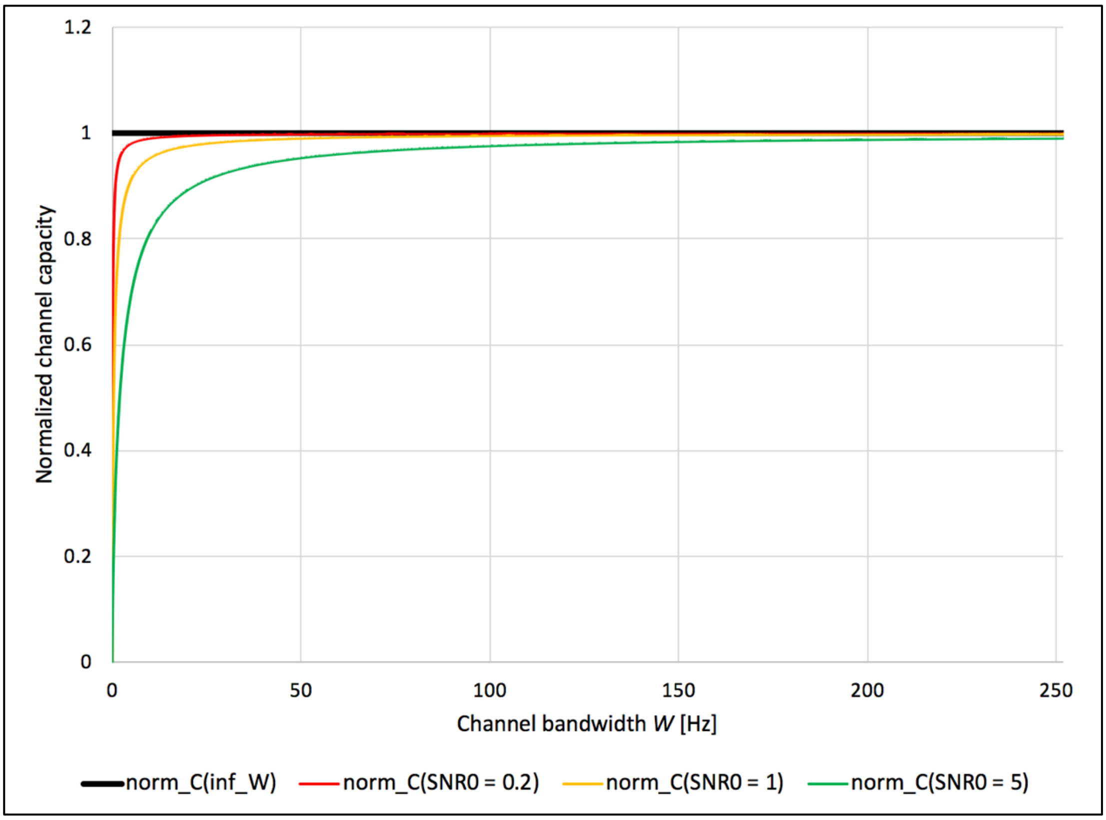

Figure 5. Normalized channel capacity C with infinite W in relation to normalized C as a function of SNR_0 (adapted from Tse & Viswanath, 2005, p. 174).

Figure 5 illustrates how average signal power to average noise power per unit of bandwidth ratio SNR_0 affects the channel capacity sensitivity to channel bandwidth W. Black horizontal line is the channel capacity normalized by itself representing the maximum capacity that is achieved as channel bandwidth W approaches infinity. We can see how at the lowest SNR_0 value 0.2 (red line) channel capacity very quickly approaches capacity limit when the channel bandwidth starts increasing from zero. This implies that channel capacity very quickly enters power-limited region and channel bandwidth has little effect. On the other hand, when SNR_0 is substantially higher at 5 (green line), the effect from channel bandwidth is retained longer as the bandwidth increases. Middle SNR_0 value 1 (orange line) yields a bandwidth sensitivity between the extreme SNR_0 values.

We now consider digital communications systems specific parts in channel capacity equations (83) and (84). By removing the impact of the digital communications systems specific parts from these equations, we intend to find a more general form expressing channel capacity. Proakis (1995, pp. 13–14) gives Nyquist rate f_N:

$$f_N = 2W, \tag{85}$$

which is the maximum symbol rate per second that can be utilized in bandwidth-limited communications systems. Simultaneously Nyquist rate is the lowest sampling frequency avoiding frequency aliasing (Proakis, 1995, p. 72). This means, as can be also seen from Chen (pp. 81–82), that we can write equation (78) by decomposing the symbol rate f_N:

$$C = f_N I_{MI}^{Max} = 2W I_{MI}^{Max}, \tag{86}$$

where the multiplier 2 is a specific feature related to communication systems allowing for two symbols per carrier wavelength to be transmitted. Additionally, in the context of finance, we don't transmit bits but our interest is in the expected continuously compounded excess growth rate which implies "transmitting" nats, i.e., utilizing a logarithm with base *e*. To compensate for these digital communications specific features in equation (84), we change the base of the logarithm to *e*, divide C_∞ by two (to remove the effect from Nyquist rate) and finally substitute equation (82) to find a more general (G) form for C_∞^G:

$$C_\infty^G = \frac{1}{2}\frac{P}{N_0}\ln e = \frac{P}{2N_0} = \frac{SNR_0}{2}, \tag{87}$$

giving the maximum average generalized information rate equal to half of the signal to noise per unit of bandwidth ratio expressed in units of nats/(1/unit of bandwidth).

The general equation (77) states that channel capacity is equal to maximum mutual information. We know from equation (30) that expected instantaneous excess growth rate g_∞^e is equal to mutual information and from equation (66) that g_∞^e finds its maximum, one half of the instantaneous Sharpe ratio squared, at full Kelly when investment fraction is f^*. Furthermore, as equation (66) concerns a continuously compounded equally weighted portfolio, we consider the compounding (and rebalancing) frequency as infinite. We therefore have channel capacity or a better describing term compounding process (CP) capacity C_∞^{CP} for the theoretical case of infinite compounding (and rebalancing) frequency:

$$C_\infty^{CP} = g_\infty^e(f^*) = \frac{SR_\infty^2}{2}. \tag{88}$$

By squaring equation (65) we have:

$$SR_\infty^2 = \frac{m_e^2}{s_e^2} = \frac{m_e^2}{Var(X^e)} = \frac{m_e^2}{Var(G_\infty^e)}, \tag{89}$$

Where $Var(X^e)$ is the variance of the excess return and $Var(G_\infty^e)$ is the variance of the excess growth. These two variances are equal when we have infinitely short compounding and rebalancing interval, i.e., instantaneous metrics.

We can consider the instantaneous expected excess return m_e as the signal we are transmitting and the standard deviation of the excess return (and excess growth) s_e as the noise. This implies square of the instantaneous expected excess return m_e^2 is the signal power and square of the standard deviation s_e^2 is the noise power. At the receiving end we see m_e^2/s_e^2 as the signal to noise ratio for the realized instantaneous arithmetic excess return $\check{m}_e$ after the communication channel. The check accent denotes "realized". Using yearly values for m_e and s_e, we have the average signal power to average noise power per unit of compounding frequency ratio for realized instantaneous arithmetic excess return:

$$SNR_0^{\widetilde{m}_e} = SR_\infty^2. \tag{90}$$

In case the rebalancing frequency is lower than infinite, we don't have instantaneous Sharpe ratio but conventional Sharpe ratio and the noise term s_e in Sharpe ratio is: $s_e = Sdev(X^e) \neq Sdev(G_\infty^e)$ implying we theoretically must use the standard deviation of the excess return, not the standard deviation of the logarithmic excess return. To be consistent with the definition of the conventional Sharpe ratio, we also need to use periodically compounding rate (instead of instantaneous rate) for the realized arithmetic excess return. The equality then becomes:

$$SNR_0^{\widetilde{m}_e} = SR^2. \tag{91}$$

The interpretation is that the square of Sharpe ratio is a SNR metric for realized excess return. It is worth noting here that Treynor and Black (1973) preferred a square of Sharpe ratio over the traditional ratio. Treynor and Black therefore were in favor of using a power domain metric over an amplitude domain metric as is conventional when applying information theory in digital communications.

As a result, equations (90) and (91) state that a square of Sharpe ratio is a SNR and equation (88) establish a square of instantaneous Sharpe ratio divided by two as the compounding process capacity C_∞^{CP}. We therefore can write the compounding process capacity for infinite compounding (and rebalancing) frequency as:

$$C_\infty^{CP} = g_\infty^e(f^*) = RP_{EW,G_\infty}(f^*) = \frac{SR_\infty^2}{2} = \frac{SNR_0^{\widetilde{m}_e}}{2}, \tag{92}$$

which is the perfect financial world counterpart for the generalized digital communications channel capacity as given by equation (87). Equation (92) therefore gives us the maximum average excess growth rate equal to half of the signal to noise per unit of compounding frequency ratio expressed in units nats/year (continuously compounded excess growth per year) equaling the maximum instantaneous geometric risk premium achievable at full Kelly allocation.

Similarly, as with channel capacity, exceeding the compounding process capacity necessarily implies lower expected instantaneous excess growth rate than would be achieved at a portfolio stock allocation level, meaning investment fraction *f* or Kelly fraction *c*, adjusted to compounding process capacity $f = f^*$ or $c = 1$.

Earlier we regarded the instantaneous geometric Sharpe ratio SR_{G_∞} as a better measure of reward to risk trade-off for a continuous-time investor compared to its arithmetic counterpart SR_∞. Following the same procedure as for $SNR_0^{\breve{m}_e}$ in equation (90), we can calculate *SNR* for the realized instantaneous excess growth rate $\breve{g}_\infty^e$ by considering the expected instantaneous excess growth rate g_∞^e as the signal instead of the expected instantaneous excess return m_e. The average signal power to average noise power per unit of compounding frequency ratio for realized instantaneous geometric excess growth rate then is:

$$SNR_0^{\breve{g}_\infty^e} = SR_{G_\infty}^2. \tag{93}$$

Intuitively, we hypothesize compounding frequency as a counterpart for channel bandwidth *W*. Both are frequencies, the former is expressed in units of 1/year while the latter is expressed as 1/second, i.e., as hertz. We can compensate for the digital communications specific features in equation (81), similarly as we did for equation (84) to acquire a more general equation (87), by changing the base of the logarithm to e and by dividing by two:

$$C^G = \frac{W}{2} ln\left(1 + \frac{P}{WN_0}\right) = \frac{W}{2} ln\left(1 + \frac{SNR_0}{W}\right). \quad (94)$$

We then substitute channel bandwidth W with compounding (C) frequency or compounding bandwidth W_C, average signal power P with m_e^2, average noise power N_0 with s_e^2 and SNR_0 with $SNR_0^{\widetilde{m}_e}$ giving compounding process capacity C^{CP} as a function of compounding bandwidth W_C:

$$\begin{aligned} C^{CP} &= \frac{W_C}{2} ln\left(1 + \frac{m_e^2}{W_C s_e^2}\right) \\ &= \frac{W_C}{2} ln\left(1 + \frac{SR^2}{W_C}\right) \\ &= \frac{W_C}{2} ln\left(1 + \frac{SNR_0^{\widetilde{m}_e}}{W_C}\right), \end{aligned} \quad (95)$$

where the Sharpe ratio is conventional Sharpe ratio instead of instantaneous Sharpe ratio.

In the discussion related to equation (30) we called the mutual information in the investing context as mutual excess capital growth which we now denote as I_{MG}. Maximum average I_{MG} is denoted as I_{MG}^{Max}. Mutual excess capital growth is measured as nats per compounding period T_C. Compounding period is the counterpart for symbol. Compounding rate (the counterpart for Nyquist rate) is denoted as f_C and is measured as T_C per year. As visualized in Figure 1, compounding process feeds back its output into input. There is one indivisible multiplicative serial compounding process implying exactly one compounding period at a time occupying the compounding process. Compounding rate therefore must be equal to compounding frequency implying $f_C = W_C$. High level version of equation (95) then can be written as:

$$C^{CP} = f_C I_{MG}^{Max} = W_C I_{MG}^{Max}. \tag{96}$$

Building on the hypothesis that compounding frequency is the counterpart for channel bandwidth W, we can interpret Figure 5 in the context of investing. Annual Sharpe ratio of about 0.45 can be considered typical for a stock market. 0.45 squared yields a $SNR_0^{\tilde{m}_e}$ of about 0.2 which corresponds to red line in the figure. Daily compounding implies compounding bandwidth $W_C = 252$ (assuming 252 trading days per year). We can see from the figure that daily compounding yields a capacity very close (99.96%) to maximum capacity achieved at infinite bandwidth which would correspond to capacity of continuously compounded stock portfolio. Empirically, we can expect the compounding frequency to be higher than daily as stocks are priced continuously in the markets. The relevant timeline for pricing the most liquid stocks is fractions of a second and even the most illiquid stocks presumably don't fall far from the daily pricing frequency. Albeit the markets aren't always open, it seems safe to assume that practical compounding process capacity is entirely in the power-limited regime where compounding bandwidth W_C doesn't play any practical role. This implies that the practical compounding process capacity is entirely determined by the (square of) instantaneous Sharpe ratio.

The square of Sharpe ratio is a function of noise power. We therefore compare the noise power component in digital communications to noise power in investing. In digital communications the noise power per unit of bandwidth, the power spectral density N_0, of the underlying thermal noise is constant with regard to channel bandwidth but is a function of temperature. In investing context, the equivalent noise power per unit of compounding bandwidth, variance s_e^2, is a function of rebalancing (RB) frequency or rebalancing bandwidth W_{RB}. Variance $s_e(W_{RB})^2$ approaches its minimum as rebalancing frequency approaches infinity.

Theoretically, as the square of Sharpe ratio is a function of rebalancing frequency, a more precise form for equation (95) determining compounding process capacity C^{CP} is given as a function of rebalancing frequency W_{RB}:

$$C^{CP} = \frac{W_C}{2} ln\left(1 + \frac{m_e^2}{W_C s_e(W_{RB})^2}\right)$$
$$= \frac{W_C}{2} ln\left(1 + \frac{SR(W_{RB})^2}{W_C}\right) \tag{97}$$
$$= \frac{W_C}{2} ln\left(1 + \frac{SNR_0^{\widetilde{m}_e}(W_{RB})}{W_C}\right).$$

As discussed earlier, however, practical rebalancing frequencies, like monthly or even yearly, don't significantly change the expected capital growth rate compared to theoretical infinite rebalancing frequency. The impact of W_{RB} therefore can be considered negligible at practical rebalancing frequencies and with typical stock market parameters. It is not clear, however, if less typical parameters associated, e.g., with different investment styles or aggressive use of leverage change the impact of rebalancing frequency.

As a sanity check, we can rearrange the equation (97):

$$C^{CP} = \frac{1}{2} SR(W_{RB})^2 \ln\left(1 + \frac{SR(W_{RB})^2}{W_C}\right)^{W_C/SR(W_{RB})^2}. \tag{98}$$

In the limit, when $W_C = f_C \to \infty$ and $W_{RB} \to \infty$, implying rebalancing bandwidth denoted as W_{RB_∞}, we substitute the Euler's number identity given in equation (15) into equation (98) and have:

$$C_\infty^{CP} = \frac{SR_\infty\left(W_{RB\infty}\right)^2}{2} = \frac{SR_\infty^2}{2} = \frac{SNR_0^{\widetilde{m}_e}}{2},$$

which is identical to equation (92) as we expect. We note that in the limit the compounding rate f_C or the equaling compounding frequency/bandwidth W_C reduce away from the compounding process capacity formula.

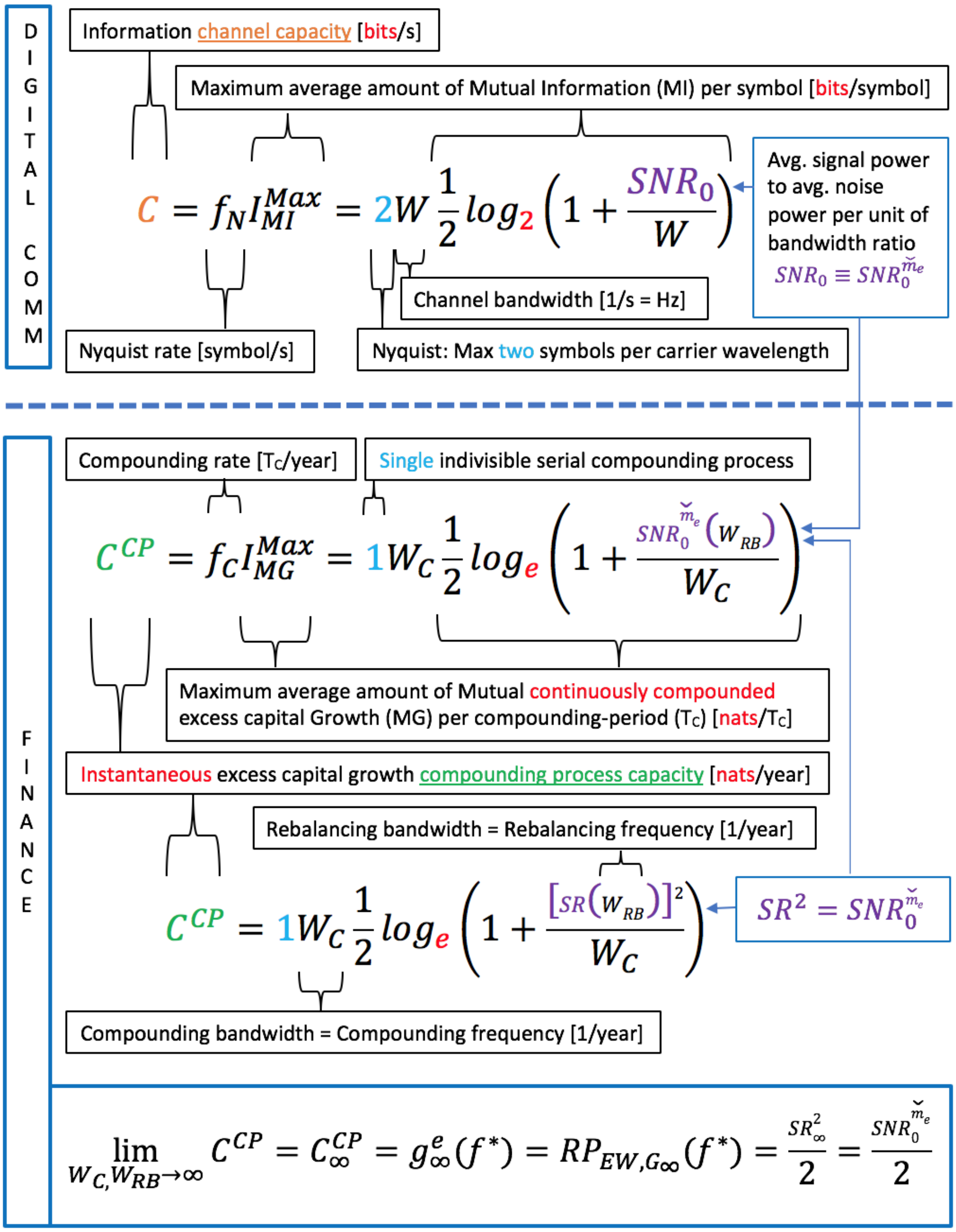

Figure 6. The Shannon limit in digital communications and in finance.

In Figure 6 we attempt to summarize what we have hypothesized about the connection between channel capacity in digital communications and compounding process capacity in the context of finance and investing. There are two main differences. The first difference is that digital communications measure bits per second while in investing we care about capital growth measured as continuously compounded yearly growth rate. Hence the base of logarithm is 2 and *e* for digital communications and finance respectively. The second main difference is that Nyquist allows packing two symbols into channel per carrier wavelength implying symbol rate twice the carrier bandwidth. In compounding there is no such trick and we need to settle for compounding rate equal to compounding frequency. On a more positive note, we can utilize the compounding bandwidth up to infinity as there is no shared resource like radio channel bandwidth in communication systems.

3.3.6 Further illustrative analogies between finance and digital communications

We will demonstrate the applicability of information theory and the concept of channel capacity, the Shannon limit, to finance by showing four analogies with digital communications: 1) diversification effect on investment portfolio and digital data transfer, 2) mutual information as a determinant of utility both in investing and data transfer, 3) correspondence of the channel capacity and compounding process capacity and similarities in the options of exploiting the capacity and 4) the correspondence, and thereby the apparent absurdness, between utilizing a single period investment assessment framework in a continuous-time world and assessing a data transfer based on transmitted, instead of received, data rate.

The first analogy relates to diversification where we show how additional diversifying receiver (RX) antennas in a mobile phone are the counterpart for additional diversifying stocks in an investment portfolio. The former increases channel capacity, the latter compounding process capacity.

It is important to notice that the compounding process capacity is a function of diversification. This is because the compounding process capacity is a function SNR which is a function of Sharpe ratio. Denominator of Sharpe ratio is standard deviation which is a function of diversification. Exactly the same phenomenon is observed and utilized in wireless digital communications where the equivalent of portfolio diversification is called RX-diversity. RX-diversity means utilizing more than one receiver antenna placed adequately apart from other antennas to ensure low cross-correlation.

In mobile phones RX-diversity suppresses the non-coherently additive noise in relation to coherently additive signal implying higher SNR leading to higher channel capacity. It is beneficial to utilize diversification as long as the marginal benefit exceeds the marginal cost. In a mobile phone this limit is quickly reached as both the physical limits and economical cost/benefit considerations limit the attractiveness of increasing channel capacity by additional diversity. However, in case of portfolio diversification, given modern cost-effective diversification technology (index funds, ETFs), there is practically no limit allowing for thousands of diversifiers in comparison to typically two to four in case of modern mobile phones. In the absence of the physical limitations and given a similar cost structure as in portfolio diversification today there is no question typical mobile phones would be equipped with vastly higher number of RX-antennas.

The second analogy with digital communications relates to utility. We will show how the utility from a data download is analogous to the utility function of a continuous-time world risk neutral investor. Additionally, we will show how the utility from a data download and the utility from investing in a continuous-time world both are functions of mutual information.

In digital communications, when downloading a large data file, the utility is the average received data rate which is the mutual information between the transmitter and the receiver. In case of errors in the communication channel the communications protocol ensures the erroneous part is quickly retransmitted meaning there is no harm from errors as long as they don’t compromise the received data rate. This means there is no risk associated with the proportion of channel capacity utilized for maximizing the information rate. In the absence of risk, the only rational choice is at all times to target utilizing the whole channel capacity maximizing the average received data rate, the utility.

We can consider data download being “risk neutral” in the spirit of a continuous-time world risk neutral investor whose utility is the average excess growth rate (mutual information) of his capital and who is indifferent to risk. For risk neutral investor the only rational choice is at all times to target utilizing the whole compounding process capacity, by investing at full Kelly allocation, maximizing the average excess growth rate of his capital. Risk averse investor cares about the average excess growth rate (mutual information) but also the risk. Therefore, risk averse investor will not attach his utility entirely to mutual information but also incorporates the associated risk. Risk averse investor takes the compounding process capacity as an upper bound for the portfolio growth rate, not as a target, when choosing the reward to risk tradeoff by choosing his investment fraction. Nevertheless, mutual information, analogously to digital communications, is part of the risk averse continuous-time world investor's utility function.

Third analogy is about the correspondence of the channel capacity of digital communications and the compounding process capacity that we have formulated. Channel capacity is well known to set a hard limit to maximum achievable information rate. However, channel capacity can alternatively be exploited to increase communication reliability which corresponds to decreasing risk in the context of investing. We will show how maximizing compounding process capacity should be the primary goal of any investor regardless of the level or existence of risk averseness.

Besides data download, there are other use cases in digital communications. Take Ultra-Reliable Low-Latency Communication (URLLC) in 5G standard for cellular communications as an example. URLLC is designed for communication with very high communication reliability (low error rate) and very low latency requirement to allow, e.g., self-driving cars to operate safely. Failing to deliver very low error rate and low latency may in such context lead to severe consequences. In other words, short-term risk becomes significant and maximizing the data rate no longer serves as the sole driver for utility. Now adequately high data rate together with very low error-rate and low latency (very low risk) constitutes the utility function. We can think URLLC use case as the counterpart for risk averse investor. Both care about sufficiently high reward delivered with sufficiently low risk.

In digital communications it is possible to increase communication reliability (decrease communication error rate) by transmitting redundant information bits along with systematic (non-redundant) information bits. The ratio between systematic information bits and total (systematic plus redundant) transmitted bits is called code rate. Increasing redundancy by adjusting code rate lower leads to lower information rate but also lower error rate. In terms of finance, this corresponds to lower reward (geometric risk premium) but also lower risk, which is achieved by lowering investment fraction. Adjusting code rate in digital communications corresponds to adjusting investment fraction in the world of investing. Theoretical maximum code rate is one (all systematic information bits) while theoretical maximum investment fraction is infinite. Both of the theoretical maximums are achieved when SNR, and consequently channel capacity, is infinite. Assuming finite signal power (finite squared excess growth rate), infinite SNR occurs when channel is perfect meaning completely noiseless (riskless) which corresponds to arbitrage in finance. Furthermore, the code rate that achieves the maximum data rate is the code rate (information allocation) corresponding to full Kelly allocation f^* in investing.

It is important to notice that channel capacity can be materialized as maximizing the information rate, as minimizing the error rate or as a combination of the two. The greater the channel capacity the greater the potential for maximizing information rate and/or minimizing the communication error rate. It follows that no matter what is the use case, the greater the channel capacity the better the outcome. Maximizing the channel capacity therefore is in the core of digital communications. Same applies to finance. No matter whether the investor is risk neutral or risk averse, the primary goal should always be maximizing the compounding process capacity and then utilizing that capacity to maximize the expected excess growth rate, to minimize the risk associated to excess growth rate or any combination of the two. In finance, the recipe for maximizing the compounding process capacity is simple: maximize Sharpe ratio by maximizing diversification. The second step after maximizing the capacity is to select the desired expected reward to risk tradeoff (average data rate to error rate tradeoff in digital communications) level by choosing the investment fraction (code rate in digital communications).

In the fourth analogy we will argue that projecting realized multiperiod returns for the future by utilizing a single period framework, which is the standard in finance, is akin to assessing a data download performance in digital communications based on average transmitted, not received, data rate.

In a data download, the received data rate, the mutual information, is equal to transmitted data rate if the channel is perfect, i.e., noise power is zero or alternatively if the transmit power has no constraint (Cover & Thomas, 2005, p. 261). Both cases imply the received signal SNR, and therefore the capacity, is infinite. We focus on the former case, the perfect channel. We can see the same in the context of investing. We know from equation (30) that the expected instantaneous excess growth rate, as given by equation (8), is equal to mutual information. From equation (8) we can see that the expected instantaneous excess growth rate, the mutual information, is equal to expected arithmetic single period return, the "transmitted" rate, when the standard deviation s_e is zero meaning the noise power in the compounding process, the channel, is zero.

In which instance then the standard deviation s_e, the risk, is zero? Naturally we can't expect to earn a risk premium without a risk implying s_e cannot be zero due to portfolio risk being zero. This rule out the possibility of a perfect noiseless compounding process (channel). However, the s_e experienced by the compounding process (the channel) can be zero if we are in a single period world that encompasses no compounding. This case also appears in our analysis if we invest in a continuous-time world but utilize a single period framework to evaluate our investment strategy or to project future investment outcomes. Projecting realized multiperiod returns for the future by utilizing a single period framework therefore is akin to assessing a data download performance in digital communications based on average transmitted, not received, data rate. Communication is mutual information and assessing communication based on how high the source rate is would be considered absurd.

As a curiosity, the standard deviation s_e experienced by the compounding process can be zero also if the beginning of period capital, which is input to compounding process, is guaranteed to have zero variance. This case corresponds to example given by Kelly (1956) describing a gambler whose wife limits his gambling funds to exactly one dollar per weekly bet placed on an event yielding the highest expected value. Constant gambling budged per bet implies that returns are not reinvested and ensures zero variance for each beginning of period capital. In this example, the gambler will enjoy arithmetic, instead of geometric, mean returns, but cannot enjoy from the effect of compounding capital. To enjoy the benefits of compounding, one has to reinvest the returns and bear the risk s_e in the compounding process implying, assuming rational investor, one has to operate below or at the compounding process capacity in accordance with his subjective risk appetite.

3.3.7 Derivation of realizable risk premium

We have established the significance of SNR in the context of expected excess growth rate and related metrics such as Sharpe ratio. Next, we show how R-squared metric can be expressed as a function of SNR or Sharpe ratio making it a SNR-based risk adjustment scale value. This risk adjustment scale value, ranging from zero to one, then can be used as a multiplier to express instantaneous geometric risk premium, the expected instantaneous excess growth rate, in risk adjusted terms. The interpretation of our R-squared is the proportion of portfolio realized instantaneous geometric risk premium (realized instantaneous excess growth) explained by instantaneous geometric risk premium (expected instantaneous excess growth). We then multiply instantaneous geometric risk premium (expected instantaneous excess growth) by R-squared, implying we derive an instantaneous geometric risk premium weighted by its average realizable proportion (the average proportion of realized instantaneous excess growth explained by its expectation), hence the term "realizable risk premium".

We will derive a formula for realizable instantaneous geometric risk premium and show that it is a function of time. This is important as it allows us to mathematically capture the well-known property that realized portfolio growth rate converges towards its expectation as investment time horizon increases. We will also show that realizable risk premium can be used as a risk adjusted risk premium metric which accounts for two dimensions: expected growth rate and the associated risk, but expresses itself in one intuitive dimension as a growth rate. Importantly, realizable risk premium is a function of diversification. We argue the metric can be considered as a diversification effect measure for risk averse continuous-time world investor. However, we argue realizable risk premium should preferably be used as a short-term measure as it does not account for the long-term compounding effect for wealth. The absolute value of typical realizable risk premium will be very small. Realizable risk premium is therefore best presented as basis points (bps) which corresponds to one hundredth of a percent. To convert the result into bps, the realizable risk premium given by the formula needs to be multiplied by ten thousand. We also define and derive a minimum investment time horizon for a stock portfolio. The minimum investment horizon is defined as the time required for SNR of the realized instantaneous excess growth rate to exceed one (zero decibels). This corresponds to time required for realizable risk premium to exceed half of the (expected) risk premium which is the point in time where less than half of the realized instantaneous excess growth rate constitute noise. Importantly, the minimum investment time horizon for a stock portfolio is a function of diversification.

Instantaneous excess growth rate will converge to its expected value as investment horizon approaches infinity. Near certainty at infinity, for any investor, is an awfully distant goal. Practical long-term investor considers long-term ranging from years to decades. Risk neutral investor is indifferent to risk, but most investors are risk averse and value lower risk over higher risk. Risk neutral investor will value the expected reward alone and is drawn to full Kelly allocation maximizing the continuous-time world investor expected reward and indirectly also maximize the rational risk to which he is indifferent to. Risk averse investor, however, faces a more complex choice-set as he values both high reward and low risk which are contradicting goals. In order to make risk averse investor choice easier to analyze, we transform the choice-set domain from reward & risk domain to risk adjusted reward domain, in other words, realizable risk premium domain. The trick is that risk adjustment is a unitless scale value implying the choice-set domain is transformed from two dimensions (reward & risk) to one dimension (realizable risk premium). This is in the spirit of what Statman (1987; 2004) did when he expressed the diversification benefit as a return difference by transforming a standard deviation difference to return difference which is easier and more intuitive to interpret. In our case the transformation simplifies the interpretation even more as it reduces the dimensions of the metrics that need to be interpreted in addition to focusing on the more intuitive metric of the two. The remaining one dimension, the realizable risk premium, is the expected instantaneous excess growth rate scaled (multiplied) by a SNR-based scale value, a coefficient of determination, R-squared.

Next, we derive the coefficient of determination, R-squared, as a function of SNR. Equation (**55**) gives us R-squared as a function of portfolio idiosyncratic variance relative to benchmark variance. Variances are power-metrics and power relatives make a SNR. Idiosyncratic variance is the noise power while benchmark variance represents the signal power. We can think of the benchmark rate of excess returns as a predictor signal for portfolio rate of excess returns, idiosyncratic excess returns as noise and their squared relation as signal to noise ratio of the benchmark rate of excess return as a portfolio rate of excess return predictor signal. We therefore have a general form for R-squared as a function of SNR:

$$R^2 = \frac{1}{1+Ivar_P/Var_{BM}} = \frac{1}{1+1/SNR}. \tag{99}$$

When R-squared is defined as above, both signal and noise being a random variable, both signal and noise accumulate non-coherently as a function of time implying SNR or R-squared, as defined in equation (99), are not functions of time.

We showed earlier in equation (90) how the average signal power to average noise power per unit of compounding frequency ratio for realized arithmetic rate of excess return $SNR_0^{\breve{m}_e}$ is a function of squared instantaneous Sharpe ratio SR_∞. In case of a Sharpe ratio, the signal is the expected rate of excess return m_e of a portfolio, which is not a random variable. We therefore don't use variance but square the expected rate of excess return to obtain signal power. Noise power is the variance of the portfolio's predicted excess growth rate $s_e^2 = Var(X^e) = Var(G_\infty^e)$. When considering expected rate of excess return of a portfolio as a predictor for realized portfolio rate of excess returns, we can write equation (99) as a function of instantaneous Sharpe ratio $SR_\infty(t)$ or as a function of average signal power to average noise power per unit of compounding frequency ratio for realized arithmetic rate of excess return $SNR_0^{\breve{m}_e}(t)$. As the signal component is an expectation, not a random variable, it accumulates coherently while noise accumulates non-coherently. This implies both the SR_∞ and $SNR_0^{\breve{m}_e}$ are now functions of time:

$$R(t)^2 = \frac{1}{1 + 1/SR_\infty(t)^2} = \frac{1}{1 + 1/SNR_0^{\widehat{m}_e}(t)}$$
$$= \frac{1}{1+1/\left(tSR_\infty^2\right)} = \frac{1}{1+1/\left(tSNR_0^{\widehat{m}_e}\right)}. \tag{100}$$

In this case, as SR_∞ and $SNR_0^{\widehat{m}_e}$ are functions of time, R-squared is a function of time and interpreted as the proportion of realized arithmetic rate of excess return determined by the expected arithmetic rate of excess return within a portfolio.

Alternatively, we can use the instantaneous geometric Sharpe ratio $SR_{G_\infty}(t)$ and its relation to average signal power to average noise power per unit of compounding frequency ratio for realized instantaneous geometric excess growth rate $SNR_0^{\breve{g}_\infty^e}(t)$ as defined in equation (93). R-squared then becomes:

$$R(t)^2 = \frac{1}{1 + 1/SR_{G_\infty}(t)^2} = \frac{1}{1 + 1/SNR_0^{\breve{g}_\infty^e}(t)}$$
$$= \frac{1}{1+1/\left(tSR_{G_\infty}^2\right)} = \frac{1}{1+1/\left(tSNR_0^{\breve{g}_\infty^e}\right)}. \tag{101}$$

In this case, R-squared is a function of time and interpreted as the proportion of realized instantaneous excess growth rate determined by the expected instantaneous excess growth rate within a portfolio.

Assuming time series for excess returns are not correlated, SNR in equations (100) and (101) are perfectly additive as a function of time meaning SNR scales linearly with slope one with time. It is therefore obvious from these equations that as time approaches infinity, SNR approaches infinity and R-squared approaches one. In other words, in the very long-term, portfolio expected rate of excess return or growth approaches as being a perfect predictor of portfolio realized rate of excess return or growth meaning the proportion determined by the expectation approaches one, which corresponds to 100%. It is not as obvious, but can be shown, that as time approaches zero, R-squared approaches SNR. In a practical short term, SNR value, which now approximates R-squared, always falls to the range from zero to one corresponding to percentage range from 0% to 100%. In other words, in the short-term, portfolio expected rate of excess return or growth is as good a predictor for portfolio realized rate of excess return or growth as its SNR as a predictor signal is. This means that the practical SNR-based scale value, a coefficient of determination, R-squared ranges from about value equal to SNR in the short term to about one in the very long-term. Practical short-term here is, e.g., a month while the long-term typically is in minimum several decades or much more depending on what the annualized SNR is.

We derive realizable instantaneous geometric risk premium $\widetilde{RP}_{EW,G_\infty}$ equaling realizable expected instantaneous excess growth rate $\tilde{g}^e_\infty(t)$, where the tilde accent denotes "realizable", for a portfolio as expected instantaneous excess growth rate g^e_∞ scaled (multiplied) by the proportion of realized instantaneous excess growth rate determined by the expected instantaneous excess growth rate which is the R-squared $R(t)^2$ from equation (101):

$$\begin{aligned}\widetilde{RP}_{EW,G_\infty} &= R(t)^2 RP_{EW,G_\infty} \\ &= \frac{RP_{EW,G_\infty}}{1+1/SR_{G_\infty}(t)^2} = \frac{RP_{EW,G_\infty}}{1+1/SNR_0^{\widetilde{RP}_{EW,G_\infty}}(t)} = \\ \tilde{g}^e_\infty(t) &= R(t)^2 g^e_\infty \\ &= \frac{g^e_\infty}{1+1/SR_{G_\infty}(t)^2} = \frac{g^e_\infty}{1+1/SNR_0^{\tilde{g}^e_\infty}(t)},\end{aligned} \tag{102}$$

where t is investment time horizon.

By using R-squared, which is a function of time, as a scale value for expected instantaneous excess growth rate, we determine an realizable expected instantaneous excess growth rate which is a function of time. Notice that expected instantaneous excess growth rate g^e_∞ is not a function of time. By measuring risk adjusted (realizable) rewards, in case of two portfolios with identical expected instantaneous excess growth rates, risk averse investor will choose the one with lower risk, i.e., the one with higher realizable expected instantaneous excess growth rate. Furthermore, in case of perfectly and imperfectly diversified portfolio the former has lower expected growth rate but also higher standard deviation. The difference in attractiveness between the portfolios therefore must be greater than the difference in expected growth rates alone. Realizable risk premium difference will reflect both the difference in expected growth rate and the difference in risk. Realizable risk premium difference therefore can be considered a diversification effect measure for risk averse continuous-time world investor.

Second scenario is a case of one portfolio but two alternative investing time horizons. In this scenario, the realizable reward is higher for longer time horizon implying an investor with fixed risk adjusted reward target may increase his investment fraction (stock allocation) when the time horizon increases. At the extreme, when investment time horizon approaches infinity, the probability for the realized growth rate deviating from the expected growth rate approaches zero and it is intuitively clear we should weight expected instantaneous excess growth rate of the portfolio very highly in relation to the predicted risk of the portfolio. In this extreme case the scale factor approaches one implying the realizable reward approaches the reward without risk adjustment, i.e., the expected instantaneous excess growth rate of the portfolio.

There is yet another scenario where realizable reward can be applied. This third scenario considers realizable reward of a portfolio as a function of stock allocation measured, e.g., as Kelly fraction. We can write the realizable expected instantaneous excess growth rate in equation (102) as a function of Kelly fraction, instantaneous Sharpe ratio and time by substituting equation (69) and equation (74). Notice that only $R(t)^2$ is a function of time:

$$\tilde{g}^e_\infty(cf^*) = R(t)^2 c\left(1-\frac{c}{2}\right)SR^2_\infty = \frac{c(1-c/2)SR^2_\infty}{1+1/SR_{G_\infty}(t)^2}$$
$$= \frac{c(1-c/2)SR^2_\infty}{1+1/\left[(1-c/2)^2 t SR^2_\infty\right]}, \tag{103}$$

which by substituting equation (90) becomes a function of Kelly fraction and $SNR_0^{\breve{m}_e}$ and time:

$$\tilde{g}^e_\infty(cf^*) = \frac{c(1-c/2)SNR_0^{\widetilde{m}_e}}{1+1/\left[(1-c/2)^2 tSNR_0^{\widetilde{m}_e}\right]}. \tag{104}$$

Alternatively, we can write the realizable expected instantaneous excess growth rate as a function of investment fraction, standard deviation, instantaneous Sharpe ratio and time by substituting equation (72) into equation (103):

$$\tilde{g}^e_\infty(f) = \frac{[1-fs_e/(2SR_\infty)]fs_eSR_\infty}{1+1/\left([1-fs_e/(2SR_\infty)]^2 tSR_\infty^2\right)}, \tag{105}$$

which at full Kelly, when $f = f^*$, by substituting equation (71) simplifies to:

$$\tilde{g}^e_\infty(f^*) = \frac{SR_\infty^2}{2\left[1+4/\left(tSR_\infty^2\right)\right]} = \frac{SNR_0^{\widetilde{m}_e}}{2\left[1+4/\left(tSNR_0^{\widetilde{m}_e}\right)\right]}. \tag{106}$$

Figure 7 gives an example how the realizable expected instantaneous excess growth rate, as given by equation (103) or (104), is maximized at Kelly fraction ranging from half to one depending on investment time horizon t. Realizable expected instantaneous excess growth rate in the figure is normalized by its maximum value to emphasize how the value in relation to maximum value evolves as a function of investment horizon. We can see from the red line that, at very short investment horizon, one month, maximum is achieved at Kelly fraction very close, just slightly greater, to value 0.5. It can be shown that expected instantaneous excess growth rate scaled (multiplied) by SNR is maximized exactly when Kelly fraction $c = 0.5$. On the other hand, we can see from the blue line that, at very long investment horizon, a thousand years, maximum is achieved at Kelly fraction very close to value one (approximately 0.98). It is obvious that as time approaches infinity, realizable reward approaches expected instantaneous excess growth rate which is maximized exactly at Kelly fraction one.

Interestingly, we always find a maximum, the absolute best risk adjusted stock allocation expressed as Kelly fraction c, ranging as $0.5 < c < 1$ as investment horizon

t range as $0 < t < \infty$ respectively. Note, however, that this assumes infinite SNR for the parameters in equation (72) used in calculating Kelly fraction c. In reality, as stressed by Thorp (2006), those parameters are noisy estimates which calls for more conservative approach for selecting the risk level c. We consider the stock allocation range $0.5 < c < 1$ as being highly risky proposition for a real-world risk averse investor. Nevertheless, this approach provides us with a method for determining a theoretically best (providing the greatest realizable expected instantaneous excess growth rate) stock allocation level (Kelly fraction) for a given portfolio (instantaneous Sharpe ratio) as a function time. The exact maximum for intermediate (between infinitely short and infinitely long) investment horizons can be found by differentiating equation (103) with respect to c.

