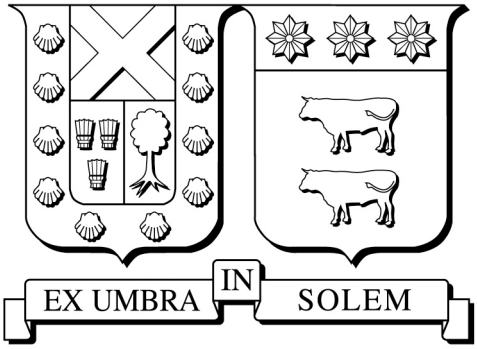
Universidad Técnica Federico Santa María
Departamento de Informática



Online learning methods for financial time series forecasting

DISSERTATION

Submitted in partial satisfaction of the requirements
for the degree of

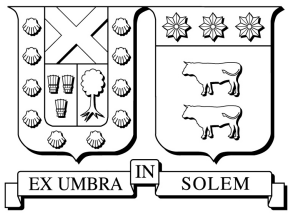
DOCTOR EN INGENIERÍA INFORMÁTICA

by

Paola Soledad Arce Azócar

Advisor
Luis Salinas Carrasco

In Valparaíso, Chile
December 2017.



Examining Committee

Dr. Luis Salinas Universidad Técnica Federico Santa María Chile	Advisor
Dr. Werner Kristjanpoller Universidad Técnica Federico Santa María Chile	Co-advisor
Dr. Claudio Torres Universidad Técnica Federico Santa María Chile	Special Reviewer
Dr. Marcelo Mendoza Universidad Técnica Federico Santa María Chile	Internal Reviewer
Dr. Christian de Peretti Ecole Centrale de Lyon France	External Peer Reviewer
Dr. Wilfredo Palma Pontificia Universidad Católica de Chile Chile	External Peer Reviewer

To my beloved family

Abstract

Financial time series are known for their non-stationary behaviour and this has motivated its study using different techniques. One relevant observation is that sometimes time series exhibit some stationary linear combinations. When this happens, it is said that those time series are cointegrated. Cointegration has been then one of the main features studied. Vector error correction model (VECM) is an econometric model which characterises the joint dynamic behaviour of a set of cointegrated variables in terms of forces pulling towards equilibrium.

Cointegration relationships are found on different time series frequencies. Cointegration in low frequency time series is motivated by a long-run equilibrium relationship between economic forces, whereas cointegration in high frequency data has its foundation in statistical arbitrage theory. However, the use of cointegration models such as VECM is mainly limited by computationally expensive routines.

In this thesis, financial time series features are studied and different algorithms were developed in order to optimise parameters and increase performance. In particular, an online algorithm based on VECM which optimises how model parameters are obtained and reduces execution times is proposed. This is achieved considering only a sliding window of the last historical data and using machine learning techniques to solve the model. Moreover, the long-run relationship between the time series is used in order to make optimisations and obtain improved execution times. Due to the large amount of financial data available and the need of quick response, the algorithm presented was optimised using high performance computing. The experiments were tested using foreign exchange rates. Results show that cointegration and high performance computing allow to obtain models with better performance accuracy and reduced execution times.

Keywords: *Computational finance - Cointegration - Time Series - Online algorithms - Regression*

Contents

List of Tables

List of Figures

Chapter 0

Introduction

"An individual economic variable, viewed as a time series, can wander extensively and yet some pairs of series may be expected to move so that they do not drift far apart."-Robert F. Engle and Clive W.J. Granger [**?**].
In this introduction we present the scope, objectives, hypothesis and organization of this thesis.

0.1. Scope of this research

The stochastic behaviour of financial time series, its incrementing amount of data available and the need of performing accurate forecasting in short periods of time has motivated researchers to create efficient and fast forecast algorithms. This study involves interdisciplinary knowledge such as: finance, scientific computing, high performance computing, machine learning among others.

Forecasting financial time series have been modelled using classical statistical approaches. More recently, machine learning models have been extensively used in forecasting. However, their main disadvantage is that getting model parameters is a computational challenge. The computational complexity of machine learning algorithms has become a limiting factor for problems that require processing large volumes of data and where response time is crucial.

Therefore, algorithms that process large amount of data in a short periods of time are required. Recently, online learning algorithms have been developed to solve large-scale problems since they process an instance at a time, updating the model at each step incrementally. This is opposed to the batch algorithms where the forecast model is built using a large collection of historical data in a training phase.

The specific scope of this study is to use financial time series features in order to design a forecasting algorithm which ensures accuracy and low response times. Cointegration is the main feature studied and it refers that one or more linear combinations of these time series are stationary even though individually they are not [**?**]. Some models, such as the Vector Error Correction Model (VECM), take advantage of this property and describe the joint behaviour of several cointegrated variables.

In this thesis, an online formulation of the VECM called Online VECM (OVECM) is proposed. OVECM is based on the consideration of a sliding window of the most recent data. The algorithm introduces matrix optimisations in order to reduce the number of operations and also takes into account the fact that cointegration vector space doesn't experience large changes with small changes in the input data. Moreover, VECM parameters are obtained using machine learning methods. Our method is later tested using four currency rates from the foreign exchange market with different frequencies. On the other hand, in order to improve VECM parameters, an adaptive VECM algorithm is presented called AVECM. AVECM allows VECM parameters to be found by maximising the number of cointegration relations for a given set of parameters on a grid search. This grid search is done in parallel.

Models effectiveness were focused on the out-of-sample forecast rather than on the in-sample fitting. This criteria allows the OVECM and AVECM prediction capability to be expressed rather than just explaining data history. Our method performance is compared with the naive forecast of the random walk model and ARIMA which are the most widely used algorithms for modelling a multivariate time series.

0.2. Research Objectives

The main motivation for this research is the development of efficient methods for forecasting financial time series.

