Spring 5-15-2018

On the Wedge between Theoretical and Actual Prices and its Implications for Investment Decisions

Luca Pezzo
Washington University in St. Louis

Washington University in St. Louis

Olin Business School

Dissertation Examination Committee:
Phillip, H. Dybvig, Chair
Guofu Zhou
Ohad Kadan
Thomas, A. Maurer
John Nachbar

On the Wedge between Theoretical and Actual Prices and its Implications for Investment Decisions

by

Luca Pezzo

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

May 2018
Saint Louis, Missouri

Contents

List of Tables

List of Figures

Acknowledgments

A special thanks goes to the graduate students and distinguished faculty within my department who have reviewed this dissertation and helped support the related research. My committee has been especially helpful in guiding the direction of the research contained within and I am eternally thankful.

Luca Pezzo

Washington University in Saint Louis

May 2018

I dedicate this dissertation to my family, friends, and colleagues. A special thank you to my wife, Elettra, for the continued support through my graduate school experience, and my son, Tommaso, for the extra boost of energy required to complete my Ph.D.

ABSTRACT OF THE DISSERTATION

On the Wedge between Theoretical and Actual Prices and its Implications for Investment Decisions

by

Luca Pezzo

Doctor of Philosophy in Finance

Washington University in St. Louis, May 2018

Research Advisor: Professor Phillip, H. Dybvig

State of the arts equilibrium models explain several financial markets' regularities but still miss many important dimensions. My research investigates the existing wedge between theoretical and actual prices and its implications for investment decisions. In the first chapter, I develop a new approach to locate and quantify the wedge between the main-stream Representative Agent pricing of the U.S. market portfolio and actual data. The determinants of the wedge are high uncertain and illiquid recessionary periods where, according to the marginal pricing rules, more efficient portfolios than the market can be formed. Since illiquidity is a major determinant, chapter two and three are devoted to the theoretical and empirical study of the impact of transaction costs on the optimal formation of equilibrium portfolios. Chapter two develops a single-period Mean-Variance theory able to solve large scale portfolio optimization problems in the presence of fixed and variable costs. Chapter three shows its relevance in the representative context of the FX markets.

Chapter 1

A Non-Parametric Test For Representative Agent Pricing

1.1 Introduction

Since the seminal work of Markowitz (1952, 1959) and Black (1972), which have laid the foundation of the CAPM, many asset pricing models in finance assume a representative agent,[1] a hypothetical unconstrained investor who holds the market portfolio. Any asset in the economy is therefore proportional to the ratio of his marginal utilities, and, as a consequence, most asset pricing tests are based on strong parametric assumptions on the agent preferences and the returns.

I make use of a result from Martin (2017) to deliver a more general test for representative agent pricing which compares the realized excess market returns with an option-implied bound on their one-period ahead risk premium. The test does not require sharp assumptions on preferences and returns and jointly applies to a non-trivial class of models (including those

[1]Recent leading examples are: the consumption ICAPM of Campbell and Viceria (1999), the external habit model of Campbell and Cochrane (1999), the long run risk models of Bansal and Yaron (2004), Bansal, Kiku, Shaliastovich, and Yaron (2014) and Campbell, Giglio, Polk, and Turley (2017), as well as the rare disaster models of Barro (2006) and J. A. Wachter (2013).

based on unobservable state variables[2]). In contrast to the standard GMM tests,[3] mine is non-parametric and only requires the time series of a proxy for the market portfolio, quotes of European puts and calls written on it, and a proxy for the risk-free rate. In particular, because no Stochastic Discount Factor (SDF hereafter) is estimated, the test avoids the usual strict point-wise restrictions on the functional form of the preferences of the representative agent. Rather, it only[4] assumes the covariance of the product of the SDF and the market return with the market return to be non-negative.

Popular representative agent pricing,[5] while holding unconditionally, is shown to be robustly rejected conditioning on highly uncertain and illiquid subperiods, which contain all the major financial crises and economic recessions in the analyzed sample. These subperiods, endogenously selected out-of-sample based on rules predicting low returns in a training sample, are defined as times where implied model-based risk premia are too high. Findings suggest that such conditional model-based implied risk premia are off by striking amounts - at least by 1.3% monthly (or 15.6% annualy) - even after controlling for risk (model based Sharpe ratios are still higher then actual ones by at least a monthly 0.2). While excessive risk aversion does not seem to explain such results, ruling out Merton (1980)'s type explanations[6], alternative explanations are found consistent with the data: either rejected models are too sensitive to market crash probabilities, which, as in the rare disaster literature (see for example Barro (2006) and J. A. Wachter (2013)), might be pushing the risk premia too high, or rejected

[2]E.g. in the rare disasters models of Barro (2006) and J. A. Wachter (2013) the probability of a rare disaster is a state variable. Because such probability is unobservable, it is difficult to test these models following the existing approaches.

[3]E.g. Hansen and Singleton (1982, 1983), Gallant and Tauchen (1989), Epstein and Zin (1991), Savov (2011) and Nagel and Singleton (2011).

[4]"I am not aware of any model that attempts to match the data quantitatively in which [this condition] does not hold. " (Martin (2017))

[5]Including those in footnote 1, for a more general description of the class refer to Section 1.2

[6]In these models the risk premium is directly proportional to the level of risk aversion.

models, typically frictionless, might not be able to account for the high informational and trading frictions characterizing the rejections.

Interestingly, the alternative marginal intermediary-based pricing setup is found more robust. In these models pricing is performed via a marginal intermediary, not necessarily a representative agent.[7] The representative model of Adrian, Etula, and Muir (2014) is found to correctly price assets even during periods where representative pricing is rejected and the marginal broker/dealer intermediary not to hold the market portfolio in equilibrium: the correlation of the equilibrium portfolio and the Standard and Poor's 500 index ($SP500$ hereafter) is around 0.20 and not even significant during representative agent pricing rejections.

Overall, these results suggest that marginal pricing generates more efficient portfolios then the representative agent market portfolio:[8] this is a key economic implication of this study and follows from the fact that during rejections there is no representative agent who holds the market in equilibrium but there are at least some marginal broker/dealers, which consistently with the data, optimally holds different portfolios. Another contribution of this paper is to provide a unified formal framework to assess the performance of consumption-based representative agent pricing in the literature, extending and complementing the existing critiques. Rejections are indeed found to be periods where actual leading frameworks perform the worst[9] and feature the worrisome properties already documented by the extant empirical literature. In particular, Muir (2017) shows how consumption based representative agent pricing has difficulty in simultaneously matching risk premia during recessions and financial crises: this is because, despite the different risk premia behavior, the consumption dynamics

[7]The main difference being that a marginal agent is not required to hold the entire market portfolio and his identity might vary over time.

[8]A portfolio is efficient if held in equilibrium (see for example Dybvig and Ross (1982)).

[9]Following the original calibrations, the Campbell and Cochrane (1999), the Bansal and Yaron (2004), and the J. A. Wachter (2013) equilibrium models performances are compared in rejection periods with respect to the rest of the sample.

is very similar. While Muir (2017) selects these periods exogenously and ex-post, my test pins them down endogenously and ex-ante as part of the rejection subsample. Moreira and Muir (2017) find profitable trading strategies which go against representative agent pricing: because actual Sharpe-ratios are lower during recessions (periods of high volatility) and higher during normal times (periods of low volatility), it is profitable to time the market by decreasing the exposure in bad times and increasing it in normal times. However, a representative agent is expected to bare more risk rather then less during bad times and be compensated accordingly. Consistently, rejection periods are characterized by higher volatility, lower Moreira-Muir exposure and Sharpe ratios implied by representative agents are too high. Finally, Amromin and Sharpe (2013) and Greenwood and Shleifer (2014), contrary to rational expectations, show how representative-agent based required market returns disagree with actual expectations of a non-trivial fraction of investors. Their results are further exacerbated while conditional on the rejection subsample.

The main technical innovation of this study is the ability of my test to endogenously deliver, upon rejection, the actual subsample originating it. Identifying periods related to the systematic failure of a non-trivial class of models, constructively enables researchers to study its characteristics and potentially design more robust models in the future. From a methodological point of view, the paper closest to mine is Nagel and Singleton (2011). As in my framework, it also provide endogenous conditioning: in a very elegant way, they optimally select a combination of a pre-determined set of GMM instruments to maximize the power of their test. The cost they have to pay for such elegance is the fact that their test only applies to pre-specified null and alternative nested linear models. In contrast, in my framework I do not have a conditioning rule intrinsically tied to the test properties but my test jointly applies to any model in which the covariance of the product of the SDF and the market return with the market return is non-negative.

The rest of this paper is structured as follows: Section 1.2 describes the empirical setup, going trhough the logic behind the non-parametric test and how to implement it. Section 1.3 describes the data used in this study. Section 1.4 shows the outcomes from the non-parametric test and describes the characteristics of the detected rejections. Section 1.5 contains the main implications: it explains how to interpret the results, what do they mean for representative agent pricing, and shows how the marginal intermediary-based pricing framework might be more efficient. Section 2.6 concludes.

1.2 Empirical Setup

1.2.1 Logic of the non-parametric test

In this subsection I first present the result from Martin (2017) and then explain how to use it to derive the non-parametric test of this study.

Let us define the gross market and risk-free returns relative to $[t : t + 1]$ as R_{t+1} and $R_{t,f}$; following Martin (2017)

Proposition 1 *Given a strictly positive SDF M_{t+1} satisfying the pricing equation*

$$E_t[M_{t+1}R_{t+1}] = 1 \tag{1.1}$$

and the Negative Covariance Condition (NCC)

$$Cov_t(M_{t+1} \times R_{t+1}, R_{t+1}) \leq 0 \tag{1.2}$$

it is possible to construct a model-free real-time lower bound, LB_t, on the market risk premium $\mathbb{E}_t[\pi_{t+1}] \equiv \mathbb{E}_t[R_{t+1} - R_{t,f}]$ by

$$LB_t = 2\left(\frac{1}{\hat{S}_t}\right)^2 \left(\int_0^{\hat{F}_t} \hat{put}_t(k)dk + \int_{\hat{F}_t}^{\infty} \hat{call}_t(k) \right) \geq 0 \tag{1.3}$$

Proof. See Appendix A.1 ■

Quantities with hats are *ex-dividend*,[10] $\hat{S}_t$ is the closing market level at time t, $\hat{F}_t$ is the forward contract on the market with unity tenor and finally $\hat{put}_t(k)$ and $\hat{call}_t(k)$ are European put and call option quotes on the market with unity tenor as a function of the common strike k. By the Put-Call parity the forward contract $\hat{F}_t \equiv \hat{F}_t(k^*)$ is the unique point $(k^*, \hat{F}_t(k^*))$ at which the call and put functions intersect so that LB_t is just a function of $\hat{S}_t, \{\hat{put}_t(k_i), \hat{call}_t(k_i)\}_{k_i \in \mathcal{K}_t}$ where $\mathcal{K}_t$ is the set of observable strikes with unit tenor at time t over which the integrals have to be approximated. The most direct interpretation of the lower bound quantity obtains when the NCC equals zero: in this case LB_t measures the market risk premium itself from the perspective of a representative agent with log-utility assuming independent market returns over time.[11]

Note that in representative agent models M_{t+1} is the ratio of the agent marginal utilities and it is strictly positive by the non-satiation requirement and (1.1) arises as part of the first order conditions, then we can interpret Proposition 1 as follows:

[10] It is possible to construct a similar lower bound which is an explicit function of market dividends (this more general task is shown in the proof of Proposition 1), nonetheless an unreported analysis (available upon request) shows how the empirical role of dividends is negligible. Therefore to avoid unnecessary complications I will stick to the baseline lower bound measure (which is the same used in Martin (2017)).

[11] See Martin (2017) Section III.

Proposition 2 *Given any representative agent model with preferences satisfying NCC there exist a model-free real-time lower bound LB_t computable via eq. (1.3)*

I will refer to the non-trivial class of models such that the NCC holds as the Martin's class: "I am not aware of any model that attempts to match the data quantitatively in which the NCC does not hold", Martin (2017). In terms of actual models it includes the leading macro-finance frameworks in Campbell and Cochrane (1999), Bansal and Yaron (2004), Bansal et al. (2014), Campbell et al. (2017), Barro (2006), and J. A. Wachter (2013). More generally it at least[12] contains any model where the representative agent preferences are:

- strictly increasing: i.e. the non-satiation requirement
- if time-separable have Relative Risk Aversion (RRA) of at least 1 at any level of wealth
- if Epstein and Zin (1989) have RRA as well as Intertemporal Elasticity of Substitution (IES) of at least 1 at any level of wealth

if the model setting is dynamic an additional requirement is also needed constraining the market return R_{t+1} to be positively associated[13] will all the other state variables.

The logic of the non-parametric test follows immediately from Proposition 2: a lower bound violation implies a joint violation of all representative agent models with preferences satisfying the NCC: i.e. a test for a lower bound violation is a non-parametric test for the Martin's class of representative agent pricing.

[12] This is because the following conditions are only sufficient.

[13] An extension of the concept of pairwise correlation to multivariate possibly non-normal settings, see footnote 9 in Martin (2017).

1.2.2 Empirical Design

In order to make the non-parametric test operational a definition of lower bound violation is needed and provided next. Based on the premises that the sample average of the excess market returns is the sample counter part of the risk premium, I define a lower bound violation as follows

Definition 1 *A lower bound violation is a subsample, an indicator function I_t^v turning 1, where the excess market return, $\pi_{t+1} \equiv R_{t+1} - R_{t,f}$, is below the lower bound, LB_t, on average*

A non-parametric test for the Martin's class of representative agent pricing is then naturally a one-sided t-test against the alternative of a lower bound violation

$$H_0 : \mathbb{E}[\pi_{t+1}|I_t^v] \geq \mathbb{E}[LB_t|I_t^v] \text{ vs. } H_1 : \mathbb{E}[\pi_{t+1}|I_t^v] < \mathbb{E}[LB_t|I_t^v]$$

Note that a lower bound violation (a sample statement) is implied by the the test alternative (a population statement) given the processes for π_{t+1} and LB_t are covariance-stationary. It is useful to define a new variable $y_{t+1} \equiv \pi_{t+1} - LB_t$ and re-write the non-parametric test more compactly as

$$H_0 : \mathbb{E}[y_{t+1}|I_t^v] \geq 0 \text{ vs. } H_1 : \mathbb{E}[y_{t+1}|I_t^v] < 0 \tag{1.4}$$

The test simply looks at the time-series of y_{t+1} in the periods selected by the subsample I_t^v and then test if its conditional mean is non-negative, given a lower bound violation is now expressed in population terms as $\mathbb{E}[y_{t+1}|I_t^v] < 0$. Finally, exploiting the properties of

conditional expectations,[14] it is instructive to re-write (1.4) in its equivalent form

$$H_0 : \mathbb{E}[y_{t+1} \times I_t^v] \geq 0 \text{ vs. } H_1 : \mathbb{E}[y_{t+1} \times I_t^v] < 0 \tag{1.5}$$

This is because equation (1.5) shows how the non-parametric test is nothing more than a standard unconditional t-test on the random variable $(y_{t+1} \times I_t^v)$, so that normal inference applies. The non-parametric test of this study will be conducted with respect to both the equivalent statements (1.5) and (1.5) using both Newey and West (1987) heteroskedasticity and autocorrelation adjusted standard errors and small-sample bootstrapped standard errors.

The non-parametric test requires a couple of assumptions which are now introduced and discussed: HP1 - regularity conditions, HP2 - objective rule to select I_t^v.

HP1: regularity conditions

Given any conditioning set I_t^v, HP1 requires the Central Limit Theorem (CLT) to hold so that a proper limiting normal distribution for $\mathbb{E}[y_{t+1}|I_t^v]$ exists. The weakest assumptions under which the CLT holds impose (a) all up the $2+\Delta$ moment of y_{t+1} (for some $\Delta > 0$) to be bounded, and (b) the process for y_{t+1} to be a strong mixing,[15] that is, a weakly dependent process in probability. A sufficient condition for strong mixing is temporal independence. For concreteness with respect to (a), I require the third moment of π_{t+1} and LB_t to be bounded, which means that they have to have well-defined skewness and implies that the third moment of y_{t+1} is bounded and in principle robust to fatter than normal tails.[16] Furthermore as

[14] Given that I_t^v is an indicator function it is non-negative preserving the sign of the inequality tested, then by assuming $P(I_t^v = 1) > 0$ we obtain equation (1.5).

[15] See for example Thm. 5.20 of White (2001).

[16] No tests can be perform to support assumption (a), this is because in general any assessment on the boundedness of a given moment of a random variable need such requirement as an assumption in order to perform the inference. Nonetheless, requiring π_{t+1} and LB_t to have finite skewnesses is a mild restriction

already noted, we also need to assume π_{t+1} and LB_t to be covariance-stationary in order to interpret $\mathbb{E}[y_{t+1}|I_t^v] < 0$ as a lower bound violation.

In an unreported analysis (available upon requests) a battery of tests is run which finds π_{t+1} and LB_t consistent with a stationary AR(0) and AR(1) processes as well as y_{t+1} consistent with a stationary AR(0) process: these results support the covariance-stationary assumption on π_{t+1} and LB_t and the temporal independence of y_{t+1} needed for the CTL.

HP2: objective rules

The non-parametric test needs a conditioning rule that selects the subsample identified by I_t^v: with little abuse of notation denote such a rule also as I_t^v. In the trivial unconditional case $I_t^v = 1$, but for any more general case we need to be careful to design a rule which is objective in order to avoid sample selection biases: i.e. the rule (i) should not be selected ad-hoc by the econometrician with the aim of maximizing ex-post the chance of getting a rejection and (ii) should not be directly linked to the test.

To tackle these kind of sample selection biases I design rules that select periods where the risk premium is "low" in a training sample, then I compute the lower bound LB_t and perform the non-parametric test in the main (subsequent) sample. In particular, note that the lower bound measure LB_t is violated in t if it is above its conditional risk premium $E_t[\pi_{t+1}]$. Of course, such quantity is unobservable, however, a rough estimate can be computed via an

which holds in many pricing specifications, even in the presence of jumps: for example in a Black and Scholes (1973) world all moments are bounded, and the same remains true if we add a jump diffusion component with constant intensity. In a framework like the time-varying rare disaster of J. A. Wachter (2013), where the intensities are time-varying according to a Cox, Ingersoll, and Ross (1985) model, it would be enough to impose a strictly positive mean-reversion and long-run mean.

econometric model of the form

$$\pi_{t+1} = f(Z_t) + e_{t+1} \tag{1.6}$$

as $\hat{\pi}_{t+1}(Z_t) \equiv \hat{f}(Z_t)$ for some function $f(\cdot)$. Then a rule can be design which turns one at t if the estimate for the risk premium $\hat{f}(Z_t) \approx E_t[\pi_{t+1}]$ is below the lower bound, LB_t, that can be computed in t, as

$$I_t^v \equiv I_t^v(Z_t, LB_t) = 1(\hat{\pi}_{t+1}(Z_t) < LB_t) \tag{1.7}$$

Such rule is (i) pinned-down by data $\{Z_t, LB_t\}$ not ad-hoc by the econometrician, and as long as (ii) the forecasting model (1.6) is specified in a training sample according to an objective criterion, say the best in-sample fit, and then I_t^v computed out-of-sample in a subsequent sample where the test is conducted, also the second objectivity requirement is met.

I construct rules of the type (1.7) in several steps

1. pre-select a vector of excess market return predictors Z_t
2. split the sample $\{1, ..., T\}$ into a training sample $TS \equiv \{1, .., T_s\}$ and a main sample $MS \equiv \{T_s + 1, .., T\}$
3. in TS: for each possible subset $W_t \subseteq Z_t$ compute the associated forecasting model for the excess market return π_{t+1} according to (1.6), and the in-sample adjusted R^2
4. in TS: rank models according to the in-sample adjusted R^2 and pick few among the best performers, say K
5. for every t in MS: using just the structures from the best K performers and $\{Z_1, ..., Z_t\}$ compute K out-of-sample forecasts for the market premium as $\{\hat{\pi}_{t+1}^k(Z_t)\}_{k=1}^K$

6. for every t in MS: compute the lower bound measure, LB_t, using observables in t via eq. (1.3)

7. for every t in MS: compute the K rules as $\{I_t^{v,k} \equiv 1(\hat{\pi}_{t+1}^k(Z_t) < LB_t)\}_{k=1}^K$

Step 1 and 2 are presented in the following data section, while the key intermediate outcomes from the other steps, as well as the potential[17] rejections $\{I_t^{v,k}\}_{k=1}^K$, are presented in the Result section.

1.3 Data

The data used in this study is at the monthly frequency and covers the United States Financial Markets over the period February 1973 to December 2014. The sample is split into a training sample $TS = \{1973:02, ..., T_s\}$ and a main sample $MS = \{T_s + 1, ..., 2014:12\}$ using the last 25 years. The choice of $T_s = 1989:12$ is due to the availability of option data (necessary for the construction of LB_t), that is, $T_s + 1 = 1990:01$ is the first date for which LB_t is computable.[18] Data is divided into two categories: (i) the main variables, namely the market return R_{t+1}, the risk-free return $R_{t,f}$ and the lower bound LB_t and (ii) the predictors in Z_t.

[17] Potential because until we perform the test we don't know if they are rejections.

[18] Choosing T_s this way also allows to maximize the statistical power of the non-parametric test: this is because to select the model for the risk premium forecasts via the proposed statistical method, I do not need to waste any single data point involving options and all available option data is used to perform the main test.

1.3.1 Main Variables

The gross total market return is defined as $R_{t+1} \equiv \frac{\hat{S}_{t+1}}{\hat{S}_t} DY_t$ where $\hat{S}$ represents the closing level of the *Standard & Poor's* 500 (SP500) index and $DY_t \equiv 1 + \frac{D_{t+1}}{\hat{S}_{t+1}}$ is the gross dividend yield with $\{D_t\}$ being the $SP500$ dividend time series (divided by 12) available on Prof. Shiller website.[19] The gross return on a risk-free investment, $R_{t,f}$, is defined as the gross 1-month yield to maturity extracted from the Center for Research in Security Prices (CRSP) continuously compounded yield curve computed over liquid secondary market transactions on U.S. Treasuries.

The time-series of the market premium lower bound, $\{LB_t\}$, is computed according to equation (1.3) in the most conservative way by linearly interpolating[20] the Chicago Board Options Exchange (CBOE) SPX options closing *bid* prices; Data from January 1990 trhough December 1995 is provided by Optsum Data, while data from January 1996 trough December 2014 is taken from OptionMetrics. For dates t in which the data is not sufficient/absent to deliver LB_t at the exact maturity of 1 month I linearly interpolate between the contemporaneous t lower bounds with the two closest maturities.

Table 1.1 summarizes the main variables: note how the estimate for the unconditional risk premium $\mathbb{E}[R_{t+1} - R_{t,f}] \times 100$ is 0.51 monthly or $0.61 \approx \mathbb{E}[\pi_{t+1}|MS]$ in the main sample only, yielding the usual annualized unconditional estimates of 6.1% and 7.3%. The last two rows already show how the lower bound LB_t is unconditionally below its risk premium in the main sample: i.e. π_{t+1} is on average above LB_t, a result we formalize later.

[19] at http://www.econ.yale.edu/ shiller/data.htm

[20] In the Appendix A.2 I show how very similar results are obtained if we use a cubic spline interpolation instead.

Figure 1.1 plots the dynamics for π_{t+1} and LB_t: as was already evident from Table 1.1 the lower bound series LB_t, even if volatile[21], is less volatile than the excess market return π_{t+1} and, by construction, never negative. The time-series of conservative lower bounds LB_t features an annualized average of 3.96% and standard deviation of 3.84%.[22] The lower bounds dynamics portrayed in Figure 1.1 are thoroughly described in Martin (2017).

1.3.2 Predictors for the market risk premium

In what follows I describe the list of pre-selected predictors Z_t which will enter the forecasting model (1.6) for the excess market return (step 1 of the procedure detailed in Section 1.2).

In principle there are infinite ways to select such predictors, giving rise to data-mining issues that can potentially bias the test.[23] To tackle this issue I discipline the choice of Z_t according to a constructive economic rationale which I explain in Appendix A.3.[24] The actual list of predictors Z_t contains the following 11 variables which proxy for the usual dimensions found in the forecasting literature.[25]

F: Ludvigson, Ma, and Ng (2016) Financial uncertainty index, computed as the average forecasting error from 150 financial time-series. It captures the underlying level of uncertainty surrounding financial markets.

[21] Note that its mean is of the same order of magnitude as its standard deviation.

[22] Numbers that, once are restricted to the appropriate sample are very similar to those in Martin (2017): the annualized sample mean and standard deviation of my lower bounds, computed using *bid* quotes, are 4.83% and 4.39%. Martin (2017)'s figures, which use mid rather than bid quotes, are 5% and 4.60%.

[23] The researcher could start with a given list Z_t, perform step 3 to 7 in Section 1.2, run the test and not reject, then he could go back to step 2, modify the list of Z_t...and repeat these steps until he finds a list Z_t which "works".

[24] A piece of evidence further supporting the claim that the test is not driven by data-snooping is given by the fact that, as further discussed later, the characteristics of the detected rejections match those scattered around the exogenous extant literature criticizing representative agent pricing.

[25] E.g. Goyal and Welch (2008), Rapach, Ringgenberg, and Zhou (2016).

SII: Rapach et al. (2016) Short Interest Index, constructed as the log of the equal-weighted mean of short interest (as a percentage of share outstanding) across all publicly listed stocks on U.S. exchanges. It captures the superior informational content of short sellers.

TAXchg: The annual percentage changes in the aggregate dollar amount paid in capital gain taxes. The dollar amounts are reported by the U.S. Department of the Treasury.

ILLIQpi: The Pastor and Stambaugh (2003) (il)liquidity index, computed ad the (negative of the) aggregate average (over a month) daily response of signed volume to next day return for all individual stocks on the New York Stock Exchange and the American Stock Exchange. It represents the % cost incurred in a 1 million 1962 USD trade in the market. Similarly to the Amihud (2002) measure, it is a price impact proxy.

ILLIQts: The W. Liu (2006) (il)liquidity index, computed as the standardized turnover-adjusted number of zero daily trading volumes over the prior 12 months. Similarly to Hou and Moskowitz (2005) measure, captures the trading speed dimension of liquidity.

MDI: The Pasquariello (2014) Market Dislocation Index, computed as the monthly average of hundreds of abnormal absolute violations (mid-quotes minus theoretical prices) of three textbook arbitrage parities in the Stock, Bond and Exchange markets. It tracks potential violations of the Law of One Price when positive.

USDg: The U.S. dollar appreciation index, computed as the percentage rate on the trade weighted dollar index available from FRED Data.[26] The index is a weighted (over the volume of bilateral transactions) average of the foreign exchange value of the U.S. dollar against the currencies of a broad group of major U.S. trading partners.

BM: The Dow-Jones Industrial Average book-to-market ratio.

[26] At https://fred.stlouisfed.org/.

M1g: The monthly percentage growth rate on the Federal Reserve M1 money supply stock. Available from FRED data.

Sent: The Baker and Wurgler (2006) Sentiment index, a composite index based on the common variation of five underlying proxies for sentiment: the closed-end fund discount, the number and average first-day returns on IPOs, the equity share in new issues, and the dividend premium. It captures miss-pricing due to subjective valuations not reflecting rational risk compensation.

The list is parsimonious yet comprehensive: parsimonious in that it conveys a wide variety of non-redundant information as certified by the average absolute correlation of 0.10 from the correlation matrix displayed in Table 1.2: note that the absolute correlation is never higher than 0.44 and greeter or equal than 0.35 only in 4 out of 55 pairs. The list is also comprehensive in that excluded popular variables are highly correlated with Z_t.[27]

1.4 Results

This section is organized in three parts: (i) first the key intermediate steps to construct the rules $\{I_t^{v,k}\}_{k=1}^{K}$ and the outcomes are presented and discussed, then (ii) the non-parametric results are shown and analyzed and (iii) finally the main characteristics of the rejection subsamples are described.

[27]E.g. U.S. inflation correlates 0.55 with $M1g$ and with BM, market volatility correlates 0.70 with F when measured as a GARCH(1,1) on the $SP500$ index and 0.82 when measured by the CBOE VIX index and BM correlates highly with the other excluded popular Goyal and Welch (2008) predictors as reported in the next table

Corr	DP	DY	EP	TBL	LTY
BM	0.90	0.90	0.82	0.69	0.71

1.4.1 Construction of the rules for the non-parametric test

Estimating the excess market return in the training sample

Given the pre-selected list of predictors in Z_t, for each subset $W_t \subseteq Z_t$, the following four flexible specifications for $f(\cdot)$ in model (1.6) are implemented in the training sample $TS = \{1973:02, ..., 1989:12\}$

Linear $f(W) = \beta_0 + \sum_{i=1}^{w} \beta_i W_i$

Pure quadratic $f(W) = \beta_0 + \sum_{i=1}^{w} \beta_i W_i + \sum_{j=1}^{w} \beta_{w+j} W_{w+j}^2$

Interaction $f(W) = \beta_0 + \sum_{i=1}^{w} \beta_i W_i + \sum_{k>l>w}^{w+\frac{w(w-1)}{2}} \beta_l W_l W_k$

Quadratic $f(W) = \beta_0 + \sum_{i=1}^{w} \beta_i W_i + \sum_{j=1}^{w} \beta_{w+j} W_{w+j}^2 + \sum_{k>l>2w}^{w+\frac{w(w-1)}{2}} \beta_{2w+l} W_l W_k$

with w representing the number of elements in W. A total of 8188 models to predict the excess market return π_{t+1} are estimated along with their adjusted R^2. To minimize a model selection purely driven by over-fitting, for each possible set of W_t, only the best model specification is retained.[28]

Figure 1.2 plots the first 100 out of the remaining 2047 models' adjusted R^2: the first 6 models immediately sets apart, a battery of Chow (1960) tests using polynomial up to the third degree highlights a structural brake in the displayed ranking at any significance level between model 6 and 7, and the Diebold and Mariano (1995) test finds the Mean Squared Error (MSE) of model 6 not statistically different from any of the first 5 but statistically lower than the MSE of model 7 at the 1% level. Thus $K = 6$ empirically.

[28] This way I avoid comparisons only made in terms of functional form.

Table 1.3 details each of the selected model in terms of predictors W_t and functional form $f(\cdot)$. 4 out of 6 models use the "Quadratic" functional form while the other 2 the "Interaction" one. In terms of selected predictors, $ILLIQts$, the W. Liu (2006) (il)liquidity measure, is never picked, MDI, the Pasquariello (2014) market dislocation index, is picked by half of the models and the Rapach et al. (2016) short interest index, SII, and the $GINIchg$ index are selected by 4 out of 6 models.

Even if the most important requirement at this level is to show that data rather than the econometrician is selecting the model specifications from (1.6) to be used in the main sample, I nonetheless conclude this subsection by listing evidence against a ranking purely driven by over-fitting: (i) the best in-sample model has the smallest number of regressions, (ii) bootstrapped adjusted R^2 and regression p-values confidence intervals for the 6 selected models are such that no model has an adjusted R^2 smaller than 10% and a regression p-value higher than 0.02 at the 5% level.

Predicting the excess market return out-of-sample

Having at disposal the first $K = 6$ model specifications and the predictors Z_t, for each t in the main sample MS we can perform step 5 of Section 1.2 and forecast π_{t+1} out-of-sample for each model $k \in K$: this yields the set of time t out-of-sample risk premium $E_t[\pi_{t+1}]$ forecasts $\{\hat{\pi}^k_{t+1}(Z_t)\}^6_{k=1}$.

Because Goyal and Welch (2008) show how a naive OLS regression of excess market returns on a large number of predictors will over-parametrize the model and lead to poor out-of-sample forecasts, I combine the information from the set of predictors to obtain optimal

forecasts using the Iterated Combination Method (ICM) of Lin, Wu, and Zhou (2016).[29] A couple of important properties of the time series of these forecasts are discussed next.

First, despite the procedure adopted so far, over-fitting might still play a role in the choice of selecting the *first* K bests models in TS to be used in MS. As a matter of facts, best performers in a given sample tend to under-perform when adopted in another sample and vice-versa with median models remaining more stable. One then might argue for a selection of the K models around the median of the in-sample R^2 ranking distribution rather than in its tail. Panel (a) in Figure 1.3 shows how this is not a concern when constructing the forecasts using the ICM approach: the graph plots the out-of sample ICM MSE on the in-sample counter-part for the 6 selected forecasting models. Models above and to the left of the 45 degree line passing through the origin performed better in-sample while those below and to the right performed better out-of-sample. As it is apparent from the graph, the selected subsample of models present similar in-sample/out-of-sample MSEs[30] and it is more or less balanced.

The second point is about the validity of the specifications of the selected models in the main sample: i.e., is the selection of the subset $W_t \subseteq Z_t$ carried over in TS according to the best fit valid in MS? to answer this question, for each k-th selected model, I compute the residuals $r^k_{t+1} \equiv \pi_{t+1} - \hat{\pi}^k_t$ and for every $z_t \in Z_t$ such that $z_t \notin W_t$ I run the following regression

$$r^k_{t+1} = \alpha + \beta z_t + u_{t+1} \tag{1.8}$$

[29]The ICM method is describe in Appendix A.4 where results from a horse-race against standard OLS forecasts are also reported. The exercise indeed confirm ex-post the superior choice of the ICM over the OLS method.

[30]The highest MSE % difference between any two models is 8.64%.

if model k is well-specified with respect to W_t in MS β should be statistically insignificant. Panel (b) of Figure 1.3 reports for each model the t-statistics associated to β in eq. (1.8). Except for the first best, all model specifications remain correct in MS. Fortunately the miss-specification in the first best model turns out to be negligible; The forth best model is exactly the correction needed to take such miss-specification into account, this is because the *only* difference between the first and the forth model is the inclusion of SII. Furthermore, and most importantly, they generate two rules that are very similar (with a correlation of 0.62), more generally, as it can be checked later, *no result* this paper finds is affected by the exclusion of SII.

The actual rules for the non-parametric test

Figure 1.4 shows the dynamics of the six best rules $\{I_t^{v,k}\}_{k=1}^6$ as well as their correlations. The top graph plots the time series of the six rules $\{I_t^{v,k}\}_{k=1}^6$ against the real GDP growth (the dotted black line). In order to make the graph more readable I multiply each rule by its associated model, i.e. rule for model k is plotted as $k \times I_t^{v,k}$ and assumes values 0 and k in the rejection periods. The six rules clearly display a counter-cyclical pattern, the least correlated rule, rule 1, displays a negative 0.16 correlation with GDP growth, the most correlated rule, rule 2, shows a negative correlation of 0.40, and the average correlation is -0.29. Furthermore, the rules are highly correlated with each other: the smallest correlation, the one between rule 1 and and rule 2, is 0.424, while the highest, the one between rule 2 and rule 3, is 0.927, and the average correlation among all rules is 0.626. The source of this high correlation comes from the high number of shared observations among the different rules: the average pairwise percentage overlap is 77.32%, with the smallest percentage, the

one between rule 1 and and rule 2, in the order of 50%, and the highest, the one between rule 2 and rule 3, in the order of 95%.

The bottom graph highlights in green the sample periods which are systematically detected by all the rejection rules for a total of 35 observations: these are times associated with negative GDP growth (correlation coefficient of -0.30), include all the major economic recessions of the last 25 years (the gray NBER recessions) as well as the 1997 Asian financial crises, the 1998 LTCM crises, the period (late 1999 to 2001) during which the dotcom bubble collapsed, a period in late 2002 when stock market was hitting new lows following the end of the dotcom boom, the quant meltdown in August 2007 and the European sovereign debt crises which accounts for the last two green stripes.

To summarize: the detected rules are very consistent with each other and unambiguously pin-down subperiods containing all the major economic recessions and financial crises. The consistency feature is particularly important in that no *apriori* structure, linking the byproduct of particular subset of instruments $W_t \subseteq Z_t$ and a given specification for the forecasting model (1.6) to yield similar result, is imposed. Once again this seem to suggest that the selected rules are driven by the intrinsic properties of the data and are uncovering systematic patterns unrelated to ad-hoc selections by the econometrician.

At this point such rules has to be considered only as potential rejection periods since no test has yet been performed conditional on them: this is indeed the topic of the next subsection.

1.4.2 The non-parametric test

Table 1.4 reports the main results from the test: each row shows the results for the k-th specification of forecasting model (1.6): four different versions of the key statistic $\mathbb{E}[y_{t+1}|I_t^{v,k}]$ associated with four different rules $I_t^{v,k}$ are reported in percentages and at the monthly frequency. Version (1) reports the outcomes of the unconditional test, version (2) corresponds to the key rule object of this study (detailed in Section 1.2) and shows the outcomes of the main non-parametric test. Versions (3) and (4) replace the lower bound series LB_t in the key rules of version (2) with its unconditional mean, $\bar{LB}$, or 0: these rules serve the purpose of understanding the actual role played by the lower bound in the main version (2). Stars, which are inversely related to the intensity of the green color highlighting the figures, represent the usual confidence levels: they are reported conservatively as the lowest confidence among the two computed from the equivalent test specifications in eq. (1.4) and (1.5) using Newey and West (1987) adjustments and the one derived from p-values adjusted for potential small-sample issues.[31]

Note how unconditionally, at any level of confidence, the lower bound holds: LB_t is on average below the risk premium by 0.281%, consistently with an unconditionally tight[32] lower bound, such estimates are not different from 0. The bulk of this article resides in

[31] P-values associated to version (2) to (4) are obtained by bootstrapping 1000000 random samples I_t^v of the same size as those derived via the rules implemented in the respective versions, computing the 5-th quantiles of the respective simulated distributions for $\mathbb{E}[y_{t+1}|I_t^v]$, and finally comparing them with the actual estimates for $\mathbb{E}[y_{t+1}|I_t^v]$ presented in Table 1.4. This exercise applied to the main version (2) reviles simulated quantiles of the order of 0.5% which means that if truly random rules where adopted (in place of the proposed ones) we would have found the results reported only 0.5% of the time; this, together with the fact that we find significant results, can again be viewed as evidence against rules purely based on over-fitting and also show how the test results are robust to non-normal, potentially fatter and asymmetric, tails in the distribution of I_t^v which might arise since I_t^v contains many recessions and crises.

[32] An unreported analysis (available upon request), consistently with the documented violations, shows that when the lower bound is conditioned on I_t^v from version (2) it becomes a much less tight and more noisy risk premium predictor.

the next set of results, mainly revealed by version (2) in the table: conditional on the rules $\{1(\hat{\pi}^k_{t+1} < LB_t)\}^6_{k=1}$ the lower bound is on average always violated at the 10% level and in 4 out of 6 cases at confidence 95%[33]. In particular the average lower bound is *above* its conditional risk premium by impressive values that range from 1.262 to 1.654. The number of observations involved is on average 66 (as reported in the table), between 23% and 29% of the main sample, with a mean of 70 if we exclude model 6. According to eq. (1.3) $LB_t \geq 0$: the results associated to version (3) and (4) speaks to the importance of its informational content in the test rejections. An *informative* lower bound is essential: $\{1(\hat{\pi}^k_{t+1} < 0)\}^6_{k=1}$ are proper subsets of $\{1(\hat{\pi}^k_{t+1} < LB_t)\}^6_{k=1}$ and they are never able to pin-down rejection, yielding test statistics which are on average 63% (figure in the bottom-right corner of the table) of those in version (2) and p-values grater than 0.10. A non-negative (informative) *dynamic* lower bound is needed: the rejection rule $\{1(\hat{\pi}^k_{t+1} < \bar{LB})\}^6_{k=1}$ associated to version (3) share an average of 91% (a minimum of 86%) of the observations with those from version (2) and significantly improve the test performance, allowing one rejection and four marginal ones with statistics that are on average 78% (figure to the left of the one located in the bottom right corner of the table) of those from version (2). This tells us that three-fourth of the main rejection magnitude (version (2) figures) is due to the dynamics of the excess market return π_{t+1} rather than that of the lower bound LB_t, nonetheless its dynamics it is not negligible yielding the extra quantum, the average residual 22% gap, needed to consistently achieve the rejections.

To sum up, the lower bound is found to hold unconditionally (corroborating the interpretation given in Martin (2017)) but is robustly rejected conditioning on the rules detailed in Section 1.2: as a matter of fact, conditional to these periods, the realized risk premium is

[33]With model 2 being borderline between 10% and 5% and model 6 generating rejections that are at the 5% if evaluated according to eq. (1.4) and at the 7% if evaluated according to eq. (1.5).

below its average lower bound by at least a huge 1.262% (15.144% annualized). The presence of the Martin (2017) lower bound LB_t is crucial to the results and, even if the major role in terms of dynamics is played by the excess market return π_{t+1}, it also have a non-negligible dynamic impact.

1.4.3 Rejection characteristics

This subsection investigates the main characteristics of the rejection periods detected by the rules $\{1(\hat{\pi}^k_{t+1} < LB_t)\}_{k=1}^6$. We already showed that they are counter-cyclical and contain the main economic recessions and financial crises in the main sample MS. Table 1.5 analyzes them using the predictors in Z_t: each row corresponds to a different predictor $z_t \in Z_t$, while different columns identifies rule n.1, $I_t^{v,1}$, trhough rule n.6, $I_t^{v,6}$. The table displays the difference in conditional means of each predictor, z_t, between the rejection subsample and the rest of the sample, $\mathbb{E}[z_t|I_t^{v,1} = 1] - \mathbb{E}[z_t|I_t^{v,1} = 0]$; only predictors with statistically different means can discriminate, and thus characterize (up to a first order approximation), the rejection periods. As it is apparent, only the Ludvigson et al. (2016) uncertainty index F and the Pastor and Stambaugh (2003) (il)liquidity index $ILLIQpi$ can consistently discriminate between rejections and the rest of the sample. As a matter of facts, rejections are periods in which uncertainty is higher by at least 3.89 VIX percentage points[34] on average and the percentage cost incurred in a 1 million transaction in the market is at least 4.387% higher.

This point is made even clearer in Figure 1.5 which plots the time series of F and $ILLIQpi$ respectively, highlighting in bold the portion of the time series which belong to the rejection subsample for the case of the rule n.1.[35] Note how, in both series, most of the spikes are

[34]Due to the high correlation of 0.82 between F and VIX I regressed the first on the latter in order to obtain interpretable magnitudes.

[35]Unreported graphs (available upon request) are very similar for all the other rules.

inside the rejections and how the general level of the series in the rest of the sample is significantly lower.

In summary, the rejection rules $\{I_t^{v,k} \equiv 1(\hat{\pi}_{t+1}^k < LB_t)\}_{k=1}^6$ pin-down periods characterized by high financial uncertainty and market illiquidity, which contain all the major financial crises and economic recessions in the main sample.

1.5 Implications

This section is also organized in three parts: (i) the first part explains the meaning of the non-parametric rejections and offer potential explanations concerning the causes of these failures, (ii) the second part turns to the implications of the rejections for consumption-based representative agent pricing, while the last part (iii) shows evidence in favor of intermediary-based pricing being a more robust setup, and more generally, marginal pricing as a setup able to generate optimal portfolios which are more efficient than the representative agent market portfolio.

1.5.1 Too high model-based representative agent risk premia

Remember from the logic of Section 1.2 that a lower bound violation implies the joint failure of the Martin's class of representative agent pricing: in particular, we have shown that even if unconditionally this class seem to hold, conditioning on the rules $\{I_t^{v,k} \equiv 1(\hat{\pi}_{t+1}^k < LB_t)\}_{k=1}^6$ described in the previous section, it is robustly rejected.

The reason why we reject these models is because their implied risk premia predictions are too high conditioning on the periods identified by the rules: the following chain of inequalities makes this point clearer

$$RP^{models}|I_t^v \equiv \mathbb{E}[\mathbb{E}_t^{models}[\pi_{t+1}]|I_t^v] \geq \mathbb{E}[LB_t|I_t^v] > \mathbb{E}[\pi_{t+1}|I_t^v] \equiv RP|I_t^v$$

by construction, the rejection subsample, I_t^v, contains subperiods where on average the lower bound from eq. (1.3)[36] is above the average excess market return π_{t+1} as shown by the strict inequality. Now, to the right of that inequality we see that the average risk premium given the subsample I_t^v is by definition the actual risk premium in that subsample, while the weak inequality to the left side follows from the pointwise definition of LB_t, which in our test act as the smallest possible bound for the Martin's class of pricing models. Finally $\mathbb{E}[\mathbb{E}_t^{models}[\pi_{t+1}]|I_t^v]$ is by definition the average risk premium implied by the Martin's class of representative agent pricing, $RP^{models}|I_t^v$. In other words, the rejected models assume conditional risk premia which are at least 1.26% (15.12% annualized) higher than the actual conditional risk premium from the data.

This difference is huge and, as shown in Appendix A.5, carries over to risk-adjustments: an analogous chain of inequalities makes this point clearer

$$SR^{models}|I_t^v \equiv \mathbb{E}\left[\frac{\mathbb{E}_t^{models}[\pi_{t+1}]}{\sigma_t(\pi_{t+1})}|I_t^v\right] \geq \mathbb{E}\left[\frac{LB_t}{\sigma_t(\pi_{t+1})}|I_t^v\right] > \frac{\mathbb{E}[\pi_{t+1}|I_t^v]}{\sigma(\pi_{t+1}|I_t^v)} \equiv SR|I_t^v$$

eq. (A.2) in Appendix A.5 shows the result from the strict inequality: the minimum implied Sharpe ratio of the Martin's class is above the analog sample count-part (this is so by at least

[36]Remember that such lower bound is computed conservatively using bid rather than mid-quotes.

0.21, or 0.72 annualized). Therefore also the Sharpe ratios implied by the rejected models are also too high.

Possible explanations

Why popular representative agent pricing delivers implied risk premia which are too high in the subsamples $\{I_t^{v,k}\}_{k=1}^6$? I first show how this fact does not seem to be driven by risk aversion and then offer two potential, non-mutually exclusive, explanations.

In models of the Merton (1980)'s type the risk premium is proportional to the level of risk aversion, it is thus possible that too high risk premia are due to too high representative agent risk aversion; if this is the case we should expect risk aversion to be on average higher during rejections. Panel A of Table 1.6 plots in blue the conditional mean of the Campbell and Cochrane (1999) proxy for time-varying risk aversion, η_t, during rejections and in red the analog conditional mean during the rest of the sample for the discussed rules, while the table at the bottom of the figure reports the difference in these means together with their levels of significance. Both means are very similar and statistically insignificant, with rule 3 and 4 even displaying opposite then expected signs.

While risk aversion in general cannot explain the rejections, I find that implied market crash probabilities are consistent with them. In frameworks such as rare disasters (e.g. Barro (2006) or J. A. Wachter (2013)) the risk premia is an increasing function of the probability of a rare disaster, while these models define a disaster as a consistent drop in consumption or GDP growths, I look at the analogous but observable behavior of the market portfolio: in particular I use the time series of implied market crash probabilities extracted from the

$SP500$ futures by Bollerslev and Todorov (2011)[37]. Panel B of Figure 1.6 plots in blue the conditional mean of the Bollerslev and Todorov (2011) left tail intensities for the $SP500$ futures, CP_t, during rejections and in red the analog conditional mean during the rest of the sample for the discussed rules, while the table at the bottom of the figure reports the difference in these means together with their levels of significance. Results show that the implied crash probabilities are on average around 3% in rejections and only 1% in the rest of the sample, and the difference is statistically significant at the 5% level in 5 out of 6 rules. This is consistent with rejected models being too sensitive to crash probabilities which, in the same spirit as in the classical rare disaster frameworks, might be pushing the risk premia too high.

The last proposed explanation is a direct consequence of the main characteristics of the detected rules. In the previous section we showed that they are characterized by high level of trading frictions (illiquidity) as well as high level of uncertainty: both being on average statistically higher than the unconditional median only during rejection times.[38] In Appendix A.3 it is shown how financial uncertainty, as measured by F, is very highly correlated with classical proxies for asymmetric information, and has itself the typical features an asymmetric information proxy should have. We therefore can conclude that another potential cause for the failure of the Martin's class is the fact that such models do not account for informational and trading frictions: as a matter of fact their absence is behind the required set of assumptions for the existence of a representative agent.[39]

[37] I am thankful to the authors for proving me with such data.

[38] This last unreported claim can be easily verified qualitatively by looking at the plots displayed in Figure 1.5 or formally via an analysis available upon request.

[39] Specifically, the absence of market frictions is required for the existence of an SDF M satisfying $1 = \mathbb{E}[MR]$, while there are currently no frameworks that can construct a representative agent starting from agents with asymmetric information (there are examples, such as Basak (2005), where a representative agent can be constructed with agents having symmetric information but different beliefs).

1.5.2 A unified setup to assess consumption-based representative pricing

As I described in Section 1.2, a key subclass of the Martin's class of models is given by the popular consumption based representative agent setups: leading macro-finance framework including external habit (Campbell and Cochrane (1999)), long-risk (e.g. Bansal and Yaron (2004), Bansal et al. (2014), Campbell et al. (2017)) and rare disaster (e.g. Barro (2006) and J. A. Wachter (2013)) models are jointly rejected conditioning on $\{I_t^{v,k}\}_{k=1}^6$. Therefore my test act as a unified formal setup to assess the characteristics of this type of pricing: in this subsection I show how the rejection characteristics are indeed able to explain, confirm and complement the critique that the extant empirical literature has with respect to consumption-based representative agent pricing.

As a starting point, an unreported analysis[40] shows that the rejection subsamples are characterized by low consumption and GDP growths, while in Appendix A.6 I document how these periods indeed coincide with instances where actual representative models perform the worst.[41]

In the rest of this subsection I link the characteristics of the rejections to the recent findings of Martin (2017), Muir (2017), Moreira and Muir (2017), and Greenwood and Shleifer (2014), confirming and extending their critiques.

[40] The analysis is available upon request and is analogous to the one performed for risk aversion and market crash probabilities in the previous subsection.

[41] The absolute pricing errors coming from the SDFs implied by Campbell and Cochrane (1999), Bansal and Yaron (2004) and J. A. Wachter (2013) using their original calibrations are on average 41% higher during rejections, with those produced by the J. A. Wachter (2013) model being less pronounced no matter the conditioning. The results are robust to nominal as well as real total market returns as proxied by the $SP500$.

One of the reason why rejected models fail might simply be due to the fact that they are not tailor-made to match option data, this is because the lower bound, LB_t, according to eq. (1.3) is just a portfolio of puts and calls on the market portfolio. However, Martin (2017) in his 2017 seminal paper shows how even models such as Bollerslev, Tauchen, and Zhou (2009), Drechsler and Yaron (2011), that explicitly address the properties of option prices are not able to replicate trhough simulations the properties of the lower bound. My test is built using such lower bound and extends Martin's concerns to the entire class of models that satisfy the NCC (eq. (1.2)) within a formal econometric setup.

Muir (2017) shows how consumption based representative agent pricing has difficulty in simultaneously matching risk premia during recessions and financial crises: this is because, despite the different risk premia behavior, the consumption dynamics is very similar. While Muir select these periods exogenously and ex-post, my test pins them down endogenously and ex-ante: as a matter of facts, we already showed that both financial crises as well as economic recessions are inside the rejection subsamples.

Moreira and Muir (2017) find profitable trading strategies which go against representative agent pricing: because actual Sharpe-ratios are lower during recessions (periods of high volatility) and higher during normal times (periods of low volatility), it is profitable to time the market by decreasing the exposure in bad times and increasing it normal times. However, a representative agent is expected to bare more risk rather then less during bad times and be compensated accordingly, therefore the authors claim their strategies to go against representative agent pricing predictions (or equivalently that implied Sharpe ratios of representative agents are too high). In the next paragraph I show how indeed rejections are characterized by higher volatility, lower Moreira-Muir exposures and implied model-based Sharpe ratio which are too high. Table 1.7 shows the average impact of volatility

(measured via the VIX index) and the Moreira and Muir (2017) exposures (computed using the procedure detailed in their paper and using a GARCH(1,1) on the $SP500$ returns[42]). Panel A shows how the rejections are characterized by higher level of volatility (on average higher then the median only during rejections), while panel B reports lower Moreira-Muir exposures (on average higher then the median only in the rest of the sample but during rejections) in the subsamples $\{I_t^{v,k}\}_{k=1}^6$. Finally, recall that during rejections Sharpe ratios implied by the Martin's class of representative agent pricing are on average higher than actual ones by at least 0.21 (or 0.72 annaulized). In summary, the rejections characteristics explain the success of the Moreira-Muir strategies and give a formal test to their claim.

Finally, Greenwood and Shleifer (2014) and Amromin and Sharpe (2013), contrary to rational expectations, show how representative-agent-based required market returns disagree with actual expectations from a non-trivial fraction of investors. Their results are further exacerbated while conditional on the rejection subsample. Greenwood and Shleifer (2014) measure the correlation between two sets of proxies:

Mod representative agent based proxies for require market returns (the dividend price ratio, DP, the Lettau and Ludvigson (2001) consumption to wealth ratio, CAY, and the negative of the Campbell and Cochrane (1999) surplus consumption ration, $-SCR$)

Dat market return actual expectations from survey data (quarterly Graham-Harvey Survey administered to CEOs of big US company, GH, monthly Gallup Survey administered to households with at least 10000 dollar invested, $Gall$)[43]

[42]Results are robust to the usage of the VIX and different rolling windows to compute the GARCH(1,1) on the $SP500$ returns.

[43]See Greenwood and Shleifer (2014) for more details.

under the null of rational expectations the correlation between the two sets of proxies should be one, but the authors find correlations that are either not statistically different from zero or negative. Table 1.8 reports a similar exercise conducted conditionally on no rejections ($I_t^v = 0$) as well as given rejections ($I_t^v = 1$): the upper panel reports the correlations between Mod and Dat conditioning on no rejections, while the bottom panel shows the differential between the conditional correlations of Mod and Dat in rejections with respect to no rejections: conditioning on no rejections already returns similar conclusions to the unconditional Greenwood and Shleifer (2014) test, while the second table shows how on average the correlation between Mod and Dat in rejections is negative and, most of the time, statistically lower than in no rejections.[44]

1.5.3 Marginal versus representative agent pricing

In this subsection I show how the leading marginal intermediary based model of Adrian et al. (2014) is robust to the non-parametric test, and how this might more generally be related to the fact that marginal pricing is able to generate equilibrium portfolios which are more efficient than the representative agent market portfolio.

Adrian et al. (2014): a robust intermediary-based setup

In intermediary based theories the Stochastic Discount Factor depends on the health of the financial sector (see Brunnermeier and Pedersen (2009), Adrian and Boyarchenko (2012),

[44]The only exception being the divided price ratio under the Graham-Harvey Survey where the conditional correlations are not statistically different one with another. But is is mainly because the no rejection conditional starting point is already very negative and statistically significant as displayed in the first raw of the upper panel.

He and Krishnamurthy (2013) and Moreira and Savov (2017)) and in place of a representative agent there is a marginal financial intermediary.[45] The model of Adrian et al. (2014) postulates a linear 1-factor structure for the SDF of a marginal broker/dealer

$$M_{t+1} = 1 - b \times LevFactor_{t+1} \tag{1.9}$$

where $LevFactor_{t+1}$ proxies for shocks to the intermediary wealth by capturing changes in its leverage. Adrian et al. show how this SDF is able to price, explaining 77% of the variation, a non-trivial cross-section of expected returns including equity portfolios sorted by size, book-to-market, and momentum, as well as the cross-section of Treasury bond portfolios sorted by maturity. I this subsection I document that (i) the model correctly prices asset even conditioning on periods where representative agent pricing is rejected, (ii) as a matter of facts, because it does not satisfy the NCC, it is not among the class of models rejected in this study, and finally (iii) the marginal broker/dealer does not hold in equilibrium the market portfolio (at least during representative agent pricing rejections).

In order to assess the correct pricing of the model, I estimate the SDF in eq. (1.9) imposing the pricing equation (1.1) via the following GMM system of equations

$$\begin{cases} \mathbb{E}\left[M_{t+1}R^j_{t+1} - 1\right] = 0\,,\ \text{j= } 1, ..., J \\ \mathbb{E}\left[M_{t+1}R^j_{t+1} - 1 | I^v_t\right] = 0\,,\ \text{j= } 1, ..., J \end{cases}$$

where I use the 41 test assets of Adrian et al. (2014) plus the market portfolio R_{t+1} for a total of $J = 42$ assets and the monthly proxy for $LevFactor_{t+1}$ constructed in Adrian et al. (2014). Results are shown in Panel A of Table 1.9: the model correctly prices the assets unconditionally as well as conditioning on subsamples $\{I^{v,k}_t\}^6_{k=1}$ where the representative

[45] Who might or not be the representative agent.

agent pricing is rejected: no matter which rule (subsample) we look at or whether we use equally weighted (EV) or value weighted (VW) portfolios to form the test assets, the J test does not reject the system of equations at the conventional 5% level and the SDF parameter b is statistically positive as it should be.[46]

Panel B of Table 1.9 reports for each rule (supsample) as well as for the case of equally weighted (EV) and value weighted (VW) test assets the NCC (translating eq.(1.2) in terms of correlations) unconditionally[47] and conditionally on the subsamples $\{I_t^{v,k}\}_{k=1}^6$: regardless of the rule or the weighting used to form the test assets all correlation are positive and very high both unconditionally and conditionally. For the NCC to hold such values should be smaller or equal to zero.

