Distinguishing skills from luck in trading and investment

Tren Ma

Submitted in fulfilment of the requirements
for the Degree of Doctor of Philosophy

Adam Smith Business School

College of Social Sciences

University of Glasgow

June 2023

Abstract

This thesis develops a series of statistical frameworks to control type I errors in picking out-performers. It consists of three independent essays which assess performance of U.S. equity mutual funds, currency trading strategies and hedge funds.

The first essay introduces a novel multiple hypothesis testing method named the functional False Discovery Rate "plus" ($fFDR^{+}$). The method incorporates informative covariates in estimating the False Discovery Rate (FDR) of predictive models' "conditional" performance. In simulations, the $fFDR^{+}$ controls well the FDR and gains considerable power over prior methods that do not account for extra information. In empirical analyses, we construct portfolios based on several covariates and show that they enhance the performance of mutual fund portfolios, highlighting the value of extra information in the multiple hypothesis testing framework.

The second essay develops the multivariate functional false discovery rate ($mfFDR$) method that accounts for multiple informative covariates to examine the conditional performance of predictive models and gain a considerably higher power than prior methods including the one in the first essay. The proposed method is then applied to control luck in detecting profitable technical trading rules using 30 developed and emerging market currencies. It selects more profitable rules than prior methods; more importantly, these rules offer better out-of-sample performance.

The third essay introduces a new procedure to control for family error rate (FWER) in picking out-performers. The method utilizes multiple side information to more precisely estimate the FWER and gains much higher power in detecting out-performers compared to existing ones. In empirical analyses, the method allows investors picking out-performing hedge funds with very low FWER. The portfolios of hedge funds selected by the method beat passive benchmarks in various settings. Further analyses show that the new method detects truly out-performing hedge fund managers who can repeat their past performance over a long horizon.

Table of Contents

Abstract **i**

Table of Contents **ii**

Acknowledgements **xi**

Declaration **xii**

Abbreviations **xiii**

Introduction **1**

1 Functional false discovery rate in mutual fund selection **5**

1.1 Introduction 5

1.2 Methods for controlling of luck with informative covariate 9

1.2.1 Functional false discovery rate ($fFDR$) 9

1.2.2 The $fFDR^+$: application in selecting out-performing funds . . . 11

1.3 Data 16

1.4 Simulation setup 17

1.4.1 The model 17

1.4.2 The distribution of fund alphas 18

1.4.3 Simulation execution 20

1.5 Analysis of $fFDR^+$ and power 21

1.5.1 False discovery rate control of $fFDR^+$ 22

1.5.2 Power analysis 25

1.5.3 Power and FDR trade-off 27

1.5.4 Varying the number of observations and funds 28

1.5.5 Estimation errors in the covariate 30
1.6 Empirical results 32
1.6.1 Five covariates proposed in the literature 32
1.6.2 The FDR^{+} and $fFDR^{+}$ portfolios 34
1.6.3 Performance comparison 36
1.7 Concluding remarks 42

2 Controlling for luck in picking trading strategies 45

2.1 Introduction 45
2.2 The use of covariates in FDR framework 49
2.2.1 The multivariate functional false discovery rate ($mfFDR$) . . . 50
2.2.2 Simulation studies 52
2.2.3 Correlation and estimation errors of covariates 61
2.2.4 Weak dependence in p-values 63
2.3 Data and strategy universe 65
2.3.1 Data 65
2.3.2 Trading rule universe 67
2.4 Measures of predictive ability and profitability 68
2.5 Empirical results 71
2.5.1 Covariates 71
2.5.2 Individual Currencies 73
2.5.3 Baskets of Currencies 78
2.6 Concluding remarks 85

3 Picking hedge fund with high confidence 87

3.1 Introduction 87
3.2 FWER and informative covariates 90
3.2.1 The use of informative covariates in controlling FWER 91
3.2.2 Application in picking outperforming funds 93
3.3 Data 93
3.4 Performance measure 94
3.5 Simulation studies 95
3.5.1 A comparison to existing methods 98

3.5.2 Performance of the $fwer^+$ under varying signals 99
3.5.3 Performance of the $fwer^+$ under insufficient, noisy and uninformative covariates 101
3.5.4 Performance of the $fwer^+$ under varying of sample size and observations 102
3.6 Empirical analysis 104
3.6.1 Covariates 105
3.6.2 Portfolios of out-performing funds 106
3.6.3 Persistent analysis 109
3.6.4 Sub-sample analysis 110
3.6.5 Boosting the informativeness of covariates 112
3.6.6 Alternative choices of benchmarks 113
3.6.7 Portfolios of the best out-performing hedge fund 114
3.7 Concluding remarks 116

Conclusion **118**

A Appendix for chapter 1 **130**
A.1 Estimating $\pi_0(z)$ and $f(p, z)$ 130
A.2 Additional simulation results 131
A.2.1 Results for balanced panel data under cross-sectional dependence 131
A.2.2 Results for unbalanced panel data 136
A.2.3 Simulation results for single normal distribution 141
A.3 Results for data sample period from 1984 144
A.4 A comparison of portfolios' trading metrics 145
A.5 Performance of $fFDR10\%$ in various periods 146
A.6 The construction of sorting portfolios 147
A.7 Wealth evolution 148
A.8 Results for alternative targets of FDR 149
A.9 Results from using an alternative proxy of covariates 149
A.10 Covariate combinations 150
A.11 Restricted data 152

B Appendix for chapter 2 **154**

B.1 The multivariate functional false discovery rate 154

B.2 Performance of $mfFDR$ under noisy covariates 156

B.3 Performance of $mfFDR$ under varying number of tests 159

B.4 Performance of $mfFDR^+$ portfolios with various FDR targets 161

B.5 Performance of $mfFDR^+$ portfolios with use of mean excess return as the testing performance metric 163

C Appendix for chapter 3 **166**

C.1 The implementation of the StepM and StepSPA procedures 166

C.2 Additional simulation studies 169

C.3 Empirical results with restriction on AUM 170

C.4 $fwer^+$ portfolios with use of individual covariates 171

C.5 Alternative choices of in-sample horizons 173

C.6 Performance of $fwer^+$ portfolios with use of simple $p-$values 174

List of Tables

1.1 Performance comparison in terms of power 26
1.2 Power comparison for varying FDR targets 28
1.3 Power comparison for varying sample size and observation length . . . 29
1.4 Power comparison for sample with small size and small number of observations 30
1.5 Power performance under noisy covariates 31
1.6 Comparison of portfolios' performances for varying time lengths of investing 38
1.7 Performance of portfolios in sub-periods 41
1.8 Performance comparison of $fFDR\tau$ portfolios and sorting based portfolios 42

2.1 Summary Statistics 66
2.2 Summary performance of trading rules in the whole sample period . . . 70
2.3 The pairwise correlation of covariates 74
2.4 Empirical power comparison 75
2.5 Sharpe ratios of portfolios before transaction costs 76
2.6 Net Sharpe ratios of portfolios 77
2.7 Correlation coefficients of covariate pairs: all currencies 79
2.8 Performance of the $mfFDR^{+}$ portfolios 80
2.9 Out-performing rate by category 81
2.10 Distribution by currency of selected trading rules in sub-periods 82
2.11 Performance of the $mfFDR^{+}$ portfolios implemented on each category 84

3.1 OOS performance of $fwer^{+}$ portfolios 108
3.2 Performance of $fwer^{+}$ in various OOS horizons. 110
3.3 OOS performance of $fwer^{+}$ portfolios in sub-samples 111
3.4 OOS performance of $fwer^{+}$ portfolios with use of new covariates 113

3.5 Performance under alternative benchmarks 114
3.6 Performance of the single-fund portfolios 115

A.1 Power comparison for discrete distribution 133
A.2 Power comparison for discrete-normal distribution mixture 134
A.3 Power comparison for mixture of two normal distributions 134
A.4 Power comparison for varying sample size and observation length . . . 135
A.5 Power comparison for sample with small size and small number of observations under cross-sectional dependence 136
A.6 Power comparison for discrete distribution 140
A.7 Power comparison for discrete-normal distribution mixture. 141
A.8 Power comparison for mixture of two normal distributions. 141
A.9 Power comparison for single normal distribution 144
A.10 Performances for varying time lengths of investing: 1984–2019 sample . 145
A.11 Comparison of performance statistics of all considered portfolios 146
A.12 Performance of portfolios prior- and post-published year of covariates. . 147
A.13 $fFDR20\%$ portfolios' performances for varying time lengths of investing 149
A.14 Performances of portfolios under an alternative proxy of covariates . . . 150
A.15 Performance of $fFDR10\%$ portfolios with combined covariates 152
A.16 Performances of portfolios under restricted data 153

B.1 Performance of $mfFDR^+$ portfolios on individual currency 162
B.2 Performance of $mfFDR^+$ portfolios on basket of currencies 163
B.3 Performance of $mfFDR^+$ portfolios on individual currency under testing mean return 164
B.4 Performance of $mfFDR^+$ portfolios under testing mean return 165

C.1 OOS performance of $fwer^+$ portfolios under restrictions on AUM . . . 170
C.2 OOS performance of $fwer^+$ portfolios with use of persistent covariates 171
C.3 OOS performance of $fwer^+$ portfolios with use of moment covariates . 172
C.4 OOS performance of $fwer^+$ portfolios with use of 24-month IS periods 173
C.5 OOS performance of $fwer^+$ portfolios with use of 48-month IS periods. 174
C.6 OOS performance of $fwer^+$ portfolios with use of simple $p-$values . . . 175

C.7 OOS performance of $fwer^+$ portfolios with use of machine learning based covariates and simple $p-$values 176
C.8 Performance of the single-fund portfolios with use of simple p-value . . 176

List of Figures

1.1 Comparison of FDR^+ and $fFDR^+$ 14

1.2 Performance of $fFDR^+$ for discrete distribution of α 23

1.3 Performance of $fFDR^+$ for discrete and normal distribution mixture of α 23

1.4 Performance of $fFDR^+$ for continuous distribution of α 24

1.5 Alpha evolution of $fFDR10\%$ and $FDR10\%$ portfolios over time . . . 37

2.1 Distribution of $p-$values 55

2.2 Distributions of $p-$values partitioned into nine groups 56

2.3 Comparison of the procedures across groups 57

2.4 Performance comparison of FDR methods 59

2.5 Performance of the $mfFDR$ under correlated covariates 62

2.6 Performance of the $mfFDR$ under correlated $p-$values 64

3.1 Distribution of hedge fund alphas 95

3.2 Performance comparison 98

3.3 Performance of the $fwer^+$ under various setting of signals. 100

3.4 Performance of the $fwer^+$ under use of insufficient information 102

3.5 Varying sample size and number of observations 103

A.1 Performance of $fFDR^+$ for discrete distribution of α 132

A.2 Performance of $fFDR^+$ for continuous distribution of α 132

A.3 Performance of $fFDR^+$ for discrete and normal distribution mixture of α 133

A.4 Performance of $fFDR^+$ under unbalanced panel data 137

A.5 Performance of $fFDR^+$ under unbalanced panel data: discrete-normal distribution 138

A.6 Performance of $fFDR^+$ under unbalanced panel data: mixture of two normals 139

A.7 Performance of the $fFDR^+$ under balanced panel data: single normal distribution . . . 142
A.8 Performance of $fFDR^+$ under unbalanced panel data: single normal distribution . . . 143
A.9 Alpha evolution with use of data from 1984 . . . 145
A.10 Evolution of wealth . . . 148
A.11 Alpha evolution of $fFDR20\%$ and $FDR20\%$ portfolios . . . 149
A.12 Alpha evolution of $fFDR10\%$ portfolios with use of an alternative proxy of covariates . . . 150
A.13 Alpha evolution of $fFDR10\%$ portfolios with combined covariates . . . 152
A.14 Alpha evolution of $fFDR10\%$ portfolios with use of restricted data . . . 153

B.1 Performance comparison under noisy covariates . . . 158
B.2 Performance comparison under varying sample size . . . 160

C.1 Performance of the $fwer^+$ under the alternative setting of out-performing funds proportion. . . . 169

Acknowledgements

I express my heartfelt gratitude to everyone who has supported me throughout my PhD journey.

First and foremost, I am deeply grateful to my supervisors, Professors Christian Ewald and Georgios Sermpinis, for their invaluable guidance and unwavering support throughout my research journey.

I also extend my thanks to the staff at the Adam Smith Business School, including Christine Athorne, Sophie Watson, and others, for their exceptional assistance. I am grateful to the School for the financial support provided during my MRes and PhD studies.

Furthermore, I am grateful to Professors Le Van Cuong and Ha Huy Khoai for providing me with recommendation letters when I applied to the MRes program. It was through their support that the whole journey began.

I would also like to thank Professors Mark P. Taylor and Ilias Filippou for their invitation to visit Olin Business School. This opportunity added a vibrant and enriching dimension to my PhD journey.

To my family, I express my deepest appreciation for your unwavering support and belief in my chosen path. And to my friends, thank you for bringing joy and creating cherished memories throughout my journey. Your presence has made this experience truly special.

Thank you all for being an integral part of my academic journey.

Tren Ma

Declaration

"I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution."

Printed Name: Tren Ma

Signature:

Abbreviations

BH	Benjamini and Hochberg
BSW	Barras, Scaillet and Wermers
CRS	Chen, Robinson and Storey
DNN	Deep neural network
FDR	False discovery rate
$fFDR$	Functional false discovery rate
$fFDR^+$	Functional false discovery rate "plus"
FWER	Family wise error rate
$fwer^+$	Family wise error rate "plus"
GB	Gradient boosting
HAC	Heteroskedasticity and auto-correlation consistent
IS	In-sample
LASSO	Least absolute shrinkage and selection operator
$mfFDR$	Multivariate functional false discovery rate
$mfFDR^+$	Multivariate functional false discovery rate "plus"
OLS	Ordinary least squares
OOS	Out-of-sample
PC1	First principal component
RF	Random fores
SR	Sharpe ratio
TNA	Total net assets under management

Introduction

Investors tend to select funds based on their past records but high past alphas might be due to luck. In order to identify the truly skilled funds, namely, the funds with genuine positive alphas due to, for instance, possession of superior information or trading skill, the investors should rely on a multiple hypothesis testing framework. They should devise an approach that controls the number of false discoveries, i.e., the number of funds that seem skilled but are truly not. Similarly, traders pick genuinely out-performing trading strategies for future use based on assessing past performance of thousands of strategies. They should control for number of lucky strategies, i.e., those are randomly buying and selling signals but turn out to be luckily outperforming in the past.

Literature in finance have developed a number of procedures that aim to control for such false discoveries. Those procedures include frameworks that control for probability of having at least one false discovery among those called discoveries, i.e, the family wise error rate (FWER), and those that control for a less stringent type I error, the expected proportion of false discovery among the discoveries, i.e., the false discovery rate (FDR) of Benjamini and Hochberg (1995). Notable contributions are the studies of White (2000), Storey (2002), Barras *et al.* (2010), Romano and Wolf (2005), Hansen (2005), Hsu *et al.* (2010) among others. While implementations of those procedures control for the type I errors and tackle successfully for data-snooping bias (see White, 2000), they are too conservative and can miss too many genuine discoveries (see Harvey and Liu, 2020). As such, we need to develop new procedures that are more powerful in detecting the genuine discoveries and, in particularly, the genuine out-performers while controlling for the type I errors. The common ground of the mentioned approaches is that they either use only $p-$value of the tests or bootstrapping procedures while ignoring additional information that can be useful.

This thesis develops a series of statistical frameworks that utilize such additional information to raise the power of testing procedures and to exploit the extra informativeness in decision-making. Put differently, the new methods allow us conduct tests conditionally on the additional information. The new procedures gain much higher power than the existing ones while controlling for the same type I error. We show that these approaches are able to transform the informativeness of the additional information to profit for investors in practice. The thesis develops three statistical frameworks which are implemented on three different financial data, first to show the advantages of the new procedures, and second, to gain economics insights. It is organised as three chapters as follows.

The first chapter contributes to literature the introduction of the functional False Discovery Rate "plus" ($fFDR^+$). It develops the functional False Discovery Rate framework ($fFDR$) of Chen *et al.* (2021a) (CRS) to control FDR in picking out-performing mutual funds. Compared to the work of CRS, the $fFDR^+$ has two distinguishing features. First, it allows us to focus on the right or left tail of the distribution and detects the significant out- or under-performers, which is important for decision makers. Second, it is robust to cross-sectional dependencies among predictive models, which are common for most problems in economics and finance. Simulation experiments show that the $fFDR^+$ controls well for FDR while gain higher power than existing approach with a gap which can be up to about 30%. In empirical experiments, we use the $fFDR^+$ to detect out-performing mutual funds for investing purpose. We find that the set of mutual funds selected as out-performers by $fFDR^+$ is usually larger and different from the one obtained by prior FDR methods. This suggests that, with more information updating, there may exist more profitable mutual funds than researchers would have expected. Based on the funds selected by $fFDR^+$, we build portfolios that consistently outperform the one generated by prior methods. The results thus highlight the economic value of extra information. The profitability of the portfolios is persistent in our sample and is even strengthened over the recent period, a finding that disagrees with part of the recent literature which suggests otherwise (see Jones and Mo, 2021).

Often, we usually have more than one informative covariates that are available and there might not be a way to efficiently combine them. The second chapter thus proposes a new methodology that accounts for more information sources in forming the rejection criterion, namely multivariate functional FDR ($mfFDR$), which enhances the statistical power while controlling the FDR. In simulations, we show that our $mfFDR$ performs well in controlling for the FDR and beats the $fFDR$ of CRS and FDR method of Storey (2002) in terms of power with gaps of about 44% and 67%, respectively. This advantage remains under weakly dependent tests and correlated or noised informative covariates data. In empirical studies, we implement the $mfFDR$ to detect genuinely profitable trading rules in a set of 635,850 technical trading rules and invest following the signals generated by the selected rules in monthly rolling fashion. For this purpose, a derivative procedure, namely $mfFDR^+$, and four covariates are introduced. We find that the out-of-sample (OOS) performance of the $mfFDR$-based portfolio with use of the four covariates is better than those based on prior methods and can gain a Sharpe ratio of 1.06 and 0.95 before and after transaction costs, respectively. In terms of empirical power, the more informative covariates that we use, the higher number of truly profitable rules are detected. This suggests that the prior methods without using covariates may underestimate technical trading rules' truly predictive ability and profits because their performance is not evaluated with comprehensive information sets. We also see that, using all four covariates under the $mfFDR$ framework offers higher profit than using the first principal component of the four as the sole covariate. This suggests that the $mfFDR$ might effectively extract non-linear information among covariates and that simple linear combinations of covariates are not effectively conveying the informativeness of all underlying covariates. These facts highlight the importance of the $mfFDR$ development.

The mentioned functional FDR approaches are so powerful and be able to detect many genuinely out-performing trading rules or funds. In some applications, such as in picking out-performing hedge funds, this might not be desirable. The reason is that, hedge funds require a significant investment, called minimum investment. When the number of funds in the portfolio is large, the total minimum investment might be also too large so that the investment becomes infeasible. In addition to this, hedge funds usually have a lock-up period during which investors are not allowed to withdraw their investment without a significant fee. The investors thus need to be more carefully in making investment decision. Alongside the FDR, another popular multiple testing type I error is the family wise error rate (FWER), which accounts for probability of having at least one false discovery. Controlling for the FWER is more stringent than controlling for FDR and therefore is more appropriate to the hedge fund selection problem. Nevertheless, the existing approaches in finance literature are too conservative when applying to this particular problem. The reason lies on the fact that hedge fund return series are typically short and thus the investors assess hedge fund performance based on short past periods. The notable developments, such as the stepwise reality check of Romano and Wolf (2005), struggle in detecting out-performing hedge funds and thus generate empty portfolios for many years. The third chapter solves this problem. More specifically, it introduces a new approach which incorporates side informative covariates in estimating FWER in detecting out-performers, namely $fwer^+$. The new approach is based on statistical framework of Zhou *et al.* (2021), but deviates to control for FWER among the discoveries in the right tail of distributions. The $fwer^+$ is so powerful that it outperforms the existing methods and allows the investors picking out-performing hedge funds with a very low level of FWER. Empirically, portolios of the hedge funds detected by the $fwer^+$ perform persistently in long out-of-sample (OOS) periods. They beat passive benchmarks with statistically significantly positive alpha. The empirical results thus highlight the admixture of informative covariates and controlling FWER and the highly potential applications of the method in real world practices.

The remainders of the thesis are structured as follows. The first chapter presents the $fFDR^+$ framework and its use in mutual fund performance assessment. The second chapter focuses on developments of the $mfFDR$ and its application in profitablity of technical trading rules in foreign exchange market. The third chapter introduces the $fwer^+$ and its implementation in the hedge fund portfolio selection. Finally, the conclusion completes the thesis.

Chapter 1

Functional false discovery rate in mutual fund selection

1.1 Introduction

Aiming to identify models with genuine predictive power from a large set of potential candidates, researchers have to resort to a multiple hypothesis testing framework to appropriately address the "data-snooping" or "p-hacking" bias that is a major challenge to social science (Sullivan *et al.*, 1999, 2001; White, 2000; Hansen, 2005). To address this challenge, researchers propose the concept of the False Discovery Rate (FDR) of Benjamini and Hochberg (1995), Storey (2002), Storey (2003), and Romano and Wolf (2005), i.e., the ratio of models that are mistakenly identified as having predictive power. Testing methods based on FDR has gained considerable attention in the literature and has been successfully applied to many areas of social science.[1]

One common feature of the methodologies in this framework is that the rejection criterion *only* depends on information of raw data and predictive models' performance metrics. However, in economics and finance research, the economic agents use all available information in assessing models' performance. Extra information sources can assist researchers to more accurately estimate FDR. Recently, Chen *et al.* (2021a), CRS henceforth, introduced the functional FDR method that embeds the role of informative covariates (i.e., variables that carry extra information) in forming null hypotheses. This advancement is important in the sense that it enables us to test the "conditional"

[1] For instance, Fan and Fan (2011) employ FDR in testing and detecting jumps; Lan *et al.* (2016) utilize such a framework to control FDR in testing coefficients in high-dimensional linear models; see also Lan and Du (2019) for extensions and applications in mutual fund selection; or Barbaglia *et al.* (2022) for applications in detecting significant sentiment variables in forecasting with economic news.

performance of predictive models, which is more consistent with the rational expectation hypothesis. To illustrate the importance of extra information in multiple testing problems, we can use mutual fund performance assessment as an example. If we use prior testing methods that do not account for extra information, we are testing an unconditional zero hypothesis, which corresponds to investors not updating information, for instance utilizing newly discovered predictors in literature, in assessing and picking mutual funds. This approach appears inappropriate because mutual funds and their managers are routinely reviewed by investors based on updated information and knowledge. In other words, a more suitable null hypothesis for a mutual fund's performance should be zero conditional on the updated information set.

Incorporating additional information in testing and asset pricing has been discussed extensively in econometrics and finance literature. For instance, Hansen (1995) introduces covariate to better estimating confident interval and raising power of unit root testing, Astill *et al.* (2023) utilize covariates to improve the power of testing for explosive series, Kelly *et al.* (2019) and Gu *et al.* (2021) develop asset pricing models where factor loadings are dynamics and estimated from asset characteristics. The combination of an FDR framework and covariates in finance literature can be traced back to the study of Barras *et al.* (2010) (hereafter BSW) where the authors introduce the FDR framework of Storey (2002) to assessing mutual fund performance. They estimate the distribution of mutual fund alphas conditionally on funds' investment style which is a covariate. However, the investigation simply relies on partitioning funds into groups based on the covariate and estimating the fund alpha distribution in each group. The studies of Storey (2002) and BSW have gained influence and popularity in economics and finance as, for instance, researchers face multiple tests in almost every area of empirical finance (see footnote 1). However, their FDR framework has been shown to be too conservative by recent work of Andrikogiannopoulou and Papakonstantinou (2019) (AP henceforth). Literature in mutual fund selection has two strands. One that picks out-performing funds with control of luck, measured by a type I error such as FDR. The other strand focuses on seeking new predictors that forecast funds' future return. A list of such predictors can be found in recent study of Jones and Mo (2021). Consequently, funds are selected based on sorting their realization of the predictors and past performance. Yet, there has not been a study that combines the two strands.

The $fFDR$ of CRS develops further the framework of Storey (2002) by incorporating an additional covariate in estimating FDR. CRS show that their new approach estimates more precisely the FDR and gains significant higher power than that of not only Storey (2002) but also other recent frameworks that incorporates additional information in estimating FDR such as those of Ignatiadis *et al.* (2016) and Ignatiadis and Huber (2021). In this study, we introduce the framework of CRS and develop it further for applications in economics and finance. Our choice of methodology is stemmed from the fact that the framework of CRS is more powerful than other developments in statistics literature and it works well for relative smaller number of hypothesis tests. The latter feature is important since in economics and finance topics we do not always have several thousands of hypothesis tests - the required input of many other methods such as Ignatiadis *et al.* (2016), Zhang *et al.* (2019) and Ignatiadis and Huber (2021).

Our main contribution is the introduction of the functional False Discovery Rate "plus" ($fFDR^+$). Compared to the work of CRS it has two distinguishing features. First, it allows us to focus on the right or left tail of the distribution and detect the significant out-performers/under-performers, which is important for decision makers (see BSW). Second, it is robust to cross-sectional dependencies among predictive models, which is common for most problems in economics and finance. For example, in mutual funds, the alphas are likely dependent due to herding and correlated trading behaviour (Wermers, 1999).

Compared to all earlier methods in the economics literature on control of the FDR, our $fFDR^+$ method incorporates extra information, has higher power, and controls for noise. It is easy to implement, does not rely on any strong assumption and can handle any continuous informative covariate. In examining our method, we use simulated mutual fund performance similarly to BSW and AP. We show that, when an informative covariate is available, our $fFDR^+$ approach detects more true positive alpha funds under different alpha distributions, balanced and unbalanced data, and both cross-sectional independence and dependence in the error terms. The gap in power between $fFDR^+$ and prior FDR methods, depending on the distribution of the fund alpha population, can be up to about 30%. Our approach is also robust to estimation errors in the covariates.

We then apply our method and construct portfolios in order to evaluate it empirically in selecting outperforming mutual funds. In particular, we explore nine informative covariates: the first set contains five covariates that have been shown in prior studies to convey information on mutual fund performance, and the second set contains four new covariates that are inspired by asset pricing models. The first set includes the R-square of the asset pricing model (e.g., Carhart four-factor model) as suggested by Amihud and Goyenko (2013), the Return Gap of Kacperczyk *et al.* (2008), the Active Weight of Doshi *et al.* (2015), the Fund Size of Harvey and Liu (2017), and the Fund Flow suggested by Zheng (1999). The second set includes the Sharpe ratio, the Beta and Treynor ratio based on the Capital Asset Pricing Model (CAPM), and the idiosyncratic volatility of the Carhart four-factor model (Sigma).

We find that the set of mutual funds selected as out-performers by $fFDR^+$ is usually larger and different from the one obtained by prior FDR methods. As already discussed, earlier studies do not account for information other than mutual funds' returns and performance metrics; thus, their null hypotheses are unconditional and neglect investors' time-varying expectation. The fact that our $fFDR^+$ discovers more outperforming funds suggests that, with more information updating, there may exist more profitable mutual funds than researchers would have expected.

Based on the funds selected by $fFDR^+$, we build portfolios that consistently outperform the one generated by prior methods. Our results highlight the economic value of extra information. In particular, the $fFDR^+$ portfolios with the R-square and Beta covariates are found to be the best with annualized alphas of 1.7%, followed by the $fFDR^+$ portfolios with the Active Weight, Fund Flow, Sigma, Treynor ratio, Fund Size, Sharpe Ratio and Return Gap covariates, separately achieving annualized alphas of at least 0.77%. We note that this profitability is persistent in our sample and is even strengthened over the recent period, a finding that disagrees with part of the recent literature which suggests otherwise (see Jones and Mo, 2021). All our $fFDR^+$ portfolios outperform the one generated by prior FDR methods and a set of portfolios created by single- and double-sorting the covariates under study.

In additional analysis, we also consider the $fFDR^+$ portfolio based on various ways of combining the nine covariates, such as the first principal component of the nine covariates (PC 1), the ordinary least squares (OLS), the least absolute shrinkage and selection operator (LASSO) of Tibshirani (1996), the ridge regression and the elastic net of Zou *et al.* (2005). We find that the elastic net delivers the best performance with an annualized alpha of 1.25%. The investors may also benefit from such combinations as they result in lower volatility in portfolio performance. This is advantageous as, in reality, investors do not know ex-ante what covariate is the best.

The chapter is organized as follows. In Section 1.2, we introduce and explain our methodology. In Section 1.3, we provide a description of our data. Section 1.4 is devoted to our simulation experiment descriptions, whereas in Section 1.5 we present in detail our simulation results. Section 1.6 focuses on the empirical part of our analysis. Section 1.7 concludes the chapter.

1.2 Methods for controlling of luck with informative covariate

1.2.1 Functional false discovery rate ($fFDR$)

Throughout this chapter, we use mutual funds to represent predictive models. We define funds' performance based on their net return, that is, the return net of trading cost, fees and other expenses except loads and taxes. A fund is deemed out-performing if it distributes to investors a net return that generates a positive alpha (i.e., a part of a return series that is unexplained by systematic risk). If the alpha is negative (zero), the fund is said to be under-performing (zero-alpha). These definitions of out-performing and under-performing funds coincide with skilled and unskilled funds in BSW, respectively, and reflect the interest of investors.

Suppose that we are assessing m funds and each of them has a net return time series. We also assume that there exists a covariate X, with observed values $(x_1, \ldots, x_m)$, that conveys information about the alpha of each fund. Associated with X, we define Z whose observed value for fund i is $z_i = rank(x_i)/m$, where $rank(x_i)$ is the ranking of x_i in the set of observed values $(x_1, \ldots, x_m)$. As X to Z is an one-one mapping and we work based on Z, we call that the covariate from now on. We introduce our notation by means of a single test, conditional on Z, for the alpha of a mutual fund:

$$H_0 : \alpha = 0, \qquad H_1 : \alpha \neq 0. \tag{1.1}$$

We denote by h the status of the null hypothesis, that is, $h = 0$ if the hypothesis $\alpha = 0$ is true and $h = 1$ if otherwise. In addition, P is the random variable representation of the p-value of the test, Z, as mentioned above, is the covariate which is uniformly distributed on $[0, 1]$, and $T = (P, Z)$. We suppose that $(h|Z = z) \sim \text{Bernoulli}(1-\pi_0(z))$, that is, conditional on $Z = z$, the fund possesses a zero alpha with probability $\pi_0(z)$; this can be constant if Z does not convey any information about the probability of the fund's alpha being zero. The estimation procedure for $\pi_0(z)$ will be discussed later on. We require that under the true null, $(P|h = 0, Z = z)$ is uniformly distributed on $[0, 1]$ regardless of the value of z; when the null hypothesis is false, the conditional density function of $(P|h = 1, Z = z)$ is $f_1(.|z)$.

To assess the performance of m funds in terms of α within our framework, we consider m conditional hypothesis tests like (1.1):

$$H_{0,i} : \alpha_i = 0, \qquad H_{1,i} : \alpha_i \neq 0, \qquad i = 1, \ldots, m, \tag{1.2}$$

where α_i is the alpha of fund i. For each i we have $T_i = (P_i, Z_i)$, and we assume that all the pairs are independent and each of them has the same distribution as (T, h).[2] Finally, we denote by $f(p, z)$ the joint density function of (P, Z). We have that

$$\mathbb{P}(h = 0|T = (p, z)) = \frac{\pi_0(z)}{f(p, z)} =: r(p, z) \tag{1.3}$$

is the posterior probability of the null hypothesis being true given that we observe $T = (p, z)$.[3]

To control the type I error, Storey (2003) introduces the "positive false discovery rate"

$$pFDR = \mathbb{E}\left(\frac{V}{R}\middle| R > 0\right), \tag{1.4}$$

where R is the number of rejected hypotheses in m tests and V the wrongly rejected ones. Chen *et al.* (2021a) show that, with a fixed set Γ in $[0, 1]^2$, if we reject hypothesis

[2] In the Appendix A.2, we show that this requirement can be eased for a typically cross-sectional dependence in mutual fund data.

[3] For more details about the role of $Z \sim \text{Uniform}(0, 1)$ and the derivation of (1.3), see Chen *et al.* (2021a).

$H_{0,i}$ whenever $T_i \in \Gamma$, then

$$pFDR(\Gamma) = \mathbb{P}(h = 0 | T \in \Gamma) = \int_{\Gamma} r(p, z) dp dz. \tag{1.5}$$

To maximize the number of rejections, we reject the hypotheses with the smallest statistic $r(p, z)$. Thus, the significance region $\{\Gamma_\theta : \theta \in [0, 1]\}$ is defined as

$$\Gamma_\theta = \{(p, z) \in [0, 1]^2 : r(p, z) \leq \theta\}, \tag{1.6}$$

where a larger θ implies more rejected hypotheses. Finally, we recall from Storey (2003) and CRS the definition of the q-value for the observed (p, z):

$$q(p, z) = \inf_{\{\Gamma_\tau | (p,z) \in \Gamma_\tau\}} pFDR(\Gamma_\tau) = pFDR(\Gamma_{r(p,z)}). \tag{1.7}$$

Given a target $\tau \in [0, 1]$, a procedure that rejects a hypothesis if and only if its q-value $\leq \tau$ guarantees that $pFDR$ is controlled at τ.

Empirically, let $\hat{\pi}_0(z)$ and $\hat{f}(p, z)$ be the estimated functions $\pi_0(z)$ and $f(p, z)$, respectively.[4] We calculate $\hat{r}(p, z) = \hat{\pi}_0(z)/\hat{f}(p, z)$ and estimate the q-value function as

$$\hat{q}(p_i, z_i) = \frac{1}{\#S_i} \sum_{k \in S_i} \hat{r}(p_k, z_k), \tag{1.8}$$

where $\#S_i$ returns the number elements of the set $S_i = \{j | \hat{r}(p_j, z_j) \leq \hat{r}(p_i, z_i)\}$ and p_i is the p-value of test i. Then, given a target $pFDR$ level $\tau \in [0, 1]$, the null hypothesis $H_{0,i}$ is rejected if and only if $\hat{q}(p_i, z_i) \leq \tau$. CRS call this procedure Functional False Discovery Rate ($fFDR$).

1.2.2 The $fFDR^+$: application in selecting out-performing funds

By applying the $fFDR$ methodology to mutual funds at a given target $pFDR$ level τ, we obtain a set that includes both significantly out-performing and under-performing funds. To further improve mutual fund selection, we propose a $fFDR$-based method that selects a group of significantly out-performing funds with control of luck. In the following section, we introduce our $fFDR^+$ and discuss its application in a mutual

[4] See Appendix A.1 for more details.

fund context.

Consider a selection of R^+ out-performing funds including V^+ wrongly selected zero-alpha or under-performing funds. We define the positive false discovery rate in those significantly out-performing funds as

$$pFDR^+ = \mathbb{E}\left(\frac{V^+}{R^+}\middle| R^+ > 0\right). \tag{1.9}$$

For m tests, let A^+ be the set of hypotheses with positive estimated alpha, i.e., $A^+ = \{i|\hat{\alpha}_i > 0\}$, where $\hat{\alpha}_i$ is the estimated alpha of fund i. At a given target τ of $pFDR^+$, by implementing the $fFDR$ procedure to control $pFDR$ at the target τ on the funds in set A^+, we obtain all the funds with positive estimated alphas (referred to as significant alphas).[5] Hence, the $fFDR$ selects positive-alpha funds with control of $pFDR$ at the given target; we call this procedure the functional FDR "plus" ($fFDR^+$).

Next, we highlight the differences between our and BSW's approaches. The starting point of both is the control of the type I error as in Benjamini and Hochberg (1995):

$$FDR = \mathbb{E}\left(\frac{V}{\max\{R,1\}}\right) = \mathbb{E}\left(\frac{V}{R}\middle| R > 0\right)\mathbb{P}(R > 0) = pFDR \cdot \mathbb{P}(R > 0), \tag{1.10}$$

where the last equality follows from (1.4). This implies that controlling for $pFDR$ at a given target τ, also controls for FDR at the same target. Furthermore, for a large number of tests, controlling for $pFDR$ and FDR is equivalent (see Storey, 2002, 2003).

Consider the m tests (1.2) in the absence of the covariate Z and let t_i be the test statistic of test i. Storey (2002) assumes that $t_1, \ldots, t_m$ are independent and the statuses of the null hypotheses $h_1, \ldots, h_m$ are independent Bernoulli random variables with $\mathbb{P}(h_i = 0) = \pi_0$. Additionally, across i, $(t_i|h_i = 0)$ and $(t_i|h_i = 1)$ are identically distributed. When we reject based on the p-values, for some $\lambda \in [0, 1)$, π_0 can be estimated by

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i|p_i > \lambda, i = 1, \ldots, m\}}{(1-\lambda)m} \tag{1.11}$$

[5] In doing so, we assume that the number of funds that are out-performing but exhibit a negative estimated alpha is negligible. This is sensible as in practice we will not select those funds anyway. In BSW, as discussed next, having a positive estimated alpha is a necessary condition for a fund to be selected as out-performer.

where # returns the number of elements in the set; this estimate is conservative biased.[6] BSW choose $\lambda = \lambda^*$ on the grid $\{0.3, 0.35, \ldots, 0.7\}$ such that the mean square error (MSE) of $\hat{\pi}_0(\lambda)$ is minimal.[7] We set $\hat{\pi}_0 = \hat{\pi}_0(\lambda^*)$.

To select out-performing funds with controlling for the FDR, BSW define the concept FDR^+ to measure the FDR in a group of funds selected as significant and positive estimated alphas as

$$FDR^+ = \mathbb{E}\left(\frac{V^+}{\max\{R^+, 1\}}\right). \tag{1.12}$$

With a significant threshold γ and a procedure which selects a fund with a positive estimated alpha whenever its p-value $\leq \gamma$, BSW estimate FDR^+ by

$$\widehat{FDR}^+_\gamma = \frac{\hat{\pi}_0 \gamma/2}{\hat{R}^+/m}, \tag{1.13}$$

where $\hat{R}^+$ is the empirical number of funds selected as out-performers, i.e., $\hat{R}^+ = \#\{i | p_i \leq \gamma, \hat{\alpha}_i > 0\}$. When using this approach to determine out-performing funds with controlling for FDR^+ at a given target τ, the threshold γ is raised gradually until the $\widehat{FDR}^+$ estimate in (1.13) reaches the target τ. We refer to this procedure as FDR^+.

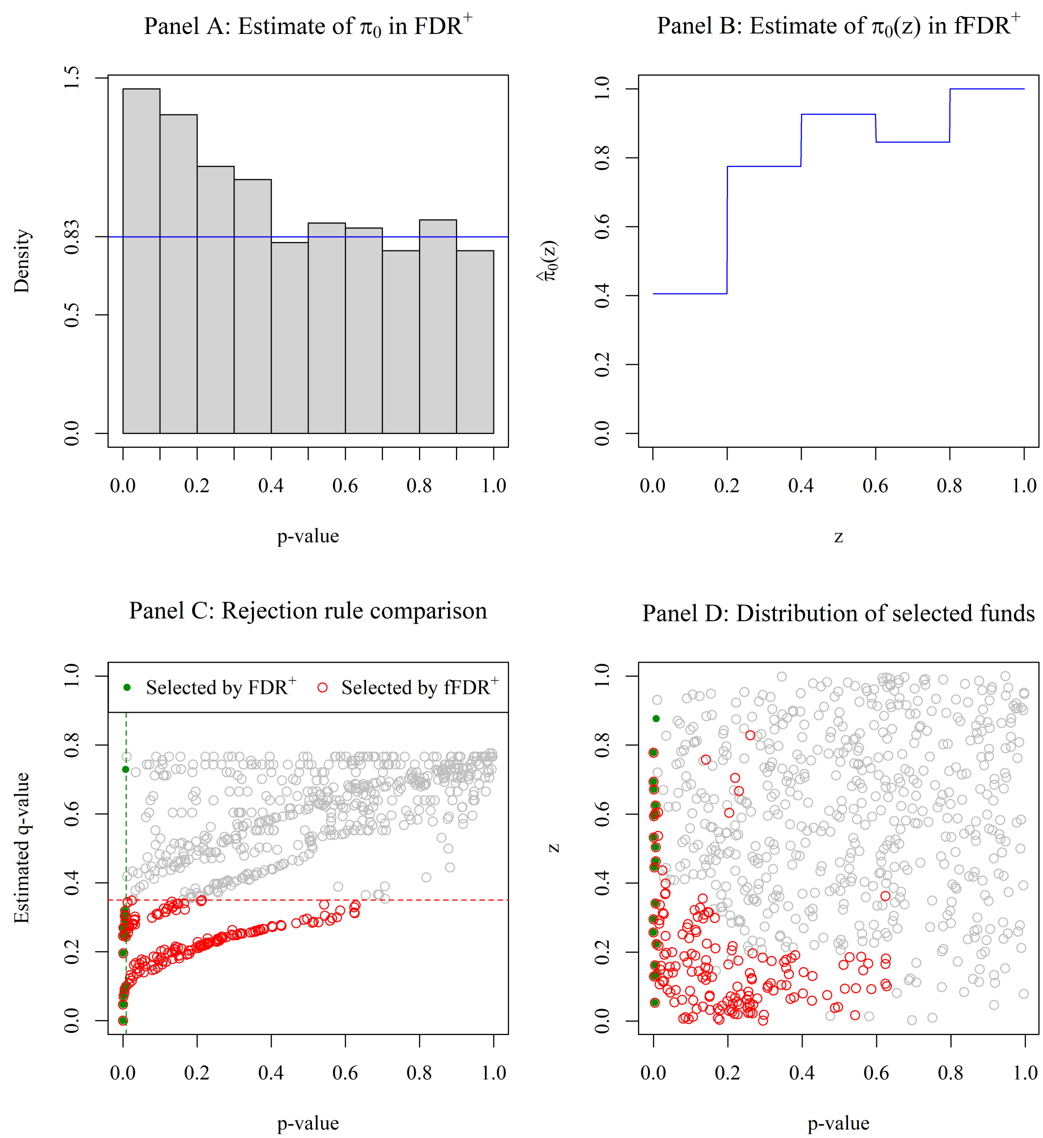
To illustrate the differences between our and BSW's procedures, we opt for a subperiod of five years from the beginning of 2001 to the end of 2004, the five-year period during which our sample reaches its highest number of funds, and implement the FDR^+ and $fFDR^+$ to detect positive alpha funds, with the alphas determined by the four-factor model of Carhart (1997). In this case, the R-square of the model is used as the covariate for $fFDR^+$.[8] In Figure 1.1, we demonstrate how the two procedures work. Based on the p-values of all the considered funds, the FDR^+ estimates the proportion of zero-alpha funds in the whole sample, as a first step, giving $\hat{\pi}_0 \approx 0.83$. It then selects the positive estimated alpha funds with smallest p-values until the estimated $\widehat{FDR}^+_\gamma$ reaches a given FDR target. For illustration, we choose the FDR target $\tau = 35\%$,

[6]To have the estimate of π_0, first, under independence, there are $m\pi_0$ funds with truly zero alpha and their p-values have a uniform distribution in $[0, 1]$. Hence, we expect $m\pi_0(1-\lambda)$ p-values in the set to fall in $[\lambda, 1]$. Second, this number can be conservatively approximated by $\#\{p_i | p_i > \lambda\}$, thus we have (1.11). With a larger λ, the estimate $\hat{\pi}_0$ is less conservative, as there are fewer p-values under the alternative belonging to $[\lambda, 1]$, but its variance is higher.

[7]In MSE $= \mathbb{E}(\hat{\pi}_0(\lambda) - \pi_0)^2$, the unknown π_0 is replaced by $\min_\lambda \hat{\pi}_0(\lambda)$ over the λ grid.

[8]The details of the funds and the calculation of the p-values are deferred to Section 1.6. Here, we focus only on illustrating the differences.

Figure 1.1: Comparison of FDR^+ and $fFDR^+$. The graphs show the differences between the two procedures with respect to their null proportion estimations and their rejection rules. Panels A and B show that π_0 is estimated as a fixed number in FDR^+ procedure (see (1.11)) but as a step function in $fFDR^+$ procedure (see Appendix A.1). Panel C shows the rejection rules of the FDR^+ and $fFDR^+$: the former selects all the funds corresponding to the points on the left of the vertical green dashed line which consists of all funds with positive estimated alphas and p-values less than 0.0086, whereas the latter all the funds corresponding to the points below the horizontal red dashed line which consists of all funds with estimated q-value (see (1.8)) less than 0.35. Panel D shows the distribution of the selected funds in Panel C with respect to the p-value and the covariate z. In Panels C and D, only funds with positive estimated alpha are shown as ultimately both methods select funds from this set. The solid green points represent funds selected by the FDR^+, whereas the red circles the funds selected by the $fFDR^+$; the green points with a red ring are the commonly selected funds.



so that both methods select a substantial number of funds.[9] Here, all the funds with p-values less than or equal to $\gamma = 0.0086$ are selected by the FDR^+. The threshold γ is depicted by the green dashed line in Panel C and all the funds corresponding

[9] If we choose any target $\tau \leq 30\%$, the FDR^+ selects no funds.

to the points on the left of the vertical line are selected. By contrast, the $fFDR^+$ considers only the set of positive estimated alpha funds and estimates the proportion of zero-alpha funds in this set as a step function of z (the quantiles of R-square).

In this experiment, we split the sample into five bins based on the ranking of the covariate z; thus, $\hat{\pi}_0(z)$ is a function with five "steps". The procedure continues with the estimation of the density function $f(p, z)$ and of the functional q-value $q(p, z)$. The $fFDR^+$ selects all the funds with estimated q-value less than or equal to 0.35: those funds correspond to the points below the red dashed line (the q-value $= 0.35$ line) in Panel C. This clearly shows that, for the same target, the $fFDR^+$ selects significantly more funds than FDR^+ (170 versus 19). More importantly, the funds selected by the FDR^+ are not merely a subset of those selected by $fFDR^+$. Panel D displays the distribution of the selected funds with respect to the p-value and z. We observe that the $fFDR^+$ assigns more weight to some funds with smaller z (thus, smaller R-square), but the weight is not equally distributed across the funds with the same level of z. As the rejection rule of $fFDR^+$ is based on the functional q-value, which is based on the estimates of $\pi_0(z)$ and $f(p, z)$, it is not possible to explain this merely by the ranking of the p-value and the covariate z, as evidenced in Panel D: the $fFDR^+$ selects some funds with p-values around 0.6 while skipping many funds with a smaller p-value at roughly the same level of z.

As shown in AP, the FDR^+ relies on an over-conservative estimate of the null proportion and utilizes only p-values and the estimated alphas. On the other hand, the $fFDR^+$ additionally uses an informative covariate about the performance of the funds and expresses the null proportion as a function of it, while accounting for the joint distribution of the p-value and the covariate. This results in a more accurately estimated FDR and, therefore, an increased power in detecting out-performing funds. We are illustrating the prominent power of the $fFDR^+$ via a set of simulation studies in the next sections. In the empirical section, we will show the actual profitability that the covariates can bring to investors while controlling for luck.

1.3 Data

We use monthly mutual fund data from January 1975 to December 2019 collected from the CRSP database.[10] As CRSP reports funds at the share class level, we use MFLINKS to acquire fund data at the portfolio level. For a fund at a given point in time with multiple share classes, we average the share classes' net return weighted by the total net asset (TNA) value at the beginning of the month.[11] The TNA at the fund level is estimated by the sum of the share classes' TNA. We omit the following month return after a missed return observation as CRSP fills this with the accumulated returns since the last non-missing month. To obtain the holdings data of the funds, which will be used to calculate our covariates, we merge the CRSP and Thomson Financial Mutual Fund Holdings databases by utilizing MFLINKS. The holdings database, which was purchased from CDA Investment Technologies Inc. (see Wermers, 2000), provides us with stock identifiers, which we use to link the funds' position with the CRSP equity files. From this equity database, we obtain information such as the price and number of shares outstanding of the stocks that the funds hold on their reported portfolio date. We use these to calculate the return gap and the active weight, which are described in more detail later.

As in BSW, we consider only open-end, U.S. equity mutual funds which are classified into three categories: Growth, Aggressive Growth and Growth & Income. We collect those funds via using their investment objective. Both CRSP and CDA provide this information; CDA is more consistent over time, hence we choose that. To obtain a precise four-factor alpha estimate, we select only funds with at least 60 monthly observations. Overall, we gather a sample of 2,224 funds which provides the empirical metrics for our simulation study.

In the empirical part, when calculating the related covariates, we additionally require each fund to hold at least 10 stocks; this is consistent with Kacperczyk *et al.* (2008) and Doshi *et al.* (2015) and is needed here as we use the return gap and active weight from their studies as two of our covariates.

[10] We are aware of the possible biases in the CRSP mutual fund data before 1984 (Fama and French, 2010) and thus conduct a robustness check using a sample from 1984 to 2019 in Appendix A.3.

[11] Since 1991, we use the monthly TNA of the fund's share classes. Before 1991, most of the funds report their TNA on a quarterly basis. For this, we follow Amihud and Goyenko (2013) to fill in the missing TNA of each fund (at the share class level) by its most recently available one.

1.4 Simulation setup

In this section, we present the details of our simulation design consisting of the choice of the model, the distributions of the alpha population, the data-generating process and the metrics that we will use to gauge the performance of the methods.

1.4.1 The model

Following the majority of the existing literature on mutual fund performance, we use the four-factor model of Carhart (1997) to compute the fund performance:

$$r_{i,t} = \alpha_i + b_i r_{m,t} + s_i r_{smb,t} + h_i r_{hml,t} + m_i r_{mom,t} + \varepsilon_{i,t}, \qquad i = 1, \ldots, m, \tag{1.14}$$

where $r_{i,t}$ is the excess net return of fund i over the risk-free rate (i.e., the one-month Treasury bill rate), $r_{m,t}$ the market's excess return on the CRSP NYSE/Amex/NASDAQ value-weighted market portfolio, $r_{smb,t}$ the Fama–French small minus big factor, $r_{hml,t}$ the high minus low factor, $r_{mom,t}$ the momentum factor and $\varepsilon_{i,t}$ the noise of fund i at time t. All factors and the one-month Treasury bill rate are obtained from French's website.[12]

Our simulations are designed similarly to BSW and AP in terms of the data-generating process accounting, in addition, for an informative covariate and considering more distribution types of the fund alpha population. Whereas BSW and AP focus on the estimated proportions of the out-performing, under-performing and zero-alpha funds, we consider the performance of the FDR^+ and $fFDR^+$. More specifically, for a given fund alpha distribution, we first generate in each iteration the true fund alpha population and a covariate that conveys information about the alpha of each fund. Second, we simulate the Fama–French factors (factors loadings) by drawing from a normal distribution with parameters equal to their sample counterparts (obtained from estimations of model (1.14)). Next, the noise is generated under both cross-sectional independence and dependence. In the first case, the noise is drawn cross-sectionally independent from a normal distribution, that is, $\varepsilon_{it} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ where, as in Barras *et al.* (2020), σ_ε is is set equal to the median of its real-data counterpart, that is,

[12]The data are obtained from https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html in June 2020.

approximately 0.0183 for our sample. The results under this assumption are reported in the next section. In the dependent case, the noise is generated as in BSW and the simulation results are deferred to Appendix A.2 . The simulated data are then used to generate the net return for each fund. Subsequently, by carrying out regression (1.14) of the generated net return on the simulated Fama–French factors, we estimate the alpha and calculate the related p-values for the tests (1.2). Finally, based on these estimated alphas, p-values and the covariate, we implement the $fFDR^+$ and FDR^+, for a given FDR target, to obtain the significantly out-performing funds. We estimate the actual false discoveries rate of the $fFDR^+$ and check if it meets the given target. We then compare the two methods in terms of power, defined as the expected ratio of the number of true positive alpha funds detected to the total number of true positive alpha funds in the population.

1.4.2 The distribution of fund alphas

We consider three different types for the distribution of fund alphas: a discrete, a discrete-continuous mixture and a continuous. A covariate Z conveys information about the alpha of each fund in the population; more specifically, a fund with $Z = z$ has a probability $\pi_0(z)$ of being zero-alpha. Also, without loss of generality, we assume that, for non-zero alpha funds, their covariates and alphas are positively correlated.[13]

First, in the discrete type, we draw alphas from three mass points $-\alpha^* < 0$, 0 and $\alpha^* > 0$ with probabilities π^-, π_0 and π^+. Thus,

$$\alpha \sim \pi^- \delta_{\alpha=-\alpha^*} + \pi_0 \delta_{\alpha=0} + \pi^+ \delta_{\alpha=\alpha^*}. \tag{1.15}$$

We consider five values for $\alpha^* \in \{1.5, 2, 2.5, 3, 3.5\}$ (the values are annualized and in %) together with six combinations of the proportions (π^+, π_0, π^-) based on $\pi^+ \in \{0.1, 0.13\}$, $\pi^-/\pi^+ \in \{1.5, 3, 6\}$ and $\pi_0 = 1 - \pi^- - \pi^+$, i.e., a total of thirty cases.[14]

In the mixed discrete-continuous distribution, we draw alphas from two components

[13] If the correlation is negative, we use instead $-Z$.

[14] The chosen π^+ values are close to those used in the recent literature: $\pi^+ = 10.6\%$ (see Harvey and Liu, 2018) and $\pi^+ = 13\%$ (see Andrikogiannopoulou and Papakonstantinou, 2016). The ratio $\pi^-/\pi^+ = 6$ is studied in AP. Aiming to extend the range of our study, we consider also the ratios 1.5 and 3.

including the mass point 0 and the normal distribution $\mathcal{N}(0,\sigma^2)$ with, respectively, probabilities $\pi_0 \in (0,1)$ and $1-\pi_0$. We have, therefore, that

$$\alpha \sim \pi_0 \delta_{\alpha=0} + (1-\pi_0)\mathcal{N}(0,\sigma^2). \tag{1.16}$$

We consider five values for $\sigma \in \{1,2,3,4,5\}$ (the values are annualized and in %) and the same six π_0 values as in the discrete distribution earlier.

Finally, in the continuous case, we draw alphas from a mixture of two normal distributions $\mathcal{N}(\mu_1,\sigma_1^2)$ and $\mathcal{N}(\mu_2,\sigma_2^2)$ with, respectively, probabilities $\pi_1 \in [0,1]$ and $\pi_2 = 1-\pi_1$, i.e.,

$$\alpha \sim \pi_1 \mathcal{N}(\mu_1,\sigma_1^2) + \pi_2 \mathcal{N}(\mu_2,\sigma_2^2). \tag{1.17}$$

When π_1 and π_2 are positive, we have indeed a mixture; we adopt from Harvey and Liu (2018) $\pi_1 = 0.3$ and $\pi_2 = 0.7$ and, to cover various scenarios of the mixture, we consider fifteen combinations based on $(\mu_1,\mu_2) \in \{(-2.3,-0.7), (-2,-0.5), (-2.5,0)\}$ and $(\sigma_1,\sigma_2) \in \{(1,0.5), (1.5,0.6), (2,1), (2.5,1.25), (3,1.5)\}$ (the values of the pairs are annualized and in %).[15]

In (1.17) $\pi_0 = 0$, whereas in (1.15) and (1.16) $\pi_0 > 0$. When $\pi_0 > 0$, we study an up-and-down shape of $\pi_0(z)$. Specifically, to guarantee $\pi_0(z) \in [0,1]$ for all z, we choose $\pi_0(z) = \min\{1, \max(f(z),0)\} \in [0,1]$, where

$$f(z) = 3.5(z-0.5)^3 - 0.5(z-0.5) + c \tag{1.18}$$

and c is chosen to satisfy $\int_0^1 \pi_0(z)dz = \pi_0$. This way we are able to investigate the effect of π_0 on the power of the methods by varying c while keeping the shape of $\pi_0(z)$ roughly unchanged.[16]

[15] Our choices are intended to be wide enough to encompass the cases of Harvey and Liu (2018): $(\pi_1,\pi_2) = (0.283, 0.717)$, $(\mu_1,\mu_2) = (-2.277,-0.685)$ and $(\sigma_1,\sigma_2) = (1.513, 0.586)$. In the Appendix A.2, we additionally present results of the case $\pi_2 = 0$, i.e., when the mixture becomes a single normal distribution.

[16] The alternative choices of a decreasing function $\pi_0(z)$ with $f(z) = -1.5(z-0.5)^3 + c$, an increasing function $\pi_0(z)$ with $f(z) = 1.5(z-0.5)^3 + c$ or a constant function $\pi_0(z) = c$ result in some discrepancies, without affecting, though, our main conclusions.

1.4.3 Simulation execution

Suppose the distribution of alpha and the form of $\pi_0(z)$ are determined. As a first step, we generate the covariate and alpha for each of the m funds. We generate the covariate vector $(z_1, z_2, \ldots, z_m)$ with each element drawn from the uniform distribution $[0, 1]$ and assign them to the funds. For the cases (1.15) or (1.16), we determine c in (1.18) such that $\int_0^1 \pi_0(z)dz = \pi_0$ for a given $\pi_0 > 0$. For each fund i, we draw h_i from the Bernoulli distribution with success probability $1 - \pi_0(z_i)$ and assign a zero alpha to fund i with $h_i = 0$. Finally, for the remaining funds, we draw true non-zero alphas from the given distribution (1.15) or (1.16) and assign them such that a fund with a smaller z has a smaller alpha. For the case (1.17), we draw alphas from the distribution and then assign them to the funds; again, a fund with a smaller z has a smaller alpha.

In the second step, we simulate the return factors from a normal distribution with parameters equal to their sample counterparts. The loadings of these factors are also drawn from a normal distribution with parameters equal to their sample counterparts obtained from the fund level estimation of equation (1.14). We consider balanced panel data for $2,000$ funds with 274 time-series observations; the number of $2,000$ is chosen to be close to our real sample of 2,224 funds, whereas the number of 274 periods is the median of our sample funds' observations. In unbalanced panel data, the number of observations for each fund is drawn randomly with replacement from the set of the number of observations of the funds in the real-data counterpart. Under cross-sectional independence, the noise term $\varepsilon_{i,t}$ is drawn from a normal distribution $\mathcal{N}(0, \sigma_\varepsilon^2)$, where, as in Barras *et al.* (2020), σ_ε is set equal to the median of its real-data counterpart, that is, approximately 0.0183 for our sample. Under cross-sectional dependence, we follow BSW and assume that all fund residuals load on a common latent factor G_t, whereas the out-performing and under-performing funds load on the specific factors G_t^+ and G_t^-, respectively. Thus,

$$\varepsilon_{i,t} = \gamma G_t + \gamma G_t^+ \mathbb{1}_{\alpha_i > 0} + \gamma G_t^- \mathbb{1}_{\alpha_i < 0} + \eta_{i,t}, \tag{1.19}$$

where $\mathbb{1}_{\alpha_i > 0}$ and $\mathbb{1}_{\alpha_i < 0}$ are, respectively, out-performing and under-performing indicators taking the value 1 if the fund i is out-performing or under-performing, and 0 otherwise. The three latent factors G_t, G_t^+ and G_t^- are assumed to be mutually orthogonal and to the four risk factors and have a normal distribution $\mathcal{N}(0, \sigma_G^2)$, where, from BSW, σ_G is set equal to the average of the monthly standard deviations of the three risk factors (size, book-to-market and momentum). The coefficient γ is set equal to the average of the loading of the three risk factors of the 2,224 funds in our sample. Finally, $\{\eta_{i,t}\}_i$ are uncorrelated and drawn from the normal distribution $\mathcal{N}(0, \sigma_\eta^2)$, where σ_η is chosen such that σ_ε is equated to the median of its real-data counterpart, as in the independent case.

In the last step, we implement the $fFDR^+$ and FDR^+ and compute their performance metrics. More specifically, based on the simulated data from the previous step, we calculate the Carhart four-factor model alpha and the corresponding p-value for each fund. We use the resulting p-value, the estimated alpha and the covariate as inputs to the $fFDR^+$ and FDR^+ procedures. At a given target of FDR, we calculate for each method a proportion of falsely classified funds $\widetilde{FDR}^+$ and a detected proportion $\widetilde{Power}^+$:

$$\widetilde{FDR}^+ = \frac{\widetilde{V}^+}{\max\left\{\widetilde{R}^+, 1\right\}} \quad \text{and} \quad \widetilde{Power}^+ = \frac{\widetilde{C}^+}{\widetilde{T}^+}, \tag{1.20}$$

where $\widetilde{R}^+$ is the number of classified out-performing funds and, among them, $\widetilde{V}^+$ funds are truly zero-alpha or under-performing funds. $\widetilde{T}^+$ is the number of truly out-performing funds in the population and, among them, $\widetilde{C}^+$ funds are classified correctly.

The previous three steps are repeated 1,000 times and we use the average $\widetilde{FDR}^+$ and $\widetilde{Power}^+$ as estimates for the actual FDR and power.

1.5 Analysis of $fFDR^+$ and power

We set the number of funds for simulations at 2,000 which is close to our sample of 2,224 funds. We demonstrate the ability of the $fFDR^+$ to control the FDR for balanced panel data, where the number of observations per fund is equal to 274, under cross-sectional independence. In the interest of space, we refer to the Appendix A.2 for the results under cross-sectional dependence as well as the unbalanced panel data cases. We then compare the powers of the $fFDR^+$ and the FDR^+ in controlling the FDR at the 10% level; we extend to higher levels and highlight the differences between the two procedures. In each simulation study, we analyze the relationship between the powers of the two methods and: i) the proportion of zero-alpha funds in the sample; ii) the magnitude and proportion of positive alpha funds in the sample. We also study the impact of the number of funds in the sample and the number of observations per fund on the power. Finally, we examine the impact of estimation errors in the covariates, in the power of our procedure.