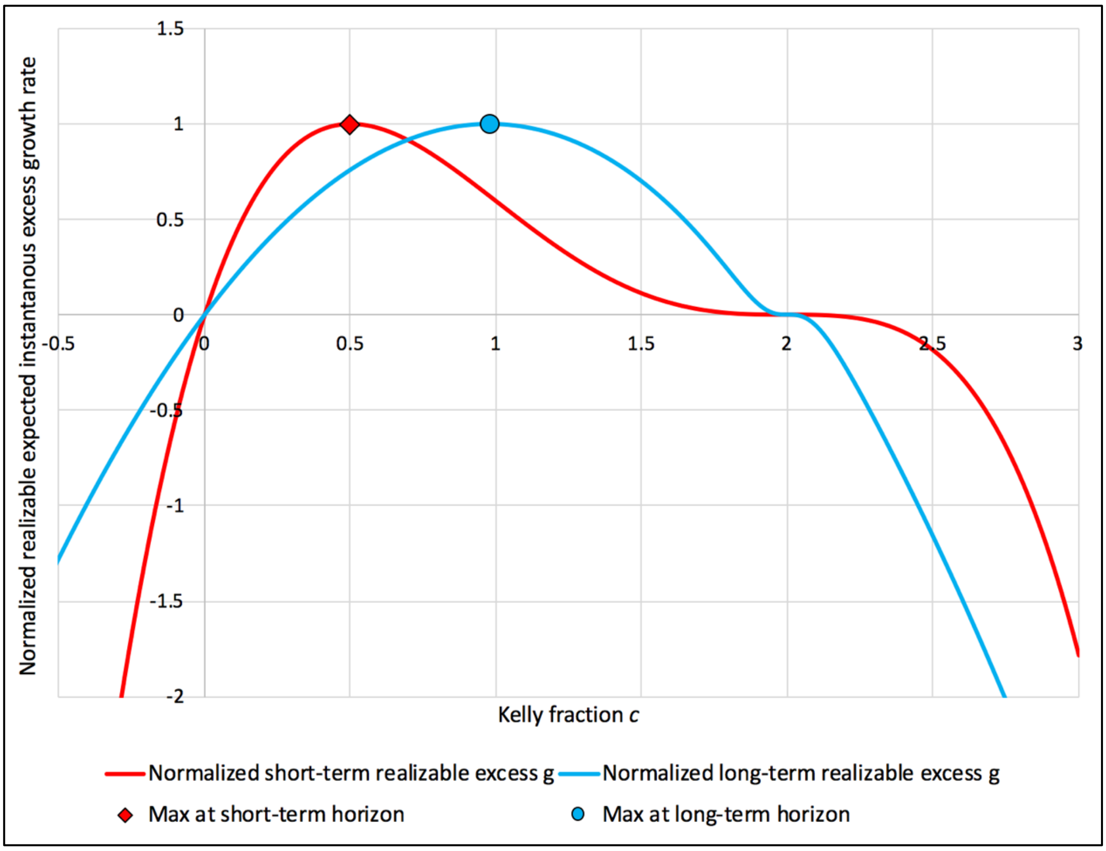

Figure 7. Normalized realizable expected instantaneous excess growth rate at very short (one month) and very long (1000 years) investment time horizons.

There are compelling arguments against the idea that stocks become less risky as the time horizon increases (see e.g. Kritzman (1994) or Kritzman and Rich (1998)). We do not state that increased SNR due to increased time horizon necessarily implies lower risk. However, we do argue that as SNR increases as a function of time we should increase the weight we give for the information expressed as the expected excess growth rate as investment time horizon increases. The picture and conclusions may, however, be different when judged based on accumulated excess wealth instead of the rate of excess growth. This is because even though the rate of excess growth converges towards its expectation as time passes, the compounding excess wealth experiences exponential growth and the distribution tends to diverge, not converge, over time. Compounding effect can be dominant for the investor experience in the long-term. If the accumulated excess wealth describes the long-term investment result after the investment period, the average and variability of the excess growth rate describe the journey to that result. A journey, and the long-term as a concept, constitutes of a series of consecutive individual short-term periods. Therefore, in the diversification context, realizable risk premium is descriptive as a measure for short-term diversification effect.

We often hear stock returns described as being noisy and, as an implication, that investor willing to invest to stocks should have a sufficiently long investment horizon. But how to quantify noisy and sufficiently long investment horizon? We have provided a method, equation (93), for assessing the noisiness in terms of SNR, the average signal power to average noise power per unit of compounding frequency ratio for realized instantaneous geometric excess growth rate, $SNR_0^{\breve{g}^e_\infty}$, to be exact. To assess the concept of sufficiently long investment horizon, we solve investment horizon t from equation (101), substitute equation (93) and (74) to obtain investment horizon as a function of R-squared, instantaneous Sharpe ratio and Kelly fraction:

$$t = \frac{1}{(1/R^2-1)(1-c/2)^2 SR_\infty^2}. \tag{107}$$

Or alternatively, substituting equation (75) instead of (74), we get investment horizon as a function of R-squared, instantaneous Sharpe ratio, investment fraction and standard deviation:

$$t = \frac{1}{(1/R^2-1)[SR_\infty - f(s_e/2)]^2}. \quad (108)$$

The sufficiently long investment horizon can be defined as time horizon which ensures at least half of the realized instantaneous excess growth rate being determined by the expected instantaneous excess growth rate of the portfolio/strategy. This corresponds to time required for realizable risk premium to equal or exceed half of the (expected) risk premium. This implies R-squared, as determined in equation (101), being at least one half, which corresponds to $SNR_0^{\breve{g}_\infty^e}$ being at least one (zero decibels). We therefore can determine the sufficiently long investment horizon as the required minimum length for the investment horizon t_{min} in years for a portfolio by substituting $R^2 = 0.5$ into equation (108):

$$t_{min} = t_{R^2=0.5} = t_{(\widetilde{RP}_{EW,G\infty}/RP_{EW,G\infty})=0.5} = t_{SNR_0^{\breve{g}_\infty^e}=1=0dB}$$
$$= \frac{1}{[SR_\infty - f(s_e/2)]^2} = \frac{1}{SNR_0^{\breve{g}_\infty^e}}. \quad (109)$$

As instantaneous Sharpe ratio SR_∞ is an increasing and standard deviation s_e a decreasing function of diversification, it is clear that the required minimum length for the investment horizon t_{min} is minimized when diversification is maximized. Similarly, it is clear that lower levels of diversification, as well as higher investment fraction f, call for longer investment horizon.

We can derive a normalized investment horizon for a given Kelly fraction c by normalizing (dividing) by the investment horizon of a full Kelly allocation. This is a general equation for normalized investment horizon and applies therefore for normalized minimum investment horizon as well. We can obtain this by utilizing equation (107) with fractional Kelly in the numerator divided by full Kelly in the denominator. The resulting equation applies generally as an investment horizon relative as both R-squared and instantaneous Sharpe ratio are reduced from the equation making it a function of Kelly fraction c and nothing else:

$$\frac{t(cf^*)}{t(f^*)} = \frac{1}{4(1-c/2)^2}. \tag{110}$$

As a sanity check, we note that utilizing the investment horizon $t(k, cf^*)$, as given by Thorp (2006, p. 408), instead of equation (107), will yield exactly the equation (110). Thorp approached the investment horizon from the angle of confidence intervals (k corresponding to a Z-score of a standard normal distribution) while our approach is based on the coefficient of determination, R-squared. Both approaches, however, result the same relationship $t(cf^*)/t(f^*)$ for the investment horizon between a fractional Kelly allocation and full Kelly allocation.

3.3.8 Decomposing risk premium

We will next decompose the instantaneous geometric risk premium of a n_P–stock portfolio to crystallize the role of diversification. The decomposition has two parts: the risk premium of an average single stock portfolio and the diversification premium from additional stocks added into the portfolio. We find that diversification, especially at the low end of its spectrum, is a dominant factor determining the size of the instantaneous geometric risk premium.

We have derived instantaneous geometric risk premium of an equally weighted portfolio RP_{EW,G_∞} in equation (8) and diversification premium DP^P for a n_P–stock portfolio in equation (34) and for a benchmark portfolio DP^{BM}in equation (38). We denote instantaneous geometric risk premium for equally weighted n_P–stock portfolio as RP^P_{EW,G_∞} and for equally weighted benchmark portfolio as RP^{BM}_{EW,G_∞}. A special case of RP^P_{EW,G_∞} is an instantaneous geometric risk premium of a single–stock portfolio which we denote as $RP^{n=1}_{EW,G_\infty}$. With these ingredients, we can decompose the instantaneous geometric risk premium of an equally weighted benchmark portfolio.

First, we write the risk premium of a n_P–stock portfolio as the risk premium of a single-stock portfolio plus the diversification premium of a n_P–stock portfolio (where $n_P > 1$) as given by equation (34). By substituting equation (38) we have:

$$RP^{P,n>1}_{EW,G_\infty} = RP^{n=1}_{EW,G_\infty} + DP^{P}_{n>1}$$
$$= RP^{n=1}_{EW,G_\infty} + \left(1 - \frac{1}{n_P} + \frac{1}{n_{BM}}\right) DP^{BM}, \tag{111}$$

where we can see that as the number of stocks in the portfolio approaches the number of stocks in the benchmark, the risk premium of the n_P–stock portfolio approaches the risk premium of the benchmark portfolio. For single stock portfolio ($n_P = 1$) the diversification premium term is zero. When the number of stocks in the portfolio is equal to number of stocks in the benchmark, we have the risk premium of the benchmark portfolio which equals the risk premium of a single-stock portfolio plus the diversification premium of a benchmark portfolio:

$$RP^{BM}_{EW,G_\infty} = RP^{n=1}_{EW,G_\infty} + DP^{BM}. \tag{112}$$

By substituting equation (113) into (111) and further substituting equation (40) we can alternatively write the risk premium of a n_P–stock portfolio as the risk premium of the benchmark portfolio plus portfolio's diversification premium difference to benchmark:

$$RP^{P,n>1}_{EW,G_\infty} = RP^{BM}_{EW,G_\infty} + \Delta DP^{BM}_{n>1}. \tag{113}$$

In the limit, when $n_{BM} \to \infty$, risk premium of a n_P–stock portfolio, as given by equation (111), simplifies to:

$$RP^{P,n_{BM}\to\infty}_{EW,G_\infty} = RP^{n=1}_{EW,G_\infty} + \left(1 - \frac{1}{n_P}\right) DP^{BM} = RP^{BM}_{EW,G_\infty} - \frac{DP^{BM}}{n_P}. \tag{114}$$

When n_{BM} is large compared to n_P, equation (114) can be used to approximate the risk premium of a n_P–stock portfolio.

An important intuition arises from equations (111) and (114). As benchmarks are typically broad, implying large n_{BM}, and their diversification premiums DP^{BM} are large, number of stocks in the portfolio n_P attain dominant role in these equations. Number of stocks in the portfolio is a measure of level of diversification. Diversification therefore, especially at the low end of its spectrum, is a dominant factor determining the size of the geometric risk premium. That being the case, all else equal, any friction limiting the level of diversification in a portfolio implies lower expected excess growth rate (lower geometric risk premium).

Diversification premium difference to benchmark is a piece in the equity premium puzzle. It is easy to imagine greater frictions for diversification in the past when today's cost-efficient diversification technology (index funds, ETFs) were not available. We can think the historical high transaction costs and large spreads to have increased the required risk premium in their own right, but they simultaneously have increased diversification costs such that investors have been forced to diversify less than perfectly. Theoretical equity premium measured for the fully diversified market portfolio therefore has been higher than practical equity premium required by (and realized for) less than perfectly diversified investors by the amount expressed by the diversification premium difference to benchmark. Today the cost for diversification is practically brought to zero implying the practical equity premium required by (and realized for fully diversified) investors is equal to the theoretical benchmark risk premium.

Assuming that the majority of investors who set the level for required risk premium diversify as broadly as is cost-efficient, we should expect the future fully diversified benchmark risk premium to roughly correspond to risk premium required by investors. All other factors except cost-efficiency of diversification being equal, we should expect fully diversified investor today to earn a risk premium similar to risk premium realized to average investor in the past with less than perfect diversification. The level of that risk premium is the historical (benchmark) risk premium minus the magnitude of diversification premium difference to benchmark in average historical investor's portfolio. We therefore should expect the future benchmark risk premium to be lower than historical benchmark risk premium. Less than perfectly diversified investors today are expected to lag the forward-looking risk premium of a fully diversified investor by the amount equal to their diversification premium difference to benchmark. In the history, full diversification was not cost-efficient and therefore not feasible, implying benchmark level risk premiums were theoretical concepts. Today, practically perfect diversification is possible and cost-efficient, implying benchmark level risk premiums are useful to describe investor's experience of the risk premium. This is one additional reason why one may not be well advised to extrapolate historical benchmark level risk premiums into the future[2]. Yet another implication from increased diversification leading to lower forward-looking market level risk premium is its impact to market valuations. All else equal, today's cost-efficient diversification is expected to increase market valuations.

3.3.9 Forward-looking geometric premiums in the presence of uncertainty about risks

There is a difference between utilizing in-sample and out-of-sample data. So far, we have built the diversification effect analysis based on in-sample data, meaning that the parameters used in the equations are measured from exhaustive historical data. These parameters represent the exact population parameter values as they occurred in the data and therefore don't exhibit uncertainty. What remains uncertain, however, is what the parameters, and consequently the diversification effect, will be in the future, in out-of-sample data.

Our focus is in in-sample data as we want to explore what explains the diversification effect. This is best achieved by using data with known population parameters. However, we eventually wish to use and project the learnings about the diversification effect into the future. This requires us to predict, or to estimate, the future parameters

[2] The first time we read and start to think about the idea of diversification impacting the level of risk premium was in one of the blog posts by a blogger/financial theorist/commentator "Jesse Livermore". Interpretation of our decomposition of the risk premium supports that idea.

somehow. This necessarily means we need to deal with the uncertainty related to future parameter estimates.

We will show that, when all other factors are equal, historical geometric risk premium can be considered the upper bound and historical diversification premium the lower bound for corresponding forward-looking premiums. This is because the forward-looking geometric metrics are impacted by both the expected risk and the uncertainty about the risk. The latter is zero in case of historical data is used, but the future is certain to entail uncertainty. We will derive forward-looking formulas for the geometric risk premium, Kelly criterion, diversification premium and the diversification premium difference to benchmark in the presence of uncertainty about the risks. The relevant risks, once again, are the systematic risk for the fully diversified portfolio and the idiosyncratic risk which determines the diversification effect. Importantly, the uncertainty about the idiosyncratic risk can be diversified away. Lastly, we will demonstrate how the uncertainty about the future fat-tailed systematic risk may be used to explain aggressive volatility spikes.

Thorp (2006, pp. 411–412) discuss the uncertainty related to parameter estimates from a Kelly investor point of view. Thorp focus on the uncertainty related to the estimating future expected returns. As the growth rate, and the geometric risk premium, is determined as a function of both the expected return and predicted variance, we will now focus on the uncertainty related to the latter, the uncertainty about the risk.

Uncertainty about the risk is particularly significant in the context of geometric metrics as any uncertainty, be it the predicted risk (predicted standard deviation) or uncertainty about the predicted risk (the standard deviation of the standard deviation estimator) decreases the expected growth rate or the expected risk premium. In case of arithmetic metrics, which are not a function of variance, there is no such effect and the uncertainty about the risk will not affect the expected return while it does make an investor less certain about the spread of the distribution of the future returns he faces.

Jacquier, Kane and Marcus (2003) study the effect of uncertainty on the future parameter estimates related to arithmetic and geometric mean returns. They find that in the very long-term both arithmetic and geometric mean rate of return estimates predicted based on historical values are upward biased forecasts of the typical future portfolio values. They also find that estimates of the arithmetic mean rate of return are more upward biased compared to their geometric counterparts. In our view the fact that historical mean rate of arithmetic returns are upward biased estimates of the typical realizable future portfolio returns is an expected result as it is the growth rate (the geometric return), not the arithmetic return that is governed by the law of large numbers (Kelly, 1956). The Jacquier et al. finding that even geometric mean rate of returns measured from historical data are upward biased estimates of the future geometric mean rate of returns is consistent with our view which we now elaborate analytically.

We can think of the future parameter estimates as randomly drawn from a distribution for associated parameter estimators which are random variables. We denote the parameter estimators with hat accent: $\widehat{m}_e$ and $\hat{s}_e$ are the estimators of m_e and s_e respectively. We can think of the average error of a parameter estimate to be the standard deviation of the parameter estimator. We assume the parameter estimators to be unbiased implying expectation of the estimation error is zero. Following the methodology that Thorp (2006, p. 406) used in his growth rate derivation, we can express the instantaneous excess growth rate incorporating the average parameter errors as weighted average over all the combinations of the average parameter estimates. We have two parameters to be estimated so the number of possible average outcomes is four. The weight of each possible average outcome is its probability 0.25. We then can write:

$$g_\infty^e = RP_{EW,G_\infty} = \\ 0.25 \begin{Bmatrix} \left(f\left[\bar{\widehat{m}}_e + Sdev(\widehat{m}_e)\right] - \frac{f^2}{2}\left[\bar{\hat{s}}_e + Sdev(\hat{s}_e)\right]^2 \right) + \\ \left(f\left[\bar{\widehat{m}}_e + Sdev(\widehat{m}_e)\right] - \frac{f^2}{2}\left[\bar{\hat{s}}_e - Sdev(\hat{s}_e)\right]^2 \right) + \\ \left(f\left[\bar{\widehat{m}}_e - Sdev(\widehat{m}_e)\right] - \frac{f^2}{2}\left[\bar{\hat{s}}_e + Sdev(\hat{s}_e)\right]^2 \right) + \\ \left(f\left[\bar{\widehat{m}}_e - Sdev(\widehat{m}_e)\right] - \frac{f^2}{2}\left[\bar{\hat{s}}_e - Sdev(\hat{s}_e)\right]^2 \right) \end{Bmatrix}. \tag{115}$$

Given the unbiased parameter estimators, we have the expected (bar accent denoting "mean") parameter estimates equal to the true parameters: $\bar{\hat{m}}_e = m_e$ and $\bar{\hat{s}}_e = s_e$. By simplifying equation (115), the instantaneous expected excess growth rate, which is equal to the instantaneous geometric risk premium, of a portfolio in the presence of uncertainty about the risk can be written as:

$$\begin{aligned} g_{\infty}^{e} = RP_{EW,G_{\infty}} &= fm_e - \frac{f^2}{2}[s_e^2 + Sdev(\hat{s}_e)^2] \\ &= fm_e - \frac{f^2}{2}[s_e^2 + Var(\hat{s}_e)] \\ &= fm_e - \frac{f^2}{2}\left[Var(G_{\infty}^{e}) + Var\left(\widehat{Sdev}[G_{\infty}^{e}]\right)\right], \end{aligned} \tag{116}$$

where we can see that the uncertainty about the excess return averages out but the uncertainty about the risk does not. The variance of the uncertainty about the risk, the variance of the estimated standard deviation of the excess growth rate, is added to the true variance of the excess growth rate. If there is no uncertainty about the risk, which is the case in an in-sample test utilizing the whole population of historical data, then the second variance term vanishes and equation (116) is equal to equation (8). We can see that $g_{\infty}^{e}[Sdev(\hat{s}_e) > 0] < g_{\infty}^{e}[Sdev(\hat{s}_e) = 0]$ meaning that any uncertainty about the risk decreases the expected instantaneous excess growth rate which is the instantaneous geometric risk premium. It appears, perhaps characteristic to geometric measures in general, that any uncertainty related to growth, be it the uncertainty about the growth itself or the uncertainty about the uncertainty about the growth, will be detrimental to expected growth rate.

By differentiating equation (116) with respect to f, we can derive a Kelly criterion in the presence of uncertainty about the risk:

$$f^* = m_e/[s_e^2 + Sdev(\hat{s}_e)^2]$$
$$= m_e/[s_e^2 + Var(\hat{s}_e)]$$
$$= m_e/\left[Var(G_\infty^e) + Var\left(\widehat{Sdev}[G_\infty^e]\right)\right], \quad (117)$$

which gives the full Kelly fraction f^* in the presence of uncertainty about the risk. If there is no uncertainty about the risk, then the second variance term vanishes and equation (117) is equal to equation (62). We can see that $f^*[Sdev(\hat{s}_e) > 0] < f^*[Sdev(\hat{s}_e) = 0]$ meaning that any uncertainty about the risk decreases the full Kelly fraction.

Diversification premium for a portfolio, as derived in equations from (31) to (35), is based on geometric metrics and therefore is subject to be affected by the uncertainty about the risk. Specifically, the uncertainty about the idiosyncratic risk. We use the result from equation (116) and adapt the derivation from equation (34) to derive the diversification premium for a portfolio of n_P stocks in the presence of uncertainty about the idiosyncratic risk (full derivation is given in Appendix 2). We denote $\widehat{Isdev}_{n=f=1}$ as the parameter estimator for the idiosyncratic standard deviation of instantaneous excess growth rate of an average single stock portfolio with 100% allocation to stocks.

$$DP_{n>1}^P = \frac{f^2}{2}\left\{\left(1 - \frac{1}{n_P} + \frac{1}{n_{BM}}\right) Ivar_{n=f=1} + \left[1 - \frac{1}{n_P} - \frac{1}{n_{BM}} + \frac{2}{\sqrt{n_P n_{BM}}}\right] Var\left(\widehat{Isdev}_{n=f=1}\right)\right\}. \quad (118)$$

By substituting $n_P = n_{BM}$ into equation (118) we find the diversification premium of a benchmark portfolio in the presence of uncertainty about the idiosyncratic risk:

$$DP^{BM} = \frac{Ivar_{n=f=1} + Var\left(\widehat{Isdev}_{n=f=1}\right)}{2} f^2, \tag{119}$$

where we can see that the diversification premium of a benchmark is an increasing function of both the variance of the excess growth rate and the variance of the average error of the estimated idiosyncratic standard deviation of the excess growth rate. Both of these underlying measures are for an average single stock portfolio with 100% stock allocation denoted with subscript $n = f = 1$. In case of no uncertainty about the risk the equation will be equal to equation (38).

Equation (118), when $n_{BM} \to \infty$, substituting equation (119) simplifies to:

$$DP^P_{n_{BM}\to\infty} = \left(1 - \frac{1}{n_P}\right) \frac{Ivar_{n=f=1} + Var\left(\widehat{Isdev}_{n=f=1}\right)}{2} f^2$$
$$= \left(1 - \frac{1}{n_P}\right) DP^{BM}. \tag{120}$$

Equation (120) approximates the diversification premium for a portfolio of n_P stocks in the presence of uncertainty about the idiosyncratic risk when n_{BM} is large, or preferably very large as there is a term $2/\sqrt{n_P n_{BM}}$ in equation (118) which descent towards zero, when $n_{BM} \to \infty$, is decelerated by the square root. We can see that in case of no uncertainty about the idiosyncratic risk the second variance term vanishes and equation (120) is equal to equation (35) as we expect. Similarly, in case of no uncertainty about the idiosyncratic risk, equation (118) is equal to equation (34).

By substituting equations (118) and (119) into equation (40) we have the diversification premium difference to benchmark in the presence of uncertainty about the idiosyncratic risk:

$$\Delta DP_{n>1}^{BM} = \frac{f^2}{2}\left\{\left(\frac{1}{n_{BM}} - \frac{1}{n_P}\right) Ivar_{n=f=1} + \left[\frac{1}{n_{BM}} - \frac{1}{n_P} + \frac{2}{\sqrt{n_P n_{BM}}}\right] Var\left(\widehat{Isdev}_{n=f=1}\right)\right\}. \tag{121}$$

When $n_{BM} \to \infty$, equation (121) simplifies to:

$$\Delta DP_{n_{BM}\to\infty}^{BM} = -\frac{Ivar_{n=f=1} + Var\left(\widehat{Isdev}_{n=f=1}\right)}{2n_P} f^2, \tag{122}$$

which approximates the diversification premium difference to benchmark in the presence of uncertainty about the idiosyncratic risk when n_{BM} is large, or preferably very large. We can see that the more uncertainty there is about the idiosyncratic risk the greater the opportunity cost of foregone diversification. Now the opportunity cost of foregone diversification is approximated as one half of the portfolio's sum of idiosyncratic variance and the variance of the idiosyncratic standard deviation estimator scaled by the squared investment fraction. This means, all else equal, that the historical opportunity cost can be considered as a lower bound for the forward-looking opportunity cost of foregone diversification.

We note that, in the spirit of information theory, we can alternatively interpret the opportunity cost of foregone diversification in terms of noise power as one half of the portfolio's sum of the noise power of the average idiosyncratic risk and the noise power of the idiosyncratic risk estimator scaled by the squared investment fraction.

Basically, any of the formulas that we have introduced that include the effect from standard deviation s_e can be transformed into a forward-looking version incorporating the uncertainty about the risk.

Contrary to geometric risk premium, which is a decreasing function of the uncertainty about the risk (risk here consists of both systematic and idiosyncratic risk), diversification premium increases as the uncertainty about the idiosyncratic risk increases. All else equal and assuming the future entails uncertainty, we can consider the historical portfolio mean growth rate, such as the historical geometric risk premium, as the upper bound for forward-looking growth rate estimate, such as forward-looking geometric risk premium. In case of diversification premium, however, we can consider the historical diversification premium as the lower bound for forward-looking diversification premium.

Finally, based on equation (112), we can write the forward-looking instantaneous geometric risk premium of the benchmark portfolio in the presence of uncertainty about the systematic risk $Var(\widehat{Sdev}_{BM})$ and uncertainty about the idiosyncratic risk $Var(\widehat{Isdev}_{n=f=1})$ as:

$$\begin{aligned} RP_{EW,G\infty}^{BM} = RP_{EW,G\infty}^{n=1} & \begin{bmatrix} Var_{BM} + Var(\widehat{Sdev}_{BM}), \\ Ivar_{n=f=1} + Var(\widehat{Isdev}_{n=f=1}) \end{bmatrix} \\ & + DP^{BM}\left[Ivar_{n=f=1} + Var(\widehat{Isdev}_{n=f=1})\right], \end{aligned} \tag{123}$$

where the terms are shown as a function of the risks and uncertainties about the risks. Note that the square brackets now indicate “function of”. $\widehat{Sdev}_{BM}$ is the parameter estimator for the standard deviation of the instantaneous excess growth rate of a benchmark. We can see that first term $RP_{EW,G\infty}^{n=1}$ is a function of systematic risk Var_{BM} plus the uncertainty about the systematic risk $Var(\widehat{Sdev}_{BM})$ as well as idiosyncratic risk $Ivar_{n=f=1}$ plus the uncertainty about the idiosyncratic risk $Var(\widehat{Isdev}_{n=f=1})$. The second term DP^{BM} is a function of the idiosyncratic risk plus its uncertainty only. Furthermore, the positive DP^{BM} perfectly offsets the negative impact arising from the idiosyncratic components in the $RP_{EW,G\infty}^{n=1}$. This means that the risk premium of a fully diversified benchmark portfolio is a function of the systematic risk Var_{BM} and its uncertainty $Var(\widehat{Sdev}_{BM})$ only.

We can think of the expected systematic risk Var_{BM} as a base rate of a forward-looking risk premium which arises from the historical realized risk premium while the uncertainty about the base rate, $Var(\widehat{Sdev}_{BM})$, is subject to any new information specific to the future systematic risk expectations. Given the fat tails of the empirical systematic risk distributions as described by Gabaix et al. (2003), we can think the term $Var(\widehat{Sdev}_{BM})$ as incorporating also the fat-tailed nature of the perceived future systematic risks which in part may help explain the sudden aggressive moves in stock market volatility as new risks with uncertain consequences suddenly arise. An important take away from equation (123) is that the uncertainty about the systematic risk is always a concern but the uncertainty about the idiosyncratic risk is completely avoidable by full diversification.

3.4 Determination of diversification metrics in a continuous-time world

We will next determine the diversification metrics which capture the effect of diversification in a continuous-time world. Diversification effect is not easily captured by one metric. There are many aspects like expected wealth after the target investment horizon, expected growth rate leading to expected wealth and the risk, the uncertainty, associated to both of these. We therefore will not try to capture the diversification effect with just one metric, but instead will have three metrics of which the two latter are determined separately for risk neutral and risk averse investors.

We consider the diversification effects in the context of in-sample data with known population parameters meaning we utilize historical parameter values which entail no uncertainty. As discussed earlier, diversification effects determined using historical parameters can be considered as lower bounds for forward-looking diversification effects which account for the uncertainties about the future risks. Therefore, when all other factors are equal, the number of stocks required for a desired level of diversification benefit in the future is expected to be at least the amount required based on our metrics which are based on historical data.

The first metric considers diversification effect around the absolute minimum acceptable investment outcome, the level of diversification ensuring positive geometric risk premium. The second, short-term, metric is based on comparing portfolio's (realizable) risk premium to that of the benchmark. The third, long-term, metric is based on comparing portfolio's (realizable) gross compound excess wealth over time to that of the benchmark.

The first metric focusses on the minimum level of diversification by comparing the expected reward from risk bearing to the expected reward from riskless investment. The reference point is the riskless investment delivering the riskless rate achieved with zero exposure to equity risk. The second and third metrics focus on the opportunity cost and risk arising from deviating from the benchmark portfolio. The reference point is the fully diversified benchmark portfolio.

Second and third metrics are further divided for two investor types: risk neutral and risk averse investor. Risk neutral investor utilize risk premium and risk averse investor realizable risk premium in the equations determining diversification metrics. The fact that realizable risk premium entails the effect from two dimensions (reward and risk) and expresses its value in one dimension (reward) enables using realizable risk premium interchangeably with risk premium in the equations.

Second and third diversification metrics are both functions of time but the effect of time is opposite. The second metric is based on ratios of (realizable) growth rates while the third metric is based on ratios of gross excess wealth compound at (realizable) expected growth rates. As time horizon lengthens, the distribution of (realizable) growth rates converges as the uncertainty around growth rates converges while the distribution of compound wealth diverges as the differences in (realizable) expected growth rates compound and expand the wealth differences exponentially. This is a common source of controversy and we attempt not to take sides but to address both aspects of the time effect.

3.4.1 Number of stocks required for positive risk premium

The first metric on the diversification effect is that an investor shall always expect a reward from bearing a risk. If riskless alternative provides a higher expected reward with no risk, there is no point investing in risky assets regardless whether the investor is risk neutral or risk averse. We therefore consider positive instantaneous geometric risk premium, meaning positive expected instantaneous excess growth rate, as our first diversification effect and determine the average minimum number of stocks in a portfolio yielding a positive geometric risk premium as a measure for the absolute minimum level of diversification.

We will show that the average minimum number of stocks in a portfolio yielding a positive geometric risk premium is simply and intuitively approximated by dividing the benchmark diversification premium by the benchmark's geometric risk premium. This provides us a tool to assess the absolute minimum level of required diversification for any strategy, a sort of level of riskiness of the strategy. Strategies, meaning the benchmark stock populations, can be, e.g., plain vanilla meaning no risk factors except equally weighted market risk or different investment styles like small cap, value etc.

From the decomposition of the risk premium of a n_P–stock portfolio in equation (111) we can solve the n_P for the case that $RP_{EW,G\infty}^{P,n>1} \geq 0$ which means that we are solving the minimum number of stocks in the portfolio that guarantees a positive risk premium. As both risk premium and diversification premium are functions of investment fraction f, we are solving $n_P(f)$:

$$n_P(f) \geq \frac{DP^{BM}(f)}{RP_{EW,G\infty}^{BM}(f) + DP^{BM}(f)/n_{BM}}, \tag{124}$$

which, by substituting equations (8), (34) and (38) and after some simplification, can alternatively be written as:

$$\begin{aligned} n_P(f) &\geq \frac{Ivar_{n=f=1}}{2\,m_e/f + Ivar_{n=f=1}/n_{BM} - Var_{BM}} \\ &= \frac{Ivar_{n=f=1}}{2\,RP_{EW,A}^{BM}/f + Ivar_{n=f=1}/n_{BM} - Var_{BM}} \\ &= \frac{Ivar_{n=f=1}}{2\sqrt{Var_{BM}}SR_{\infty}^{BM}/f + Ivar_{n=f=1}/n_{BM} - Var_{BM}}, \end{aligned} \tag{125}$$

where $RP_{EW,A}^{BM}$ is the arithmetic risk premium of an equally weighted benchmark portfolio. Equation (34) requires $n_P > 1$ meaning that equation (125) is accurate when $n_P > 1$. Approximate equations (126) and (127) do not have this limitation.

In the limit, when $n_{BM} \to \infty$, minimum average number of stocks required to achieve a positive risk premium, as given in equation (124), simplifies to:

$$n_P^{n_{BM}\to\infty}(f) \geq \frac{DP^{BM}(f)}{RP_{EW,G\infty}^{BM}(f)}, \tag{126}$$

which, by substituting equations (8) and (38) and after some simplification, can alternatively be written as:

$$n_P^{n_{BM}\to\infty}(f) \geq \frac{Ivar_{n=f=1}}{2RP_{EW,A}^{BM}/f - Var_{BM}} = \frac{Ivar_{n=f=1}}{2\sqrt{Var_{BM}}SR_{\infty}^{BM}/f - Var_{BM}}. \tag{127}$$

When n_{BM} is large, equations (126) and (127) can be used to approximate the minimum number of stocks required with a given investment fraction, $n_P(f)$, to achieve a positive risk premium.

We find that the approximation of the average required minimum number of stocks in the portfolio to achieve a positive risk premium, i.e., a positive expected instantaneous excess growth rate, simplifies to remarkably simple and intuitive equation (126) which implies simply dividing the benchmark diversification premium by the benchmark risk premium.

As typical number of stocks required for positive instantaneous geometric risk premium is very low (often close to one) and our empirical benchmark portfolios are large, we choose to use approximate equations (126) and (127) over the exact equations.

3.4.2 Number of stocks required for a proportion of benchmark risk premium

The second diversification metric, risk premium ratio, is intended to serve as a short-term diversification metric. In case of risk neutral investor, risk premium ratio is used while for risk averse investor we use realizable risk premium ratio. Risk averse investor cares about two aspects simultaneously: the expected excess growth rate, i.e., the geometric risk premium, and the associated risk. Both of these aspects are captured by one metric, the realizable risk premium. The diversification metric is the ratio between portfolio's (realizable) risk premium and that of the benchmark. This ratio tells us how much of the maximum diversification benefit on average is realizable by the portfolio. Note that, in case the portfolio's risk premium is negative the risk premium ratio (assuming benchmark risk premium is positive) is also negative.

We will show that in the very long-term the realizable risk premium ratio approaches the ratio between portfolio's (expected) risk premium and that of the benchmark, and in the short-term the ratio between portfolio's SNR weighted risk premium and that of the benchmark. This means that the risk is important in the short-term while the expected reward dominates in the long-term.

The realizable risk premium is a function of time as it converges towards the (expected) risk premium as time horizon increases. It does not, however, sufficiently reflect the effect of time on investor's portfolio as is does not experience the exponential growth like wealth does. The realizable risk premium can be considered as a proxy for the investor's experience of the journey towards the target wealth as it combines in one metric both the average direction (expected growth rate) and the risk (the volatility) along the way.

By choosing not to diversify maximally, an investor accepts both lower expected growth rate and higher risk both of which are reflected in the realizable risk premium. By determining the realizable risk premium of a n_P–stock portfolio divided by the realizable risk premium of a fully diversified benchmark as a short-term diversification effect, we capture the opportunity cost of foregone diversification measured in the units of risk premium. Even a very long-term investor lives his life in short-term increments. Therefore, the short-term effects, including short-term diversification effect, are meaningful to any risk averse investor.

Based on equations (101) and (102), substituting (75) and (113), we can derive target realizable risk premium ratio for a portfolio with $n_P > 1$ (tilde accent denotes "realizable"):

$$\begin{aligned}
T\widetilde{RP}R_{n>1} &= \widetilde{RP}^{P}_{EW,G\infty} / \widetilde{RP}^{BM}_{EW,G\infty} = \frac{R(t)^2_P}{R(t)^2_{BM}} \frac{RP^{P}_{EW,G\infty}}{RP^{BM}_{EW,G\infty}} \\
&= \frac{1 + 1/\left[t\left(SR^{BM}_{G\infty}\right)^2\right]}{1 + 1/\left[t\left(SR^{P}_{G\infty}\right)^2\right]} \left(1 + \frac{\Delta DP^{BM}_{n>1}}{RP^{BM}_{EW,G\infty}}\right) = \\
&\frac{1 + 1/\left(t\left[(SR^{BM}_{\infty})^2 - f\sqrt{Var_{BM}}SR^{BM}_{\infty} + \frac{f^2}{4}Var_{BM}\right]\right)}{1 + 1/\left(t\left[\frac{Var_{BM}}{Var_P(n_P)}(SR^{BM}_{\infty})^2 - f\sqrt{Var_{BM}}SR^{BM}_{\infty} + \frac{f^2}{4}Var_P(n_P)\right]\right)} \\
&\times \left(1 + \frac{\Delta DP^{BM}_{n>1}(f, n_P)}{f\sqrt{Var_{BM}}SR^{BM}_{\infty} - \frac{f}{2}Var_{BM}}\right),
\end{aligned} \tag{128}$$

which contains two terms which are functions of n_P: portfolio variance $Var_P(n_P)$ and diversification premium difference to benchmark $\Delta DP^{BM}_{n>1}(f, n_P)$ determined by equations (33) and (40), respectively. The resulting equation is complicated and solving n_P will not lead to simple and intuitive result. We therefore use computer to solve n_P. Regardless the complexity of the resulting equation, it still shows that after plugging in the size of the benchmark n_{BM} and deciding the investment fraction and target realizable risk premium ratio, we only need three parameters to predict the required number of stocks to achieve the target realizable risk premium ratio. The three parameters are: instantaneous Sharpe ratio of a benchmark portfolio (geometric risk premium would be equally good), the variance of a benchmark portfolio and the average idiosyncratic variance of a single stock portfolio. The last parameter is acquired as an output from the regression described by equation (47) and the two former parameters are calculated from the time series data of the benchmark portfolio.

In the limit, when $t \to \infty$, equation (128) simplifies to target risk premium ratio, the short-term diversification effect metric for a risk neutral investor:

$$\underset{t\to\infty}{T\widetilde{RP}R_{n>1}} = TRPR_{n>1} = \frac{RP^{P}_{EW,G\infty}}{RP^{BM}_{EW,G\infty}} = 1 + \frac{\Delta DP^{BM}_{n>1}}{RP^{BM}_{EW,G\infty}}$$
$$= 1 + \frac{\Delta DP^{BM}_{n>1}(f, n_P)}{f\sqrt{Var_{BM}}SR^{BM}_{\infty} - \frac{f^2}{2}Var_{BM}}, \tag{129}$$

where n_P, substituting equations (40) and (65), has a simple solution and becomes:

$$n_P = 1 \Big/ \left[\frac{1}{n_{BM}} + \frac{\left(2f\sqrt{Var_{BM}}SR^{BM}_{\infty} - f^2 Var_{BM}\right)(1 - TRPR_{n>1})}{f^2 Ivar_{n=f=1}}\right]. \tag{130}$$

Equation (129) show how the realizable target risk premium ratio $\underset{t\to\infty}{T\widetilde{RP}R_{n>1}}$ approaches (expected) target risk premium ratio ($TRPR_{n>1}$) as investment time horizon t approaches infinity. Equation (130) shows the number of stocks required to achieve desired diversification benefit level implying there is a diversification benefit for a risk neutral continuous-time world investor. This is remarkable compared to one period world where arithmetic expected return is constant regardless the level of diversification, implying there is no such concept as diversification benefit for a risk neutral investor. This means that diversification is beneficial to continuous-time world investor regardless the level or the existence of risk averseness. In fact, in the absence of stock picking skill, rational risk neutral continuous-time world investor will always diversify maximally to maximize his expected portfolio growth for a given investment fraction f (which, in the absence of leverage constraints, will be Kelly allocation f^*).

At the other extreme, when $t \to 0$, substituting (75) and (113) we find that equation (128) becomes:

$$T\widetilde{RP}R_{\substack{n>1\\t\to 0}} = \frac{SNR_0^{\check{g}_\infty^e,P} RP_{EW,G_\infty}^P}{SNR_0^{\check{g}_\infty^e,BM} RP_{EW,G_\infty}^{BM}} = \frac{\left(SR_{G_\infty}^P\right)^2 RP_{EW,G_\infty}^P}{\left(SR_{G_\infty}^{BM}\right)^2 RP_{EW,G_\infty}^{BM}} = \frac{\left[\frac{Var_{BM}(SR_\infty^{BM})^2}{Var_P(n_P)} - f\sqrt{Var_{BM}}SR_\infty^{BM} + \frac{f^2}{4}Var_P(n_P)\right] \times \left[f\sqrt{Var_{BM}}\left(SR_\infty^{BM} - \frac{f}{2}\sqrt{Var_{BM}}\right) + \Delta DP_{n>1}^{BM}(f,n_P)\right]}{f\sqrt{Var_{BM}}\left(SR_\infty^{BM} - \frac{f}{2}\sqrt{Var_{BM}}\right)^3} \tag{131}$$

which shows that in the short-term, realizable risk premium ratio approaches the ratio of the SNR weighted risk premiums. SNR (the squared geometric Sharpe ratio) is the dominant term, meaning that risk dominates the metric in the short-term. In the long-term, as shown by equation (129), the metric is dominated by the risk premium ratio. Equation (131) contains two terms which are functions of n_P: portfolio variance $Var_P(n_P)$ and diversification premium difference to benchmark $\Delta DP_{n>1}^{BM}(f,n_P)$ determined by equations (33) and (40), respectively. Computer can be used to solve n_P from the equation.

3.4.3 Number of stocks required for a proportion of benchmark wealth over time

The third, long-term, diversification metric is based on the fact that at the end of the day, at the end of the investment horizon that is, an investor will not eat the growth rate of his portfolio but the accumulated wealth. Excess growth rate is the means while the accumulated excess wealth is the end. By defining a diversification effect based on gross compound excess wealth of a portfolio, the gross compound wealth in excess of what is expected from compounding riskless rate, instead of expected instantaneous excess growth rate, we avoid confusing the means to the end that really matters. By (realizable) gross compound excess wealth, we mean the gross wealth that is continuously compounded using (realizable) risk premium as growth rate. *Gross* means the initial investment is included in the final wealth. In case of long-term risk averse investor, realizable risk premium replaces risk premium (which is used in case of long-term risk neutral investor) in the equations. Our third diversification metric is best suitable for a long-term investor who seeks to benefit from the compounding effect on accumulated capital.

We will derive two versions of the gross excess wealth-based diversification effects: risk neutral and risk averse versions. Risk neutral version will not account for the risk and is based solely on gross excess wealth accumulated by compounding geometric risk premium. Geometric risk premium is the expected excess growth rate of the portfolio which can be considered to be realizable at infinitely long investment time horizon. However, realistic investment time horizons are nowhere close to infinity implying the risk, not only the expected excess growth rate, becomes significant. Risk averse version of the metric will account for the risk by scaling the risk premium based on how much of the realized risk premium at a given time horizon is explained by the (expected) risk premium.

We will see that in the continuous-time word, even a risk neutral investor benefits greatly from diversification. In fact, a risk neutral investor with no stock picking skill should diversify maximally to maximize the expected growth of his wealth. Risk neutral investor with stock picking skill should optimize his level of diversification to serve the competing goals: goal of increasing diversification to decrease the diversification premium difference to benchmark and goal of decreasing diversification to allow for better analysis per stock in the portfolio.

In some cases, an investor with very high tolerance for short-term volatility who can be considered as risk neutral in the short-term may be risk averse in the long-term. For example, a person saving for retirement can be risk neutral in a sense that he is psychologically immune to short-term stock market volatility. However, if the retirement saver defines his investing to be success in case he has accumulated a liability matching amount, then long-term risk matters regardless the risk neutrality in the short-term.

3.4.3.1 Risk neutral wealth ratio

We will first define a minimum target wealth ratio and then the number of stocks n_P required to achieve this ratio as our diversification metric. The minimum target wealth ratio is maintained continuously as a function of time and is defined as the portfolio's gross excess wealth compound at its instantaneous expected excess growth rate divided by the benchmark's gross excess wealth compound at its instantaneous expected excess growth rate. We will then derive formulas for the diversification effect, the required number of stocks n_P, based on three alternative regression outputs: the idiosyncratic variance $Ivar_{n=f=1}$, alpha $\alpha_{n=f=1}$ and finally R-squared $R^2_{n=f=1}$ together with benchmark variance Var_{BM}. The formulas decompose the diversification effect and allow us to acquire intuition on the aspects that constitute the diversification effect. We will show that the main aspect is the overall size of the exposure to the costly idiosyncratic risk. More detailed, we will show that the main drivers for a higher required level of diversification are low number and homogeneity of portfolio's risk (investment style) exposures, preference for high risk investment styles, long investment time horizon, and perhaps most notably, high investment fraction meaning large fraction of the capital allocated to stocks.