Finally, Panel C of Table 1.9 shows the correlations, unconditionally and conditionally on the subsamples $\{I_t^{v,k}\}_{k=1}^6$, of the implied equilibrium portfolio held by the marginal broker/dealer and the market portfolio: while unconditionally the two are 0.235 and 0.231 positively correlated, conditioning on the rejections $\{I_t^{v,k}\}_{k=1}^6$ in general they are uncorrelated. If the broker/dealer was holding the market portfolio in equilibrium the correlations should have been much closer to one.[48]

[46]Meaning that periods where leverage is low are associated with periods where the marginal utility of the intermediary is high, this is a standard prediction of intermediary-based modes where the leverage is driven by debt considerations as in Brunnermeier and Pedersen (2009) or Adrian and Boyarchenko (2012).

[47]The unconditional figures are slightly different because they are computed as a second stage using the SDF parameter values of b in eq. (1.9) from Panel A.

[48]Not exactly one since the $SP500$, following the Roll (1977)'s critique, might not be a perfect proxy for the truly unobservable market portfolio.

Is marginal pricing able to generate more efficient equilibrium portfolios?

We have seen that the subsamples $\{I_t^{v,k}\}_{k=1}^6$ give hard times to the Martin's class of representative agent pricing but are not an issue for the leading marginal intermediary setup of Adrian et al. (2014). In particular, the conditional failure of representative agent pricing implies that there is no agent holding the market portfolio (as proxied by the $SP500$) in equilibrium, thus, at least for the rejection subperiods, the market portfolio is not efficient in the sense used for example in Dybvig and Ross (1982). On the other hand, in the setup of Adrian et al. (2014), (i) there is no explicit structure that constrains the intermediary broker/dealer to be a representative agent,[49] (ii) the total financial wealth held by broker/dealers is only around 3% (as reported in He and Krishnamurthy (2013) with respect to the year 2010) and (iii) the marginal broker/dealer always (no matter the conditioning) hold some optimal portfolio, other than the market, in equilibrium.

These results suggests the marginal intermediary-based setup of Adrian et al. (2014) to be more robust than those in the Martin's class of representative pricing, and perhaps find more generally, marginal pricing, as able to generate more efficient equilibrium portfolios than the representative agent market portfolio. This latter claim as well as the potential reason why this might indeed be the case are left to be further investigated by future research.

1.6 Conclusion

This paper makes use of a new result by Martin (2017) to deliver a more general and constructive non-parametric test for a non-trivial class of representative agent pricing models.

[49] In particular no assumptions constraining preferences to be homotetic or identical across broker/dealers are imposed.

The test is more general in that it does not impose sharp restrictions on the preferences of the agent and only uses the time series of the market return, its options' quotes and a proxy for the risk-free rate. It is more constructive because, upon rejection, endogenously returns the conditional subsample originating the rejection: a feature which enable the econometrician to study the characteristics of periods associated with the systematic failure of asset pricing models, and theorists to potentially design more robust setups in the future.

Popular representative agent pricing is rejected conditionally on high uncertain and illiquid periods, which contains all the major financial crises and economic recessions in the analyzed sample. These subperiods, endogenously selected based on rules predicting low returns in a training sample, are defined as times where implied model-based risk premia are too high. Findings suggest that such conditional model-based implied risk premia are off by striking amounts - at least by 1.3% monthly (15.6% annually) - even after risk adjustments (model based Sharpe ratios are still higher then actual ones by at least a monthly 0.2). While excessive risk aversion does not seem to explain such results, ruling out Merton (1980)'s type explanations, two alternative channels are found consistent with the data: either rejected models are too sensitive to market crash probabilities, which, as in the rare disaster literature, might be pushing the risk premia too high, or rejected models, typically frictionless, might not be able to account for the high informational and trading frictions characterizing the rejections. Interestingly, the alternative marginal intermediary-based pricing setup is found more robust. In these models pricing is performed via a marginal intermediary, not necessarily a representative agent: the representative model of Adrian et al. (2014) is found to correctly price assets even during periods where representative pricing is rejected and the marginal broker/dealer intermediary not to hold the market portfolio in equilibrium.

This findings are overall suggestive of marginal intermediary pricing being able to generate more efficient portfolios (in the sense used for example in Dybvig and Ross (1982)) than the popular representative agent market portfolio. Further investigation of this claim, as well as the potential reason why it might be the case, are left for future research.

Figure 1.1: Main variables

The figure plots the excess market return $\pi_{t+1} \equiv R_{t+1} - R_{t,f}$ and the lower bound measure LB_t computed according to (1.3), using linear interpolation and bid quotes.

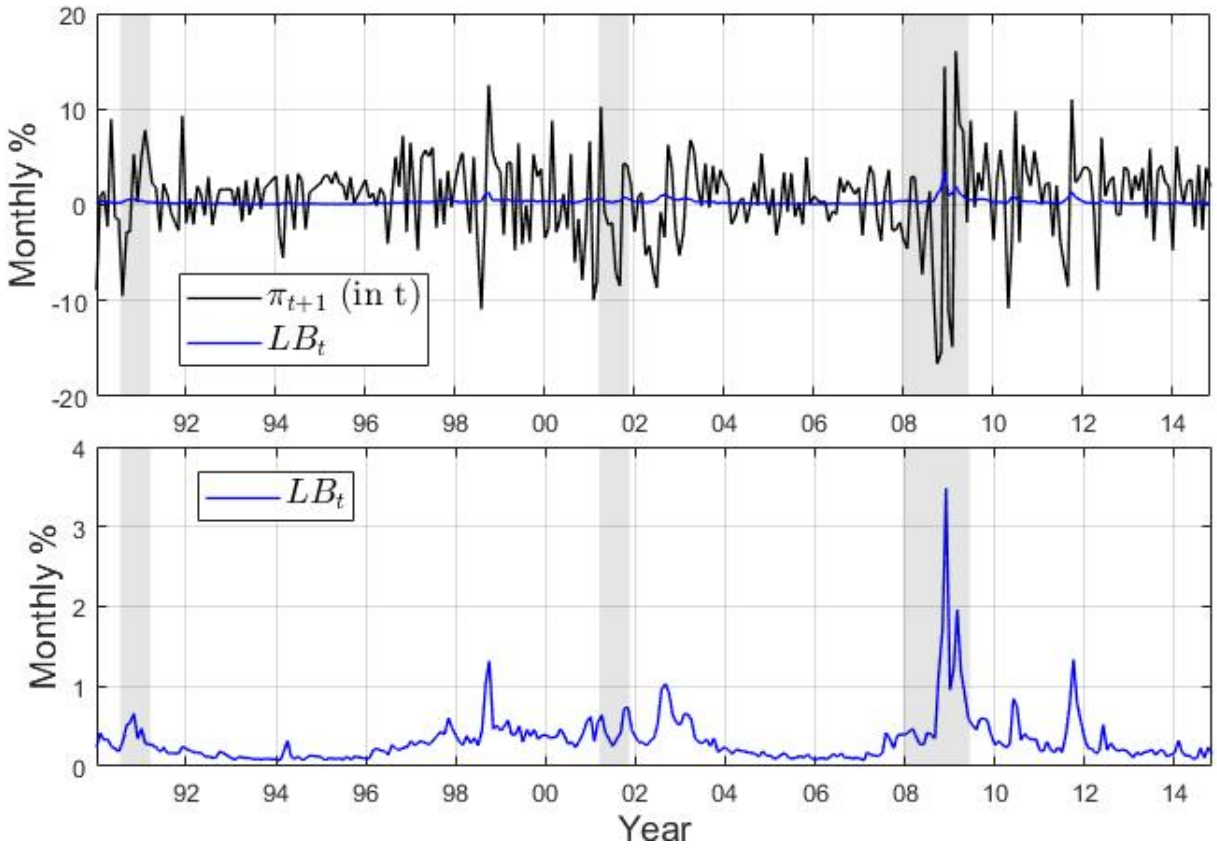


Figure 1.2: First 100 models in training sample according to in-sample fit

The figure shows the first 100 models ranked by their adjusted in-sample R^2 along with a third order polynomial fit. Chow (1960) tests using linear, quadratic or cubic specifications unambiguously identify a brake in correspondence of model 6. The sample is the training one starting from February 1973 and ending December 1989.

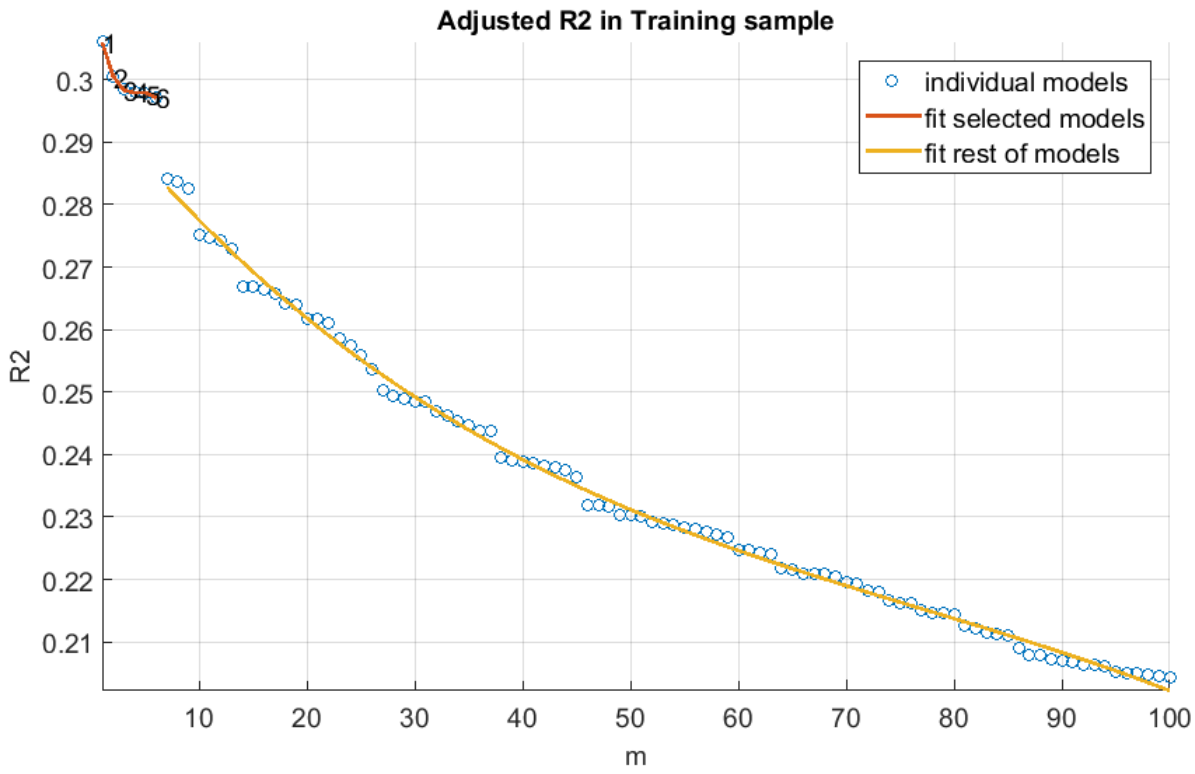
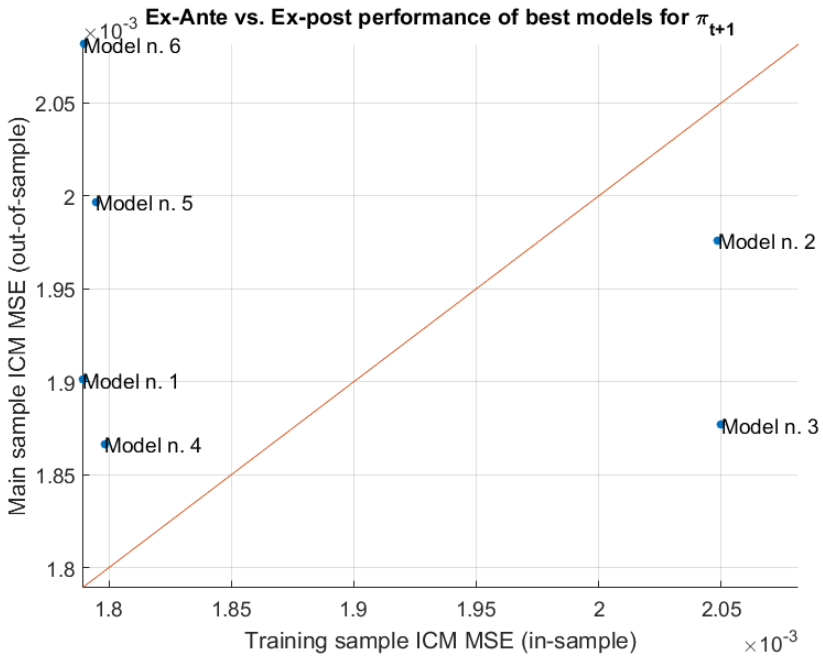


Figure 1.3: Tackling over-fitting in the design of the rules for the non-parametric test

Panel a of the figure plots the out-of sample ICM MSE on the in-sample counter-part for the 6 selected forecasting models. Models above and to the left of the 45 degree line passing through the origin performed better in-sample while those below and to the right performed better out-of-sample. Panel b reports for each selected model in the training sample starting in February 1972 and ending in December 1989 the t-statistics associated to β in equation (1.8) with $r^i_{t+1} \equiv \pi_{t+1} - \hat{\pi}_t$ where $\hat{\pi}_t$ is the forecast of π_{t+1} given one of the six specifications of model (1.6) performed in the main sample starting in January 1990 and ending in December 2014.

Panel (a)



Panel (b)

$$r^i_{t+1} = \alpha + \beta z_t + u_{t+1}, z_t \in Z, z_t \notin W$$

	SII	ILLIQts	MDI	GINIchg
Rule 1	-3.15***	-0.762	-0.5559	-
Rule 2	-	-1.2339	-	-1.81*
Rule 3	-	-1.0336	0.4791	-1.91*
Rule 4	-	-0.8182	-0.539	-
Rule 5	-	-1.0468	-	-
Rule 6	-1.72*	-0.8274	-	-

Figure 1.4: Subsamples originating the non-parametric test rejections

The top graph plots the time series of the six objective rules (subsamples) $I_t^v(W_t, LB_t)$ against the real GDP growth: in order to make the graph more readable I multiply each rule by its associated model, i.e. rule for model j is plotted as $j \times I_t^v(W_t, LB_t)$ and assumes values 0 and j in the rejection periods. The bottom graph shows in green the sample periods which are systematically detected by all the rejection rules for a total of 35 observations.

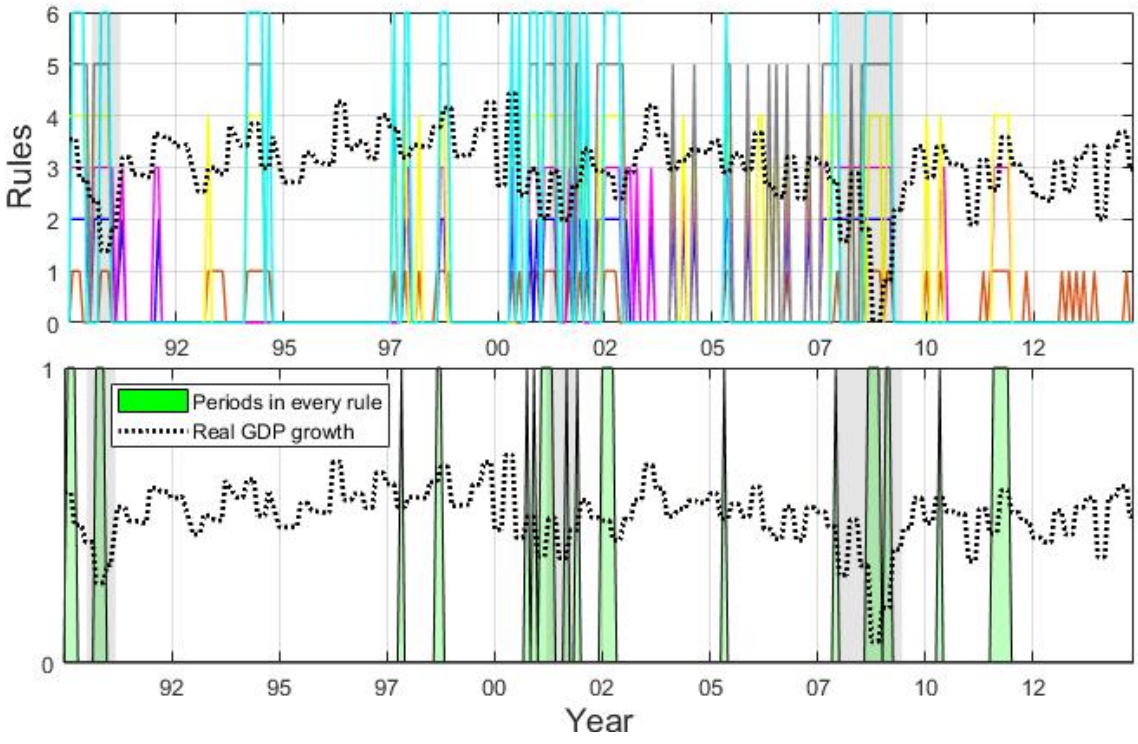


Figure 1.5: Dynamics of uncertainty (F) and illiquidity (ILLIQpi)

The dynamics for the Ludvigson et al. (2016) financial uncertainty index F and the Pastor and Stambaugh (2003) (il)liquidity index are plotted. In bold the periods selected by the representative rule n.1, $I_t^{v,1}$, are highlighted in correspondence of each time-series.

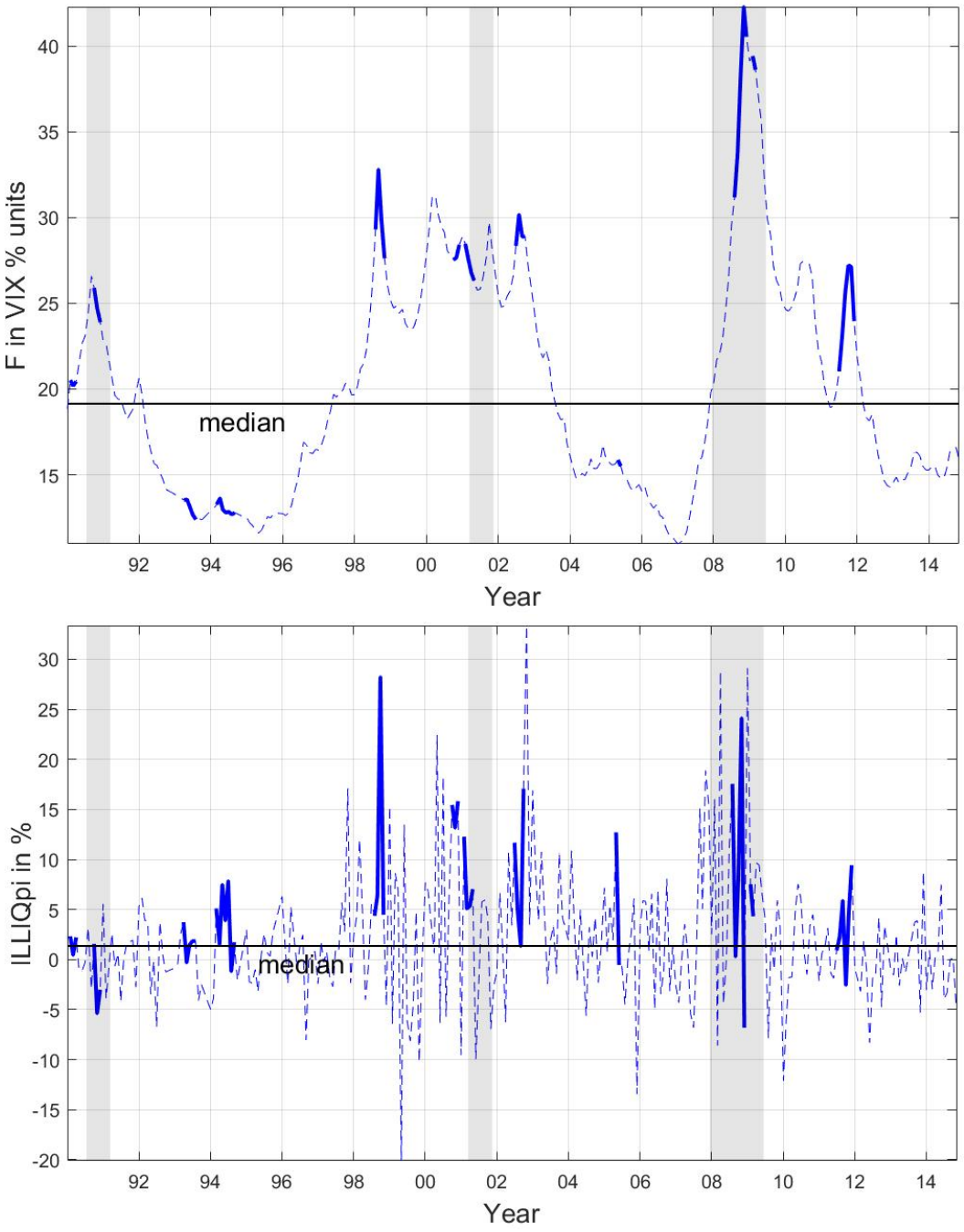


Table 1.1: Summary Statistics on Main Variables

The table summarizes the main variables: $R_{t+1} - 1$ is the total net return on the Standard and Poor's 500, $R_{t,f} - 1$ is the 1-month yield to maturity on U.S. Treasuries, $\pi_{t+1} = R_{t+1} - R_{t,f}$ is the excess market return and LB_t is the market premium lower bound measure computed through (1.3) using linear interpolation and bid quotes. Observations are at the monthly frequency (not annualized). The lower bound and excess market return statistics are computed in the main sample January 1990 trhough December 2014 while the market and the risk-free return are computed over the entire sample February 1973 trhough December 2014.

Variable	**Mean**	**Std.Dev.**	**Min**	**Max**	**N. Obs.**	Sample
$(R_{t+1} - 1) \times 100$	0.93	4.57	-21.62	17.05	492	All
$(R_{t,f} - 1) \times 100$	0.42	0.29	0.000	1.38	492	All
$\pi_{t+1} \times 100$	0.61	4.53	-16.62	16.04	289	Main
$LB_t \times 100$	0.33	0.32	0.07	3.48	289	Main

Table 1.2: Pearson correlation matrix for the candidate predictors

The table displays Pearson correlation coefficients for the candidate predictors (instruments) Z, described in Section 1.3, over the entire sample from February 1973 to December 2014, the overall average absolute correlation is 0.10.

Corr	F	SII	$TAXchg$	$ILLIQpi$	$ILLIQts$	MDI	$USDg$	BM	$M1g$	$Sent$
F	1									
SII	-0.03	1								
$TAXchg$	-0.16	0.03	1							
$ILLIQpi$	0.35	0.06	-0.04	1						
$ILLIQts$	-0.05	0.01	0.05	0.10	1					
MDI	0.44	0.04	-0.10	0.23	-0.04	1				
$USDg$	0.00	-0.14	0.02	-0.02	0.09	0.10	1			
BM	0.02	-0.41	0.05	0.06	0.00	0.05	-0.11	1		
$M1g$	0.15	-0.03	-0.09	-0.07	-0.05	0.19	-0.05	0.20	1	
$Sent$	-0.08	0.08	0.23	-0.15	0.13	-0.19	0.10	-0.42	-0.06	1
$GINIchg$	-0.13	-0.01	0.24	-0.04	0.06	-0.10	-0.03	-0.04	-0.05	0.22

Table 1.3: Best selected models to predict the excess market return

The table describes the characteristics of the first 6 models to predict the excess market return π, ranked by their adjusted in-sample R^2 in the training sample from February 1973 through December 1989.

W	F	SII	$TAXchg$	$ILLIQpi$	$ILLIQts$	MDI	$USDg$	BM	$M1g$	$Sent$	$GINIchg$	$f(\cdot)$
1st	X		X	X			X	X	X	X	X	Int
2nd	X	X	X	X		X	X	X	X	X		Quad
3rd	X	X	X	X			X	X	X	X		Quad
4th	X	X	X	X			X	X	X	X	X	Int
5th	X	X	X	X		X	X	X	X	X	X	Quad
6th	X		X	X		X	X	X	X	X	X	Quad

Table 1.4: Non-parametric Test

Each row shows the results for the k-th specification of forecasting model (1.6): four different versions of the key statistic $\mathbb{E}[y_{t+1}|I_t^{v,k}]$ associated with four different rules $I_t^{v,k}$ are reported in percentages and at the monthly frequency. Version (1) reports the outcomes of the unconditional test, version (2) corresponds to the key rule object of this study (detailed in Section 1.2) and shows the outcomes of the main non-parametric test. Finally versions (3) and (4) replace the lower bound series LB_t in the key rules of version (2) with its unconditional mean, $\bar{LB}$, or zero: these rules serve the purpose of understanding the actual role played by the lower bound in the main version (2). Stars, which are inversely related to the intensity of the green color highlighting the figures, represent the usual confidence levels: they are reported conservatively as the lowest confidence among the two computed from the equivalent test specifications in eq. (1.4) and (1.5) and the one derived from p-values adjusted for potential small-sample issues. At the bottom of the table the average number of observations per given rule (a given version), and the ratio of the magnitudes of figures in version (3) and (4) with respect to the baseline version (2) are reported.

(n)	(1)	(2)	(3)	(4)
$E\left[y_{t+1}\middle\vert I_t^{v,k}\right](\%)$	$I_t^{v,k}=1$	$I_t^{v,k}=1(\hat{\pi}_{t+1}<LB_t)$	$I_t^{v,k}=1(\hat{\pi}_{t+1}<\overline{LB})$	$I_t^{v,k}=1(\hat{\pi}_{t+1}<0)$
Model 1 (k=1)	0.281	-1.283**	-0.697	-0.895
Model 2 (k=2)	0.281	-1.262*	-1.150*	-0.403
Model 3 (k=3)	0.281	-1.399**	-1.068*	-0.565
Model 4 (k=4)	0.281	-1.364**	-1.108**	-1.107
Model 5 (k=5)	0.281	-1.479**	-1.219*	-0.781
Model 6 (k=6)	0.281	-1.654*	-1.363*	-1.565
Avg.	0.281	-1.407	-1.101	-0.886
Avg. obs.	289	66	68.5	44.3
Avg. (n)/(2)			0.782	0.630

Table 1.5: Rejection characteristics

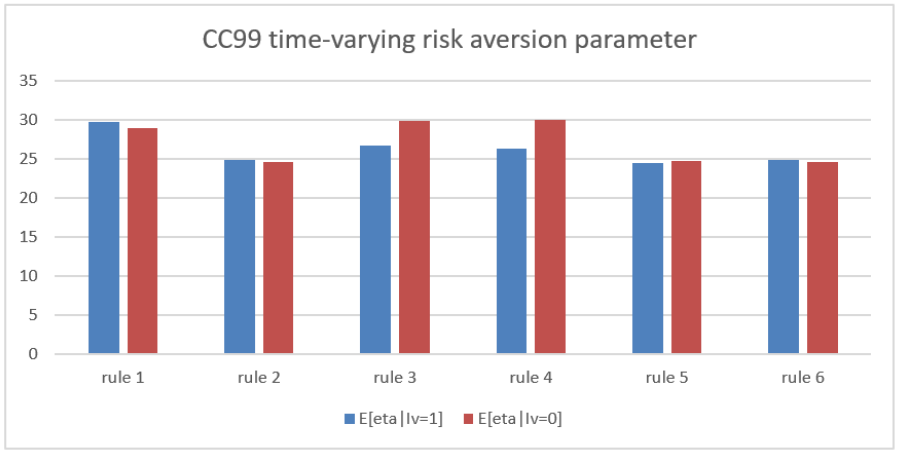
Each row corresponds to a different predictor $z_t \in Z_t$, while different columns identifies rule n.1, $I_t^{v,1}$, trhough rule n.6, $I_t^{v,6}$. The table displays the difference in conditional means of each predictor, z_t, between the rejection subsample and the rest of the sample, $\mathbb{E}[z_t|I_t^{v,1}=1]-\mathbb{E}[z_t|I_t^{v,1}=0]$; only predictors with statistically different means can discriminate, and thus characterize (up to a first order approximation), the rejection periods.

$E[z_t\|I_t^v=1]-E[z_t\|I_t^v=0]$		Rejection subsample					
		$I_t^{v,1}$	$I_t^{v,2}$	$I_t^{v,3}$	$I_t^{v,4}$	$I_t^{v,5}$	$I_t^{v,6}$
z	F (VIX % units)	3.893***	5.973***	6.027***	3.994***	5.073***	6.867***
	ILLIQpi (%)	4.387***	6.444***	6.638***	5.919***	6.716***	7.227***
	MDI	0.113*	0.133***	0.147***	0.089*	0.119**	0.170**
	Sent	0.244	-0.023	0.04	0.217	0.179	0.281
	BM	0.005	0.029	0.023	0.011	0.015	0.000
	M1g (%)	0.440**	0.384***	0.321*	0.299	0.338*	0.299
	USDg (%)	0.794***	0.366	0.535**	0.333	0.322	0.661**
	TAXchg (%)	8.859*	-10.848*	-6.257	-1.341	-3.367	-2.956
	ILLIQts	0.012**	0.011*	0.008	0.012**	0.011*	0.012**
	SII	-0.121	0.819***	1.010***	0.770***	0.503***	0.021
	GINIchg (%)	0.297*	-0.117	-0.150	0.223*	0.259*	0.364**

Table 1.6: Potential explanations for Representative Agent pricing failures

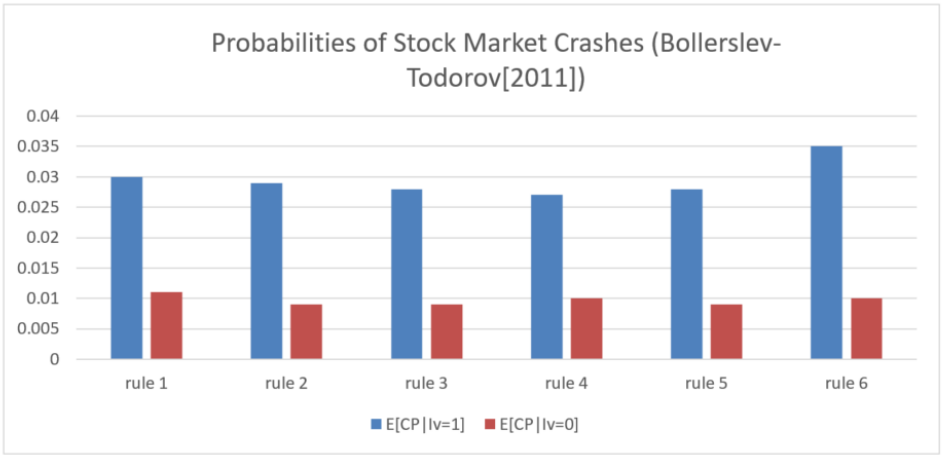
Panel A plots in blue the conditional mean of the Campbell and Cochrane (1999) proxy for time-varying risk aversion, η_t, during rejections and in red the analog conditional mean during the rest of the sample for the discussed rules, while the table at the bottom of the figure reports the difference in these means together with their levels of significance. Similarly panel B reports in blue the conditional mean of the Bollerslev and Todorov (2011) left tail intensities for the Standard and Poor's futures, CP_t, during rejections and in red the analog conditional mean during the rest of the sample for the discussed rules, while the table at the bottom of the figure reports the difference in these means together with their levels of significance.

Panel A



	rule 1	rule 2	rule 3	rule 4	rule 5	rule 6
E[eta|Iv=1]-E[eta|Iv=0]	0.891	0.213	-3.149	-3.653*	-0.321	0.249

Panle B

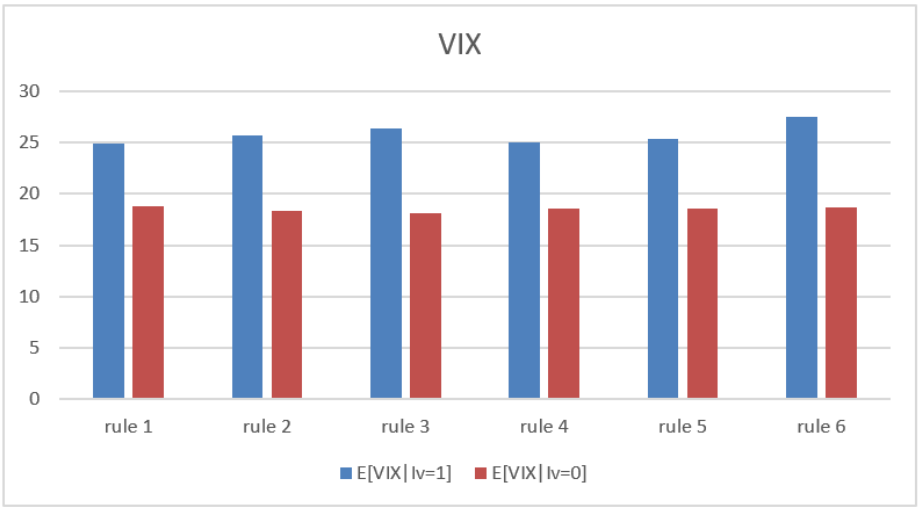


	rule 1	rule 2	rule 3	rule 4	rule 5	rule 6
E[CP|Iv=1]-E[CP|Iv=0]	0.019*	0.019**	0.02**	0.017**	0.019**	0.025**

Table 1.7: Justifying failures in light of the results of Moreira and Muir (2017)

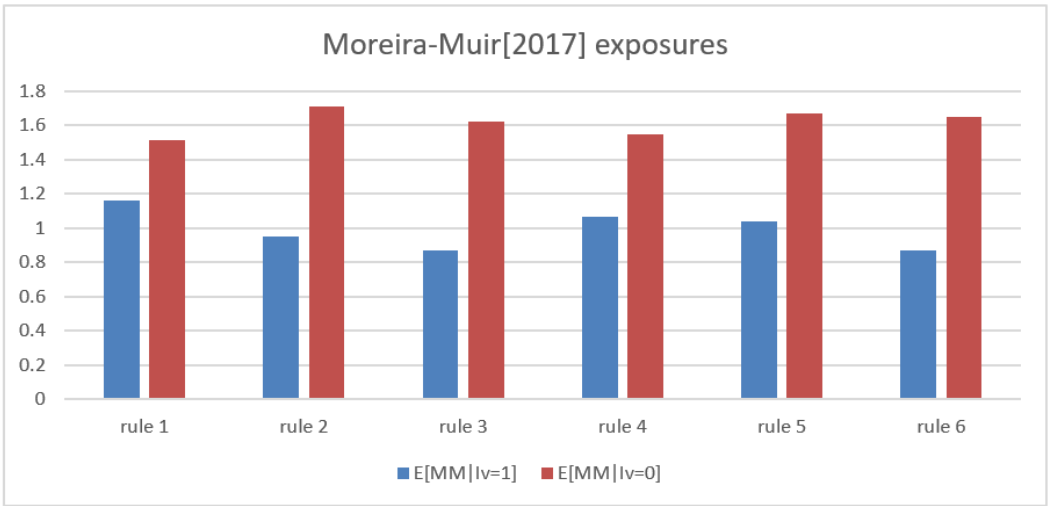
Panel A shows how the rejections are characterized by higher level of volatility (on average higher then the median only during rejections), while panel B reports lower Moreira and Muir (2017) exposures (on average higher then the median only in the rest of the sample but during rejections) in the subsamples $\{I_t^{v,k}\}_{k=1}^{6}$.

Panel A



	rule 1	rule 2	rule 3	rule 4	rule 5	rule 6
E[VIX|Iv=1]-E[VIX|Iv=0]	6.082***	7.413***	8.335***	6.441***	6.806***	8.84***

Panel B



	rule 1	rule 2	rule 3	rule 4	rule 5	rule 6
E[MM|Iv=1]-E[MM|Iv=0]	-0.352*	-0.758***	-0.750***	-0.480***	-0.626***	-0.777***

Table 1.8: Greenwood and Shleifer (2014) test for rational expectations

The upper panel reports the correlations between model-based representative agent proxies for require market returns (the dividend price ratio, DP, the Lettau and Ludvigson (2001) consumption to wealth ratio, CAY, and the negative of the Campbell and Cochrane (1999) surplus consumption ration, $-SCR$) and proxies for actual market return expectations from survey data (quarterly Graham-Harvey Survey administered to CEOs of big US company, GH, monthly Gallup Survey administered to households with at least 10000 dollar invested, $Gall$, see Greenwood and Shleifer (2014) for more details) conditioning on no rejections ($I_t^v = 0$), while the bottom panel shows the differential between the conditional correlations of these variables in rejections ($I_t^v = 1$) with respect to no rejections ($I_t^v = 0$).

corr(Dat,Mod | Iv=0)		rule 1	rule 2	rule 3	rule 4	rule 5	rule 6	avg.
Dat = GH	Mod = DP	-0.603	-0.741	-0.659	-0.614	-0.680	-0.634	-0.655
	Mod = CAY	-0.076	0.132	0.058	-0.037	0.013	0.007	0.016
	Mod = -SCR	-0.535	-0.433	-0.575	-0.564	-0.401	-0.415	-0.487
Dat = Gall	Mod = DP	-0.635	-0.641	-0.631	-0.667	-0.634	-0.617	-0.637
	Mod = CAY	-0.203	-0.131	-0.116	-0.202	-0.206	-0.147	-0.168
	Mod = -SCR	-0.494	-0.559	-0.531	-0.527	-0.530	-0.507	-0.525

corr | Iv=1 - corr | Iv=0		rule 1	rule 2	rule 3	rule 4	rule 5	rule 6	avg.
Dat = GH	Mod = DP	-0.157	0.398	0.078	-0.068	0.128	-0.021	0.060
	Mod = CAY	-0.572	-0.666	-0.528	-0.515	-0.511	-0.965	-0.626
	Mod = -SCR	0.106	-0.158	-0.081	0.103	-0.254	-0.247	-0.089
Dat = Gall	Mod = DP	-0.202	-0.075	-0.097	-0.032	-0.115	-0.221	-0.124
	Mod = CAY	-0.261	-0.297	-0.376	-0.270	-0.237	-0.368	-0.301
	Mod = -SCR	-0.308	-0.199	-0.205	-0.213	-0.260	-0.319	-0.251

Sig. at 1% Sig. at 5% Sig. at 10%

Table 1.9: The robust intermediary-based setup of Adrian, Etula and Muir (2014)

Panel A displays, across rules (subsamples) and using equally weighted (EV) versus value weighted (VW) test assets as in Adrian et al. (2014), the pricing delivered by eq. (1.9): a correct pricing should display a positive and significant coefficient b and should not be rejected by the GMM J-test. Panel B reports the NCC condition of eq. (1.2) translated in terms of correlations both unconditionally and conditionally upon the rejection rules (the reason why the unconditional values slightly differ across rules is due to the fact that the correlations have been computed as a second stage using the b coefficients estimated in Panel A). Finally Panel C reports the unconditional versus conditional correlations between the market return R_{t+1} and a proxy for the Adrian et al. (2014) broker/dealer equilibrium portfolio $LevFactor_{t+1}$.

Panel A

b

	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Rule 6
EW	0.857***	0.825***	0.923***	0.873***	0.791***	0.822***
VW	0.790***	0.715***	0.827***	0.792***	0.702***	0.728***
	J-test never rejected at the 5% level					

Panle B

	$corr(M_{t+1} \times R_{t+1}, R_{t+1})$						$corr(M_{t+1} \times R_{t+1}, R_{t+1} \mid I_t^v)$					
	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Rule 6	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Rule 6
EW	0.797***	0.826***	0.771***	0.791***	0.838***	0.828***	0.849***	0.871***	0.826***	0.849***	0.884***	0.871***
VW	0.829***	0.868***	0.815***	0.828***	0.873***	0.865***	0.866***	0.894***	0.849***	0.876***	0.902***	0.890***

Panel C

	$corr(LevFactor_{t+1}, R_{t+1})$						$corr(LevFactor_{t+1}, R_{t+1} \mid I_t^v)$					
	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Rule 6	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Rule 6
EW	0.235***						0.178	0.137	0.222*	0.153	0.071	0.044
VW	0.231***						0.193	0.156	0.230**	0.18	0.118	0.064

Chapter 2

Mean-Variance Portfolio Rebalancing with Transaction Costs

2.1 Introduction

Optimal portfolio rebalancing given transaction costs is a complex problem. Even with only two assets, solving for the optimal strategy in a continuous-time model involves either a primal free boundary problem (see, for example, Davis and Norman (1990), Dumas and Luciano (1991), H. Liu and Loewenstein (2002), Shreve and Soner (1994) and Taksar, Klass, and Assaf (1988)) or its dual formulation (e.g. Goodman and Ostrov (2010), see Schachermayer (2017) for a comprehensive summary of this approach). When there are more securities or time is discrete, models have been solved only in the extreme case of uncorrelated returns and constant absolute risk aversion (H. Liu (2004)) or with numerical or heuristic approximations (Leland (2000), Balduzzi and Lynch (1999, 2000), Donohue and Yip (2003), Han (2005), Muthuraman and Kumar (2006), Irle and Prelle (2008), Lynch and Tan (2009), Myers (2009) or Dumas and Buss (2017)). Furthermore, except for the case of uncorrelated securities, when solutions are available they only involve two and rarely three risky assets.

In this paper, we study the single period investment decisions of a mean-variance investor only using basic calculus and we are able to: (i) derive exact solutions in the presence of many assets, with generic correlation structure, under proportional and fixed costs, (ii) validate key characteristics of the optimal solutions only conjectured in previous studies, (iii) uncover new economic insights behind the optimal strategies, and (iv) provide useful algorithms for actual large scale problems.

Mean-variance analysis was originated by Markowitz (1952, 1959), who described the basic formulations and the quadratic programming tools used to solve them. The theory was further described by Tobin (1958), who focused on macroeconomic implications of the theory. Early discussions of transaction costs often focused on the intuition that small investors who face high costs will choose a smaller and less diversified portfolio than will a large investor with smaller costs. This intuition has been formalized by a constraint on the number of securities in the portfolio (Jacob (1974)), a fixed cost for each security included in the portfolio (Brennan (1975), Goldstein (1979), and Mayshar (1979, 1981)) or a study of benefits of adding securities without modeling the costs (Mao (1970, 1971)). Unfortunately, this type of assumptions tend to produce a somewhat messy combinatoric problem looking at all possible subsets to include, and their static perspective does not seem suited to questions about rebalancing. The current analysis differs in two important ways from the traditional mean-variance literature on transaction costs. First, the traditional literature considered the purchase of a portfolio from scratch, while the current analysis considers rebalancing from any starting portfolio. Second, we also consider variable costs rather than only fixed costs (both security specific and overall) and their combination (variable costs and other institutional features were included in choice problems of Pogue (1970), but without any analysis of the solution).

The two setups that are closest to our analysis are the continuous time multi-asset models with constant investment opportunity sets presented in Leland (2000) and H. Liu (2004). Leland (2000) provides an heuristic approach to minimize proportional transaction costs. There is no explicit utility maximization and the solution is conjectured to contain a non-trading region with the shape of a parallelogram (or its higher dimensional analogs). Imposing mean-variance preferences in a one-period model, our setup endogenously generates the same qualitative properties. H. Liu (2004) solves proportional and fixed transaction costs problems in the presence of many uncorrelated risky assets. His no-trading regions are rectangles (or higher dimensional analogs) which arise as special cases in our setup when the covariance matrix of the risky securities is diagonal.

A solution to a portfolio optimization problem can be thought of as the set of all potential trades (mappings from the initial pre-trades allocations to the final post-trades ones) that are optimal given the preferences of the hypothetical investor. Absent costs any trade leads to the same ideal allocation (so that usually in these frameworks a solution is directly defined as such allocation). When trading involves transaction costs, there exists a set of initial allocations which are too close to the ideal one to justify any trade, so that it is optimal not to trade at all. This set is referred to as the no trade region: its shape, as well as the type of trades which are optimal from initial allocations outside of it, are fully characterized by the specific structure of the costs.

In the variable cost models, it is optimal to trade only to the closest boundary of the non-trading region, since trading further would incur additional costs that are not justified.[50]

[50]Masters (2003) contains a mean-variance-style analysis with a single risky security and variable trading costs in which it is claimed that it is not optimal to trade to the boundary of the non-trading region. However, this is because the paper computed the non-trading region incorrectly as the set of portfolios from which it would be worth trading to the ideal point that would be chosen absent costs. The error is that there are portfolios from which it pays to trade partway to the ideal portfolio but not all the way.

With proportional costs, the cost of trading is additive (if all trades are in the same direction) or less than additive (if the second trade reverses the first trade in some securities). If a candidate trade does not take us to the no trade region, we could add on the additional trade we would make from that point and be better off. Or, if a candidate trade takes us beyond the boundary of the trading region, is better to trade along the line to the boundary because the part of the trade beyond the boundary is not justified. These arguments do not work for fixed costs, because they rely implicitly on costs being additive for sequential trade along a line, and on costs being no more than additive for sequential trade that are not along a line.

In the models with fixed costs, any trade moves to inside the no trade region if it is optimal to trade at all: this is because fixed costs once triggered become sunk costs. With only an overall fixed cost, any nonzero trade moves to an ideal portfolio that would be held absent costs. This ideal portfolio is in the interior of the no trade region because the value of trading from nearby is too small to cover the fixed cost. With security-specific fixed costs, any trade will take us to the interior of the no trade region. However, different starting portfolios will cause us to trade to different target portfolios.

In models featuring both fixed (overall and asset-specific) and proportional costs the no trade region and the optimal trades look very similar to the ones generated by only proportional costs for initial allocations enough far away from the ideal allocation. This is because the impact of proportional over fixed costs increases with the distance from the ideal allocation (and only proportional costs depends on it) so that when the position is far enough proportional costs are just what matter. For initial positions closer to the ideal one, the impact of the fixed component(s) become more and more relevant up to a locus of points, the actual boundaries of the no trade region, where we are indifferent between not trading

at all or paying the fixed cost(s) and trade further up to where it is optimal according to the proportional components of the costs.

The analysis in this article can accommodate multiple risky assets, trading of individual securities or bundles or pairs, and trading futures or swaps as well as stocks. In particular, while applying our model to trading in futures and its underlying, we show how to improve over traditional futures overlay strategies when it is possible to take advantage of a superior return from holding the underlying. We also show how engaging in cheaper bundles trading makes us better off by squeezing the non-trading region towards the ideal allocation absent costs.

Our models are obtained by solving and comparing a finite combination of standard, strictly convex, quadratic programs. Some programs have closed-form solutions while the simple structure of the remaining ones allows a one-to-one mapping between their first order conditions and linear complementarity representations, for which efficient and fast converging algorithms exist. This is why we are able to provide algorithms for large scale problems involving many risky securities.

Even if simple and exact, a mean-variance setup remains a myopic approximation of the true dynamic strategy, nonetheless optimizing over transaction costs using our algorithms is very useful in practice. Maurer and Pezzo (2018) show the empirical relevance of applying our framework in the context of the FX markets: taking into account costs while optimizing over 29 developed and emerging currencies from 1976 to 2016 leads to an economically large and statistically significant improvement in the out-of-sample performance with a Sharpe ratio increment of approximately 30%. They also show how the majority of such increment, 70%, is due to the proper treatment of correlations. In other words, modeling correlations

is important, a strategy based on a framework like that of H. Liu (2004), where assets are considered uncorrelated, delivers no significant improvement.

The rest of the paper is organized as follows: Section 2.2 gives an overview of the general framework, in Section 2.3 we provide graphical examples covering all the main features of our framework including the new insights on futures overlay strategies, bundle tradings and the impact of a benchmark. Section 2.4 contains the formal characterization of the problem: we proof the existence and uniqueness of solutions and discusses their structures including the steps to constructs all the basic examples of Section 2.3 and the comparative statics analysis. Section 2.5 describes the algorithm to numerically solve our models for a large number of risky assets. Section 2.6 concludes while the Appendix contains the proof of the existence and uniqueness of solutions.

2.2 The Mean-Variance Framework

There are $n+1$ asset returns which realizes at the end of the period and initial financial wealth normalized to 1. By default the first asset, asset 0, also referred to as cash, is risk-free and can be understood as the bank account used to trade in all other risky assets. At the beginning of the period, starting from an initial allocation for the risky assets (which can be a vector of zeros), θ^0, the investor has to choose the vector of portfolio weights θ in order to maximize

$$U(\theta) = r + \theta'(\mu - r1) - \frac{\lambda}{2}\theta'\mathbf{V}\theta - \frac{\kappa}{2}(\theta - \theta^B)'\mathbf{V}(\theta - \theta^B) - c(\theta, \theta^0) \tag{2.1}$$

The first three terms in the utility function are standard for mean-variance optimization. The first two terms r and $\theta'(\mu - r1)$ give the expected return; First the risk-free rate and then the net change in expected return from holding a nontrivial risky portfolio. The third term $\frac{\lambda}{2}\theta'\mathbf{V}\theta$ is the utility penalty for taking on variance. The constant $\lambda > 0$ is the coefficient of risk aversion; the larger the value of λ, the more reluctant the investor is to take on risk in exchange for return, and $\theta'\mathbf{V}\theta$ is the portfolio's variance. [51]

The fourth term $\frac{\kappa}{2}(\theta - \theta^B)'\mathbf{V}(\theta - \theta^B)$ is a penalty for tracking error. This term is perhaps controversial because it depends on a benchmark, θ^B, and not just on the distribution of returns.[52] The dependence on the benchmark would be unnecessary and probably damaging in an ideal world,[53] but does arise in practice and should be very familiar to practitioners.

The last term is the cost function $c(\theta, \theta^0)$: trading does not come for free and re-balancing the initial position θ^0 to the new position θ entails resource dissipation. We will model such costs as proportional (to the size of the trade), fixed (per trade and independent of size) or a combination of the two as detailed in the following sections.

2.3 Examples

We first provide the reader with the main insights of our framework with a graphical overview: Example 1 trough 5 illustrate the class of problems analyzed in this paper while Example 6

[51] Including 2 in the denominator makes the units the same as absolute risk aversion in a multivariate normal model with exponential utility, and also cancels when we look at the first-order conditions.

[52] If you do not like this term, you can always restrict attention to $\kappa = 0$.

[53] Indeed Roll (1992) shows how the mean-variance frontier obtained by minimizing the variance of the tracking error subject to a target expected return over a given benchmark is dominated by the standard mean-variance frontier.

trough 8 show interesting applications which exploit the analytical form of our setup. The formal problem characterization is postponed to the next section.

2.3.1 Example 1: Proportional Costs

A typical case with proportional transaction costs is shown in Figure 2.1. If the initial allocation θ^0 is in the nontrading region, the area of the parallelogram, then there is no trade whose benefit covers the cost and it is best not to trade. The right boundary of the nontrading region is part of the line along which we are just indifferent about selling Security 1, $s1$, and it is optimal to sell Security 1 if we start to the right of this boundary. The left boundary of the nontrading region is part of the line along which we are just indifferent about purchasing Security 1, $b1$, and it is optimal to purchase Security 1 to the left of this boundary. The boundaries for purchasing and selling are different because the costs put a wedge between the marginal valuations at market prices and valuation the the prices net of costs.

Similar to the case of Security 1, we sell Security 2, $s2$, if we start above the top boundary and we buy Security 2, $b2$, if we start below the bottom boundary. If we start in the regions further away from the corners (not directly to the right, left, top, or bottom of any of the sides of the nontrading region), then we trade in both securities up to the nearest corner.

It may not be obvious that the correct trades are as shown by the arrows in Figure 2.1. For example, could there be some points in the region above the top boundary of the nontrading region from which we trade to the upper right corner of the region? The answer is no, because if we buy Security 1 at all, we must end up on the corresponding boundary. In this case, any net purchase of Security 1 must be to the left boundary.

Absent the positive correlation between the returns on the two securities, the non-trading region would have been a square (or a rectangle given diverse security-specific costs) with sides parallel to the axes. With positive correlation, the two securities are substitutes, and over- or under-weighting in one security is more serious if we have the same over- or under-weighting in the other security. This is why the nontrading region is larger along the -45 direction in which the over- and under-weightings cancel than along the 45 direction in which the over- and under-weightings are reinforced.

2.3.2 Example 2: Overall Fixed Cost

If there is an overall fixed cost, then if we trade the cost is the same whatever trade we make. Therefore, if we are going to trade, we always trade to the same ideal portfolio.

The overall fixed cost is illustrated in Figure 2.2. The non-trading boundary is bounded by an ellipse. From outside this region, it is optimal to trade to the ideal point, since it is no more costly to trade to the ideal portfolio than to trade to a less-preferred portfolio. If the asset returns were uncorrelated with symmetric covariances, the non-trading region would be a circle. In this example, everything is symmetric but there is correlation. The correlation means that the two assets are substitutes and it is not so bad if we have too little of one asset if we have too much of the other. As in the case of proportional costs, this is why we are quicker to trade if we are over-weighted in both assets than if we are over-weighted in one and under-weighted in the other.

2.3.3 Example 3: Asset-Specific Fixed Costs

What may be more plausible than an overall fixed cost is a fixed cost for each security. Arguably, a security-specific fixed cost comes from a due diligence requirement to monitor or document any security in the portfolio.

Figures 2.3 illustrate the inaction region in the presence of asset specific fixed costs. Near the ideal point in the middle, θ^*, is the non-trading region. The North-West and South-East corners lie on an ellipse reminiscent (but in general not equal to) of the no-trading region of an overall fixed cost problem while the pairs of parallel outer thick black dashed straight lines and the two intersecting blue and red lines come from the asset-specific component of the fixed costs. For analogous reasons as those discussed in the overall fixed cost example, for any initial position inside the region defined by the above mentioned corners and dashed line it is optimal not to trade at all. Outside of this area we trade as follows: From the regions in the corners, we trade both securities to the ideal point θ^*. From the regions on the right and left, we trade Security 1 but not Security 2 to the thick blue line going through the ideal point. This would be a vertical line if we had no correlation, but has negative slope in our case. Similarly, from the regions above and below, we trade only Security 2 to the thick red line running through the ideal point.

2.3.4 Example 4: Overall Fixed and Asset-Specific Proportional Costs

Figure 2.4 shows the different (no-)trading regions for the case in which the investor both pays an overall fixed cost to enter the market as well as asset specific proportional costs

to (re)balance his portfolio to the desired allocation. The parameters, except for the fixed cost $K = 0.00075$, are those of Figure 2.1 and the inner parallelogram is indeed the same: this is the region where the investor will optimally trade to *if* the fixed cost of entering the market is not too big. There is in fact a threshold, pictured as a dash-line, surrounding the proportional no-trading region; If the initial position θ^0 is inside this line the dis-utility from the fixed cost to enter the market is more than the benefit of trading to the border of the no-trading region (inner parallelogram) so that it is optimal not to trade at all. Outside of this line it is optimal to trade up to the border of the no trading region as in the proportional only case: in particular, there are four corridors and four corners delimited by wavy dashed lines. Inside each corridor it is optimal to trade (buy or sell) along straight lines only one asset at a time, while at each corner it is optimal to trade (buy/sell) two assets until the nearest no-trading region corner is reached.

2.3.5 Example 5: Asset-Specific Fixed and Proportional Costs

A similar but slightly more complex pattern emerges when we look at the the case of asset-specific fixed and proportional costs. Figure 2.5 visualizes of the (no-)trading regions for the case of both asset specific fixed and proportional costs. The parameters, except for the fixed cost $K = 0.00075$, are those of Figure 2.1 and the inner parallelogram is indeed the same: this is the region where the investor will optimally trade to *if* the fixed costs are not too big. Each fixed cost, albeit asset-specific, is taken as identical in the figure. The presence of asset specific costs rather than an overall fixed cost, on top of the asset specific proportional components, has similar effects, and an analogous intuition, as those described in Figure 2.4 in the presence of an overall fixed cost in place of the asset-specific fixed ones with two notable differences: 1 - the corridors inside which only one asset at time is traded up to the

border of the no trading region are wider pushing the corner regions further away from the unconstrained optimum. 2 - the thresholds where the investor is indifferent between paying the fixed costs and trade to the border of the no trading region or do not trade at all are different with respect to Figure 2.4 and feature a couple of intersections along the 45 degree line passing through the unconstrained optimum.