In general, the results show that the $fFDR^+$ controls well the FDR at any given targets. When the FDR target is set at 10%, the $fFDR^+$ detects more positive alpha funds than the FDR^+ with a difference in power up to 30%, depending on cases and parameters of the distributions. When we raise the FDR target to higher levels, the difference is even higher in favour of the $fFDR^+$. The results are consistent regardless of the number of funds in the sample, the structure of the panel data and the dependence of the cross-sectional error terms.

In an empirical setting, the informative covariates are estimated quantities. This is translated to an estimation noise that may affect the power of our procedure. Our simulations reveal that our method is robust in terms of power up to moderate to high estimation noise.

1.5.1 False discovery rate control of $fFDR^+$

For varying targets of FDR $\in \{5\%, 10\%, \ldots, 90\%\}$, we implement the simulation procedure in Section 1.4 with balanced panel data. Figures 1.2, 1.3 and 1.4 exhibit our results for the generated data under cross-sectional independence.

In Figure 1.2, we show our results for the discrete distribution (1.15) for varying α^*. The upper three subplots correspond to $\pi^+ = 0.1$, whereas the lower three subplots to $\pi^+ = 0.13$. From left to right, the ratio π^-/π^+ increases from 1.5 to 6 (with the null proportion π_0 decreasing accordingly). For example, the top-left subplot exhibits the actual FDR (vertical axis) and the given targets of FDR (horizontal axis) with the alphas drawn from a discrete population of which 75%, 10% and 15% are, respectively, zero-, positive- and negative-alpha funds. A point on or below the 45°-line indicates that the $fFDR^+$ controls FDR well for the given level; this is the case for $\alpha^* = 1.5$ at all the FDR targets. For $\alpha^* = 3.5$, the FDR is slightly not met for targets in the interval $(0.1, 0.8)$. In general, we witness slight failure of the $fFDR^+$ to control for FDR when

Figure 1.2: Performance of $fFDR^{+}$ for discrete distribution of α. The graphs show the performance of the $fFDR^{+}$ in terms of FDR control when alphas are drawn from a discrete distribution. The simulated data are balanced panels with cross-sectional independence.

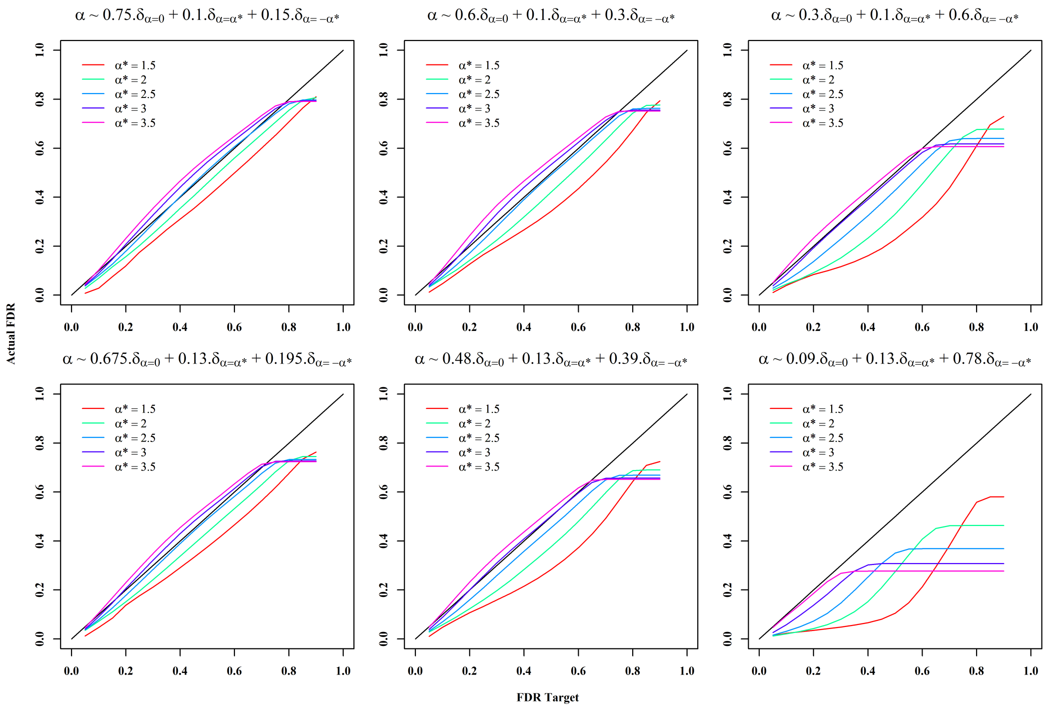
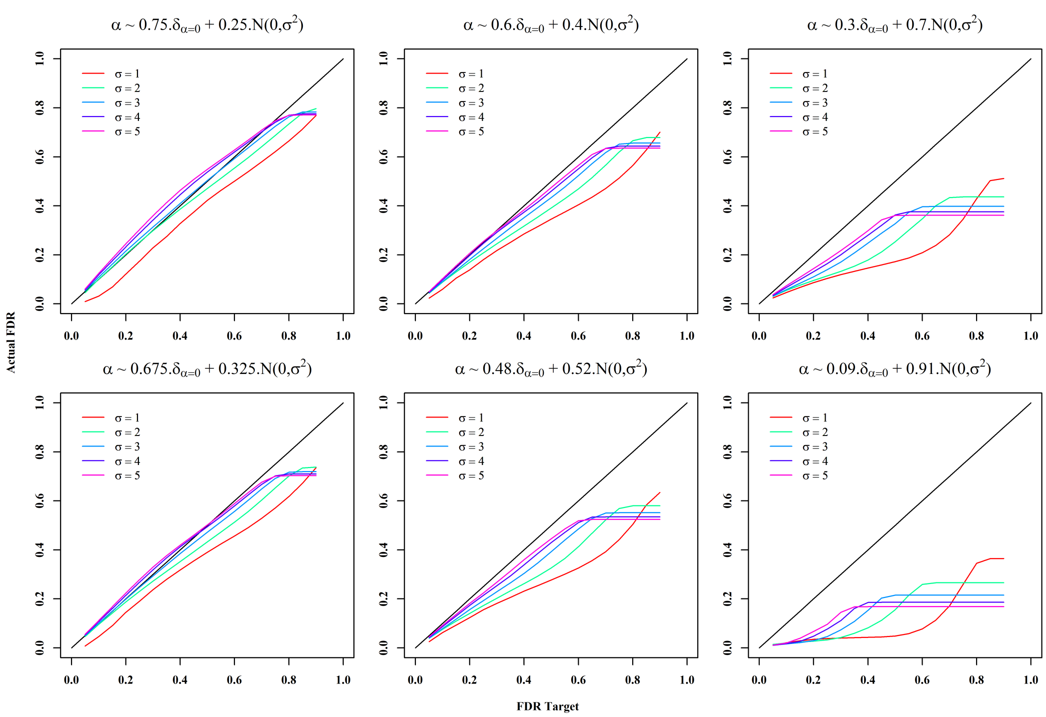


Figure 1.3: Performance of $fFDR^{+}$ for discrete and normal distribution mixture of α. The graphs show the performance of the $fFDR^{+}$ in terms of FDR control when alphas are drawn from a mixture of discrete and normal distributions. The simulated data are balanced panels with cross-sectional independence.

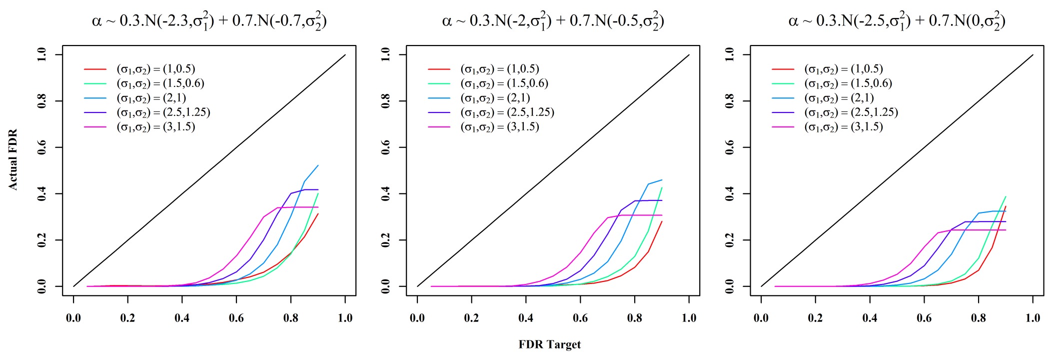


α^{*} is abnormally high. In the last case with smallest π_0, the FDR is controlled well.

In Figure 1.3, we study the case of the fund alpha population described by the mixed discrete-continuous distribution (1.16). We organize our results based on the same null proportions π_0 as in Figure 1.2 and present these for varying σ. We observe that the FDR target is slightly unmet only for extreme values of σ when the null proportion is very high and this effect is also milder compared to the discrete distribution cases.

Finally, in Figure 1.4, we report the results for the continuous distribution (1.17) for varying μ or (μ_1, μ_2) and σ or (σ_1, σ_2). We find that the $fFDR^+$ controls FDR well at all targets.

Figure 1.4: Performance of $fFDR^+$ for continuous distribution of α. The graphs show the performance of the $fFDR^+$ in terms of FDR control when alphas are drawn from a continuous distribution which is a mixture of two normals. The simulated data are balanced panels with cross-sectional independence.



In summary, our simulations are based on proposed fund alpha distributions from the recent literature, from the least realistic cases, with all the out-performing and under-performing funds assumed to have the same mass alpha value, to the more realistic ones, where the alpha is drawn from a continuous distribution, in which no fund has exact zero but rather mostly negative alpha. Our results suggest that, for the continuous distribution, the proposed $fFDR^+$ approach controls well for FDR at any given target.

In the Appendix A.2 we repeat the exercise for balanced data under cross-sectional dependence and unbalanced data under both cross-sectional independence and dependence. Our findings remain robust.

1.5.2 Power analysis

Next, we study the power of our $fFDR^+$ approach in detecting truly positive alpha funds, calculated as described in Section 1.4, and compare it with the FDR^+ of BSW for FDR control at 10%. Although the magnitude of our results varies with different targets of FDR, our main conclusion of the power superiority of the $fFDR^+$ remains.

In Panel A of Table 1.1, we report for the discrete distribution (1.15). For $(\pi^+, \pi_0, \pi^-) = (10, 75, 15)\%$ with highest π_0 and smallest $\alpha^* = 1.5$, both the $fFDR^+$ and FDR^+ achieve similar powers, i.e., 0.3% and 0.4%, respectively. This is expected in this particular case as the number and magnitude of the true positive alphas are small, while we are controlling for FDR at 10%.[17] The superiority of the $fFDR^+$ is more perceptible and stabler for larger α^*. This discrepancy depends not only on the magnitude and proportion of positive alphas, but also on the proportion of zero alphas. This is because both procedures use the null proportion (π_0 in FDR^+ and $\pi_0(z)$ in $fFDR^+$) to estimate the FDR. With the same magnitude and proportion of positive alphas, the small proportion of zero alphas implies the higher power of both the $fFDR^+$ and FDR^+. The effect of the null proportion on the gap of $fFDR^+$ over FDR^+ is stronger when the magnitude of positive alphas is not too high. The gap varies by case and may even exceed 30% (when $\pi^+ = 10\%, \pi_0 = 30\%$ and $\alpha^* = 2.5$).[18]

Panel B exhibits the power upshots for the case of the fund alpha population described by the distribution mixture (1.16). This implies the dependence of the proportion and magnitude of positive alphas on the proportion of the zero-alpha funds and the σ value for non-zero alphas. We expect a higher power for both methods for a smaller zero-alpha proportion and/or a higher value of σ. We find that the $fFDR^+$ is more powerful than FDR^+. More specifically, for the balanced data under cross-sectional independence and $\pi_0 = 75\%$, the power of the $fFDR^+$ (FDR^+) increases from 0.3% to 60.8% (0.2% to 52.2%) with increasing σ from 1 to 5. For given, say, $\sigma = 2$, the power of the $fFDR^+$ (FDR^+) increases from 15.4% to 38% (8.2% and 22%) with reducing

[17] As will be shown later, with a higher FDR target (such as 30%), the power of the $fFDR^+$ exceeds that of FDR^+ by 6%. Considering a higher target than 10% is sensible for trading and diversification purposes as otherwise very few or no out-performing funds are selected. In the study of BSW, the estimated FDR in the empirical application is about 41.5% on average (depending on portfolio).

[18] As shown in the Appendix A.2, the relevant reports vary slightly when the simulated data are generated with alternative forms of $\pi_0(z)$ mentioned in footnote 16, with unbalanced panel or with cross-sectional dependence, however the overall picture remains the same.

π_0. The gap is generally evident for $\sigma > 1$ with power differences around 10% but which can also reach up to 16%.

Finally, in Panel C, we study the outcome from using the mixture of normals (1.17)

Table 1.1: Performance comparison in terms of power (%). The table compares the power of the $fFDR^+$ and FDR^+ at FDR target of 10% when the alphas of 2,000 funds are drawn from a discrete distribution, i.e. $\alpha \sim \pi^+\delta_{\alpha=\alpha^*} + \pi_0\delta_{\alpha=0} + \pi^-\delta_{\alpha=-\alpha^*}$ (Panel A), a discrete-normal distribution mixture, i.e. $\alpha \sim \pi_0\delta_{\alpha=0} + (1-\pi_0)\mathcal{N}(0,\sigma^2)$ (Panel B), and a mixture of two normal distributions, i.e. $\alpha \sim 0.3\mathcal{N}(\mu_1,\sigma_1^2) + 0.7\mathcal{N}(\mu_2,\sigma_2^2)$ (Panel C) with various setting of parameters. The simulated data are a balanced panel with 274 observations per fund and are generated with cross-sectional independence.

Panel A: discrete distribution.						
(π^+,π_0,π^-)	Procedure	$\alpha^*=1.5$	$\alpha^*=2$	$\alpha^*=2.5$	$\alpha^*=3$	$\alpha^*=3.5$
$(10,75,15)\%$	$fFDR^+$	0.3	5.1	21.8	45.7	67.3
	FDR^+	0.4	2.1	12.1	32.3	53.5
$(10,60,30)\%$	$fFDR^+$	1.1	10.3	33.1	58.5	77.5
	FDR^+	0.4	2.3	13.8	35.9	57.4
$(10,30,60)\%$	$fFDR^+$	3.5	22.9	52.9	76.6	89.7
	FDR^+	0.4	3.3	21.4	47.8	69.6
$(13,67.5,19.5)\%$	$fFDR^+$	0.8	8.8	30.1	55.1	75.1
	FDR^+	0.4	3.1	17.6	39.7	60.9
$(13,48,39)\%$	$fFDR^+$	2.3	16.4	43	68.1	84.3
	FDR^+	0.5	4	21.8	46.1	66.8
$(13,9,78)\%$	$fFDR^+$	6.4	34	67.6	89.2	97.5
	FDR^+	0.5	6.9	37.2	69.2	88
Panel B: discrete-normal distribution mixture.						
π_0	Procedure	$\sigma=1$	$\sigma=2$	$\sigma=3$	$\sigma=4$	$\sigma=5$
75%	$fFDR^+$	0.3	15.4	36.1	51.1	60.8
	FDR^+	0.2	8.2	26.7	41.7	52.4
60%	$fFDR^+$	1.2	21.6	42.8	57.1	66.1
	FDR^+	0.2	11.4	31.5	46.6	56.9
30%	$fFDR^+$	4	31.6	54	67.2	74.8
	FDR^+	0.4	17.5	40.5	55.6	65.4
67.5%	$fFDR^+$	0.8	18.9	40	54.5	63.7
	FDR^+	0.2	9.9	29.6	44.5	55
48%	$fFDR^+$	2.3	25.9	47.8	61.6	70.4
	FDR^+	0.3	13.9	35.4	50.5	60.5
9%	$fFDR^+$	6	37.9	60.6	73.6	80.9
	FDR^+	0.5	22	47.1	62.7	72.2
Panel C: mixture of two normal distributions.						
				(σ_1,σ_2)		
(μ_1,μ_2)	Procedure	$(1,0.5)$	$(1.5,0.6)$	$(2,1)$	$(2.5,1.25)$	$(3,1.5)$
		$\pi^+=6\%$	$\pi^+=10.4\%$	$\pi^+=20.7\%$	$\pi^+=25.5\%$	$\pi^+=29.1\%$
$(-2.3,-0.7)$	$fFDR^+$	0	0.3	4.5	12.9	22.5
	FDR^+	0	0	0.3	1.9	7.1
		$\pi^+=11.8\%$	$\pi^+=16.9\%$	$\pi^+=26.4\%$	$\pi^+=30.5\%$	$\pi^+=33.4\%$
$(-2,-0.5)$	$fFDR^+$	0	0.4	5.9	15.1	24.8
	FDR^+	0	0.1	0.4	2.9	9
		$\pi^+=35.2\%$	$\pi^+=36.4\%$	$\pi^+=38.2\%$	$\pi^+=39.8\%$	$\pi^+=41.1\%$
$(-2.5,0)$	$fFDR^+$	0.1	0.6	8.3	17.8	27.6
	FDR^+	0	0	0.6	4.2	11.4

with $\pi_1 = 0.3$, $\pi_2 = 0.7$ and non-positive means (μ_1, μ_2) to limit the likelihood of a positive alpha. The proportion of positive alphas ranges from 6% to 41.1%. For small (σ_1, σ_2) values, the positive alphas are also small in magnitude and, consequently, the power is negligible. When (σ_1, σ_2) are higher than $(2, 1)$, the power of both methods as well as their discrepancy increase significantly and favourably for $fFDR^+$ reaching up to 16%.

Our analysis has shown that, when controlling for FDR at an as low level as 10%, both the $fFDR^+$ and FDR^+ have good power for large (in magnitude) alphas. When this happens, the gain in power of the $fFDR^+$ over FDR^+ can vary depending on the underlying fund alpha distribution: 10% to 16% (continuous distribution) and 20% to 30% (discrete distribution). On the other hand, when the zero-alpha proportion is high and the proportion and magnitude of positive alphas is small, the power of both methods reduces.

Finally, as we demonstrate in the Appendix A.2 that our conclusions are not affected by the data structure (balanced versus unbalanced panel) or dependencies.

1.5.3 Power and FDR trade-off

In what follows, we study the impact on power when controlling for FDR at different (higher than 10% level) targets. Our results show clear differences between the $fFDR^+$ and FDR^+ and, in support of the former, even for cases of negligible power for a 10% target. Constructing mutual fund portfolios at higher FDR levels is sensible as otherwise we may end up with empty portfolios. Investors have to face a trade-off between the power in detecting out-performing funds and the FDR threshold, which we discuss next.

We focus on cases of very low power when the FDR is controlled at 10%. For brevity, we present in Table 1.2 our results for only balanced data under cross-sectional independence and FDR targets up to 90%, noting that these are largely unchanged for unbalanced data. In particular, for the underlying discrete fund alpha distribution, the $fFDR^+$ gains rapidly power with increasing FDR targets, peaking at 40% in excess of the FDR^+ when the target is set at 70%. For the continuous distribution, the power of the FDR^+ changes very slowly and is persistently negligible (mixture of normals) even for FDR controlled at 90%. On the other hand, the $fFDR^+$ detects abundant positive

Table 1.2: Power comparison (in %) for varying FDR targets (%). The table presents some selected cases of low powers of the $fFDR^+$ and FDR^+ at FDR target of 10%. We consider a discrete distribution: $\alpha \sim 0.75\delta_{\alpha=0} + 0.1\delta_{\alpha=1.5} + 0.15\delta_{\alpha=-1.5}$; a discrete-normal mixture: $\alpha \sim 0.75\delta_{\alpha=0} + 0.2\mathcal{N}(0, 1.5^2)$; and a two-normal mixture: $\alpha \sim 0.3\mathcal{N}(-2.3, 1^2) + 0.7\mathcal{N}(-0.7, 0.5^2)$. The simulated data are balanced panels with cross-sectional independence.

		FDR target								
Distribution	Procedure	10	20	30	40	50	60	70	80	90
Discrete	$fFDR^+$	0.3	2.5	8	18.1	32.3	48.5	64.3	76.3	85
	FDR^+	0.4	0.9	2	3.9	7.4	14	24.7	41.5	65.1
Mixture of discrete and normal	$fFDR^+$	0.3	1.3	3.2	6.5	11.8	19.8	31.3	46.3	64.1
	FDR^+	0.2	0.4	0.7	1.1	1.7	2.7	4.9	10.4	26.5
Mixture of normals	$fFDR^+$	0	0.1	0.4	1.2	2.7	5.9	11.7	21.3	35.3
	FDR^+	0	0	0	0.1	0.1	0.1	0.1	0.1	0.1

alpha funds with a power that can reach up to 46% in excess of the FDR^+ (mixture of two normal distributions with 90% target).

1.5.4 Varying the number of observations and funds

Hitherto, we have assumed a sample with $m = 2{,}000$ funds, which reflects our actual dataset for the whole period from 1975 to 2019. When constructing a portfolio, we usually use sub-periods of five years and the number of alive funds in these sub-periods naturally varies. In this section, we investigate the impact of varying number of observations T per fund and the number of funds m on the power.

In Table 1.3, we present the outcomes for different underlying distributions of fund alphas, when we control FDR at a 10% target and use balanced panel data with cross-sectional independence. We vary m from 500 to 3,000 and T from 120 months (i.e., 10 years) to 420 months (i.e., 35 years). It is evident from the reports that the power of the $fFDR^+$ increases at a much faster pace with increasing T. With rising m, the power of the $fFDR^+$ slightly decreases, whereas such is observed for the FDR^+ mainly in Panel C. This is not a substantial concern though, as in reality we do not have a very large number of alive funds in a given sub-period. Overall, the power of the $fFDR^+$ in excess of the FDR^+ can reach 30%.

Apparently for $T = 120$, both procedures have low power. Empirically, when constructing a portfolio of mutual funds, we usually use in-sample sub-periods of 5 years. In these cases, the investors may have to raise the FDR target to a higher level

Table 1.3: Power comparison (in %) for varying sample size and observation length. The table compares the power of the $fFDR^+$ and FDR^+ in a balanced panel data with varying number of observations per fund (T) and number of funds (m). We present three cases where alphas of m funds are drawn from i) discrete distribution: $\alpha \sim 0.1\delta_{\alpha=2} + 0.3\delta_{\alpha=0} + 0.6\delta_{\alpha=-2}$ (Panel A); ii) discrete-normal mixture: $\alpha \sim 0.3\delta_{\alpha=0} + 0.7\mathcal{N}(0, 2^2)$ (Panel B); and mixture of two normal distributions: $\alpha \sim 0.3\mathcal{N}(-2, 2^2) + 0.7\mathcal{N}(-0.5, 1)$ (Panel C). The simulated data are balanced panels with cross-sectional independence.

		Number of observations per fund					
m	Procedure	$T = 120$	$T = 180$	$T = 240$	$T = 300$	$T = 360$	$T = 420$
		Panel A: Discrete distribution					
500	$fFDR^+$	2.7	8.5	19.6	31.8	44.6	54.8
	FDR^+	0.6	1.4	3	5.3	10.6	18.4
1000	$fFDR^+$	1.5	6	16.3	29.4	42.4	52.9
	FDR^+	0.4	0.8	2.1	4.9	10.6	19.1
2000	$fFDR^+$	1.2	5.7	15.4	28	40.6	51.4
	FDR^+	0.2	0.6	1.5	4.8	11.2	20.4
3000	$fFDR^+$	1.1	5.4	15	27.6	39.3	50.8
	FDR^+	0.2	0.5	1.6	4.9	11.8	20.7
		Panel B: Mixture of Discrete and Normal distributions					
500	$fFDR^+$	12.4	21.3	29.1	35.2	40.5	44.9
	FDR^+	2.4	7.5	14.1	20	25.3	29.8
1000	$fFDR^+$	11.7	21	28.1	34.7	40	44.5
	FDR^+	2.1	7.8	14.1	20.1	25.2	29.7
2000	$fFDR^+$	11.4	20.5	28.1	34.1	39.3	43.7
	FDR^+	2.2	7.9	14.2	19.9	25.1	29.7
3000	$fFDR^+$	11.2	20.4	27.8	33.9	39	43.6
	FDR^+	2.3	8	14.1	20	25.2	29.7
		Panel C: Mixture of Normal distributions					
500	$fFDR^+$	1.3	3	5.3	8	10.9	13.4
	FDR^+	0.2	0.3	0.5	0.8	1.3	1.8
1000	$fFDR^+$	0.9	2.4	4.8	7.6	10.1	12.8
	FDR^+	0.1	0.2	0.4	0.6	1.1	1.6
2000	$fFDR^+$	0.7	2.2	4.5	6.9	9.6	12
	FDR^+	0.1	0.1	0.3	0.5	1	1.6
3000	$fFDR^+$	0.7	2.2	4.3	6.8	9.3	11.9
	FDR^+	0	0.1	0.2	0.4	0.9	1.5

as explained in the previous section.[19] In Table 1.4, we focus the spotlight on (small) $m = 500$ and $T = 60$ (i.e., 5 years). It is shown there that both methods yield even lower power at the FDR target of 10%. By increasing the target, the power of the $fFDR^+$ in detecting out-performing funds rises faster than that of the FDR^+, especially for the discrete and mixed normal distributions.

[19]In fact, in order to construct non-empty FDR^+ portfolios based on five-year in-samples, BSW introduce a procedure where they allow the estimate of FDR^+ to be above 70% for several years.

Table 1.4: Power comparison (in %) for varying FDR targets for sample with small size and small number of observations. In this table, we consider three distributions as in Table 1.3 for samples consisting of $m = 500$ funds (balanced panels with cross-sectional independence) with $T = 60$ observations per fund (5 years).

		FDR target								
Distribution	Procedure	10	20	30	40	50	60	70	80	90
Discrete	$fFDR^+$	0.5	2.2	5.8	12.2	20.9	30.8	41.5	53.5	66.3
	FDR^+	0.2	0.5	0.7	0.9	1.3	1.7	2.1	2.6	3.6
Mixture of discrete and normal	$fFDR^+$	2.4	7.4	14.4	23	32.7	42.9	53.2	63.5	68.4
	FDR^+	0.4	0.9	1.6	3	5.6	10.4	18.9	32.2	47.3
Mixture of normals	$fFDR^+$	0.2	1	2.9	6.2	11.1	18	26.7	37.5	51
	FDR^+	0.1	0.1	0.2	0.3	0.4	0.5	0.8	1	1.5

1.5.5 Estimation errors in the covariate

In the main settings of simulations, we consider a simple covariate where in the set of *non-zero* alpha funds, the ranking of funds' alpha is the same as that of funds' covariate. This does not hold in the whole population. Put differently, one cannot simply rank the funds based on a covariate to distinguish the out-performing ones from the zero-alpha and the under-performing ones. In this section, we further study the behaviour of our $fFDR^+$ approach by adding to the original covariate a noise reflecting potential estimation biases, as all covariates in the real data are calculated based on a certain sample period. More specifically, instead of using the covariate Z as in our previous simulations, we use $Z' = (z'_1, \ldots, z'_m)$ given by

$$z'_i = z_i + \eta_i, \tag{1.21}$$

where η_i denotes the noise and is generated independently from a normal distribution $N(0, \sigma_\eta^2)$. Alternatively Z' can be viewed as a realization of some informative covariate which aims to capture Z. Depending on the scale of the estimation error, the realized covariate could have different levels of information. We do not know actual estimation errors in covariates in reality. Thus, we simulate low to high noise in our covariates. More specifically, we consider two different values of σ_η including $\sigma_1 = 0.5/\sqrt{12}$ and $\sigma_2 = 1/\sqrt{12}$. These values are based on the fact that the covariate $Z \sim U[0, 1]$, which has a standard deviation of $1/\sqrt{12}$. We confirm that the $fFDR^+$ controls well for the FDR in this setting and the figures are virtually the same as those presented in the previous sections in the original setting. This is the most important characteristic of

$fFDR^+$ we should expect, that is, ability to control well for the risk even when the new information contains noise.

In Table 1.5 we provide further information by presenting the power (at FDR target of 10%) of the $fFDR^+$. Comparing with Table 1.1, the power is lower but still remarkably higher than that of the FDR^+ with a varying gap across cases of the alpha distribution and the choice of σ_η. As will be shown in our empirical analysis, the $fFDR^+$ with use of each covariate gains significant power over the FDR^+. Therefore, we could assume that covariates in our application have relatively less noise than ones in this simulation.

Table 1.5: Power (in %) of $fFDR^+$ under noised covariates. The data are generated as in tables 1–3 except the use of a new covariate containing a noise: $Z' = Z + \eta$ instead of Z. The noise is drawn independently form normal distribution $\eta \sim N(0, \sigma_\eta^2)$ where σ_η taking value in $\{\sigma_1 = 0.5/\sqrt{12}, \sigma_2 = 1/\sqrt{12}\}$.

Panel A: Discrete distribution.										
	$\alpha^* = 1.5$		$\alpha^* = 2$		$\alpha^* = 2.5$		$\alpha^* = 3$		$\alpha^* = 3.5$	
(π^+, π_0, π^-)	σ_1	σ_2	σ_1	σ_2	σ_1	σ_2	σ_1	σ_2	σ_1	σ_2
(10,75,15)%	0.3	0.3	4.7	3.9	19.6	16.9	41.9	37.5	63.7	59.3
(10,60,30)%	1	0.7	8.7	6.5	28	23.1	52	45.7	71.8	66.3
(10,30,60)%	2.6	1.5	16.4	12	43.7	36.1	69.8	61.8	85.7	79.9
(13,67.5,19.5)%	0.7	0.6	8.2	6.7	27.5	23.6	50.8	45.9	71.2	66.6
(13,48,39)%	1.9	1.3	14	10.7	38.2	32.1	62.8	56	80.6	75.2
(13,9,78)%	5.1	3.3	27.8	21.7	62.3	55.2	87.6	82.4	96.6	94.2
Panel B: Mixture of a discrete and a normal distributions.										
	$\sigma = 1$		$\sigma = 2$		$\sigma = 3$		$\sigma = 4$		$\sigma = 5$	
π_0	σ_1	σ_2	σ_1	σ_2	σ_1	σ_2	σ_1	σ_2	σ_1	σ_2
75%	0.2	0.1	14	11.9	33.9	31.7	48.3	46.4	58.1	56.4
60%	0.6	0.3	19.3	16.4	39.6	36.9	53.8	51.3	62.5	60.5
30%	2.2	1.2	28.2	23.9	49.2	45.4	62	59	70.5	68
67.5%	0.4	0.2	16.8	14.3	36.8	34.4	51.2	49	60.7	58.7
48%	1.2	0.7	22.9	19.4	43.6	40.3	57.1	54.4	65.6	63.3
9%	3.6	2.1	34	29	56.1	51.7	68.5	65.4	75.9	73.8
Panel C: Mixture of two normal distributions.										
	(σ_1, σ_2)									
	$(1, 0.5)$		$(1.5, 0.6)$		$(2, 1)$		$(2.5, 1.25)$		$(3, 1.5)$	
(μ_1, μ_2)	σ_1	σ_2	σ_1	σ_2	σ_1	σ_2	σ_1	σ_2	σ_1	σ_2
(-2.3,-0.7)	0	0	0	0	1.2	0.6	6.5	4.2	15	11.2
(-2,-0.5)	0	0	0	0	2.1	1.1	8.9	6	18.2	13.9
(-2.5,0)	0	0	0.1	0.1	4.3	2.4	12.2	8.3	22.1	17.1

Concluding this section, we recollect that the simulated power of $fFDR^+$ in detecting out-performing funds is found to be larger than FDR^+'s. This persists for different fund alpha distributions, balanced and unbalanced data, cross-sectional dependence of error terms accounted for or not. This power advantage depends on the magnitude and proportion of positive alphas as well as the proportion of zero alpha in the population, the number of funds in the sample, estimation errors in the covariates, and the average number of observations per fund. When the last factor is small, leading to a diminished power for both procedures, we can uplift the FDR to a reasonable level so that the $fFDR^+$ detects remarkable number of out-performing funds. In our empirical application of the next section, we show how the investors can benefit from this.

1.6 Empirical results

1.6.1 Five covariates proposed in the literature

We start our empirical investigation of the $fFDR^+$ approach by considering five covariates that may convey information about the performance of mutual funds. They are shown to be persistent and, therefore, can predict the performance of mutual funds. We also propose four new covariates based on asset pricing models.

First, we study the R-square of Amihud and Goyenko (2013), which is estimated from the Carhart four-factor model and measures the activeness of a fund. If a fund replicates the market, the R-square will be close to one; if, instead, it is more active, it will have a small R-square and in this case, according to the authors, funds tend to perform better.

The second covariate is the Fund Size of Harvey and Liu (2017). This takes into account both the fund size, which is the total net assets under management (TNA) of a fund, and the industry size, which is the total assets under management of all active mutual funds in the sample (sum of TNA). More specifically, for fund i at time t, it is defined as

$$\text{Fund Size}_{i,t} = \ln \frac{\text{TNA}_{i,t}}{\text{IndustrySize}_t} - \ln \frac{\text{TNA}_{i,0^*}}{\text{IndustrySize}_{0^*}}, \tag{1.22}$$

where $t = 0^*$ corresponds to the time of the first TNA observation in our sample. The Fund Size reflects the growth in scale of a fund relative to the whole active mutual fund market. Harvey and Liu (2017) show a significant negative relationship between Fund Size and funds' performance.[20]

[20]Pastor *et al.* (2015) and Chen *et al.* (2004) as well as Zhu (2018), respectively, argue that the

The third covariate is the Return Gap of Kacperczyk *et al.* (2008), which is intended to reflect the unobserved actions of the funds. Mutual funds usually disclose their portfolio holdings and return periodically, e.g., quarterly or semi-annually. The investors are unaware of the funds' trading activities in the period of consecutive reports. The Return Gap of a fund is defined as the difference between the return that is disclosed by the fund and the return that the fund would have based on disclosure of its last portfolio holdings. Kacperczyk *et al.* (2008) show that the funds' performance can be predicted by their past return gaps; mutual funds with higher past return gap tend to perform better in the future.

Our fourth covariate is the Active Weight of Doshi *et al.* (2015), which aims to gauge the fund's activeness level and is given by the sum of the absolute differences of the stock value weights and the actual weights that the fund assigns to the stocks in its portfolio holdings. In their research, they show that funds with higher active weight tend to perform better. To obtain meaningful values for the active weight and the return gap, as in Kacperczyk *et al.* (2008) and Doshi *et al.* (2015), we require each mutual fund to hold at least 10 stocks in its portfolio at any time.

The fifth covariate is the Fund Flow. The interaction of fund flow and funds' performance has been studied quite extensively such as in Sirri and Tufano (1998), Berk and Green (2004), Harvey and Liu (2017) and Capponi *et al.* (2020), among others. Zheng (1999), in particular, discovers that funds receiving money perform better than those that lose money. The author also shows that investors can earn abnormal returns by using small funds' flow information. Here, we follow Bris *et al.* (2007) and define Fund Flow at time t as

$$\text{Fund Flow}_t = \frac{\text{TNA}_t - (1 + r_t)\text{TNA}_{t-1}}{(1 + r_t)\text{TNA}_{t-1}}, \tag{1.23}$$

where r_t is the return of the fund in the period $t - 1$ to t.

In addition to the aforementioned well-known covariates, we propose four new co-

industry size and the fund size (approximated by the logarithm of the fund's TNA) have a negative impact on the funds' performance. We use the Fund Size of Harvey and Liu (2017) as it incorporates information of both covariates. Other studies on the relationship between fund size and performance and funds' holding liquidity (e.g., Yan, 2008) or funds' merger (i.e., McLemore, 2019) document the same conclusion.

variates that are based on asset pricing models and are available for all funds in our sample. These are the Sharpe ratio, the Beta and Treynor ratio obtained from the Capital Asset Pricing Model, and the idiosyncratic volatility (Sigma) of the Carhart four-factor model. The Sharpe and Treynor ratios are risk-adjusted performance measures of funds, whereas the Beta and Sigma reflect systematic and idiosyncratic risk, respectively. These metrics reveal aspects of the past mutual funds' performance and, thus, may assist in identifying out-performing and under-performing funds. Asset pricing metrics are regularly used by wealth managers and academics in the fields of trading, asset pricing and investors' performance, but are overlooked in the mutual funds literature.[21]

1.6.2 The FDR^+ and $fFDR^+$ portfolios

In this section, we illustrate how $fFDR^+$ helps to identify out-performing mutual funds using a portfolio approach following BSW. More specifically, at the end of year t, we select a group of funds to invest in year $t+1$ based on historical information from the last five years ($t-4$ to t). In order to implement $fFDF^+$ and FDR^+, we require the observed values of the covariates of each fund, the estimated alpha and the p-value of each test. We execute, first, the Carhart four-factor model over the 5-year period to estimate the alpha.

The informative value of the Return Gap, Active Weight, Fund Flow and Fund Size on funds' performance is persistent, i.e., the choice between using the most recent (final-year) observations for these covariates or their average values over the whole in-sample (five years) is of less importance, as demonstrated by our robustness check in Appendix A.9.[22] Although the predictability of the covariates may last for a long horizon of up to five years, we expect their informative values to decrease with time; hence, forming portfolios based on their recent realizations is preferred to their average values of the whole last five years' time. Because of this, Return Gap, Active Weight, Fund Flow and Fund Size are calculated based on data in the final year of the in-sample (i.e., we use

[21]For instance, Clifford *et al.* (2021) study the relation between idiosyncratic volatility and mutual funds flows but they do not focus on using this informative covariate as a factor for funds selection.

[22]Readers may refer to Kacperczyk *et al.* (2008), Doshi *et al.* (2015), Zheng (1999) and Harvey and Liu (2017) for the studies of the persistence of the Return Gap, Active Weight, Fund Flow and Fund Size, respectively. It should also be noted, that in our $fFDR$ framework, all covariates are transformed to uniform with only the ranking of the covariates across the funds counting.

the exposure of the fund flow in year t for the Fund Flow, the value at the end of year t for the Fund Size, whereas for the Active Weight and the Return Gap we use their average exposures in year t). The R-square, Sharpe Ratio, Beta, Sigma and Treynor ratio are based on the whole five years. We calculate our p-values in a similar fashion to BSW. For the funds that suffer from heteroskedasticity or autocorrelation, we calculate the t-statistics based on the heteroskedasticity and autocorrelation-consistent standard deviation estimator of Newey and West (1987).[23] For each fund, we implement 10,000 bootstrap replications to estimate the distribution of the t-statistic and subsequently calculate the bootstrapped p-value for the fund.[24]

As required by our method, the p-values of any truly zero-alpha funds, given a covariate value, should be uniformly distributed. Although it is difficult for us to validate this requirement in reality as we never know which funds are truly zero-alpha, it appears intuitive for us to assume that this condition is satisfied. Consider, for example, the R-square. We expect the truly zero-alpha funds to invest randomly in the stock market, thus they should possess an R-square value of roughly equal to one. Conditional on a specific R-square value that a truly zero-alpha fund could have, i.e., close to one, if the fund is truly zero-alpha then its p-value should follow a uniform distribution like any usual true null hypothesis test.[25]

Next, we describe the selection process of out-performing funds to invest in year $t+1$ given an FDR target τ in $(0,1)$. First, we recall the relevant selection process for BSW's "$FDR\tau$" portfolio. For each γ on the grid $\{0.01, 0.02, ..., 0.6\}$, we calculate the $\widehat{FDR}^+_\gamma$ given by (1.13). Then, we find γ^* such that $\widehat{FDR}^+_{\gamma^*}$ is closest to τ; this is the significant threshold for BSW's portfolio, that is, all the positively estimated alpha funds in the in-sample window with p-values $\leq \gamma^*$ will be included in the $FDR\tau$ portfolio. This guarantees the non-empty property of the portfolio but does not always meet the FDR target τ, thereby $\widehat{FDR}^+_{\gamma^*}$ may be much higher than τ.

[23] We check heteroskedasticity, autocorrelation and ARCH effect by using White, Ljung-box and Engle tests, respectively. We see that a half of funds in our sample suffer from at least one of the mentioned effects.

[24] The bootstrapping procedure may result in duplicated bootstrapped p-values. For this, we use an adequate number of replications to reduce that effect and obtain good estimates of $\pi_0(z)$ and $f(p,z)$.

[25] Indeed, the p-value of each test i is defined as $p_i = 1 - F(|t_i|)$, where $F(|t_i|) = \mathbb{P}(|\mathcal{T}_i| < |t_i| | \alpha_i = 0)$ and $\mathcal{T}_i$ is the conventional t statistic of test i and t_i its estimated value. If hypothesis $\alpha_i = 0$ is true, conditional on a specific covariate value, the p-value of test i is uniformly distributed since $\mathbb{P}(P_i < p_i) = \mathbb{P}(1 - F(|\mathcal{T}_i|) < p_i) = \mathbb{P}(|\mathcal{T}_i| > F^{-1}(1-p_i)) = 1 - \mathbb{P}(|\mathcal{T}_i| < F^{-1}(1-p_i)) = 1 - F(F^{-1}(1-p_i)) = p_i$.

Second, we select out-performing funds for a $fFDR$-based portfolio, namely, "$fFDR\tau$". To establish comparable $fFDR\tau$ and $FDR\tau$ portfolios, we implement the $fFDR^+$ (with a particular covariate) to control $pFDR^+$ at a target τ^* that reflects the FDR level controlled by the $FDR\tau$ portfolio but has to be less than one.[26] As the FDR of the $FDR\tau$ portfolio is controlled at level $\widehat{FDR}^+_{\gamma^*}$ which may be greater than one or less than τ, we set: $\tau^* = \tau$ if $\widehat{FDR}^+_{\gamma^*} \leq \tau < 1$; $\tau^* = \widehat{FDR}^+_{\gamma^*}$ if $\tau < \widehat{FDR}^+_{\gamma^*} < 1$.[27] If $\widehat{FDR}^+_{\gamma^*} \geq 1$, we just select all the funds in the $FDR\tau$ portfolio.

For both the $fFDR\tau$ and $FDR\tau$ portfolios, we invest equally in the selected funds in the following year. If a selected fund does not survive for a month during the year, then its weights are redistributed to the remaining (surviving) funds.

As aforementioned, at the beginning of each year we select funds in to a portfolio by using the previous five consecutive years as in-sample. To be eligible for this, a fund needs to have 60 observations in the in-sample. We start constructing our portfolios from December 1981.[28]

1.6.3 Performance comparison

In this section, we assess the portfolios' performance based on their alphas. We demonstrate the advantage of the $fFDR^+$ in picking out-performing funds and the efficient use of the covariates' information. We estimate the alpha evolution and the average alphas of our $fFDR\tau$ portfolios based on the nine covariates and compare with those of the $FDR\tau$ portfolio. We also explore the performance of $fFDR\tau$ portfolios after linearly combining the nine covariates and using their first principal component, an ordinary least squares regression, a least absolute shrinkage and selection operator,

[26]If we implement the $fFDR^+$ and FDR^+ to strictly control FDR at a target, say, $\tau = 10\%$ or $\tau = 20\%$, both result in empty portfolios for many years. With BSW's $FDR\tau$ portfolios, the problem is solved. In BSW's study, for the $FDR10\%$ portfolio, the empirical $\widehat{FDR}^+_{\gamma^*}$ is always greater than 10% with an average of 41.5%. For our data, among the thirty eight times of portfolio construction, with target $\tau = 20\%$ (10%) the $\widehat{FDR}^+_{\gamma^*}$ is less than τ on eight (zero) occasions and greater than one on five occasions for both targets.

[27]We could have set $\tau^* = \widehat{FDR}^+_{\gamma^*}$ for both cases. However, it seems fairer to set $\tau^* = \tau$ if $\widehat{FDR}^+_{\gamma^*} \leq \tau$ since both portfolios initially aim to control FDR at τ.

[28]As Fama and French (2010) point out possible biases in the CRSP mutual fund data before 1984, we conduct a robustness check using a sample from 1984 to 2019; based on our results, presented in Appendix A.3, our conclusions remain unchanged.

a ridge regression and an elastic net.[29]

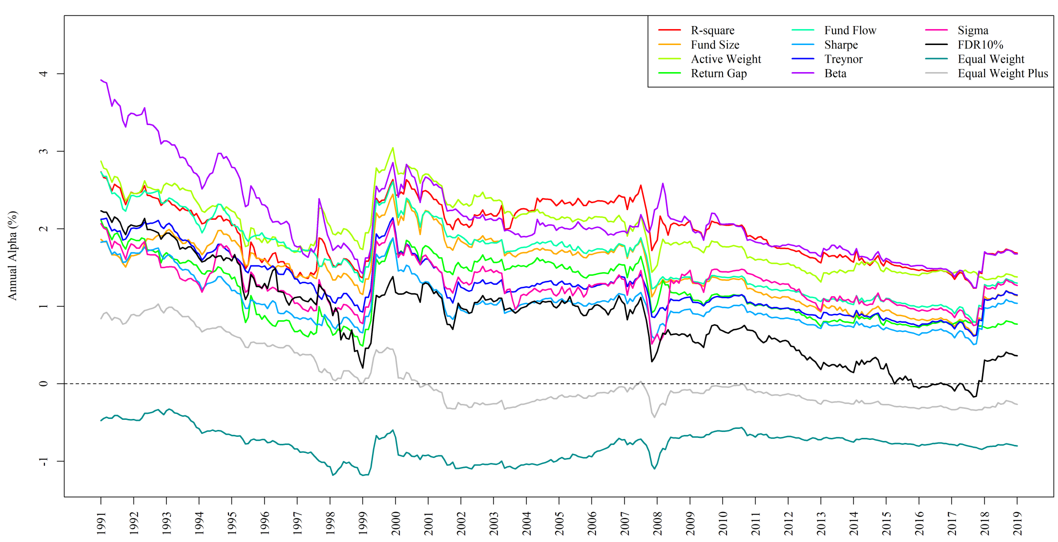
We focus on portfolios with small FDR targets of $\tau = 10\%$. We repeat all estimations with $\tau = 20\%$ in Appendix A.8. Our results remain unchanged for all exercises.

1.6.3.1 The alpha evolution

For each portfolio, we obtain its alpha evolution by calculating the Carhart four-factor alpha using its returns from January 1982 up to the end of each month from December 1991 onwards. In addition to the aforementioned portfolios, we construct two naive benchmark equally weighted portfolios, without control for the FDR: one that simply includes all the mutual funds in the in-sample window to be invested in the following year; and, another that contains only those with positive estimated alphas. We name these two portfolios Equal Weight and Equal Weight Plus.

We present all the alpha evolution in Figure 1.5. It is obvious from it that the $FDR10\%$ portfolio gains higher alphas than the equally weighted portfolio and all the $fFDR10\%$ portfolios outperform the $FDR10\%$. Ultimately, at the end of 2019, the $fFDR10\%$ portfolios with the R-square and Beta covariates are found to be the best with annualized alphas of about 1.7%, followed by the $fFDR10\%$ portfolios with

Figure 1.5: Alpha evolution of $fFDR10\%$ and $FDR10\%$ portfolios over time. The graph presents the evolution of annualized alphas (in %) of the nine $fFDR10\%$ portfolios corresponding to the nine covariates, the portfolio $FDR10\%$ of BSW and the two equally weighted portfolios.



[29] In Appendix A.4 we provide a detailed comparison of all the $fFDR\tau$ portfolios in regard to several trading metrics, whereas in Appendix A.7 the performance in terms of wealth evolution is presented.

the Active Weight, Fund Flow, Sigma, Treynor ratio, Fund Size, Sharpe ratio and Return Gap covariates achieving annualized alphas of at least 0.77%. By contrast, the $FDR10\%$, without the use of covariate information, winds up with a small positive alpha of 0.36%. It is noteworthy that all $fFDR10\%$ and the $FDR10\%$ portfolios seem to rebounce in terms of performance over the last two years of our sample.

1.6.3.2 The average alpha

The alpha evolution in the previous section is calculated based on the portfolio returns from the start of 1982 up to a time point of interest. This may represent limited information in the case of investors with a different investment period of, say, five or ten years. For this, in Table 1.6, we report the average alpha that the investors will gain if they invest for $n \in \{5, 10, 15, 20, 30, 35, 38\}$ consecutive years: for each portfolio, we calculate its "n-year"alpha based on the portfolio returns over a period of $12n$ consecutive months, we repeat by shifting every time one month forward, and eventually present the average alpha. We report the $fFDR10\%$ for each covariate and the $FDR10\%$. We note that the last case, $n = 38$, corresponds to the alphas for the whole period from January 1982 to December 2019 and are the last points in the plots in Figure 1.5.

Table 1.6: Comparison of portfolios' performances for varying time lengths of investing. In this table, we consider 10 portfolios including nine $fFDR10\%$ portfolios corresponding to the nine covariates and the $FDR10\%$ portfolio of BSW. We compare the average alphas of the portfolios that are kept in periods of exactly n consecutive years. For example, consider $n = 5$. For each portfolio, we calculate the alpha for the first 5 years based on the portfolios' returns from January 1982 to December 1986. Then, we roll forward by a month and calculate the second alpha. The process is repeated and the last alpha is estimated based on the portfolios' returns from January 2015 to December 2019. The average of these alphas is presented in the first rows of the table.

	$fFDR10\%$									$FDR10\%$
n	R-square	Fund Size	Active Weight	Return Gap	Fund Flow	Sharpe	Treynor	Beta	Sigma	
5	1.49	0.87	1.24	0.56	0.92	0.57	0.73	1.09	1.19	0.12
10	1.48	0.85	1.18	0.51	0.93	0.65	0.76	1.2	1.06	0.05
15	1.7	0.94	1.4	0.72	1.06	0.79	0.88	1.2	1.09	0.14
20	1.84	1.05	1.59	0.91	1.15	0.91	0.96	1.31	1.17	0.26
25	1.61	0.9	1.36	0.67	0.99	0.8	0.86	1.24	1.09	0.13
30	1.41	0.78	1.23	0.54	0.95	0.78	0.86	1.2	1.01	0.01
38	1.69	1.14	1.38	0.77	1.3	1.04	1.15	1.67	1.27	0.36

We find that the $fFDR10\%$ portfolios outperform the $FDR10\%$ for all considered covariates and for all n. Although these results should be interpreted with caution (some covariates were not well known in the literature at the start of our sample, such as the

Active Weight and the Fund Size which were published in 2015 and 2017, respectively), they do indicate the stability of our approach for different investment horizons.

1.6.3.3 Sub-period performance

In the alpha evolution in Figure 1.5, we note that the performance of our portfolios varies over time. By construction, this figure contain returns which start from January 1982 and are not representative of the recent mutual fund performance. In order to investigate the contribution of the returns in different periods to the performance of the portfolios, we split the whole period into four non-overlapping sub-periods: 1982–1991 (P1), 1992–2001 (P2), 2002–2011 (P3) and 2012–2019 (P4). We repeat the exercise for each sub-period and present in Table 1.7 the average 5-year alpha and alpha of portfolios (with an FDR target $\tau = 10\%$) in the sub-period.

In terms of alphas and average 5-year alphas, it is clear that all the portfolios perform well in the first two sub-periods before suffering a decline in the third sub-period. On P3, we observe negative average 5-year alphas for the $FDR10\%$ portfolio and the $fFDR10\%$ portfolios with Active Weight and Return Gap covariates. On the last sub-period, this decrease continues for $FDR10\%$, whilst all of the $fFDR10\%$ portfolios witness rebounds. We note that all the $fFDR10\%$, except the ones with Return Gap and Active Weight covariates, achieve both positive alpha and average 5-year alpha in all the sub-periods. The t-statistic columns for the whole sub-period alpha, show that most portfolios have significantly positive alphas in the first sub-period. Interestingly, for the Sharpe ratio, we witness the highest reports in the last sub-period (which is also slightly shorter), whereas the lowest ones appear in the third sub-period which covers the global financial crisis of 2007–2008. From the realizations of the equally weighted portfolio, that is, the portfolio that selects all the eligible funds in the in-sample windows and invests them equally in the following year, we infer that the high Sharpe ratio in the final sub-period partially comes from the whole mutual fund market. The Equal Weight Plus portfolio, which invests in all funds with positive estimated alphas in the previous five years, is always better than the Equal Weight one. This simple screening portfolio even outperforms the $FDR10\%$ in the last two sub-periods. The alphas of the $fFDR10\%$ portfolios, by contrast, are nuanced depending on the covariate used; most of them beat the equally weighted one in all the sub-periods and for all the metrics (with notable exceptions of the Active Weight and Return Gap covariates in the third sub-period).

The implications of these results are as follows. First, we note that the R-square, Return Gap, Active Weight, Fund Flow and Fund Size retain their predictive abilities for mutual fund performance in recent years. From the five traditional covariates, the R-square, Fund Size and Fund Flow still have predictive abilities even after their respective publication dates.[30] Our results disagree partly with the findings of Jones and Mo (2021) who argue that published predictors are losing value in the recent period due to increases in arbitrage activities. Second, we note that our four new covariates contain valuable information on mutual funds' performance that in recent years can surpass the conventional covariates in some cases (see, for example, the performance of the $fFDR10\%$ portfolios in P4 with the Sigma and the Return Gap). Third, they further verify that our approach can resolve the identification issues in mutual funds due to noise/luck where other approaches (such as BSW) fail to.

To further support the aforementioned argument on identification issues, we compare the performance of the portfolios formed in the $fFDR$ framework with a traditional sorting portfolio formation. If a covariate has a highly linear relation with the performance of mutual funds, then forming a portfolio based on sorting the funds on the covariate should be sufficient. We construct single- and double-sorting portfolios similarly to Kacperczyk *et al.* (2008) and Doshi *et al.* (2015), and Amihud and Goyenko (2013), respectively.[31]

The performance in terms of alpha of those portfolios from 1982 to 2019 is presented in Table 1.8. Our results show that most of the sorting portfolios, except the Active Weight and Sharpe ratio, have negative or negligible positive alphas at the end of 2019, which contrasts to the assumption of a linear relationship between the covariate and the funds' performance. Obviously, sorted portfolios perform better if they are based on the correct sign of the correlation between the underlying covariate and the funds' future performance.

[30] Appendix A.5 shows that three out of the five covariates still gain significant alphas in the post-published period.

[31] For further details on the construction of these portfolios we refer the reader to Appendix A.6.

Table 1.7: Performance of portfolios in sub-periods. The table displays the performance of the nine $fFDR10\%$ portfolios corresponding to the nine covariates, the $FDR10\%$ and equally weighted portfolios in sub-periods (P1: 1982–1991, P2: 1992–2001, P3: 2002–2011 and P4: 2012–2019) in terms of the average 5-year alpha (annualized, in %), the annualized alpha (in %) of the whole sub-period, the corresponding t-statistic (with use of Newey–West heteroskedasticity and autocorrelation-consistent standard error) and the annual Sharpe ratio.

	Average 5-year alpha				Whole sub-period alpha				Whole sub-period t-statistic				Annual Sharpe Ratio			
Portfolio	P1	P2	P3	P4	P1	P2	P3	P4	P1	P2	P3	P4	P1	P2	P3	P4
R-square	3.18	2.37	1.29	1.97	2.74	2.74	2.14	3.21	2.81	1.64	0.71	1.59	0.65	0.7	0.26	1.37
Fund Size	2.01	1.74	0.23	1.44	1.86	2.27	0.53	3.07	2.18	1.18	-0.29	1.56	0.62	0.61	0.2	1.37
Active Weight	3.01	3.1	-0.48	1.19	2.87	3.11	-0.01	0.56	2.47	1.85	-0.54	0.53	0.65	0.74	0.19	1.17
Return Gap	2.29	0.91	-0.43	0.55	2.11	1.78	0.17	0.09	2.3	1.04	-0.3	0.09	0.6	0.61	0.2	1.12
Fund Flow	2.65	0.73	0.06	1.82	2.73	1.32	0.54	3.44	2.22	0.74	-0.06	1.77	0.66	0.62	0.22	1.42
Sharpe	1.45	0.7	0.57	1.11	1.83	0.87	0.94	2.99	1.97	0.59	0.25	1.46	0.64	0.72	0.25	1.37
Treynor	1.77	0.73	0.62	1.37	2.12	0.98	0.93	3.19	2.03	0.63	0.19	1.61	0.64	0.69	0.24	1.38
Beta	3.52	0.72	0.45	2.02	3.92	1.58	1.33	3.65	2.15	0.64	0.06	1.94	0.65	0.45	0.21	1.43
Sigma	2.19	1.66	1.6	2.36	2.07	1.66	2.03	3.63	1.88	0.91	0.84	1.93	0.59	0.64	0.29	1.38
$FDR10\%$	2.7	0.6	-0.47	-0.35	2.23	1.2	0.09	1.63	2.01	0.83	-0.33	0.69	0.6	0.65	0.19	1.09
Equal Weight	-0.45	-1.65	0.29	-1.56	-0.48	-1.28	0.2	-1.34	-1.11	-1.53	-0.36	-2.65	0.48	0.54	0.23	1.01
Equal Weight Plus	0.76	-0.96	0.26	-0.65	0.84	-1.01	0.4	-0.38	1.17	-1.12	-0.36	-0.62	0.55	0.54	0.21	1.11

Table 1.8: Performance comparison of $fFDR\tau$ portfolios and portfolios based on sorting on covariates (single-sorting) as well as based on both covariates and past alpha (double-sorting). The table shows the portfolios' annual Carhart four-factor alpha (in %) for the period January 1982 to December 2019. At the end of each year, for the single-sorting 10% portfolio, funds are sorted by the covariate. Depending on whether the relationship of the covariate and the fund performance is positive or negative, the funds in the top or bottom 10% are chosen to invest in the following year. For the double-sorting 10% portfolio, the funds chosen in the single-sorting 10% are ranked based on the past five-year alpha and then only 10% of the funds in the top are selected. *Note.* As documented in the literature, the R-square and Fund Size (Fund flow, Return Gap and Active Weight) have a negative (positive) effect on the mutual funds' performance. The single- and double-sorting portfolios constructed based on this assumption appear italicized.

Portfolio	R-square	Fund Size	Active Weight	Return Gap	Fund Flow	Sharpe	Treynor	Beta	Sigma
Panel A: Performance of $fFDR$10% and $fFDR$20% portfolios									
$fFDR$10%	1.69	1.14	1.38	0.77	1.30	1.04	1.15	1.67	1.27
$fFDR$20%	1.84	1.16	1.45	0.82	1.28	1.02	1.10	1.77	1.61
Panel B: Assuming a positive effect of the covariate on performance of the fund									
Single sort 10%	-1.07	-0.64	*-0.63*	*-1.46*	*-1.02*	0.13	-0.07	-2.11	-2.40
Double sort 10%	-1.03	0.03	*1.43*	*-0.40*	*0.33*	0.18	0.44	0.30	0.97
Single sort 20%	-1.17	-0.75	*-0.67*	*-1.15*	*-0.75*	-0.17	-0.28	-1.80	-1.69
Double sort 20%	-0.60	-0.18	*1.15*	*-0.07*	*0.11*	0.01	-0.10	-0.64	-0.53
Panel C: Assuming a negative effect of the covariate on performance of the fund									
Single sort 10%	*-0.89*	*-0.83*	-1.40	-1.45	-1.00	-1.96	-2.28	0.49	-0.50
Double sort 10%	*-1.72*	*0.30*	-1.39	-0.37	0.31	1.86	0.80	0.18	0.47
Single sort 20%	*-0.86*	*-1.01*	-1.14	-1.34	-1.04	-1.49	-1.49	0.21	-0.67
Double sort 20%	*-0.34*	*0.25*	-1.20	0.04	-0.01	0.47	0.16	0.19	-0.03

The portfolios based on $fFDR$ gain significant positive alphas and beat the corresponding sorted portfolios. These results further validate the advantage of our method in exploiting the non-linear relationship of the covariates, luck and funds' performance. The inability of the traditional sorted portfolios, that dominate the related literature, to reflect the predictive value of the covariates under study is thus noteworthy.

In Appendix A.10, we implement an exercise to combine the covariates to a new one via linear regression and shrinkage method. We see that these simple linear combinations of the covariates does not improve the performance of the $fFDR$ based portfolios. This result further supports the assumption of the non-linear relationship between the considered covariates and the performance of mutual funds. As further robustness checks, in Appendix A.11, we demonstrate that our findings are robust with respect to a data subset where we require a minimum of $15 million in TNA for a fund to be considered.

1.7 Concluding remarks

In this chapter, we introduce the $fFDR^+$, a novel multiple hypothesis testing framework, that incorporates informative covariates to raise the power of detecting outperformers, and apply it to mutual fund investing. First, we conduct simulation experiments to assess how well our method performs in controlling FDR and raising power compared to the FDR^+ method of BSW. We then construct empirical portfolios based on our new method and nine covariates. We study five covariates, which, based on earlier contributions, convey information about mutual funds' performance and propose four new ones based on asset pricing models. We show how the admixture of control for FDR and incorporated covariates advances the generation of more positive and higher alphas than a portfolio that controls FDR only or a portfolio based on sorting on the covariate and the past funds' performance.

The implications of our study are both methodological and empirical. The methodological literature in the field of selecting out-performing mutual funds is rich and expanding. In addition to the influential and well-cited study of BSW, other notable contributions are due to Kosowski *et al.* (2006), Andrikogiannopoulou and Papakonstantinou (2016), Harvey and Liu (2020) and Grønborg *et al.* (2021). All these have their merits and the authors present several promising empirical findings. In our study we focus on the FDR, whilst we defer an examination of their power relative to ours to future research. Nevertheless, we ought to note three main distinguishing features of our method. First, it allows the use of more data in the form of informative covariates, whilst the vast majority of others are limited to funds' past returns and their cross dependencies. Second, it is simple to implement and computationally less intensive than some of the most recent ones (e.g., the double bootstrap of Harvey and Liu, 2020). Third, our work can be extended to other problems in which statistical power weighs more than conservatism (i.e., the FDR threshold is higher), such as in the selection of hedge funds and bond funds or the assessment of trading strategies.

