From surveys and theory to statistical methods. Which predictors can we trust? and when?

Marc Van Uffelen[1]

[1]USI Lugano and SFI

Abstract

Given the largely efficient nature of the stock market, forecasting its returns remains inherently challenging. However, prior literature demonstrates that such predictions are not entirely infeasible. In this study, I examine the temporal performance of over 20 market return predictors over the period from December 2001 to December 2018. Among these predictors, some are based on surveys, while some are based on economical findings. Many of them are constructed using statistical methods, including learning techniques. Notably, during the Global Financial Crisis, none of the predictors exhibit significant predictive power. However, outside of this period, certain predictors, including some based on survey data, demonstrate a higher degree of potential informativeness.

Keywords: Asset pricing, Quantitative Methods, Machine Learning, Predictions.

This version: November 2024

Marc Van Uffelen is at USI Lugano and with the SFI. The author thanks audiences at various conferences and seminars for their comments and suggestions. Email: marc.jean-paul.van.uffelen@usi.ch.

I. Introduction

The ability to predict stock market returns has long captivated both academic researchers and practitioners. Fama [1970], Malkiel [2003] and many more studies show that financial markets are mostly efficient. On the other hand, some studies, such as Fama and French [1989] and Goyal et al. [2024], suggest that return predictability is possible under certain conditions. I seek to contribute to this ongoing discussion by highlighting the importance of investigating the performance of various market return predictors across different economic periods, including the Global Financial Crisis (GFC).

I analyze and compare over 20 different predictors, over a period spanning from December 2001 to December 2018. Among these predictors, some are based on surveys, while some are based on economic findings. Surveys provide insights into investor sentiment, with several studies, such as Greenwood and Greenwood and Shleifer [2013] and Nagel and Xu [2021], examining their predictive potential. Additionally, I utilize factor models, including the well-known Capital Asset Pricing Model (CAPM) by Sharpe [1964] and Lintner [1965], as well as the Fama and French [1993] three-factor model, which are widely recognized for their ability to characterize stock market behavior. I construct additional predictors from these models to evaluate their predictive capabilities.

Furthermore, a variety of statistical techniques, including Lasso regression, as discussed in studies such as Freyberger et al. [2020] and Lee et al. [2022], and Conditional Mean Embedding (CME), examined in works such as Li et al. [2022] and Tamás and Csáji [2024], are applied to predictive variables to derive new predictors and assess their informativeness regarding future market movements. This approach builds on prior work, such as that of Campbell and Thompson [2008] and Goyal et al. [2024], but introduces the novel aspect of explicitly accounting for temporal variations in predictor performance.

There are several existing studies that compare market return predictors or their equivalents. For instance, Bianchi et al. [2024] compare the performance of market risk premium predictors derived from theoretical frameworks, statistical methods, or a combination of both. The unique contribution of this study lies in its focus on the temporal variation in predictor performance, rather than evaluating static performance. This analysis is conducted within an online learning framework, allowing for a dynamic understanding of how predictor effectiveness evolves over time.

Although such studies remain relatively uncommon, some financial research has begun to apply online learning techniques to various topics. For example, Mohri [2018] employs online learning to explore statistical arbitrage strategies, while Yuan et al. [2024] apply it in the context of high-frequency trading analysis. To the best of my knowledge, this study is the first to utilize online learning for the analysis and comparison of market return predictors. Accordingly, I elucidate the advantages of adopting this framework, particularly in capturing the dynamic and evolving nature of predictive abilities over time.

Survey-based predictors, particularly the CFO survey, which is conducted by Duke University's Fuqua School of Business, demonstrate considerable informativeness during stable periods, suggesting that the insight of market participants is indeed informative about future market developments. However, during times of extreme market instability, such as the GFC, even expert-based predictions struggle to provide significant insight, underscoring the challenge of capturing and understanding rapid changes through human judgment alone. In fact, non of the predictors in this study demonstrate any kind of informativeness during the GFC.

Factor model-based predictors appear to generally perform poorly across different market conditions. Even during stable periods, these models fail to consistently provide significant predictive power, which suggests that their explanatory power in cross-sectional asset pricing does not translate well into time-series forecasting for aggregate market returns.

On the other hand, statistical methods seem to offer more promising results, particularly those that help mitigate overfitting, such as Lasso and Ridge regressions, as well as CME. The traditional OLS regression, in contrast, performs significantly worse, especially following major market shocks, such as the collapse of Lehman Brothers. This underperformance reinforces the critical need for regularization and more sophisticated statistical approaches that are less prone to overfitting, especially when dealing with complex financial data.

The rest of this paper is structured as follows. Section II presents the key definitions and concepts of this paper. Section III shows which predictors I analyze and compare, and which methods I use for the comparison. In section IV, I present and discuss the main findings. Section V concludes.

II. Market Return Predictors

I begin by offering a precise definition of the term "predictor," followed by an introduction to the various types of predictors that are examined and compared in this study.

A. Defining predictors

Let R_s be the simple monthly market return of the month s. The goal in this paper is to investigate how to best determine an estimation of

$$\mathbb{E}_t\left[R_{t+1}\right] \tag{1}$$

where t corresponds to the current time, and $t+1$ corresponds to the next month. 1 is supposed to correspond to the optimal way to predict the next month's market return, based on all available information at time t. Using all information available at time t is not practically feasible. Therefore, rather than trying to find (1) directly, a better approach would seem to be to select some practically obtainable information set $I_{P,t}$ and determine

$$P_t := \mathbb{E}\left[R_{t+1}|I_{P,t}\right] \tag{2}$$

where P_t would be defined as predictor, with $I_{P,t}$ as the information used at time t, in order to determine P_t. Unfortunately, the optimal way to use $I_{P,t}$ is usually unclear. Therefore, a *Predictor* P_t is going to be defined as a fixed way to estimate (2), where the choice of $I_{P,t}$ is fixed. In other words, a predictor P_t is defined as an estimator of the conditional expectation of the simple monthly market return at time $t+1$, which is constructed based on a usable set of information available at time t.

B. Surveys

Adam et al. [2021] and many other studies use surveys to try to better understand investors' insight, or lack thereof, about future states of the market. In those surveys, individual and/or professional investors are typically asked to make a multitude of personal predictions, some of which are connected to how the market will develop. Such surveys include the CFO, UBS, Schiller and Livingston surveys. Some of these surveys ask participants to predict future market levels, while others ask for returns predictions. Each of these predictions can ultimately be converted into a predictor, as in (2). Therefore surveys are also investigated as potential market return predictors. How I do so specifically is detailed in section III.

C. *Factor model inspired predictors*

The literature provides a multitude of models, which can either be used to predict or explain market returns or risk premia, or to predict or explain the cross-section of individual stock returns, including the famous capital asset pricing model (CAPM) by Sharpe [1964] and Lintner [1965]. As the market level can be considered as a special type of stock price, one can try to use such models to predict monthly market returns as well. In the case of factor models, such as the one provided by Fama and French [1993], one can try to estimate (1) through the following idea

$$P_t = \beta_t^\top F_t \approx \mathbb{E}_t \left[R_{t+1} \right] \tag{3}$$

where F_t corresponds to the model's estimated factors at time t, and β_t represents the corresponding estimated factor loadings. Such predictors are also included in this study, in order to gain a better understanding of how much insight these models can provide in the main quest of predicting returns.

D. *Statistical methods and predictive variables*

Over the years, especially in recent times, academic researchers and market participants have been trying to predict different types of financial values by looking at the information within a wide range of sources, using tools provided by econometricians or other data analysis experts. Such tools include the traditional regression, or some more advanced methods, such as the famous LASSO regression, the Ridge regression, or even kernel methods. In this study, such tools are referred to as statistical methods, and I use some of them to construct market return predictors. The main idea behind this type of construction is straight forward; I look at a wide range of information sources, such as variables considered in Goyal et al. [2024], and apply such statistical methods unto these sources. For example, let us assume that certain macro-economically relevant variables A and B contain relevant information about future market returns. One can then try to estimate the following.

$$\mathbb{E} \left[R_{t+1} | \sigma \left(A_t, B_t \right) \right] \tag{4}$$

where $\sigma(A_t, B_t)$ corresponds to the entire information provided by A and B until time t. For a predictor P_t which estimates (4) through a fixed statistical method, $I_{P,t}$ is then equal to $\sigma(A_t, B_t)$. What will be considered as *Predictive Variable* is any variable that could have been included in, or added to $\sigma(A_t, B_t)$. Generally speaking, for the construction of some of the predictors that this paper investigates, groups of predictive variables are paired with statistical methods. As a predictive variable can be any kind of variable that contains information about monthly market returns, predictors can also be seen as predictive variables, but a generic predictive variable will not always fit the definition of a predictor, as defined above.

III. Empirical Measurements

Now that predictors are properly defined, an important question arises. Which predictor is a better approximation of (1), and how can that be determined? I cannot explore the performance of every single possible market return predictor, as such choices are infinite. This paper will focus on three main types of predictors: survey implied predictors, factor model derived predictors, and predictors constructed through a variety of predictive variables and statistical methods. This section describes what data I use, which predictors exactly I look into, and different methods I use to compare the performance of the different considered predictors.

A. Data description

In my analysis, I consider the aggregate S&P 500 composite index as a reasonable proxy for the aggregate market portfolio. In order to use multiple surveys without sacrificing too many observations, my predictor sample spans the period from December 2001 to December 2018, while my predictive variable sample spans the period from December 1996 to December 2018, as the statistical predictors I use require additional observations for training. The market return that the predictors try to predict is assumed to be the monthly return of the S&P500, including dividends, which can be directly obtained from the calculations by the Center for Research in Security Prices (CRSP).

For one of the survey predictors, I rely on The CFO Survey by Duke University's Fuqua School of Business, which is conducted every quarter. The survey provides an estimation for the expected yearly returns, which I convert into monthly simple returns, and in order to have monthly observations to compare to other predictors, I consider that the return prediction of each quarter represents the prediction of each month within that quarter. The monthly predictor I thereby construct from The CFO Survey can therefore be expected to perform less well than if the survey were conducted on a monthly basis.

Another survey I rely on is the Livingston Survey by the Federal Reserve Bank of Philadelphia. The survey is conducted twice a year: in June and in December. It provides an estimation for the level of the S&P500 stock price index in the end of the next half of the year. By using the level of the index at the time when the survey was conducted, this estimation can be converted into an estimated return over half a year, which I in turn convert into a monthly simple return. In order to obtain a monthly predictor, I assume that the monthly return prediction derived from each survey remains constant every month until the next survey is conducted. In this case as well, the performance of the predictor is likely to be a lower bound of the performance that would be observed if the survey were conducted on a monthly basis instead.

From the Federal Reserve Bank of St. Louis (FRED), I obtain data for monthly observations of the US Personal income (pi), and the US Personal consumption expenditures (pce). Quarterly observations of the US GDP are also obtained from the FRED. These observations can all be considered as predictive variables.

Goyal et al. [2024] provide a detailed study on the predictive abilities of numerous variables. These variables can be obtained from the personal website of Amir Goyal[1], where he provides data for the paper. Predictive variables I obtain from there include the price level (price) of the S&P500, its return without dividends (retx), its return with dividends (ret), its 12 months dividends (d12), its 12 months earnings (e12), AAA bond yields (AAA), the corporate bond return (corpr), and the risk-free return (Rfree). Other variables it provides include the long-term bond rate of return (ltr), the term-spread (tms), the default yield (dfy), as in Fama and French [1989]; the T-Bill rate (tbl), as in Campbell [1987]; the dividend-price ratio (d/p), the dividend-yield (d/y), the earnings-price ratio (e/p), the dividend-payout ratio (d/e), as in Campbell and Shiller [1988]; the inflation rate (infl), as in Fama and Schwert [1977]; net issuing activity (ntis), as in Boudoukh et al. [2007]; the stock volatility (svar), as in Guo [2006]; book-market (b/m), as in Kothari and Shanken [1997] and Pontiff and Schall [1998]; the first principal component of 14 technical indicators (tchi), as in Neely et al. [2014]; the log of the number of zero returns (lzrt), as in Chen et al. [2018]; the output gap (ogap), as in Cooper and Priestley [2009]; the price of West-Texas Intermediate crude

[1]Amir Goyal is a professor at the University of Lausanne and Swiss Finance Institute.

oil (wtexas), as in Driesprong et al. [2008]; the average of monthly skewness values across firms (skvw), as in Jondeau et al. [2019]; the time-varying tail risk measure (tail) from Kelly and Jiang [2014]; the single factor extracted from the cross-section of book-to-market ratios (fbm), by Kelly and Pruitt [2013]; the scaled current difference to the 52-week high of the Dow Jones index (dtoy), and the current difference to the lifetime high of the Dow Jones index (dtoat), as in Li and Yu [2012]; the ygap variable (ygap), as computed as in Goyal et al. [2024] and strongly motivated by Maio [2013]; the cross-sectional standard deviation on the set of 100 size and book-to-market portfolios (rdsp), as in Maio [2016]; the average correlation among the 500 largest stocks (avgcor), as in Pollet and Wilson [2010].

B. *Statistical predictor selection*

Using the predictive variables presented in A, and different statistical methods, an exorbitant number of predictors could be constructed. Investigating the performance of all such possible predictors would not be feasible. Therefore, I instead define groups of these aforementioned variables, based on economic intuition, and perform different statistical methods on those, to construct a more reasonable number of statistical predictors.

Group 1 (G1stx): This group is based on stock cross-section and stock market related variables, and includes price, d12, e12, ret, retx, d/e, d/p, d/y, e/p, bm, ntis, svar, lzrt, skvw, tail, fbm, rdsp, avgcor, tchi, dtoy and dtoat.