2.3.6 Example 6: Futures Overlay

In this example we suggest an improvement over traditional future overlay strategies.

It is increasingly common for plan sponsors to use futures as well as (or instead of) equities for managing exposure to market risk. One popular example of a transaction-cost-aware strategy is to use futures as an inexpensive way of keeping effective asset allocation in line with a benchmark or ideal allocation. For example, if we think the ideal weighting in equities is 60%, then as the market rises we become over-weighted and as the market falls we become under-weighted (since the fixed-income part of the portfolio moves less than proportionately with moves in the equity market). Maintaining a weighting near the ideal weighting by trading equities is very expensive. A futures overlay might correct for minor deviations from the ideal weighting by trading in futures, which are highly correlated with equities but much cheaper to trade.

For a futures strategy, we normally think that the correlation between the equity position and futures is close to one, so that holding futures and bonds is a close substitute for trading the underlying equities. We also usually believe that futures are much less expensive to trade than the underlying, which is why it is appealing to consider substituting futures trades for

trades in equities. The expected returns ("alphas") are however not usually discussed much, but they turn out to be very important.

Generally, we might expect the return on the underlying equities to be higher than the return on synthetic equity due to the benefits of active management. Or, moving somewhat outside the model, the extra return on the underlying may be due to the cost of rolling the futures or the tax-timing advantages to equity. Figure 2.6 illustrates an example in which equities, Security 1, have a significantly higher expected return than the synthetic equity strategy using futures, Security 2. In this case, there is a trade-off between transaction costs and expected return and it is optimal to use futures to substitute for trading in the underlying only for some trade. In practice, most plan sponsors use a "symmetric" futures overlay that uses futures to the same extent for correcting over- and under-exposure to the market. However, the analysis here prescribes an asymmetric strategy that makes good economic sense. If the market exposure must be reduced, we sell futures,[54] which allows us to keep the extra return on the underlying equities. On the other hand, if the market exposure must be increased, we buy equities, which have the extra return, rather than futures, which don't.

2.3.7 Example 7: Bundles

If it is possible to purchase a bundle of securities more cheaply than its constituents, then it might be possible to shrink the non-trading region and get closer, at least in some directions, to the unconstrained optimum.

In the presence of proportional costs only Figure 2.7 illustrates an example identical to that in Figure 2.1 except the additional opportunity of buying or selling a 50-50 mix of the two

[54]This is the normal situation but extreme cases may be different. For example, in Figure 2.6 we would sell equities rather than futures if we found ourselves 210% long equities and 140% short futures.

securities with a transaction cost of 0.0025. The dashed red lines represents the no trading region in the absence of the bundle, while the solid lines are the new region boundaries.

Adjacent the region in which we sell the bundle, there are regions where we sell the bundle and one of the securities. Similarly, adjacent the region in which we buy the bundle, there are regions where we buy the bundle and one of the securities. To keep the graph simple, we have considered a case with only two underlying securities, but bundles trading would obviously be more useful and more interesting with many securities. For example, trading a bundle might be a cost-effective way of aligning market or sector exposures with a target. With many securities in the bundle, the trading pattern could be more complex, for example, with simultaneous buys and sells of different individual securities to compensate for imbalances caused by trading the bundle.

In the presence of fixed costs we can reach a similar conclusion: Figure 2.8 compares the inaction regions of Example 2.4 (red) and 2.5 (blue). Even if no explicit bundle of assets is available, for situations in which it is optimal to simultaneously trade both assets (in case of negative correlation the no trade regions stretches more along the 45 line) one can think of the overall cost as the fixed cost for accessing the bundle consisting of asset 1 and 2 thus coming closer to the unconstrained optimum.

2.3.8 Example 8: Following a Benchmark with Proportional and Fixed Costs

This last application confirms and generalizes the i.i.d. approximate framework of Leland (2000): found managers' performances are usually evaluated relative to a benchmark, θ^B, and price fluctuations generate an additional trade-off. While previous examples show that

for positions that are not too distant from the ideal portfolio it is optimal not to trade, if managers' performance are evaluated against a benchmark portfolio there might be additional incentives to trade sooner to keep the strategy relatively near the benchmark. Leland (2000) studies this situation in a dynamic setting where returns are i.i.d, in the presence of proportional costs only and in the absence of standard preferences. Our one-period framework is able to put such analysis in the context of (explicit) mean-variance preferences and a more general cost function (featuring a fixed component on top of the asset-specific proportional costs).

Figure 2.9 shows two typical such situations: with proportional and an overall fixed cost (upper plot) and with proportional and asset specific costs (bottom plot). By comparing these graphs with the equivalent formulations in the absence of a benchmark, Figure 2.4 and 2.5 respectively, we notice how the trade-off works: the presence of the benchmark θ^B shrinks and shift the no trading regions towards the benchmark.

2.4 Analytical Characterizations

In this section we formally analyze the different models introduced earlier. The following are the set of weak technical assumptions that our framework requires in order to be well-defined:

A1: the variance-covariance matrix, V, of risky returns net of any liquidation costs is positive definite[55]

[55]This assumption might be relaxed in an even more flexible setup able to handle all the modeling situations that require perfect positive or negative correlation between assets. Important examples include the design of a model to study the closed-end fund puzzle or the more applied need to model situations in which the same asset is traded in different exchanges (e.g. as an ordinary stock in Country X and as an A.D.R. in the U.S.). This objective is left for future research.

A2: the risk and tracking error aversion parameters λ and κ are non-negative and at least one of them is strictly positive

A3: the cost function $c(\theta, \theta^0) \equiv c(P, S, \theta^0)$ is

$$c(P,S,\theta^0) = \begin{cases} P'C^P + S'C^S \text{ with } C^P > 0, C^S > 0 \text{ if costs are proportional} \\ 1_{[\theta \neq \theta^0]} k \text{ with } k > 0 \text{ if cost is overall fixed} \\ \sum_i^n 1_{[\theta_i \neq \theta_i^0]} k_i \text{ with } k_i > 0 \text{ if costs are asset-specific fixed} \\ P'C^P + S'C^S + 1_{[\theta \neq \theta^0]} k \text{ if costs are proportional and overall fixed} \\ P'C^P + S'C^S + \sum_i^n 1_{[\theta_i \neq \theta_i^0]} k_i \text{ if costs are proportional and asset-specific fixed} \end{cases}$$

where $P \geq 0$ is the n-dimensional vector of risky assets' purchases, $S \geq 0$ is the n-dimensional vector of risky assets' sales, C^P is the n-dimensional vector of proportional purchase costs and C^S is the n-dimensional vector of proportional sale costs.

The cost function detailed in $A3$ covers the typical cost structures used in the literature. Note that such cost function is defined in terms of P and S rather then asset weights θ, as part of the proof of Theorem 2 below we show that a necessary condition for any solution θ is that $c(P, S, \theta^0) = c(\theta, \theta^0)$. That is, the cost function can equivalently be represented as a (piece-wise linear) function of the n-dimensional vector of risky weights θ.

Under these assumptions we can proof the existence of a solution θ^* for the class of models having objective function defined by (2.1)

Theorem 2 *Given* $A1 - A3$ *the problem*

$$\max_{\theta \in \mathbb{R}^n} U(\theta)$$

has solution $\theta^* = \{\theta^{PC}, \theta^{\setminus PC}\}$. θ^{PC} *is the unique solution if costs are only proportional, for all other cases the solution,* $\theta^{\setminus PC}$*, is still unique for initial positions outside or inside the no trading region. For initial positions* exactly on *the borders of the no-trading region the investor will choose* $\theta^{\setminus PC}$*, depending on the actual costs composition, as either* $\hat{\theta}^{PC}$ *(the solution to a proportional costs problem in which trades are allowed only for a subset of the assets) or*

$$\theta^u = \frac{V^{-1}}{\kappa + \lambda}(\mu - r1 + \kappa V \theta^B) \tag{2.2}$$

which is the unique optimum in the absence of costs. Furthermore the solution θ^* *is such that* $U(\theta^*) \leq J(\theta^u)$ *where*

$$J(\theta) = r + \theta'(\mu - r1) - \frac{\lambda}{2}\theta' \mathbf{V} \theta - \frac{\kappa}{2}(\theta - \theta^B)' \mathbf{V} (\theta - \theta^B) \tag{2.3}$$

is the hypothetical utility function in the absence of costs.

Proof. See Appendix B.1 ■

The proof, in conjunction with the algorithm described in section 2.5, gives the recipe to numerically solve the class of problems covered in this paper. Computationally our framework scale up nicely with the number n of risky assets for which a solution is computationally feasible. In contrast with the literature, in which solutions are available only up to 2 or 3 risky assets, our one-period setup accommodates large number of assets, especially when costs are not asset-specific fixed.[56]

[56] In this case, since the solution is found by enumerating the potential alternative investments resulting from any possible combination of the available assets, the number of sub-problems to solve grows at a rate of 2^n with the number n of risky assets, limiting the applicability to a number n of the order of 20.

The rest of this section covers the in-depth characterization of the solutions for the different class of proposed models.

2.4.1 Proportional Costs

The Problem

The cost function is defined as $c(\theta, \theta^0) \equiv P'C^P + S'C^S$. The investor has to choose for each risky asset by how much to increase or decrease the initial position θ^0: the n chosen increments are stored in the vector of risky purchases, P, while the n chosen decrements are stored in the vector of risky sales, S. Incrementing the assets' positions by P entails a cost, expressed in return units, of $P'C^P$: i.e. each long trade of risky asset i with size P_i is taxed at a rate of C_i^P. Analogously decreasing the assets' positions by S entails a cost of $S'C^S$: each short trade of risky asset i with size S_i is taxed at a rate of C_i^S. Thus, different assets can have different costs for purchasing and selling, which are paid at end-of-period (or equivalently are measured in future value units). Since utility is also measured in end-of-period return units, marginal utilities and costs are in identical units.

The problem can be formally described as

Problem 3

$$\max_{P,S} \quad U(\theta) = r + \theta'(\mu - r1) - \frac{\lambda}{2}\theta'\mathbf{V}\theta - \frac{\kappa}{2}(\theta - \theta^B)'\mathbf{V}(\theta - \theta^B) - P'C^P - S'C^S$$

subject to

$$\theta = \theta^0 + P - S$$

$$P \geq 0$$

$$S \geq 0$$

where we use the following notation:

θ^0*:* $n \times 1$ *vector of initial risky asset weights*

P : *vector of risky assets' purchases*

S : *vector of risky assets' sales*

r*: risk-free rate of interest*

μ*:* $n \times 1$ *vector of expected risky asset rates net of any liquidation costs*

λ*: risk aversion parameter*

κ*: tracking error parameter*

$\mathbf{V}$*:* $n \times n$ *covariance matrix of risky rates net of any liquidation costs*

θ^B*:* $n \times 1$ *vector of benchmark portfolio weights*

C^P*:* $n \times 1$ *vector of proportional transaction costs for purchases*

C^S*:* $n \times 1$ *vector of proportional transaction costs for sales.*

Characterization of the solution

This subsection derive by construction the unique solution, θ, in the space of asset weights.

The utility function $U(\theta)$ and the constraint vector $\theta = \theta^0 + P - S$, which yield the end-of-period portfolio weights, θ, as a function of the chosen sales S and purchases P, define a quadratic programming which is easy to solve and in which any candidate solution (P, S) has

to satisfy the First Order Conditions (FOCs) of Problem 3.[57] We are assuming that transaction costs are the *only* source of market frictions. In practice, we could add nonnegativity constraints for portfolio positions, no-borrowing constraints or constraints on proportions in individual stocks or industries, and the problem would still be easy to solve, but including such considerations here would only be a distraction from our main message.

If we substitute in Problem 3 the constraints into the objective function we can write the Lagrangian as

$$L(P, S, \lambda_P, \lambda_S) = U(\theta) + \lambda_P' P + \lambda_S' S$$

then the Khun-Tucker (KT) conditions are given by

$$\frac{\partial L(P, S, \lambda_P, \lambda_S)}{\partial P} = \frac{\partial U(\theta)}{\partial P} + \lambda_P = 0 \tag{2.4}$$

$$\frac{\partial L(P, S, \lambda_P, \lambda_S)}{\partial S} = \frac{\partial U(\theta)}{\partial S} + \lambda_S = 0 \tag{2.5}$$

with the complementarity slackness conditions

$$\lambda_P' P = 0 \tag{2.6}$$

$$\lambda_S' S = 0 \tag{2.7}$$

[57] In the proof of Theorem 2, Problem 3 is rewritten in the equivalent standard quadratic programming form of

$$\max_x \quad \tilde{U}(x) = a'x + \frac{1}{2}x'Qx$$

subject to

$$x \geq 0$$

and the Hessian Q is shown to be Semi-Positive Definite implying that any candidate x satisfying the first order conditions is a solution of the problem.

and the Lagrange multipliers

$$\lambda_P \geq 0$$

$$\lambda_S \geq 0$$

In particular, due to the nonegativity of the Lagrange multipliers, equations (2.4) and (2.5) imply

$$\frac{\partial U(\theta)}{\partial P} \leq 0$$

$$\frac{\partial U(\theta)}{\partial S} \leq 0$$

once we substitute the actual partial derivatives we obtain

$$m \equiv \mu - r - \lambda \mathbf{V}\theta - \kappa \mathbf{V}(\theta - \theta^B)$$

satisfying for each risky asset i with $i = 1, ..., n$

$$m_i \in [-C_i^S, C_i^P] \tag{2.8}$$

the term m_i defines the marginal utility of holding asset i in the absence of transaction costs (therefore the marginal utility of shorting asset i is given by $-m_i$). At the unconstrained optimum such marginal utility should equal zero for any asset, however in the presence of proportional transaction costs equation (2.8) shows that for each asset it lies in in the compact interval $[-C_i^S, C_i^P]$ around zero instead. Remember that in our framework no trade in any risky asset i occurs without paying either C_i^S or C_i^P, thus m_i tells us that is optimal not to trade asset i if the marginal utility m_i lies in $[-C_i^S, C_i^P]$.

If we now substitute the definition of marginal utility in equations (2.4) and (2.5), solve for the Lagrange multipliers and plug them into equations (2.6) and (2.7), we obtain the revisited complementarity slackness conditions for each asset i

$$(m_i - C_i^P)P_i = 0 \tag{2.9}$$

$$(-m_i - C_i^S)S_i = 0 \tag{2.10}$$

let's focus our attention on equation (2.9): from equation (2.8) when $m_i - C_i^P \leq 0$ the marginal utility of holding an extra unit of asset i is smaller or equal to the marginal cost and we then know that $P_i = 0$ is optimal. On the other hand, when $m_i - C_i^P > 0$ we are outside of the no trading region, equation (2.8) is violated meaning we are not at optimum and there exist a choice of either $S_i > 0$ or $P_i > 0$ or both for some S_i and P_i such that the investor is better off. In particular $m_i - C_i^P > 0$ implies $-m_i - C_i^S < 0$ thus equation (10) implies $S_i = 0$. We infer that the only choice available to increase utility is to buy more of asset i, i.e. set $P_i > 0$. Note from the definition of m_i that it is continuous and decreasing in P_i, so that the investor has to keep on buying more and more of asset i until $m_i = C_i^P$ and condition (2.8) is satisfied. Thus equation (2.9) says that if we are not inside the no trading region because we are not holding enough of asset i we should buy more of it to the point in which the marginal benefit equal the marginal cost and we are on the border of the no-trading region. Similarly the other complementary condition (2.10) says that if we are not inside the no trading region because we are holding too much of asset i we should keep selling it to the point in which the marginal benefit of shorting, $-m_i$, equal the marginal cost C_i^S and again we are exactly on the border of the no-trading region.

The two conditions together imply that it is never optimal to buy and sell any asset i at the same time, i.e. either $S_i > 0$ or $P_i > 0$ but not both. Thus there exist only one combination $(P, S) \in \mathbb{R}^n_+ \times \mathbb{R}^n_+$ satisfying the FOCs of Problem 2.4.1 and the constraint $\theta = \theta^0 + P - S$: if this combination is feasible then $\theta \in \mathbb{R}^n$ would be the unique solution of Problem 2.4.1 in the space of asset weights. The only reason why (P, S) might not be feasible is if there exists at least one asset i for which setting P_i or S_i to $+\infty$ is optimal. By contradiction suppose there exist such an asset i: because $\lim_{P_i \to +\infty} \theta' V \theta = \lim_{S_i \to +\infty} \theta' V \theta = +\infty$, it follows that $U(\theta) = -\infty$. Thus $P_i = +\infty$ or $S_i = +\infty$ is never optimal.

The no-trading region and the optimal policy

Absent costs, a standard mean-variance problem with a nonsingular covariance matrix and strictly concave preferences over mean and standard deviation has a unique optimal portfolio θ^u and it is optimal to trade directly to the optimum whatever the initial portfolio θ^0. As a result, there is only a single starting point, the unique optimum θ^u, from which the agent would not trade. With transaction costs, however, there is a whole set of portfolios θ, including θ^u, from which there would be no trade. Although possibly not at the ideal portfolio θ^u, any trade from this region would generate a benefit too small to cover the transaction costs.

As shown above any optimal portfolio θ in $\mathbb{R}^n$ for which there is no trade has to satisfy the first order conditions: that is it has to be inside the interval defined by (2.8) for each i and in case it is outside for some component j (asset j) the complementary slackness conditions (2.6) and (2.7) tell us it needs to be pushed back at the boundary of the interval of asset j by either buying or selling more of asset j but not both. Thus any optimal portfolio θ satisfy

condition (2.8) for any asset i suitably re-arranged:

$$\mathbf{V}_i\theta \geq \frac{1}{\kappa+\lambda}(\kappa\mathbf{V}_i\theta^B + (\mu_i - r - C_i^P)) \tag{2.11}$$

$$\mathbf{V}_i\theta \leq \frac{1}{\kappa+\lambda}(\kappa\mathbf{V}_i\theta^B + (\mu_i - r + C_i^S)) \tag{2.12}$$

where V_i is the i-th row of V. The set of portfolios given by (2.11) defines a half-space such that it is always optimal not to buy asset i (it may be optimal in some part to actually sell asset i); this is because (2.11) is equivalent to the statement $m_i \leq C_i^P$. Similarly equation (2.12) defines a half-space in $\mathbb{R}^n$ such that it is never optimal to sell asset i (in some part is actually optimal to buy asset i). The intersection of these half-spaces for all asset i characterizes the no trading region. Thus, in contrast to other studies[58] that assume linear edges for the no trading region in the presence of proportional transaction costs, the linearity of the inaction region boundaries endogenously arises in our setup.

The no trading region is an n dimensional object, with each dimension associated to a risky asset i for $i = 1, ..., n$. Figure 2.10 illustrates the no trading region for the case of 2 and 3 risky assets. More generally, there are $2 \times n$ half-spaces, a pair for each asset i. Note that for a given asset i, conditions (2.11) and (2.12) with equality defines a pair of $n-1$ dimensional hyperplanes which, together with the inequalities' signs, fully characterize the respective half-spaces. In particular, the half-spaces are of the form $\mathbf{V}_i\theta + b_i^P \geq 0$ and $\mathbf{V}_i\theta + b_i^S \leq 0$ with scalars b_i^P and b_i^S being the negative of the RHS of (2.11) and (2.12). Written in this standard way it is easy to see that V_i defines the hyperplanes' orientations while the b coefficients characterize their directions in the space. Because the orientation is the same, the hyperplanes are parallel to each other and their relative distance is given by $\frac{|b_i^P - b_i^S|}{||\mathbf{V}_i||}$; simple algebra shows that $b_i^P - b_i^C = \frac{C_i^P + C_i^S}{\kappa+\lambda}$ which is strictly grater than zero given $A2$

[58] e.g. H. Liu (2004) and Leland (2000)

and at least one between C_i^P and C_i^P strictly positive, meaning that the hyperplanes do not overlap. Therefore, the opposite signs in (2.11) and (2.12) reveal that the non-overlapping hyperplanes define overlapping half-spaces pointing in opposite directions covering a non-empty corridor in $\mathbb{R}^n$. The convex hull generated by the intersection of this n corridors defines the boundaries of the no-trading region. The existence of such convex hull is due to the fact that the covariance matrix V is positive definite, assumption $A1$, and thus full rank. This way each of its row, V_i, is linearly independent and thus each hyperplane pair, and thus corridor, has a unique orientation implying that there will always be $2 \times n$ intersections originated from the n pairs of parallel half-spaces and 2^n intersections where n half-planes meet. The $2 \times n$ intersections represent the no-trading region edges, while the 2^n intersections are the region's corners. Finally, each of the $2 \times n$ half-spaces portions delimited by the $2 \times n$ edges originates the $2 \times n$ faces of the region.

We can also say more about the shape of this region: the $2 \times n$ edges are intersections of parallel hyperplanes thus are linear and symmetric. This also implies that the $2 \times n$ faces, being portions of the original half-spaces delimited by the edges, are linear too. Thus the no-trading region is a parallelogram in the presence of two risky assets and its analogous in higher dimensions.

In terms of trading rules (optimal policy) by carefully inspecting (2.11) and (2.12) we notice that each pair of parallel faces (edges excluded) is a function of a single asset only. The trading rule with respect to asset i, for each initial allocation lying outside the interior of such faces along the hyperplanes' normal vector, is to either buy or sell *exclusively* asset i until the new allocation lies *on* the nearest of the two faces; this is by far the most common trading that can ever occur since the faces of the no-trading region represent the majority of the entire region. On the other extreme, for allocations lying in the portion of $\mathbb{R}^n$ outside

a region corner defined by the intersection of the n hyperplanes forming that corner, a trade would entail either buying or selling (but not both at the same time) each of the n risky asset; this type of trade is the least common among all since the probability of being in the portion of $\mathbb{R}^n$ outside one of the 2^n corners is low. Finally, in-between the two extremes, are the situations in which the initial allocation is outside the trading region in the space generated by the intersections of up to $n-1$ hyperplanes, the ones defining the region's edges (corners exuded). In these cases the trading rules entail either buying or selling (but not both at the same time) up to $n-1$ risky assets, each asset associated with one of the intersecting hyperplanes.

In summary, the FOCs completely characterize the optimal trading rule: equation (2.8) for all assets n defines the no-trading region while conditions (2.6) and (2.7) tell us the optimal directions, along straight lines, in which the trades should occur in order to approach the boundaries of the no trading region.

How the shape of the no-trading region changes as the number of asset increases

Davis and Norman (1990) show how the no trading region for the case of only one asset is an interval on the real line. Our framework, via (2.11) and (2.12), formally confirms the linearity of its boundaries for the case of any arbitrary number of risky assets. In particular, as we saw in Figure 2.10, going from 1 to 3 assets makes the interval first to become a parallelogram and then a parallelepiped. In sharp contrast with H. Liu (2004), Figure 2.11 reminds us that, as long as the correlation among assets is not zero, we cannot reduce the dimensionality of a problem involving n assets to a problem involving $n-k$ assets and a problem involving k assets; in light blue is the slice of the no trading region evaluated in correspondence of the optimal unconstrained allocation for asset 3 while in red is the optimal

inaction region for the same problem only involving asset 1 and asset 2. In the H. Liu (2004) world, in the absence of correlation, the two parallelogram would become two rectangles (squares) perfectly overlapped (a claim we verified in our framework as well). Thus Figure 2.11 illustrates the inseparability of the problem once the correlations among assets are taken into account: this is because, as (2.11) and (2.12) show, the weight of any asset is a function of the covariance with itself and any other (risky) asset.

2.4.2 Fixed Costs

We consider two different models with fixed costs. In one case, there is a fixed cost for any change in position. In the other case, there is a cost for each risky security traded.

Overall fixed cost: The Problem

The assumption of an overall fixed cost, k, is that we incur a fixed cost of "going to the market". The cost function is defined as $c(\theta, \theta^0) \equiv 1_{[\theta \neq \theta^0]} k$ and the choice problem is

Problem 4

$$\underset{\theta \in \mathbb{R}^n}{max} \;\; U(\theta) = r + \theta'(\mu - r1) - \frac{\lambda}{2}\theta' \mathbf{V} \theta - \frac{\kappa}{2}(\theta - \theta^B)' \mathbf{V} (\theta - \theta^B) - 1_{[\theta \neq \theta^0]} k$$

Overall fixed cost: Characterization of the Solution

Although Problem 4 is a nonconvex problem, there are only two cases to consider and its solution is simple. If there is a trade, it is to the same ideal point, θ^u described in equation

(2.2), whatever the initial position. Therefore, the no-trade region can be computed as the set of points, portfolios such that $\theta = \theta^0$, where the improvement in value from going to the ideal point does not exceed the fixed cost. This set of portfolios is characterized by the equation

$$r + \theta'(\mu - r1) - \frac{\lambda}{2}\theta'\mathbf{V}\theta - \frac{\kappa}{2}(\theta - \theta^B)'\mathbf{V}(\theta - \theta^B) \geq$$
$$r + \theta^{u\prime}(\mu - r1) - \frac{\lambda}{2}\theta^{u\prime}\mathbf{V}\theta^u - \frac{\kappa}{2}(\theta^u - \theta^B)'\mathbf{V}(\theta^u - \theta^B) - K$$

and it is the area inside an ellipse (As we showed in Example 2) or its analogous counterpart when we deal with more than 2 risky assets.

Asset-specific fixed Costs: The Problem

What may be more plausible than an overall fixed cost is a fixed cost for each security. Arguably, a security-specific fixed cost comes from a due diligence requirement to monitor or document any security in the portfolio, although a serious consideration of this motivation probably leads us to informational or strategic considerations outside the current framework.[59]

The cost function is defined as $c(\theta, \theta^0) \equiv \sum_{i+1}^{n} 1_{[\theta_i \neq \theta_i^0]} k_i$ and the choice problem is

Problem 5

$$\max_{\theta} \quad U(\theta) = r + \theta'(\mu - r1) - \frac{\lambda}{2}\theta'\mathbf{V}\theta - \frac{\kappa}{2}(\theta - \theta^B)'\mathbf{V}(\theta - \theta^B) - \sum_{i+1}^{n} 1_{[\theta_i \neq \theta_i^0]} k_i$$

[59]For example, why would we have to monitor a position unless new information arrival is possible and subsequent trade is possible? Perhaps a regulator requires documentation of the trade and a due diligence study of the firm issuing each share of stock we hold, even though we know we are not going to learn anything from the exercise. Another question is why we don't have to do monitoring or due diligence on a stock we already hold and choose not to sell. Or, it may be that our broker offers to make any trade in a single maturity, whatever the size, for the same fixed price. It seems much easier to make an argument for why there are variable costs.

where k_i is the fixed cost incurred to trade (risky) security i

Asset-specific fixed Costs: Characterization of the solution

The solution to Problem 5 is more of a combinatoric problem, since each possible set of included portfolios gives a different piece of the overall nonconcave objective function. The existence theorem and the algorithm of Section 2.5 enable us to numerically solve this problem for $n > 2$ risky assets. However, in a particular small example we can analytically and graphically (Figure 2.12 shows the "architecture" behind Example 3) construct the solution, since the set of boundary points where two subsets are equally preferred is a conic section.

Next we characterize the solution for the case of $n = 2$ risky assets. There are four basic regions (which can also be subdivided by the direction of trade):

Region a**:** no trade

Region b**:** trade security 1 but not security 2

Region c**:** trade security 2 but not security 1

Region d**:** trade both securities

For all the cases, it simplifies the algebra to write the objective function in terms of deviations from the ideal portfolio θ^u as we do in the proof of Theorem 2 and, in order to convey the main intuition, set $\kappa = 0$ as well. Let $\gamma = \theta - \theta^u$ (and $\gamma^0 = \theta^0 - \theta^u$), then we can rewrite the objective of Problem 5 as

$$U(\gamma) = U^u - \frac{\lambda}{2}\gamma' \mathbf{V} \gamma - \sum_{i+1}^{n} 1_{[\gamma_i \neq \gamma_i^0]} k_i$$

where

$$U^u \equiv r + \frac{1}{2\lambda}(\mu - r1)'V^{-1}(\mu - r1) \tag{2.13}$$

In region a, there is no trade so the cost is zero, and the value is the value of the initial position θ^0:

$$U_a = U^u - \frac{\lambda}{2}\gamma^{0\prime}\mathbf{V}\gamma^0$$

this is a strictly concave quadratic function of γ with a maximum of U^u achieved at $\gamma = 0$.

In region b, we trade security 1, incurring a cost k_1, and the value is

$$U_b = \max_{\gamma_1}\{U^u - \frac{\lambda}{2}(V_{11}\gamma_1^2 + 2V_{12}\gamma_1\gamma_2^0 + V_{22}\gamma_2^2) - k_1\} = U^u - \frac{\lambda}{2}\left(V_{22} - \frac{V_{12}^2}{V_{11}}\right)\gamma_2^{0\,2} - k_1$$

where $\gamma_1 = -(V_{12}/V_{11})\gamma_2$ achieves the maximum. In this case, the value function is strictly concave in γ_2 and constant in γ_1. It achieves a maximum of $U^u = -k_1$ on the line $\gamma_2 = 0$.

Region c is symmetric to Region b but with the securities swapped, so we have

$$U_c = U^u - \frac{\lambda}{2}\left(V_{11} - \frac{V_{12}^2}{V_{22}}\right)\gamma_1^{0\,2} - k_2$$

In region d, we go to the ideal point and incur both costs k_1 and k_2, and the value is

$$U_d = U^u - k_1 - k_2$$

the value is constant independent of the starting position γ^0.

Having computed the values in each region, it is straightforward to compute the candidate boundaries, where $U_a = U_b$, $U_a = U_c$, etc. For the example in Figure 2.12, all these candidate

boundaries are lines except the locus $U_a = U_d$ which is an ellipse. (In general, even in many dimensions, the boundary solves a quadratic or linear equation.) Near the ideal point (marked as a red point), not trading is optimal so we mark these regions with a. If we start in a region a and cross a boundary not involving a (say $U_b = U_c$), we remain in region a. However, if we cross a boundary involving a (say $U_a = U_d$), then we switch regions (in this case to d). (In principle, there could be a degenerate case in which two regions are equal on the boundary but the same is better on both sides. However, this does not happen in our current example.) Going through this exercise confirms the regions in Figure 2.3.

2.4.3 Fixed and Proportional Costs

We now combine the previous analysis to characterize two types of more complex, (perhaps) more realistic, settings. One in which there is an overall fixed cost on top of asset-specific proportional costs and another where also the fixed component becomes asset-specific. We start by analyzing the asset-specific framework in that the overall fixed setup is a special case and, while discussing the latter, we make a comparison between the two frameworks.

Asset-Specific fixed and Proportional Costs: The Problem

The cost function is defined as $c(\theta, \theta^0) \equiv P'C^P + S'C^S + \sum_{i+1}^{n} 1_{[\theta_i \neq \theta_i^0]} k_i$ and the choice problem is

Problem 6

$$\max_{P,S} \; U(\theta) = r + \theta'(\mu - r1) - \frac{\lambda}{2}\theta'\mathbf{V}\theta - \frac{\kappa}{2}(\theta - \theta^B)'\mathbf{V}(\theta - \theta^B) - P'C^P - S'C^S - \sum_{i+1}^{n} 1_{[\theta_i \neq \theta_i^0]} k_i$$

subject to

$$\theta = \theta^0 + P - S$$

$$P \geq 0$$

$$S \geq 0$$

Asset-Specific fixed and Proportional Costs: Characterization of the solution

As in the case of Problem 5, the presence of asset-specific fixed costs makes Problem 6 more of a combinatorial problem. The existence theorem and the algorithm of Section 2.5 enable us to numerically solve this problem for $n > 2$ risky assets. Nonetheless, in order to convey the main intuition, it is useful to analytically and graphically solve the problem for the case of $n = 2$ risky assets.

For the ease of exposition we set κ to 0[60] and rewrite the problem in positions $\gamma = \theta - \theta^u$ relative to the unconstrained optimum θ^u defined in (2.2) as

$$U = U^u - \frac{\lambda}{2}\gamma'\mathbf{V}\gamma - P'C^P - S'C^S - \sum_{i+1}^{n} 1_{[\gamma_i \neq \gamma_i^0]} k_i$$

where U^u is defined as in (2.13).

Analogously to Figure 2.12, Figure 2.13, shows the steps to construct Example 5 (i.e. Figure 2.5). Each asset i now have a fixed and a proportional component; The fixed part is the sunk cost the investor has to pay to be allowed to trade that asset while the proportional part is

[60]The steps for solving the problem with $\kappa > 0$ are the same and yield a smaller no trading region with the same shape only shifted towards the target portfolio θ^B. This is because the optimization takes now into account, in the spirit of Leland (2000), the trade off between trading costs (a smaller no trade region) and tracking error benefits (a post-trade position closer to the target θ^B). This is what we discussed in Example 8 and shown in bottom graph of Figure 2.9.

the constant rate[61] charged to the trade that moves the initial position of asset i, θ_i^0, to the post-trade position θ_i. The longer the distance the higher the impact of the proportional component. Thus from any initial position θ^0 enough far away from the unconstrained optimum, the red dot in Figure 2.13, the fixed component of any asset is negligible with respect to the proportional one. This tells us that from any such initial position we are just solving a proportional asset specific problem. As a matter of fact the FOCs of Problem 6 are identical to those of Problem 3 and define the inner parallelogram around the unconstrained optimum in Figure 2.13. We therefore know that, for any given asset i trade, it is optimal either to buy or to sell but not both, trades outside the inner parallelogram edges are straight lines up to the closest edge involving only one asset at a time and trades outside of corners involve two assets at a time and always end up at the closest corner.

As we move to initial positions θ_0 closer to the unconstrained optimum, the impact of fixed over proportional costs increase up to the point where the investor is indifferent between paying the fixed costs and trade to the closest boundary of the inner parallelogram or do not trade.

Suppose the initial position θ^0 is somewhere outside the left edge of the inner no trading region, let U^{P_1} be the investor utility of paying the fixed cost k_1 and buy additional units of asset 1 only ($\gamma_2 = \gamma_2^0$) up to the left edge, i.e.

$$\begin{aligned} U^{P_1} &\equiv \max_{\gamma_1}\{U^u - \frac{\lambda}{2}[\gamma_1, \gamma_2^0]\mathbf{V}\begin{bmatrix}\gamma_1 \\ \gamma_2^0\end{bmatrix} - (\gamma_1 - \gamma_1^0)C_1^P - k_1\} \\ &= U^u - \frac{\lambda}{2}(V_{11} - \frac{V_{12}^2}{V_{11}})(\gamma_2^0)^2 + \frac{V_{12}^2}{V_{11}}C_1^P\gamma_2^0 + C_1^P\gamma_1^0 + \frac{(C_1^P)^2}{2\lambda V_{11}} \end{aligned}$$

[61] C_i^S for a sell trade and C_i^P for a buy trade.

where we used the fact that either $S_1 > 0$ or $P_1 > 0$ but not both and $\theta_1 = \theta_1^0 + P_1 - S_1$. Recall that, exactly as in Problem 5, the utility of not trading is

$$U^{NT} = U^u - \frac{\lambda}{2}\gamma^{0\prime}\mathbf{V}\gamma^0$$

The locus of points θ_0 where the investor is indifferent between paying the fixed cost k_1 and trade to left inner edge or do not trade is given by $U^{P_1}(\gamma_0) = U^{NT}(\gamma_0)$. (which can be written as a function of θ^0 via $\theta^0 = \gamma_0 + \theta^u$) This corresponds to the leftmost dashed blue line in Figure 2.13. To the left of this line the investor is better off trading up to the parallel thick blue line to the immediate right, which defines the left edge of the inner parallelogram. For initial positions θ_0 to the right of the leftmost dashed blue line, the fixed cost k_1 is too high to enter any trade so that the investor does not move.[62]

Suppose now the initial position θ^0 is somewhere outside the right edge of the inner no trading region, let U^{S_1} be the investor utility of paying the fixed cost k_1 and sell additional units of asset 1 only ($\gamma_2 = \gamma_2^0$) up to the right edge, i.e.

$$U^{S_1} \equiv \max_{\gamma_1}\{U^u - \frac{\lambda}{2}[\gamma_1, \gamma_2^0]\mathbf{V}\begin{bmatrix}\gamma_1 \\ \gamma_2^0\end{bmatrix} - (\gamma_1^0 - \gamma_1)C_1^S - k_1\}$$

where we used the fact that either $S_1 > 0$ or $P_1 > 0$ but not both and $\theta_1 = \theta_1^0 + P_1 - S_1$. The locus of points θ_0 where the investor is indifferent between paying the fixed cost k_1 and trade to right inner edge or do not trade is given by $U^{S_1}(\gamma_0) = U^{NT}(\gamma_0)$. (which can be written as a function of θ^0 via $\theta^0 = \gamma_0 + \theta^u$) This corresponds to the rightmost dashed blue line in Figure 2.13. To the right of this line the investor is better off trading up to the inner bold

[62]The fact that to the left we trade up to the inner edge and to the right we do not move and not the other way around follows from the fact that the utility increases from every direction as we move closer to the unconstrained optimum.

blue line defining the right inner edge. For initial positions θ_0 to the left of the rightmost dashed blue line, the fixed cost k_1 is too high to enter any trade so that the investor does not move.[63]

The loci of points θ_0 where the investor is indifferent between paying the fixed cost k_2 and buy additional units of asset 2 up to the South inner edge or do nothing, $U^{P_2}(\gamma_0) = U^{NT}(\gamma_0)$, and where the investor is indifferent between paying the fixed cost k_2 and sell additional units of asset 2 up to the North inner edge or do nothing, $U^{S_2}(\gamma_0) = U^{NT}(\gamma_0)$ are derived in a symmetric fashion with the two securities swapped and are shown by the two parallel dashed red lines in figure 2.13. Analogous arguments as above show that outside this lines the investor is better off trading up to the closest thick inner red border while in the corridors between the two dashed red lines is better not to trade at all.

What happens at and outside corners is what is still left to complete the picture. Each corner is characterized by three elements: (i) the locus of points θ_0 where the investor is indifferent between trading in the two assets simultaneously (e.g. the North-West corner involve the additional purchase of asset 1 and the additional sale of asset 2) and go to the nearest inner corner (one of the intersection of the solid blue and red lines) or do not trade, which is an ellipse, and (ii) and (iii), the horizontal and vertical black dashed lines defining the loci of points θ_0 where the investor is indifferent between trading the two assets to the nearest inner corner or trade only one asset to the nearest inner edge.

Let us focus on the North-West corner first: define U^{P_1,S_2} as the investor's utility of trading from any (feasible) initial position θ_0 to the inner North-West corner (the intersection of the

[63] A similar argument to that of the previous note proof that this is indeed the right thing to do.

left solid blue line with the upper thick red line), θ^{Γ}

$$U^{P_1,S_2}(\gamma_0) \equiv U^u - \frac{\lambda}{2}\gamma^{\Gamma\prime}\mathbf{V}\gamma^{\Gamma} - (\gamma_1^{\Gamma} - \gamma_1^0)C_1^P - (\gamma_2^0 - \gamma_2^{\Gamma})C_2^S - k_1 - k_2$$

the locus of points θ_0 where the investor is indifferent between trading in the two assets simultaneously (additional purchases in asset 1 and additional sales in asset 2) and go to θ^{Γ} or do not trade is $U^{P_1,S_2}(\gamma_0) = U^{NT}(\gamma_0)$ (written as a function of θ^0 via $\theta^0 = \gamma_0 + \theta^u$) which corresponds to the North-West ellipse in Figure 2.13. Inside the ellipse the investor is better off not to trade while outside it is optimal to trade up to θ^{Γ}. The uppermost horizontal dashed black line is the locus of points such that the investor is indifferent between trading asset 1 and 2 to θ^{Γ} or buy additional units of asset 1 until the inner left solid blue line, i.e. $U^{P_1,S_2}(\gamma_0) = U^{P_1}(\gamma_0)$ (written as a function of θ^0 via $\theta^0 = \gamma_0 + \theta^u$). Below this line it is better to only buy asset 1 up to the inner left bold blue line while above the best option is to buy more of asset 1 and sell more of asset 2 and go to θ^{Γ}. Note that this line marks the end of the inner left bold solid blue line as well as the outer left bold dashed blue line.

Similarly the leftmost vertical dashed black line is the locus of points such that the investor is indifferent between trading asset 1 and 2 to θ^{Γ} or sell additional units of asset 2 until the inner North solid red line, i.e. $U^{P_1,S_2}(\gamma_0) = U^{S_2}(\gamma_0)$ (written as a function of θ^0 via $\theta^0 = \gamma_0 + \theta^u$). To the right of this line it is better to only sell asset 2 up to the inner North bold red line while to the left the best option is to buy more of asset 1 and sell more of asset 2 and go to θ^{Γ}. Note that this line marks the end of the inner North bold solid red line as well as the outer North bold dashed red line.

The piece of North-West ellipse in-between the above mentioned vertical and horizontal lines as well as the part of the horizontal line to the left of the ellipse and the part of the vertical

line above the ellipse define the North-West corner and are marked in bold dashed black. From any point θ^0 further North-West it is optimal to buy asset 1 and sell asset 2 up to θ^Γ, (the North-West corner of the inner parallelogram) from any point θ^0 below the horizontal black bold dashed portion it is optimal to trade up to the left inner bold blue line, from any point θ^0 to the right of the vertical black bold dashed portion it is optimal to trade up to the inner North bold red line; Finally, from any point θ^0 in the approximately triangular region delimited by the intersection of the horizontal and vertical black dashed lines and the piece of ellipse between these two lines is better not to trade.[64]

A symmetrical analysis where the two securities are swapped analogously describe the South-East corner of Figure 2.13.

Next, let us focus on the South-West corner: due to the positive correlation of asset 1 and asset 2, the second lowermost horizontal black line (the thick one defined by $U^{P_1,P_2}(\gamma_0) = U^{P_1}(\gamma_0)$) and the second leftmost black vertical line (the thick one defined by $U^{P_1,P_2}(\gamma_0) = U^{P_2}(\gamma_0)$) intersects outside[65] the South-West ellipse (implicitly defined by $U^{P_1,P_2}(\gamma_0) = U^{NT}(\gamma_0)$). This is also a feature of the North-East corner and it is in opposition with what happens at the other two corners. If the correlation where negative the opposite would have occurred. As in other corners note how the horizontal dashed black line marks the South end of the inner left solid blue line as well as the South end of the outer left solid dashed bold blue line, while the vertical line marks the left end of the inner South solid red line as well as the left end of the outer South dashed bold red line. This time the corner is defined by a

[64]The reader might wonder why the corner includes the piece of ellipse in-between the two bold dashed black lines rather than the portion of the outer left dashed blue line up to the intersection with the outer North dashed red line and that red line up to the intersection with the ellipse: the reason why this is not the case is because the region above the horizontal dashed line and to the left of the vertical dashed line is where it is optimal to trade both assets while the outer left dashed blue line or the outer North dashed red line only involve one asset at a time.

[65]Not inside as in the respective cases of the North-West and and South-East corners.

dashed bold upside down reflected L. For any point θ_0 more South-East it is optimal to both buy more of asset 1 and asset 2 and go to the South-East corner of the inner parallelogram (defined by the intersection of the inner left solid bold blue line and the inner South solid bold red line), θ^L, for any point θ_0 above the bold portion of the horizontal dashed line it is optimal to only buy asset 1 until the inner left bold solid blue line is reached, for any point θ_0 to the right of the bold portion of the vertical dashed line is optimal to buy more of asset 2 until the inner South bold solid red line. Finally, for any point in the quadrilateral region formed by the intersection of the horizontal and vertical dashed lines, the vertical line and the outer South dashed red line, the outer South dashed red line and the the outer left dashed blue line, and the outer left dashed blue line with the horizontal dashed black line, it is optimal not to trade.

A symmetrical analysis where the two securities are both sold instead of bought analogously describes the North-East corner of Figure 2.13.

This analysis verified the optimal policy graphically described in Figure 2.5.

Overall fixed and Proportional Costs: The Problem

The cost function is defined as $c(\theta, \theta^0) \equiv P'C^P + S'C^S + 1_{[\theta \neq \theta^0]}k$ and the choice problem is

Problem 7

$$\max_{P,S} \quad U(\theta) = r + \theta'(\mu - r1) - \frac{\lambda}{2}\theta'\mathbf{V}\theta - \frac{\kappa}{2}(\theta - \theta^B)'\mathbf{V}(\theta - \theta^B) - P'C^P - S'C^S - 1_{[\theta \neq \theta^0]}k$$

subject to

$$\theta = \theta^0 + P - S$$

$$P \geq 0$$

$$S \geq 0$$

Overall fixed and Proportional Costs: Characterization of the solution

The existence theorem and the algorithm of Section 2.5 enable us to numerically solve this problem for $n > 2$ risky assets. Nonetheless, in order to convey the main intuition, it is useful to analytically and graphically solve the problem for the case of $n = 2$ risky assets.

For the ease of exposition we set κ to 0[66] and rewrite the problem in positions $\gamma = \theta - \theta^u$ relative to the unconstrained optimum θ^u defined in (2.2) as

$$U = U^u - \frac{\lambda}{2}\gamma' \mathbf{V} \gamma - P'C^P - S'C^S - 1_{[\gamma \neq \gamma^0]}k$$

where U^u is defined as in (2.13). In light of the previous subsection, solving this problem is easy and a graphical representation is provided in Figure 2.14 (which provides the structure behind Example 4 - i.e. Figure 2.4) for the case of positive correlation between asset 1 and 2.

The steps and the intuition are exactly the same as those of the previous section; The inner parallelogram and the trades properties are the same since the FOCs are the same and we still have a locus of points θ_0 such that the investor is indifferent between paying the fixed cost to enter the market and trade up to the closest boundary of the inner parallelogram or do not trade. What is a bit different is the geometry of this locus of points.

[66]The steps for solving the problem with $\kappa > 0$ are the same and, for the same reasons discussed in the asset-specific case, yield a smaller no trading region with the same shape only shifted towards the target portfolio θ^B as discussed in Example 8 and shown in the upper graph of Figure 2.9.

While the utility of not trading is always the same, for any trade from initial positions θ_0 entailing only one asset at a time the investor's utility of paying the fixed cost k and going to the closest edge of the inner parallelogram, U^{T_i} with $T \in \{P, S\}$, is the same as the asset-specific fixed costs except the fact that k replaces k_i. For any trade from initial positions θ_0 involving both assets the investor's utility of paying the fixed cost k and going to the closest corner of the inner parallelogram, U^{T_i,R_j} with $T, R \in \{P, S\}$ and $i, j \in \{1, 2\}$, is the same as Problem 5 except the fact that k replaces $k_1 + k_2$.

It follows that the locus of points θ_0 such that $U^{P_1} = U^{NT}$ and that implied by $U^{S_1} = U^{NT}$ are the two parallel outer blue dashed lines while the locus of points θ_0 such that $U^{P_2} = U^{NT}$ and that implied by $U^{S_2} = U^{NT}$ are the two parallel outer red dashed lines. Outside of the bold dashed portion of the corridors formed by each pair of parallel lines it is optimal to trade up to the closest inner parallelogram edge, while inside those corridors the investor is better off not trading.

With respect to corners the situation for the North-West and South-East is analogous to that of Problem 6, while that concerning the South-West and North-East corner is a bit different. What is different is that all the intersections of the horizontal and vertical dashed lines that implicitly define the loci of points such that the investor is indifferent between trading one or two assets are exactly at the corners of the inner parallelogram as in Problem 3 rather than outside as in Problem 6. The fact that such intersections are inside the North-West and South-East ellipses defining the loci of initial positions θ_0 where $U^{P_1,S_2} = U^{NT}$ and $U^{S_1,P_2} = U^{NT}$ is consistent with the asset specific fixed costs case and result in qualitatively similar corners' shapes. The latter fact is nonetheless in contrast with what happens in the South-West and North-East corners where the intersections of the horizontal and vertical lines also occurs outside of the South-West and North-East ellipses in the present setup.

Another remarkable difference is given by the fact that the area covered by the ellipses at corners is now much wider with respect to the one covered by the regions where it is optimal to trade only one asset at a time. In relative terms the present structure of the costs favors the trades in both assets because the investor only need to pay the fixed cost once. This last important point is made even clearer in Figure 2.15 panel (a).

Panel (a) of Figure 2.15 simultaneously plots the inaction regions for the asset specific fixed and proportional cost (in blue) and for the overall fixed and asset-specific proportional costs (in red) where all the fixed costs are set to 0.00075. The fact the the red region is contained in the blue one[67] reinforces our intuition.

Thus, as expected, given all fixed cost are the same, in the presence of asset specific fixed costs rather than an overall fixed one the investor can only do worse except in the overlapping portion of regions entailing trades in only one asset at a time. Interestingly, as graphed in panel (b) of Figure 2.15, the exact opposite occurs if the investor has the chance of trading the same assets for half of the overall fixed cost; In this case the investor, regardless the correlation[68] is always better off trading each asset with a specific fixed cost which is half of the overall fixed one and is indifferent in the overlapping portion of corners involving the simultaneous additional purchases of one asset and additional sales of the other if the correlation is positive or the simultaneous additional purchases or sales of both if the correlation is negative. When the correlation is zero the investor is indifferent only at corners (a zero probability event).

We conclude this subsection with a last comparison between the overall and asset specific fixed costs in the presence of proportional cost. Panel (c) of Figure 2.15 shows a situation

[67] A result which is independent from the correlation structure. (analogous comparisons in the case of zero and negative correlations give the same result.)

[68] Analogous comparisons in the case of zero and negative correlations give the same result.

which is in-between those illustrated in the two previous panels: we compare the no trading regions when the fixed cost of asset 1, k_1, and asset 2, k_2, are 75% of the overall fixed cost. This time there is no scenario which dominates, it is better to have a lower asset specific cost when it is optimal only to trade one asset while it is better to have the higher overall fixed cost when it is optimal either to simultaneously buy and sell the two assets if the correlation is positive, or simultaneously buy or sell both assets if the correlation is negative. When the correlation is zero it always better to have the higher overall fixed cost while trading simultaneously the two assets.

As we discussed in Example 7, the benefits described in the first and last situation here are similar to those described in the presence of proportional costs only when cheaper bundles of assets are available in that one can think of the overall cost as the fixed cost for accessing the bundle consisting of asset 1 and 2.

2.4.4 Comparative statics

Comparative statics for the case of n assets and proportional costs are readily available from conditions (2.11) and (2.12).

Recall that these conditions with equality define a pair of $n-1$ dimensional hyperplanes and if we let scalars b_i^P and b_i^S be the negative of the RHS of (2.11) and (2.12) we can write each pair as $\mathbf{V}_i\theta + b_i^P \geq 0$ and $\mathbf{V}_i\theta + b_i^S \leq 0$.[69] The i-th pair define a corridor in $\mathbb{R}^n$ outside of which it is optimal to trade asset i only up to the closest of the two planes and the intersection of the n corridors forms the no trading region which is a parallelogram (or its higher dimensional analog).

[69] Also recall that V_i corresponds to the i-th row of V.

Each pair of planes share the same orientation $\mathbf{V}_i\theta$ and a different constant (b_i^P or b_i^S respectively), thus the planes are parallel and any change in the correlation structure as well as in the assets variances primarily affects the orientation of each pair of corridors in $\mathbb{R}^n$. Specifically, changes in the correlations affects *simultaneously* all the corridors while changes in the variance of asset i *only* affect the orientation of the i-th corridor.

Changes in any other parameter, i.e. $\lambda, \kappa, r, \mu_i, C_i^P, C_i^S$ and θ^B, will not affect the shape of the no trading region, rather will make it shrink/expand and/or shift. This is because any such parameters enter each pairs of hyperplanes defined by (2.11) and (2.12) only trhough their intercepts b_i^P and b_i^S. In particular higher (lower) proportional costs C_i^P and/or C_i^S increase (decrease) the width of the i-th corridor while higher (lower) risk aversion λ and/or tracking error aversion κ[70] will shrink (expand) the no trading region as a whole: this is because the i-th corridor width is given by $\frac{|b_i^P - b_i^S|}{||\mathbf{V}_i||}$ and $b_i^P - b_i^C = \frac{C_i^P + C_i^S}{\kappa + \lambda}$. Changes in the i-th mean of asset i, μ_i, will cause a parallel shift of the i-th corridor only. Finally changes in the composition of the benchmark θ^B will simultaneously cause the no trading region to shift and shrink/expand according to the new directions represented by the modified benchmark θ^B.

From the analysis of this entire section we also know that adding a fixed cost (overall or asset specific) creates a surface which surrounds the no-trading region where the investor is indifferent between paying the additional fixed cost and trade to the closest boundary of the no-trading region or do not trade at all. Not considering corners, this amounts to imposing the surface of an outer parallelogram (or its higher dimension analog) which has at its center the proportional cost only inaction region. Because the latter object is proportional to the former, the comparative statics for all parameters except the just discussed fixed (asset

[70] Positive (negative) changes in κ will also make the no trading region move closer (further) from θ^B.

specific or overall) cost component remain the same as those for the proportional costs only framework.

We also learned that corners are always formed by the intersections of a locus where the investor is indifferent between trading in all n assets or not trade at all which is an ellipse (or its higher dimension analog) and n loci where the investor is indifferent among trading in $n-1$ assets or the remaining one which are hyperplanes parallel and perpendicular to the Cartesian axes.

2.5 Algorithm

In this section we introduce a powerful algorithm to solve problems of the form

$$\max_{x} \quad \tilde{U}(x) = a'x + \frac{1}{2}x'Qx$$

subject to

$$x \geq 0$$

where Q is semi-positive definite and symmetric. As we show in the proof of Theorem 1, for any finite number of risky assets n the building blocks of our solution strategy for the class of problems having objective function (2.1), can all be re-written in that form.

As shown in Cottle, Pang, and Stone (1992), the first order conditions from the above quadratic program can be expressed via the system

$$\begin{cases} x \geq 0 \\ a + Qx \geq 0 \\ x'(a + Qx) = 0 \end{cases} \tag{2.14}$$

Finding a vector $x \geq 0$ satisfying the above system is referred to as the linear complementarity problem $LCP(a, Q)$. The attractiveness of the LCP framework is the availability of efficient iterative schemes converging to the solution(s) of the problem which are essentially based upon the characteristics of the matrix M.

Given a decomposition of the matrix Q as $Q = B + C$ it is easy to verify[71] that $LCP(a, Q)$ can be re-expressed as $LCP(a^x, B)$ defined as

$$\begin{cases} x \geq 0 \\ a^z + Bx \geq 0 \\ x'(a^z + Bx) = 0 \end{cases}$$

with $a^z \equiv a + Cx$ and that a solution x to $LCP(a, Q)$ is a fixed point for $LCP(a^x, B)$. The algorithm that solves the above quadratic program, which is given next, is nothing more than an iterative scheme to find the fixed point x of $LCP(a^x, B)$.

In order to solve the original quadratic program of this section we exploit the following theorem

[71] See Chapter 1 of Cottle et al. (1992).

Theorem 8 *Let $Q = B + C$ be positive semi-definite such that B and $B - M$ are positive-definite and B is a diagonal matrix, if a is such that $LCP(a, Q)$ admits solutions, then the following algorithm produces a sequence $\{x^v\}$ which converges to some solution of $LCP(a, Q)$*

Step 1: *Set arbitrary $x^v \geq 0$, with $v = 0$*

Step 2: *Given x^v compute the new vector x^{v+1} as*

$$x^{v+1} = max(0, x^v - B^{-1}(a + Qx^v))$$

Step 3: *Given a pre-determined tolerance $\epsilon \geq 0$ stop, otherwise go back to Step* 1

Proof. See Theorem 5.6.1., Algorithm 5.2.1 and section 5.10 in Cottle et al. (1992). ■

It is important to notice that Theorem 8 perfectly fits our needs. Q is positive semi-definite, exploiting the notion of diagonally dominant matrix it is straightforward to use Q to construct a diagonal positive definite matrix B such that $B - M$ is also positive-definite.[72] Remember that $LCP(a, Q)$ are just the FOCs for the quadratic program; Such program when applied to the building blocks of Theorem 2 it is showed to have a unique and feasible solution, thus as long as $A1 - A3$ hold a is such that $LCP(a, Q)$ admits solutions. Then the above algorithm produces a sequence $\{x^v\}$ which converges to some solution of $LCP(a, Q)$, but we also know from Theorem 2 that such solution is unique so that the algorithm generates an iterative scheme that uniquely solves the quadratic program of interest.

[72] A square matrix A is (strictly) diagonally dominant if $|a_{ii}|(>) \geq \sum_{j \neq i} |a_{ij}|$ for all i and any such matrix is positive semi-definite (definite). It is thus enough to define the typical diagonal element of B as $B_{ii} > M_{ii} + \sum_{j \neq i} |M_{ij}|$.