The empirical implications of our study are also of interest to academics and practitioners. We demonstrate that the five traditional mutual fund covariates can offer substantial profits in more recent periods. However, the relationship between these covariates, luck and funds' performance is non-linear. To fully exploit them, one should rely on powerful methods that control luck and noise. Our method ensures that. We also introduce four new covariates and find that their performance in our context is strong and surpasses that of traditional covariates; a finding that is expected to be of interest to investment managers who are constantly looking for valuable covariates in portfolio selection. From practical point of view, the five new covariates alongside the R-square provide investors a set of informative covariates that are easily calculated from funds' return and freely available risk factors.

As with any methodological approach, there are caveats with our $fFDR$ procedure. In particular, this requires large datasets and gains higher power as the FDR threshold increases (see Sections 1.5.3 and 1.5.4). This implies that our approach should not be applied in problems which require a small FDR target (i.e., when the risk of a false discovery can lead to disastrous outcomes). As in our context of mutual funds' performance, it is difficult to explore covariates that seem promising (see, for example, the list of covariates studied in Jones and Mo, 2021) but with limited data availability.

We aspire that the $fFDR$ and $fFDR^+$ methods will become essential tools for people confronted by multiple competing factors, funds or models. The fields of finance and economics are extending towards big datasets and the literature is filled with predictors that may have value in economic variables of interest. Our approach can contribute to the evaluation of all these predictors and be a valuable arrow in the quiver of both academics and practitioners.

Chapter 2

Controlling for luck in picking trading strategies

2.1 Introduction

Technical analysis and trading rules are widely used in foreign exchange (FX) trading (Allen and Taylor, 1990; Taylor and Allen, 1992; Menkhoff and Taylor, 2007). Meese and Rogoff (1983) and Chinn and Meese (1995) document that major exchange rates follow a random walk. In contrast, Levich and Thomas (1993) and Neely *et al.* (2009) argue that the profitability of technical trading rules existed in the 1970s and 1980s but declined in the early 1990s, and Neely *et al.* (1997) and Neely (2002) provide evidence showing that such profitability cannot be attributed to systematic risk or government interventions.

However, given that (i) technical trading rules are not theoretically motivated and that (ii) there exists an enormous number of such rules, the profitability of technical analysis in FX markets is likely subject to data snooping issues.[1] Among all tools to mitigate data snooping bias in social sciences, the multiple testing frameworks based on Romano and Wolf (2005), and Bajgrowicz and Scaillet (2012) offer computationally feasible solutions and has thus been widely applied to research questions involving a large number of predictive models.[2] One common feature of the methodologies in this framework is that the rejection criterion *only* depends on information created by predictive models and does not account for external information.

[1] Data-snooping is a statistical bias that appears when a dataset is used more than once, for inference and model selection. It can lead to results that seem statistically significant but are due to luck and misuse of data analysis.

[2] Aiming to control for data-snooping in assessing trading rules, numerous multiple testing procedures have been proposed such as the contributions of Sullivan *et al.* (1999, 2001), White (2000), Hansen (2005), Barras *et al.* (2010), Bajgrowicz and Scaillet (2012) and Hsu *et al.* (2010) among others.

Using the aforementioned multiple testing methods, some recent studies show that technical trading rules' profitability has declined since the 1990s (Qi and Wu, 2006; Hsu *et al.*, 2016). Nevertheless, such a decline pattern is based on null hypotheses being unconditional zero, which corresponds to investors *not* acquiring external information in assessing technical rules' profitability. Despite its prevalence and convenience, such unconditional testing may not be ideal because currency traders are assessing and picking trading rules based on multiple performance metrics. Thus, the value of technical analysis for FX traders may not be appropriately estimated in an unconditional testing framework as has been done in literature.

In this chapter, we propose a new methodology that accounts for more information sources (which can be captured by "informative covariates") in forming the rejection criterion, which enhances the statistical power given the same false discovery rate (FDR) defined by Benjamini and Hochberg (1995) (BH henceforth). We name our method as multivariate functional FDR ($mfFDR$). Conceptually speaking, embedding informative covariates (and new information they carry) in multiple testing enables us to form conditional null hypotheses, in which predictive models' profits are zeros conditional on updated information. Such a conditional setting is more consistent with the rational expectation hypothesis and better captures market participants' time-varying standards.

Via simulations, we show that our $mfFDR$ method performs well in controlling for the FDR under various settings. Its performance in terms of power is impressive and beats that of prior methods with gaps of about 67% and 44%, respectively.[3] In addition, the proposed procedure performs well under weak signal-to-noise data, i.e. the data where the true alternative hypothesis tests have high $p-$values due to high noise. It is also robust under weakly dependent data and when informative covariates are correlated or contain estimation errors.

We implement the $mfFDR$ method to detect genuinely profitable trading rules in a set of more than 21,000 technical trading rules. Using daily data over a maximum

[3]As benchmarks, prior methods include the FDR approach of Storey (2002) and the functional FDR ($fFDR$) in Chen *et al.* (2021a) (CRS henceforth) that allows *only one* informative covariate. These two methods have been applied to stock index and mutual fund performance by Barras *et al.* (2010), Bajgrowicz and Scaillet (2012), and Chapter 1.

of 50 years for 30 U.S. dollar exchange rates, we use our $mfFDR$ method to select rules that outperform zero in a rolling 12-month window. The informative covariates we consider include (i) the auto-correlation of a trading rule's excess return; and (ii) the estimates of the alpha, beta, and R-square from a regression of the trading rule's excess return on the excess return of a passive buy-and-hold strategy or a currency market factor.

We then collect outperforming rules to construct a monthly portfolio for a currency or a basket of currencies and track the performance of these portfolios with and without transaction costs. We find that it gains positive profits in 28 over 30 considered currencies for an extended period from 1973 to 2020 after transaction costs, and 16 among them are statistically significant.

To examine the advantage of the $mfFDR$ method, the number of rules selected by the $mfFDR$ is larger than those based on prior methods ($fFDR$ and FDR). More importantly, we find that the out-of-sample (OOS) performance of the $mfFDR$-based portfolio is better than those based on prior methods in two aspects. First, the $mfFDR$-based portfolio with the use of four covariates beats all $fFDR$-based with the use of every single covariate. Second, using a linear combination of the four covariates, such as the first principal component (PC1) does not improve the performance of the $fFDR$-based approach compared to using individual underlying covariates. Finally, when we construct a larger portfolio, which is based on a pool of all trading rules applied on all 30 currencies (about 635,850 trading rules = 21,195 $\times$ 30), we find that it gains a Sharpe ratio of 1.06 and 0.95 before and after transaction costs, respectively.

These empirical results highlight the value of directly incorporating more information in multiple testing and offer the following insights. First, considering more covariates in the $mfFDR$ enhances the OOS performance of detected out-performing trading rules. Second, the $mfFDR$ outperforms the $fFDR$ with linear combinations of multiple covariates, suggesting that the $mfFDR$ might effectively extract non-linear information among covariates. Third, prior methods based on the unconditional null hypothesis, such as FDR, may underestimate technical trading rules' true predictive ability and profits because their performance is not evaluated with comprehensive information sets.

Our further empirical analyses aim to present the evolution of the profitability of technical analysis. We split our sample into five decades (1973-1980, 1981-1990, 1991-2000, 2001-2010, and 2011-2020) and examine the performance of technical trading rules using the $mfFDR$ in each decade. We find a decent proportion of profitable technical trading rules in the most recent three decades. Those detected out-performing rules are still profitable OOS for the most recent decade. Cross-currency analyses show that the profitable rules distribute significantly across currencies become less commonly in developed currencies since 1990. Cross-category analyses further show that categories of rules perform differently. First, the moving average category is most profitable group overtime both in-sample and OOS. Second, they generate trading signals with different frequency from which we imply useful guidance for traders.

This study contributes to the finance and econometrics literature as follows. From a methodological perspective, we add an important dimension – the conditioning information – to the correction for data snooping biases.[4] Under the rational expectation hypothesis, the performance of predictive models (and associated null hypotheses) should reflect researchers' and industry practitioners' time-varying expectations. The $mfFDR$ methodology we propose allows researchers to utilize comprehensive information from multiple covariates to test conditional null hypotheses that appear to be more realistic to both academia and industry. By designing and implementing suitable Monte Carlo simulations, we illustrate that our $mfFDR$ approach actually controls for FDR under various settings of signal strength. This is robust under weakly dependent data, correlated covariates, and even those covariates with estimation errors. In addition, our simulations suggest that the $mfFDR$ method has higher power than prior methods (FDR and $fFDR$) that do not update for sufficient information.

From an empirical perspective, we perform the most comprehensive study of technical trading rules in FX markets to timely assess the predictability of such rules and provide further insights on FX traders' "obstinate passion" in technical analysis (Menkhoff and Taylor, 2007). Our analyses are based on constructing 635,850 trading rules, using

[4]Some prior literature highlights the data snooping issues, which leads to the development of some methodologies to guard against such biases. While several multiple testing procedures have been proposed in the past (White, 2000; Hansen, 2005; Barras *et al.*, 2010; Bajgrowicz and Scaillet, 2012; Hsu *et al.*, 2010), they only consider unconditional null hypotheses and use information from predictive models' performance metrics.

FX data of all 30 currencies for a long period (some are as long as almost five decades), and implementing large-scale multiple testing for the profitability of technical analysis. As a matter of fact, we find that the number of outperforming technical rules is comparable in the 2001-2010 and 2011-2020 periods. This novel evidence thus challenges the following two prior beliefs or common perceptions in the literature: (i) the profits of technical analysis are illusive and driven by data snooping biases; and (ii) such profits are a short-term phenomenon and only exist in relatively immature markets (e.g., Lo, 2004; Neely *et al.*, 2009).

The chapter is organized as follows. The next section develops the $mfFDR$ framework and designs simulations to show its performance. Section 2.3 presents descriptions of the data and trading rule universe and Section 2.4 is devoted to the trading rule's performance measure. In Section 2.5, we present the empirical results where $mfFDR$-based portfolios are constructed on individual currencies and a basket of 30 currencies. Finally, Section 2.6 concludes the chapter.

2.2 The use of covariates in FDR framework

In this section, we introduce the $mfFDR$ that estimates the false discovery rate as a function of more than one informative covariate. Our approach develops the frameworks in Chapter 1 on the $fFDR$ with a single covariate. First, we present the setting of our method and its implementation in the context of data snooping and FX trading. Second, we conduct simulations demonstrating the value of multiple informative covariates in controlling the false discovery rate and the superior power of the $mfFDR$ compared to the related existing approaches. Third, we validate the performance of our method under certain dependence structures of data which are typical in finance. We confirm that our approach retains its power and control of the false discovery rate when the statistics are weakly dependent and when informative covariates are correlated and contain estimation errors - all these features are common in most financial data and topics.

2.2.1 The multivariate functional false discovery rate ($\boldsymbol{mfFDR}$)

Suppose we have n trading rules, each producing an excess return. Assume that there are covariates that convey information about the performance of each trading rule, which is measured by a metric ϕ. As detailed in the following sections, in this study, we use the Sharpe ratio (SR) as our main performance metric. To assess the performance of each strategy, conditional on the realization of the covariates, we conduct a hypothesis test

$$H_0 : \phi = 0, \qquad H_1 : \phi \neq 0. \tag{2.1}$$

The n trading rules produce n conditional tests as in test (2.1). Our aim is to detect a maximal number of strategies having significant none-zero ϕ while controlling for FDR, the expected proportion of false discoveries among the hypothesis tests called significant as introduced in BH.

For the convenience of notation, let us consider the single test (2.1). To formulate the assumptions, we assume there are d covariates, represented by random variables $Z^1, \ldots, Z^d$, conveying information about the probability of a hypothesis being true null as well as the distribution of the $p-$value of the false null hypothesis. Let us denote by P the random variable representing the $p-$value of the test and $\boldsymbol{Z} = (Z^1, \ldots, Z^d)$. Also, let h be the true status of the hypothesis, that is, $h = 0$ if the null hypothesis is true and $h = 1$ if the null hypothesis is false. To indicate a particular test corresponding to a trading rule i, we add the subscript i to all mentioned notations, i.e. p_i, h_i and $\boldsymbol{Z}_i$.

To control the FDR at a level $\tau \in (0, 1)$, the conventional decision rule is a null hypothesis i is rejected if and only if $p_i \leq \hat{\Theta}$ where $\hat{\Theta}$ is a common threshold for all hypotheses, determined via a data-driven manner depending on particular procedures.[5] In contrast, in this study, the threshold depends on realized values of covariates $\boldsymbol{Z}$ and varies across hypotheses, i.e. the rejection condition becomes $p_i \leq \hat{\Theta}(\boldsymbol{z}_i)$ and will be expressed in an implicit formula as presented below.

For each j ($j = 1, \ldots, d$), we transform the observed value of the covariate Z_i^j to a form so that Z_i^j uniformly distributed on $[0, 1]$, via using $z_i^j = r_i^j/n$ where r_i^j

[5] For instance, the procedure of BH ranks the hypotheses based on $p-$values from smallest to highest. Denote by $\tilde{p}_j$s the $p-$values after ranking, the BH procedure seeks for $j^* = \max\{j|\tilde{p}_j \leq j \times \tau/N\}$ where $N = n$, then sets $\Theta = \tilde{p}_{j^*}$. Storey (2002) uses the same procedure with $N =$ (number of $p_i > \lambda)/(1 - \lambda)$ for some $\lambda \in (0, 1)$.

is the rank of the observed value of Z_i^j in the set of observed values of $Z_1^j, \dots, Z_n^j$, $j = 1, \dots, d$. In the followings, Z^js are the ones in the transformed forms. We assume that, conditional on $\boldsymbol{Z} = \boldsymbol{z}$, the hypothesis is a priori true null with probability $\pi_0(\boldsymbol{z})$, i.e. $(h|\boldsymbol{Z} = \boldsymbol{z}) \sim Bernoulli(1 - \pi_0(\boldsymbol{z}))$. Its estimation procedure will be discussed in below.

To formulate the theoretical framework, we require $\boldsymbol{Z}$ and P satisfying: i) Z^js are mutually independent, $j = 1, \dots, d$; ii) Conditional on $\boldsymbol{Z} = \boldsymbol{z}$, when the null hypothesis is true, P is uniformly distributed on the interval $[0, 1]$ and when the null hypothesis is false, P has a distribution determined by some density function $f_{alt}(p|\boldsymbol{z})$. To develop the theoretical framework, we assume that the aforementioned n tests are independent replications of the test (2.1), i.e., the triple h, $p-$value and covariates of the tests are independent, and each of them has the same distribution as the triple $(h, P, \boldsymbol{Z})$. In the next Sections, we show that our method can also be applied to scenarios when there is a dependence among test statistics and covariates.

The gist of our method is as follows. Given a target τ of FDR, we do not decide to reject the null hypotheses based on their $p-$values solely. We instead reject a null hypothesis by a rule based on both the $p-$value and the covariates. Thereby, we define each hypothesis with an observed $(p, \boldsymbol{z})$, a posterior probability of being null denoted by $r(p, \boldsymbol{z})$. If there are any significant hypotheses, the hypothesis with the smallest $r(p, \boldsymbol{z})$ will be selected first, then the second smallest one and so on. Each time a hypothesis is added to the significant set, the FDR is raised. We stop the procedure when the FDR target is reached. More specifically, the posterior probability of being true null is

$$r(p, \boldsymbol{z}) = \mathbb{P}(h = 0|(P, \boldsymbol{Z}) = (p, \boldsymbol{z})) \tag{2.2}$$

which can be developed further as

$$r(p, \boldsymbol{z}) = \frac{\mathbb{P}(h = 0|\boldsymbol{Z} = \boldsymbol{z})}{\mathbb{P}((P, \boldsymbol{Z}) = (p, \boldsymbol{z}))} = \frac{\pi_0(\boldsymbol{z})}{f(p, \boldsymbol{z})} \tag{2.3}$$

where $f(p, \boldsymbol{z})$ is the joint density function of the $p-$value and covariates.[6]

[6] The equation (2.3) is obtained by using the fact that $\mathbb{P}(h = 0, P = p, \boldsymbol{Z} = \boldsymbol{z}) = \mathbb{P}(P = p|(h = 0, \boldsymbol{Z} = \boldsymbol{z})).\mathbb{P}(h = 0|\boldsymbol{Z} = \boldsymbol{z}).\mathbb{P}(\boldsymbol{Z} = \boldsymbol{z}) = \mathbb{P}(h = 0|\boldsymbol{Z} = \boldsymbol{z})$ where the first and last factors equal one as they are density functions of uniform distributions.

Empirically, $r(p, \boldsymbol{z})$ is estimated by $\hat{r}(p, \boldsymbol{z}) = \hat{\pi}_0(\boldsymbol{z})/\hat{f}(p, \boldsymbol{z})$ where $\hat{\pi}_0(\boldsymbol{z})$ and $\hat{f}(p, \boldsymbol{z})$ are estimators of $\pi_0(\boldsymbol{z})$ and $f(p, \boldsymbol{z})$, respectively. For the sake of space, we refer to the details of the estimation procedures in Appendix B.1.

Consequently, the rejection region has a form $\Gamma(\theta) = \{(p, \boldsymbol{z}) | \hat{r}(p, \boldsymbol{z}) \leq \theta\}$ where $\theta \in (0, 1)$ is satisfying

$$\int_{\Gamma(\theta)} \hat{r}(p, \boldsymbol{z}) dp d\boldsymbol{z} \leq \tau. \tag{2.4}$$

The left side of the equation (2.4) is the estimate of the FDR corresponding with the rejection region $\Gamma(\theta)$. Hence, we choose a maximal threshold $\theta = \theta^*$ such as the condition (2.4) holds. Thus, a hypothesis is significant if and only if its observed $(p, \boldsymbol{z})$ belongs to the set $\Gamma(\theta^*)$. This rejection rule implicitly stands for the mentioned condition $p_i \leq \hat{\Theta}(\boldsymbol{z}_i)$.[7] To make this procedure more intuitive and efficient, the rejection rule is implemented via the "functional $q-$value" introduced in CRS. For each hypothesis i with observed $(p, \boldsymbol{z}) = (p_i, \boldsymbol{z}_i)$, we determine its $q-$value as the estimate of the FDR when we reject the null of all hypotheses j having $\hat{r}(p_j, \boldsymbol{z_j}) \leq \hat{r}(p_i, \boldsymbol{z}_i)$. Thus, at the given target τ of FDR, a null hypothesis is rejected if and only if its corresponding $q-$value $\leq \tau$.

We name the proposed procedure, where the $\boldsymbol{z}$ is a vector of more than one covariate, as $mfFDR$. When $d = 1$, the $mfFDR$ is the $fFDR$ of CRS. In the followings, we present an intuitive illustration of the value of informative covariates in the $mFDR$ framework, and we compare the performance of the proposed method to others in terms of FDR control and power.

2.2.2 Simulation studies

We consider the simplest case where we have two informative covariates $\boldsymbol{Z} = (U, V)$. We simulate $n = 10,000$ hypotheses where the proportion of the null hypotheses is approximately 0.8.[8] Suppose that the two covariates convey information about the hypotheses. We will demonstrate that by using $mfFDR$, which utilizes both covariates

[7] That is, the null hypothesis i is rejected if and only if $\hat{r}(p_i, \boldsymbol{z}_i) \leq \theta^*$ - an implicit form of $p_i \leq \hat{\Theta}(\boldsymbol{z}_i)$.

[8] This setting of π_0 is chosen similar to studies in statistics literature such as Storey (2003) and CRS. Our conclusions on the superior power of the $mfFDR$ over other methods under studying remain unchanged under different choices of π_0.

as inputs, we obtain a higher power in detecting false null hypotheses than the $fFDR$, which uses only one of the two covariates.

The data-generating process in our $mfFDR$ simulations is as follows. In each iteration, we independently draw the elements of the covariates $U = (u_1, \ldots, u_n)$ and $V = (v_1, \ldots, v_n)$ from the uniform distribution $U(0, 1)$. For each hypothesis i we draw its null status h_i from a Bernoulli distribution with a probability of being null $\mathbb{P}[h_i = 0|(u, v) = (u_i, v_i)] = \pi_0(u_i, v_i)$, i.e. $h_i \sim Bernoulli(1 - \pi_0(u_i, v_i))$, where $\pi_0(u, v)$ has one of the two following forms

- $\pi_0(u, v) = \sin[\pi(u + v)/2]$, i.e. a sine function;
- $\pi_0(u, v) = 1 - (u^4 + v^4)/2$, i.e. a monotonic function (concerning each covariate).

We obtain two bundles of (U, V, H) that correspond to the two forms of the $\pi_0(u, v)$, where $H = (h_1, \ldots, h_n)$.

Next, for each triple (U, V, H), we generate $p-$values for tests such that for each true null hypothesis, i.e. the one with $h_i = 0$, its $p-$value is drawn from the uniform distribution $U[0, 1]$, whereas the $p-$value of a false null hypothesis the distribution with density function $f_{alt}(p|(u, v))$. We specify $f_{alt}(p|(u, v))$ by using a Beta distribution $Beta(\alpha, \beta)$, i.e. $f_{alt}(p|(u, v)) \propto p^{\alpha-1}(1 - p)^{\beta-1}$ where the α and β are positive real parameters determining the shape of the distribution. Here, similar to CRS, we set the β as a function of the covariates, specifically as $\beta = 3 + 1.5(u + v)$. Aiming to study the method's performance in various circumstances, we consider three cases of $\alpha \in \{0.5, 1, 1.5\}$ which we name as strong, weak and very weak signal, respectively (thus, we have three specifications of a Beta distribution). In the strong signal case, the false null hypotheses are more easily distinguished from the true null ones than those in the weak and very weak signal cases.

To illustrate clearer how different the three cases are, Figure 2.1 depicts in Panel A (Panel B and C) the distribution of $p-$values of the false null hypotheses drawn from the strong (weak and very weak) signal case on the left, and of the whole population on the right of the panel. In these panels, the probability of being null each hypothesis is generated from the same sine form of $\pi_0(u, v)$ and the $p-$values of the true null hypotheses are drawn from the uniform distribution $[0, 1]$. In the strong signal case

($\alpha = 0.5$), the $p-$values of the false null ones are mostly concentrated near the zero point. In contrast, those $p-$values under the weak signal setting are less condensed at the zero point and dispersed remarkably up to 0.6. In the very weak case, the peak departs from the zero point. The false null hypotheses in the weak and the very weak cases are more difficult to be detected. For example, if we reject a null hypothesis whenever its $p-$value is less than 0.05, then we detect half of the false null hypotheses in the strong signal case while in the very weak signal case, the detected portion is much smaller.[9] In the context of the technical trading rule, a weak signal means that the truly out/under-performing rules have small absolute Sharpe ratios, and thus they have large $p-$values, which make them more difficult to be detected from a random walk.

The task is to detect the false null hypotheses from the simulated sample with control of the FDR at given targets by using only the $p-$value and the covariates. In our first experiment, we illustrate the role of informative covariates in detecting false null hypotheses. Then we benchmark our procedure, the $mfFDR$ with both U and V as covariates, against the $fFDR$ with only U as a covariate, the FDR of Storey (2002) which we notate as standard FDR ($StdFDR$), and the FDR of BH. By comparing the obtained results with the null status of the hypotheses (i.e., the H) we calculate a false discovery proportion, which is the ratio of the number of true null hypotheses falsely rejected over the number of the discoveries, and a correct detection proportion, which is the ratio of the number of false null hypotheses detected over the number of false null ones in the population.

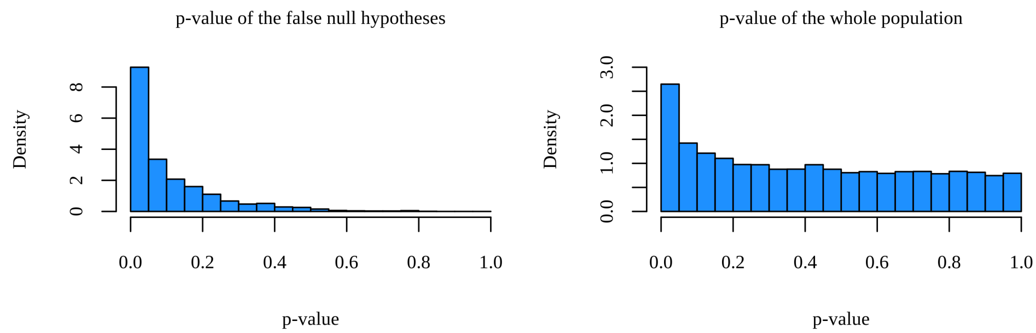
2.2.2.1 The role of informative covariates in detecting false nulls

We analyse a simulated sample to show the usefulness of informative covariates and how they work in the FDR framework. Particularly, we select the one depicted in Panel A of Figure 2.1. This sample is generated under the strong signal setting with the sine form of the $\pi_0(u, v)$. To see how the covariates convey the information of the hypotheses, we partition the sample into nine groups based on dividing each of the u

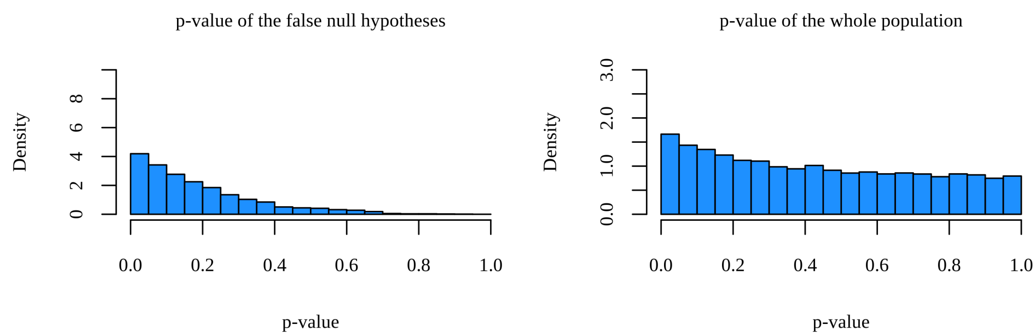
[9] We also note that the same number of true null hypotheses will be wrongly rejected at this threshold for all cases. Consequently, the false discovery proportion is much higher in the very weak signal case.

Figure 2.1: Distribution of $p-$values. The figure shows the distribution of $p-$values of false null hypotheses component and of the whole population in three scenarios: the $p-$values of the false null hypotheses are drawn from a strong signal case (Panel A), a weak signal case (Panel B) and a very weak signal case (Panel C).

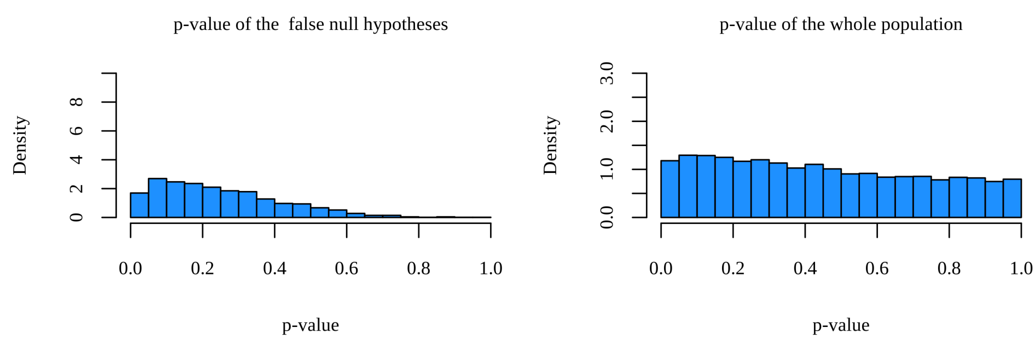
Panel A: Distribution of $p-$value under the strong signal case ($\alpha = 0.5$).



Panel B: Distribution of $p-$value under the weak signal case ($\alpha = 1$).



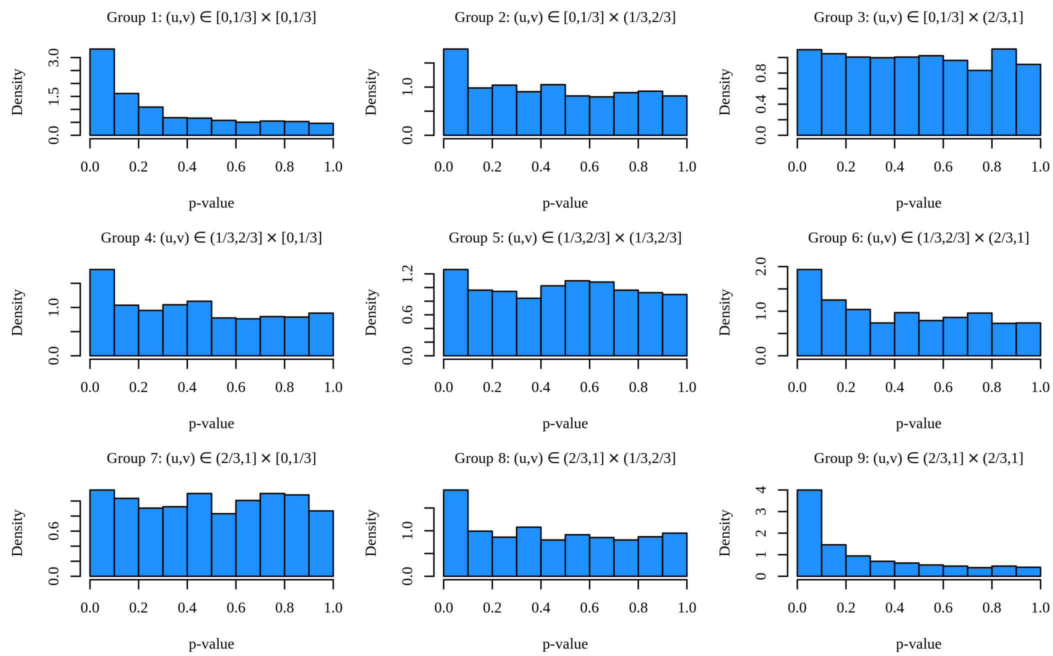
Panel C: Distribution of $p-$value under the very weak signal case ($\alpha = 1.5$).



and v into three equal segments. For convenience, the groups are named from 1 to 9, corresponding to the nine areas of (u, v). In doing so, each group contains roughly the same number of hypotheses. The $p-$value distributions of the groups are depicted in Figure 2.2.

To control the FDR at a target τ, which is $\tau = 0.2$ in this particular example, the BH and $StdFDR$ reject null hypotheses based on the $p-$values by seeking a threshold at which all null hypotheses with smaller $p-$value are rejected. Thus, this threshold is

Figure 2.2: Distributions of $p-$values partitioned into nine groups. The figure shows the distributions of $p-$values of nine groups which are partitioned from the sample in Panel A of the Figure 2.1 based on the value of the covariates u and v. Each sub-figure represents a group of hypotheses corresponding to the values of (u, v) shown in its title.

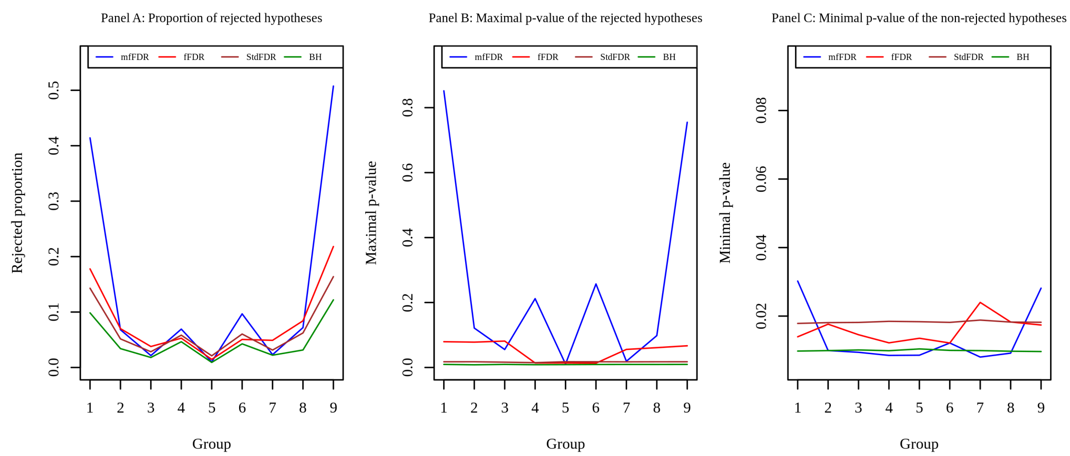


fixed for all groups. However, as shown in Figure 2.2, groups 1, 2, 4, 6, 8 and 9 contain much more false null hypotheses than others.[10] More specifically, as the $p-$values of the true null hypotheses are uniformly distributed, a $p-$value histogram such as the one of group 3 indicates that most of the hypotheses in this group are true null. In other words, the true null proportion in this group is very high. In contrast, the null proportion in group 1, for instance, is much lower. Hence, when the purpose is to maximize the number of discoveries (while controlling for the FDR at the given target), instead of rejecting all null hypotheses up to a single threshold as in a traditional approach, say 0.05, we could simply use a threshold of 0.2 in group 1 and 0.01 for group 3. This is reflected clearer in Figure 2.3.

In Panel A of Figure 2.3, each line represents the proportion of null hypotheses rejected by the four procedures across groups 1 to 9. Here, the rejected proportion is the ratio of the number of null hypotheses rejected in a group over the number of null hypotheses in the group. It is clear that all procedures reject more null hypotheses

[10]Since the $p-$values of the tests are independent, the ones of true null hypotheses are uniformly distributed and therefore producing a flat histogram. Consequently, a group with a more skewed histogram contains more true alternatives.

Figure 2.3: Comparison of the procedures across groups. Panel A of the figure presents the proportion of rejections of each procedure, whereas Panel B the corresponding maximal $p-$value of those null hypotheses rejected in each group. Panel C, in contrast, shows the minimal $p-$value of those hypotheses whose nulls are not rejected. The partition of hypotheses into groups is described in Figure 2.2.



in groups 1 and 9 than they do in groups 3, 5 and 7. Especially, we witness that the rejected proportions of the $mfFDR$ and $fFDR$ are much higher than those of the BH and $StdFDR$ in groups 1 and 9, while the figures of the former are slightly less than that of the $StdFDR$ in group 5. In Panel B, we show the maximal $p-$value of those null hypotheses rejected by each procedure. As discussed, a null hypothesis is rejected by the BH and $StdFDR$ if its $p-$value is less than some threshold (these thresholds are 0.01 for the BH and 0.018 for the $StdFDR$, which virtually coincide with the green and brown lines in Panel B, respectively). Hence, the maximal $p-$values of the null hypotheses rejected in the nine groups for BH and $StdFDR$ are roughly the same. In contrast, $mfFDR$ rejects some null hypotheses having $p-$value up to more than 0.8 in group 1, while in group 5 it does not reject any null hypotheses having $p-$value more than 0.011, which is less than the significant threshold of the $StdFDR$. Thus, a few null hypotheses are rejected by the $StdFDR$ but not by the $mfFDR$. Also, as Panels B and C noted, it is not uncommon in a group that a null hypothesis with a high $p-$value is rejected while another with a lower $p-$value is not. For instance, in group 1, there is a null hypothesis with a $p-$value of 0.03 that is not rejected by the $mfFDR$, while in the same group, there are null hypotheses with $p-$values of more than 0.8 that are rejected.

In summary, our experiments indicate that the $mfFDR$ rejects more null hypotheses in groups where the false null ones are rich. It is worth mentioning that, while partitioning the hypotheses into several groups as presented above illustrates the role of informative covariates in the $mfFDR$ framework, the $mfFDR$ method does not rely on grouping hypotheses into just a few groups.[11] Loosely, it can be understood that, the $mfFDR$ method treats each hypothesis as a group and establishes a particular rejection threshold for each group, i.e., a rejection threshold for each hypothesis. This is implementable using the $q-$value concept presented in Appendix B.1.

2.2.2.2 Performance of the $mfFDR$: FDR control and power comparison

In this section, we assess the two most important criteria of an FDR procedure: the control of the FDR and the power. To do so, we implement the $mfFDR$ and benchmark procedures at FDR targets $\tau \in \{0.05, 0.1, \ldots, 0.95\}$ over 1000 iterations and average the false discovery proportions and the correct detection proportions to have estimates of the actual FDR and the power, i.e. the expectation of the mentioned correct detection proportion, respectively. In total, we are studying six cases corresponding to the combinations of the two forms of the function $\pi_0(u, v)$ and the three aforementioned specifications of the Beta distribution. We also demonstrate that our method retains its power when the covariates are correlated with each other and are subject to estimation errors. Finally, we illustrate the FDR control and the power of the $mfFDR$ when the $p-$values are weakly dependent which will be the case when each hypothesis test represents a technical trading rule that stands for a specific combination of parameters. To assess the first criterion, we compare the estimated actual FDR of the $mfFDR$ to the given FDR targets, while for the second one, we compare the power of the $mfFDR$ against that of the $fFDR$, $StdFDR$, and BH approaches.

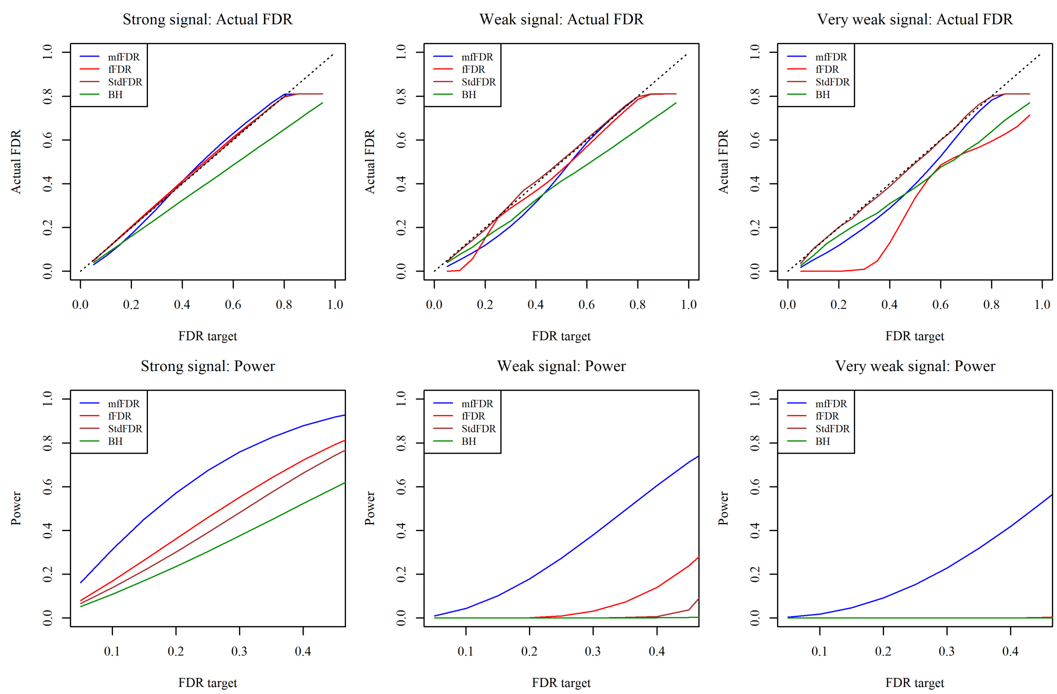
In Figure 2.4, we show the estimate of the actual FDR and the power of all procedures under the sine (Panel A) and monotonic (Panel B) form of $\pi_0(\boldsymbol{z})$. In each panel, the top three sub-figures exhibit the estimated actual FDR corresponding to the three cases of the signal (strong, weak and very weak). In each of those sub-figures, each line presents the estimated actual FDR of a procedure at the given FDR targets. Ideally, a

[11] Ignatiadis *et al.* (2016) and Ignatiadis and Huber (2021) recently introduce a group weighting approach where they partition hypotheses to groups based on a single covariate and determine the rejection threshold (of $p-$value) in each group. This approach, however, is less powerful than the $fFDR$ of CRS.

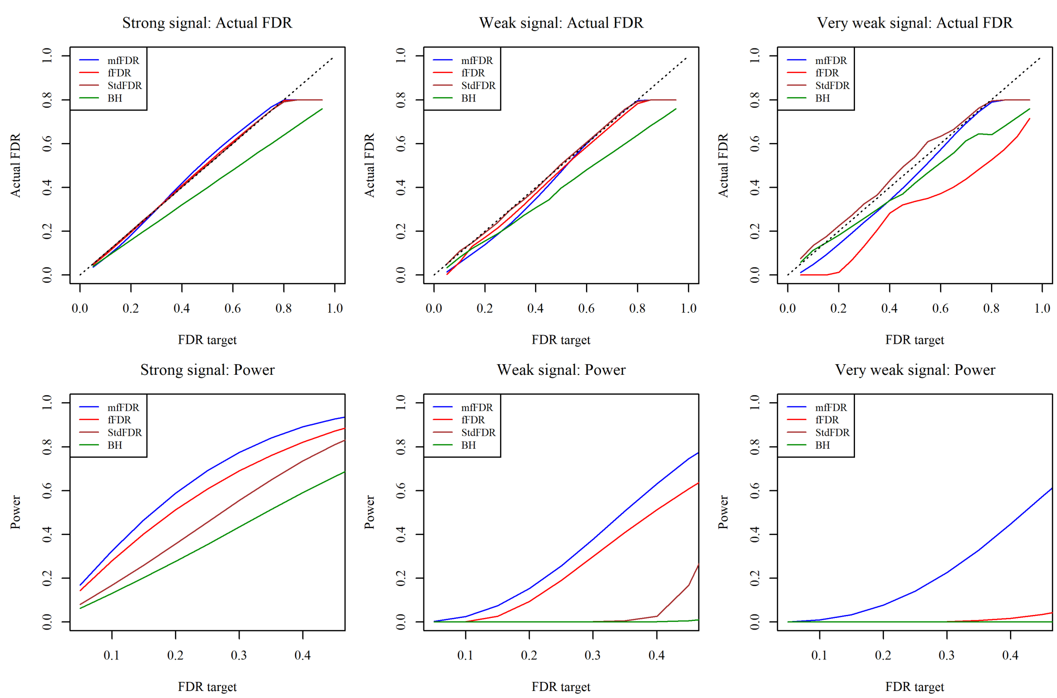
procedure perfectly (strictly) controls the FDR at a target if its estimated actual FDR at that target lies on (below) the 45° dotted line. For instance, in the top left sub-figure,

Figure 2.4: Performance comparison of FDR methods. The figure exhibits the performance comparison of the $mfFDR$, the $fFDR$, the Standard FDR of Storey ($SdtFDR$) and the FDR procedure in BH. Panel A shows the performance when the $\pi_0(u, v)$ has a sine form, whereas Panel B is the monotonic one. In each panel, the top three sub-figures exhibit the estimated actual FDR and three bottom sub-figures present the power.

Panel A: $\pi_0(u, v)$ is a sine function.



Panel B: $\pi_0(u, v)$ is a monotonic function.



all procedures either perfectly or strictly, control for FDR at target 20%. As controlling FDR is a sample property, it is acceptable to observe a point positioning slightly above the dotted line since we estimate the actual FDR over only 1000 iterations. In general, from the sub-figures, we see that all procedures control well for the FDR at any given targets and in all considered cases.

In terms of power, it makes sense to focus on only the FDR targets less than 0.5 (i.e. up to 0.45 in our simulation). Hence, the three bottom sub-figures in each panel present the power of the four methods for only those targets. In all cases, the lines representing the power of the $mfFDR$ are always at the top regardless of the form of $\pi_0(u, v)$ as well as the strength of the signal. In other words, the $mfFDR$ beats its benchmarks in terms of power. Apparently, all procedures have higher power when the signal is strong. In this case, the $mfFDR$ has gaps up to about 21%, 29% and 37% compared with the $fFDR$, *StdFDR* and BH procedures, respectively. Those figures are 44%, 67% and 76% (57%, 62% and 62%) for the weak (very weak) signal case. At the FDR target of 20%, which will be used later in our main analysis, the gap of the $mfFDR$ over $fFDR$ varies from 10% to 20%. Finally, the weak and the very weak signal cases highlight the benefit of using the $mfFDR$ when the data has a low signal-to-noise ratio. While the $fFDR$, *StdFDR*, and BH procedures can hardly detect a single false null hypothesis even at the FDR target of 20% (see the very weak signal case), the $mfFDR$ quickly gains a significant power of more than 10%.

To sum up, we have developed $mfFDR$, which enables multiple informative covariates to detect false null hypotheses and compare our method to the FDR procedures of Storey (2002) and BH, which do not use covariates. We have shown that the power of the former largely surpasses that of the latter ones while controlling well for the FDR at any given target. In other words, we can detect more outperforming technical rules under a testing framework that is conditional on multiple information (i.e., $mfFDR$) than the ones that are unconditional or use less information. On the other hand, we also show that when more than one informative covariate that are mutually independent, the $mfFDR$ gains remarkably higher power than the $fFDR$, which is the $mfFDR$ with $d = 1$, especially when the signal of the false null hypothesis is weak.

2.2.3 Correlation and estimation errors of covariates

Hitherto, we have studied the performance of the $mfFDR$ where the covariates are independently drawn from a uniform distribution. In financial data, this is unlikely the case. To consider this issue, in this section, we design a simple model where the covariates are positively correlated.[12] Given a correlation coefficient of r, the two covariates studied in the previous section are transformed (with the use of Cholesky factorization) into two new covariates having a correlation coefficient of approximately r. The simulated data are then generated similarly to the previous section. For the interest of space, we present the results for only the sine form of the null proportion function.[13] Aiming to study the impact of the correlation in covariates on the performance of the $mfFDR$, we consider varying values of r from 0.1 to 0.8.

Figure 2.5 depicts the performance of the $mfFDR$ in terms of FDR control (Panel A) and its power compared to others (Panels B and C). For the former aspect, the FDR is well controlled when $r < 0.7$ regardless of signal level. This coefficient range covers most cases in our real data in which, as shown in Section 2.5, more than 95% of the empirical coefficients have an absolute value less than 0.7. When $r \geq 0.7$, the FDR is controlled well under the weak and very weak signals and asymptotically controlled in the strong case.

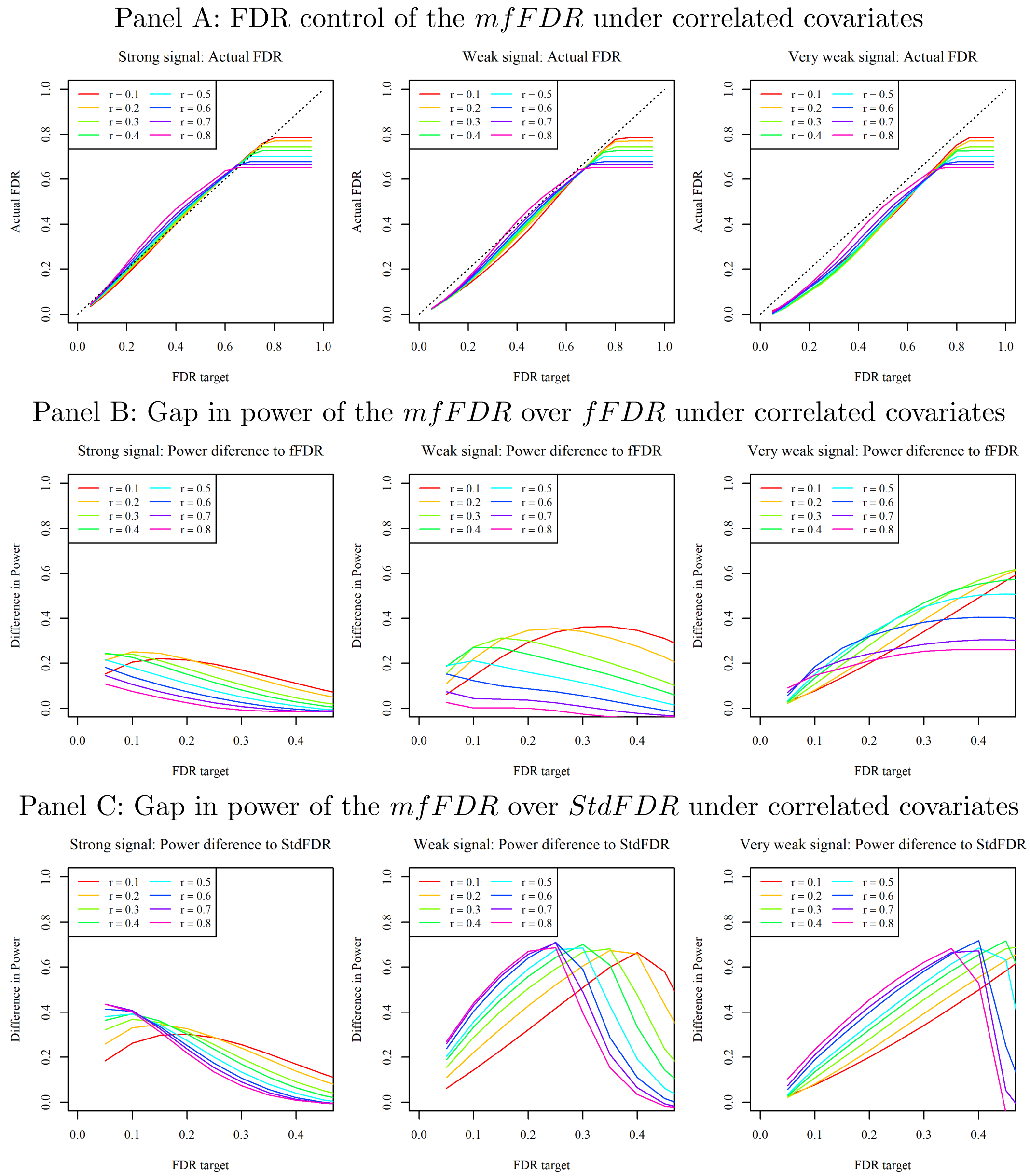
It is noted that the powers of the $mfFDR$ under different correlation coefficients are incomparable due to the differences in the level of signals. That is, the signal in higher correlation settings might be stronger due to the transformed covariates as the $p-$values of tests are generated based on them. Consequently, to assess the performance in terms of power, we calculate the gaps in the power of the $mfFDR$ over the $fFDR$ and $StdFDR$ under each case of r and depict them in Panels B and C, respectively. In the sub-figures of these panels, a line above zero indicates a higher power of the $mfFDR$ compared to its benchmarks. This is the case for all of the considered settings, with the peak varying across cases and could reach about 70% or 30%, under the FDR target of 20%, when the benchmark is the $StdFDR$ or $fFDR$, respectively.

Alongside the concern in the correlation of covariates, the estimation errors of co-

[12]The performances of the procedures under negatively correlated covariates are similar.

[13]These two new covariates are not uniformly distributed on $[0, 1]$, thus, the null proportion function is slightly modified, $\pi_0(u, v) = \min\{1, \max\{0, \sin(\pi(u + v)/2)\}\}$, so that its values are in $[0, 1]$.

Figure 2.5: Performance of the $mfFDR$ under correlated covariates. Panel A exhibits the performance of the procedures in terms of FDR control whereas Panels B and C present the power differences of the $mfFDR$ over the $fFDR$ and the FDR procedure of Storey (2002) ($StdFDR$), respectively.



variates (i.e., the noise in the estimating covariates) also potentially affect the performance of the $mfFDR$. In empirical finance, the covariates will be estimated quantities and are thus subject to estimation errors. To address this concern, we additionally conduct simulations where the covariates used as input contain noise. For the interest of space, we defer the results of this experiment to Appendix B.2. Generally, we find that $mfFDR$ still controls FDR well. Perhaps not surprisingly, its power is lower than the prior case with uncorrelated covariates, as presented in the previous section.

Nevertheless, the power of the $mfFDR$ is still remarkably higher than that of other methods. The $mfFDR$ method is robust to correlation and estimation biases.

2.2.4 Weak dependence in p-values

In developing a theoretical $mfFDR$ framework, we assume that the tests are independent replications of the test (2.1). In financial applications, this scenario is unlikely to be the case. In this particular study, we will consider hypothesis tests comparing the Sharpe ratios of technical trading rules against zero. In a particular category of trading rules, the rules with close parameters tend to have highly correlated returns. This leads to a weak dependency among the testing statistics or $p-$values of the corresponding hypotheses. This section shows that our method is robust under this type of dependence. Specifically, we are generating data such that the hypotheses are partitioned into groups with the same size k. In each group, the $p-$values of testing hypotheses are mutually dependent at the same level, which is characterized by a covariance matrix $\mathbf{\Sigma}$. The $p-$values corresponding to the hypotheses from different groups are independent.

The data-generating process is as follows. The simulated covariates $\pi_0(u, v)$ and true status of null hypotheses H are as described in Section 2.2.2. For the sake of space, we present only the results for the sine form of $\pi_0(u, v)$ and with a strong signal setting of $p-$values. To account for the dependence in hypotheses, we first partition the true null hypotheses into groups of size k. Secondly, we generate the $z-$scores for the null hypotheses of each group from the multivariate normal distribution $\mathcal{N}(0, \mathbf{\Sigma})$. We process similarly for the false null ones, but the $z-$scores of each group are drawn from $\mathcal{N}(2, \mathbf{\Sigma})$. To simplify, the matrix $\mathbf{\Sigma} = (\Sigma_{ij})_{k\times k}$ is set as $\Sigma_{ii} = 1$ and $\Sigma_{ij} = c$ for $i \neq j = 1, \ldots, k$ for some c. By considering various values of the parameters k and c, we reveal the impact of the dependence at different levels on the performance of the $mfFDR$. Here, we choose $c \in \{0, 0.25, 0.5, 0.75\}$ where the case $c = 0$ indicates the absence of the dependence among $p-$values and will be used as a benchmark for a comparison purpose, and $k \in \{10, 100, 500\}$. The parameter k represents the dependence scale. A larger k indicates the presence of more hypotheses that are mutually dependent.[14] Finally, the $p-$value of each (two-sided) test is calculated from its $z-$score by using

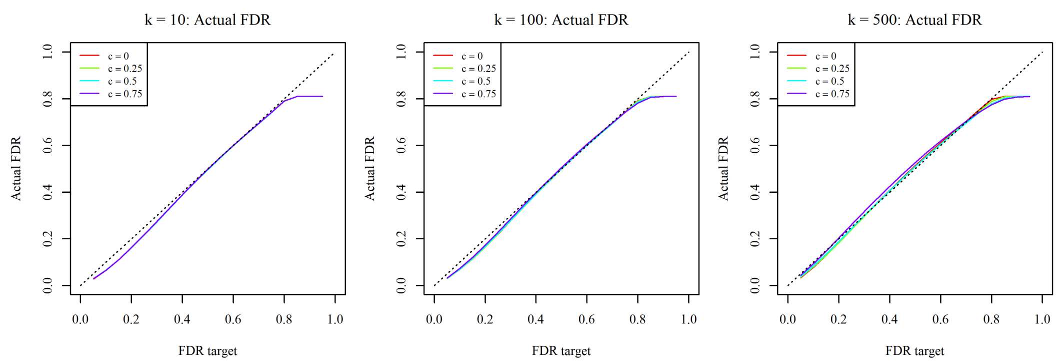
[14] This type of weak, dependent setting is also studied in Storey (2003) with $k = 10$. Here we extend the cases of k to study its impact on the method's performance.

the cumulative distribution function of the standard normal distribution. Note that, in this setting, the covariates convey only the information on the probability of being true null of the hypotheses (i.e. via $\pi_0(u, v)$) and not on the $p-$values (i.e. we do not generate $p-$values via $f_{alt}(p|(u, v))$, but via the dependent $z-$scores, which does not depend on the covariates) as in the simulations conducted in the previous sections.

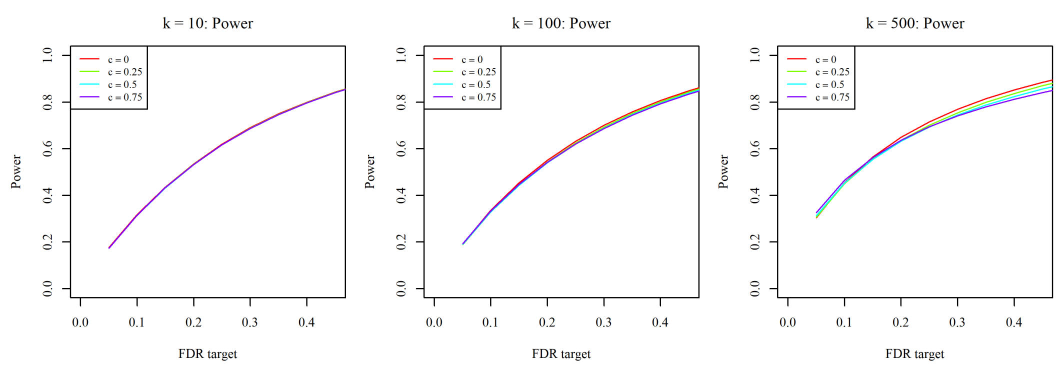
Figure 2.6 presents the performance of the $mfFDR$ under the settings of c and k. Panel A of the figure shows that the dependence does not affect the FDR control of the method. In terms of power in Panel B, when k is small, the first two sub-figures show an insignificant difference between the power of the $mfFDR$ with different levels of dependence among $z-$scores in each group. Therefore, the lines corresponding to different values of c are virtually identical and covered by a purple line. Their powers are only distinguishable in the third sub-figure, where k is large and for high targets of FDR. In this case, the power is decreasing concerning c, i.e. the level of dependence level among the $z-$scores.

Figure 2.6: Performance of the $mfFDR$ under correlated $p-$values. Panel A (B) presents the FDR control (power) of the $mfFDR$ under different level of dependencies among $p-$value of the tests.

Panel A: FDR control of the $mfFDR$ under correlated $p-$values



Panel B: Power of the $mfFDR$ under correlated $p-$values



In this section, our simulations cover most concerns about applying our $mfFDR$

method to economic and financial data. We find that the $mfFDR$ performs well in terms of FDR control and power with correlated covariates, under correlated $p-$values, and with estimation errors in covariates.

2.3 Data and strategy universe

2.3.1 Data

We collect daily spot and 1-month forward exchange rates data against the U.S. dollar for 30 countries, including nine developed markets (Australian dollar, Canadian dollar, German mark/euro, Japanese yen, New Zealand dollar, Norwegian krone, Swedish krona, Swiss franc, and U.K. pound) and 21 emerging markets (Argentine peso, Brazilian real, Chilean peso, Colombian peso, Czech koruna, Hungarian forint, Indian rupee, Indonesian rupiah, Israeli shekel, Korean won, Mexican peso, Philippine peso, Polish zloty, Romanian new leu, Russian ruble, Singaporean dollar, Slovak koruna, South African rand, Taiwanese dollar, Thai baht, and Turkish lira). The sample periods for developed market currencies start on January 4, 1971, and end on December 31, 2020. The sample periods for emerging market countries start from various dates due to data available on exchange and short-term interest rates. Israel has the earliest starting date among emerging market currencies (January 1978) and is followed by South Africa (January 1981), Singapore (January 1982) and Taiwan (October 1983); all emerging market data end in December 2020. Our data on exchange rates and short-term interest rates were kindly supplied by the London branch of the asset manager BlackRock and are based on midday quotations in the London market.

Table 2.1 presents summary statistics of gross currency returns and short-term interest rates of developed and emerging economies. We define the gross currency return that is the return from buying a foreign currency unit and holding it for one day, $r_t = ln(s_t/s_{t-1})$, where s_t represents the spot exchange rate on the day t. We define the spot exchange rate as units of U.S. dollars per foreign currency, so an increase of s is associated with an appreciation of the foreign currency. We report the mean, the standard deviation, the minimum and the maximum and the corresponding sample period. We find that gross returns tend to be more volatile for emerging economies offering higher minimum and maximum values.

Table 2.1: Summary Statistics. The table presents summary statistics of gross returns on foreign currencies and short-term interest rates. We report the mean, the volatility, the minimum and the maximum. We also report the sample period for each currency in our sample. We show results for developed and emerging economies.