Group 2 (G2bonds): This group is related to bonds and government securities, and includes AAA, corpr, ltr, tbl, Rfree, tms and dfy.

Group 3 (G3macro): This group is based on variables related to the macro-economy, commodities and sentiment, and includes pi, pce, ogap, infl and ygap and wtexas.

Group 4 (G4stx+ff): This group includes the entirety of Group 1, and combines it with all of the factors present in the model by Fama and French [2015], along with the momentum factor introduced by Jegadeesh and Titman [1993], and Carhart [1997].

Group 5 (G5all): This group contains the variables of all aforementioned groups.

The statistical methods I apply on these groups to construct predictors are the ordinary least squares (OLS) regression (reg), the Lasso regression (las), the Ridge regressions (rid), and the conditional mean embedding with a cosine kernel (CME). While there exist various other methods, as the number of observations is rather small from the perspective of a machine learning algorithm, these four should be sufficient to provide a good understanding of how well statistical methods can perform in this case of monthly return predictions. For each of these methods, I consistently use a training period of five years to estimate the model parameters and/or to cross-validate the hyper parameters, with a rolling window of one month.

Additionally, I construct three predictors, respectively based on the three factor model by Fama and French [1993] (FF3), the four factor model by Carhart [1997] (FF4), and the five factor model by Fama and French [2015] (FF5), using the idea in (3).

C. Performance measurement

In order to understand how well a predictor performs, usually a benchmark predictor is used. However, since the goal is to compare potential predictors in general, arbitrarily choosing one of them as benchmark would seem counterintuitive. Therefore ideally, there should be a way to compare all of the predictors at once. On top of that, from Fama and French [1989] and other studies, we know that the distribution of returns changes over time, which suggests that the best predictor may not remain the same at all times, which is why I also want to continuously perform comparisons over time, to capture possible changes in the ranking. There are different ways to do so, and the one I choose to use is by applying an algorithm from the Online Learning (OL) literature. Within the given context, OL can be interpreted as follows.

In order to more easily differentiate between the different predictors under treatment, let

$$\mathcal{H} := \{P_h | h \in H\} \tag{5}$$

be the set of predictors that are to be compared, and let $P_{h,t}$ be the prediction at time t of predictor P_h, with $h \in H$. H is used in (5) in order to highlight the flexibility of comparing different groups of predictors, as $\mathcal{H}$ can, for example, be the set of all statistically based predictors. As can for example be seen in Bubeck [2011], in the OL literature, the predictors in $\mathcal{H}$ would typically be called "experts". The main idea is to use $\mathcal{H}$ to construct a new hybrid predictor

$$\hat{P}_t^{\mathcal{H}} = \sum_{h \in H} w_{h,t} P_{h,t} \tag{6}$$

where the ideal convex weights $w_{h,t}$ are to be found. This means that we have $w_{h,t} > 0$ and $\sum_{h \in H} w_{h,t} = 1$ for any $h \in H$ and point in time t. In the end, it is these weights that can serve as relative performance measure, which makes them even more important than $\hat{P}_t^{\mathcal{H}}$ for this study.
In order to construct $w_{h,t} > 0$ within the OL framework, first a convex loss function needs to be chosen. Since the goal is to optimally approximate a conditional expectation, the most commonly used loss function L is the squared error, which for any predictor P_t at time t is defined as

$$L\left(P_t, R_{t+1}\right) = \left(P_t - R_{t+1}\right)^2 \tag{7}$$

where R_{t+1} is the observable simple market return of the next month that P_t tries to predict.
Then let the regret $r_{h,t}$ of $\hat{P}^{\mathcal{H}}$ relative to P_h at time t be defined as

$$r_{h,t} := L\left(\hat{P}_t^{\mathcal{H}}, R_{t+1}\right) - L\left(P_{h,t}, R_{t+1}\right) \tag{8}$$

This naturally leads to defining the cumulative regret $\mathcal{R}_{h,t}$ of $\hat{P}^{\mathcal{H}}$ until time t, relative to P_h as

$$\mathcal{R}_{h,t} := \sum_{\tau=t_0}^{t} r_{h,\tau} \tag{9}$$

where t_0 is the starting date of the time series of the predictors to be compared.
We can use this to define the maximal cumulative regret $\mathcal{R}_{\mathcal{H},t}$ of $\hat{P}^{\mathcal{H}}$ at time t as

$$\mathcal{R}_{\mathcal{H},t} := \max_{h \in H} \mathcal{R}_{h,t} \tag{10}$$

According to Cesa-Bianchi and Lugosi [2006], Bubeck [2011], and many others in the OL literature, a great way to construct $\hat{P}_t^{\mathcal{H}}$ is by setting the weights $w_{h,t}$ as

$$w_{h,t} := \frac{\exp\left(\eta \mathcal{R}_{h,t}\right)}{\sum_{j \in H} \exp\left(\eta \mathcal{R}_{j,t}\right)} \tag{11}$$

with $\eta \in \mathbb{R}$. By choosing the loss function L such that it is bounded by $[0, 1]$, which can be achieved by computing squared errors of monthly returns, as realistic monthly returns don't exceed 50% or go below -50%, it can be proven that the following holds

$$\mathcal{R}_{\mathcal{H},t} \leq \frac{\log|\mathcal{H}|}{\eta} + \frac{t\eta}{2} \tag{12}$$

where $|\mathcal{H}|$ is the number of elements in $\mathcal{H}$. This means that by choosing $\eta = \sqrt{2\frac{\log|\mathcal{H}|}{t}}$, we can obtain

$$\mathcal{R}_{\mathcal{H},t} \leq \sqrt{2t \log|\mathcal{H}|} \tag{13}$$

This in turn implies that the average regret $\frac{\mathcal{R}_{\mathcal{H},t}}{t}$ converges to 0 at least as fast as $\sqrt{2\frac{\log|\mathcal{H}|}{t}}$. In other words, if given enough observations, $\hat{P}_t^{\mathcal{H}}$ eventually becomes at least as good as the best predictor in $\mathcal{H}$.

As mentioned above, for this study, the key part of the algorithm lies in the weights $w_{h,t}$, as a higher weight implies a better overall performance up until time t, and the predictor $\hat{P}_t^{\mathcal{H}}$ is not actually necessary to measure the performance. There are also different options to compare the predictors in $\mathcal{H}$. However, this OL framework has several advantages. Firstly, when $0 \leq w_{h,t} \leq 1$ is smaller than $\frac{1}{|\mathcal{H}|}$, then we can infer that the performance of P_h is worse than average, up until time t. Secondly, this method can include any number of benchmark predictors, without requiring any. Additionally, by carefully choosing the content of $\mathcal{H}$, one can use $\hat{P}_t^{\mathcal{H}}$ as representative predictor of $\mathcal{H}$, even when $\sqrt{2\frac{\log|\mathcal{H}|}{t}}$ has not yet reached 0. For example, by setting $\mathcal{H}_1$ as the set of statistical predictors that can be defined using G1stx, $\hat{P}_t^{\mathcal{H}_1}$ can be used a representative predictor, in order to better understand the overall informativeness that G1stx seems to provide. I use this concept in section IV. Moreover, the OL framework can also automatically adapt to changing time indices. For example, in the case of different regimes, one could analyze the performance of the predictors in $\mathcal{H}$ specifically during a regime $\mathcal{T}_1$ by applying the OL algorithm on only the indices t that fall in regime $\mathcal{T}_1$. I also apply a similar idea in section IV.

In the literature, a more traditional method to determine the statistical performance of a predictor is to use the R^2. More specifically, Campbell and Thompson [2008], Goyal et al. [2024] and many others use the out-of-sample-Campbell-Thompson (OOSCT) R^2, which is defined as follows.

$$R^2_{t_0,T}(P) = 1 - \frac{\sum_{t=t_0}^{T} (R_{t+1} - P_t)^2}{\sum_{t=t_0}^{T} \left(R_{t+1} - \bar{R}_t\right)^2} \tag{14}$$

where $\{t_0, ..., T\}$ is the range of time points at which the predictor P is estimated, and $\bar{R}_t := \frac{1}{t-t_0} \sum_{\tau=t_0}^{t} R_\tau$ is the prevailing mean at time t.

This measure essentially compares the performance of P by comparing it to a benchmark, which is set as the time series of prevailing means. This choice of benchmark is arbitrary, and can not easily be adapted to analyzing performances under different regimes. Therefore, I choose to use the OL framework instead, for the main study, but for some of the main results I do report the OOSCT R^2, in order to facilitate the comparison to related studies.

IV. Main Findings

According to the Federal Reserve History, the GFC lasted from December 2007 to June 2009. As can be seen in A, this period critically affects the overall performance of the predictors. Therefore, in order to gain a better understanding of this affect, I analyze not only the performance over the entire sample period, but also the performance during, and after the GFC, which indeed proves to show differing trends.

A. Main Empirical Findings

Using the OL framework, I start by analyzing the different predictor types separately. I then proceed to perform an analysis on the aggregations of the predictor types. Finally, I also identify the best predictors over the sample period, and compare them among each other.

A.1. Surveys

Following the dotcom bubble burst, starting from March 2002, U.S. indices were steadily falling, reaching severe lows in September and October 2002, and this downturn ended with another final low in March 2003. Figure (1) refers to this period as "post dotcom", and while we can see that the average performance of the CFO survey is consistently above that of the Livingston survey, the out-performance of CFO appears to be especially significant during this period, and around the time of May 2010, which coincides with the trillion dollar 2010 flash crash that hit the U.S. market. However, the performances of both surveys do not appear to diverge, in the end of the sample period. This indicates that the performances outside of those critical moments do not differ as much. In fact, outside of the two above mentioned instances, the OL weights almost seem to converge. The overall OOSCT R^2 of the CFO survey is 2.00%, and that of the Livingston survey is -11.35%.

A.2. Factor Models

As can be seen in Figure (2), the average performance of FF3 appears to remain consistently below that of FF4, while the average performance of FF5 is the lowest. Their respective overall OOSCT R^2s are -18.55%, -7.03% and -22.22% respectively.

A.3. Statistical Methods

Applying different statistical methods onto G1stx, figure (3) shows that Lasso, Ridge and CME all perform similarly well, while OLS performs significantly worse. The difference becomes especially clear starting from September 2008, which coincides with the month during which the famous U.S. bank Lehman Brothers filed for bankruptcy. By a slight margin, CME seems to have the best average performance. The OOSCT R^2 of the OLS regression is -431.10%, while the OOSCT R^2s of Lasso, Ridge and CME are -4.31%, -3.99% and -1.87%, respectively.

Applying different statistical methods onto G2bonds, figure (4) shows that Lasso and Ridge perform similarly well, while OLS performs significantly worse. CME seems to have the best average performance. The performance of OLS reaches its peak around September 2009, and then proceeds to rapidly decline until April 2009, which coincides with the month during which a G20 summit took place, where the leaders of the world's leading economies agreed to pledge to triple funding for the International Monetary Fund, and to increase trade financing, among other things. After April 2009, the performances appear to not differ as much. The OOSCT R^2 of the OLS regression is -26.34%, while the OOSCT R^2s of Lasso, Ridge and CME are -5.20%, -4.55% and -0.38%, respectively.

Applying different statistical methods onto G3macro, figure (5) shows that Lasso and Ridge perform similarly well, while CME performs worse. The performance of OLS remains significantly worse than the others. The performance of OLS reaches its peak around September 2008, and then proceeds to steadily decline. The OOSCT R^2 of the OLS regression is -34.15%, while the OOSCT R^2s of Lasso, Ridge and CME are -3.96%, -3.30% and -6.65%, respectively.

Applying different statistical methods onto G4stx+ff, figure (6) shows that Lasso, Ridge and CME all perform similarly well, while OLS performs significantly worse. The difference becomes especially clear starting from September 2008. By a slight margin, CME seems to have the best average performance. The OOSCT R^2 of the OLS regression is -458.18%, while the OOSCT R^2s of Lasso, Ridge and CME are -4.39%, -3.24% and -0.34%, respectively.

Applying different statistical methods onto G5all, figure (7) shows that Lasso, Ridge and CME all perform similarly well, while OLS performs significantly worse. The difference becomes especially clear starting from October 2002, which corresponds to the month during which the indices reached their bottom, during the 2002 downturn period. The difference becomes even significantly larger starting from September 2008. By a slight margin, CME seems to have the best average performance. The OOSCT R^2 of the OLS regression is -578.61%, while the OOSCT R^2s of Lasso, Ridge and CME are -5.10%, -4.50% and 0.35%, respectively.

A.4. Aggregate Predictors

I construct aggregate predictors for each of the five groups in III, by using (6), given by the OL framework. However, given the poor performance from using OLS regressions, I exclude them from the construction, as they seem to wrongly convey the informativeness of their corresponding predictive variables. Figure (8) shows that G3macro performs significantly better than the other groups, until around January 2009, which coincides with the month before a significant stimulus package was signed into law by the U.S. president, during the GFC. After January 2009, the groups G2bonds and G3macro, which are the top predictors until then, become the two worst performing predictors overall. The OOSCT R^2 of G1stx is -1.70%, while the OOSCT R^2s of the groups G2bonds, G3macro, G4stx+ff and G5all are -2.21%, -1.75%, -1.60% and -1.67%, respectively.