The specific objectives of this research are,

- $\mathcal{O}_1$: *A review of the literature on time series analysis models including machine learning techniques.*
- $\mathcal{O}_2$: *Development of a set of known features of the studied time series and the application to improve forecasting.*
- $\mathcal{O}_3$: *Development of parallel and efficient algorithms to ensure quick response times .*
- $\mathcal{O}_4$: *Deep mathematical analysis of the proposal and financial concepts involved.*
- $\mathcal{O}_5$: *Design and implement representative set of experiments in order to show when and why the proposal performs better.*

0.3. Research Hypothesis

Cointegration concept was introduced by Engle and Granger in 1987 [**?**] and implies that one or more linear combinations of non-stationary variables are stationary even though individually they are not. Moreover Stock and Watson in 1988 [**?**] observed that cointegration reflects the common stochastic trends providing a useful way to understand cointegration relationships. These common stochastic trends can be also interpreted as a long-run equilibrium relationships.

Vector error correction model (VECM) introduces this long-run relationship among a set of cointegrated variables as an error correction term. VAR model expresses future values as a linear combination of variables past values. However, VAR model cannot be used with non-stationary variables. VECM is a linear model but in terms of variable differences. If cointegration exists, variable differences are stationary and they introduce an error correction term which adjusts coefficients to bring the variables back to equilibrium. In finance, many economic time series turn to be stationary when they are differentiated and cointegration restrictions often improves forecasting [**?**]. Therefore, VECM has been widely adopted.

Both VECM and VAR model parameters are obtained using ordinary least squares (OLS) method. OLS has two main problems: is sensitive to errors on input data and involves many calculations. The former problem is commonly solved using Ridge Regression (RR) [**?**] which introduces a regularization parameter that leads to an unbiased estimation with better generalisation capability. The second problem of computational complexity depends on the number of past values and observations considered. Recently, online learning algorithms have been proposed to solve problems with large data sets because of their simplicity and their ability to update the model when new data is available.

The main research hypothesis explored in this dissertation is the following:

An online learning algorithm based on cointegration and high performance computing will allow faster forecasting algorithms for financial time series to be obtained while maintaining good accuracy levels.

0.4. Organisation of this Thesis

Chapter 1 contains relevant finance concepts required to understand financial time series models such as market hypothesis, frequency, order book generation, market microstructure. Chapter 2 discuss main characteristics and concepts of financial time series including classic models such as ARMA, ARIMA, GARCH and more recent ones such as VECM, VAR and volatility models. Chapter 3 gives an introduction to machine learning and its variant online learning. Chapter 4 presents the first proposal called

AVECM which includes a new parallel method to choose VECM parameters based on the maximisation of the percentage of cointegration. In Chapter 5 a second approach is presented called OVECM, which is an online version of OVECM for high frequency data. Chapter 6 presents a discussion of found results and summarise the main conclusions of this thesis. It also presents some research directions for future study.

Chapter 1

High Frequency Trading

High frequency trading (HFT) strategies are based on buy and sell assets in short periods of time gaining a small profit in every transaction despite the transaction costs. The key is the amount of transactions that HFT algorithms are capable to execute. In 2016 nearly 50% of US market trades were HFT. HFT has motivated computer-driven strategies capable of processing large amount of data in short periods of time.

1.1. High Frequency Trading Definition

HFT is not a strategy but a technology that implements different trading strategies. The aim of HFT is to benefit from market short-term pricing inefficiencies [**?**]. HFT is characterised for the use of high-speed and sophisticated quantitative and algorithmic computer applications for modelling and executing orders efficiently. In order to make fast decisions, HFT firms require speed access to trading servers and sometimes they are physically near so they can minimise network latencies.

High frequency trades are short-term positions that commonly end the trading day avoiding leaving opened positions to the next business day. HFT is frequently associated to algorithm trading strategies, but the former is focused into reduce the market impact of large institutional positions and the later refers to trade execution strategies that are typically used by fund managers to buy or sell large amounts of assets. The duration of the positions in HFT it is not well-defined. Figure 1.1 shows a survey done to traders about the holding time of a position opened. Most of them agreed that HFT refers to positions between 1-second and 10-minutes. Overnight positions are avoided since they are riskier and also have more expensive fees, HFT firms end the trading day closing all opened positions.

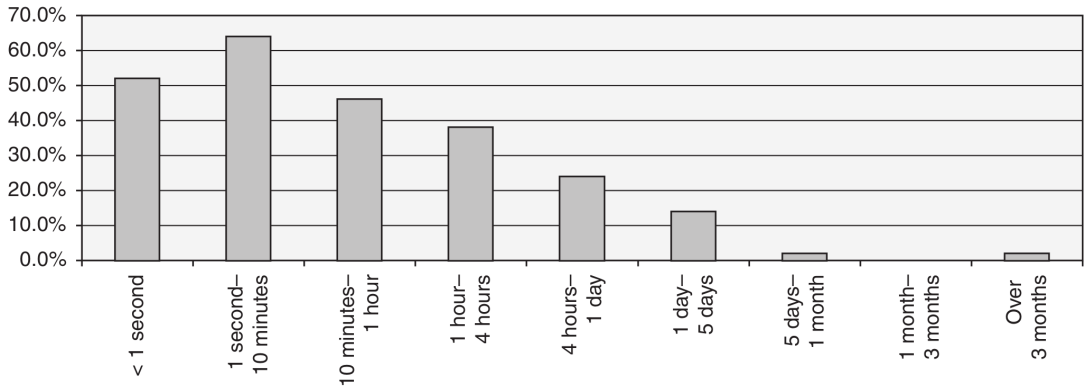


FIGURE 1.1. Holding time of an opened position of a high frequency trade. Source [?].

In 2016 HFT represented nearly 50% of equity trades in the US showing a consistent fall since 2009, which was its best year, until 2013 where this rate remains the same until 2016. Figure 1.2 shows HFT trades percentage of US equity shares.