Countries	Gross returns on foreign currencies				Short-term interest rates				Sample Period
	mean(%)	vol	min	max	mean(%)	vol(%)	min(%)	max(%)	
Developed									
Australia	-0.0028	0.0068	-0.1925	0.1073	0.0191	0.0119	0.0001	0.1709	1/4/1971-12/31/2020
Canada	-0.0018	0.0041	-0.0434	0.0505	0.0172	0.0125	0.0005	0.0596	1/4/1971-12/31/2020
Germany/E.U.	-0.0017	0.0058	-0.0421	0.0462	0.0092	0.0078	-0.0020	0.0344	1/4/1971-12/31/2020
Japan	0.0095	0.0063	-0.0626	0.0950	0.0108	0.0124	-0.0006	0.0406	1/4/1971-12/31/2020
New Zealand	-0.0033	0.0073	-0.2050	0.0995	0.0194	0.0111	-0.0007	0.1286	1/4/1971-12/31/2020
Norway	-0.0014	0.0066	-0.0814	0.0646	0.0163	0.0106	0.0023	0.1372	1/4/1971-12/31/2020
Sweden	-0.0036	0.0065	-0.1507	0.0555	0.0171	0.0139	-0.0011	0.2090	1/4/1971-12/31/2020
Switzerland	0.0122	0.0073	-0.0892	0.1267	0.0045	0.0098	-0.0086	0.2276	1/4/1971-12/31/2020
U.K.	-0.0044	0.0058	-0.0848	0.0467	0.0160	0.0112	0.0004	0.0460	1/4/1971-12/31/2020
U.S.	-	-	-	-	0.0128	0.0097	-0.0000	0.0466	1/4/1971-12/31/2020
Emerging									
Argentina	-0.0579	0.0098	-0.3418	0.1712	0.0343	0.0368	0.0035	0.4144	4/1/1991-12/31/2020
Brazil	-0.0238	0.0097	-0.1080	0.1178	0.0428	0.0283	0.0051	0.2792	1/6/1992-12/31/2020
Chile	-0.0070	0.0064	-0.1160	0.1114	0.0014	0.0016	0.0000	0.0214	1/1/1991-12/31/2020
Colombia	-0.0224	0.0062	-0.0593	0.0562	0.0315	0.0240	0.0052	0.0908	1/3/1986-12/31/2020
Czech	0.0040	0.0073	-0.0707	0.0522	0.0104	0.0133	0.0002	0.2628	1/4/1978-12/31/2020
Hungary	-0.0177	0.0080	-0.0842	0.0520	0.0297	0.0227	0.0021	0.0834	1/2/1987-12/31/2020
India	-0.0178	0.0046	-0.1281	0.0376	0.0254	0.0149	0.0002	0.1944	9/1/1994-12/31/2020
Indonesia	-0.0276	0.0130	-0.3576	0.2361	0.0296	0.0242	0.0000	0.2054	1/2/1997-12/31/2020
Israel	-0.0683	0.0059	-0.1725	0.0645	0.0555	0.0962	0.0002	0.6309	4/27/1993-12/31/2020
South Korea	-0.0047	0.0078	-0.1809	0.2012	0.0154	0.0138	0.0012	0.0676	7/4/1994-12/31/2020
Mexico	-0.0349	0.0114	-0.2231	0.2231	0.0403	0.0416	0.0077	0.3387	1/3/1994-12/31/2020
Philippines	-0.0096	0.0044	-0.0860	0.1015	0.0198	0.0165	0.0014	0.1962	4/22/1992-12/31/2020
Poland	-0.0117	0.0079	-0.0715	0.0670	0.0212	0.0198	-0.0004	0.0792	6/4/1991-12/31/2020
Romania	-0.0368	0.0091	-0.3887	0.0953	0.0407	0.0518	0.0006	0.3667	1/3/1992-12/31/2020
Russia	-0.0515	0.0132	-0.3863	0.2779	0.0288	0.0391	0.0022	0.3208	1/1/1987-12/31/2020
South Africa	-0.0286	0.0097	-0.1030	0.1440	0.0289	0.0115	0.0000	0.0588	6/4/1993-12/31/2020
Singapore	0.0043	0.0033	-0.0276	0.0414	0.0053	0.0050	0.0000	0.0181	1/2/1981-12/31/2020
Slovakia	0.0021	0.0063	-0.1097	0.0462	0.0130	0.0164	0.0001	0.2068	1/4/1982-12/31/2020
Taiwan	0.0035	0.0029	-0.0420	0.0430	0.0082	0.0067	0.0004	0.0483	10/4/1983-12/31/2020
Thailand	-0.0022	0.0055	-0.2077	0.0635	0.0102	0.0107	0.0004	0.0664	1/2/1991-12/31/2020
Turkey	-0.1002	0.0109	-0.3348	0.1256	0.0833	0.0643	0.0130	1.0328	2/1/1990-12/31/2020

Another important aspect of our analysis is the short-term interest rate as it affects the overall return of the trading strategies even though technical analysis focuses more on exchange rate fluctuations. We convert our annualized short-term interest rate (i^a), which is the overnight rate, into daily data $i_t = ln(1 + i^a)/360$. In Table 2.1, we report the mean, standard deviation, minimum and maximum values of short-term interest rates for both developed and emerging countries. We find that short-term interest rates for developed countries range between 0.45 to 1.94 basis points on average. We also find that short-term interest rates are higher for most of the currencies of the emerging markets category and they exhibit a particular variation. The highest interest rate in this group is for Turkey which is 8.33 basis points and the lowest is for Chile which offers 0.14 basis points.

2.3.2 Trading rule universe

Based on past daily spot exchange rates, a trading rule determines positions (long, short or neutral) that traders should take in the next day. In this study, we assess the universe of trading rules used in Hsu *et al.* (2016), which consists of the following five category trading rules widely used by traders.

The filter trading rules: The rules generate a long (short) position whenever the closing exchange spot rate has risen (fallen) by a given percentage above (below) its most recent high (low). This family rule is generally based on the momentum of the exchange rate where traders believe the rising (falling) rate continues rise (or fall). The threshold percentages are set so that the traders are not misled by the small fluctuations.

Moving average trading rules: These rules generate long or short positions based on comparing the closing spot exchange rate to one or three simple moving averages of given different lengths or comparing the moving averages of two different lengths. For example, the simplest moving average rule generates a long (short) position when the spot exchange rate moves up (down) at least a certain per cent above (below) the moving average of a specific length.

Relative strength indicators: A relative strength indicator is a popular form of an oscillator, which aims to identify imminent market corrections after rapid exchange rate movements. Generally, the indicator of a given length has values from 0 to 100 and generates an oversold or overbought signal when it crosses a predetermined lower or upper extremity, respectively.

Support-resistance rules: These rules rely on determining a support or resistance level for which the exchange rate appears to have difficulty in penetrating in a previously given number of days and a premise that when the closing exchange rate breaches the level, it will trigger further movement in the same direction.

Channel breakouts: The rules establish time-varying support and resistance levels, forming a trading channel with upper and lower bounds. Once a bound is breached, a long or short position is initiated in a similar way as the support-resistance rules.

Given the described family rules, a number of their variants are generated by varying plausible parameters, the delay time and the position's holding time. Ultimately, we obtain 2,835 filter rules, 12,870 moving average rules, 600 relative strength indicators, 1,890 support-resistance rules and 3,000 channel breakout rules, making up 21,195 trading rules in total. Readers are referred to Appendix A in Hsu *et al.* (2016) for the detailed specifications of our technical rules.

2.4 Measures of predictive ability and profitability

In this study, we distinguish the excess return and net excess return gained by a trading strategy before and after counting for transaction cost, respectively. The excess return from buying one unit of foreign currency (against the U.S. dollar) and holding it for one day is calculated as the summation of returns due to appreciation/depreciation of the foreign currency and the return obtained from lending the money in foreign currency, minus the benchmark return which is the return would gain if the money is deposited in the U.S., that is

$$r_t = \ln(s_t/s_{t-1}) + \ln(1 + i^*_{t-1}) - \ln(1 + i_{t-1}) \tag{2.5}$$

where s_t and s_{t-1} are spot rates at the midday of the days t and $t-1$, respectively, and i_{t-1} and i^*_{t-1} are daily interest rates on U.S. dollar deposits and foreign currency deposits on the day $t-1$, respectively.

The daily excess return of the trading rule j, earned from day $t-1$ to day t, $R_{j,t}$, is determined as

$$R_{j,t} = S_{j,t-1}.r_t \tag{2.6}$$

where $S_{j,t-1}$ is a position guided by the trading rule based on past data up to day $t-1$ and taking value in the interval $[-1, 1]$. For most of the trading rules in our universe, $S_{j,t-1}$ takes value $+1$ for the long, (-1) for the short and 0 for the neutral position on the foreign currency. For some moving average trading rules, where the signal guides for positioning one-third of funds on a long position while the rest is kept in U.S. dollar or on a short position while the rest is kept in foreign currency, $S_{j,t-1}$ takes value $+1/3$ or $-1/3$, respectively.[15]

In this study, we assess the performance of a trading rule j based on its Sharpe ratio. From day T_1 to day T_2, it is defined as

$$SR_j = \frac{\overline{R}_j}{\sigma_j}, \tag{2.7}$$

where $\overline{R}_j$ and σ_j are the mean and standard deviation of excess returns of the trading rule over the mentioned period, that is $\overline{R}_j = \sum_{t=T_1}^{T_2} R_t/N$, $\sigma_j = \sqrt{\sum_{t=T_1}^{T_2}(R_{j,t} - \overline{R}_j)^2/(N-1)}$ where $N = T_2 - T_1 + 1$, respectively.

When a trading rule changes its guiding signal, a new or existing position is triggered or closed and transaction costs occur consequently. The net excess return on a daily basis, therefore, is the daily excess return less the cost caused by the transactions and the net excess return and Sharpe ratio after transaction costs of a trading rule are calculated accordingly. As in Neely and Weller (2013), we use one-third of the quoted one-month forward rate bid-ask spread in each currency as an estimate of the one-way transaction costs on any particular day. For periods before the forward data is available, we set fixed transaction costs for each period in the same way as in Neely and Weller (2013): for developed country currencies we set the transaction cost at a flat 5 basis points in the 1970s, 4 basis points in the 1980s and 3 basis points in the 1990s, and for emerging market currencies we set the daily cost at one-third of the average of the first 500 bid-ask observations.

[15]Generally, with 1 dollar fund, the excess return for a position guided by a signal $S_{t-1} \in [-1, 1]$ is $S_{t-1}.r_t$. When $S_{t-1} > 0$, we invest S_{t-1} U.S. dollar on the foreign currency and the remaining $1 - S_{t-1}$ on U.S. bond. When $S_{t-1} < 0$, we borrow an amount equivalent to $|S_{t-1}|$ U.S. dollar in the foreign currency and convert it to U.S. dollar, then we invest the total of $|S_{t-1}| + 1$ dollar on U.S. bond.

Table 2.2 captures important features of trading rules in terms of performance corresponding to each exchange rate, after transaction costs. It is noted that the number of significant rules in this particular table is merely based on a conventional $p-$value of 5%. If all rules are independent random walks, one should expect there are roughly 530 significantly out-performing rules for each exchange rate.[16] With both performance

Table 2.2: Summary performance of trading rules in the whole sample period. The table shows the number of significant rules, which are those having bootstrap p-value < 0.05 (under conventional individual tests) and positive estimated performance, the highest performing rule in each currency with its performance. The column "Best rule" presents the class with the technical rules that provide the highest performance metric among all trading rules in a currency in the sample period. See Appendix A of Hsu *et al.* (2016) for the details of the various trading rules and corresponding class codes (F3, MA4, etc). We report two performance metrics - the Sharpe ratio and the mean of net excess return (both are annualized) which is the excess return after transaction cost."*", "**" and "***" respectively indicate statistical significance at levels of 10%, 5% and 1%.

	Mean net return			Sharpe ratio		
	Number of significant rules	Highest return (%)	Best rule	Number of significant rules	Highest ratio	Best rule
Australia	900	5.73***	F3	1155	0.54***	F3
Canada	196	2.67***	MA4	217	0.41***	MA4
Germany/E.U.	8585	5.33***	F3	8536	0.66***	MA5
Japan	7862	6.13***	MA4	7954	0.61***	MA4
New Zealand	6204	6.14***	MA1	6747	0.56***	MA1
Norway	443	4.24***	SR1	449	0.41***	SR1
Sweden	6827	5.99***	MA4	7038	0.59***	MA5
Switzerland	601	5.05***	F3	599	0.48***	CB1
U.K.	2161	4.58***	MA2	2217	0.54***	CB1
Argentina	9254	11.22***	SR1	10973	0.77***	MA5
Columbia	3915	11.04***	MA1	4025	1.17***	MA5
India	1844	4.44***	F3	2287	0.67***	F3
Indonesia	1744	10.71***	MA4	2401	0.56***	MA5
Israel	15271	9.26***	MA1	15452	1.22***	MA5
Philippines	6401	4.74***	SR1	6444	0.69***	MA5
Romania	67	5.53***	F3	118	0.52***	MA5
Russia	3897	15.63**	F1	6246	0.83***	MA5
Slovak	1331	5.69***	MA4	1348	0.58***	MA4
Brazil	4938	10.97***	SR1	6623	0.77***	MA5
Chile	1545	7.16***	SR1	1431	0.85***	MA5
Czech	107	5.87***	MA4	114	0.52***	MA5
Hungary	34	5.39***	SR1	80	0.45***	RSI
Korea	200	8.71***	F3	456	0.7***	F3
Mexico	8	4.66**	SR1	190	0.38***	CB1
Poland	39	6.3***	F3	74	0.51***	F3
Singapore	739	2.45***	CB1	673	0.56***	MA5
South Africa	914	7.82***	CB1	1013	0.53***	CB1
Taiwan	9782	4.3***	MA3	9827	1***	MA5
Thailand	3850	6.85***	MA1	3842	0.87***	MA5
Turkey	13516	16.07***	F3	14568	1.1***	MA5

[16] That is, as the $p-$values of true null hypotheses have a uniform distribution, there are approximate $21195 \times 0.05/2 \approx 530$ significantly out-performing rules.

measures under study, there are 22 over 30 exchange rates having more than 530 significant rules. This fact suggests that there could indeed exist a portion of technical trading rules that are truly profitable. All the best rules of the 30 exchange rates are significant with five of them gaining more than 10% per annum and four with a Sharpe ratio of at least one. However, we do not know if these rules are truly profitable or just lucky before we implement formal tests to correct data snooping biases.

2.5 Empirical results

In this section, we discuss our covariates in technical trading rules and the information they potentially convey. Based on the $mfFDR$ test incorporating these covariates, we select outperforming rules in each currency separately and in all currencies combined. We then construct rolling portfolios by the selected rules and show their in-sample and OOS performance. If the performance of technical trading rules in the foreign currency market is predictable to some degree, investing in technical rules that are truly profitable in an in-sample period is expected to generate an OOS profit. In particular, we demonstrate the OOS performance of the $mfFDR$ in detecting profitable rules using the following procedure: we implement the $mfFDR$ on the group of *positive* estimated SR trading rules to control for FDR at the given target. We name this procedure the multivariate functional false discovery rate "plus" ($mfFDR^+$). Similarly, we consider $fFDR^+$ of Chapter 1 as a benchmark that accommodates one covariate and focuses on positive performance. For the interest of space, we present the results with control of FDR at a target of 20% in the main text. In Appendix B.4, we show that our conclusions are robust for other FDR targets.

2.5.1 Covariates

We apply within our $mfFDR$ framework a set of informative continuous variables that are derived from technical rules' return series. These covariates reflect the performance persistence, financial risk, activeness and out-performance compared to a passive strategy of the trading rules under study. The first covariate we consider is the autocorrelation of a trading rule's return (ρ) that reflects its performance persistence. For the remaining covariates, we estimate the following equation for each trading rule in a particular currency using a rolling window

$$r_{i,t} = \alpha_{i,bh} + \beta_{i,bh} r_{bh,t} + \varepsilon_{bh,t} \tag{2.8}$$

where the $r_{i,t}$ is the excess return of the trading rule i, $r_{bh,t}$ is the excess return of the buy-and-hold strategy in a currency and $\epsilon_{bh,t}$ is the noise at day t. The $\alpha_{i,bh}$, $\beta_{i,bh}$ and the R–squared of the regression (2.8), respectively, represent the alpha, the riskiness level, and the variation of the trading rule i compared to the buy-and-hold strategy. We denote the three covariates as α_{bh}, β_{bh} and R^2_{bh} in the following context.

When the target is to select outperforming trading rules from the universe of all trading rules of all currencies (or else, when we examine all currencies together), which consists of $21195 \times 30 = 635850$ rules, we compare a particular trading rule to the average currency excess return factor of Lustig *et al.* (2011), denoted by RX. That is the return of a strategy that invests equally weighted on all 30 currencies in our sample (henceforth, currency market factor). In these cases, instead of using regression (2.8), the latter three covariates are obtained from the regression model

$$r_{i,t} = \alpha_{i,mk} + \beta_{i,mk} RX_t + \varepsilon_{mk,t} \tag{2.9}$$

where the RX_t is the currency market factor on the day t. Analogously, we denote the three new covariates as α_{mk}, β_{mk} and R^2_{mk}, respectively.

While the mentioned covariates have been studied in asset pricing literature such as Amihud and Goyenko (2013) use the R–square a predictor in mutual fund performance and Frazzini and Pedersen (2014) use the beta as a factor in various financial assets, this is the first time in FX literature the informativeness of those covariates in decision making is studied, especially in context of technical trading rule. In the following analyses, the informativeness of the covariates will be examined and the benefit of using multiple covariates under the $mfFDR$ framework will be discussed. We first focus on constructing a portfolio on each currency and use ρ, α_{bh}, β_{bh} and R^2_{bh}. Then, we examine all currencies together and consider ρ, α_{mk}, β_{mk}, and R^2_{mk}. Finally, we investigate how profitable trading rules vary across currencies.

2.5.2 Individual Currencies

As a first exercise, we examine each currency individually. Therefore, the covariates of $mfFDR^+$ are $R^2_{bh}, \alpha_{bh}, \beta_{bh}$ and ρ as described in Section 2.5.1. We use daily data and construct portfolios in a monthly frequency as follows: for each currency at the end of each month, we utilize the daily data in the most recent 12 months data as the in-sample period to calculate the $p-$value and covariates of the trading rules. We then implement the $mfFDR^+$ (as explained at the beginning of Section 2.5) to detect outperforming strategies in the in-sample period at the FDR target of 20%. We combine the signals of these outperforming rules to determine the position of that day by neutralizing the opposite ones. For instance, suppose there are 100 trading rules selected as profitable right before a day. Among them, 20 of them indicate a buy signal with a weight $+1$, 30 of them indicate sell signals with a weight $-1/3$, and the remaining 50 provide neutral signals. After combining, we have $10(= 20 \times (+1) + 30 \times (-1/3))$ long signals out of 100 profitable rules. Thus, the trader takes a long $1/10(= 10/100)$ position in the foreign currency.[17] We follow the signals of these portfolios to determine the position of each trading day in the following month (i.e., the OOS period). We then compare the performance metrics of these portfolios to those based on rules selected by FDR^+ and $fFDR^+$ with one covariate. The truly profitable rules are detected based on the use of the excess return (before transaction costs) and we mainly assess the OOS performance of the portfolio's net excess return (after transaction costs).[18]

Table 2.3 provides the quantiles of pairwise correlation coefficients of input covariates and those after combining all coefficients of the pairs as a pool. Those numbers capture an overview of the dependencies among the input covariates. The final row indicates that a majority of 98% coefficients are in $[-0.69, 0.59]$ for which the $mfFDR$

[17]This approach is based on the idea of the $1/N$ portfolio strategy, where we invest equally funds into each of the selected rules. We choose this method for three reasons; First, DeMiguel *et al.* (2007) show that such an approach is hard to be beaten by more sophisticated approaches (e.g. approaches weight funds differently on selected rules). Second, it reflects directly the performance of its components (e.g., the selected rules). Lastly, the funds allocated to opposite signals should be neutralized to avoid transaction costs. In a different context, Burnside *et al.* (2011) follow a similar approach to construct carry trade and momentum portfolios and Filippou *et al.* (2018) build an equally-weighted portfolio of a trading rule that is based on political risk.

[18]We assess the rules before transaction cost in in-sample period since the selected rules will be combined to have a single trading signal in a particular day in OOS period. As such, the frequency of the combined rule will be reduced. Thus, a rule that generates high transaction cost might not does so after combining.

controls very well for any FDR targets as shown in simulations in Section 2.2.3. We emphasize that there are only 1% of the coefficients having an absolute value from or above 0.7.

Table 2.3: The pairwise correlation of covariates. The table presents the quantiles of the correlation coefficients (calculated monthly with the use of one-year in-sample data) of six covariate pairs which are combinations of the four covariates: $R^2_{bh}, \alpha_{bh}, \beta_{bh}$ and ρ. The final row shows the numbers for the set of all correlated pairs.

Covariate pairs	min	1%	5%	25%	50%	75%	95%	99%	max
R^2_{bh}, α_{bh}	-0.73	-0.44	-0.32	-0.12	0.05	0.21	0.40	0.54	0.70
R^2_{bh}, β_{bh}	-0.82	-0.75	-0.68	-0.51	-0.25	0.18	0.59	0.70	0.81
R^2_{bh}, ρ	-0.92	-0.68	-0.40	-0.09	0.02	0.12	0.28	0.47	0.91
α_{bh}, β_{bh}	-0.81	-0.54	-0.39	-0.20	-0.06	0.09	0.30	0.45	0.72
α_{bh}, ρ	-0.86	-0.45	-0.22	-0.10	-0.04	0.02	0.12	0.21	0.72
β_{bh}, ρ	-0.52	-0.32	-0.19	-0.06	0.01	0.08	0.22	0.34	0.58
All pairs	-0.92	-0.69	-0.51	-0.15	-0.02	0.10	0.36	0.59	0.91

In Table 2.4, we present the average number of trading rules selected by FDR^+ and $mfFDR^+$, controlling for FDR at 20%, at the end of each month in different sample periods. As mentioned earlier, we measure technical trading rules' in-sample performance using the Sharpe ratio based on the portfolios' excess returns. We find that controlling at the same level of FDR, the $mfFDR^+$ detects more outperforming rules than FDR^+ does. This finding strongly holds for all currencies both in the whole period and in the sub-periods.

In Table 2.5, we present the OOS performance of the portfolios including the FDR^+, $fFDR^+$ and $mfFDR^+$ in terms of Sharpe ratio based on the portfolios' excess return before transaction costs. We also study the performance of $fFDR^+$ having as covariate the first principal component (PC1) of the four aforementioned covariates. We find that most portfolios are able to produce positive Sharpe ratios. It is noteworthy that the $mfFDR^+$ surpasses the other two methods. On the other hand, the performance of the $fFDR^+$ portfolio with PC1 as a covariate does not surpass that of all $fFDR^+$-based portfolios with the original individual covariates. On average, it beats β_{bh} and ρ but not the other two. This combines with fact that the $mfFDR^+$ portfolio based on four covariates outperforms the $fFDR^+$ portfolio based on PC1 suggest that the former is effective in extracting useful information, which might not be effectively incorporated via linear combinations, from different covariates in detecting outperforming rules. This is possible because the $mfFDR^+$ accounts for the interactions of covariates via the multivariate null proportion and joint density functions.

Table 2.4: Empirical power comparison. The table presents the average number of trading rules in portfolios based on the FDR^+ and $mfFDR^+$ at beginning of each month, controlling for FDR at 20%. The first five columns report the numbers in five sub-periods while the last column shows those numbers across the months from the first time forming portfolios till December 2020.

	Period											
	1973-1980		1981-1990		1991-2000		2001-2010		2011-2020		1973-2020	
Currency	FDR^+	$mfFDR^+$	FDR^+	$mfFDR^+$	FDR^+	$mfFDR^+$	FDR^+	$mfFDR^+$	FDR^+	$mfFDR^+$	FDR^+	$mfFDR^+$
Australia	4138	8710	2451	8804	359	6615	1104	8095	114	6191	1530	7636
Canada	4603	8089	1349	6810	335	5826	370	6497	1099	5751	1433	6533
Germany/E.U.	2835	8244	3865	10832	496	7225	655	7707	864	6012	1697	7984
Japan	6121	11179	3453	9309	1173	8791	25	6211	724	6499	2145	8282
New Zealand	3227	10921	4483	10066	1141	8163	2232	8298	137	5047	2195	8388
Norway	2218	7348	3914	10194	127	6278	1243	6592	990	5615	1675	7192
Sweden	2485	9306	3050	10704	2142	8036	688	8146	1288	6079	1906	8410
Switzerland	3255	9296	1531	8960	1053	6513	259	6299	353	4684	1210	7056
U.K.	6303	10826	2889	9646	321	4828	22	6389	349	5413	1803	7280
Argentina					1674	2132	1565	3390	8213	10230	4011	5524
Columbia					3711	9102	3723	9918	1821	6589	3010	8463
India					1305	9112	3399	9262	387	5249	1716	7774
Indonesia			20	3936	988	5291	2801	9641	3084	8593	2090	7493
Israel	15551	3570	9365	5464	1882	8918	1542	7876	620	6223	3658	7021
Philippines			10529	13022	4671	9466	3644	8230	357	6962	3348	8509
Romania					11420	14408	1915	5954	896	6175	2312	6823
Russia					6787	11236	3806	9383	2597	8139	3836	9198
Slovak					345	5616	1643	8526	890	5813	1061	6818
Brazil					4185	5536	1350	10093	650	8451	1583	8580
Chile					5425	10570	3694	8572	210	6692	2635	8214
Czech					16	5919	2646	8221	1670	5788	1619	6726
Hungary					331	6843	473	6934	133	4711	310	6096
Korea					3098	9697	1899	8850	124	5346	1547	7761
Mexico			63	1388	628	5454	905	5326	147	6579	526	5518
Poland					2427	7506	1617	6950	359	5365	1298	6447
Singapore			1060	6313	2422	6740	1014	6655	1068	5400	1415	6268
South Africa			2045	9723	2817	8579	1661	8192	187	6635	1649	8199
Taiwan			8678	9708	4078	9220	2860	8490	116	6551	3278	8317
Thailand					1247	7120	2917	9792	2483	7851	2284	8330
Turkey					13393	11654	1477	8028	1935	9302	5334	9596

Table 2.5: Sharpe ratios of portfolios before transaction costs. The table presents the annualized Sharpe ratios before transaction costs of seven portfolios based on the FDR^+, the $fFDR^+$ and $mfFDR^+$ controlling FDR at 20%. For the $fFDR^+$, we first consider four covariates: α_{bh}, β_{bh}, R^2_{bh}, ρ and the first principal component of the four mentioned covariates. For the $mfFDR^+$ we study $d = 4$ with all four covariates. The last row is the average Sharpe ratio across currencies. The numbers in parentheses are the corresponding $p-$values. "*", "**" and "***" respectively indicate statistical significance at levels of $10\%, 5\%$ and 1%.

Currency	FDR^+	$fFDR^+$					$mfFDR^+$
		α_{bh}	β_{bh}	R^2_{bh}	ρ	PC1	
Australia	0.16 (0.15)	0.07 (0.60)	0.14 (0.30)	0.14 (0.30)	0.14 (0.30)	0.12 (0.34)	0.18 (0.17)
Canada	0.09 (0.53)	0.00 (0.94)	0.11 (0.44)	0.10 (0.52)	0.09 (0.59)	0.11 (0.46)	0.17 (0.23)
Germany/E.U.	0.24 (0.06)*	0.24 (0.11)	0.35 (0.01)***	0.34 (0.01)***	0.23 (0.13)	0.31 (0.03)**	0.49 (0.00)***
Japan	0.49 (0.00)***	0.37 (0.01)***	0.27 (0.07)*	0.31 (0.03)**	0.15 (0.29)	0.30 (0.03)**	0.45 (0.00)***
New Zealand	0.27 (0.00)***	0.22 (0.08)*	0.21 (0.07)*	0.29 (0.01)***	0.27 (0.02)**	0.34 (0.00)***	0.34 (0.00)***
Norway	0.11 (0.44)	0.17 (0.22)	0.07 (0.61)	0.25 (0.08)*	0.02 (0.87)	0.02 (0.90)	0.18 (0.17)
Sweden	0.34 (0.00)***	0.46 (0.00)***	0.32 (0.02)**	0.30 (0.02)**	0.28 (0.03)**	0.32 (0.02)**	0.40 (0.00)***
Switzerland	0.02 (0.89)	0.22 (0.13)	0.13 (0.35)	0.15 (0.27)	0.05 (0.70)	0.16 (0.28)	0.26 (0.06)*
U.K.	0.24 (0.09)*	0.09 (0.52)	0.22 (0.13)	0.19 (0.20)	0.20 (0.17)	0.12 (0.37)	0.29 (0.05)**
Argentina	0.39 (0.12)	0.42 (0.01)***	0.42 (0.01)***	0.41 (0.02)**	0.47 (0.01)***	0.41 (0.02)**	0.35 (0.06)*
Columbia	1.02 (0.00)***	0.59 (0.00)***	0.51 (0.02)**	0.64 (0.00)***	0.75 (0.00)***	0.52 (0.02)**	0.60 (0.00)***
India	0.37 (0.01)***	0.25 (0.15)	0.17 (0.32)	0.29 (0.08)*	0.28 (0.09)*	0.35 (0.04)**	0.34 (0.03)**
Indonesia	0.40 (0.04)**	0.34 (0.02)**	0.21 (0.17)	0.25 (0.08)*	0.22 (0.16)	0.30 (0.03)**	0.34 (0.02)**
Israel	1.10 (0.00)***	0.94 (0.00)***	0.78 (0.00)***	0.75 (0.00)***	0.81 (0.00)***	0.81 (0.00)***	0.54 (0.00)***
Philippines	0.51 (0.00)***	0.61 (0.00)***	0.59 (0.00)***	0.56 (0.00)***	0.60 (0.00)***	0.63 (0.00)***	0.62 (0.00)***
Romania	0.07 (0.72)	0.08 (0.65)	0.09 (0.62)	0.10 (0.58)	-0.03 (0.94)	0.03 (0.88)	0.23 (0.19)
Russia	0.44 (0.08)*	0.41 (0.06)*	0.39 (0.06)*	0.42 (0.05)**	0.42 (0.06)*	0.43 (0.05)**	0.51 (0.01)***
Slovak	0.04 (0.77)	-0.01 (0.98)	0.20 (0.31)	0.12 (0.51)	0.08 (0.67)	0.14 (0.46)	0.18 (0.35)
Brazil	0.08 (0.65)	0.24 (0.20)	0.02 (0.91)	0.10 (0.63)	0.06 (0.72)	0.11 (0.60)	0.36 (0.07)*
Chile	0.56 (0.01)***	0.65 (0.00)***	0.44 (0.02)**	0.46 (0.01)***	0.61 (0.01)***	0.56 (0.01)***	0.46 (0.02)**
Czech	0.11 (0.51)	-0.01 (0.98)	0.05 (0.72)	0.27 (0.11)	-0.08 (0.68)	0.00 (0.93)	0.12 (0.46)
Hungary	0.01 (0.89)	-0.01 (1.00)	-0.03 (0.89)	0.03 (0.85)	-0.04 (0.90)	-0.06 (0.76)	-0.11 (0.60)
Korea	0.41 (0.07)*	0.33 (0.09)*	0.23 (0.22)	0.22 (0.23)	0.18 (0.34)	0.27 (0.16)	0.29 (0.16)
Mexico	0.17 (0.20)	0.17 (0.25)	-0.05 (0.76)	0.04 (0.75)	0.06 (0.65)	0.10 (0.52)	0.10 (0.49)
Poland	-0.07 (0.70)	0.07 (0.67)	0.05 (0.76)	0.00 (0.98)	0.10 (0.60)	0.15 (0.45)	0.02 (0.92)
Singapore	0.14 (0.38)	0.14 (0.40)	0.12 (0.53)	0.21 (0.20)	0.13 (0.45)	0.03 (0.88)	0.32 (0.02)**
South Africa	0.16 (0.33)	0.21 (0.18)	0.09 (0.53)	0.12 (0.41)	0.22 (0.14)	0.28 (0.07)*	0.23 (0.14)
Taiwan	0.77 (0.00)***	0.70 (0.00)***	0.55 (0.00)***	0.69 (0.00)***	0.52 (0.00)***	0.70 (0.00)***	0.73 (0.00)***
Thailand	0.34 (0.07)*	0.51 (0.01)***	0.27 (0.13)	0.29 (0.12)	0.37 (0.04)**	0.23 (0.22)	0.47 (0.02)**
Turkey	0.60 (0.00)***	0.62 (0.00)***	0.61 (0.00)***	0.60 (0.00)***	0.64 (0.00)***	0.62 (0.00)***	0.51 (0.00)***
Average	0.32	0.30	0.25	0.29	0.26	0.28	0.33

Table 2.6: Net Sharpe ratios of portfolios. The table presents the annualized Sharpe ratio after transaction cost of seven portfolios based on the FDR^+, the $fFDR^+$ and $mfFDR^+$ controlling FDR at 20%. For the $fFDR^+$, we first consider four covariates: α_{bh}, β_{bh}, R^2_{bh}, ρ and the first principal component of the four mentioned covariates. For the $mfFDR^+$ we study $d = 4$ with all four covariates. The second last row is the average of the Sharpe ratio across currencies. The last row is the $t-$statistic of the test comparing the (paired) means of the portfolios $fFDR^+/mfFDR^+$ to the portfolios FDR^+. The numbers in parentheses are the corresponding $p-$values. "*", "**" and "***" respectively indicate statistical significance at levels of $10\%, 5\%$ and 1%.

Currency	FDR^+	$fFDR^+$					$mfFDR^+$
		α_{bh}	β_{bh}	R^2_{bh}	ρ	PC1	
Australia	-0.02 (0.42)	-0.01 (0.97)	0.11 (0.42)	0.11 (0.42)	0.11 (0.42)	0.06 (0.63)	0.14 (0.29)
Canada	-0.25 (0.80)	-0.10 (0.50)	0.06 (0.69)	0.05 (0.80)	0.03 (0.89)	0.00 (0.98)	0.11 (0.46)
Germany/E.U.	0.13 (0.02)**	0.20 (0.17)	0.32 (0.02)**	0.31 (0.02)**	0.19 (0.19)	0.27 (0.07)*	0.46 (0.00)***
Japan	0.38 (0.06)*	0.31 (0.02)**	0.24 (0.11)	0.28 (0.06)*	0.11 (0.43)	0.25 (0.09)*	0.41 (0.00)***
New Zealand	0.00 (0.05)**	0.10 (0.43)	0.15 (0.24)	0.24 (0.05)**	0.19 (0.11)	0.24 (0.05)**	0.28 (0.02)**
Norway	-0.08 (0.14)	0.07 (0.64)	0.03 (0.88)	0.21 (0.14)	-0.04 (0.78)	-0.06 (0.64)	0.13 (0.36)
Sweden	0.17 (0.04)**	0.39 (0.01)***	0.28 (0.05)**	0.25 (0.04)**	0.24 (0.07)*	0.26 (0.05)**	0.36 (0.01)***
Switzerland	-0.12 (0.41)	0.16 (0.26)	0.10 (0.44)	0.11 (0.41)	0.01 (0.91)	0.10 (0.47)	0.22 (0.12)
U.K.	0.09 (0.30)	0.02 (0.87)	0.18 (0.22)	0.15 (0.30)	0.15 (0.29)	0.06 (0.64)	0.25 (0.08)*
Argentina	0.27 (0.03)**	0.36 (0.04)**	0.38 (0.02)**	0.37 (0.03)**	0.43 (0.01)***	0.35 (0.04)**	0.31 (0.11)
Columbia	0.71 (0.01)***	0.44 (0.04)**	0.40 (0.07)*	0.51 (0.01)***	0.59 (0.00)***	0.35 (0.10)*	0.53 (0.01)***
India	0.25 (0.12)	0.18 (0.29)	0.13 (0.45)	0.25 (0.12)	0.24 (0.15)	0.30 (0.07)*	0.30 (0.07)*
Indonesia	-0.09 (0.29)	0.23 (0.12)	0.12 (0.40)	0.15 (0.29)	0.11 (0.47)	0.19 (0.18)	0.24 (0.09)*
Israel	0.73 (0.00)***	0.75 (0.00)***	0.68 (0.00)***	0.63 (0.00)***	0.68 (0.00)***	0.65 (0.00)***	0.39 (0.01)***
Philippines	-0.42 (0.05)**	0.36 (0.05)**	0.40 (0.03)**	0.35 (0.05)**	0.39 (0.03)**	0.35 (0.07)*	0.46 (0.01)***
Romania	-0.31 (0.81)	-0.13 (0.53)	0.03 (0.86)	0.04 (0.81)	-0.13 (0.58)	-0.11 (0.60)	0.12 (0.51)
Russia	0.39 (0.07)*	0.38 (0.07)*	0.37 (0.07)*	0.40 (0.07)*	0.40 (0.07)*	0.40 (0.05)**	0.48 (0.01)***
Slovak	-0.12 (0.62)	-0.09 (0.68)	0.17 (0.38)	0.10 (0.62)	0.05 (0.76)	0.08 (0.63)	0.13 (0.48)
Brazil	-0.05 (0.71)	0.20 (0.29)	0.00 (0.98)	0.07 (0.71)	0.02 (0.94)	0.06 (0.78)	0.33 (0.10)*
Chile	0.22 (0.07)*	0.52 (0.01)***	0.35 (0.07)*	0.36 (0.07)*	0.47 (0.03)**	0.43 (0.05)**	0.38 (0.06)*
Czech	-0.11 (0.17)	-0.12 (0.55)	0.01 (0.89)	0.23 (0.17)	-0.14 (0.42)	-0.08 (0.70)	0.06 (0.65)
Hungary	-0.23 (0.97)	-0.09 (0.67)	-0.06 (0.75)	0.00 (0.97)	-0.08 (0.67)	-0.11 (0.55)	-0.17 (0.41)
Korea	0.15 (0.41)	0.18 (0.33)	0.15 (0.39)	0.14 (0.41)	0.09 (0.62)	0.14 (0.42)	0.23 (0.25)
Mexico	0.04 (0.97)	0.10 (0.49)	-0.09 (0.58)	0.00 (0.97)	0.02 (0.86)	0.03 (0.85)	0.04 (0.73)
Poland	-0.23 (0.89)	-0.01 (0.98)	0.03 (0.87)	-0.03 (0.89)	0.06 (0.77)	0.09 (0.64)	-0.03 (0.89)
Singapore	-0.38 (0.67)	-0.19 (0.25)	-0.02 (0.93)	0.08 (0.67)	-0.03 (0.87)	-0.20 (0.21)	0.15 (0.32)
South Africa	-0.27 (0.99)	0.01 (0.93)	-0.02 (0.93)	0.00 (0.99)	0.08 (0.62)	0.10 (0.54)	0.10 (0.49)
Taiwan	0.44 (0.00)***	0.56 (0.00)***	0.45 (0.00)***	0.59 (0.00)***	0.42 (0.01)***	0.57 (0.00)***	0.66 (0.00)***
Thailand	0.11 (0.30)	0.37 (0.05)**	0.18 (0.35)	0.19 (0.30)	0.27 (0.13)	0.07 (0.75)	0.38 (0.04)**
Turkey	0.36 (0.00)***	0.54 (0.00)***	0.53 (0.00)***	0.52 (0.00)***	0.56 (0.00)***	0.54 (0.00)***	0.45 (0.00)***
Average	0.06	0.19	0.19	0.22	0.18	0.18	0.26
t-statistic		3.9	3.5	4.8	3.6	3.4	5.1

Table 2.6 presents the portfolios' excess returns after transaction costs and shows a similar pattern. On average, the Sharpe ratio of the FDR^+ portfolio now is indistinguishable from zero. The final row of the table presents the $t-$statistic comparing the mean of Sharpe ratios of the $fFDR^+$ and $mfFDR^+$ portfolios across currencies to that of the FDR^+. All covariate-augmented portfolios statistically significantly outperform the FDR^+. The $mfFDR^+$ using all four covariates performs the best. This portfolio gains positive a Sharpe ratio for all currencies except the Hungarian forint and Polish zloty. 16 of these positive Sharpe ratios are statistically significant.

The implications from Tables 2.5 and 2.6 are threefold. First, the covariates are informative in a way that they help us to discover more out-performing rules that are able to deliver profits after trading costs are accounted for. Second, considering more covariates indeed enhances the OOS performance of technical trading rules portfolio, which justifies the advantage of $mfFDR^+$ and the importance of conditional hypotheses. Third, the fact that our method outperforms the $fFDR^+$ suggests that the information contents of four covariates may be non-linear and cannot be summarized by principal components, which again supports the strength of $mfFDR^+$.

2.5.3 Baskets of Currencies

In the prior section, we select out-performing trading rules separately in each currency. The profitability of technical trading rules is found to vary by currency and time period. It, therefore, might be beneficial for a trader to assess the performance of trading rules across currencies simultaneously. In doing so, the trader will be able to diversify and/or switch her/his funds across currencies. Thus, we assume that the trader can trade all currencies and select, with control of luck, out-performing rules in a pool of technical trading rules across currencies. Depending on the availability of the data, the pool consists of 190,755 ($= 9 \times 21,195$) to 635,850 ($= 30 \times 21,195$) trading rules. We construct a monthly rolling portfolio as in the previous section with the use of the new trading rule set. The four input covariates of the $mfFDR^+$ now are ρ, α_{mk}, β_{mk} and R^2_{mk} as described in Section 2.5.1. The quantiles of correlation coefficients of covariate pairs are shown in Table 2.7. Here, again a majority of about 98% of the coefficients are in $[-0.73, 0.53]$.

Table 2.7: Summary of correlation coefficients of covariate pairs: the case of all currencies. The table presents the quantiles of the correlated coefficient of six covariate pairs which are combinations of the four covariates: $R^2_{mk}, \alpha_{mk}, \beta_{mk}$ and ρ. The final row shows the numbers for the set of all correlated pairs.

Covariate pairs	min	1%	5%	25%	50%	75%	95%	99%	max
R^2_{mk}, α_{mk}	-0.44	-0.38	-0.27	-0.08	0.02	0.10	0.37	0.67	0.71
R^2_{mk}, β_{mk}	-0.81	-0.76	-0.70	-0.52	-0.23	0.12	0.50	0.54	0.55
R^2_{mk}, ρ	-0.26	-0.24	-0.18	-0.08	-0.02	0.07	0.25	0.30	0.36
α_{mk}, β_{mk}	-0.93	-0.90	-0.68	-0.31	-0.16	-0.03	0.31	0.53	0.63
$\alpha_{mk}\rho$	-0.20	-0.15	-0.11	-0.03	0.04	0.12	0.26	0.33	0.36
β_{mk}, ρ	-0.40	-0.36	-0.27	-0.11	-0.03	0.05	0.15	0.20	0.21
All pairs	-0.93	-0.73	-0.55	-0.16	-0.03	0.07	0.32	0.52	0.71

At beginning of each month, the trading rules selected by the $mfFDR^+$ procedure (control for FDR at 20%) are pooled together based on currency. For instance, suppose the $mfFDR^+$ identifies a set of out-performing rules which contains only rules from two currencies namely A and B. The numbers of these out-performing rules in the two currencies are k_A and k_B, respectively. The wealth is then split into k ($= k_A + k_B$) portions and k_A (k_B) of them are invested on the corresponding rules in currency A (B). We then calculate the performance of this portfolio for each month.

The performance of the $mfFDR^+$-based portfolios in terms of annualized Sharpe ratios and mean returns before and after transaction costs are exhibited in Table 2.8. The first row of each table reveals the performances over the whole sample period, from 1973 to the end of 2020. We find impressive Sharpe ratios with values of about 1.06 and 0.95 before and after transaction costs, respectively. The one-way break-even point (reported in the rightmost column), which is a fixed transaction cost at which the Sharpe ratio of the $mfFDR^+$ portfolio is set to zero, is as high as 60 basis points. The net Sharpe ratio indicates that in the whole sample period, the $mfFDR^+$ portfolio that consists of all currencies performs much better than any considered FDR^+-, $fFDR^+$-, and $mfFDR^+$-based portfolios that are traded on a single currency. The next rows of the table break down the performance of the portfolio into sub-periods of roughly ten years. While we find deterioration in Sharpe ratios over time, the $mfFDR^+$ portfolio is still fairly profitable even in the most recent decade (2011-2020). For instance, the net Sharpe ratio is 0.21 with a break-even cost of 14 basis points.

To better understand the evolution of time-series variation of technical profitability we conduct the following two analyses. The first one examines the proportion of techni-

Table 2.8: Performance of the $mfFDR^+$ portfolios. The table shows the annualized Sharpe ratios (SR) and mean returns (before and after transaction cost) of implementing the $mfFDR^+$ on all strategies in all currencies to control the FDR at 20%. The last column is the related break-even point.

Period	Excess SR	Net SR	Excess Return (%)	Net Return (%)	Break-even (bps)
Whole Period	1.06	0.95	3.80	3.40	60
1973-1980	1.45	1.35	4.47	4.18	69
1981-1990	2.08	1.97	7.30	6.93	128
1991-2000	0.92	0.81	4.15	3.64	72
2001-2010	0.69	0.53	2.37	1.82	34
2011-2020	0.29	0.21	0.88	0.63	14

cal rules being selected as profitable by the $mfFDR^+$ in each sub-period. The second one examines the portion of technical rules being selected as truly profitable by the $mfFDR^+$ in each currency in each sub-period, i.e., how prevalent the out-performing rules are in each category. [19]

For the first purpose, we calculate the ratio of the numbers of trading rules selected as profitable by $mfFDR^+$ in each month divided by the overall number of the input technical rules which varies from 190,755 (= 9×21,195) at the beginning of the sample period, when only nine currencies are considered, to 635,850 (= 30×21,195) at the end of the sample. These ratios are averaged so we have a selected ratio per month. We report the results for both the whole sample period and sub-periods in the first column of Table 2.9. We observe that overall, there are 27% of rules are profitable. This number is 36% in the first decade, then reduced overtime to 18% in the most recent decade (2011-2020).

Table 2.9 also provides the selected ratio in each category of trading rules. This measure shows how rich each technical rule category is in terms of containing outperforming rules and thus being useful for practitioners. We observe that, in the whole sample period, the moving average is the most profitable category with 34% of the set to be outperforming, followed by filter rule, support-resistance, RSI and channel breakout categories with 21%, 19%, 13% and only 8%, respectively. When breaking down the rates into sub-periods, more useful facts are revealed. Although the portions vary across periods, the cross-category distribution of the ratios appears consistent over

[19] We note that, the two analyses reflect how prevalent the out-performing rules (in in-sample of one-year periods) are under our framework, they do not reflect how well those rules perform individually in OOS. In our study, the OOS performance is checked after the selected rules are combined so that the transaction cost will be reduced by neutralizing the opposite trading signals.

Table 2.9: Out-performing rate by category. The table shows ratios of selected technical rules in each category selected by the $mfFDR^+$ under controlling for FDR at 20%. It exhibits the average of the ratios of technical rules (in %) in each category (Channel breakout (CB), Filter rule (FR), Moving average (MA), Relative strength relative (RSI) and Support-resistance (SR)) and in a whole pool of strategies (All) have been selected to invest each month over the whole period (first row) and sub-periods (remaining rows).

	Selected strategy rate					
Period	All	CB	FR	MA	RSI	SR
Whole period	27	8	21	34	13	19
1973-1980	36	11	30	45	19	26
1981-1990	36	10	27	46	15	26
1991-2000	23	7	18	29	12	16
2001-2010	23	7	19	29	11	16
2011-2020	18	6	13	23	11	13

time.

Tables 2.8 and 2.9 collectively offer three important insights: first, a substantial part of technical trading rules are still able to predict FX rates in recent years. Second, we still find substantial profits from technical rules in the most recent decade (2011-2020). And third, the moving average category is the richest source of out-performing rules.

2.5.3.1 Cross-currency distribution of the out-performing rules

It is well-known that technical profitability could decline over time with the improvement of efficiency in foreign exchange markets (especially in developed currencies). As documented in prior studies, the profitability of technical trading rules seems to vanish in developed currencies in the recent decades (Qi and Wu, 2006; Neely *et al.*, 2009); nevertheless, it still exists in several emerging currencies (Hsu *et al.*, 2016). It is therefore important for us to examine the evolution of the ratio of selected profitable trading rules (based on Sharpe ratios) across currencies over time. Table 2.10 presents the share of profitable rules selected by the $mfFDR^+$ among currencies in each decade. In the period from 1973 to 1980, when only developed currencies (and a short period of Israeli shekel) are considered, the most profitable currency is the Japanese yen. From 1981 to 1990, the Israeli shekel is the most profitable. And in the most recent decade, it is Argentine Peso. It is also clear that from 1990 the contribution of the selected profitable rules in developed currencies to the whole out-performing portfolios slowly decline. Overall, consistent with the literature, profitable technical rules are more

Table 2.10: Distribution by currency of trading rules selected by the $mfFDR^+$ (control for FDR at 20%) in sub-periods. The table displays the average proportion (%) of trading rules selected by the $mfFDR^+$ in each sub-period. For instance, the average number of outperforming rules selected at the end of each month in the last sub-period, 2011-2020, by the $mfFDR^+$ is 112,647 rules (the bottom number of the final column) and, on average, 3.1% of those rules are applied on Australian dollar (the top number of the final column). The bold numbers are the highest ones in a sub-period.

	Period				
Countries	1973-1980	1981-1990	1991-2000	2001-2010	2011-2020
Australia	9.6	6.7	3.0	3.5	3.1
Canada	9.3	4.2	2.5	2.4	2.3
Germany/E.U.	9.0	9.3	3.9	3.3	3.0
Japan	**14.8**	7.9	4.8	2.2	3.2
New Zealand	12.1	8.4	3.7	3.9	2.3
Norway	7.7	8.1	3.5	2.9	2.8
Sweden	9.8	9.1	4.4	3.5	3.1
Switzerland	11.4	7.9	4.0	2.6	2.3
U.K.	12.9	7.7	2.7	2.4	2.5
Argentina			1.1	2.8	**10.1**
Columbia			2.8	4.8	3.2
India			3.7	3.4	2.6
Indonesia		1.7	2.8	4.1	3.7
Israel	3.2	**10.9**	4.4	2.7	2.3
Philippines		2.9	6.1	3.2	3.1
Romania			1.5	3.4	3.2
Russia			2.9	4.0	4.5
Slovak			1.1	4.0	2.9
Brazil			1.2	**4.9**	5.0
Chile			2.2	3.5	3.0
Czech			1.7	3.9	3.2
Hungary			2.5	3.4	2.6
Korea			3.1	3.7	2.2
Mexico		0.2	3.0	1.9	3.3
Poland			1.6	3.2	2.9
Singapore		2.9	3.2	2.1	2.2
South Africa		6.2	4.3	4.1	4.1
Taiwan		5.9	4.7	2.6	2.5
Thailand			1.9	3.8	3.2
Turkey			**11.6**	3.8	5.4
Average number of selected rules (per month)	69,028	93,317	122,476	146,532	113,160

popularly found in emerging currencies and are less commonly in developed ones since 1990, likely due to market efficiency (Neely *et al.*, 2009).

2.5.3.2 Portfolios of rules conditional on category

Table 2.9 clearly show that the performance of rules are different across category. Put differently, the category is itself an informative covariate. It is therefore interesting to see how $mfFDR$ performs conditionally on the category. However, this covariate is not continuous and thus cannot be an input of the $mfFDR$. To study its informativeness, we therefore repeat the experiment with basket of currencies but on each of the category. In doing so, we are able to answer two questions: first, which category is most profitable OOS; and second, which categories traders should consider or avoid when constructing a portfolio. More specifically, for each category of trading rules we pool the rules across currencies belonging to the category to form a new set of rules. Thus, we have five new sets: the largest set is that of the Moving Average category with number of rules varying from 9×2,870 =25,830 to 30×2,870 = 86,100 rules while the smallest set is that of RSI category containing from 9×600 = 54,000 to 30×600 =18,000 rules. We construct five $mfFDR$-based portfolios which control FDR at 20%, each for one of the new sets.

In Table 2.11, we present the OOS performance of those selected rules. We learn several facts from the results. First, generally the portfolio conditional on filter rules is the most profitable OOS before transaction cost and remains so alongside with moving average after transaction cost. Second, the fact that the break-even points corresponding with filter and support-resistance categories are lower than those of channel breakout and moving average ones implies that the former ones are trading in higher frequencies, and thus generate more trading cost. The implication is that, in the time when transaction cost is high, the traders should avoid using or containing the filter and support-resistance rules in the portfolio. Third, the RSI category performs worst with negative profit for all sub-samples and thus for whole sample period. This implies that the performance of the selected out-performing RSI rules is not persistent. Thus, if traders construct a $mfFDR$ portfolio based on all rules, they should exclude the RSI rules from the pool.

As a robustness check, we repeat the exercises with the use of mean excess return as the performance metric in hypothesis testing. We find a consistent pattern, which is presented in Appendix B.5 due to the interest in space.

As final remarks, we observe from Table 2.9 that the proportion of profitable rules detected by the $mfFDR^+$ are significant across decades though slightly decreasing. From Table 2.8, we see that the profit that the selected rules generate also decline though

Table 2.11: Performance of the $mfFDR^+$ portfolios implemented on each category. The table shows the annualized Sharpe ratios (SR) and mean returns (before and after transaction cost) of portfolios generated by implementing the $mfFDR^+$ on each category of trading rule (on all currencies) to control the FDR at 20%. The last column is the related break-even point. In each panel, we present the mentioned metrics for the portfolio implemented on a category in whole sample period (first row) and sub-samples (next five rows).

Period	Excess SR	Net SR	Excess Return (%)	Net Return (%)	Break-even (bps)
			Panel A: Channel Breakout Rule		
Whole Period	0.88	0.80	2.64	2.40	72
1973-1980	0.82	0.76	2.05	1.88	62
1981-1990	2.03	1.95	5.81	5.57	184
1991-2000	0.98	0.89	3.45	3.14	99
2001-2010	0.47	0.36	1.43	1.11	37
2011-2020	0.12	0.06	0.33	0.18	7
			Panel B: Filter Rule		
Whole Period	1.26	0.91	4.41	3.19	23
1973-1980	1.77	1.53	5.62	4.86	36
1981-1990	2.17	1.81	7.14	5.97	41
1991-2000	0.98	0.66	4.90	3.27	27
2001-2010	1.02	0.44	2.95	1.28	14
2011-2020	0.67	0.37	1.71	0.95	7
			Panel C: Moving Average		
Whole Period	1.01	0.92	3.80	3.47	71
1973-1980	1.39	1.32	4.49	4.24	89
1981-1990	1.99	1.91	7.42	7.12	156
1991-2000	0.92	0.83	4.23	3.82	89
2001-2010	0.65	0.53	2.43	1.97	41
2011-2020	0.18	0.11	0.57	0.37	9
			Panel D: RSI		
Whole Period	-0.46	-0.96	-0.50	-1.04	-
1973-1980	-0.73	-1.16	-0.63	-0.99	-
1981-1990	-0.20	-0.70	-0.25	-0.84	-
1991-2000	-1.04	-1.67	-1.05	-1.70	-
2001-2010	-0.15	-0.73	-0.19	-0.90	-
2011-2020	-0.40	-0.74	-0.40	-0.74	-
			Panel E: Support-resistance		
Whole Period	1.00	0.80	3.13	2.48	32
1973-1980	1.37	1.20	3.64	3.18	41
1981-1990	2.06	1.83	6.02	5.34	64
1991-2000	0.87	0.68	3.68	2.87	41
2001-2010	0.55	0.24	1.54	0.67	15
2011-2020	0.34	0.19	0.87	0.49	8

still profitable in recent decade. There are several potential reasons for these facts. First, the introduction of high frequency trading with the involvement of complicated trading algorithm implemented by high performance computer. It is worth to remark that the trading frequency of rules under studying is daily through beginning to the end of the sample. Experiments with use of hourly data to generate a set of hourly rules can improve the performance of the portfolios. This is devoted for the future study. Second, alongside with the high frequency trading, the introduction of machine learning and in particular deep learning might explain the mentioned pattern of the profit generated by the portfolios. Technical analysis can be divided into two types that are quantitative and qualitative. This study focuses on only the former type. The latter type aims to detect profitable price pattern. The improvements of computer power and developments of deep learning allow the qualitative technical analysis being computationally conducted. For instance, Jiang *et al.* (forthcoming) show that the convolutional neural network, a deep leaning model, can detect profitable candlestick chart patterns across global stock markets. Such technique might have been implemented by investment institutions such as hedge funds during recent decades across financial markets including FX.

2.6 Concluding remarks

We introduce the $mfFDR$ testing method, which estimates the FDR as a function of multiple covariates, for detecting the false null hypotheses with control of FDR for large-scale multiple testing problems. We show that the method works well in controlling FDR and gains a considerably higher power than existing methods in detecting false null hypotheses. Our use of multiple informative covariates helps us examine predictors' conditional performance by incorporating a comprehensive set of information and is applicable to important finance research questions.

Empirically, we apply the $mfFDR$ method to a large universe of technical trading rules to detect truly profitable ones with control of data snooping biases. We first study the trading rules of each currency individually. With the use of multiple informative covariates, the results show that the $mfFDR$-based portfolio is much more powerful than the existing methods that are not using covariates. More importantly, this method quantifies the conditional performance of technical rules, which is more realistic because currency traders and portfolio managers review and select trading strategies based not only on a single performance metric but on a set of comprehensively updated information.

We find that the $mfFDR$-based portfolio of selected rules generates positive profits higher than those based on prior data snooping control methods with and without using a sole covariate. We then study 30 currencies together where more than 600 thousand trading rules are generated. We implement the $mfFDR$ method on this set of rules to construct a portfolio that generates a Sharpe ratio of roughly one for roughly 50 years. Moreover, this portfolio generates out-of-sample profits even over the recent decades.

The development of the $mfFDR$ framework will contribute to complex problems in Finance, Economics and other fields of Social Sciences that are plagued by multiple competing models and hypotheses. It is a powerful framework that is easy to implement and robust to noise, making our method a promising tool in decision-making in the era of big data.

Chapter 3

Picking hedge fund with high confidence

3.1 Introduction

The literature on hedge fund performance has extensively focused on understanding the risk-return characteristics of hedge funds. Those studies cover the persistent performance of hedge fund managers (see, e.g., Agarwal and Naik, 2000; Baquero *et al.*, 2005; Kosowski *et al.*, 2007; Sun *et al.*, 2018), and the relationship between characteristics of the funds and their performance. Based on these findings, investors can create portfolios by sorting hedge funds based on their characteristics and past performance. However, forming portfolios in this manner has several caveats that make them impractical in practice. Firstly, due to the large number of funds available, sorting portfolios, such as those created through quintile partition, can become large in size. Each hedge fund typically requires a significant minimum investment, which implies a huge investment requirement. Second, if only a few funds is chosen, via the rankings of the characteristics and other potential predictors, it is likely that some lucky funds are selected without proper control. Third, existing approaches in fund performance literature that controls for lucky funds focus on false discovery rate (FDR), which is expected proportion of non-outperforming funds among those selected as out-performers. In hedge fund application, controlling for this type I error is too loose as there exist many outperforming hedge funds (see, e.g., Chen *et al.*, 2017). Consequently, portfolios of hedge funds are again too large in size and more importantly, there are virtually some non-outperforming funds in portfolio at all time. To have high confidence in hedge fund portfolio selection, investors require a tool to control for a more stringent error. For this purpose, controlling for the family wise error rate (FWER), which is probability of selecting at least one non-outperforming fund in forming the portfolio, provides a well-suited solution. To explain this, suppose the investors control for FDR at 5% when forming their portfolio. Then they would expect there are always about 5% of funds in the portfolio are non-outperforming. In contrast, if they control for FWER at 5%, the chance of having non-outperforming funds in the portfolio is 5%. Loosely speaking, if they form such portfolios yearly over 100 years, they should expect only about 5 years where their portfolio containing some non-outperforming funds. Thus, when the investors opt to control the FWER instead of the FDR, they gain a much higher level of confidence in their investment decision, especially when substantial amounts of capital is involved, as is often the case in hedge fund investments.

Literature in controlling for FWER in detecting out-performers is rich with notable contributions of White (2000), Romano and Wolf (2005), Hansen (2005) and Hsu *et al.* (2010). The main focuses are developing testing procedures that control for FWER while enhancing the performance in terms of power. All of the mentioned works rely on bootstrapping procedures and exploit only raw information such as return of funds or trading strategies. Investors' flows chase for funds that are truly out-performing, i.e., the ones that generate positive alpha - the excess return adjusted for some passive benchmark. However, hedge fund return series are usually short. Consequently, the investors assess funds based on a short periods of time, typically 24 or 36 months (see, e.g., Kosowski *et al.*, 2007; Cumming *et al.*, 2012; Chen *et al.*, 2017). Given the small number of observations, the mentioned existing FWER methods are struggling in detecting even a small number of the out-performing hedge funds. A more powerful procedure is therefore in high demand.

This chapter fills the gap by introducing a new approach to control for FWER. The new approach is based on statistical framework of Zhou *et al.* (2021), which estimates the FWER as a function of multiple informative covariates. The new approach deviates from the framework of Zhou *et al.* (2021) by specifically aiming to control the FWER among the discoveries in the right tail. It is well-suited to controlling for luck in the hedge fund portfolio selection. First, it has a high power in detecting out-performing hedge funds, which allows investors picking fund with a very low FWER. Secondly, while effectively controlling for such a low error rate the approach is able to select a reasonable number of funds, thereby making the investment size more feasible. Third, and more importantly, the hedge funds identified by the new approach, though based on assessment over a short period, perform persistently in some long out-of-sample (OOS) periods. This makes the method a high potential in real world practice.

Distinguished from existing methods that solely rely on funds' adjusted returns or alpha, our approach harnesses additional information to enhance the detection power. More specifically, we assess performance of a fund via testing its alpha against zero. We implement the framework of Zhou *et al.* (2021) with use of the additional information to detect all funds that having significant non-zero alpha. As we are focusing on the right tail of the alpha distribution, we select as out-performing funds the subset of those detected non-zero alpha funds having positive estimated alpha. We name this procedure the *FWER "plus"* ($fwer^+$). We show that the procedure controls well for FWER at any given targets, and when an informative covariate is available, it gains an impressive power in detecting truly positive alpha funds. Controlling for FWER at 5% target, our procedure outperforms the stepwise approaches of Romano and Wolf (2005) and Hsu *et al.* (2010), the two most powerful ones in recent finance and economics applications, with gaps that varies from 1% to 15% depending on the number of observations per fund and the informativeness level of the covariates.

In empirical experiments, we first construct yearly rolling portfolios of out-performing hedge funds with control for FWER at small targets spanning from 0.1% to 5%. This choice of rolling window is to overcome the lockup period which might create restrictions on trading or incur additional cost. We use 20 covariates that are available and easily calculated from the hedge funds return. The $fwer^+$ with use of single or multiple covariates always detects non-empty group of out-performing funds despite of the small targets of FWER and the choice of short in-sample (IS) periods such as 24, 36 or 48 months. We then invest in the selected funds in the following year as OOS and roll forward till the end of 2021. The portfolios gain statistically significant positive alphas which spans from 4% to more than 5% per annum. We see that, the portfolios controlling for smaller target of FWER tend to gain higher alphas, which transform to Sharpe ratio of more than 2. These results are robust to the choice of IS periods and asset pricing models. To further examine the persistence in the performance of the funds detected by the $fwer^+$, we consider the choices of longer OOS periods. We see that even with OOS of four years, the $fwer^+$ portfolios still gain roughly as high alpha as the choice OOS of one year. This suggests that the $fwer^+$ helps the investors selecting truly out-performing funds based on assessing funds over a relative short past performance. We then enhance the informativeness of the covariates via using machine learning techniques to forecast future funds' return and use them as new covariates.[1] The $fwer^+$ portfolios with use of those new covariates are be able to generate Sharpe ratio of more than 2.5. Finally, we construct portfolios of only a single fund who performs best among those selected by the $fwer^+$ in in-sample period. These portfolios gain slightly lower alpha than those in previous exercise but perform impressively in terms of Sharpe ratio which can reach 5.3.