I then construct an aggregate predictor for the surveys, and one for the factor models, and proceed to add them to comparison. Figure (9) shows that factor models perform significantly worse in general, but even more so around the 2002 downturn, and around January 2009. G3macro still appears to be outperforming until January 2009. Surveys start off by showing poor performance, during the 2002 downturn, and then steadily improve, relative to the other aggregate predictors, until they become the best performing predictor for some time. However, shortly before the 2010 flash crash, their drops down again, and towards the end of the sample period, all aggregate predictors appear to perform similarly well, apart from the factor models. The OOSCT R^2 of the surveys is -1.88%, while the OOSCT R^2s of the factor models is -12.41%.

In order to better understanding just how well or poorly these aggregate predictors truly are, I add two benchmark predictors to the comparison, which make no use of expertise or any special type of predictive variable. The first benchmark predictor is the one which continuously blindly predicts 0 returns. The second benchmark predictor is a time series of prevailing 5-years monthly return averages. It is similar to the benchmark predictor used in the computation of the OOSCT R^2, but it can be used more flexibly in the OL framework. In Figure (10), I exclude the aggregate factor model predictor from the comparison, because it seems evident that all other predictors in the comparison perform significantly better, and being able to take a closer look at the other predictors' performances seems to be more valuable. Figure (10) shows that the 0-predictor actually outperforms all aggregate predictors, especially after January 2009, and also significantly outperforms the 5-year average predictor, which performs similarly well to the aggregate predictors. The OOSCT R^2 of the 0-predictor is 0.02%, while the OOSCT R^2s of the 5-year average predictor is -1.46%.

A.5. Best Individual Predictors

Table (II) contains the OOSCT R^2 for every predictor mentioned above, for the period from December 2001 to November 2007, which corresponds to the available sample before the start of the GFC. Apart from G4stx+ff, the table confirms that for each group of predictive variables, as in III.B, there exists a statistical method which constructs a predictor which outperforms the benchmark models before the crisis. Figure (11) illustrates this observation. Additionally, it shows a severe decline in the performance of these statistical predictors, starting from January 2009, relative to the zero predictor. Only the CFO survey seems to consistently perform well. In order to better understand the impact of the GFC, and how it affects all predictors, I separately analyze the predictors' performances during that time.

A.6. *During the GFC*

Figures (12) and (13) show that the 0-predictor's relative performance is rarely below any other predictor's, during the GFC, and strongly increases after January 2009. On the other hand, the prevailing 5-year average predictor performs particularly poorly before January 2009, compared to every aggregate statistical predictor, and compared to the survey based predictors. Towards the end of the crisis, the relative performance of the 5-year average predictor returns to the predictor sample's average. Table (III) reports the OOSCT R^2s during the time of the GFC, and confirms what can already be seen from the figures. Despite having no informativeness by construction, the OOSCT R^2 of the 0-predictor is the largest among all predictors, and is equal to 5.27%.

A.7. *After the GFC*

By the end of the sample period after the GFC, the only predictors standing out are those constructed through regressions, and the predictor based on the Livingston survey. Those predictors all display poor performance. Figure (15) shows that the 0-predictor's relative performance is below average, after the GFC, although not severely, especially during the last few years of the sample data. The performance of the 5-year average predictor remains close to, but slightly above average. Until the end of 2015, factor models seem to perform well, and figure (13) suggests that this is mainly influenced by FF4's performance. In fact, the OOSCT R^2 of FF3 and FF5 are negative, as can be seen in table (IV). Overall, during the period after the GFC, no predictor performs particularly well, although table(IV) does show that all predictors, that use a Lasso regression on stock market related variables, slightly outperform the 5-year average benchmark predictor.

B. *Result Interpretation*

Generally, the CFO survey demonstrates considerable informativeness during stable periods, as it exhibits the best predictive performance, when looking at the sample period while excluding the GFC. In fact, outside of the GFC, the CFO survey consistently outperforms the benchmark predictors. This suggests that expert judgment, gathered systematically, can serve as a valuable predictor when markets are less volatile. Moreover, the Livingston survey, which is conducted less frequently, shows weaker performance than the CFO survey, especially during times of sudden market downturns. Hence, the frequency of survey data collection may play a significant role, as a higher granularity of expert insights may yield better predictive performance. However, as my analysis only includes two different surveys, this idea needs to be studied more, before reaching any definitive conclusion.

Overall, the factor model based predictors all perform rather poorly. From the end of the GFC until the end of 2015, FF4 seems to perform well on average, but this is probably only due to the fact that the stock market was mostly steadily rising again, which would naturally make the momentum factor seem more informative than it necessarily is. Figure (16) seems to confirm this idea, as it can be seen that the performance of FF4 suddenly and severely drops in August 2011, which coincides with a month during which the stock market experienced a sudden drop. There are no such further drops before 2018. One reasonable explanation for the slowly decreasing relative performance of FF4 during the years of 2014 and 2015 would be that other predictors simply took more time to detect the upward trend of the stock market. FF4 stops performing better than the other predictors around December 2015, which coincides with the implementation of a long awaited interest rate increase. Perhaps this increase can be associated with a more general awareness of the upward trend. The fact that G4stx+ff does not outperform G1stx, despite using the same predictive variables and more, shows that neither of the factors really seems to contain any information about future returns, that the predictive variables of G1stx do not already possess. This also supports the idea that factor models do not seem to be well suited for return predictions. In other words, the explanatory power of factor models does not seem to translate into predictive power.

Lasso regressions, Ridge regressions and conditional mean embeddings all mitigate the risk of overfitting. The same cannot be said about OLS regressions. The fact that OLS regressions across all groups of predictive variables severely underperform, compared to their more sophisticated counterparts, especially after the bankruptcy of Lehman Brother, reinforces the critical need for regularization and more sophisticated statistical approaches that are less prone to overfitting, especially when dealing with complex financial data. In other words, using OLS regressions to determine the informativeness of predictive variables appears to be inappropriate, especially during severe market downturns. Figure (7) highlights this point, as the drops of the OLS regression's overall performance appear specifically around two major events of market downturns. With this idea in mind, as can be seen in figure (4), G2bonds seems to capture a deceleration in market decline after the G20 summit.

Before the stimulus signed in February 2009, figures (9) and (10) show that G3macro performs particularly well, which indicates that the informativeness of the macro-economic predictive variables is rather high around this period, compared to others. Considering that statistical methods in this study have a five years training period, the fact that G3macro starts to perform poorly after the fiscal stimulus suggests that certain global market properties may have changed. Looking at the development of the U.S. GDP in figure (17), this explanation seems plausible. This would also explain the complete prediction failures of all predictors during several months following the signature of the stimulus contract, as the methods would have had to readapt to the new properties. While it is true that some predictors have positive OOSCT R^2s around that time, the mere fact that a completely uninformative predictor such as the 0-predictor outperforms all other predictors shows that they all fail to display any type of relevant informativeness. This last point highlights two other key observations. The first observation is that an OOSCT R^2 of even more than 5% may not indicate any informativeness during times when the implicit benchmark predictor can be expected to perform poorly. The second observation is that the informativeness of predictors can be overseen when we do not dissect the sample period adequately. In fact, while G3macro performs very well before the GFC, and even at the beginning of the crisis, looking at the OOSCT R^2s over the entire sample period only would suggest that the group possesses no informativeness, as it does not even outperform the 0-predictor.

Additionally, I would like to point out the difference between the performances before the GFC, and the performances during the period after the crisis. The relative informativeness of most predictors seems weak after the crisis. This could be related to the stability around that time. As the market was steadily growing after a period of intense downfall, most predictors are unsurprisingly able to detect such a simple trend. This type of stability was not present after the recession following the dotcom bubble. Accordingly, clearer performance differences can be observed in the sample period before the GFC, which in turn helps to identify the informativeness of the different predictors more clearly. Apart from G4stx+f, all other groups of predictive variables have a statistical method which allows them to provide better predictions than both benchmark predictors, before the GFC. CME helps to indicate that the stock market, as well as the bond and security related predictive variables contain some level of informativeness, as can be seen from table (11). Meanwhile applying the Ridge regression onto the predictive macro-economic variables of G3macro shows signs of significant informativeness from the latter. Applying CME onto G5all, which contains all the predictive variables of this study, confirms that at least some of these variables contain valuable information about future returns.

Looking at the entire sample period, despite the disturbance of the GFC, the CFO survey predictor, as well as the predictor which applies CME onto G5all, outperform the two benchmark predictors, signifying that both survey predictors and statistical predictors show some level of predictive power. Interestingly, the insight that can be gained from both seem to differ, as combining the two with arbitrarily chosen weights can happen to result in an OOSCT R^2 which is greater than any individual predictor analyzed above. For example, as can be seen in table (I), by assigning 80% to the CFO survey, and 20% to the predictor that applies CME onto G5all, I obtain an OOSCT R^2 of 2.11%, surpassing even the 2.00% of the CFO survey predictor.

Lastly, I would like to point out the surprisingly strong performance of the 0-predictor, which performs better than all other predictors after the stimulus contract's signature, even though none of the monthly returns between January 2009 and April 2009 had an absolute value below 8%. Both survey predictors also perform better than the average, around that time. Interestingly, their predictions remain close to 0 as well. Therefore it seems that a more conservative approach during such volatile times may be better than trusting any predictor that suggests otherwise, even when the predictions of such a conservative approach can be predictably far from the real value.

V. Conclusion

This study analyzes the performances of survey-based, factor model-based, and statistically constructed predictors from December 2001 to December 2018, and highlights the importance of understanding the context-dependent nature of market return predictors. Interestingly, a simple 0-predictor (which continuously predicts zero returns) displays surprisingly robust performance over the sample period, especially during the GFC. This finding illustrates that, during periods of extreme uncertainty, the value of more conservative and less complex approaches should not be underestimated. This counterintuitive result underscores a key takeaway: during highly volatile and unpredictable times, simple, baseline predictions can sometimes outperform more complex, supposedly sophisticated models, as they avoid being misled by rapidly changing and noisy information.

On the other hand, while expert surveys can provide valuable insights during stable periods, and advanced statistical methods can outperform during certain volatile phases, no single predictor appears to be universally superior. Adaptability is crucial, and a multi-faceted approach that leverages the strengths of various methods, while being responsive to the changing market landscape, appears to be the most prudent strategy. Future research should focus on developing adaptive predictive models that can integrate different methods dynamically, offering flexibility in response to the diverse conditions observed in financial markets, and more research should be conducted on the predictive powers that can be derived from financial models, as factor models do not cover such models exhaustively.

References

Klaus Adam, Dmitry Matveev, and Stefan Nagel. Do survey expectations of stock returns reflect risk adjustments? *Journal of Monetary Economics*, 117:723–740, 2021. ISSN 0304-3932. doi: https://doi.org/10.1016/j.jmoneco.2020.04.010. URL https://www.sciencedirect.com/science/article/pii/S030439322030057X.

Daniele Bianchi, Alexandre Rubesam, and Andrea Tamoni. It takes two to tango: Economic theory and model uncertainty for equity premium prediction. Working Paper, 09 2024.

Jacob Boudoukh, Roni Michaely, Matthew Richardson, and Michael R. Roberts. On the importance of measuring payout yield: Implications for empirical asset pricing. *Journal of Finance*, 62:877–915, 2007.

Sébastien Bubeck. *Introduction to Online Optimization*. Princeton University, Department of Operations Research and Financial Engineering, 2011.

John Campbell and Samuel B. Thompson. Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, 21(4):1509–1531, 2008. URL https://EconPapers.repec.org/RePEc:oup:rfinst:v:21:y:2008:i:4:p:1509-1531.

John Y. Campbell. Stock returns and the term structure. *Journal of Financial Economics*, 18: 373–399, 1987.

John Y. Campbell and Robert J. Shiller. Stock prices, earnings, and expected dividends. *Journal of Finance*, 43:661–676, 1988.

Mark M. Carhart. On persistence in mutual fund performance. *Journal of Finance*, 52(1):57–82, 1997.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, 2006.

Yu Chen, Gregory W. Eaton, and Bradley S. Paye. Micro (structure) before macro? the predictive power of aggregate illiquidity for stock returns and economic activity. *Journal of Financial Economics*, 130:48–73, 2018.

Ian Cooper and Richard Priestley. Time-varying risk premiums and the output gap. *Review of Financial Studies*, 22:2801–2833, 2009.

Gerben Driesprong, Ben Jacobsen, and Benjamin Maat. Striking oil: Another puzzle? *Journal of Financial Economics*, 89:307–327, 2008.

Eugene F. Fama. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417, 1970. ISSN 00221082, 15406261. URL http://www.jstor.org/stable/2325486.

Eugene F. Fama and Kenneth R. French. Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics*, 25(1):23–49, 1989. ISSN 0304-405X. doi: https://doi.org/10.1016/0304-405X(89)90095-0. URL https://www.sciencedirect.com/science/article/pii/0304405X89900950.

Eugene F. Fama and Kenneth R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993. ISSN 0304-405X. doi: https://doi.org/10.1016/0304-405X(93)90023-5. URL https://www.sciencedirect.com/science/article/pii/0304405X93900235.

Eugene F. Fama and Kenneth R. French. A five-factor asset pricing model. *Journal of Financial Economics*, 116:1–22, 2015.

Eugene F. Fama and G. William Schwert. Asset returns and inflation. *Journal of Financial Economics*, 5:115–146, 1977.

Joachim Freyberger, Andreas Neuhierl, and Michael Weber. Dissecting characteristics nonparametrically. *The Review of Financial Studies*, 33(5):2326–2377, 04 2020. ISSN 0893-9454. doi: 10.1093/rfs/hhz123. URL https://doi.org/10.1093/rfs/hhz123.

Amit Goyal, Ivo Welch, and Athanasse Zafirov. A comprehensive 2022 look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 37(11):3490–3557, 09 2024. ISSN 0893-9454. doi: 10.1093/rfs/hhae044. URL https://doi.org/10.1093/rfs/hhae044.

