HEC Paris

MSc in International Finance

Statistical Forecasting methods and Portfolio Allocation

By

Ricardo Furquim

ÉCOLE DES HAUTES ETUDES COMMERCIALES DE PARIS
Department of Finance

A dissertation submitted to HEC Paris in accordance with the requirements of the degree of MASTER OF SCIENCE in International Finance.

JUNE 2017

Advisor: Prof. Dr. Hugues Langlois

Abstract

Finance is normally seen as an unpredictable world where the more information you can guess, the more money you get. In this random scenario, market players are constantly seeking more robust and efficient methods to allocate wealth. Although least square regressions are a classical milestone, they may not perform well on current real world scenarios. On this paper we present some alternative statistical forecasting methods and we analyze their performance on portfolio allocation using real data.

Keywords: LASSO, Adaptive LASSO, Elastic net, Ridge penalization, Portfolio Allocation, GARCH.

Dedication and acknowledgements

To my family.

TABLE OF CONTENTS

List of Tables

Table **Page**

List of Figures

CHAPTER 1

INTRODUCTION

A substantial part of the innovation explosion we have been experiencing is propelled by the surge on computational capabilities, allied with better statistical models and more efficient algorithms. The core objective of this paper is to analyze the performance of learning and forecasting methods on a new context: financial market. First we present the portfolio allocation framework. On chapter 2 we briefly describe each forecasting algorithm. Finally on chapter 3 we merge both domains and we discuss the performance of these algorithms on real data.

Before digging on abstract concepts and models, we want to reinforce what Paul Wilmott and Emanuel Derman explain on their manifesto ([DW09]) : quantitative finance and applied mathematics are related, but must not be handled on the same manner. On their words "Building financial models is challenging and worthwhile: you need to combine the qualitative and the quantitative, imagination and observation, art and science, all in the service of finding approximate patterns in the behavior of markets and securities". Financial modelers must never forget, no matter how elegant or complex a model is, "we has not created the world, and it doesn't satisfy our equations".

1.1 Portfolio Allocation

Quantitative portfolio allocation methods usually follows a similar template ([Mar52], [BKM08], [Ang14], [LL17]). The idea is to maximize your **preference metric**, subject to your **constraints**, considering your **view of the world**. So, before we define an investment strategy, we should ask some questions:

1. **Preference metric:** What do I care about? What do I want to maximize? What is my

investment horizon? Will a outflow of capital be necessary during my investment horizon? Do I like risks or I prefer to avoid them? If I do not like risks, how much would make me bear a given risk?

2. **Constraints:** Do I have legal constraints? Financial constraints? What is the universe of assets/instruments available for me? May I leverage my investments? Sell short?

3. **View of the world:** Markets are inefficient? How returns are generated?

1.1.1 Utility function

Functions usually may be seen as a mathematical entity that transforms inputs in outputs. In our framework, the utility function $U(v)$, as the name suggests, transform investor's final wealth (v) on the utility of this wealth. Although this function may change a lot for different investors, they should take into account at least two concepts:

1. **Return:** Any rational rational investor prefers more money than less. The utility is a increasing function, the utility of 10 million dollar is expected to be bigger than the utility of 10 bucks.

$$\frac{\partial U}{\partial v} \geq 0$$

2. **Risks:** If a rational investor has to chose between two different opportunities with the same expected return, which one would he/she prefer?

   - Investment **A** offers a certain return r_{CE};
   - Investment **B** offers a stochastic return r_B with mean $\mathbb{E}[r_B] = r_{CE}$ and variance $\sigma_B^2 > 0$;

   The answer depends on investor's level of risk aversion:

   - Risk Averse: Investor prefers investment **A** because you have the same expected payoff, but no risk. This behavior implies a **concave** utility function: $\frac{\partial^2 U}{\partial v^2} < 0$;
   - Risk Neutral: Investor is indifferent as both opportunities have the same expected payoff: $\frac{\partial^2 U}{\partial v^2} = 0$;
   - Risk Seeking: Investor prefers investment **B**, as it has the same expected return of A and provides a non null probability of getting better returns: $\frac{\partial^2 U}{\partial v^2} \geq 0$;

Considering this framework an interesting question arises: why a risk-averse investor would invest on risky assets? How much would he/she demand for bearing risks? The answer of these questions defines the **certainty equivalent**.

1.1.2 Certainty Equivalent and mean-variance preference

Defining $\overline{V}_T := \mathbb{E}[V_T]$ and applying Taylor's formula around $\overline{V}_T$, we get

$$U(V_T) \simeq U(\overline{V}_T) + U'(\overline{V}_T)\ (V_T - \overline{V}_T) + \frac{1}{2}\, U''(\overline{V}_T)\ (V_T - \overline{V}_T)^2$$

$$\mathbb{E}[U(V_T)] \simeq U(\overline{V}_T) + \frac{1}{2}\, U''(\overline{V}_T)\ \sigma^2_{V_T}, \quad \text{with } \sigma^2_{V_T} := \mathbb{E}\left[(V_T - \overline{V}_T)^2\right] \tag{1.1}$$

Applying the mean value theorem ([Rud87]), there exist $p > 0$ such that $U(\overline{V}_T - p) = \mathbb{E}[U(V_T)]$. Intuitively, we define the certainty equivalent of V_T as $V_{eq} := \mathbb{E}[V_T] - p$. Using again Taylor's expansion:

$$U(\overline{V}_T - p) \simeq U(\overline{V}_T) - p\, U'(\overline{V}_T) \tag{1.2}$$

Combining equations 1.1 and 1.2:

$$p = \frac{1}{2}\sigma^2_{V_T} A(\overline{V}_T), \text{where } A(\overline{V}_T) := -\frac{U''(\overline{V}_T)}{U'(\overline{V}_T)} \tag{1.3}$$

Considering a constant risk-aversion $A = A(\overline{V}_T)$, we finally get the **mean variance preference:** on this framework, maximize the expected utility is equivalent to the following problem:

$$\max \left\{ \mathbb{E}[V_T] - \frac{A}{2}\, \sigma^2_{V_T} \right\} \tag{1.4}$$

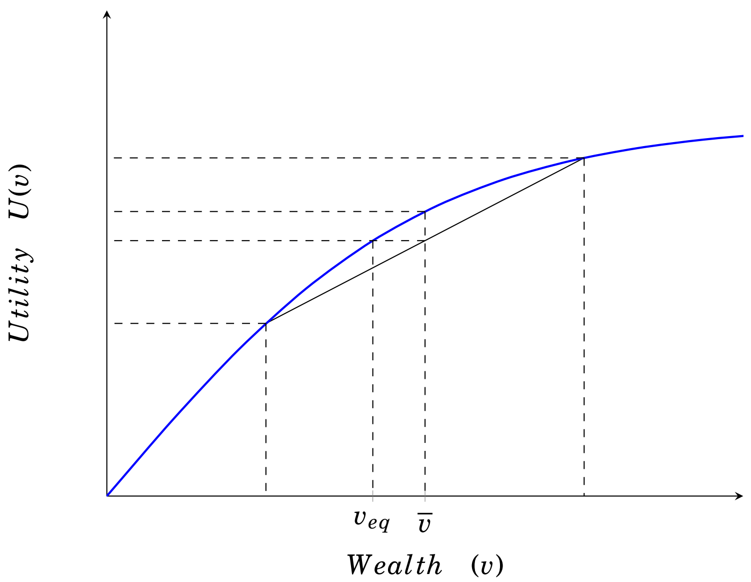

Figure 1.1: Wealthy

1.2 Markowitz Portfolio Selection

Now we are almost ready define the optimization problem we want to solve, i.e. determining the best investment for a given investor. But first we need to consider the following definitions and assumptions:

- Investment universe: n risky assets S_t^i, $i \in \{1,..,n\}$, with known initial prices S_0^i, but random returns $r_i = S_T^i/S_0^i$. One risk-free asset S_t^0, which grows at a known constant rate $r_f = S_T^0/S_0^0$;

- Discrete time horizon: we invest all our initial wealth (V_0) at $t = 0$ and we hold the portfolio until $t = T$. We allocate a proportion w_i of our wealth in each risk asset, and $1 - \sum_{i=0}^n w_i$ on the risk-free asset. The **strategy** is defined by the vector of weights w.

- Self-financing portfolio: there is no inflow or outflow of capital between $t = 0$ and $t = T$.

- The expectation of returns $\mathbb{E}[r]$ and its covariance matrix Σ exists and are finite;

- The risk premium vector is, by definition, $\mu = \mathbb{E}[r] - r_f$;

- Constant absolute risk aversion A;

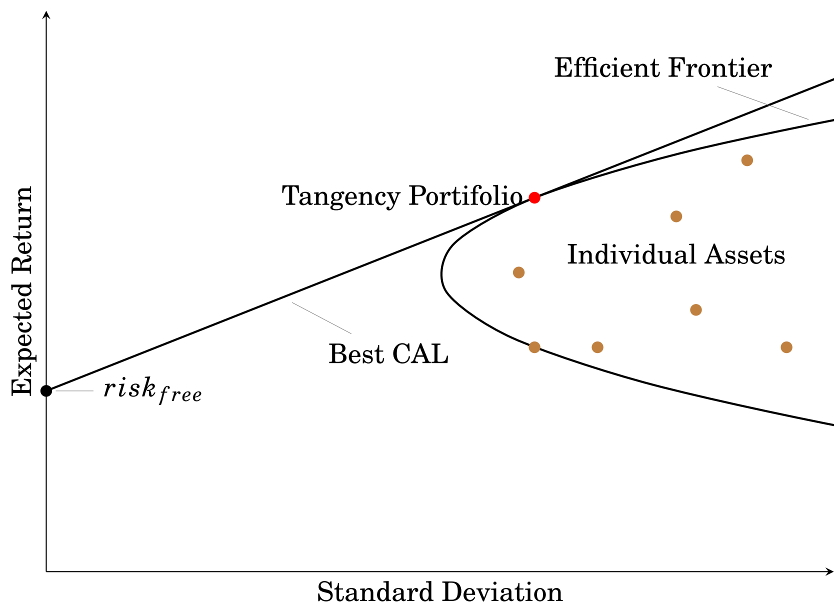

Figure 1.2: Hiperbole

1.2.1 Maximizing Mean-Variance metric

Under the aforesaid assumptions and definitions, we may write the total return of the portfolio $r_P := V_T/V_0$ as:

$$\begin{aligned} r_P = V_T/V_0 &= \sum_{i=0}^{n} w_i\, r_i + \left(1 - \sum_{i=0}^{n} w_i\right) r_f \\ &= r + \sum_{i=0}^{n} w_i\,(r_i - r_f) \\ &= r + w^T\,(r - r_f) \end{aligned}$$

Then,

$$\mathbb{E}[r_P] = r_f + w^T\,\mu \quad \text{and} \quad \mathbb{V}ar[r_P] = w^T\,\Sigma\,w \tag{1.5}$$

Rewriting problem 1.4,

$$\begin{aligned} w^* := \arg\max_{w} \quad & \mathbb{E}[r_p] - \frac{1}{2} A\,\mathbb{V}ar[r_p] \\ \arg\max_{w} \quad & r_f + w^T\,\mu - \frac{1}{2} A\,w^T\,\Sigma\,w \end{aligned}$$

Finally, we may find that "the best strategy" is determined by:

$$w^* = \frac{\Sigma^{-1}\,\mu}{A} \tag{1.6}$$

1.2.2 Out-of-sample Portfolio Selection Procedure

Looking at equation 1.6, investing seems a quite straightforward job. You just have three tasks:

- Determine your level of risk aversion A;
- Estimate the correlation matrix Σ;
- Estimate the risk premium vector.