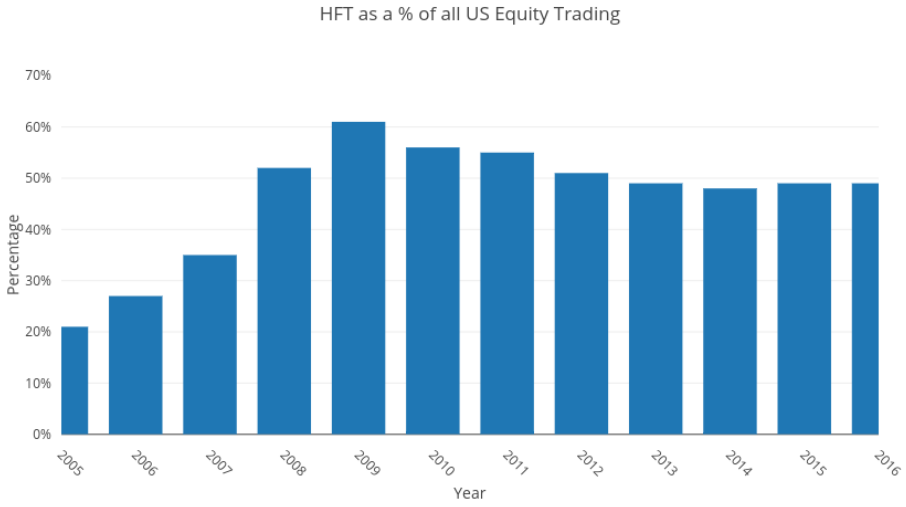


FIGURE 1.2. HFT market share in the US. Source: TABB Group.

US market revenues have fallen dramatically. Figure 1.3 shows how revenues in US stocks have fallen the last years. One of the reasons of this fall is that exchange markets have adapted, having now faster, more transparent and efficient market structures than before. This has been possible due to its investment in technology enhancing reliability and stability of transactions. Even though this fall, HFT is still a major component of regulated markets and will probably remain as a topic of interest for researchers in the near future.

FIGURE 1.3. Revenue in the US. Source: TABB Group.

On the other hand, HFT has been criticised on qualitative issues concerning fairness and systemic risk. However, HFT has lead to beneficial impacts such as reducing spreads (difference between buyers (bid) and sellers (ask) prices), increased liquidity, allowing more efficient price formation, reduced transaction costs and lower market volatility.

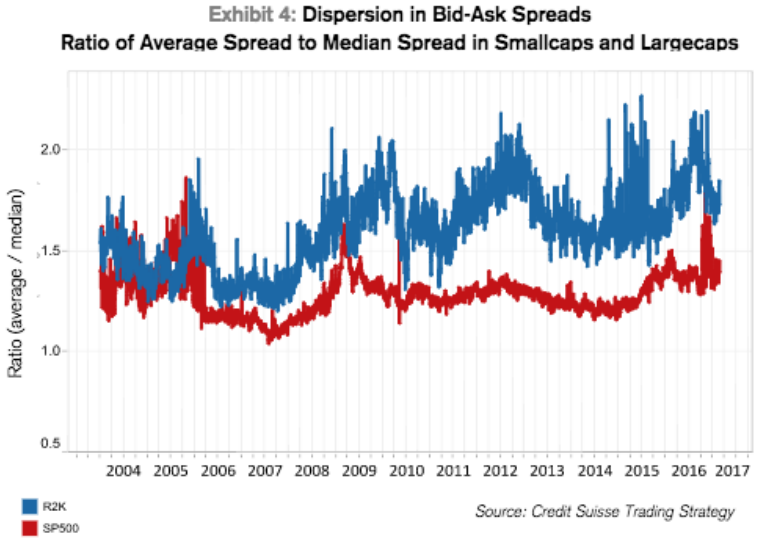


FIGURE 1.4. Bid-ask spreads for large and small caps stock US. Source: Credit Suisse.

In figure 1.4 shows the Bid-ask spread for large and small caps. A large cap refers to an asset with market capitalisation value of more than $10 billion dollars.

In theory, Bid-ask spreads for large caps and small caps move on the same direction, meaning they both are affected by volatility. This scenario is only true until 2009, where the large caps (in red) gets tight spreads and small caps have wider spreads, suggesting a concentrating of trading in the most liquid, biggest stocks.

1.2. Financial Markets

A financial market is any marketplace where buyers and sellers participate trading different assets such as equities, bonds, currencies and derivatives (future or options). One of the main objectives of financial markets is to set prices for global trade. A financial market has many components but the most commonly used are money markets and capital markets. Money markets are used for short-term basis, usually for assets up to one year, for greater periods, capital markets are used. Capital markets include the stock or equity market and the bond or debt market and their movements are the most widely followed [**?**].

Figure 1.5 shows the typical structure of capital markets existed from the early 1929s through much of the 1990s where the broker-dealers played the central and most profitable role. At the core are the exchanges or inter-dealer networks (foreign exchange trading). Exchanges are the centralised marketplaces for transacting. Broker-dealers perform two functions: trading for their own accounts and transacting for their customers. Broker-dealers use inter-dealer brokers to quickly find the best price for a particular asset among the network of other broker-dealers. Occasionally, broker-dealers also deal directly with other broker-dealers, particularly for less liquid instruments. Broker-dealers clients are institutional investors, corporate clients, commercial clients, and high-net-worth individuals [**?**].

This centralised structure existed until computer technology allowed a better communication structure. Today financial markets are more decentralised providing more liquidity. Exchanges and inter-dealer brokers were replaced by liquidity pools or Electronic Communication Networks (ECNs) which are able to transmit and order quickly matching buyers and sellers optimally. There are also dark liquidity pools where trader identity and orders remain anonymous.

Figure 1.6 shows current structure of capital markets including ECNs and dark pools. In this structure, ECNs, Exchanges, dark pools, broker-dealers and retail brokerages can execute orders. However, there are some institutional clients that have also become broke-dealers.