Robin Greenwood and Andrei Shleifer. Expectations of returns and expected returns. *Review of Financial Studies*, 27, 01 2013. doi: 10.1093/rfs/hht082.

Hui Guo. On the out-of-sample predictability of stock market returns. *Journal of Business*, 79: 645–670, 2006.

Eric Jondeau, Qian Zhang, and Xiaohu Zhu. Average skewness matters. *Journal of Financial Economics*, 134:29–47, 2019.

Bryan Kelly and Hao Jiang. Tail risk and asset prices. *Review of Financial Studies*, 27:2841–2871, 2014.

Bryan Kelly and Seth Pruitt. Market expectations in the cross-section of present values. *Journal of Finance*, 68:1721–1756, 2013.

S. Kothari and Jay Shanken. Book-to-market, dividend yield, and expected market returns: A time-series analysis. *Journal of Financial Economics*, 44:169–203, 1997.

Ji Hyung Lee, Zhentao Shi, and Zhan Gao. On lasso for predictive regression. *Journal of Econometrics*, 229(2):322–349, 2022. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2021.02.002. URL https://www.sciencedirect.com/science/article/pii/S030440762100049X.

Jin Li and Jianfeng Yu. Investor attention, psychological anchors, and stock return predictability. *Journal of Financial Economics*, 104:401–419, 2012.

Zhu Li, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. Optimal rates for regularized conditional mean embedding learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 4433–4445. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/1c71cd4032da425409d8ada8727bad42-Paper-Conference.pdf.

John Lintner. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics*, 47(1):13–37, 1965. ISSN 00346535, 15309142. URL http://www.jstor.org/stable/1924119.

Paulo Maio. The "fed model" and the predictability of stock returns. *Review of Finance*, 17: 1489–1533, 2013.

Paulo Maio. Cross-sectional return dispersion and the equity premium. *Journal of Financial Markets*, 29:87–109, 2016.

Burton G. Malkiel. The efficient market hypothesis and its critics. *Journal of Economic Perspectives*, 17(1):59–82, March 2003. doi: 10.1257/089533003321164958. URL https://www.aeaweb.org/articles?id=10.1257/089533003321164958.

Christopher Mohri. Online learning algorithms for statistical arbitrage, 2018. URL https://arxiv.org/abs/1811.00200.

Stefan Nagel and Zhengyang Xu. Asset pricing with fading memory. *The Review of Financial Studies*, 35(5):2190–2245, 08 2021. ISSN 0893-9454. doi: 10.1093/rfs/hhab086. URL https://doi.org/10.1093/rfs/hhab086.

Christopher J. Neely, David E. Rapach, Jun Tu, and Guofu Zhou. Forecasting the equity risk premium: The role of technical indicators. *Management Science*, 60:1772–1791, 2014.

Joshua M. Pollet and Mungo Wilson. Average correlation and stock market returns. *Journal of Financial Economics*, 96:364–380, 2010.

Jeffrey Pontiff and Lawrence D. Schall. Book-to-market ratios as predictors of market returns. *Journal of Financial Economics*, 49:141–160, 1998.

William F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3):425–442, 1964. doi: https://doi.org/10.1111/j.1540-6261.1964.tb02865.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1964.tb02865.x.

Ambrus Tamás and Balázs Csanád Csáji. Recursive estimation of conditional kernel mean embeddings. *Journal of Machine Learning Research*, 25(264):1–35, 2024. URL http://jmlr.org/papers/v25/23-0168.html.

Zixuan Yuan, Junming Liu, Haoyi Zhou, Denghui Zhang, Hao Liu, Nengjun Zhu, and Hui Xiong. Lever: Online adaptive sequence learning framework for high-frequency trading. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6547–6559, 2024. doi: 10.1109/TKDE.2023.3336185.

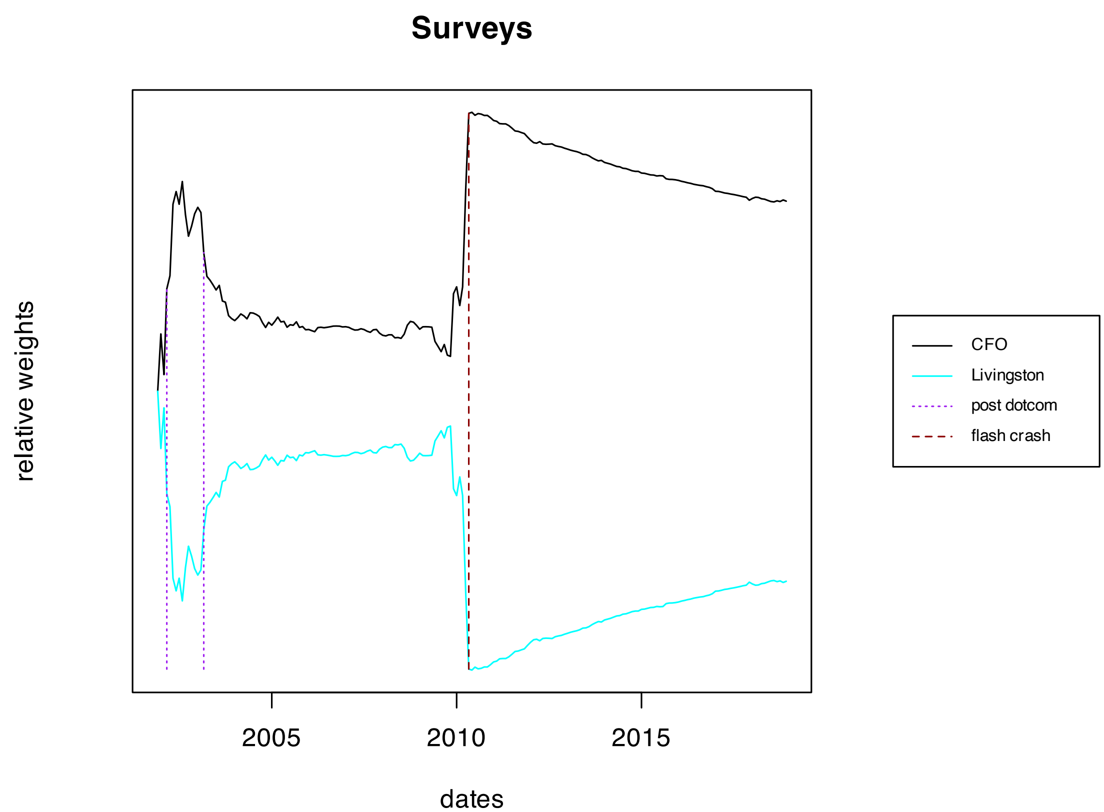

Figure 1. This figure depicts the weights obtained by applying the OL algorithm onto the two survey predictors, over the sample period from December 2001 to December 2018. "post dotcom" refers to the period from March 2002 to March 2003, which corresponds to a period of U.S. market downfall following the dotcom bubble burst. "flash crash" refers to the month of May 2010, during which the U.S. market experienced a trillion dollar flash crash.

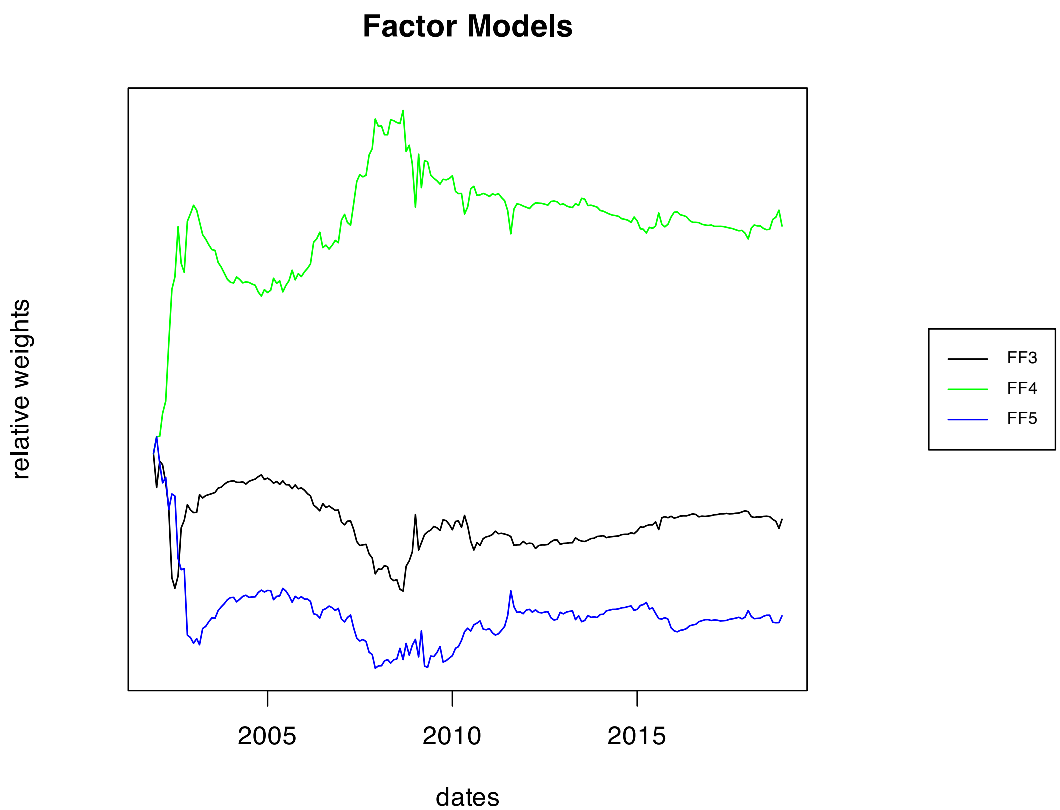

Figure 2. This figure depicts the weights obtained by applying the OL algorithm onto the factor-model predictors FF3, FF4 and FF5, over the sample period from December 2001 to December 2018.

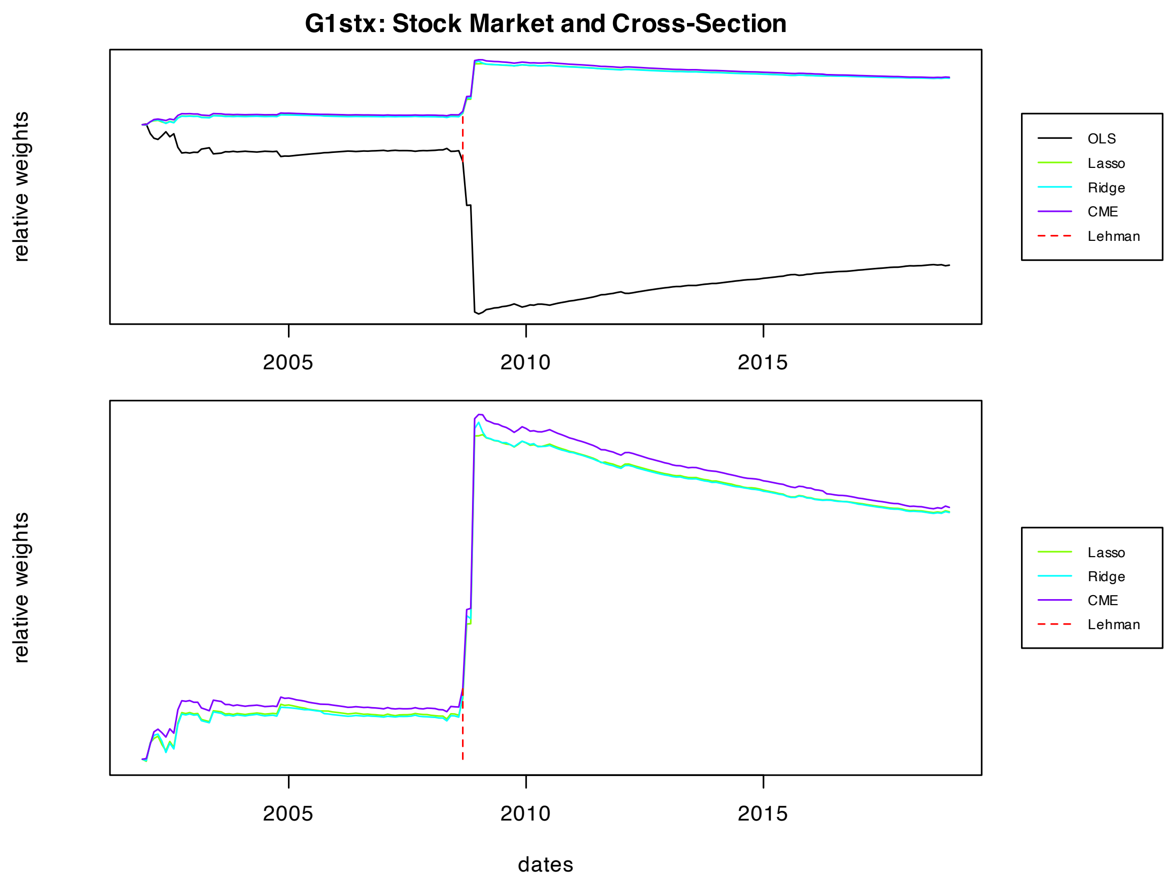

Figure 3. The top depicts the weights obtained by applying the OL algorithm over the sample period from December 2001 to December 2018, onto a group of four predictors that are constructed by applying different statistical methods onto the predictive variables of G1stx, which correspond to the predictive variables that are related to the stock market and the cross-section of stocks. The four statistical methods used are the OLS regression, the Lasso regression, the Ridge regression, and the CME. The bottom depicts the exact same weights as the top, but without showing the weights assigned to the OLS regression, in order to facilitate the differentiation between the weights assigned to the other methods. "Lehman" refers to the month of September 2008, during which the bank Lehman Brothers filed for bankruptcy.