But, in practice, those tasks are not as simple as they may appear at first glance. Risk aversion is a discretionary decision, but it may changes over time. Covariance matrix may present singularities and heteroskedasticity. And last but not least, to estimate the risk premium we usually need to rely on a time-series predictive model:

$$r_{t+1} - r_{f,t+1} = \mu + \theta\,X_t + \epsilon_{t+1} \tag{1.7}$$

Where, $(r_{t+1})_{t\in\mathbb{N}}$ is the vector of returns on t, which is know only on $t+1$. $\left(r_{f,t+1}\right)_{t\in\mathbb{N}}$ is the risk-free rate on for period $t+1$ known at time t. X_t are all available predictors at time t and θ its predictive coefficients. We are going to analyze statistical methods we may use to solve this problem on chapter 2.

Chapter 2

Statistical Methods

As detailed on chapter 1, under Markowitz' assumptions, optimal investing consists on applying eq. (1.6). Risk-premium and covariance matrix, however, are not directly observable. Both needs to be estimated through an statistical model. Moreover, the kind of prediction error you should expect is deeply model-dependent. On this chapter we are going to briefly detail statistical methods for estimating risk-premium and variance.

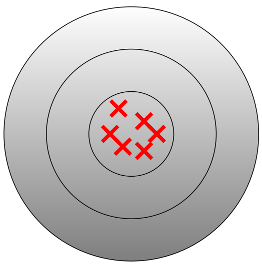

Figure 2.1: Low bias, low variance

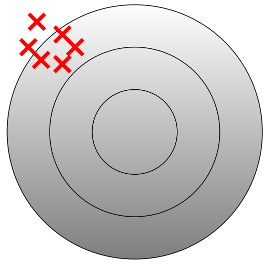

Figure 2.2: High bias, low variance

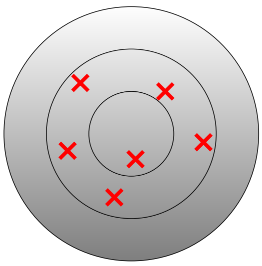

Figure 2.3: Low bias, high variance

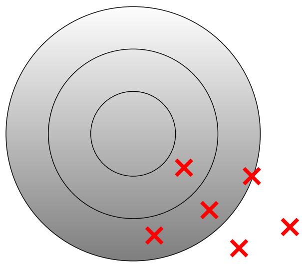

Figure 2.4: High bias, high variance

As detailed on [SW11], selecting statistical models and estimating parameters are always subject to two kind of error: bias and variance. The more complex the model you use to fit your data, the better your results over *in-sample*data, but the bigger the risk of over-fitting and getting poor results on *out-of-sample* data. We are always restricted by a bias-variance trade-off.

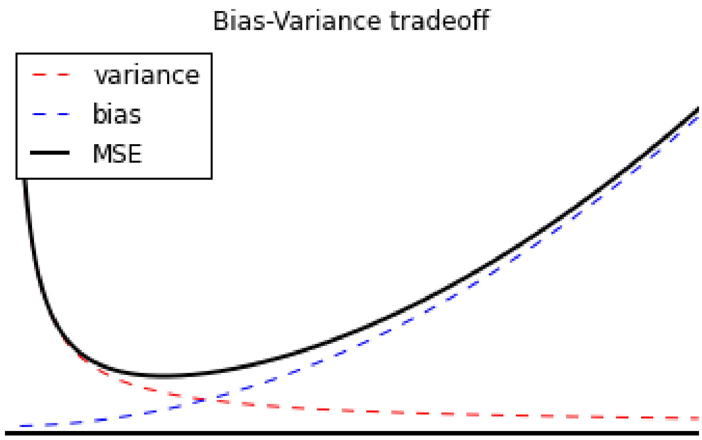

Figure 2.5: MSE and bias-variance trade off.

2.1 GLS - Generalized least square

Generalized Least Squares (GLS) was first described by Alexander Aiken in 1934 ([Ait35]). This technique allows estimating the unknown parameters in a linear regression model even when there is a certain degree of correlation between the residuals. It is relevant to reinforce that in those cases the ordinary least squares and weighted least squares can be statistically inefficient, or even give misleading inferences.

Consider a typical linear regression problem: $\{y_i, x_{ij}\}_{i=1,...,n}$ is the observed data; with $y = (y1,...,yn)^T$ and $X = (x_1^T,...,x_n^T)^T$ being,respectively, the response and the predictor values vectors.

The model assumes that the conditional mean of y given X is a linear function of X, whereas the conditional variance of the error term given X is a known nonsingular matrix Ω. In mathematical terms:

$$y = X\beta + \epsilon, \quad \mathbb{E}[\epsilon|X] = 0, \quad Var[\epsilon|X] = \Omega \tag{2.1}$$

where $\beta \in \mathbb{R}^k$ is the regression coefficient (a vector of unkown constants that must be estimated from the data).

GLS estimates β by minimizing the squared Mahalanobis length ([DMJRM00]) of the residual vector:

$$\hat{\beta}_{GLS} = argmin\left((y - X\beta)\Omega^{-1}(y - X\beta)\right) \tag{2.2}$$

Since the objective is a quadratic form in β, the estimator has thus an explicit formula:

$$\hat{\beta}_{GLS} = (X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}y \tag{2.3}$$

The GLS is unbiased, consistent, efficient, and asymptotically normal:

$$\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{d} \mathcal{N}\left(0,(X^T\Omega^{-1}X)^{-1}\right) \tag{2.4}$$

One can notice that GLS is equivalent to applying ordinary least squares to a linearly transformed version of the data. Intuition for the previous sentence can be found by factorizing $\Omega = CC^T$, using, for instance Cholesky decomposition ([Haz01]). Then, multiplying the original equation by C^{-1} we obtain the equivalent linear model $y^* = X^*$, where $y^* = C^{-1}y$, $X^* = C^{-1}X$ and $\epsilon^* = C^{-1}\epsilon$. In this new model, $Var[\epsilon^*|X] = C^{-1}\Omega(C^{-1})^T = I$ (the identity matrix) and β can thus be estimated by Ordinary Least Squares (OLS), which requires the same minimization as in equation (2.2).

This has the effect of standardizing the scale of the errors and "de-correlating" them. Since OLS is applied to data with homoscedastic errors, the Gauss–Markov theorem ([Ode83]) applies, and therefore the GLS is the best linear unbiased estimator for β. For a more mathematical and detailed approach please refer to [Ait35] and [Woo15].

2.1.1 Feasible GLS - FGLS

Instead of assuming the structure of heteroskedasticity, we may estimate the structure of heteroskedasticity from OLS. This method is called Feasible GLS (FGLS). First, we estimate $\hat{\Omega}$ from OLS, and, second, we use $\hat{\Omega}$ instead of Ω:

$$\hat{\beta}_{GLS} = (X^T\hat{\Omega}a^{-1}X)^{-1}X^T\hat{\Omega}^{-1}y \tag{2.5}$$

There are many ways to estimate FGLS. One flexible approach is discussed in [Woo15].

2.2 Ridge Regression

A ridge regression is an ordinary least squares estimation, with a constraint on the sum of the squared coefficients. One of the original motivations for the ridge regression was to ensure matrix invertibility when estimating a linear regression. However, it is most frequently used to reduce the variance of parameter estimates. The constraint is applied to the regression through the value chosen for the tuning parameter λ. As λ increases the regression parameters $\beta_1, \ldots, \beta_p$ are forced to smaller values with lower variance. Some explanatory variables may be completely excluded from the model as their parameters are forced to zero.

The choice of λ can be made through minimizing prediction error or cross validation. A ridge regression estimation is dependent upon the scale and intercept of the model. As a consequence, variables are typically centered and standardized prior to model estimation.

The original publication of the ridge regression was Hoerl and Kennard (1970)([HK70]), and discussions can be found in several texts ([FHT10]; [You12]). There are several variations of the classical ridge regression. These often involve multiple λ's or different approaches to the standardization of the explanatory variables.

In mathematical terms, the classical ridge regression estimator can be defined as:

$$\hat{\beta}_{Ridge} = argmin\left(\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2\right) \tag{2.6}$$

Equation 2.6 is equivalent to:

$$\hat{\beta}_{Ridge} = argmin\left(\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2\right) \tag{2.7}$$

subject to $\sum_{j=1}^{p} \beta_j^2 \leq s^*$, with s^* having an one-to-one correspondence to λ. Moreover, it is equivalent to:

$$\hat{\beta}_{Ridge} = (X^T X + \lambda I)^{-1} X^T Y \tag{2.8}$$

From (2.8) one can see that a ridge regression can be expressed as modifying an OLS by adding λ to variance terms in the variance-covariance matrix of the explanatory variables. This modification impacts both the variance and covariance terms in the variance-covariance matrix, but disproportionately down weights the covariance terms.

As the ridge regression is equivalent to applying a prior to beta parameters with zero covariance and positive variance (refer to [HK70] and [FHT10]), the ridge estimation also reduces the impact of the covariance terms between explanatory variables. Intuition for the previous sentence can be found by considering the variance-covariance matrix for OLS parameters $\left(\sigma^2(X^T X)^{-1}\right)$.

Unlike OLS, the ridge estimator relies on a tuning parameter λ. The choice of this tuning parameter can pose a dilemma. Like any model parameter, the choice is an attempt to minimize a loss function, such as mean squared error which is unobservable. As a consequence there are several commonly used criteria that people use as an approximation. Examples include: minimize forecast error and various cross validation approaches; with Generalized Cross Validation being a common choice.

2.3 LASSO - Least absolute shrinkage and selection operator

A lasso regression is very similar to a ridge regression. The difference is that the constraint is applied to the sum of absolute parameter estimates rather than the sum of their squares.

This minor difference has significant repercussions in terms of the resulting estimates. Due to the absolute penalty, the lasso is less sensitive to the standardization of explanatory variables and has a stronger tendency to push coefficients to zero, giving a easier interpretable model ([Tib96]).

The lasso estimator is defined as:

$$\hat{\beta}_{Lasso} = argmin\left(\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p}|\beta_j|\right) \tag{2.9}$$

that is equivalent to:

$$\hat{\beta}_{Lasso} = argmin\left(\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2\right) \tag{2.10}$$

subject to $\sum_{j=1}^{p}|\beta_j|^2 \leq t$, where $t > 0$ is a tuning parameter which controls the amount of shirnkage applied to the estimates. Smaller values of t (larger valeus of λ) result intuitively in more zero β coefficients.