Importantly, growth rate is indifferent to time while accumulated wealth experiences the passage of time very vividly in the form of exponential growth, compounding. We aim at capturing the effect of time in our diversification effect metric. The effect of time becomes eminent as a tiny difference in excess growth rate compounds to increasingly meaningful difference in the accumulated gross excess wealth over time. We therefore determine the gross compound excess wealth at the end of the investment horizon as our third diversification effect and determine the gross compound excess wealth of a portfolio divided by the gross compound excess wealth of a fully diversified benchmark as a normalized diversification benefit, a measure of exhausted diversification benefit potential. This allows us to assign a minimum target level for the normalized diversification benefit at the end of the investment horizon, the minimum target for gross compound excess wealth ratio, and to calculate the average number of stocks required to achieve that target. When we say "at the end of the investment horizon" we mean at any given time in the future meaning that the minimum target for gross compound excess wealth ratio is maintained continuously.

We denote this minimum target for gross compound excess wealth ratio as TWR where T denote target, W wealth and R ratio. The highest possible value for TWR is one which implies full diversification meaning that investment portfolio is identical to benchmark portfolio. When a lower than full diversification is desired, then the expected excess growth rate, the geometric risk premium, for a randomly picked portfolio will necessarily be lower compared to benchmark and TWR need to be set accordingly to a value lower than one. For example, $TWR = 0.9$ would imply that investor is targeting in minimum 90% of the gross compound excess wealth of the benchmark portfolio at any given time in the future. We say in minimum 90% as 90% is the expected result when the investor chooses the stocks randomly. We can assume 90% of the gross compound excess wealth of the benchmark portfolio applies for an investor with no stock picking skill[3], which given the stock market efficiency makes a good base assumption. This implies that the investor needs a considerable amount of skill just to climb from the 90% level to the 100% level equaling the gross compound excess wealth of the fully diversified benchmark portfolio. This is the skill that needs to be consumed to compensate for the portfolio diversification premium difference to benchmark ΔDP^{BM}, which was derived in equation (40).

First, we derive the minimum target for gross compound excess wealth ratio TWR. The ratio is simply portfolio's gross compound excess wealth after investment time horizon t compound at the expected instantaneous excess growth rate divided by the corresponding measure for the fully diversified benchmark:

$$TWR = \frac{e^{tg_{\infty}^{e,P}}}{e^{tg_{\infty}^{e,BM}}} = e^{t\left(g_{\infty}^{e,P} - g_{\infty}^{e,BM}\right)} = e^{t\Delta DP^{BM}}. \tag{132}$$

Next, we take $\ln(\cdot)$ both sides and substitute equation (40), which assumes $n_P > 1$, yielding $\ln TWR$:

[3] In reality an investor with no skill often is an investor with psychological biases which may imply less than 90% of the accumulated excess wealth of the benchmark portfolio.

$$\ln TWR_{n>1} = t\left(g_{\infty}^{e,P} - g_{\infty}^{e,BM}\right) = t\Delta DP_{n>1}^{BM}$$
$$= -\frac{Ivar_{n=f=1}}{2}\left(\frac{1}{n_P} - \frac{1}{n_{BM}}\right) f^2 t, \tag{133}$$

which is the difference in expected excess capital growths between the investment portfolio and the benchmark.

We can solve the n_P giving the average number of stocks required to achieve the minimum target for gross compound excess wealth ratio $TWR_{n>1}$ as a function of idiosyncratic variance $Ivar_{n=f=1}$ estimated based on regression equation (47):

$$n_P = \frac{Ivar_{n=f=1}}{Ivar_{n=f=1} f^2 t / n_{BM} - 2\ln TWR_{n>1}} f^2 t, \tag{134}$$

which in the limit, when $n_{BM} \to \infty$, substituting equation (39) gives:

$$n_P^{n_{BM}\to\infty} = \frac{Ivar_{n=f=1}}{-2\ln TWR} f^2 t = \frac{1}{-\ln TWR} DP_{f=1}^{BM} f^2 t. \tag{135}$$

When n_{BM} is large compared to resulting n_P, equation (135) can be used to approximate the average number of stocks required to achieve the minimum target for gross compound excess wealth ratio TWR.

Equation (135) simply and intuitively determines the average number of stocks required as a function of targeted diversification benefit level $1/(-\ln TWR)$, diversification premium of an unlevered benchmark portfolio $DP_{f=1}^{BM}$, square of an investment fraction f^2 and finally time t. Remarkably, after deciding the desired diversification benefit level target, fraction of portfolio allocated to stocks and the investment time horizon the average number of stocks required in a portfolio is a function of diversification premium of the unlevered benchmark portfolio alone.

Alternatively, we can solve the n_P giving the average number of stocks required to achieve the minimum target for gross compound excess wealth ratio TWR, by substituting equation (50) into equation (133), as a function of alpha $\alpha_{n=f=1}$ estimated based on regression equation (46):

$$n_P = \frac{1}{1/n_{BM} + \ln TWR_{n>1}/\alpha_{n=f=1} f^2 t}, \tag{136}$$

or by substituting equation (56) into equation (133) as a function of R-squared $R^2_{n=f=1}$ estimated based on regression equation (47):

$$n_P = \frac{1}{1/n_{BM} + 2\ln TWR_{n>1}/\left[\left(1 - 1/R^2_{n=f=1}\right) Var_{BM} f^2 t\right]}. \tag{137}$$

In the limit, when $n_{BM} \to \infty$, equations (136) and (137) simplify to:

$$n_P^{n_{BM}\to\infty} = \frac{\alpha_{n=f=1}}{\ln TWR} f^2 t, \tag{138}$$

and:

$$n_P^{n_{BM}\to\infty} = \frac{\left(1 - 1/R^2_{n=f=1}\right) Var_{BM}}{2\ln TWR} f^2 t, \tag{139}$$

respectively.

When n_{BM} is large compared to resulting n_P,, equations (138) and (139) can be used to approximate the average number of stocks required to achieve the minimum target for gross compound excess wealth ratio TWR.

By restructuring equation (139) we can maximize our intuition about what affects the average number of stocks required to achieve the minimum target for gross compound excess wealth ratio TWR. The restructured equation has six terms:

$$n_P^{n_{BM}\to\infty} = \frac{1}{2}\frac{1}{-\ln TWR}\left(\frac{1}{R^2_{n=f=1}} - 1\right) Var_{BM} f^2 t. \tag{140}$$

The first term $\frac{1}{2}$ is a constant. The second term $\frac{1}{-\ln TWR}$ is a multiplier entailing the impact from the diversification target. The term, when the target is in its reasonable range, say from 90% to 99% of the maximum diversification benefit, is roughly approximated by $1/(1 - TWR)$. This yields approximate multipliers 10, 20 and 100 for targets 0.9, 0.95 and 0.99 respectively. The third term $\left(\frac{1}{R^2_{n=f=1}} - 1\right)$ is the size, in relation to systematic risk component Var_{BM} size, of one–stock portfolio risks not explained by the benchmark, i.e., the relative size of idiosyncratic, costly risk. R-squared determines how representative the benchmark is to individual stock risks, i.e., what proportion of the individual stock risks is explained by the benchmark. Fourth component Var_{BM} is the absolute size of the systematic risk. Fifth component $f^2$is the square of the investment fraction. Finally, the sixth component is the length of the investment time horizon.

The interpretation is that number of stocks required is higher when the diversification target level is higher or when the multiplication of third and fourth term yielding the absolute size of the idiosyncratic risk is high, i.e., when relative idiosyncratic variance size is high and/or the systematic variance is high. Investment fraction’s contribution to the required number of stocks is important to the second power and investment horizon length scales the requirement linearly with slope equaling one.

A more practical interpretation may give investor some valuable viewpoint: The lower the level of diversification, the more skill is needed just to overcome the portfolio growth rate difference to benchmark to reach the level from where additional skill starts to benefit the selected strategy. The more exposure an investor have to compensated risk factors (e.g. by exposure to investing styles), the lower the proportion of costly idiosyncratic risk and the greater the expected portfolio growth rate. The riskier the benchmark, i.e., the population of the selected stock picking universe, the more severe the impact from the idiosyncratic risk exposure is. In other words, stock picking from a low systematic risk universe (benchmark) requires less diversification compared to stock picking from high systematic risk universe. If an investor intends to leverage his portfolio, diversification becomes of extreme importance. The longer the time horizon, the more important diversification becomes.

Low level of diversification combined with lack of compensated investing style(s), high systematic risk benchmark, leverage and long investment horizon is the recipe for financial underperformance and, especially with leverage, financial ruin.

Finally, the most concise interpretation what affects the number of stocks required in a portfolio is provided by equation (135). We can see that it is all about total idiosyncratic variance which consist of average unlevered idiosyncratic variance of a single stock portfolio levered by the square of investment fraction and time. In other words, it all depends on the size of the exposure to the costly idiosyncratic risk.

3.4.3.2 Risk averse wealth ratio

Equations in section 3.4.3.1 are based on gross excess wealth accumulated by compounding instantaneous geometric risk premium over time. This, however, will not account for the risk and can be considered as risk neutral metric. To account for the risk and to derive risk averse version for the diversification effect expressed as gross excess wealth ratio, we derive equations utilizing realizable gross excess wealth.

Target realizable gross compound excess wealth ratio $T\widetilde{W}R_{n>1}$ for a portfolio with $n_P > 1$ (tilde accent denotes "realizable") can be derived similarly as we derived target realizable risk premium ratio in equation (128). The difference is that we now replace risk premium ratio with gross compound excess wealth ratio and, by substituting equation (132), arrive to equation:

$$T\widetilde{W}R_{n>1} = \widetilde{W}^{P}_{EW,G\infty} / \widetilde{W}^{BM}_{EW,G\infty} = \frac{R(t)^2_P}{R(t)^2_{BM}} TWR_{n>1}$$

$$= \frac{1 + 1/\left[t\left(SR^{BM}_{G\infty}\right)^2\right]}{1 + 1/\left[t\left(SR^{P}_{G\infty}\right)^2\right]} e^{t\Delta DP^{BM}_{n>1}} =$$

$$\frac{1 + 1/\left(t\left[(SR^{BM}_{\infty})^2 - f\sqrt{Var_{BM}}SR^{BM}_{\infty} + \frac{f^2}{4}Var_{BM}\right]\right)}{1 + 1/\left(t\left[\frac{Var_{BM}}{Var_P(n_P)}(SR^{BM}_{\infty})^2 - f\sqrt{Var_{BM}}SR^{BM}_{\infty} + \frac{f^2}{4}Var_P(n_P)\right]\right)} \times e^{t\Delta DP^{BM}_{n>1}(f,n_P)}, \quad (141)$$

which contains two terms which are functions of n_P: portfolio variance $Var_P(n_P)$ and diversification premium difference to benchmark $\Delta DP^{BM}_{n>1}(f, n_P)$ determined by equations (33) and (40), respectively. n_P can be solved by computer.

Target realizable gross compound excess wealth ratio accounts for the risk (noise) associated with realizable risk premium which compounds to gross excess wealth over time. The greater the proportion of the realized risk premium explained by noise, the lower the weight given to risk premium. After the risk weighting, resulting risk adjusted risk premium is used as the excess growth rate when compounding the gross excess wealth over time. This process accounts for the spread of the gross excess wealth distribution. Diversification makes a great difference to the resulting wealth distribution and the spread in the distribution can be considered as caused by noise and representing risk.

3.5 Hypotheses about the empirical diversification effects

We will determine seven empirically testable hypothesis based on the derived theory.

Hypothesis 1: Taking less risk, by investing a fraction of the portfolio in riskless rate, can increase the geometric risk premium for poorly diversified portfolios. This result, and the geometric risk premium in general, is explained by the fractional Kelly criterion.

Hypothesis 1 is based on section 3.3.4 including several equations which support this hypothesis. In particular equation (62) shows how full Kelly point is a decreasing function of portfolio's variance. Variance will be high for poorly diversified portfolios which implies low full Kelly point. Figure 2 shows how geometric risk premium has a maximum at fully Kelly point and increasing stock allocation beyond the full Kelly point will decrease the risk premium.

Hypothesis 2: Diversification is a negative price lunch implying that diversification premium exists.

Hypothesis 2 is a direct consequence of equation (8) where we can see that any additional variance caused by less than perfect diversification will decrease the geometric risk premium. Equation (34) determines diversification premium which is a formal proof that diversification is a negative price lunch as described in section 3.3.3.

Hypothesis 3: One half of the portfolio's idiosyncratic variance closely approximates the magnitude of diversification premium difference to benchmark, the opportunity cost of foregone diversification.

Hypothesis 4: Diversification premium difference to benchmark for a portfolio is a function of portfolio's squared investment fraction.

Hypotheses 3 and 4 are based on equation (42) as explained in section 3.3.3.

Hypothesis 5: For a risk neutral long-term investor, the number of stocks required to make a diversified portfolio is approximately directly proportional to investment time horizon length.

Hypothesis 6: For a risk neutral long-term investor, the number of stocks required to make a diversified portfolio is an increasing, approximately squared, function of investment fraction.

Hypotheses 5 and 6 are based on equation (135) as explained in section 3.4.3.

Hypothesis 7: Number of stocks required to make a diversified portfolio is a function of investment style.

Hypothesis 7 is based on equation (140) which shows that both exposure to risk factors and the variance of the benchmark portfolio affect the number of stocks required to make a diversified portfolio.

4 DATA AND RESEARCH METHODOLOGY

4.1 Data description

4.1.1 Three distinct periods and stock populations

The stock data used in empirical analysis is retrieved from the merged Center for Research in Securities Prices (CRSP) and Compustat database. The data contains monthly U.S. stock market return data from January 1926 to June 2018. One-month Treasury bill data, our proxy for riskless rate, is retrieved from Kenneth French's website. Combining the stock data with available riskless rate data yields monthly excess return data spanning from July 1926 to Jun 2018 totaling 92 years. We include all common stocks which comprises of CRSP share codes 10, 11 and 12. Delisting returns are taken into account when calculating stock specific monthly excess returns. Generally, our stock return data selection follows Bessembinder (2018), with the exception that our time span extends 18 months longer.

In addition, we utilize annual accounting-based data for U.S. stock market from Compustat. The data from Compustat is merged with the CRSP/Treasury bill based monthly excess return data and selected risk factor data. The risk factor data is retrieved from Kenneth French's website. Combined CRSP/Compustat/Treasury bill/risk factor data is utilized selectively for the time period spanning from August 1962 to June 2018.

Initially, as visualized by Figure 8, starting from July 1926, our stock excess return data contains the universe of NYSE securities. Starting from August 1962, the universe of NYSE Amex securities is included alongside NYSE securities. The stock universe of Nasdaq securities was included into CRSP database in 1973. Since January 1973, the CRSP monthly database includes the whole universe for common stocks comprising of all three major stock exchanges in the U.S.: NYSE, NYSE Amex and Nasdaq. Our primary interest therefore is the data spanning from January 1973 to June 2018, totaling 45.5 years and containing the whole U.S. common stock universe. We call this time span as "the modern era of U.S. stock returns".

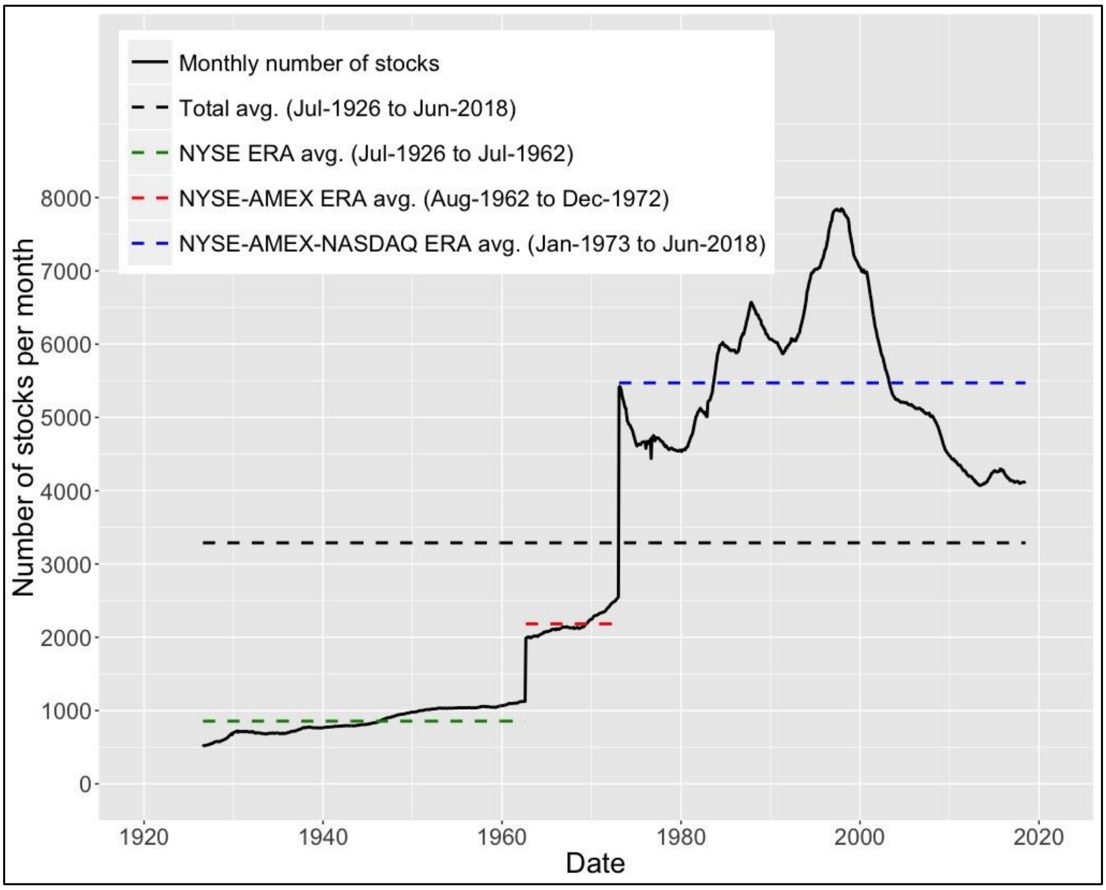

Figure 8. Number of stocks per month over the period from Jul-1926 to Jun-2018.

We can see from Figure 8 that there are three distinct time periods which have very different average number of stocks. The average number of stocks is 856, 2183, 5472 and 3289 for period 1, period 2, period 3 and the whole time span, respectively. The monthly number of stocks varies substantially in the third period which includes the whole universe of common stocks.

In Figure 9 we can see how the average firm size characteristics change when we change from period to another. We follow Fama and French (2008) and use NYSE breakpoints to distinguish between small and big stocks (the 50th percentile) and microcaps (the 20th percentile). In addition, we show megacaps as the 80th percentile. We assign the stocks to percentiles monthly.

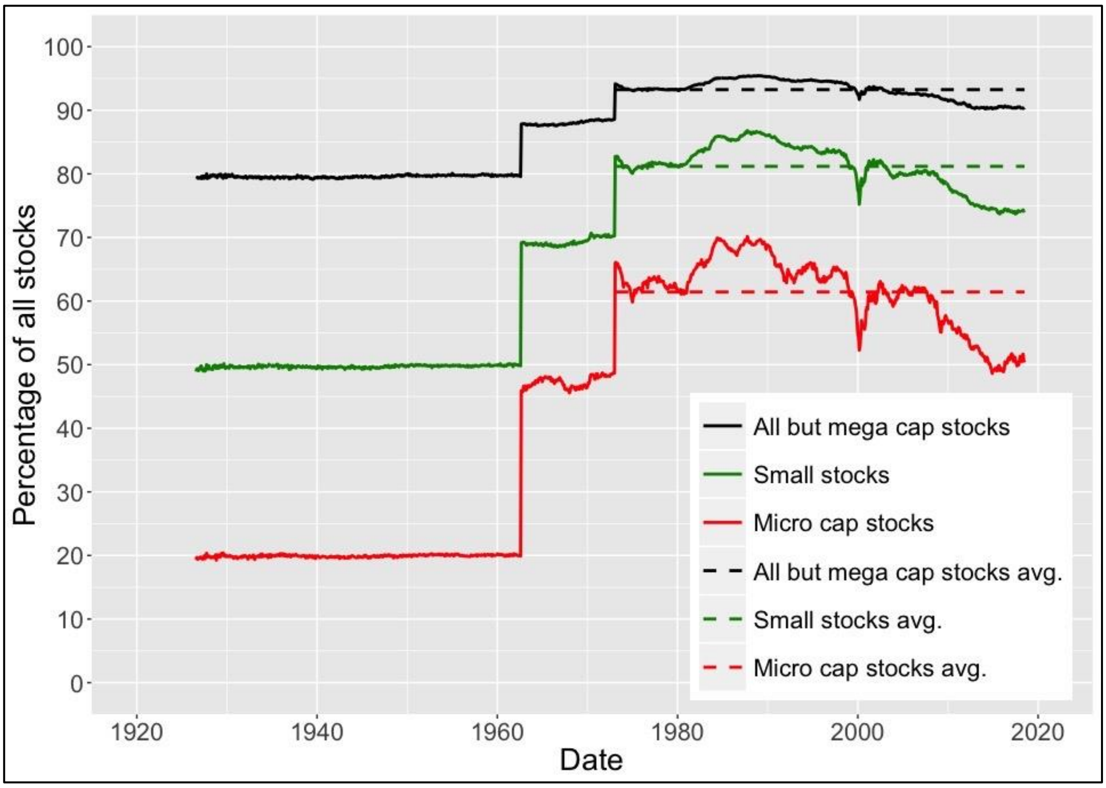

Figure 9. Monthly firm size representation based on NYSE breakpoints over the period from Jul-1926 to Jun-2018.

In the first, the NYSE, period we can see that the NYSE breakpoints accurately describe the firm size characteristics. Measured as number of stocks, big stocks represent 50%, small stocks 30% and microcaps 20% of the stock universe. Megacaps are the top 20%. This is as expected as we only have the NYSE stocks in our stock population. In the second, the NYSE Amex, period we see a big change as smaller Amex firms are added into our population of stocks. Now microcaps populate close to half of all stocks and big stocks represent about 30% of all stocks. Yet another big move takes place as on average tiny Nasdaq stocks are added into database at the beginning of the third period in January 1973. The modern era of U.S. stock returns is dominated by microcaps which on average represent more than 60% of all stocks while big stocks are left with 20% and megacaps just slightly more than 5% of the stock population. In addition, we see that the representation of firm sizes fluctuates significantly in the third period.

Looking at Figure 8 and Figure 9, it is clear that the time span from July 1926 to June 2018 is not a period of one common stock population. We can see three distinct stock populations which differ both by the number of stocks and by the firm size characteristics.

It is worth noting that just based on seeing these two figures, Figure 8 and Figure 9, the number of stocks required to make a diversified portfolio can be expected to differ significantly depending which time period the data is from. This potentially is one reason why the early literature settles to lower number of stocks compared to more recent studies.

4.1.2 Descriptive statistics

Next, in Table 1, we show some descriptive statistics from the whole period from July 1926 to June 2018 and the three identified sub periods. We show statistics for number of months and stocks (# designates "number of"), for equally weighted market portfolio and for average single stock portfolio. The idiosyncratic variance and idiosyncratic standard deviation, Ivar and Isdev, respectively, are calculated with normalized data based on regression equation (47). This means that single stock portfolio statistics are not for pooled data over the time period, but for average monthly values. This ensures each month has equal weight regardless the number of stocks that month. We can see from the data that the last period has the greatest average number of stocks per month. By using pooled data, the last period would greatly dominate the statistics from that pool. By eliminating the effect of number of stocks from the single stock portfolio data, we eliminate the dominance of highly populated months and simultaneously ensure the statistics is compatible with our random sampling (bootstrapping) method. We use bootstrapping to create random portfolios from monthly data and compare the characteristics of bootstrapped portfolios. Bootstrapping results reflect equal weight for each month regardless the number of stocks populating each month.

We can see from the data that number of stocks is the greatest in the last period. This correlates with the idiosyncratic variance for single stock portfolios, which is in line with the dominance of microcap stocks populating the last period. Just by comparing the single stock standard deviation and idiosyncratic standard deviation (or idiosyncratic variance) figures between periods, it is possible to conclude the last period, by a wide margin, must have the greatest potential for diversification benefit. It also means we have three different periods with different stock populations and analyzing them as one period extending from 1926 to 2018 is not the most informative choice. We therefore will mostly focus on the most recent period from January 1973 to June 2018, which also contains most of the monthly return data, close to three million samples.

Table 1. Descriptive statistics for empirical data.

Metric	Jul-1926 to Jun-2018	Jul-1926 to Jul-1962	Aug-1962 to Dec-1972	Jan-1973 to Jun-2018
#Months	1104	433	125	546
Avg. #stocks per month	3289	856	2183	5472
#Distinct stocks	26373	1645	3511	24952
#Distinct returns	3630744	370453	272815	2987476
EW MKT geom. RP	0.0885	0.1038	0.0824	0.0778
EW MKT avg. return	0.1516	0.1628	0.1447	0.1442
EW MKT avg. RF rate	0.0329	0.0131	0.0442	0.0460
EW MKT geom. RP Sdev	0.2451	0.3003	0.1923	0.2041
EW MKT SR	0.4841	0.4984	0.5225	0.4812
Avg. 1-stock geom. RP Ivar	0.2125	0.1360	0.1158	0.2953
Avg. 1-stock geom. RP Isdev	0.4610	0.3688	0.3403	0.5434
Avg. 1-stock geom. RP Sdev	0.5221	0.4755	0.3909	0.5805
Avg. 1-stock SR	0.2272	0.3147	0.2571	0.1692

4.2 Research methodology

4.2.1 Log-normal simulator

4.2.1.1 Motivation and description of the log-normal simulator

Log-normal simulator is implemented to allow for large samples of data to be created using parameters extracted from empirical data. We have 92 years of empirical data of which last 45.5 years is our primary stock population. In many cases, single 45.5-year history is too little to confidently draw conclusions from statistics. This is where the log-normal simulator steps in. Using the simulator, we can create as many, as long and as broad alternative histories as we like, limited only by simulation time constraints.

The log-normal simulator utilizes the fact that systematic and idiosyncratic variances are orthogonal by definition. We first create alternative histories (set of periods) for the benchmark by randomly drawing equally weighted monthly market (EW MKT) returns from a log-normal distribution. Empirical EW MKT mean return and standard deviation are used as parameters. We will use each randomly drawn logarithmic EW MKT return (logarithm of one plus return) as an arithmetic mean return for the individual stock returns for that particular period. For each period, monthly individual stock returns are then randomly drawn from log-normal distribution by utilizing the period specific mean return drawn in the first step and forming the standard deviation by summing the empirical systematic and idiosyncratic variances and taking a square root. By selecting the number of periods and number of stocks per period, we can draw as long (number of periods) and broad (number of stocks per period) alternative histories as we like.

4.2.1.2 Simulation to select the regression output for empirical tests

We can use three alternative inputs to equations determining diversification premium and diversification premium difference to benchmark. Inputs can be idiosyncratic variance, alpha or R-squared, which is used together with benchmark variance, obtained from a regression utilizing exhaustive single stock log return data. This is described in details in sections 3.3.1 and 3.3.2.

Based on our equations, the three alternative regression outputs are asymptotically identical. This means they are supposed to converge to same value given an infinite amount of data. In order to test the convergence properties of these three candidate metrics with finite data, we utilize a log-normal simulator to create data for different time period lengths. We need to have a large number of very long time series to be sure about the convergence which means empirical data with one realization of maximum length of 92 years or 1104 months will not be enough.

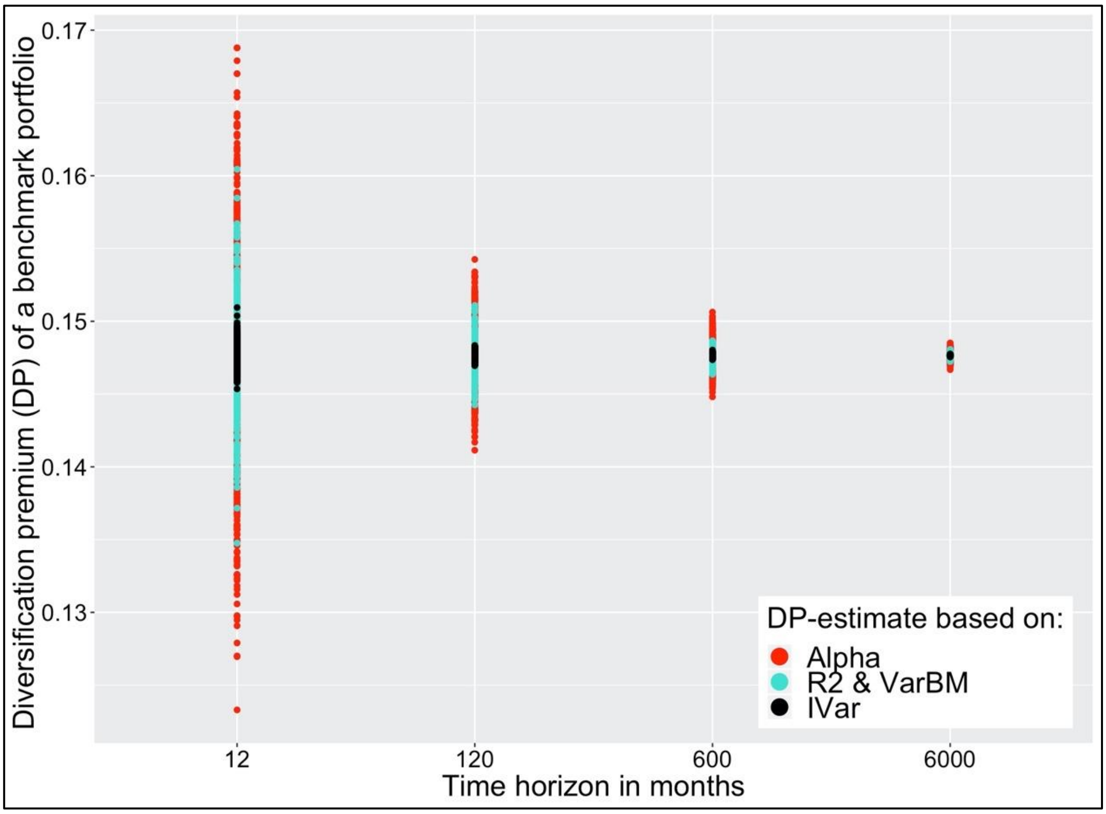

Figure 10. Convergence of a diversification premium of a benchmark portfolio by utilizing alternative estimates.

Figure 10 shows how the estimate of diversification premium of a benchmark portfolio converges when alternative estimates based on idiosyncratic variance, R-squared and benchmark variance or alpha are used. The simulation uses monthly individual stock return data for time period lengths of 1, 10, 50 and 500 years. The parameters of the data are acquired from the empirical modern era of U.S. stock returns. Each estimate is calculated 500 times per time horizon length. Each estimate is calculated based on simulation which randomly creates 5472 individual monthly log-normally distributed stock returns for each month belonging to the time period.

Firstly, we can see from Figure 10 that all three estimates indeed do seem to converge to common value. Secondly, there are clear differences in how quickly the estimates converge. Alpha based estimate, as in equation (51), is the slowest to converge while idiosyncratic variance-based estimate, as in equation (39), is the quickest. R-squared and benchmark variance-based estimate, as in equation (57), is in between. Based on this result we choose to use the idiosyncratic variance-based metric in our empirical tests as it should be the most accurate given the finite empirical data.

As we will use idiosyncratic variance-based metric, it implies that the regression specification to be used in empirical tests is given by equation (47).

4.2.1.3 Simulating forward-looking geometric premiums with uncertain risk

We hypothesized in section 3.3.9 that, because of uncertainty about future risks, forward-looking geometric risk premium will be overestimated and diversification premium underestimated when estimates are based on historical averages. We now test and confirm this hypothesis with log-normal simulator.

In case of geometric risk premium, we create the uncertainty about risk by not using the expected value of standard deviation in simulation, but instead introducing a standard deviation for the standard deviation. This is achieved by subtracting 0.05 from the standard deviation in half of simulation runs and adding 0.05 in other half of the runs. This creates a standard deviation of 0.05 for the standard deviation. This is repeated using values 0, 0.05, 0.1 and 0.2. Value 0 corresponds to case where we have no uncertainty about risk and the expected value for the standard deviation is used. Historical values from January 1973 to June 2018 are used for excess return and standard deviation in this simulation. We use the log-normal simulator to create 1000 years of return history 100 000 times for both low and high standard deviation values and use 5472 monthly firm returns. Historical idiosyncratic variance is used as a parameter. We average the simulation results and end up with aggregated 200 000 000 years of simulated return history where we calculate the geometric risk premium. Geometric risk premium converges extremely slowly to its expectation, but using such very long data set we achieve convergence. Results of the simulation are shown in Figure 11. Simulation is run for each investment fraction value with step size 0.01

yielding up to about 500 simulations per one curve. We can see that risk premium is lower when standard deviation of standard deviation is higher. Predicted and simulated risk premium curves are practically identical and predicted full Kelly point locates at maximum risk premium. This simulation shows that equations (116) and (117), predicting geometric risk premium and full Kelly point in the presence of uncertainty about risk, respectively, work.

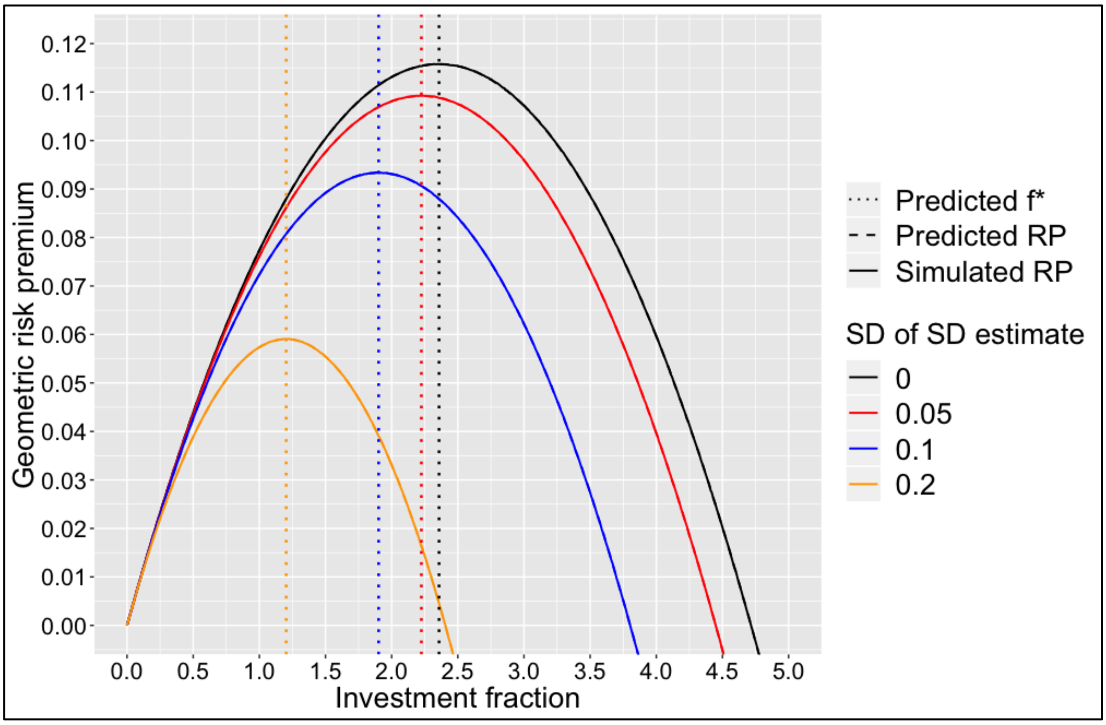

Figure 11. Risk premium and full Kelly fraction in the presence of uncertainty about risk.

We run similar simulation for diversification premium. We repeat the simulation for three investment fraction values. This time, for the lowest non-zero standard deviation for standard deviation value, 0.05, we create the uncertainty by drawing the standard deviation from normal distribution with standard deviation 0.05. For the larger values, 0.1 and 0.2, we use the same method as in risk premium simulation to avoid negative standard deviation values. We simulate 20 times 500 years with 5472 monthly firm returns resulting to 10 000 years of aggregated data. We have a lot of data, but diversification premium still converges much quicker compared to geometric risk premium. Simulation calculates diversification premium using regression equation (47). We can see from Figure 12 that predicted diversification premium match with simulated value. Based on this, equation (119), predicting diversification premium in the presence of uncertainty about risk, works.

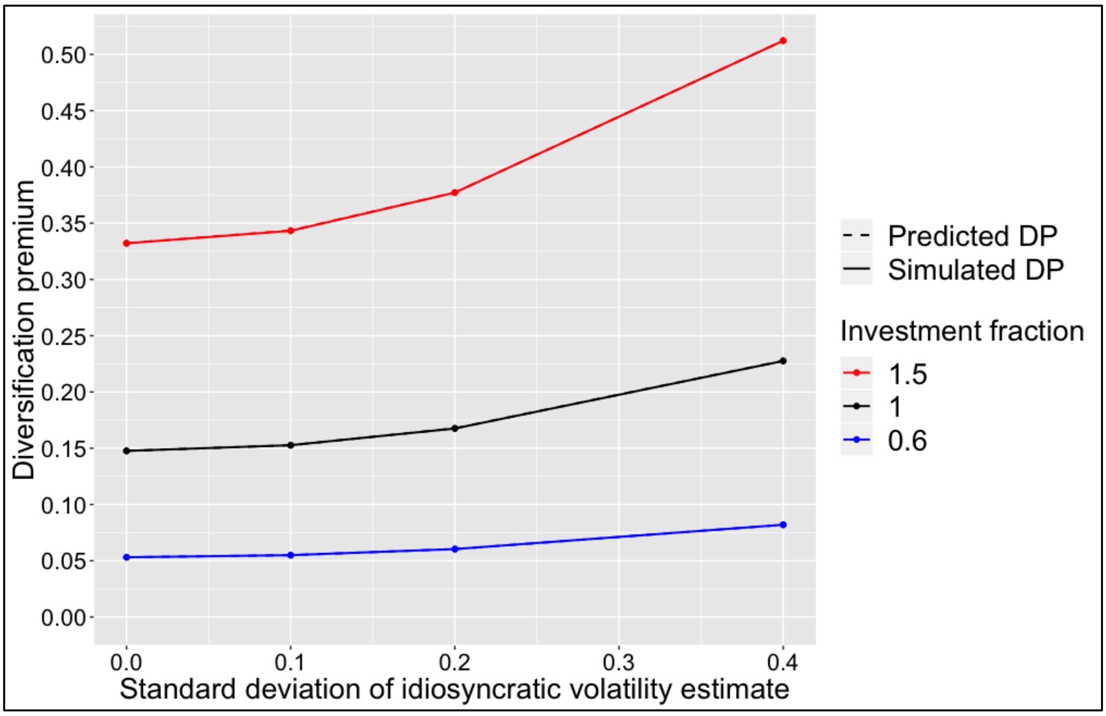

Figure 12. Diversification premium as a function of uncertainty about idiosyncratic risk.

4.2.2 Empirical data

The empirical part of the research is based on randomly forming stock portfolios of different sizes from selected benchmark stock populations and utilizing equations derived in section 3 to measure diversification effects. Portfolios are formed monthly. All portfolios, including benchmarks, are formed by weighting the stocks equally and by rebalancing the weights on a monthly basis.

The method used to form random portfolios from empirical data is bootstrapping without replacement. Bootstrapping without replacement means random picking without replacing the picked stock return in the data. This implies each stock can be picked only once per month per portfolio. Investment fraction is simply a multiplier for the monthly excess returns. This implies we assume the cost of borrowing to be equal to interest rate of monthly T-bills. This is not a realistic assumption for an average investor implying the geometric risk premiums for leveraged portfolios are theoretical upper bounds.

Research methodology is based on the properties of the geometric risk premium. We utilize the geometric risk premium as given in equation (8), implied results from section 3.3 and diversification effect metrics defined and derived in section 3.4. All these are explicitly described by equations.

We use monthly empirical data for individual stock returns and riskless rate. If we denote monthly return for i-th stock in the month t as $m\_ret_{t,i}$ and monthly riskless rate in the month t as m_rf_t, we calculate monthly excess return for i-th stock in month t as

$$m\_eret_{t,i} = \exp\left[\ln\left(1 + m\_ret_{t,i}\right) - \ln(1 + m\_rf_t)\right] - 1.$$

Monthly excess return for i-th stock, $m\_eret_{t,i}$, is used for empirical calculations. For example, bootstrapping process randomly pick monthly excess returns to form portfolios of selected size. Scaling by investment fraction f is implemented in the empirical data by multiplying each $m\_eret_{t,i}$ by f.

The empirical parameters used in equation (8) (and in many other equations we have derived) are the mean instantaneous (arithmetic) excess return per year, m_e, and the standard deviation for the continuously compounded excess growth per year, s_e.

Mean portfolio instantaneous (arithmetic) return per year, m, is calculated from monthly returns by first calculating monthly means over all portfolio's stock returns per month ($mean_i$) and then the mean over monthly means ($mean_t$) as follows:

$$m = 12 * ln\left(1 + mean_t[mean_i(m_{ret_{t,i}})]\right).$$

Mean instantaneous riskless rate per year, r, is calculated from monthly riskless rates as a mean over all months as follows:

$$r = 12 * ln[1 + mean_t(m\_rf_t)].$$

Mean instantaneous (arithmetic) excess return per year, m_e, is calculated by subtracting mean instantaneous riskless rate from mean instantaneous return:

$$m_e = m - r.$$

The portfolio standard deviation for the continuously compounded excess growth per year, s_e, is calculated from portfolio's monthly excess returns by first calculating mean over all stocks per month and then calculating the standard deviation for the logarithmic monthly means:

$$s_e = \sqrt{12} * Sdev\big[ln(1 + mean_i[m\_eret_{t,i}])\big].$$

By first calculating the mean over all stocks per month, when calculating m and s_e, ensures that each month is equally weighted regardless the number of stocks in the data per month. When bootstrapping is used to form the portfolios, then equal number of stocks per month is guaranteed, but for example for benchmark portfolios the number of stocks vary per month.

Any relevant additional information, specific to each empirical test, will be given in the related section where test and the results are described.

5 EMPIRICAL DIVERSIFICATION EFFECTS

5.1 Fractional Kelly criterion explaining geometric risk premium

5.1.1 Predicted vs. realized geometric risk premiums

We know from equation (114) that approximate instantaneous geometric risk premium of a portfolio is equal to instantaneous geometric risk premium of the benchmark portfolio plus portfolio's diversification premium difference to benchmark. By further substituting equations (8) and (38) we can write the approximate instantaneous geometric risk premium equation as:

$$RP_{EW,G\infty}^{P,n_{BM}\to\infty} = RP_{EW,G\infty}^{BM} - \frac{DP^{BM}}{n_P}$$
$$= f\sqrt{Var_{BM}}SR_{\infty}^{BM} - \frac{f^2}{2}\left(Var_{BM} + \frac{Ivar_{n=f=1}}{n_P}\right). \qquad (142)$$

The above equation is the same form as fractional Kelly equation (73). The variance of the portfolio is the sum of systematic and idiosyncratic variances. Based on this, we can use the fractional Kelly criterion in the form of equation (142) to predict the instantaneous geometric risk premium of a n_P–stock portfolio. Remarkably, we only need three empirical parameters to predict the instantaneous geometric risk premium at any given investment fraction for any given portfolio size (if we used the exact equation (111), we would additionally need average monthly number of stocks in the benchmark portfolio). The needed three parameters are the instantaneous Sharpe ratio of a benchmark portfolio (instantaneous excess return would be equally good), the variance of a benchmark portfolio and the average idiosyncratic variance of a single stock portfolio. The last parameter is acquired as an output from the regression described by equation (47) and the two former parameters are calculated from the benchmark portfolio's time series data. In case of benchmark portfolios, we set the idiosyncratic variance of the portfolio to zero instead of setting n_P to average number of stocks in the benchmark portfolio. This has very small impact on prediction accuracy as the number of stocks in our benchmarks is so large, ranging from 1001 to 5472.