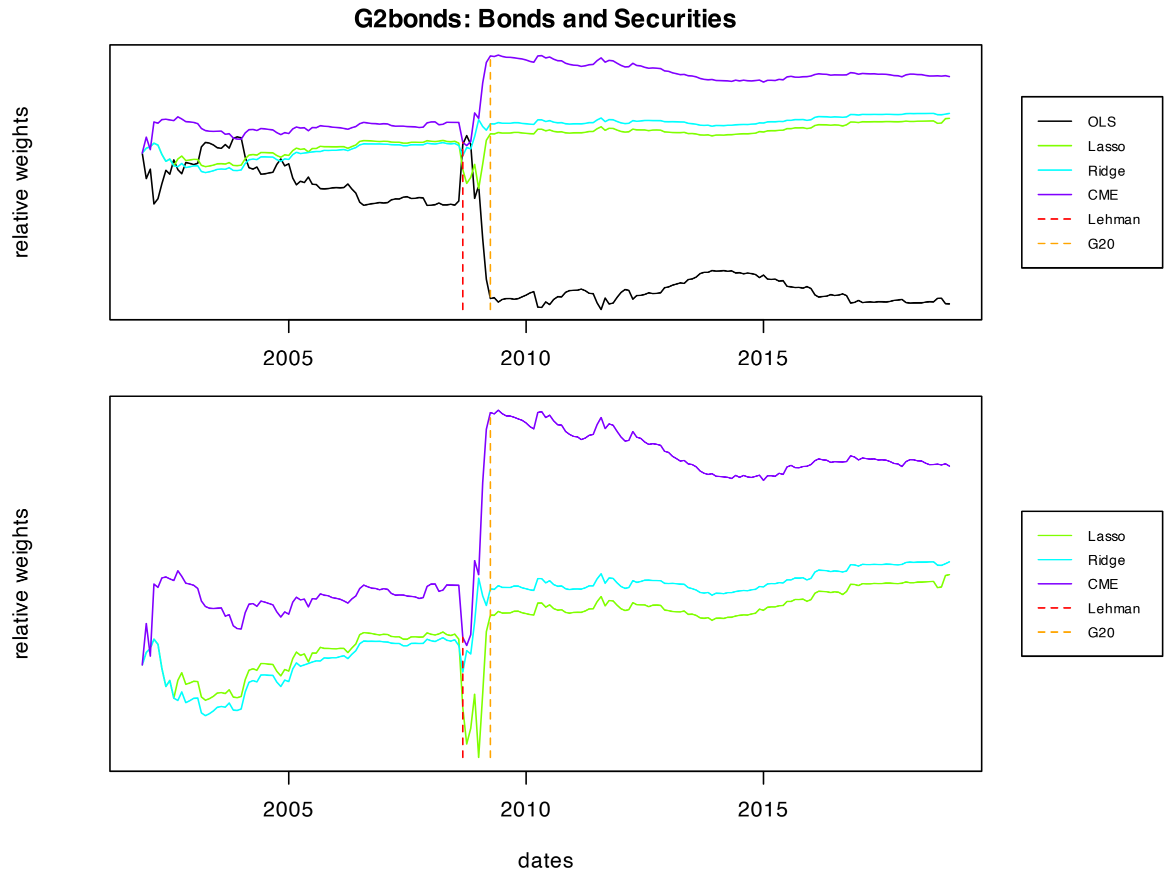

Figure 4. The top depicts the weights obtained by applying the OL algorithm over the sample period from December 2001 to December 2018, onto a group of four predictors that are constructed by applying different statistical methods onto the predictive variables of G2bonds, which correspond to the predictive variables that are related to bonds and government securities. The four statistical methods used are the OLS regression, the Lasso regression, the Ridge regression, and the CME. The bottom depicts the exact same weights as the top, but without showing the weights assigned to the OLS regression, in order to facilitate the differentiation between the weights assigned to the other methods. "Lehman" refers to the month of September 2008, during which the bank Lehman Brothers filed for bankruptcy. "G20" refers to the month of April 2009, during which a G20 summit took place, where global leaders agreed to push certain financial regulations.

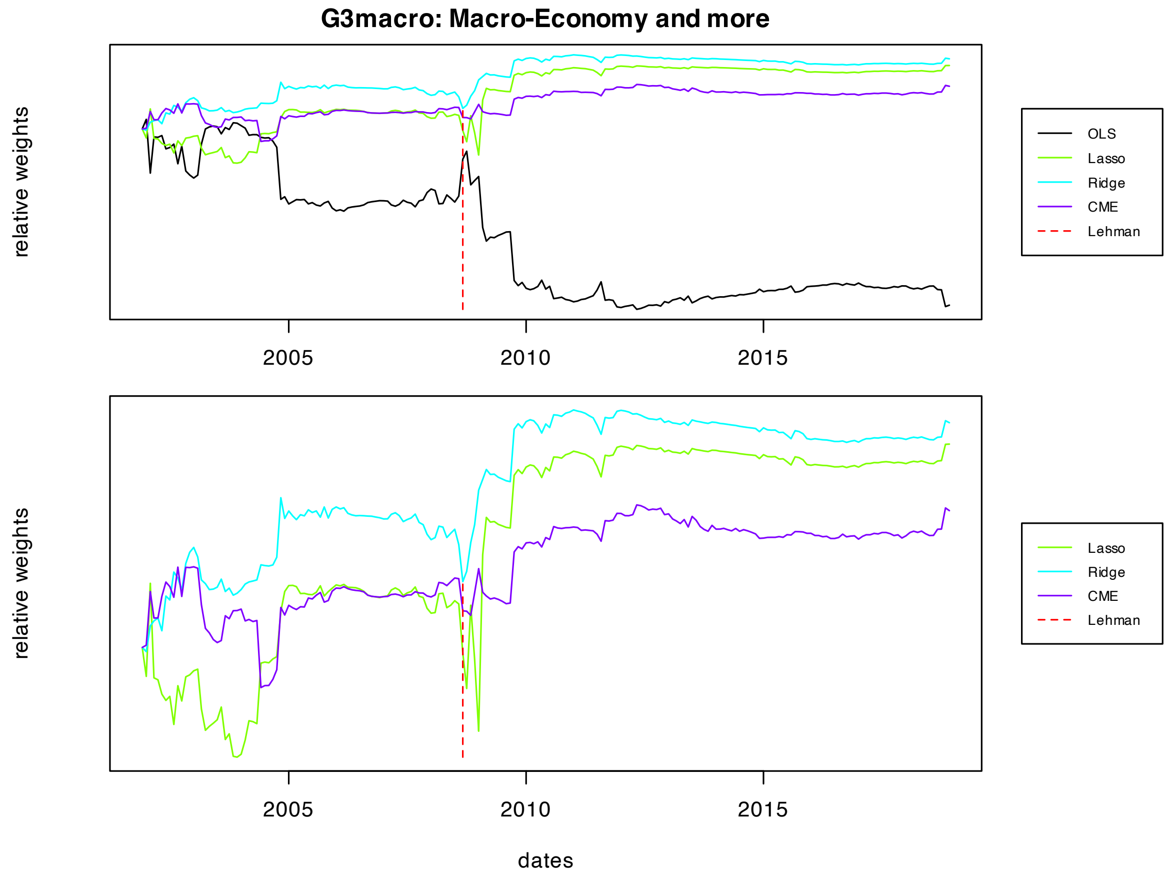

Figure 5. The top depicts the weights obtained by applying the OL algorithm over the sample period from December 2001 to December 2018, onto a group of four predictors that are constructed by applying different statistical methods onto the predictive variables of G3macro, which correspond to the predictive variables that are related to the macro-economy, commodities and sentiment. The four statistical methods used are the OLS regression, the Lasso regression, the Ridge regression, and the CME. The bottom depicts the exact same weights as the top, but without showing the weights assigned to the OLS regression, in order to facilitate the differentiation between the weights assigned to the other methods. "Lehman" refers to the month of September 2008, during which the bank Lehman Brothers filed for bankruptcy.

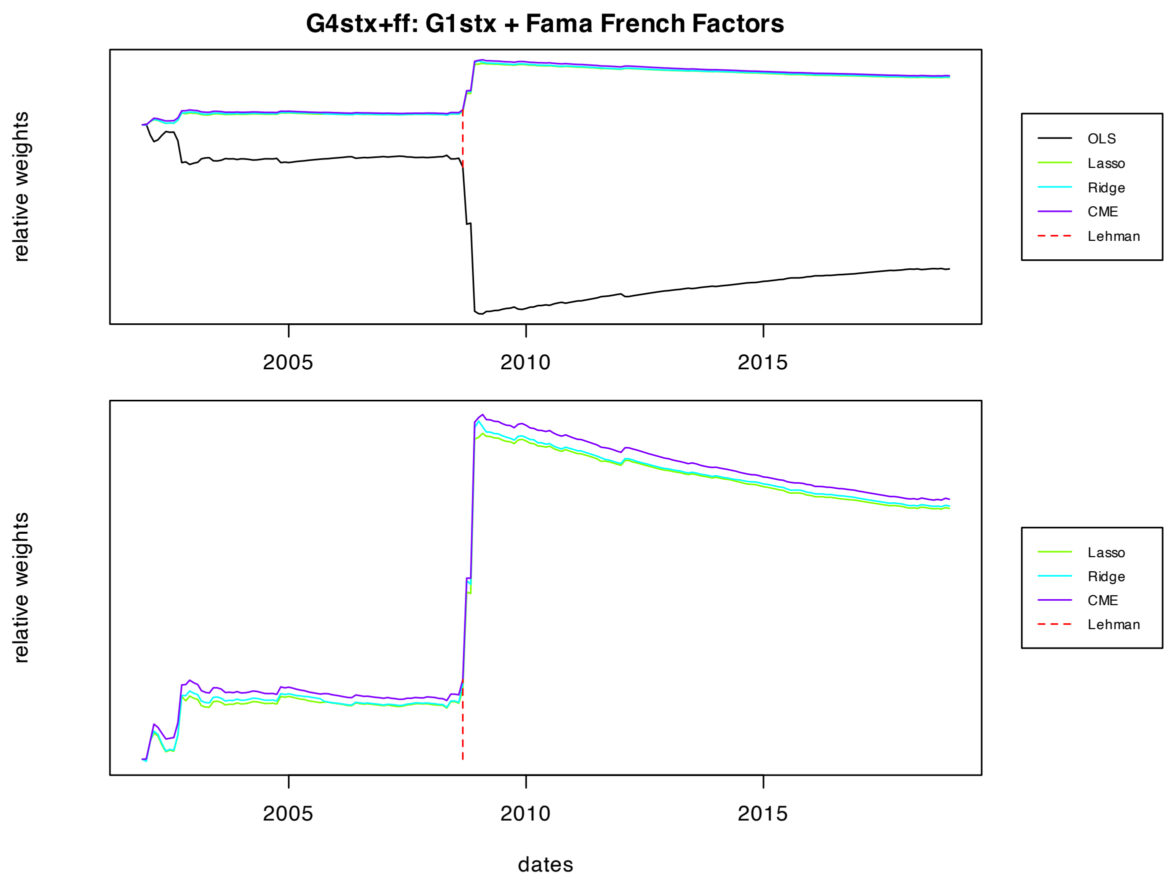

Figure 6. The top depicts the weights obtained by applying the OL algorithm over the sample period from December 2001 to December 2018, onto a group of four predictors that are constructed by applying different statistical methods onto the predictive variables of G4stx+ff, which correspond to the predictive variables that are related to the stock market and the cross-section of stocks, as well as all the factors present in FF4 and/or FF5. The four statistical methods used are the OLS regression, the Lasso regression, the Ridge regression, and the CME. The bottom depicts the exact same weights as the top, but without showing the weights assigned to the OLS regression, in order to facilitate the differentiation between the weights assigned to the other methods. "Lehman" refers to the month of September 2008, during which the bank Lehman Brothers filed for bankruptcy.