Equity market and foreign exchange market (Forex) are the most popular markets for high frequency trading strategies [**?**]. In the Equity market, stocks such as futures and options, exchange-traded funds (ETFs) among others financial instruments can be

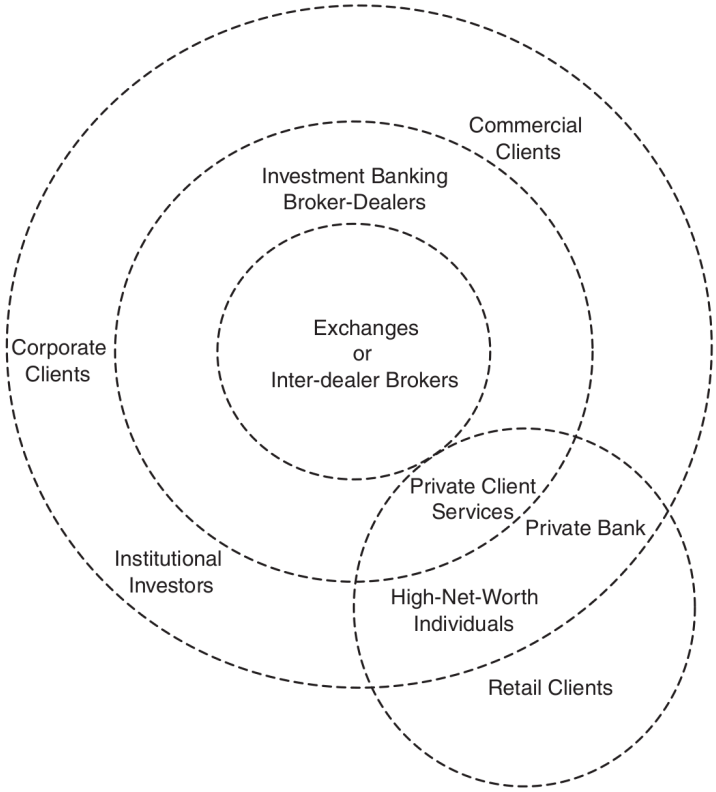


FIGURE 1.5. Old structure of capital markets. Source [**?**].

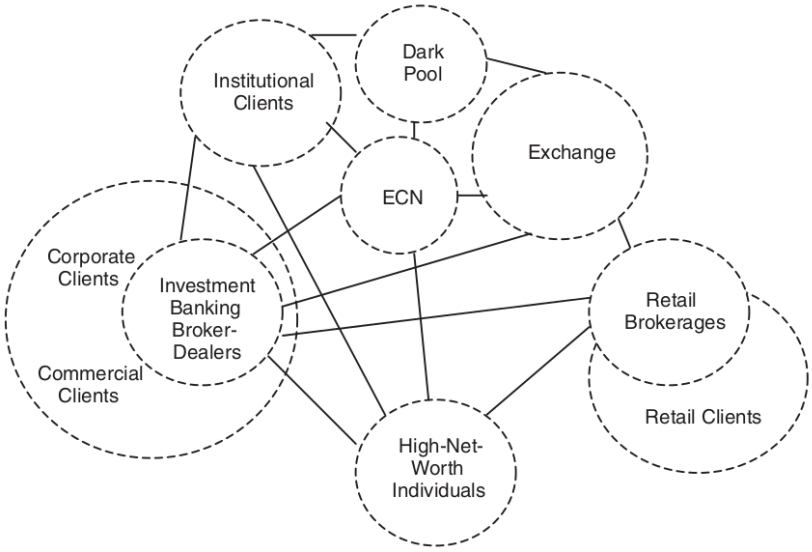


FIGURE 1.6. Actual structure of capital markets. Source [**?**].

traded. Additionally, in the Forex market, interest rates denominated pair currencies such as EURUSD (Euro to UD dollar) or USDJPY (US dollar to Japanese yen) are traded. Traders of this market are diverse, some of them are high frequency traders, long-term investors and corporations. The Forex market used to be centralised, only commercial banks had exclusive access to inter-dealer networks. Today, Forex market is decentralised and has become the biggest market in terms of volume of trading. The main participants are international banks geographically dispersed with continuous 24 hours operations excepting weekends. Figure 1.7 shows different market trading hours in GMT for London, New York, Sydney and Tokyo.

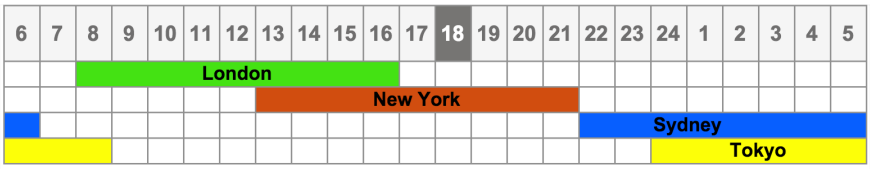


FIGURE 1.7. Forex market trading hours (GMT). Source [**?**].

There are three periods of overlaps between different trading times:

- New York and London from 13:00 GMT until 17:00 GMT
- Tokyo and London from 8:00 GMT until 9:00 GMT
- Sydney and Tokyo from 23:00 GMT to 7:00 GMT

These overlap hours are very important since the highest volume of trades are expected. Therefore, most HFT strategies are executed in these hours.

1.3. PRICE FORMATION PROCESS

A trade consists at least of two orders: one to enter the market (buy order) and exit the market (sell order). Traders have different types of order they can execute:

Market order: allows to buy or sell an asset at the current best price whatever the price is. The transaction price is determined by the market considering previous executed orders and volume requested.

Limit order: allow to specify the price you want to buy or sell an asset. Depending on the market conditions, this type of orders could never been executed.

Stop order: are suitable for investors who are unable to monitor their investments for a period of time. Stop order allows to specify the price that a position should be closed; this price is called stop price. When the stop price is reached a market order is executed, this means that market price could not be exactly the same specified in the stop order.