We will next use fractional Kelly based equation (142) to predict the instantaneous geometric risk premium as a function of investment fraction for different portfolio sizes in the time period from January 1973 to June 2018, in the modern era of U.S. stock returns. According to Goetzmann and Kumar (2008) and Polkovnichenko (2005), typical individual investor is poorly diversified. We will therefore use relatively small portfolio sizes in our tests. Small portfolio sizes also allow us to best observe the effect of diversification.

Bootstrapping without replacement is used to form portfolios from empirical data. We continue leveraging the bootstrapped portfolios as long as we encounter the first occasion where the loss exceeds 100% for a portfolio. This should theoretically never happen based on the Kelly criterion. However, theoretically we are rebalancing continuously at infinite frequency and there are no jumps in returns. In reality in our test we are rebalancing once per month and the empirical excess returns have fat tails meaning they entail sudden and large moves. In the figures we can observe where we first encountered a loss greater than 100% as lines for each portfolio size end at that point. This is one additional way to see the effect of diversification. The more diversification, the more leverage the portfolio tolerates before we first observe that some investor suddenly loses more than he has. As an example, single stock portfolio does not tolerate leverage at all. This is because some stocks have a monthly excess return of around -99% and when leveraged they immediately exceed the 100% loss threshold. Single stock portfolio curves therefore typically end exactly at investment fraction one.

In the figures, we denote the predicted geometric risk premiums with dashed lines and the geometric risk premiums for bootstrapped, randomly from empirical data formed portfolio's, with solid lines. Portfolio sizes are color coded. Note that one stock portfolios are exhaustive meaning that there is no bootstrapping but all available data is used to calculate the geometric risk premiums. The tick-marks on the x-axis in the figures are set to investment fraction values 0.6, 1.0, 1.5, etc. The reason why we always have these values is that 0.6 corresponds to classic 60/40 allocation, 1.0 is the

100% stock allocation and 1.5 corresponds to 50% leveraged portfolio which we will consider as an example of leveraged portfolio in our later analysis.

For bootstrapping we form 50 000 portfolios each month. For the benchmark portfolio (BM), we show the average number of stocks per month calculated over the whole time period. As there is only one benchmark portfolio, we don't use bootstrapping but use directly the risk premium calculated for the benchmark.

In Figure 13 we have the instantaneous geometric risk premiums calculated from the population of all stocks. We first observe the parabolic shape of the curves as expected based on the fractional Kelly criterion as depicted in Figure 2 and Figure 11.

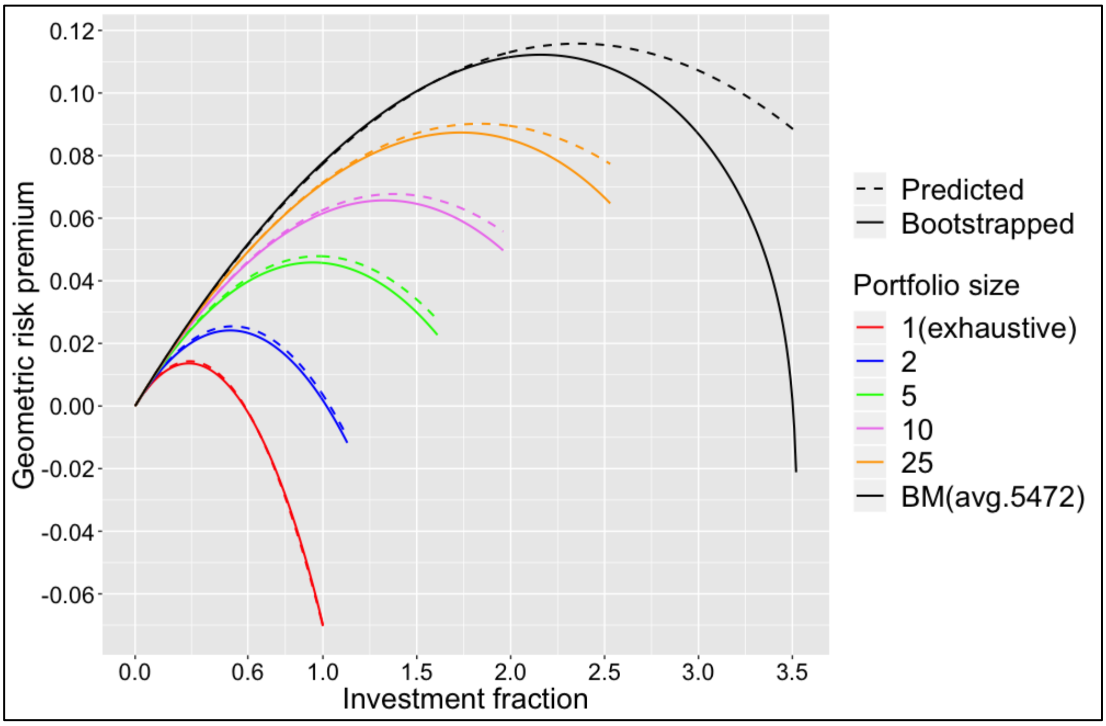

Figure 13. Geometric risk premium for all stocks between Jan-1973 and Jun-2018.

Predictions in Figure 13 are extremely accurate for the smallest portfolio sizes, especially for one stock portfolio size. Overall the predicted values match well with the bootstrapped values, but we can observe two deviations. First is that after investment fraction value of about 1.5, the bootstrapped values start to increasingly deviate from the predicted values. This is best visible for the benchmark portfolio. Second deviation from prediction is that for the intermediate portfolio sizes, from 2 or

5 to 25, we can see that there is some difference between predicted and realized risk premiums already at lower investment fractions.

Our hypothesis is that the first deviation from the prediction is related to rebalancing frequency. The Kelly criterion, and therefore the predicted value, assumes infinite rebalancing frequency and we rebalance monthly. We suspect that the increasing leverage as the investment fraction is increased, especially given the fat tails of the excess returns, contributes to the difference between the predicted and the realized risk premium. For the second deviation, we hypothesize it is caused by the fat tails of the monthly excess return distribution. The fat tail hypothesis will be studied in section 5.2.2.

Loud and clear, the Figure 13 says asset allocation, expressed by investment fraction in our two asset tests, and diversification make all the difference. Comparing the (red) single stock portfolio curve to the (black) benchmark portfolio curve tells it all. In a continuous-time world, as opposed to single period world, asset class risk premium of stocks is not generalizable to less than perfectly diversified stock portfolio's risk premium.

For the smallest portfolio sizes, one, two and five stocks, the geometric risk premium is lower at investment fraction one than at some lower investment fraction value. Against the core principle of the single period world, and finance in general, that risk and reward go hand in hand, poorly diversified continuous-time world investor on average earns a higher reward by investing part of his portfolio into riskless rate instead of stocks. In other words, higher reward is expected by taking less risk.

We can see from the two-stock portfolio size curve that it goes to zero at investment factor slightly greater than one. This means that positive geometric risk premium is achieved at portfolio size slightly smaller than two stocks. Using equation (126) we find that 1.91 stocks leads on average to zero geometric risk premium which agrees with our visual estimation from Figure 13.

One clear message is that too much leverage is too much. Every portfolio size has its limit, the Shannon limit, for maximum reward. Even the maximally diversified benchmark portfolio which has a predicted theoretical limit at 11.58% at investment fraction 2.36 (at full Kelly) and realized limit at 11.22% at investment fraction 2.16. However, we can see how the compounding process capacity, the Shannon limit, is very much a function of diversification. For comparison to fully diversified benchmark, single stock portfolio's maximum predicted theoretical risk premium is 1.43% at investment fraction 0.29 while realized maximum is 1.36% at the same investment fraction 0.29. Exceeding the full Kelly allocation increases the risk while decreases the reward. Especially we note that the realized reward is even lower than the predicted reward when full Kelly allocation is exceeded. The real-world reward to risk ratio for monthly rebalanced portfolios therefore is even worse than predicted reward to risk ratio for corresponding portfolios rebalanced at theoretical infinite frequency.

At investment fraction one the risk premium for the fully diversified benchmark portfolio is 7.74%. Corresponding realized value is 7.78%. This sounds as a rather large risk premium, but we need to bear in mind that this is the risk premium of an equally weighted portfolio dominated by microcap stocks. Value weighted portfolios would probably have a lower risk premium for the same time period. Also, we need to keep in mind that such equally weighted microcap loaded portfolio may not be entirely realistic benchmark as such fund or ETF would probably run into trouble with the low liquidity of the microcaps. Nevertheless, such equally weighted benchmark portfolio provides us a fair comparison in a sense that we tend to treat stock portfolios as equally weighted when we consider diversification benefits.

As we use equally weighted portfolios, the modern era of U.S. stock returns is dominated by microcap stocks as is visible in Figure 9. To have a better understanding of the risk premium, we present corresponding figures for three size groups: microcap stocks, small stocks and big stocks.

Average monthly number of big stocks plus small stocks plus microcap stocks is less than total average monthly number of stocks because there is one-month lag when assigning stocks to size groups. This means that not only the first month (Jan-1973) of the whole timespan is removed but also the first month of each individual stock added to data set during the time span.

In Figure 14 we have the geometric risk premium for microcap stocks. Predicted risk premiums are well in line with their bootstrapped counterparts, with similar caveats as in Figure 13. We can see that the risk premium of the two-stock portfolio is slightly negative. This implies that the number of stocks required to achieve a positive risk premium should be slightly higher than two. For portfolio sizes one, two and five stocks, the geometric risk premium is lower at investment fraction one than at some lower investment fraction value.

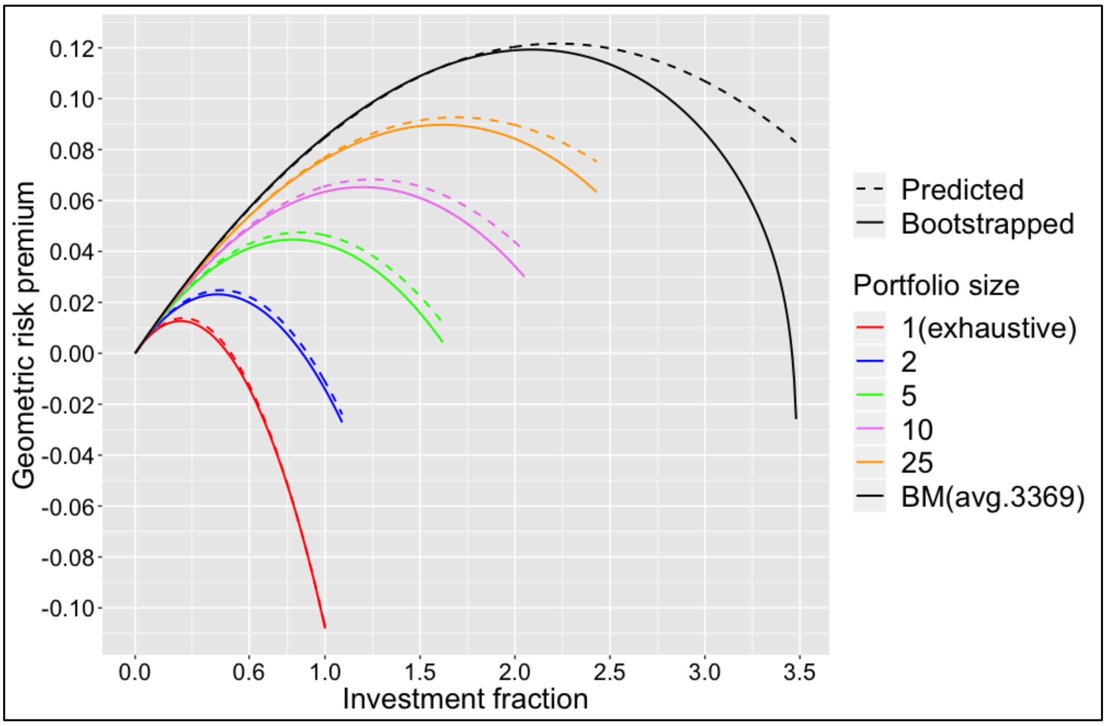

Figure 14. Geometric risk premium for microcap stocks between Feb-1973 and Jun-2018.

Figure 15 shows the risk premium for small stocks. Based on the figure, we expect that on average more than one stock is required to achieve a positive risk premium at investment fraction one. For portfolio sizes one and two stocks, the geometric risk premium is lower at investment fraction one than at some lower investment fraction value.

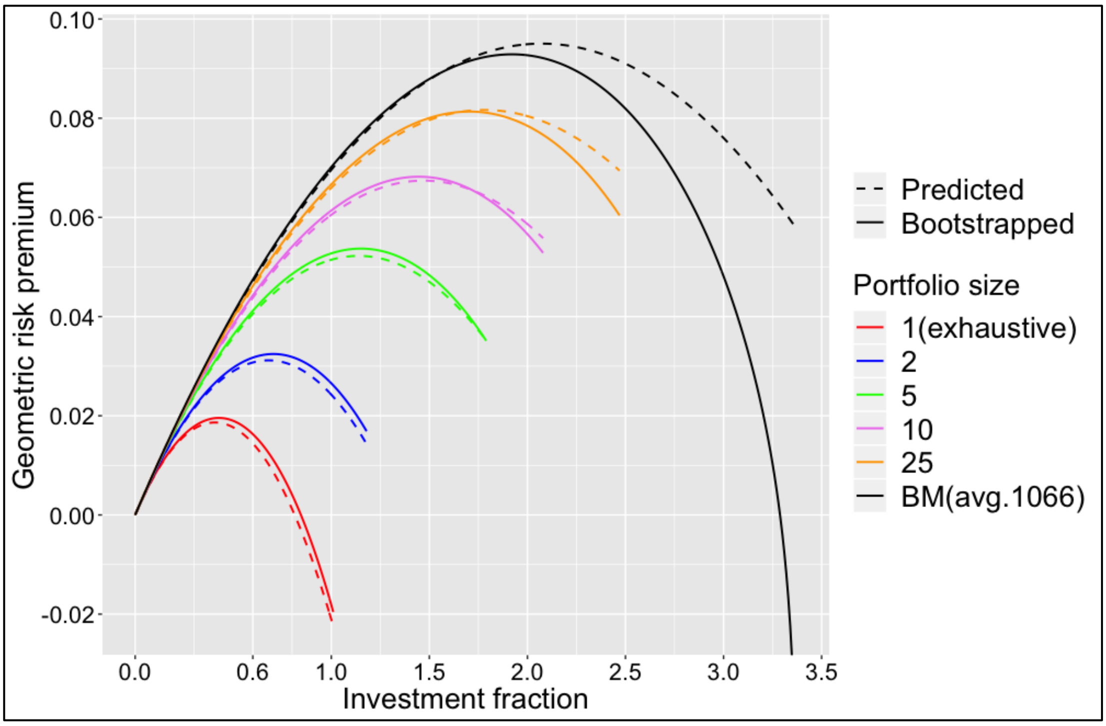

Figure 15. Geometric risk premium for small stocks between Feb-1973 and Jun-2018.

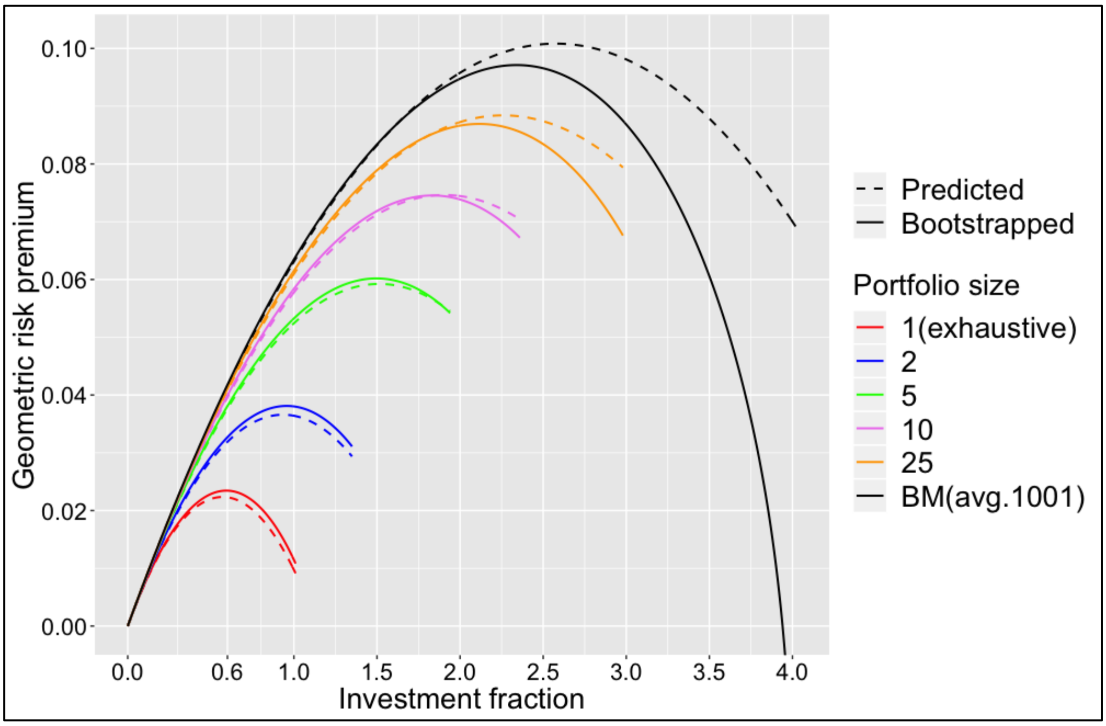

Figure 16. Geometric risk premium for big stocks between Feb-1973 and Jun-2018.

In Figure 16 we see the risk premiums for big stocks. The risk premium of the single stock portfolio is positive at investment fraction one. This means that the average number of stocks required to achieve a positive risk premium is less than one. For portfolio sizes one and two stocks, the geometric risk premium is lower at investment fraction one than at some lower investment fraction value.

Table 2 summarizes the parameters used in predicting the risk premiums in Figure 13 to Figure 16. Some key risk premium metrics from these figures are given in Table 3, where we have used the fractional Kelly criterion to predict the values and compare the predicted values to actually realized values for benchmark and single stock portfolio. Additionally, we show the minimum number of stocks required on average to achieve a positive risk premium calculated using equation (127).

Table 2. Parameters used as inputs to risk premium predictions.

	All stocks	Microcaps	Small stocks	Big stocks
SR_{∞}^{BM}	0.4812	0.4932	0.4360	0.4491
Var_{BM}	0.0416	0.0491	0.0442	0.0304
$Ivar_{n=f=1}$	0.2953	0.3838	0.1813	0.1067

We can see from Table 3 that overall the predictions are accurate. The predicted geometric risk premiums are particularly precise at investment fraction 1, at 100% stock allocation. There is some deviation in the predictions at the maximum risk premium and the associated full Kelly fraction. Our hypothesis is that this deviation is at least partly explained by our less than optimal rebalancing frequency combined with the sudden large changes by the fat tailed stock excess returns. The minimum number of stocks required on average to achieve a positive risk premium seems accurate when comparing to figures.

Table 3. Summary of predicted vs. realized risk premium metrics from Jan-1973 to Jun-2018.

	All stocks	Microcaps	Small stocks	Big stocks
Predicted RP_{EW,G_∞}^{BM}	7.74%	8.48%	6.96%	6.31%
Realized RP_{EW,G_∞}^{BM}	7.78%	8.52%	7.01%	6.35%
Predicted $RP_{EW,G_\infty}^{BM}(f^*)$	11.58%	12.16%	9.50%	10.08%
Realized $RP_{EW,G_\infty}^{BM}(f^*)$	11.22%	11.93%	9.23%	9.71%
Predicted $BM\ f^*$	2.36	2.22	2.07	2.58
Realized $BM\ f^*$	2.16	2.09	1.92	2.34
Predicted $RP_{EW,G_\infty}^{n=1}(f^*)$	1.43%	1.38%	1.85%	2.23%
Realized $RP_{EW,G_\infty}^{n=1}(f^*)$	1.36%	1.27%	1.96%	2.34%
Predicted 1-stock f^*	0.29	0.25	0.41	0.57
Realized 1-stock f^*	0.29	0.24	0.43	0.59
Min #Stocks for $RP_{EW,G_\infty}^{P} > 0$	1.91	2.26	1.30	0.85

Table 3 shows that at 100% stock allocation level bearing systematic risk is compensated. Microcap stocks has the highest and big stocks the lowest risk premium RP_{EW,G_∞}^{BM}. This is as we would expect as the microcap stocks are the riskiest and big stocks the least risky as shown by benchmark variances in Table 2. However, at full Kelly allocation f^* we find that the maximum risk premium, the Shannon limit, is not the lowest for big but for small stocks. This is explained by the relatively low volatility of the big stocks that allows for more leverage to achieve a higher full Kelly fraction. Notice that to predict the maximum risk premium, we only need to input the instantaneous Sharpe ratio into equation (66). Single stock portfolio's full Kelly fraction and maximum risk premium show how microcap single stock portfolio does not tolerate more than 24% stock allocation at maximum which is expected to deliver a dismal 1.27% risk premium. The greater the market cap, the greater the investment fraction that portfolio tolerates and the more the portfolio is expected to deliver at maximum allocation. Similar story is told by the minimum number of stocks required on average for positive risk premium. One stock portfolio delivers a positive risk premium for big stocks while small stocks require two and microcaps three stocks. For all stocks category the numbers are close to microcap numbers as most of the stocks are microcaps. A minimum of two stocks is needed to secure a positive risk premium for all stocks category.

In Figure 17 we show only the smallest portfolio sizes ranging from one to five stocks. This time also single stock portfolio size is bootstrapped instead of using exhaustive data.

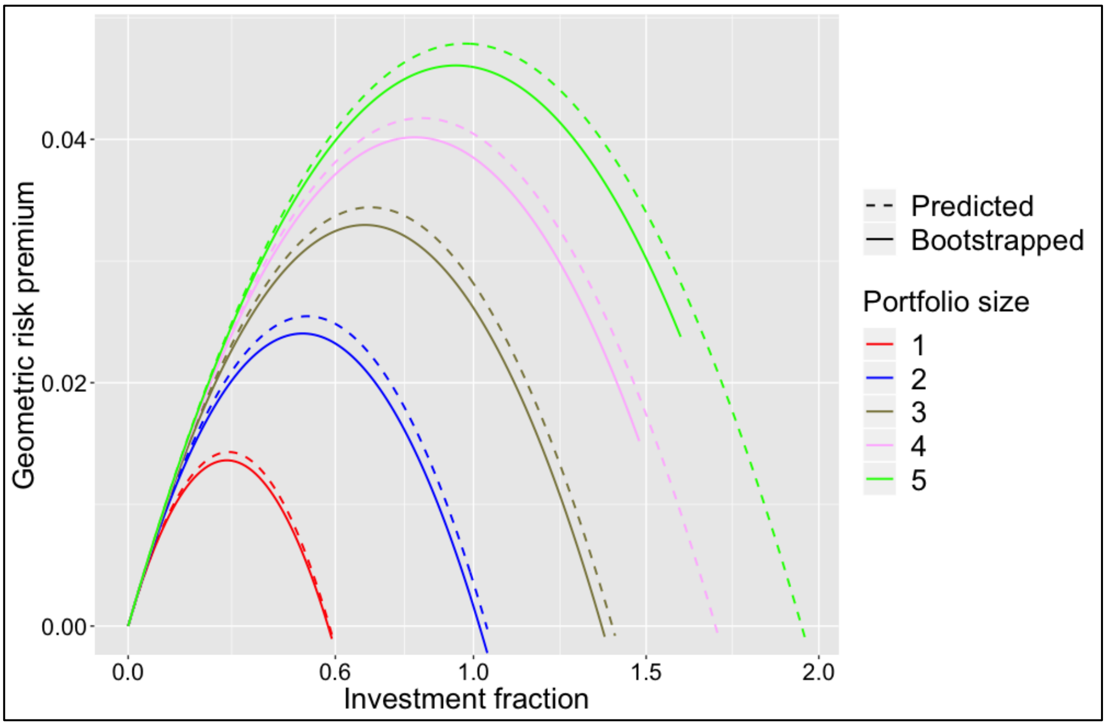

Figure 17. Geometric risk premium for poorly diversified portfolios between Jan-1973 and Jun-2018.

We can observe that at this poor diversification, each of the portfolio sizes will have lower geometric risk premium at investment fraction one (at 100% stock allocation) compared to some lower stock allocation. In other words, taking less risk, by investing a fraction of the portfolio in riskless rate, will yield a greater geometric risk premium.

The interpretation of figures from Figure 13 to Figure 17 and the Table 3 seem to support our *hypothesis 1*:

Hypothesis 1: Taking less risk, by investing a fraction of the portfolio in riskless rate, can increase the geometric risk premium for poorly diversified portfolios. This result, and the geometric risk premium in general, is explained by the fractional Kelly criterion.

5.1.2 Predicting required number of stocks for positive risk premium

Next, we show that our prediction for required number of stocks for positive risk premium works. To test the accuracy of approximate equation (127), we first solve f:

$$f = \frac{2\sqrt{Var_{BM}}SR_{\infty}^{BM}}{Var_{BM}+Ivar_{n=f=1}/n_P}. \tag{143}$$

Equation (143) should now give the investment fraction where geometric risk premium goes to zero.

Accuracy of the equation is tested by the data used in Figure 17. We use the data for bootstrapped risk premium which is measured for each investment fraction value using step size 0.01 until risk premium becomes negative. For 4 and 5 stock portfolios, the bootstrapped data does not extend to negative risk premium. Portfolio sizes 1, 2, and 3 stocks can be used to evaluate the accuracy of predicted investment fraction f when risk premium is zero. Predicted value is given by equation (143) and is compared against the bootstrapped investment fraction when risk premium first gets negative.

Result of the test is given in Table 4. We can see that the prediction is very accurate at portfolio size of one stock and remains relatively accurate for portfolio sizes two and three stocks. Similarly, as we saw with risk premium predictions, there is some deviation between predicted and bootstrapped value as investment fraction increases. Based on this test, we conclude that equation (127) accurately predicts the required number of stocks to achieve a positive risk premium.

Table 4. Predicted versus bootstrapped investment fraction for zero risk premium.

	Number of stocks		
	1	2	3
Predicted f	0.58	1.04	1.40
Bootstrapped f	0.58	1.02	1.38

5.2 Diversification is a negative price lunch

5.2.1 Diversification premium difference to benchmark in different time periods

In 5.1 we saw how the parabolic curves for instantaneous geometric risk premiums for portfolios of different sizes all were below the parabola of the benchmark portfolio. We now turn our focus in the diversification premium difference to benchmark which is the portfolio risk premium minus the benchmark portfolio risk premium.

We will show predicted and bootstrapped diversification premium differences for various portfolio sizes in different time periods. The predicted metric, for portfolios greater than one stock, is based on equation (40) and, after deciding the investment fraction and portfolio size of interest, only needs two inputs, the idiosyncratic variance of a single stock portfolio and number of stocks in the benchmark portfolio. For single stock portfolios, we use equation (41) which only requires the idiosyncratic variance of a single stock portfolio in addition to investment fraction as an input. Idiosyncratic variance of a single stock portfolio is acquired as an output from regression specified by equation (47).

We will show the results for the whole time period extending from July 1926 to June 2018 and for the three subperiods shown in Figure 8. In addition, to demonstrate the effect of fat tails, we will show the results for two additional sets of data from the third time period, January 1973 to June 2018, where we cut the tails of the distribution.

For each of the diversification premium difference to benchmark figures, we will show the portfolio sizes up to 500 or 1000 stocks, depending on the minimum monthly number of stocks for the period. 1000 will be used as the maximum whenever the monthly minimum exceeds that amount. Similarly, as with the figures in the section 5.1, we will plot the lines until the investment fraction is high enough for the first bootstrapped portfolio to exceed 100% loss.

Figure 18 shows the diversification premium difference to benchmark for the whole 92-year time period. Bootstrapping without replacement is implemented by creating 25 000 random portfolios per portfolio size each month. Given that the time period consists of three distinct stock populations as shown by Figure 8 and Figure 9, the predicted value is surprisingly close to bootstrapped realized values.

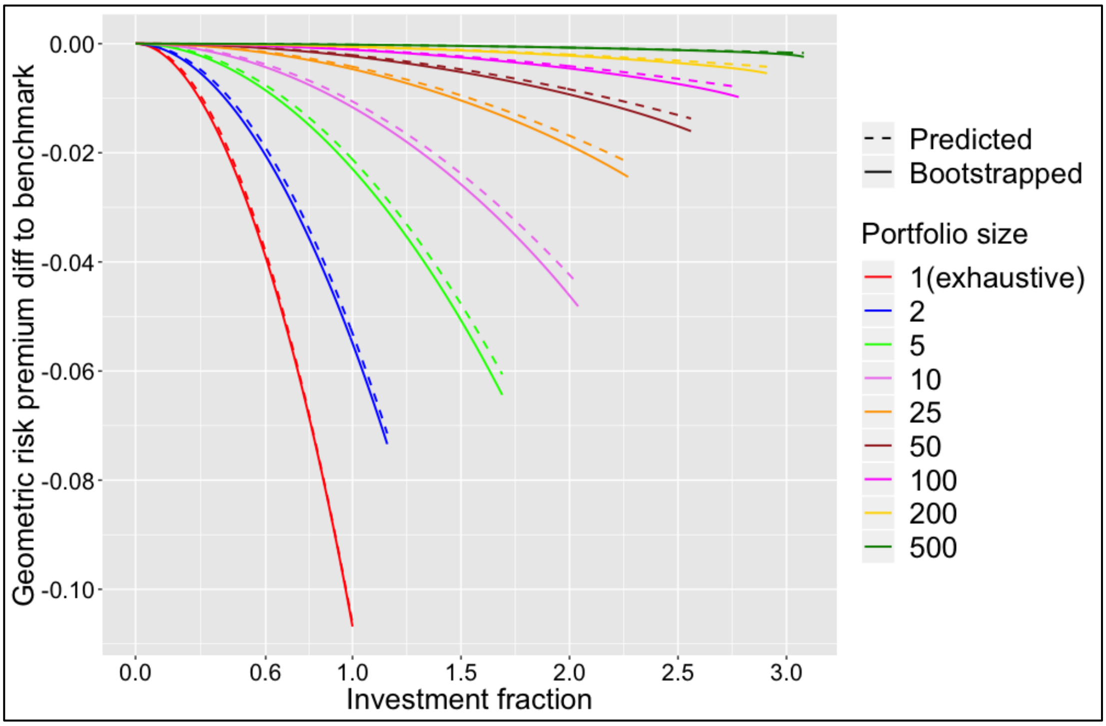

Figure 18. Diversification premium difference to benchmark between Jul-1926 and Jun-2018.

The single stock portfolio prediction, for which the realized value is based on exhaustive data rather than bootstrapped, is particularly accurate. Also, the larger portfolio sizes seem very accurate although small differences are difficult to spot from figure.

In Figure 19 we show the diversification premium difference to benchmark for the NYSE stock era. 75 000 random portfolios are created per portfolio size each month.

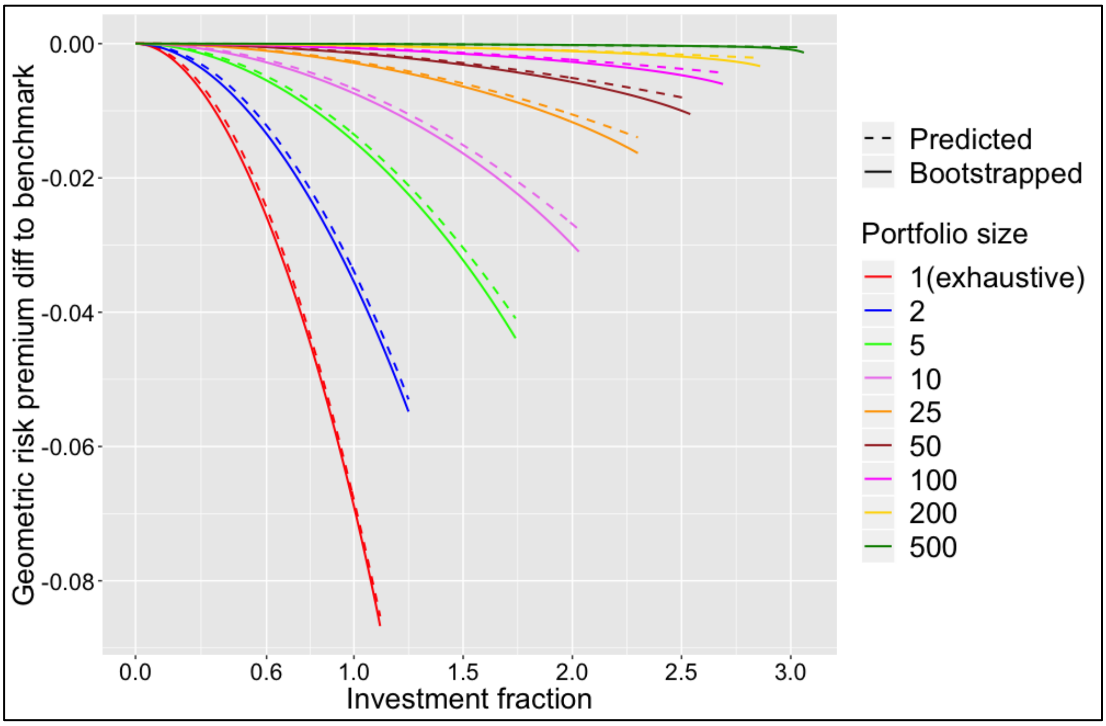

Figure 19. Diversification premium difference to benchmark between Jul-1926 and Jul-1962.

The era contains a tail event, the Great Depression, and on average large firm sizes compared to forthcoming time periods. Despite the massive tail event, the predicted values still appear to explain the realized diversification premium difference to benchmark rather accurately. On the other hand, when the tail event occurs on the market level, stocks on average "ride on top of the market wave" and therefore may not be that susceptible to the tail event. Because of this "market neutral" nature of the diversification premium, it can be hypothesized to be resilient against market volatility implying diversification may actually be highly beneficial regardless the market moves. This would be against the conventional wisdom that diversification is beneficial until it is most needed during a severe bear market. The reference for the value of diversification should be relative not absolute. Using relative reference, as in our case, is about comparing the outcomes between perfect and less than perfect diversification. Using absolute reference, as in the case of the conventional wisdom, is about assessing the outcome of perfect diversification when the market falls. If diversification indeed could prevent bear markets, would there be such thing as equity risk or equity risk premium?

In Figure 20 we can see the diversification premium difference to benchmark for the NYSE Amex stock era. 100 000 random portfolios are created per portfolio size each month during this relatively short period of time.

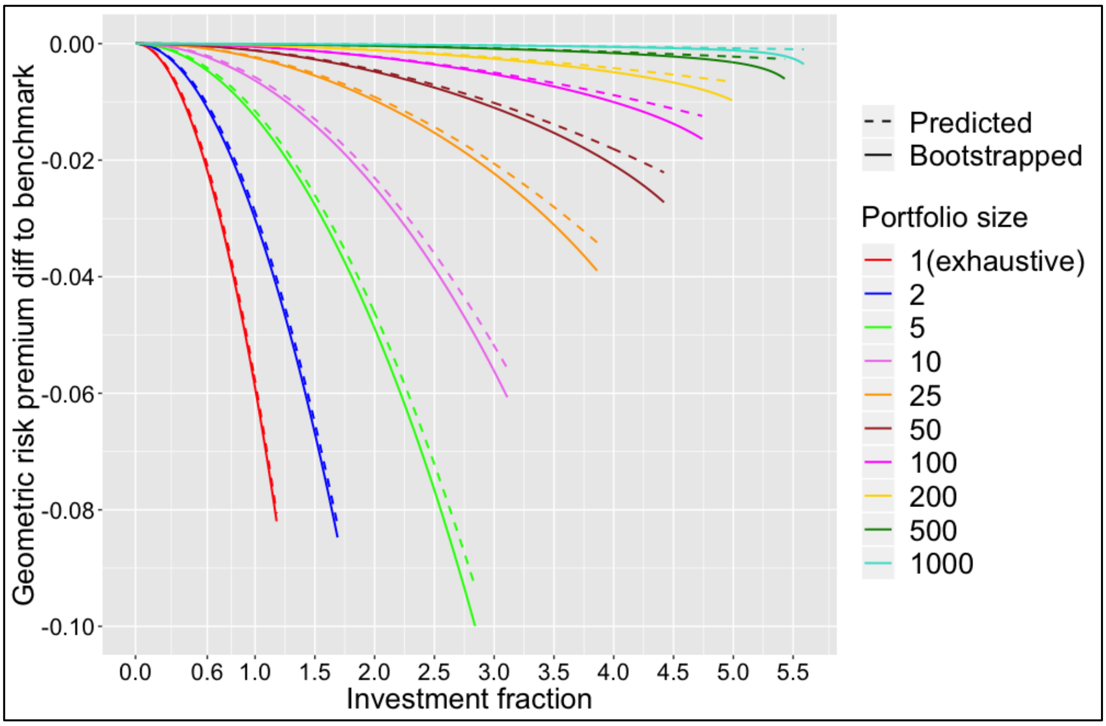

Figure 20. Diversification premium difference to benchmark between Aug-1962 and Dec-1972.

Our main interest is the modern era of U.S. stock returns, the 45.5-year period extending from the beginning of the 1973 until June 2018. This period is characterized and dominated by microcap stocks which on average are very volatile and exhibit very large idiosyncratic variance. Another characteristic of the microcaps is fat tailed and positively skewed monthly excess returns. Returns between portfolios loaded with volatile, skewed and fat tailed microcap stocks can differ wildly, which should put our diversification metrics to real test. Fortunately, there are plenty of stocks and monthly excess returns, 2 987 476 in total. The large number of data points ensures relatively accurate predictions. However, regardless the huge data, we still see from Figure 21 that the realized metrics for the intermediate portfolio sizes don’t exactly align with the predictions.

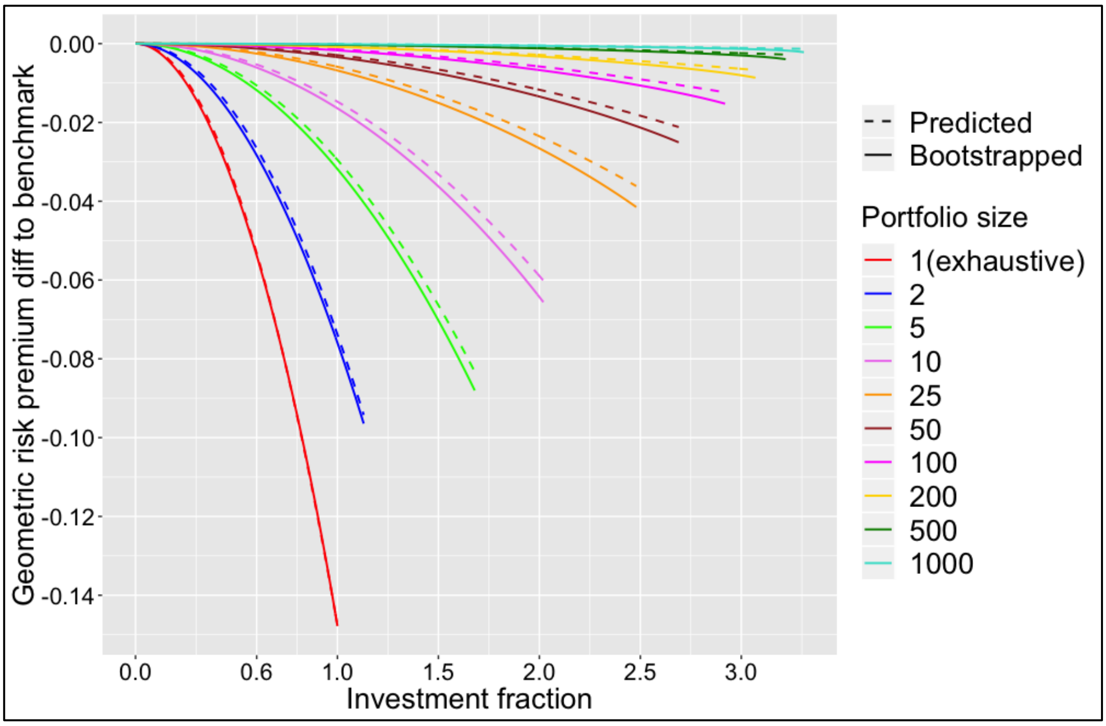

Figure 21. Diversification premium difference to benchmark between Jan-1973 and Jun-2018.

Table 5 summarizes the diversification premium difference to benchmark metrics at investment fraction one for different time periods. Predicted and realized values are shown for different portfolio sizes. Predicted and realized values are typically almost identical for portfolio size one. Predictions are fairly accurate also for larger portfolio sizes, but we notice a trend that the absolute value of the realized diversification premium difference to benchmark is typically slightly greater than the absolute value of the predicted metric. We suspect this is mainly due to fat tailed monthly excess returns and will further investigate this hypothesis in section 5.2.2.

Table 5. Summary of predicted vs. realized diversification premium differences to benchmarks.

Diversification premium difference to benchmark [pp]	Jul-1926 to Jun-2018	Jul-1926 to Jul-1962	Aug-1962 to Dec-1972	Jan-1973 to Jun-2018
Predicted $\Delta DP_{n=1}^{BM}$	-10.626	-6.799	-5.790	-14.765
Realized $\Delta DP_{n=1}^{BM}$	-10.682	-6.890	-5.904	-14.783
Predicted $\Delta DP_{n=10}^{BM}$	-1.059	-0.672	-0.576	-1.474
Realized $\Delta DP_{n=10}^{BM}$	-1.166	-0.741	-0.634	-1.642
Predicted $\Delta DP_{n=25}^{BM}$	-0.422	-0.264	-0.229	-0.588
Realized $\Delta DP_{n=25}^{BM}$	-0.467	-0.288	-0.246	-0.682
Predicted $\Delta DP_{n=100}^{BM}$	-0.103	-0.060	-0.055	-0.145
Realized $\Delta DP_{n=100}^{BM}$	-0.113	-0.070	-0.055	-0.171

By observing the diversification premium difference to benchmark metrics in the latest time period, from January 1973 to June 2018, we can see the head start that the investor has given to fully diversified benchmark investor by being content with low diversification. Ten-stock and twenty-five-stock portfolios are often considered as sufficiently diversified. Based on Table 5, however, in the absence of stock picking skill, a ten-stock portfolio on average has lost more than 1.6 percentage points in risk premium compared to fully diversified benchmark portfolio. A twenty-five-stock portfolio has lost close to 0.7 percentage points. A difference of 1.6 percentage points is comparable to the cost of an expensive active mutual fund.

In addition to our empirical results, we can apply the concept of diversification premium difference to benchmark to empirical results in Tidmore et al. (2019) study. Tidmore et al., as described in section 2.2, measure the difference in expected (geometric) excess return between a portfolio of selected size and the fully diversified benchmark portfolio and call this difference as average expected excess return. This is the same metric as our diversification premium difference to benchmark. Tidmore et al. find that for a single stock portfolio the expected excess return is -9.9 percentage points. It is not clear whether this is continuously or annually compounded rate, but it does not make a practical difference. Alpha in our regression equation (46) estimates single stock portfolio's geometric mean return difference to geometric mean return of a fully diversified benchmark portfolio. The -9.9 percentage points for a single stock portfolio in the Tidmore et al. study is the empirically measured return difference to benchmark and can be considered as approximately equal to alpha in equation (46). We therefore can consider the -9.9 percentage points to represent the alpha in our approximate alpha-based equation (54) for diversification premium difference to benchmark. Table 6 shows the relationship between our predicted (based on the -9.9 percentage points acquired for single stock portfolio) diversification premium difference to benchmark and empirical results from Tidmore et al. study. Equation (54) accurately explains the empirical Tidmore et al. expected excess return for different portfolio sizes.