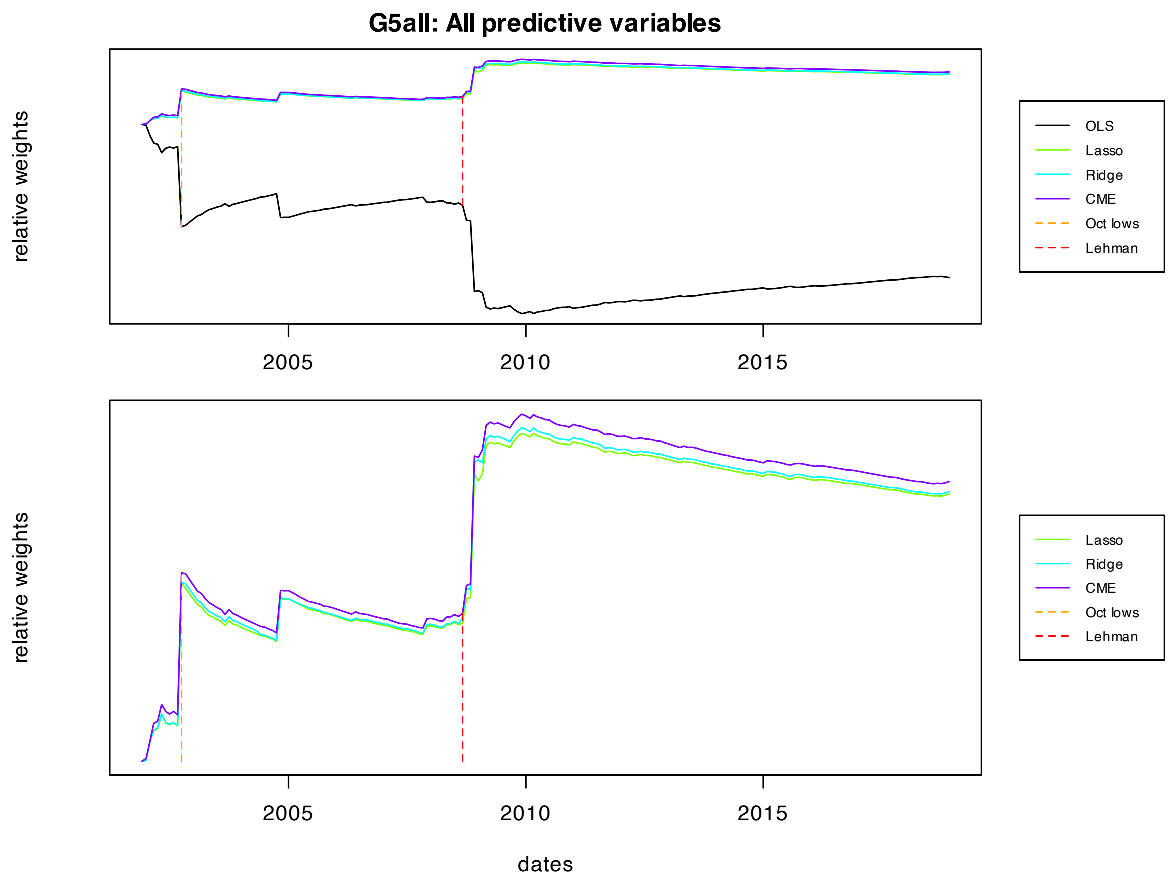

Figure 7. The top depicts the weights obtained by applying the OL algorithm over the sample period from December 2001 to December 2018, onto a group of four predictors that are constructed by applying different statistical methods onto the predictive variables of G5all, which correspond to the set of all predictive variables considered in the study. The four statistical methods used are the OLS regression, the Lasso regression, the Ridge regression, and the CME. The bottom depicts the exact same weights as the top, but without showing the weights assigned to the OLS regression, in order to facilitate the differentiation between the weights assigned to the other methods. "oct lows" refers to the month of October 2002, during which the U.S. market reached its deepest low of a period of downfall from March 2002 to March 2003."Lehman" refers to the month of September 2008, during which the bank Lehman Brothers filed for bankruptcy.

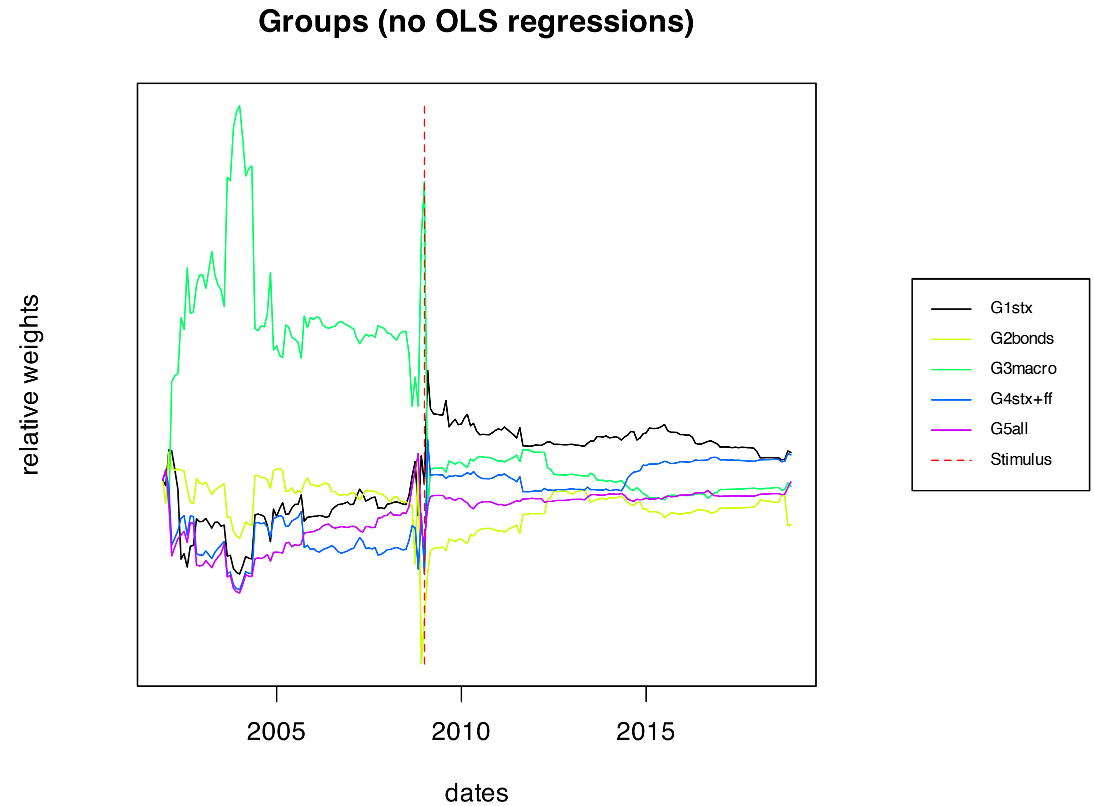

Figure 8. This figure depicts the weights obtained by applying the OL algorithm onto five aggregate predictors, over the sample period from December 2001 to December 2018. The aggregate predictors correspond to the groups G1stx, G2bonds, G3macro, G4stx+ff and G5all, respectively. They are constructed by applying the OL algorithm over the sample period from December 2001 to December 2018, onto a group of three predictors that are constructed by applying different statistical methods onto the corresponding sets of predictive variables. The three statistical methods used are the Lasso regression, the Ridge regression, and the CME. "Stimulus" refers to the month of January 2009, the month before a stimulus package was signed into law by the U.S. president.

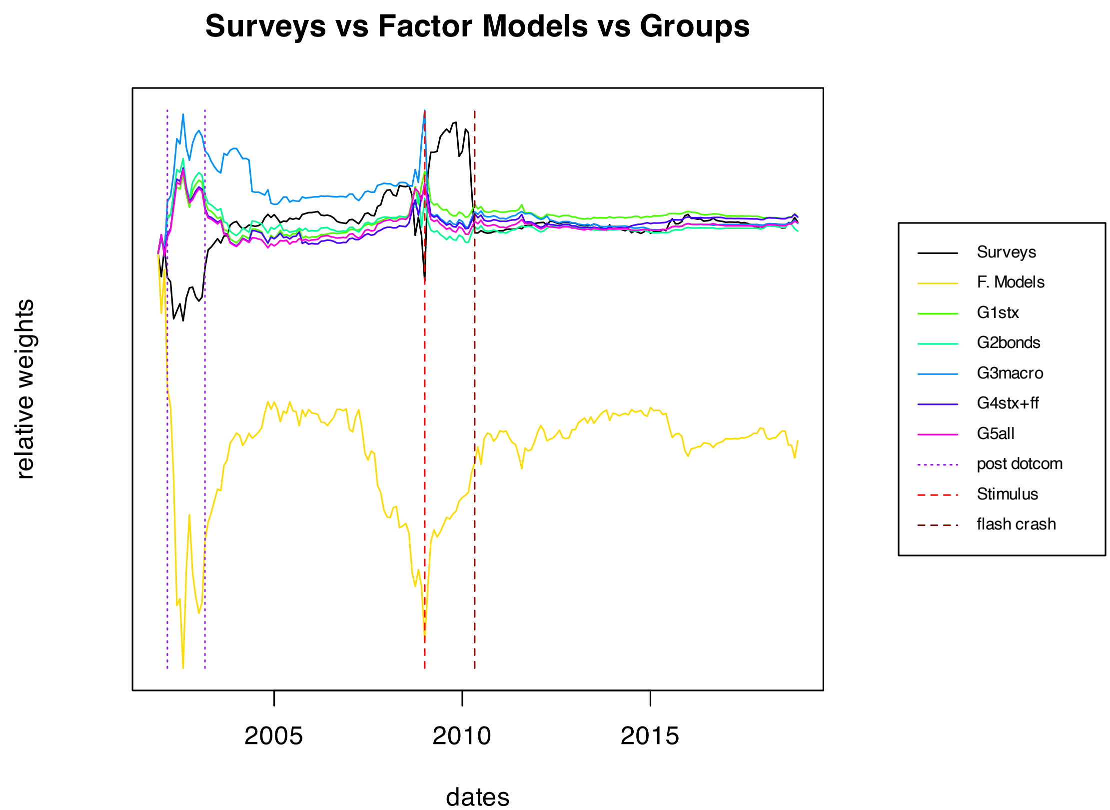

Figure 9. This figure depicts the weights obtained by applying the OL algorithm onto seven aggregate predictors, over the sample period from December 2001 to December 2018. One of these aggregate predictors is constructed by applying the OL algorithm onto the two survey predictors, over the sample period from December 2001 to December 2018. Another one of these aggregate predictors is constructed by applying the OL algorithm onto the factor-model predictors FF3, FF4 and FF5, over the sample period from December 2001 to December 2018. The other five aggregate predictors are the same as the ones in Figure 8. "post dotcom" refers to the period from March 2002 to March 2003, which corresponds to a period of U.S. market downfall following the dotcom bubble burst. "Stimulus" refers to the month of January 2009, the month before a stimulus package was signed into law by the U.S. president. "flash crash" refers to the month of May 2010, during which the U.S. market experienced a trillion dollar flash crash.

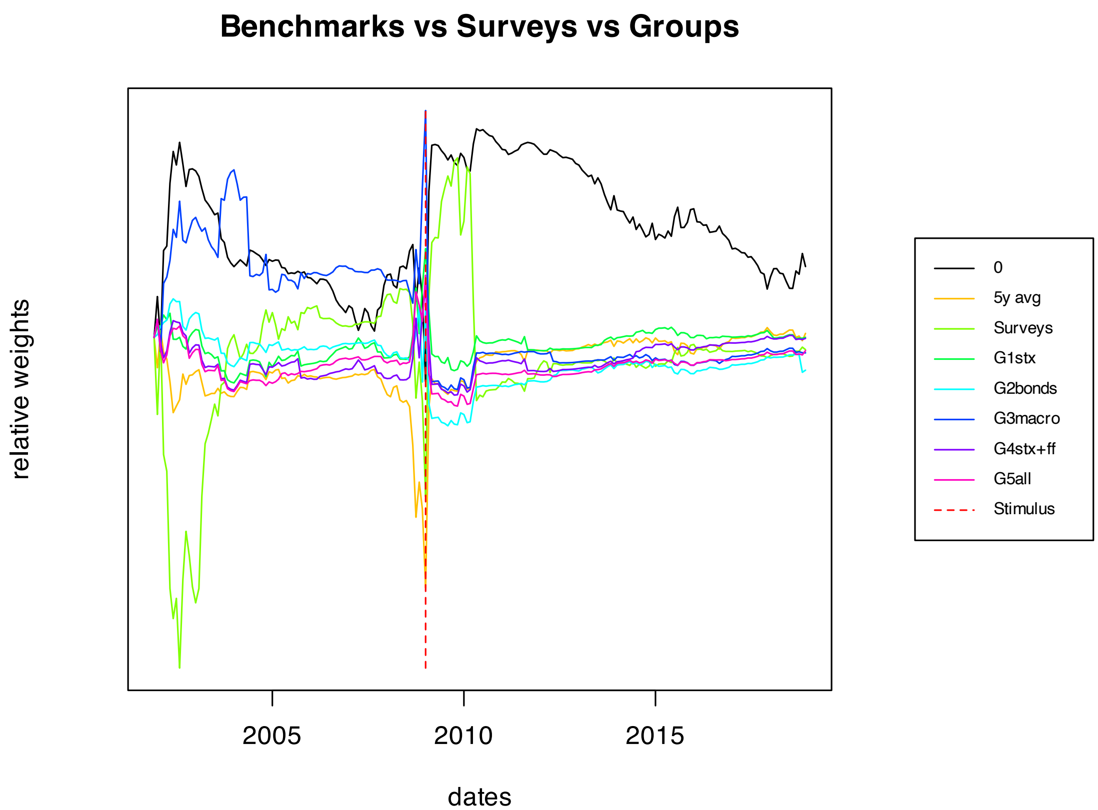

Figure 10. This figure depicts the weights obtained by applying the OL algorithm onto eight different predictors, over the sample period from December 2001 to December 2018. One of these predictors is the 0-predictor. Another one of these predictors is the prevailing 5-year average of the monthly market return. The other six predictors are the same ones as in Figure 9, without the aggregate factor-models predictor. "Stimulus" refers to the month of January 2009, the month before a stimulus package was signed into law by the U.S. president.