This chapter thus contributes to literature in both econometrics and economics aspects. First, it provides a powerful approach to detecting out-performers, which is in demand not only in hedge funds context but also in many other topics in economics and finance. Second, it shows that the performance of the genuine hedge funds detected by the new procedure persists. By using the new method, the investors can detect skilled hedge fund managers who are able to repeat their performance for long future horizons.

The chapter is organised as follows. Section 3.2 introduces our $fwer^+$ procedure. Section 3.3 describes our hedge fund dataset while Section 3.4 discusses on our choice of funds' performance measure. Section 3.5 provides simulations to show the performance of the $fwer^+$. Section 3.6 is devoted for empirical analyses while Section 3.7 concludes our chapter.

3.2 FWER and informative covariates

Suppose that we are assessing n hedge funds based on a performance metric ϕ. We test for each fund i a hypothesis

$$H_0 : \phi_i = 0 \quad \text{against} \quad H_1 : \phi_i \neq 0. \tag{3.1}$$

[1]The idea of enhancing the forecasting power via combining predictors has a long history of developments. Wang *et al.* (2023) provide a comprehensive review on this topic. In this chapter, we are studying the machine learning methods that have been recently gained attention in asset pricing literature such as Gu *et al.* (2020) and Wu *et al.* (2021).

where ϕ_i is the true but unknown value ϕ of the fund i, $i = 1, \ldots, n$.

This study focuses on detecting out-performing hedge funds based on their alpha, i.e., the metric ϕ_i is alpha of the fund i, $i = 1, \ldots, n$. We aim to detect the funds having positive ϕ, with controlling for the probability of selecting at least one non-outperforming fund at a predetermined target $\tau \in (0, 1)$. Formally, let R be number of funds selected as out-performing funds and among them F funds actually having $\phi \leq 0$. We are attempting to control a type I error which is defined as FWER $= \mathbb{P}(F \geq 1)$ at the target τ.

Literature in detecting out-performers with controlling for FWER focuses on one-sided test with a composite null, and numerous testing procedures have been developed. Most recent contributions are the procedures of Romano and Wolf (2005), Hansen (2005) and Hsu *et al.* (2010). The common ground of those procedures are based on bootstrapping with use of funds' returns (or more generally, some relative performance variable). These procedures suffer from the computational burden and are not using additional information, which are informative and available alongside the tests. In picking out-performing funds, which typically assesses funds' performance based on short time series of return, these approaches appear to be lack of power.

The low power issue of multiple testing procedures has been gained attention across areas of science. Recent developments in statistics attempt to incorporate additional information to raise the power of detecting false nulls, such as the contributions of Ignatiadis and Huber (2021) and Zhou *et al.* (2021). Nevertheless, the implementation of those innovative approaches in the selecting out-performers has not been addressed. In this study we are proposing a simple procedure to further develop the framework of Zhou *et al.* (2021) to solving the low power problem of the existing approaches.[2] In the follows, we summarise the framework and subsequently propose our procedure.

3.2.1 The use of informative covariates in controlling FWER

Suppose we set a target of FWER at $\tau \in (0, 1)$ and there is a set of d informative covariates $Z^1, \ldots, Z^d$ carrying the information on probability that the null of tests (3.1)

[2]The framework of Zhou *et al.* (2021) is more sufficient in terms of computation and has been shown to be more powerful than that of Ignatiadis and Huber (2021).

being true. Each Z^k is a column vector $(Z_1^k, \ldots, Z_n^k)'$, $k = 1, \ldots, d$. For convenience, we denote $\boldsymbol{Z} = (Z^1, \ldots, Z^d)$ and thus $\boldsymbol{Z}_i$ means $(Z_i^1, \ldots, Z_i^d)$. For each $i = 1, \ldots, n$, let P_i be the random variable representing the $p-$value of the test (3.1) corresponding to the fund i and p_i be its realization. Conditional on $\boldsymbol{Z}_i = \boldsymbol{z}_i$, we denote the prior probability of the null hypothesis i being true by $\pi_0(\boldsymbol{z}_i)$. We model the distribution of P_i as a mixture of two groups in which the weights of the first and the second group is $\pi_0(\boldsymbol{z}_i)$ and $1 - \pi_0(\boldsymbol{z}_i)$, respectively. Let f_0 and f_{alt} be the density functions of the first and second group, respectively. Formally, we have

$$P_i|(\boldsymbol{Z}_i = \boldsymbol{z}_i) \sim \pi_0(\boldsymbol{z}_i) f_0(\cdot) + (1 - \pi_0(\boldsymbol{z}_i)) f_{alt}(\cdot) \tag{3.2}$$

Thus, the covariates $\boldsymbol{Z}$ carry their information through the $\pi_0(\boldsymbol{z}_i)$ which takes value differently across the tests. In contrast, the f_0 and f_{alt} respectively are the density functions of those $p-$values under true nulls and false nulls, and they are in the same form for all tests, i.e., not depending on i. In this model, the two density functions do not depend on $\boldsymbol{Z}$ neither. We assume the $p-$value of a true null is uniformly distributed, i.e., $f_0(p) = 1\ \forall p$.

In conventional approaches, the rejection region is determined by a common threshold T which is fixed for all tests, i.e, a hypothesis i is rejected if and only if $p_i \leq \Theta$. The idea now is to determine for each hypothesis i a threshold which is a function of $\boldsymbol{z}_i$ denoted by $\Theta(\boldsymbol{z}_i)$, i.e., the null of hypothesis i is rejected if and only if $p_i \leq \Theta(\boldsymbol{z}_i)$. For this purpose, we assume the $f_{alt}(p)$ to be a strictly decreasing function of p and follow the developments of Zhou *et al.* (2021), the mentioned threshold is defined as

$$\Theta(\boldsymbol{z}_i) = f_{alt}^{-1}\left(\frac{\pi_0(\boldsymbol{z}_i)\theta^*}{1 - \pi_0(\boldsymbol{z}_i)}\right) \tag{3.3}$$

where $\theta^* = \min\left\{\ \theta > 0 :\ \sum_{i=1}^n \pi_0(\boldsymbol{z}_i) f_{alt}^{-1}\left(\frac{\pi_0(\boldsymbol{z}_i)\theta}{1-\pi_0(\boldsymbol{z}_i)}\right) \leq \tau\ \right\}$ and f_{alt}^{-1} is the inverse function of f_{alt}.

In practice, the $\pi_0(\boldsymbol{z})$ is modelled as a logit function which has a form of $\pi_0(\boldsymbol{z}_i) = 1/(1 + e^{-b_0 - \boldsymbol{b}'\boldsymbol{z}_i})$, where $\boldsymbol{b} = (b_1, \ldots, b_d)$ is the column vector of the coefficients of the d covariates, while $f_{alt}(p)$ is modelled as a beta distribution $f_{alt}(p) = kp^{k-1}$ for $k \in (0, 1)$. The parameters $b_0, \ldots, b_d$ and k are estimated via an expectation and maximization algorithm.[3]

3.2.2 Application in picking outperforming funds

As we aim to picking truly positive alpha funds, we actually need to controlling for FWER in the group of the selected out-performing funds. In this section, we propose a simple procedure to applying the framework of Zhou *et al.* (2021) to the problem.

Given a target $\tau \in (0, 1)$ of FWER, our procedure, namely $fwer^+$, consists of two steps. First, we implement FWER procedure of Zhou *et al.* (2021) on the population of funds with controlling FWER at the target τ. The procedure will detect a set of abnormal funds, say A, which includes both under- and out-performing funds. Second, we pick from this set only a subset consisting of those funds having positive estimated alpha, say A^+. Since the probability of having at least one zero-alpha funds in A is less than τ, this also conservatively holds for the set A^+ as it is a subset of A. Assuming that there are no truly under-performing funds in A that are very lucky and selected into A^+, then the set A^+ consists of out-performing funds with FWER being controlled at the target τ. As controlling for FWER is stringent, this assumption is likely to be valid.

As will be shown in Section 3.5, the $fwer^+$ controls well for FWER at any given targets and, when informative covariates are available, it is more powerful than existing methods.

3.3 Data

Our hedge fund data is collected from Lipper TASS database. Following previous research, we impose screenings to deal with common sample biases (see, Fung and Hsieh (2001); Bali *et al.* (2012); Chen *et al.* (2021b)). We include only US dollar-based hedge funds in our sample to avoid duplicate funds listed in different currencies. We do not consider funds that have not reported any data during the study period as well as we include both "live" and "graveyard" funds from January 1994 to account for survivorship bias. To address the back-fill bias issue, we exclude the first 12 months of returns for each fund. To control for multi-period sampling bias, we require all funds

[3] For the details of developments and algorithms, readers are referred to Zhou *et al.* (2021).

have at least 36 months of return history. For each fund, we consider only months where the fund's net-of-fee return and asset under management data are available. After the above restrictions, we end up with a sample of 5,314 funds covering the period January 1994 to December 2021.[4]

3.4 Performance measure

Following majority of the existing literature on hedge fund performance, we use the seven-factor model alpha of Fung and Hsieh (2004) as our baseline performance measure of a fund.[5] For each fund i we regress

$$r_{i,t} = \alpha_i + \boldsymbol{F_t}\boldsymbol{\beta}_i + \varepsilon_{i,t} \tag{3.4}$$

where $r_{i,t}$ is the excess return, $\boldsymbol{F_t} = [F_t^1, F_t^2, \dots, F_t^7]$ is the 1×7 matrix of the seven risk factors, $\boldsymbol{\beta}_i = [\beta_i^1, \beta_i^2, \dots, \beta_i^7]'$ is the 7×1 matrix of coefficients, and $\varepsilon_{i,t}$ is the noise of the fund i at month t. The seven factors include an equity market factor which is the S&P500 return minus risk-free rate; the Wilshire small cap minus large cap return; the change in the constant maturity yield of the 10-year Treasury ($\Delta 10Y$); the change in the spread between Moody's Baa yield and the 10-year Treasury ($\Delta CredSpr$); and 3 trend-following factors for bonds ($BDTF$), currency ($FXTF$), and commodities ($CMTF$).

Later, we additionally conduct robustness check for our analyses under the use of different models in measuring the fund's alpha. They include the four-factor model of Carhart (1997), a six-factor model where we add the two risk factors $\Delta 10Y$ and $\Delta CredSpr$ into the four-factor model, and a nine-factor model where three more risk factors including $BDTF$, $FXTF$ and $CMTF$ are added into the six-factor model. The four risk factors in the four-factor model consist of the market's excess return on the CRSP NYSE/Amex/NASDAQ value-weighted market portfolio, the Fama–French small minus big factor, the high minus low factor, the momentum factor.[6]

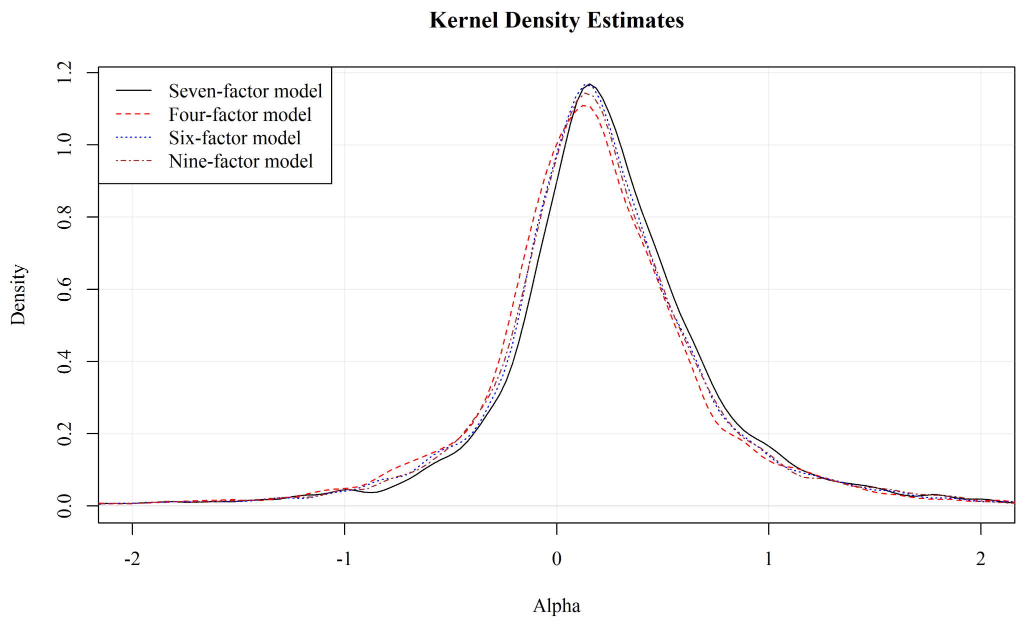
[4]Following hedge fund literature (see Chen *et al.* (2023)) we exclude monthly net-of-fee returns that are below -90% or excess 300%.

[5]See, e.g., Kosowski *et al.* (2007) and Chen *et al.* (2017).

[6]We follow Hsieh's and French's websites to collect the seven risk factors of the Fung and Hsieh (2004) seven-factor model, the one-month Treasury bill rate and four risk factors of the Carhart (1997) four-factor model.

Figure 3.1 depicts the distributions of hedge fund alphas under the use of the mentioned four models via estimating their kernel density curves. We see that all curves have a peak at some positive alpha point. This is similar to reports in hedge fund literature such as those of Chen *et al.* (2023). It indicates the presence of a majority positive alpha hedge funds under the considering factor models.

Figure 3.1: Distribution of hedge fund alphas. The figure shows the kernel density estimates of funds' alpha with use of different factor models including four-, six-, seven-, and nine-factor models. We require at least 36 observations per fund and for each model we regress excess return of each individual fund on the model's risk factors to obtain its alpha and then estimate the kernel density of the alpha population.



3.5 Simulation studies

In this section, we conduct a set of simulations to show: i) our proposed approach works well in terms of controlling FWER and outperforms the existing methods in terms of power; ii) the excellent performance of the $fwer^+$ under variants of important factors.

As presented in Section 3.2.1, the FWER is estimated based on a mixture model assumption in which the informativeness of the covariates is conveyed via the null proportion $\pi_0(\boldsymbol{z})$. Thus, the relationship between $\boldsymbol{Z}$ and the non-zero alpha is not be reflected in the model.[7] This implies that, in simulation design, the non-zero alpha

[7]This differs from other models, such as Chen *et al.* (2021a), where they take into count the joint

component can be freely generated and not depending on the value $\boldsymbol{Z}$. Thus, the value of $\boldsymbol{Z}$ will be generated in the first step. Based on these values and assumption on $\pi_0(\boldsymbol{z})$ we assign the true nulls, i.e., determine which funds have zero-alpha. The remaining funds will be assigned with non-zero alpha values from a predetermined distribution. Remark that, the signals of being false null, i.e. the magnitude of the non-zero alphas, are transformed to rejection rule via the estimated f_{alt}.

To estimate necessary parameters for data generating process, we use data of all the 5,314 funds in our sample and the risk factors of our baseline model. Specifically, we utilize the seven risk factors data which start from January 1995 to December 2021 and estimate their mean and matrix of correlation coefficients. We calculate the factor loadings ($\boldsymbol{\beta}_i$) for the funds by regressing each of the 5,131 funds on the risk factors. We design simulations to cover different scenarios in applications. We generate balanced panel data of n funds with T observations per fund. As hedge fund literature focuses on constructing portfolio with use of 24-, 36- and 48- month IS periods, we first consider $T \in \{24, 36, 48\}$.

In each iteration, we conduct 7 steps as followings.

1. We generate two covariates $\boldsymbol{Z} = (U, V)$ from a bivariate normal distribution with mean 0 and standard deviation of 1 and with specific correlation coefficient ρ of U and V. We consider two cases of $\rho \in \{0, 0.5\}$.

2. The $\pi_0(u, v)$ has a logit form $1/(1 + e^{-b_0 - b_1 u - b_2 v})$ where the triple (b_0, b_1, b_2) is one of the three cases $(0.5, 1, 1)$, $(0.5, 1.5, 1.5)$ and $(0.5, 2, 2)$. These three cases cover a weak, moderate and strong relationship between covariates and the probability of a fund being zero alpha, respectively. The choice of $b_0 = 0.5$ is to generate a set of simulated hypotheses with a null proportion, denoted by π_0, of 60%. For non-zero alpha funds, we generate data sets such that a half of them have positive alpha.[8]

   Given a specific choice of $\pi_0(u, v)$, we determine funds having zero-alpha funds

distribution of alternative hypothesis' $p-$value and covariates. This way, the link between the non-zero alpha and value of covariates is reflected in the models.

[8]By using procedures of Storey (2002) and Barras *et al.* (2010) we find the estimated proportions of zero-alpha (π_0) and out- and under-performing funds (π^+ and π^-) are 59%, 41% and 0%. To cover general scenarios in applications we consider $\pi_0 = 60\%$ and $\pi^+ = \pi^- = 20\%$. We conduct robustness check for alternative choice of the proportions, see footnote 10.

as follows. For each fund i, we draw a random value from Bernoulli distribution which takes value 0 with probability of $\pi_0(u_i, v_i)$. Funds with drawn values of 0 are assigned as zero-alpha funds. We assign randomly a half of the remaining funds with alpha of $\alpha > 0$ and the rests with alpha of $-\alpha$ where the monthly alpha $\alpha \in \{0.5\%, 1\%, 1.5\%\}$. The choice of 0.5% is close to the third quantile of the estimated alphas in our data sample while other values are chosen under assumption that the true α is some value in between the third quantile and the maximum of the estimated alphas.

3. We generate the risk factors, $\boldsymbol{F}^s$ and their loadings $\boldsymbol{\beta}^s$ from normal distributions such that their parameters are the same as those of the real sample counterpart in whole sample period (i.e., from January 1995 to December 2021).

4. We generate the simulated excess return of each fund via the following formula

$$R_{i,t} = \alpha_i^s + \boldsymbol{F}_t^s \boldsymbol{\beta}_i^s + \varepsilon_{i,t} \tag{3.5}$$

   where the noise $\epsilon_{i,t}$ is drawn independently from a normal distribution $N(0, \sigma^2)$ with σ to be set at 2.2% as the median of standard deviation of error terms estimated from the real sample fund-by-fund regressions.

5. For each fund, we then regress its simulated returns on the seven simulated factors, $\boldsymbol{F}_t^s$, to obtain its estimated $\hat{\alpha}$ and the $p-$value of testing its alpha against 0.

6. We implement the stepwise reality check (StepM) of Romano and Wolf (2005), stepwise superior predictive ability (Stepwise-SPA) of Hsu *et al.* (2010) and $fwer^+$ procedures, controlling for FWER at predetermined targets $\tau \in (0, 1)$, to detect truly positive alpha funds with use of the $\hat{\alpha}$s, calculated $p-$values and simulated covariates.[9] We consider $\tau \in \{0.1\%,\ 1\%,\ 2\%,\ \ldots,\ 20\%\}$.

7. By comparing the simulated α^s and the selected out-performing funds, we record the family wise error (FWE) which takes value 1 if there is at least one of the funds in the negative or zero α^s groups classified as out-performers. We also calculate the detected proportion which is the ratio of truly out-performing funds

[9] The detail of the StepM and StepSPA procedures under our framework is presented in Appendix C.1.

detected by each procedure.

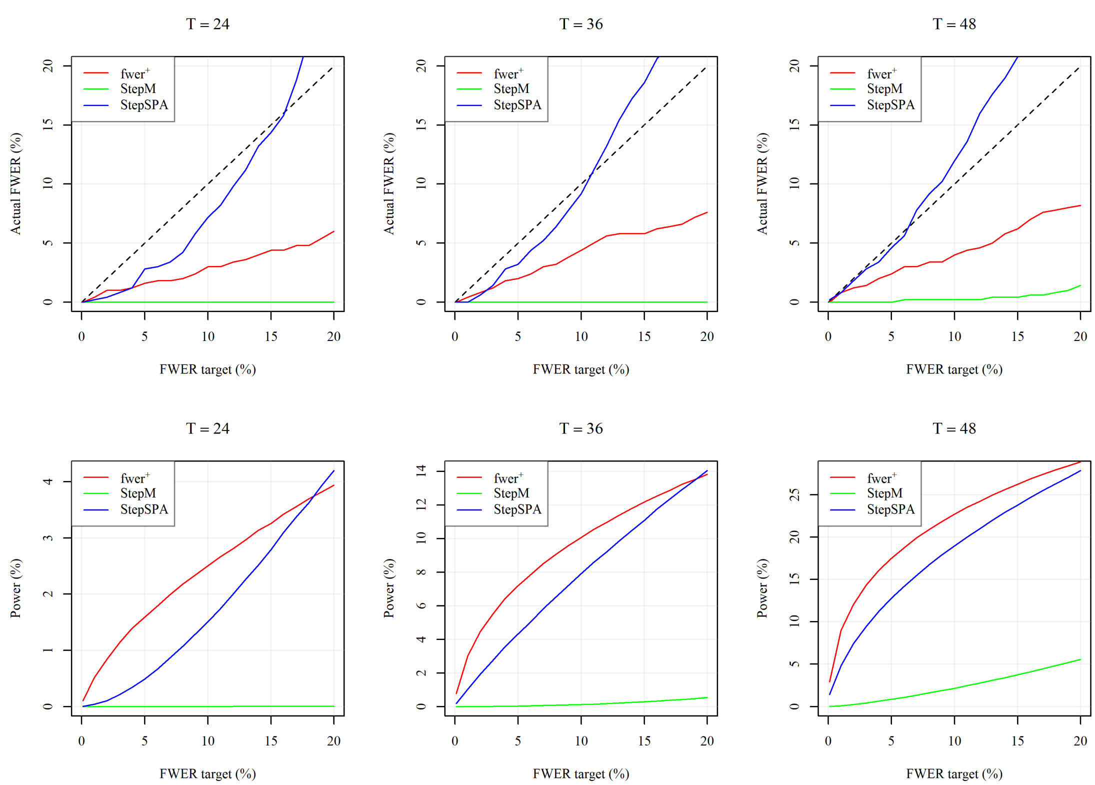
We repeat the steps 1 to 7 across 1000 iterations and calculate estimates of the actual FWER as the ratio of number of times we observe FWE = 1 over 1000, i.e. the frequency of error, and the power as average of the detected proportion recorded in the step 7 above.

3.5.1 A comparison to existing methods

In order to compare the performance of the $fwer^+$ to existing procedures, the StepM of and StepSPA, we opt a specific simulated data setting with $n = 1000$ funds and alpha magnitude of non-zero alpha fund $\alpha = 1$.

The performance of the procedures are presented in Figure 3.2 where all numbers are in percentage. In each of the top three sub-figures, we depict the estimated actual FWER given the targets where each curve represents a procedure. At a specific target,

Figure 3.2: Performance comparison. The figure compares the $fwer^+$ and existing procedures including StepM of Romano and Wolf (2005) and StepSPA of Hsu *et al.* (2010) in terms of FWER control (top three sub-figures) and power (bottom three sub-figures). The simulated data are balanced panels of 1000 funds with T observations per fund. From the left to the right, T takes values 24, 36 and 48. The input covariates U, V of the $fwer^+$ are independent.



a procedure controls well for FWER if the corresponding represented point on the curve at that target is lying below or on the dashed 45° line. We see that the $fwer^+$ and StepM procedures control well for considering FWER targets regardless the number of observations per fund whereas the StepSPA starts to lose its controlling of FWER when the target is higher than 15%, 10% and 5% in $T = 24, 36$ and 48 settings, respectively.

In terms of power, provided that the FWER is controlled well, the $fwer^+$ always performs better than the other two with gaps depending on the FWER target and number of observations per fund T. For instance, at target $\tau = 5\%$ and $T = 36$, the gaps in power of the $fwer^+$ compared to the StepM and StepSPA are 7.5% and 3.5%, respectively. Those numbers are larger (smaller) for $T = 48$ ($T = 24$) case which are about 15% and 5% (1.5% and 1%), respectively.

3.5.2 Performance of the $fwer^+$ under varying signals

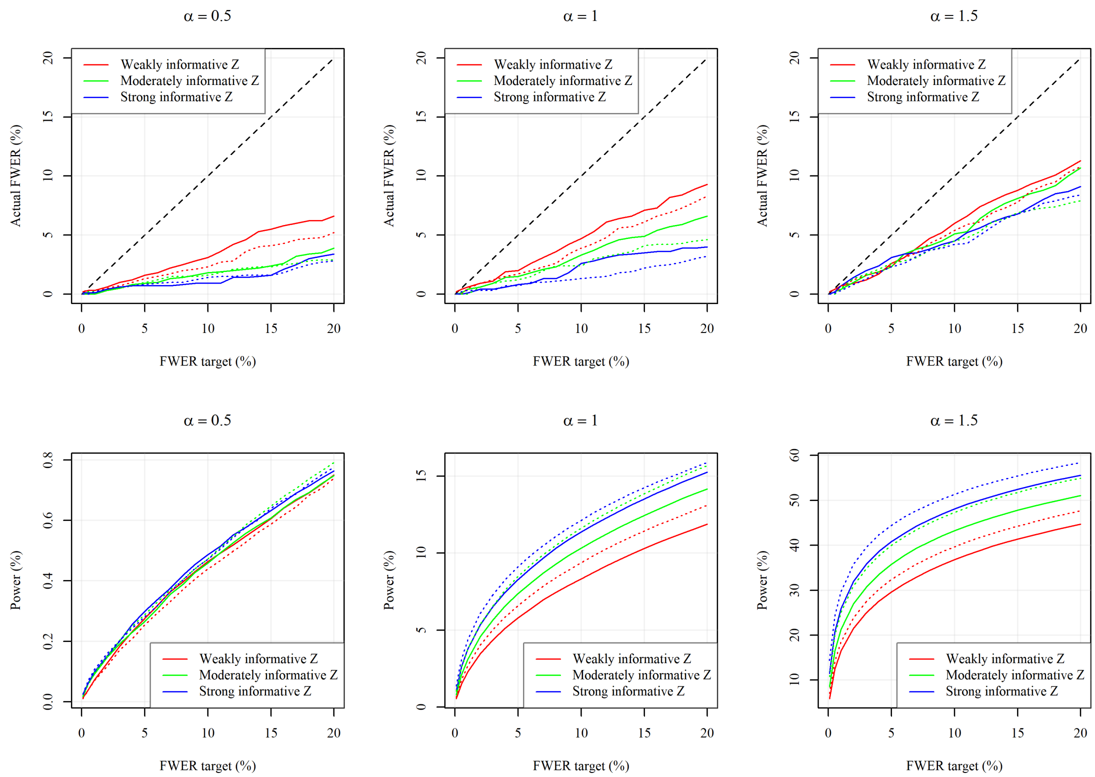
In applications, the parameters of the input data are varying. For instance, in our portfolio construction, which will be presented in Section 3.6, we need to assess funds' performance based on a short window of 36 months. The magnitude of alpha of non-zero alpha funds and the informativeness level of covariates are varying across different periods. To study impacts of these factors on performance of the $fwer^+$, in this section we vary the alpha of the non-zero alpha funds and the informativeness of the covariate.

As such, we fix number of funds $n = 1000$ and $T = 36$, which are close to the representative IS sample in our baseline empirical experiment, while α is varying from 0.5% to 1.5% and the relationship between the covariates and the prior null are weak, moderate and strong.

We report in Figure 3.3 the performance of the $fwer^+$ procedure under both independent and correlated covariates settings.[10] The top three sub-figures show the estimated actual FWER at the given targets whereas the bottom three sub-figures the power. From left to right, each sub-figure represents for a setting of non-zero alpha magnitude including 0.5%, 1% and 1.5%. In each of the top three sub-figures, the red- (green- and blue-) solid curves present for the estimated actual FWER of the $fwer^+$

[10] We additionally conduct simulations with $\pi^+ = 40\%, \pi^- = 0\%$ and present the results in Section Appendix C.2 and the results are roughy the same.

Figure 3.3: Performance of the $fwer^+$ under various setting of signals. The figure shows impact of signals, i.e., the magnitude of true non-zero alpha and informativeness of covariates, on the performance of the $fwer^+$ in terms of FWER control (top three sub-figuers) and power (bottom three figures). The simulated data are balanced panels of $n = 1000$ funds where each of them has $T = 36$ observations. The funds population consists of around 60%, 20% and 20% zero-alpha, under- and out-performing funds, respectively. The out-performing (under-performing) funds in population have alpha of α $(-\alpha)$ which varies in $\{0.5\%, 1.0\%, 1.5\%\}$. We consider three settings of the two covariates $\boldsymbol{Z} = (u, v)$ including weakly, moderately and strongly informative. The covariates can be independent (solid curves) or correlated with a correlation coefficient of 0.5 (dotted curves).



under a setting of the weakly (moderately and strongly) informative and independent covariates (i.e., $\rho = 0$). The dotted curve of the same color as the solid one is the estimated actual FWER of the corresponding informativeness level under the correlated covariates setting (i.e., $\rho = 0.5$). It is clear that the $fwer^+$ controls well for FWER at any given targets, regardless the dependence of covariates, as all points of the curves are below or on the 45° line.

The presentations in the bottom three sub-figures are similar but representing for the power. We observe that the stronger the informativeness of the covariates the higher power the $fwer^+$ gains. Moving from the left to the right sub-figures, the alpha magnitude of the out-performing funds is increasing and the $fwer^+$ gains higher power. This is unsurprising as the out-performing funds are being easier to be detected. When the two covariates are correlated, the power might be slightly lower (e.g., $\alpha = 0.5$ case), or higher (e.g., $\alpha = 1$ and 1.5 cases). This indicates that the dependence among covariates does not significantly effect the performance of the $fwer^+$.

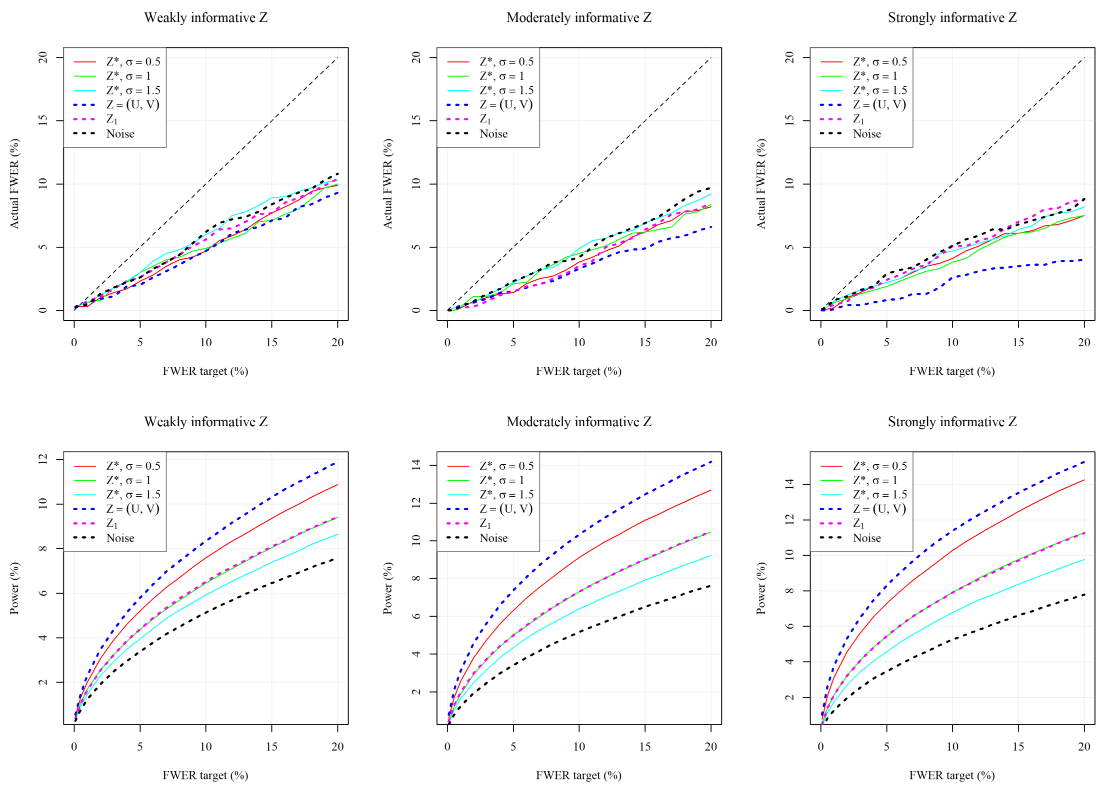
3.5.3 Performance of the $fwer^+$ under insufficient, noisy and uninformative covariates

We have investigated performance of the $fwer^+$ under the use of two covariates $\boldsymbol{Z} = (U, V)$. Often, in real applications we do not know how many covariates actually convey information on the performance of funds. Consequently, it happens the case we use less covariates than we should, or we might use the covariates that are estimated with noise or, even worse, unrelated to the funds' performance (i.e, uninformative covariates).

For the first scenario, we implement the $fwer^+$ with use of only one of the two covariates $Z_1 = U$. For the covariates estimated with noise case, we generate two new covariates $\boldsymbol{Z}^* = (U + \eta, V + \zeta)$ where η and ζ are noise drawn from normal distribution $N(0, \sigma^2)$. We investigate different levels of the noise via varying the $\sigma \in \{0.5, 1.0, 1.5\}$. Finally, for the uninformative covariate case, i.e., a covariate that is totally noise drawn from $N(0, 1)$ is used as a single input covariate. The performance of the $fwer^+$ for all mentioned scenarios are depicted in Figure 3.4. In this figure, we add the performance of the $fwer^+$ with use of the informative covariates $\boldsymbol{Z}$ for comparison purpose.

We see that in all scenarios, the FWER is controlled well at all considering targets. This is an excellent property of the $fwer^+$. The uninformative case implies that it is safe, in terms of controlling FWER, to implement the $fwer^+$ even if we wrongly include an unrelated covariate. Unsurprisingly, in terms of power, the $fwer^+$ performs best when we use the truly and sufficiently information while it is least powerful in case the covariate is irrelevant or uninformative. The power of the $fwer^+$ with use of the covariates estimated with noise lies in between the two extreme cases and decreases with respect to the level of the noise, i.e., the magnitude of σ. This also implies that, adding into a given informative covariates set an uninformative one might damage the power of the $fwer^+$.

Figure 3.4: Performance of the $fwer^+$ under use of insufficient information. The figure shows impact of using insufficient covariates, covariates containing different levels of noise and uninformative covariate on the performance of the $fwer^+$ in terms of FWER control (top three sub-figuers) and power (bottom three figures). The simulated data are balanced panels of $n = 1000$ funds where each of them has $T = 36$ observations. The funds population consists of around 60%, 20% and 20% zero-alpha, under- and out-performing funds, respectively. The out-performing (under-performing) funds in population have alpha of 1%. We consider three settings of the two covariates $\boldsymbol{Z} = (U, V)$ including weakly, moderately and strongly informative. The simulated data are generated based on $\boldsymbol{Z}$ via $\pi_0(\boldsymbol{Z})$. In noisy covariates cases, instead of using $\boldsymbol{Z}$, the $fwer^+$ uses $\boldsymbol{Z}^* = (U + \eta, V + \zeta)$ where $\eta, \zeta \sim N(0, \sigma^2)$ and $\sigma \in \{0.5, 1.0, 1.5\}$. In insufficient covariates case, the $fwer^+$ uses only $Z_1 = U$ while in the uninformative case it uses only one covariate which is a noise drawn from $N(0, 1)$ without any connection to $\pi_0(\boldsymbol{Z})$. We include performance of the $fwer^+$ with use of $\boldsymbol{Z}$ for comparison purpose.

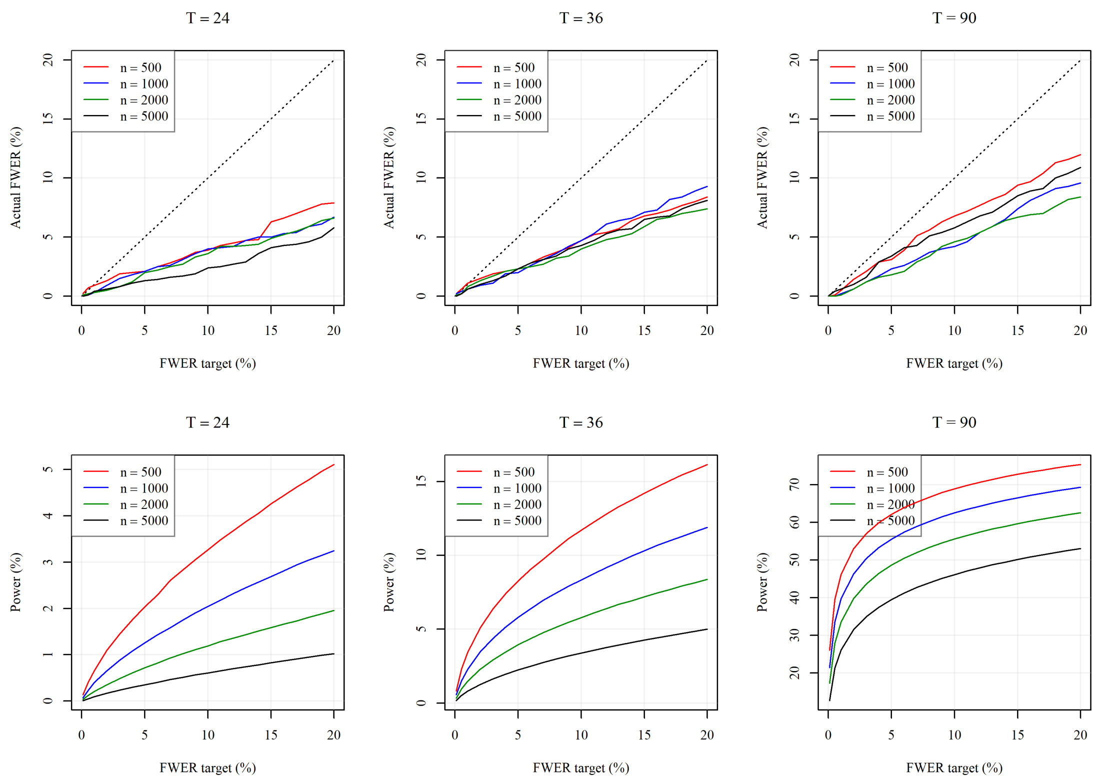


3.5.4 Performance of the $fwer^+$ under varying of sample size and observations

Hitherto, we have investigated the performance of the $fwer^+$ under different scenarios of the informative covariates as well as the strength of the signals of the out-performing funds. In this section, we further investigate the impact of sample size, i.e., number of funds in sample, and number of observations per fund on the performance of the procedure.

As such, we consider balanced panel data with varying the number of funds n and the number of observations per fund T. As the IS horizons for portfolio selection in hedge fund literature are typically 24 or 36 months we consider $T = 24$, and 36. We additionally experiment with much longer time series of $T = 90$, which is also the median number of observations per fund in our whole sample data. The n is also varying to cover all cases of our application in empirical experiments which spreads from around 500 to 2000. We also add a case $n = 5000$ which is close to our whole sample size. For the interest of space, in this set of simulations, we present results for data generated under the independent and weakly informative covariates and with $\alpha = 1$ setting.[11] The results are depicted in Figure 3.5.

Figure 3.5: Varying sample size and number of observations. The figure presents the performance of the $fwer^+$ under varying sample size (n) and number of observations per fund (T). The simulation data are balanced panels with T observations per fund under weakly informative and independent covariates.



In sub-figures of the Figure 3.5, the number of observations per fund is increasing from left to right. In each sub-figure, we present the results corresponding to different setting in number of funds, $n = 500, 1000, 2000$ and 5000. From the top three sub-figures, we again witness the excellent performance of the $fwer^+$ in terms of FWER control. It is clear from the bottom three sub-figures that, the power gains are higher

[11] Our conclusions are robust to other settings such as dependent and moderately and strongly informative covariates and the results are available upon request.

for the data with longer time series (i.e., larger T). This is consistent with the fact that the out-performing funds are easier to be detected if they outperform in a longer period. When the sample size is larger, the number of out-performing funds detected by $fwer^+$ is higher. For instance, in the case $T = 90$, at target 5%, the $fwer^+$ detects about 400 and 200 for the case sample size $n = 5000$ and $n = 2000$, respectively. Those numbers transform to roughly 40% and 50% in terms of power, respectively. These observations imply that the number of out-performing funds detected by the $fwer^+$ is not increasing proportionately to the number of truly out-performing funds in the population. Consequently, the $fwer^+$ is more powerful when the sample size is smaller. We note that this property is also found throughout developments of FWER frameworks, from the Bonferroni correction to other recent proposals (see simulations of Hansen, 2005).[12] This is a good property since the procedure can be applied in a wider problem both with small and large number of hypotheses.

In conclusion, the simulations show the excellent performance of the $fwer^+$ in terms of controlling for the FWER in various scenarios of data. We witness the higher power of the proposed procedure when we have one of the followings: i) the stronger the relationship between the covariates and the prior null; ii) the larger magnitude of out-performing funds' alpha; iii) the more sufficient set of informative covariates; iv) the out-performing funds do well in a longer period (larger T); and v) the smaller number of funds in the population. In Appendix C.2, we show that our conclusions are robust to alternative setting of the out-performing funds proportion.

3.6 Empirical analysis

In this section, we use the $fwer^+$ procedure to detect out-performing funds based on past short IS performance and invest in those detected funds in a rolling forward fashion. We describe the covariates that we are studying, the formation of our $fwer^+$-based portfolios and show their performance in various choice of IS horizons and models that we use to assessing the performance of funds.

[12] We recall that Bonferroni correction simply rejects all null hypotheses having $p-$value $\leq \tau/n$. When n is increasing, the threshold τ/n becomes smaller and the power declines rapidly.

3.6.1 Covariates

As hedge fund data reveal little information on funds' holdings, we focus on the covariates that are calculated based on excess return of the funds. Since we are assessing the performance of a fund via testing its adjusted return - the alpha, we include covariates that are potentially adding information alongside the alpha itself. There exists a number of such additional information that have been shown to be linked with the performance of hedge funds.

First, Titman and Tiu (2011) regress individual hedge fund returns on a group of risk factors and find that funds with low R-squares gain higher alpha. The authors further document that those low R-square funds charge higher incentive and management fees. Thus, the R-square of the funds is not only conveying the fund's managerial skill but also some other fund's characteristics. We use the R-square of the considering factor model as a covariate.

Second, as documented in Boyson (2008), funds' performance is more consistent among the younger and smaller funds. As investors' flows chase funds outperforming in the past, funds become larger and more passive. Thus, the size, i.e., the asset under management (AUM) of funds have a link with the funds' performance, and is chosen as one of our covariates.[13]

Third, Khandani and Lo (2011) argue that fund's excess return auto-correlation can measure the illiquidity in hedge funds and find a significant link between the auto-correlation of a fund and its expected return. Thus we consider as our covariates the first, second and third degrees of auto-correlation coefficients, which are denoted by ACF1, ACF2 and ACF3 respectively, of the fund's past 12-, 24- and 36-month excess return. These make up nine covariates and constitute our "persistent covariates" group.

Last, we study the risk measures based on fund's excess return as they are potentially informative. For example, Liang and Park (2007) document that downside risk measures incorporating higher moments help explain the cross-sectional variation of hedge fund performance and have predictive power. Wu *et al.* (2021) also find that the kurtosis of the excess return is an important variable in forecasting future hedge fund

[13] In our experiment, we follow literature to use the logarithm of fund's AUM instead of the AUM itself.

return. Thereby, we study risk metrics that consist of the variance (total volatility), kurtosis and skewness the fund's excess return over the past 12-, 24- and 36-month periods. These make up nine covariates and form our "moment covariates" group.

3.6.2 Portfolios of out-performing funds

In this section, we use the $fwer^+$ to construct portfolios of hedge funds based on assessing their short-term performance over a past period. We investigate OOS performance of the $fwer^+$ portfolios under various FWER target $\tau \in (0, 1)$.

We first describe our baseline portfolios. At the end of each year from 1997, we use most recent three years data up to that point of time as the IS period to calculate needed information. Specifically, we assess funds based on alpha of the 7-factor model and conduct for each fund the test of its alpha against zero, calculate $p-$value and estimate the mentioned covariates with use of only the data in IS period. A fund is eligible if it has returns data for all months of the IS period and the data of all considering covariates at the portfolio constructing time.[14] We implement the $fwer^+$ to picking out-performing funds with control for FWER at the given target τ. We then invest equally weighted in those selected funds in the following year. The performance of the portfolio in this OOS year is recorded. If there are no funds selected, we invest on bond to earn a return at the interest rate. When a selected fund stops reporting its returns during the OOS year, we redistribute fund equally into the remaining funds in the portfolio.[15] Our portfolios are rolling forward yearly. The first OOS period is the year 1998, which we invest in the funds selected based on the data in the IS period from January 1995 to December 1997. The final OOS period is the year 2021, which we invest in the funds selected based on using data from January 2018 to December 2020. Thus, each of our portfolios has OOS returns spanning over 24 years, from January 1998 to December 2021.

[14]We additionally conduct exercises where we restrict to consider only the funds that have at least 5 million USD in AUM and find that our empirical conclusions remain unchanged. For the interest of space, the results are presented in Appendix C.3.

[15]The IS horizon, which is used to estimate alpha (and covariates), could be 24 months as in Chen *et al.* (2017) and Kosowski *et al.* (2007) or 36 months as in Cumming *et al.* (2012). The OOS period is also varying in literature, Chen *et al.* (2017) use 3, 6, 9, 12, 24 and 36 months while Kosowski *et al.* (2007) use 12 months. Practically, hedge fund is a long-term investment vehicle and there is usually a lock-up period which is varying up to one year depending on funds. Thus, in this study we use the holding period of at least one year.

To further study the empirical performance of the $fwer^+$, we construct portfolios that control for FWER at different targets $\tau \in \{0.01\%, 1\%, 5\%\}$. The input $p-$value is calculated with use of heteroskedasticity and auto-correlation consistent (HAC) correction of Newey and West (1987).[16]

As benchmarks, we conduct two equally weighted portfolios as followings. At the end of each year from 1997, the first (second) equally weighted portfolio, denoted by EW (EW^+), simply selects all funds that are eligible (eligible and having positive estimated alpha) in the IS period to invest with equal weights in the following year. We repeat yearly until the end of 2020 to have a set of funds to invest in the year 2021.

As the numbers of covariates in the persistent and moment groups are large, we construct representative covariates that are the first principal component (PC1) of each group.[17] We thus have four covariates and construct our $fwer^+$ portfolios with use each of the four. Their OOS performance metrics are reported in panels A to D of Table 3.1. The metrics include annualized alpha as well as its HAC correction $t-$statistics and $p-$value, annualized excess return and Sharpe ratio. As a measurement of empirical power, we report the average, minimum, maximum and standard deviation of the number of out-performing funds detected by the $fwer^+$.

As our first observation, the $fwer^+$ detects non-empty group of out-performing funds in all 24 times of portfolio constructions even when controlling for FWER at 0.01%, which is very low target. This reflects the superior power of the $fwer^+$ procedure and thus it allows investors picking funds with high confidence, i.e., with very low error. More importantly, all portfolios with use one of the considering covariates gain positive abnormal alpha from around 4.2% to 5.3% in OOS period which are statistically significant with $t-$statistics varying from around 6 to 8. Of the four considering

[16]For the purpose of selecting out-peroming funds with low FWER targets, bootstrapped $p-$value has a limitation since the $p-$value is lower bounded at $1/(B+1)$ where B is number of bootstrapped iterations. Consequently, the highly out-performing funds with truly smaller $p-$value lose their advantage to be selected and empty portfolios are generated as a result. We therefore use $p-$value calculated from $t-$score with HAC standard error correction. In small sample size time series, the HAC correction might be biased as documented in Boudoukh *et al.* (2022) and Muller (2014). We further conduct experiments based on $p-$value calculated without using HAC correction and report the results in Appendix C.6. We see that the performances of portfolios are better and our main conclusions are even stronger.

[17]We report comprehensively the OOS performance of the $fwer^+$ portfolios with use of each of individual covariates Appendix C.4. We see that portfolios with use of individual covariates in the same group perform similarly. This suggests the use of PC1s as the representative covariates.

covariates, the PC1 of the persistent group performs best followed by others which are somewhat similar. It is clear that portfolios which control for a lower FWER target tend to perform better and can gain a Sharpe ratio of more than 2.

Next, we construct our $fwer^+$ portfolios with use of multiple covariates. As such, we use the R-square, AUM and the two PC1s of the persistent and moment groups as the four input covariates. As shown in Panel E of Table 3.1, those portfolios gain higher power than those with use of a sole underlying covariate. This is consistent with the fact shown in our simulation, that is, the more input informative covariates we use the higher power the $fwer^+$. The OOS performance of these portfolios is roughly at the average performance of the portfolios based on each of the four underlying covariates. The results suggest that using more covariates does not necessarily imply a higher alpha. This is not implausible because the $fwer^+$ is more powerful with more covariates and

Table 3.1: OOS performance of $fwer^+$ portfolios. Panels A to D of the table report OOS performance metrics of the $fwer^+$ portfolios with use each of R-square, AUM, and PC1s of moment and persistent group as the sole input covariate. The performance metrics include annualized alpha as well as its $t-$statistic and $p-$value, excess return and Sharpe ratio and summary on number of out-performing funds detected by the $fwer^+$. Panel E reports these metrics of the $fwer^+$ portfolio with use of all four mentioned covariates whereas panel F the performance metrics of the equally weighted (EW) and equally weighted plus (EW^+) portfolios.

τ (%)	Alpha (%)	$t-$statistic	$p-$value	Return (%)	Sharpe Ratio	Number of detected funds			
						Average	Min	Max	Std
Panel A: $fwer^+$ with use of R-square as the covariate									
0.01	4.98	7.5	0.00	5.42	2.17	16	4	25	6
1.00	4.44	7.2	0.00	4.94	1.59	33	10	55	13
5.00	4.45	6.8	0.00	4.95	1.58	47	14	88	20
Panel B: $fwer^+$ with AUM as the covariate									
0.01	4.87	7.2	0.00	5.33	2.09	16	4	25	7
1.00	4.33	6.9	0.00	4.82	1.55	35	8	64	15
5.00	4.25	6.6	0.00	4.80	1.48	49	14	95	23
Panel C: $fwer^+$ with use of PC1 of moment group as the covariate									
0.01	4.85	7.0	0.00	5.26	2.03	16	4	29	7
1.00	4.48	7.4	0.00	4.99	1.63	34	10	65	15
5.00	4.21	6.5	0.00	4.75	1.53	49	14	97	22
Panel D: $fwer^+$ with use of PC1 of persistent group as the covariate									
0.01	5.27	8.4	0.00	5.64	2.38	15	4	24	6
1.00	4.60	7.5	0.00	5.10	1.69	33	10	56	13
5.00	4.41	6.9	0.00	4.88	1.59	46	15	84	19
Panel E: $fwer^+$ with use of the R-square, AUM and PC1s of the two groups as the covariates									
0.01	5.12	8.2	0.00	5.47	2.32	17	4	31	7
1.00	4.47	7.2	0.00	4.97	1.62	37	10	69	16
5.00	4.10	6.1	0.00	4.66	1.44	52	14	104	25
Panel F: equally weighted portfolios									
EW	2.58	2.9	0.00	4.65	0.72	1067	350	1570	361
EW^+	3.00	3.7	0.00	4.77	0.80	761	273	1418	324

it might select some more smaller truly positive alpha funds.

As benchmarks, we report the performance of the equally weighted portfolios in panel F of the same table. We see that all the considering $fwer^+$ portfolios outperform the equally weighted ones. It is also noted that, the equally weighted portfolio does not select all funds but ones that pass the screening based on the number of observations, and the equally weighted plus one further requires funds having positive estimated alpha in the IS period. Thus it is not surprised that those portfolios also gain significantly positive alpha.

Overall, the $fwer^+$ portfolios perform well with all of the considering covariates. The $fwer^+$ shows its power in detecting outperforming funds even when we control for a very small error. The selected funds perform persistently in the OOS period and those selected with lower FWER targets tend to perform better on average.

3.6.3 Persistent analysis

As documented in Section 3.6.2, the performance of the funds selected by $fwer^+$ is persistent at least over the rolling OOS of one year. In this section, we provide further evidence on this advantage of the $fwer^+$ portfolios. Thereby, we examine the performance of those funds selected by the $fwer^+$ over longer OOS horizons. As such, we implement the $fwer^+$ every m years and we hold the detected funds over m years where $m = 2, 3$ and 4. For the interest of space, we report in Table 3.2 the performance of only the $fwer^+$ portfolios with use of R-square, AUM, and PC1s of moment and persistent group as the four input covariates. In long horizons, the attrition rate becomes important since the selected funds might not survive throughout the holding periods, leading to potential empty portfolios. We thus report the summary of monthly portfolio size rather than that of the number of funds selected by the $fwer^+$ as in previous discussions. We see that all considering portfolios are non-empty throughout the holding periods even with the holding horizon of four years.

In terms of alpha and Sharpe ratio, we see that the $fwer^+$ portfolios with 2- and 4-year holding horizons perform as well as those with one year holding whereas those of 3-year holding horizon are slightly worse. Given the use of only 3-year IS periods, the persistence in performance of the 4-year holding horizon portfolios is impressive.

Table 3.2: Performance of $fwer^+$ in various OOS horizons. The table reports performance metrics of the $fwer^+$ portfolios with use of R-square, AUM, and PC1s of moment and persistent group as the four input covariate with different OOS holding horizons. In OOS horizon of 2 (3 and 4) years, outperforming funds are selected by the $fwer^+$ every 2 (3 and 4) years and invested in the following 2 (3 and 4) years. The performance metrics include annualized alpha as well as its $t-$statistic and $p-$value, excess return, Sharpe ratio and summary on monthly portfolio size. Panels A, B, and C report these metrics for portfolios with holding horizons of 2, 3, and 4 years, respectively.

τ (%)	Alpha (%)	$t-$statistic	$p-$value	Return (%)	Sharpe Ratio	Portfolio Size			
						Average	Min	Max	Std
Panel A: 2-year OOS horizon									
0.01	5.36	8.0	0.00	5.71	2.13	17	3	31	7
1.00	4.60	6.9	0.00	5.22	1.66	36	9	69	16
5.00	4.55	6.8	0.00	5.32	1.52	52	17	104	24
Panel B: 3-year OOS horizon									
0.01	4.95	7.8	0.00	5.40	2.03	16	2	31	7
1.00	3.55	5.3	0.00	4.30	1.24	39	7	60	17
5.00	3.38	5.4	0.00	4.16	1.19	55	15	98	25
Panel C: 4-year OOS horizon									
0.01	4.99	8.2	0.00	5.27	2.07	17	1	31	9
1.00	4.31	7.2	0.00	4.85	1.63	33	3	69	19
5.00	4.27	6.9	0.00	4.96	1.50	45	9	104	28

Holding for a longer period also implies a less re-balanced cost. However, investors also face a risk of funds' attrition which might lead to a low diverse portfolio. As shown in summary of portfolio size columns, the minimum portfolio size is reducing with respect to the holding horizon. Nevertheless, in this particular case, the investors will not face any diversification problem if they set a target of FWER at 5%.

3.6.4 Sub-sample analysis

In this section, we further investigate the performance of the $fwer^+$ portfolios in sub-periods. We partition the whole OOS period, which spans from 1998 to 2021, into five non-overlapping sub-periods: 1998–2001, 2002–2006, 2007–2011, 2012–2016, and 2017–2021. Of those sub-periods, only the first one lasts for four years, others are five-year periods. We calculate the performance metrics of the portfolios in each sub-period and report them in each panel of Table 3.3.

The table shows that the $fwer^+$ portfolios gain positive alpha and Sharpe ratio in all sub-periods. Except the period 2007–2011, which covers the global financial crisis 2007–2008, the portfolios' alphas are statistically significant at all considering FWER targets. Compared to equally weighted portfolios, the $fwer^+$ portfolios gain higher alpha for four over five sub-periods. In only the first sub-period, the $fwer^+$ portfolios

Table 3.3: OOS performance of $fwer^+$ portfolios in sub-samples. Table report OOS performance metrics of the $fwer^+$ portfolios with use of R-square, AUM, and PC1s of moment and persistent group as the four input covariate in five non-overlapping sub-periods. For each sub-period we construct the equally weighted (EW) and equally weighted plus (EW^+) portfolios as benchmarks. The performance metrics include annualized alpha as well as its $t-$statistic and $p-$value, excess return, Sharpe ratio and summary on monthly portfolio size. We report in each panel the performance of the portfolios in the sub-period shown in its title.

τ (%)	Alpha (%)	$t-$statistic	$p-$value	Return (%)	Sharpe Ratio	Portfolio size			
						Average	Min	Max	Std
Panel A: Period 1998–2001									
0.01	4.96	3.9	0.00	3.75	0.97	22	18	30	4
1.00	5.01	4.1	0.00	3.81	1.00	50	40	60	8
5.00	5.10	4.3	0.00	3.95	0.96	67	55	84	11
EW	6.01	2.9	0.01	4.63	0.59	526	333	724	133
EW^+	5.54	3.3	0.00	4.24	0.61	401	258	571	111
Panel B: Period 2002–2006									
0.01	4.83	5.6	0.00	6.59	3.65	18	8	27	7
1.00	4.81	6.6	0.00	6.73	3.99	48	31	69	14
5.00	5.17	6.7	0.00	6.96	4.07	72	46	106	24
EW	3.90	3.2	0.00	7.25	1.49	1056	766	1364	188
EW^+	3.99	3.0	0.00	7.57	1.56	824	676	977	110
Panel C: Period 2007–2011									
0.01	2.86	1.9	0.06	2.70	1.10	10	2	22	7
1.00	2.63	1.7	0.10	1.87	0.44	26	7	61	18
5.00	2.44	1.6	0.12	1.60	0.35	41	10	87	27
EW	1.82	1.0	0.32	0.39	0.09	1393	1191	1544	78
EW^+	2.20	1.2	0.24	0.72	0.13	1125	927	1399	115
Panel D: Period 2012–2016									
0.01	7.07	6.3	0.00	8.18	4.06	14	7	20	4
1.00	6.58	7.7	0.00	7.99	4.86	35	27	41	4
5.00	6.81	7.9	0.00	8.40	4.95	51	35	61	9
EW	0.46	0.4	0.67	5.18	1.23	1135	924	1422	126
EW^+	1.84	1.7	0.09	5.49	1.56	774	616	1028	117
Panel E: Period 2017–2021									
0.01	4.69	6.0	0.00	5.17	3.30	18	7	27	6
1.00	3.29	3.3	0.00	4.12	1.33	24	10	36	6
5.00	3.08	2.9	0.01	3.88	1.35	28	12	42	8
EW	-0.52	-0.4	0.70	5.95	0.82	763	540	962	120
EW^+	0.83	0.7	0.48	5.87	1.04	410	267	561	96

gain lower alpha but their $t-$statistic are higher.[18] Also, the $fwer^+$ portfolios always have a higher Sharpe ratio than the equally weighted portfolios regardless the considered FWER targets and sub-periods.

The $fwer^+$ portfolios perform best during the period 2012–2016 with alphas roughly 7% and Sharpe ratios spanning from 4 to roughly 5. The most recent sub-period of our sample, the $fwer^+$ portfolios perform as well as the average of the whole sample reported in previous section and the Sharpe ratio can reach 3.3.

[18]It is worth to note that, the $fwer^+$ aims to select highly significant alpha funds, which are reflected via the significant of the tests, i.e., the $t-$statistics.

3.6.5 Boosting the informativeness of covariates

Hitherto, we have utilized only the informativeness of the covariates' variation. Yet, the covariates are likely containing noise and thus their informativeness is affected. In this section, we show that the performance of the $fwer^+$ portfolios can be improved via boosting the informativeness of the covariates. The main idea is to generating new covariates that target for future funds' expected returns. Thereby, we first use machine learning models to predict future return of funds and then we use the predicted returns as covariates. More specifically, we are considering four well-known machine learning models including the least absolute shrinkage and selection operator (LASSO, see Tibshirani 1996), random forest (RF, see Breiman 2001), stochastic gradient boosting (GB, see Friedman 2002) and deep neural network (DNN, see LeCun *et al.* 2015).

Formally, the relationship of the funds' cumulative future return during period $t+1,\ldots,t+h$ and the information of covariates at the end of month t can be modelled as

$$\tilde{R}_{i,t\to t+h} = f_t(X_{i,t}) + \tilde{\epsilon}_{i,t} \tag{3.6}$$

where $\tilde{R}_{i,t\to t+h}$ is cumulative return of fund i from month $t+1$ to $t+h$, $X_{i,t}$ is the realized covariates of the fund i measured at the end of month t, the function f_t describes the relationship of the $X_{i,t}$ and the future accumulated return over h months $\tilde{R}_{i,t\to t+h}$ whereas $\tilde{\epsilon}_{i,t}$ is the noise.

Consistent with the choice of our IS horizon and rolling window, to predict the return of year corresponding to period from month $t+1$ to $t+12$ for some t, we use data at the end of each previous three years until the end of month t to train the model (3.6) and use it to predict future return. That is, we train the model by using target variable $\tilde{R}_{i,k\to k+12}$ and features $X_{i,k}$ with $k = t-36, t-24$ and $t-12$ across funds considered in the IS period. We fit into training model the data of features $X_{i,t}$ at the end of month t to acquire the predicting accumulated future return for period $t+1$ to $t+12$.[19] As our AUM is available from December 1997, our first predicted returns is for the year 1999. This predicted return is calculated from data up to December 1998 and used as the input covariates of the $fwer^+$ to select funds invested in the year 1999. We rolling forward and re-balance the portfolios yearly in the same fashion as the $fwer^+$

[19] We follow Wu *et al.* (2021) in tuning the hyperparameters of the models.

portfolios described in previous section.