2.6 Conclusion

We have used a mean-variance analysis of portfolio rebalancing given transaction costs to illustrate a number of important economic features in a context that is simple to understand and solved completely. The single-period case is suggestive of good strategies in more realistic cases, and is a useful benchmark for comparisons.

Figure 2.1: Mean-Variance Problem with Proportional Transaction Costs

With proportional costs, the non-trading region is the area of a parallelogram. Outside the non-trading region, it is optimal to trade (along the arrows) to the boundary of the non-trading region. If returns were uncorrelated, then the non-trading region would be a square with sides parallel to the axes. In this example, returns are correlated and the two securities are substitutes and over-weighting in one security is less likely to result in a trade if we are under-weighted in the other security.

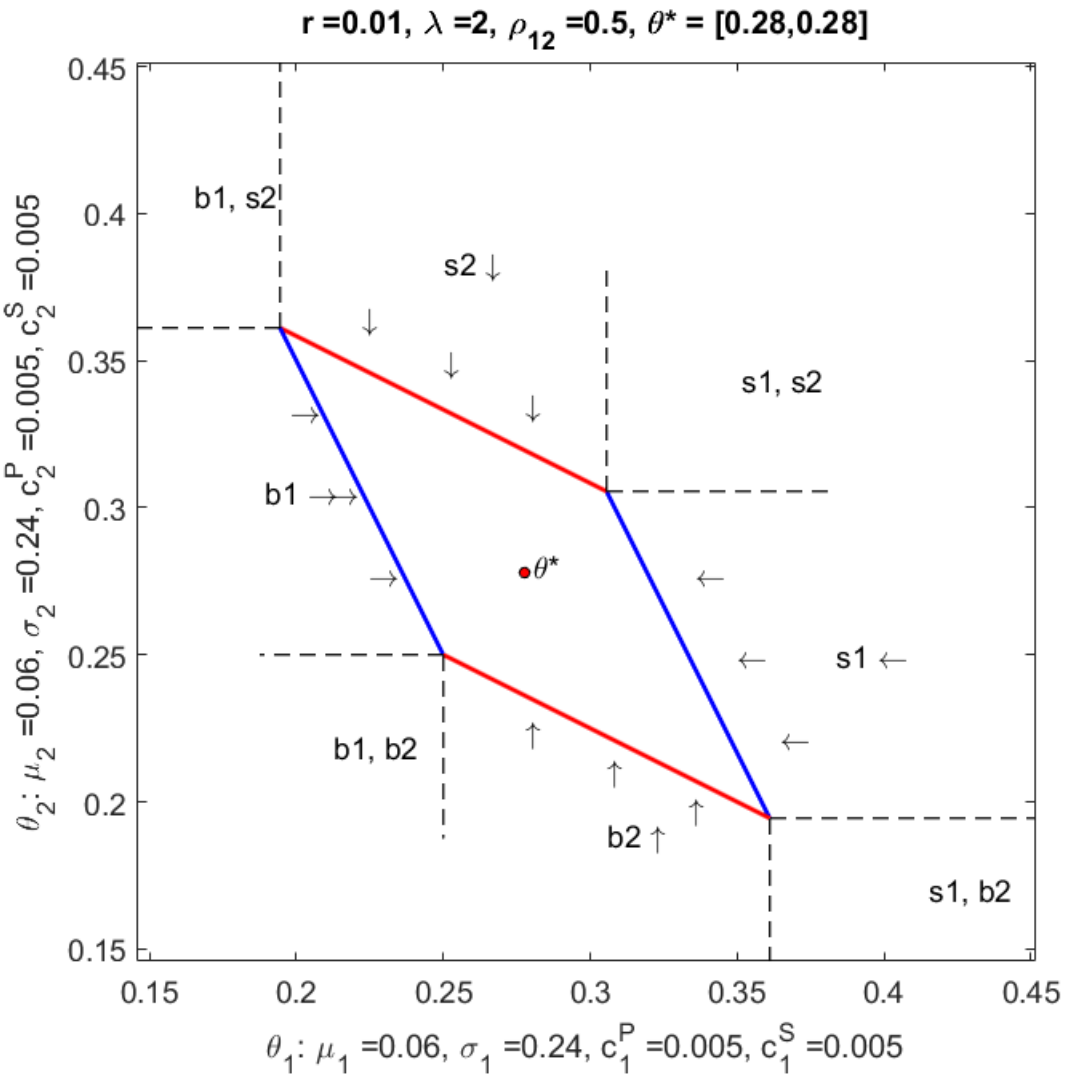


Figure 2.2: Mean-Variance Problem with an Overall Fixed Transaction Cost

With an overall fixed cost, either there is trade immediately to the ideal point or it is not worth trading at all. The nontrading region is the area of an ellipse. As with proportional costs, correlation between the assets implies that it is more damaging (and more likely to do trade) when both asset positions are out of line in the same direction.

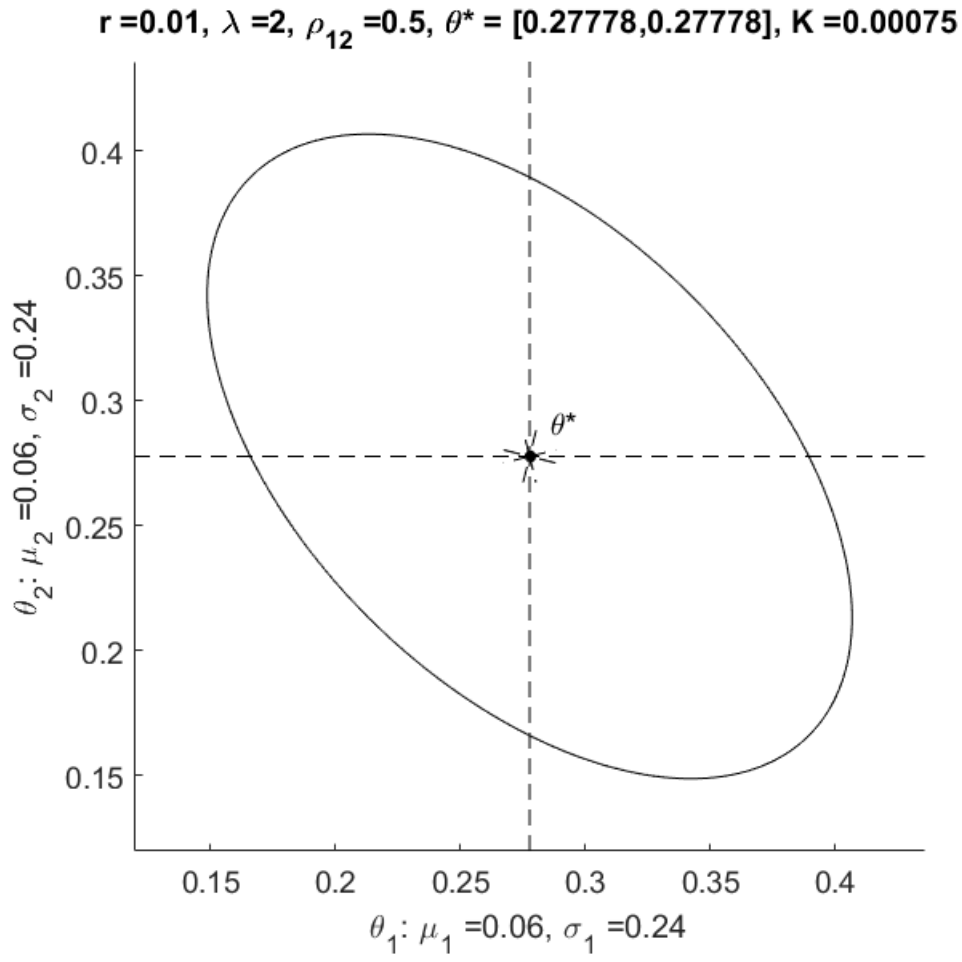


Figure 2.3: Mean-Variance Problem with Asset-Specific Fixed Transaction Costs

Near the ideal point in the middle, θ^*, is the non-trading region. The North-West and South-East corners lie on an ellipse reminiscent of the no-trading region of an overall fixed cost problem while the pairs of parallel outer thick black dashed straight lines and the two intersecting blue and red lines come from the asset-specific component of the fixed costs. For analogous reasons as those discussed in the overall fixed cost example, for any initial position inside the region defined by the above mentioned corners and dashed line it is optimal not to trade at all. Outside of this area we trade as follows: from the regions in the corners, we trade both securities to the ideal point θ^*. From the regions on the right and left, we trade Security 1 but not Security 2 to the thick blue line going through the ideal point. This would be a vertical line if we had no correlation, but has negative slope in our case. Similarly, from the regions above and below, we trade only Security 2 to the thick red line running through the ideal point.

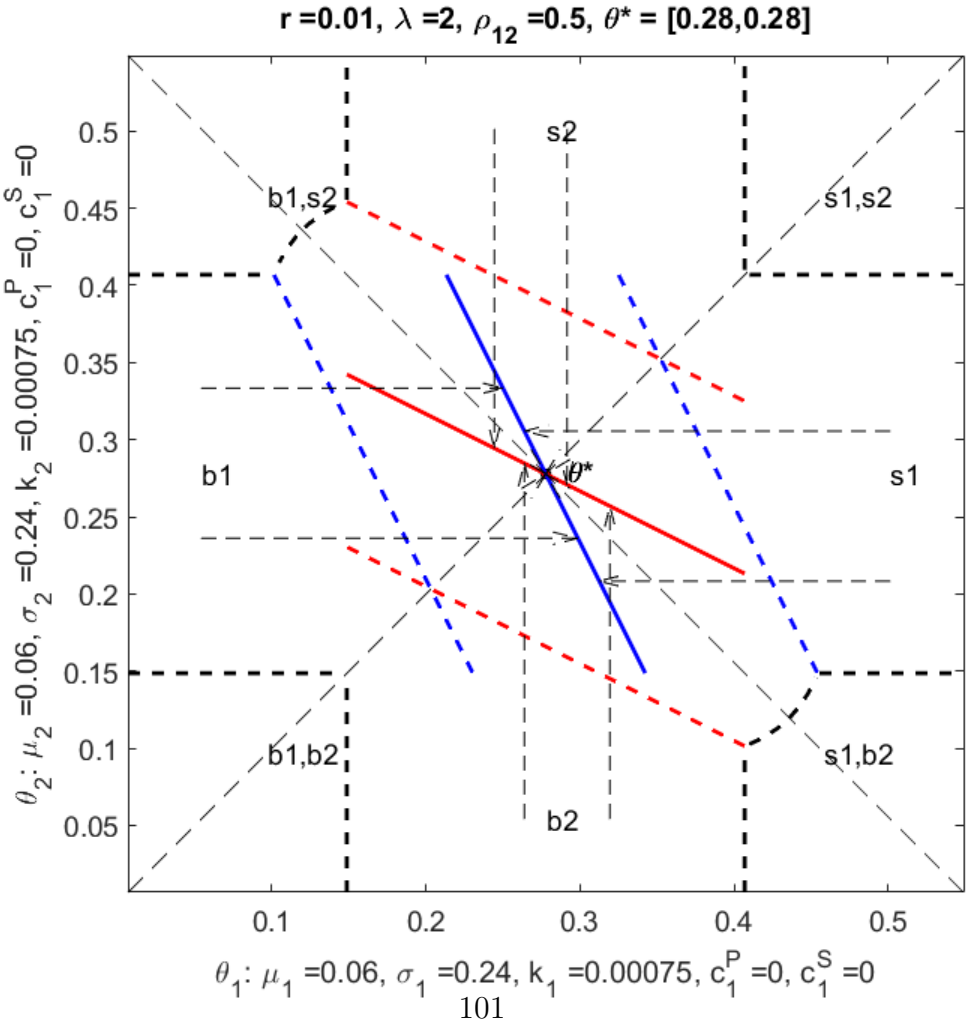


Figure 2.4: Mean-Variance Problem with Overall Fixed and Asset-Specific Proportional Costs

The figure shows the different (no-)trading regions for the case in which the investor both pays an overall fixed cost to enter the market as well as asset specific proportional costs to (re)balance his portfolio to the desired allocation. The parameters, except for the fixed cost $K = 0.00075$, are those of Figure 2.1 and the inner parallelogram is indeed the same: this is the region where the investor will optimally trade to *if* the fixed cost of entering the market is not too big. There is in fact a threshold, pictured as a dash-line, surrounding the proportional no-trading region; If the initial position θ^0 is inside this line the dis-utility from the fixed cost to enter the market is more than the benefit of trading to the border of the no-trading region (inner parallelogram) so that it is optimal not to trade at all. Outside of this line it is optimal to trade up to the border of the no trading region as in the proportional only case: in particular, there are four corridors and four corners delimited by wavy dashed lines. Inside each corridor it is optimal to trade (buy or sell) along straight lines only one asset at a time, while at each corner it is optimal to trade (buy/sell) two assets until the nearest no-trading region corner is reached.

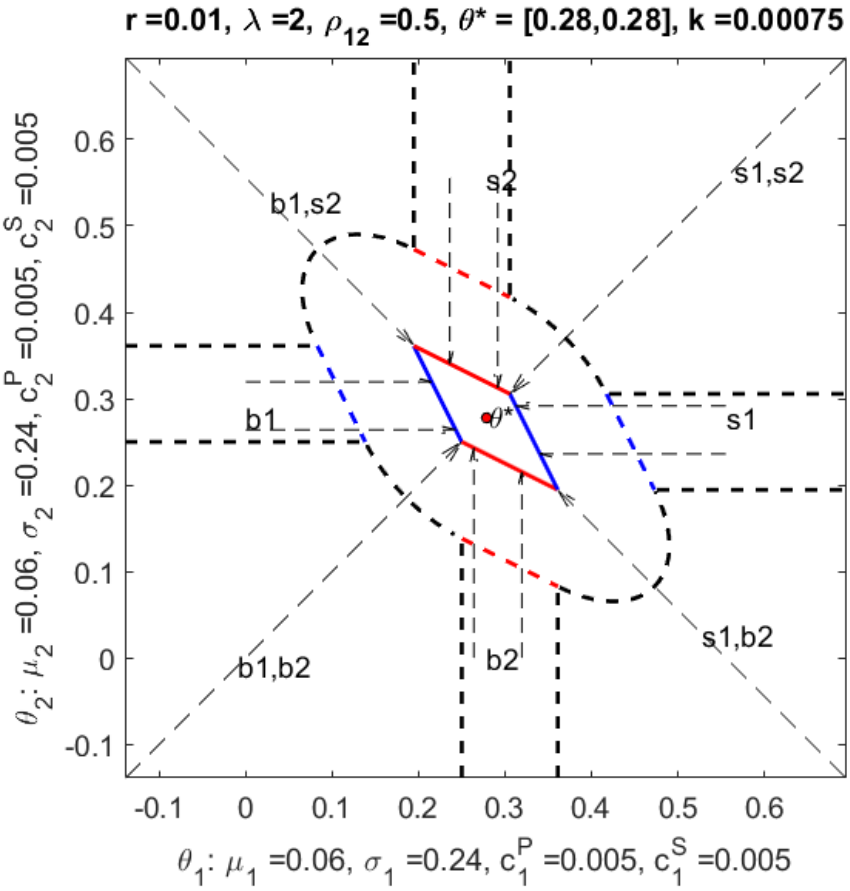


Figure 2.5: Mean-Variance Problem with Asset-Specific Fixed and Proportional Costs

Parameters, except for the fixed cost $K = 0.00075$, are those of Figure 2.1 and the inner parallelogram is indeed the same: this is the region where the investor will optimally trade to *if* the fixed costs are not too big. Each fixed cost, albeit asset-specific, is taken as identical in the figure. The presence of asset specific costs rather than an overall fixed cost, on top of the asset specific proportional components, has similar effects, and an analogous intuition, as those described in Figure 2.4 with two notable differences: 1 - the corridors inside which, only one asset at time is traded up to the border of the no trading region, are wider pushing the corner regions further away from the unconstrained optimum. 2 - the thresholds where the investor is indifferent between paying the fixed costs and trade to the border of the no trading region or do not trade at all are different, featuring a couple of intersections along the 45 degree line passing through the unconstrained optimum.

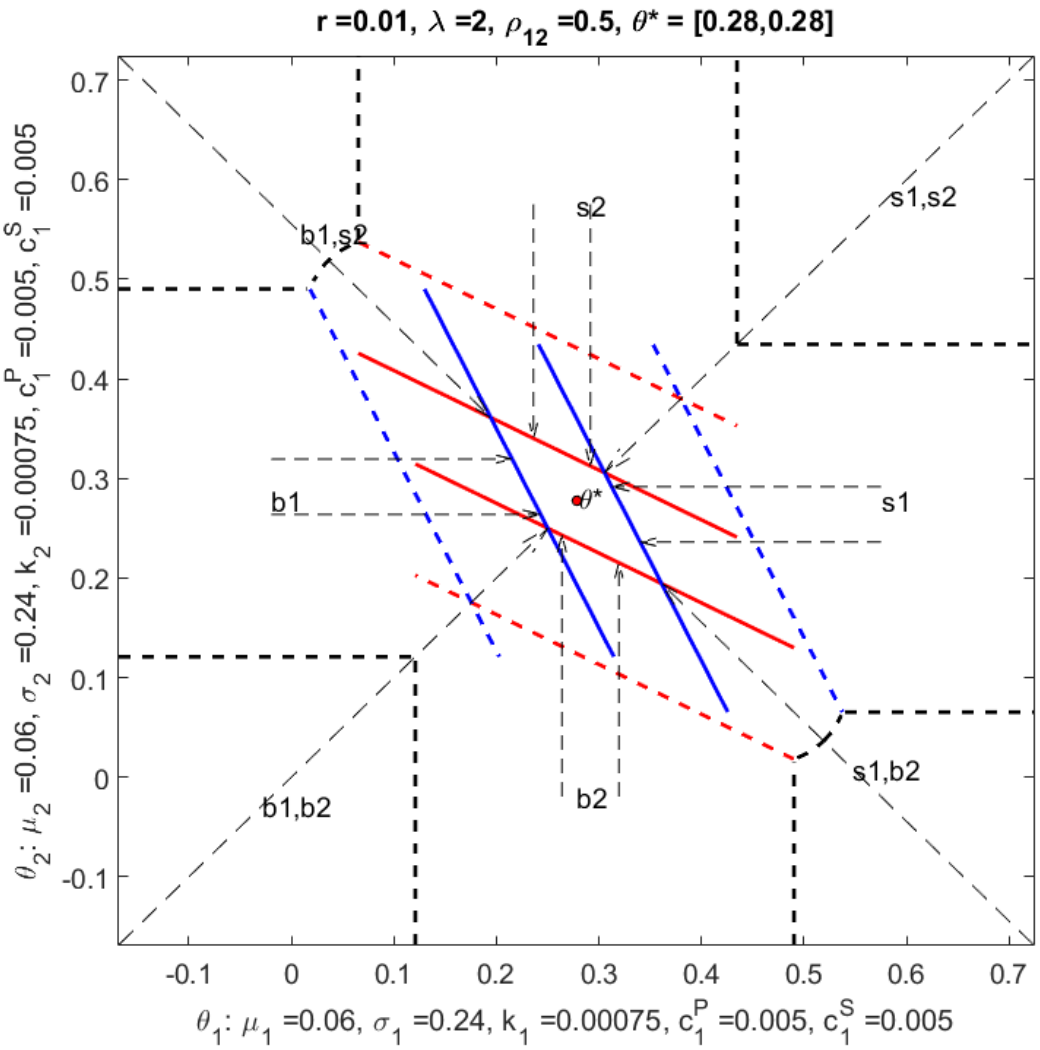


Figure 2.6: Asymmetric Futures Overlay Strategies

Security 1, the equities, has a significantly higher expected return ("alpha") than Security 2, futures. The optimal strategy is an asymmetric "futures overlay" strategy typically selling futures to correct for overexposure to market risk but buying underlying equities to correct for underexposure to market risk. This asymmetry is due to the fact that selling futures allows us to keep the alpha on the exposure we are eliminating, while buying equities allows us to gain alpha on the exposure we are taking on.

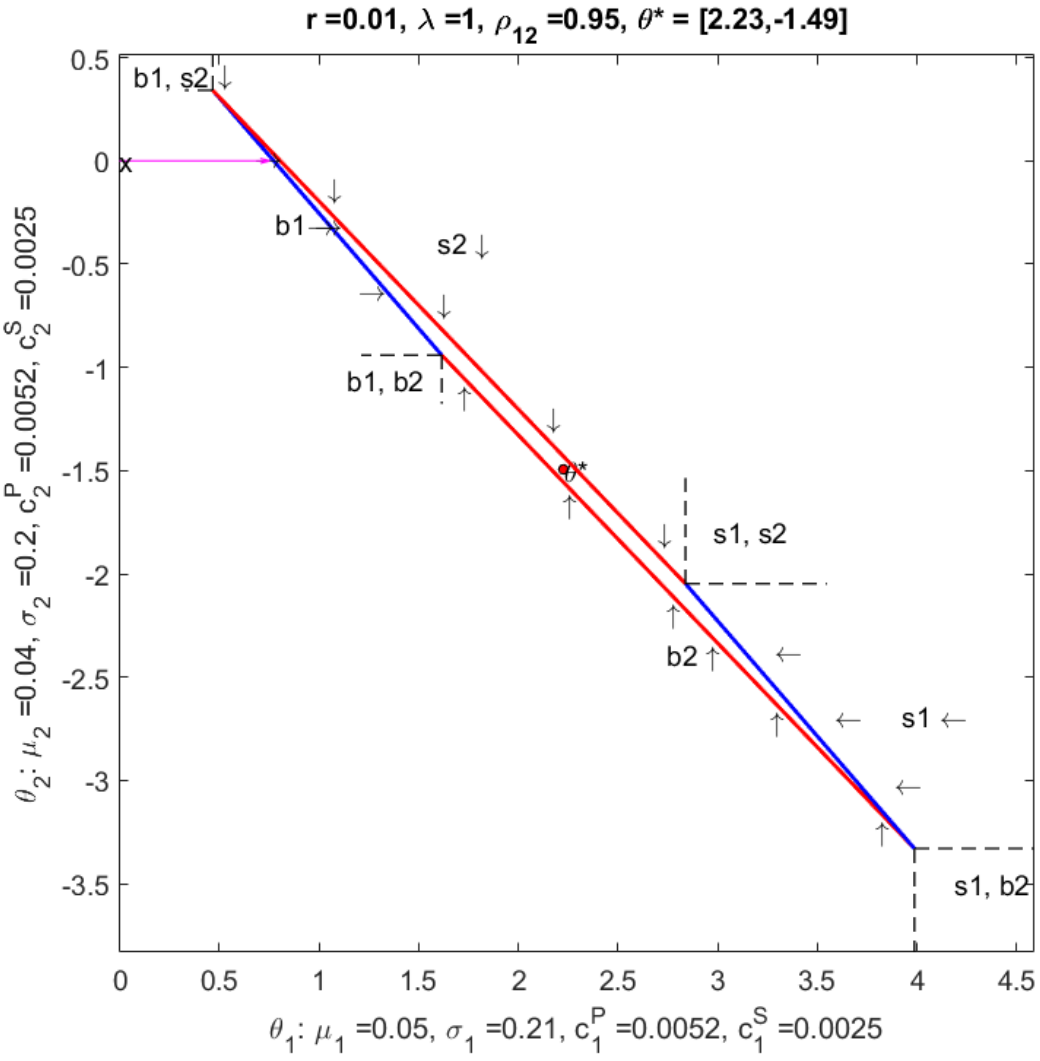


Figure 2.7: Bundles Trading

With bundle trading it is possible to shrink the no-trading region to get closer, at least in some directions, to the unconstrained optimum. The new region have additional sides. This is the same case as in Figure 2.1 but with an additional opportunity to trade a 50-50 portfolio of the two assets at a cost of 0.0025.

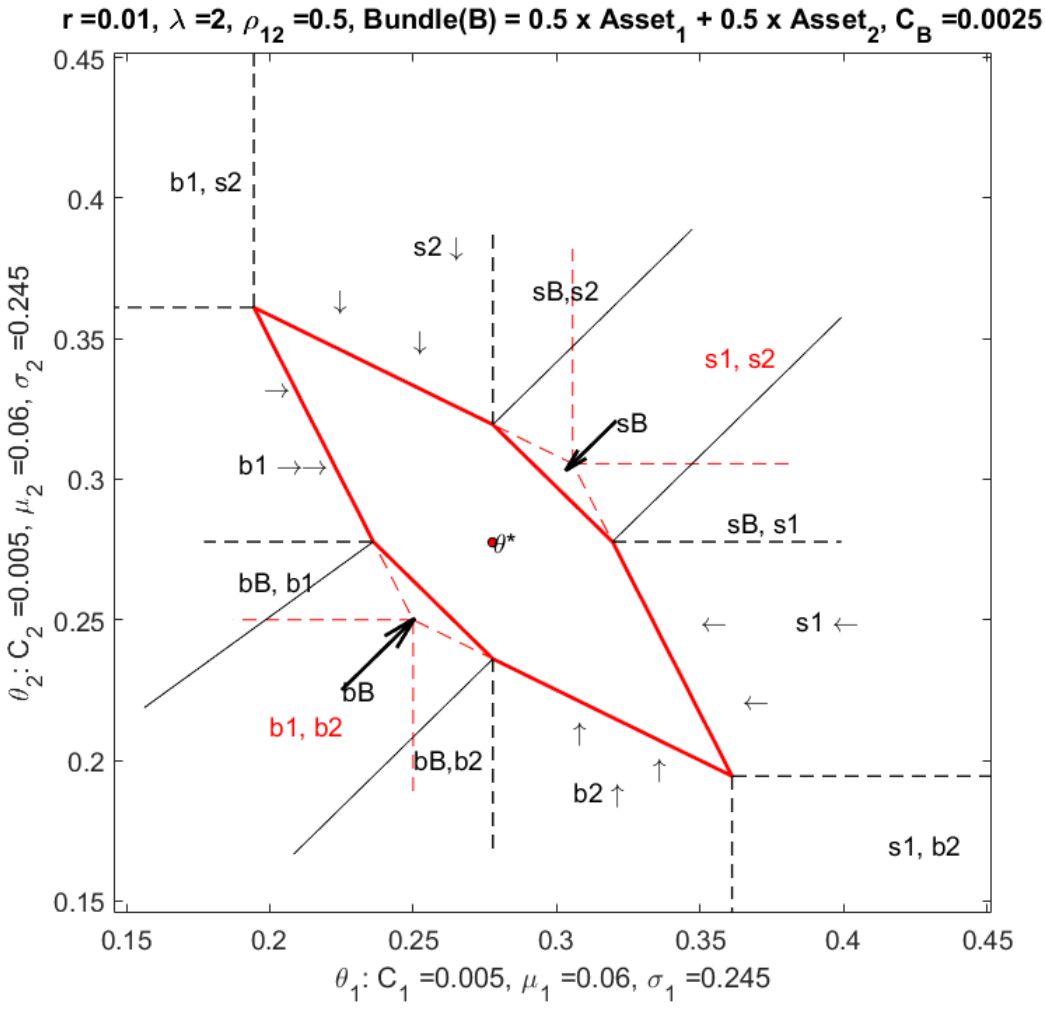


Figure 2.8: Shrinking the No Trade Region: Overall vs. Asset-Specific Fixed Costs

The figure compares the inaction regions of Example 2.4 (red) and 2.5 (blue). Even if no explicit bundle of assets is available, for situations in which it is optimal to simultaneously trade both assets (in case of negative correlation the no trade regions stretches more along the 45 line) one can think of the overall cost as the fixed cost for accessing the bundle consisting of asset 1 and 2 thus coming closer to the unconstrained optimum.

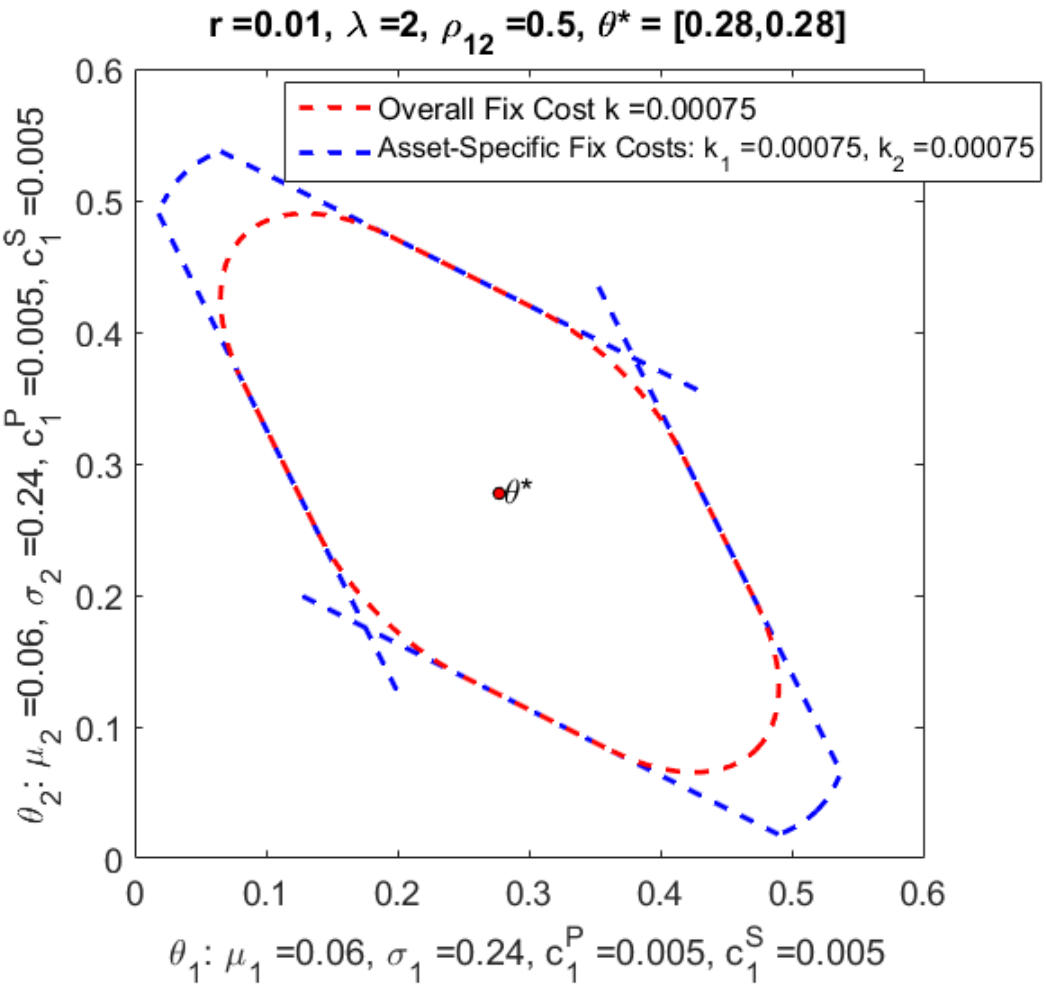


Figure 2.9: Optimal Trading in the Presence of a Benchmark

Trades with proportional and an overall fixed cost (top) and with proportional and asset specific costs (bottom). Benchmark θ^B shrinks and shift the no trading regions towards it (Compare with Figure 2.4 and 2.5).

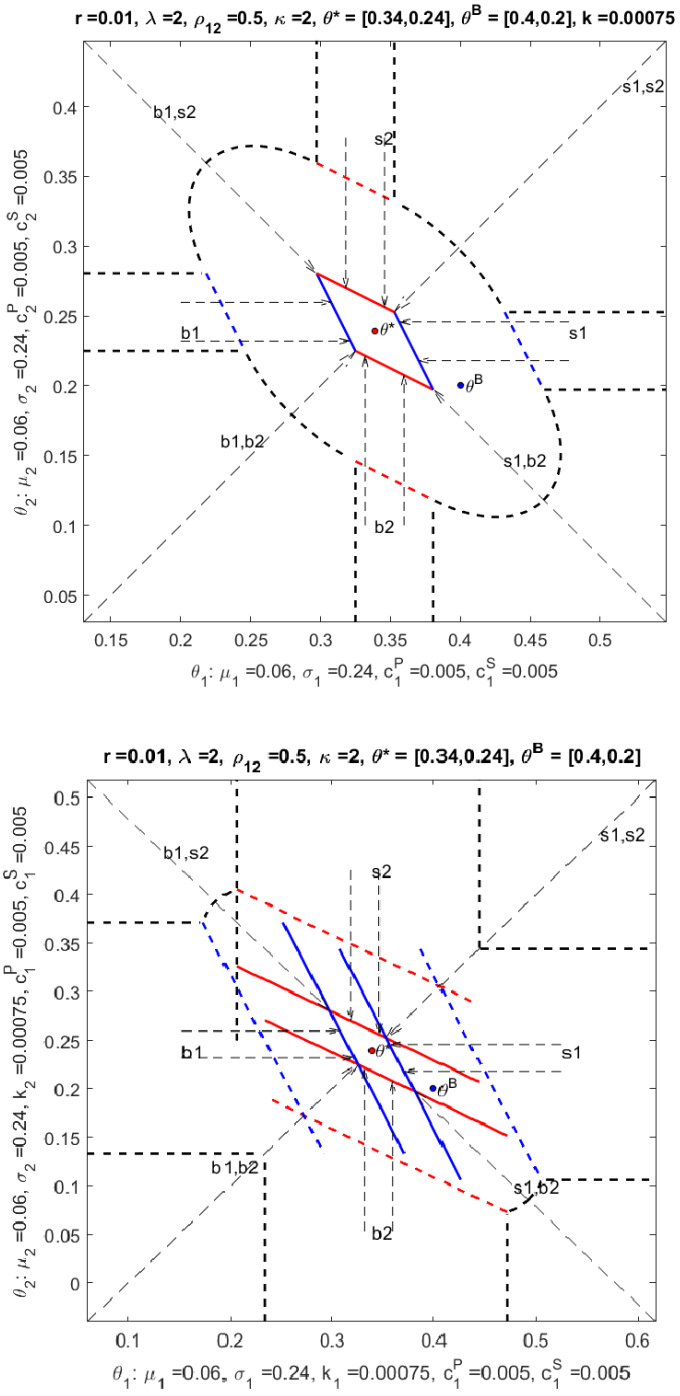
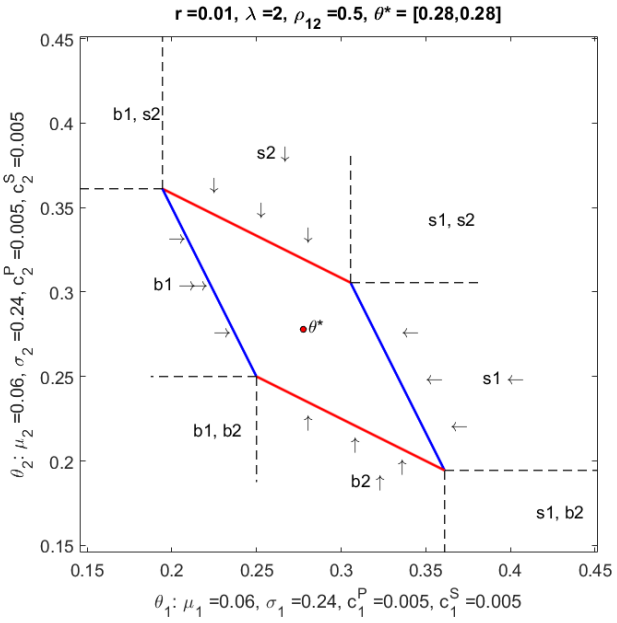


Figure 2.10: No Trade Region with Proportional Costs: Case of 2 and 3 Risky Securities

The no trading regions with proportional transaction costs for two typical cases with 2 and 3 positively correlated assets.



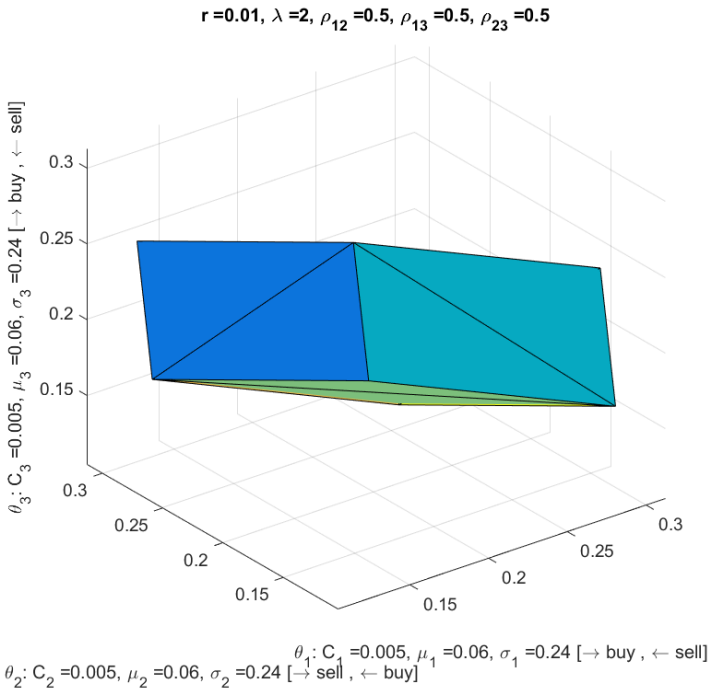


Figure 2.11: Indivisibility of the Mean-Variance Problem with Transaction Costs

In contrast to the H. Liu (2004) framework, in the presence of nonzero correlations among assets it is not possible to separately solve n different sub-problems involving the allocation between a risky asset and the risk free. In light blue it is plotted the slice of the optimal no trading region (the parallelepiped of Figure 2.10) evaluated at the optimal level of asset 3 together with the no trading region of a problem only involving asset 1 and asset 2. If the assets where uncorrelated the two plotted parallelograms would be perfectly overlapped squares as in H. Liu (2004) but in general they are different.

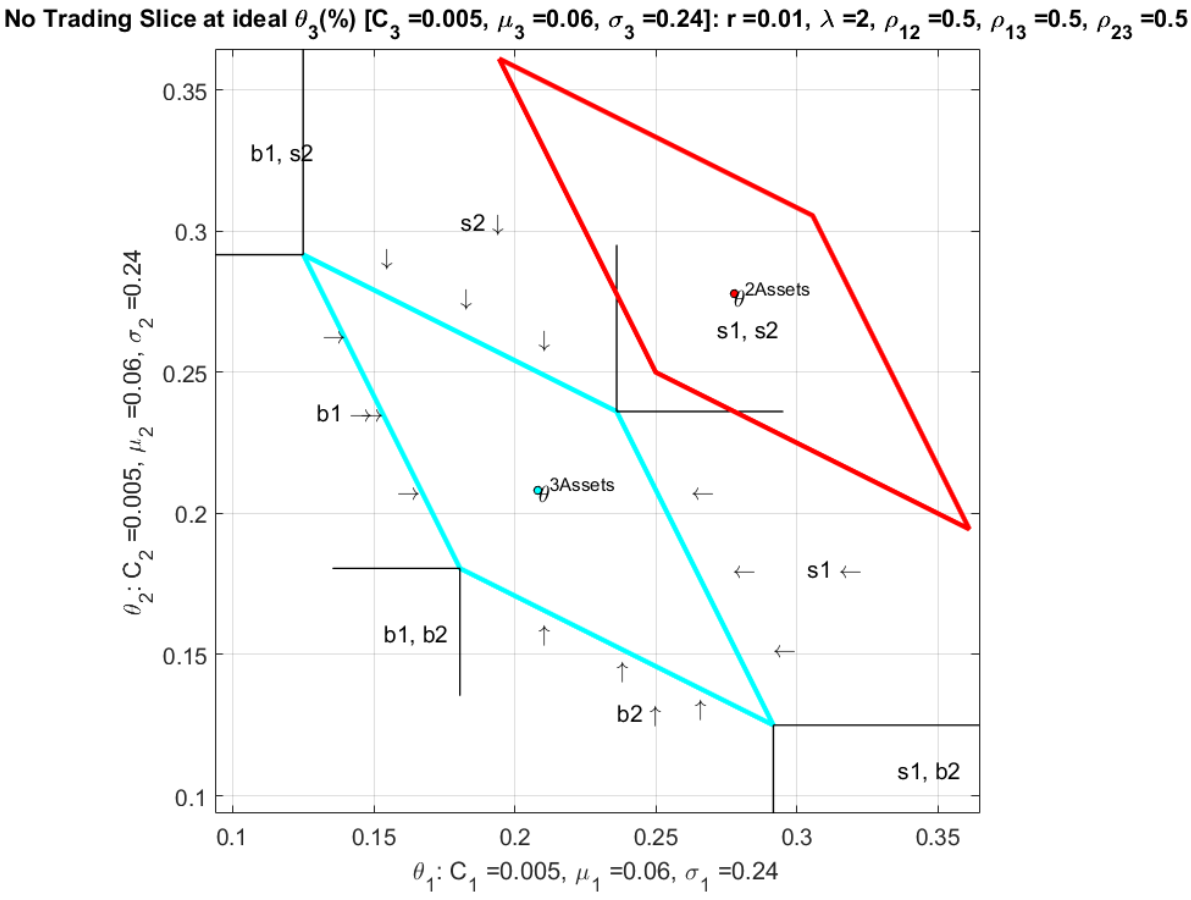


Figure 2.12: Architecture Behind the Asset-Specific Fixed No Trade Region

This figure illustrates the construction of the different trading regions for the case of security-specific fixed costs and it shows the "architecture" behind Figure 2.3.

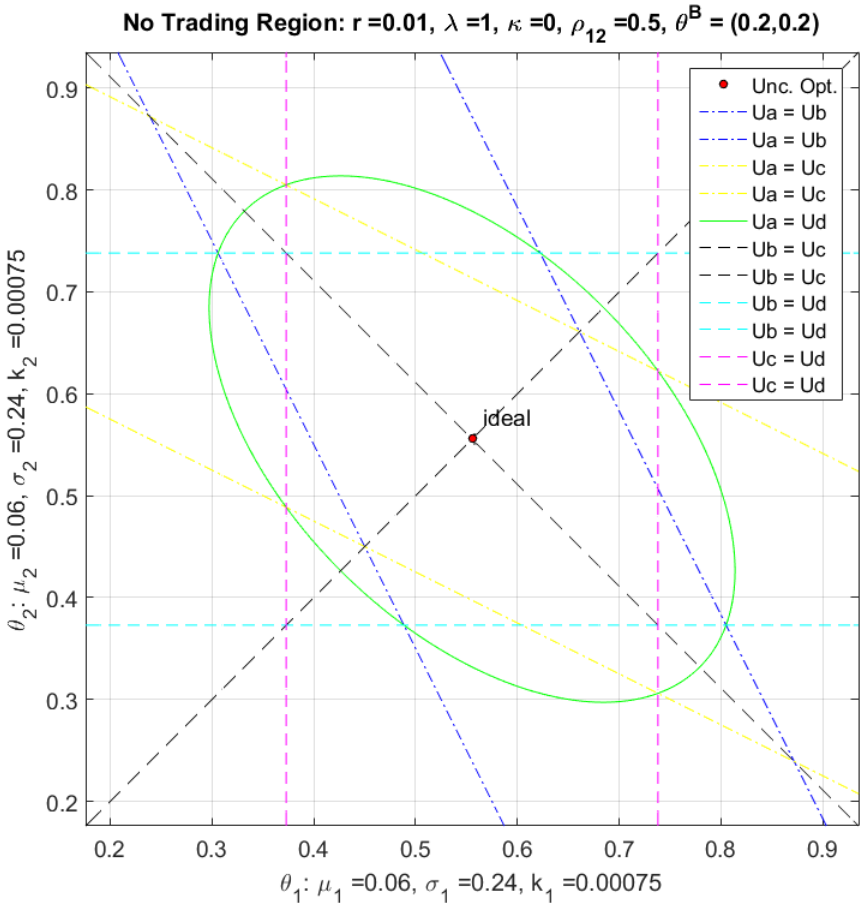


Figure 2.13: Architecture Behind the Asset-Specific Fixed and Proportional No Trade Region

This figure illustrates the construction of the different trading regions for the case of security-specific fixed and proportional costs and it shows the "architecture" behind Figure 2.5.

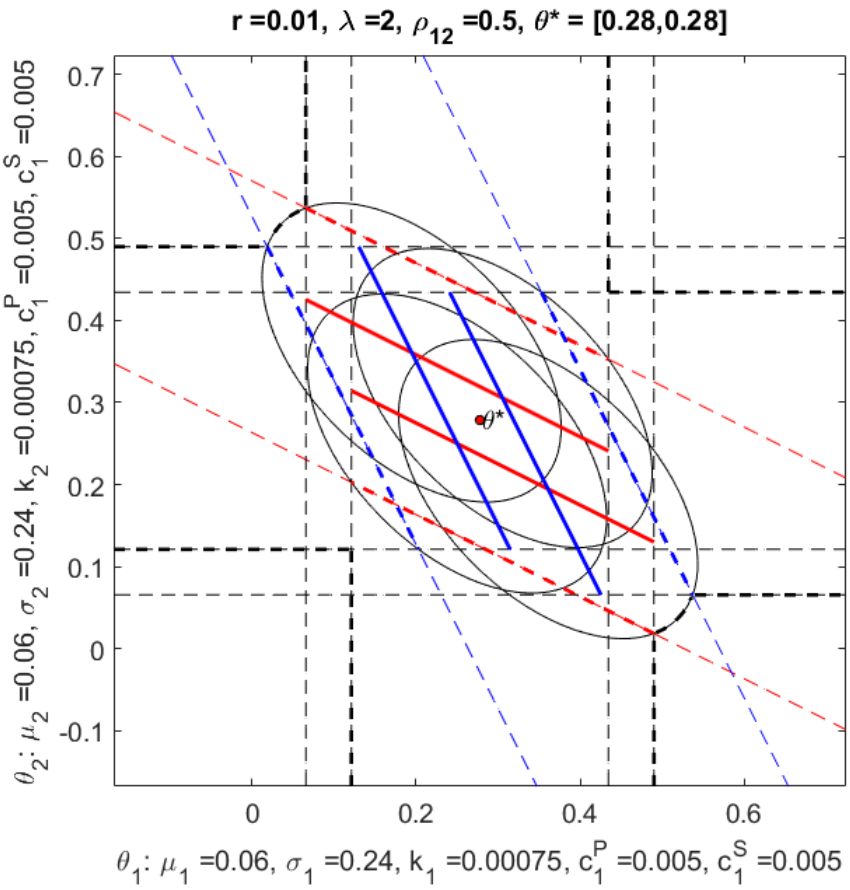


Figure 2.14: Architecture Behind the Overall Fixed and Proportional No Trade Region

This figure illustrates the construction of the different trading regions for the case of asset-specific proportional costs with an overall fixed costs and it shows the "architecture" behind Figure 2.4.

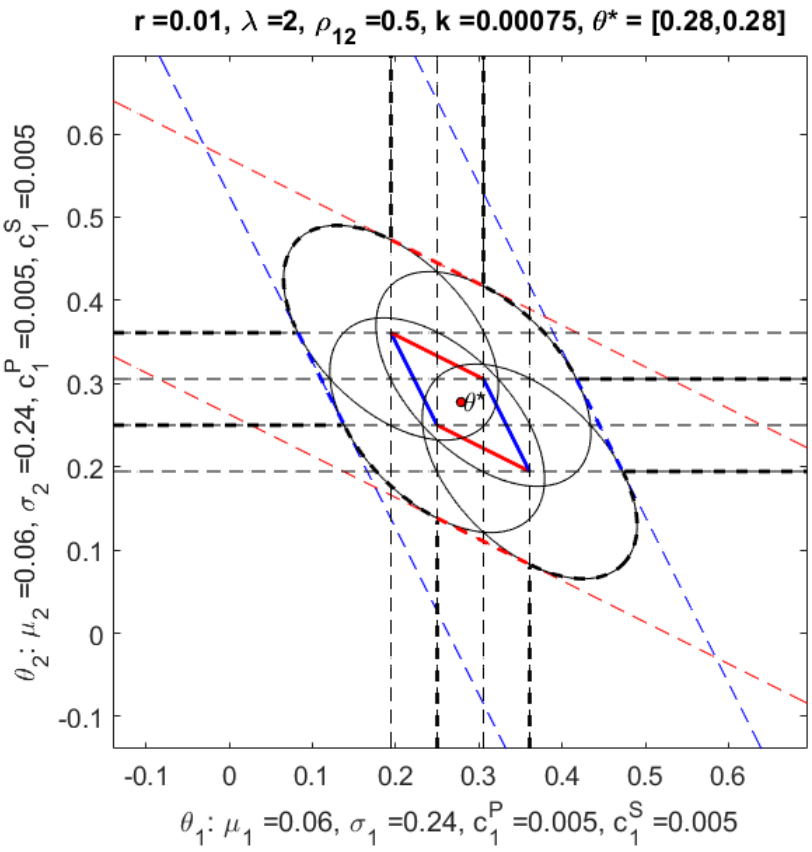
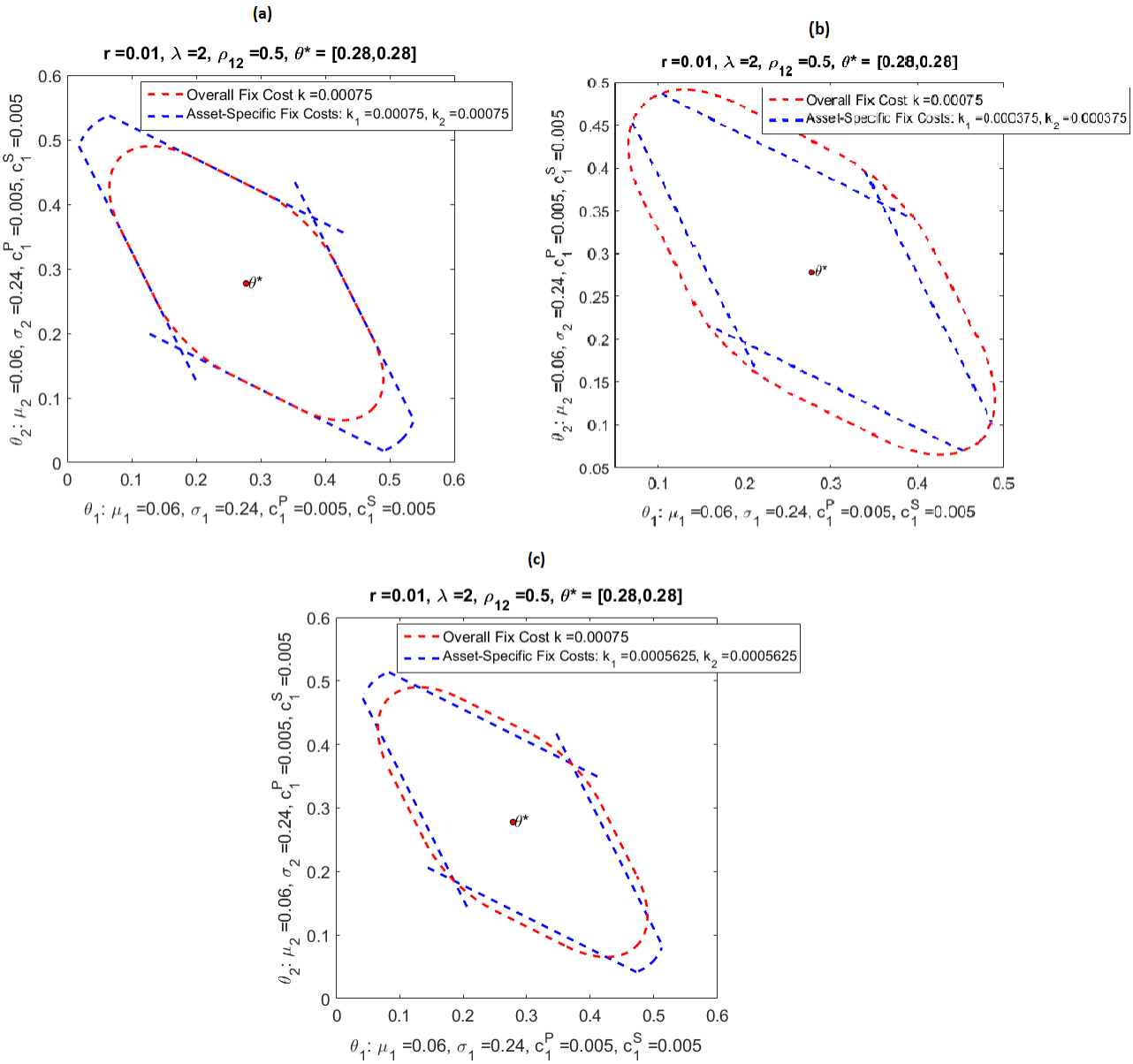


Figure 2.15: No Trade Region Under Different Types of Fixed Costs

This figure compares the optimal trading in the presence of asset specific proportional costs and an overall fixed cost versus asset specific proportional and fixed costs for different fixed costs. In panel (a) all fixed cost are equal to 0.00075, in panel (b) the asset specific fixed costs k_1 and k_2 are half the overall fixed cost k, while in panel (c) the asset specific fixed costs k_1 and k_2 are 0.75 the overall fixed cost k.



Chapter 3

Importance of Transaction Costs for Asset Allocations in FX Markets

3.1 Introduction

A large body of theoretical research studies the implications of transaction costs on the optimal portfolio choice. However, it is unclear whether accounting for transaction costs when optimizing a portfolio is empirically relevant. Even though a portfolio optimized over transaction costs is theoretically different from a portfolio that ignores costs, the out-of-sample performance of these two portfolios may or may not be significantly different.

Using foreign exchange (FX) market returns of 29 developed and emerging currencies from 1976 to 2016, we show that taking transaction costs into account in a mean-variance portfolio optimization leads to an economically large and statistically significant improvement in the out-of-sample performance. We document that the out-of-sample Sharpe ratio after costs is 0.7 for a mean-variance efficient portfolio which ignores transaction costs (MV), while the Sharpe ratio is 0.9 for a portfolio which takes costs into account in the optimization (MV_{TC}). Other moments of the return distributions and in particular the crash risk exposures are

similar across the two strategies. To our knowledge we are the first to empirically quantify the substantial out-of-sample benefit of accounting for transaction costs in the construction of mean-variance optimized portfolios. This is an important contribution to the literature and has interesting implications for practitioners.

We employ the algorithm proposed by Dybvig and Pezzo (2018) to construct MV_{TC}. They characterize the theoretical shape of the no trading region for multiple risky assets and explain how it depends on the cost structure in a single period mean-variance framework. We quantify and assess the empirical importance of four theoretical predictions.

First, we expect MV to outperform MV_{TC} if the performance is measured in returns before transaction costs.[73] This first prediction is empirically irrelevant in FX markets. The Sharpe ratios before transaction costs are almost identical, i.e., 0.99 for MV and 1 for MV_{TC}.

Second, we expect transaction costs to be larger for MV than for MV_{TC}. This second prediction is empirically important. MV_{TC} pays 1.28% of the portfolio value per year in transaction costs which is substantially lower than the 3.71% paid by MV.

Third, we expect MV_{TC} to outperform MV after transaction costs. Moreover, the outperformance is expected to depend on the size of the no trading region of MV_{TC}, which in turn, is expected to be increasing in the size of transaction costs and in the correlation between assets. This third prediction is important in the data. MV_{TC} has a Sharpe ratio after transaction costs of 0.9, while the Sharpe ratio of MV is only 0.7. This is driven by the significant reduction of unnecessary trading, which substantially lowers the turnover and transaction costs and increases the performance after costs of MV_{TC} compared to MV. Therefore, optimizing over transaction costs is particularly important if costs are large.

[73]This is because theoretically MV is the mean-variance portfolio with the highest Sharpe ratio.

Fourth, if assets are positively correlated, then we expect the no trading region of MV_{TC} to be larger than the one of $MV_{TC\backslash Corr}$, a strategy which accounts for transaction costs in the optimization but for simplicity assumes that assets are uncorrelated when constructing the no trading region. Thus, transaction costs of MV_{TC} are expected to be lower than the costs of $MV_{TC\backslash Corr}$, and we expect MV_{TC} to outperform $MV_{TC\backslash Corr}$. This fourth prediction is also empirically relevant. The no trading region of MV_{TC} is larger than the one of $MV_{TC\backslash Corr}$. Transaction costs paid by $MV_{TC\backslash Corr}$ are 2.56% per year and its Sharpe ratio after costs is only 0.76, which is inferior to its counter-part in MV_{TC}. Thus, accounting for correlations between assets is important for the superior performance of MV_{TC}. In contrast, the $MV_{TC\backslash Corr}$ does not significantly outperform MV. This result has important theoretical implications: it invalidates the setup of H. Liu (2004), based on the assumption of uncorrelated assets. Unfortunately, this is the only framework that, so far, can solve continuous-time portfolio optimizations in the presence of transaction costs with more than two or three risky assets.