For a more detailed analysis and a historic review of the method refer to, respectively, to [Tib96] and [Tib11]

2.4 ALASSO - Adaptive LASSO

Fan and Li ([FL01]) studied a class of penalization methods including the lasso one. They showed that the lasso method leads to estimators that may suffer an appreciable bias. Furthermore they conjectured that the oracle properties do not hold for the lasso. Hence Zou ([Zou06]) proposes to consider the following modified lasso criterion, called adaptive lasso:

$$\hat{\beta}_{ALasso} = \left(\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 + \lambda_n \sum_{j=1}^{p}\frac{|\beta_j|}{|\tilde{\beta}_j|^\gamma}\right) \tag{2.11}$$

where $(\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^T$ denotes a preliminary estimate of β_0 (the ones estimated by OLS $\hat{\beta}^{ols}$, for instance).

This modification allows to produce sparse solutions more effectively than lasso. Precisely, Zou ([Zou06]) shows that the adaptive lasso enjoys the **oracle properties**.

DEFINITION : Denote by $\hat{\beta}(\delta)$ the coefficient estimator produced by a fitting procedure δ. We call δ an oracle procedure if $\hat{\beta}(\gamma)$ (asymptotically) has the following oracle properties:

- Identifies the right subset model, $\{j : \hat{\beta}_j \neq 0\} = \mathscr{A}$
- Has the optimal estimation rate, $\sqrt{n}\left(\hat{\beta}_{\mathscr{A}} - \hat{\beta}_{\mathscr{A}^*}\right) \xrightarrow{d} \mathscr{N}(0, \Sigma^*)$, where $\Sigma*$ is the covariance matrix knowing the true subset model

2.5 Elastic net

Although the lasso has shown success in many situations, it has some limitations. Consider the following three scenarios.

(a) In the "$p > n$" case (high-dimensional data with few examples), the lasso selects at most n variables before it saturates, because of the nature of the convex optimization problem. This seems to be a limiting feature for a variable selection method. Moreover, the lasso is not well defined unless the bound on the L1-norm of the coefficients is smaller than a certain value.

(b) If there is a group of variables among which the pairwise correlations are very high, then the lasso tends to select only one variable from the group and does not care which one is selected.

(c) For usual "$n > p$" situations, if there are high correlations between predictors, it has been empirically observed that the prediction performance of the lasso is dominated by ridge regression ([Tib96]).

Hui Zou and Trevor Hastie proposed then in their paper [ZH05]) a new regularization: the elastic net. This new technique aims to work as well as the lasso whenever the lasso does the best, and fix the problems that were highlighted above, i.e. it should mimic the ideal variable selection method in scenarios (a) and (b), especially with microarray data, and it should deliver better prediction performance than the lasso in scenario (c).

Similar to the lasso, the elastic net simultaneously does automatic variable selection and continuous shrinkage, and it can select groups of correlated variables. In the authors words "it is like a stretchable fishing net that retains 'all the big fish'". Simulation studies and real data examples show that the elastic net often outperforms the lasso in terms of prediction accuracy.

Mathematically, the elastic net adds a quadratic part to the LASSO's classical penalty:

$$\hat{\beta}_{ENet} = argmin\left(|y - X\beta|^2 + \lambda_2|\beta|^2 + \lambda_1|\beta|_1\right), \tag{2.12}$$

where $|\beta|^2 = \sum_{j=1}^{p} \beta_j^2$ and $|\beta|_1 = \sum_{j=1}^{p} |\beta_j|$

Notice that this quadratic part makes the loss function strictly convex, and it therefore has a unique minimum. Also observe that the method includes the lasso and ridge regression ($\lambda_2 = 0$ and $\lambda_1 = 0$ respectively). To understand better the relations among those methods, please observe the Figure 2.6 that shows the penalty contour of those techniques.

The naive version of elastic net method finds an estimator in a two-stage procedure : first for each fixed λ_2 it finds the ridge regression coefficients, and then does a lasso type shrinkage. This two-stage procedure (a ridge-type direct shrinkage followed by a lasso-type thresholding) can be view at Figure 2.7 . Besides, the figure shows the operational characteristics of the three penalization methods in an orthogonal design, which helps to gain a more intuitive comprehension of the methods.

This kind of estimation incurs,though, a double amount of shrinkage, which leads to increased bias and poor predictions. To improve the prediction performance, the authors rescale the

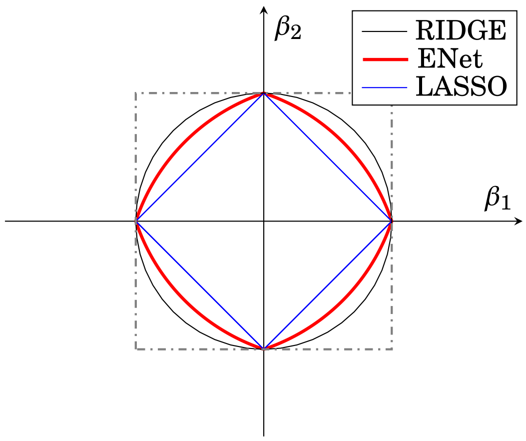

Figure 2.6: (in black, the shape of the RIDGE penalty;red, contour of the ENet penalty; blue, contour of the elastic LASSO with$\lambda_1 = \lambda_2 = 0.5$): we see that singularities at the vertices and the edges are strictly convex.

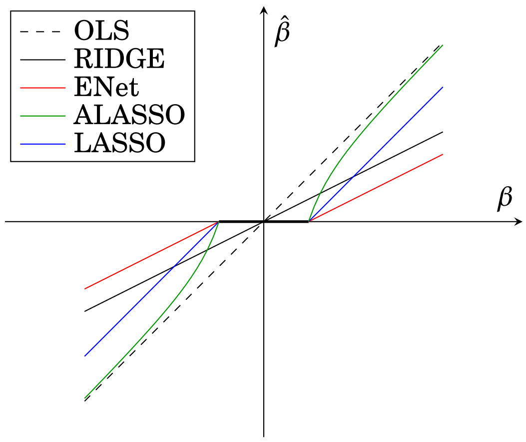

Figure 2.7: Exact solutions for the lasso, ridge and the naive elastic net in an orthogonal design: the shrinkage parameters are $\lambda_1 = 2$ and $\lambda_2 = 1$

coefficients of the naive version of elastic net by multiplying the estimated coefficients by $(1+\lambda_2)$ ([ZH05]). For a deeper description of this optimization and the implementations please refer to [ZH05] and[FHT10].

2.6 ARMA - Autoregressive moving average

Since ARMA is a mixed model, having a basic knowledge about the autoregressive and the moving average models is fundamental to obtain a complete understanding of ARMA behaviour and applications. Considering that, we show at the next subsubsections quick reviews of those models main concepts.

2.6.1 Autoregressive (AR) Models

The autoregressive model is basically an extension of the random walk that includes terms further back in time, it is essentially a regression model where the previous terms are the predictors.

DEFINITION : A time series model, x_t, is an autoregressive model of order p, AR(p), if:

$$x_t = \alpha_1 x_{t-1} + ... + \alpha_p x_{t-p} + \omega_t \tag{2.13}$$

$$x_t = \sum_{i=1}^{p} \alpha_i x_{t-i} + \omega_t \tag{2.14}$$

where $\{\omega_t\}$ is white noise and $\alpha_i \in \mathbb{R}$, with $\alpha_p \neq 0$ for a p-order autoregressive process. It is thus straightforward to make predictions with the AR(p) model, for any time t, as once we have the α_i coefficients determined, our estimate simply becomes:

$$\hat{x}_t = \alpha_1 x_{t-1} + ... + \alpha_p x_{t-p} \tag{2.15}$$

Hence we can make n-step ahead forecasts by producing $\hat{x}_t, \ldots, \hat{x}_{t+n}$.

2.6.2 Moving Average (MA) Models

A Moving Average model is similar to an Autoregressive model, except that instead of being a linear combination of past time series values, it is a linear combination of the past white noise terms.Intuitively, this means that the MA model sees such random white noise "shocks" directly at each current value of the model. This is in contrast to an AR(p) model, where the white noise "shocks" are only seen indirectly, via regression onto previous terms of the series.

DEFINITION : $\{x_t\}$ is a moving average model of order q, MA(q), if:

$$x_t = \omega_t + \beta_1 \omega_{t-1} + ... + \beta_q \omega_{t-q} \tag{2.16}$$

where ω_t is white noise with $\mathbb{E}(\omega_t) = 0$ and variance σ^2.

Now that we have considered autoregressive processes and moving average processes, we know that:

- The former model considers its own past behavior as inputs for the model and as such attempts to capture market participant effects, such as momentum and mean-reversion in stock trading.

- The latter model is used to characterize "shock" information to a series, such as a surprise earnings announcement or unexpected event.

An ARMA model attempts, hence, to capture both of these aspects when modeling financial time series.

The ARMA(p,q) model is a linear combination of two linear models and thus is itself still linear:

DEFINITION : A time series model, $\{x_t\}$, is an autoregressive moving average model of order p, q, ARMA(p,q), if:

$$x_t = \alpha_1 x_{t-1} + ... + \alpha_p x_{t-p} + \omega_t + \beta_1 \omega_{t-1} + ... + \beta_q \omega_{t-q} \tag{2.17}$$

$$x_t = \sum_{i=1}^{p} \alpha_i x_{t-i} + \sum_{i=1}^{q} \beta_i x_{t-i} + \omega_t \tag{2.18}$$

where ω_t is white noise with $\mathbb{E}(\omega_t) = 0$ and variance σ^2.

We can straightforwardly see that by setting $p \neq 0$ and $q = 0$ we recover the AR(p) model. Similarly if we set $p = 0$ and $q \neq 0$ we recover the MA(q) model.

One of the key features of the ARMA model is that it is parsimonious and redundant in its parameters. That is, an ARMA model will often require fewer parameters than an AR(p) or MA(q) model alone. Note also that an ARMA model does not take into account volatility clustering, a key empirical phenomena of many financial time series.

2.7 GARCH - Generalized autoregressive conditional heteroskedasticity

The Autoregressive Conditional Heteroskedastic (ARCH) class of models was introduced by Engle ([Eng82]) to accommodate the possibility of serial correlation in volatility. The ARCH(q) model considers the conditional variance as time dependent $Var(\omega_t|\omega_{t-1}) = h_t$.

$$h_t = c + \sum_{i=1}^{q} \alpha_i \omega_{t-i}^2 \tag{2.19}$$

The Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) is an extended ARCH model proposed by Bollerslev ([Bol86]) where ω_t^2 takes the form:

$$\omega_t^2 = v_t^2 h_t \tag{2.20}$$

where $\sigma_v^2 = 1$ is basically a white noise process and, more important:

$$h_t = c + \sum_{i=1}^{p} \alpha_i^{'} h_{t-1} + \sum_{i=1}^{q} \beta_i^{'} \omega_{t-1}^2 \tag{2.21}$$

As it can be easily seen from (2.21), any GARCH model where $p = 0$, i.e., a GARCH(0,q) , becomes an ARCH(q) model. Therefore, in a GARCH model, as it incorporates mean reversion, the dynamics of ω_t^2 can then be explained through past volatility shocks ω_{t-i}^2.