Table 3.4 reports the OOS performance of the $fwer^+$ portfolios with use of the predicted return of each considering machine learning model as a covariate. We see that the performances of the portfolios are generally better than the portfolios with four covariates presented in previous section. The portfolios' alpha range from 4.4% to 5.4% and Sharpe ratio from 1.31 to 2.79. On average across the considering FWER targets, the DNN model seems to be the best with an annualized alpha varying from 4.54% to 5.37% and an annualized Sharpe ratio that can reach 2.7. These numbers are generally higher than those of the $fwer^+$ portfolio with use of the four covariates reported in Table 3.1. This supports for benefit of using advanced machine learning techniques in forecasting hedge funds' return.

Table 3.4: OOS performance of $fwer^+$ portfolios with use of new covariates. Panel A (B, C and D) reports OOS annualized alpha as well as its $t-$statistic and $p-$value, excess return, Sharpe ratios and summary on number of out-performing funds selected by the $fwer^+$ with use of funds' future return predicted by LASSO (GB, RF and DNN) model at given FWER targets τ.

τ (%)	Alpha (%)	$t-$statistic	$p-$value	Return (%)	Sharpe Ratio	Number of detected funds			
						Average	Min	Max	Std
Panel A: $fwer^+$ with use of future return predicted by the LASSO model as the sole covariate									
0.01	5.42	9.3	0.00	5.83	2.79	15	4	25	7
1.00	4.50	7.6	0.00	5.04	1.73	33	10	59	14
5.00	4.40	7.0	0.00	4.96	1.69	46	14	86	20
Panel B: $fwer^+$ with use of future return predicted by the GB model as the sole covariate									
0.01	5.18	8.6	0.00	5.56	2.64	14	0	25	7
1.00	4.68	8.4	0.00	5.21	1.85	30	1	56	15
5.00	4.35	7.1	0.00	4.88	1.70	42	1	87	23
Panel C: $fwer^+$ with use of future return predicted by the RF model as the sole covariate									
0.01	5.42	9.4	0.00	5.87	2.52	15	4	25	7
1.00	4.87	7.0	0.00	5.66	1.38	30	5	56	14
5.00	4.52	6.2	0.00	5.28	1.31	43	7	89	21
Panel D: $fwer^+$ with use of future return predicted by the DNN model as the sole covariate									
0.01	5.37	9.3	0.00	5.78	2.72	15	4	24	6
1.00	4.84	9.2	0.00	5.39	1.96	32	10	55	14
5.00	4.54	7.4	0.00	5.12	1.76	45	14	87	20

3.6.6 Alternative choices of benchmarks

In this section, we show that the performance of the $fwer^+$ portfolios is robust to alternative benchmarks used in fund performance literature. More specifically, we are considering three alternative factor models including the four-factor model of Carhart (1997), the six-factor and nine-factor models described in Section 3.4. Those factor models are also considered in hedge fund performance (see, e.g., Bali *et al.*, 2012)

Table 3.5: Performance under alternative benchmarks. The table reports the OOS performance of the $fwer^+$ portfolios constructed by selecting truly positive alpha under alternative benchmarks. The $fwer^+$ uses all of the considering four covariates (the R-square, AUM, and two PC1s of the persistent and moment groups) as inputs. Panel A (B and C) presents annualized alpha of corresponding factor model as well as its $t-$statistic and $p-$value, excess return, Sharpe ratios and summary on the number of funds selected by the $fwer^+$ under the use of the four- (six- and nine-) factor model as the benchmark.

τ (%)	Alpha (%)	$t-$statistic	$p-$value	Return (%)	Sharpe Ratio	Number of selected funds			
						Average	Min	Max	Std
Panel A: four-factor model									
0.01	5.26	7.7	0.00	5.75	2.61	14	4	28	7
1.00	5.16	7.1	0.00	6.07	2.26	28	7	60	13
5.00	4.61	5.4	0.00	5.58	1.81	37	12	81	18
Panel B: six-factor model									
0.01	5.43	8.5	0.00	5.85	2.62	16	4	32	7
1.00	4.73	6.4	0.00	5.42	1.84	31	9	57	13
5.00	4.62	6.5	0.00	5.34	1.73	41	13	75	17
Panel C: nine-factor models									
0.01	5.26	7.9	0.00	5.63	2.23	14	3	27	7
1.00	4.88	8.2	0.00	5.42	2.01	30	8	61	13
5.00	4.15	6.4	0.00	4.83	1.54	42	12	79	18

though less common compared to our baseline, i.e., the seven-factor model.

For each of the alternative benchmarks, we repeat the exercises presented in previous sections. For the interest of space, we present in Table 3.5 the performance of the $fwer^+$ portfolios with use of the R-square of the considering model, AUM, and two PC1s of the moment and consistent groups. Overall, the OOS alphas of the $fwer^+$ portfolios are varying across the benchmark but all are statistically significantly positive. We see that, the $fwer^+$ portfolios under the four- and six-factor models gain highest annualized alpha which varie from 4.61% to 5.43%. However, as the alphas are of different factor models, it is not appropriate to compare the portfolios based on different models on this metric. Interestingly, comparing all considered models, including the seven-factor presented in Panel E of the Table 3.1, the highest Sharpe ratio is gained under the use of the four-factor model. Overall, all of our conclusions on the power of the $fwer^+$ as well as the ability in detecting truly out-performing hedge funds remain.

3.6.7 Portfolios of the best out-performing hedge fund

We have conducted portfolios of hedge funds with control for FWER at certain targets under consideration of various performance assessments. We have witnessed the ability of the $fwer^+$ in detecting out-performing funds based on utilizing short IS data windows. In this section, we further construct the portfolios consisting of only a single fund selected by the $fwer^+$. Given an FWER target τ, as the FWER of the group of funds detected by the $fwer^+$ is controlled at the target τ, it is held so for any subgroup of the detected funds. Instead of investing on all funds selected by the $fwer^+$, our new single-fund portfolio is established by investing only the fund selected by the $fwer^+$ that performs best in the IS period, i.e., one that has highest $t-$score among those selected by the $fwer^+$. As our $fwer^+$ portfolios are non-empty all the time regardless the considering FWER targets, and the portfolio with higher FWER target contains portfolios with lower targets, the choices of the considering FWER targets (0.1%,1% and 5%) will not effect on the best fund. We see that, the best funds are also unchanged under the choices of the considering covariates. In contrast, different choices of the factor model result on different best funds. We thus report in Table 3.6 the performance of the portfolio without showing the FWER target and covariates. We report results for all considering factor models. We see that all portfolios performs impressive, especially in terms of Sharpe ratio with the best reaching 5.3. In terms of alpha, the single-fund portfolios do not perform slightly worse those $fwer^+$ portfolios with use of all four covariates at FWER target 0.1% reported in tables 3.1 and 3.5. On the downside, as presented in the rightmost column, the portfolios are empty for 4 to 11 months over 288 months of the investing period.

Table 3.6: Performance of the single-fund portfolios. The table reports the OOS performance of the portfolio that consists of the fund performed best in IS period among those selected by the $fwer^+$.

Model	Alpha (%)	$t-$statistic	$p-$value	Return (%)	Sharpe Ratio	Empty rate (%)
4 factors	5.16	14.8	0.00	5.36	4.86	11/288
6 factors	5.08	15.2	0.00	5.27	5.33	7/288
7 factors	5.01	10.6	0.00	5.32	3.50	4/288
9 factors	5.16	15.0	0.00	5.37	5.31	7/288

We have assessed the performance of hedge funds based on past 36-month IS periods. As robustness checks of for this choice of the IS horizon, we additionally conduct experiments with use of 24- and 48-month IS periods. For the interest of space, the results are presented in Appendix C.5. Generally, with use of 24-month IS periods, the alphas of the $fwer^+$ portfolios are slightly higher than those reported for 36-month IS case though the Sharpe ratios are slightly lower. In contrast, the Sharpe ratios of the 48-month case are comparable to the baseline while the alphas are slightly lower.

Overall, all proposed portfolios beat passive benchmarks. From investors' perspective, the higher performing portfolios the better. For that purpose, using four factor model seems to be the best. We also see that controlling for a lower FWER target virtually results on a higher Shape ratio. At the extremely low FWER target, the portfolio that contains the fund performs best in IS among the detected out-performing funds performs best among the considered FWER targets. Our analyses on the persistent performance also point out that holding over a reasonably long horizon is not a bad idea. In fact, Tables 3.1 and 3.2 show that the portfolios constructed by assessing funds based on past three years and holding those selected out-performing ones over the following two years turn out to be the best. The choice of investors apparently depends on their risk averse level as the lower FWER target or the longer holding period tends to produce the less diverse portfolios. The performance of the portfolios might be improved if the covariates are more informative and thus the choice of covariates as well as the methods to boost their informativeness level are also of the investors. With the aim to point out the benefit of controlling FWER admixture with covariates, this study does not seek for the best covariates neither the best method to combine them.

3.7 Concluding remarks

We have introduced the $fwer^+$ to control FWER in picking out-performers. The procedure utilizes additional information in estimating the FWER. Via simulations we show that when informative covariates are available the method gains significant higher power than existing methods which are not using the covariates.

Empirical experiments in hedge funds context show that the method is so powerful that it can detect out-performing funds even with a very low target of FWER. The portfolios of the detected funds are able to generate statistically significantly positive alpha and the performance of those funds are persistent for a long period. This is robust to various choices of IS horizons and asset pricing models. All experiments suggest a powerful and promising tool for investors who desire to picking hedge funds with high confidence.

From practical point of view, investors need to consider further other aspects in order to implement the proposed portfolios in reality. First, they are constructed with the assumption that there are no restrictions in investing in the hedge funds selected by the $fwer^+$. This might not always be the case since the selected hedge funds might close to new investors. Second, the portfolios generate significant positive alphas with a high Sharpe ratio but relatively low annual excess return. Thus, the proposed yearly rolling portfolios will be appropriate to investment portion that favor low risk with stable adjusted return. To improve the excess return, investors can construct portfolios with higher frequency, for instance, by using a rolling forward quarterly instead of yearly, on a new set funds excluding those that require a lockup period of more than three months.

The new method has highly potential applications in problems where multiple testing is adopted, especially for the context that require a low level of error. In similar applications to our study, the method can be used in picking out-performing mutual fund, bond fund and trading strategies. It can be also used to guard the data snooping in predictive model and factor selection.

Conclusion

The thesis has developed methods to control for data-snooping bias in detecting out-performers, particularly, in assessing out-performing mutual funds, trading strategies and hedge funds. The methods incorporate multiple information in controlling two type I errors in multiple hypothesis testings, the FDR and the FWER, and contribute to solving the low power issue of existing methods. They have been shown to be working well under typical dependence in financial data and gain considerably higher power than existing approaches in literature. We have explored the benefits of using the methods in the mentioned financial topics.

In the first chapter, we introduce a framework, the $fFDR^{+}$, that incorporates a single informative covariate to estimates more precisely the FDR in detecting out-performing mutual funds. The method performs better than existing approaches in both simulations and empirical experiments. It provides a powerful tool for investors in assessing fund performance and for researchers in tackling with in conducting multiple testings. The method selects successfully out-performing mutual fund managers whose persistently outperform passive benchmarks in OOS period of 38 years. We witness that the portfolios constructed by the $fFDR^{+}$ generate positive and higher alphas than those established by an FDR method that does not account for additional information, or by sorting on the informative covariate and past performance of funds. The findings show the economics value of the covariates under studying even for recent decade, which is different from recent findings in mutual fund literature.

However, as we usually have more than one informative covariates available and in most cases, it is inefficient to combine them into only a single one. Thus, it is on demand to develop the $fFDR^{+}$ to a framework that further utilizes directly multiple information. Motivated by this, the second chapter develops the $fFDR$ method of Chen *et al.* (2021a) to a framework which is named multivariate functional false discovery rate ($mfFDR^{+}$). The new framework adapts more than one informative covariates and gains much higher power than the $fFDR$ and other existing approaches. This advantage is robust to various settings of dependence structures such as correlated covariates, weak dependence among tests and even when the covariates are estimated with noise. In order to use the $mfFDR$ to picking out-performers with control for FDR, a procedure named $mfFDR^{+}$ is introduced. The implementation of this procedure in detecting out-performing technical trading rules in foreign exchange rates shows the prevalence of profitable rules and the $mfFDR^{+}$ can form portfolios of rules that gain Sharpe ratio of about one during an OOS period of roughly 50 years. The empirical results highlight the importance of testing the performance of technical rules conditional on a set of comprehensively updated information.

In line with the developments of the two methods in chapters 1 and 2, the third chapter introduces a procedure, namely $fwer^{+}$, which incorporates multiple informative covariates in estimating the FWER in picking out-performers. The procedure gains significant higher power than existing methods which control for the same type I error but are not using the informative covariates. The implementation of the new procedure in picking out-performing hedge funds shows the benefits of using additional information in two aspects. First, it is powerful enough to detect non-empty group of out-performing funds though using short in-sample windows. Second, the portfolio of funds selected by the new procedure beats passive benchmark by generating statistically significant positive alpha, which transforms to Sharpe ratios of about 3.

The new methods have their limitations and there are still space for further improvements. First the $mfFDR$ as well as its derivative the $mfFDR^{+}$ and the special case $fFDR^{+}$ need a sufficient number of tests in order to control FDR perfectly. Thus they are most suitable for applications where the number of tests are large, usually more than 1000. The $fwer^{+}$ performs well with even less number of tests as it does not need to estimate multivariate density functions as the mentioned functional FDR methods. The present $fwer^{+}$ framework estimates the FWER via a null proportion, which is function of the covariates, and the density function of $p-$value under alternative. It is possible to improve the power of the $fwer^{+}$ via another statistical framework which estimates the density function of the $p-$value conditional on the covariates. This direction is beyond the scope of this thesis and is intended for future study.

The developments of the $fFDR^{+}$, $mfFDR$, $mfFDR^{+}$ and $fwer^{+}$ contribute to literature in multiple testing in Finance, Economics and other fields of Social Sciences. Although this thesis focuses on applications in picking out-performing funds and trading rules, the methods are highly promising in many other areas where researchers tackle with multiple testing problems such as selecting predictive models and detecting predictors which explain the cross-sectional stock returns.

References

Agarwal, V. and Naik, N. Y. (2000) Multi-period performance persistence analysis of hedge funds. *The Journal of Financial and Quantitative Analysis*, **35**, 327–342.

Allen, H. and Taylor, M. P. (1990) Charts, noise and fundamentals in the london foreign exchange market. *The Economic Journal*, **100**, 49–59.

Amihud, Y. and Goyenko, R. (2013) Mutual fund's R^2 as predictor of performance. *The Review of Financial Studies*, **26**, 667–694.

Andrikogiannopoulou, A. and Papakonstantinou, F. (2016) Estimating mutual fund skill: A new approach. Working paper.

Andrikogiannopoulou, A. and Papakonstantinou, F. (2019) Reassessing false discoveries in mutual fund performance: Skill, luck, or lack of power? *The Journal of Finance*, **74**, 2667–2688.

Astill, S., Taylor, A. R., Kellard, N. and Korkos, I. (2023) Using covariates to improve the efficacy of univariate bubble detection methods. *Journal of Empirical Finance*, **70**, 342–366.

Bajgrowicz, P. and Scaillet, O. (2012) Technical trading revisited: False discoveries, persistence tests, and transaction costs. *Journal of Financial Economics*, **106**, 473–491.

Bali, T. G., Brown, S. J. and Caglayan, M. O. (2012) Systematic risk and the cross section of hedge fund returns. *Journal of Financial Economics*, **106**, 114–131.

Baquero, G., ter Horst, J. and Verbeek, M. (2005) Survival, look-ahead bias, and persistence in hedge fund performance. *The Journal of Financial and Quantitative Analysis*, **40**, 493–517.

Barbaglia, L., Consoli, S. and Manzan, S. (2022) Forecasting with economic news. *Journal of Business & Economic Statistics*, **0**, 1–12.

Barras, L., Scaillet, O. and Wermers, R. (2010) False discoveries in mutual fund performance: Measuring luck in estimated alphas. *The Journal of Finance*, **65**, 179–216.

Barras, L., Scaillet, O. and Wermers, R. (2020) Reassessing false discoveries in mutual fund performance: Skill, luck, or lack of power? A reply. *The Journal of Finance*. URL https://doi.org/10.37214/jofweb.2. Replications and Corrigenda (web-only).

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**, 289–300.

Berk, J. B. and Green, R. C. (2004) Mutual fund flows and performance in rational markets. *Journal of Political Economy*, **112**, 1269–1295.

Boudoukh, J., Israel, R. and Richardson, M. (2022) Biases in long-horizon predictive regressions. *Journal of Financial Economics*, **145**, 937–969.

Boyson, N. M. (2008) Hedge fund performance persistence: A new approach. *Financial Analysts Journal*, **64**, 27–44.

Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 5–32.

Bris, A., Gulen, H., Kadiyala, P. and Rau, P. R. (2007) Good stewards, cheap talkers, or family men? The impact of mutual fund closures on fund managers, flows, fees, and performance. *The Review of Financial Studies*, **20**, 953–982.

Burnside, C., Eichenbaum, M. S. and Rebelo, S. (2011) Carry trade and momentum in currency markets. Tech. rep., National Bureau of Economic Research.

Capponi, A., Glasserman, P. and Weber, M. (2020) Swing pricing for mutual funds: Breaking the feedback loop between fire sales and fund redemptions. *Management Science*, **66**, 3581–3602.

Carhart, M. M. (1997) On persistence in mutual fund performance. *The Journal of Finance*, **52**, 57–82.

Chen, J., Hong, H., Huang, M. and Kubik, J. D. (2004) Does fund size erode mutual fund performance? the role of liquidity and organization. *The American Economic Review*, **94**, 1276–1302.

Chen, X., Robinson, D. G. and Storey, J. D. (2021a) The functional false discovery rate with applications to genomics. *Biostatistics*, **22**, 68–81.

Chen, Y., Cliff, M. and Zhao, H. (2017) Hedge funds: The good, the bad, and the lucky. *The Journal of Financial and Quantitative Analysis*, **52**, 1081–1109.

Chen, Y., Han, B. and Pan, J. (2021b) Sentiment trading and hedge fund returns. *Journal of Finance*, **76**, 2001–2033.

Chen, Y., Li, S. Z., Tang, Y. and Zhou, G. (2023) Anomalies as new hedge fund factors: A machine learning approach. *Working paper*.

Chinn, M. D. and Meese, R. A. (1995) Banking on currency forecasts: How predictable is change in money? *Journal of International Economics*, **38**, 161–178.

Clifford, C. P., Fulkerson, J. A., Jame, R. and Jordan, B. D. (2021) Salience and mutual fund investor demand for idiosyncratic volatility. *Management Science*, **67**, 5234–5254.

Cumming, D., Dai, N., Haß, L. H. and Schweizer, D. (2012) Regulatory induced performance persistence: Evidence from hedge funds. *Journal of Corporate Finance*, **18**, 1005–1022.

DeMiguel, V., Garlappi, L. and Uppal, R. (2007) Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The Review of Financial Studies*, **22**, 1915–1953.

Doshi, H., Elkamhi, R. and Simutin, M. (2015) Managerial activeness and mutual fund performance. *The Review of Asset Pricing Studies*, **5**, 156–184.

Fama, E. F. and French, K. R. (2010) Luck versus skill in the cross-section of mutual fund returns. *The Journal of Finance*, **65**, 1915–1947.

Fan, Y. and Fan, J. (2011) Testing and detecting jumps based on a discretely observed process. *Journal of Econometrics*, **164**, 331–344.

Filippou, I., Gozluklu, A. E. and Taylor, M. P. (2018) Global political risk and currency momentum. *Journal of Financial and Quantitative Analysis*, **53**, 2227–2259.

Frazzini, A. and Pedersen, L. H. (2014) Betting against beta. *Journal of Financial Economics*, **111**, 1–25.

Friedman, J. H. (2002) Stochastic gradient boosting. *Computational Statistics & Data Analysis*, **38**, 367–378.

Fung, W. and Hsieh, D. A. (2001) Empirical characteristics of dynamic trading strategies: the case of hedge funds. *Review of Financial Studies*, **10**, 275–302.

Fung, W. and Hsieh, D. A. (2004) Hedge fund benchmarks: A risk-based approach. *Financial Analysts Journal*, **60**, 65–80.

Geenens, G. (2014) Probit transformation for nonparametric kernel estimation on the unit interval. *Journal of the American Statistical Association*, **109**, 346–358.

Grønborg, N. S., Lunde, A., Timmermann, A. and Wermers, R. (2021) Picking funds with confidence. *Journal of Financial Economics*, **139**, 1–28.

Gu, S., Kelly, B. and Xiu, D. (2020) Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, **33**, 2223–2273.

Gu, S., Kelly, B. and Xiu, D. (2021) Autoencoder asset pricing models. *Journal of Econometrics*, **222**, 429–450.

Hansen, B. E. (1995) Rethinking the univariate approach to unit root testing: Using covariates to increase power. *Econometric Theory*, **11**, 1148–1171.

Hansen, P. R. (2005) A test for superior predictive ability. *Journal of Business & Economic Statistics*, **23**, 365–380.

Harvey, C. R. and Liu, Y. (2017) Decreasing returns to scale, fund flows, and performance. Working paper.

Harvey, C. R. and Liu, Y. (2018) Detecting repeatable performance. *The Review of Financial Studies*, **31**, 2499–2552.

Harvey, C. R. and Liu, Y. (2020) False (and missed) discoveries in financial economics. *The Journal of Finance*, **75**, 2503–2553.

Hsu, P.-H., Hsu, Y.-C. and Kuan, C.-M. (2010) Testing the predictive ability of technical analysis using a new stepwise test without data snooping bias. *Journal of Empirical Finance*, **17**, 471–484.

Hsu, P.-H., Taylor, M. P. and Wang, Z. (2016) Technical trading: Is it still beating the foreign exchange market? *Journal of International Economics*, **102**, 188–208.

Ignatiadis, N. and Huber, W. (2021) Covariate powered cross-weighted multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **83**, 720–751.

Ignatiadis, N., Klaus, B., Zaugg, J. B. and Huber, W. (2016) Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods*, **13**, 577–580.

Jiang, J., Kelly, B. and Xiu, D. (forthcoming) (re-)imag(in)ing price trends. *The Journal of Finance.*

Jones, C. S. and Mo, H. (2021) Out-of-sample performance of mutual fund predictors. *The Review of Financial Studies*, **34**, 149–193.

Jones, C. S. and Shanken, J. (2005) Mutual fund performance with learning across funds. *Journal of Financial Economics*, **78**, 507–552.

Kacperczyk, M., Sialm, C. and Zheng, L. (2008) Unobserved actions of mutual funds. *The Review of Financial Studies*, **21**, 2379–2416.

Kelly, B. T., Pruitt, S. and Su, Y. (2019) Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, **134**, 501–524.

Khandani, A. E. and Lo, A. W. (2011) Illiquidity premia in asset returns: An empirical analysis of hedge funds, mutual funds, and us equity portfolios. *The Quarterly Journal of Finance*, **01**, 205–264.

Kosowski, R., Naik, N. Y. and Teo, M. (2007) Do hedge funds deliver alpha? a bayesian and bootstrap analysis. *Journal of Financial Economics*, **84**, 229–264.

Kosowski, R., Timmermann, A., Wermers, R. and White, H. (2006) Can mutual fund "stars" really pick stocks? new evidence from a bootstrap analysis. *The Journal of Finance*, **61**, 2551–595.

Lan, W. and Du, L. (2019) A factor-adjusted multiple testing procedure with application to mutual fund selection. *Journal of Business & Economic Statistics*, **37**, 147–157.

Lan, W., Zhong, P.-S., Li, R., Wang, H. and Tsai, C.-L. (2016) Testing a single regression coefficient in high dimensional linear models. *Journal of Econometrics*, **195**, 154–168.

LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep learning. *Nature*, **521**, 436–444.

Levich, R. M. and Thomas, L. R. (1993) The significance of technical trading-rule profits in the foreign exchange market: a bootstrap approach. *Journal of International Money and Finance*, **12**, 451–474.

Liang, B. and Park, H. (2007) Risk measures for hedge funds: a cross-sectional approach. *European Financial Management*, **13**, 333–370.

Lo, A. W. (2004) The adaptive markets hypothesis. *The Journal of Portfolio Management*, **30**, 15–29.

Loader, C. (1999) *Local Regression and Likelihood*. Springer.

Lustig, H., Roussanov, N. and Verdelhan, A. (2011) Common Risk Factors in Currency Markets. *The Review of Financial Studies*, **24**, 3731–3777.

McLemore, P. (2019) Do mutual funds have decreasing returns to scale? evidence from fund mergers. *Journal of Financial and Quantitative Analysis*, **54**, 1683–1711.

Meese, R. A. and Rogoff, K. (1983) Empirical exchange rate models of the seventies: Do they fit out of sample? *Journal of International Economics*, **14**, 3–24.

Menkhoff, L. and Taylor, M. P. (2007) The obstinate passion of foreign exchange professionals: Technical analysis. *Journal of Economic Literature*, **45**, 936–972.

Muller, U. K. (2014) Hac corrections for strongly autocorrelated time series. *Journal of Business & Economic Statistics*, **32**, 311–322.

Neely, C., Weller, P. and Dittmar, R. (1997) Is technical analysis in the foreign exchange market profitable? a genetic programming approach. *The Journal of Financial and Quantitative Analysis*, **32**, 405–426.

Neely, C. J. (2002) The temporal pattern of trading rule returns and exchange rate intervention: intervention does not generate technical trading profits. *Journal of International Economics*, **58**, 211–232.

Neely, C. J. and Weller, P. A. (2013) Lessons from the evolution of foreign exchange trading strategies. *Journal of Banking and Finance*, **37**, 3783–3798.

Neely, C. J., Weller, P. A. and Ulrich, J. M. (2009) The adaptive markets hypothesis: Evidence from the foreign exchange market. *The Journal of Financial and Quantitative Analysis*, **44**, 467–488.

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

Newton, M. A., Noueiry, A., Sarkar, D. and Ahlquist, P. (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, **5**, 155–176.

Pastor, L., Stambaugh, R. F. and Taylor, L. A. (2015) Scale and skill in active management. *Journal of Financial Economics*, **116**, 23–45.

Politis, D. N. and Romano, J. P. (1994) *Journal of the American Statistical Association*, **89**, 1303–1313.

Qi, M. and Wu, Y. (2006) Technical trading-rule profitability, data snooping, and reality check: Evidence from the foreign exchange market. *Journal of Money, Credit and Banking*, **38**, 2135–2158.

Romano, J. P. and Wolf, M. (2005) Stepwise multiple testing as formalized data snooping. *Econometrica*, **73**, 1237–1282.

Sirri, E. R. and Tufano, P. (1998) Costly search and mutual fund flows. *The Journal of Finance*, **53**, 1589–1622.

Storey, J. D. (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **64**, 479–498.

Storey, J. D. (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics*, **31**, 2013–2035.

Storey, J. D., Akey, J. M. and Kruglyak, L. (2005) Multiple locus linkage analysis of genomewide expression in yeast. *PLOS Biology*, **3**, 1380–1390.

Storey, J. D. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, **100**, 9440–9445.

Sullivan, R., Timmermann, A. and White, H. (1999) Data-snooping, technical trading rule performance, and the bootstrap. *The Journal of Finance*, **54**, 1647–1691.

Sullivan, R., Timmermann, A. and White, H. (2001) Dangers of data mining: The case of calendar effects in stock returns. *Journal of Econometrics*, **105**, 249–286.

Sun, Z., Wang, A. W. and Zheng, L. (2018) Only winners in tough times repeat: Hedge fund performance persistence over different market conditions. *Journal of Financial and Quantitative Analysis*, **53**, 2199–2225.

Taylor, M. P. and Allen, H. (1992) The use of technical analysis in the foreign exchange market. *Journal of International Money and Finance*, **11**, 304–314.

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

Titman, S. and Tiu, C. (2011) Do the Best Hedge Funds Hedge? *The Review of Financial Studies*, **24**, 123–168.

Wang, X., Hyndman, R. J., Li, F. and Kang, Y. (2023) Forecast combinations: An over 50-year review. *International Journal of Forecasting*, **39**, 1518–1547.

Wermers, R. (1999) Mutual fund herding and the impact on stock prices. *The Journal of Finance*, **54**, 581–622.

Wermers, R. (2000) Mutual fund performance: An empirical decomposition into stock-picking talent, style, transactions costs, and expenses. *The Journal of Finance*, **55**, 1655–1695.

White, H. (2000) A reality check for data snooping. *Econometrica*, **68**, 1097–1126.

Wu, W., Chen, J., Yang, Z. B. and Tindall, M. L. (2021) A cross-sectional machine learning approach for hedge fund return prediction and selection. *Management Science*, **67**, 4577–4601.

Yan, X. S. (2008) Liquidity, investment style, and the relation between fund size and fund performance. *Journal of Financial and Quantitative Analysis*, **43**, 741–767.

Zhang, M. J., Xia, F. and Zou, J. (2019) Fast and covariate-adaptive method amplifies detection power in large-scale multiple hypothesis testing. *Nature Communications*, **10**, 1–11.

Zheng, L. (1999) Is money smart? A study of mutual fund investors' fund selection ability. *The Journal of Finance*, **54**, 901–933.

Zhou, H., Zhang, X. and Chen, J. (2021) Covariate adaptive familywise error rate control for genome-wide association studies. *Biometrika*, **108**, 915–931.

Zhu, M. (2018) Informative fund size, managerial skill, and investor rationality. *Journal of Financial Economics*, **130**, 114–134.

Zou, H., and Hastie, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **67**, 301–320.

Appendix A

Appendix for chapter 1

A.1 Estimating $\pi_0(z)$ and $f(p, z)$

Let $\{(p_i, z_i)\}_{i=1}^m$ be the collection of p-value and covariate realizations of the different funds under consideration, with $\{z_i\}_{i=1}^m$ transformed in uniform distribution $[0, 1]$ (see Section 1.2.1). We create fund bins $\{K_b\}_{b=1}^n$, where K_b contains a fund i if $z_i \in ((b-1)/n, b/n]$ and for each bin K_b we estimate a common $\pi_0(z)$ for all the funds i in the bin. For some common $\lambda \in (0, 1)$, we estimate the $\pi_0(z)$ in each bin b by

$$\hat{\pi}_{0,b}(\lambda) = \frac{\#\{p_i > \lambda, z_i \in K_b\}}{(1-\lambda)\#K_b}, \quad b = 1, 2, \ldots, n. \tag{A.1}$$

We determine λ by minimizing the mean integrated square error (MISE):

$$\text{MISE}(\lambda) = \text{bias}^2 + \text{variance} = \left(\int_0^1 \phi(z,\lambda)dz - \pi_0\right)^2 + \int_0^1 [\hat{\pi}_0(z,\lambda) - \phi(z,\lambda)]^2 dz \tag{A.2}$$

We estimate π_0 using the smoothing spline method of Storey and Tibshirani (2003, Remark B).[1] Similarly to CRS, we calculate $\hat{\pi}_0(z_i, \lambda) = \hat{\pi}_{0,b}(\lambda)$ for each grid value $\lambda \in \Lambda = \{0.4, 0.5, \ldots, 0.9\}$, $i = 1, \ldots, m$ and $b = 1, 2, \ldots, n$, the $\hat{\pi}_0(z_i, \lambda)$ and, subsequently, $\int_0^1 \hat{\pi}_0(z, \lambda)dz = \sum_{i=1}^m \hat{\pi}_0(z_i, \lambda)/m$. The unknown $\phi(z, \lambda)$ is estimated by $\hat{\phi}(\lambda, z) = \hat{\pi}_0(z, \Lambda_{\min}) - c_\lambda(1 - \hat{\pi}_0(z, \Lambda_{\min}))$, where c_λ is chosen such that $\int_0^1 \hat{\phi}(\lambda, z)dz = \int_0^1 \hat{\pi}_0(\lambda, z)dz$. We then obtain the optimal $\lambda^* = \arg\min_\lambda \text{MISE}(\lambda)$.

To estimate the joint density function $f(p, z)$, CRS use a local likelihood kernel density estimation (KDE) method with a probit transformation (Geenens, 2014).

[1] On rare occasions when the sample size m is small, the smoothing spline method does not work adequately. In these cases, we use the bootstrap method of Barras *et al.* (2010, Appendix A.1).

Specifically, let $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} dx$ and Φ^{-1} its inverse. Using $z_i' = \Phi^{-1}(z_i)$ and $p_i' = \Phi^{-1}(p_i)$, we obtain a "pseudo-sample" $\{(p_i', z_i')\}_{i=1}^n$, i.e., we transform the variables (p, z) to (p', z'); we denote by $\tilde{f}(p', z')$ the joint density function of (p', z'), which CRS estimate using the local likelihood KDE method. The bandwidth of the KDE is chosen locally via a k-Nearest-Neighbor approach using generalized cross-validation; this step can be implemented easily via the freely available R package `locfit`. The desired density function is then estimated as $\hat{f}(p, z) = \frac{\tilde{f}(p', z')}{\phi(p')\phi(z')}$ where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

Additionally, $f(p, z)$ may be non-increasing in p for each fixed z. CRS implement one more step which modifies the $\hat{f}(p, z)$ so that monotonicity is ensured. In our simulations, we use all the aforementioned techniques. In the empirical part, the monotonicity is switched off as this property is unknown in our data. For more details, readers are referred to CRS and their R package `fFDR`, Geenens (2014) as well as to the references therein.

A.2 Additional simulation results

To complement Section 1.5 of Chapter 1, we show here the performance of the $fFDR^+$ in terms of FDR control and power under several settings. We first show the results corresponding to the balanced panel data under cross-sectional dependence. Next, we present results for unbalanced panel data under both cross-sectional independence and dependence. Finally, to cover all distributions studied in the literature, we exhibit simulation results for the case where alphas are drawn from a single normal distribution.

A.2.1 Results for balanced panel data under cross-sectional dependence

We start by presenting in Figures A.1–A.3 the cases where the data are generated as balanced panels under cross-sectional dependent errors. The comparisons in terms of power between $fFDR^+$ and FDR^+ are shown in Tables A.1–A.5.

Figure A.1: Performance of $fFDR^{+}$ for discrete distribution of α. The graphs show the performance of the $fFDR^{+}$ in terms of FDR control when alphas are drawn from a discrete distribution. The simulated data are balanced panels with cross-sectional dependence.

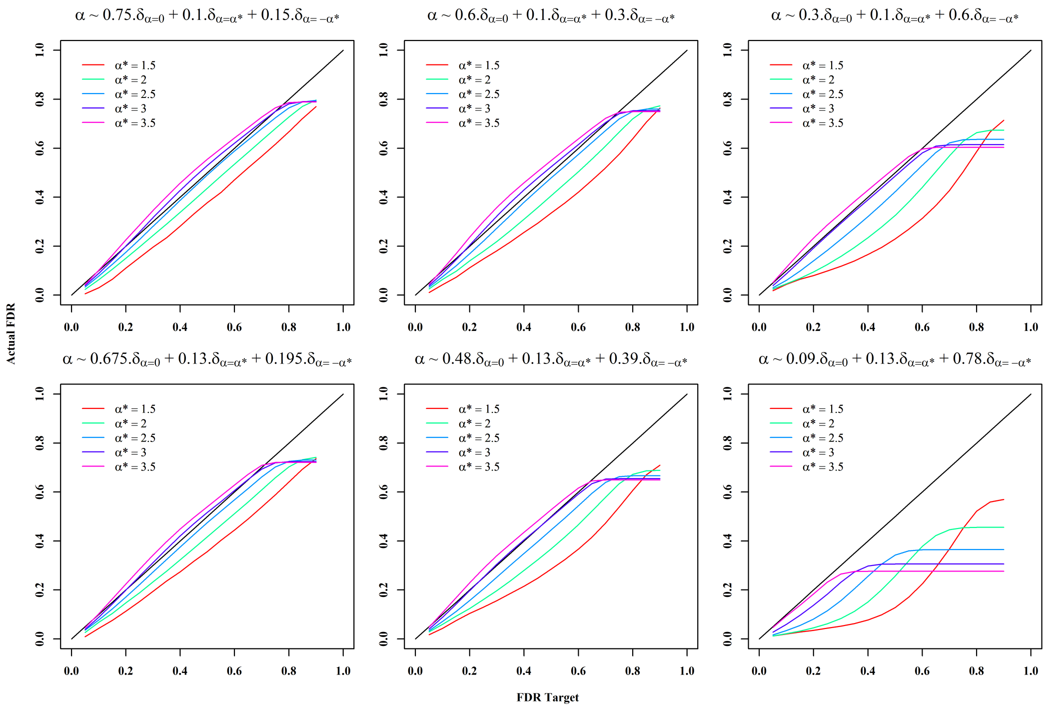


Figure A.2: Performance of $fFDR^{+}$ for continuous distribution of α. The graphs show the performance of the $fFDR^{+}$ in terms of FDR control when alphas are drawn from a continuous distribution which is a mixture of two normals. The simulated data are balanced panels with cross-sectional dependence.

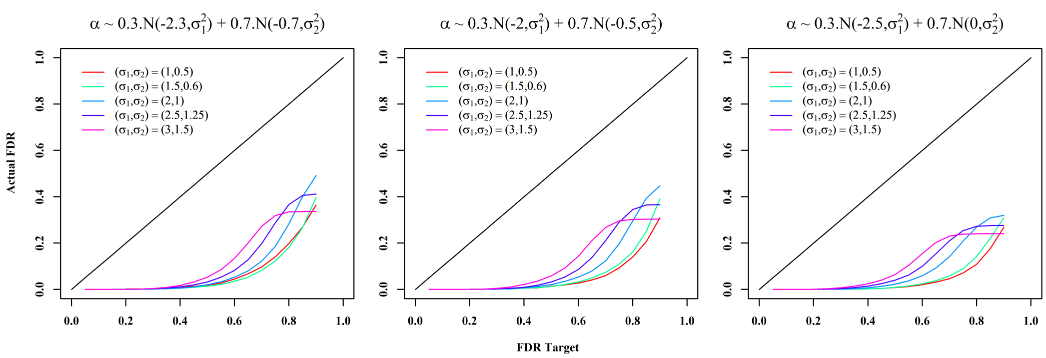


Figure A.3: Performance of $fFDR^+$ for discrete and normal distribution mixture of α. The graphs show the performance of the $fFDR^+$ in terms of FDR control when alphas are drawn from a mixture of discrete and normal distributions. The simulated data are balanced panels with cross-sectional dependence.

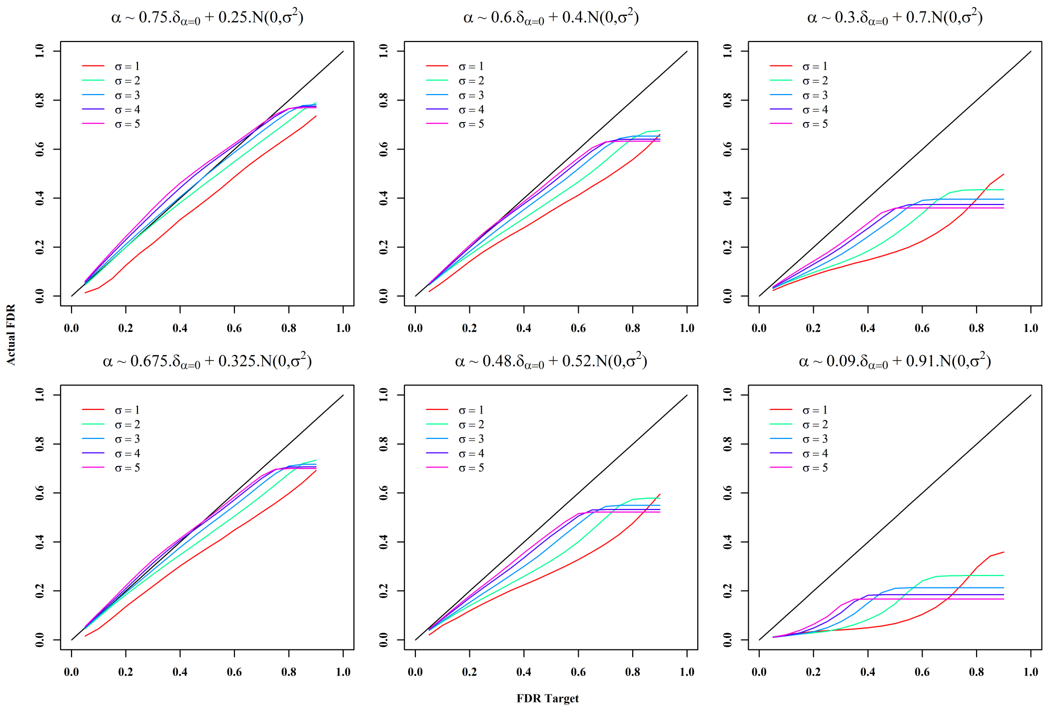


Table A.1: Power comparison (in %) for discrete distribution. The table compares the power of the $fFDR^+$ and FDR^+ at FDR target of 10% when the alphas of 2,000 funds are drawn from a discrete distribution: $\alpha \sim \pi^+\delta_{\alpha=\alpha^*} + \pi_0\delta_{\alpha=0} + \pi^-\delta_{\alpha=-\alpha^*}$ with varying α^* (annualized, in %) and proportions (π^+, π_0, π^-). The simulated data are a balanced panel with 274 observations per fund and generated with cross-sectional dependence.

(π^+, π_0, π^-)	Procedure	$\alpha^* = 1.5$	$\alpha^* = 2$	$\alpha^* = 2.5$	$\alpha^* = 3$	$\alpha^* = 3.5$
$(10, 75, 15)\%$	$fFDR^+$	0.8	6.1	21.3	43.6	65.5
	FDR^+	0.5	2.6	12.1	30.5	51.9
$(10, 60, 30)\%$	$fFDR^+$	1.9	11.2	32.3	56.6	76
	FDR^+	0.5	3	14.1	34.3	56
$(10, 30, 60)\%$	$fFDR^+$	4.6	23.1	51.5	75.4	89.1
	FDR^+	0.5	4	20.7	46.6	68.8
$(13, 67.5, 19.5)\%$	$fFDR^+$	1.5	9.7	29	53.1	73.7
	FDR^+	0.7	4.1	17	38	59.2
$(13, 48, 39)\%$	$fFDR^+$	3.4	17.1	41.3	66.3	83.3
	FDR^+	0.6	4.6	20.7	44.4	65.3
$(13, 9, 78)\%$	$fFDR^+$	8.5	34.2	67.9	89	97.2
	FDR^+	0.7	8.2	37.1	69.1	87.9

Table A.2: Power comparison (in %) for discrete-normal distribution mixture. The table compares the power of the $fFDR^+$ and FDR^+ at FDR target of 10% when alphas of 2,000 funds are drawn from a discrete-normal distribution mixture: $\alpha \sim \pi_0\delta_{\alpha=0} + (1-\pi_0)\mathcal{N}(0, \sigma^2)$ with varying σ (annualized, in %) and null proportion π_0. The simulated data are a balanced panel with 274 observations per fund and generated with cross-sectional dependence.

π_0	Procedure	$\sigma = 1$	$\sigma = 2$	$\sigma = 3$	$\sigma = 4$	$\sigma = 5$
75%	$fFDR^+$	0.5	15.7	36.2	51.2	60.8
	FDR^+	0.2	8.4	26.6	41.7	52.3
60%	$fFDR^+$	1.6	21.5	42.8	57.1	66.3
	FDR^+	0.3	11.4	31.3	46.5	56.8
30%	$fFDR^+$	4.7	32.4	54.5	67.6	75
	FDR^+	0.6	17.9	40.8	55.8	65.4
67.5%	$fFDR^+$	1	18.7	39.4	54	63.3
	FDR^+	0.2	9.8	29	44	54.5
48%	$fFDR^+$	2.5	25.5	47.3	61.3	70.2
	FDR^+	0.3	13.5	34.6	49.8	59.9
9%	$fFDR^+$	6.7	38	60.7	73.6	80.7
	FDR^+	0.7	22	46.9	62.5	72

Table A.3: Power comparison (in %) for mixture of two normal distributions. The table compares the power of the $fFDR^+$ and FDR^+ at FDR target of 10% when alphas of 2,000 funds are drawn from a mixture of two normal distributions: $\alpha \sim 0.3\mathcal{N}(\mu_1, \sigma_1^2) + 0.7\mathcal{N}(\mu_2, \sigma_2^2)$ with varying standard deviation pairs (σ_1, σ_2) and mean pairs (μ_1, μ_2) (both parameters' pairs are annualized and in %). The simulated data are a balanced panel with 274 observations per fund and generated with cross-sectional dependence.

		(σ_1, σ_2)				
(μ_1, μ_2)	Procedure	$(1, 0.5)$	$(1.5, 0.6)$	$(2, 1)$	$(2.5, 1.25)$	$(3, 1.5)$
		$\pi^+ = 6\%$	$\pi^+ = 10.4\%$	$\pi^+ = 20.7\%$	$\pi^+ = 25.5\%$	$\pi^+ = 29.1\%$
$(-2.3, -0.7)$	$fFDR^+$	0.1	0.4	5	13.6	23.3
	FDR^+	0	0	0.3	2.2	7.4
		$\pi^+ = 11.8\%$	$\pi^+ = 16.9\%$	$\pi^+ = 26.4\%$	$\pi^+ = 30.5\%$	$\pi^+ = 33.4\%$
$(-2, -0.5)$	$fFDR^+$	0.1	0.6	6.5	15.8	25.5
	FDR^+	0	0.1	0.5	3.2	9.1
		$\pi^+ = 35.2\%$	$\pi^+ = 36.4\%$	$\pi^+ = 38.2\%$	$\pi^+ = 39.8\%$	$\pi^+ = 41.1\%$
$(-2.5, 0)$	$fFDR^+$	0.4	1	9.2	18.6	28.3
	FDR^+	0	0.1	1	4.7	11.7

Table A.4: Power comparison (in %) for varying sample size and observation length. The table compares the power of the $fFDR^+$ and FDR^+ in a balanced panel data with varying number of observations per fund (T) and number of funds (m). We present three cases where alphas of m funds are drawn from i) discrete distribution: $\alpha \sim 0.1\delta_{\alpha=2} + 0.3\delta_{\alpha=0} + 0.6\delta_{\alpha=-2}$ (Panel A); ii) discrete-normal mixture: $\alpha \sim 0.3\delta_{\alpha=0} + 0.7\mathcal{N}(0, 2^2)$ (Panel B); and mixture of two normal distributions: $\alpha \sim 0.3\mathcal{N}(-2, 2^2) + 0.7\mathcal{N}(-0.5, 1)$ (Panel C). For each alpha population, we generate data with cross-sectional dependence.

		Number of observations per fund					
m	Procedure	$T = 120$	$T = 180$	$T = 240$	$T = 300$	$T = 360$	$T = 420$
		Panel A: Discrete distribution					
500	$fFDR^+$	3.7	9.4	19.9	31	43.5	54.5
	FDR^+	0.7	1.4	3.2	6.2	12	18.9
1000	$fFDR^+$	2.2	8.3	17.1	29.8	40.4	52.9
	FDR^+	0.4	1.1	2.6	5.9	11.3	19.9
2000	$fFDR^+$	2.1	7.3	16.5	26.8	40.6	50.6
	FDR^+	0.2	0.9	2.5	5.5	11.9	19.9
3000	$fFDR^+$	1.9	7	16	27.8	39.5	48.9
	FDR^+	0.2	0.7	2.2	5.9	12.3	19.6
		Panel B: Mixture of Discrete and Normal distributions					
500	$fFDR^+$	13	22	29.2	35.8	40.6	45.5
	FDR^+	3	8.1	13.8	20	25.3	29.9
1000	$fFDR^+$	12.5	21.2	29.1	35.1	39.8	44.2
	FDR^+	2.9	8.2	14.6	20.3	24.9	29.6
2000	$fFDR^+$	12.1	20.9	28.4	34.9	39.4	44.3
	FDR^+	2.7	8.2	14.4	20.4	25	29.8
3000	$fFDR^+$	11.8	20.8	28.3	34.4	39.9	43.7
	FDR^+	2.7	8.3	14.4	20.1	25.6	29.6
		Panel C: Mixture of Normal distributions					
500	$fFDR^+$	1.7	3.5	6.4	8.2	11.2	14.2
	FDR^+	0.2	0.3	0.6	0.9	1.4	2
1000	$fFDR^+$	1.2	3.2	5.6	8.6	10.8	13.3
	FDR^+	0.1	0.2	0.4	0.9	1.2	1.9
2000	$fFDR^+$	1.1	2.8	4.9	7.6	10.1	12.8
	FDR^+	0.1	0.2	0.3	0.7	1.1	2
3000	$fFDR^+$	1.1	2.8	5	7.6	10.3	12.6
	FDR^+	0.1	0	0.3	0.6	1.2	1.9

Table A.5: Power comparison (in %) for varying FDR targets (in %) for sample with small size and small number of observations under cross-sectional dependence. In this table, we consider three distributions as in Table A.4 for samples consisting of $m = 500$ funds (balanced panels) with $T = 60$ observations per fund (5 years).

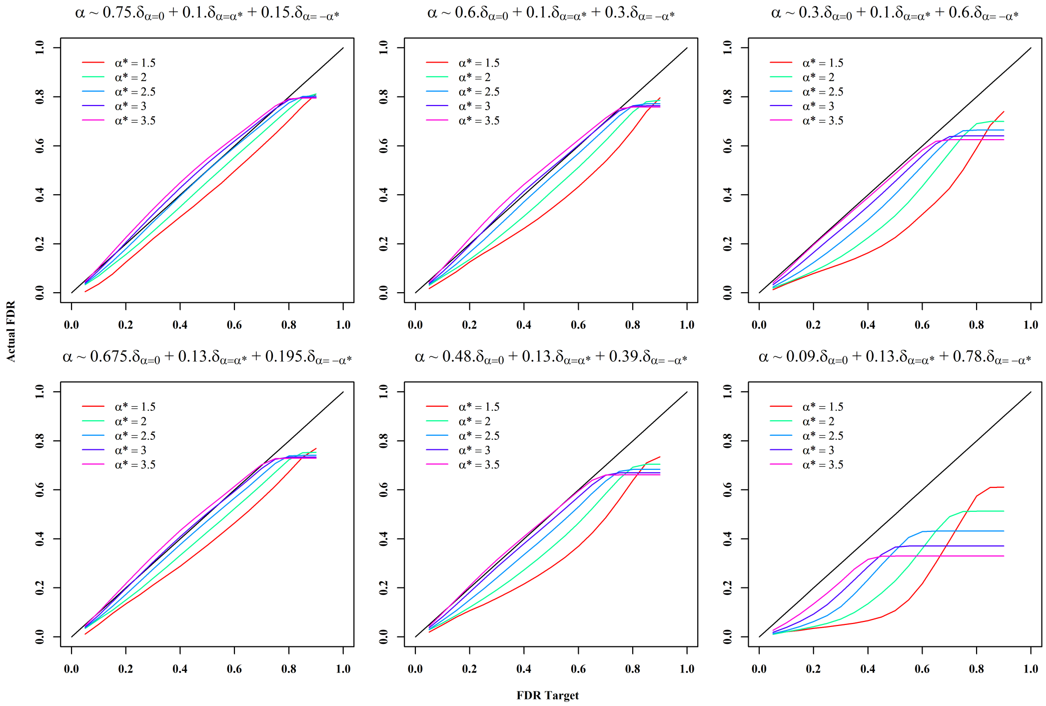
Distribution	Procedure	FDR target								
		10	20	30	40	50	60	70	80	90
Discrete	$fFDR^+$	0.8	3	7.3	13.6	21.9	31.4	41	51.3	63.3
	FDR^+	0.3	0.5	0.8	1	1.4	1.9	2.8	4	5.9
Mixture of discrete and normal	$fFDR^+$	3.1	8.5	15.4	23.5	32.3	41.4	50.8	60.9	67.2
	FDR^+	0.4	1.2	2.7	5.2	8.6	14.5	22.3	32.5	41.3
Mixture of normals	$fFDR^+$	0.4	1.8	4.3	8.1	13.4	29.8	27.7	37.6	50.7
	FDR^+	0.1	0.1	0.3	0.4	0.5	0.8	1.4	2.5	4.1

A.2.2 Results for unbalanced panel data

In this section, we present the performance of the $fFDR^+$ under both cross-sectional independence and dependence. Figures A.4–A.6 depict the FDR control of the $fFDR^+$, while the power comparisons are given in Tables A.6–A.8.

Figure A.4: Performance of $fFDR^{+}$ under unbalanced panel data. Figure shows the performance of $fFDR^{+}$ in terms of FDR control when alphas are drawn from the discrete distribution with unbalanced panel data.

Panel A: Cross-sectional Independent Data



Panel B: Cross-sectional Dependent Data

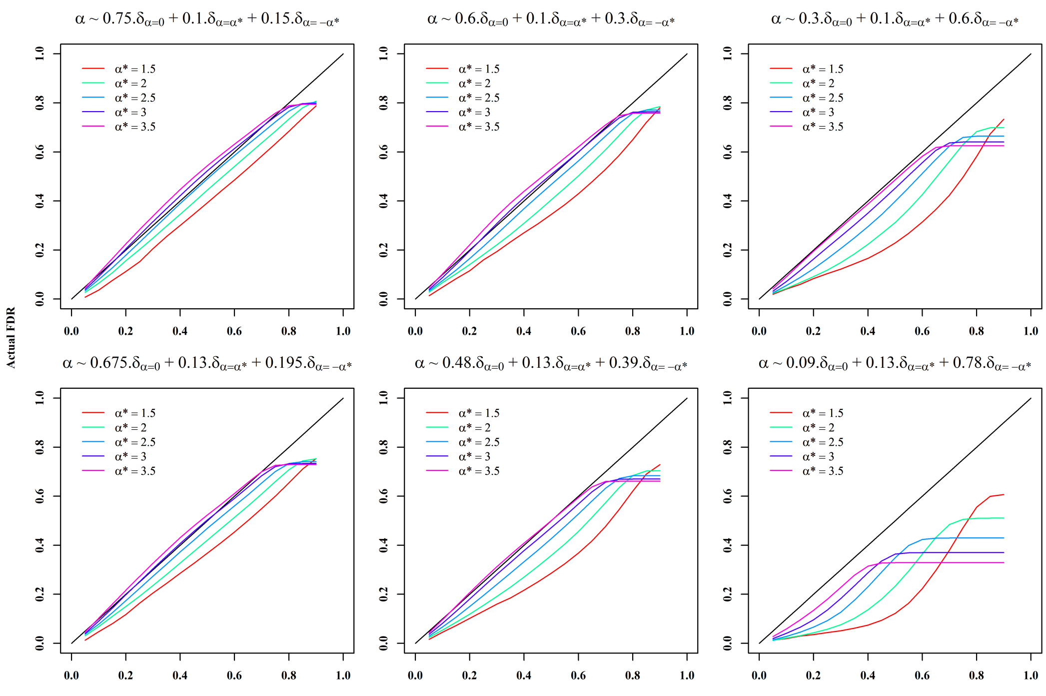
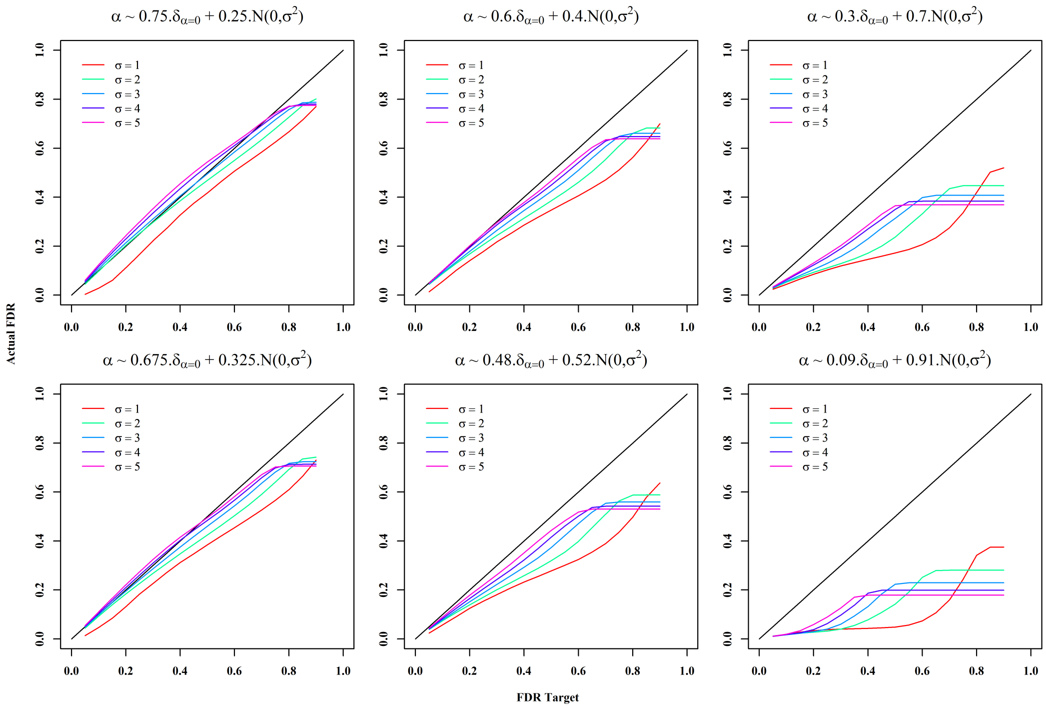


Figure A.5: Performance of $fFDR^{+}$ under discrete-normal distribution with unbalanced panel data. The figure shows the performance of $fFDR^{+}$ in terms of FDR control when alphas are drawn from the discrete-normal distribution mixture with unbalanced panel data.

Panel A: Cross-sectional Independent Data



Panel B: Cross-sectional Dependent Data

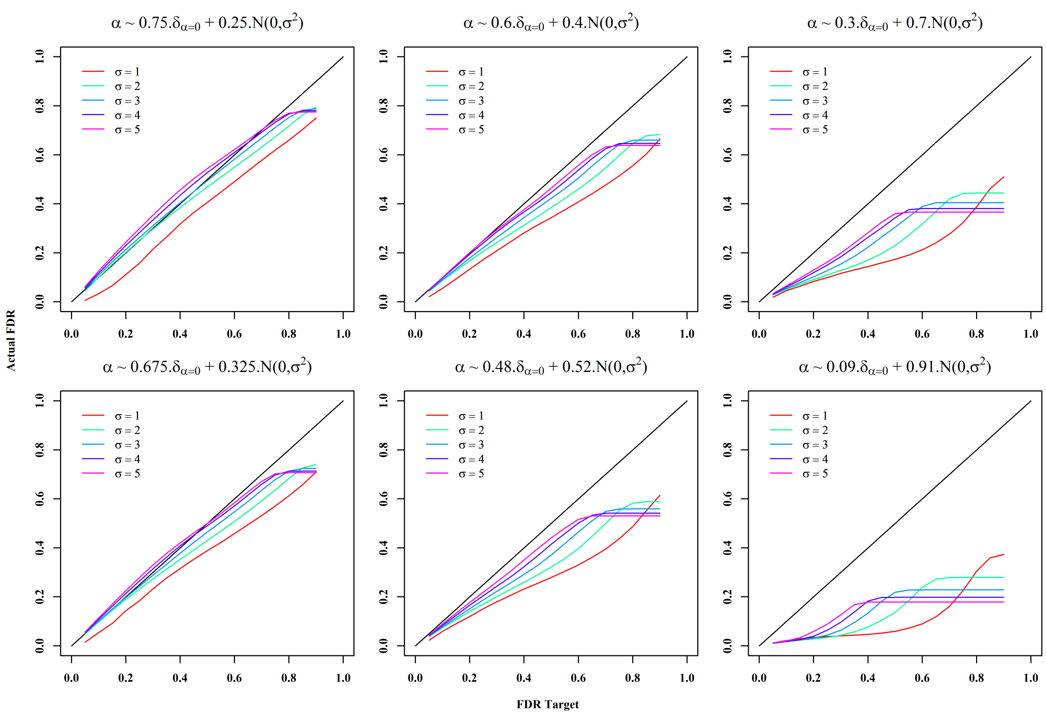
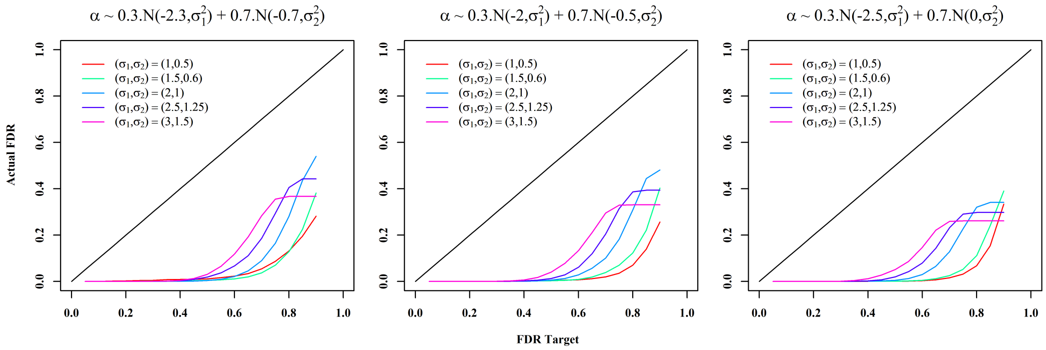


Figure A.6: Performance of $fFDR^{+}$ under mixture of two normals with unbalanced panel data. The figure shows the performance of $fFDR^{+}$ in terms of FDR control when alphas are drawn from the mixture of two normals with unbalanced panel data.