Figure 3.1 further illustrates that our mean-variance efficient portfolios dominate other popular currency strategies (DOL, $DDOL$, HML, MOM, VAL) in terms of out-of-sample Sharpe ratios before and after transaction costs. DOL invests equally in all bilateral carry trades. $DDOL$ takes a long position in DOL if the median exchange rate forward discount is positive, and a short position otherwise. HML sorts bilateral carry trades according to the forward discount into quintiles and short sells the bottom and invests in the top quintile. MOM sorts bilateral carry trades according to their past 12 month performance into quintiles and short sells the bottom and invests in the top quintile. VAL sorts bilateral carry trades according to the power purchase parity adjusted exchange rate into quintiles and short sells the top quintile (overvalued currencies with high real exchange rates) and invests in the bottom quintile (undervalued currencies with low real exchange rates).

An empirical challenge when constructing mean-variance efficient portfolios is that we need sensible estimates of conditional expected returns and the covariance matrix. If estimation errors are large, then a mean-variance optimization often leads to extreme portfolio weights and a poor out-of-sample performance (Brandt (2005)). For instance, DeMiguel, Garlappi, and Uppal (2009) show that in US stock markets an equally weighted portfolio outperforms optimized portfolios out-of-sample. Fortunately, estimation errors are less severe in FX markets. The set of excess returns is described by bilateral carry trades, i.e., uncovered positions in forward exchange rates. First, forward discounts are good proxies for conditional expected excess returns of carry trades because exchange rate growths are well-described by a random walk (Meese and Rogoff (1983)). Second, there is a strong factor structure to describe the covariance matrix (Lustig, Roussanov, and Verdelhan (2011)). This is helpful to reduce estimation errors. These properties are exploited in several recent papers and mean-variance optimized portfolios in FX markets are shown to be very profitable out-of-sample (Baz, Breedon, Naik, and Peress (2001), Della Corte, Sarno, and Tsiakas (2009), Daniel, Hodrick, and Lu (2017), Ackermann, Pohl, and Schmedders (2016), Maurer, To, and Tran (2018)). We follow this literature and construct mean-variance optimized portfolios in FX markets to determine the importance of transaction costs.

FX markets are more liquid and have a higher trading volume than stock markets. Moreover, carry trade strategies are known to outperform stock markets over the past 4 decades. Therefore, FX markets do not only provide a useful environment to study mean-variance efficient portfolios but they are also among the most important asset markets to investors.

Our results have important practical implications. First, accounting for costs when optimizing a portfolio is beneficial and improves the out-of-sample performance. Second, transaction costs are declining over time, and thus, traders who specialize in developed currencies may

be tempted to ignore transaction costs when constructing mean-variance efficient portfolios. However, even if transaction costs are low during normal times, they substantially increase during crises and become relevant (Karnaukh, Ranaldo, and Soederlind (2015)). Third, many currency traders have shifted their focus to emerging and frontier markets because exchange rate forward discounts among developed currencies are close to zero for the past decade. Transaction costs in emerging and frontier markets are generally larger than the costs considered in our analysis, and thus, the implications of transaction costs on the optimal portfolio choice are even more important for these traders.

Our paper is related to the literature on portfolio optimization in the presence of transaction costs. Due to the complexity of the problem most of the literature solves frameworks with only two assets, either directly (Taksar et al. (1988), Davis and Norman (1990), Dumas and Luciano (1991), Shreve and Soner (1994), Balduzzi and Lynch (1999, 2000), H. Liu and Loewenstein (2002), Dumas and Buss (2017)) or through the indirect martingale approach (e.g. Goodman and Ostrov (2010), see Schachermayer (2017) for a comprehensive summary of this approach). H. Liu (2004) solves a multi-asset model but requires the simplifying assumption that assets are uncorrelated and preferences exponential. Leland (2000), Donohue and Yip (2003), Muthuraman and Kumar (2006), Irle and Prelle (2008), Myers (2009) and Lynch and Tan (2009) propose numerical or heuristic approximations. However, except for the case of uncorrelated assets, these numerical solutions are only feasible for a maximum of two and in rare cases three risky assets. Dybvig and Pezzo (2018) provide an algorithm to solve a general multi-asset model with correlated assets restricting to a single period mean-variance framework.

3.2 Theory framework

We quantify the empirical importance of transaction costs in a mean-variance portfolio optimization. Therefore, we implement a generalized version of the algorithm proposed by Dybvig and Pezzo (2018) to construct optimal portfolios using FX market returns of 29 developed and emerging currencies from 1976 to 2016. We use bid-ask spreads as a proxy for proportional transaction costs and assume that there are no fixed costs to trade. We denote the mean-variance efficient strategy without optimizing over transaction costs by MV, and the strategy that takes transaction costs into account in the optimization problem by MV_{TC}.

The investment opportunity set at time t consists of one risk-free asset with risk-free rate of return $r_{f,t}$ and N risky assets with conditional expected excess returns over the risk-free rate (or risk premia) $\mu_{\mathbf{t}}^{\mathbf{e}}$ and conditional covariance matrix $\mathbf{V_t}$. If there are no transaction costs, then an investor with mean-variance preferences with risk aversion λ selects the N-vector of risky asset portfolio weights $\theta_{\mathbf{t}}^{\mathbf{MV}} = \arg\max_{\{\theta_{\mathbf{t}} \in \mathbb{R}^N\}} \left\{ {\theta_{\mathbf{t}}}' \mu_{\mathbf{t}}^{\mathbf{e}} - \frac{\lambda}{2} {\theta_{\mathbf{t}}}' \mathbf{V_t} \theta_{\mathbf{t}} \right\}$ to maximize her utility. The optimal investment in the N risky assets is $\theta_{\mathbf{t}}^{\mathbf{MV}} = \frac{1}{\lambda} {\mathbf{V_t}}^{-1} \mu_{\mathbf{t}}^{\mathbf{e}}$ and the investment in the risk-free asset is $\theta_{0,t}^{MV} = 1 - \mathbf{1}'_{\{N \times 1\}} \theta_{\mathbf{t}}^{\mathbf{MV}}$, where $\mathbf{1}_{\{N \times 1\}}$ is a N-vector with all elements equal to 1 Ma1952. We denote this strategy by MV.

Next, we describe the optimization problem if the investor takes into account transaction costs and we denote this strategy by MV_{TC}. Let $\theta_{\mathbf{t}}^{\mathbf{0}}$ be the N-vector of "initial" weights before trading at time t. The initial weights are equal to the portfolio chosen at time $t-1$ and held until time t. Further, let $\theta_{\mathbf{t}}^{\mathbf{0}+} = \max\{\theta_{\mathbf{t}}^{\mathbf{0}}, 0\}$ describe the initial long and $\theta_{\mathbf{t}}^{\mathbf{0}-} = \min\{\theta_{\mathbf{t}}^{\mathbf{0}}, 0\}$ the initial short positions[74]. The investor has to choose by how much to increase ($\mathbf{\Delta_t^{P+}} \geq 0$) or decrease ($\mathbf{\Delta_t^{S+}} \geq 0$) her long positions, and by how much to increase ($\mathbf{\Delta_t^{S-}} \geq 0$) or

[74]Note that $\theta_{\mathbf{t}}^{\mathbf{0}-} \leq 0$ and a large absolute value (i.e. a small value) means that there are many short positions.

decrease ($\mathbf{\Delta_t^{P-}} \geq 0$) her short positions. The allocation after trading at time t is given by the weights vector $\theta_\mathbf{t} = \theta_\mathbf{t}^+ + \theta_\mathbf{t}^-$ where $\theta_\mathbf{t}^+ = \theta_\mathbf{t}^{\mathbf{0}+} + \mathbf{\Delta_t^{P+}} - \mathbf{\Delta_t^{S+}}$ and $\theta_\mathbf{t}^- = \theta_\mathbf{t}^{\mathbf{0}-} + \mathbf{\Delta_t^{P-}} - \mathbf{\Delta_t^{S-}}$ describe the long and short positions. Element i of N-vector $\mathbf{C_t^{P+}}$ denotes the cost of an increase in the long position of asset i per dollar at time t. Similarly, vector $\mathbf{C_t^{S+}}$ describes the per dollar cost of closing long positions, and $\mathbf{C_t^{S-}}$ respectively $\mathbf{C_t^{P-}}$ the per dollar costs to open respectively close short positions. Costs are proportional, asset specific and depend on whether we open or close a long or a short position. We assume there are no fixed transaction costs. Therefore, the trades $\mathbf{\Delta_t^{P+}}$, $\mathbf{\Delta_t^{S+}}$, $\mathbf{\Delta_t^{P-}}$ and $\mathbf{\Delta_t^{S-}}$ reduce the portfolio return by $\mathbf{\Delta_t^{P+\prime} C_t^{P+}} + \mathbf{\Delta_t^{P-\prime} C_t^{P-}} + \mathbf{\Delta_t^{S+\prime} C_t^{S+}} + \mathbf{\Delta_t^{S-\prime} C_t^{S-}}$. Our setting is a straightforward extension of the case studied by (Dybvig & Pezzo, 2018) where costs to adjust long and short positions are identical, i.e., $\mathbf{C_t^{P+}} = \mathbf{C_t^{P-}}$ and $\mathbf{C_t^{S+}} = \mathbf{C_t^{S-}}$. The optimization problem is:

Problem 9 (Strategy MV_{TC})

$$\max_{\{\mathbf{\Delta_t^{P+}} \geq 0, \mathbf{\Delta_t^{P-}} \geq 0, \mathbf{\Delta_t^{S+}} \geq 0, \mathbf{\Delta_t^{S-}} \geq 0\}} \left\{ \theta_\mathbf{t}{}' \mu_\mathbf{t}^\mathbf{e} - \frac{\lambda}{2} \theta_\mathbf{t}{}' \mathbf{V_t} \theta_\mathbf{t} - \mathbf{\Delta_t^{P+\prime} C_t^{P+}} - \mathbf{\Delta_t^{P-\prime} C_t^{P-}} - \mathbf{\Delta_t^{S+\prime} C_t^{S+}} - \mathbf{\Delta_t^{S-\prime} C_t^{S-}} \right\}$$

$$\begin{aligned} s.t. \quad \theta_\mathbf{t} &= \theta_\mathbf{t}^+ + \theta_\mathbf{t}^- \\ \theta_\mathbf{t}^+ &= \theta_\mathbf{t}^{\mathbf{0}+} + \mathbf{\Delta_t^{P+}} - \mathbf{\Delta_t^{S+}}, \;\; \mathbf{\Delta_t^{P+}} \geq 0, \;\; \mathbf{\Delta_t^{S+}} \leq \theta_\mathbf{t}^{\mathbf{0}+}, \;\; \theta_\mathbf{t}^{\mathbf{0}+} = \max\left\{\theta_\mathbf{t}^\mathbf{0}, 0\right\} \\ \theta_\mathbf{t}^- &= \theta_\mathbf{t}^{\mathbf{0}-} + \mathbf{\Delta_t^{P-}} - \mathbf{\Delta_t^{S-}}, \;\; \mathbf{\Delta_t^{P-}} \leq -\theta_\mathbf{t}^{\mathbf{0}-}, \;\; \mathbf{\Delta_t^{S-}} \geq 0, \;\; \theta_\mathbf{t}^{\mathbf{0}-} = \min\left\{\theta_\mathbf{t}^\mathbf{0}, 0\right\}. \end{aligned}$$

We provide an algorithm to solve Problem 9 in Appendix C.1.2.

The mean-variance setup without transaction costs ($\mathbf{C_t^{P+}} = \mathbf{C_t^{P-}} = \mathbf{C_t^{S+}} = \mathbf{C_t^{S-}} = 0$) is a special case of Problem 9. The portfolio of strategy MV is independent of the initial position $\theta_\mathbf{t}^\mathbf{0}$, and it is always optimal to trade all the way to $\theta_\mathbf{t}^\mathbf{MV}$. In contrast, if there are transaction costs ($\mathbf{C_t^{P+}} > 0$, $\mathbf{C_t^{P-}} > 0$, $\mathbf{C_t^{S+}} > 0$, or $\mathbf{C_t^{S-}} > 0$), then $\theta_\mathbf{t}^\mathbf{MV_{TC}}$ crucially depends on the

origin θ_t^0. Intuitively, there is a trade-off between paying transaction costs (which are linear in portfolio weight changes) and utility gains (which are convex) when moving towards θ_t^{MV}. If the initial allocation θ_t^0 is close enough to θ_t^{MV}, it is optimal not to trade at all since the marginal cost required to move towards θ_t^{MV} is higher than the marginal utility gain. Thus, there is a no trading region. If the initial allocation θ_t^0 is far enough from θ_t^{MV}, then it is optimal to move towards θ_t^{MV} but only until $\theta_t^{MV_{TC}}$ which lies on the boundary of the no trading region. This is because the marginal utility gain from moving towards θ_t^{MV} is decreasing while the marginal transaction cost is constant.

Figure 3.2 illustrates the optimal solution to Problem 9 in a setting with two risky assets (and one risk-free asset) and $C_t^{P+} = C_t^{P-} = C_t^{S+} = C_t^{S-} > 0$. The horizontal axis describes the weight placed on asset 1 and the vertical axis the weight on asset 2. The weight on the risk-free asset is 1 minus the sum of the weights on the two risky assets. The blue dot labeled θ_t^{MV} is the optimal portfolio if there were no transaction costs. The blue parallelogram surrounding θ_t^{MV} defines the no trading region when the two assets are positively correlated. If the initial allocation θ_t^0 is inside the no trading region (i.e., within the blue parallelogram), then there is no trade and $\theta_t^{MV_{TC}} = \theta_t^0$, because the marginal cost to trade towards θ_t^{MV} exceeds the marginal utility gain.

If the initial portfolio θ_t^0 lies outside of the no trading region, then the investor wants to move towards θ_t^{MV} but stops trading once she reaches the boundary of the no trading region. The arrows indicate the direction of trade and the arrow heads show how far to trade. Suppose the initial portfolio θ_t^0 lies in the bottom, right corner of the figure (anywhere below and to the right of the bottom, right corner of the blue parallelogram). Then, the arrows indicate that the investor sells asset 1 ($\Delta_{1,t}^S > 0$) until she reaches the vertical line (extending from the bottom, right corner of the parallelogram) and buys asset 2 ($\Delta_{2,t}^P > 0$) until she reaches

the horizontal line (extending from the bottom, right corner of the parallelogram).[75] Thus, the optimal portfolio $\theta_{\mathbf{t}}^{\mathbf{MV_{TC}}}$ is exactly on the bottom, right corner of the no trading region parallelogram.

Next, suppose that the initial portfolio $\theta_{\mathbf{t}}^{\mathbf{0}}$ lies below the no trading region and between the two vertical lines extending downward from the bottom, left and right corners of the parallelogram. Then, the arrows indicate that the investor does not change her position in asset 1 but only buys asset 2 and $\theta_{\mathbf{t}}^{\mathbf{MV_{TC}}}$ lies on the boundary of the no trading region parallelogram vertically above $\theta_{\mathbf{t}}^{\mathbf{0}}$.

Analogous arguments apply for initial portfolios $\theta_{\mathbf{t}}^{\mathbf{0}}$ farther to the left or above the no trading region. Thus, if the initial portfolio $\theta_{\mathbf{t}}^{\mathbf{0}}$ lies outside of the no trading region, the optimal portfolio $\theta_{\mathbf{t}}^{\mathbf{MV_{TC}}}$ always lies on an edge (and often exactly in one of the corners) of the no trading region parallelogram.

Finally, if we change the setting to two uncorrelated assets, then the no trading region reduces to the red checkered square. For instance, H. Liu (2004) assumes uncorrelated assets, which simplifies the optimization problem, to derive a solution for the optimal trading strategy.[76] We denote this approximate solution by $MV_{TC\setminus Corr}$ and provide details about the solution algorithm in Appendix C.1.2.

[75] Note that we do not distinguish between opening long and closing short positions and closing long and opening short position in our illustration because the costs of both actions are identical in this example. For a more general example where the costs are not identical we refer to Appendix C.1.1.

[76] Constructing optimal trading strategies in the presence of transaction costs and non-zero correlations is complex. Dynamic optimization models can only be solved heuristically or using numerical approximations for two or three risky assets (Balduzzi and Lynch (1999, 2000), Leland (2000), Donohue and Yip (2003), Han (2005), Muthuraman and Kumar (2006), Irle and Prelle (2008), Lynch and Tan (2009), Myers (2009)). Assuming uncorrelated assets greatly simplifies the problem because the original problem can then be split into independent sub-problems, each one handling one asset at a time.

If the two assets are positively correlated, then the no trading region of MV_{TC} is larger along the $-45°$ line than the one of $MV_{TC\backslash Corr}$. This is because the two assets are substitutes if they are positively correlated, while they are not substitutable if they are uncorrelated. Note that if the two assets were perfect substitutes (i.e. a correlation equal to 1), then selling asset 1 would be identical (in terms of risk exposure) to buying asset 2. In the same spirit, if the two assets are (imperfect) substitutes (i.e. correlation between 0 and 1), then there is less benefit in selling one and at the same time buying the other asset than if they are not substitutable at all (i.e. correlation equal to 0). Since an initial position $\theta_{\mathbf{t}}^{\mathbf{0}}$ close to the $-45°$ line requires the investor to buy one and sell the other asset, the marginal utility gain from trading towards $\theta_{\mathbf{t}}^{\mathbf{MV}}$ is smaller and the no trading region larger if the two assets are positively correlated than if they are uncorrelated. Conversely, a similar argument can be applied to the case of a negative correlation, and the no trading region of MV_{TC} is larger along the $45°$ line but smaller than the one of $MV_{TC\backslash Corr}$ (see Appendix C.1.1 for more details).

In summary, constructing a portfolio without taking into account transaction costs in the optimization leads to more trading and higher costs than what is optimal. Second, if assets are positively correlated but an investor assumes that assets are uncorrelated (to simplify the optimization problem and obtain an approximate solution for the optimal trading strategy), then the constructed portfolio also leads to more trading and higher transaction costs than what is optimal. In the following we show that taking into account transaction costs in the optimal trading strategy is quantitatively important in FX markets.

We obtain four theoretical predictions. First, we expect MV to outperform MV_{TC} if the performance is measured in returns before transaction costs. This is because, by definition, MV is the optimal portfolio when evaluated before transaction costs. Second, we expect

transaction costs to be larger for MV than for MV_{TC}. Third, we expect MV_{TC} to outperform MV after transaction costs. Moreover, the size of these differences between MV and MV_{TC} are expected to depend on the size of the no trading region of MV_{TC}. In turn, the no trading region is expected to be increasing in the size of transaction costs and in the correlation between assets. Fourth, if assets are positively correlated, then we expect the no trading region of MV_{TC} to be larger than the one of $MV_{TC\setminus Corr}$. Thus, transaction costs of MV_{TC} are expected to be lower than the costs of $MV_{TC\setminus Corr}$, and we expect MV_{TC} to outperform $MV_{TC\setminus Corr}$. In the following we quantify and assess the empirical importance of these four predictions.

3.3 FX Markets

The investment strategies MV (ignoring transaction costs in the optimization), MV_{TC} (taking into account costs in the optimization) and $MV_{TC\setminus Corr}$ (taking into account costs in the optimization but assuming assets are uncorrelated) are based on a mean-variance optimization (see Section 3.2 for details). In order to construct mean-variance efficient portfolios that perform well out-of-sample, we need sensible estimates of conditional expected returns and the covariance matrix. Estimation errors are a well-known problem in the portfolio optimization literature and can lead to a bad out-of-sample performance of optimized portfolios (Brandt (2005)). For instance, DeMiguel et al. (2009) show that in the US stock market an equally weighted portfolio outperforms mean-variance optimized portfolios out-of-sample due to estimation errors. FX markets are special because exchange rate changes are hard to predict (Meese and Rogoff (1983)), and current exchange rate forward discounts (in the forward exchange rate market) are good proxies of conditionally expected excess returns of bilateral

carry trades (i.e. the excess returns of uncovered positions in forward exchange rates). This fact has been exploited in several recent papers and mean-variance efficient portfolios in FX markets are shown to be very profitable out-of-sample (Baz et al. (2001), Della Corte et al. (2009), Daniel et al. (2017), Ackermann et al. (2016), Maurer et al. (2018)). We follow this literature and implement MV, MV_{TC} and $MV_{TC\setminus Corr}$ in FX markets to quantify the importance of accounting for transaction costs in the construction of optimized portfolios.

3.3.1 Investment Opportunity Set in FX Markets

We denote spot and 1-month forward exchange rates as USD (US-dollar) per unit of currency i at time t by $X_{i,t}$ and $F_{i,t}$. Following the literature, we define the 1-month realized bilateral carry trade return between currency i and the USD (denominated in USD) by

$$CT_{i,t+1} \equiv \ln\left(\frac{X_{i,t+1}}{F_{i,t}}\right) = fd_{i,t} + \Delta x_{i,t+1},$$

where $fd_{i,t} = \ln\left(\frac{X_{i,t}}{F_{i,t}}\right)$ (known at time t) is the forward discount, and $\Delta x_{i,t+1} = \ln\left(\frac{X_{i,t+1}}{X_{i,t}}\right)$ (realized at time $t+1$) is the exchange rate growth. $CT_{i,t+1}$ is the excess return (over the risk-free rate in USD) of entering an uncovered long position in the 1-month forward exchange rate contract.[77]

We use the bilateral carry trade returns for N currencies (against the USD) as our universe of N risky assets. Due to data availability the number of currencies N changes through

[77] Under the premise of the covered interest rate parity (CIP), the forward discount is equal to the interest rate differential $fd_{i,t} = \ln\left(\frac{R_{i,t}}{R_{US,t}}\right)$ where $R_{US,t}(= e^{r_{f,t}})$ and $R_{i,t}$ are 1-month risk-free interest rates in the USD and currency i, and the carry trade return is equivalent to borrow $\frac{1}{R_{US,t}}$ USD and lend $\frac{1}{R_{US,t}X_{i,t}}$ units of currency i. Note that we do not require the CIP to hold for the construction of our portfolios or the out-of-sample performance analysis. We implement all carry trade returns using forward and spot exchange rates and do not need information about interest rates.

time; for notational simplicity, we drop the time subscript for N. The excess returns from time t to $t+1$ of strategies MV, MV_{TC} and $MV_{TC\backslash Corr}$ are $CT'_{t+1}\theta_{\mathbf{t}}^{\mathbf{MV}}$, $CT'_{t+1}\theta_{\mathbf{t}}^{\mathbf{MV_{TC}}}$ and $CT'_{t+1}\theta_{\mathbf{t}}^{\mathbf{MV_{TC\backslash Corr}}}$, where CT_{t+1} is the vector of excess returns of all N bilateral carry trades.

The constructions of MV, MV_{TC} and $MV_{TC\backslash Corr}$ require estimates of conditional expected excess returns $\mu_{\mathbf{t}}^{\mathbf{e}}$ and the covariance matrix $\mathbf{V_t}$. We follow the literature and use the current forward discount $fd_{i,t}$ as a proxy for conditional expected excess return $\mu_{\mathbf{i,t}}^{\mathbf{e}}$ (Baz et al. (2001), Della Corte et al. (2009), Daniel et al. (2017), Ackermann et al. (2016), Maurer et al. (2018)). This is motivated by the empirical finding that exchange rate changes are difficult to predict over a short horizon, i.e., $E_t\left[\Delta x_{i,t+1}\right] \approx 0$ (Meese and Rogoff (1983)).

To estimate the conditional covariance matrix $\mathbf{V_t}$ we follow the literature on portfolio optimization under parameter uncertainty and use the shrinkage method of Ledoit and Wolf (2003). In particular, our estimate of $\mathbf{V_t}$ is a convex combination of the sample covariance matrix of daily exchange rate growths and the covariance matrix implied by a single index model with the first principal component of daily exchange rate growths as the factor. Both the sample covariance matrix and the principal component analysis use daily exchange rate growths within a 9 month window preceding month t such that our estimate uses only information available prior to t and the subsequent portfolio construction is out-of-sample. The first principal component is well-known to capture most of the time-series variation in exchange rate growths (Lustig et al. (2011)), and thus, using it as the target in the shrinkage estimation is a natural choice.

Using shrinkage to estimate $\mathbf{V_t}$ is similar to the estimation by Maurer et al. (2018) based on principal component analysis and removing components which capture only a small fraction of the common variation in exchange rate growths. Both approaches mitigate estimation errors and avoid the presence of near-arbitrage opportunities in the underlying model (Ross

(1976), Kozak, Nagel, and Santosh (2015)). Moreover, trading strategies based on either approach are very profitable out-of-sample. We choose shrinkage over the principal component analysis based approach because a positive definite covariance matrix red is required in our algorithm to solve Problem 9 (see Appendix C.1.2 for details).

The constructions of MV_{TC} and $MV_{TC\backslash Corr}$ further require estimates of transaction costs. We compute carry trade returns before and after transaction costs. We use mid exchange rate quotes for $X_{i,t}$ and $F_{i,t}$ to compute returns before transaction costs. To account for transaction costs we use bid-ask quotes, indicated by superscripts ^{b} and ^{a}. Since it is relatively cheap to roll a contract over from month to month, the literature typically assumes no roll-over fees and only accounts for transaction costs if there is a change in a position (Menkhoff, Sarno, Schmeling, and Schrimpf (2012), Della Corte, Ramadorai, and Sarno (2016), Maurer et al. (2018)). Alternatively, we could quantify full round-trip costs (i.e. assume that a position is completely closed and re-opened every month), which would lead to substantially larger transaction costs and a quantitatively larger effect in our analysis. Full round-trip costs are considered too conservative and larger than the trading costs paid in practice. Our estimates of the per dollar transaction costs to open new long positions ($\mathbf{C_{i,t}^{P+}}$), close existing long positions ($\mathbf{C_{i,t}^{S+}}$), open new short positions ($\mathbf{C_{i,t}^{S-}}$) and close existing short positions ($\mathbf{C_{i,t}^{P-}}$) are

$$\begin{aligned}
\mathbf{C_{i,t}^{P+}} &\equiv \ln\left(\frac{X_{i,t+t}}{F_{i,t}}\right) - \ln\left(\frac{X_{i,t+1}}{F_{i,t}^{a}}\right) = \ln\left(\frac{F_{i,t}^{a}}{F_{i,t}}\right) \\
\mathbf{C_{i,t}^{S+}} &\equiv \ln\left(\frac{X_{i,t}}{F_{i,t-1}}\right) - \ln\left(\frac{X_{i,t}^{b}}{F_{i,t-1}}\right) = \ln\left(\frac{X_{i,t}}{X_{i,t}^{b}}\right) \\
\mathbf{C_{i,t}^{S-}} &\equiv -\ln\left(\frac{X_{i,t+1}}{F_{i,t}}\right) + \ln\left(\frac{X_{i,t+1}}{F_{i,t}^{b}}\right) = \ln\left(\frac{F_{i,t}}{F_{i,t}^{b}}\right) \\
\mathbf{C_{i,t}^{P-}} &\equiv -\ln\left(\frac{X_{i,t}}{F_{i,t-1}}\right) + \ln\left(\frac{X_{i,t}^{a}}{F_{i,t-1}}\right) = \ln\left(\frac{X_{i,t}^{a}}{X_{i,t}}\right).
\end{aligned}$$

Figure 3.3 plots the time-series of the cross-sectional average of annualized costs (cents per dollar trade) for a set of 29 developed and emerging currencies (green solid line), a subsets of 14 emerging currencies (red dashed line), and a subset of 15 developed currencies (black dotted line).[78] As expected, transaction costs to trade emerging currencies are substantially larger than developed currencies. Transaction costs generally decrease over time, except during FX market crises, which do not necessarily coincide with NBER recessions (grey shaded areas). The costs reach low levels between 0.04 and 0.06 cents per dollar trade in the final year of our sample. Notice, however, that these low numbers do not necessarily imply that transaction costs are unimportant nowadays. Since forward discounts in FX markets of developed and many emerging currencies have approached zero in the past decade, carry traders have often started to shift their focus towards carry trades in frontier markets, which feature substantially higher transaction costs.

Notice that all strategies (MV, MV_{TC}, $MV_{TC\backslash Corr}$) use information (i.e. estimates for $\boldsymbol{\mu}_{\mathbf{t}}^{\mathbf{e}}$, $\mathbf{V_t}$, and $\mathbf{C}_{\mathbf{i,t}}^{\mathbf{z}}$ $\forall z \in \{P+, S+, P-, S-\}$) available at the end of month t to construct a portfolio which we then hold until the end of the subsequent month $t+1$. Thus, all returns are out-of-sample and none of the trading strategies suffers from a look-ahead bias.

3.3.2 Data

We collect daily spot and 1-month forward bid, ask and mid exchange rates from Barclays Bank International and Reuters via Datastream. We use quotes of the last day of the month to compute monthly returns $CT_{i,t+1}$. A concern with currencies of emerging countries is that there are capital controls and major trading frictions. Menkhoff et al. (2012) and

[78] The cross-sectional average of costs is computed as $\frac{1}{4\times N}\sum_{i=1}^{N}\left(\mathbf{C}_{\mathbf{i,t}}^{\mathbf{P}}+\mathbf{C}_{\mathbf{i,t}}^{\mathbf{P-}}+\mathbf{C}_{\mathbf{i,t}}^{\mathbf{S+}}+\mathbf{C}_{\mathbf{i,t}}^{\mathbf{S-}}\right)$, where N is the number of exchange rates for which we have data available at time t.

Della Corte et al. (2016) suggest to exclude countries with a negative score on the capital account openness index of Chinn and Ito (2006).[79] Following this literature, we include currencies of 29 countries in our analysis. According to Lustig et al. (2011) 15 of them are classified as "developed", while the remaining 14 are "emerging" countries. The 15 developed countries are: Australia, Belgium, Canada, Denmark, Euro Area, France, Germany, Italy, Japan, Netherlands, New Zealand, Norway, Sweden, Switzerland, United Kingdom. The 14 emerging countries are: Brazil, Czech Republic, Greece, Hungary, Iceland, Ireland, Mexico, Poland, Portugal, Singapore, South Africa, South Korea, Spain, Taiwan. The Euro was introduced in January 1999 and we exclude all countries which have joined the Euro after that date and only keep the Euro as a currency.

Exchange rates of all 29 currencies are quoted against the USD for the sample starting on October 11th, 1983 and ending on March 2nd, 2016. We are able to extend our sample further back to January 2nd, 1976 for the following subset of 14 countries with exchange rates quoted against the GBP (Great British Pound): Austria, Canada, France, Germany, Ireland, Italy, Japan, Netherland, Norway, Portugal, Spain, Sweden, Switzerland, USA. For the period from January 2nd, 1976 to October 11th, 1983 we convert all data to exchange rates quoted against the USD using mid exchange rate quotes of USD/GBP.

[79] We further exclude a currency at time t if more than 20% of its daily exchange rate growths are missing over the past 9 months, or if the absolute value of the annualized forward discount $12 \times |fd_{i,t}|$ is larger than 25%. Forward discounts of more than 25% are rare and we believe that such large values likely indicate non-tradable outliers in the data, the presence of severe trading frictions, sizable sovereign default risk or an extraordinary large expected currency devaluation. Under these conditions, a currency trader is likely not able or willing to consider a currency as part of the investment opportunity set.

3.4 Results

Our main result is that the out-of-sample performance after transaction costs of mean-variance efficient portfolios substantially improves if transaction costs are taken into account in the optimization. We document a statistically significant and economically large outperformance (after transaction costs) of MV_{TC} over MV. Moreover, we quantify the empirical importance of the four theoretical predictions discussed in Section 3.2.

3.4.1 Performance Before Transaction Costs

Prediction 1: *MV is expected to outperform MV_{TC} if the performance is measured in returns before transaction costs.*

Table 3.1 quantifies the difference between MV and MV_{TC} and summarizes the monthly out-of-sample excess returns of both strategies for our full set of 29 currencies (columns 1 and 2) and the subset of 15 developed currencies (columns 3 and 4) from January 1976 to February 2016. The first panel of Table 3.1 reports the Sharpe ratios (SR) and average excess returns (Mean) before transaction costs. The annualized Sharpe ratios of MV and MV_{TC} are almost identical: 0.99 and 1.00 for the set of all 29 currencies, and 0.87 and 0.82 for the set of 15 developed currencies. The difference in Sharpe ratios between MV and MV_{TC} is not significant (neither in the set of all 29 nor the 15 developed currencies). The average annual return before transaction costs of MV_{TC} is about 1.5% lower than the average return of MV (denoted by ΔMean in Table 3.1), but the volatility (Vol) of MV_{TC} is also proportionally lower, which implies almost identical Sharpe ratios across the two strategies.

The top panel of Figure 3.4 displays the cumulative returns of MV (black dashed line) and MV_{TC} (red solid line) before transaction costs for our full set of 29 currencies. The two time-series closely track each other and the returns of the two strategies are almost identical at every point in time. We highlight two crash periods: (i) the 1992 European Monetary System (ERM) crisis, which led to a temporary suspension of the Italian Lira and the UK Sterling from the ERM, and (ii) the 1997 Asian financial crisis and 1998 default of Russia. In both periods the before transaction cost returns of MV and MV_{TC} are almost identical, i.e., our results are not affected by these extreme events. In Section 3.5 we provide results from a robustness analysis where we exclude these crises.

To conclude, we do not find a significant difference in the performance before transaction costs between MV and MV_{TC}. Although MV_{TC} trades less actively than MV due to the no trading region and generally holds an ex-ante sub-optimal position[80], the ex-post performance before transaction costs is almost identical. That is, while it is theoretically true that MV_{TC} is sub-optimal in terms of a before transaction costs evaluation, this first theoretical prediction is empirically irrelevant.

3.4.2 Transaction Costs

Prediction 2: *Transaction costs paid by MV are expected to be higher than by* MV_{TC}.

The second panel in Table 3.1 reports the average transaction costs paid per year as a percentage of the portfolio value (or alternatively as a reduction in the portfolio return). The costs paid by MV are substantial, i.e., 3.71% for the set of 29 currencies and 1.97% for the set of 15 developed currencies. That is, 20%-30% of MV's expected return is lost to

[80]Sub-optimal if there are no transaction costs.

transaction costs. The costs paid by MV_{TC} are less than half the size of the costs paid by MV, i.e., 1.28% for the set of 29 currencies and 0.8% for the set of 15 developed currencies. These savings in transaction costs are economically large. Moreover, the difference in costs between MV and MV_{TC} is highly statistically significant for both the set of 29 currencies and the subset of 15 developed currencies.

Figure 3.5 visualizes this striking result by plotting the time-series of cumulative transaction costs (top panel) and the monthly costs (bottom panel) paid by MV (black dashed line) and MV_{TC} (red solid line) for our full set of 29 currencies. The spread between the cumulative costs of MV and MV_{TC} is steadily increasing, while the monthly costs incurred by MV are without exception always larger than the costs of MV_{TC}. Therefore, our second theoretical prediction that MV is subject to larger transaction costs than MV_{TC} is empirically important.

3.4.3 Performance After Transaction Costs

Prediction 3: *MV_{TC} is expected to outperform MV after transaction costs. Moreover, the outperformance is expected to be more substantial if transaction costs are large.*

The third panel in Table 3.1 compares returns after transaction costs. The Sharpe ratios after transaction costs are highlighted in boldface. For the full set of 29 currencies, the annualized Sharpe ratio of MV is 0.7 and the one of MV_{TC} is 0.9. The difference of $\Delta\text{SR} = 0.19$ is economically meaningful. MV_{TC} is compensated by an almost 2% higher annual expected return than MV per 10% return volatility (which is roughly equal to the unconditional volatility of a typical carry trade strategy). We further find that this difference is statistically significant with a p-value of 0.007. We employ the test proposed by Ledoit and Wolf (2008),

which uses block bootstrapping and is robust to heteroskedasticity and cross- and auto-correlation.[81] The bottom panel in Figure 3.4 illustrates this striking dominance of MV_{TC} by plotting cumulative returns after transaction costs. The spread in cumulative returns after costs is steadily opening. Neither the aforementioned crises have a noteworthy effect nor are our result driven by outliers.[82] This result suggests that our third theoretical prediction is empirically important. Optimizing over transaction costs when constructing mean-variance efficient portfolios substantially improves the out-of-sample performance.

For the set of 15 developed currencies, we also find that MV_{TC} outperforms MV after transaction costs. The Sharpe ratios are 0.75 for MV_{TC} and 0.7 for MV. While the difference in Sharpe ratios $\Delta\text{SR} = 0.05$ is smaller than in the case of the full set of 29 countries, it is still economically important. Per 10% volatility, MV_{TC} earns 0.5% more per year than MV. The difference in Sharpe ratios is not statistically significant with a p-value of 0.385. This may be due to the low power, i.e., transaction costs are relatively small among the developed currencies and thus, we would need a lot of data to identify a statistically significant difference. The finding that ΔSR is larger for the full set of 29 currencies is consistent with the fact that transaction costs are larger among emerging than developed currencies. Indeed, we expect that the implications of transaction costs are more important if average costs and the no trading region of MV_{TC} are large.

In addition to the Sharpe ratio analysis, we investigate the (ex-post) utility gain when switching from MV to MV_{TC}. The last four rows of Table 3.1 report the annualized return or certainty equivalent CE_λ a mean-variance investor with risk aversion $\lambda \in \{1, 5, 10, 50\}$ is

[81] We choose a block size of 10 observations for the block bootstrapping. This is a conservative value and our results are stronger if we use smaller block sizes which are closer to what Ledoit and Wolf (2008) suggest in their illustrations.

[82] As a robustness we repeat our analysis excluding the two crises and find similar results. The robustness results are in Section 3.5.

willing to give up in order to switch from MV to MV_{TC}. In parenthesis next to CE_λ we report the percentage of months with a certainty equivalent larger than 0 (% of $CE_\lambda > 0$), or equivalently the months in which the investor with risk aversion λ (ex-post) prefers MV_{TC} over MV. The monthly certainty equivalent at time t is calculated using the realized return in month t as a proxy for the conditional expected return and the daily returns within the month to estimate the conditional variance. For our set of 29 currencies, a log investor ($\lambda = 1$) is willing to give up 1.16% to switch from MV to MV_{TC}, and $CE_\lambda > 0$ in 75% of all months. For an investor with λ equal to 5, 10 or 50, CE_λ increases to 1.89%, 2.8% or 10.1% and the percentage of monthly observations with $CE_\lambda > 0$ increases to 80%, 82% or 83%. For the set of 15 developed currencies, the certainty equivalents are smaller, i.e., for $\lambda \in \{1, 5, 10, 50\}$, $CE_\lambda \in \{-0.31\%, 0.22\%, 0.87\%, 6.11\%\}$ and the percentage of $CE_\lambda > 0$ are 50%, 63%, 66% and 72%. The monotonically increasing relation highlights that more risk averse investors have a stronger desire to manage transaction costs efficiently.

We further investigate how much less MV_{TC} is trading compared to MV due to the no trading region. Therefore, we plot the time-series of the turnover $\sum_i \|\theta_{i,t} - \theta_{i,t-1}\|$ of MV (black dashed line) and MV_{TC} (red solid line) in the top panel in Figure 3.6 for our full set of 29 currencies. The turnover of MV is on average 2.5 times larger than the turnover of MV_{TC}. In the bottom panel of Figure 3.6, we report the average portfolio holdings and 1-standard deviation error bars of MV (downward pointing triangles and thin black lines) and MV_{TC} (upward pointing triangles and thick red lines). The average portfolio holdings are similar across the two strategies but the standard deviation is substantially larger for MV, which indicates more trading activity.

All other moments of returns after transaction costs are comparable across MV and MV_{TC}. Table 3.1 lists the monthly return skewness (Skew), kurtosis (Kurt), the percentage of positive monthly returns (% Positive), the maximum draw down (MDD), which measures the maximum loss from peak to trough of the strategy in the entire sample, and the autocorrelation (AC). If anything the skewness and the MDD of MV_{TC} are more favorable than the ones of MV, suggesting that MV_{TC} has less crash risk exposure than MV.

Finally, we plot the time-series of the notional value or total dollar exposure $\sum_i \|\theta_{i,t}\|$ of MV (black dashed line) and MV_{TC} (red solid line) for our full set of 29 currencies in Figure 3.7. The notional value is slightly smaller for MV_{TC} than for MV and almost always below 15. Only during the 1997 Asian financial crisis and the 1998 default of Russia the notional value spiked to levels of around 35. Margin requirements in FX derivatives markets are low and implementing a strategy with a notional value of 35 is typically unproblematic.

To sum up, we recall that the performance before transaction costs of MV and MV_{TC} are almost identical, which implies that our first theoretical prediction is empirically irrelevant. However, MV faces substantially larger transaction costs than MV_{TC}, and in turn, MV_{TC} substantially outperforms MV after transaction costs. Thus, the second and third theoretical predictions are empirically important. The results are driven by the significant reduction of unnecessary trading, which substantially lowers the turnover and transaction costs and increases the performance after transaction costs of MV_{TC} compared to MV. Accounting for transaction costs in the portfolio optimization is particularly important if costs are large.

These results have important practical implications. First, accounting for costs when optimizing a portfolio is beneficial and improves the out-of-sample performance. Second, transaction costs are declining over time, and thus, traders who specialize in developed currencies

may be tempted to ignore transaction costs when constructing mean-variance efficient portfolios. However, even if transaction costs are low during normal times, they substantially increase during crises and become relevant (Karnaukh et al. (2015)). Third, many currency traders have shifted their focus to emerging and frontier markets because exchange rate forward discounts among developed currencies are close to zero for the past decade. Transaction costs in emerging and frontier markets are generally larger than the costs considered in our analysis, and thus, the implications of transaction costs on the optimal portfolio choice are even more important for these traders.

3.4.4 Importance of Correlations between Assets

Prediction 4: *If assets are positively correlated, then the no trading region of MV_{TC} is expected to be larger than the one of $MV_{TC\setminus Corr}$.*[83] *Moreover, transaction costs of MV_{TC} are expected be lower than the costs of $MV_{TC\setminus Corr}$, and we expect MV_{TC} to outperform $MV_{TC\setminus Corr}$ after costs.*

Figure 3.8 shows the time-series of the average conditional correlation of each exchange rate growth i with all other exchange rate growths, $\rho_{i,t} = \frac{1}{N-1}\sum_{j=1}^{N-1} Corr_t\left(\Delta x_{i,t}, \Delta x_{j,t}\right)$ for our full set of 29 currencies. To estimate the conditional correlation $Corr_t\left(\Delta x_{i,t}, \Delta x_{j,t}\right)$ between exchange rate growths i and j in month t we use daily exchange rate growths within the month. The bold black line is the average of all correlations $\rho_t = \frac{1}{N-1}\sum_{i=1}^{N} \rho_{i,t}$ in month t. Correlations $\rho_{i,t}$ are almost always positive and on average close to 0.5. The average correlation ρ_t is always between 0.1 and 0.8.

[83] Recall that $MV_{TC\setminus Corr}$ is the strategy which optimizes over transaction costs similar to MV_{TC} but assumes that assets are uncorrelated to simplify the construction of the no trading region and obtain an approximate solution.

Table 3.2 summarizes the monthly excess returns of MV, $MV_{TC\backslash Corr}$ and MV_{TC} for our full set of 29 currencies from 1976 to 2016.[84] Note that MV and MV_{TC} are also described above and in Table 3.1. Consistent with the previous finding, the average returns and Sharpe ratios before transaction costs are almost identical across the three strategies. $MV_{TC\backslash Corr}$ has transaction costs of 2.56% per year, which is a 1.14% saving in costs compared to MV but 1.24% larger than the costs incurred by MV_{TC}. After transaction costs, the Sharpe ratio of $MV_{TC\backslash Corr}$ is 0.76, which is 0.06 higher than the ratio of MV but 0.14 lower than the ratio of MV_{TC}. The difference in Sharpe ratios between MV and $MV_{TC\backslash Corr}$ is not statistically significant (p-value of 0.428) but the difference between $MV_{TC\backslash Corr}$ and MV_{TC} is significant (p-value of 0.084). Therefore, accounting for correlations in the optimization is important to significantly increase the Sharpe ratio when optimizing over transaction costs. Employing the approximate solution $MV_{TC\backslash Corr}$ to optimize over transaction costs adds not much benefit.

The certainty equivalent CE_λ an investor with risk aversion $\lambda \in \{1, 5, 10, 50\}$ is willing to pay to switch from $MV_{TC\backslash Corr}$ to MV is mostly slightly negative (i.e., $MV_{TC\backslash Corr}$ is preferred to MV), but when $\lambda = 50$, it is positive (i.e., MV is preferred to $MV_{TC\backslash Corr}$). The percentages of months where MV is preferred to $MV_{TC\backslash Corr}$ are 29%, 32%, 35% and 44% when λ is 1, 5, 10 and 50. In contrast, the certainty equivalent is substantially larger for a switch from $MV_{TC\backslash Corr}$ to MV_{TC}, and it is increasing in the risk aversion λ of the investor. The CE_λ to switch from $MV_{TC\backslash Corr}$ to MV_{TC} are 0.26%, 1.15%, 2.27% and 11.19% when λ is 1,5,10 and 50. The corresponding percentages of months when MV_{TC} is preferred to $MV_{TC\backslash Corr}$ are 57%, 68%, 72% and 78%.

[84] We focus on our full set of 29 currencies and do not report the the results for our subset of 15 developed currencies. Latter results are available on request.

We conclude that the no trading regions of $MV_{TC\setminus Corr}$ and MV_{TC} are not only theoretically but also quantitatively very different. Accounting for correlations in the optimization over transaction costs is empirically important and the out-of-sample outperformance of MV_{TC} over $MV_{TC\setminus Corr}$ is economically and statistically significant. On the other hand, the outperformance of $MV_{TC\setminus Corr}$ over MV is empirically small. Therefore, it is not beneficial to account for transaction costs in the portfolio optimization while imposing the simplifying assumption that assets are uncorrelated when constructing the no trading region. This empirical finding is an important contribution to the literature: it invalidates the general usefulness of the continuous time framework of H. Liu (2004), where assets are considered uncorrelated. This is unfortunate since H. Liu (2004)'s setup provides so far the only available model able to deliver a solution in dynamic portfolio optimization settings in the presence of transaction costs with many risky assets.

3.4.5 Size of the No Trading Region and Trade Aggressiveness

In our theoretical discussion in Section 3.2 we establish that if the investor optimizes over transaction costs (i.e. MV_{TC} or $MV_{TC\setminus Corr}$), she trades from her initial position $\theta_{\mathbf{t}}^{\mathbf{0}}$ towards $\theta_{\mathbf{t}}^{\mathbf{MV}}$ but stops at the boundary of the no trading region. We measure the size of the no trading region by $1 - \overline{TA}\left(\theta_{\mathbf{t}}^{\mathbf{S}}\right)$, where the trade aggressiveness $\overline{TA}\left(\theta_{\mathbf{t}}^{\mathbf{S}}\right)$ of strategy $S \in \{MV_{TC}, MV_{TC\setminus Corr}\}$ is defined as the ratio of the turnover of strategy S and MV,

$$\overline{TA}\left(\theta_{\mathbf{t}}^{\mathbf{S}}\right) = \frac{\sum_i \|\theta_{\mathbf{i,t}}^{\mathbf{S}} - \theta_{\mathbf{i,t}}^{\mathbf{0}}\|}{\sum_i \|\theta_{\mathbf{i,t}}^{\mathbf{MV}} - \theta_{\mathbf{i,t}}^{\mathbf{0}}\|} \in [0,1]\,.$$

The turnover as a distance measure is suitable because we want to quantify the total amount of trade. Normalizing by the turnover of MV helps to put this distance into perspective.

A large $\overline{TA}(\theta_{\mathbf{t}}^{\mathbf{S}})$ indicates that the investor trades aggressively and chooses a position $\theta_{\mathbf{t}}^{\mathbf{S}}$ close to $\theta_{\mathbf{t}}^{\mathbf{MV}}$, which in turn implies that the no trading region is small. In the extreme case where $\overline{TA}(\theta_{\mathbf{t}}^{\mathbf{S}}) = 1$, $\theta_{\mathbf{t}}^{\mathbf{S}} = \theta_{\mathbf{t}}^{\mathbf{MV}}$ and there does not exist a no trading region. In contrast, a small value indicates that the investor does not trade aggressively and $\theta_{\mathbf{t}}^{\mathbf{S}}$ is far away from $\theta_{\mathbf{t}}^{\mathbf{MV}}$, which in turn means that the no trading region is large. In the extreme case where $\overline{TA}(\theta_{\mathbf{t}}^{\mathbf{S}}) = 0$, strategy S does not trade at all, $\theta_{\mathbf{t}}^{\mathbf{S}} = \theta_{\mathbf{t}}^{\mathbf{0}}$, and the initial position lies within the no trading region. Thus, $\overline{TA}\left(\theta_{\mathbf{t}}^{\mathbf{S}}\right)$ measures how aggressive an investor trades from the initial position θ_t^0 towards the optimum without transaction costs θ_t^{MV}, and $1 - \overline{TA}\left(\theta_{\mathbf{t}}^{\mathbf{S}}\right)$ quantifies the size of the no trading region of strategy S.

In the one period model discussed in Section 3.2 we choose the initial position $\theta_{\mathbf{t}}^{\mathbf{0}}$ exogenously. However, in our empirical implementation the initial position in month t is equal to the portfolio allocation chosen in month $t-1$. Thus, the initial position for strategy S is $\theta_{\mathbf{t-1}}^{\mathbf{S}}$, while it is $\theta_{\mathbf{t-1}}^{\mathbf{MV}}$ for strategy MV. These initial positions are generally not identical, and therefore, our actual trade aggressiveness measure $TA(\theta_{\mathbf{t}}^{\mathbf{S}}) = \frac{\sum_i \|\theta_{\mathbf{i,t}}^{\mathbf{S}} - \theta_{\mathbf{i,t-1}}^{\mathbf{S}}\|}{\sum_i \|\theta_{\mathbf{i,t}}^{\mathbf{MV}} - \theta_{\mathbf{i,t-1}}^{\mathbf{MV}}\|} \approx \overline{TA}\left(\theta_{\mathbf{t}}^{\mathbf{S}}\right)$ is not anymore bounded above by 1.

Table 3.3 summarizes the monthly realizations of $TA\left(\theta_{\mathbf{t}}^{\mathbf{S}}\right)$ for strategies $S \in \{MV_{TC}, MV_{TC\setminus Corr}\}$ for our full set of 29 currencies from 1976 to 2016.[85] On average the trade aggressiveness of strategy MV_{TC} is 0.41. That is, the investor reduces trading by 59% compared to MV. Moreover, the 5- and 95-percentiles of the trade aggressiveness of MV_{TC} are 0.14 and 0.74, which means that the trading activity of MV_{TC} is most of the time substantially lower compared to MV. An investor, who follows strategy $MV_{TC\setminus Corr}$ on average has a trade aggressiveness of 0.98, i.e., reduces trading by only 2% compared to MV. The 5- and 95-percentiles are 0.69 and 1.39. The value larger than 1 indicates that $MV_{TC\setminus Corr}$

[85] We focus on our full set of 29 currencies and do not report the the results for our subset of 15 developed currencies. Latter results are available on request.

sometimes trades even more than MV. As explained above, this is possible because the initial position of $MV_{TC\setminus Corr}$ can be quite different from the initial position of MV. Our empirical measure is not bounded above by 1 as it is in the single period setting where both strategies have the same initial position.

The difference between the two measures, $\Delta\text{TA} = TA\left(\theta_{\mathbf{t}}^{\mathbf{MV_{TC\setminus Corr}}}\right) - TA\left(\theta_{\mathbf{t}}^{\mathbf{MV_{TC}}}\right)$, is of particular interest. It measures how much more aggressively an investor, who implements $MV_{TC\setminus Corr}$, trades compared to an investor, who invests in MV_{TC}. We expect $\Delta\text{TA} > 0$ if correlations between assets are (predominantly) positive, and $\Delta\text{TA} < 0$ if correlations between assets are (predominantly) negative. This is because the no trading region of MV_{TC} is larger (smaller) than the one of $MV_{TC\setminus Corr}$ when correlations are positive (negative) (the intuition is that when assets are positively correlated they act as substitutes while if they are negatively correlated they function as complements, see the discussion in Section 3.2). We find that ΔTA is on average 0.57 and statistically significantly different from 0 (top panel in Table 3.3). The median of ΔTA is 0.54 and the 5- and 95-percentiles are 0.2 and 1.03. This implies that the no trading region of MV_{TC} is generally larger than the one of $MV_{TC\setminus Corr}$, which is consistent with the fact that correlations between exchange rates are on average positive (Figure 3.8).

Figure 3.9 further analyzes the time series of ΔTA (solid line). The horizontal dashed line highlights its sample median. The gray shaded areas indicate NBER recessions. Although MV_{TC} generally trades less aggressive than $MV_{TC\setminus Corr}$, there are a couple of monthly observations where the opposite is true. Finally, we highlight the Asian financial crisis in 1997, when ΔTA spikes. The striking increase is not surprising since correlations typically sharply increase during crises.

3.4.6 Heuristic Adjustment of Static Solution to Approximate the Dynamic Problem

MV_{TC} is the optimal solution in a single period but not necessarily in a multi-period framework. Suppose we extend our model to T periods. The investor trades in every period t and her utility at time t is $U_t = E_t\left[\sum_{\tau=t}^{T} \beta^{\tau-t} u_\tau\right]$ where $E_t[.]$ is the conditional expectation operator, $\beta \in [0,1]$ is a subjective time discount factor of future period τ mean-variance utility $u_\tau = \theta_\tau{}'\mu_\tau^{\mathbf{e}} - \frac{\lambda}{2}\theta_\tau{}'\mathbf{V}_\tau\theta_\tau - \mathbf{\Delta}_\tau^{\mathbf{P}+\prime}\mathbf{C}_\tau^{\mathbf{P}+} - \mathbf{\Delta}_\tau^{\mathbf{P}-\prime}\mathbf{C}_\tau^{\mathbf{P}-} - \mathbf{\Delta}_\tau^{\mathbf{S}+\prime}\mathbf{C}_\tau^{\mathbf{S}+} - \mathbf{\Delta}_\tau^{\mathbf{S}-\prime}\mathbf{C}_\tau^{\mathbf{S}-}$. Moreover, suppose the investment opportunity set is constant, i.e., $\mu_\tau^{\mathbf{e}} = \mu^{\mathbf{e}}$, $\mathbf{V}_\tau = \mathbf{V}$, $\mathbf{C}_\tau^{\mathbf{z}} = \mathbf{C}^{\mathbf{z}}$ $\forall z \in \{P+, S+, P-, S-\}$. If there are no transaction costs ($\mathbf{C}^{\mathbf{P}+} = \mathbf{C}^{\mathbf{S}+} = \mathbf{C}^{\mathbf{P}-} = \mathbf{C}^{\mathbf{S}-} = 0$), then it is well-known that the optimal solution in every period t is the same as the solution in the single period model, $\theta_{\mathbf{t}}^{\mathbf{MV^T}} = \theta^{\mathbf{MV}} = \frac{1}{\lambda}\mathbf{V}^{-1}\mu^{\mathbf{e}}$, where the superscript T indicates that the portfolio is the solution to the T-period setting. In contrast, if there are positive transaction costs ($\mathbf{C}^{\mathbf{P}+} > 0, \mathbf{C}^{\mathbf{S}+} > 0, \mathbf{C}^{\mathbf{P}-} > 0, \mathbf{C}^{\mathbf{S}-} > 0$), then in general the optimal solution is not equal to the single period solution, $\theta_{\mathbf{t}}^{\mathbf{MV_{TC}^T}} \neq \theta^{\mathbf{MV_{TC}}}$.

Unfortunately, we do not have reliable algorithms to solve the multi-period model in the presence of many assets.[86] Dybvig and Pezzo (2018) only provide a solution to the single period model. Intuitively, we expect the no trading region of the multi-period strategy MV_{TC}^T to be smaller than for the single period strategy MV_{TC}. The marginal utility gain from moving towards $\theta^{\mathbf{MV}}$ should be larger in the multi-period than in the single period model because the benefit of being close to $\theta^{\mathbf{MV}}$ can be reaped for multiple periods instead

[86] Our results show how important it is to properly account for correlations among assets: a fact that empirically invalidate the H. Liu (2004)' setup, based on the assumption of no correlation among the available assets. H. Liu (2004)'s model is the only framework that so far has been able to deliver solutions for dynamic portfolio optimizations in the presence of transaction costs with many assets. Due to the complexity of the general problem, the current literature on transaction costs only provide heuristic/approximate solutions when dealing with two and in rare cases three risky assets.

of only once. This intuition carries over to settings with stochastic changes in the investment opportunity set (though there is an additional level of complexity due to hedging demands). In particular, we expect the size of the no trading region to depend inversely on the persistence in the state variables. Two extreme cases are (i) independent shocks to the investment opportunity set where the no trading region is expected to be large, and (ii) the constant investment opportunity set where we expect a relatively small no trading region.