In terms of commissions, limit and stop order are more expensive than market orders and there is no certainty of execution.

Moreover, all orders can specify other parameters such as:

Fill or Kill: is an order that must be executed immediately or being cancelled, no partial fulfilments are allowed.

Day: the order is only valid during the day.

Good till cancelled: a order is active until the investor decides to cancel it or the trade is executed.

In order to determine the execution price, buyers and sellers orders are placed in an order book which help to determine which order can be fulfilled. Figure 1.8 illustrates how buyers and sellers are ordered. The ask or offer price is the current lower price a seller is willing to accept for a good. The bid price corresponds to the current highest price a buyer is willing to pay for a good. Orders with the same price are prioritised by arrival time and placed on top of the book. The difference between current ask and bid price is called spread and their average is called mid-price [**?**].

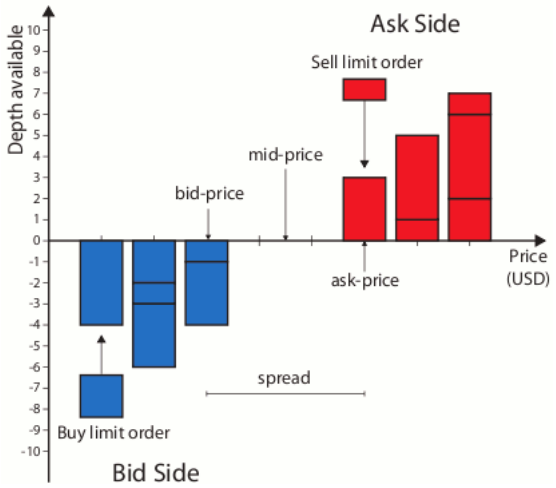


FIGURE 1.8. Order book. Buyers and sellers are ordered according to the bid or ask price and the market determines the mid-price and transacted volume. Source [**?**].

Today, order books are available and they are a popular source of study among researchers. The mid-price and spread modelling are some of the problems related with this area.

1.4. Efficient market hypothesis

Efficient Market Hypothesis (EMH) was developed independently by Paul Samuelson and Eugene Fama in the 1960s. EMH also known as the random walk theory states that current stock prices fully reflect available information related to its value and there is no way to earn excess profits [**?**]. EMH requires that all the participants have rational expectations, and investors reactions be random and follow a normal distribution pattern. Thus, any one can be wrong about the market, but the market is always right as a whole. There are three common forms of EMH: weak-form efficiency, semi-strong efficiency and strong-form efficiency.

The weak form claims that prices already reflect all past publicly available information. Therefore, future prices cannot be predicted based on analysis of historical data. This implies that future prices movements are determined entirely by information not contained in the past prices and participants are unable to systematically profit from market inefficiencies. However, many studies have shown a marked tendency for the stock markets to trend over time periods. Various explanations for such large and apparently non-random price movements have been promulgated. Even Fama has accepted price anomalies which do not follow the weak-form efficiency hypothesis [**?**].

The semi-strong form of the EMH claims that prices instantly change to reflect new public (not private) information such that no excess return can be obtained.

The strong form of the EMH additionally claims that prices instantly reflect even hidden, private or "insider" information.

EMH is related with two approaches to investment analysis: fundamental and technical analysis. Fundamental analysts base their predictions of stock price behaviour on fundamental factors such as internal information of a company, its industry or the economy. Technical analysts, by contrast, consider that all this financial information is already included in the prices and believe that future stock prices can be predicted studying the historical market behaviour. A market technician bases his predictions on historical patterns of prices of volume changes.

Under the weak form of EMH, strategies based on technical analysis will not be able to produce excess returns. However, the weak form accepts that some forms of fundamental analysis may still provide excess return. Similarly, the semi-strong form of EMH neither technical or fundamental analysis can produce excess return.

If the stock market efficiently digests all available information, there is little justification for seeking excess returns gains from investing. However, EMH doesn't lessen the importance of investing only change its philosophy. The only way an investor can possibly obtain higher returns is by purchasing riskier investments.

Researches can determine now how efficient a financial market is, i.e how efficiently information is processed.

EMH is based on rational human behaviour and its validity has been criticised by psychologists and behavioural economists who argue that the EMH is based on counterfactual assumptions regarding human behaviour, that is, rationality. Recent advances in evolutionary psychology and the cognitive neuroscience may be able to reconcile the EMH with behavioural anomalies. On the other hand, if the EMH is true, the market really walks randomly and therefore there shouldn't be any difference between experienced and novice traders. Kim Man Lui proved the contrary in a controlled experiment [**?**].

Chapter 2

Financial Time Series

A time series is a collection of observations in time (discrete or continuous). The analysis of time series main objective is to find possible internal structure in the data such as autocorrelation, trend or seasonal variation. Some of application and uses of time series analysis are data compression, explanatory variables (relationships with other variables, seasonal factors, etc.), signal processing and forecasting (predict future values), which is the focus of this thesis. There are two main approaches to study financial time series: to study directions of financial rates and to explain financial rate volatility. This chapter reviews the most relevant techniques in the rich and rapidly growing field of time series analysis considering these two approaches.

2.1. Characteristics of Financial Time Series

There are several characteristics and concepts of financial time series and here we will discuss the more important ones.

2.1.1. Stylised facts of asset returns. There are several known features exhibited by financial instruments called stylised facts, which have been empirically studied and some of them have been documented only recently and accepted as truth. Stylised facts are usually formulated in terms of qualitative properties of asset returns and may not be precise enough to distinguish among different parametric models. [**?**].