For a more detailed and mathematical approach of ARCH and GARCH models refer to [Eng01] and [BD16].

CHAPTER 3

EMPIRICAL APPLICATION

As highlighted on chapter 1, our investment strategy (section 1.2.2) depends on the statistical approach we chose for estimating parameters. Although we have several different methods, real data does not have to comply [DW09]. In this chapter we analyze the performance of our investment strategy on historical data using the statistical methods we brifly discussed on chapter 2.

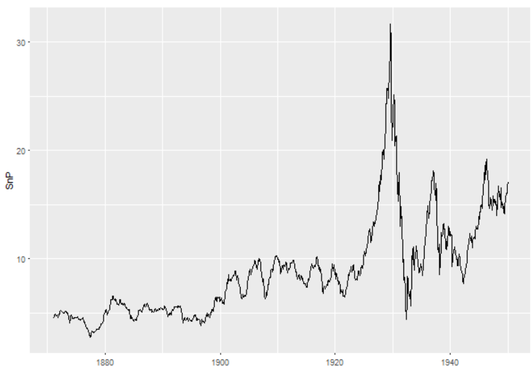

Figure 3.1: S&P index from January 1871 up to January 1950.

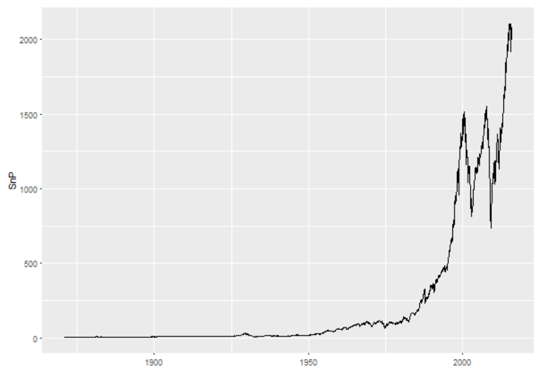

Figure 3.2: S&P index from January 1871 up to December 2015.

3.1 Data

We considered monthly returns of S&P index from January 1871 up to December 2015. We also considered a several predictors: TODO:Cite data source (ask Prof. Langlois)

- **log(DP)**: Log dividend-price ratio: log of the 12-month moving sum of dividends paid divided by the price.

- **log(DY)**: Log dividend yield: log of the 12-month moving sum of dividends divided by the lagged price.
- **log(EP)**: Log earnings-price ratio: log of the 12-month moving sum of earnings divided by the price.
- **log(DE)**: Log dividend-payout ratio: log of the 12-month moving sum of dividends divided by the 12-month moving sum of earnings.
- **SVAR**: Stock variance: monthly sum of squared daily returns.
- **BM**: Book-to-market ratio: ratio of the accounting book value to the market price.
- **NTIS**: Net equity expansion: ratio of the 12-month moving sum of net equity issues to the total end-of-year market capitalization of the market.
- **TBL**: Treasury bill rate: interest rate on a three-month Treasury bill (secondary market).
- **LTY**: Long term yield: long-term government bond yield.
- **LTR**: Long-term retur: return on long-term government bonds.
- **TMS**: Term spread: long-term yield minus the Treasury bill rate.
- **DFY**: Default yield spread: difference between BAA- and AAA-rated corporate bond yields.
- **DFR**: Default return spread: long-term corporate bond return minus the long term government bond return.
- **INFL**: Inflation: We use 1-month lagged inflation to account for the delay in CPI releases.

3.2 Returns: Time series approach

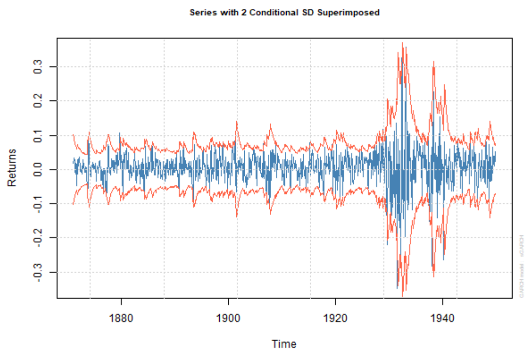

Figure 3.3: Index return and with 2 conditional standard deviations from January 1871 up to January 1950.

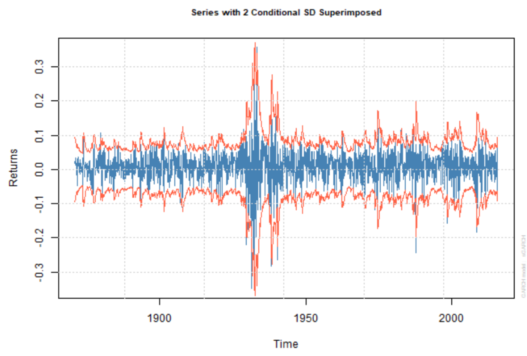

Figure 3.4: Index return and with 2 conditional standard deviations from January 1871 up to January 1950.

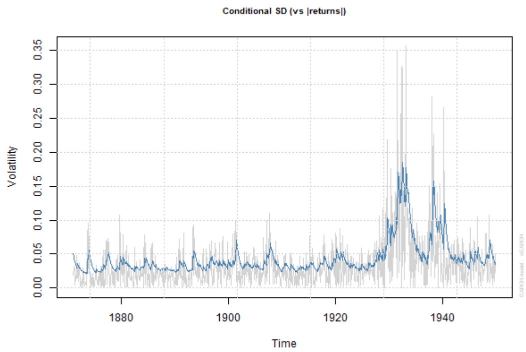

Figure 3.5: GARCH Conditional volatility vs |returns| from from January 1871 up to January 1950.

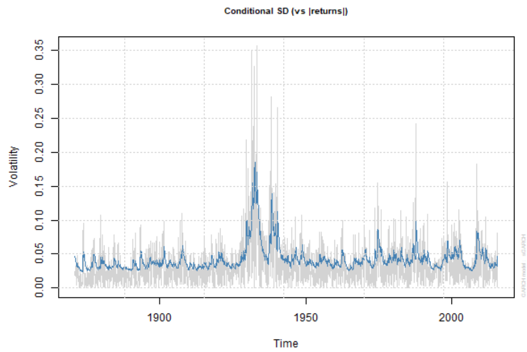

Figure 3.6: GARCH Conditional volatility vs |returns| from from January 1871 up to December 2015.

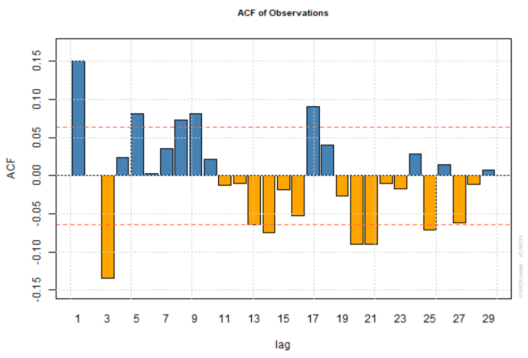

Figure 3.7: ACF of returns from January 1871 up to January 1950.

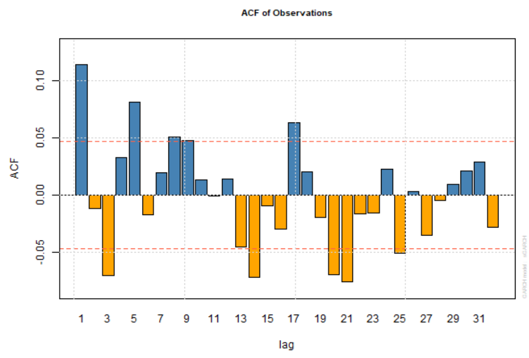

Figure 3.8: ACF of returns from January 1871 up to December 2015.

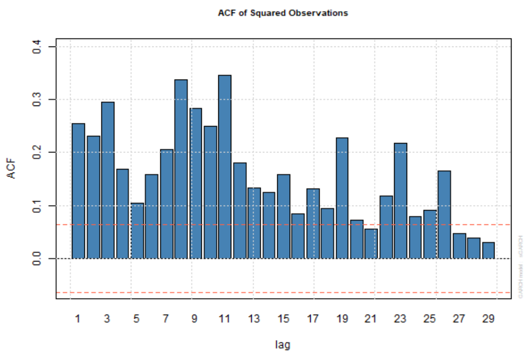

Figure 3.9: ACF of squared returns from January 1871 up to January 1950.

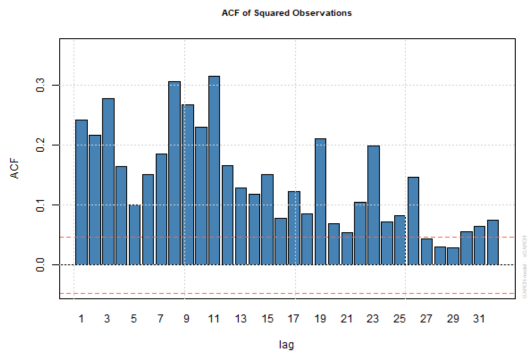

Figure 3.10: ACF of squared returns from January 1871 up to December 2015.

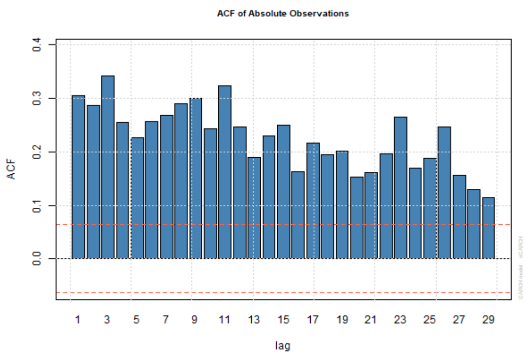

Figure 3.11: ACF of absolute returns from January 1871 up to January 1950.

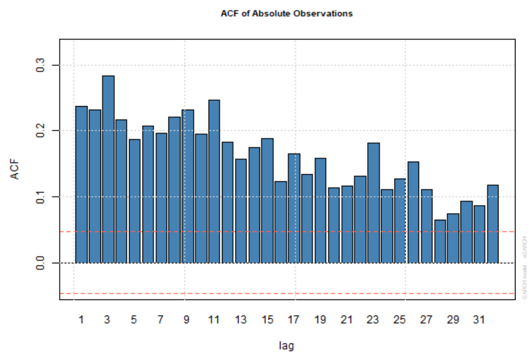

Figure 3.12: ACF of absolute returns from January 1871 up to December 2015.

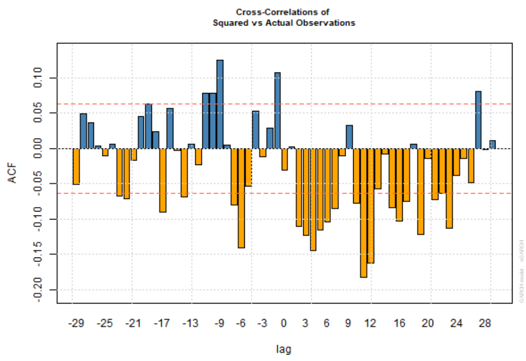

Figure 3.13: Cross correlation from January 1871 up to January 1950.