Following our intuition, we propose the following heuristic solution to the multi-period model. We define the cost multiplier $\mathbf{M_i}(c,a) = c + a \times \rho\left(\frac{\mu^{\mathbf{e}}_{\mathbf{i,t}}}{\sigma^2_{i,t}}\right)$, where $\rho(x_t)$ is the first autocorrelation operator of the time-series x_t and $\sigma^2_{i,t}$ is the ith diagonal element of $\mathbf{V_t}$, and the adjusted transaction costs associated with asset i are $\mathbf{C}^z_{i,t}(c,a) = \mathbf{M_i}(c,a)\mathbf{C^z_{i,t}}$ $\forall z \in \{P+, S+, P-, S-\}$. We conjecture that the solution $\theta^{\mathbf{MV^M_{TC}}}(c,a)$ of Problem 9 with the adjusted transaction costs approximates the true solution $\theta^{\mathbf{MV^T_{TC}}}_{\mathbf{t}}$ in the multi-period model. Notice that $\theta^{\mathbf{MV^M_{TC}(c=1,a=0)}} = \theta^{\mathbf{MV_{TC}}}$ and $\theta^{\mathbf{MV^M_{TC}(c=0,a=0)}} = \theta^{\mathbf{MV}}$ nests the single period model solutions with and without transaction costs.

We empirically assess the importance of our proposed approximate solution of the multi-period model for our full set of 29 currencies from 1976 to 2016.[87] We construct $\theta^{\mathbf{MV^M_{TC}(c,a)}}$ for $c \in [0, +\infty)$ and $a \in (-\infty, +\infty)$ and compute out-of-sample returns. Figure 3.10 plots the annualized out-of-sample Sharpe ratios against parameters $(c,a) \in [0,2] \times [0,1]$, which represent the neighborhood of the global optimum. The point highlighted by a blue arrow indicates MV_{TC} with a Sharpe ratio of 0.90. The point highlighted by a red arrow indicates $MV^M_{TC}(c = 0.7, a = 0.8)$ with a Sharpe ratio of 0.92. This portfolio yields the highest Sharpe ratio for any combination of (c,a). The point highlighted by a black arrow indicates $MV^M_{TC}(c = 1.3, a = 0)$ with a Sharpe ratio of 0.91. This portfolio yields the highest Sharpe

[87] We focus on our full set of 29 currencies and do not report the the results for our subset of 15 developed currencies. Latter results are available on request.

ratio for any value c and $a = 0$. Points indicated by blue crosses are portfolios $MV_{TC}^{M}(c, a)$ with Sharpe ratios which are statistically significantly different from the Sharpe ratio of MV_{TC} (using the test of Ledoit and Wolf (2008) and a p-value of 0.05). Figure 3.10 suggests that the Sharpe ratio is not sensitive to changes in parameters c and a within a large neighborhood around the maximum with $(c, a) = (0.7, 0.8)$, including strategy MV_{TC} with $(c, a) = (1, 0)$. Thus, our proposed heuristic approximation of the multi-period model solution does not improve the out-of-sample performance over the single period model solution MV_{TC}.[88]

A potential reason for the just discussed heuristic approximation not to find significant results is the lack of consideration of any mean-reversion effects in the optimal weights $\theta^{\mathbf{MV_{TC}^{T}}}$. Under the assumption of a constant investment opportunity set there exists a true (but unobservable) stationary optimum $\theta^{\mathbf{MV}}$, which can be thought of as the vector of long run mean weights absent costs. We take the average of the optimal weights $\{\theta_{\mathbf{t}}^{\mathbf{MV}}\}_t$ in our sample from January 1976 to February 2016 as such a proxy. By construction, the actual time-series of $\theta_{\mathbf{t}}^{\mathbf{MV}}$ hoovers around it, more or less closely (depending on the width of the no-trading region) followed by the time series of $\theta_{\mathbf{t}}^{\mathbf{MV_{TC}}}$. In a dynamic setting at generic date t we should expect any re-balancing decision $\theta_{\mathbf{t}}^{\mathbf{MV_{TC}^{T}}}$ to be sub-optimal if it does not bring the inherited position $\theta_{\mathbf{t-1}}^{\mathbf{MV_{TC}^{T}}}$ any closer to our proxy for the true long-run optimum $\theta^{\mathbf{MV}}$. We therefore expect two different multipliers of the form $\mathbf{M}(c_1, a_1)$ and $\mathbf{M}(c_2, a_2)$ to be respectively smaller and bigger than the vector of ones $1_{\{N\times N\}}$ when they push $\theta_{\mathbf{t}}^{\mathbf{MV_{TC}}}$ towards respectively away from $\theta^{\mathbf{MV_{TC}}}$. Accordingly, we construct the new approximate solution by designing costs of the form $\mathbf{C_{i,t}^{z}}(c_1, a_1, c_2, a_2) = [\alpha_{i,t}\mathbf{M_i}(c_1, a_1) + (1 - \alpha_{i,t})\mathbf{M_i}(c_2, a_2)]\,\mathbf{C_{i,t}^{z}}$

[88] However, we cannot exclude the possibility that the true solution of the multi-period model outperforms MV_{TC} in out-of-sample tests, because we do not know how different our conjectured approximate solution is from the true solution.

$\forall z \in \{P+, S+, P-, S-\}$ where $\alpha_{i,t}$ is an indicator that turns on when $\theta_{\mathbf{i,t}}^{\mathbf{MV_{TC}}}$ pushes $\theta_{\mathbf{i,t-1}}^{\mathbf{MV_{TC}}}$ towards $\theta_{\mathbf{i}}^{\mathbf{MV_{TC}}}$.

Again, we empirically assess the importance of the new approximate dynamic solution for $c_i \in [0, +\infty)$ and $a_i \in (-\infty, +\infty)$ with $i \in \{1, 2\}$ in terms of the out-of-sample returns for our full set of 29 currencies from 1976 to 2016.[89] The new maximal Sharpe ratio is 0.92, the same generated by the former heuristic, and it is not statistically different from the 0.90 Sharpe ratio of our MV_{TC} strategy.

Overall, based on the insights from our heuristic analysis, we expect the myopic MV_{TC} strategy not to be very far from the true optimal dynamic strategy.

3.5 Robustness

Table 3.4, 3.5 and 3.6 provide robustness results of our main findings in Table 3.1. Our focus is on our full set of 29 developed and emerging currencies because transaction costs are generally larger and more relevant than in the subset of 15 developed currencies (see discussion in Section 3.4 for details). Results for our subset of 15 developed currencies are available on request.

3.5.1 Sample without Crises, 1976-2016

The first two columns of Table 3.4 report the out-of-sample performance of MV and MV_{TC} for our full set of 29 currencies from 1976 to 2016 but excluding observations during the 1992

[89] We focus on our full set of 29 currencies and do not report the the results for our subset of 15 developed currencies. Latter results are available on request.

ERM crisis, the 1997 Asian Financial crisis and 1998 Russian default. These portfolios are not actually traded since we would not have been able to predict these crises in real time. However, the analysis is still useful to understand whether any of our results are driven by these periods.

Not surprisingly the performance of both strategies improves when we remove the observations during the crises. The Sharpe ratios before transaction costs of MV and MV_{TC} increase from 0.99 and 1.00 to 1.15 and 1.17. The Sharpe ratios before transaction costs across the two strategies are almost identical. On the other hand, the excluded crises do not impact the costs: the implementation of MV costs 3.63% per year and MV_{TC} 1.22% (including crises 3.71% and 1.28% respectively). MV_{TC} saves a significant amount compared to MV and the difference in costs between MV and MV_{TC} is highly statistically significant. The Sharpe ratios after transaction costs are higher in the sample without crises than in the full sample. MV earns a Sharpe ratio of 0.81 and MV_{TC} 1.04. The difference between the strategies is 0.23 and statistically significant with a p-value of 0.005, which is similar to the result from full sample. To sum up, the first theoretical prediction is empirically irrelevant while the second and third predictions are economically large and statistically significant. Therefore, our conclusions drawn from our sample without crises are the same as the conclusions in Section 3.4.

3.5.2 Sample from November 1983 to February 2016

Our main analysis uses the sample from January 2nd, 1976 to March 2nd, 2016. The data before October 11th, 1983 is quoted against the Great British Pound (GBP), and we convert all data to exchange rates quoted against the USD (using mid quotes between the USD and

GBP). The data quoted against the GBP are less reliable compared to the later sample quoted against the USD. Moreover, the bid and ask quotes after converting the 1976-1983 data to quotes against the USD do not exactly reflect the true bid and ask quotes against the USD, i.e., they are the bid and ask quotes against the GBP converted by the mid quote between USD and GBP. We show that our results are robust independent of whether we use the full sample from 1976 to 2016 or the shorter sample from 1983 to 2016.

Columns 3-4 of Table 3.4 summarize the out-of-sample excess returns of MV and MV_{TC} for our full set of 29 currencies from November 1983 to February 2016. The Sharpe ratios before transaction costs of MV and MV_{TC} are 0.87 and 0.88, which is 0.12 lower than in the full sample from 1976 to 2016. The costs paid by MV and MV_{TC} are 2.80% and 1.08% per year. These numbers are lower than in the full sample, which is consistent with the fact that average transaction costs are decreasing over time (Figure 3.3). MV_{TC} costs less than 40% of MV to implement. The difference in costs between MV and MV_{TC} is highly statistically significant. The Sharpe ratios after transaction costs are also lower in the sample starting in 1983 than in the full sample. The Sharpe ratio of MV is 0.66 and the one of MV_{TC} is 0.79. The difference between the strategies is 0.14 and statistically significant with a p-value of 0.025. Therefore, consistent with our main results, the first theoretical prediction is empirically irrelevant while the second and third predictions are important.

3.5.3 NBER Recessions

Next we investigate the impact of recessions. We confirm our results in both NBER recessions and non-recession periods. Table 3.5 summarizes the monthly excess returns of MV and MV_{TC} during NBER recessions (columns 1-2) and during non-recession periods (columns

3-4) for our full set fo 29 currencies from 1976 to 2016. Sharpe ratios before transaction costs are twice in non-recession periods than during recessions. This difference is driven by a large difference in average returns while volatilities are almost constant across recession and non-recession periods. The difference in Sharpe ratios before transaction costs between MV and MV_{TC} are close to zero in recession and non-recession periods. Transaction costs in and out of recessions are identical. MV_{TC} saves on average 2.4% in costs compared to MV. The difference in costs between MV and MV_{TC} is highly statistically significant. MV_{TC} outperforms MV and the difference in Sharpe ratios is 0.22 in recessions and 0.19 during non-recession periods. This difference in Sharpe ratios after transaction costs is only significant in non-recession periods (p-value of 0.006), while the p-value during recessions is 0.366. The p-value in recession periods is relatively large because we only have 56 monthly observations during recessions and the power of the test is low. However, the economic magnitude of our result is identical in and out of recessions. We conclude, that our findings in Section 3.4 are present both in and out of recession periods.

3.5.4 Subsamples before and after the Introduction of the Euro

The introduction of the Euro non-trivially affected the investment opportunity set in FX markets. Our results from Section 3.4 are present in the subsamples before and after the introduction of the Euro. The results are stronger in the earlier subsample, which is mostly due to the general decline in average transaction costs over time (Figure 3.3).

Table 3.6 summarizes the monthly excess returns of MV and MV_{TC} for our full set of 29 currencies before (columns 1-2) and after (columns 3-4) the introduction of the Euro on January 2nd, 1999. In both samples, there is no difference in Sharpe ratios before transaction

costs between MV and MV_{TC}. Sharpe ratios are slightly larger in the sample after the introduction of the Euro. Transaction costs are substantially larger in the pre-Euro sample. Costs incurred by MV and MV_{TC} are 5.58% and 1.19% per year (a difference of 3.67%) in the pre-Euro, and 1.18% and 0.44% per year (a difference of 0.74%) in the post-Euro sample. The difference in costs between MV and MV_{TC} is highly statistically significant in both pre- and post-Euro samples. The Sharpe ratios after transaction costs of MV and MV_{TC} are 0.69 and 0.91 in the pre-Euro and 0.95 and 1.09 in the post-Euro sample. The difference of 0.22 in the pre-Euro sample is economically and statistically significant with a p-value of 0.004. The difference of 0.14 in the post-Euro sample is economically large but not statistically significant (p-value of 0.307). The decline in the difference in Sharpe ratios after costs between MV_{TC} and MV from the pre- to the post-Euro sample is mostly due to the strong decline in average transaction costs. However, this does not mean that optimizing over transaction costs is useless in the post-Euro era. The superior performance of MV_{TC} over MV is steady over the entire period. The cumulative returns after costs of MV_{TC} are always above those of MV and the spread is monotonically increasing. To conclude, the main results of Section 3.4 are confirmed in both subsamples before and after the introduction of the Euro.

3.6 Conclusion

Using foreign exchange (FX) market returns for 29 developed and emerging currencies from 1976 to 2016, we show that taking transaction costs into account in a mean-variance portfolio optimization leads to an economically large and statistically significant improvement in the out-of-sample performance.

We present four main findings. First, we document that the out-of-sample Sharpe ratios before transaction costs of MV (which is the mean-variance efficient portfolio which ignores transaction costs in the optimization) and MV_{TC} (which takes costs into account in the optimization) are identical and equal to 1. Second, MV_{TC} pays 1.28% of the portfolio value per year in transaction costs which is substantially lower than the 3.71% paid by MV. Third, MV_{TC} has an out-of-sample Sharpe ratio after transaction costs of 0.9, while the Sharpe ratio of MV is only 0.7. Other moments of the return distribution are similar across the two strategies. Thus, taking costs into account in the optimization significantly improves the out-of-sample performance after transaction costs. Fourth, transaction costs paid by $MV_{TC\setminus Corr}$ (the strategy which accounts for transaction costs in the optimization but for simplicity assumes that assets are uncorrelated when constructing the no trading region) are 2.56% per year and its Sharpe ratio after costs is only 0.76, which is significantly inferior to its counter-part in MV_{TC}. Thus, accounting for correlations between assets is important for the superior performance of MV_{TC}. In contrast, the approximate solution $MV_{TC\setminus Corr}$ has no benefit. This result invalidates the insights coming from the continuous-time model of H. Liu (2004), the only setup so far able, by assuming independence across assets, to deliver solutions to large scale dynamic portfolio optimizations in the presence of transaction costs.

Our results have important practical implications. First, accounting for costs when optimizing a portfolio is beneficial and improves the out-of-sample performance. Second, transaction costs are declining over time, and thus, traders who specialize in developed currencies may be tempted to ignore transaction costs when constructing mean-variance efficient portfolios. However, even if transaction costs are low during normal times, they substantially increase during crises and become relevant (Karnaukh et al. (2015)). Third, many currency traders have shifted their focus to emerging and frontier markets because exchange rate forward discounts among developed currencies are close to zero for the past decade. Transaction

costs in emerging and frontier markets are generally larger than the costs considered in our analysis, and thus, the implications of transaction costs on the optimal portfolio choice are even more important for these traders.

Figure 3.1: Importance of Transaction Costs in FX Markets

Annualized out-of-sample Sharpe ratios before (blue bars to the left) and after (green bars in the middle) transaction costs of various currency trading strategies and transaction costs (yellow bar to the right) paid by them. MV is the mean-variance optimized portfolio without taking into account transaction costs in the optimization. $MV_{TC\backslash Corr}$ is the mean-variance optimized portfolio which optimizes over transaction costs but makes the simplifying assumption that assets are uncorrelated. MV_{TC} is the mean-variance optimized portfolio which optimizes over transaction costs. DOL invests equally in all bilateral carry trades. $DDOL$ takes a long position in DOL if the median exchange rate forward discount is positive, and a short position otherwise. HML sorts bilateral carry trades according to the forward discount into quintiles and shorts the bottom and invests in the top quintile. MOM sorts bilateral carry trades according to their past 12 month performance into quintiles and shorts the bottom and invests in the top quintile. VAL sorts bilateral carry trades according to the power purchase parity adjusted exchange rate into quintiles and shorts the top quintile (overvalued currencies with high real exchange rates) and invests in the bottom quintile (undervalued currencies with low real exchange rates). The data are monthly returns for our full set of 29 currencies from January 1976 to February 2016.

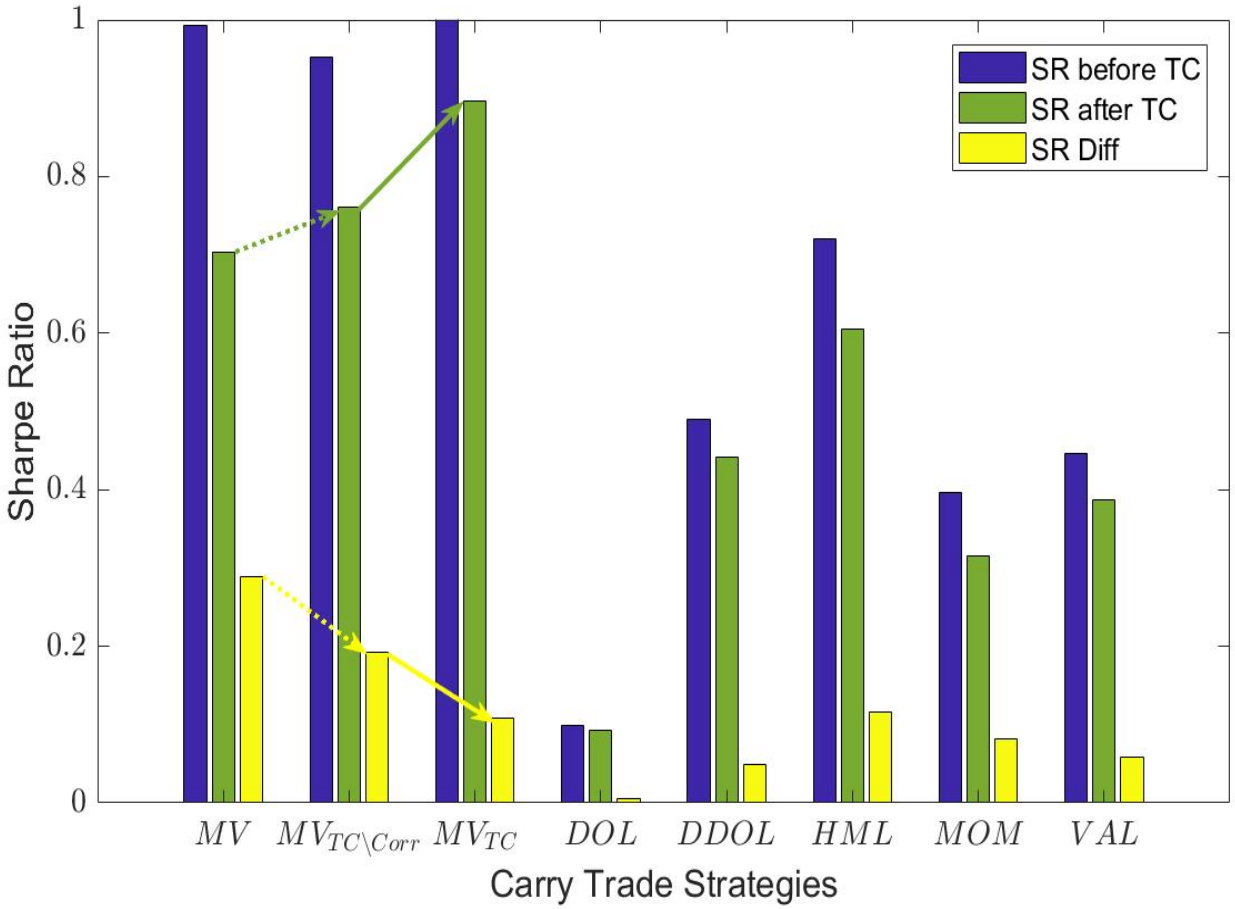


Figure 3.2: Mean-Variance Problem with TC: Case of 2 Risky Assets

The investment opportunity set consists of two risky assets which are positively correlated. The horizontal axis measures the weight a portfolio places on asset 1, and the vertical axis the weight on asset 2. The point labeled $\theta^{\mathbf{MV}}$ is the optimal portfolio if there were no transaction costs. The blue parallelogram illustrates the no trading region of MV_{TC}, which optimizes over transaction costs. The red checkered square (within the blue parallelogram) determines the no trading region of $MV_{TC\setminus Corr}$, which optimizes over transaction costs but assumes that the two assets are uncorrelated. If the initial position is within the no trading region, then the investor does not trade. If it is outside, then the investor trades along vertical and horizontal lines (indicated by black arrows) towards $\theta^{\mathbf{MV}}$ until to the boundary of the no trading region. $\Delta_i^P > 0$ respecively $\Delta_i^S > 0$ indicate the regions where the investor increases respectively decreases her position in asset i $\forall i \in \{1, 2\}$.

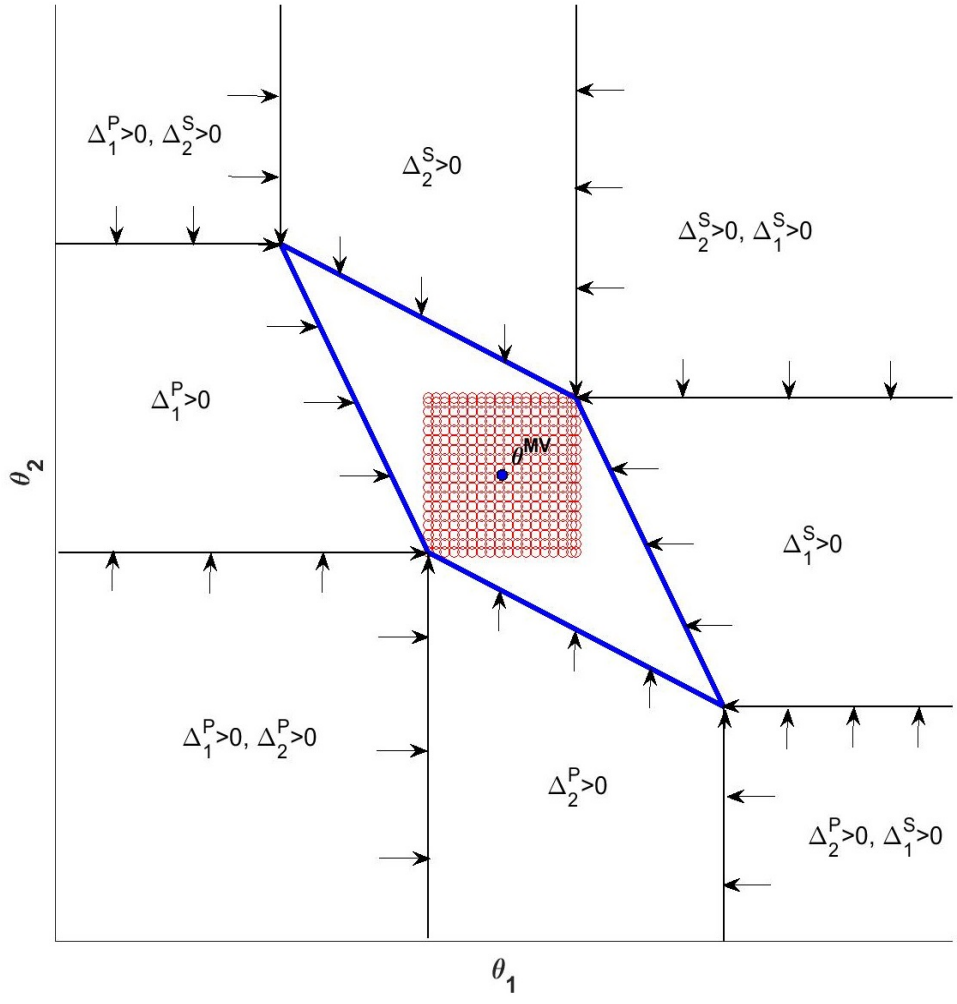


Figure 3.3: Average Annualized Transaction Costs

Average (across currencies) annualized costs (in percentage points) to change a position in a bilateral carry trade for the the full set of 29 currencies (green solid line), the subset of 15 developed currencies (black dotted line), and the subset of 14 emerging currencies (red dashed line) from January 1976 to February 2016. Grey shaded areas indicate NBER recessions.

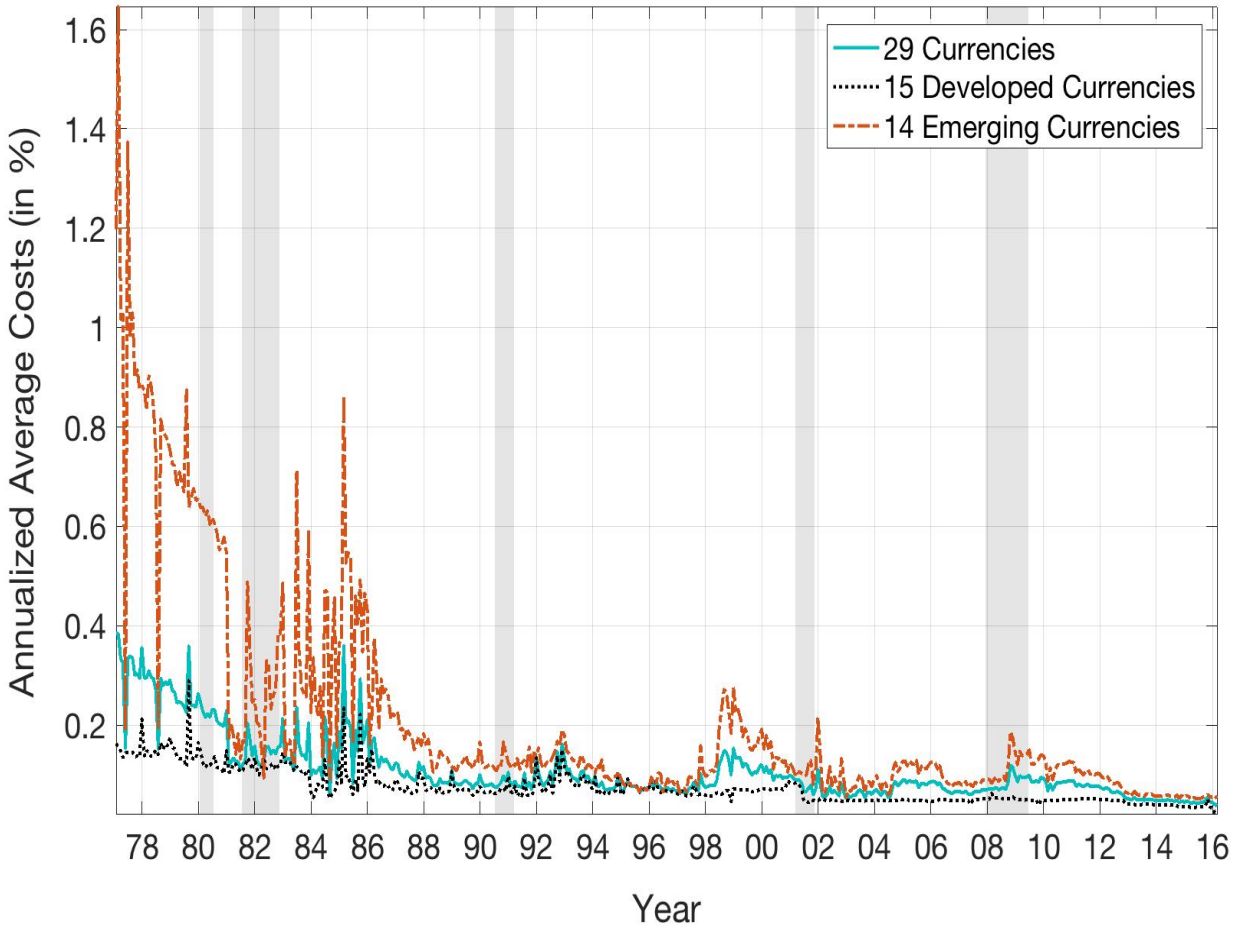
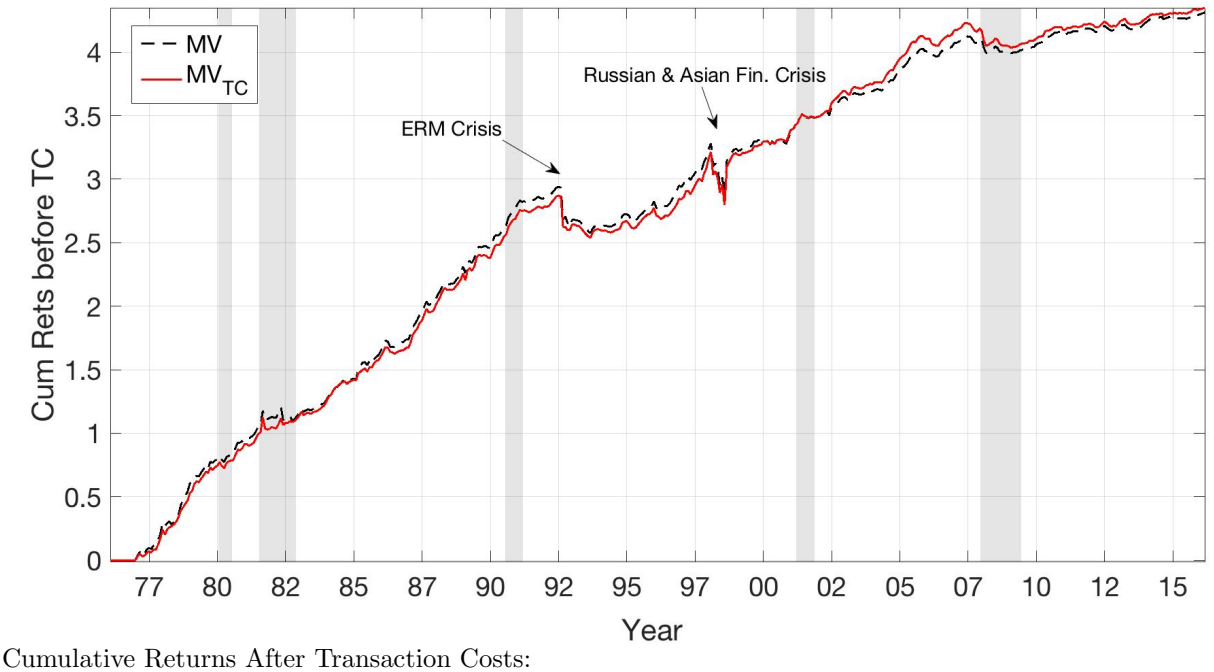


Figure 3.4: Cumulative Returns of MV and MV_{TC}

Time series of cumulative returns of MV (black dashed line) and MV_{TC} (red solid line) for our set of 29 currencies from January 1976 to February 2016. Returns before transaction costs are shown in the top panel, and returns after transaction costs in the bottom panel. Grey shaded areas indicate NBER recessions.

Cumulative Returns Before Transaction Costs:



Cumulative Returns After Transaction Costs:

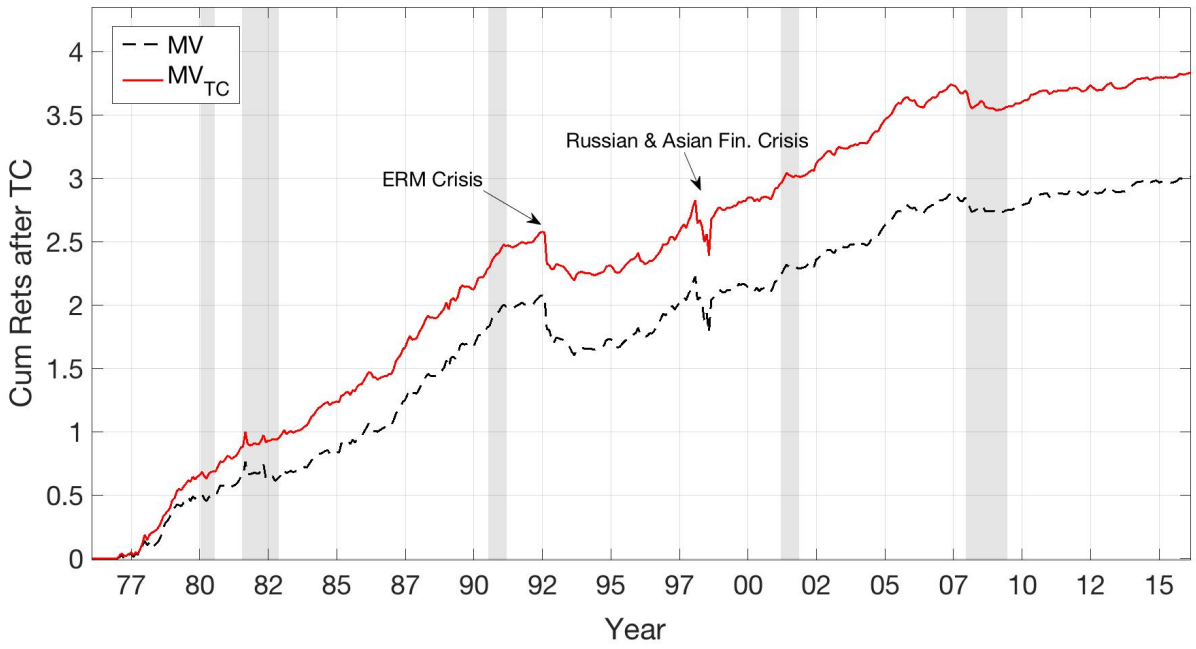
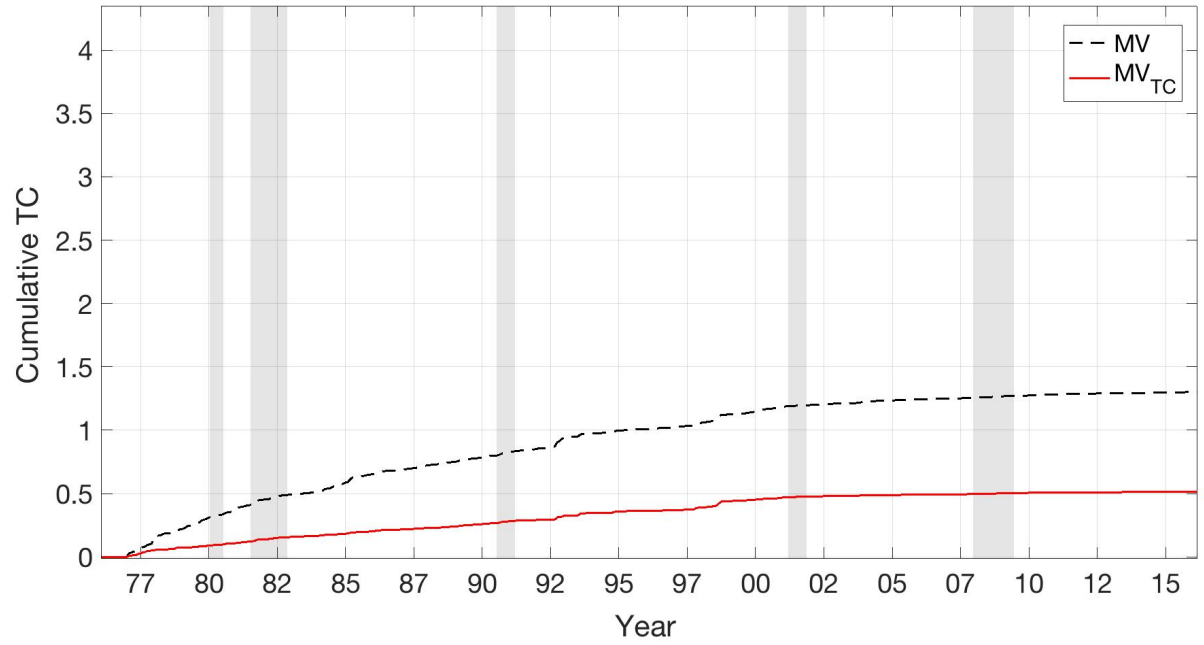


Figure 3.5: Transaction Costs of MV and MV_{TC}

Time series of transaction costs of MV (black dashed line) and MV_{TC} (red solid line) for our set of 29 currencies from January 1976 to February 2016. Cumulative costs are shown in the top panel, and monthly costs in the bottom panel. Grey shaded areas indicate NBER recessions.

Cumulative Transaction Costs:



Monthly Transaction Costs:

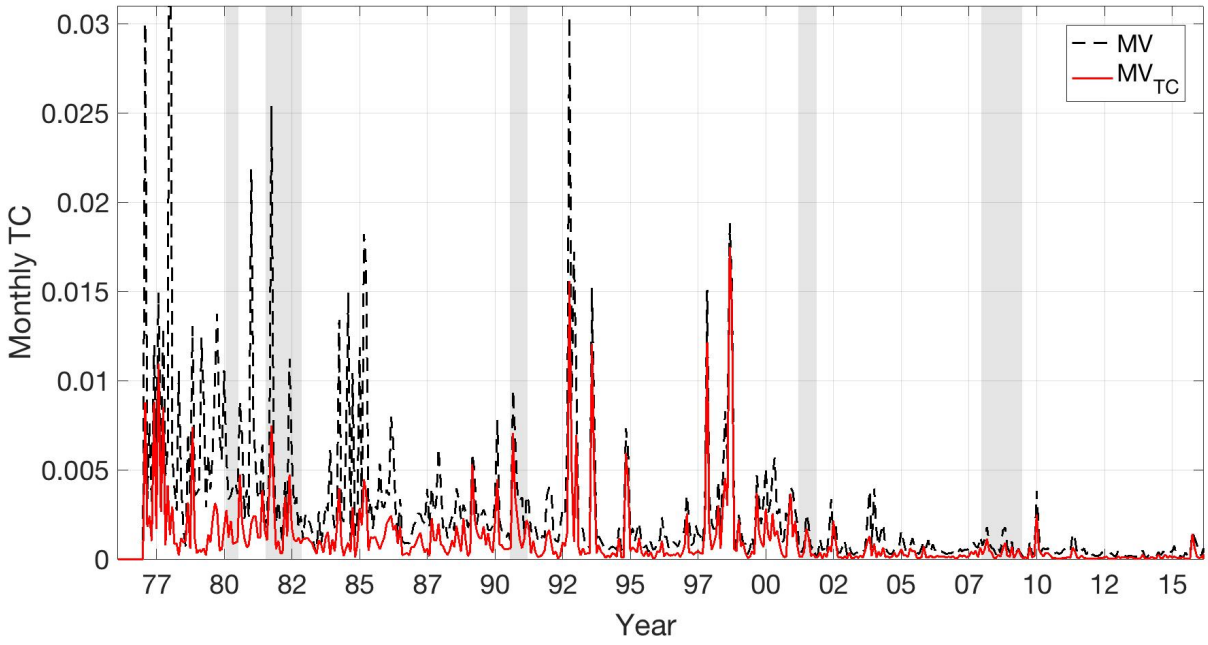
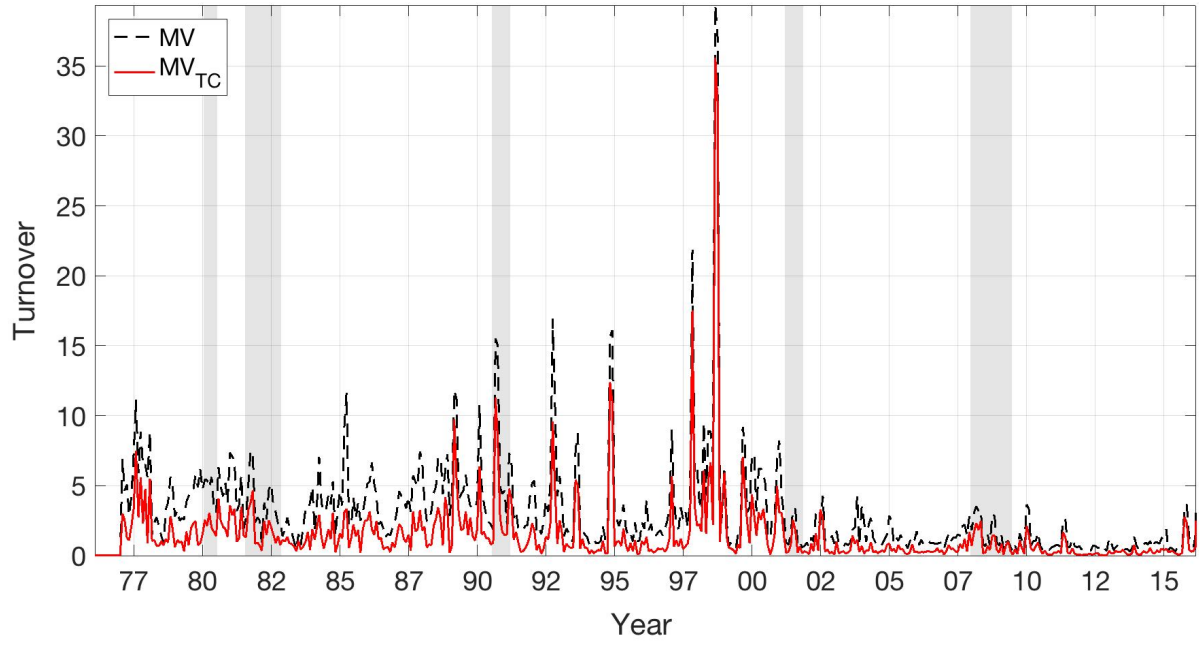


Figure 3.6: Trading Activity of MV and MV_{TC}

Top panel: Time series of the turnover $\sum_i \|\theta_{i,t} - \theta_{i,t-1}\|$ of MV (black dashed line) and MV_{TC} (red solid line) for our set of 29 currencies from January 1976 to February 2016. Grey shaded areas indicate NBER recessions. Bottom panel: Average portfolio weights and 1-standard deviation error bars of MV (downward pointing triangle and thin black line) and MV_{TC} (upward pointing triangle and thick red line).

Turnover:



Average Portfolio Weights and Standard Deviation Bars:

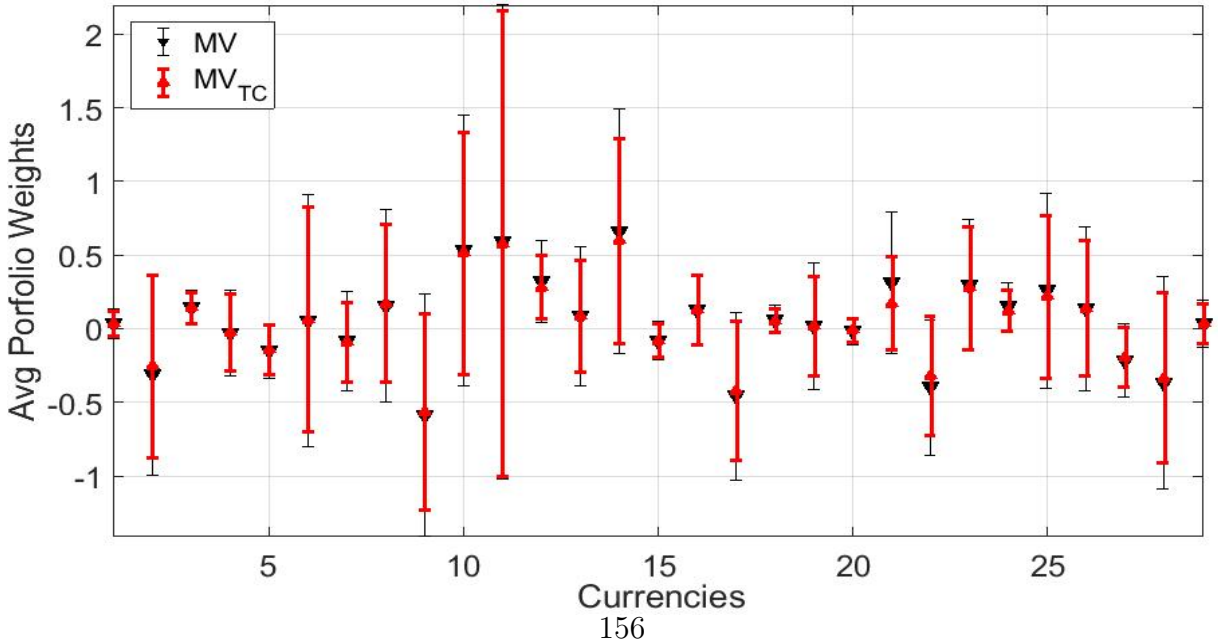


Figure 3.7: Notional Value of MV and MV_{TC}

Time series of the notional value or total dollar exposure $\sum_i \|\theta_{i,t}\|$ of MV (black dashed line) and MV_{TC} (red solid line) for our set of 29 currencies from January 1976 to February 2016. Grey shaded areas indicate NBER recessions.

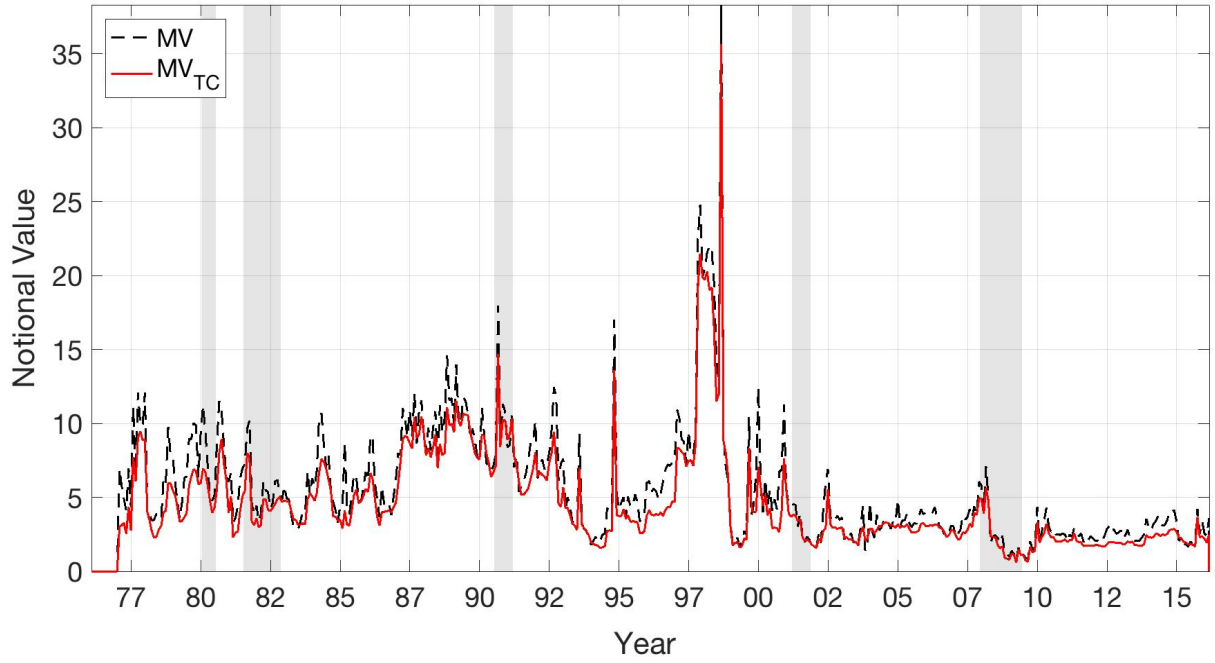


Figure 3.8: Average Correlations

Time-series of the average conditional correlation of each exchange rate growth i with all other exchange rate growths for our full set of $N = 29$ currencies, $\rho_{i,t} = \frac{1}{N-1}\sum_{j=1}^{N-1} Corr_t\left(\Delta x_{i,t}, \Delta x_{j,t}\right)$ estimated using daily data within each month from January 1976 to February 2016. The bold black line captures the time-series of the cross-sectional average across all correlations, $\rho_t = \frac{1}{N-1}\sum_{i=1}^{N} \rho_{i,t}$. Grey shaded areas indicate NBER recessions.

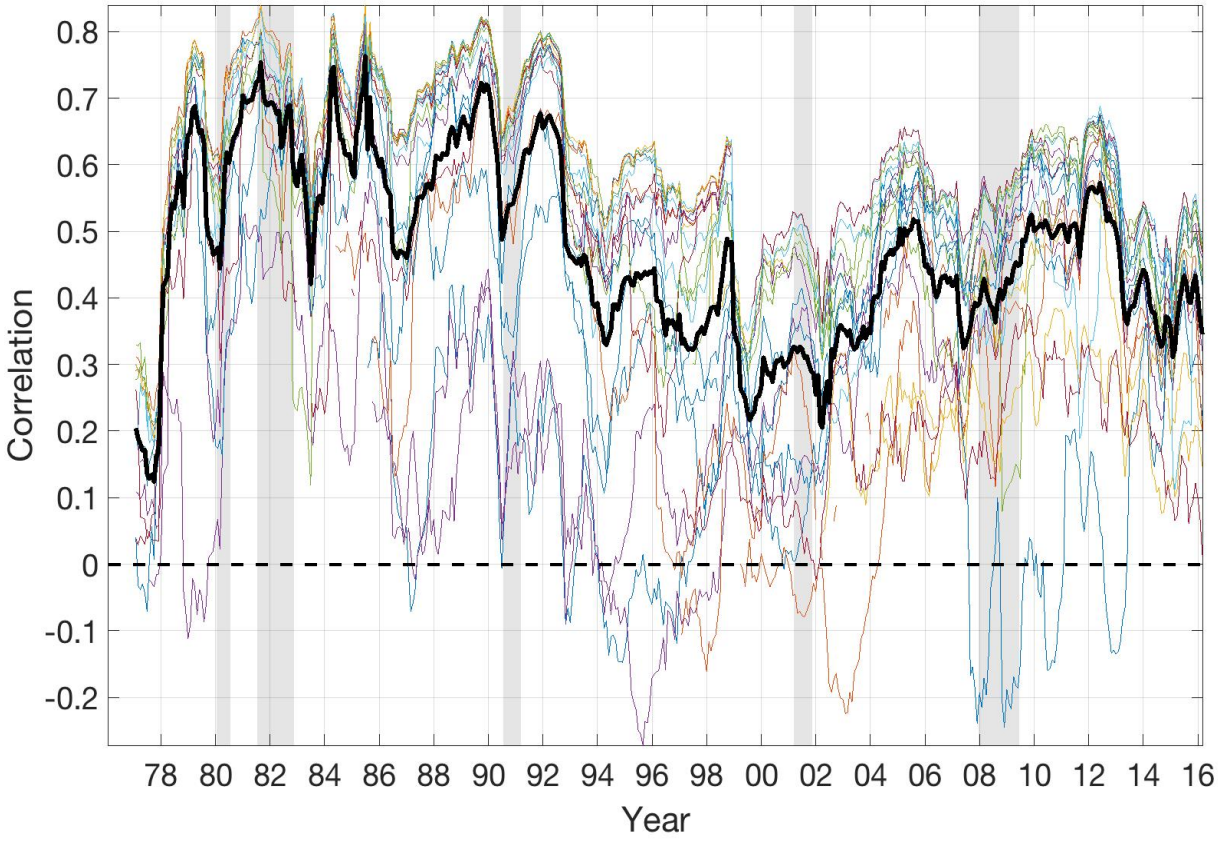


Figure 3.9: Difference in Trade Aggressiveness (ΔTA)

Time-series of the difference in trade aggressiveness between $MV_{TC\setminus Corr}$ and MV_{TC}, $\Delta\text{TA} = \text{TA}\left(\theta^{\mathbf{MV_{TC\setminus Corr}}} - \theta^{\mathbf{MV_{TC}}}\right)$ (solid line), where the trade aggressiveness is defined as $TA(\theta^{\mathbf{S}}_{\mathbf{t}}) = \frac{\sum_i \|\theta^{\mathbf{S}}_{\mathbf{i,t}} - \theta^{\mathbf{S}}_{\mathbf{i,t-1}}\|}{\sum_i \|\theta^{\mathbf{MV}}_{\mathbf{i,t}} - \theta^{\mathbf{MV}}_{\mathbf{i,t-1}}\|}$. The horizontal black dashed line is the sample median of ΔTA. The gray shaded areas indicate NBER recessions. The data refers to our full set of 29 currencies from January 1976 to February 2016. Reported values are in percentage points.

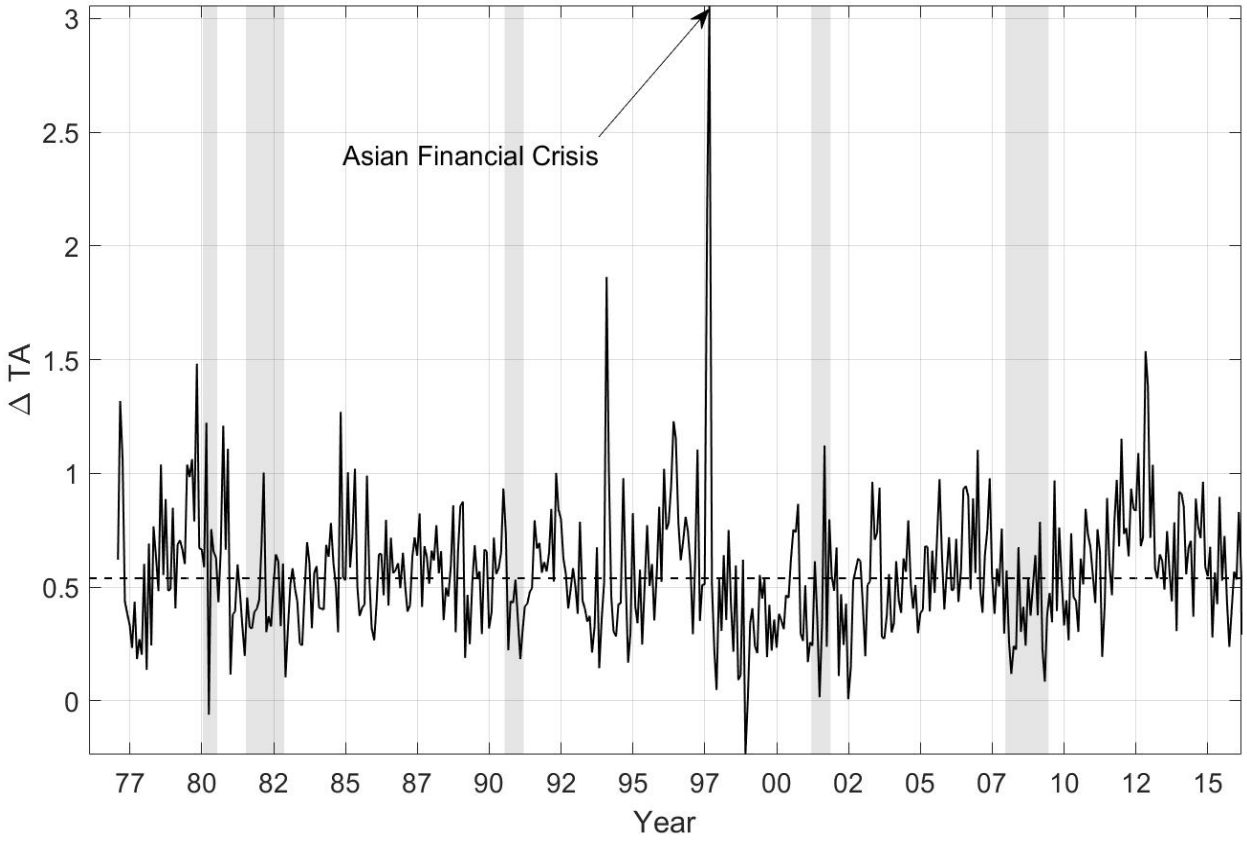


Figure 3.10: Sharpe Ratios of Approximate Solutions in Multi-Period Model

Annualized out-of-sample Sharpe ratios of $\theta^{\mathbf{MV_{TC}^{M}}}(c,a)$ for $(c,a) \in [0,2]\times[0,1]$ for our full set of 29 currencies from 1976 to 2016 (see details in Section 3.4.6). The point highlighted by a blue arrow indicates MV_{TC} with a Sharpe ratio of 0.90. The point highlighted by a red arrow indicates $\theta^{\mathbf{MV_{TC}^{M}}}(c=0.7,a=0.8)$ with a Sharpe ratio of 0.92, which is the highest Sharpe ratio for any combination of (c,a). The point highlighted by a black arrow indicates $\theta^{\mathbf{MV_{TC}^{M}}}(c=1.3,a=0)$ with a Sharpe ratio of 0.91, which is the highest Sharpe ratio for any value c and $a=0$. Points indicated by blue crosses are portfolios $\theta^{\mathbf{MV_{TC}^{M}}}(c,a)$ with Sharpe ratios which are statistically significantly different from the Sharpe ratio of $\theta^{\mathbf{MV_{TC}}}$ (using the test of Ledoit and Wolf (2008) and a p-value of 0.05).