Panel A: Cross-sectional Independent Data



Panel B: Cross-sectional Dependent Data

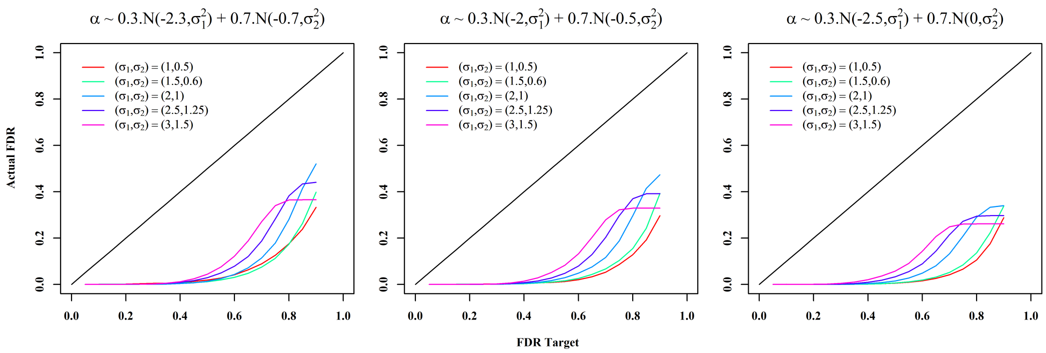


Table A.6: Power comparison (in %) for discrete distribution. The table compares the power of the $fFDR^+$ and FDR^+ at FDR target of 10% when the alphas of 2,000 funds are drawn from a discrete distribution: $\alpha \sim \pi^+\delta_{\alpha=\alpha^*} + \pi_0\delta_{\alpha=0} + \pi^-\delta_{\alpha=-\alpha^*}$ with varying α^* (annualized, in %) and proportions (π^+, π_0, π^-). The simulated data are an unbalanced panel with the number of observations of each fund drawn randomly with replacement from the real-data counterpart. We study the simulated data with both cross-sectional independence (left-hand side) and cross-sectional dependence (right-hand side).

		Cross-sectional Independence					Cross-sectional Dependence				
(π^+, π_0, π^-)	Procedure	$\alpha^* = 1.5$	$\alpha^* = 2$	$\alpha^* = 2.5$	$\alpha^* = 3$	$\alpha^* = 3.5$	$\alpha^* = 1.5$	$\alpha^* = 2$	$\alpha^* = 2.5$	$\alpha^* = 3$	$\alpha^* = 3.5$
$(10, 75, 15)\%$	$fFDR^+$	0.4	5.5	21.1	41.3	58.9	0.6	5.9	20.6	40.3	57.8
	FDR^+	0.4	2.6	12.7	29.9	46.7	0.4	2.9	13	29.3	45.9
$(10, 60, 30)\%$	$fFDR^+$	1.1	10.3	30.6	51.8	68	1.5	10.5	29.8	50.7	66.9
	FDR^+	0.5	2.9	14.6	32.6	49.9	0.5	3.2	14.3	31.9	49
$(10, 30, 60)\%$	$fFDR^+$	3.2	19.8	46.6	66.8	79.8	3.9	19.8	45.6	66	79.4
	FDR^+	0.5	3.6	19.1	40	58.1	0.5	4	18.9	39.5	57.6
$(13, 67.5, 19.5)\%$	$fFDR^+$	0.9	8.9	27.7	48.5	65.1	1.2	9.2	27.1	47.5	64.1
	FDR^+	0.5	3.9	17.4	35.6	52.3	0.6	4.2	17	34.9	51.5
$(13, 48, 39)\%$	$fFDR^+$	2.2	15.5	37.8	58.8	73.7	2.9	15.5	37.1	57.8	73
	FDR^+	0.5	4.5	20.3	39.8	56.9	0.7	4.8	19.5	39	56
$(13, 9, 78)\%$	$fFDR^+$	6.2	27.5	60.2	78.1	88.7	7.5	29.2	60	78.4	88.9
	FDR^+	0.6	6.8	29.5	54.2	72.5	0.8	7.7	30	54.7	72.8

Table A.7: Power comparison (in %) for discrete-normal distribution mixture. The table compares the power of the $fFDR^+$ and FDR^+ at FDR target of 10% when alphas of 2,000 funds are drawn from a discrete-normal distribution mixture: $\alpha \sim \pi_0 \delta_{\alpha=0} + (1-\pi_0)\mathcal{N}(0, \sigma^2)$ with varying σ (annualized, in %) and null proportion π_0. The simulated data are an unbalanced panel with the number of observations of each fund drawn randomly with replacement from the real-data counterpart. We study the simulated data with both cross-sectional independence (left-hand side) and cross-sectional dependence (right-hand side).

		Cross-sectional Independence					Cross-sectional Dependence				
π_0	Procedure	$\sigma=1$	$\sigma=2$	$\sigma=3$	$\sigma=4$	$\sigma=5$	$\sigma=1$	$\sigma=2$	$\sigma=3$	$\sigma=4$	$\sigma=5$
75%	$fFDR^+$	0.3	13.9	31.9	45.9	55.7	0.4	13.9	31.7	45.7	55.5
	FDR^+	0.2	7.8	23.5	37.1	47.5	0.3	7.9	23.3	36.9	47.2
60%	$fFDR^+$	1.3	19.2	38	51.8	60.9	1.3	19	37.8	51.7	60.9
	FDR^+	0.3	10.4	27.8	41.9	52.2	0.3	10.3	27.6	41.7	52
30%	$fFDR^+$	3.5	27.6	48.3	61.9	70.3	3.6	27.4	48	61.4	70.1
	FDR^+	0.4	15.4	35.7	50.5	60.6	0.5	15.2	35.4	50.2	60.3
67.5%	$fFDR^+$	0.8	16.8	35.2	48.9	58.4	0.9	16.9	35.1	49	58.5
	FDR^+	0.3	9.2	25.9	39.6	49.8	0.3	9.2	25.7	39.6	49.9
48%	$fFDR^+$	2.1	22.9	42.5	56.1	65.2	2.3	22.9	42.4	56	65.1
	FDR^+	0.3	12.4	31.2	45.6	55.7	0.4	12.5	31.1	45.4	55.4
9%	$fFDR^+$	5.3	33.3	54.9	68.2	76.7	5.6	33.5	55	68.2	76.7
	FDR^+	0.6	19.1	41.6	57	67.2	0.7	19.1	41.5	57	67.2

Table A.8: Power comparison (in %) for mixture of two normal distributions. The table compares the power of the $fFDR^+$ and FDR^+ at FDR target of 10% when alphas of 2,000 funds are drawn from a mixture of two normal distributions: $\alpha \sim 0.3\mathcal{N}(\mu_1, \sigma_1^2) + 0.7\mathcal{N}(\mu_2, \sigma_2^2)$ with varying standard deviation pairs (σ_1, σ_2) and mean pairs (μ_1, μ_2) (both parameters' pairs are annualized and in %). The simulated data are an unbalanced panel with the number of observations of each fund drawn randomly with replacement from the real-data counterpart. We study the simulated data with both cross-sectional independence (left-hand side) and cross-sectional dependence (right-hand side).

		Cross-sectional Independence					Cross-sectional Dependence				
(μ_1, μ_2)	Procedure	σ^1	σ^2	σ^3	σ^4	σ^5	σ^1	σ^2	σ^3	σ^4	σ^5
$(-2.3, -0.7)$	$fFDR^+$	0	0.2	3.8	11.3	19.5	0	0.3	4.2	11.7	20.1
	FDR^+	0	0	0.3	1.8	6.3	0	0	0.3	2	6.4
$(-2, -0.5)$	$fFDR^+$	0	0.4	5	13	21.7	0.1	0.5	5.5	13.7	22.2
	FDR^+	0	0.1	0.4	2.7	7.8	0	0.1	0.5	2.9	8.1
$(-2.5, 0)$	$fFDR^+$	0.1	0.5	7.3	15.4	24	0.3	0.8	7.8	16	24.6
	FDR^+	0	0.1	0.6	3.9	9.9	0	0.1	0.9	4.2	10.3

where $\sigma^1 = (1, 0.5), \sigma^2 = (1.5, 0.6), \sigma^3 = (2, 1), \sigma^4 = (2.5, 1.25), \sigma^5 = (3, 1.5)$.

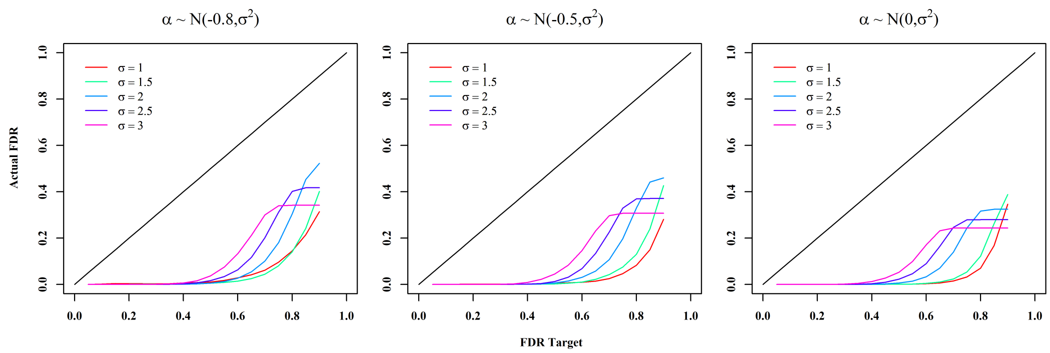
A.2.3 Simulation results for single normal distribution

In this section, we present the simulation results for a special case of continuous distribution where the mixture (1.17) has only one component. Specifically, we consider the case $\pi_2 = 0$, $\alpha \sim \mathcal{N}(\mu, \sigma^2)$ and, based on Jones and Shanken (2005) and Fama and French (2010), we use $\mu \in \{-0.8, -0.5, 0\}$ and $\sigma \in \{1, 1.5, 2, 2.5, 3\}$ (the presented values of both parameters are annualized and in %).[2]

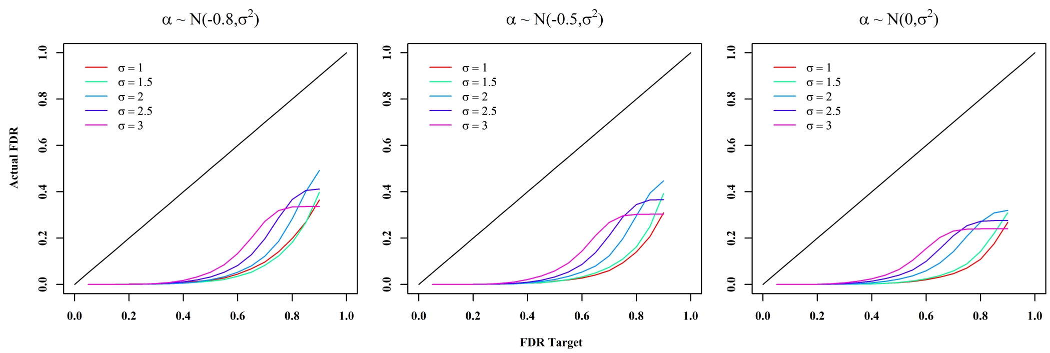
Figures A.7 and A.8 present the performance of the $fFDR^{+}$ procedure when the alphas are drawn from balanced and unbalanced panel data, respectively. It is shown that the FDR is controlled at any given target.

Figure A.7: Performance of the $fFDR^{+}$ under single normal distribution with balanced panel data. The figure shows the performance of the $fFDR^{+}$ in terms of FDR control when alphas are drawn from the single normal distribution with balanced panel data.

Panel A: Cross-sectional Independent Data



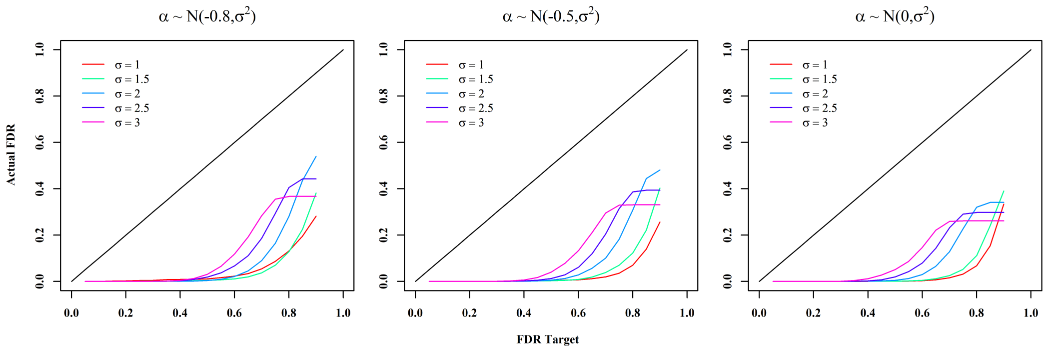
Panel B: Cross-sectional Dependent Data



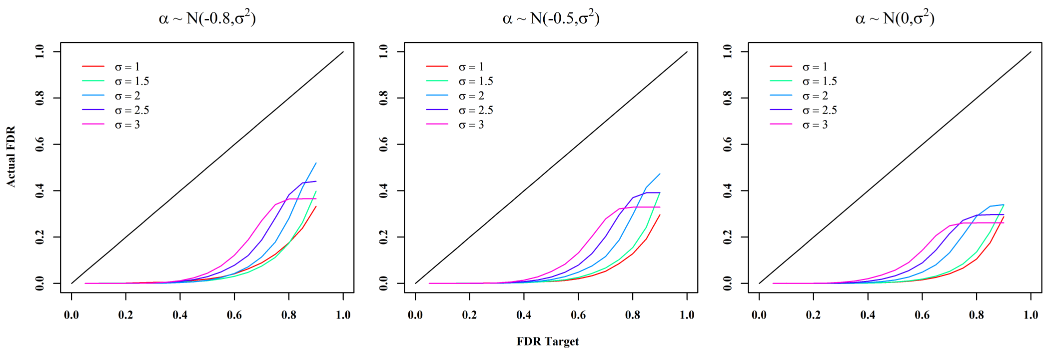
[2] Jones and Shanken (2005) assume that the fund alphas are drawn from a normal distribution and their estimates for the mean and standard deviation are based on prior beliefs. They find that the mean is 1.3%-1.4% per annum before expenses (about 2%) and the standard deviation is 1.5%-1.8%. In addition, Fama and French (2010) assume that the fund (gross) alpha population has a normal distribution centered at 0.

Figure A.8: Performance of the $fFDR^+$ under single normal distributions with unbalanced panel data. The figure shows the performance of the $fFDR^+$ in terms of FDR control when alphas are drawn from the single normal distribution with unbalanced panel data.

Panel A: Cross-sectional Independent Data



Panel B: Cross-sectional Dependent Data



In Table A.9, we focus on comparing the performance of $fFDR^+$ and FDR^+ in terms of power. As π^+ depends on both the mean μ and variance σ^2 of the distribution, we need to distinguish the value of π^+ from the pairs (μ, σ). We provide in Panel A additional information about π^+, which helps us assess the impact of the magnitude of positive alphas on the power. For instance, for $\pi^+ \approx 40\%$, the power of the two procedures for $(\mu, \sigma) = (-0.8, 3)$ is significantly higher than for $(\mu, \sigma) = (-0.5, 2)$. We observe a boost in power for both methods with increasing σ (for given non-positive μ), resulting in larger proportion and magnitude of positive alphas. In all the cases under consideration, the $fFDR^+$ dominates FDR^+ in terms of power and this gap soon becomes omnipresent for $\sigma \geq 1.5$ reaching up to 18%.

Table A.9: Power comparison (in %) for single normal distribution. The table compares the power of the $fFDR^+$ and FDR^+ at FDR target of 10% when alphas of 2,000 funds are drawn from a normal distribution: $\alpha \sim \mathcal{N}(\mu, \sigma^2)$ with varying standard deviation σ and mean μ (both parameters are annualized and in %). In Panel A the simulated data are a balanced panel with 274 observations per fund, whereas in Panel B an unbalanced panel with the number of observations of each fund drawn randomly with replacement from the real-data counterpart. For each type of panel data, we generate data cross-sectional independence (left-hand side) and with cross-sectional dependence (right-hand side).

		Cross-sectional Independence					Cross-sectional Dependence				
		σ					σ				
μ	Procedure	1	1.5	2	2.5	3	1	1.5	2	2.5	3
		Panel A: Balanced Data									
	π^+	21.2	29.7	34.5	37.4	39.5	21.2	29.7	34.5	37.4	39.5
-0.8	$fFDR^+$	1.6	14.1	30.5	44.4	55.1	2.1	14.9	30.9	44.7	55.4
	FDR^+	0.1	1.7	12.6	27.2	40.6	0.1	2.1	12.8	27.4	40.7
	π^+	30.9	36.9	40.1	42.1	43.4	30.9	36.9	40.1	42.1	43.4
-0.5	$fFDR^+$	3	17.6	33.8	47.3	57.7	3.8	18.3	34.5	47.8	58.1
	FDR^+	0.1	3.6	16.5	31.3	44.1	0.2	4	16.7	31.5	44.3
	π^+	50	50	50	50	50	50	50	50	50	50
0	$fFDR^+$	7.9	24.8	40.7	52.8	62.4	8.9	25.7	41.3	53.3	62.7
	FDR^+	0.6	9.1	24.2	38.7	50.3	1	9.5	24.6	38.9	50.5
		Panel B: Unbalanced Data									
-0.8	$fFDR^+$	1.4	12.1	26.5	39.5	50.1	1.7	12.7	27.1	39.8	50.2
	FDR^+	0.1	1.7	10.8	23.2	35.2	0.1	2	11.2	23.5	35.4
-0.5	$fFDR^+$	2.6	15.2	29.8	42.5	52.6	3.1	15.8	30.2	42.8	52.7
	FDR^+	0.2	3.4	14.1	26.8	38.6	0.2	3.7	14.5	27.2	38.8
0	$fFDR^+$	6.8	21.6	36	47.8	56.9	7.4	22.4	36.4	48	57.1
	FDR^+	0.6	8.1	20.8	33.6	44.5	0.9	8.5	21.2	33.9	44.6

A.3 Results for data sample period from 1984

Given potential biases in the mutual fund data for the period before 1984, we construct portfolios using a data sample from 1984 as a robustness check. We start by using the first five years' data, spanning from January 1984 to December 1988, to calculate the inputs of the procedures. The detected out-performing funds are equally invested in 1989. Then, the five years of data from January 1985 to December 1989 are used for the recalculation of the inputs of the procedures to detect out-performing funds invested in 1990, and so on. The process is yearly rolled over until the end of the sample. Thus, the OOS returns of the portfolios start from January 1989 to December 2019. At the end of each month from December 1998, i.e. when the portfolios' return series reach a length of at least ten years, we calculate the portfolios' alpha based on the returns from January 1989 to that month and present that in Figure A.9. We also report the average n-year alpha with the length of investing n from 5 to 31 years in Table A.10.

Figure A.9: Alpha evolution of $fFDR10\%$ and $FDR10\%$ portfolios over time with use of data from 1984. The graph presents the evolution of annualized alphas (in %) of the nine $fFDR10\%$ portfolios corresponding to the nine covariates, the portfolio $FDR10\%$ of BSW and the two equally weighted portfolios.

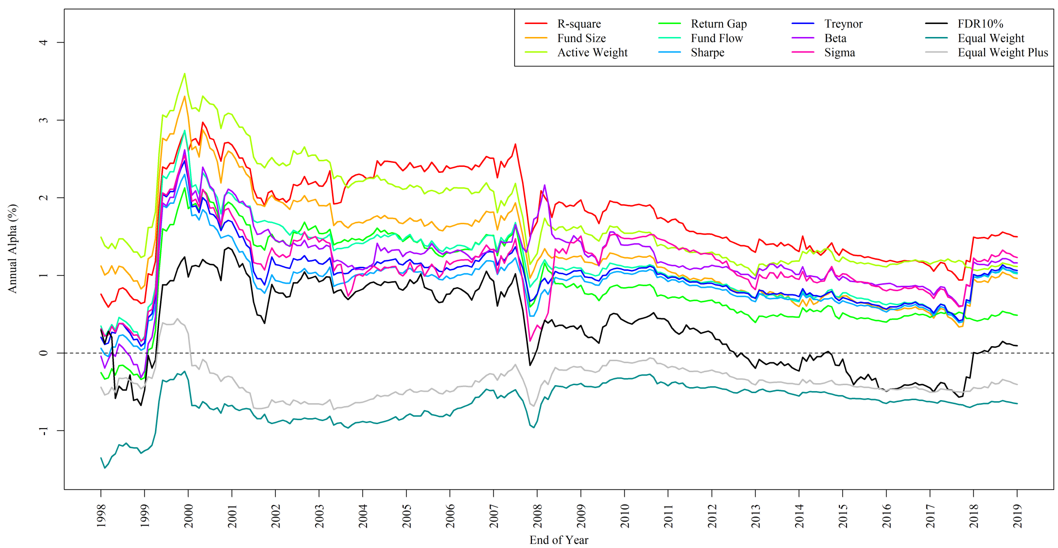


Table A.10: Comparison of portfolios' performances for varying time lengths of investing: results for sample data from 1984 to 2019. In this table, we consider 10 portfolios including nine $fFDR10\%$ portfolios corresponding to the nine covariates and the $FDR10\%$ portfolio of BSW. We compare the average alphas of the portfolios that are kept in periods of exactly n consecutive years. For example, consider $n = 5$. For each portfolio, we calculate the alpha for the first 5 years based on the portfolios' returns from January 1989 to December 1993. Then, we roll forward by a month and calculate the second alpha. The process is repeated and the last alpha is estimated based on the portfolios' returns from January 2015 to December 2019. The average of these alphas is presented in the first rows of the table.

	$fFDR10\%$									
n	R-square	Fund Size	Active Weight	Return Gap	Fund Flow	Sharpe	Treynor	Beta	Sigma	$FDR10\%$
5	1.17	0.62	0.88	0.22	0.5	0.35	0.47	0.53	1.02	-0.45
10	1.43	0.61	0.99	0.46	0.58	0.5	0.51	0.91	1.04	-0.37
15	1.64	0.6	1.09	0.65	0.69	0.65	0.63	0.96	1.03	-0.17
20	1.61	0.65	1.29	0.7	0.77	0.79	0.75	1.07	1.17	-0.12
25	1.28	0.53	1.12	0.43	0.61	0.59	0.57	0.9	0.93	-0.33
30	1.45	0.93	1.07	0.43	1.02	1.05	1.05	1.13	1.21	0.03
31	1.5	0.96	1.11	0.49	1.02	1.04	1.06	1.16	1.23	0.1

A.4 A comparison of portfolios' trading metrics

Next, we evaluate our portfolios in regard to a set of trading metrics, including the annualized estimated alpha $\hat{\alpha}$ of the Carhart four-factor model, its bootstrap p-value and t-statistic (with use of heteroskedasticity and autocorrelation-consistent standard error), the annual standard deviation of the four-factor model residuals ($\hat{\sigma}_\varepsilon$), the geometric mean return in excess of the one-month T-bill rate, the annual Sharpe ratio and the annual Information Ratio $\hat{\alpha}/\hat{\sigma}_\varepsilon$. All metrics are presented in Table A.11. We find that the $fFDR10\%$ portfolio based on the R-square covariate is the best for all considered metrics.

Table A.11: Comparison of performance statistics of all considered portfolios ($\tau = 10\%$). The table compares the portfolios with regard to metrics including the annual Carhart four-factor alpha ($\hat{\alpha}$, in %) with its bootstrap p-value and t-statistic (with use of Newey–West heteroskedasticity and autocorrelation-consistent standard error), the annual standard deviation of the four-factor model residuals ($\hat{\sigma}_\varepsilon$, in %), the mean return in excess of the one-month T-bill rate (in %), the annual Sharpe ratio and the annual Information Ratio ($IR = \hat{\alpha}/\hat{\sigma}_\varepsilon$).

Covariate	$\hat{\alpha}$ (p-value)	t-statistic	$\hat{\sigma}_\varepsilon$	Mean Return	Sharpe Ratio	IR
R-square	1.69 (0.06)	1.85	4.42	7.93	0.61	0.38
Fund Size	1.14 (0.2)	1.32	4.02	7.34	0.56	0.28
Active Weight	1.38 (0.1)	1.72	3.79	8	0.6	0.36
Return Gap	0.77 (0.34)	0.99	3.81	7.38	0.55	0.2
Fund flow	1.3 (0.14)	1.56	3.78	7.75	0.6	0.34
Sharpe	1.04 (0.2)	1.33	3.37	7.77	0.62	0.31
Treynor	1.15 (0.15)	1.45	3.49	7.65	0.6	0.33
Beta	1.67 (0.07)	1.78	4.92	7.28	0.55	0.34
Sigma	1.27 (0.26)	1.16	5.01	7.69	0.57	0.25
$FDR10\%$	0.36 (0.72)	0.37	4.75	6.5	0.52	0.08
Equal Weight	-0.8 (0.03)	-2	1.86	6.3	0.5	-0.43
Equal Weight Plus	-0.26 (0.48)	-0.56	2.18	6.7	0.52	-0.12

A.5 Performance of $fFDR10\%$ in various periods

In this section, we present the alpha of the $fFDR10\%$ portfolios in periods before and after the covariates were published. The first line of Table A.12 shows that all covariates gain positive alpha for the period January 1982 to the end of the prior-published year. The last line of the table indicates that three of the five previously known covariates still gain significant alpha in the post-published period.

Table A.12: Performance of $fFDR10\%$ portfolios in various periods prior- and post-published year of the covariates. The table shows the annualized alpha (in %) of the $fFDR10\%$ portfolio corresponding to each covariate in specific periods, with $[a, b]$ denoting a period extending from the beginning of year a over to the end of year b. For instance, the first value in the R-square column, that is, 1.75, is the alpha of the $fFDR10\%$ with R-square covariate for the period from the beginning of 1982 to the end of 2012 (i.e., $n-1 = 2012$, where $n = 2013$ is the published year of the covariate). The middle value in the column is the Carhart four-factor alpha of the portfolio for year n, which contains only 12 months corresponding to 12 data points of returns.

	$fFDR10\%$				
Period	R-square $n = 2013$	Fund Size $n = 2017$	Active Weight $n = 2015$	Return Gap $n = 2008$	Fund flow $n = 1999$
$[1982, n-1]$	1.75	0.82	1.46	1.53	1.6
$[n-10, n-1]$	1.20	-2.04	-1.27	3.00	0.35
$[n-5, n-1]$	-2.11	0.11	-1.54	1.71	-0.65
$[n-4, n-1]$	-1.81	-0.12	-0.76	0.62	-1.01
$[n-3, n-1]$	-2.50	0.22	2.79	0.20	-1.44
$[n-2, n-1]$	-2.44	0.50	3.09	0.82	-1.83
$[n-1, n-1]$	-0.92	-2.1	8.00	-0.04	-0.87
$[n, n]$	-4.77	-2.39	-3.22	2.67	-0.40
$[n+1, n+1]$	4.27	1.33	-1.81	2.70	20.76
$[n+1, n+2]$	-0.21	5.45	-0.91	-0.95	6.85
$[n+1, n+3]$	1.45	-	-0.53	-1.73	4.33
$[n+1, n+4]$	1.82	-	-0.05	-0.81	2.36
$[n+1, n+5]$	3.03	-	-	-2.13	1.90
$[n+1, n+10]$	-	-	-	-0.59	2.47
$[n+1, 2019]$	3.73	5.45	-0.05	-0.3	1.31

A.6 The construction of sorting portfolios

Here, we describe the constructions of the single- and double-sorting portfolios which are traditionally conducted in the literature. Specifically, the single-sorting portfolios based on a covariate are as in Kacperczyk *et al.* (2008) and Doshi *et al.* (2015), and the double-sorting based on a covariate and the past alpha are as in Amihud and Goyenko (2013).

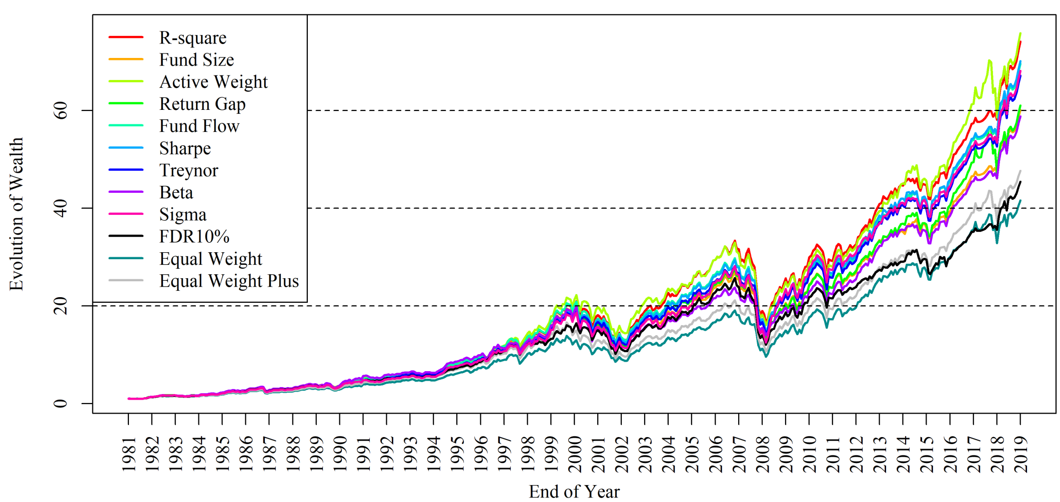
To construct the single-sorting portfolio for each covariate, at the end of each year from 1981, all the mutual funds are sorted into deciles (quintiles) according to the given covariate. For the covariate that has a negative/positive relationship with the performance of the funds, the funds in the bottom/top decile (quintile) are selected. These form a portfolio to be invested in the following year. To form the double-sorting portfolio, the funds selected in the single-sorting portfolio are again sorted into decile (quintile) according to the past alpha. The funds in the top decile (quintile) form the portfolio to be invested in the following year. This process is rolled forward until the end of the sample period.

For consistency with the $fFDR$ portfolios, we use the same type of 5-year rolling window, i.e., each time we use the aforementioned observed covariates and the alpha and covariates calculated based on the last five years.

A.7 Wealth evolution

In Figure 1.5 in Chapter 1, we study the alpha evolution of the portfolios over time. However, an investor may be interested in the gain in value. Figure A.10 shows the growth of 1 dollar that the investor invests in each portfolio at the beginning of 1982. Ultimately, at the end of 2019, this amount grows to about 74 dollars if she chooses the $fFDR10\%$ portfolio with R-square as the covariate, as opposed to just 45, 47 and 41 dollars with the $FDR10\%$, the equal weight plus and equally weighted portfolios, respectively. This exercise reveals the potential profitability of an investor who had the perfect oracle in 1982 on the methods and the covariate that would be presented over the next 30 years.

Figure A.10: Evolution of wealth. The graph plots the evolution of 1 dollar invested at the beginning of 1982 in the nine $FDR10\%$ portfolios corresponding to the nine covariates, the $fFDR10\%$, the Equal Weight and Equal Weight Plus portfolios.



A.8 Results for alternative targets of FDR

In this section, we repeat the exercise with the FDR target of 20%. Figure A.11 presents the alpha evolution of the individual covariates and Table A.13 shows the average n-year alpha of those portfolios.

Figure A.11: Alpha evolution of $fFDR20\%$ and $FDR20\%$ portfolios over time. The graph presents the evolution of annualized alpha of the nine $fFDR20\%$ portfolios corresponding to the nine covariates, the $FDR20\%$ of BSW and the two equally weighted portfolios.

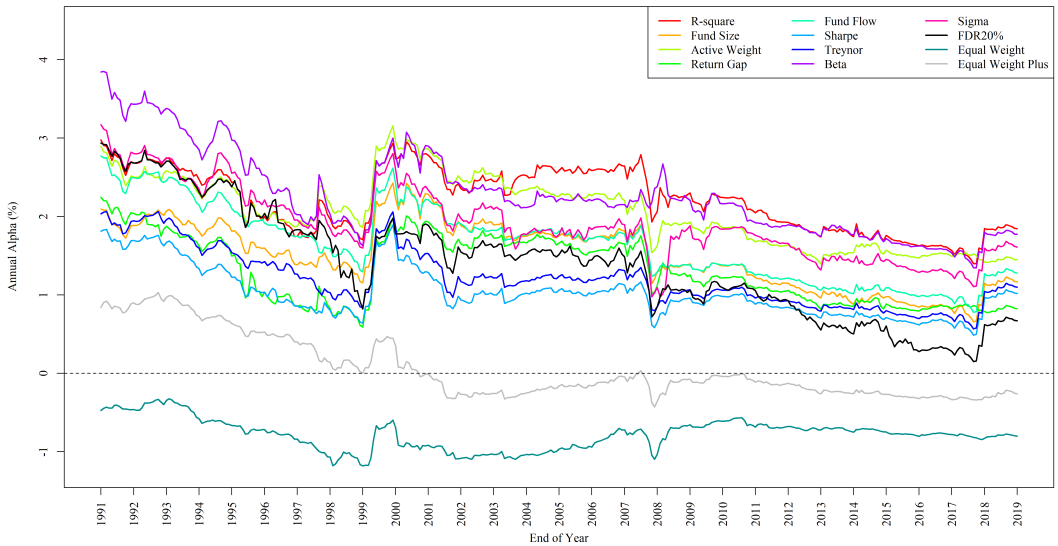


Table A.13: Comparison of portfolios' performances for varying time lengths of investing. In this table, we consider 10 portfolios including nine $fFDR20\%$ portfolios corresponding to the nine covariates and the $FDR20\%$ portfolio of BSW. We compare the average alphas (annualized and in %) of the portfolios that are kept in periods of exactly n consecutive years. For example, consider $n = 5$. For each portfolio, we calculate the alpha for the first 5 years based on the portfolios' returns from January 1982 to December 1986. Then, we roll forward by a month and calculate the second alpha. The process is repeated and the last alpha is estimated based on the portfolios' returns from January 2015 to December 2019. The average of these alphas is presented in the first row in the table.

	$fFDR20\%$									$FDR20\%$
n	R-square	Fund Size	Active Weight	Return Gap	Fund Flow	Sharpe	Treynor	Beta	Sigma	
5	1.6	0.8	1.23	0.61	0.89	0.58	0.65	1.15	1.4	0.41
10	1.63	0.82	1.21	0.61	0.93	0.65	0.7	1.33	1.2	0.34
15	1.84	0.92	1.46	0.82	1.06	0.79	0.83	1.34	1.22	0.41
20	1.97	1.05	1.66	1.03	1.15	0.9	0.92	1.44	1.28	0.53
25	1.75	0.9	1.42	0.78	0.99	0.79	0.82	1.37	1.18	0.42
30	1.55	0.81	1.28	0.67	0.95	0.76	0.8	1.35	1.16	0.31
38	1.84	1.16	1.45	0.82	1.28	1.02	1.1	1.77	1.61	0.67

A.9 Results from using an alternative proxy of covariates

In this section, we present in Figure A.12 the alpha evolution of $fFDR10\%$ portfolios and in Table A.14 their average $n-$year alpha where the proxy for each covariate is based on whole data in the in-sample period instead of the data in final year as in Chapter 1. We see that the performance of the portfolios does not vary significantly.

Figure A.12: Alpha evolution of $fFDR10\%$ portfolios with use of an alternative covariate proxy. The graph presents the evolution of annualized alpha (in %) of the nine $fFDR10\%$ portfolios (corresponding to the nine covariates), the portfolio $FDR10\%$ of BSW and the two equally weighted portfolios. The proxy for each covariate (except the R-square and the four covariates obtained from the asset pricing models) is its average realizations in the five years in-sample period.

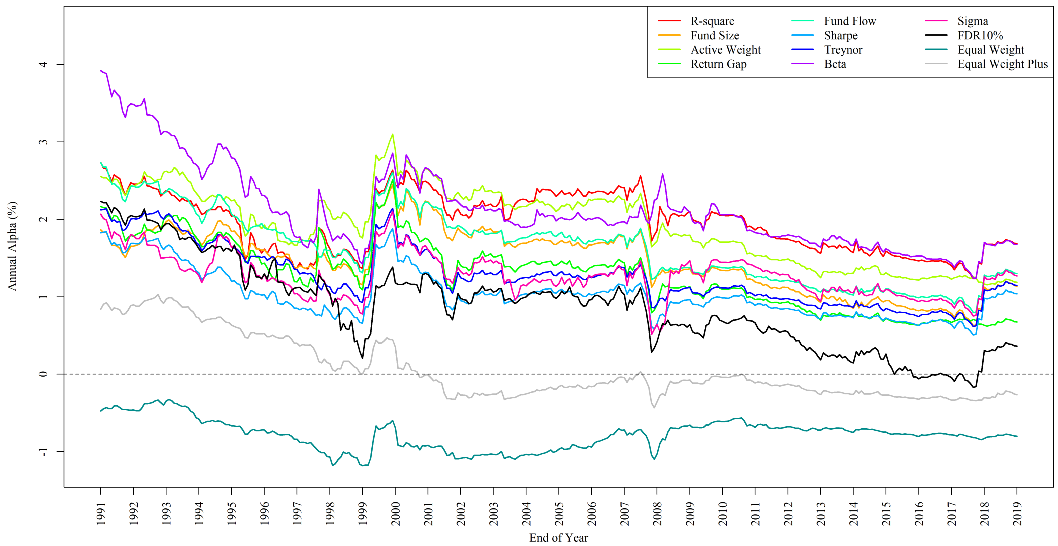


Table A.14: Comparison of portfolios' performances for varying time lengths of investing: alternative proxy of covariates. We consider 10 portfolios including nine $fFDR10\%$ portfolios (corresponding to the nine covariates) and the $FDR10\%$ portfolio of BSW. We compare the average alphas (annualized, in %) of the portfolios that are kept for periods of exactly n consecutive years. The proxy for each covariate (except the R-square and the four covariates obtained from the asset pricing models) is its average realizations in the five years in-sample period.

	$fFDR10\%$									$FDR10\%$
n	R-square	Fund Size	Active Weight	Return Gap	Fund Flow	Sharpe	Treynor	Beta	Sigma	
5	1.49	0.87	1.28	0.69	0.92	0.57	0.73	1.09	1.19	0.12
10	1.48	0.85	1.25	0.63	0.93	0.65	0.76	1.2	1.06	0.05
15	1.7	0.94	1.42	0.72	1.06	0.79	0.88	1.2	1.09	0.14
20	1.84	1.05	1.53	0.79	1.15	0.91	0.96	1.31	1.17	0.26
25	1.61	0.9	1.29	0.61	0.99	0.8	0.86	1.24	1.09	0.13
30	1.41	0.78	1.11	0.5	0.95	0.78	0.86	1.2	1.01	0.01
38	1.69	1.14	1.19	0.67	1.3	1.04	1.15	1.67	1.27	0.36

A.10 Covariate combinations

So far, we have considered the effect from the information brought in by each single covariate. In what follows, we explore the effect from combining the information from the different covariates and potential consequent performance improvement. More specifically, we create a new covariate given by the linear combination of the underlying covariates. More specifically, for each fund i at time t, we have

$$\begin{aligned}\text{New Covariate}_{t,i} &= c_{1t}\text{R-square}_{t,i} + c_{2t}\text{Active Weight}_{t,i} + c_{3t}\text{Return Gap}_{t,i} \\ &\quad + c_{4t}\text{Fund Size}_{t,i} + c_{5t}\text{Fund Flow}_{t,i} + c_{6t}\text{Sharpe Ratio}_{t,i} \\ &\quad + c_{7t}\text{Treynor Ratio}_{t,i} + c_{8t}\text{Sigma}_{t,i} + c_{9t}\text{Beta}_{t,i}. \qquad \text{(A.3)}\end{aligned}$$

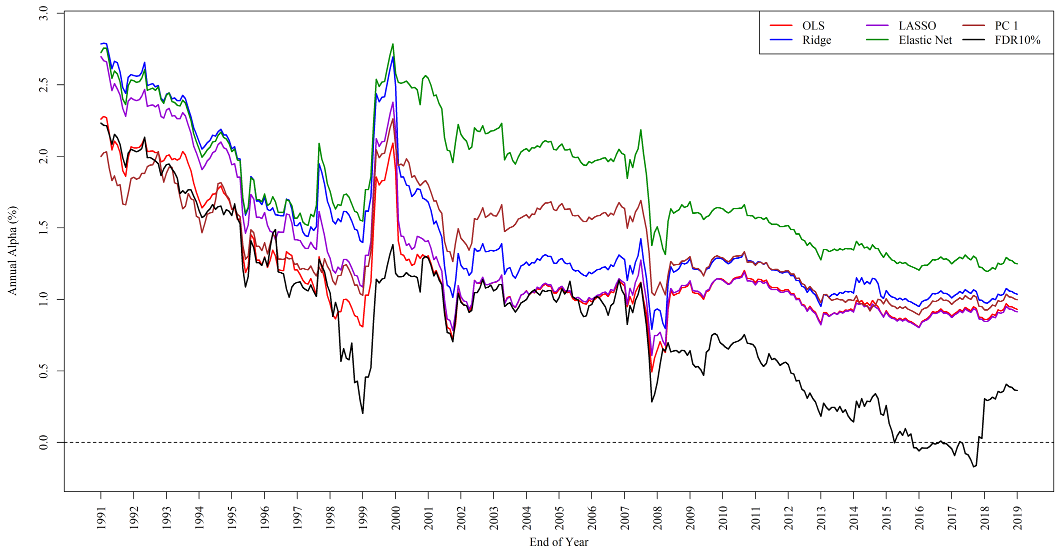
We consider two approaches to estimating the coefficients $c_{1t}, \ldots, c_{9t}$ in (A.3). First, we use as our new covariate the first principal component of all nine (standardized) covariates. By transforming the covariates to their principal components, their information about the performance of a fund is preserved and conveyed. We use the first principal component as it captures most of the variation of the covariates. Second, we use a linear model that regresses the fund returns for year k on the observed value of the covariates in year $k-1$, where $k \in \{t, t-1, t-2, t-3\}$. Then, we predict the return for year $t+1$ based on the estimated regression model and the covariates in year t. This is equivalent to using equation (A.3) with the regression's estimated coefficients as the $c_{1t}, \ldots, c_{9t}$. We use ordinary least squares (OLS), the least absolute shrinkage and selection operator (LASSO) of Tibshirani (1996) , ridge regression and the elastic net of Zou *et al.* (2005).[3]

Figure A.13 exhibits the performance of the $fFDR\tau$ portfolios with the newly created covariates in terms the alpha evolution.[4] We find that the portfolio based on the covariate obtained from the elastic net performs best amongst the combined covariates at $\tau = 10\%$.

[3]For each method (except OLS), the covariates are standardized before being used in the estimation. We use cross-validation to determine the parameters in the LASSO, ridge and elastic net methods.

[4]There are a few years where LASSO (two years) and the elastic net (three years) shrink all the regression coefficients to zero. In these cases, the new covariate is equal to zero for all funds and, to avoid an empty portfolio, we simply select all the funds in the $FDR\tau$ portfolio.

Figure A.13: Alpha evolution of $fFDR10\%$ portfolios with combined covariates. The graph shows the alpha evolution of the $fFDR10\%$ portfolios with each using a covariate obtained from either the principal component method or regression method; for the former, the covariate is the first principal component (PC 1) of the five covariates, whereas for the latter the new covariate is a linear combination of the five underlying covariates with the weights obtained based on one of the OLS, LASSO, Ridge and elastic net regressions.



Aiming to acquire a more complete portrayal of the various covariates combinations, we study also the portfolios' alphas for various time lengths of investing. Table A.15 shows the average n-year alphas of the $fFDR10\%$ portfolios.

Table A.15: Performance of $fFDR10\%$ portfolios with combined covariates for varying time lengths of investing. The table displays the average n-year alpha (annualized and in %) of the $fFDR10\%$ portfolios which use covariates obtained by the first principal component (PC 1), the OLS, LASSO, Ridge and elastic net (see descriptions in Figure **??**). The average n-year alpha of each portfolio is calculated as per the description in Table 1.6.

n	OLS	Ridge	LASSO	Elastic Net	PC 1
5	0.78	1.02	0.8	1.2	0.76
10	0.81	1.03	0.81	1.36	0.94
15	0.91	1.07	0.89	1.5	1.17
20	1.06	1.15	1	1.67	1.31
25	0.96	1.07	0.9	1.44	1.13
30	0.94	1.05	0.89	1.32	1.02
38	0.93	1.04	0.91	1.25	1

The elastic net performs also better for all time lengths. However, this best combined covariate does not beat the R-square and Beta under the $fFDR$ framework.

A.11 Restricted data

As supplementary to our empirical study of Section 1.6, we repeat here our experiments for a data subset where a mutual fund enters the sample when its TNA reaches \$15 million (adjusted for inflation as of January 2019). This choice of threshold is consistent with Pastor *et al.* (2015) and Zhu (2018). Figure A.14 shows the alpha evolution of the $fFDR10\%$ portfolios based on each individual covariate and Table A.16 reports their average n-year alpha.

Figure A.14: Alpha evolution of $fFDR10\%$ portfolios with data restricted. The graph presents the evolution of annualized alpha (in %) of the nine $fFDR10\%$ portfolios (corresponding to the nine covariates), the portfolio $FDR10\%$ of BSW and the two equally weighted portfolios. The portfolios are constructed based on the data that contains only funds' returns from when their TNA reaching at least \$15 million.

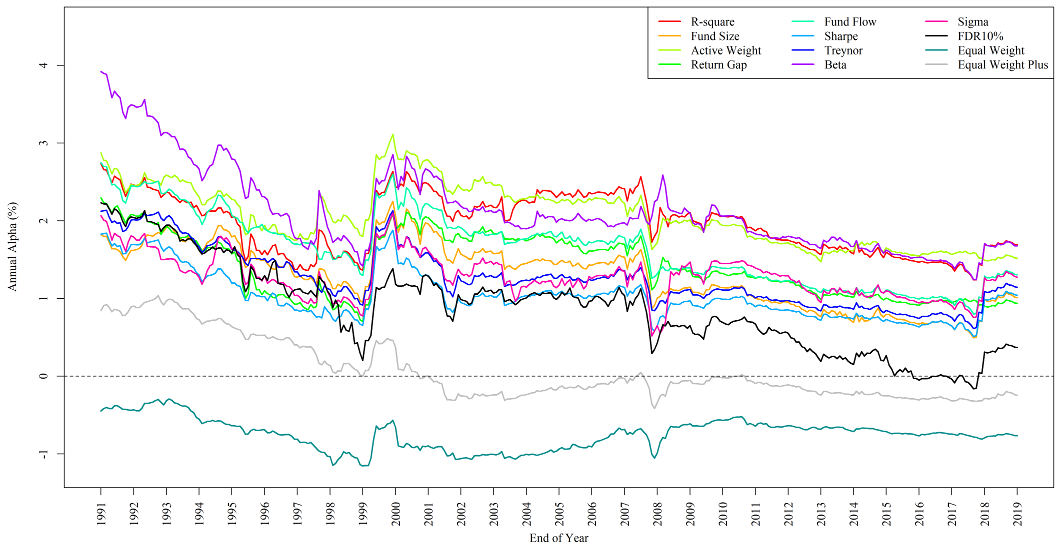


Table A.16: Comparison of portfolios' performances for varying time lengths of investing: restricted data. We consider 10 portfolios including nine $fFDR10\%$ portfolios (corresponding to the nine covariates) and the $FDR10\%$ portfolio of BSW. We compare the average alphas (annualized, in %) of the portfolios that are kept for periods of exactly n consecutive years. The portfolios are constructed based on the data that contains only funds' returns from when their TNA reaching at least \$15 million.

	$fFDR10\%$									$FDR10\%$
n	R-square	Fund Size	Active Weight	Return Gap	Fund Flow	Sharpe	Treynor	Beta	Sigma	
5	1.5	0.81	1.39	0.62	0.93	0.57	0.73	1.09	1.19	0.13
10	1.48	0.68	1.36	0.63	0.93	0.65	0.75	1.2	1.06	0.06
15	1.7	0.73	1.6	0.85	1.06	0.8	0.87	1.2	1.1	0.15
20	1.84	0.82	1.79	1.07	1.14	0.91	0.96	1.31	1.18	0.27
25	1.62	0.71	1.56	0.82	0.98	0.81	0.86	1.24	1.09	0.14
30	1.42	0.63	1.41	0.69	0.95	0.79	0.86	1.2	1.01	0.02
38	1.69	1.01	1.52	0.94	1.3	1.04	1.14	1.68	1.27	0.37

Appendix B

Appendix for chapter 2

B.1 The multivariate functional false discovery rate

To complement Section 2.2.1 of the manuscript, we present the procedures for estimating the $\pi_0(\boldsymbol{z}), f(p, \boldsymbol{z})$ and $q-$value.

To estimate $\pi_0(\boldsymbol{z})$ one can extend the "bin" approach presented in HKMS, where we partition m tests into some groups based on covariates and estimate $\pi_0(\boldsymbol{z})$ as a constant in each group. In this study, we additionally utilize another approach which is based on density estimation as follows. Firstly, for some $\lambda \in [0, 1)$ we define:

$$\pi_0(\boldsymbol{z}, \lambda) = \frac{\mathbb{P}(P > \lambda | \boldsymbol{Z} = \boldsymbol{z})}{1 - \lambda}. \tag{B.1}$$

Conditional on $\boldsymbol{Z} = \boldsymbol{z}$, $\mathbb{P}(P > \lambda) \geq \mathbb{P}(P > \lambda | h = 0).\mathbb{P}(h = 0)$. Thus, we have

$$\begin{aligned} \mathbb{P}(P > \lambda | \boldsymbol{Z} = \boldsymbol{z}) &\geq \mathbb{P}(P > \lambda | \boldsymbol{Z} = \boldsymbol{z}, h = 0).\mathbb{P}(h = 0 | \boldsymbol{Z} = \boldsymbol{z}) \\ &= (1 - \lambda)\mathbb{P}(h = 0 | \boldsymbol{Z} = \boldsymbol{z}) \\ &= (1 - \lambda)\pi_0(\boldsymbol{z}) \end{aligned} \tag{B.2}$$

where the second step comes from the uniform distribution of $P|(\boldsymbol{Z} = \boldsymbol{z}, h = 0)$. This turns out $\pi_0(\boldsymbol{z}, \lambda) \geq \pi_0(\boldsymbol{z})$, i.e. $\pi_0(\boldsymbol{z}, \lambda)$ is an conservative estimate of $\pi_0(\boldsymbol{z})$. If we express $\pi_0(\boldsymbol{z}, \lambda)$ as

$$\pi_0(\boldsymbol{z}, \lambda) = \mathbb{P}(\boldsymbol{Z} = \boldsymbol{z} | P > \lambda).\frac{\mathbb{P}(P > \lambda)}{1 - \lambda} \tag{B.3}$$

then, the first term $\frac{\mathbb{P}(P>\lambda)}{1-\lambda}$ is conservatively estimated by $\hat{\pi}_0(\lambda) = \frac{\#\{p_i>\lambda\}}{n(1-\lambda)}$, here $\#$ returns the number of elements in the set, as in Storey (2002). The remain $\mathbb{P}(\boldsymbol{Z} =$

$\boldsymbol{z}|P > \lambda)$, which is the density function of $\boldsymbol{Z}$ conditional on $p-$value $> \lambda$, will be estimated as a function $\hat{h}_\lambda(\boldsymbol{z})$ such that

$$\hat{\pi}_0(\boldsymbol{z}, \lambda) = \hat{h}_\lambda(\boldsymbol{z}).\hat{\pi}_0(\lambda) \tag{B.4}$$

is a conservative estimate of $\pi_0(\boldsymbol{z})$.

Next, to estimate the density functions $\hat{h}_\lambda(\boldsymbol{z})$ and $f(p, \boldsymbol{z})$ we use a local likelihood kernel density estimation (KDE) approach in which a probit transformation in Geenens (2014) is adopted. More specifically, let $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} dx$ and Φ^{-1} its inverse. We transform the variables $(p, \boldsymbol{z})$ to $(\tilde{p}, \tilde{\boldsymbol{z}})$ by using $\tilde{p}_i = \Phi^{-1}(p_i)$ and $\tilde{z}_i^k = \Phi^{-1}(z_i^k)$, $k = 1, \ldots, d$. We denote by $\tilde{f}(\tilde{p}, \tilde{\boldsymbol{z}})$ and $\tilde{h}_\lambda(\tilde{\boldsymbol{z}})$, respectively, the joint density function of $(\tilde{p}, \tilde{\boldsymbol{z}})$ and the conditional density function of $\tilde{z}$ on the group of null hypotheses having $p-$value $< \lambda$, $\{i|P_i > \lambda\}$. We estimate them by using the local likelihood KDE method in Loader (1999). When the number of variables in the estimating function is greater than two (i.e., $d \geq 3$ for $\hat{h}_\lambda(\tilde{\boldsymbol{z}})$ and $d \geq 2$ for $f(p, \tilde{\boldsymbol{z}})$), we drop the cross-product terms from the local model. This allows us to overcome the curse of dimensionality as we are working with one-dimensional integrals instead of a multivariate one. The estimation can be implemented easily via the freely available R package `locfit`. When the dimension is less than three, the bandwidth of the KDE is chosen locally via a k-Nearest-Neighbor approach using generalized cross-validation similar to CRS.

The desired density functions $\hat{h}_\lambda(\boldsymbol{z})$ and $f(p, \boldsymbol{z})$, respectively, are then estimated, as $\hat{h}_\lambda(\boldsymbol{z}) = \frac{\tilde{h}_\lambda(\tilde{\boldsymbol{z}})}{\Pi_{k=1}^{d} \phi(\tilde{z}^k)}$ and $\hat{f}(p, \boldsymbol{z}) = \frac{\tilde{f}(\tilde{p}, \tilde{\boldsymbol{z}})}{\phi(\tilde{p}) \Pi_{k=1}^{d} \phi(\tilde{z}^k)}$ where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

For each λ, the $\hat{h}_\lambda(\boldsymbol{z})$ is plugged to (B.4) to acquire the estimate $\hat{\pi}_0(\boldsymbol{z}, \lambda)$. The λ is chosen in the set $\{0.4, \ldots, 0.8\}$ such that the mean integrated squared error defined as in CRS is minimal.

As argued in CRS, if the $f(p, z)$ is non-increasing with respect to p for each fixed $\boldsymbol{z}$, then we should adjust the $\hat{f}(p, z)$ so that it carries this property. In practice, this can be acquired by resetting the value of $\hat{f}(p_i, z_i)$

$$\hat{f}(p_i, z_i) := \min\left\{\hat{f}(p_i, z_i), \min\left\{\hat{f}(p_l, z_l) | p_l < p_i, \|z_i - z_l\| \leq \epsilon\right\}\right\}$$

where $\|.\|$ is the Euclid distance and ϵ is a small positive number which is set at 0.01

in this study.

Finally, we calculate $\hat{r}(p, \boldsymbol{z}) = \hat{\pi}_0(\boldsymbol{z})/\hat{f}(p, \boldsymbol{z})$ and estimate the so-called $q-$value function (see CRS) as

$$\hat{q}(p_i, \boldsymbol{z}_i) = \frac{1}{\# S_i} \sum_{k \in S_i} \hat{r}(p_k, \boldsymbol{z}_k), \tag{B.5}$$

where $S_i = \{k | \hat{r}(p_k, \boldsymbol{z}_k) \leq \hat{r}(p_i, \boldsymbol{z}_i)\}$, p_i is the $p-$value of test i and $\boldsymbol{z}_i = (z_i^1, \ldots, z_i^d)$ is the covariate bundle of strategy i.[1]

We recall here the "positive false discovery rate", a type I error introduced in Storey (2003), $pFDR = \mathbb{E}\left(\frac{V}{R} \middle| R > 0\right)$ where R is the number of rejected null hypotheses in n tests and V the wrongly rejected ones. For a given target $\tau \in [0, 1]$ of $pFDR$, a null hypothesis $H_{0,i}$ is rejected if and only if $\hat{q}(p_i, \boldsymbol{z}_i) \leq \tau$. One can replicate the arguments in CRS to show that the $mfFDR$ procedure controls at the target τ of $pFDR$. We emphasize that control for $pFDR$ is equivalent to control for the FDR at the same level when n is large which is the case in this study.

B.2 Performance of $mfFDR$ under noisy covariates

As mentioned in Section 2.2.3, the covariates are estimated quantities and thus, have inherited some noise that might affect the power of our method. In this section, we address this concern by considering a simple setting where the input covariates contain noise. More specifically, we use the covariates $u = (u_1, \ldots, u_n), v = (v_1, \ldots, v_n)$ as in our previous simulations in Section 2.2.2 to generate the simulated data, but now we use inputs of $mfFDR$ the observed $u' = (u'_1, \ldots, u'_n)$ and $v' = (v'_1, \ldots, v'_n)$ defined as

$$u'_i = u_i + \eta_i \tag{B.6}$$

and

$$v'_i = v_i + \epsilon_i \tag{B.7}$$

where η_i and ϵ_i are noise generated independently from a normal distribution $N(0, \sigma^2)$, $i = 1, \ldots, n$. To study the performance of the $mfFDR$, we consider three different values of σ including $\sigma_1 = 0.5/\sqrt{12}$, $\sigma_2 = 1.0/\sqrt{12}$ and $\sigma_3 = 1.5/\sqrt{12}$. The determina-

[1] This estimation is proposed by Newton *et al.* (2004) and Storey *et al.* (2005) and subsequently adopted in CRS.

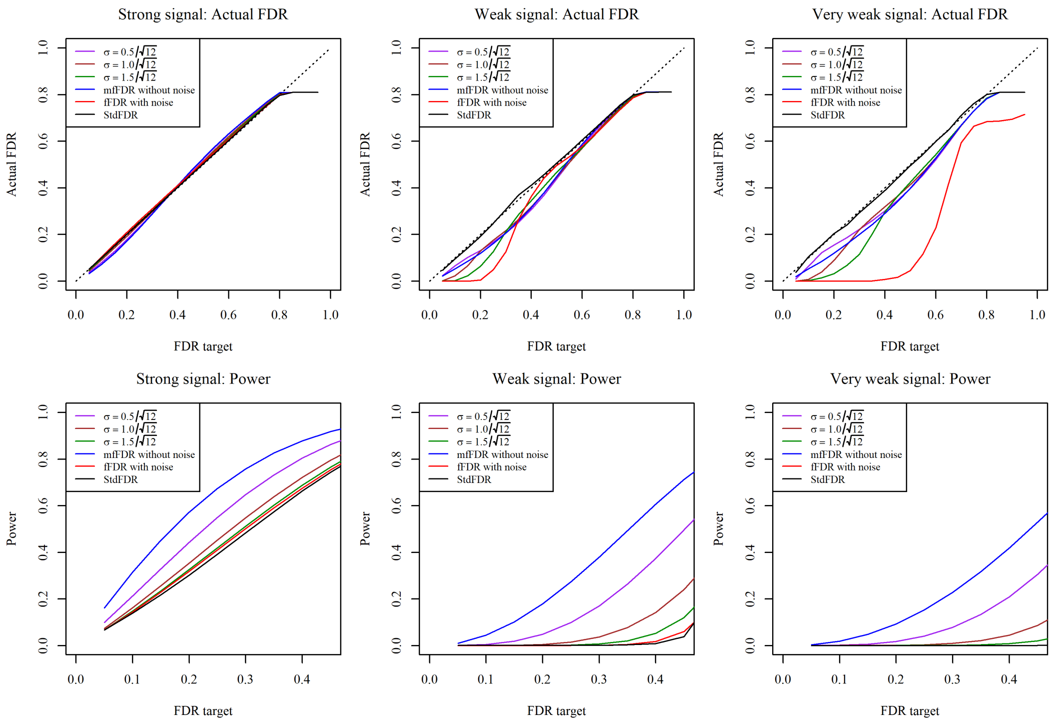
tion of those values are based on the fact that the covariates $u, v \sim U[0,1]$, which has a standard deviation of $1/\sqrt{12}$.

The magnitude of the noise in this setting reflects the level of informativeness of the covariates (the higher variation of the noise is, the less informative the covariates are) or the strength of the relationship between the observed covariates and the true status of hypotheses (the higher variation of the noise is, the weaker the relationship is).

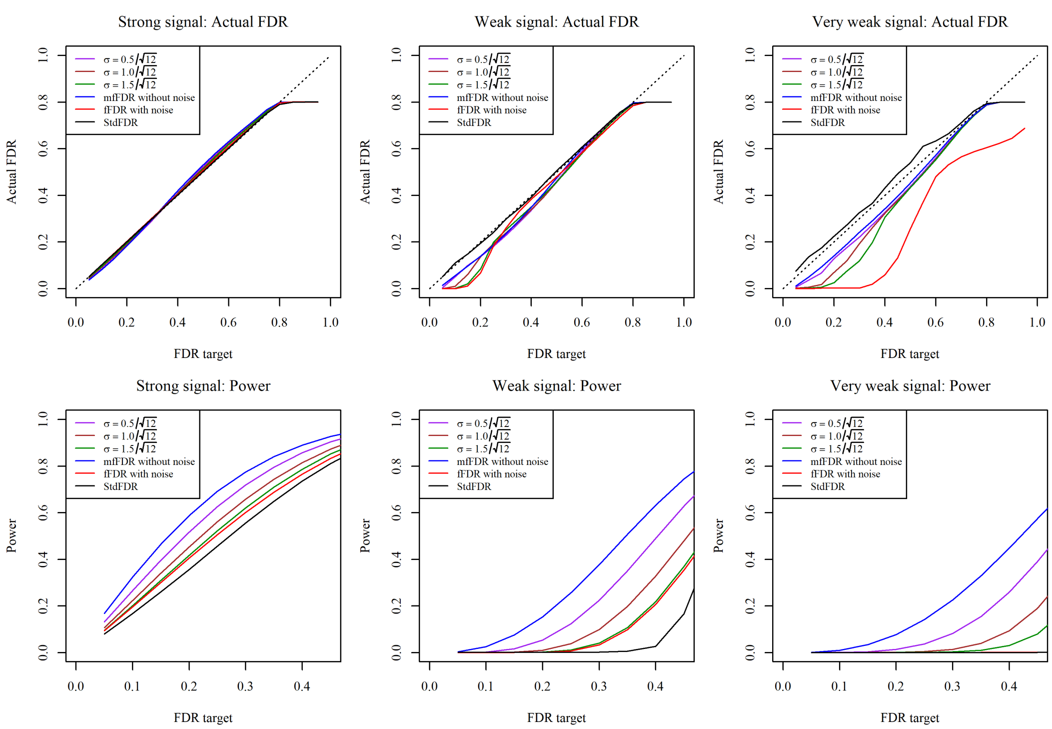
In Figure B.1, we present the performance of $mfFDR$ in terms of FDR control and its power (at FDR target up to 0.45). We see that the FDR is controlled well at any given target from 0.05 to 0.95. To have a complete picture on the impact of the noise on the power of the covariate augmented methods, we add the performance of the $fFDR$ when its covariate is u' and compare it to the $mfFDR$ using original (u, v) and the $StdFDR$. Evidently, the power of the $mfFDR$ with noise in covariates is lower than the case with original ones but still remarkably higher than that of the $fFDR$ (with noised covariate u') and $StdFDR$.