Table 6. Tidmore et al. results explained by approximate diversification premium difference to benchmark equation.

	Diversification premium difference to benchmark [pp]	
Portfolio size	Prediction based on eq. (54)	Tidmore et al. result
5	-2.0	-2.0
10	-1.0	-1.0
15	-0.7	-0.7
30	-0.3	-0.4
50	-0.2	-0.2
100	-0.1	-0.1
200	0.0	-0.1
500	0.0	0.0

5.2.2 The effect of fat tails

In sections 5.1 we hypothesized that the difference between the predicted and realized values would be due to low rebalancing frequency and the fat tails of the excess return data. In case of diversification premium difference to benchmark the rebalancing frequency possibly partially cancels out as we subtract the benchmark portfolio risk premium from the investment portfolio risk premium as both portfolios suffer from less than optimal rebalancing frequency. However, we suspect that the effect of the fat tailed excess returns does play a role in the difference between the predicted and realized values.

Our observation that the predicted diversification premium difference to benchmark systematically underestimates the empirically realized value is consistent with Taleb (2020, pp. 152–153). Taleb notes that the risk of a fat-tailed return distribution is underestimated (implying the required level of diversification is underestimated) by frameworks assuming thin-tailed return distributions. Our framework does not assume thin-tailed return distribution, but it does assume theoretical infinite rebalancing frequency. We hypothesize that the fat-tailed return distribution emphasizes the impact from less than optimal rebalancing frequency as the tails of the return distribution have one month, instead of an infinitely short period, to deviate from the equal weighting of the portfolio.

To test the hypothesis that the difference is caused by the fat tails, we cut the tails of the distribution. We have two tests. The first is to cut the permille tails of the distribution. This means removing the 0.1st percentile (the worst) and the 99.9th percentile (the best) monthly excess returns from the data. We do this for each month. The second test is similar except now we cut the percent tails meaning the 1st percentile and the 99th percentile. We use the data from January 1973 to June 2018 with an average of 5472 monthly stock excess returns. In the first test we remove on average 6 worst and 6 best monthly excess returns from each month's data. This is closer the 1.1 permille, but close enough. In the second test, we remove on average 55 worst and 55 best monthly excess returns from each month's data. Both tests utilize 50 000 randomly created portfolios each month.

Figure 22 shows the result from cutting the permille tails. Comparing to Figure 21, we can see that the match is clearly better now after cutting the extreme tails. Another striking observation from the data is that the realized geometric risk premium for the benchmark portfolio now, after cutting the permille tails, is 6.24% while it was 7.74% before the tails were cut. Removing just one permille of the worst and best excess returns each month lowers the geometric risk premium by 1.5 percentage points. This illustrates the enormous effect the positive tail has to the mean. It also shows how diversification is about including those rare outsized winners rather than excluding the extreme losers.

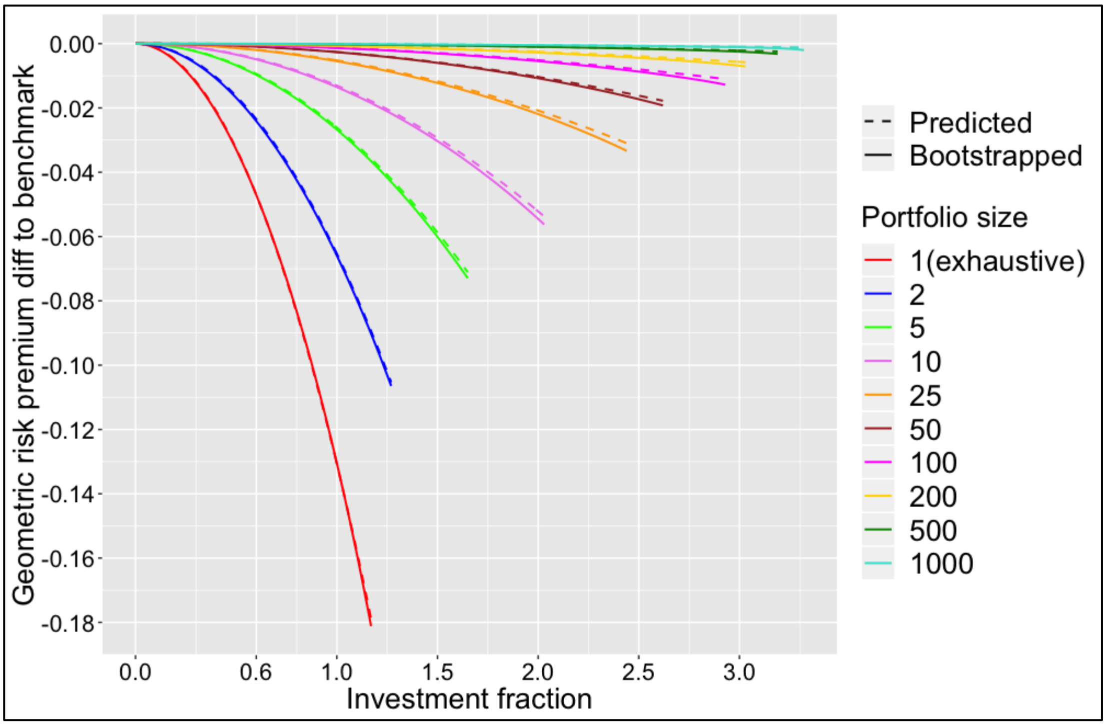

Figure 22. Diversification premium difference to benchmark between Jan-1973 and Jun-2018 for data with permille tails cut.

Comparing Figure 23 to Figure 22 we see that cutting the percent tails leads to even better match. Judging by the eye, the match is extremely good. The realized geometric risk premium for the benchmark portfolio now is 3.08%, 4.66 percentage points less than with the tails. Of course, it is not very likely that one ends up avoiding the percent tails of the excess return distribution every month, but nevertheless this shows how the fat positive tail of the distribution dominates the mean geometric excess return.

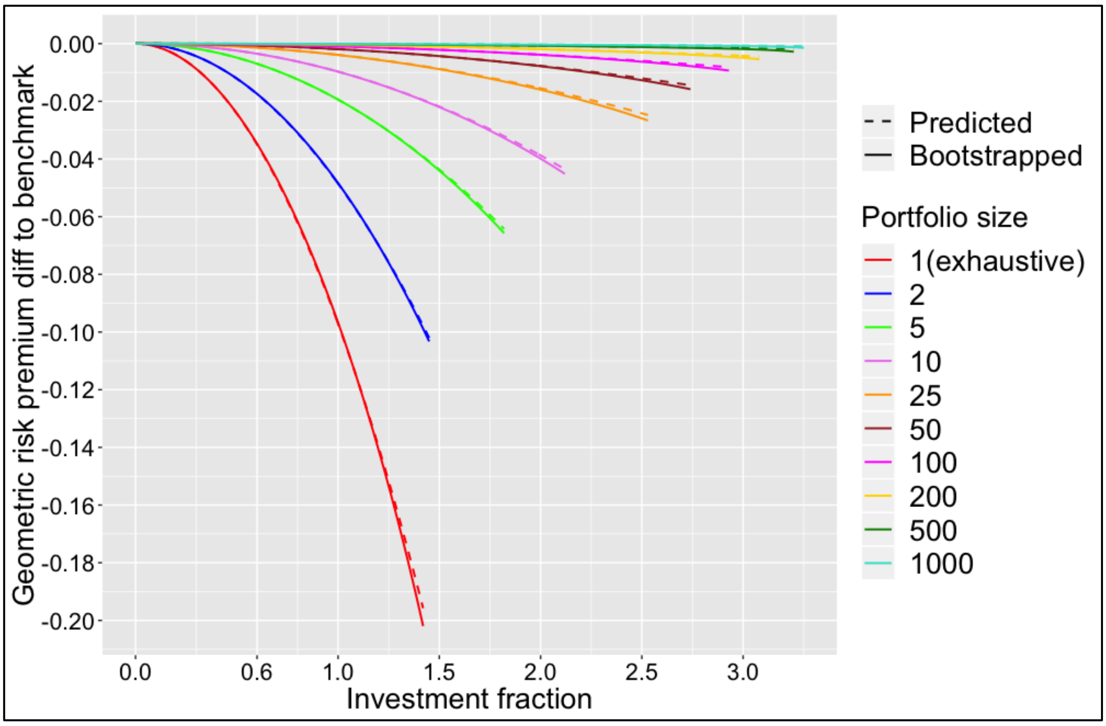

Figure 23. Diversification premium difference to benchmark between Jan-1973 and Jun-2018 for data with percent tails cut.

Table 7 summarizes the effect of fat tails. Removing the permille tails helps the realized diversification premium difference to benchmark to converge more quickly and the realized values are close to predicted values. When percent tails are removed, we can see that the predictions become extremely accurate. Prediction error for percent tails cut data is ranging from 0.000 to 0.037 percentage points (pps), which equals a range of 0 to 3.7 basis points (bps). One basis point (bp) is one hundredth part of a percentage point (pp). We also notice how cutting the tails decreases the diversification premium to benchmark. In other words, the fat, positively skewed tails of the excess return distribution emphasize the importance of diversification.

Table 7. The effect of cutting the tails of the monthly excess return distribution.

Diversification premium difference to benchmark [pp]	Jan-1973 to Jun-2018 Full data	Jan-1973 to Jun-2018 Permille tails cut	Jan-1973 to Jun-2018 Percent tails cut
Predicted $\Delta DP_{n=1}^{BM}$	-14.765	-13.067	-9.710
Realized $\Delta DP_{n=1}^{BM}$	-14.783	-13.046	-9.673
Predicted $\Delta DP_{n=10}^{BM}$	-1.474	-1.304	-0.969
Realized $\Delta DP_{n=10}^{BM}$	-1.642	-1.352	-0.972
Predicted $\Delta DP_{n=25}^{BM}$	-0.588	-0.520	-0.387
Realized $\Delta DP_{n=25}^{BM}$	-0.682	-0.547	-0.393
Predicted $\Delta DP_{n=100}^{BM}$	-0.145	-0.128	-0.095
Realized $\Delta DP_{n=100}^{BM}$	-0.171	-0.136	-0.095

5.2.3 Statistical evidence

To find some statistical evidence that diversification is a negative price lunch, implying diversification premium exists, we turn to regression analysis. We use the bootstrapped (except for single stock portfolio, for which we use exhaustive data) data from January 1973 to Jun 2018. Data contains 546 months and 50 000 bootstrapped portfolios per month. We use simple OLS regression where the left-hand side is logarithmic excess return of a portfolio of selected size and the explanatory variable is the logarithmic excess return of a benchmark portfolio. Intercept captures the alpha which is the mean difference in logarithmic excess returns between the portfolio of interest and benchmark portfolio. In other words, the alpha is the diversification premium difference to benchmark. For exhaustive single stock portfolio data, which has varying number of stocks per month, we normalize the logarithmic returns based on equation (44) to ensure equal weight for each month.

The null hypothesis is that alpha is zero. Alternative hypothesis is that alpha is different from zero. We summarize the regression results in Table 8. The null hypothesis is convincingly rejected. Statistical significance is extraordinarily high, which is attributable to our very large sample size and the fact that we are testing a mathematical inevitability. At lowest, the t-statistic for the alpha is 11.61. This occurs when portfolio size is 1000 meaning that even this large portfolio still has a statistically highly significant diversification premium difference to benchmark. Economically the alpha for a 1000-stock portfolio is hardly significant at 1.3 basis points. We can see that the beta is very close to one at small portfolio sizes and approaches exactly one the larger the portfolio size gets. R-squared tells us that more than 90% of the portfolio variance is explained by the benchmark only after portfolio size is 100 or higher. This is one indication of the prevalence of the costly idiosyncratic variance at low portfolio sizes.

Table 8. Regression results for portfolio sizes ranging from one to thousand stocks.

Portfolio size	1(exhaustive)	2	5	10	25
Alpha	-1.474E-01	-7.528E-02	-3.144E-02	-1.619E-02	-6.716E-03
	(1.094E-03)	(2.551E-04)	(1.660E-04)	(1.197E-04)	(7.687E-05)
Beta	0.9939	0.9917	0.9951	0.9970	0.9986
	(1.517E-03)	(3.590E-04)	(2.337E-04)	(1.684E-04)	(1.082E-04)
R^2	0.1257	0.2184	0.3991	0.5621	0.7573
N	2987476	27300000	27300000	27300000	27300000
Portfolio size	50	100	200	500	1000
Alpha	-3.387E-03	-1.681E-03	-8.055E-04	-2.873E-04	-1.301E-04
	(5.464E-05)	(3.865E-05)	(2.714E-05)	(1.670E-05)	(1.120E-05)
Beta	0.9994	0.9996	0.9998	0.9999	1.0000
	(7.690E-05)	(5.439E-05)	(3.820E-05)	(2.350E-05)	(1.577E-05)
R^2	0.8608	0.9252	0.9617	0.9851	0.9933
N	27300000	27300000	27300000	27300000	27300000

Our conclusion from the Table 8 is that alpha is always convincingly negative implying diversification premium, and therefore diversification difference to benchmark, exists and is statistically highly significant throughout the whole spectrum of tested portfolio sizes. We therefore consider *hypothesis 2* as confirmed.

Hypothesis 2: Diversification is a negative price lunch implying that diversification premium exists.

5.2.4 The difference in lunch pricing illustrated

We use an example from the data demonstrating what the negative price lunch means. We have 20 000 bootstrapped portfolios created for each month during the 45.5 years period from January 1973 to June 2018. We link the monthly portfolio excess returns to form 20 000 individual, randomly created, excess return histories. In Figure 24 we show the distributions for the annualized randomly created 45.5-year return histories for three portfolio sizes. Number of stocks in the portfolio, n_P, is 1, 5 or 100 stocks. Initially the standard deviations in the figure may appear as low, but we need to remember that annual standard deviations are mitigated by a factor of square root of 45.5. Vertical dashed lines mark the mean values. We have the annualized geometric (logarithmic) excess return distribution and the annualized arithmetic excess return distribution on the left and right, respectively.

Using the lunch analogy, we can put it like this: on the left in Figure 24 we see a continuous-time world investor and, on the right, a one period world investor at their lunches, respectively. In the upmost figures we witness a fasting meaning no lunch at all, in the middle an order from the appetizer menu is enjoyed and in the lowest figures a lunch buffet is consumed. The figures on the right visualize how the diversification is a free lunch in the one period world as the expected arithmetic excess return is not compromised when lunch is consumed but the risk decreases as the lunch portion increases. Unchanged expected reward equals zero cost. The figures on the left vividly demonstrate what the negative price lunch in the continuous-time world is all about. Like in one period world, the risk decreases as the lunch portion is increased, but now simultaneously the expected geometric excess return increases as a greater portion of lunch is consumed. In a continuous-time world, an investor lowers the risk and increases the expected reward simultaneously by accepting more diversification. Increased expected reward equals negative price.

Being advised by Markowitz is good. You will be encouraged to enjoy a decent size free lunch, which will save you from expected starvation. But internalizing some of the work by Shannon, Kelly and Thorp is even better. It may convince you to enjoy a larger portion as you learn you will be paying a negative expected price.

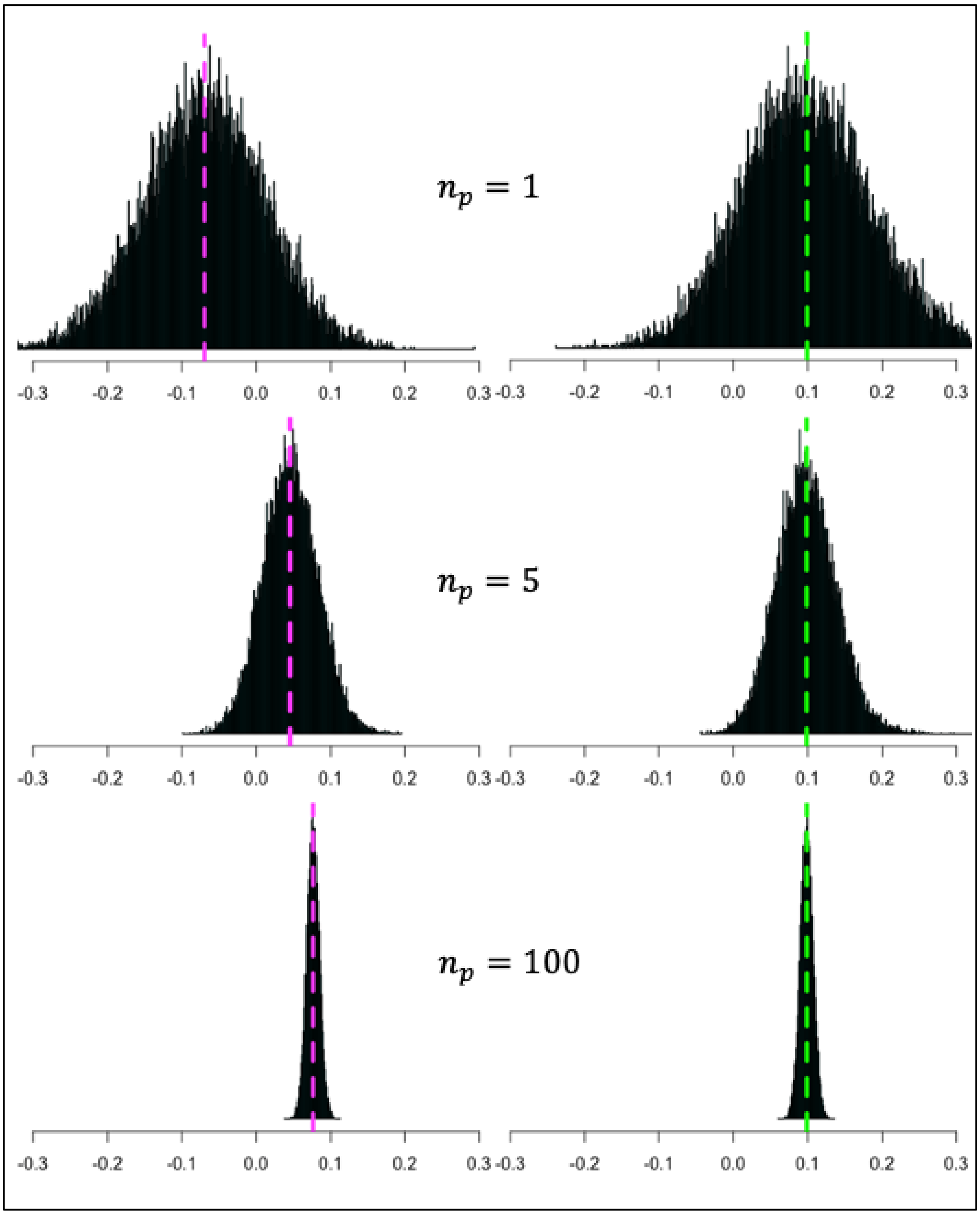

Figure 24. Negative price lunch on the left as opposed to free lunch on the right.

5.3 Opportunity cost of foregone diversification

5.3.1 One half of portfolio's idiosyncratic variance

Our *hypothesis 3* states that opportunity cost of foregone diversification is approximated by the one half of portfolio's idiosyncratic variance, which is equal to the magnitude of the diversification premium difference to benchmark. We next show some statistics supporting the hypothesis.

First, however, we can visually evaluate if the hypothesis seems to hold. The series of figures from Figure 18 to Figure 23 show us the predicted and realized magnitudes of diversification premium difference to benchmark for different portfolio sizes. We know from equation (42) that the magnitude of diversification premium difference to benchmark is approximated by one half of portfolio's idiosyncratic variance when benchmark is broadly diversified. We can see from the figures that realized opportunity cost of foregone diversification typically is well approximated by the prediction.

To assess the hypothesis using some statistics, we use the prediction error and prediction error to prediction percentage ratio. By showing these statistics, we can evaluate both the absolute and the relative error as a function of portfolio size at investment fraction one. Prediction error is defined as predicted diversification premium difference to benchmark minus the corresponding realized metric.

Table 9 presents the metrics for the full data and the data with percent tails cut spanning from January 1973 to June 2018. For the full data, prediction errors are always below twenty percent and the highest at intermediate portfolio sizes. Cutting the fat tails has a large effect. Error is typically less than one percent and less than eight percent at highest.

Table 9. Prediction errors for diversification premium difference to benchmark.

	Jan-1973 to Jun-2018, full data		Jan-1973 to Jun-2018, percent tails cut	
Portfolio size	Prediction error	Prediction error to prediction ratio %	Prediction error	Prediction error to prediction ratio %
1(exhaustive)	1.73E-04	0.12	-3.69E-04	-0.38
2	2.12E-03	2.88	-1.77E-04	-0.37
5	2.33E-03	7.88	-9.81E-06	-0.05
10	1.68E-03	11.41	3.20E-05	0.33
25	9.43E-04	16.04	6.61E-05	1.71
50	5.05E-04	17.25	-1.89E-05	-0.98
100	2.60E-04	17.95	-3.78E-06	-0.40
200	1.12E-04	15.71	-3.88E-06	-0.83
500	2.30E-05	8.57	6.88E-06	3.91
1000	1.12E-05	9.28	6.20E-06	7.85

Based on visual evaluation of the figures and by evaluating the prediction errors, we find supporting evidence for *hypothesis 3*.

Hypothesis 3: One half of the portfolio's idiosyncratic variance closely approximates the magnitude of diversification premium difference to benchmark, the opportunity cost of foregone diversification.

5.3.2 A function of squared investment fraction

Next, we assess the effect of investment fraction to the diversification premium difference to benchmark. Based on *hypothesis 4*, we expect the opportunity cost of foregone diversification to be a function of squared investment fraction.

Visually, based on figures ranging from Figure 18 to Figure 23, we can see that the predicted values match well with their realized counterparts.

In Table 10 we have the relative prediction errors for the full data and percent tails cut data as a function of three investment fractions: 0.6, 1.0 and 1.5. The data is again from the period from January 1973 to June 2018. NA in the table means no data available. We observe that the relative error is relatively constant as a function of investment fraction. We can see that without the fat tails the prediction error is very small. We interpret this as evidence that our realized diversification difference to benchmark indeed is in line with the corresponding predicted metric which in turn assumes scalability as a function of squared investment fraction.

Table 10. Prediction errors to prediction ratios [%] for different investment fractions.

	Jan-1973 to Jun-2018, full data			Jan-1973 to Jun-2018, percent tails cut		
	Prediction error to prediction ratio [%]			Prediction error to prediction ratio [%]		
Portfolio size	$f = 0.6$	$f = 1.0$	$f = 1.5$	$f = 0.6$	$f = 1.0$	$f = 1.5$
1(exhaustive)	1.10	0.12	NA	-0.69	-0.38	NA
2	6.66	2.88	NA	-0.33	-0.37	NA
5	11.31	7.88	5.85	-0.03	-0.05	0.82
10	14.44	11.41	9.30	0.48	0.33	0.90
25	19.66	16.04	13.67	2.53	1.71	1.82
50	20.37	17.25	15.23	-1.88	-0.98	-0.06
100	20.92	17.95	16.07	-0.80	-0.40	0.24
200	16.95	15.71	14.77	-1.90	-0.83	0.13
500	4.38	8.57	10.40	5.84	3.91	3.37
1000	5.49	9.28	10.89	13.15	7.85	5.60

Table 11 shows predicted and realized diversification premium differences to benchmarks as a function of investment fraction for selected portfolio sizes. The data is from January 1973 to June 2018. In parenthesis we have the diversification premium difference to benchmark ratio between the investment fraction of interest and investment fraction one. As we assume a squared relationship, we expect the ratio to be 0.36 at investment fraction 0.6 and 2.25 at investment fraction 1.5. We can see that for predicted values this is accurate. The small variation is due to rounding. For realized values the ratio is fairly close to expected value. For percent tails cut data this would be very accurate. Comparing the diversification premium difference to benchmark across investment fractions gives us some insight about the significance of the investment fraction defining the diversification effect. For example, at portfolio size of ten stocks the realized opportunity cost of foregone diversification is more than 1.6 percentage points. At investment fraction 0.6 the corresponding value is much more tolerable at about 0.6 percentage points. However, leveraging the portfolio to investment fraction 1.5 implies opportunity cost to rise to more than 3.6 percentage points. Especially when leveraging the portfolio, one is well advised to diversify broadly.

Table 11. Diversification premium differences to benchmarks for different investment fractions.

	Diversification premium difference to benchmark [pp]		
	$f = 0.6$ (/ $f = 1.0$)	$f = 1.0$	$f = 1.5$ (/ $f = 1.0$)
Predicted $\Delta DP_{n=1}^{BM}$	-5.316 (0.360)	-14.765	NA
Realized $\Delta DP_{n=1}^{BM}$	-5.374 (0.364)	-14.783	NA
Predicted $\Delta DP_{n=10}^{BM}$	-0.531 (0.360)	-1.474	-3.316 (2.250)
Realized $\Delta DP_{n=10}^{BM}$	-0.607 (0.370)	-1.642	-3.625 (2.208)
Predicted $\Delta DP_{n=25}^{BM}$	-0.212 (0.361)	-0.588	-1.323 (2.250)
Realized $\Delta DP_{n=25}^{BM}$	-0.253 (0.371)	-0.682	-1.504 (2.205)
Predicted $\Delta DP_{n=100}^{BM}$	-0.052 (0.359)	-0.145	-0.326 (2.248)
Realized $\Delta DP_{n=100}^{BM}$	-0.063 (0.368)	-0.171	-0.379 (2.216)

Based on visual evaluation of the figures and by evaluating the prediction errors as a function of investment fraction, we find convincing supportive evidence for *hypothesis 4*.

Hypothesis 4: Diversification premium difference to benchmark for a portfolio is a function of portfolio's squared investment fraction.

5.4 Risk premium ratio as a diversification effect measure

5.4.1 Utilizing realizable risk premium ratio to account for the risk

We defined realizable risk premium in section 3.3.7 and determined it as our short-term diversification effect measure in section 3.4.2. Any long-term period consists of series of short-term periods meaning short-term, at least to risk averse investor, is always meaningful regardless of the investment horizon. We will now show step by step how the realizable risk premium is constructed. In addition, we will show that our theoretical prediction relatively accurately corresponds to corresponding metric measured from empirical data.

Realizable risk premium captures the effect of risk and reward in a single number by scaling the geometric risk premium (the expectation) by the proportion the realized geometric risk premium is explained by its expectation. As shown in section 3.3.7, in the short term, this is equivalent to scaling the geometric risk premium by its SNR.

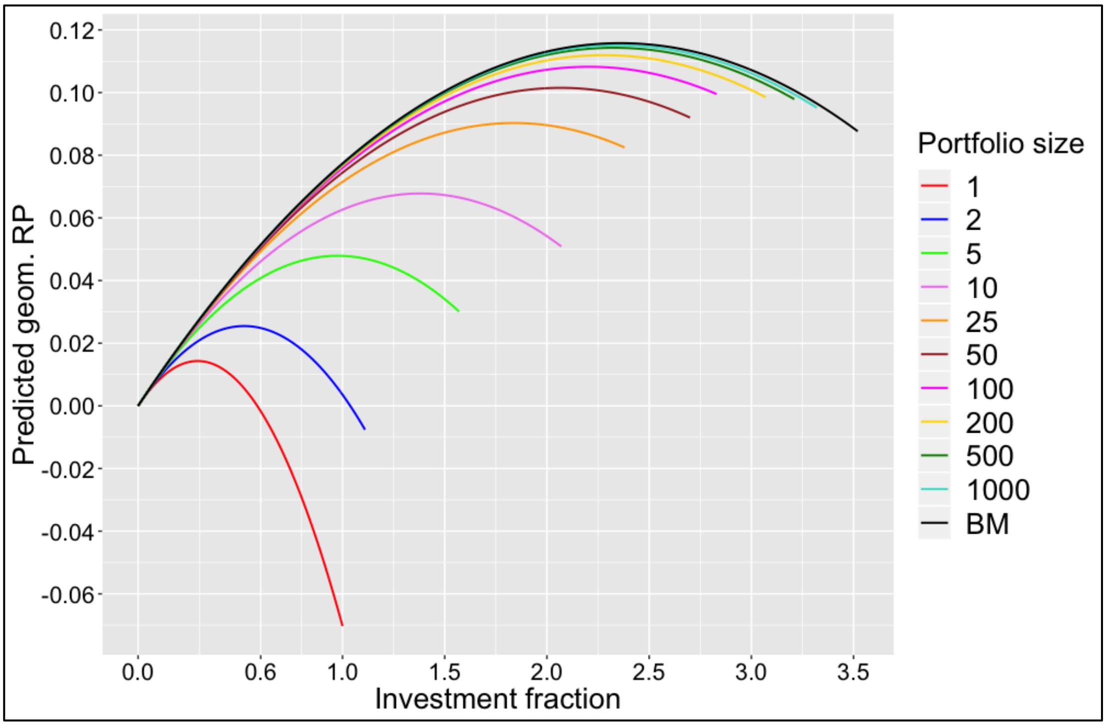

Figure 25. Predicted geometric risk premium.

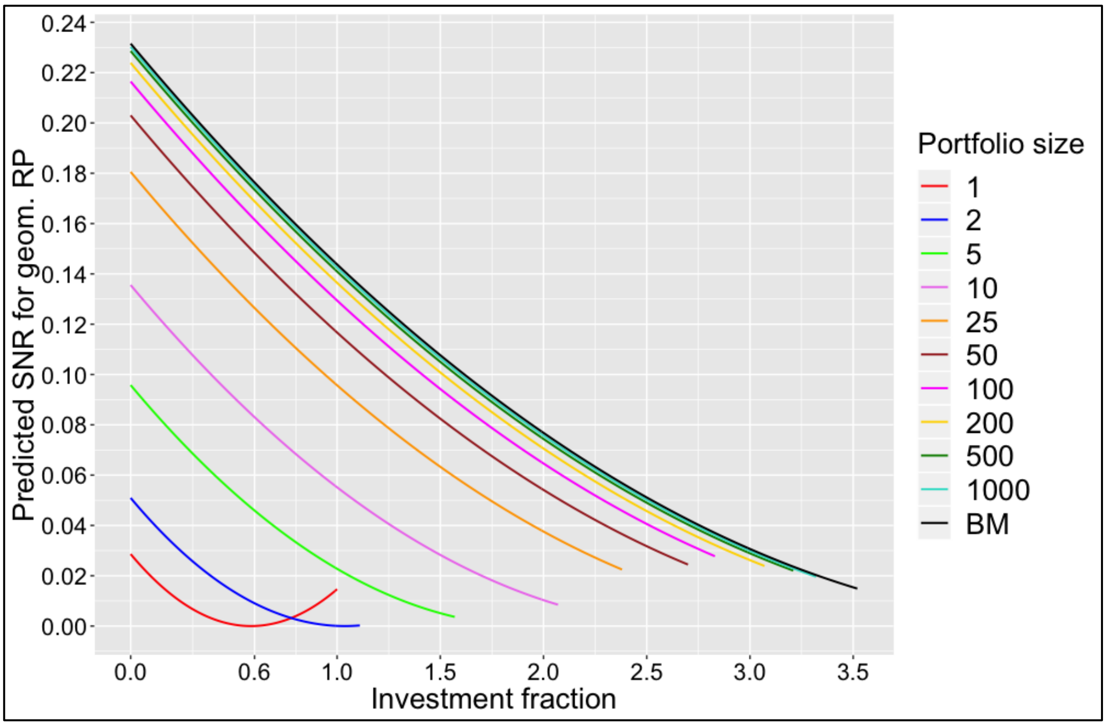

Figure 26. Predicted SNR for the geometric risk premium.

We show the two components, geometric risk premium (Figure 25) and its SNR (Figure 26), as a function of investment fraction and finally the realizable risk premium (Figure 27), which results from multiplying these two measures together.

Comparing figures Figure 25 and Figure 26, we observe that the SNR of the geometric risk premium for an investment portfolio is markedly lower than for SNR for the benchmark even for relatively large portfolio sizes, whereas for geometric risk premium the difference to benchmark is small for all but the smallest portfolio sizes. This means that the SNR of the geometric risk premium tends to dominate the short-term realizable risk premium metric. In the short-term, it is the risk (the noise power component in the SNR measure) that dominates.

Interestingly, we see from the one-stock curve in Figure 26 how the SNR starts to get higher shortly after the risk premium goes negative. This may seem odd at first, but it just means that the negative risk premium becomes the more certain to realize in the short-term the more we exceed the full Kelly point multiplied by two.

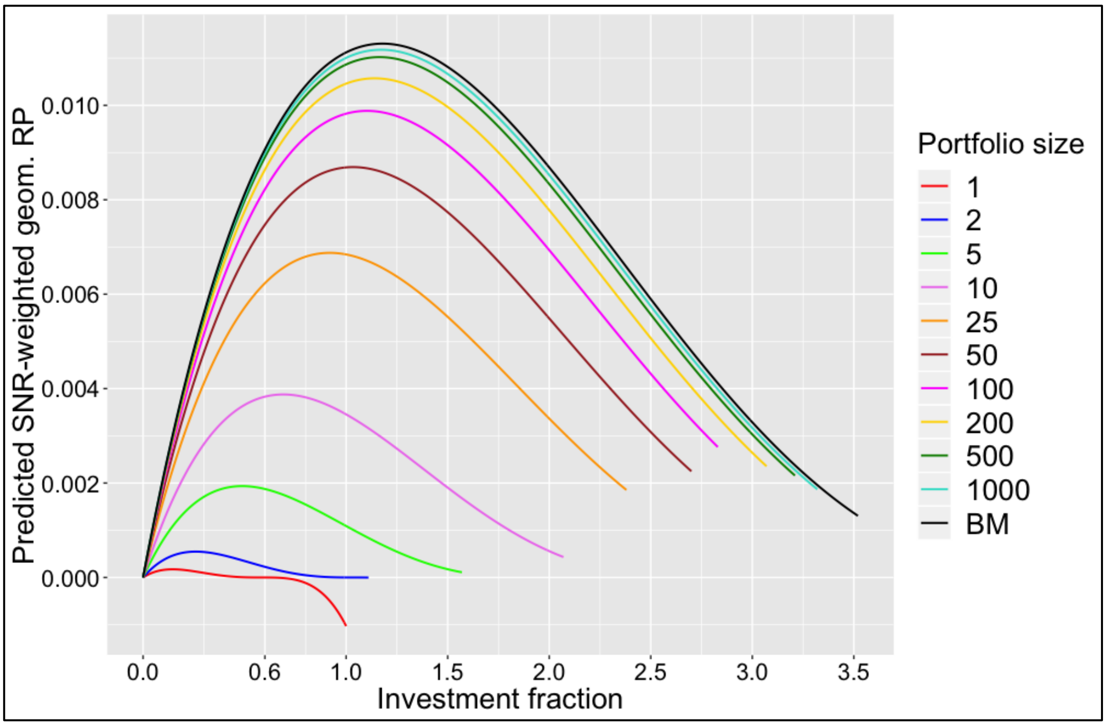

Figure 27. Predicted short-term realizable risk premium.

Figure 27 shows the resulting short-term realizable risk premium. The maximum is slightly higher than one percentage point (100 basis points) in annualized terms. Maximum occurs at half Kelly fraction, which corresponds to investment fraction slightly greater than one for the benchmark portfolio. For less than perfectly diversified portfolios, half Kelly fraction and maximum realizable risk premium occurs at lower investment fraction. The maximum realizable risk premium for ten-stock portfolio is below 40 basis points at investment fraction substantially below one, which shows how the risk dominates this short-term metric. Even for 200-stock portfolio, we still see a clearly lower value compared to benchmark.

Figure 28 shows the predicted and bootstrapped risk premium ratio for different portfolio sizes as a function of investment fraction. Risk premium ratio is the risk premium (shown in Figure 25) of a portfolio of selected size divided by the risk premium of a fully diversified benchmark portfolio. Notice, that the value can be negative when portfolio risk premium is negative. However, the metric only makes sense as long as the benchmark risk premium is positive. In other words, the metric makes sense only up to the investment fraction where benchmark risk premium turns negative, which theoretically occurs exactly at two times the benchmark full Kelly fraction. Predicted risk premium ratio is calculated based on equation (129).

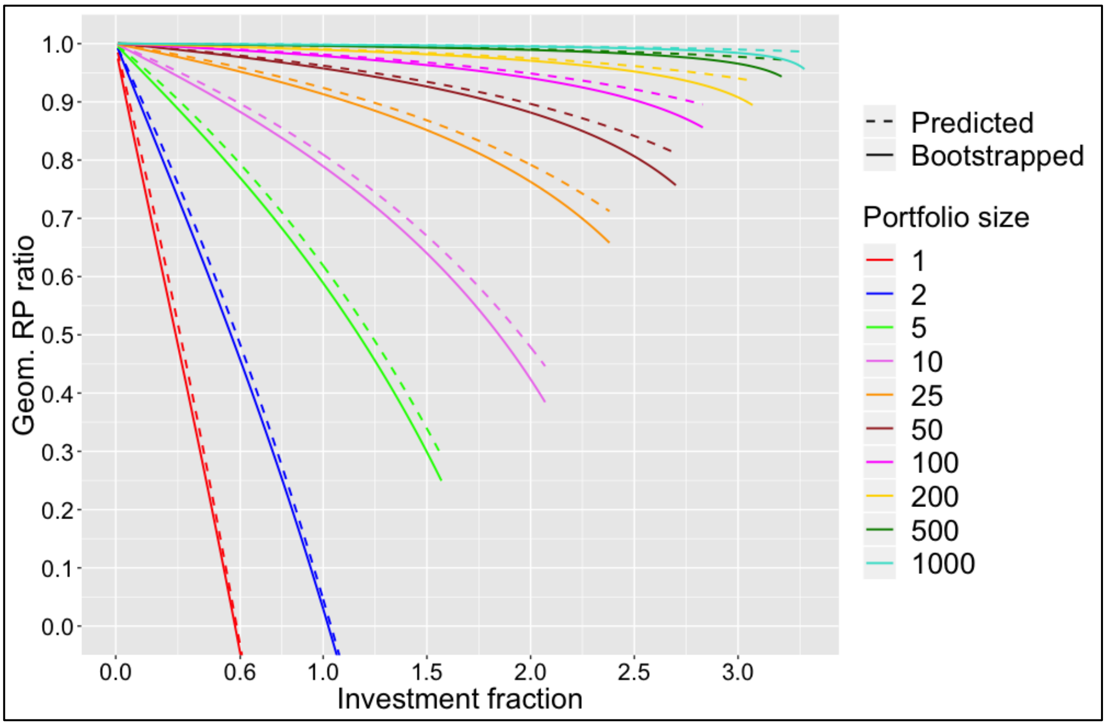

Figure 28. Predicted vs. bootstrapped risk premium ratio.

We can see that equation (129) predicts bootstrapped risk premium ratio very accurately. There is some deviation especially at very high investment fractions, which we attribute to fat tailed return distribution combined with less than infinite rebalancing frequency. Based on visual examination of the figure, close to 25 stocks are required to achieve 90% of the maximum risk premium ratio at investment fraction one.

Figure 29 shows the instantaneous realizable risk premium ratio. This is the same ratio as in Figure 28, except that now infinitely short-term realizable risk premium is used in place of infinitely long-term realizable risk premium, which is better known as risk premium. Predicted realizable risk premium is calculated based on equation (131). The metric only makes sense as long as the benchmark risk premium is positive.

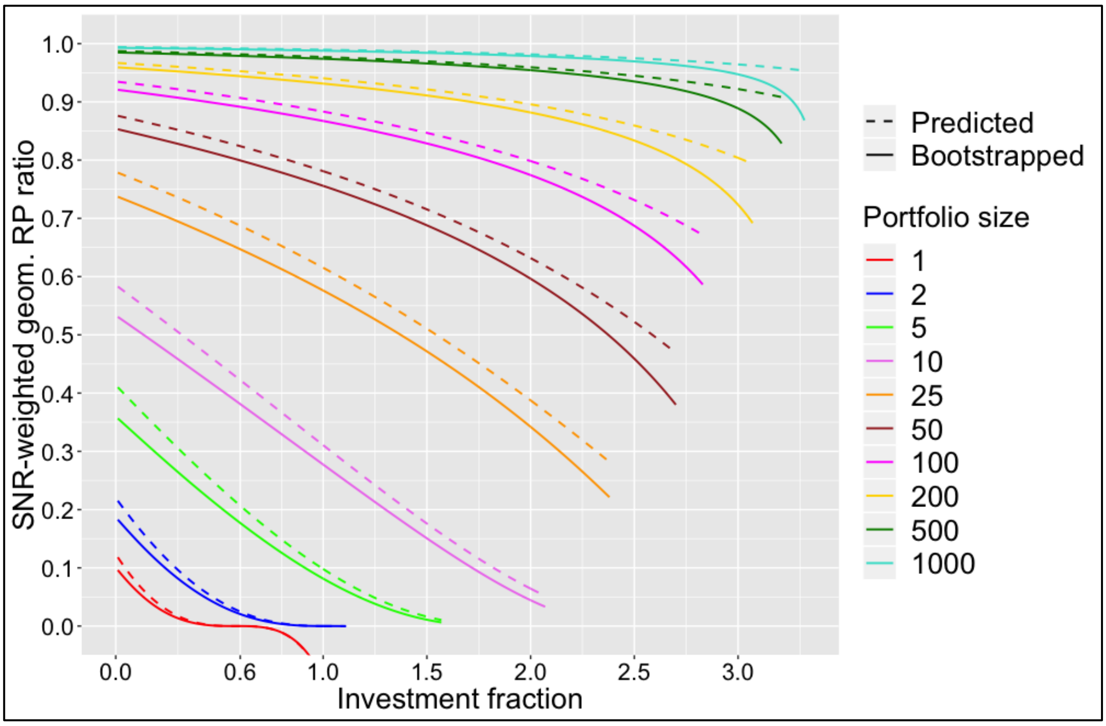

Figure 29. Predicted vs. bootstrapped short-term realizable risk premium ratio.

Equation (131) predicts realizable risk premium relatively accurately. Now both determinants of the realizable risk premium, the SNR of the risk premium and the risk premium, suffer from inaccuracy introduced by fat tails. Consequently, we see a larger deviation from predicted value compared to risk premium ratio, where only one ratio suffered from the inaccuracy. Based on visual approximation, more than 100 stocks are required to achieve 90% of the maximum realizable risk premium ratio at investment fraction one.

In Figure 29, the short-term is infinitely short-term (instantaneous), which not very practical. However, practical short-terms, such as one-month period, result to very similar realizable risk premium. Figure 30 shows the instantaneous (SNR weighted) and monthly (R-squared weighted) realizable risk premiums which are shown to be practically identical. We find that practical short-term periods ranging between infinitely short and several months all lead to very similar result. We conclude that SNR weighted risk premium can be used to model and to predict practical short-term realizable risk premium.