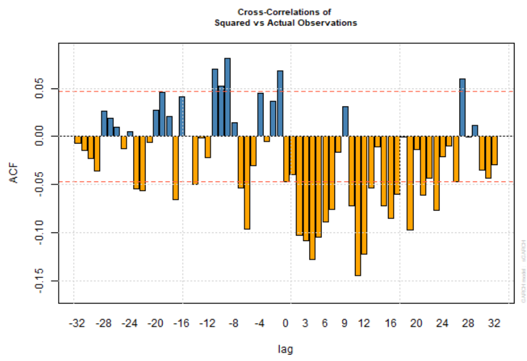

Figure 3.14: Cross correlation from January 1871 up to December 2015.

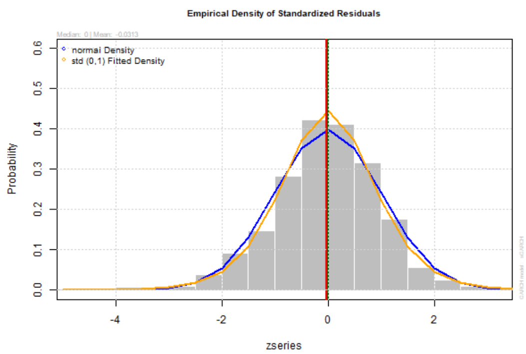

Figure 3.15: Empirical Density of Standardized Residuals from January 1871 up to January 1950.

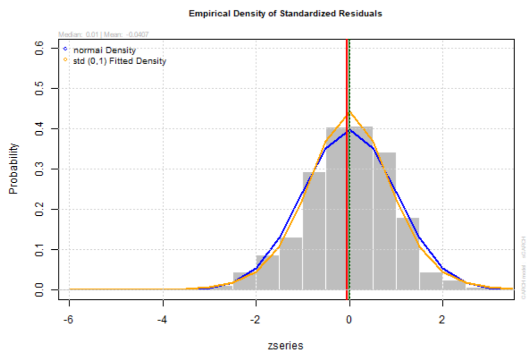

Figure 3.16: Empirical Density of Standardized Residuals from January 1871 up to December 2015.

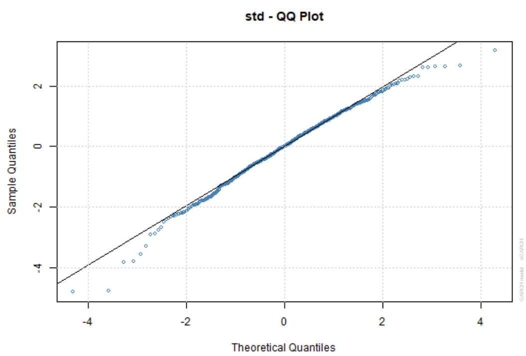

Figure 3.17: QQPlot for Standardized Residuals from January 1871 up to January 1950.

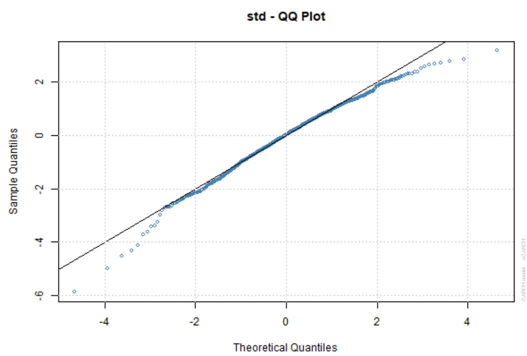

Figure 3.18: QQPlot for Standardized Residuals from January 1871 up to December 2015.

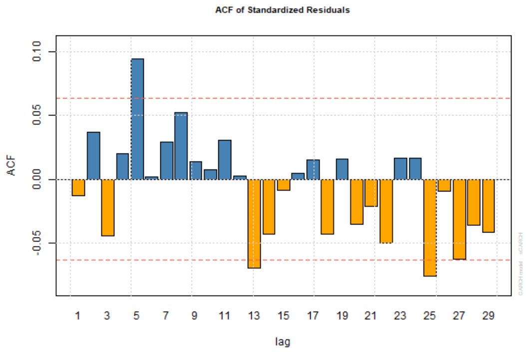

Figure 3.19: ACF of Standardized Residuals from January 1871 up to January 1950.

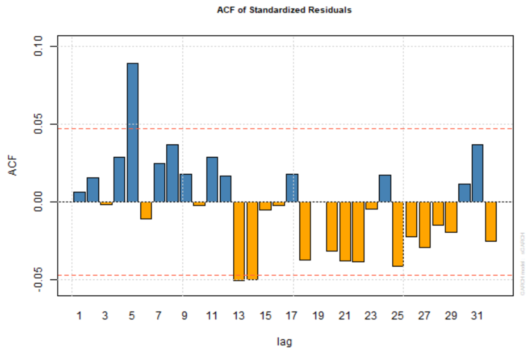

Figure 3.20: ACF of Standardized Residuals from January 1871 up to January 1950

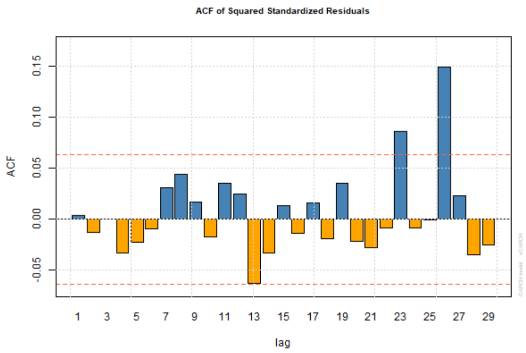

Figure 3.21: ACF of Squared Standardized Residuals from January 1871 up to January 1950.

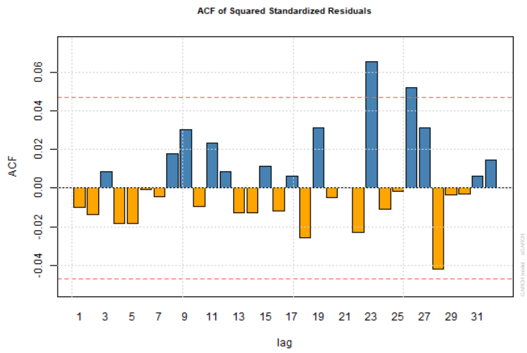

Figure 3.22: ACF of Squared Standardized Residuals from January 1871 up to December 2015.

TODO: Colocar a tabela dos testes statiticos do garch aqui

3.3 Portfolio performance

TODO: Colocar a tabela dos CE usando o GARCH pra prever a variancia

Forecasting method	RA = 3	RA = 5
LM-log_DP	7.134	6.014
LM-log_DY	5.928	5.290
LM-log_EP	7.009	5.940
LM-log_DE	7.737	6.382
LM-SVAR	6.271	5.500
LM-BM	4.519	4.452
LM-NTIS	7.457	6.212
LM-TBL	8.564	6.880
LM-LTY	7.314	6.131
LM-LTR	8.290	6.711
LM-TMS	9.321	7.329
LM-DFY	7.552	6.268
LM-DFR	8.298	6.721
LM-INFL	8.413	6.788
LM-ER.L1	7.652	6.333
LM-ER.L2	7.739	6.380
LM-ER.L3	4.141	4.216
LM-ER.L12	7.547	6.266
LM-ALL	-12.056	-5.506
LASSO	8.527	6.859
RIDGE	8.527	6.859
A.LASSO	8.527	6.859
ENET_0.5	8.527	6.859

Table 3.1: Results table

3.4 Risk premium: a machine learning approach

3.4.1 LASSO

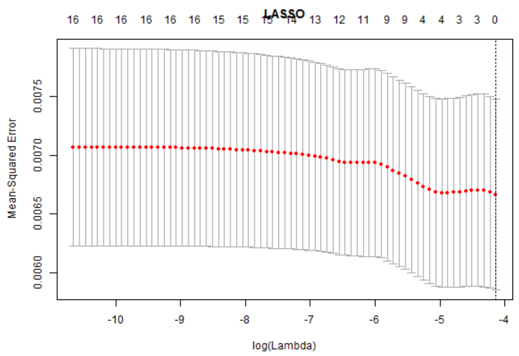

Figure 3.23: Cross validation for monthly data from Jan-1871 up to Jan-1950.

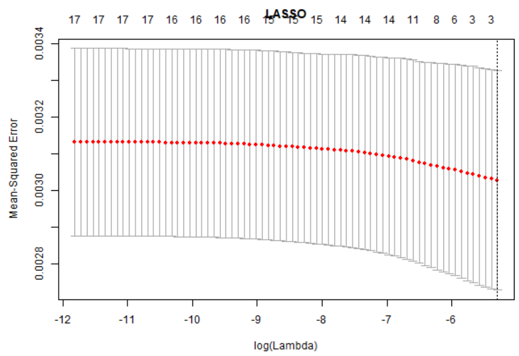

Figure 3.24: Cross validation for monthly data from Jan-1871 up to Dec-2015.

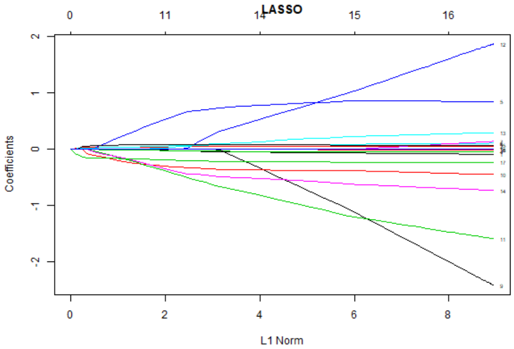

Figure 3.25: Coefficients vs $\mathbb{L}_1$-norm for LASSO model using data from Jan-1871 up to Jan-1950.

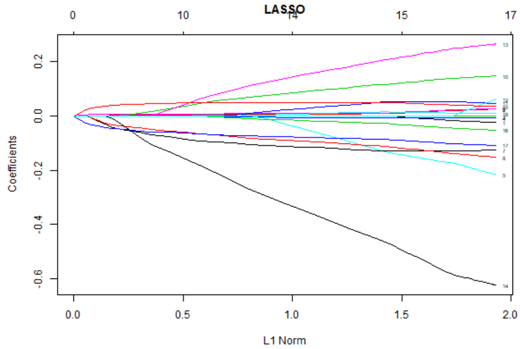

Figure 3.26: Coefficients vs $\mathbb{L}_1$-norm for LASSO model using data from Jan-1871 up to Dec-2015.