Some stylised facts which are common to a wide set of financial assets [**?**] are:

Dependence: autocorrelation function (ACF) in returns is largely insignificant. Returns of a time series y_t is defined as $r_t = y_t - y_{t-1}$. The ACF measures the linear predictability of the time series y_t at time t using only the value at time

s. ACF is defined as:

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}$$

where $\gamma(s,t)$ is the auto-covariance function that measures the linear dependence between two points on the same series observed at different times. It is defined as the second moment product:

$$\gamma_y(s,t) = E[(y_s - \mu_s)(y_t - \mu_t)]$$

The ACF in the absolute and squared returns is always positive, significant and decays slowly. In addition, the ACF in the absolute returns is generally higher than the ACF in the corresponding squared returns.

Figure 2.1 shows the ACF of the returns of SPY stock where the correlations are significant even for very long lags, this implies a long-memory process.

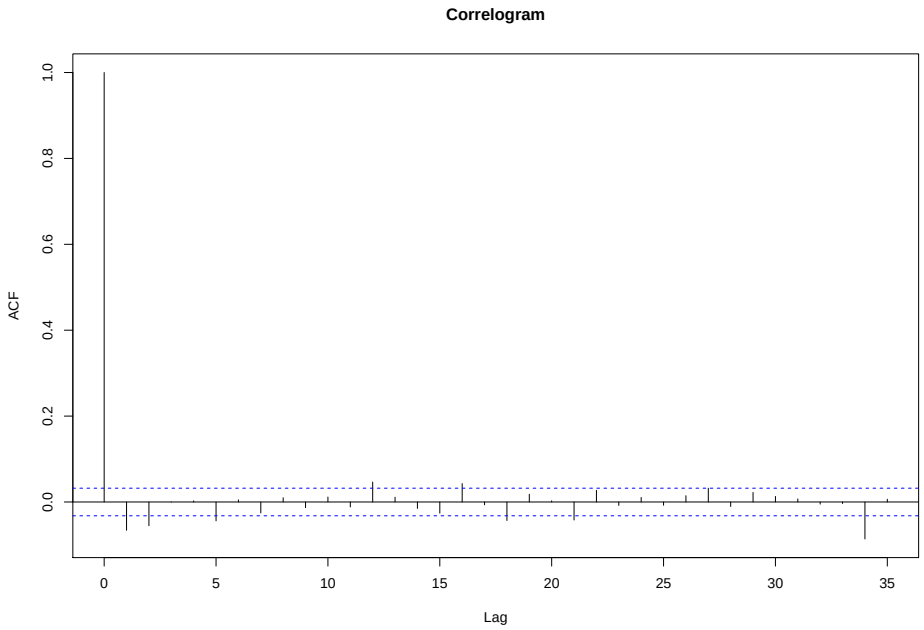


FIGURE 2.1. SPY returns ACF

Distribution: The distribution of returns is approximately symmetric and has high kurtosis (i.e fat tails and a peaked centre compared with the normal distribution). The returns distribution tails are larger than what is hypothesised by common data generation process (generally normal distribution assumption). In the markets, fat tails are an undesirable feature because of the additional risk they imply. However, distribution of returns whose were obtained from higher frequencies looks more like a normal distribution. In the figure 2.2 is

shown the SPY returns distribution based on daily dates from the period 1st July 1998 to 4th April 2013. The distribution was compared against the normal distribution which clearly doesn't fit the data.

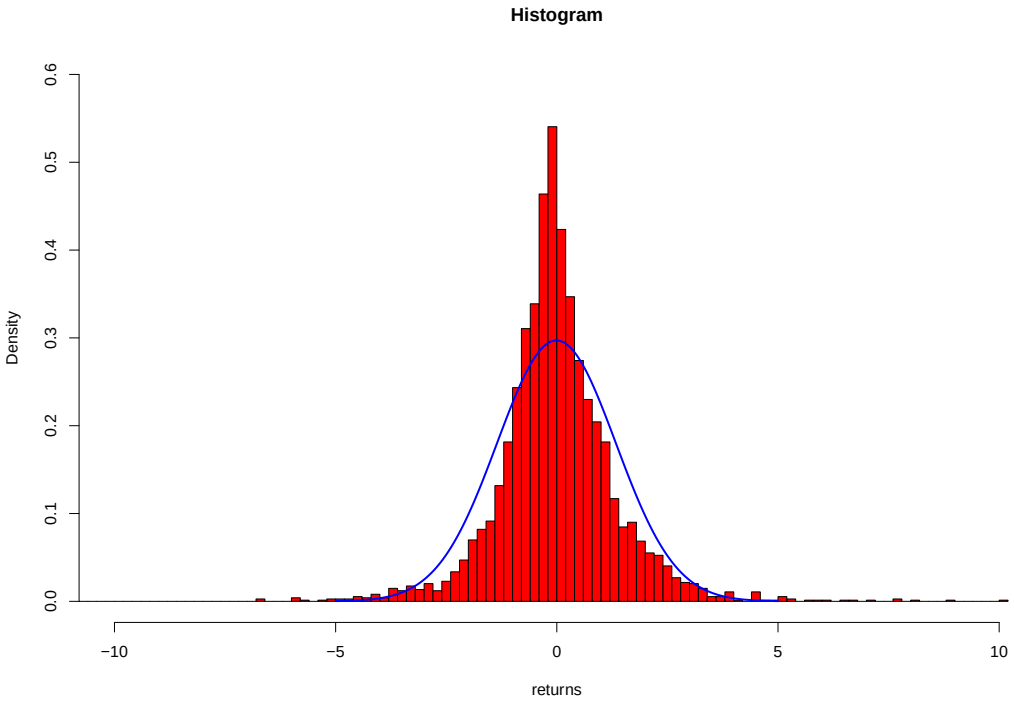


FIGURE 2.2. SPY returns distribution

Heterogeneity: despite the fact that financial returns are non-stationary, stationary periods can be observed. Economist speaks in terms of a structural break [**?**], in machine learning this is known as drift which means that the statistical properties of the target variable change over time [**?**], [**?**].

Non-Linearity: financial returns may be non-linear in mean and/or non-linear in variance.