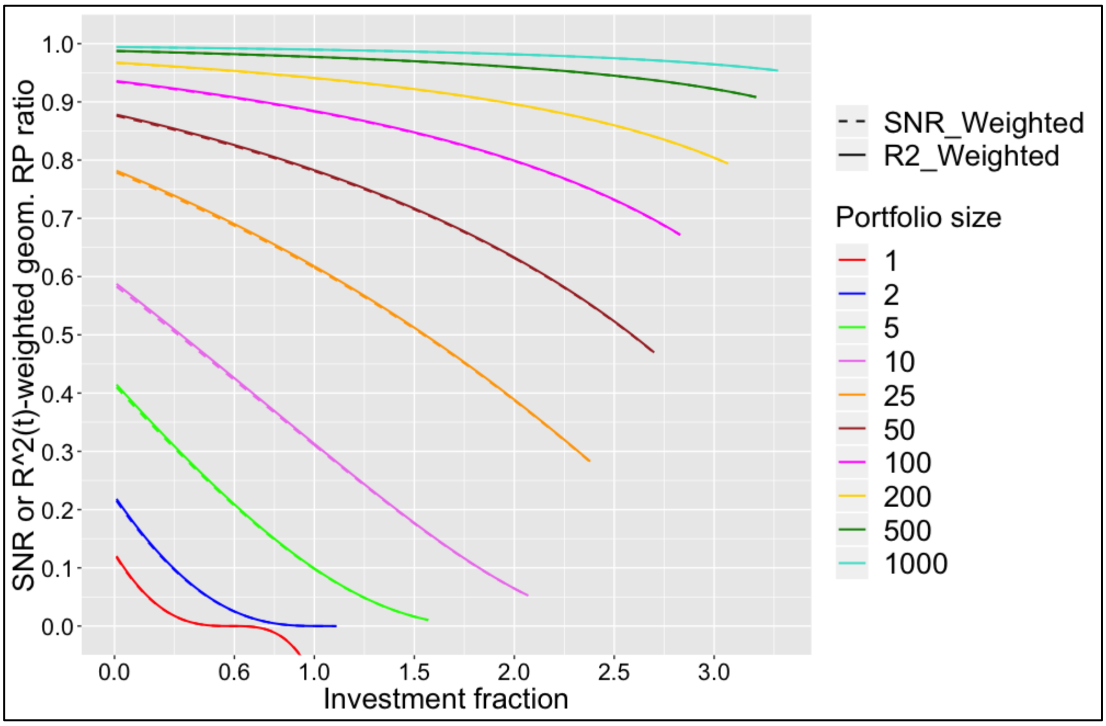

Figure 30. Instantaneous vs. monthly predicted realizable risk premium ratio.

5.4.2 Minimum investment horizon length for different portfolio sizes

We utilized the SNR of the geometric risk premium to determine realizable risk premium. Now we will show how the effect of time on SNR can be used to determine the minimum investment horizon as a function of portfolio size and investment fraction.

We assume time periods are uncorrelated implying SNR is directly proportional to investment time horizon length and we can utilize equation (109) to calculate minimum investment time horizon which equals the time required to have the SNR of the geometric risk premium equal one. SNR equaling one means half of the realized risk premium is explained by noise (risk) while the other half is explained by signal (reward, i.e., the risk premium). When SNR is greater than one, more than half of the realized risk premium is explained by the (expected) risk premium. Equation (109) shows that the minimum investment time horizon is simply the reciprocal of SNR.

Figure 31 shows the empirical (calculated based on realized empirical risk premium SNR) minimum investment time horizons for different portfolio sizes and for the fully diversified benchmark. We can see that diversification greatly affects the required investment time horizon length. As an example, at investment fraction one, benchmark requires 6.9 years, while a 10-stock portfolio requires 19.6 years to have in minimum half of the realized risk premium explained by the (expected) risk premium. With leverage, the difference is even more dramatic. At investment fraction 1.5, corresponding numbers are 9.5 and 40.3 years.

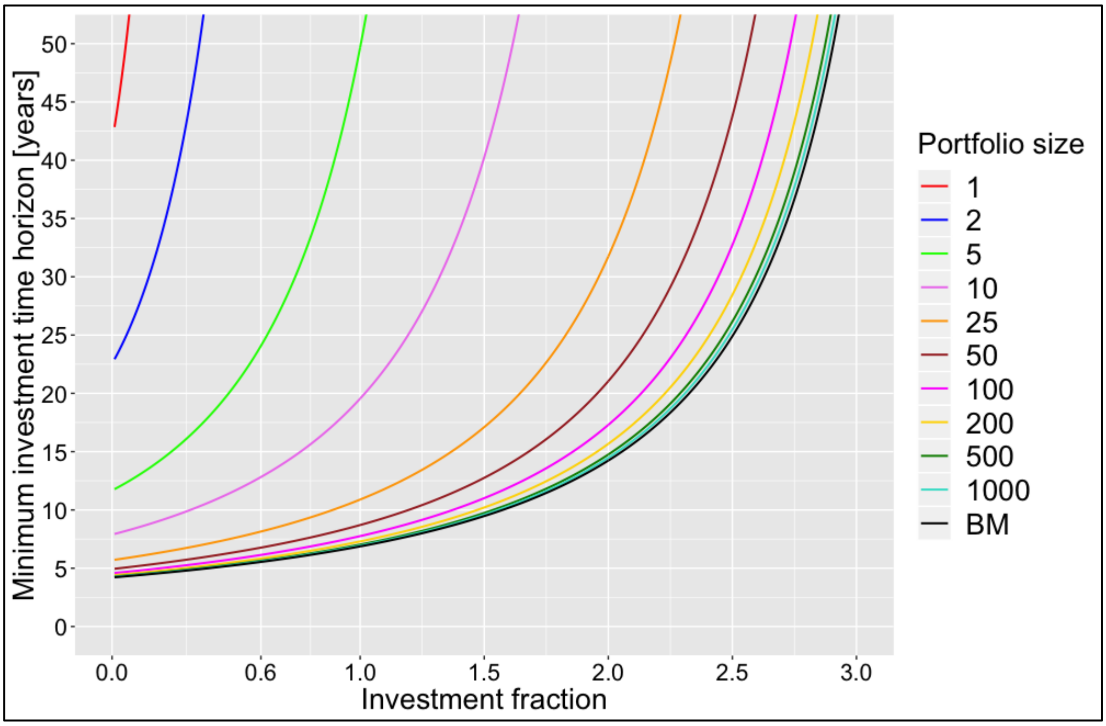

Figure 31. Minimum empirical investment time horizon.

5.4.3 Predicting required number of stocks for maintaining constant risk premium ratio

Next, the required level of diversification to achieve target risk premium ratio for selected investment fraction will be calculated and shown as a function of time. We will show separate metrics for risk neutral and risk averse investors. Risk averse metric will further be divided to short-term and long-term versions, while risk neutral metric is for short-term only.

Risk neutral metric is risk premium ratio calculated based on equation (129). Target risk premium ratio in Figure 32 is 0.90 meaning that 90% of the maximum risk premium ratio is achieved at about 20-stock portfolio size. Short-term risk averse metric is based on equation (131), which defines short-term as infinitely short. Long-term risk averse metric is based on equation (128), which takes time horizon as an input. The long-term metric is a function of time and as the SNR of the realized risk premium increase when time horizon increase, the required number of stocks to achieve 90% of the maximum realizable risk premium ratio decreases as a function of time.

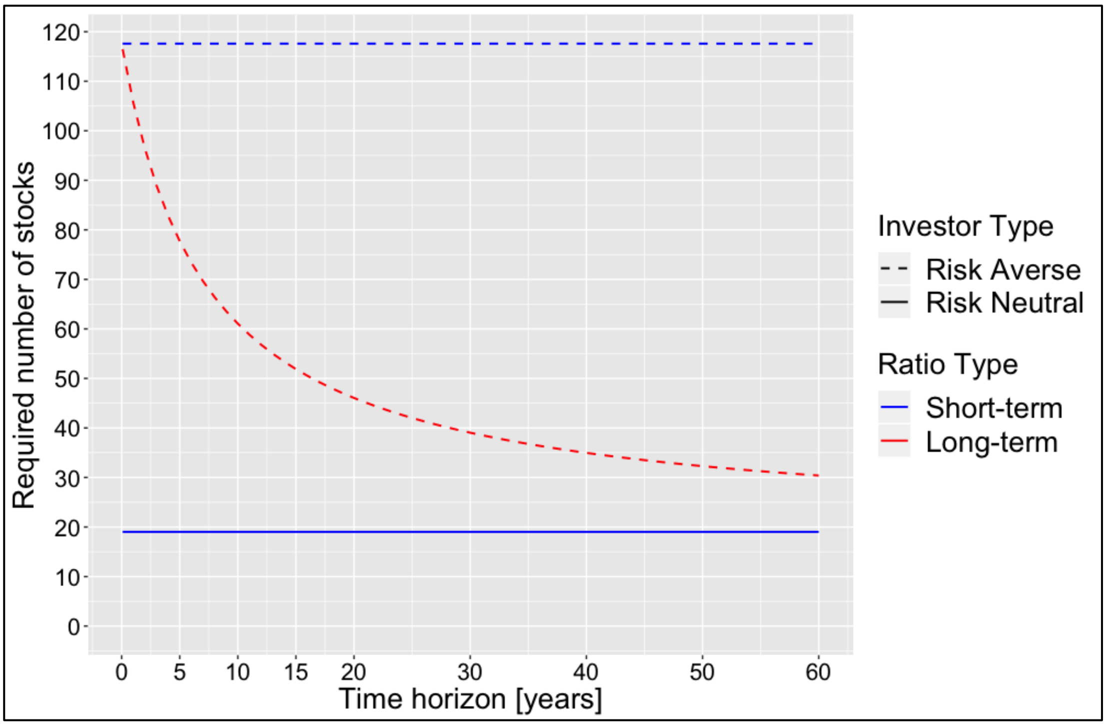

Figure 32. Required number of stocks for 0.90 risk premium ratio when $f = 1$.

Figure 32 shows the discussed three (realizable) risk premium ratio-based diversification metrics for investment fraction one as a function of time. Equations (130) (risk neutral), (131) (short-term risk averse) and (128) (long-term risk averse) are used to calculate the required number of stocks n_P. In case of equations (131) and (128), n_P is solved by computer. Close to 20 and 120 stocks are required to achieve 90% of the maximum short-term diversification benefit for risk neutral and risk averse investor, respectively. As the annualized standard deviation decreases as a function of a square root of time, the long-term realizable risk premium ratio based required number of stocks for risk averse investor decreases as a time horizon lengthens. However, we consider risk premium ratio primarily as a short-term metric and do not include the long-term risk averse metric in the set of our final diversification metrics. We do, however, utilize the long-term realizable risk premium when calculating realizable gross excess wealth ratio in section 5.6.3.

5.5 The effect of diversification on gross compound excess wealth illustrated

We will illustrate how thinking solely in terms of growth rates can be misleading. Although our focus is in geometric risk premium, we need to keep in mind the risk premium is not the final goal of a long-term investor. The final goal is the accumulated gross excess wealth over time, the total wealth accumulated by compounding geometric risk premium over time. We will measure the accumulated *excess* wealth, the wealth in excess of what compounding riskless rate over time produces, to assess the benefit of bearing risk.

As time horizon lengthens, the distribution of geometric rate of excess return converges towards its expected value, the geometric risk premium. It is tempting to interpret this convergence as a favorable effect of time as the uncertainty about the excess growth rate decreases as time horizon increases. However, when assessed in terms of what really counts to an investor, the amount of wealth after the investment horizon, we find the distribution of the final excess wealth, as opposed to the distribution of the excess growth rate, diverges as the time horizon lengthens. The effect of time therefore can be considered as increasing the uncertainty about the investor's excess wealth at the end of the investment horizon.

Next, we illustrate the diversification effect expressed as the gross compound excess wealth difference to benchmark over time. We use *gross* excess wealth implying the initial investment is included in final wealth. Four figures, from Figure 33 to Figure 36, show the cross-sectional distribution for the 50 000 bootstrapped total gross excess returns (wealths) at the end of the 45.5 year investment horizon. As this is a cross-sectional distribution, the variance is entirely idiosyncratic and the systematic variance plays no role. We can think of the whole cross-sectional distribution as riding on an uncorrelated common systematic benchmark-beta wave, the time series variance of the benchmark.

We have four metrics shown in the figures. Portfolio mean (over the 50 000 portfolios) realized gross excess return is expected to correspond to benchmark gross compound excess wealth. Benchmark gross compound excess wealth is the realized benchmark gross excess wealth. Portfolio median (over the 50 000 portfolios) realized gross excess return is expected to match closely with the portfolio gross compound excess wealth which is the continuously compound excess wealth using geometric risk premium (mean excess growth rate over the 50 000 portfolios) as growth rate. Bin width in the histograms is 0.1 and y-axis shows the number of occurrences per bin.

Figure 33 shows the cross-sectional distribution for bootstrapped single stock portfolios. We don’t show the portfolio gross compound excess wealth nor the portfolio median realized gross excess return as these metrics are very close to zero (at 0.042 corresponding to -95.8% excess return) and would hide the bulk of the distribution also concentrated close to zero. The distribution is fat-tailed and extremely positively skewed. In other words, typical gross excess return is very, very different from the mean gross excess return dominated by rare very high returns. Furthermore, despite the very large number (50 000) of bootstrapped return histories, the portfolio mean fails to converge to its expectation, the benchmark gross compound excess wealth. This is expected in the light of Taleb, Bar-Yam and Cirillo (2020) showing how the mean converges very slowly and is typically underestimated in the presence of fat-tails accompanied with positive skewness.

It seems clear that assessing the long-term expected return for a single stock portfolio based on the expected return of a benchmark makes no practical sense at all. This implies that utilizing a one period model, which implicitly assumes benchmark properties (apart from variance) as generalizable to any portfolio size, to assess diversification effect must be practically insufficient.

Figure 33 is a manifestation of the ergodicity problem described by Peters (2019). The experience of a typical (median) individual long-term single stock investor is completely different compared to the experience of aggregated (mean) single stock investors. We live our lives as individuals and hence our experience is better described by the excess wealth accumulated using geometric, instead of arithmetic, risk premium as a growth rate.

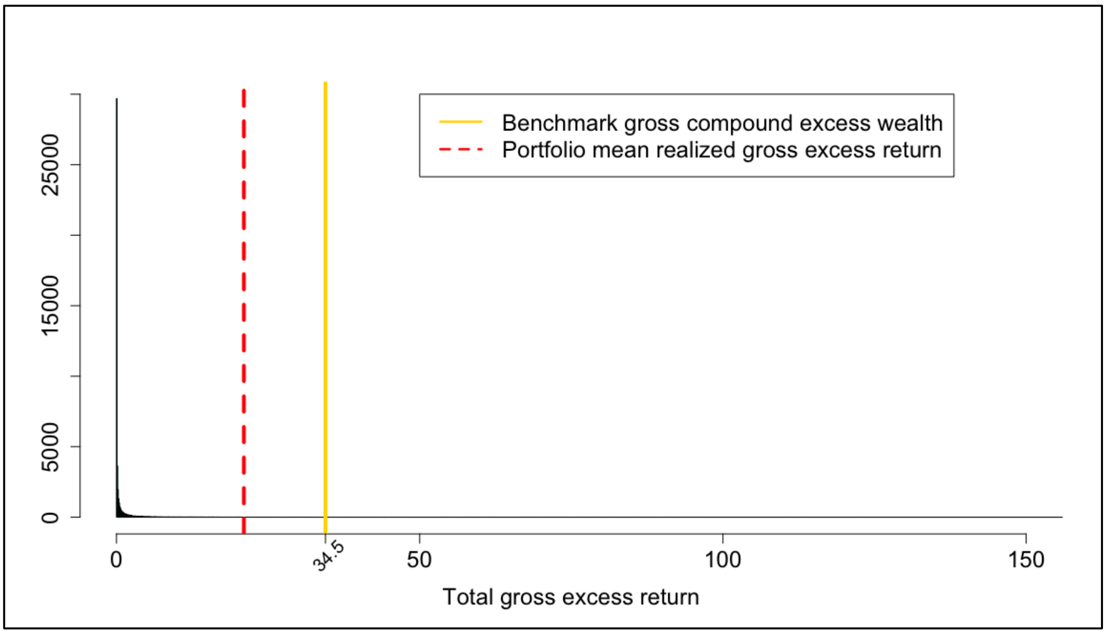

Figure 33. Cross-sectional gross excess return distribution for bootstrapped 1-stock portfolios.

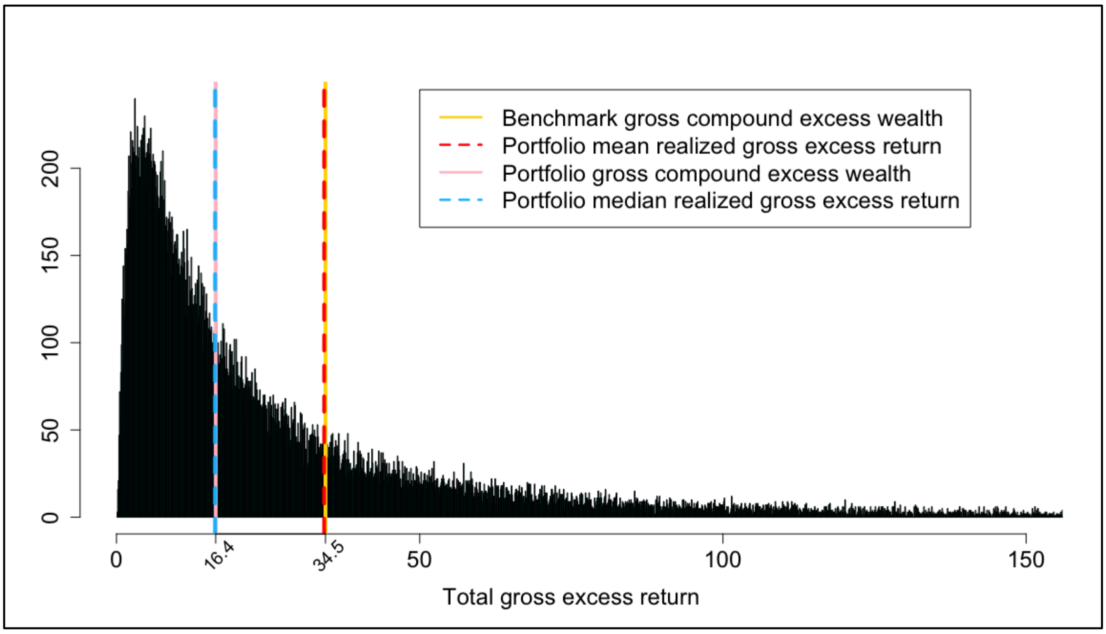

Figure 34. Cross-sectional gross excess return distribution for bootstrapped 10-stock portfolios.

Figure 34 shows the distribution for ten-stock portfolios. Considering the conventional wisdom that about ten stocks make a diversified portfolio, this figure speaks entirely different language. Skewness of the distribution is still very well visible. Typical portfolio total gross excess return, measured as portfolio gross compound excess wealth or as a median of the distribution, is less than half of the benchmark gross compound excess wealth. Now we notice that the mean has converged to its expectation, i.e., the portfolio mean realized gross excess return and benchmark gross compound excess wealth are equal in visual examination.

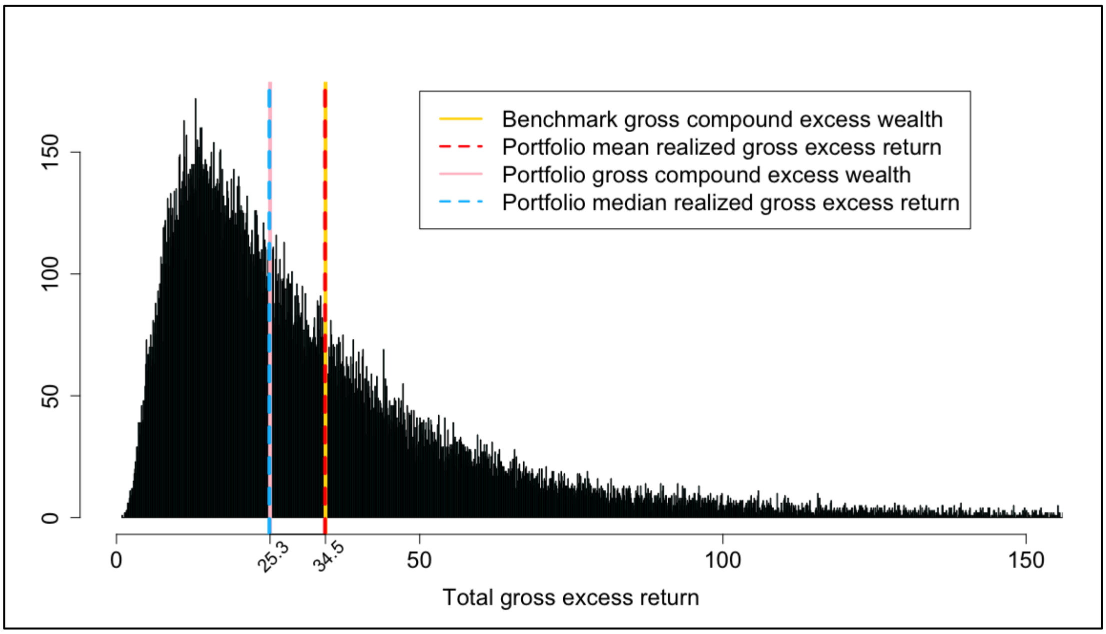

Figure 35. Cross-sectional gross excess return distribution for bootstrapped 25-stock portfolios.

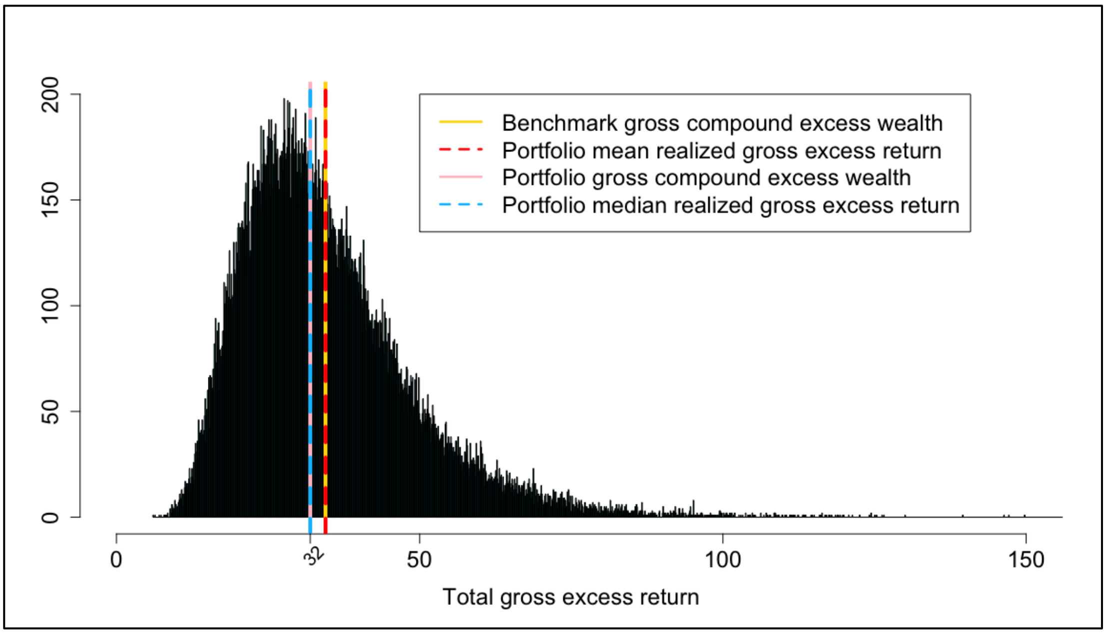

Figure 36. Cross-sectional gross excess return distribution for bootstrapped 100-stock portfolios.

Figure 35 and Figure 36 show the distributions for twenty-five and hundred stock portfolios, respectively. 25 stocks still show a large difference between typical and mean values and there is a small, but clearly observable, difference even in the 100-stock distribution.

These distributions are from a 45.5-year time span. As time horizon lengthens, the standard deviation and the skewness both increase as a function of a square root of time. Increase in variance, including the costly idiosyncratic variance, is directly proportional to time. This implies that the cross-sectional excess return distribution diverges as a function of time, which means the difference between the excess wealth of an average portfolio and its benchmark increases as a function of time. In other words, the diversification becomes more important as a function of time. Equations (134) and (135) show that to maintain the desired wealth ratio (portfolio gross compound excess wealth ratio to corresponding benchmark metric), which is equivalent to maintaining the shape of the distribution, number of stocks needs to increase directly proportionally to increase in investment time horizon length.

The magnitude of the diversification effect is very different depending whether we consider it in the context of time indifferent excess growth rate or in the context of accumulated excess wealth over time. We illustrated the former diversification effect in the time period from January 1973 to June 2018 in Figure 21 and for selected portfolio sizes in Table 5. The diversification premium difference to benchmark is quickly decimated as the number of stocks in the portfolio is increased from one to ten, twenty-five or hundred stocks. Realized diversification premium difference to benchmark in Table 5 is around -14.8pp, -1.6pp, -0.7pp and -0.17pp, for the portfolio sizes respectively. Judging solely based on the excess growth rates, it would be easy to conclude that 10, 25 or certainly 100 stocks are sufficient as the growth rate difference to benchmark seems tolerable. The effect of time, however, may call for reconsideration.

In Table 12 we summarize some key metrics from the figures when considering the diversification effect in the context of gross compound excess wealth over time. The data used in the table is identical to those used in assessing the diversification premium difference to benchmark in Figure 21 and in Table 5. Additionally, the portfolio sizes match with those in Table 5.

The data in the table shows that the effect of diversification on gross compound excess wealth ratio, the portfolio gross excess wealth divided by that of the benchmark, is dramatic. For a 10-stock portfolio, less than half of the benchmark gross wealth no longer sounds as tolerable as the about 1.6pp annualized growth rate difference. 10-stock portfolio lose to benchmark in close to three quarters of all portfolios and to risk free rate about in one out of a hundred portfolios. About one out of ten portfolios produce more than twice the benchmark wealth and the best portfolio out of 50 000 produce close to fifty times the benchmark wealth.

Table 12. Key cross-sectional gross excess wealth metrics.

	Portfolio size			
	1	10	25	100
Gross excess wealth ratio	0.001	0.475	0.735	0.927
Gross excess wealth < BM [%]	96.7	73.3	65.5	57.8
Gross excess wealth < RF [%]	80.9	1.0	0.0	0
Gross excess wealth > 2*BM [%]	2.1	11.7	9.8	2.5
Max gross excess wealth ratio	5303.9	49.4	24.1	5.4

For a single stock portfolio, it is realistic to assume that practically all wealth is lost in the long-run. On average, when compounding the geometric risk premium over time to form the final gross excess wealth, about one permille of the benchmark gross wealth is expected in 45.5-year period. Furthermore, more than four out of five portfolios are expected to accumulate less wealth than compounding riskless rate would. However, for those into lottery, the best single stock portfolio provides a phenomenal wealth, more than 5300 times that of the benchmark. It is noteworthy, that single stock portfolio is not more likely than the 100-stock portfolio to produce a wealth greater than twice the benchmark wealth, but does produce some extreme winners which contribute enormously to the mean. Mean return therefore bears no information, expect reminding that some rare lottery winners exist, for typical investor.

5.6 Gross compound excess wealth ratio as a diversification effect measure

5.6.1 The effect of time

Next, we will show how the effect of time on maintaining constant gross compound excess wealth ratio is approximately directly proportional to required portfolio size. This linear relationship holds very accurately in the range from one to about one hundred stocks and remains relatively accurate until portfolio size is several hundreds of stocks. This implies the longer the investment time horizon, the more important diversification becomes.

As illustrated in section 5.5, the wealth distribution after a long investment horizon is very much a function of diversification. The gross excess wealth ratio decreases while the asymmetry and fat tailedness, skewness and excess kurtosis of the excess gross wealth distribution, respectively, increase as a function of idiosyncratic variance of the portfolio's growth rate. The level of idiosyncratic variance is affected both by the level of diversification and time. It is well known that standard deviation is a function of a square root of time implying variance is directly proportional to time. Furthermore, skewness is a function of a square root of time. Controlling the effect of time is beyond us, but we can control the level of diversification to affect the wealth distribution at the end of the investment horizon.

As time irresistibly marches on, we need to increase our level of diversification as a function of time if we intend to keep the expected wealth distribution asymmetry unchanged meaning if we intend to prevent the gross compound excess wealth ratio from decreasing as a function of time. As shown in section 5.5, increasing idiosyncratic variance implies diverging gross excess wealth difference between investment portfolio and benchmark. As idiosyncratic variance of a portfolio increases directly proportionally to time and decreases approximately inversely proportionally to number of stocks in the portfolio, the effect of time on wealth distribution is compensated by increasing the level of diversification approximately directly proportionally to increase in investment time horizon. This is shown by approximate equation (135) and by the corresponding exact equation (134), which accounts for the finite average number of stocks in the benchmark.

Due to this approximately inverse proportionality between time and number of stocks in a portfolio, we can think of the 45.5 year investment time horizon gross excess wealth ratios in Table 12 alternatively so that the one stock portfolio wealth ratio is interpreted as the ten-stock portfolio wealth ratio after 10*45.5=455 years. Or we can interpret the hundred-stock portfolio wealth ratio as the ten-stock portfolio wealth ratio after 45.5/10=4.55 years.

We will test the accuracy of the exact equation (134) empirically using the data from January 1973 to June 2018. Empirical portfolio size per time horizon is selected based on equation (134). Solid lines in Figure 37 and Figure 39 are based on exact equation (134) while the dashed lines are based on approximate equation (135). The data is from a 45.5 period which equals 546 months. To test the effect of time, we accept only (selected) time periods by which the 546 months is divisible to ensure identical full data is always utilized for each tested sub-period length. We select a target wealth ratio (TWR) to present the targeted gross compound excess wealth ratio between an investment portfolio and the benchmark at the end of an investment time horizon. Our equations predict the approximate (exactly linear relationship with time) and exact (approximately linear relationship with time) number of stocks required to keep the wealth ratio at target at as a function of investment time horizon.

Figure 37 shows the result for TWR 0.90, 0.95 and 0.99, which correspond to wealth target of 90%, 95% and 99% of the benchmark gross excess wealth at any investment time horizon length. Y-axis on the left is the average gross compound excess wealth ratio while y-axis on the right is the portfolio size. Table 13 provides the portfolio sizes predicted by the exact equation (and used to form empirical portfolios) and the difference between exact portfolio size and linear approximation for each TWR. We can see from the figure that TWR are stable and horizontal except for the first point on the 0.90 TWR curve which is explained by inaccuracy from rounding the decimal number provided by the formula to integer (1.5 rounded to 2). Overall, all three targeted gross excess wealth ratios remain constant as the investment time horizon lengthens implying the exact equation (134) predicts the effect of investment time horizon well empirically. Approximate (dashed line) portfolio size curves are exactly linear in the figure. When we compare the exact (solid line) portfolio size curves to approximate curves within each TWR, we can see that the curves are almost exactly aligned when portfolio size is about one hundred or less. It means that portfolio sizes smaller than about one hundred, which the range meaningful to most stock pickers, are in the range where linear approximation very accurately captures the effect of time. Linear approximation remains reasonably accurate for larger portfolio sizes too, but we can see there is a noticeable difference when portfolio size is in the range of several hundreds of stocks. As the portfolio sizes approach the benchmark size, linear approximation will cease to work.

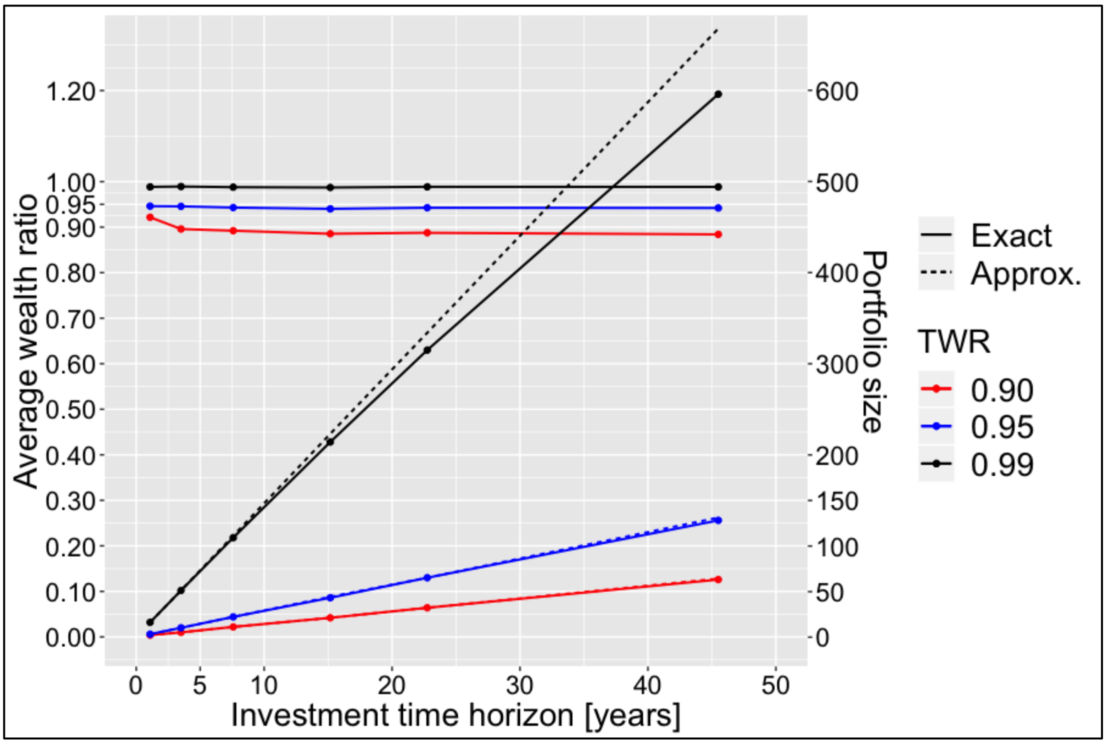

Figure 37. Approximately linear relationship between time and portfolio size to maintain constant TWR.

Table 13. Predicted required number of stocks to maintain constant wealth ratio as a function of time.

	Predicted exact num. of stocks (approx. - exact)		
Time horizon [years]	TWR = 0.90	TWR = 0.95	TWR = 0.99
1.08	2 (0)	3 (0)	16 (0)
3.50	5 (0)	10 (0)	51 (0)
7.58	11 (0)	22 (0)	109 (2)
15.17	21 (0)	43 (1)	214 (9)
22.75	32 (0)	65 (0)	315 (19)
45.50	63 (1)	128 (3)	596 (72)

In Figure 38 we can observe the TWR curves (solid lines) shown in Figure 37 more accurately. Table 14 shows the same numerically. We can see that realized average gross excess wealth ratios are slightly below the targeted ratios. This is in line what we saw in Figure 21 where realized diversification premium differences to benchmark were slightly larger compared to corresponding predictions. Cutting the fat tails helped to converge the curves in case of diversification premium differences to benchmark metric. Figure 23 shows the result after cutting the percent tails from each month's data. For TWR tests, we show the effect of cutting the percent tails in Figure 38 using dashed lines. Cutting the tails makes the realized average gross excess wealth ratios extremely accurate. The remaining noticeable inaccuracy in the short time intervals (small portfolio sizes) is due to rounding. Our takeaway is that in the presence of fat tails the number of stocks required to maintain the desired TWR, as given by equations (134) and (135), is the lower bound and required portfolio sizes using monthly rebalancing are expected to be slightly higher than predicted by these equations. Again, we see how the fat tails make diversification more important.

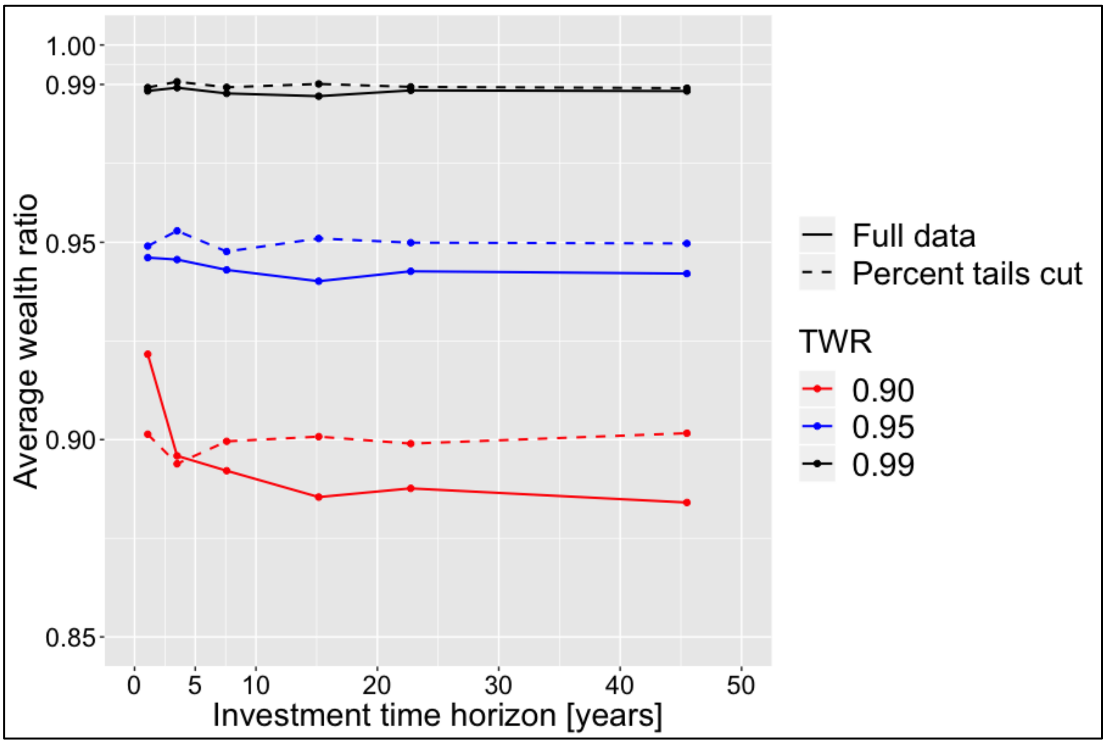

Figure 38. The effect of fat tails on realized target gross excess wealth ratio.

Table 14. Realized average wealth ratios for full and percent tails cut data.

	Realized avg. wealth ratio: Full data; Percent tails cut data		
Time horizon [years]	TWR = 0.90	TWR = 0.95	TWR = 0.99
1.08	0.9216; 0.9014	0.9461; 0.9490	0.9884; 0.9892
3.50	0.8959; 0.8939	0.9456; 0.9529	0.9892; 0.9907
7.58	0.8921; 0.8996	0.9430; 0.9477	0.9877; 0.9893
15.17	0.8854; 0.9008	0.9401; 0.9510	0.9870; 0.9901
22.75	0.8877; 0.8990	0.9427; 0.9499	0.9885; 0.9894
45.50	0.8841; 0.9016	0.9421; 0.9497	0.9883; 0.9891

We consider the empirical results shown in this section to be highly supportive to *hypothesis 5*.

Hypothesis 5: For a risk neutral long-term investor, the number of stocks required to make a diversified portfolio is approximately directly proportional to investment time horizon length.

5.6.2 The effect of investment fraction

Time is approximately linearly related to required portfolio size to maintain a constant gross excess wealth ratio. Next, we will show that in the investment fraction dimension we need to multiply the portfolio size by approximately squared investment fraction to maintain the wealth ratio constant. This is in line with what was shown in 5.3.2, where we showed that the diversification premium difference to benchmark is a function of squared investment fraction.

In Figure 39 we test the accuracy of equation (134) empirically by setting the investment time horizon constant (45.5 years) and changing the investment fraction in the range from 0.6 to 1.5. For portfolio sizes, we show both the prediction by exact (solid line) and approximate (dashed line) equations, (134) and (135) respectively. Empirical portfolio size per investment fraction is selected based on equation (134). Table 15 shows the portfolio sizes predicted by the exact equation and the difference between exact portfolio size and quadratic approximation for each TWR. We can see that the approximation (squared investment fraction relationship) is very accurate in portfolio sizes below about one hundred stocks and remains reasonably accurate up to portfolio sizes of several hundreds of stocks. Overall, realized TWR remains very stable as a function of investment fraction implying equation (134) predicts the effect of investment fraction well empirically. Take TWR 0.90 as an example and we can see how big a difference the investment fraction makes. The difference in required portfolio size between investment fractions 0.6 and 1.5 is more than six-fold (6.09). Using quadratic approximation, the expected difference would be $1.5^2/0.6^2 = 6.25$.

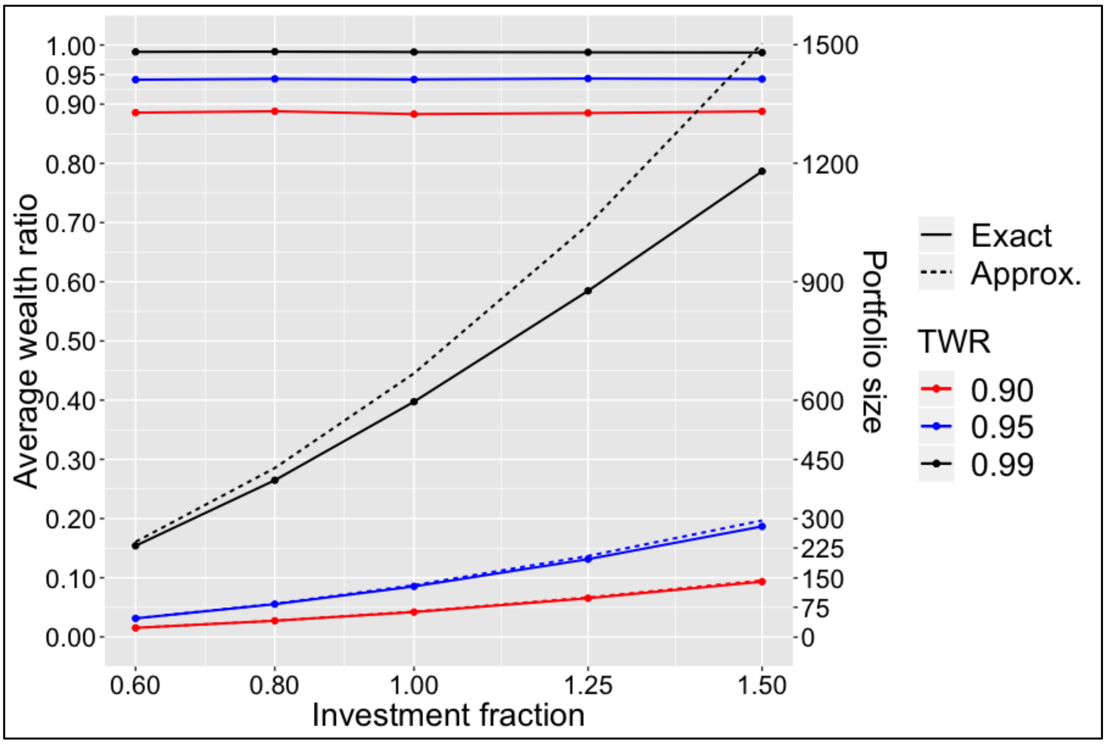

Figure 39. Approximately quadratic relationship between investment fraction and portfolio size to maintain constant TWR.

Table 15. Predicted required number of stocks to maintain constant wealth ratio as a function of investment fraction.

	Predicted exact num. of stocks (approx. - exact)		
Investment fraction	TWR = 0.90	TWR = 0.95	TWR = 0.99
0.6	23 (0)	47 (0)	231 (10)
0.8	41 (0)	83 (1)	397 (31)
1.0	63 (1)	128 (3)	596 (72)
1.25	98 (2)	197 (8)	877 (167)
1.5	140 (3)	280 (15)	1180 (324)

We consider results shown in Figure 39 and Table 15 as supportive to *hypothesis 6*.

Hypothesis 6: For a risk neutral long-term investor, the number of stocks required to make a diversified portfolio is an increasing, approximately squared, function of investment fraction.

5.6.3 Predicting required number of stocks for maintaining constant wealth ratio

Next, we will show our long-term diversification metrics for risk neutral and risk averse investors work with empirical data. We use wealth ratio-based metrics to describe the long-term diversification effect. Risk neutral investor metric is gross compound excess wealth ratio, while realizable gross compound excess wealth ratio is used for risk averse investor.