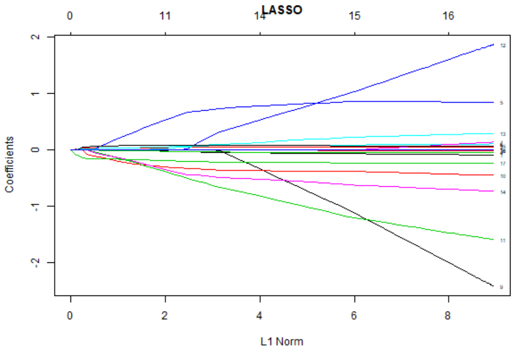

Figure 3.27: LASSO lambda 1950

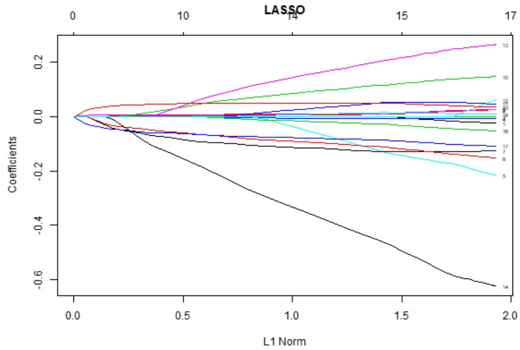

Figure 3.28: LASSO lambda 2015.

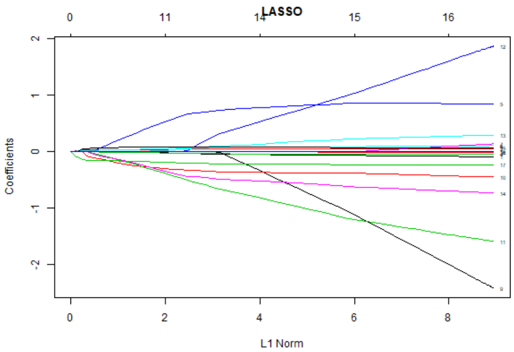

Figure 3.29: Coefficients vs explained variance for LASSO model using data from Jan-1871 up to Jan-1950.

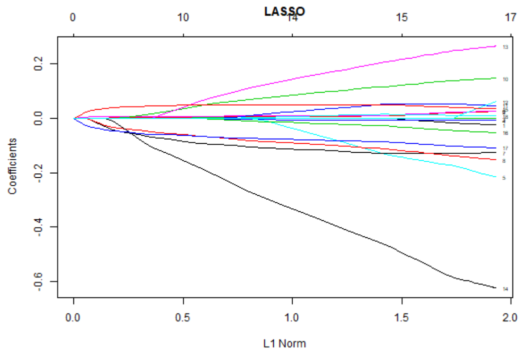

Figure 3.30: Coefficients vs explained variance for LASSO model using data from Jan-1871 up to Dec-2015.

3.4.2 RIDGE

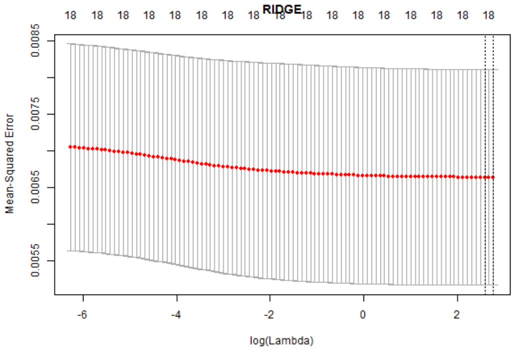

Figure 3.31: RIDGE cv 1950.

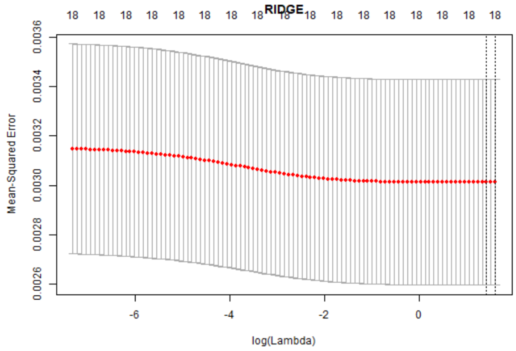

Figure 3.32: RIDGE cv 2015.

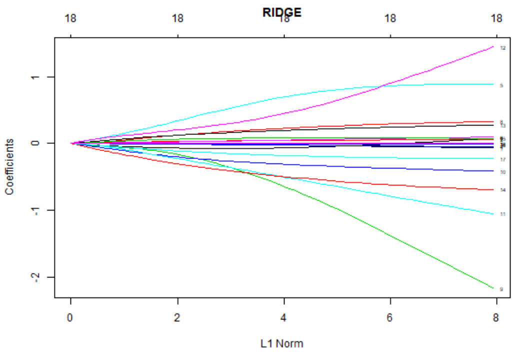

Figure 3.33: RIDGE norm 1950.

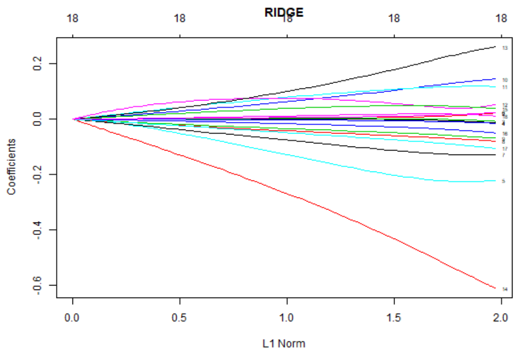

Figure 3.34: RIDGE norm 2015.

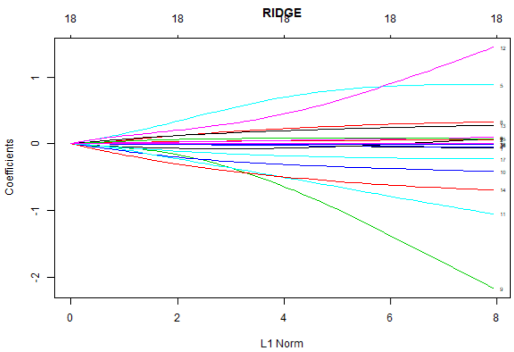

Figure 3.35: RIDGE lambda 1950.

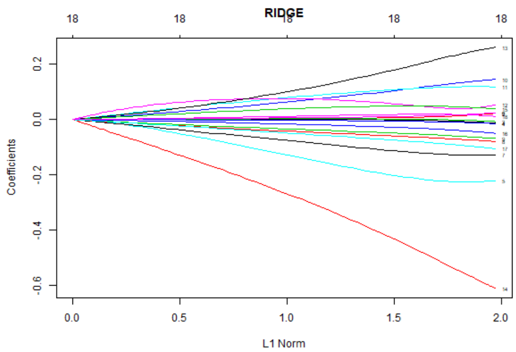

Figure 3.36: RIDGE lambda 2015.

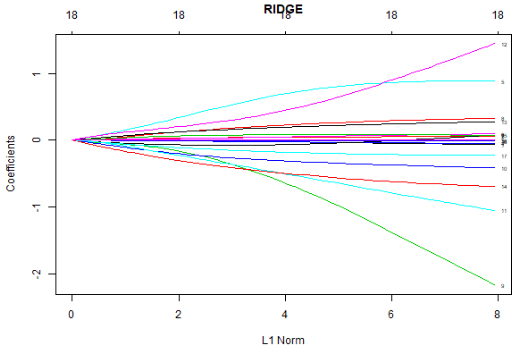

Figure 3.37: RIDGE dev 1950.

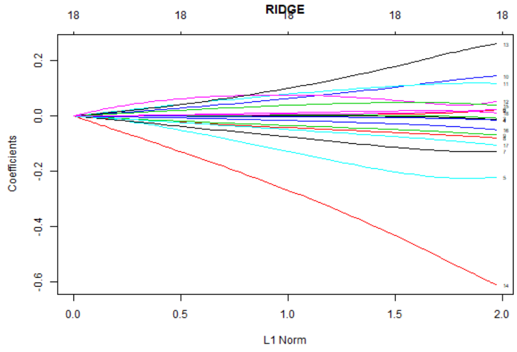

Figure 3.38: RIDGE dev 2015.

3.4.3 ENET

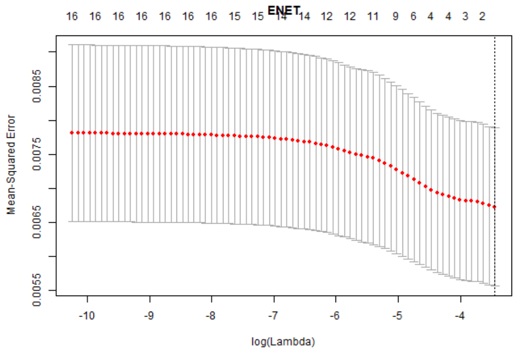

Figure 3.39: ENET cv 1950.

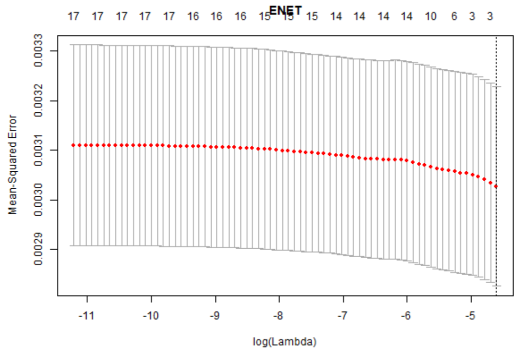

Figure 3.40: ENET cv 2015.

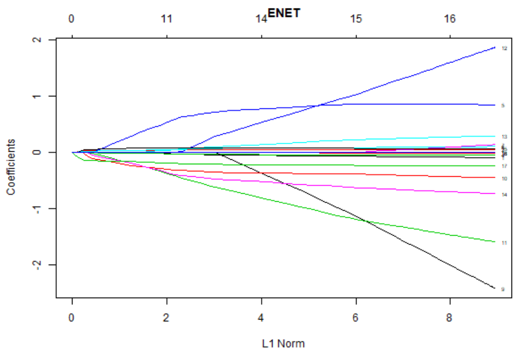

Figure 3.41: ENET norm 1950.

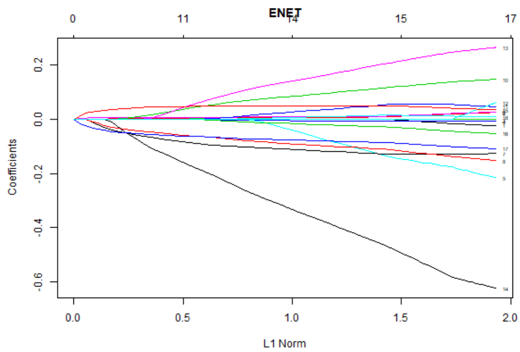

Figure 3.42: ENET norm 2015.

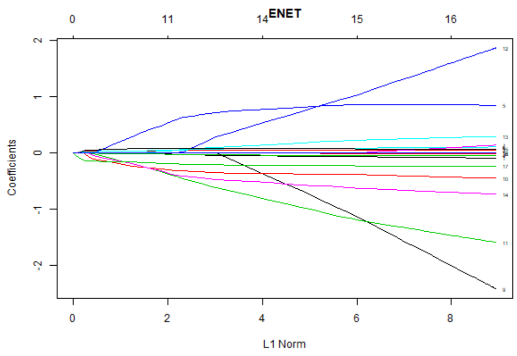

Figure 3.43: ENET lambda 1950.