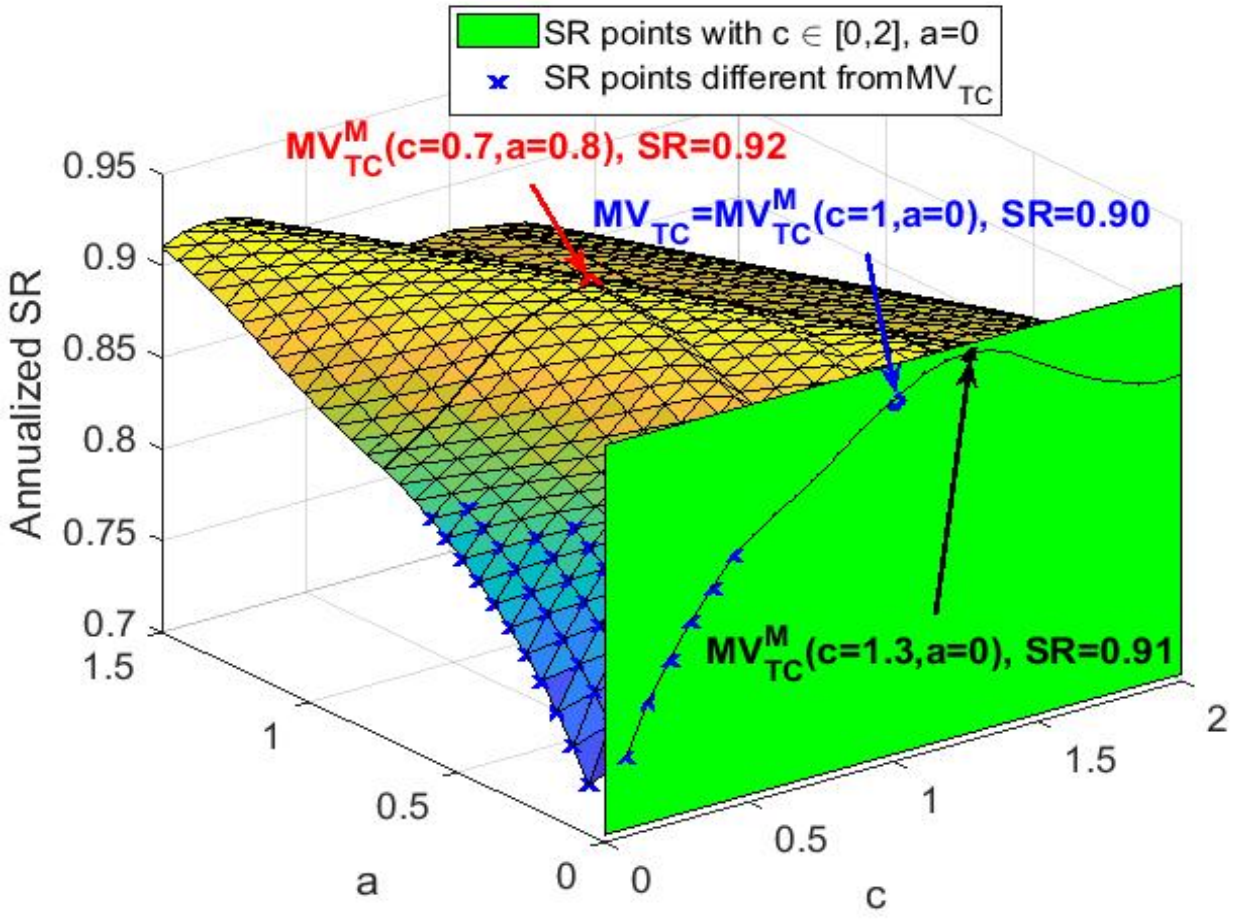


Table 3.1: Mean-Variance Strategies: MV vs. MV_{TC}

Summary statistics of monthly excess returns of MV and MV_{TC}, described in Section 3.2. First two columns report results for all 29 currencies, last two columns for 15 developed currencies. The sample period is 1976-2016. SR is the annualized Sharpe ratio, Mean the annualized average return (in percentage points), Mean Costs the average annualized transaction costs measured in percentage of the portfolio value, Vol the annualized standard deviation (in percentage points), Skew the skewness, Kurt the kurtosis, % Positive the percentage of positive monthly returns, MDD the Maximum Draw Down, AC the autocorrelation, CE_λ the annualized rate of return (Certainty Equivalent) an investor with mean-variance preferences and risk aversion λ is willing to give up in order to switch from strategy MV to strategy MV_{TC}. % of $CE_\lambda > 0$ indicates the percentage of months with positive CE_λ. ΔMean, ΔMean Costs, ΔSR are the differences in the Mean, Mean Costs, SR between MV_{TC} and MV. Standard errors of ΔSR are estimated using block bootstrapping with block sizes of 10 observations to account for heteroskedasticity, cross- and auto-correlation (Ledoit and Wolf (2008)). Standard errors of ΔMean Costs are estimated using Newey and West (1987) to account for heteroskedasticity and auto-correlation. ***, **, * indicate a statistical significance at the 1%, 5%, 10% level of ΔSR and ΔMean Costs. We only report the p-value for ΔSR after costs.

	All 29 Currencies		15 Developed Currencies	
	MV	MV_{TC}	MV	MV_{TC}
Before Transaction Costs:				
SR	0.99	1.00	0.87	0.82
Mean	12.26	10.81	9.09	7.49
ΔMean	-	-1.45	-	-1.60
Transaction Costs:				
Mean Costs	3.71	1.28	1.97	0.80
ΔMean Costs	-	-2.42***	-	-1.17***
After Transaction Costs:				
SR	**0.70**	**0.90**	**0.70**	**0.75**
ΔSR	**-**	**0.19*****	**-**	**0.05**
(p-value)	-	**(0.007)**	**-**	**(0.385)**
Mean	8.55	9.53	7.13	6.69
Vol	12.15	10.62	10.23	8.93
Skew	-1.20	-0.65	2.96	4.22
Kurt	28.58	30.73	82.53	100.15
% Positive	66.38	67.87	66.60	67.02
MDD	-54.10	-43.47	-43.57	-29.43
AC	-0.07	-0.05	-0.03	-0.03
$CE_{\lambda=1}$ (% of $CE_{\lambda=1} > 0$)	-	1.16 (75%)	-	-0.31 (50%)
$CE_{\lambda=5}$ (% of $CE_{\lambda=5} > 0$)	-	1.89 (80%)	-	0.22 (63%)
$CE_{\lambda=10}$ (% of $CE_{\lambda=10} > 0$)	-	2.80 (82%)	-	0.87 (66%)
$CE_{\lambda=50}$ (% of $CE_{\lambda=50} > 0$)	-	10.10 (83%)	-	6.11 (72%)

Table 3.2: Mean-Variance Strategies: Importance of Correlations

Summary statistics of monthly excess returns of MV, MV_{TC} and $MV_{TC\setminus Corr}$ (described in Section 3.2) for all 29 currencies from 1976 to 2016. SR is the annualized Sharpe ratio, Mean the annualized average return (in percentage points), Mean Costs the average annualized transaction costs measured in percentage of the portfolio value, Vol the annualized standard deviation (in percentage points), Skew the skewness, Kurt the kurtosis, % Positive the percentage of positive monthly returns, MDD the Maximum Draw Down, AC the autocorrelation. CE_λ is the annualized rate of return (Certainty Equivalent) an investor with mean-variance preferences and risk aversion λ is willing to give up in order to switch from strategy $MV_{TC\setminus Corr}$ to strategy MV or MV_{TC}. % of $CE_\lambda > 0$ indicates the percentage of months with positive CE_λ. ΔMean, ΔMean Costs, ΔSR are the differences in the Mean, Mean Costs, SR between MV_{TC} and MV. Standard errors of ΔSR are estimated using block bootstrapping with block sizes of 10 observations to account for heteroskedasticity, cross- and auto-correlation (Ledoit and Wolf (2008)). Standard errors of ΔMean Costs are estimated using Newey and West (1987) to account for heteroskedasticity and auto-correlation. ***, **, * indicate a statistical significance at the 1%, 5%, 10% level of ΔSR and ΔMean Costs. We only report the p-value for ΔSR after costs.

	MV	$MV_{TC\setminus Corr}$	MV_{TC}
Before Transaction Costs:			
SR	0.99	0.95	1.00
Mean	12.26	12.05	10.81
ΔMean	0.21	-	-1.24
Transaction Costs:			
Mean Costs	3.71	2.56	1.28
ΔMean Costs	1.14***	-	-1.28***
After Transaction Costs:			
SR	**0.70**	**0.76**	**0.90**
ΔSR	**-0.06**	**-**	**0.14***
(p-value)	**(0.428)**	**-**	**(0.084)**
Mean	8.55	9.49	9.53
Vol	12.15	12.50	10.62
Skew	-1.20	-0.10	-0.65
Kurt	28.58	28.06	30.73
% Positive	66.38	66.81	67.87
MDD	-54.10	-50.00	-43.47
AC	-0.07	-0.06	-0.05
$CE_{\lambda=1}$ (% of $CE_{\lambda=1} > 0$)	-0.90 (29%)	-	0.26 (57%)
$CE_{\lambda=5}$ (% of $CE_{\lambda=5} > 0$)	-0.73 (32%)	-	1.15 (68%)
$CE_{\lambda=10}$ (% of $CE_{\lambda=10} > 0$)	-0.53 (35%)	-	2.27 (72%)
$CE_{\lambda=50}$ (% of $CE_{\lambda=50} > 0$)	1.09 (44%)	-	11.19 (78%)

Table 3.3: Trade Aggressiveness (TA)

Summary statistics of monthly trade aggressiveness $TA(\theta^{\mathbf{S}}_{\mathbf{t}}) = \frac{\sum_i \|\theta^{\mathbf{S}}_{\mathbf{i,t}} - \theta^{\mathbf{S}}_{\mathbf{i,t-1}}\|}{\sum_i \|\theta^{\mathbf{MV}}_{\mathbf{i,t}} - \theta^{\mathbf{MV}}_{\mathbf{i,t-1}}\|}$ for strategies $S \in \{MVTC, MV_{TC\setminus Corr}\}$. Strategies $MV, MVTC$ and $MV_{TC\setminus Corr}$ are described in Section 3.2. Mean TA reports the time-series average of monthly $TA(\theta^{\mathbf{S}}_{\mathbf{t}})$. In parenthesis below we report p-values, which are calculated using standard errors robust to heteroskedasticity and auto-correlation (Newey and West (1987)). Median TA reports the median of monthly $TA(\theta^{\mathbf{S}}_{\mathbf{t}})$. 5-%ile and 95-%ile report the 5 and 95 percentiles of the monthly $TA(\theta^{\mathbf{S}}_{\mathbf{t}})$ distribution. The data is our full set of 29 currencies from January 1976 to February 2016. ***, **, * indicate a statistical significance at the 1%, 5%, 10% level.

	MV_{TC}	$MV_{TC\setminus Corr}$	ΔTA
Mean TA	0.41***	0.98***	0.57***
(p-value)	(0.000)	(0.000)	(0.000)
95-%ile	0.74	1.39	1.03
Median TA	0.40	0.96	0.54
5-%ile	0.14	0.69	0.20

Table 3.4: Mean-Variance Strategies: 1976-2016 without Crises & 1983-2016

Summary statistics of monthly excess returns of MV and MV_{TC} for all 29 currencies. First two columns report results for the sample 1976-2016 excluding the 1992 ERM crisis and the 1997 Asian financial and 1998 Russian crisis, last two columns report results for 1983-2016. SR is the annualized Sharpe ratio, Mean the annualized average return (in percentage points), Mean Costs the average annualized transaction costs measured in percentage of the portfolio value, Vol the annualized standard deviation (in percentage points), Skew the skewness, Kurt the kurtosis, % Positive the percentage of positive monthly returns, MDD the Maximum Draw Down, AC the autocorrelation, CE_λ the annualized Certainty Equivalent an investor with mean-variance preferences and risk aversion λ is willing to give up in order to switch from strategy MV to strategy MV_{TC}. % of $CE_\lambda > 0$ indicates the percentage of months with positive CE_λ. ΔMean, ΔMean Costs, ΔSR are the differences in the Mean, Mean Costs, SR between MV_{TC} and MV. Standard errors of ΔSR are estimated using block bootstrapping with block sizes of 10 observations to account for heteroskedasticity, cross- and auto-correlation (Ledoit and Wolf (2008)). Standard errors of ΔMean Costs are estimated using Newey and West (1987) to account for heteroskedasticity and auto-correlation. ***, **, * indicate a statistical significance at the 1%, 5%, 10% level of ΔSR and ΔMean Costs.

	1976-2016 without Crises		1983-2016	
	MV	MV_{TC}	MV	MV_{TC}
Before Transaction Costs:				
SR	1.15	1.17	0.87	0.88
Mean	11.96	10.50	10.91	9.73
ΔMean	-	-1.47	-	-1.18
Transaction Costs:				
Mean Costs	3.63	1.22	2.80	1.08
ΔMean Costs	-	-2.41***	-	-1.72***
After Transaction Costs:				
SR	**0.81**	**1.04**	**0.66**	**0.79**
ΔSR	**-**	**0.23*** **	**-**	**0.14** **
(p-value)	-	**(0.005)**	**-**	**(0.025)**
Mean	8.33	9.28	8.11	8.65
Vol	10.31	8.92	12.36	10.94
Skew	-1.20	-1.86	-1.76	-0.69
Kurt	14.81	15.21	31.75	32.96
Positive	66.17	67.87	65.81	66.58
MDD	-51.42	-43.47	-54.10	-43.47
AC	0.01	0.05	-0.06	-0.04
$CE_{\lambda=1}$ (% of $CE_{\lambda=1} > 0$)	-	1.09 (75%)	-	0.71 (72%)
$CE_{\lambda=5}$ (% of $CE_{\lambda=5} > 0$)	-	1.66 (80%)	-	1.39 (77%)
$CE_{\lambda=10}$ (% of $CE_{\lambda=10} > 0$)	-	2.37 (82%)	-	2.23 (80%)
$CE_{\lambda=50}$ (% of $CE_{\lambda=50} > 0$)	-	8.09 (83%)	-	9.00 (81%)

Table 3.5: Mean-Variance Strategies: NBER Recessions vs. Non-Recessions

Summary statistics of monthly excess returns of MV and MV_{TC} for all 29 currencies for the sample 1976-2016. First two columns report results for NBER recession periods, last two columns report results for non-recession periods. SR is the annualized Sharpe ratio, Mean the annualized average return (in percentage points), Mean Costs the average annualized transaction costs measured in percentage of the portfolio value, Vol the annualized standard deviation (in percentage points), Skew the skewness, Kurt the kurtosis, % Positive the percentage of positive monthly returns, MDD the Maximum Draw Down, AC the autocorrelation, CE_λ the annualized rate of return (Certainty Equivalent) an investor with mean-variance preferences and risk aversion λ is willing to give up in order to switch from strategy MV to strategy MV_{TC}. % of $CE_\lambda > 0$ indicates the percentage of months with positive CE_λ. ΔMean, ΔMean Costs, ΔSR are the differences in the Mean, Mean Costs, SR between MV_{TC} and MV. Standard errors of ΔSR are estimated using block bootstrapping with block sizes of 10 observations to account for heteroskedasticity, cross- and auto-correlation (Ledoit and Wolf (2008)). Standard errors of ΔMean Costs are estimated using Newey and West (1987) to account for heteroskedasticity and auto-correlation. ***, **, * indicate a statistical significance at the 1%, 5%, 10% level of ΔSR and ΔMean Costs. We only report the p-value for ΔSR after costs.

	NBER Recessions		Non-Recessions	
	MV	MV_{TC}	MV	MV_{TC}
Before Transaction Costs:				
SR	0.47	0.51	1.07	1.07
Mean	5.83	5.54	13.10	11.50
ΔMean	-	-0.29	-	-1.60
Transaction Costs:				
Mean Costs	3.96	1.60	3.67	1.24
ΔMean Costs	-	-2.36***	-	-2.43***
After Transaction Costs:				
SR	**0.14**	**0.37**	**0.78**	**0.97**
ΔSR	**-**	**0.22**	**-**	**0.19*****
(p-value)	-	**(0.366)**	**-**	**(0.006)**
Mean	1.87	3.94	9.43	10.26
Vol	12.98	10.81	12.03	10.59
Skew	-0.55	-0.14	-1.30	-0.72
Kurt	6.71	6.37	32.47	34.28
Positive	55.36	55.36	67.87	69.57
MDD	-17.29	-17.38	-54.10	-43.47
AC	-0.15	-0.06	-0.06	-0.06
$CE_{\lambda=1}$ (% of $CE_{\lambda=1} > 0$)	-	2.28 (70%)	-	1.01 (76%)
$CE_{\lambda=5}$ (% of $CE_{\lambda=5} > 0$)	-	3.10 (75%)	-	1.73 (80%)
$CE_{\lambda=10}$ (% of $CE_{\lambda=10} > 0$)	-	4.13 (77%)	-	2.62 (83%)
$CE_{\lambda=50}$ (% of $CE_{\lambda=50} > 0$)	-	12.37 (82%)	-	9.79 (83%)

Table 3.6: Mean-Variance Strategies: Pre- vs. Post-Euro

Summary statistics of monthly excess returns of MV and MV_{TC} for all 29 currencies. First two columns report results for pre-Euro period (1976-1999), last two columns report results for post-Euro period (1999-2016). SR is the annualized Sharpe ratio, Mean the annualized average return (in percentage points), Mean Costs the average annualized transaction costs measured in percentage of the portfolio value, Vol the annualized standard deviation (in percentage points), Skew the skewness, Kurt the kurtosis, % Positive the percentage of positive monthly returns, MDD the Maximum Draw Down, AC the autocorrelation, CE_λ the annualized rate of return (Certainty Equivalent) an investor with mean-variance preferences and risk aversion λ is willing to give up in order to switch from strategy MV to strategy MV_{TC}. % of $CE_\lambda > 0$ indicates the percentage of months with positive CE_λ. ΔMean, ΔMean Costs, ΔSR are the differences in the Mean, Mean Costs, SR between MV_{TC} and MV. Standard errors of ΔSR are estimated using block bootstrapping with block sizes of 10 observations to account for heteroskedasticity, cross- and auto-correlation (Ledoit and Wolf (2008)). Standard errors of ΔMean Costs are estimated using (Newey & West, 1987) to account for heteroskedasticity and auto-correlation. ***, **, * indicate a statistical significance at the 1%, 5%, 10% level of ΔSR and ΔMean Costs. We only report the p-value for ΔSR after costs.

	Pre-Euro		Post-Euro	
	MV	MV_{TC}	MV	MV_{TC}
Before Transaction Costs:				
SR	1.05	1.05	1.14	1.16
Mean	15.98	13.82	7.25	6.75
ΔMean	-	-2.16	-	-0.50
Transaction Costs:				
Mean Costs	5.58	1.91	1.18	0.44
ΔMean Costs	-	-3.67***	-	-0.74***
After Transaction Costs:				
SR	**0.69**	**0.91**	**0.95**	**1.09**
ΔSR	**-**	**0.22*****	**-**	**0.14**
(p-value)	-	**(0.004)**	**-**	**(0.307)**
Mean	10.40	11.91	6.07	6.32
Vol	15.06	13.07	6.37	5.80
Skew	-1.17	-0.70	-0.28	-0.55
Kurt	21.11	23.27	5.09	6.72
Positive	67.17	70.94	65.37	63.90
MDD	-54.10	-43.47	-17.76	-20.75
AC	-0.11	-0.11	0.21	0.31
$CE_{\lambda=1}$ (% of $CE_{\lambda=1} > 0$)	-	1.81 (80%)	-	0.29 (69%)
$CE_{\lambda=5}$ (% of $CE_{\lambda=5} > 0$)	-	2.97 (85%)	-	0.41 (72%)
$CE_{\lambda=10}$ (% of $CE_{\lambda=10} > 0$)	-	4.41 (88%)	-	0.58 (75%)
$CE_{\lambda=50}$ (% of $CE_{\lambda=50} > 0$)	-	16.00 (87%)	-	1.89 (78%)

References

Ackermann, F., Pohl, W., & Schmedders, K. (2016). Optimal and Naive Diversification in Currency Markets. *Management Science*.

Adrian, T., & Boyarchenko, N. (2012). Intermediary leverage cycles and financial stability. *Federal Reserve Bank of New York Staff Report 567*.

Adrian, T., Etula, E., & Muir, T. (2014). Can time-varying risk of rare disasters explain aggregate stock market volatility? *The Journal of Finance*, *69*(6), 2557–2596.

Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets*, *5*(1), 31–56.

Amromin, G., & Sharpe, S. A. (2013). From the horse's mouth: Economic conditions and investor expectations of risk and return. *Management Science*, *60*(4), 845–866.

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, *61*(4), 1645–1680.

Balduzzi, P., & Lynch, A. W. (1999). Transaction costs and predictability: some utility cost calculations. *Journal of Financial Economics*, *52*(1), 47–78.

Balduzzi, P., & Lynch, A. W. (2000). Predictability and transaction costs: The impact on rebalancing rules and behavior. *The Journal of Finance*, *55*(5), 2285–2309.

Bansal, R., Kiku, D., Shaliastovich, I., & Yaron, A. (2014). Volatility, the macroeconomy, and asset prices. *The Journal of Finance*, *69*(6), 2471–2511.

Bansal, R., Kiku, D., & Yaron, A. (2011). An empirical evaluation of the long-run risks model for asset prices. *Working Paper, Duke University and University of Pennsylvania*.

Bansal, R., & Yaron, A. (2004). Risks for the long run: A potential resolution of asset pricing puzzles. *The Journal of Finance*, *59*(4), 1481–1509.

Barro, R. J. (2006). Rare disasters and asset markets in the twentieth century. *The Quarterly Journal of Economics*, *121*(3), 823–866.

Basak, S. (2005). Asset pricing with heterogeneous beliefs. *Journal of Banking and Finance*, *29*(11), 2849–2881.

Baz, J., Breedon, F., Naik, V., & Peress, J. (2001). Optimal Portfolios of Foreign Currencies. *The Journal of Portfolio Management*, *28*(1), 102-111.

Bertaut, C., & Judson, R. (2014). Estimating u.s. cross-border securities positions: New data and new methods. *Board of Governors pf the Federal Reserve System, International Finance Discussion Paper N. 113*.

Bhamra, H. S., & Uppal, R. (2013). Asset prices with heterogeneity in preferences and beliefs. *Working paper*.

Black, F. (1972). Capital market equilibrium with restricted borrowing. *The Journal of Business*, *45*(3), 444–455.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *The Journal of Political Economy*, *81*(3), 637–654.

Bollerslev, T., Tauchen, G., & Zhou, H. (2009). Expected stock returns and variance risk premia. *The Review of Financial Studies*, *22*, 4463–4492.

Bollerslev, T., & Todorov, V. (2011). Tails, fears, and risk premia. *The Journal of Finance*, *66*(6), 2165–2211.

Brandt, M. W. (2005). Portfolio choice problems. In *Handbook of financial econometrics* (p. 269-336). Amsterdam: Elsevier.

Brennan, M. (1975). The optimal number of securities in a risky asset portfolio when there are fixed costs of transacting: Theory and some empirical results. *Journal of Financial and Quantitative Analysis*, *10*, 483–496.

Brunnermeier, M., & Pedersen, L. (2009). Market liquidity and funding liquidity. *Review of Financial Studies*, *22*, 2201–2238.

Campbell, J. Y., & Cochrane, L. H. (1999). By force of habit: A consumption-based explanation of aggregate stock market behavior. *The Journal of Political Economy*, *107*(2), 205–251.

Campbell, J. Y., Giglio, S., Polk, C., & Turley, R. (2017). An intertemporal capm with stochastic volatility. *Forthcoming, Journal of Financial Economics*.

Campbell, J. Y., & Viceria, L. M. (1999). Consumption and portfolio decisions when expected returns are time-varying. *The Quarterly Journal of Economics*, *114*(2), 433–495.

Chinn, M. D., & Ito, H. (2006). What Matters for Financial Development? Capital Controls, Institutions, and Interactions. *Journal of Development Economics*, *81*(1), 163-192.

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, *28*(3), 591–605.

Christie, S. (2005). Is the sharpe ratio useful in asset allocations? *MAFC Research Paper No. 31, Applied Finance Centre, Macquarie University*.

Cottle, R. W., Pang, J.-S., & Stone, R. E. (1992). The linear complementarity problem. In *The linear complementarity problem* (p. 761).

Cox, J. C., Ingersoll, J. E., & Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrica*, *53*, 385–407.

Crucini, M. J., & Shintani, M. (2008). Persistence of the law of one price deviations: evidence from micro-data. *Journal of Monetary Economics*, *55*, 629–644.

Daniel, K., Hodrick, R. J., & Lu, Z. (2017). The Carry Trade: Risks and Drawdowns. *Critical Finance Review*, *6*(2), 211-262.

Davis, M. H. A., & Norman, A. R. (1990). Portfolio selection with transaction costs. *Mathematics of Operations Research*, *15*(4), 676–713.

Della Corte, P., Sarno, L., & Tsiakas, I. (2009). An Economic Evaluation of Empirical Exchange Rate Models. *Review of Financial Studies*, *22*(9), 3491-3530.

Della Corte, P., Ramadorai, T., & Sarno, L. (2016). Volatility Risk Premia and Exchange Rate Predictability. *Journal of Financial Economics*, *120*, 21-40.

DeMiguel, V., Garlappi, L., & Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *Review of Financial Studies*, *22*, 1915-1953.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *13*, 253–253.

Donohue, C., & Yip, K. (2003). Optimal portfolio rebalancing with transaction costs: Improving on calendar- or volatility-based strategies. *Journal of Portfolio Management*, *29*, 49–63.

Drechsler, I., & Yaron, A. (2011). What's vol got to do with it. *The Review of Financial Studies*, *24*, 1–45.

Dumas, B., & Buss, A. (2017). The dynamic properties of financial-market equilibrium with trading fees. *Working Paper*.

Dumas, B., & Luciano, E. (1991). An exact solution to a dynamic portfolio choice problem under transaction costs. *The Journal of Finance*, *46*, 577–595.

Dybvig, P. H., & Pezzo, L. (2018). *Mean-Variance Porfolio Rebalancing with Transaction Costs* (Working paper). Washington University.

Dybvig, P. H., & Ross, S. A. (1982). Portfolio efficient sets. *Econometrica*, *50*(6), 1125–1546.

Dybvig, P. H., & Ross, S. A. (1987). Arbitrage. In *New palgrave, a dictionary of economics* (pp. 100–106).

Epstein, L. G., & Zin, S. E. (1989). Substitution, risk aversion, and the temporal behavior of consumption and asset returns: A theoretical framework. *Econometrica*, *57*(4), 937–969.

Epstein, L. G., & Zin, S. E. (1991). Substitution, risk aversion, and the temporal behavior of consumption and asset returns: An empirical analysis. *The Journal of Political Economy*, *99*(2), 263–286.

Gallant, A., & Tauchen, G. (1989). Seminonparametric estimation of conditionally constrained heterogeneous processes: Asset pricing applications. *Econometrica*, *57*(5), 1091–1120.

Goldstein, A. (1979). *A model of capital asset pricing with transaction costs* (Unpublished doctoral dissertation). Yale University.

Goodman, J., & Ostrov, D. N. (2010). Balancing small transaction costs with loss of optimal allocation in dynamic stock trading strategies. *SIAM Journal on Applied Mathematics*, *70*(6), 1977–1998.

Goyal, A., & Welch, I. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, *21*(4), 1455–1508.

Greenwood, R., & Shleifer, A. (2014). Expectations of returns and expected returns. *Review of Financial Studies*, *27*(3), 714–746.

Grossman, S. J., & Stigliz, J. E. (1980). On the impossibility of informationally efficient markets. *The American Economic Review*, *70*(3), 393–408.

Han, J. (2005). *Dynamic portfolio management: An approximate linear programming approach* (Unpublished doctoral dissertation). Stanford University.

Hansen, L. P., & Singleton, K. J. (1982). Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica*, *50*(5), 1269–1286.

Hansen, L. P., & Singleton, K. J. (1983). Stochastic consumption, risk aversion, and the temporal behavior of asset returns. *The Journal of Political Economy*, *91*(2), 249–265.

Harrison, J., & Kreps, D. (1979). Martingales and arbitrage in multiperiod securities markets.

Journal of Economic Theory, *20*, 381–408.

He, Z., & Krishnamurthy, A. (2013). Intermediary asset pricing. *American Economic Review*, *103*, 732–770.

Horvth, J., Rtfai, A., & Dome, B. (2008). The border effect in small open economies. *Economic Systems*, *32*, 33–45.

Hou, K., & Moskowitz, J., Tobias. (2005). Rare disasters and asset markets in the twentieth century. *The Review of Financial Studies*, *18*(3), 981–1020.

Irle, A., & Prelle, C. (2008). A renewal theoretic result in portfolio theory under transaction costs with multiple risky assets. *Working Paper*.

Jacob, N. L. (1974). A limited-diversification portfolio selection model for the small investor. *The Journal of Finance*, *29*, 874–856.

Karnaukh, N., Ranaldo, A., & Soederlind, P. (2015). Understanding FX Liquidity. *The Review of Financial Studies*, *28*(11), 3073-3108.

Kozak, S., Nagel, S., & Santosh, S. (2015). *Interpreting Factor Models* (Working paper).

Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, *53*(6), 1315–1335.

Ledoit, O., & Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, *10*, 603-621.

Ledoit, O., & Wolf, M. (2008). Robust Performance Hypothesis Testing with the Sharpe Ratio. *Journal of Empirical Finance*, *15*, 850-859.

Leland, H. (2000). Optimal portfolio implementation with transaction costs and capital gain taxes. *Working Paper*.

Lettau, M., & Ludvigson, S. (2001). Consumption, aggregate wealth, and expected stock returns. *The Journal of Finance*, *56*(3).

Lin, H., Wu, C., & Zhou, G. (2016). From the horse's mouth: Economic conditions and investor expectations of risk and return. *Management Science, forthcoming*.

Liu, H. (2004). Optimal consumption and investment with transaction costs and multiple risky assets. *The Journal of Finance*, *59*, 289–338.

Liu, H., & Loewenstein, M. (2002). Optimal portfolio selection with transaction costs and finite horizons. *Review of Financial Studies*, *15*, 805–835.

Liu, W. (2006). A liquidity-augmented capitl asset pricing model. *Journal of Financial Economics*, *82*, 631–671.

Lo, A. W. (2002). Statistics of sharpe ratio. *Financial Analysis Journal*, *58*, 36–52.

Ludvigson, S. C., Ma, S., & Ng, S. (2016). Uncertainty and business cycles: Exogenous impulse or endogenous response? *Working Paper, NYU*.

Lustig, H., Roussanov, N., & Verdelhan, A. (2011). Common Risk Factors in Currency Returns. *Review of Financial Studies*, *24*(11), 3731-3777.

Lynch, A., & Tan, S. (2009). Multiple risky assets, transaction costs and return predictability: Allocation rules implications for u.s. investors. *Journal of Financial and Quantitative Analysis*, *45*, 1015–1053.

Mao, J. C. T. (1970). Essentials of portfolio diversification strategy. *The Journal of Finance*, *25*, 1109–1121.

Mao, J. C. T. (1971). Security pricing in an imperfect capital market. *Journal of Financial and Quantitative Analysis*, *6*, 1105–1116.

Markowitz, H. M. (1952). Portfolio selection. *The Journal of Finance*, *7*(1), 77–91.

Markowitz, H. M. (1959). Portfolio selection. In *Portfolio selection.*

Martin, I. (2017). What is the expected return on the market? *The Quarterly Journal of Economics*, *132*(1), 367–433.

Masters, S. J. (2003). Rebalancing: Establishing a consistent framework. *Journal of Portfolio Management*, *29*, 52–57.

Maurer, T. A., & Pezzo, L. (2018). Importance of transaction costs for asset allocations in fx markets. *Working Paper*.

Maurer, T. A., To, T.-D., & Tran, N.-K. (2018). *Optimal Factor Strategy in FX Markets* (Working paper). Washington University.

Mayshar, J. (1979). Transaction costs in a model of capital market equilibrium. *Journal of Political Economy*, *87*, 673–700.

Mayshar, J. (1981). Transaction costs and the pricing of assets. *The Journal of Finance*, *36*, 583–597.

Meese, R., & Rogoff, K. (1983). Empirical Exchange Rate Models of the Seventies: Do They Fir Out of Sample? *Journal of International Economics*, *14*, 3-24.

Menkhoff, L., Sarno, L., Schmeling, M., & Schrimpf, A. (2012). Currency Momentum Strategies. *Journal of Financial Economics*, *106*, 620-684.

Mertens, E. (2002). Comments on variance of the iid estimator in lo (2002). *Working Paper*.

Merton, R. C. (1980). On estimating the expected return on the market. *Journal of Financial Economics*, *8*, 323–361.

Moeller, B., S., Schlingemann, P., F., & Stulz, M., R. (2007). How do diversity of opinion and information asymmetry affect acquirer returns? *Review of Financial Studies*, *20*(6), 2047–2078.

Moreira, A., & Muir, T. (2017). Volatility managed portfolios. *The Journal of Finance*, *72*(4), 1611–1644.

Moreira, A., & Savov, A. (2017). The macroeconomics of shadow banking. *The Journal of Finance*, *0*.

Muir, T. (2017). Financial crises and risk premia. *The Quarterly Journal of Economics*, *132*(2), 765–809.

Muthuraman, K., & Kumar, S. (2006). Multidimensional portfolio optimization with proportional transaction costs. *Mathematical Finance*, *16*(2), 301–335.

Myers, J. D. (2009). *Portfolio optimization with transaction costs and preconceived portfolio weights* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Nagel, S., & Singleton, K. J. (2011). Asset pricing with garbage. *The Journal of Finance*, *66*(3), 873–909.

Newey, W., & West, K. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, *55*(3), 703–708.

Opdyke, J. (2007). Comparing sharpe ratios: so where are the p-values? *Journal of Asset Management*, *8*(5), 308–336.

Pasquariello, P. (2014). Financial market dislocations. *The Review of Financial Studies*, *27*, 1868–1914.

Pastor, L., & Stambaugh, R. F. (2003). Liquidity risk and expected stock returns. *The Journal of Political Economy*, *111*(3), 642–685.

Pippenger, J., & Phillips, L. (2008). Some pitfalls in testing the law of one price in commodity markets. *Journal of International Money and Finance*, *27*, 915–925.

Pogue, G. A. (1970). An extension of the markowitz portfolio selection problem to include variable transactions' costs, short sales, leverage policies and taxes. *The Journal of Finance*, *25*, 1005–1027.

Rapach, D., Ringgenberg, M. C., & Zhou, G. (2016). Short interest and aggregate stock returns. *Journal of Financial Economics*, *171*(1), 46–65.

Roll, R. (1977). A critique of the asset pricing theory's tests part i: On past and potential testability of the theory. *Journal of Financial Economics*, *4*(2), 129–176.

Roll, R. (1992). A mean/variance analysis of tracking error. *The Journal of Portfolio Management*, *18*(4), 13–22.

Ross, S. A. (1973). Return, risk and arbitrage. *Wharton Discussion Paper, I. Friend and J. Bicksler, eds., Risk and Return in Finance*, 189–217.

Ross, S. A. (1976). The Arbitrage Theory of Capital Asset Pricing. *Journal of Economic Theory*, *13*, 341-360.

Ross, S. A. (1978). A simple approach to the valuation of risky streams. *The Journal of Business*, *51*, 453–475.

Savov, A. (2011). Asset pricing with garbage. *The Journal of Finance*, *56*(1), 177–201.

Schachermayer, W. (2017). Asymptotic theory of transaction costs. In *Asymptotic theory of transaction costs* (p. 160).

Shreve, S. E., & Soner, H. M. (1994). Optimal investment and consumption with transaction costs. *The Annals of Applied Probability*, *4*(3), 609–692.

Taksar, M., Klass, M. J., & Assaf, D. (1988). A diffusion model for optimal portfolio selection in the presence of brokerage fees. *Mathematics of Operations Research*, *13*, 277–294.

Tobin, J. (1958). Liquidity preference as behavior towards risk. *Review of Economic Studies*, *25*, 65–86.

Wachter, J., & Seo, S. B. (2016). Option prices in a model with stochastic disaster risk. *Working Paper*.

Wachter, J. A. (2013). Can time-varying risk of rare disasters explain aggregate stock market volatility? *The Journal of Finance*, *68*(3), 987–1035.

Walker, D. (2015). Quarterly update: Foreign ownership of u.s. assets. *Council on Foreign Relations, www.cfr.org*.

White, H. (2001). Asymptotic theory for econometrician. In *Asymptotic theory for econometrician* (p. 64).

Yu, J. (2011). Disagreement and return predictability of stock portfolios. *Journal of Financial Economics*, *99*, 162–183.

Appendix A

A Non-Parametric Test For Representative Agent Pricing

A.1 Appendix A - Martin (2017) Lower Bound Existence Proof

First I show why LB_t is a lower bound for the market risk premium $\mathbb{E}_t[R_{t+1} - R_{t,f}]$ then I derive equation (1.3).

Suppose there exist a stochastic discount factor $M_{t+1} > 0$ satisfying the pricing equation (1.1), then by the Fundamental Theorem of Asset Pricing (FTAP, Ross (1973, 1978), Harrison and Kreps (1979), Dybvig and Ross (1987)) there exist an equivalent risk-neutral measure Q such that $R_f = \mathbb{E}[R^i]$ for any gross return R^i (thus for the market return R as well).

By definition the conditional risk neutral variance for the market return at horizon $t+1$ can be written as

$$Var_t^Q(R_{t+1}) \equiv E_t^Q[R_{t+1}^2] - E_t^Q[R_{t+1}]^2$$

where R_{t+1} is the gross cum-dividend market return. Still from FTAP we can go back and forth from the physical probability measure and the risk-neutral one, thus $E_t^Q[R_{t+1}^2] = E_t[R_{t,f}M_{t+1}R_{t+1}^2]$ and by definition of the risk-neutral measure, $E_t^Q[R_{t+1}]^2 = {R_{t,f}}^2$, hence

$$Var_t^Q(R_{t+1}) = E_t[R_{t,f}M_{t+1}R_{t+1}^2] - R_{t,f}^2$$

dividing the above equation by the gross risk-free return $R_{t,f}$ and rearranging

$$\frac{Var_t^Q(R_{t+1})}{R_{t,f}} = E_t[R_{t+1} - R_{t,f}] + Cov_t(M_{t+1}R_{t+1}, R_{t+1})$$

if $Cov_t(M_{t+1}R_{t+1}, R_{t+1}) \leq 0$, which together with $M_{t+1} > 0$ defines the NCC, then $LB_t \equiv \frac{Var_t^Q(R_{t+1})}{R_{t,f}}$ is a lower bound for $RP_t \equiv E_t[R_{t+1} - R_{t,f}]$.

Next, I derive equation (1.3). From the definition of variance, using hats to denotes ex-dividend quantities and letting S be the cum-dividend market level

$$\begin{aligned} Var_t^Q(R_{t+1}) \equiv & \quad E_t^Q\left[\left(\frac{S_{t+1}}{S_t}\right)^2\right] - E_t^Q\left[\frac{S_{t+1}}{S_t}\right]^2 \\ = & \quad E_t^Q\left[\left(\frac{\hat{S}_{t+1}}{\hat{S}_t}DY_t\right)^2\right] - {R_{t,f}}^2 \\ = & \quad \frac{(DY_t)^2 R_{t,f}}{(\hat{S}_t)^2} E_t^Q\left[\frac{\hat{S}_{t+1}^2}{R_{t,f}}\right] - {R_{t,f}}^2 \end{aligned}$$

by no arbitrage (see Martin (2017)), since the options are written on $\hat{S}_t$

$$E_t^Q\left[\frac{\hat{S}_{t+1}^2}{R_{t,f}}\right] = 2\int_0^\infty \hat{call}_t(k)dK = 2\left(\int_0^{\hat{F}_t} \hat{call}_t(k)dK + \int_{\hat{F}_t}^\infty \hat{call}_t(k)dK\right)$$

since deep-in-the-money call options are neither liquid in practice nor intuitive to think about, it is convenient to split the range of integration for $E_t^Q\left[\frac{\hat{S}_{t+1}^2}{R_{t,f}}\right]$ into two and use the put-call parity to replace in-the-money call prices with out- of-the-money put prices. Assume that Market Dividends are paid as lump sums D_{t+1} at the and of the period $[t : t+1]$ but before $t+1$, then the following is true

$$max(S_{t+1} - D_{t+1} - k, 0) = max(k - S_{t+1} + D_{t+1}, 0) + (S_{t+1} - D_{t+1}) - k$$

since $\hat{S}_{t+1} = S_{t+1} - D_{t+1}$

$$max(\hat{S}_{t+1} - k, 0) = max(k - \hat{S}_{t+1}, 0) + (S_{t+1} - D_{t+1}) - k$$

by linearity of the pricing equation

$$\hat{call}_t(k) = \hat{put}_t(k) + \hat{S}_t - PV(D_{t+1}) - \frac{k}{R_{t,f}}$$

where $PV(D_{t+1}) = \mathbb{E}_t^Q\left[\frac{D_{t+1}}{R_{t,f}}\right] = (1 - DY_t)\mathbb{E}_t^Q\left[\frac{\hat{S}_{t+1}}{R_{t,f}}\right] = \frac{DY_t - 1}{DY_t}\hat{S}_t$ and the last equality comes from $R_{t,f} = \mathbb{E}_t^Q\left[\frac{S_{t+1}}{S_t}\right]$. Applying the put-call parity

$$\begin{aligned}\int_0^{\hat{F}_t} \hat{call}_t(k)dK = & \int_0^{\hat{F}_t} \hat{put}_t(k)dK + \hat{F}_t\left(\hat{S}_t - \frac{DY_t - 1}{DY_t}\hat{S}_t\right) - \frac{\hat{F}_t^2}{2R_{t,f}} \\ = & \int_0^{\hat{F}_t} \hat{put}_t(k)dK + \hat{F}_t\left(\frac{\hat{S}_t}{DY_t} - \frac{\hat{F}_t}{2R_{t,f}}\right)\end{aligned}$$

which implies

$$E_t^Q\left[\frac{\hat{S}_{t+1}^2}{R_{t,f}}\right] = 2\left[\int_0^{\hat{F}_t} \hat{put}_t(k)dK + \hat{F}_t\left(\frac{\hat{S}_t}{DY_t} - \frac{\hat{F}_t}{2R_{t,f}}\right) + \int_{\hat{F}_t}^{\infty} \hat{call}_t(k)dK\right]$$

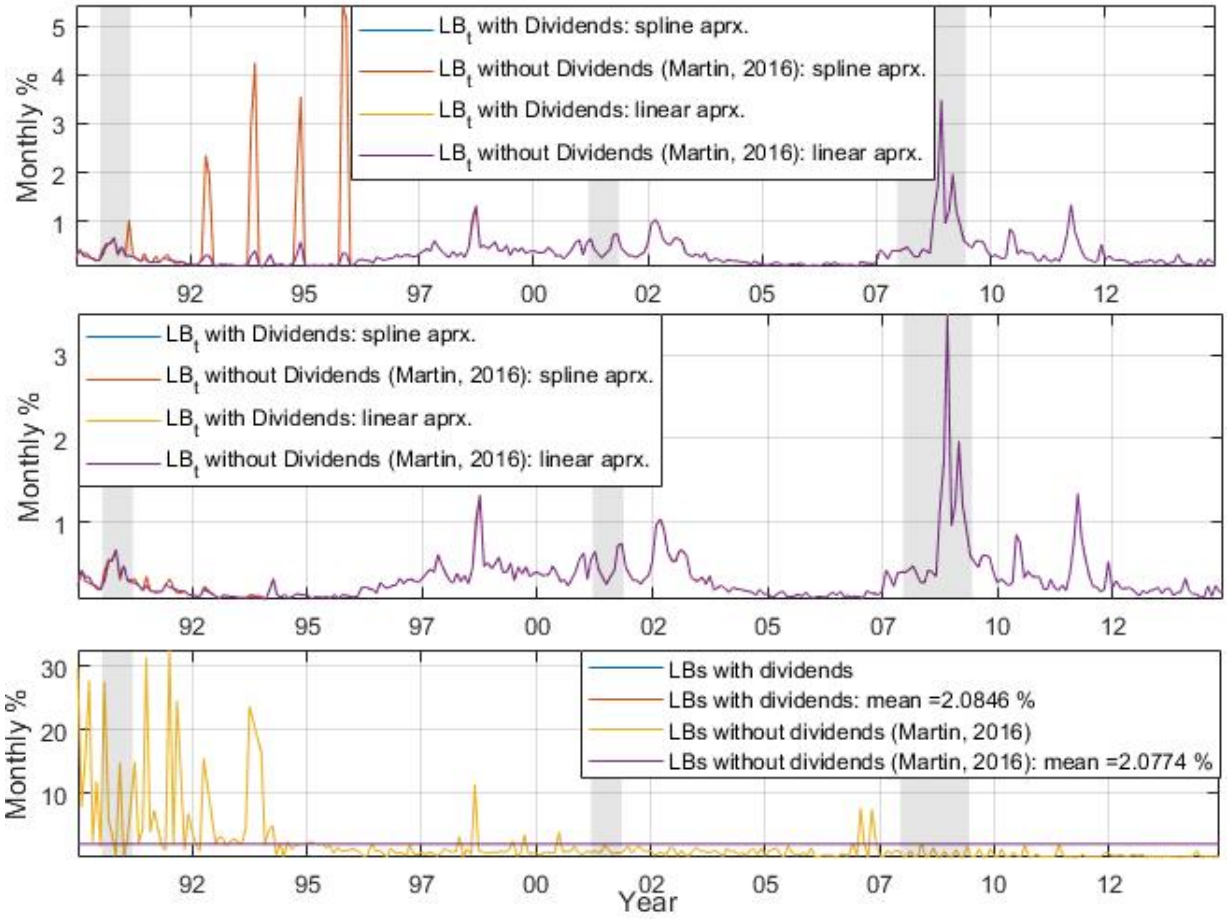
plugging $E_t^Q\left[\frac{\hat{S}_{t+1}^2}{R_{t,f}}\right]$ in $Var_t^Q(R_{t+1}) = \frac{(DY_t)^2 R_{t,f}}{(\hat{S}_t)^2} E_t^Q\left[\frac{\hat{S}_{t+1}^2}{R_{t,f}}\right] - {R_{t,f}}^2$ delivers equation (1.3)

$$LB_t = 2\frac{(DY_t)^2}{(\hat{S}_t)^2}\left(\int_0^{\hat{F}_t} \hat{put}_t(k)dK + \hat{call}_t(k)dK\right) \tag{A.1}$$

setting $DY = 1$ delivers the original Martin (2017)' measure used in the current study.

A.2 Appendix B - Linear vs. cubic spline Lower Bound approximations

In order to compute the lower bound measure at time t, LB_t, according to equation (1.3) I use the SPX options (Put and Call) bid quotes at horizon 1 month for the different available strikes as at the of the first business day of month t from Optsum and Optionmetrics. To approximate the integral in (1.3) we first need to interpolate the functions $\hat{put}(k)$ and $\hat{call}(k)$ over a continuum of strikes. In the study, following Martin (2017), I have used a linear interpolation. Another popular interpolant option is the cubic spline.



The figure shows the time-series of lower bounds in the main sample MS computed with the linear as well as the cubic-spline method with and without dividends as an explicit argument of LB_t (see proof of Proposition 1 for an explicit formula of LB_t as a function of dividends: i.e. eq.(A.1)): the top panel uses the full available data: note how, independently from the presence of dividends, the spline and the linear interpolation almost perfectly overlap except for isolated points in the pre-1996 period. I adopt the most conservative of the approaches by excluding from the main sample all instances in which the spline and the linear interpolation differ proportionally (with respect to the linear scheme) by more than 50%. The new resulting sample (which is the one used in the main analysis), along with the

lower bound estimates, is shown in the middle panel. All the estimates are now very close; The same point can be more precisely appreciated by looking at the bottom graph which plots the absolute percentage difference between the lower bound measures (with respect to the linear scheme) when computed using the linear as opposed to the spline approximation for the case the bound features dividends and for the case it does not. Again, it is impossible to distinguish between the case in which dividends are included from the case in which they are not, furthermore, the maximum discrepancy is now around 30% and on average the two scheme only differ by 2%.

A.3 Appendix C - Rationale behind the choice of the market risk premium predictors

In order to tackle the potential data-snooping issue discussed in Section 1.3, the selection of the actual list Z_t is disciplined by a constructive economic rationale. In particular, the objective of this study is to find, if any, violations of the Martin's class of representative agent pricing: a key subclass is represented by the Consumption-based Representative Agent Models (as we show in Section 1.2 all the mainstream CRAMs are inside the Martin's class). Any violation will thus contain instances of simultaneous failure of CRAMs; because these are systematic failures they must be associated with the failure of, at least one, of the key assumptions of these models. The key assumptions of CRAMs are: ($CRA1$) the existence of a representative agent, ($CRA2$) the absence of market frictions, ($CRA3$) the absence of arbitrage, and ($CRA4$) the presence of (closed) and real economies. Thus, predictors in Z_t will be selected as proxies against $CRA1$-$CRA4$. In other words, the predictors in Z_t can be viewed as state variables that should not be able to predict the excess market return π_{t+1}

under the null that CRAMs hold. While the selection does not require each predictor to be uniquely associated with a single dimension going against the key CRAM assumptions, to streamline the exposition, I offer a possible mapping of the variables into proxies against each of the assumptions separately.

A representative agent exists (assumption $CRA1$) under symmetric information,[90] market completeness, and in general[91] independence of preferences from wealth distribution. Because market completeness is not easy to proxy for, I focus on proxies for wealth distribution and asymmetric information. In particular, I exploit the percentage changes in the $GINI$ index, $GINIchg$, from the United States Census Bureau to capture the wealth distribution effect, while proxying for informational asymmetries through the Rapach et al. (2016) short interest index SII and the Ludvigson et al. (2016) financial uncertainty index F. SII seizes " the superior informational content of short seller in anticipating future aggregate cash flows and associated market returns" , while F, originally designed to capture the latent degree of unpredictability in financial markets, has zero correlation with SII[92] and can be considered as an asymmetric information proxy.[93]

[90]This is true even in frameworks such as Basak (2005) or Bhamra and Uppal (2013) that allow aggregation under heterogeneity in beliefs.

[91]Unless preference are homotetic and identical.

[92]See Table 1.2

[93]Empirically Moeller, Schlingemann, and Stulz (2007) and the literature therein refer to uncertainty and analyst forecasts dispersion measures as asymmetric information proxies, while from a theoretical point of view, standard predictions of asymmetric information models, such as Grossman and Stigliz (1980) or Kyle (1985), dictates a positive correlation with price impact measures, proxied in this study by the Pastor and Stambaugh (2003) (il)liquidity index $ILLIQpi$. The following table displays the correlation coefficients over the overlapping period $[1992-2005]$ of F and classical information asymmetry proxies such as the $I/B/E/S$ analysts earning growth forecast dispersions, ($AnlystForecastsDispIBES$ for the $SP500$ and $AnlystForecastsDispYu$ for an average of individual stocks forecasts as in Yu (2011)) and market volatility proxies such as the CBOE VIX index or a GARCH(1,1) on the $SP500$ return $GARCH$

Corr	$ILLIQpi$	$AnlystForecastsDispIBES$	$AnlystForecastsDispYu$
F	0.30	0.67	0.63
$GARCH$	0.17	0.72	0.35
VIX	0.31	0.62	0.44

Turning to trading frictions, the opposite of $CRA2$, I track this dimension either by looking at the impact of taxes, through the annual percentage changes in the aggregate dollar amount paid in capital gain taxes, $TAXchg$, or by employing two popular (il)liquidity indexes: the Pastor and Stambaugh (2003) index, $ILLIQpi$, designed to capture the percentage cost incurred in a 1 million 1962 USD trade in the market,[94] and the mimicking portfolio for the W. Liu (2006) index, $ILLIQts$, constructed with the aim of seizing the trading speed dimension of liquidity.[95]

The absence of arbitrage, assumption $CRA3$, is a sufficient condition for the Law of One Price (LoP) to hold thus if LoP fails there is arbitrage. Following the standard equilibrium framework I look at price relations in the financial rather than the commodity markets[96] and track significant departures from the LoP through the Pasquariello (2014) market dislocation index MDI which measures abnormal discrepancies between actual (mid-quote) and theoretical prices using three textbook arbitrage parities in stock, foreign exchange, and money markets: the Covered Interest Rate Parity, the Triangular Arbitrage Parity and the American Depository Receipt Parity. I also add two more general and popular mispricing proxies: the Baker and Wurgler (2006) sentiment index, $Sent$, designed to capture miss-pricing due to subjective valuations not reflecting rational risk compensation, and the Dow-Jones Industrial Average book-to-market ratio BM.

Finally $CRA4$ dictates CRA models to be embedded in real and closed economies. I take into account the effect of nominal forces and the impact of foreign markets by including

F and $GARCH$ are the only proxies available throughout the required sample period $\{1973:2-2014:12\}$. Note how F is the measure of uncertainty which simultaneously correlates the most with the price impact proxy $ILLIQpi$ and with the analyst forecast dispersions.

[94] Similarly to the Amihud (2002) proxy, it is a price impact measure.

[95] Another proxy constructed with the same aim is designed in Hou and Moskowitz (2005).

[96] In the context of commodity markets Horvth, Rtfai, and Dome (2008), Pippenger and Phillips (2008) and Crucini and Shintani (2008) find contrasting results concerning the validity of the LoP.

the growth rate of the U.S. money supply, $M1g$,[97] and the rate at which the U.S. dollar appreciate, $USDg$,[98] into the list of candidates.

A.4 Appendix D - ICM vs. OLS horse-race to forecast the market risk premium

As shown by Goyal and Welch (2008), a naive OLS regression of excess market returns on a large number of predictors will over-parametrize the model and lead to poor out-of-sample forecasts, I therefore combine the information from the set of predictors to obtain optimal forecasts using the Iterated Combination Method (ICM) of Lin et al. (2016). First, predictive regressions are run on each predictor and a constant to obtain individual forecasts. Then a weighted average of the mean of all of the individual forecasts and the prevailing mean of the excess market return using all observations till time t, serves as the t forecast. This methodology basically amounts to a weighted average of a shrinked OLS regression, in which the out-of-diagonal elements in the regressors' matrix are set to zero and the regressors' coefficients are divided by the number of regressors, and the prevailing dependent variable mean.[99]

[97] For the sake of parsimony and due to the high correlation of 0.55 with inflation I do not include the latter.

[98] The index is a weighted (over the volume of bilateral transactions) average of the foreign exchange value of the U.S. dollar against the currencies of a broad group of major U.S. trading partners. The index captures the impact of foreign financial markets on the domestic stock market through the weights: since the third quarter of 1982 the U.S. runs a deficit in the current account (see Balance on Current Account, available through FRED at https://fred.stlouisfed.org/series/NETFI.html), and, as reported by Bertaut and Judson (2014) on behalf of the Board of Governors of the Federal Reserve System, the excess of imports over exports has been funded primarily by foreign acquisitions of U.S. securities. See also Walker (2015).