Figure 40 shows the predicted and bootstrapped gross compound excess wealth ratios for different portfolio sizes as a function of investment fraction. Investment time horizon is 45.5 years. Predicted values are based on equation (132). Bootstrapped values closely follow predicted values. There are deviations for moderate portfolio sizes and at high investment fraction values, which we attribute to fat-tailed return distribution and less than infinite rebalancing frequency. At investment fraction one, the resulting gross excess wealth ratios for portfolio sizes 1, 10, 25 and 100 stocks are identical with the values given in Table 12. Corresponding cross-sectional gross excess return distributions, for those four portfolio sizes, can be seen in figures ranging from Figure 33 to Figure 36. We can see from Figure 40 how investment fraction dramatically affects the gross excess wealth ratio at small portfolio sizes.

In Figure 41, we have predicted and bootstrapped realizable gross compound excess wealth ratios for different portfolio sizes as a function of investment fraction. The difference to gross compound excess wealth ratios in Figure 40 is that now realizable risk premium is used in the place of risk premium. Realizable risk premium, given by equation (128), is an increasing function of time. Predicted realizable gross excess wealth ratio is based on equation (141). Bootstrapped values closely follow predicted values. Deviations are similar to deviations in Figure 40. Similarly, as with the risk neutral metric, investment fraction dramatically affects the realizable gross excess wealth ratio at small portfolio sizes.

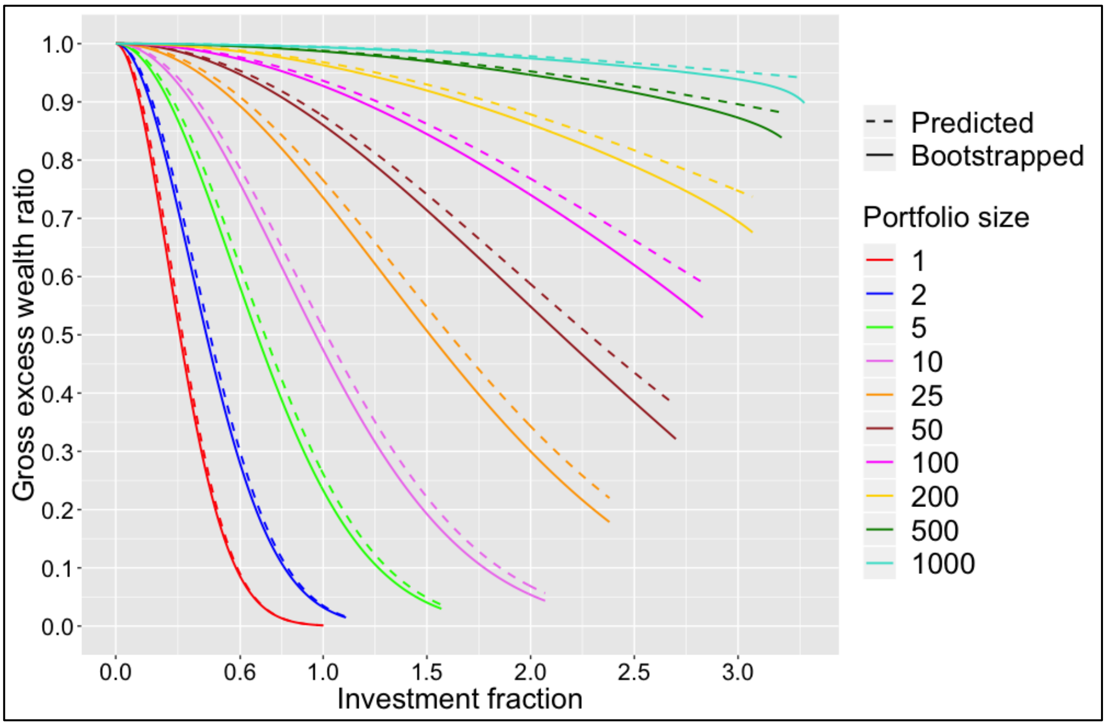

Figure 40. Predicted vs. bootstrapped gross excess wealth ratio at 45.5-year investment horizon.

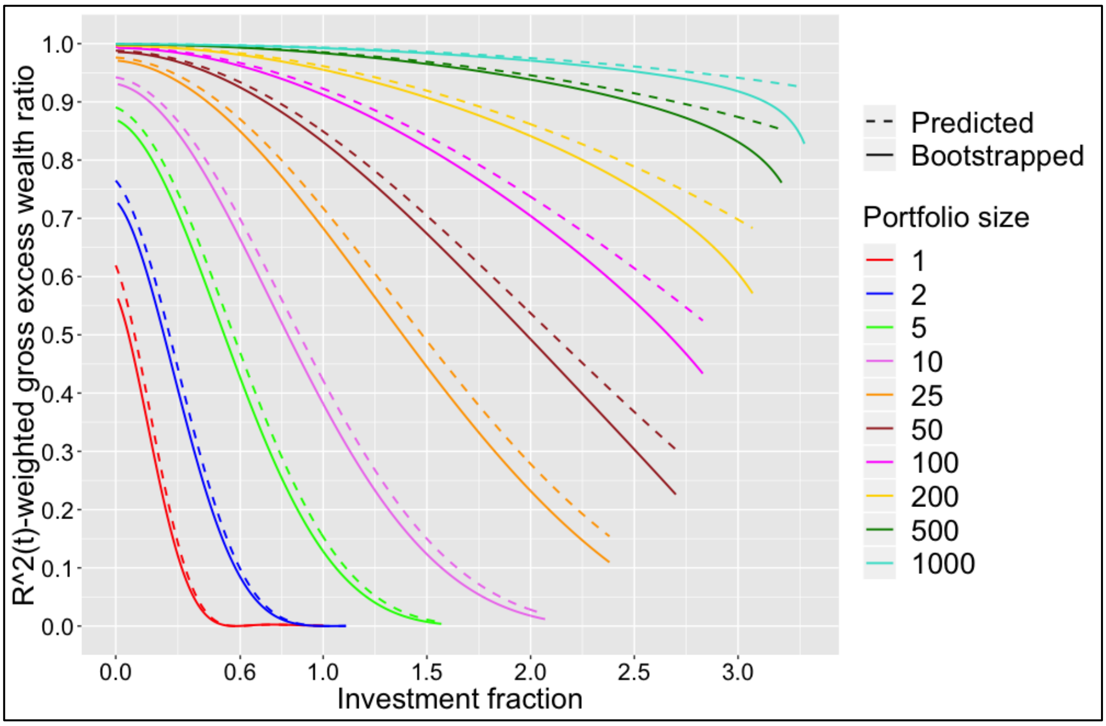

Figure 41. Predicted vs. bootstrapped realizable gross excess wealth ratio at 45.5-year horizon.

By utilizing gross compound excess wealth ratio and realizable gross compound excess wealth ratio for risk neutral and risk averse investors, respectively, we calculate the required number of stocks to achieve 90% of the maximum diversification benefit and show the result for investment fraction one in Figure 42. Equations (134) and (141) are used to calculate the required number of stocks n_P. In case of equation (141), n_P is solved by computer. We can see that the level of required diversification for risk neutral investor is approximately directly proportional to investment time horizon length. The longer the investment time horizon, the more important diversification is. For risk averse investor, the required level of diversification first decreases as a function of time as the realizable risk premium ratio decrease as a function of time (as shown in Figure 32). However, as the realizable gross compound excess wealth ratio asymptotically approach gross compound excess wealth ratio, which is an increasing function of time, eventually realizable gross compound excess wealth ratio starts to increase as a function of time. Risk averse metric finds its minimum at about 55 stocks around 15 years mark before starting to increase towards benchmark portfolio size as time horizon increases towards infinity.

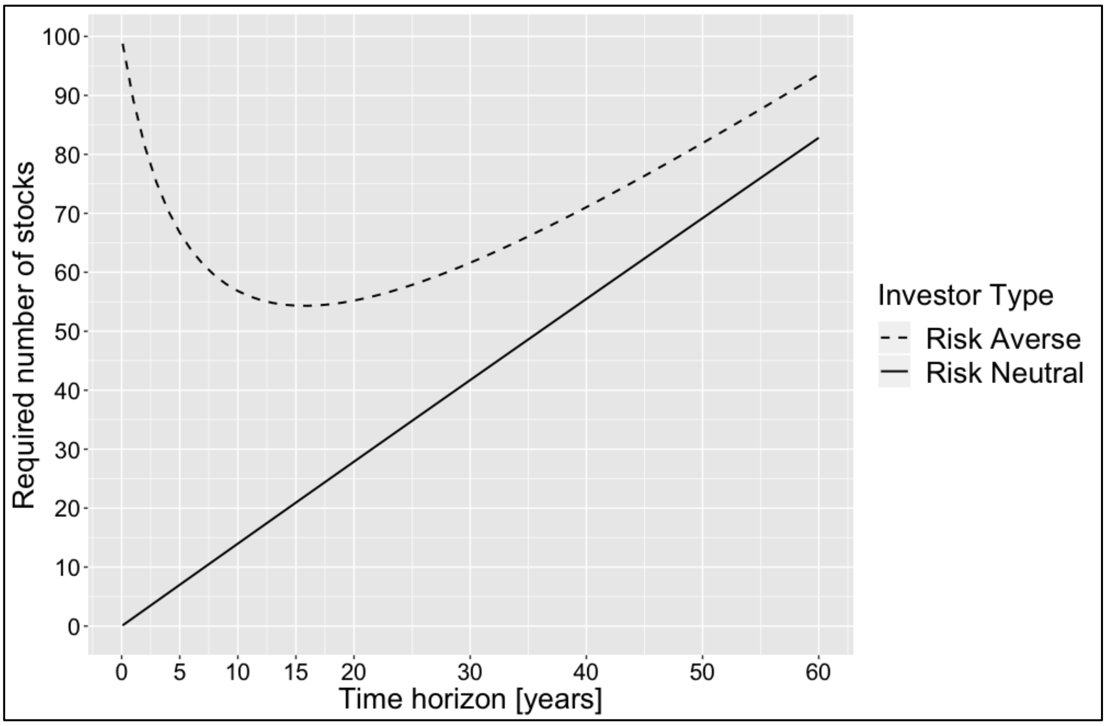

Figure 42. Required number of stocks for 0.90 gross excess wealth ratio when $f = 1$.

5.7 How many stocks make a diversified portfolio in a continuous-time world?

Finally, we show all of the selected diversification metrics in one figure. First metric, number of stocks required for positive risk premium (as defined in section 3.4.1 and shown to work empirically in section 5.1.2), is shown in black. Second metric, number of stocks required for a proportion of benchmark risk premium (as defined in section 3.4.2 and shown to work empirically in section 5.4) is shown in blue. Third metric, number of stocks required for a proportion of benchmark wealth over time (as defined in section 3.4.3 and shown to work empirically in section 5.6) is shown in red. For the second and third metrics, we show risk averse and risk neutral investor variants in dashed and solid line types, respectively. Risk premium-based metrics are used to describe short-term diversification effects, while gross excess wealth-based metrics are for the mid to long-term. Empirical data is the 45.5-year time span from January 1973 to June 2018.

Based on tests not shown here, the empirical number of stocks for the second and third metric, both for risk averse and risk neutral investors, is about 15% higher than predicted by our equations and shown in the figures and tables in this section. The systematic underestimation by our theoretical framework is attributable to less than infinite rebalancing frequency combined with fat-tailed return distribution as discussed in section 5.2.2.

We define the number of stocks required by risk neutral investor as the maximum between the risk premium and wealth-based metrics (the maximum among solid lines in the figures). In the short-term, risk premium ratio determines the required level of diversification. In the long-term, as compounding starts to have an effect, gross excess wealth ratio dominates and determines the required number of stocks. Similarly, we define the number of stocks required by risk averse investor as the maximum between the realizable risk premium and realizable wealth-based metrics (the maximum among dashed lines in the figures). Realizable risk premium ratio dominates and determines the required level of diversification for risk averse investor until the very long-term (close to hundred years).

However, realizable gross excess wealth ratio-based metric for risk averse investor (red dashed line in the figures) has a special interpretation. Long-term investor targeting a liability matching portfolio (e.g., an investor saving for retirement) can be thought to have succeeded if he achieves the targeted portfolio size at targeted date and failed if he falls short of the targeted portfolio size. Such investor does not particularly care about the upside after achieving the targeted portfolio size. A long-term investor targeting liability matching portfolio, even if he has a high tolerance for short-term volatility, can be thought to be risk averse in the long-term as he greatly dislikes the volatility (likes the predictability) of the ending wealth. Such investor is risk averse in the long-term even if he is totally unaware of the market moves (and therefore has infinite risk tolerance in the short-term) before the target date.

Red dashed line can be interpreted as a diversification metric for long-term risk averse investor while blue dashed line can be interpreted as a diversification metric for short-term risk averse investor. The red dashed line decreases in the mid-term as realizable risk premium increases, but rises in the long-term as the compounding of diversification premium difference to benchmark starts to have an effect. The blue dashed line is about the bumpiness of the ride, the short-term volatility, which investor must survive regardless the targeted investment horizon length.

Figure 43, Figure 44 and Figure 45 show the required number of stocks to achieve 90% of the maximum diversification benefit at investment fractions 1.0, 0.6 and 1.5, respectively. Comparing these figures shows how increasing investment fraction requires more diversification. First time horizon value in the figures is one month.

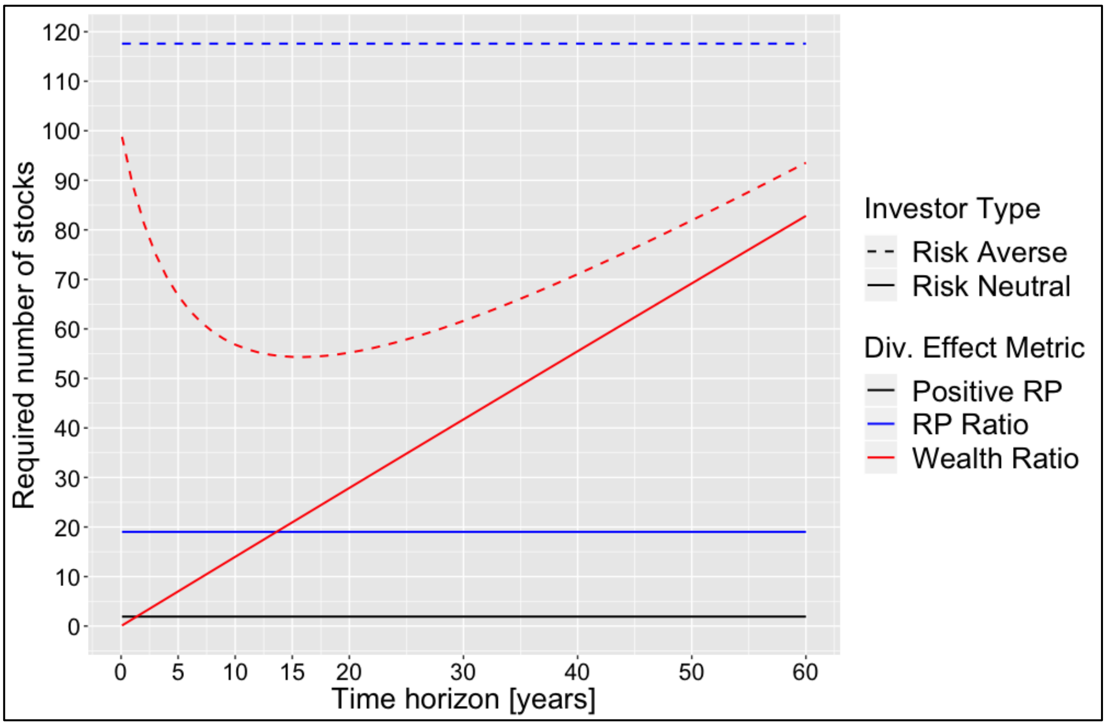

Figure 43. Required number of stocks for 90% diversification benefit when $f = 1$.

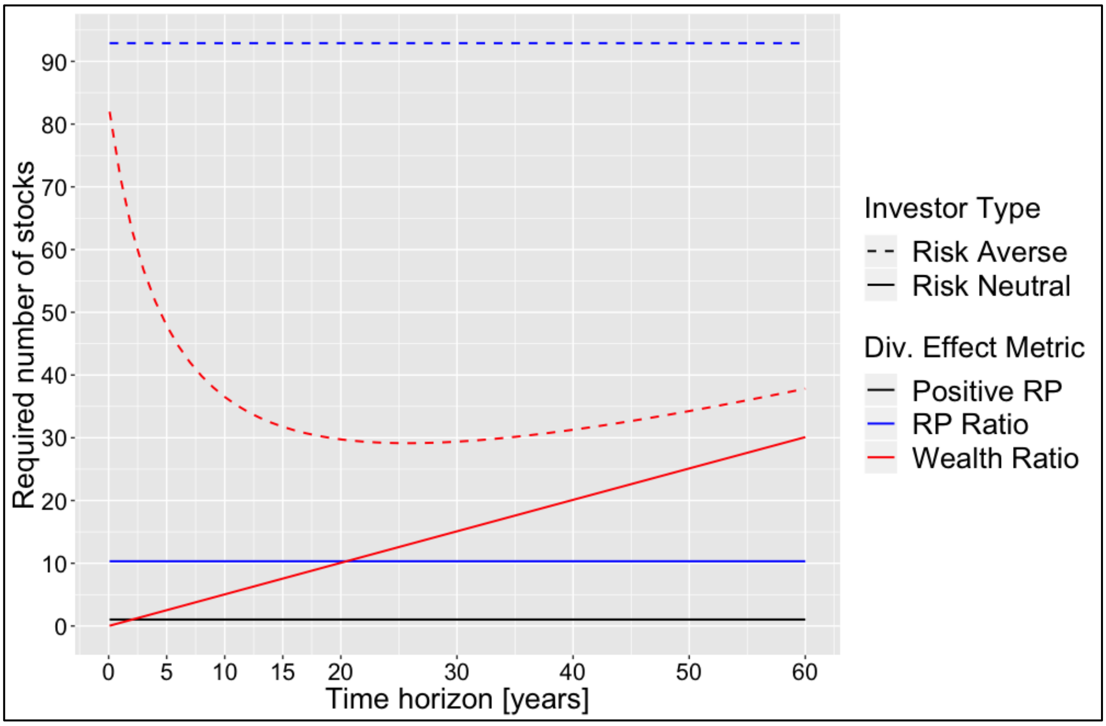

Figure 44. Required number of stocks for 90% diversification benefit when $f = 0.6$.

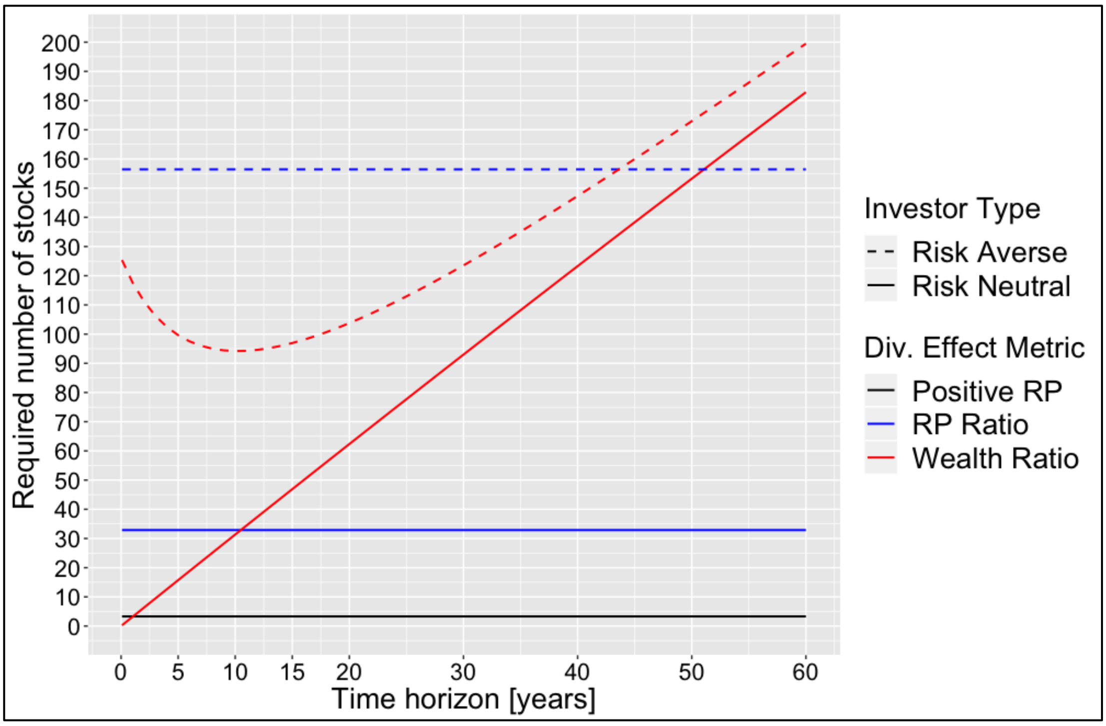

Figure 45. Required number of stocks for 90% diversification benefit when $f = 1.5$.

Figure 46 and Figure 47 show the required number of stocks at investment fraction one to achieve 95% and 99% of the maximum diversification benefit, respectively. Required number of stocks (for all metrics except the black solid line) to achieve 90% of the maximum diversification benefit are multiplied roughly by two and ten when moving from 90% to 95% and 99% of the diversification benefit, respectively. In 99% figure, we see that wealth ratio-based metric for risk neutral investor (red solid line) starts to lose its linearity as the resulting portfolio size starts to approach benchmark portfolio size.

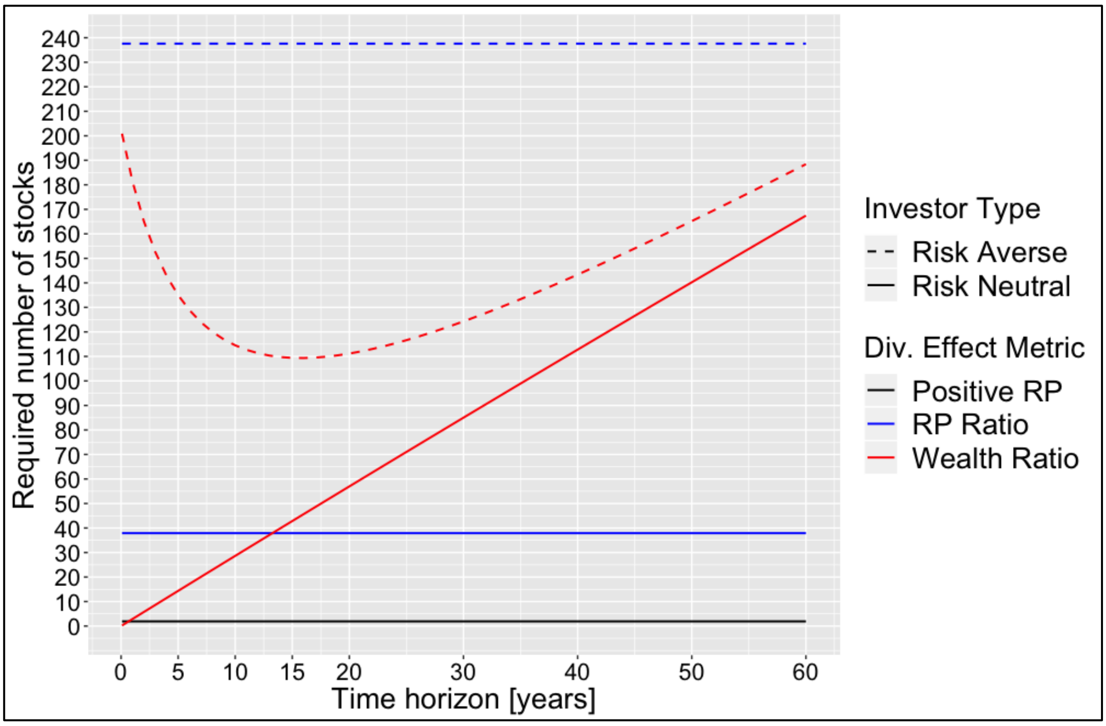

Figure 46. Required number of stocks for 95% diversification benefit when $f = 1$.

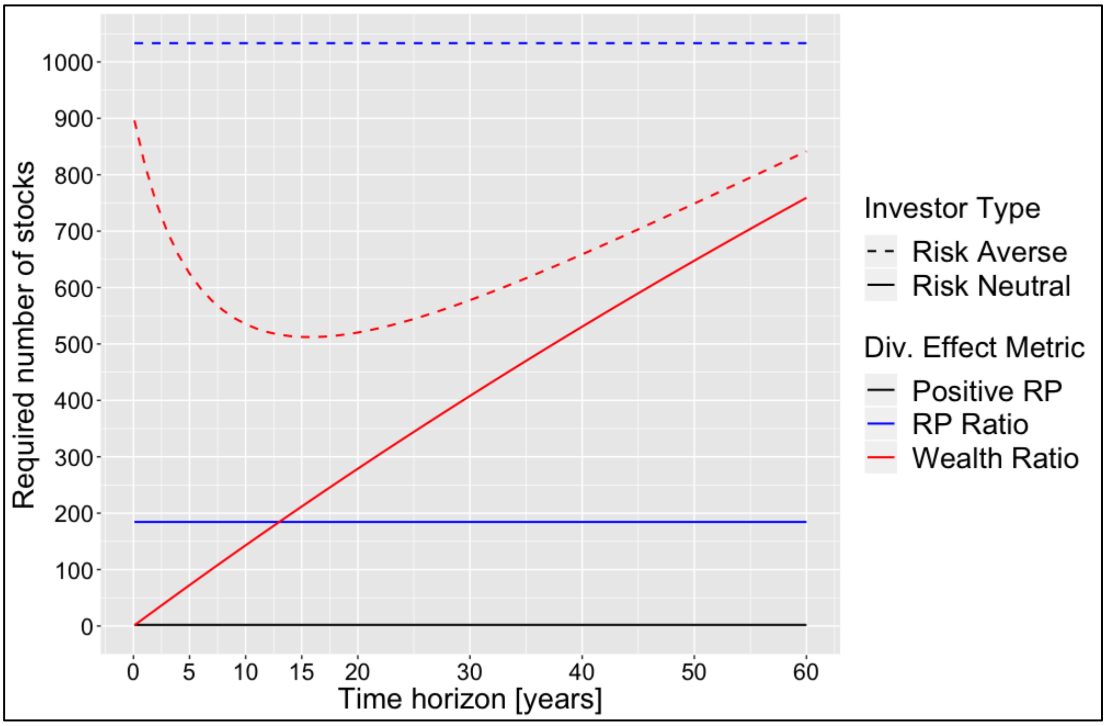

Figure 47. Required number of stocks for 99% diversification benefit when $f = 1$.

Table 16 and Table 17 summarize the required level of diversification numerically for the short and long-term, respectively. Both tables show the required level of diversification as a function of three different investment fractions and for risk averse and risk neutral investor separately. Table 16 additionally shows the minimum number of stocks required for positive risk premium, while Table 17 shows the long-term required level of diversification for four investment time horizons.

It is evident from these tables that investment fraction plays a significant role as a determinant of diversification effects in a continuous-time world. Similarly, for a long-term risk neutral investor, it is evident that the required level of diversification increases approximately directly proportional to investment time horizon. Risk averse investor always requires more diversification compared to risk neutral investor, but the difference decreases as a function of time. The remarkable property of diversification in a continuous-time world is that long-term risk neutral investor benefits from it almost as much as risk averse investor. In a single period world, there is no such thing as diversification benefit for a risk neutral investor.

Table 16. Required number of stocks for the short-term.

Investment fraction	0.6	1	1.5
Positive RP	1.03	1.91	3.31
	Panel A: 90% of the maximum diversification benefit		
Risk averse RP ratio	92.9	118	156
Risk neutral RP ratio	10.3	19.0	32.9
	Panel B: 95% of the maximum diversification benefit		
Risk averse RP ratio	190	238	312
Risk neutral RP ratio	20.6	37.9	65.4
	Panel C: 99% of the maximum diversification benefit		
Risk averse RP ratio	854	1033	1293
Risk neutral RP ratio	101	184	312

Table 17. Required number of stocks for the long-term.

Investment fraction	0.6	1	1.5
Time horizon [years]	10/20/40/60	10/20/40/60	10/20/40/60
	Panel A: 90% of the maximum diversification benefit		
Risk averse wealth ratio	36.5/29.7/31.3/37.8	56.8/55.2/71.0/93.6	94.2/104/147/200
Risk neutral wealth ratio	5.0/10.1/20.1/30.1	14.0/27.9/55.5/82.8	31.4/62.3/123/183
	Panel B: 95% of the maximum diversification benefit		
Risk averse wealth ratio	74.2/60.1/63.4/76.8	115/111/143/188	188/208/294/394
Risk neutral wealth ratio	10.3/20.6/41.1/61.5	28.6/57.0/113/167	64.0/127/247/363
	Panel C: 99% of the maximum diversification benefit		
Risk averse wealth ratio	358/291/307/370	535/520/659/842	838/912/1226/1552
Risk neutral wealth ratio	52.4/104/204/300	143/279/531/759	312/590/1065/1456

5.8 Investing style driving diversification effects

5.8.1 Firm size makes a difference

The time period from January 1973 to June 2018 is dominated by microcap stocks. On average, more than 60% of the stocks are microcaps. Diversification premium and consequently all of our diversification metrics are very much functions of firm size. We show that diversification premium increase monotonically as a function of firm size (measured as market capitalization) decile and that especially big, but also small, stocks require substantially less diversification compared to microcap stocks.

Figure 48 shows the diversification premium for benchmark portfolio as a function of firm size decile. Simultaneously, geometric risk premiums are shown for benchmark and single stock portfolios. Standard deviation for the benchmark geometric risk premium is shown as well. Stock deciles are formed monthly based on previous month market capitalization.

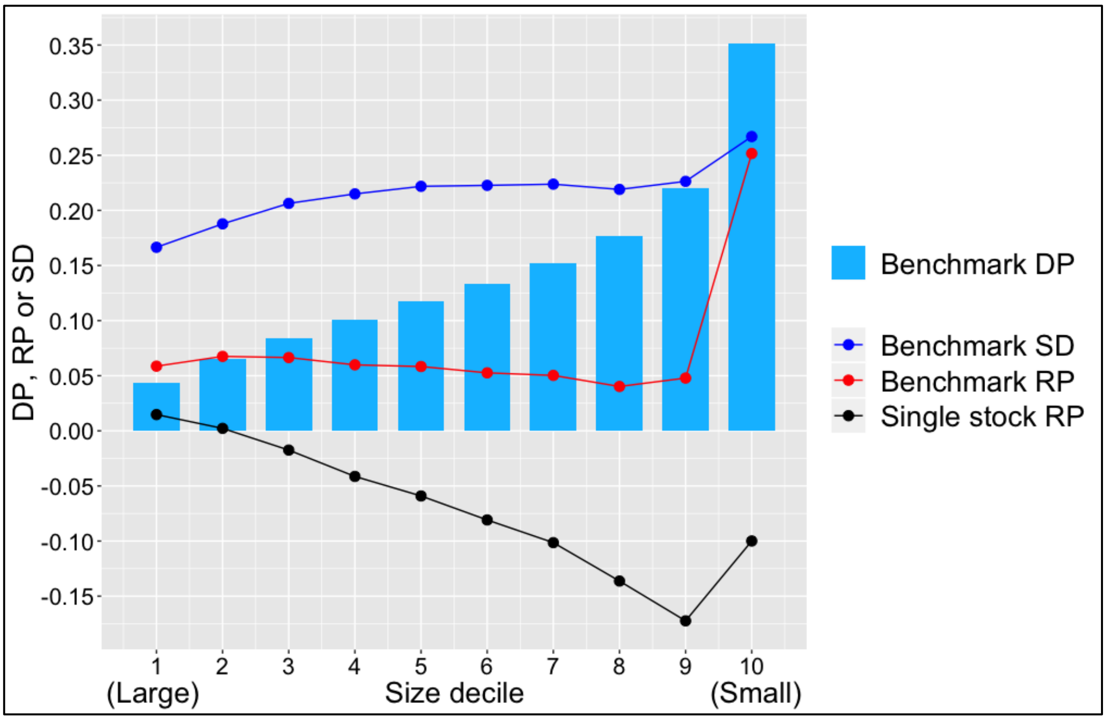

Figure 48. Diversification premium as a function of firm size.

We can see from Figure 48 that diversification premium increase monotonically as a function of firm size decile implying the cost of foregone diversification increase as a function of decreasing firm size. Single stock risk premium is positive for the two largest deciles and negative for the remaining 8 deciles. Two first size deciles roughly correspond to big stocks universe which on average consist of about 1000 firms. By randomly picking a single stock from the 8 smallest deciles, investor is expected to earn less than riskless rate. Diversification premium is a function of idiosyncratic variance. Idiosyncratic variance seems to somewhat correlate with systematic variance (standard deviation of the benchmark). Small stock premium for equally weighted portfolios appears to be fully explained by the very large risk premium for the smallest decile.

Figure 49, Figure 50 and Figure 51 show the required number of stocks to achieve 90% of the maximum diversification benefit at investment fraction one for microcap, small and big stocks, respectively. On average, close to three times more stocks are typically required for microcap portfolio compared to big stock portfolio.

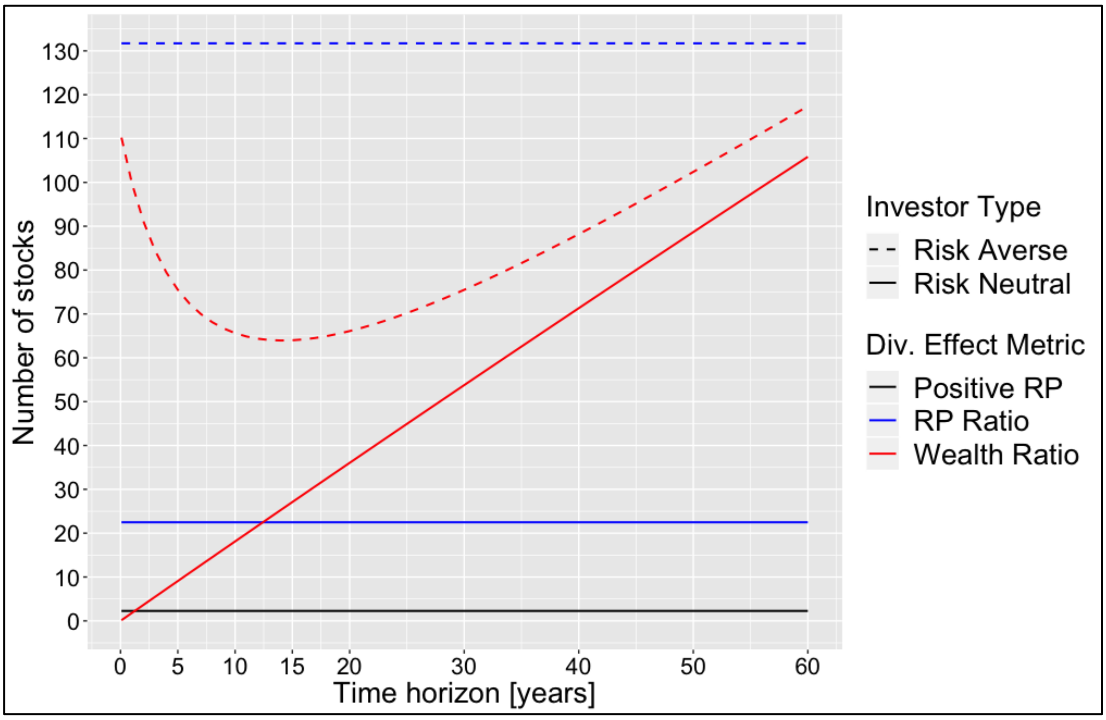

Figure 49. Required number of microcap stocks for 90% diversification benefit when $f = 1$.

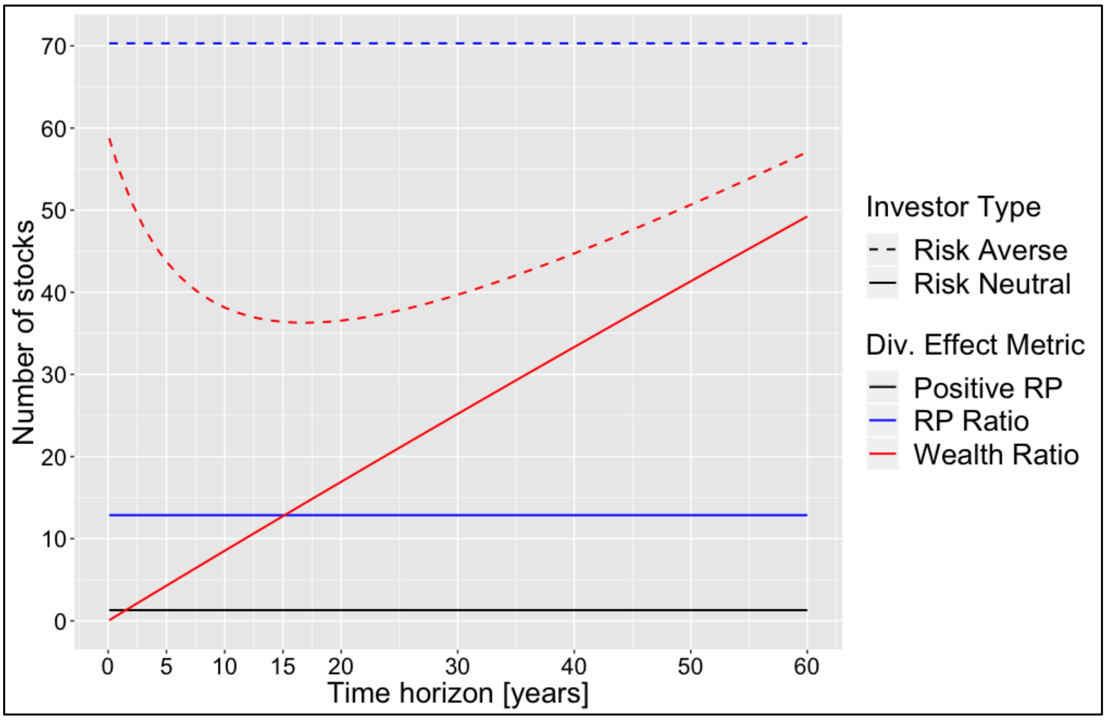

Figure 50. Required number of small stocks for 90% diversification benefit when $f = 1$.

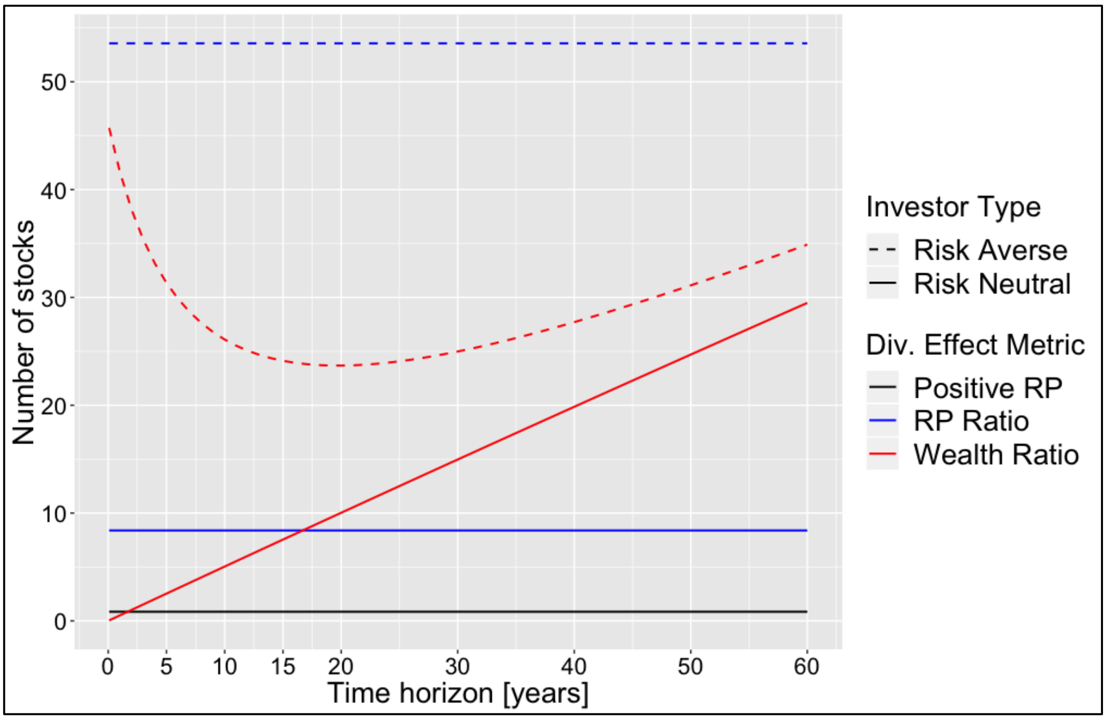

Figure 51. Required number of big stocks for 90% diversification benefit when $f = 1$.

Table 18 and Table 19 summarize the short-term and long-term required number of stocks to achieve 90% of the maximum diversification benefit for each firm size group at investment fraction one.

Table 18. Short-term required number of stocks per firm size for 90% diversification benefit at investment fraction one.

	Short-term diversification metric		
Firm size	Positive RP	Risk neutral RP ratio	Risk averse RP ratio
Big stocks	0.85	8.4	53.6
Small stocks	1.30	12.9	70.3
Microcap stocks	2.26	22.5	132

Table 19. Long-term required number of stocks per firm size for 90% diversification benefit at investment fraction one.

	Long-term diversification metric [10/20/40/60 years]	
Firm size	Risk neutral wealth ratio	Risk averse wealth ratio
Big stocks	5.0/10.0/19.9/29.5	26.1/23.7/27.7/34.9
Small stocks	8.5/16.9/33.3/49.2	38.1/36.6/44.7/57.1
Microcap stocks	18.1/36.0/71.3/106	65.6/66.1/88.2/117

5.8.2 Factor exposures determining diversification effects

In addition to firm size, we find that other commonly known investing styles significantly affect the required level of diversification. Exposure to high ROE (high earnings to book, i.e., high E/B), high momentum (high MOM), value stocks (high book to price, i.e., high B/P) or any combination of these (high earnings to price, i.e., high E/P) or high E/P combined with high MOM (high E/P&MOM) require significantly less diversification compared to portfolio with no factor exposure or opposite exposures to these factors. Notice that earnings yield E/P = (E/B)*(B/P) implying that high E/P is a combination of high ROE and value strategies. E stands for earnings, B for book value of equity, P for price and MOM for momentum.

We form the style portfolios by selecting highest and lowest 30% of the stocks to high and low style portfolios respectively. For example, value style portfolio (high B/P) is formed by selecting 30% of the stocks with highest B/P value each month, while growth style portfolio (low B/P) is formed by selecting the 30% of the stocks with the lowest B/P value each month. Six months lagged values are used for accounting variables (book value of equity and earnings) to avoid look ahead bias. Momentum is calculated based on last 12 month (excluding the last month) cumulative return. Book value of equity is defined following Brandt, Santa-Clara & Valkanov (2009) as total assets minus liabilities plus balance sheet deferred taxes and investment tax credits minus preferred stock value.

Stocks are ranked to 20 quantiles based on each style factor. Each quantile is given a quantile score corresponding to quantile rank. We combine earnings yield and momentum styles by summing the earnings yield and momentum quantile scores and by selecting the high earnings yield & high momentum style (high E/P&MOM) stocks as the top 30% (quantiles 15-20) of stocks with the new combined score. Low earnings yield & low momentum style (low E/P&MOM) is the 30% (quantiles 1-6) of stocks with the lowest score.