Figure B.1: Performance comparison under noisy covariates. The figure compares the performance of the $mfFDR$ and $fFDR$ with noisy covariates and the Standard FDR of Storey ($SdtFDR$) procedures. Here the the input covariates for the $mfFDR$ are $u' = u + \varepsilon, v' = v + \eta$, where $\varepsilon, \eta \sim N(0, \sigma^2)$ and $\sigma \in \{0.5/\sqrt{12}, 1.0/\sqrt{12}, 1.5/\sqrt{12}\}$, whereas the $fFDR$ the u'. Panel A shows the performance when the $\pi_0(u, v)$ has a sine form whereas Panel B is the monotonic one.

Panel A: $\pi_0(u, v)$ is a sine function.



Panel B: $\pi_0(u, v)$ is a monotonic function.



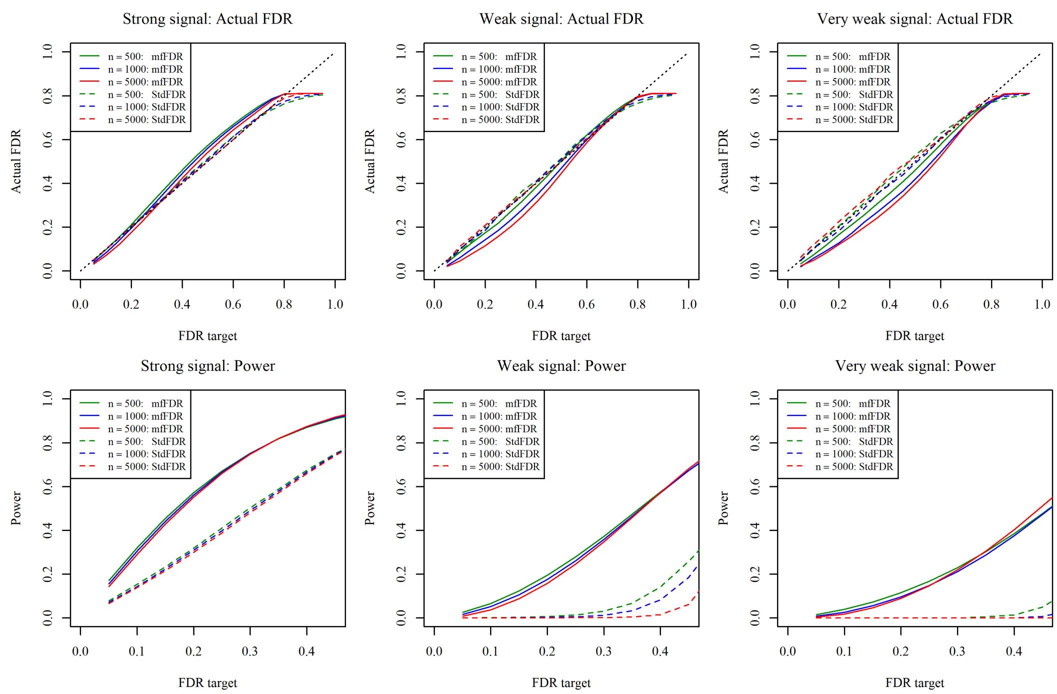
B.3 Performance of $mfFDR$ under varying number of tests

In the main manuscript, we conduct simulation studies with population of n =10,000 tests. This number is chosen since it is close to that of the actual input of $mfFDR$ in the empirical experiments. Aiming for a wider range in application of the $mfFDR$ framework, we additionally conduct a robustness check with use of smaller numbers of tests. Particularly, we repeat the simulation with: $n = 500, 1000$ and 5000 tests.

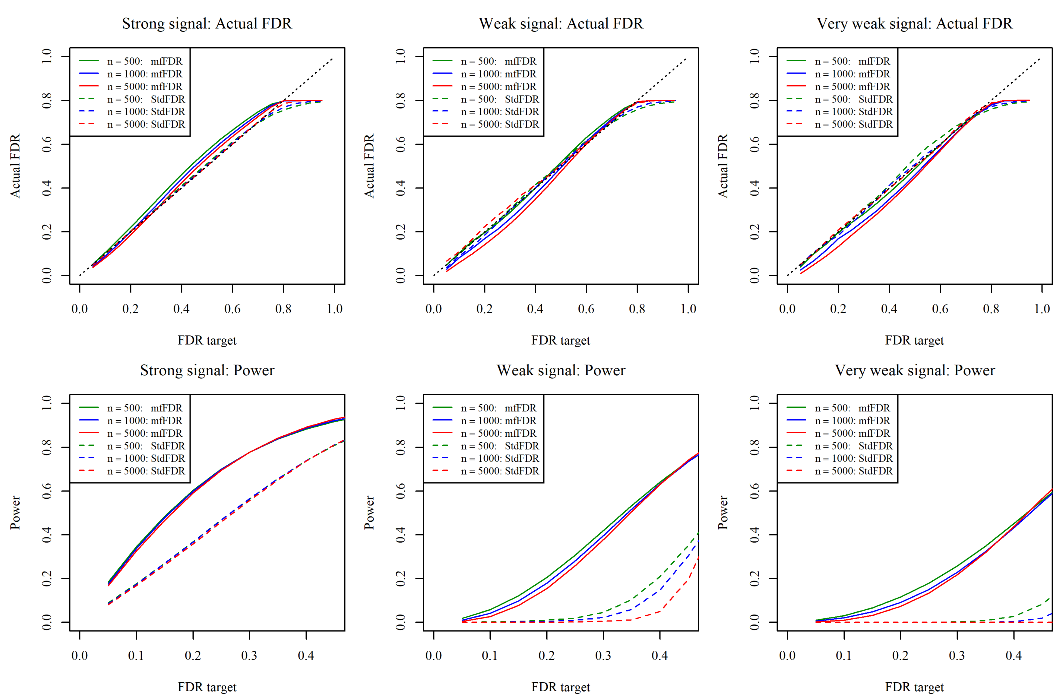
In Figure B.2, we present the performance of the $mfFDR$ against its benchmark, the $StdFDR$, in terms of FDR control and power. First, under the weak and very weak signal cases, the $mfFDR$ controls well for any given FDR targets. Second, for strong signal data, the $mfFDR$ slightly violates the FDR control at high targets, especially when the number of tests is small. More specifically, when $n = 500$, the $mfFDR$ strictly controls well for FDR at targets up to 0.2. When $n = 3000$ and $n = 5000$, these numbers are 0.3 and 0.4. Thus, if the number of considering tests is smaller than 500 and the aim is controlling for a FDR target higher than 0.2, the method should be used on weak or very weak signal rather than strong signal data. Although, it should be noted that controlling FDR at a high target with a small number of tests, has no practical value in finance, economics and most fields of science. Last, in terms of power, the $mfFDR$ is superior to that of the $StdFDR$ in all cases regardless the number of tests.

Figure B.2: Performance comparison under varying sample size. The figure compares the performance of the $mfFDR$ against the Standard FDR of Storey ($SdtFDR$) with varying population size (n). Panel A shows the performance when the $\pi_0(u,v)$ has a sine form whereas in Panel B is the monotonic one.

Panel A: $\pi_0(u,v)$ is a sine function.



Panel B: $\pi_0(u,v)$ is monotonic with respect to each covariate.



B.4 Performance of $mfFDR^+$ portfolios with various FDR targets

In this section, we present the performance of the $mfFDR^+$ based portfolios both when it is applied on individual currency as well as on all currencies on various FDR targets. More specifically, we repeat the experiment in the main manuscript with FDR target τ varying from 10% to 40%. The results for the former portfolios are presented in Table B.1 while those of all currencies together one are exhibited in Table B.2. The results corresponding to $\tau = 0.2$ is represented for conveniences in comparison. Overall, the performance of the portfolios are stable across the considered targets. When all currencies are examined together, we observe higher Sharpe ratio and smaller annual return for higher target. This observation implicitly indicates that the volatility of the portfolio's daily return is higher when we control FDR at small target.

Table B.1: Performance of *mfFDR* based portfolios on individual currency with varying *FDR* target. The table shows annualized Sharpe ratios of the *mfFDR* based portfolio with FDR target $\tau = \{0.1, 0.2, 0.3, 0.4\}$ based on portfolios' returns before (left side) and after transaction cost (right side). The final row shows the average Sharpe ratio across 30 portfolios corresponding to the 30 currencies. The numbers in parentheses are the corresponding $p-$values. "*", "**" and "***" respectively indicate statistical significance at levels of $10\%, 5\%$ and 1%.

Countries	Before transaction cost				After transaction cost			
	$\tau = 0.1$	$\tau = 0.2$	$\tau = 0.3$	$\tau = 0.4$	$\tau = 0.1$	$\tau = 0.2$	$\tau = 0.3$	$\tau = 0.4$
Australia	0.17 (0.21)	0.18 (0.17)	0.18 (0.17)	0.18 (0.17)	0.13 (0.34)	0.14 (0.29)	0.14 (0.30)	0.14 (0.31)
Canada	0.19 (0.17)	0.17 (0.23)	0.17 (0.25)	0.17 (0.24)	0.13 (0.37)	0.11 (0.46)	0.11 (0.47)	0.11 (0.46)
Germany/E.U.	0.48 (0.00)***	0.49 (0.00)***	0.48 (0.00)***	0.48 (0.00)***	0.45 (0.00)***	0.46 (0.00)***	0.45 (0.00)***	0.45 (0.00)***
Japan	0.46 (0.00)***	0.45 (0.00)***	0.45 (0.00)***	0.44 (0.00)***	0.43 (0.00)***	0.41 (0.00)***	0.41 (0.00)***	0.41 (0.00)***
New Zealand	0.32 (0.01)***	0.34 (0.00)***	0.35 (0.00)***	0.35 (0.00)***	0.27 (0.02)**	0.28 (0.02)**	0.29 (0.01)***	0.29 (0.01)***
Norway	0.20 (0.15)	0.18 (0.17)	0.18 (0.19)	0.18 (0.18)	0.15 (0.29)	0.13 (0.36)	0.13 (0.38)	0.13 (0.37)
Sweden	0.37 (0.00)***	0.40 (0.00)***	0.40 (0.00)***	0.40 (0.00)***	0.33 (0.02)**	0.36 (0.01)***	0.36 (0.00)***	0.36 (0.00)***
Switzerland	0.26 (0.06)*	0.26 (0.06)*	0.25 (0.07)*	0.25 (0.07)*	0.22 (0.11)	0.22 (0.12)	0.22 (0.12)	0.21 (0.13)
U.K.	0.29 (0.05)**	0.29 (0.05)**	0.28 (0.05)**	0.28 (0.06)*	0.26 (0.07)*	0.25 (0.08)*	0.24 (0.10)*	0.24 (0.10)*
Argentina	0.35 (0.06)*	0.35 (0.06)*	0.34 (0.06)*	0.35 (0.06)*	0.32 (0.10)*	0.31 (0.11)	0.29 (0.11)	0.30 (0.12)
Columbia	0.56 (0.00)***	0.60 (0.00)***	0.62 (0.00)***	0.61 (0.00)***	0.48 (0.01)***	0.53 (0.01)***	0.55 (0.00)***	0.54 (0.01)***
India	0.35 (0.03)**	0.34 (0.03)**	0.34 (0.03)**	0.34 (0.04)**	0.31 (0.06)*	0.30 (0.07)*	0.30 (0.07)*	0.30 (0.07)*
Indonesia	0.33 (0.02)**	0.34 (0.02)**	0.33 (0.02)**	0.33 (0.02)**	0.23 (0.10)*	0.24 (0.09)*	0.24 (0.09)*	0.24 (0.10)*
Israel	0.52 (0.00)***	0.54 (0.00)***	0.54 (0.00)***	0.54 (0.00)***	0.37 (0.02)**	0.39 (0.01)***	0.39 (0.01)***	0.39 (0.01)***
Philippines	0.64 (0.00)***	0.62 (0.00)***	0.62 (0.00)***	0.62 (0.00)***	0.48 (0.01)***	0.46 (0.01)***	0.46 (0.01)***	0.46 (0.01)***
Romania	0.19 (0.31)	0.23 (0.19)	0.24 (0.16)	0.25 (0.15)	0.09 (0.65)	0.12 (0.51)	0.13 (0.47)	0.13 (0.46)
Russia	0.47 (0.01)***	0.51 (0.01)***	0.52 (0.01)***	0.52 (0.01)***	0.44 (0.02)**	0.48 (0.01)***	0.49 (0.01)***	0.49 (0.01)***
Slovak	0.16 (0.40)	0.18 (0.35)	0.19 (0.33)	0.19 (0.31)	0.11 (0.51)	0.13 (0.48)	0.13 (0.46)	0.14 (0.43)
Brazil	0.36 (0.07)*	0.36 (0.07)*	0.35 (0.08)*	0.35 (0.08)*	0.33 (0.10)*	0.33 (0.10)*	0.32 (0.10)*	0.32 (0.10)*
Chile	0.49 (0.01)***	0.46 (0.02)**	0.45 (0.02)**	0.45 (0.02)**	0.41 (0.04)**	0.38 (0.06)*	0.37 (0.06)*	0.37 (0.06)*
Czech	0.07 (0.64)	0.12 (0.46)	0.13 (0.43)	0.14 (0.41)	0.01 (0.85)	0.06 (0.65)	0.07 (0.61)	0.08 (0.60)
Hungary	-0.12 (0.56)	-0.11 (0.60)	-0.11 (0.60)	-0.11 (0.62)	-0.17 (0.39)	-0.17 (0.41)	-0.16 (0.41)	-0.16 (0.43)
Korea	0.30 (0.15)	0.29 (0.16)	0.29 (0.16)	0.29 (0.16)	0.24 (0.23)	0.23 (0.25)	0.23 (0.26)	0.23 (0.26)
Mexico	0.10 (0.48)	0.10 (0.49)	0.09 (0.50)	0.10 (0.49)	0.05 (0.71)	0.04 (0.73)	0.04 (0.74)	0.04 (0.73)
Poland	-0.02 (0.90)	0.02 (0.92)	0.02 (0.88)	0.03 (0.85)	-0.07 (0.73)	-0.03 (0.89)	-0.03 (0.91)	-0.02 (0.93)
Singapore	0.28 (0.05)**	0.32 (0.02)**	0.32 (0.02)**	0.32 (0.02)**	0.11 (0.52)	0.15 (0.32)	0.15 (0.32)	0.15 (0.33)
South Africa	0.19 (0.19)	0.23 (0.14)	0.25 (0.10)*	0.25 (0.09)*	0.07 (0.61)	0.10 (0.49)	0.12 (0.43)	0.12 (0.43)
Taiwan	0.73 (0.00)***	0.73 (0.00)***	0.73 (0.00)***	0.73 (0.00)***	0.66 (0.00)***	0.66 (0.00)***	0.65 (0.00)***	0.65 (0.00)***
Thailand	0.46 (0.02)**	0.47 (0.02)**	0.46 (0.02)**	0.46 (0.02)**	0.37 (0.05)**	0.38 (0.04)**	0.38 (0.04)**	0.38 (0.04)**
Turkey	0.51 (0.00)***	0.51 (0.00)***	0.51 (0.00)***	0.50 (0.00)***	0.45 (0.00)***	0.45 (0.00)***	0.45 (0.00)***	0.44 (0.00)***
Average	0.32	0.33	0.33	0.33	0.25	0.26	0.26	0.26

Table B.2: Performance of $mfFDR^+$ portfolios on basket of currencies. The table shows the annualized Sharpe ratios and mean returns (before and after transaction cost) and break-even point (bps) by implementing the $mfFDR^+$ on all strategies in all currencies to control the FDR at 10% (Panel A), 20% (Panel B), 30% (Panel C) and 40% (Panel D). The selected out-performing strategies then are combined by currency. The fund is allocated to trade on each of currencies having out-performing strategies with weighted by size of the selected out-performing trading rules. The first row presents the numbers of whole sample period while the rests are those of sub-periods.

Period	Excess Sharpe Ratio	Net Sharpe Ratio	Excess Return	Net Return	Break-even
		Panel A: FDR target of 10%			
Whole Period	0.98	0.88	4.36	3.91	63
1973-1980	1.40	1.31	4.89	4.58	76
1981-1990	2.11	2.01	8.28	7.89	141
1991-2000	0.81	0.72	4.93	4.38	84
2001-2010	0.62	0.47	2.53	1.89	34
2011-2020	0.33	0.25	1.28	0.99	10
		Panel B: FDR target of 20%			
Whole Period	1.06	0.95	3.80	3.40	60
1973-1980	1.45	1.35	4.47	4.18	69
1981-1990	2.08	1.97	7.30	6.93	128
1991-2000	0.92	0.81	4.15	3.64	72
2001-2010	0.69	0.53	2.37	1.82	34
2011-2020	0.29	0.21	0.88	0.63	14
		Panel C: FDR target of 30%			
Whole Period	1.11	0.99	3.54	3.15	57
1973-1980	1.49	1.39	4.29	4.01	76
1981-1990	2.02	1.91	6.69	6.33	118
1991-2000	1.00	0.87	3.77	3.28	66
2001-2010	0.72	0.56	2.28	1.75	34
2011-2020	0.32	0.23	0.83	0.60	8
		Panel D: FDR target of 40%			
Whole Period	1.12	1.00	3.34	2.97	54
1973-1980	1.47	1.37	4.11	3.84	75
1981-1990	1.98	1.87	6.30	5.94	112
1991-2000	1.06	0.92	3.57	3.10	63
2001-2010	0.73	0.57	2.20	1.70	34
2011-2020	0.29	0.20	0.72	0.50	7

B.5 Performance of $mfFDR^+$ portfolios with use of mean excess return as the testing performance metric

As a robustness check, we repeat all experiments presented in main manuscript with use of mean excess return as performance metric in hypothesis testing (ϕ). Table B.3 present the OOS performance of the $mfFDR$-based portfolios, both before and after transaction cost, when implementing the method on individual currencies to control FDR at targets of 10%, 20%, 30% and 40%.

Table B.3: Performance of $mfFDR$ based portfolios on individual currency with varying FDR target when mean return is used as the testing performance metric. The table shows annualized Sharpe ratios of the $mfFDR$ based portfolio with FDR target $\tau = \{0.1, 0.2, 0.3, 0.4\}$ based on portfolios' returns before (left side) and after transaction cost (right side). The final row shows the average Sharpe ratio across 30 portfolios corresponding to the 30 currencies. The numbers in parentheses are the corresponding $p-$values. "*", "**" and "***" respectively indicate statistical significance at levels of $10\%, 5\%$ and 1%.

Countries	Before transaction cost				After transaction cost			
	$\tau = 0.1$	$\tau = 0.2$	$\tau = 0.3$	$\tau = 0.4$	$\tau = 0.1$	$\tau = 0.2$	$\tau = 0.3$	$\tau = 0.4$
Australia	0.20 (0.11)	0.20 (0.10)*	0.20 (0.11)	0.19 (0.11)	0.16 (0.19)	0.16 (0.20)	0.15 (0.21)	0.15 (0.22)
Canada	0.19 (0.15)	0.18 (0.17)	0.18 (0.17)	0.18 (0.17)	0.13 (0.31)	0.12 (0.39)	0.12 (0.41)	0.12 (0.38)
Germany/E.U.	0.45 (0.00)***	0.46 (0.00)***	0.45 (0.00)***	0.45 (0.00)***	0.42 (0.00)***	0.42 (0.00)***	0.42 (0.00)***	0.41 (0.01)***
Japan	0.46 (0.00)***	0.45 (0.00)***	0.45 (0.00)***	0.45 (0.00)***	0.43 (0.00)***	0.41 (0.01)***	0.42 (0.01)***	0.41 (0.01)***
New Zealand	0.33 (0.01)***	0.35 (0.00)***	0.35 (0.00)***	0.36 (0.00)***	0.27 (0.03)**	0.29 (0.02)**	0.30 (0.02)**	0.30 (0.02)**
Norway	0.21 (0.09)*	0.19 (0.14)	0.18 (0.14)	0.18 (0.14)	0.16 (0.20)	0.14 (0.27)	0.13 (0.29)	0.13 (0.29)
Sweden	0.37 (0.01)***	0.40 (0.01)***	0.40 (0.01)***	0.40 (0.01)***	0.32 (0.01)***	0.35 (0.01)***	0.35 (0.01)***	0.35 (0.01)***
Switzerland	0.30 (0.04)**	0.30 (0.02)**	0.30 (0.03)**	0.29 (0.03)**	0.26 (0.06)*	0.26 (0.06)*	0.26 (0.07)*	0.26 (0.07)*
U.K.	0.32 (0.02)**	0.30 (0.04)**	0.30 (0.05)**	0.30 (0.05)**	0.28 (0.05)**	0.27 (0.06)*	0.26 (0.07)*	0.26 (0.07)*
Argentina	0.22 (0.29)	0.24 (0.25)	0.24 (0.23)	0.24 (0.23)	0.18 (0.37)	0.19 (0.33)	0.20 (0.31)	0.20 (0.31)
Columbia	0.52 (0.01)***	0.57 (0.00)***	0.57 (0.00)***	0.57 (0.00)***	0.45 (0.03)**	0.50 (0.01)***	0.51 (0.01)***	0.50 (0.01)***
India	0.37 (0.02)**	0.35 (0.03)**	0.35 (0.03)**	0.35 (0.03)**	0.33 (0.05)**	0.31 (0.06)*	0.31 (0.06)*	0.30 (0.06)*
Indonesia	0.33 (0.02)**	0.34 (0.02)**	0.33 (0.02)**	0.33 (0.02)**	0.24 (0.10)*	0.24 (0.10)*	0.24 (0.10)*	0.24 (0.10)*
Israel	0.30 (0.04)**	0.31 (0.03)**	0.31 (0.03)**	0.31 (0.03)**	0.12 (0.38)	0.13 (0.32)	0.13 (0.33)	0.13 (0.33)
Philippines	0.66 (0.00)***	0.64 (0.00)***	0.63 (0.00)***	0.63 (0.00)***	0.49 (0.01)***	0.47 (0.01)***	0.46 (0.02)**	0.47 (0.02)**
Romania	0.16 (0.39)	0.20 (0.28)	0.20 (0.25)	0.20 (0.26)	0.05 (0.78)	0.08 (0.65)	0.09 (0.62)	0.09 (0.62)
Russia	0.40 (0.09)*	0.41 (0.08)*	0.42 (0.08)*	0.42 (0.08)*	0.38 (0.10)*	0.39 (0.08)*	0.40 (0.08)*	0.40 (0.08)*
Slovak	0.14 (0.42)	0.15 (0.40)	0.16 (0.37)	0.17 (0.36)	0.10 (0.58)	0.10 (0.56)	0.11 (0.54)	0.12 (0.50)
Brazil	0.35 (0.08)*	0.35 (0.09)*	0.34 (0.09)*	0.34 (0.09)*	0.32 (0.11)	0.31 (0.11)	0.31 (0.11)	0.31 (0.11)
Chile	0.44 (0.03)**	0.40 (0.04)**	0.40 (0.05)**	0.40 (0.05)**	0.36 (0.08)*	0.32 (0.11)	0.31 (0.12)	0.31 (0.12)
Czech	0.07 (0.64)	0.11 (0.51)	0.12 (0.47)	0.12 (0.46)	0.01 (0.88)	0.05 (0.71)	0.06 (0.67)	0.06 (0.64)
Hungary	-0.13 (0.49)	-0.12 (0.54)	-0.12 (0.55)	-0.12 (0.57)	-0.18 (0.35)	-0.18 (0.36)	-0.17 (0.37)	-0.17 (0.38)
Korea	0.32 (0.12)	0.30 (0.14)	0.29 (0.15)	0.29 (0.15)	0.26 (0.19)	0.24 (0.21)	0.24 (0.21)	0.24 (0.21)
Mexico	0.08 (0.55)	0.08 (0.55)	0.08 (0.57)	0.08 (0.56)	0.03 (0.78)	0.03 (0.80)	0.03 (0.81)	0.03 (0.80)
Poland	-0.05 (0.78)	-0.01 (0.98)	-0.01 (0.98)	0.00 (0.98)	-0.10 (0.59)	-0.05 (0.78)	-0.06 (0.78)	-0.05 (0.81)
Singapore	0.27 (0.05)**	0.32 (0.03)**	0.32 (0.03)**	0.32 (0.03)**	0.11 (0.54)	0.15 (0.35)	0.14 (0.37)	0.14 (0.37)
South Africa	0.17 (0.24)	0.21 (0.16)	0.23 (0.12)	0.24 (0.11)	0.05 (0.71)	0.09 (0.53)	0.11 (0.46)	0.11 (0.45)
Taiwan	0.78 (0.00)***	0.77 (0.00)***	0.77 (0.00)***	0.77 (0.00)***	0.70 (0.00)***	0.70 (0.00)***	0.69 (0.00)***	0.69 (0.00)***
Thailand	0.46 (0.02)**	0.47 (0.01)***	0.47 (0.01)***	0.47 (0.01)***	0.37 (0.05)**	0.39 (0.04)**	0.39 (0.04)**	0.38 (0.04)**
Turkey	0.48 (0.00)***	0.48 (0.00)***	0.48 (0.00)***	0.48 (0.00)***	0.42 (0.00)***	0.42 (0.00)***	0.42 (0.00)***	0.42 (0.00)***
Average	0.31	0.31	0.31	0.31	0.24	0.24	0.24	0.24

Table B.4 exhibits the results when implementing the method on all currencies together.

Table B.4: Performance of $mfFDR^+$ portfolios with use of mean return as the testing performance metric. The table shows the annualized Sharpe ratios and mean returns (before and after transaction cost) and break-even point (bps) by implementing the $mfFDR^+$ on all strategies in all currencies to control the FDR at 10% (Panel A), 20% (Panel B), 30% (Panel C) and 40% (Panel D). The selected out-performing strategies then are combined by currency. The fund is allocated to trade on each of currencies having out-performing strategies with weighted by size of the selected out-performing trading rules. The first row presents the numbers of whole sample period while the rests are those of sub-periods.

Period	Excess Sharpe Ratio	Net Sharpe Ratio	Excess Return	Net Return	Break-even
		Panel A: FDR target of 10%			
Whole Period	1.00	0.90	4.48	4.04	65
1973-1980	1.43	1.34	5.01	4.69	78
1981-1990	2.09	1.99	8.36	7.96	141
1991-2000	0.85	0.76	5.27	4.71	89
2001-2010	0.63	0.47	2.51	1.89	34
2011-2020	0.35	0.28	1.37	1.08	11
		Panel B: FDR target of 20%			
Whole Period	1.08	0.97	3.87	3.47	60
1973-1980	1.45	1.36	4.47	4.18	75
1981-1990	2.07	1.97	7.33	6.95	128
1991-2000	0.95	0.84	4.31	3.80	75
2001-2010	0.69	0.53	2.35	1.80	34
2011-2020	0.34	0.26	1.03	0.78	10
		Panel C: FDR target of 30%			
Whole Period	1.12	1	3.55	3.17	57
1973-1980	1.47	1.38	4.24	3.97	75
1981-1990	2.02	1.91	6.71	6.34	119
1991-2000	1.02	0.89	3.85	3.37	68
2001-2010	0.72	0.55	2.24	1.73	34
2011-2020	0.34	0.25	0.88	0.65	9
		Panel D: FDR target of 40%			
Whole Period	1.13	1	3.36	2.99	55
1973-1980	1.46	1.36	4.06	3.80	75
1981-1990	1.99	1.87	6.33	5.97	113
1991-2000	1.08	0.94	3.62	3.16	65
2001-2010	0.73	0.56	2.18	1.68	33
2011-2020	0.31	0.22	0.76	0.54	8

Appendix C

Appendix for chapter 3

C.1 The implementation of the StepM and StepSPA procedures

In this section we present the StepM and StepSPA procedures for multiple tests where the testing metric is the alpha of an asset pricing model. In line with frameworks of Romano and Wolf (2005) and Hansen (2005), Hsu *et al.* (2010) we first consider performance of n funds and conduct for each fund i a hypothesis test:

$$H_0 : \mu_i \leq 0 \qquad H_1 : \mu_i > 0 \tag{C.1}$$

where μ_i is the expectation of a time-varying metric $d_{i,t}$ which represents for the performance of the fund i relative to a benchmark at time t, $i = 1, \ldots, n$. The relative performance can be expressed in a form of $d_{i,t} = L_{0,t} - L_{i,t}$ where $L_{0,t}$ and $L_{i,t}$ are values of a loss function measured at time t of the benchmark and fund i, respectively. The choice of the loss function is flexible and depends on the goal of researchers.

For instance, in the framework of Hsu *et al.* (2010), where they assess the performance of trading rules, the $L_{i,t}$ is set to be -1 multiplied by the return of a trading rule i in excess of interest rate in day t. The benchmark strategy is one that earns the interest rate, whose $L_{0,t} = 0$ which is -1 multiplied by 0 (the benchmark return excess of the interest rate). The $d_{i,t}$ turns out to be the excess return of the strategy i and μ_i is its expected return.

In our framework, the testing performance is the alpha of a fund, the choice of the loss function will be different. Suppose we are testing the alpha of the model (3.4). We consider funds surviving through periods t from 1 to T and assess their performances in this period. As the adjusted return of a fund i is $r_{i,t} - \boldsymbol{F}_t\hat{\boldsymbol{\beta}}_i$ where $\hat{\boldsymbol{\beta}}_i$ is the estimate of $\boldsymbol{\beta}_i$, we define a loss function as $L_{i,t} = -[r_{i,t} - \boldsymbol{F_t}\hat{\boldsymbol{\beta}}_i]$. This is a natural setting as the smaller the loss $L_{i,t}$, the better the performance of the fund. The benchmark is the portfolio that invests on the considering risk factors and thus the adjusted return is 0 and its loss is $L_{0,t} = 0$. We have that the expectation of $L_{0,t} - L_{i,t} = r_{i,t} - \boldsymbol{F_t}\hat{\boldsymbol{\beta}}_i$ is the alpha of the fund i. Thus, in our framework

$$d_{i,t} = r_{i,t} - \boldsymbol{F_t}\hat{\boldsymbol{\beta}}_i = \hat{\alpha}_i + \hat{\varepsilon}_{i,t} \tag{C.2}$$

The StepM and StepSPA procedures rely on a bootstrapped resampling where the stationary bootstrap procedure of Politis and Romano (1994), with average length $1/q$ where $q \in (0, 1)$, is adopted.

First we estimate the variance $\hat{\omega}_i^2$ of $d_{i,t}$ as in Hansen (2005). More specifically, let $\bar{d}_i$ be the average of $d_{i,1}, \ldots, d_{i,T}$. Then,

$$\hat{\omega}_i^2 = \hat{\gamma}_{i,0} + 2\sum_{t=1}^{T-1} \kappa(T, t)\hat{\gamma}_{i,t} \tag{C.3}$$

where $\hat{\gamma}_{i,t} = 1/T. \sum_{k=1}^{T-t}(d_{i,k} - \bar{d}_i)(d_{i,k+t} - \bar{d}_i)$, $t = 0, \ldots, T - 1$ and $\kappa(T, t) = \frac{T-t}{T}(1 - q)^t + \frac{t}{T}(1 - q)^{T-t}$.

For each bootstrapped iteration b, a cross-sectionally and jointly bootstrapped return of each fund i and risk factors are generated. We calculate $\bar{d}_{i,b} = \sum_{t=1}^{T} d_{i,t,b}/T$ where $d_{i,t,b}$ is the relative performance obtained by implementing (C.2) on the bootstrapped return of the fund i.[1]

For the (studentized) StepM procedure, we calculate the variance $\hat{\omega}_{i,b}^2$ for the bootstrapped differential $d_{i,t,b}$ via using (C.3) where $d_{i,t}$ and $\bar{d}_i$ are replaced by $d_{i,t,b}$ and $\bar{d}_{i,b}$, respectively.

After B iterations, we establish a bootstrapped critical point for StepM, $c^*_{\tau,StepM}$, as the $(1-\tau)^{th}$ quantile of the bootstrapped population $\max_i[(\bar{d}_{i,b} - \bar{d}_i)/\hat{\omega}_{i,b}]$, $b = 1, \ldots, B$.[2]

[1]We bootstrap from fund returns and risk factors as adjusted return of a fund i is changing via both its return and the estimated $\boldsymbol{\beta}_i$ calculated with use of the return.

[2]Our implementation of the StepM procedure is similar to Hsu *et al.* (2010) which differs from

In the (studentized) StepSPA, we define $\hat{\mu}_i = \bar{d}_i.\mathbf{1}_{\{\sqrt{T}\bar{d}_i \leq -\hat{\omega}_i\sqrt{2\log\log T}\}}$ where $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function, $i = 1, \ldots, n$. and the bootstrapped critical point is defined as $c^*_\tau = \max\{c_\tau, 0\}$ where c_τ is the $(1-\tau)^{th}$ quantile of the bootstrapped population $\sqrt{T}\max_i[(\bar{d}_{i,b} - \bar{d}_i + \hat{\mu}_i)/\hat{\omega}_i]$, $b = 1, \ldots, B$.

Both the StepM and StepSPA are processed with the same steps but different in the statistics and bootstrapped critical point. In particular, the StepSPA procedure is as followings:

- Sort $\sqrt{T}\bar{d}_i/\hat{\omega}_i$ in a descending order.
- Select the top k funds if $\sqrt{T}\bar{d}_k/\hat{\omega}_k > c^*_\tau$. If there is no hypothesis rejected then stop the procedure. Otherwise,
    1. Remove the selected funds to obtain a sub-sample. Recalculate the critical c^*_τ with use of the sub-sample, denoted by c^s_τ.
    2. The top k' funds in the sub-sample with $\sqrt{T}\bar{d}_{k'}/\hat{\omega}_{k'} > c^s_\tau$ are selected. If there is no hypothesis rejected then stop the procedure. Otherwise go to step 3.
    3. Repeat the steps 1 and 2 above until there is no hypothesis that can be rejected.

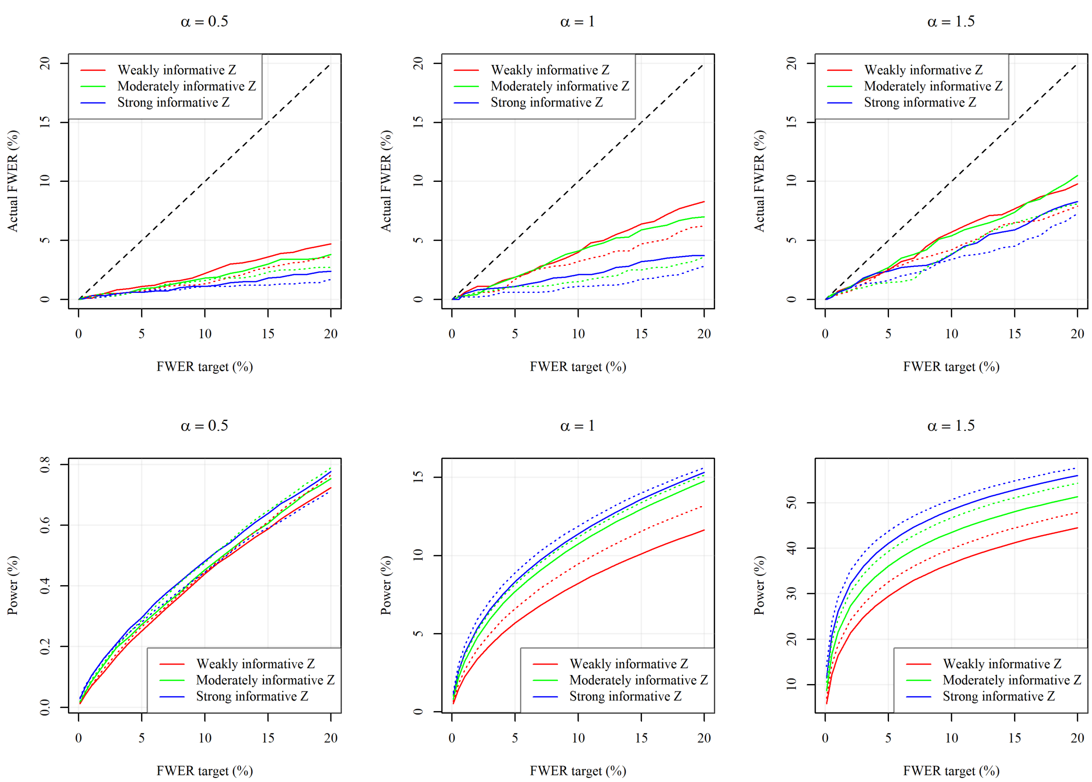
In the StepM procedure, the statistics $\sqrt{T}\bar{d}_k/\hat{\omega}_k$ and bootstrapped critical point c^*_τ are replaced by $\bar{d}_k/\hat{\omega}_k$ and $c^*_{\tau,StepM}$, both in the whole sample and sub-sample, respectively.

the original procedure of Romano and Wolf (2005) in three aspects. First, they use a circular block bootstrap while we use the stationary bootstrap. Second, they adopt data-driven algorithm in determining the block size of bootstrap, while we use a fixed value $q = 0.9$ following literature. Third, they use bootstrapped standard errors whereas we adopt HAC estimators as described in C.3. These differences might affect the finite sample performance of the StepM performance reported in our simulations results.

C.2 Additional simulation studies

To complement our simulation studies presented in Section 3.5 of Chapter 3, this section presents simulation results for a different setting of the triple (π^+, π_0, π^-). More specifically, we additionally consider the case $\pi^+ = 40\%, \pi_0 = 60\%$ and $\pi^- = 0\%$. Generally, the results are very similar to the case $\pi^+ = 20\%, \pi_0 = 60\%$ and $\pi^- = 20\%$ and for the interest of space, we present the main results in Figure C.1 where 18 variant settings are considered. These variants cover various signal settings in terms of α magnitude, the covariate signal strength and the correlation among the covariates.

Figure C.1: Performance of the $fwer^+$ under the alternative setting of out-performing funds proportion. The figure shows impact of signals, i.e., the magnitude of true non-zero alpha and informativeness of covariates, on the performance of the $fwer^+$ in terms of FWER control (top three sub-figuers) and power (bottom three figures). The simulated data are balanced panel of $n = 1000$ funds where each of them has $T = 36$ observations. The funds population consists of around 60%, 0% and 40% zero-alpha, under- and out-performing funds, respectively. The out-performing (under-performing) funds in population have alpha of α $(-\alpha)$ which varies in $\{0.5\%, 1.0\%, 1.5\%\}$. We consider three settings of the two covariates $\boldsymbol{Z} = (u, v)$ including weakly, moderately and strongly informative. The covariates can be independent (solid curves) or correlated with a coefficient of 0.5 (dotted curves).



C.3 Empirical results with restriction on AUM

In this section we present the performance of the $fwer^+$ portfolios when we restrict to considering only funds having at least 5 millions of AUM at the time we construct the portfolios. Table C.1 reports OOS performance of the $fwer^+$ portfolios with use of the R-square, AUM, and two PC1s of the moment and persistent groups as solely input covariate (Panels A to D) and four input covariates (Panel E). As benchmarks, we also report in Panel F the performance of the equally weighted portfolios which simply select all eligible funds in IS period (EW) and subset of those additionally having positive estimated IS alpha (EW^+). Generally, the performances of the portfolios remain similar to those with only requirement on the availability of the AUM reported in the Table 3.1 of Chapter 3.

Table C.1: OOS performance of $fwer^+$ portfolios under restrictions on AUM. Panels A to D of the table report OOS performance metrics of the $fwer^+$ portfolios with use each of R-square, AUM, and PC1s of moment and persistent group as the sole input covariate. The performance metrics include annualized alpha as well as its $t-$statistic and $p-$value, excess return, Sharpe ratio and summary on number of out-performing funds detected by the $fwer^+$ procedure. Panel E reports these metrics of the $fwer^+$ portfolio with use of all four mentioned covariates whereas panel F the performance metrics of the equally weighted (EW) and equally weighted plus (EW^+) portfolios.

τ (%)	Alpha (%)	$t-$statistic	$p-$value	Return (%)	Sharpe Ratio	Number of detected funds			
						Average	Min	Max	Std
Panel A: $fwer^+$ with use of R-square as the covariate									
0.01	4.94	7.2	0.00	5.35	2.09	16	4	26	7
1.00	4.39	7.3	0.00	4.88	1.58	34	9	54	13
5.00	4.38	6.8	0.00	4.90	1.56	47	15	85	19
Panel B: $fwer^+$ with AUM as the covariate									
0.01	4.80	6.8	0.00	5.23	2.00	16	4	26	7
1.00	4.37	7.1	0.00	4.86	1.55	35	9	61	15
5.00	4.19	6.5	0.00	4.75	1.46	49	15	94	23
Panel C: $fwer^+$ with use of PC1 of moment group as the covariate									
0.01	4.78	6.7	0.00	5.15	1.96	17	4	30	7
1.00	4.54	7.7	0.00	5.03	1.66	35	9	66	15
5.00	4.30	6.7	0.00	4.86	1.53	49	15	98	22
Panel D: $fwer^+$ with use of PC1 of persistent group as the covariate									
0.01	5.21	8.1	0.00	5.56	2.29	15	4	25	6
1.00	4.54	7.3	0.00	5.03	1.65	33	9	53	13
5.00	4.37	6.9	0.00	4.83	1.57	46	14	82	19
Panel E: $fwer^+$ with use of the R-square, AUM and PC1s of the two groups as the covariates									
0.01	5.00	7.6	0.00	5.32	2.16	17	4	30	7
1.00	4.43	7.0	0.00	4.93	1.55	38	10	69	16
5.00	4.38	6.5	0.00	4.97	1.48	53	14	106	25
Panel F: equally weighted portfolios									
EW	2.50	2.8	0.01	4.58	0.70	1018	336	1533	353
EW^+	2.91	3.5	0.00	4.68	0.78	739	266	1389	319

C.4 $fwer^+$ portfolios with use of individual covariates

In this section, we present the OOS performance metrics of the $fwer^+$ portfolios with use of underlying individual covariates. Particularly, we present in Table C.2 and C.3 the metrics corresponding to individual covariates in the persistent and moment groups, respectively.

Table C.2: OOS performance of the $fwer^+$ portfolios with use of persistent covariates. The table reports the OOS performance of the $fwer^+$ portfolios with use of individual covariates in persistent group. Panels A to I report OOS annualized alpha as well as its $t-$statistic and $p-$value, excess return and Sharpe ratios and summary on size of the $fwer^+$ portfolios with use of the sole covariate at various given FWER target τ.

τ (%)	Alpha (%)	$t-$statistic	$p-$value	Return (%)	Sharpe Ratio	Portfolio Size			
						Average	Min	Max	Std
Panel A: $fwer^+$ with use as the covariate the ACF1 of past 12-month excess returns									
0.01	5.38	8.1	0.00	5.81	2.31	15	4	25	6
1.00	4.77	7.6	0.00	5.34	1.73	33	10	56	13
5.00	4.54	6.9	0.00	5.09	1.63	46	11	84	19
Panel B: $fwer^+$ with use as the covariate the ACF1 of past 24-month excess returns									
0.01	5.18	8.0	0.00	5.59	2.26	15	4	25	6
1.00	4.33	6.9	0.00	4.84	1.55	33	10	56	14
5.00	4.37	6.9	0.00	4.83	1.60	46	15	82	19
Panel C: $fwer^+$ with use as the covariate the ACF1 of past 36-month excess returns									
0.01	5.24	8.1	0.00	5.64	2.26	15	4	25	6
1.00	4.44	7.2	0.00	4.94	1.59	34	10	56	13
5.00	4.45	7.0	0.00	4.91	1.61	46	15	87	19
Panel D: $fwer^+$ with use as the covariate the ACF2 of past 12-month excess returns									
0.01	4.89	7.0	0.00	5.36	2.07	15	4	26	6
1.00	4.40	7.3	0.00	4.90	1.60	33	8	54	14
5.00	4.43	6.9	0.00	4.94	1.60	46	14	88	19
Panel E: $fwer^+$ with use as the covariate the ACF2 of past 24-month excess returns									
0.01	4.91	7.2	0.00	5.38	2.11	15	4	25	6
1.00	4.48	7.3	0.00	5.01	1.62	33	10	55	14
5.00	4.41	6.8	0.00	4.94	1.58	46	14	87	20
Panel F: $fwer^+$ with use as the covariate the ACF2 of past 36-month excess returns									
0.01	4.93	7.2	0.00	5.37	2.06	15	3	25	6
1.00	4.33	6.6	0.00	4.87	1.53	34	10	55	13
5.00	4.22	6.4	0.00	4.73	1.50	47	14	88	19
Panel G: $fwer^+$ with use as the covariate the ACF3 of past 24-month excess returns									
0.01	5.26	8.3	0.00	5.64	2.35	15	4	23	6
1.00	4.59	7.4	0.00	5.09	1.66	33	10	54	13
5.00	4.28	6.7	0.00	4.76	1.52	46	15	87	20
Panel H: $fwer^+$ with use as the covariate the ACF3 of past 12-month excess returns									
0.01	5.24	8.4	0.00	5.62	2.39	15	4	25	6
1.00	4.42	7.1	0.00	4.93	1.61	33	10	55	13
5.00	4.15	6.6	0.00	4.60	1.47	46	15	83	19
Panel I: $fwer^+$ with use as the covariate the ACF3 of past 36-month excess returns									
0.01	4.96	7.4	0.00	5.42	2.15	15	4	25	7
1.00	4.56	7.5	0.00	5.07	1.65	33	10	55	13
5.00	4.22	6.6	0.00	4.68	1.51	46	14	87	20

Table C.3: OOS performance of the $fwer^+$ portfolios with use of moment covariates. The table reports the OOS performance of the $fwer^+$ portfolios with use of individual covariates in moment group. Panels A to I report OOS annualized alpha as well as its $t-$statistic and $p-$value, excess return, Sharpe ratio and a summary on the size of the $fwer^+$ portfolios with use of the each covariate in moment group.

τ (%)	Alpha (%)	$t-$statistic	$p-$value	Return (%)	Sharpe Ratio	Portfolio Size			
						Average	Min	Max	Std
Panel A: $fwer^+$ with use as the covariate the kurtosis of past 12-month excess returns									
0.01	4.94	7.2	0.00	5.38	2.11	16	4	25	6
1.00	4.37	7.2	0.00	4.87	1.63	32	9	55	14
5.00	4.39	7.0	0.00	4.88	1.58	45	13	86	20
Panel B: $fwer^+$ with use as the covariate the kurtosis of past 24-month excess returns									
0.01	4.89	7.0	0.00	5.35	2.06	15	4	26	6
1.00	4.50	7.8	0.00	5.00	1.68	33	9	55	13
5.00	4.18	6.4	0.00	4.70	1.48	45	14	86	20
Panel C: $fwer^+$ with use as the covariate the kurtosis of past 36-month excess returns									
0.01	4.94	7.2	0.00	5.40	2.10	15	4	25	6
1.00	4.35	7.3	0.00	4.85	1.60	33	10	55	13
5.00	4.17	6.5	0.00	4.67	1.48	45	13	87	20
Panel D: $fwer^+$ with use as the covariate the skewness of past 12-month excess returns									
0.01	4.84	6.9	0.00	5.26	2.05	15	4	25	7
1.00	4.38	7.2	0.00	4.87	1.58	33	9	57	13
5.00	4.30	6.8	0.00	4.78	1.56	45	13	86	20
Panel E: $fwer^+$ with use as the covariate the skewness of past 24-month excess returns									
0.01	5.17	7.6	0.00	5.57	2.26	15	4	26	7
1.00	4.25	6.9	0.00	4.69	1.56	34	10	56	13
5.00	4.33	6.7	0.00	4.83	1.56	46	14	88	20
Panel F: $fwer^+$ with use as the covariate the skewness of past 36-month excess returns									
0.01	5.43	7.9	0.00	5.85	2.35	15	3	26	7
1.00	4.29	7.1	0.00	4.75	1.57	32	10	55	14
5.00	4.25	6.3	0.00	4.75	1.49	46	14	87	21
Panel G: $fwer^+$ with use as the covariate the variance of past 12-month excess returns									
0.01	4.90	7.0	0.00	5.31	2.03	16	4	29	7
1.00	4.51	7.5	0.00	5.01	1.65	34	10	64	15
5.00	4.15	6.2	0.00	4.68	1.49	48	13	90	22
Panel H: $fwer^+$ with use as the covariate the variance of past 24-month excess returns									
0.01	4.86	7.1	0.00	5.26	2.04	16	4	30	7
1.00	4.49	7.4	0.00	4.99	1.63	34	10	65	15
5.00	4.24	6.5	0.00	4.79	1.51	49	14	92	22
Panel I: $fwer^+$ with use as the covariate the variance of past 36-month excess returns									
0.01	4.91	7.3	0.00	5.34	2.12	16	4	27	7
1.00	4.45	7.2	0.00	4.96	1.60	35	10	65	15
5.00	4.36	6.8	0.00	4.90	1.58	49	14	97	23

Generally, we see that the performances of the $fwer^+$ portfolios with use of different individual covariates of the same sub-groups but differing in estimation windows tend to be similar. This supports the use of a representative covariate such as PC1 for each group as presented in Chapter 3.

C.5 Alternative choices of in-sample horizons

Literature in hedge fund performance construct portfolios based on assessing funds' performance over a short past performance, i.e, a short IS horizon. The most common choices are 24 and 36 months. Beside, a choice of 48 month is also considered. As robustness checks, we repeat the discussed experiments with the choices of 24- and 48-month IS horizons and present the OOS performance in Tables C.4 and C.5, respectively.

In both cases, our conclusions on both power and performance remain as in the 36-month IS case. On average, the $fwer^+$ gain higher power for a longer IS period. We also observe that the $fwer^+$ portfolios with a longer IS period tend to gain higher Sharpe ratio but lower alpha. Nevertheless, the differences are small.

Table C.4: OOS performance of the $fwer^+$ portfolios with use of 24-month IS periods. Panels A to D of the table report OOS performance metrics of the $fwer^+$ portfolios with use each of R-square, AUM, and PC1s of moment and persistent group as the sole input covariate. The performance metrics include annualized alpha as well as its $t-$statistic and $p-$value, excess return and Sharpe ratio of the portfolios and summary on number of funds selected by the $fwer^+$. Panel E reports these metrics of the $fwer^+$ portfolio with use of all four mentioned covariates as inputs whereas panel F the performance metrics of the equally weighted (EW) and equally weighted plus (EW^+) portfolios.

τ (%)	Alpha (%)	$t-$statistic	$p-$value	Return (%)	Sharpe Ratio	Number of selected funds			
						Average	Min	Max	Std
Panel A: $fwer^+$ with use of R-square as the covariate									
0.01	5.15	6.4	0.00	5.40	1.70	9	1	18	5
1.00	4.67	5.7	0.00	5.12	1.56	21	7	47	10
5.00	3.80	4.8	0.00	4.32	1.24	30	11	72	15
Panel B: $fwer^+$ with AUM as the covariate									
0.01	5.67	6.1	0.00	5.99	1.68	10	1	20	5
1.00	4.40	5.1	0.00	4.85	1.44	22	7	52	11
5.00	3.78	4.5	0.00	4.31	1.22	31	12	79	17
Panel C: $fwer^+$ with use of PC1 of moment group as the covariate									
0.01	6.03	7.0	0.00	6.39	1.91	10	1	18	5
1.00	4.82	6.1	0.00	5.34	1.68	22	7	52	11
5.00	3.92	4.9	0.00	4.46	1.33	31	10	75	17
Panel D: $fwer^+$ with use of PC1 of persistent group as the covariate									
0.01	5.71	6.2	0.00	6.04	1.66	9	1	18	5
1.00	4.68	5.6	0.00	5.10	1.54	21	7	47	10
5.00	4.06	5.1	0.00	4.53	1.39	30	11	72	15
Panel E: $fwer^+$ with use of the R-square, AUM and PC1s of the two groups as the covariates									
0.01	5.57	7.2	0.00	5.84	1.95	10	2	20	5
1.00	4.59	5.6	0.00	5.08	1.56	24	7	56	12
5.00	3.85	4.9	0.00	4.40	1.28	34	12	79	17
Panel F: equally weighted portfolios									
EW	2.85	3.2	0.00	5.00	0.78	1098	500	1570	335
EW^+	3.24	4.0	0.00	5.13	0.87	783	323	1418	313

Table C.5: OOS performance of $fwer^+$ portfolios with use of 48-month IS periods. Panels A to D of the table report OOS performance metrics of the $fwer^+$ portfolios with use each of R-square, AUM, and PC1s of moment and persistent group as the sole covariate. The performance metrics include annualized alpha as well as its $t-$statistic and $p-$value, excess return and Sharpe ratio of the portfolios and summary on number of funds selected by the $fwer^+$. Panel E reports these metrics of the $fwer^+$ portfolio with use of all four mentioned covariates as inputs whereas panel F the performance metrics of the equally weighted (EW) and equally weighted plus (EW^+) portfolios.

τ (%)	Alpha (%)	$t-$statistic	$p-$value	Return (%)	Sharpe Ratio	Number of selected funds			
						Average	Min	Max	Std
Panel A: $fwer^+$ with use of R-square as the covariate									
0.01	4.82	7.3	0.00	5.18	2.34	18	3	33	8
1.00	4.19	6.2	0.00	4.69	1.72	38	13	74	19
5.00	3.85	4.9	0.00	4.44	1.46	52	18	100	27
Panel B: $fwer^+$ with AUM as the covariate									
0.01	4.76	8.2	0.00	5.13	2.48	19	3	35	8
1.00	4.18	6.2	0.00	4.70	1.69	41	12	80	21
5.00	3.84	5.0	0.00	4.45	1.40	56	17	111	31
Panel C: $fwer^+$ with use of PC1 of moment group as the covariate									
0.01	4.73	7.0	0.00	5.09	2.12	19	2	34	8
1.00	4.34	6.0	0.00	4.86	1.68	38	4	77	20
5.00	4.11	5.4	0.00	4.73	1.56	54	8	105	30
Panel D: $fwer^+$ with use of PC1 of persistent group as the covariate									
0.01	4.92	7.6	0.00	5.29	2.38	18	3	34	8
1.00	4.20	6.1	0.00	4.71	1.70	37	13	70	19
5.00	3.99	5.1	0.00	4.59	1.50	52	15	100	27
Panel E: $fwer^+$ with use of the R-square, AUM and PC1s of the two groups as the covariates									
0.01	4.55	7.3	0.00	4.86	2.10	20	1	37	9
1.00	4.43	6.7	0.00	4.98	1.81	43	2	82	23
5.00	4.04	5.0	0.00	4.70	1.46	59	2	118	34
Panel F: equally weighted portfolios									
EW	2.77	3.1	0.00	4.94	0.76	898	337	1271	283
EW^+	3.09	3.7	0.00	5.00	0.84	672	263	1153	263

C.6 Performance of $fwer^+$ portfolios with use of simple $p-$values

As mentioned in Chapter 3, there might be concern in use of $p-$values with HAC correction given the short IS time series. In this section, we present OOS performance of the $fwer^+$ portfolios with use of simple $p-$values, i.e., the p-values calculated without using the HAC correction. The results are shown in Table C.6. The performance metrics gained by the portfolios tend to be higher than those gained by the portfolios constructed with use of HAC correction. As also shown in Table C.7, when the new covariates obtained by the four famous machine learning techniques are used, the $fwer^+$ portfolios perform impressively with Sharpe ratio of more than 2 at all considering FWER targets and more than 3 at the lowest considering FWER target. In Table C.8 we report the performance of the single-fund portfolio under different factor models. We see that there are many more months where the portfolios are empty and performance are lower than those observed with use of the HAC correction.

Table C.6: OOS performance of $fwer^+$ portfolios with use of simple $p-$values. Panels A to D of the table report OOS performance metrics of the $fwer^+$ portfolios with use each of R-square, AUM, and PC1s of moment and persistent group as the sole covariate. The performance metrics include annualized alpha as well as its $t-$statistic and $p-$value, excess return and Sharpe ratio and summary on number of out-performing funds detected by the $fwer^+$. Panel E reports these metrics of the $fwer^+$ portfolio with use of all four mentioned covariates as inputs whereas panel F the performance metrics of the equally weighted (EW) and equally weighted plus (EW^+) portfolios.

τ (%)	Alpha (%)	$t-$statistic	$p-$value	Return (%)	Sharpe Ratio	Number of detected funds			
						Average	Min	Max	Std
Panel A: $fwer^+$ with use of R-square as the covariate									
0.01	5.45	9.2	0.00	5.71	2.60	13	2	24	6
1.00	4.90	7.7	0.00	5.40	2.09	28	9	57	11
5.00	4.56	7.0	0.00	5.11	1.74	39	13	84	18
Panel B: $fwer^+$ with AUM as the covariate									
0.01	5.32	8.3	0.00	5.56	2.29	14	2	26	7
1.00	4.70	7.5	0.00	5.20	1.98	29	9	60	13
5.00	4.40	6.2	0.00	5.03	1.59	41	10	97	22
Panel C: $fwer^+$ with use of PC1 of moment group as the covariate									
0.01	5.37	8.7	0.00	5.62	2.45	14	2	30	7
1.00	4.51	6.6	0.00	4.99	1.91	29	9	69	13
5.00	4.81	6.8	0.00	5.45	1.65	41	11	108	22
Panel D: $fwer^+$ with use of PC1 of persistent group as the covariate									
0.01	5.39	8.6	0.00	5.63	2.42	13	2	23	6
1.00	4.78	7.2	0.00	5.28	1.96	27	9	59	12
5.00	4.69	7.0	0.00	5.26	1.83	37	10	83	17
Panel E: $fwer^+$ with use of the R-square, AUM and PC1s of the two groups as the covariates									
0.01	5.55	8.7	0.00	5.84	2.48	15	3	30	7
1.00	5.10	7.6	0.00	5.70	1.81	32	10	72	15
5.00	4.99	8.0	0.00	5.69	1.62	47	13	115	25
Panel F: equally weighted portfolios									
EW	2.58	2.9	0.00	4.65	0.72	1067	350	1570	361
EW^+	3.00	3.7	0.00	4.77	0.80	761	273	1418	324

Table C.7: OOS performance of $fwer^+$ portfolios with use of machine learning based covariates and simple $p-$values. Panel A (B, C and D) reports OOS annualized alpha as well as its $t-$statistic and $p-$value, excess return and Sharpe ratios and summary on the size of the $fwer^+$ portfolios with use of funds' future return predicted by LASSO (GB, RF and DNN) model at given FWER targets τ.

τ (%)	Alpha (%)	$t-$statistic	$p-$value	Return (%)	Sharpe Ratio	Number of detected funds			
						Average	Min	Max	Std
Panel A: $fwer^+$ with use of future return predicted by the LASSO model as the sole covariate									
0.01	5.96	11.8	0.00	6.18	3.33	13	2	22	6
1.00	5.49	9.8	0.00	5.98	2.58	25	8	49	10
5.00	5.16	8.6	0.00	5.76	2.24	35	10	81	16
Panel B: $fwer^+$ with use of future return predicted by the GB model as the sole covariate									
0.01	5.75	11.4	0.00	6.01	3.22	13	2	23	7
1.00	5.01	7.9	0.00	5.54	2.39	26	9	56	12
5.00	4.93	7.6	0.00	5.55	2.13	36	10	80	19
Panel C: $fwer^+$ with use of future return predicted by the RF model as the sole covariate									
0.01	6.61	9.0	0.00	6.81	2.46	12	2	22	6
1.00	5.23	8.2	0.00	5.75	2.38	25	9	50	10
5.00	5.12	7.8	0.00	5.82	2.01	35	10	81	17
Panel D: $fwer^+$ with use of future return predicted by the DNN model as the sole covariate									
0.01	5.91	11.7	0.00	6.20	3.28	13	2	23	7
1.00	5.08	8.3	0.00	5.61	2.39	26	5	56	12
5.00	5.16	8.2	0.00	5.80	2.20	36	9	77	18

Table C.8: Performance of the single-fund portfolios with use of simple p-value. The table reports the OOS performance of the portfolio that consists of the fund performed best in IS period among those selected by the $fwer^+$.

Model	Alpha (%)	$t-$statistic	$p-$value	Return (%)	Sharpe Ratio	Empty rate (%)
Panel A: any underlying covariates						
4 factors	4.90	8.8	0.00	5.12	2.69	6/288
6 factors	4.81	9.4	0.00	4.89	2.57	17/288
7 factors	4.90	9.9	0.00	4.92	2.59	17/288
9 factors	4.81	9.2	0.00	4.92	2.59	17/288