Calendar effects: are also called seasonal effects and they are cyclical anomalies in returns where the cycle is based on the calendar. Some of known calendar effects are: intraday effect, weekend effect, Monday effect, intramonth effect, the January effect and Holiday effect. The most important calendar anomalies are the January effect and the weekend effect.

2.1.2. Stationary. A strictly stationary times series y_t is one for which the probabilistic behaviour of every collection of values $\{y_{t_1}, y_{t_2}, \ldots, y_{t_L}\}$ is identical to that of the time shifted set, more precisely:

$$P\{y_{t_1} \leq c_1, \ldots, y_{t_L} \leq c_L\} = P\{y_{t_1+h} \leq c_1, \ldots, y_{t_L+h} \leq c_L\} \quad \forall L \in \mathbb{N}, \forall h \in \mathbb{Z}$$

where $c_1, \ldots, c_L$ are constants. This definition is too strong and difficult to assess from a single data set. The weak version of this definition imposes conditions only on the first two moments.

A weakly stationary time series is a process which mean, variance and auto covariance do not change over time:

$$\begin{aligned} E(y_t) &= \mu \quad \forall t \in \mathbb{N} \\ E(y_t^2) &= \sigma^2 \quad \forall t \in \mathbb{N} \\ \lambda(s,t) &= \lambda(s+h, t+h) \quad \forall s, t \in \mathbb{N}, \forall h \in \mathbb{Z} \end{aligned}$$

with $\lambda(s,t) = E[(y_s - \mu)(y_t - \mu)]$

One of the consequences of stationary processes is how they recover from a shock. A shock represents an unexpected change in a variable or a in its error term in a particular period of time. For stationary time series, shocks to the system will gradually die away. That is, a shock during time t will have a smaller effect in time $t+1$, a smaller effect on $t+2$ and so on. For non-stationary data, the persistence of shocks will always have permanent effects because is not a mean-reverting process.

2.1.3. Non-stationary processes. There are different types of non-stationary time series models often found in economics:

(a) **Deterministic trend:** Deterministic trend or trend stationary processes have the following form:

$$y_t = f(t) + \epsilon_t \quad ,$$

where t is the time trend and ϵ_t represents a stationary error term (with mean 0 and variance σ^2) and $f(t)$ is a deterministic function of time:

- If $f(t) = \alpha + \beta t$ we have a linear trend model which is widely used.
- If $f(t) = \alpha \exp^r t$ we have an exponential growth curve.
- If $f(t) = c_1 + c_2 t + c_3 t^2$ we have a quadratic trend model
- If $f(t) = \frac{1}{k+\alpha\beta^t}$ we have a logistic curve

(b) **Stochastic trend:** Stochastic trend processes are also called unit root or difference stationarity processes and have the following form:

$$y_t = \mu + y_{t-1} + \epsilon_t$$

where ϵ_t is a stationary process. When $\mu = 0$ the process is called pure random walk and when $\mu \neq 0$ the process is called random walk with drift.

Alternatively this process can be expressed using the lag operator L such as:

$$(1 - L)y_t = \mu + \epsilon_t$$

This process is also called unit root because the root of the characteristic equation $(1 - z = 0)$ is the unity.

A random walk with or without drift can be transformed to a stationary process by differencing the time serie once. The disadvantage of differencing is that the process loses one observation each time the time series is differentiated.

Apart from a stochastic trend, many economic financial time series seem to involve an exponential trend, this is the reason why researchers often take the logarithmic transformation before doing analysis.

Other less common forms of non-stationarity are structural break in mean and structural break in variance.

2.1.4. Spurious regression. The use of non-stationary data can lead to spurious regressions. Spurious regression dates back to Yule in 1926 [**?**]. If two stationary variables are generated as independent random series and are trending over time, when one of those variables is regressed on the other, they could have a high coefficient of determination (R^2) even if the two are totally unrelated. So, if standard regression techniques are applied to non-stationary data, it could look good under standard measures but valueless [**?**]. The regression measure R^2 gives high values in presence of spurious relationships. This measure would means that the model explains all the variability of the response data around its mean which is not true.

The R^2 measure is calculated as follows:

$$R^2 = 1 - \frac{\sum_{t=1}^{T} \epsilon_t^2}{\sum_{t=1}^{T} (y_t - \bar{y})^2} \tag{2.1}$$

In presence of spurious relationships the denominator of equation 2.1 becomes very large because a large weight is placed on extreme observations on either side of the mean $\bar{y}$.

Example of spurious regression An example is shown below. Two random walks series $x_{1,t}$ and $x_{2,t}$ with a small drift ($\alpha = 0.09$) are generated and then regressed one on the other:

$$x_{i,t} = \alpha + x_{i,t-1} + \epsilon_i, t \qquad \text{where} \qquad \mathcal{N}(0,1), \quad i = 1, 2 \tag{2.2}$$

Figure 2.3 shows the time series $x_{1,t}$ and $x_{2,t}$ and figure 2.4 shows the regression between them and how they seem to be related.

These two random walks, which are independent by construction, appear to be related with a $R^2 = 0.883$. However, what it is happening is that they are drifting in the same direction. Drift can be removed using returns or first differences. If after taking