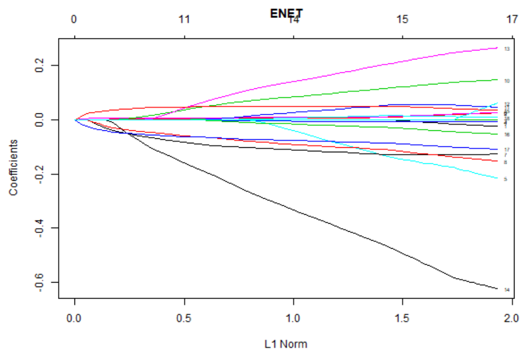

Figure 3.44: ENET lambda 2015.

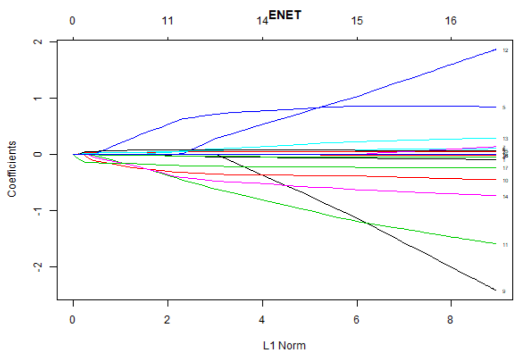

Figure 3.45: ENET dev 1950.

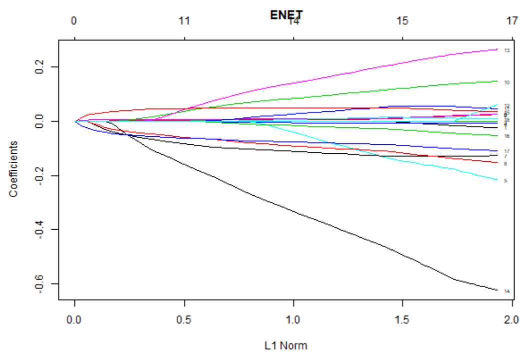

Figure 3.46: ENET dev 2015.

3.4.4 ALASSO

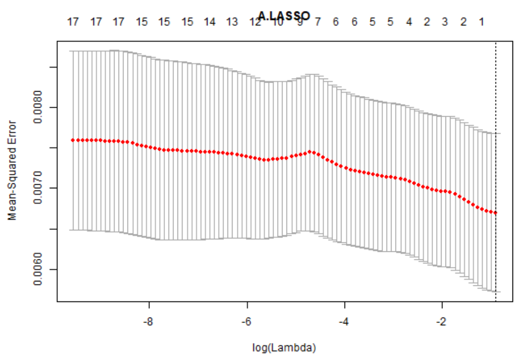

Figure 3.47: ALASSO CV 1950.

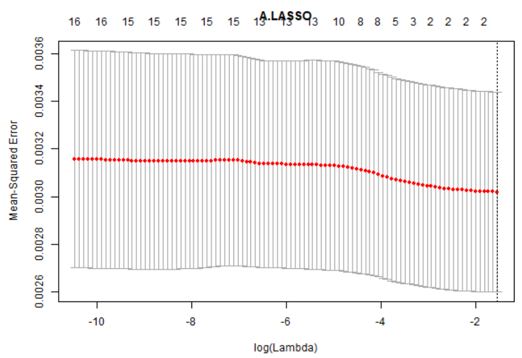

Figure 3.48: ALASSO cv 2015.

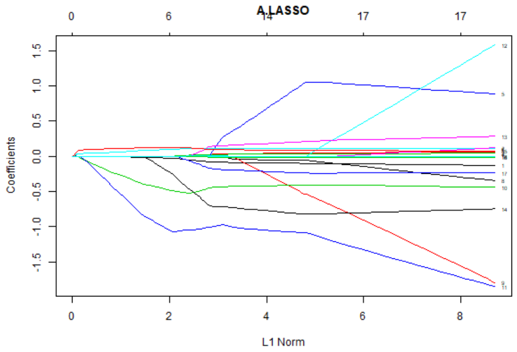

Figure 3.49: ALASSO norm 1950.

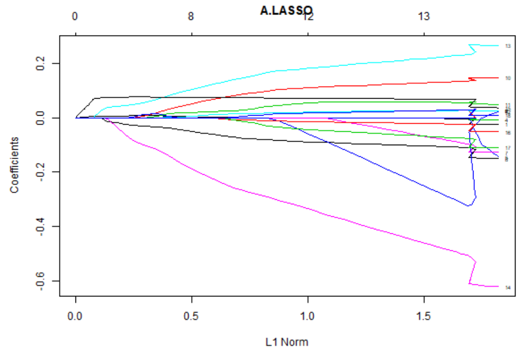

Figure 3.50: ALASSO norm 2015.

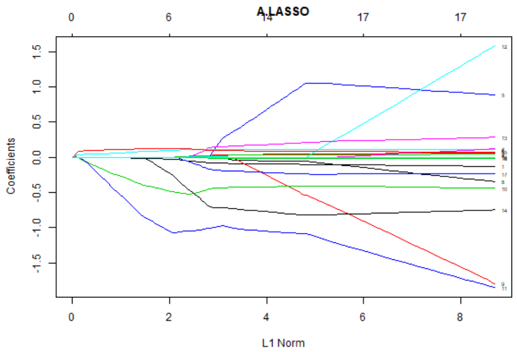

Figure 3.51: ALASSO lambda 1950.

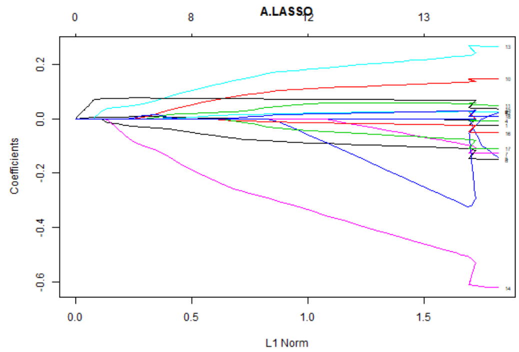

Figure 3.52: ALASSO lambda 2015.

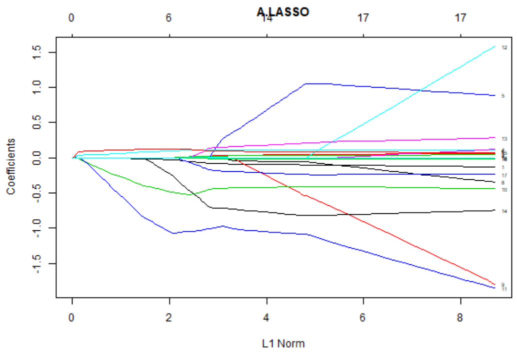

Figure 3.53: ALASSO dev 1950.

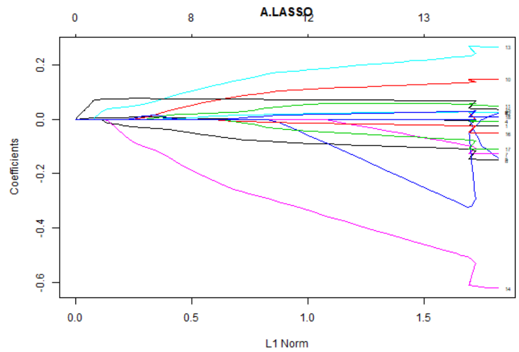

Figure 3.54: ALASSO dev 2015.

Chapter 4

Conclusions and Further Research

Quantitative Finance aims to describe real-market dynamics through mathematical equations. In other words, it is about translating a set of hypotheses about market and agent behavior into numerical prediction. As we have seen there are several different approaches, which one with pros and cons.

Albeit LASSO, Adaptive LASSO, RIDGE and Elastic Net methods seems much more robust than a simple least-square regression, it has not been translated on a much better certainty equivalent. As pointed out, the percentage of explained variance is too small for all regression methods, corroborating with efficiency market hypothesis. It certainly the Variable selection and forecasting

As we also verified is that returns presents strong heteroskedasticity. Statistically, GARCH surely offers a better volatility estimation and forecasting than sample covariance. But it also does not always implies a bigger certainty equivalent, especially when used together with the regularization methods. There are some interesting approaches that try to combine variance forecasting and penalized regression [AK16] [WRC16].

Quantitative investing is a vivid research theme. Finance, however, is not only about mathematics, it is also about people [DW09]. Anyone applying quantitative analysis on finance must always remember that, as people, market's behavior is constantly changing.

APPENDIX A

APPENDIX A

Begins an appendix

Table A.1: My caption

Risk Premium	Variance forecasting method			
Forecasting method	Sample Var		GARCH(1,1)	
	RA = 3	RA = 5	RA = 3	RA = 5
LM-log_DP	7.134	6.014	7.453	6.207
LM-log_DY	5.928	5.290	5.925	5.290
LM-log_EP	7.009	5.940	7.090	5.991
LM-log_DE	7.737	6.382	7.700	6.363
LM-SVAR	6.271	5.500	6.206	5.463
LM-BM	4.519	4.452	3.018	3.554
LM-NTIS	7.457	6.212	6.688	5.752
LM-TBL	8.564	6.880	8.785	7.015
LM-LTY	7.314	6.131	7.281	6.113
LM-LTR	8.290	6.711	7.964	6.517
LM-TMS	9.321	7.329	9.191	7.253
LM-DFY	7.552	6.268	7.497	6.237
LM-DFR	8.298	6.721	8.203	6.667
LM-INFL	8.413	6.788	8.572	6.886
LM-ER.L1	7.652	6.333	7.494	6.241
LM-ER.L2	7.739	6.380	7.940	6.503
LM-ER.L3	4.141	4.216	3.397	3.770
LM-ER.L12	7.547	6.266	7.715	6.368
LM-ALL	-12.056	-5.506	-1.7082	-8.522
LASSO	8.527	6.859	8.484	6.836
RIDGE	8.527	6.859	8.484	6.836
A.LASSO	8.527	6.859	8.484	6.836
ENET_0.5	8.527	6.859	8.484	6.836

Table A.2: PART 1

GARCH Model Fit

Conditional Variance Dynamics				
GARCH Model	sGARCH(1,1)			
Mean Model	ARFIMA(1,0,1)			
Distribution	std			

Optimal Parameters				
	Estimate	Std, Error	t value	Pr(>|t|)
mu	0	0.00136	5.35688	0
ar1	0.08201	0.119617	0.68561	0.492961
ma1	0.168475	0.118448	1.42236	0.154921
omega	0.000073	0.000024	3.01779	0.002546
alpha1	0.166888	0.034522	4.83423	0.000001
beta1	0.800838	0.035149	22.78385	0
shape	8.224285	2.028826	4.05372	0.00005

Robust Standard Errors				
	Estimate	Std, Error	t value	Pr(>|t|)
mu	0.007286	0.001495	4.87464	0.000001
ar1	0.08201	0.087392	0.93841	0.348031
ma1	0.168475	0.099839	1.68747	0.091514
omega	0.000073	0.000022	3.34553	0.000821
alpha1	0.166888	0.032735	5.0982	0
beta1	0.800838	0.033019	24.25395	0
shape	8.224285	2.134533	3.85297	0.000117