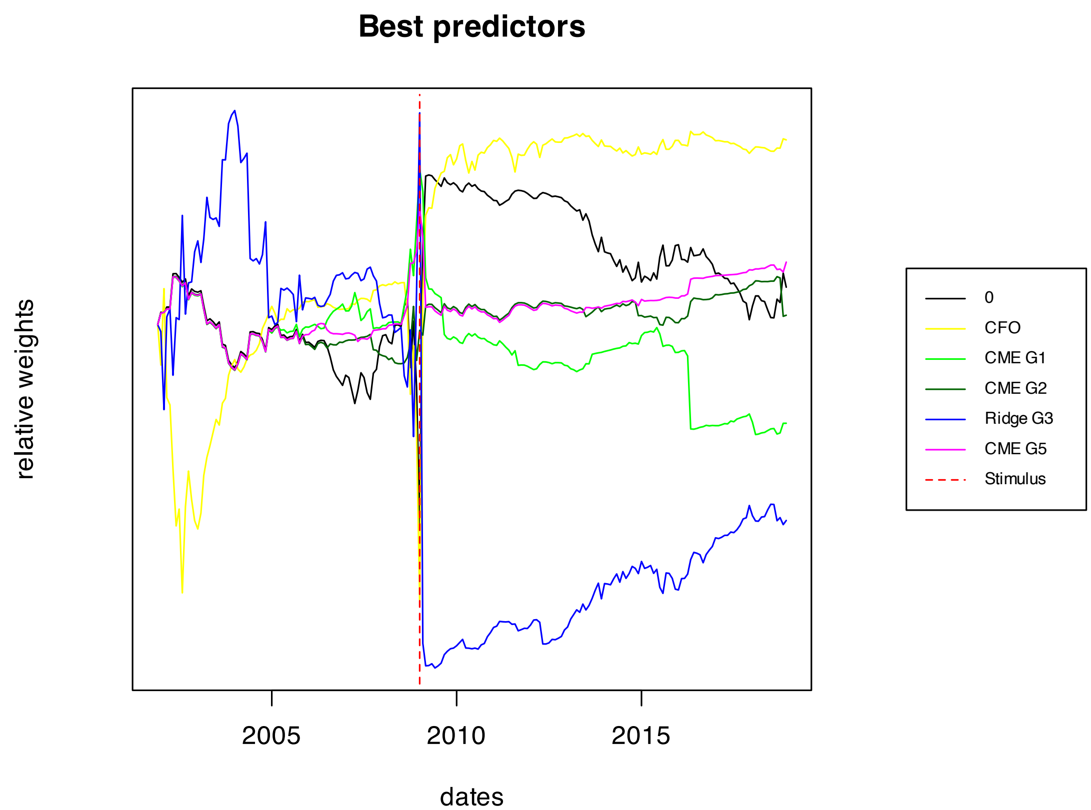

Figure 11. This figure depicts the weights obtained by applying the OL algorithm onto seven different predictors, over the sample period from December 2001 to December 2018. These seven predictors are the 0-predictor, the predictor based on the CFO survey, the predictor which applies CME onto the predictive variables from G1stx, the predictor which applies CME onto the predictive variables from G2bonds, the predictor which applies the Ridge regression onto the predictive variables from G3macro, and the predictor which applies CME onto the predictive variables from G5all. "Stimulus" refers to the month of January 2009, the month before a stimulus package was signed into law by the U.S. president.

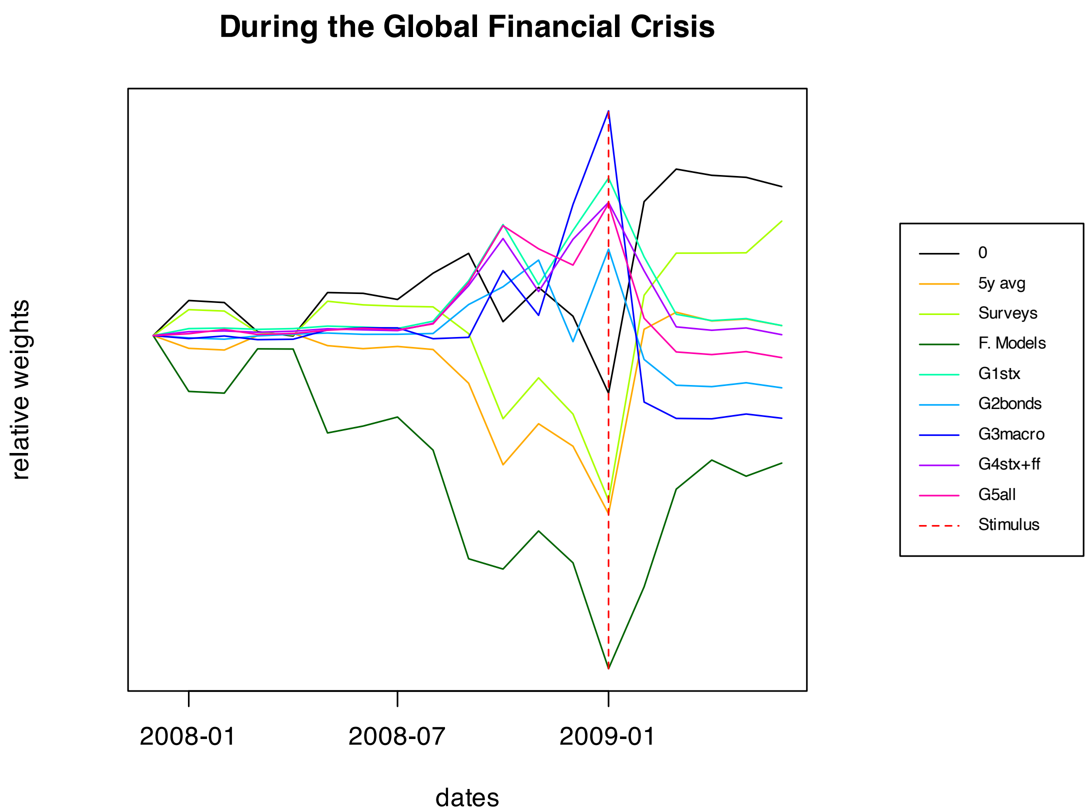

Figure 12. This figure depicts the weights obtained by applying the OL algorithm onto nine different predictors, over the sample period from December 2007 to June 2009. One of these predictors is the 0-predictor. Another one of these predictors is the prevailing 5-year average of the monthly market return. The other seven predictors are the same ones as in Figure 9. "Stimulus" refers to the month of January 2009, the month before a stimulus package was signed into law by the U.S. president.

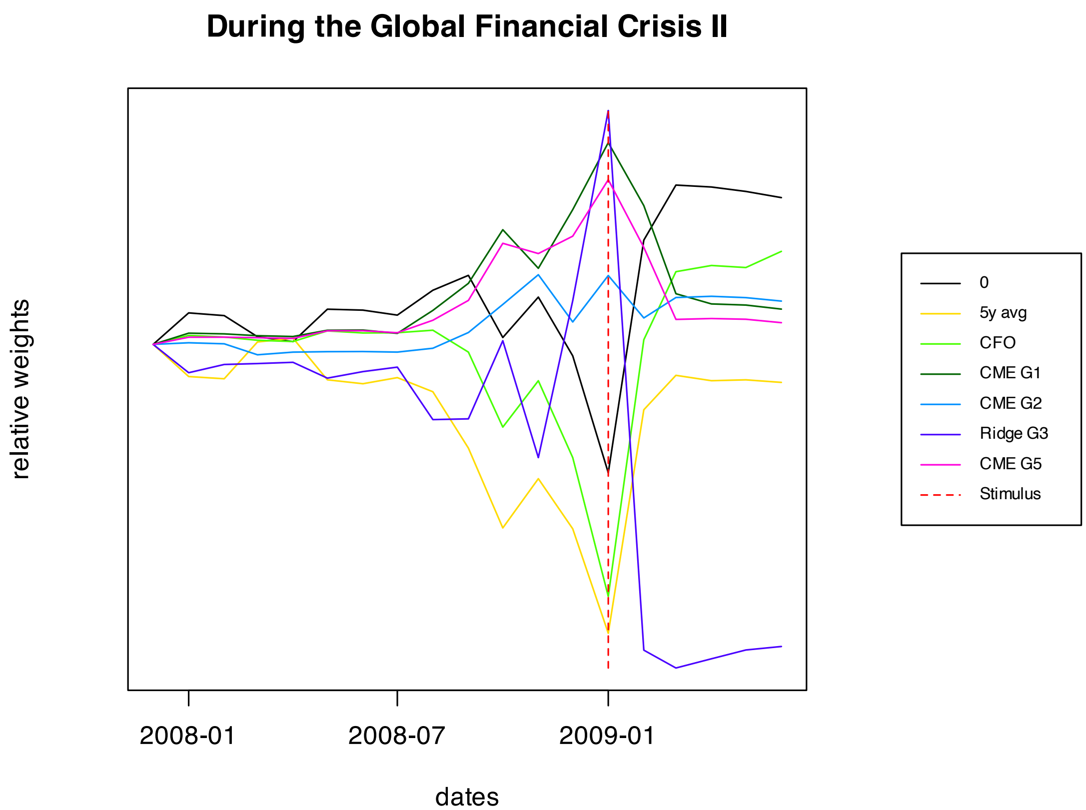

Figure 13. This figure depicts the weights obtained by applying the OL algorithm onto seven different predictors, over the sample period from December 2007 to June 2009. These seven predictors are the 0-predictor, the prevailing 5-year average of the monthly market return, the predictor based on the CFO survey, the predictor which applies CME onto the predictive variables from G1stx, the predictor which applies CME onto the predictive variables from G2bonds, the predictor which applies the Ridge regression onto the predictive variables from G3macro, and the predictor which applies CME onto the predictive variables from G5all. "Stimulus" refers to the month of January 2009, the month before a stimulus package was signed into law by the U.S. president.

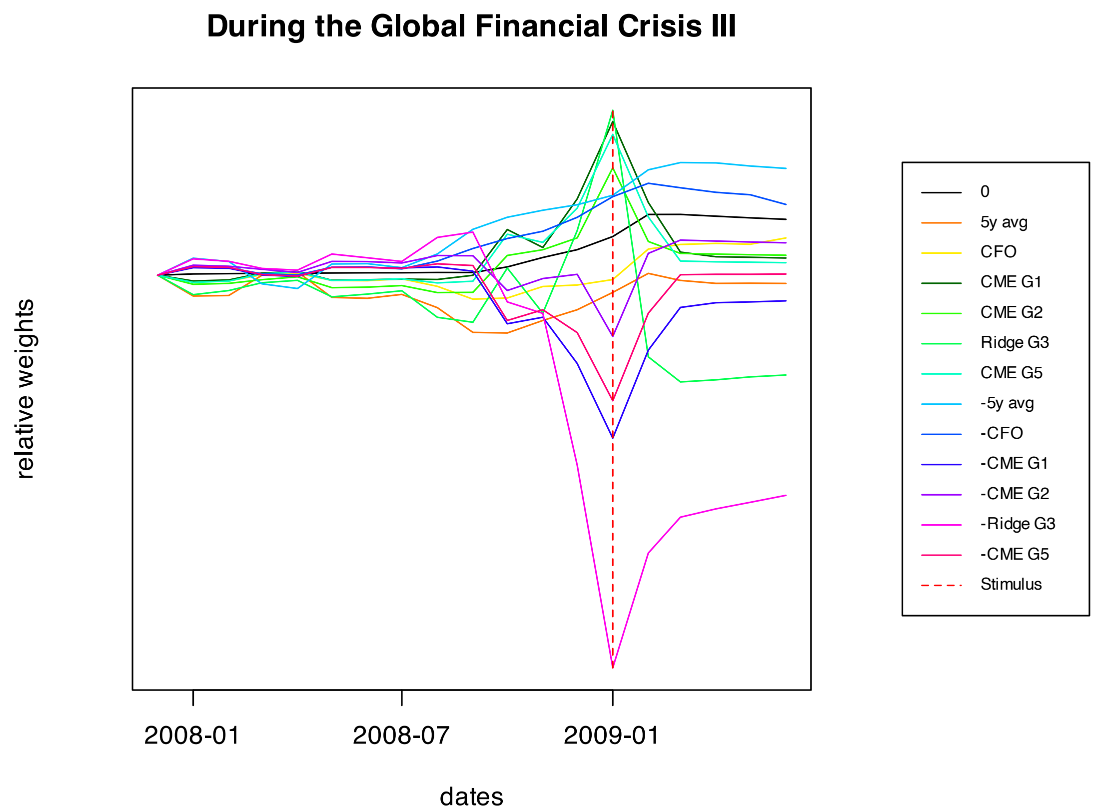

Figure 14. This figure depicts the weights obtained by applying the OL algorithm onto twelve different predictors, over the sample period from December 2007 to June 2009. One of these twelve predictors is the prevailing 5-year average of the monthly market return. The other eleven predictors are the same ones as in Figure 11, along with the respective minus versions of them. Note that the minus version of a predictor predicts the additive inverse of the original predictor, and that the minus version of the 0-predictor is still the 0-predictor. "Stimulus" refers to the month of January 2009, the month before a stimulus package was signed into law by the U.S. president.