[99] The weights are designed to minimize the out-of-sample mean squared error and increase the out of sample R^2. (See Lin et al. (2016))

The next table compares the performance of the ICM and OLS approaches for the 6 selected specifications

$$\pi_{t+1} = \alpha + \beta\pi_{t+1}^{M} + \varepsilon_{t+1} \quad M \in \{OLS, ICM\} \quad (*)$$

		α	β	R2	MSE	DM tstat	Obs.
			(1)				
Traning (in-sample)	OLS	0.000	1.000***	0.440	0.001	-2.913	192
	ICM	0.000	1.000***	0.077	0.002	-2.913	192
Main (out-of-sample)	OLS	0.000	0.301***	0.054	0.003	3.466	289
	ICM	0.000	1.002***	0.072	0.002	3.466	289
			(2)				
Traning (in-sample)	OLS	0.000	1.000***	0.423	0.001	-2.918	203
	ICM	0.000	1.000***	0.061	0.002	-2.918	203
Main (out-of-sample)	OLS	0.000	0.305***	0.086	0.003	3.375	230
	ICM	0.000	0.666***	0.080	0.002	3.375	230
			(3)				
Traning (in-sample)	OLS	0.000	1.000***	0.383	0.001	-2.679	203
	ICM	0.000	1.000***	0.061	0.002	-2.679	203
Main (out-of-sample)	OLS	0.005**	0.299***	0.069	0.003	3.216	289
	ICM	0.000	0.930***	0.083	0.002	3.216	289
			(4)				
Traning (in-sample)	OLS	0.000	1.000***	0.464	0.001	-3.131	191
	ICM	0.000	1.000***	0.073	0.002	-3.131	191
Main (out-of-sample)	OLS	0.004	0.346***	0.083	0.003	3.044	289
	ICM	0.003	1.155***	0.091	0.002	3.044	289
			(5)				
Traning (in-sample)	OLS	0.000	1.000***	0.538	0.001	-3.409	191
	ICM	0.000	1.000***	0.074	0.002	-3.409	191
Main (out-of-sample)	OLS	0.005*	0.267***	0.115	0.003	1.911	230
	ICM	0.002	0.614***	0.082	0.002	1.911	230
			(6)				
Traning (in-sample)	OLS	0.000	1.000***	0.497	0.001	-3.058	191
	ICM	0.000	1.000***	0.077	0.002	-3.058	191
Main (out-of-sample)	OLS	0.004	0.313***	0.122	0.003	1.845	230
	ICM	0.002	0.518***	0.074	0.002	1.845	230

for a given model (a specific panel among (1) trhough (6)) the in-sample (training) and out-of-sample (main) time series of next month excess return estimates[100] π_{t+1}^{M} are produced with $M \in \{OLS, ICM\}$. The performances of the two different methods are judged using the out-of-sample mean squared error statistic, MSE, and the following regression benchmark

$$\pi_{t+1} = \alpha + \beta \pi_{t+1}^{M} + \varepsilon_{t+1}$$

in terms of the produced coefficient and R^2. A good performance entails a (relatively) small MSE, a (relatively) high R^2, $\alpha = 0$ and $\beta = 1$. No matter which model we look at OLS beats ICM in-sample (especially in terms of R^2 and MSE[101]) but ICM consistently out-perform OLS exactly where we care the most: out-of-sample in the period from January 1990 to December 2014. The ICM method, while producing similar out-of-sample R^2, generates a MSE 1.5 smaller, non significant αs, and βs which on average are much closer to 1 and 2.7 times bigger. In particular. according to the DM statistic, we always reject at the 95% the null of out-of-sample OLS MSE grater than the ICM ones. These results ex-post validate the choice of adopting the Lin et al. (2016) ICM method to generate the out-of-sample forecasts for the market excess return.

[100]Such estimates produced by either methods in either sample and for each model are not spurious: their first autocorrelation parameters safely lies at least 2 standard errors below 0.95.

[101]The Diebold and Mariano (1995) (DM) statistic has to be interpreted as the usual t-statistic design to compare the OLS and ICM MSEs.

A.5 Appendix E - Lower Bound violations imply too high model-based Sharpe ratios

This subsection shows how the rejection periods are times in which representative agents with preferences satisfying the NCC have ex-ante market Sharpe ratios systematically higher then ex-post (empirical) counterparts and how this can be seen as a direct consequence of the non-parametric test outcomes.

Since NCC is a restriction imposed on the SDF, it is linked to the preferences of any Martin's model representative agent; In particular, Martin (2017) shows how for the representative log investor the NCC holds with equality and $LB_t = \mathbb{E}_t^{log}[\pi_{t+1}]$. Given an objective proxy[102] for the conditional market volatility at time t, $\sigma_t(R_{t+1}) \equiv \sigma_t(\pi_{t+1})$, we can compute the time-series of conditional ex-ante Sharpe Ratios from the log investor perspective as $SR_t^{Ex-Ante} \equiv \frac{LB_t}{\sigma_t(R_{t+1})}$ and thus get an estimated average for the rejection periods as $\mathbb{E}[SR_t^{Ex-\hat{Ante}}|I^v = 1]$. Note that since the conditional log investor risk premium is the lowest possible among the Martin's models, any representative agent from these models will have an (average) market assessment in the rejection periods of *at least* $\mathbb{E}[SR_t^{Ex-\hat{Ante}}|I^v = 1]$. It is therefore instructive to compare $\mathbb{E}[SR_t^{Ex-\hat{Ante}}|I^v = 1]$ with the estimate for the ex-post Sharpe Ratio: i.e. $(SR^{Ex-\hat{Post}}|I^v = 1) \equiv \frac{\mathbb{E}[\hat{\pi_{t+1}}|I^v=1]}{\sigma(\hat{\pi_{t+1}}|I^v=1)}$. Results from this exercise are illustrated in the next table

[102] Computed as the out-of-sample prediction from a $GARCH(1,1)$ model on the market return using a rolling window of either 120, 60 or 30 observations.

SHARPE RATIO TEST, IV = 1

Estiamate	$\widehat{SR^{Ex-post}}\|I_t^v - E[\widehat{SR_t^{Ex-ante}}\|I_t^v]$	$\widehat{SR^{Ex-post}}\|I_t^v$	$E[\widehat{SR_t^{Ex-ante}}\|I_t^v]$	$E[\widehat{\pi_{t+1}}\|I_t^v] - E[\widehat{LB_t}\|I_t^v]$ (%)	$\sigma(\widehat{\pi_{t+1}}\|I_t^v) - E[\widehat{\sigma_{t;t+1}}\|I_t^v]$(%)
A: Rule 1					
	-0.216**	-0.119	0.097***	-1.283**	1.704
Pr(Estimate>0) - (base)	0.049	0.165	1.000	0.042	0.853
Pr(Estimate>0|unconditional moments)	0.047	0.160	-	-	-
B: Rule 2					
	-0.221**	-0.111	0.110***	-1.262**	1.544
Pr(Estimate>0) - (base)	0.046	0.182	1.000	0.053	0.825
Pr(Estimate>0|unconditional moments)	0.044	0.178	-	-	-
C: Rule 3					
	-0.233**	-0.127	0.110***	-1.399**	1.485
Pr(Estimate>0) - (base)	0.030	0.140	1.000	0.031	0.818
Pr(Estimate>0|unconditional moments)	0.027	0.134	-	-	-
D: Rule 4					
	-0.233**	-0.132	0.102***	-1.364**	1.691
Pr(Estimate>0) - (base)	0.035	0.138	1.000	0.033	0.859
Pr(Estimate>0|unconditional moments)	0.031	0.131	-	-	-
E: Rule 5					
	-0.252**	-0.147	0.105***	-1.479**	1.564
Pr(Estimate>0) - (base)	0.029	0.117	1.000	0.028	0.808
Pr(Estimate>0|unconditional moments)	0.026	0.111	-	-	-
F: Rule 6					
	-0.256*	-0.145	0.111***	-1.654**	1.766
Pr(Estimate>0) - (base)	0.053	0.163	1.000	0.050	0.823
Pr(Estimate>0|unconditional moments)	0.048	0.156	-	-	-
G: SHARPE RATIO TEST, IV = 0 (Rule 1)					
	0.221	0.287	0.066	0.008	0.006
Pr(Estimate>0) - (base)	0.999	1.000	1.000	1.000	0.960
Pr(Estimate>0|unconditional moments)	0.999	-	-	-	-

the first column contains the estimates for $(SR^{Ex-Post}|I^v = 1) - \mathbb{E}[SR_t^{Ex-Ante}|I^v = 1]$ as well as two different ways[103] to compute its p-value against the alternative of $(SR^{Ex-Post}|I^v =$

[103]The econometric challenge here is to make inference on the ex-post Sharpe Ratio estimate while simultaneously taking into account the parameter uncertainty surrounding the estimates of $\mathbb{E}[SR_t^{Ex-Ante}|I^v = 1]$ and $(SR^{Ex-Post}|I^v = 1)$. With respect to the first challenge I use the results from Lo (2002), Mertens (2002), Christie (2005) and Opdyke (2007) who derive the normal limiting distribution for the Sharpe Ratio measure only imposing stationarity and ergodicity on π_{t+1}: this gives us a way to compute the standard error of $(\hat{SR^{Ex-Post}}|I^v = 1)$ as $SE(SR^{Ex-Post}) = \sqrt{\frac{1-\lambda_3(\hat{SR^{Ex-Post}}|I^v=1)+0.25(\lambda_4-1)(\hat{SR^{Ex-Post}}|I^v=1)^2}{n_{I^v=1}-1}}$ with λ_3 and λ_4 representing the skewness and kurtosis of $(\pi_{t+1}|I^v = 1)$ and $n_{I^v=1}$ being the number of rejection observations. Since the estimate for $\mathbb{E}[SR_t^{Ex-Ante}|I^v = 1]$ is a canonical OLS coefficient from regressing a constant on the available time series $(SR_t^{Ex-Ante}|I^v = 1)$, its standard error, which we denote as $SE(\mathbb{E}[SR_t^{Ex-Ante}|I^v = 1])$, is an ordinary Newey and West (1987) corrected standard error. The second challenge is therefore tackled by estimating the standard error for $(SR^{Ex-Post}|I^v = 1) - \mathbb{E}[SR_t^{Ex-Ante}|I^v = 1]$, as $\sqrt{SE(SR^{Ex-Post})^2 + 2SE(SR^{Ex-Post})SE(\mathbb{E}[SR_t^{Ex-Ante}|I^v = 1]) + SE(\mathbb{E}[SR_t^{Ex-Ante}|I^v = 1])^2}$ which correctly accounts for the potential correlation among the two estimates. I use such standard error to compute

$1) < \mathbb{E}[SR_t^{Ex-Ante}|I^v = 1]$, each row refers to a different rule except the last raw which reports the analogous results while conditioning on the non-rejection periods for the representative case of the first rule I_t^v.[104]

The ex-ante Sharpe Ratios of the Martin's Models are systematically above their ex-post counter parts in the rejection periods: $(\hat{SR^{Ex-Post}}|I^v) - \mathbb{E}[\hat{SR_t^{Ex-Ante}}|I^v]$ is always negative only when $I^v = 1$ (5 out of 6 case at the 5% level and 1 out of 6 at the 10%). As column 2 and 3 show, this patter is generated by economically negative, albeit insignificant, average ex-post Sharpe ratios and statistically positive ex ante counter-parts. Note how the situation reverses when we condition on $I^v = 1$: there things seem to work with ex-ante Sharpe ratios implied by the Martin's class solidly below their ex-post realizations.

These results can be viewed as a *direct* implication of the non-parametric test outcomes: this is because

$$\mathbb{E}[\hat{SR_t^{Ex-Ante}}|I^v] \equiv \mathbb{E}\left[\frac{\hat{LB_t}}{\hat{\sigma_t(\pi_{t+1})}}|I^v\right] \approx \frac{\mathbb{E}[\hat{LB_t}|I^v]}{\mathbb{E}[\hat{\sigma_t(\pi_{t+1})}|I^v]}$$

where the approximation is justified by the fact that, across the six analyzed cases, the first estimate is on average 1.07 the second with a maximum of 1.11, and these differences are never statistically significant. Then $\mathbb{E}[\hat{SR_t^{Ex-Ante}}|I^v = 1] > (\hat{SR^{Ex-Post}}|I^v = 1)$ is approximately equal to

$$\frac{\mathbb{E}[\hat{LB_t}|I^v]}{\mathbb{E}[\hat{\sigma_t(\pi_{t+1})}|I^v]} > \frac{\mathbb{E}[\hat{\pi_{t+1}}|I^v = 1]}{\hat{\sigma(\pi_{t+1}}|I^v = 1)} \tag{A.2}$$

In light of this approximation column 4 and 5 of the above table compare the numerators and the denominators of eq. (A.2): column 4 says that $\mathbb{E}[\pi_{t+1}|I^v = 1] - \mathbb{E}[LB_t|I^v = 1] \equiv \mathbb{E}[y_{t+1}|I^v = 1]$ is significantly samller than zero while according to column 5 we cannot reject

the first (base) p-value. As a robustness check I also compute the the p-value in the case λ_3 and λ_4 represent the skewness and kurtosis of π_{t+1} over the entire available sample: this is the second p-value reported in the table.

[104]The other rules yield virtually identical results.

the null of $\sigma(\pi_{t+1}|I^v = 1) = \mathbb{E}[\sigma_t(\pi_{t+1})|I^v]$. This means that the results from this subsection are implied by the lower bound LB_t being on average above the risk premium in the rejection periods which is the exact same statement made by the non-parametric test.

A.6 Appendix F - Performance of actual representative models

The non-parametric test rejects the entire class of Martin's models conditional on periods t such that $I_t^v = 1$. One of the advantage of the test is its ability to make inference over equilibrium models which are usually difficult to test either because based on unobservable state variables or because such variables, when available, are very noisy in the data. This section, subject to the just mentioned data caveat, can be viewed as a robustness check on the ability of the non-parametric test to correctly identify periods where indeed actual mainstream models of the Martin's class perform worse as well as a useful exercise to quantify their performance.

I compare three cornerstone consumption-based models belonging to the Martin's class: the Campbell and Cochrane (1999) external habit model, $CC99$, the Bansal and Yaron (2004) long run risk model, $BY04$, and the J. A. Wachter (2013) time-varying rare disaster model,[105] $W13$. To ensure, by the logic of the test, that bed performance conditional on $I^v = 1$ come from the failure of the models' FOCs, I use the original calibrations of these models (for which Martin (2017) proofs the NCC holds) as shown in the following table

[105] I am thankful to Professor Wachter for providing the original code used to perform the simulations in her model.

Model	State Variables	Parameters	Original	Main sample
CC99	Cosumption growth	g: Mean consumption growth	0.002	0.002
		σ: Standard deviation of consumption growth	0.004	0.003
	Surplus Consumption ratio	ϕ: Persistence coefficient	0.989	0.942
		δ: Subjective time discount factor	0.990	0.990 (Assumed)
		γ: Utility curvature	2.000	2 (Assumed)
		$M_t^{CC99} = e^{\log(\delta) - \gamma \log\left(S_{t+1}/S_t\right) - \gamma \log\left(C_{t+1}/C_t\right)}$		
BY04	Mean Cosumption growth (recovered with BKY 2007)	γ: Coeffieicent of relative risk aversion	10.000	10 (Assumed)
		ψ: Elasticity of Intertemporal Substitution	1.500	1.5 (Assumed)
		δ: Subjective time discount factor	0.998	0.998 (Assumed)
		ρ: Autocorrelation of the long-run component	0.979	0.943
		φe: Time invariant volatility component of long run component	0.044	0.050
	Cosumption growth volatility (recovered with BKY 2007)	μc: Time invariant component of consumption growth	0.002	0.002
		σ: Long run standard deviation og consumption growth	0.008	0.003
		v: Mean reversion of consumption variance	0.987	0.957
		σω:Time invariant volatility component of consumption growth variance	0.000	0.000
		$M_t^{BY04} = e^{\theta \log(\delta) - \frac{\theta}{\psi} \log\left(C_{t+1}/C_t\right) + (\theta-1) \log\left(r^{C}{}_{t+1}/r^{C}{}_{t}\right)}, \quad \theta = \frac{1-\gamma}{1-1/\psi}$		
W13	Consumption growth	γ: Coeffieicent of relative risk aversion	3.000	3 (Assumed)
		β: Rate of time preference (annualized)	0.012	0.012 (Assumed)
		μ: Average consumption growth (annualized)	0.025	0.023
		σ: Volatility of cosumption growth (annualized)	0.020	0.010
	Disaster intensity (recovered with BT 2011)	λ: Average probability of a rare disaster (annualized)	0.036	0.021 (with 0.0355 in 95% C.I.)
		κ: Mean reversion in the disaster intesity process (annualized)	0.080	0.069 (with 0.08 in 95% C.I.)
		σλ: Volatility of the disaster intensity porcess (annualized)	0.067	0.094 (with 0.067 in 95% C.I.)
		$M_t^{W13} = e^{\eta/12 - \gamma \log\left(C_{t+1}/C_t\right) + b(1-\beta)^{\Delta\lambda_t}/12}$		

each model is reported detailing its state variables, parameters and SDF functional form.[106] The last 2 columns display the parameters' values in the original calibration as well as the estimates for the current (main) sample of this study. Consumption is computed as in $CC99$ by the sum of non-durables and services, the state variables in $BY04$ are recovered using the procedure detailed in Bansal, Kiku, and Yaron (2011), and the time-varying disaster probabilities (intensities) are proxied using the monthly average daily $SP500$ crash probabilities as computed in Bollerslev and Todorov (2011) (BT).[107]

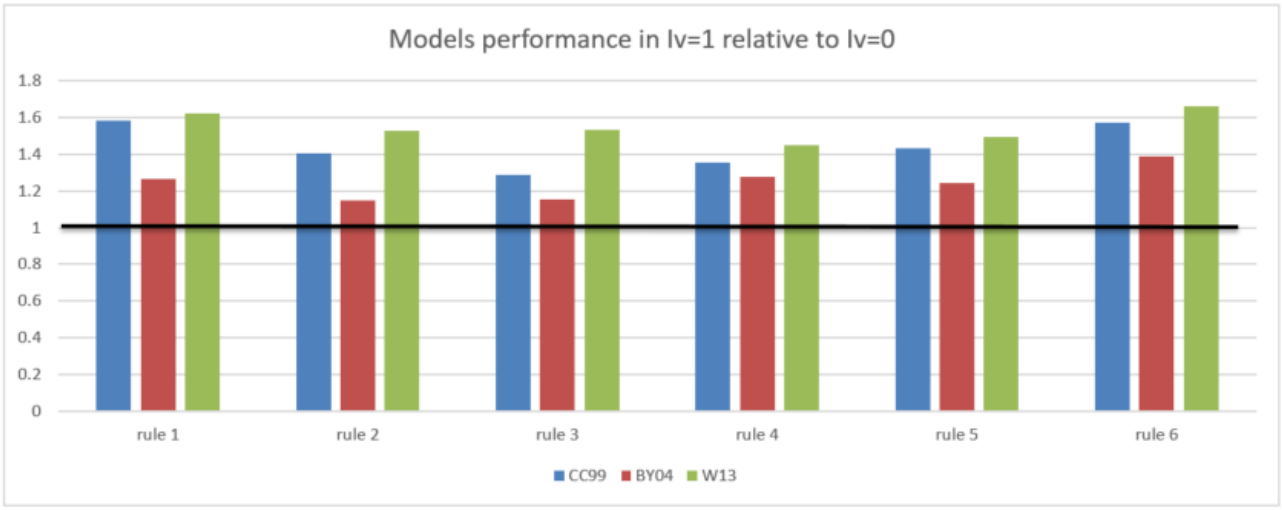
[106]The SDF formula for the case of $W13$ is derived using J. A. Wachter (2013) and the appendix of J. Wachter and Seo (2016).

[107]Because, as reported, the 95% confidence interval of such estimates contains the original values of the parameters when applied to consumption crashes via simulations in $W13$, for reasons already explained, I use the original values.

Having at disposal the SDFs time series I can compute the pricing errors for model i at time t as $PE_t^i \equiv M_t^i R_t - 1$. Specifically, I evaluate the performance of these models via the estimates of the following metric

$$|PE^i \ ratio| \equiv \frac{E[PE_{t+1}^i | I_t^v = 1]}{E[PE_{t+1}^i | I_t^v = 0]}$$

and report the result in the following graph



|PE ratio|	rule 1	rule 2	rule 3	rule 4	rule 5	rule 6	Avg.
CC99	1.583>1 at ***	1.403>1 at **	1.289>1 at *	1.353>1 at **	1.434>1 at **	1.575>1 at **	1.440
BY04	1.264>0 at **	1.147	1.154	1.278>1 at *	1.243>1 at *	1.390>1 at **	1.246
W13	1.620>0 at **	1.529>1 at **	1.535>1 at **	1.450>1 at **	1.494>1 at **	1.664>1 at **	1.549

for each of the six rules each model's $|PE \ ratio|$ is strictly grater than 1; The average absolute pricing errors during rejection periods, $I_t^v = 1$, are on average 41% higher than in the rest of the sample, with the average ratio being 44%,25% and 55% for $CC99, BT04$ and $W13$ respectively. The table immediately below the bar graph reveals how in most cases the

PE ratios are statistically grater than 1.[108] The displayed results use the real gross market return (deflated by the CPI), but are robust to the usage of the nominal gross market return as well.

Finally, an unreported analysis,[109] also rank these models in terms of their performance. The ranking is found independent of the conditioning, ($I_t^v = 1$ or $I_t^v = 0$), and tells us that the absolute pricing errors from $W13$ are on average statistically smaller than those produced by $CC99$ and $BY04$ which are on average indistinguishable.

[108] Standard errors are computed using the delta method.
[109] Available upon request.

Appendix B

Mean-Variance Portfolio Rebalancing with Transaction Costs

B.1 Existence of Solutions to the Mean-Variance Setup Under Transaction Costs

This appendix contains the proof of Theorem 2.

We start by assuming the cost function described in $A3$ to be such that $c(P, S, \theta^0) = c(\theta, \theta^0)$ and verify the claim later. Let us define

$$U^u \equiv r + \frac{1}{2(\kappa + \lambda)}(\kappa V\theta^B + \mu - r1)'V^{-1}(\kappa V\theta^B + \mu - r1) - \frac{\kappa}{2}\theta^{B\prime}V\theta^B$$

and

$$\gamma \equiv \theta - \theta^u$$

where θ^u is the ideal portfolio absent costs, defined in equation (2.2).

We observe that

$$U(\theta) = U(\gamma) = U^u - \frac{\kappa + \lambda}{2}\gamma' V \gamma - c(\gamma, \gamma^0)$$

where $c(\gamma, \gamma^0) = c(\theta, \theta^0)$ follows from the definition of γ and the fact that $c(P, S, \theta^0) = c(\theta, \theta^0)$. The problem class to solve is thus

$$\max_{\gamma \in \mathbb{R}^n} U(\gamma)$$

subject to

$$\gamma = \gamma^0 + P - S$$

$$P \geq 0$$

$$S \geq 0$$

Note that since the cost function is non-negative by construction $U(\theta) = U(\gamma) \leq U^u - \frac{\kappa+\lambda}{2}\gamma' V \gamma \equiv J(\gamma) = J(\theta) \leq J(\theta^u)$ therefore any solution θ^*, if it exists, is such that $U(\theta^*) \leq J(\theta^u)$.

From mathematical convenience the problem can be equivalently restated as

Problem 10

$$\underset{\gamma \in \mathbb{R}^n}{max}\ \tilde{U}(\gamma)$$

subject to

$$\gamma = \gamma^0 + P - S$$

$$P \geq 0$$

$$S \geq 0$$

where

$$\tilde{U}(\gamma) \equiv -\frac{\kappa+\lambda}{2}\gamma' V\gamma - c(\gamma,\gamma^0)$$

The aim of Theorem 2 is to show the existence of solutions to Problem 10. We proof this statement by first looking at four special sub-problems.

Sub-problem 1: Optimum Absent Costs

$$\max_{\gamma\in\mathbb{R}^n} \quad -\frac{\kappa+\lambda}{2}\gamma' V\gamma - K$$

This is the classical unconstrained mean-variance problem in which we subtract a constant, K, from the objective function. For this problem K can be thought as $c(\gamma,\gamma^0)$, where $c(\gamma,\gamma^0) = c(\theta,\theta^0) = c(0,\theta^0) = c(\theta^0) = c(0,0,\theta^0) = c(P,S,\theta^0)$ which is trivially linear in γ and θ. The unique solution, $\gamma = 0$, of this problem is the unconstrained optimum corresponding to $\theta = \theta^u$ and an optimal objective function value of $-K$. The result follows from the fact that the objective function is continuous, strictly concave and bounded above (since $\mathbf{V}$ is positive-definite), and $\gamma = 0$ is feasible.

The next couple of sub-problems benefit from the following lemma

Lemma 1 *Given the positive definite variance covariance matrix V, the matrix $\bar{I}'V\bar{I}$, with $\bar{I} \equiv [-I_n, I_n]$ where I_n is the $n \times n$ identity matrix, is positive semi-definite*

Proof. Since $\bar{I}'V\bar{I}$ is a square matrix we just need to show that its eigenvalues are non-negative. By definition any eigenvalue λ of $\bar{I}'V\bar{I}$ is such that $\bar{I}'V\bar{I}x = \lambda x$ where x is the associated non-zero eigenvector. Note that $\bar{I}'V\bar{I} = \begin{bmatrix} V & -V \\ -V & V \end{bmatrix}$. Then we can re-write the

eigenvalue representation of the matrix $\bar{I}'V\bar{I}$ as $\begin{bmatrix} V & -V \\ -V & V \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$. We now have two cases:

$\lambda = 0$ **case:** in this case any non-zero $x = x_1 = x_2$ is a solution

$\lambda \neq 0$ **case:** the solution is $x_1 = -x_2$ which implies $2Vx_1 = \lambda x_1$, but V is p.d. implying $0 < 2x_1'Vx_1 = \lambda x_1'x_1$ which is only possible if $\lambda > 0$

■

Sub-problem 2

$$\max_{P,S} \quad -\frac{\kappa + \lambda}{2}\gamma' V \gamma - P'C^P - S'C^S - K$$

subject to

$$\gamma = \gamma^0 + P - S$$

$$P \geq 0$$

$$S \geq 0$$

This problem is the proportional cost setup of Problem 3 in which a constant, K, is subtracted from the objective function: thus we already know (see Section 2.4.1) the existence of a unique solution $\theta = \theta + P - S$ such that for any risky asset i either $P_i \geq 0$ or $S_i \geq 0$ but $P_i > 0$ and $S_i > 0$ cannot happen. In this sub-problem $c(P, S, \theta^0) \equiv P'C^P + S'C^S + K$ which is linear in P and S, furthermore, due to the FOCs for every i, P_i and S_i are linear functions of θ_i trhough $\theta_i = \theta_i^0 + P_i 1_{[S_i=0]} + S_i 1_{[P_i=0]}$ thus $c(P, S, \theta^0) = c(P(\theta), S(\theta), \theta^0) = c(\theta, \theta^0)$ with the latter still being a linear function of its arguments.

The above problem can be more compactly re-written in the standard quadratic form

$$\max_{x} \quad \tilde{U}(x) = a'x + \frac{1}{2}x'Qx$$

subject to

$$x \geq 0$$

where $x' \equiv [P', S']$, $c' \equiv [C^{P'}, C^{S'}]$, $a' \equiv [c' - (\kappa + \lambda)\gamma^{0'}V\bar{I}]$ and $Q \equiv (\kappa + \lambda)\bar{I}'V\bar{I}$. From lemma 1, Q is positive semi-definite thus, from standard quadratic programing theory,[110] any x satisfying the FOCs is a solution. As shown in Section 2.5, such x, for any arbitrary $n \in \mathbb{N}$ solves $LCP(a, Q)$. Furthermore, from Section 2.4.1 we know such x always exists and it is unique so that the algorithm of Section 2.5 will provide it.

Sub-problem 3

$$\max_{P,S} \quad -\frac{\kappa + \lambda}{2}\gamma'V\gamma - P'C^P - S'C^S - (\bar{K}'D_j)1$$

subject to

$$\gamma = \gamma^0 + P - S$$

$$D_j P = 0$$

$$D_j S = 0$$

$$P \geq 0$$

$$S \geq 0$$

[110] See for e.g. Cottle et al. (1992).

where $\bar{K}' = [k_1, ..., k_n]$, $D_j \equiv \begin{bmatrix} \mathbf{1_j} & 0 \\ 0 & \mathbf{0_{-j}} \end{bmatrix}$ with $\mathbf{1_j}$ being a $1 \times j > 0$ vector of ones and $\mathbf{0_{-j}}$ being a $1 \times (n-j)$ vector of zeros; $\mathbf{1_j}$ and $\mathbf{0_{-j}}$ together form the main diagonal of D_j. The vector $\mathbf{1_j}$ contains the j assets which are constrained not to be traded while $\mathbf{0_{-j}}$ contains the $(n-j)$ assets that are allowed to be traded in this sub-problem; Note that setting $j = 0$ reduces the setup to the one of the previous sub-problem.

Because the domain of this sub-problem is a linear sub-space of that of sub-problem 2, given $\gamma' = [\gamma'_j = \gamma_j^{0\prime}, \gamma_{-j}]$, γ_{-j} is the optimal solution of sub-problem 2 restricted to the $(n-j)$ traded assets with $K \equiv (\bar{K}' D_j)1$.

Sub-problem 4

$$\max_{\gamma_{-j} \in \mathbb{R}^{n-j}} \quad -\frac{\kappa + \lambda}{2} \gamma' V \gamma - (\bar{K}' D_j)1$$

where $\gamma' = [\gamma'_j = \gamma_j^{0\prime}, \gamma_{-j}]$ and we solve for the optimal weights difference γ_{-j} corresponding to the subset of assets $i \in \{j+1, ..., n\}$ that are allowed to trade. The cost function for this sub-problem is $c(P, S, \theta^0) = (\bar{K}' D_j)1 \equiv K$ which is the same as the one discussed in sub-problem 1. Thus $c(P, S, \theta^0) = c(\theta, \theta^0) = c(\gamma, \gamma^0)$ and it is trivially linear in all its arguments. In terms of solutions we note that by decomposing the variance-covariance matrix V as $V \equiv \begin{bmatrix} V_j & V_{j,-j} \\ V'_{j,-j} & V_{-j} \end{bmatrix}$ we can solve the equivalent problem instead

$$\min_{\gamma_{-j} \in \mathbb{R}^{n-j}} \quad (\kappa + \lambda) \gamma_j^{0\prime} V_{j,-j} \gamma_{-j} + \frac{\kappa + \lambda}{2} \gamma'_{-j} V_{-j} \gamma_{-j}$$

since V is positive definite, also V_{-j} is positive definite. Thus this is an unconstrained minimization problem with a strictly concave continuous and differentiable objective function which unique (feasible) solution is $\gamma^*_{-j} = (V_{-j})^{-1} V'_{j,-j} \gamma^0_j$.

In summary, sub-problems 1-4 admit a unique fisible solution θ^*, or the equivalent characterization γ^* in the space of weights difference. Moreover the cost function $c(P, S, \theta^0) = c(\theta, \theta^0) = c(\gamma, \gamma^0)$ and it is linear in its arguments.

Putting the pieces together we can now proof the existence of solutions γ^* for Problem 10 which is equivalent to the existence of solutions θ^* for the class of problems having objective function 2.1 under assumption $A1 - A3$. By doing so we also provide the algorithm to solve the general problem. We proof Theorem 2 case-wise:

Proportional costs: Look at Problem 10 with $c(\gamma, \gamma^0) = P'C^P + S'C^S$: this is sub-problem 3 with $K = 0$ thus we know that for this case there exist a unique feasible solution, γ^*, or $\theta^{\setminus PC}$ in the space of weights θ, easily computable for large n via the algorithm presented in Section 2.5.

Overall fixed cost: Look at Problem 10 with $c(\gamma, \gamma^0) = 1_{[\gamma \neq \gamma^0]} k$ and $P = S = 0$ and solve for γ: only two things can happen:

1. The investor does not trade: $\gamma = \gamma^0$, $c(\gamma, \gamma^0) = 0$. This is feasible and yields an utility of $\tilde{U}(\gamma^0) = -\frac{\kappa+\lambda}{2}\gamma^{0\prime} V \gamma^0$
2. The investor trades to the unconstrained optimum: $\gamma = \gamma^u = 0$, $c(\gamma, \gamma^0) = k > 0$. This (feasible) strategy yields an utility of $\tilde{U}(\gamma^u) = -k$ and corresponds to the solution of sub-problem 1

The solution to the overall fixed cost problem is

$$\gamma^* = argmax(\tilde{U}(\gamma^0), \tilde{U}(\gamma^u))$$

or, in the space of weights θ

$$\theta^{\backslash PC} = \gamma^* + \theta^u$$

such solution exists because the arguments of the $argmax$ function are well-defined.

Asset-specific fixed costs: Look at Problem 10 with $c(\gamma, \gamma^0) = \sum_i^n 1_{[\gamma_i \neq \gamma_i^0]} k_i$, $P = S = 0$ and solve for γ: a *finite* number of cases might occur:

1. The investor does not trade: $\gamma = \gamma^0$, $c(\gamma, \gamma^0) = 0$ which is (trivially) linear in γ. This is feasible and yields an utility of $\tilde{U}(\gamma^0) = -\frac{\kappa+\lambda}{2}\gamma^{0\prime} V \gamma^0$
2. The investor trades in a subset of cardinality $(n-j) > 0$ of assets: This is sub-problem 4 for which we know that the unique (feasible) optimal solution is $\gamma^{j*} \equiv \begin{bmatrix} \gamma_j^0 \\ (V_{-j})^{-1} V'_{j,-j} \gamma_j^0 \end{bmatrix}$ and the optimal value is $\tilde{U}(\gamma^{j*}) = -\frac{\kappa+\lambda}{2}\gamma^{j*\prime} V \gamma^{j*} - (\bar{K}' D_j)1$. Note that there are $N = \sum_{j=1}^{n-1} \binom{n}{j}$ different subsets j_k with $k = 1, ..., N$, so that we have that many sub-problems 4 to solve, each one yielding solution γ^{j_k*}.
3. The investor trades in all assets: this is sub-problem 1. Thus the unique (feasible) optimum is $\gamma = \gamma^u = 0$, yielding an utility of $\tilde{U}(\gamma^u) = -k$.

The solution to the asset-specific fixed costs case is

$$\gamma^* = argmax(\tilde{U}(\gamma^0), \tilde{U}(\gamma^{j_1*}), ..., \tilde{U}(\gamma^{j_N*}), \tilde{U}(\gamma^u))$$

or, in the space of weights θ

$$\theta^{\backslash PC} = \gamma^* + \theta^u$$

which exists because the arguments of the $argmax$ function are well-defined. The computation of such solution is practically feasible only for relatively small number

of risky assets: with 10 assets there are 1022 sub-problems, with 20 already 1049574 many, while with 30 a total of 536870910. In general the number of sub-problems grows at a rate of 2^n with the n number of risky assets. This type of problem does not scale up as nicely as the previous ones. Yet, it nonetheless represents a big improvement over the current state-of-the-art in the literature where this kind of setups are only solvable numerically with $n \leq 3$.

Proportional and overall fixed costs: Look at Problem 10 with $c(\gamma, \gamma^0) = P'C^P + S'C^S + 1_{[\gamma \neq \gamma^0]}k$, solve for P and S to get to $\gamma = \gamma^0 + P - S$: only two things can happen:

1. The investor does not trade: $\gamma = \gamma^0$, $P = S = 0$ and $c(\gamma, \gamma^0) = 0$ which is (trivially) linear in γ. This is feasible and yields an utility of $\tilde{U}(\gamma^0) = -\frac{\kappa+\lambda}{2}\gamma^{0\prime}V\gamma^0$
2. The investor trades: $\gamma \neq \gamma^0$, this is sub-problem 2 with $K = k$. The solution γ^T exists and it is unique, and is computable through the algorithm of Section 2.5. The cost function is linear in γ (and θ).

The solution to the proportional and overall fixed costs case is

$$\gamma^* = argmax(\tilde{U}(\gamma^0), \tilde{U}(\gamma^T))$$

or, in the space of weights θ

$$\theta^{\setminus PC} = \{\theta^0, \theta^{PC}\}$$

with $\theta^{PC} = \gamma^T + \theta^u$. This solution exists because the arguments of the $argmax$ function are well-defined. Because the overall fixed cost applies to all assets we avoid the combinatorial issues arising with asset specific fix cost and our setup can thus deliver the solution for large number of risky assets n.

Proportional and asset-specific fixed costs: Look at Problem 10 with $c(\gamma, \gamma^0) = P'C^P + S'C^S + \sum_i^n 1_{[\gamma_i \neq \gamma_i^0]} k_i$, solve for P and S to get to $\gamma = \gamma^0 + P - S$: a *finite* number of cases might occur:

1. The investor does not trade: $\gamma = \gamma^0$, $P = S = 0$ and $c(\gamma, \gamma^0) = 0$ which is (trivially) linear in γ. This is feasible and yields an utility of $\tilde{U}(\gamma^0) = -\frac{\kappa+\lambda}{2}\gamma^{0\prime} V \gamma^0$
2. The investor trades in a subset of cardinality $(n - j) > 0$ of assets: this is sub-problem 3 for which we know that the unique (feasible) optimal solution is γ^{j*}, , or θ^{PC} in the space of weights θ, computable with the aid of Section 2.5 algorithm, the cost function is linear in γ and the optimal value is $\tilde{U}(\gamma^{j*})$. Note that there are $N = \sum_{j=1}^{n-1} \binom{n}{j}$ different subsets j_k with $k = 1, ..., N$, so that we have that many sub-problems 3 to solve, each one yielding solution γ^{j_k*} (respectively $\theta^{PC_{j_k*}}$).
3. The investor trades in all assets: this is sub-problem 2 with $K \equiv \sum_i^n k_i$: thus the unique (feasible) optimum is γ^{TAll}, yielding an utility of $\tilde{U}(\gamma^{TAll})$ with a linear cost function.

The solution to the proportional and asset-specific fixed costs case is

$$\gamma^* = argmax(\tilde{U}(\gamma^0), \tilde{U}(\gamma^{j_1*}), ..., \tilde{U}(\gamma^{j_N*}), \tilde{U}(\gamma^{TAll}))$$

or, in the space of weights θ

$$\theta^{\backslash PC} = \{\theta^0, \theta^{PC_{j_1*}}, ..., \theta^{PC_{j_N*}}, \theta^{PC_{TAll}}\}$$

with $\theta^{PC_{(.)}} = \gamma^{(.)} + \theta^u$. The solution exists because the arguments of the $argmax$ function are well-defined. Since this problem involves asset specific fixed costs, as

shown above, combinatorial issues limit the actual number of risky assets for which the problem is computationally feasible.

Appendix C

Importance of Transaction Costs for Asset Allocations in FX Markets

C.1 Details on Portfolio Optimization Problem

C.1.1 Characterization of the No Trading Region

In the main text we provide a graphical visualization of the no trading region of the optimal trading strategy for the case of 2 risky assets when $\mathbf{C_t^{P+}} = \mathbf{C_t^{P-}} = \mathbf{C_t^{S+}} = \mathbf{C_t^{S-}}$ (Figure 3.2). Figure C.1 generalizes the cost structure in this illustration. From left to right Figure C.1 illustrates the no trading regions in the case of asset correlations equal to (1) $\rho = 0.5$, (2) $\rho = 0$, and (3) $\rho = -0.5$. We choose the other parameters of the investment opportunity set such that the 2 risky assets match the mean values of our full set of 29 currencies from 1976 to 2016. In particular, we set $\mu_{\mathbf{t}}^{\mathbf{e}} = 2.4\%$, $\sigma_{\mathbf{t}} = 10\%$ (the diagonal elements of $\mathbf{V_t}$), $\mathbf{C_t^{P+}} = 1.45\%$ for the costs of increasing long positions, $\mathbf{C_t^{S+}} = 0.71\%$ for the costs of decreasing long positions, $\mathbf{C_t^{P-}} = 0.71\%$ for the costs of reducing short positions, and

$\mathbf{C_t^{S-}} = 1.45\%$ for the costs of increasing short positions.We set the coefficient of risk aversion $\lambda = 5$.

The no trading regions are described as follows: blue for MV_{TC}, red for $MV_{TC\setminus Corr}$, black for MV_{TC} if $\mathbf{C_t^{P+}} = \mathbf{C_t^{P-}} = 1.45\%$ and $\mathbf{C_t^{S+}} = \mathbf{C_t^{S-}} = 0.71\%$, and yellow MV_{TC} if $\mathbf{C_t^{P+}} = \mathbf{C_t^{P-}} = 0.71\%$ and $\mathbf{C_t^{S+}} = \mathbf{C_t^{S-}} = 1.45\%$. The arrows indicate the optimal actions $\mathbf{\Delta_t^{P+}}$, $\mathbf{\Delta_t^{S+}}$, $\mathbf{\Delta_t^{P-}}$, $\mathbf{\Delta_t^{S-}}$ from any initial position $\theta_\mathbf{t}^\mathbf{0}$ outside the blue no trading region.

Under the black no trading region parallelogram (MV_{TC}), it is optimal to either not trade at all (if the initial position is inside the parallelogram), trade only 1 asset at a time, along vertical or horizontal straight lines up to the closest edge of the parallelogram, or trade in both assets in the regions beyond the corners and outside the parallelogram up to the closest corner. Under MV_{TC} with $\mathbf{C_t^{P+}} = \mathbf{C_t^{S-}} = 1.45\%$ and $\mathbf{C_t^{S+}} = \mathbf{C_t^{P-}} = 0.71\%$ (i.e. the blue no trading region), the closest to the actual MV_{TC} performed in the data, the same trading behavior occurs most of the time. The two no trading regions and optimal trading activity only differ for the case of $\rho = 0.5$ in the neighborhood of the lower right and upper-left corners of the black parallelogram, where the borders of the blue no trading region are horizontal and vertical respectively, with trades in only one asset at a time proceeding vertically or horizontally inside the area of the black no trading region. These two additional edges of the blue no trading region are on a horizontal respectively vertical line that passes through the origin.

Consider for instance an initial position $\theta_\mathbf{t}^\mathbf{0}$ to the left of the vertical line that passes through the origin. The initial weight of asset 1 is negative and below the optimal level. The investor would like to increase her position in this asset, ideally moving horizontally all the way to the nearest edge of the yellow parallelogram, given the costs $\mathbf{C_t^{P-}} = 0.71\%$. But as soon as the weight becomes positive the no trading region switches to the black parallelogram with

the higher costs $\mathbf{C_t^{P+}} = 1.45\%$. In other words, given the costs increase to $\mathbf{C_t^{P+}} = 1.45\%$ when the position on asset 1 switches from negative to positive but the marginal benefit of moving closer to $\theta^{\mathbf{MV}}$ remains unchanged, the investor stops trading earlier than what she has originally intended when facing the the lower costs $\mathbf{C_t^{P-}} = 0.71\%$.

Because the all trading regions in case (2) with $\rho = 0$ and case (3) with $\rho = -0.5$ lie in the positive quadrant, the blue and the black no trading regions coincide while the yellow is shifted towards the upper-right corner. Finally, in case (2) with $\rho = 0$ the red and the blue no trading regions coincide because they are solving the same problem.

C.1.2 Algorithms

Problem 9

Following the solution approach of Dybvig and Pezzo (2018), we can rewrite Problem 9 as a standard quadratic program of the form

$$\min_{\mathbf{x}} \quad \mathbf{q}'\mathbf{x} + \frac{1}{2}\mathbf{x}'\mathbf{H}\mathbf{x}$$

subject to

$$\underline{\mathbf{x}} \leq \mathbf{x} \leq \overline{\mathbf{x}}$$

where $\mathbf{q}' \equiv \mathbf{b}' + \lambda\hat{\theta}_{\mathbf{t}}^{\mathbf{0}\prime}\mathbf{Q}\bar{\bar{\mathbf{I}}} - \mu_{\mathbf{t}}^{\mathbf{e}}\bar{\mathbf{I}}\bar{\bar{\mathbf{I}}}$ with $\mathbf{b}' \equiv [\mathbf{C}_t^{P+\prime}, \mathbf{C}_t^{P-\prime}, \mathbf{C}_t^{S+\prime}, \mathbf{C}_t^{S-\prime}]$, $\hat{\theta}_{\mathbf{t}}^{\mathbf{0}\prime} \equiv [\theta_{\mathbf{t}}^{\mathbf{0}+\prime}, \theta_{\mathbf{t}}^{\mathbf{0}-\prime}]$, $\mathbf{Q} \equiv \bar{\mathbf{I}}'\mathbf{V}_t\bar{\mathbf{I}}$, $\bar{\mathbf{I}} \equiv [I_n, I_n]$, $\bar{\bar{\mathbf{I}}} \equiv [I_{2n}, -I_{2n}]$ and I_a is the $a \times a$ identity matrix, and the hessian $\mathbf{H}$ is given by $\lambda\bar{\bar{\mathbf{I}}}'\mathbf{Q}\bar{\bar{\mathbf{I}}}$. The program returns the solution $\mathbf{x} \equiv [\mathbf{\Delta_t^{P+\prime}}, \mathbf{\Delta_t^{P-\prime}}, \mathbf{\Delta_t^{S+\prime}}, \mathbf{\Delta_t^{S-\prime}}]$ from

which the optimal portfolio $\theta_t^{MV_{TC}}$ is obtained by

$$\theta_t^{MV_{TC}} = \theta_{\mathbf{t}}^{\mathbf{0}} + \bar{\bar{\mathbf{I}}}\mathbf{x}.$$

We solve this strictly convex quadratic program using the Matlab Optimization ToolBox.

Simplifying Problem 9: Uncorrelated Assets

H. Liu (2004) suggests that the assumption of uncorrelated assets greatly reduces the complexity to optimize a portfolio subject to transaction costs. This is because with uncorrelated assets we can solve N independent problems each one associated with only one asset. We continue to use the true correlation matrix to compute $\theta_{\mathbf{t}}^{\mathbf{MV}}$ but impose the assumption of uncorrelated assets when we construct the no trading region surrounding $\theta_{\mathbf{t}}^{\mathbf{MV}}$.

We proceed in two steps. First, we solve two sub-problems. The frist one assumes that the costs of opening new long or closing existing short positions both are $\mathbf{C_t^{P,1}} \equiv \mathbf{C_t^{P-}}$, and closing existing long or opening new short positions both are $\mathbf{C_t^{S,1}} \equiv \mathbf{C_t^{S+}}$. The second sub-problem assumes that the costs of opening new long or closing existing short positions both are $\mathbf{C_t^{P,2}} \equiv \mathbf{C_t^{P+}}$, and closing existing long or opening new short positions both are $\mathbf{C_t^{S,2}} \equiv \mathbf{C_t^{S-}}$. Both sub-problems ignore correlations between assets when we construct the no trading region around $\theta_{\mathbf{t}}^{\mathbf{MV}}$. Given $\mu_{\mathbf{t}}^{\mathbf{e}}$, $\mathbf{V_t}$ and $\theta_{\mathbf{t}}^{\mathbf{0}}$ and the generic costs $\mathbf{C_t^{P,j}}$ and $\mathbf{C_t^{S,j}}$ for sub-problem $j \in \{1, 2\}$, the First Order Conditions (FOCs) are DyPe2018

$$\underline{\theta}_{\mathbf{t}}^{(\mathbf{j})} \equiv \frac{\mathbf{V_t^{-1}}}{\lambda}(\mu_{\mathbf{t}}^{\mathbf{e}} - \mathbf{C_t^{P,j}}) \leq \theta_{\mathbf{t}}^{(\mathbf{j})} \leq \frac{\mathbf{V_t^{-1}}}{\lambda}(\mu_{\mathbf{t}}^{\mathbf{e}} + \mathbf{C_t^{S,j}}) \equiv \overline{\theta}_{\mathbf{t}}^{(\mathbf{j})}.$$

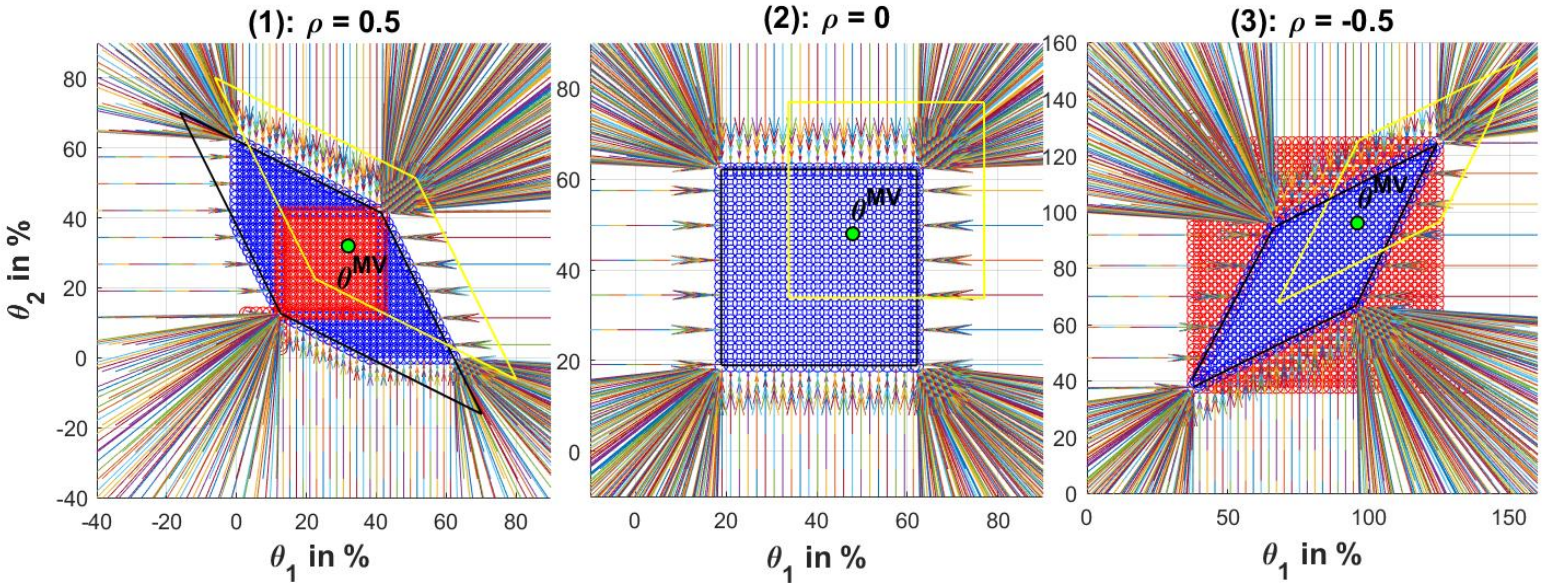
When the correlations between assets are ignored, the solution is given for each asset i by

$$\theta_{\mathbf{i,t}}^{(\mathbf{j})} = \begin{cases} \theta_{\mathbf{i,t}}^{\mathbf{0}} & \text{if} \quad \underline{\theta}_{\mathbf{i,t}}^{(\mathbf{j})} \leq \theta_{\mathbf{i,t}}^{\mathbf{0}} \leq \overline{\theta}_{\mathbf{i,t}}^{(\mathbf{j})} \\ \overline{\theta}_{\mathbf{i,t}}^{(\mathbf{j})} & \text{if} \quad \theta_{\mathbf{i,t}}^{\mathbf{0}} > \overline{\theta}_{\mathbf{i,t}}^{(\mathbf{j})} \\ \underline{\theta}_{\mathbf{i,t}}^{(\mathbf{j})} & \text{if} \quad \theta_{\mathbf{i,t}}^{\mathbf{0}} < \underline{\theta}_{\mathbf{i,t}}^{(\mathbf{j})}. \end{cases}$$

In the second step, conditional on the initial position $\theta_{\mathbf{i,t}}^{\mathbf{0}}$ we decide for each asset i which of the two sub-problem solutions $\theta_{\mathbf{i,t}}^{(\mathbf{1})}$ or $\theta_{\mathbf{i,t}}^{(\mathbf{2})}$ is the correct solution for $\theta_{\mathbf{i,t}}^{\mathbf{MV_{TC \setminus Corr}}}$. Because (in the data) $\mathbf{C}_{\mathbf{i,t}}^{\mathbf{S}+} \leq \mathbf{C}_{\mathbf{i,t}}^{\mathbf{S}-}$ and $\mathbf{C}_{\mathbf{i,t}}^{\mathbf{P}-} \leq \mathbf{C}_{\mathbf{i,t}}^{\mathbf{P}+}$, it follows that if $\theta_{\mathbf{i,t}}^{\mathbf{0}} > 0$ and $\theta_{\mathbf{i,t}}^{(\mathbf{1})} \geq 0$ or $\theta_{\mathbf{i,t}}^{\mathbf{0}} < 0$ and $\theta_{\mathbf{i,t}}^{(\mathbf{1})} \leq 0$ then $\theta_{\mathbf{i,t}}^{\mathbf{MV_{TC \setminus Corr}}} = \theta_{\mathbf{i,t}}^{(\mathbf{1})}$, otherwise $\theta_{\mathbf{i,t}}^{\mathbf{MV_{TC \setminus Corr}}} = \theta_{\mathbf{i,t}}^{(\mathbf{2})}$.

Figure C.1: No Trading Regions: General Cost Structures and Correlations between Assets

No trading regions for setting of two risky assets with correlations (1) $\rho = 0.5$, (2) $\rho = 0$, and (3) $\rho = -0.5$, and $\mu_{\mathbf{t}}^{\mathbf{e}} = 2.4\%$, $\sigma_{\mathbf{t}} = 10\%$ (the diagonal elements of $\mathbf{V_t}$), $\mathbf{C_t^{P+}} = 1.45\%$, $\mathbf{C_t^{S+}} = 0.71\%$, $\mathbf{C_t^{P-}} = 0.71\%$, $\mathbf{C_t^{S-}} = 1.45\%$, and $\lambda = 5$. The no trading regions are: blue for MV_{TC}, red for $MV_{TC\backslash Corr}$, black for MV_{TC} if $\mathbf{C_t^{P+}} = \mathbf{C_t^{P-}} = 1.45\%$ and $\mathbf{C_t^{S+}} = \mathbf{C_t^{S-}} = 0.71\%$, and yellow MV_{TC} if $\mathbf{C_t^{P+}} = \mathbf{C_t^{P-}} = 0.71\%$ and $\mathbf{C_t^{S+}} = \mathbf{C_t^{S-}} = 1.45\%$. The arrows indicate the optimal actions $\mathbf{\Delta_t^{P+}}$, $\mathbf{\Delta_t^{S+}}$, $\mathbf{\Delta_t^{P-}}$, $\mathbf{\Delta_t^{S-}}$ from any initial position $\theta_{\mathbf{t}}^{\mathbf{0}}$ outside the blue no trading region.



C.2 Data Sources: Spot and Forward Exchange Rates

In Table C.1 we list the Datastream mnemonics for spot and forward exchange rate quotes against the GBP, whereas those against the USD are listed in Table C.2. To obtain mid-, bid- and ask-exchange rates, the suffixes (ER), (EB) and (EO) are added to the corresponding mnemonics.

Table C.1: Datastream mnemonics for currency quotes against the British pound

Currency	Spot rate	Forward rate	Quote convention
Canadian dollar	CNDOLLR	CNDOL1F	FCU/GBP
Danish krone	DANISHK	DANIS1F	FCU/GBP
French franc	FRENFRA	FRENF1F	FCU/GBP
German mark	DMARKER	DMARK1F	FCU/GBP
Irish punt	IPUNTER	IPUNT1F	FCU/GBP
Italian lira	ITALIRE	ITALY1F	FCU/GBP
Japanese yen	JAPAYEN	JAPYN1F	FCU/GBP
Netherlands guilder	GUILDER	GUILD1F	FCU/GBP
Norwegian krone	NORKRON	NORKN1F	FCU/GBP
Portuguese escudo	PORTESC	PORTS1F	FCU/GBP
Spanish peseta	SPANPES	SPANP1F	FCU/GBP
Swedish krona	SWEKRON	SWEDK1F	FCU/GBP
Swiss franc	SWISSFR	SWISF1F	FCU/GBP
U.S. dollar	USDOLLR	USDOL1F	FCU/GBP

Table C.2: Datastream mnemonics for currency quotes against the U.S. dollar

Currency	Spot rate	Forward rate	Quote convention
Australian dollar	BBAUDSP	BBAUD1F	FCU/USD
Brazilian real	BRACRU$	USBRL1F	FCU/USD
British pound	BBGBPSP	BBGBP1F	USD/FCU
Canadian dollar	BBCADSP	BBCAD1F	FCU/USD
Czech koruna	CZECHC$	USCZK1F	FCU/USD
Danish krone	BBDKKSP	BBDKK1F	FCU/USD
Euro	BBEURSP	BBEUR1F	FCU/USD
French franc	BBFRFSP	BBFRF1F	FCU/USD
German mark	BBDEMSP	BBDEM1F	FCU/USD
Greek Drachma	GREDRA$	USGRD1F	FCU/USD
Hungarian forint	HUNFOR$	USHUF1F	FCU/USD
Icelandic krona	ICEKRO$	USISK1F	FCU/USD
Irish punt	BBIEPSP	BBIEP1F	USD/FCU
Italian lira	BBITLSP	BBITL1F	FCU/USD
Japanese yen	BBJPYSP	BBJPY1F	FCU/USD
Mexican peso	MEXPES$	USMXN1F	FCU/USD
Netherland guilder	BBNLGSP	BBNLG1F	FCU/USD
New Zealand dollar	BBNZDSP	BBNZD1F	FCU/USD
Norwegian krone	BBNOKSP	BBNOK1F	FCU/USD
Polish zloty	POLZLO$	USPLN1F	FCU/USD
Portuguese escudo	PORTES$	USPTE1F	FCU/USD
Singapore dollar	BBSGDSP	BBSGD1F	FCU/USD
South Africa rand	BBZARSP	BBZAR1F	FCU/USD
South Korean won	KORSWO$	USKRW1F	FCU/USD
Spanish peseta	SPANPE$	USESP1F	FCU/USD
Swedish krona	BBSEKSP	BBSEK1F	FCU/USD
Swiss france	BBCHFSP	BBCHF1F	FCU/USD
Taiwan new dollar	TAIWDO$	USTWD1F	FCU/USD

On the Wedge between Theoretical and Actual Prices and its Implications for Investment Decisions, Pezzo, PhD in Business Administration 2018