LogLikelihood	1741.927

Information Criteria	
Akaike	-3.6602
Bayes	-3.6243
Shibata	-3.6603
Hannan-Quinn	-3.6465

Table A.3: PART 2

Weighted Ljung-Box Test on Standardized Residuals				
	Statistic	P-value		
Lag[1]	0.1548	0.69403		
Lag[2*(p+q)+(p+q)-1][5]	4.1689	0.04276		
Lag[4*(p+q)+(p+q)-1][9]	8.6213	0.03365		
Degrees of freedom	2			
H0	No serial correlation			
Weighted Ljung-Box Test on Standardized Squared Residuals				
	Statistic	P-value		
Lag[1]	0.01484	0.903		
Lag[2*(p+q)+(p+q)-1][5]	0.68158	0.9264		
Lag[4*(p+q)+(p+q)-1][9]	1.93729	0.9124		
Degrees of freedom	2			
Weighted ARCH LM Tests				
	Statistic	Shape	Scale	P-Value
ARCH Lag[3]	4.41E-05	0.5	2.000	0.9947
ARCH Lag[5]	1.16E+00	1.44	1.667	0.6855
ARCH Lag[7]	1.71E+00	2.315	1.543	0.7794
Nyblom stability Test				
Joint Statistic	3.7663			
Individual Statistics:				
mu	0.3984			
ar1	1.0071			
ma1	1.2957			
omega	0.5301			
alpha1	0.7504			
beta1	1.0679			
shape	0.257			
	Asymptotic Critical Values			
	10%	5%	1%	
Joint Statistic	1.69	1.9	2.35	
Individual Statistic	0.35	0.47	0.75	
Sign Bias Test				
	t-value	Prob Sig		
Sign Bias	0.9765	0.3291		
Negative Sign Bias	0.1353	0.8924		
Positive Sign Bias	0.1996	0.8419		
Joint Effect	2.3192	0.5089		
Adjusted Pearson Goodness-of-Fit Test				
	Group	Statistic	P-value(g-1)	
1	20	19.43	0.4298	
2	30	23.65	0.7462	
3	40	30.9	0.8192	
4	50	49.68	0.4461	

Table A.4: My caption

GARCH Model Fit

Conditional Variance Dynamics				
GARCH Model	sGARCH(1,1)			
Mean Model	ARFIMA(1,0,1)			
Distribution	std			
Optimal Parameters				
	Estimate	Std. Error	t value	Pr(>|t|)
mu	0.008764	0.000954	9.185740	0.000000
ar1	-0.022857	0.187352	-0.122000	0.902898
ma1	0.148857	0.185528	0.802340	0.422355
omega	0.000084	0.000022	3.809370	0.000139
alpha1	0.137867	0.022983	5.998650	0.000000
beta1	0.820473	0.026174	31.347100	0.000000
shape	8.529831	1.569887	5.433400	0.000000
Robust Standard Errors				
	Estimate	Std. Error	t value	Pr(>|t|)
mu	0.008764	0.001093	8.01785	0
ar1	-0.022857	0.158463	-0.14424	0.885309
ma1	0.148857	0.162991	0.91328	0.361094
omega	0.000084	0.00002	4.16775	0.000031
alpha1	0.137867	0.02311	5.96573	0
beta1	0.820473	0.024436	33.57618	0
shape	8.529831	1.761054	4.84359	0.000001
LogLikelihood	3156.245			
Information Criteria				
Akaike	-3.624			
Bayes	-3.602			
Shibata	-3.624			
Hannan-Quinn	-3.6159			

Table A.5: My caption

Weighted Ljung-Box Test on Standardized Residuals				
	Statistic	P-value		
Lag[1]	0.07628	0.78241		
Lag[2*(p+q)+(p+q)-1][5]	3.77233	0.115322		
Lag[4*(p+q)+(p+q)-1][9]	10.18022	0.007717		
Degrees of freedom	2			
H0	No serial correlation			
Weighted Ljung-Box Test on Standardized Squared Residuals				
	Statistic	P-value		
Lag[1]	0.1717	0.6786		
Lag[2*(p+q)+(p+q)-1][5]	0.8593	0.8906		
Lag[4*(p+q)+(p+q)-1][9]	1.5878	0.9481		
Degrees of freedom	2			
Weighted ARCH LM Tests				
	Statistic	Shape	Scale	P-Value
ARCH Lag[3]	1.25E-01	0.5	2.000	0.7233
ARCH Lag[5]	9.42E-01	1.44	1.667	0.7502
ARCH Lag[7]	1.06E+00	2.315	1.543	0.9039
Nyblom stability Test				
Joint Statistic	5.8605			
Individual Statistics:				
mu	0.6834			
ar1	3.6944			
ma1	3.669			
omega	0.6664			
alpha1	0.3356			
beta1	0.6775			
shape	0.1339			
	Asymptotic Critical Values			
	10%	5%	1%	
Joint Statistic	1.69	1.9	2.35	
Individual Statistic	0.35	0.47	0.75	
Sign Bias Test				
	t-value	Prob Sig		
Sign Bias	2.5636	0.0104437		
Negative Sign Bias	0.1621	0.8712142		
Positive Sign Bias	1.009	0.3131328		
Joint Effect	18.66	0.0003214		
Adjusted Pearson Goodness-of-Fit Test				
	Group	Statistic	P-value(g-1)	
1	20	42.94	0.00132	
2	30	49.19	0.011019	
3	40	68.26	0.002576	
4	50	73.22	0.014063	

BIBLIOGRAPHY

[Ait35] AC Aitkin.
On least squares and linear combination of observations.
In *Proc. Roy. Soc. Edin. A*, volume 55, pages 42–48, 1935.

[AK16] Francesco Audrino and Simon D Knaus.
Lassoing the har model: A model selection perspective on realized volatility dynamics.
Econometric Reviews, 35(8-10):1485–1521, 2016.

[Ang14] Andrew Ang.
Asset management: A systematic approach to factor investing.
Oxford University Press, 2014.

[BD16] Peter J Brockwell and Richard A Davis.
Introduction to time series and forecasting.
springer, 2016.

[BKM08] Zvi Bodie, Alex Kane, and Alan J Marcus.
Investments (ed.), 2008.

[Bol86] Tim Bollerslev.
Generalized autoregressive conditional heteroskedasticity.
Journal of econometrics, 31(3):307–327, 1986.

[CL+13] A Chatterjee, SN Lahiri, et al.
Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap.
The Annals of Statistics, 41(3):1232–1259, 2013.

[DMJRM00] Roy De Maesschalck, Delphine Jouan-Rimbaud, and Désiré L Massart.
The mahalanobis distance.
Chemometrics and intelligent laboratory systems, 50(1):1–18, 2000.

[DW09] Emanuel Derman and Paul Wilmott.
The financial modelers' manifesto.

In *SSRN: http://ssrn. com/abstract*, volume 1324878, 2009.

[EHJ+04] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al.
Least angle regression.
The Annals of statistics, 32(2):407–499, 2004.

[Eng82] Robert F Engle.
Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation.
Econometrica: Journal of the Econometric Society, pages 987–1007, 1982.

[Eng01] Robert Engle.
Garch 101: The use of arch/garch models in applied econometrics.
The Journal of Economic Perspectives, 15(4):157–168, 2001.

[FHT01] Jerome Friedman, Trevor Hastie, and Robert Tibshirani.
The elements of statistical learning, volume 1.
Springer series in statistics Springer, Berlin, 2001.

[FHT10] Jerome Friedman, Trevor Hastie, and Rob Tibshirani.
Regularization paths for generalized linear models via coordinate descent.
Journal of statistical software, 33(1):1, 2010.

[FL01] Jianqing Fan and Runze Li.
Variable selection via nonconcave penalized likelihood and its oracle properties.
Journal of the American statistical Association, 96(456):1348–1360, 2001.

[Gef14] Deborah Gefang.
Bayesian doubly adaptive elastic-net lasso for var shrinkage.
International Journal of Forecasting, 30(1):1–11, 2014.

[GS07] VW Griffis and JR Stedinger.
The use of gls regression in regional hydrologic analyses.
Journal of Hydrology, 344(1):82–95, 2007.

[Haz01] M Hazewinkel.
Cholesky factorization.
Encyclopedia of mathematics, 2001.

[HK70] Arthur E Hoerl and Robert W Kennard.
Ridge regression: Biased estimation for nonorthogonal problems.
Technometrics, 12(1):55–67, 1970.

[KS14] Hyun Hak Kim and Norman R Swanson.
Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence.
Journal of Econometrics, 178:352–367, 2014.

[LL17] Hugues Langlois and Jacques Lussier.
Rational Investing: The Subtleties of Asset Management.
Columbia University Press, 2017.

[LLZ+11] Sophie Lambert-Lacroix, Laurent Zwald, et al.
Robust regression through the huber's criterion and adaptive lasso penalty.
Electronic Journal of Statistics, 5:1015–1053, 2011.

[Mar52] Harry Markowitz.
Portfolio selection.
The journal of finance, 7(1):77–91, 1952.

[Ode83] Patrick L Odell.
Gauss–markov theorem.
Encyclopedia of statistical sciences, 1983.

[RSTZ15] David Rapach, Jack Strauss, Jun Tu, and Guofu Zhou.
Industry interdependencies and cross-industry return predictability.
2015.

[Rud87] Walter Rudin.
Real and complex analysis.
Tata McGraw-Hill Education, 1987.

[SW11] Claude Sammut and Geoffrey I Webb.
Encyclopedia of machine learning.
Springer Science & Business Media, 2011.

[Tib96] Robert Tibshirani.
Regression shrinkage and selection via the lasso.
Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288, 1996.

[Tib11] Robert Tibshirani.
Regression shrinkage and selection via the lasso: a retrospective.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(3):273–282, 2011.

[WD12] Jens Wagener and Holger Dette.
Bridge estimators and the adaptive lasso under heteroscedasticity.
Mathematical Methods of Statistics, 21(2):109–126, 2012.

[Woo15] Jeffrey M Wooldridge.
Introductory econometrics: A modern approach.
Nelson Education, 2015.

[WRC16] Ines Wilms, Jeroen Rombouts, and Christophe Croux.
Lasso-based forecast combinations for forecasting realized variances.
2016.

[You12] James Younker.
Ridge Estimation and its Modifications for Linear Regression with Deterministic or Stochastic Predictors.
PhD thesis, Université d'Ottawa/University of Ottawa, 2012.

[ZH05] Hui Zou and Trevor Hastie.
Regularization and variable selection via the elastic net.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2):301–320, 2005.

[ZHX+14] Xiangzhou Zhang, Yong Hu, Kang Xie, Shouyang Wang, EWT Ngai, and Mei Liu.
A causal feature selection algorithm for stock prediction modeling.
Neurocomputing, 142:48–59, 2014.

[Zou06] Hui Zou.
The adaptive lasso and its oracle properties.
Journal of the American statistical association, 101(476):1418–1429, 2006.