Figure 52 summarizes the diversification premium for benchmark portfolio as a function of investing style. Additionally, geometric risk premiums are shown for benchmark and single stock portfolios and standard deviation for the benchmark geometric risk premium is given. There is a dramatic difference in the diversification premium between the high and low exposures to tested investing styles. High earnings yield with high momentum (high E/P&MOM) requires the least diversification while low earnings yield (low E/P) requires the most. The difference is visible also in the average single stock risk premium which is positive for the high style exposures and very negative for the low exposures. This means that investing in a single high E/P&MOM style stock is expected to deliver a risk premium of 6.9% while investing in a single low E/P&MOM style stock is expected to deliver a risk premium of -21.4%. Style portfolio risk premium starts to increase from these figures as portfolios become more diversified achieving the benchmark risk premium (single stock risk premium + diversification premium) at full diversification. Similarly, as with the size deciles, we can see that there is some correlation between systematic variance (benchmark SD) and idiosyncratic variance (diversification premium). Also, the styles with high benchmark risk premium tend to require less diversification, which differs from the size decile test where high benchmark risk premium for the highest decile was associated with the greatest diversification premium.

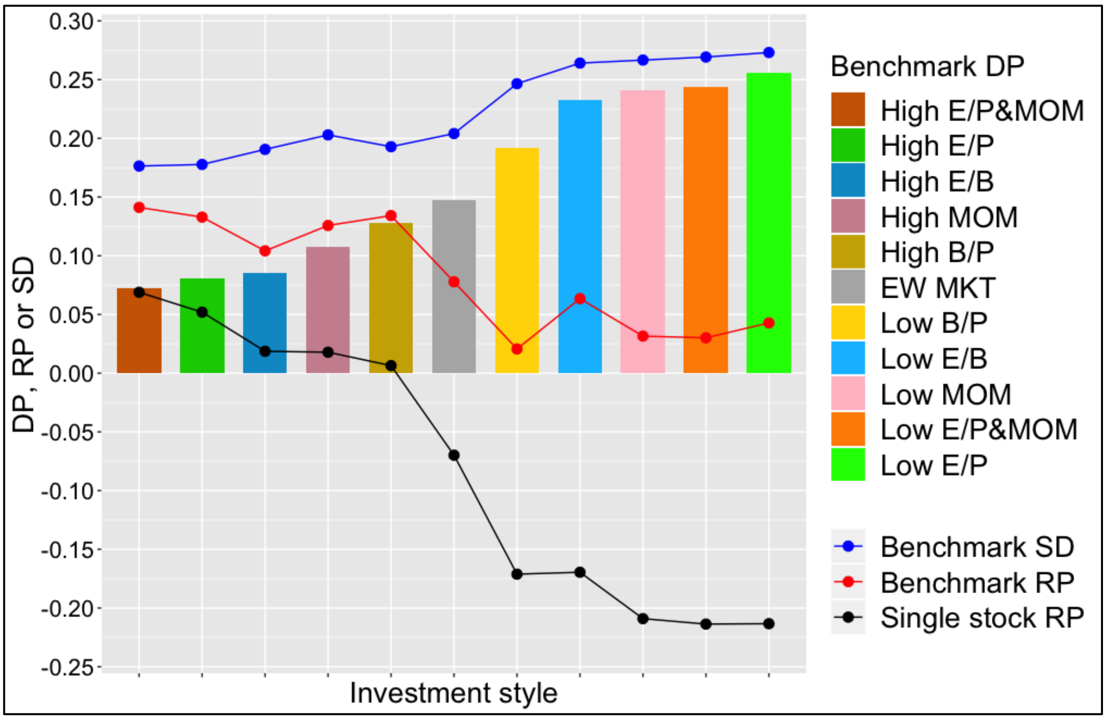

Figure 52. Diversification premium as a function of investment style.

Figure 53 and Figure 54 show the required level of diversification to achieve 90% of the maximum diversification benefit at investment fraction one for high earnings yield and low earnings yield styles, respectively. Low earnings yield style requires three to ten times more diversification depending on the diversification metric. Long-term risk averse wealth ratio (red dashed line) approaches its risk neutral counterpart (red solid line) very quickly and very slowly for high and low earnings yield styles, respectively. Long-term risk averse liability matching portfolio targeting investor who tolerates short-term risk cares mostly about the red dashed line. In case of high earnings yield style, the red dashed line in the long-term is very close to red solid line, the long-term risk neutral investor metric. For this particular investor type combined with this particular investing style, the risk seems to dissipate in the long-term.

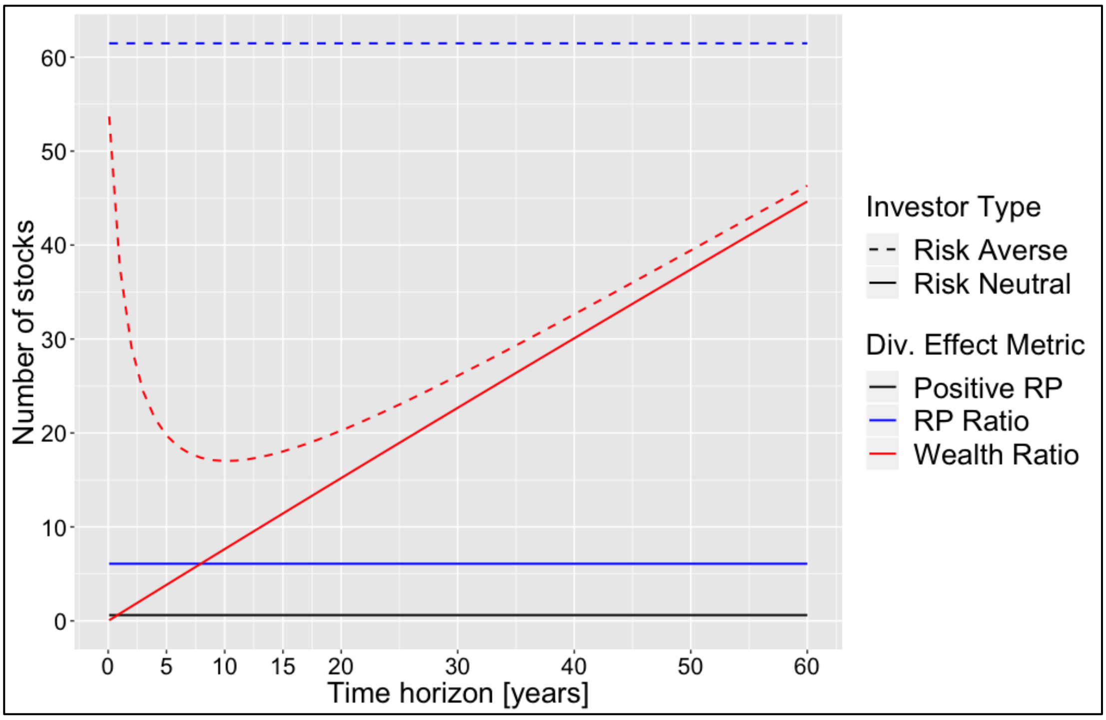

Figure 53. Required number of high E/P stocks for 90% diversification benefit when $f = 1$.

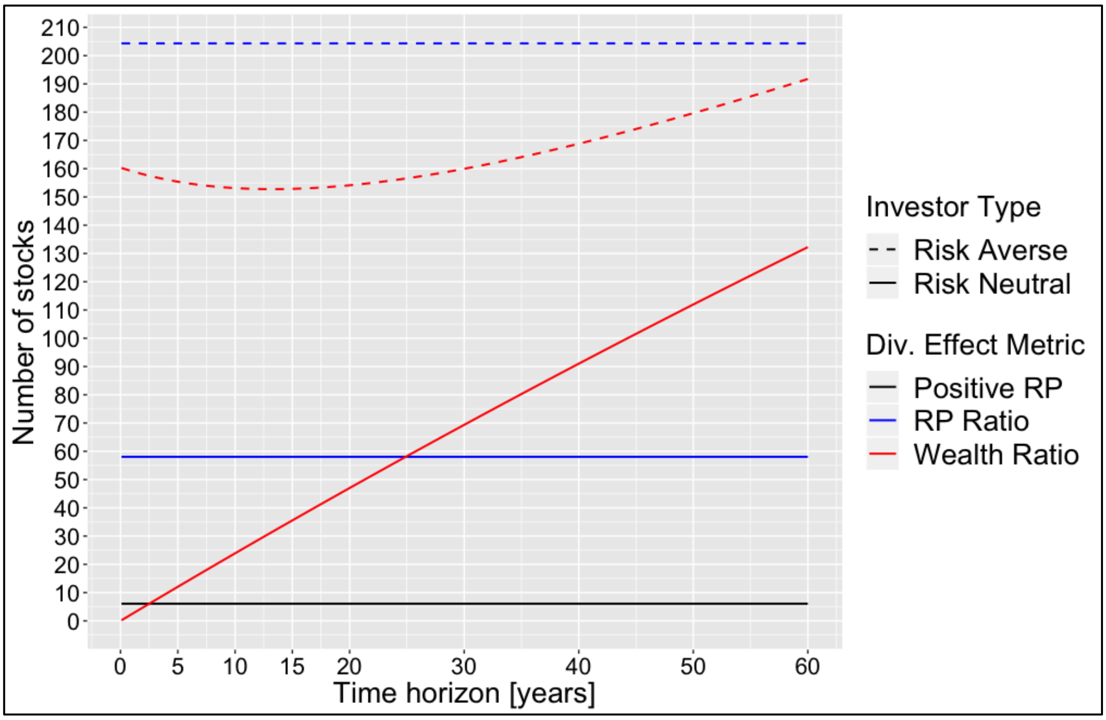

Figure 54. Required number of low E/P stocks for 90% diversification benefit when $f = 1$.

Table 20 and Table 21 summarize the short-term and long-term required number of stocks to achieve 90% of the maximum diversification benefit for each investing style at investment fraction one. We can see that the required level of diversification correlates strongly with the diversification premium. All of the metrics indicate much greater required level of diversification for the low exposures to selected investment styles compared to corresponding high exposures.

Table 20. Short-term required number of stocks per investing style for 90% diversification benefit at investment fraction one.

	Short-term diversification metric		
Investing style	Positive RP	Risk neutral RP ratio	Risk averse RP ratio
High E/P&MOM	0.51	5.1	54.8
High E/P	0.61	6.1	61.5
High E/B	0.82	8.2	63.5
High MOM	0.86	8.6	69.3
High B/P	0.95	9.5	84.7
EW MKT	1.91	19.0	118
Low B/P	9.61	90.0	272
Low E/B	3.70	35.9	149
Low MOM	7.62	72.5	238
Low E/P&MOM	8.15	77.0	246
Low E/P	6.05	58.0	204

Basically, selecting a style with low diversification requirement implies that stock picker is exposed to risk factors associated to the style already at low levels of diversification. On the other hand, selecting a style with high diversification requirement implies that stock picker is exposed relatively less to risk factors, but more to firm specific, idiosyncratic factors. For example, an investor investing to high earnings yield stocks very quickly gets exposed to the risks and rewards associated with that style whereas the success of an investor investing to low earnings yield stocks depends more on his firm selection than style selection at low levels of diversification.

Further study is required to better understand the drivers for the different diversification needs behind investing styles. For example, DuPont analysis could reveal in more detail why high ROE style and its derivatives such as high earning yield style require so little diversification compared to low ROE and low earnings yield styles.

Table 21. Long-term required number of stocks per investing style for 90% diversification benefit at investment fraction one.

	Long-term diversification metric [10/20/40/60 years]	
Investing style	Risk neutral wealth ratio	Risk averse wealth ratio
High E/P&MOM	6.8/13.6/26.9/40.0	14.3/17.6/28.9/41.3
High E/P	7.6/15.2/30.1/44.6	17.0/20.3/32.6/46.3
High E/B	8.1/16.0/31.7/46.9	23.4/24.8/36.3/50.0
High MOM	10.2/20.2/39.9/59.0	24.1/28.0/43.8/61.6
High B/P	12.0/23.8/46.9/69.2	26.5/31.7/50.8/71.8
EW MKT	14.0/27.9/55.5/82.8	56.8/55.2/71.0/93.6
Low B/P	18.0/35.5/69.2/101	212/216/226/239
Low E/B	21.7/42.8/82.8/120	100/100/119/145
Low MOM	22.5/44.3/86.1/125	183/186/198/217
Low E/P&MOM	22.8/44.8/86.8/126	190/194/206/224
Low E/P	23.9/47.0/91.0/132	153/154/169/192

We consider the results shown in section 5.8 as supportive to *hypothesis 7*.

Hypothesis 7: Number of stocks required to make a diversified portfolio is a function of investment style.

5.9 Evidence on the consistency of the historical diversification premium

We have shown how geometric risk premium can be decomposed into single stock risk premium and diversification premium. Furthermore, we have shown that the diversification premium varies substantially as a function of firm size and investing style. Diversification premium is an average over time. But how consistent diversification premium is as a function of time? The answer is very consistent.

Figure 55 shows the equally weighted market risk premium (blue line) decomposed to single stock risk premium (black line) and diversification premium (red line) as a function of time. Premiums are monthly values (not annualized). We get the blue line by summing black and red lines. What is immediately clear is that it is the single stock risk premium that is responsible for the large variance (risk) of the market risk premium. Diversification premium's contribution to the variance of the market risk premium is negligible. The time series variance of the diversification premium is very small compared to time series variance of risk premium. In other words, compared to risk premium, diversification premium is very consistent as a function of time. And unlike risk premium, diversification premium is always positive.

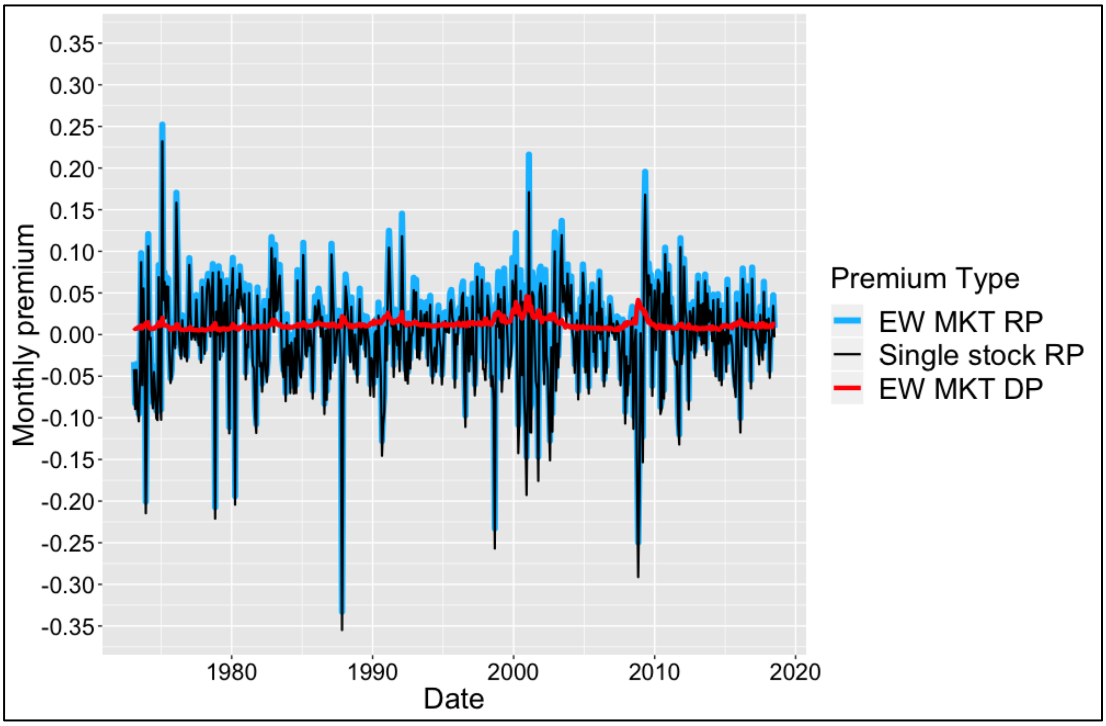

Figure 55. Decomposed EW market geometric risk premium.

Figure 56 allows us to have a more detailed picture of the same diversification premium as in Figure 55. In addition to diversification premium (red line), we show drawdowns (blue line) and bear markets (with black line) as a function of time. We consider bear market to begin when cumulative monthly equally weighted market return falls 20% from recent highs and to end when cumulative return rises 20% from the recent lows. Crossing the 20% mark must last at least three consecutive months. We can see that diversification premium is relatively low when the market is calm and drawdowns small. Around large drawdowns, diversification premium is relatively high. There appears to be a correlation such that diversification is most beneficial when uncertainty is high and the market is volatile.

We find that diversification works when it is the most needed meaning the bear markets. Diversification premium (annualized) is on average 19.4 percentage points compared to 13.3 percentage points during bear and bull markets, respectively. There are 129 and 417 bear and bull market months, respectively.

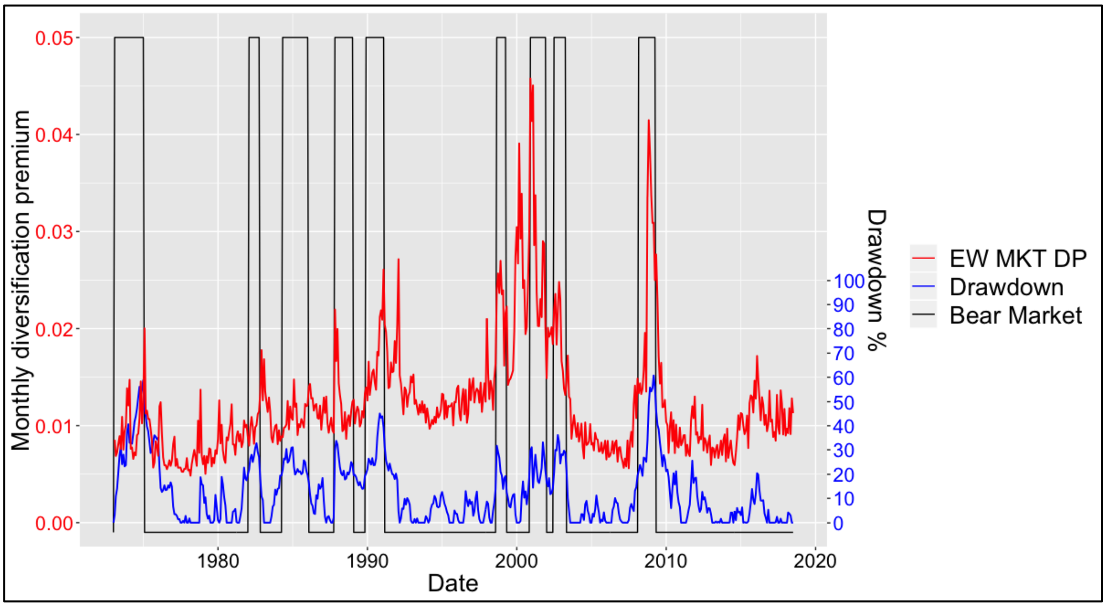

Figure 56. Association between diversification premium, drawdowns and bear markets.

Finally, Figure 57 compares diversification premium ratios for opposite investing styles as a function of time. Ratios are based on monthly values. We have shown in Figure 52 that diversification premium, calculated as an average over time, is much higher for low exposures to E/B (ROE), MOM (momentum) and B/P (value) compared to high exposures to corresponding factors. Diversification premium ratio has straightforward interpretation as telling how much more diversification is required for low end exposure compared to high end exposure for a selected style to keep the cost of foregone diversification equal. For example, based on the black line, monthly high ROE style requires roughly 1 to 8.5 times less diversification compared to low ROE style. Diversification premium ratios do vary as a function of time, but we can see that there are very few values below one. This means that the high exposure to a factor very consistently requires less diversification compared to low exposure to that particular factor.

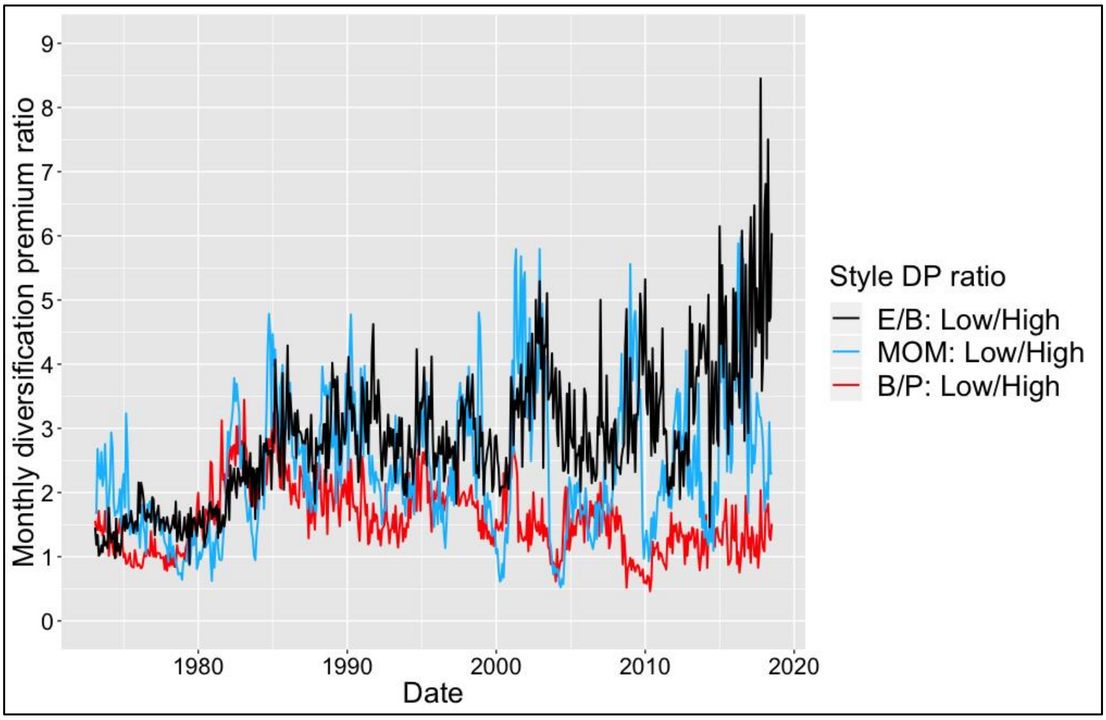

Figure 57. Diversification premium ratios within factors.

In Table 22 we compare how consistent the diversification premium difference is compared to risk premium difference between opposite styles within one factor. For example, it is well known that even though value has outperformed growth in the very long-run, there are long periods of time when growth has outperformed. Calculated based on monthly risk premium difference (value RP - growth RP), we find that value has outperformed on slightly more than six out of ten months. However, when comparing the diversification premium ratio, we find that value has required less diversification than growth on close to nine out of ten months. The difference in consistency is the most striking for high ROE versus low ROE style comparison where we find that high ROE has required less diversification in 544 out of 546 months. We conclude that the difference in diversification requirement has been far more consistent and reliable than the difference in risk premiums between investing styles.

Table 22. Comparing consistencies of risk premium difference and diversification premium ratio.

	Investing style		
Metric [#months]	ROE	Momentum	Value
RP difference > 0	58.1%	65.2%	61.7%
DP ratio > 1	99.6%	92.5%	86.6%

6 CONCLUSIONS

In this thesis we searched for the answer to question: *how many stocks make a diversified portfolio in a continuous-time world?* The question is thoroughly researched, but in the context of a theoretical single period model. Our contribution is to bring the question to the realm of continuous-time world which we show to fundamentally change the answer.

As the single period framework is inadequate in continuous-time setting, we derive a completely new, information theory based, theoretical framework to assess diversification effects in a continuous-time world. The fundamental difference between single period and continuous-time worlds is the use of arithmetic versus geometric mean returns, respectively. Academic world typically utilizes arithmetic returns, but the vast majority of investment practitioners care about geometric rates of returns making our results of interest to large audience.

Perhaps the most fundamental difference between single period and continuous-time worlds is that in a single period world there is no such concept as diversification benefit for risk neutral investor, but continuous-time world risk neutral investor benefits from diversification as geometric mean return is an increasing function of level of diversification. Consequently, we show that diversification in a continuous-time world is a negative price lunch as opposed to free lunch in a single period world.

Further fundamental differences arise when we explore the effect of asset allocation (or leverage) and investment time horizon length to required level of diversification. Diversification effect in single period model is indifferent to fraction of assets allocated to stocks and, by construction, does not consider the effect of time. In a continuous-time world, we show that asset allocation is a dominant determinant for the required level of diversification implying leveraged portfolios call for very wide diversification while 60/40 portfolio requires significantly less diversification than a portfolio with 100% stock allocation. For a long-term risk neutral investor, we use wealth ratio-based diversification metric and show that the required level of diversification increases directly proportionally to investment time horizon. Furthermore, different investment styles have vastly different requirements for the level of diversification. Small stocks, particularly microcaps, require more diversification than big stocks. High ROE, high momentum, value, high earnings yield and high earnings yield combined with high momentum require substantially less diversification compared to equally weighted market portfolio and particularly compared to opposite styles.

The conventional wisdom, based on the early empirical results utilizing single period model, states that no more than about ten stock are required to make a sufficiently diversified portfolio. We find that in a continuous-time world, risk averse investor with 100% stock allocation requires more than hundred stocks to exhaust 90% of the diversification benefit potential, while risk neutral investor requires about twenty stocks in the short-term and substantially more in the long-term.

In a single period world one stock portfolio is sufficiently diversified for risk neutral investor. In a continuous-time world rational risk neutral investor without stock picking skill will diversify to all stocks as maximum diversification implies maximum geometric risk premium. In this case the number of stocks that makes a sufficiently diversified portfolio is all stocks.

The number of stocks required to make a diversified portfolio is predicted by the developed theoretical framework using empirical parameters as inputs. The predictions are shown to be consistent and accurate, but with a slight tendency for underestimation. The underestimation is attributable to fat-tailedness of the monthly portfolio risk premium distributions combined with monthly rebalancing frequency. Fat-tailedness further increases the need for diversification, but also implies that we need to be careful and require very large sample sizes before drawing conclusions.

Investors can benefit from these results by better understanding the factors affecting the required level of diversification. Instead of one simple answer, our study provides the tools to assess the need for diversification based on investor’s individual circumstances and investing targets. Stock pickers, in particular, may benefit from understanding the differences in diversification benefits between investing styles.

Finally, our thesis offers one piece to equity premium puzzle. This piece is the diversification premium difference to benchmark, which is the difference in (geometric) risk premiums between a less than perfectly diversified portfolio and perfectly diversified benchmark portfolio. Before the easy and cheap access to diversification available today, investors were not able to diversify broadly because of the associated costs. The aggregate of less than perfectly diversified investors set the required level for risk premium. Historical theoretical risk premium for fully diversified benchmark, not realizable for less than perfectly diversified investors, therefore is higher (by the magnitude of average diversification premium difference to benchmark) than practical risk premium required by (and realized for) investors.

REFERENCES

Baz, J. & Guo, H. (2017). *An asset allocation primer: Connecting Markowitz, Kelly and risk parity*. (PIMCO quantitative research). Retrieved from https://global.pimco.com/en-gbl/insights/viewpoints/research/an-asset-allocation-primer-connecting-markowitz-kelly-and-risk-parity

Bernstein, W. J. & Wilkinson, D. J. (1997). *Diversification, rebalancing and the geometric efficient frontier* (SSRN Working Paper No. 53503). Retrieved from https://ssrn.com/abstract=53503

Bessembinder, H. (2018). Do stocks outperform treasury bills? *Journal of Financial Economics*, *129*(3), 440–457. doi:10.1016/j.jfineco.2018.06.004

Bessembinder, H., Chen, T-F., Choi, G., & Wei, K. C. J. (2019). *Do global stocks outperform US treasury bills?* (SSRN Working Paper No. 3415739). Retrieved from https://ssrn.com/abstract=3415739

Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities. *The Journal of Political Economy*, *81*(3), 637–654. doi:10.1086/260062

Brandt, M. W., Santa-Clara, P., & Valkanov, R. (2009). Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns. *The Review of Financial Studies*, *22*(9), 3411–3447. doi:10.2139/ssrn.661343

Brooks, C. (2014). *Introductory econometrics for finance* (3rd ed). Cambridge, United Kingdom: Cambridge University Press.

Campbell, J. Y., Lettau, M., Burton, G. M. & Xu, Y. (2001). Have individual stocks become more volatile? An empirical exploration of idiosyncratic risk. *Journal of Finance*, *56*(1), 1–43. doi:10.1111/0022-1082.00318

Chen, S. *ELEC3203 Digital Coding and Transmission – Overview & Information Theory* (Lecture 6). Retrieved from http://www.southampton.ac.uk/~sqc/ELEC3203/InfThe-L6.pdf

Cover, T. M. & Thomas, J. (2005). *Elements of information theory* (2nd ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.

Dimson, E., Marsh, P. & Staunton, M. (2003). Global evidence on the equity risk premium. *Journal of Applied Corporate Finance, 15*(4), 27–38. doi:10.1111/j.1745-6622.2003.tb00524.x

Dimson, E., Marsh, P. & Staunton, M. (2011). *Equity premia around the world*. (SSRN Electronic Journal). Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1940165. doi:10.2139/ssrn.1940165

Domian, D. L., Louton, D. A. & Racine, M. D. (2007). Diversification in portfolios of individual stocks: 100 stocks are not enough. *Financial Review*, *42*(4), 557–570. doi:10.1111/j.1540-6288.2007.00183.x

Elton, E. J. & Gruber, M. J. (1977). Risk reduction and portfolio size: An analytical solution. *Journal of Business*, *50*(4), 415–437. doi:10.1086/295964

Erb, C. B., Harvey, C. R. (2006). The strategic and tactical value of commodity futures. *Financial Analyst Journal*, *62*(2), 69–97. doi:10.2469/faj.v62.n2.4084

Evans, J. L. & Archer, S. H. (1968). Diversification and the reduction of dispersion: An empirical analysis. *Journal of Finance*, *23*(5), 761–767. doi:10.1111/j.1540-6261.1968.tb00315.x

Fama, E. F. & French, K. R. (2008). Dissecting anomalies. *The Journal of Finance*, *63*(4), 1653–1678. doi:10.1111/j.1540-6261.2008.01371.x

Fisher, L. & Lorie, J. H. (1970). Some studies of variability of returns on investments in common stocks. *Journal of Business*, *43*(2), 99–134. doi:10.1086/295259

Gabaix, X., Gopikrishnan, P., Plerou, V. & Stanley, HE. (2003). A theory of power law distributions in financial market fluctuations. *Nature*, *423*(6937), 267–270. doi:10.1038/nature01624

Goetzmann, W. N. & Kumar, A. (2008). Equity portfolio diversification. *Review of Finance, 12*(3), 433–463. doi:10.1093/rof/rfn005

Hudson, R. S. & Gregoriou, A. (2015). Calculating and comparing security returns is harder than you think: A comparison between logarithmic and simple returns. *International Review of Financial Analysis*, *38*(C), 151–162. doi:10.1016/j.irfa.2014.10.008

Hull, J. C. (2015). *Options, futures and other derivatives* (9th ed, Global ed). Harlow, England: Pearson Education Limited.

Jaconetti, C. M., Kinniry, F. M. & Zilbering, Y. (2015). *Best practices for portfolio rebalancing* (Vanguard Research). Retrieved from https://www.vanguard.com/pdf/ISGPORE.pdf

Jacquier, E., Kane, A. & Marcus, A. J. (2003). Geometric or arithmetic mean: A reconsideration. *Financial Analysts Journal*, *59*(6), 46–53. doi:10.2469/faj.v59.n6.2574

Kelly, J. L. (1956). A new interpretation of information rate. *Bell System Technical Journal*, *35*(4), 917–926. doi:10.1002/j.1538-7305.1956.tb03809.x

Kritzman, M. (1994). What practitioners need to know about time diversification. *Financial Analyst Journal*, *50*(1), 14–18. doi:10.2469/faj.v50.n1.14

Kritzman, M. & Rich, D. R. (1998). Beware of Dogma. *Journal of Portfolio Management*, *24*(4), 66–77. doi:10.3905/jpm.1998.409647

Kuhn, D. & Luenberger, D. G. (2010). Analysis of the rebalancing frequency in log-optimal portfolio selection. *Quantitative Finance*, *10*(2), 221–234. doi:10.1080/14697680802629400

Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance*, *7*(1), 77–91. doi:10.1111/j.1540-6261.1952.tb01525.x

Markowitz, H. M. (1959). *Portfolio selection*. New York: John Wiley & Sons.

Markowitz, H. M. (1976). Investment for the long run: New evidence for an old rule. *The Journal of Finance*, *31*(5), 1273–1286. doi:10.1111/j.1540-6261.1976.tb03213.x

Newbould, G. D. & Poon, P. S. (1993). The minimum number of stocks needed for diversification. *Financial Practice and Education*, *3*(2), 85–87.

Peters, O. (2019). The ergodicity problem in economics. *Nature Physics*, *15*(12), 1216–1221. doi:10.1038/s41567-019-0732-0

Polkovnichenko, V. (2005). Household portfolio diversification: A case for rank-dependent preferences. *The Review of Financial Studies, 18*(4), 1467–1502. doi:10.1093/rfs/hhi033

Proakis, J. G. (1995). *Digital communications* (3rd ed. Global International Editions 1995). Singapore: McGraw-Hill Book Co.

Samuelson, P. A. (1971). The "fallacy" of maximizing the geometric mean in long sequences of investing or gambling. *Proceedings of the natural academy of sciences*, *68*(10), 2493–2496. doi:10.1073/pnas.68.10.2493

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, *19*(3), 425–442. doi:10.1111/j.1540-6261.1964.tb02865.x

Sharpe, W. F. (1966). Mutual Fund Performance. *The Journal of Business*, *39*(1), 119–138.

Sharpe, W. F. (1994). The Sharpe ratio. *Journal of Portfolio Management*, *21*(1), 49–58. doi:10.3905/jpm.1994.409501

Statman, M. (1987). How many stocks make a diversified portfolio? *Journal of Financial and Quantitative Analysis*, *22*(3), 353–363. doi:10.2307/2330969

Statman, M. (2004). The diversification puzzle. *Financial Analyst Journal*, *60*(4), 44–53. doi:10.2469/faj.v60.n4.2636

Surz, R. J. & Price, M. (2000). *The truth about diversification by the numbers*. Retrieved from: http://ppca-inc.com/Articles/DiversByNumbers.pdf

Taleb, N. N. (2020). *Statistical consequences of fat tails: Real world preasymptotics, epistemology, and applications*. STEM Academic Press.

Taleb, N. N., Bar-Yam, Y. & Cirillo, P. (2020). *On single point forecasts for fat-tailed variables*. Retrieved from: https://arxiv.org/pdf/2007.16096.pdf

Thorp, E. O. (1971). Portfolio choice and the Kelly criterion. *Proceedings of the Business and Economic Section of the American Statistical Association,* 215–224.

Thorp, E. O. (2006). The Kelly criterion in blackjack sports betting, and the stock market. In S. A. Zenios & W. T. Ziemba (Eds.), *Handbook of asset and liability management, volume 1* (pp. 385-428). Amsterdam, The Netherlands: Elsevier, North-Holland.

Thorp, E. O. (2008, May and September). Understanding the Kelly criterion. *Wilmott magazine,* columns from the series *A mathematician on Wall Street*.

Tidmore, C., Kinniry, F. M., Renzi-Ricci, G. & Cilla, E. (2019). *How to increase the odds of owning the few stocks that drive returns*. Retrieved from: https://www.vanguardinvestments.dk/documents/institutional/increase-odds-of-owning-less-stock-that-drive-returns.pdf

Tobin, J. (1958). Liquidity preference as behavior towards risk. *The Review of Economic Studies*, *25*(2), 65–86. doi:10.2307/2296205

Treynor, J. L. & Fisher, B. (1973). How to use security analysis to improve portfolio selection. *The Journal of Business*, *46*(1), 66–86. doi:10.1086/295508

Tse, D. & Viswanath, P. (2005). *Fundamentals of Wireless Communication*. Cambridge University Press. Retrieved from: https://web.stanford.edu/~dntse/wireless_book.html

Appendix 1

Derivation of the instantaneous geometric risk premium using Euler's number identity approach

We continue the derivation from equation (19):

$$g^e(f) = 0.5\begin{pmatrix} \sum_{i=1}^{n} \ln\left[1 + f\left(\frac{m_{e,i}}{n} + \frac{s_{e,i}}{\sqrt{n}}\right)\right] + \\ \sum_{i=1}^{n} \ln\left[1 + f\left(\frac{m_{e,i}}{n} - \frac{s_{e,i}}{\sqrt{n}}\right)\right] \end{pmatrix}$$

$$\begin{aligned} g^e(f) &= \frac{1}{2}\sum_{i=1}^{n}\left(\ln\left[1 + f\left(\frac{m_{e,i}}{n} + \frac{s_{e,i}}{\sqrt{n}}\right)\right] + \ln\left[1 + f\left(\frac{m_{e,i}}{n} - \frac{s_{e,i}}{\sqrt{n}}\right)\right]\right) \\ &= \frac{n}{2}\left(\ln\left[1 + f\left(\frac{m_e}{n} + \frac{s_e}{\sqrt{n}}\right)\right] + \ln\left[1 + f\left(\frac{m_e}{n} - \frac{s_e}{\sqrt{n}}\right)\right]\right) \\ &= \frac{n}{2}\ln\left(\left[1 + f\left(\frac{m_e}{n} + \frac{s_e}{\sqrt{n}}\right)\right]\left[1 + f\left(\frac{m_e}{n} - \frac{s_e}{\sqrt{n}}\right)\right]\right) \\ &= \frac{n}{2}\ln[b(f)c(f)] \end{aligned}$$

$$\begin{aligned} b(f)c(f) &= 1 + f\frac{m_e}{n} - f\frac{s_e}{\sqrt{n}} + f\frac{m_e}{n} + f^2\frac{{m_e}^2}{n^2} - f^2 s_e\frac{m_e}{n^{3/2}} + f\frac{s_e}{\sqrt{n}} \\ &\quad + f^2 s_e\frac{m_e}{n^{3/2}} - f^2\frac{{s_e}^2}{n} \\ &= 1 + 2f\frac{m_e}{n} + f^2\frac{{m_e}^2}{n^2} - f^2\frac{{s_e}^2}{n} \\ &= 1 + \frac{1}{n}\left[2fm_e + f^2\left(\frac{{m_e}^2}{n} - {s_e}^2\right)\right] \\ &= 1 + \frac{1}{n}a(f) \end{aligned}$$

$$a(f) = 2fm_e + f^2\left(\frac{{m_e}^2}{n} - {s_e}^2\right)$$

$$g^e(f) = \frac{n}{2}\ln[b(f)c(f)] = \frac{n}{2}\ln\left[1 + \frac{1}{n}a(f)\right]$$

$$= \frac{1}{2} a(f) \frac{n}{a(f)} \ln \left[1 + \frac{a(f)}{n} \right]$$

$$= \frac{a(f)}{2} \ln \left(\left[1 + \frac{a(f)}{n} \right]^{\frac{n}{a(f)}} \right)$$

We can substitute:

$$x = \frac{n}{a(f)}$$

And have:

$$g^e(f) = \frac{a(f)}{2} \ln \left(\left[1 + \frac{1}{x} \right]^x \right)$$

By utilizing Euler's number identity given in equation (15), when $n \to \infty$ implying $x \to \infty$, we can write:

$$\lim_{n \to \infty} g^e(f) = g^e_\infty = \frac{\lim_{n \to \infty} a(f)}{2} \lim_{n \to \infty} \ln \left(\left[1 + \frac{a(f)}{n} \right]^{\frac{n}{a(f)}} \right)$$

$$= \frac{\lim_{n \to \infty} a(f)}{2} \lim_{x \to \infty} \ln \left(\left[1 + \frac{1}{x} \right]^x \right) = \frac{\lim_{n \to \infty} a(f)}{2} \ln(e)$$

$$= \frac{2fm_e - f^2 {s_e}^2}{2} = fm_e - \frac{f^2 {s_e}^2}{2}$$

Appendix 2

Derivation of diversification premium in the presence of uncertainty about idiosyncratic risk

We denote $\widehat{Sdev}_{BM}$, $\widehat{Sdev}_P$ and $\widehat{Sdev}_{n=f=1}$ as the parameter estimator for the standard deviation of instantaneous excess growth rate of a benchmark, portfolio and single stock portfolio with 100% allocation to stocks respectively. In addition, we denote $\widehat{Isdev}_P$ and $\widehat{Isdev}_{n=f=1}$ as the parameter estimator for the idiosyncratic standard deviation of instantaneous excess growth rate of a portfolio and single stock portfolio with 100% allocation to stocks respectively.

$$DP^P = \frac{f^2}{2}\left(\left[Var_{n=f=1} + Var\left(\widehat{Sdev}_{n=f=1}\right)\right] - \left[Var_P + Var\left(\widehat{Sdev}_P\right)\right]\right)$$

Assumption: zero correlation between $\widehat{Sdev}_{BM}$ and $\widehat{Isdev}_P$.

$$\begin{aligned} DP^P &= \frac{f^2}{2}\left(\left[Var_{BM} + Ivar_{n=f=1} + Var\left(\widehat{Sdev}_{BM}\right) + Var\left(\widehat{Isdev}_{n=f=1}\right)\right]\right. \\ &\qquad \left. - \left[Var_{BM} + Ivar_P + Var\left(\widehat{Sdev}_{BM}\right) + Var\left(\widehat{Isdev}_P\right)\right]\right) \\ &= \frac{f^2}{2}\left(\left[Ivar_{n=f=1} + Var\left(\widehat{Isdev}_{n=f=1}\right)\right] - \left[Ivar_P + Var\left(\widehat{Isdev}_P\right)\right]\right) \end{aligned}$$

From equation (33) we have:

$$Ivar_P = \frac{Ivar_{n=f=1}}{n_P} - \frac{Ivar_{n=f=1}}{n_{BM}} = \left(\frac{1}{n_P} - \frac{1}{n_{BM}}\right) Ivar_{n=f=1}$$

Similarly, we can derive:

$$\widehat{Ivar}_P = \left(\frac{1}{n_P} - \frac{1}{n_{BM}}\right)\widehat{Ivar}_{n=f=1} \Rightarrow \widehat{Isdev}_P = \left(\frac{1}{\sqrt{n_P}} - \frac{1}{\sqrt{n_{BM}}}\right)\widehat{Isdev}_{n=f=1}$$

$$DP^P = \frac{f^2}{2}\Bigg[\Big(Ivar_{n=f=1} + Var\big(\widehat{Isdev}_{n=f=1}\big)\Big) - \left(\left[\frac{1}{n_P} - \frac{1}{n_{BM}}\right] Ivar_{n=f=1} + Var\left[\left(\frac{1}{\sqrt{n_P}} - \frac{1}{\sqrt{n_{BM}}}\right)\widehat{Isdev}_{n=f=1}\right]\right)\Bigg]$$

$$Var\left[\left(\frac{1}{\sqrt{n_P}} - \frac{1}{\sqrt{n_{BM}}}\right)\widehat{Isdev}_{n=f=1}\right] = \left(\frac{1}{\sqrt{n_P}} - \frac{1}{\sqrt{n_{BM}}}\right)^2 Var\big(\widehat{Isdev}_{n=f=1}\big)$$

$$DP^P = \frac{f^2}{2}\left[\left(1 - \frac{1}{n_P} + \frac{1}{n_{BM}}\right) Ivar_{n=f=1} + \left[1 - \left(\frac{1}{\sqrt{n_P}} - \frac{1}{\sqrt{n_{BM}}}\right)^2\right] Var\big(\widehat{Isdev}_{n=f=1}\big)\right]$$

$$= \frac{f^2}{2}\left[\left(1 - \frac{1}{n_P} + \frac{1}{n_{BM}}\right) Ivar_{n=f=1} + \left[1 - \frac{1}{n_P} - \frac{1}{n_{BM}} + \frac{2}{\sqrt{n_P n_{BM}}}\right] Var\big(\widehat{Isdev}_{n=f=1}\big)\right]$$