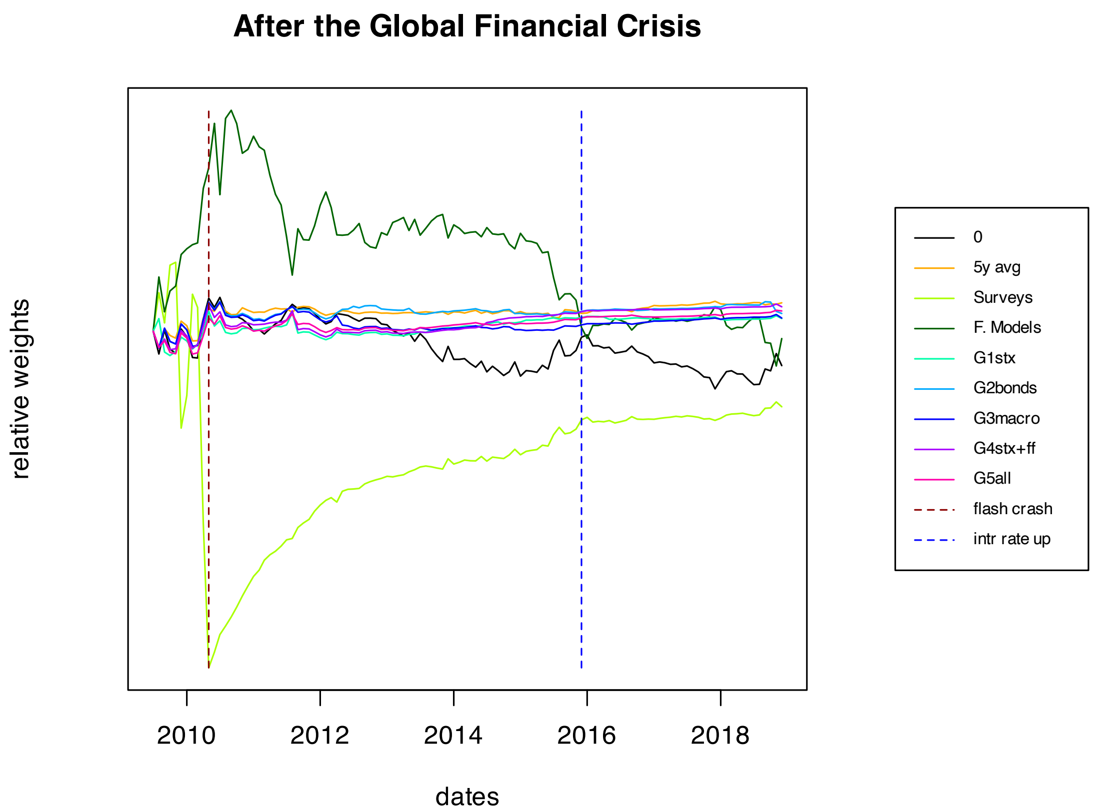

Figure 15. This figure depicts the weights obtained by applying the OL algorithm onto nine different predictors, over the sample period from July 2009 to December 2018. One of these predictors is the 0-predictor. Another one of these predictors is the prevailing 5-year average of the monthly market return. The other seven predictors are the same ones as in Figure 9. "flash crash" refers to the month of May 2010, during which the U.S. market experienced a trillion dollar flash crash. "intr rate up" refers to the month of December 2015, during which the U.S. Federal Reserve implemented the first interest rate increase in over 9 years.

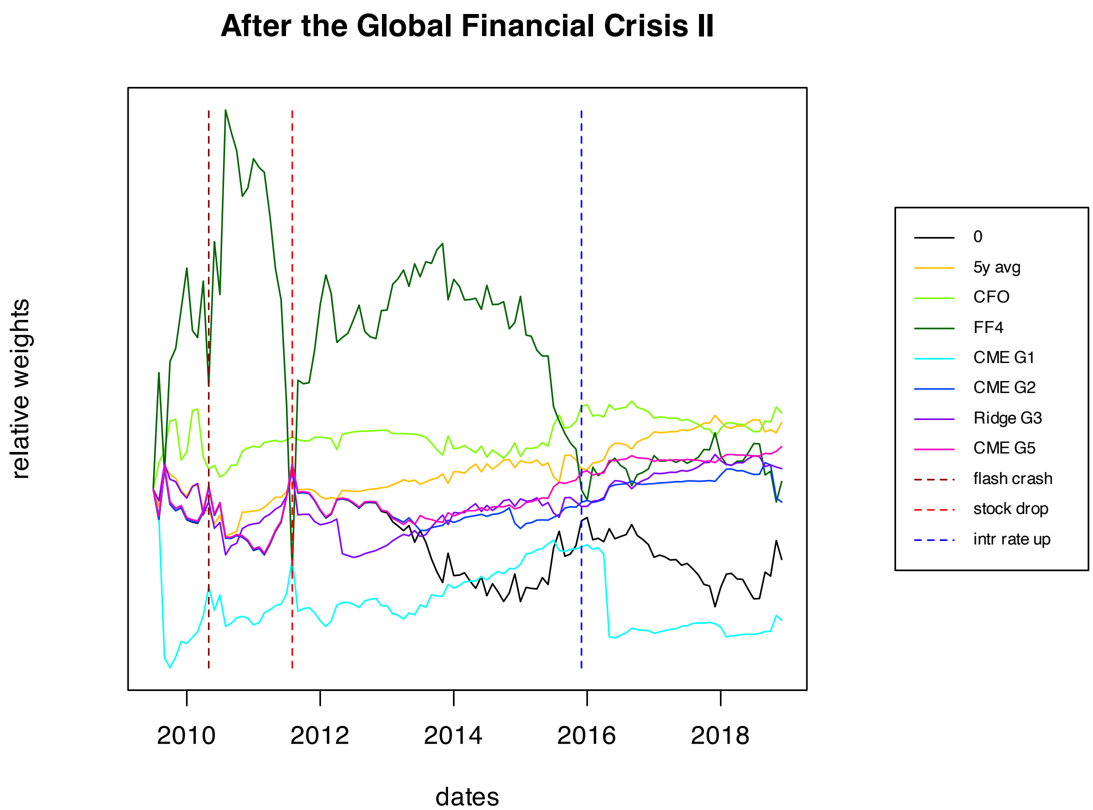

Figure 16. This figure depicts the weights obtained by applying the OL algorithm onto eight different predictors, over the sample period from July 2009 to December 2018. These eight predictors are the 0-predictor, the prevailing 5-year average of the monthly market return, the predictor based on the CFO survey, the FF4 predictor, the predictor which applies CME onto the predictive variables from G1stx, the predictor which applies CME onto the predictive variables from G2bonds, the predictor which applies the Ridge regression onto the predictive variables from G3macro, and the predictor which applies CME onto the predictive variables from G5all. "flash crash" refers to the month of May 2010, during which the U.S. market experienced a trillion dollar flash crash. "stock drop" refers to the month of August 2011, during which the U.S. market, as well as many other markets, experienced a sharp drop in stock prices. "intr rate up" refers to the month of December 2015, during which the U.S. Federal Reserve implemented the first interest rate increase in over 9 years.

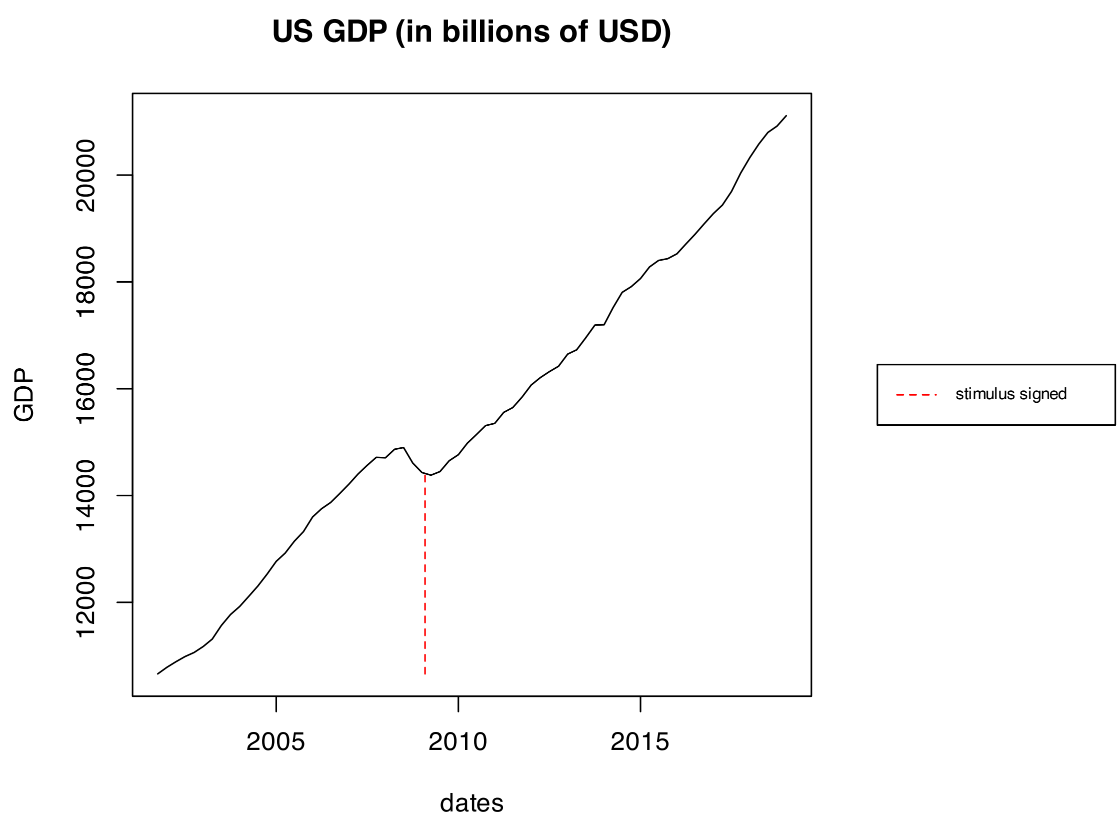

Figure 17. This figure depicts the level of the U.S. GDP, in billions of U.S. Dollars, from December 2001 until December 2018. "stimulus signed" refers to the month of February 2009, during which a stimulus package was signed into law by the U.S. president.

Predictor	R^2 **overall (%)**
cfo	2.00
liv	-11.35
ff3	-18.55
ff4	-7.03
ff5	-22.22
reg1	-431.10
las1	-4.31
rid1	-3.99
cme1	-1.87
reg2	-26.34
las2	-5.20
rid2	-4.55
cme2	-0.38
reg3	-34.15
las3	-3.96
rid3	-3.30
cme3	-6.65
reg4	-458.18
las4	-4.39
rid4	-3.24
cme4	-0.34
reg5	-578.61
las5	-5.10
rid5	-4.50
cme5	0.35
zero	0.02
5y avg	-1.46
G1stx	-1.70
G2bonds	-2.21
G3macro	-1.75
G4stx+ff	-1.60
G5all	-1.67
surveys	-1.88
ff	-12.41
80%cfo+20%cme5	2.11

Table I. This table contains the respective OOSCT R^2s of all considered predictors, evaluated on the time interval from December 2001 until December 2018. "reg" refers to the OLS regression. "las" refers to the Lasso regression. "rid" refers to the Ridge regression. The number corresponds to which group of predictive variables (see Section III) the statistical method is applied onto.

Predictor	R^2 **before crisis (%)**
cfo	7.65
liv	-3.11
ff3	-42.04
ff4	-5.77
ff5	-50.17
reg1	-168.33
las1	0.01
rid1	-0.79
cme1	6.64
reg2	-14.63
las2	1.79
rid2	1.18
cme2	5.94
reg3	-22.47
las3	1.35
rid3	7.80
cme3	2.08
reg4	-220.80
las4	-0.24
rid4	-0.20
cme4	5.36
reg5	-595.86
las5	0.18
rid5	0.58
cme5	6.57
zero	5.39
5y avg	1.64
G1stx	2.67
G2bonds	3.30
G3macro	6.84
G4stx+ff	2.16
G5all	3.08
surveys	4.21
ff	-27.27

Table II. This table contains the respective OOSCT R^2s of all considered predictors, evaluated on the time interval from December 2001 until November 2007. "reg" refers to the OLS regression. "las" refers to the Lasso regression. "rid" refers to the Ridge regression. The number corresponds to which group of predictive variables (see Section III) the statistical method is applied onto.

Predictor	R^2 during crisis (%)
cfo	2.84
liv	3.62
ff3	-11.90
ff4	-12.68
ff5	-15.36
reg1	-1113.05
las1	-11.21
rid1	-7.56
cme1	0.64
reg2	-44.52
las2	-14.86
rid2	-11.89
cme2	0.82
reg3	-29.64
las3	-10.62
rid3	-12.16
cme3	-17.48
reg4	-1134.31
las4	-11.21
rid4	-6.26
cme4	0.35
reg5	-1041.75
las5	-13.82
rid5	-9.02
cme5	0.17
zero	5.27
5y avg	-2.62
G1stx	-3.09
G2bonds	-5.52
G3macro	-6.94
G4stx+ff	-3.48
G5all	-4.00
surveys	3.35
ff	-11.01

Table III. This table contains the respective OOSCT R^2s of all considered predictors, evaluated on the time interval from December 2007 until June 2009. "reg" refers to the OLS regression. "las" refers to the Lasso regression. "rid" refers to the Ridge regression. The number corresponds to which group of predictive variables (see Section III) the statistical method is applied onto.

Predictor	R^2 **after crisis (%)**
cfo	3.01
liv	-20.38
ff3	-3.61
ff4	1.16
ff5	-4.68
reg1	-111.04
las1	2.81
rid1	1.58
cme1	-3.41
reg2	-15.02
las2	2.24
rid2	2.08
cme2	0.13
reg3	-37.21
las3	2.43
rid3	1.18
cme3	0.63
reg4	-126.02
las4	2.79
rid4	2.04
cme4	0.85
reg5	-233.38
las5	2.69
rid5	0.64
cme5	1.79
zero	-1.59
5y avg	2.49
G1stx	1.69
G2bonds	1.78
G3macro	1.69
G4stx+ff	2.44
G5all	2.11
surveys	-3.79
ff	0.71

Table IV. This table contains the respective OOSCT R^2s of all considered predictors, evaluated on the time interval from July 2009 until December 2018. "reg" refers to the OLS regression. "las" refers to the Lasso regression. "rid" refers to the Ridge regression. The number corresponds to which group of predictive variables (see Section III) the statistical method is applied onto.