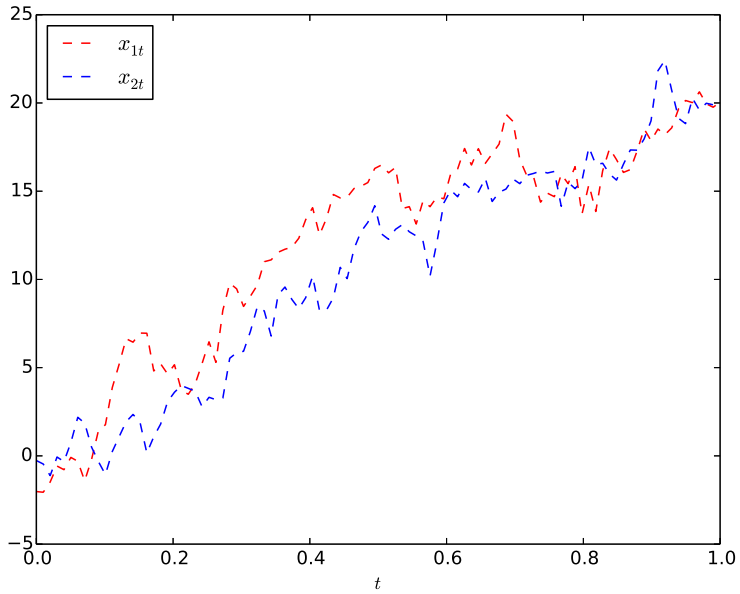


FIGURE 2.3. Two random walks time series $x_{1,t}$ and $x_{2,t}$

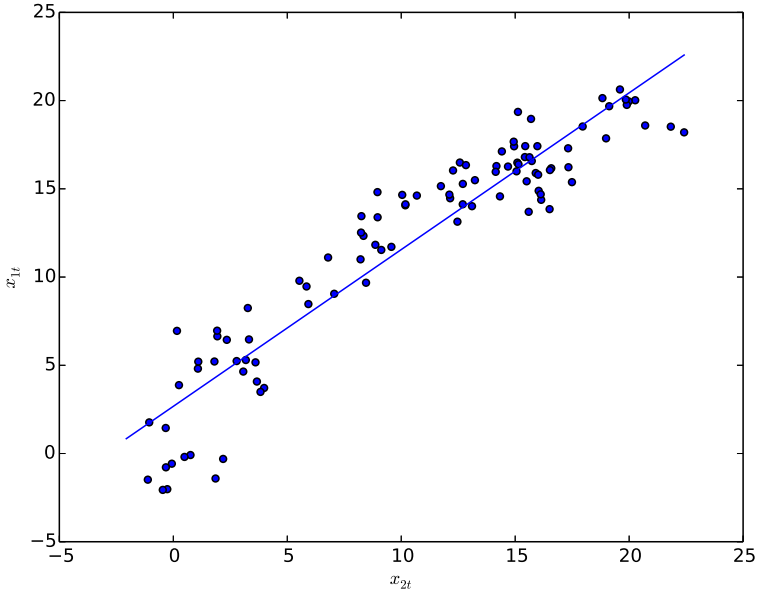


FIGURE 2.4. Regression between two random walks time series

returns or first differences the regression fit is still good we will say that the variables are cointegrated, this means that both series may drift but their residuals will not. A more formal definition of cointegration is given in the section 2.1.6.

2.1.5. Integration. Following Johansen [?] we shall say that a stochastic process Y_t which satisfies $Y_t - E(Y_t) = \sum_{i=0}^{\infty} C_i\, \varepsilon_{t-i}$ is called $I(0)$, and then we shall write $Y_t \sim I(0)$, whenever $\sum_{i=0}^{\infty} C_i \neq 0$ and $\sum_{i=0}^{\infty} C_i\, z^i$ converges for $z \in \mathbb{C}$ with $|z| < 1$. It is understood that the condition $\varepsilon_t \sim iid(0, \sigma^2)$ holds.

A vector time series $\mathbf{y}_t$ is said to be *integrated of order* d, and then we shall write $\mathbf{y}_t \sim I(d)$, whenever after d times (discrete) differentiation a stationary process is obtained [?]; more precisely, whenever $(1-L)^d\, \mathbf{y}_t \sim \mathrm{I}(0)$, where L is the usual lag operator: $(1-L)\, \mathbf{y}_t = \Delta \mathbf{y}_t = \mathbf{y}_t - \mathbf{y}_{t-1}$ for all t.

Note that this definition includes the scalar case as time series of vectors of dimension 1; in this scalar case we will write the time series in non-bold format.

2.1.6. Cointegration. Cointegration concept was introduced by Engle in 1987 [?] and implies that one or more linear combinations of non-stationary variables are stationary even though individually they are not.

Let $\mathbf{y}_t^{\nu}$, $\nu = 1, \ldots, p$, be a set of p vector time series of order $I(1)$. They are said to be *cointegrated* if a vector $\beta = [\beta(1), \ldots, \beta(p)]^{\top} \in \mathbb{R}^p$ exists, such that the time series,

$$\mathbf{Z}_t := \sum_{\nu=1}^{p} \beta(\nu)\, \mathbf{y}_t^{\nu} \sim \mathrm{I}(0)\,. \tag{2.3}$$

In other words, a set of $I(1)$ time series is said to be cointegrated if a linear combination of them exists, which is I(0).

Stock and Watson in 1988 [?] observed that cointegration reflects the common stochastic trends providing a useful way to understand cointegration relationships. These common stochastic trends can be also interpreted as a long-run equilibrium relationship.

The idea of cointegration was immediately adopted in finance since it could represent their long-run relationship implied by economic theory [?], [?]. Economic theory suggest that economic time series are mean-reverting process and therefore, it reflects the idea of that some set of variables cannot wander too far from each other.

On the other hand, the efficient markets hypothesis, also known as the random walk theory states that current stock prices fully reflect available information related to its value and there is no way to earn excess profits [?]. This means that if we have stock prices from a jointly efficient market, they cannot be cointegrated [?], [?]. However, [?] claims that cointegration is directly at odds with market efficiency, even though, there is no evidence that cointegration among asset prices have implications about market efficiency [?].

Despite the fact that cointegration on closing daily rates of currency pairs has not been found [?], [?], different time series frequencies can have different behaviours [?]. Pair trading is a very common example of cointegration application [?] but cointegration can also be extended to a larger set of variables [?],[?].

2.1.7. Johansen method. Johansen in 1988 [**?**] suggests a method for determining cointegration vectors.

There are two statistics for cointegration under Johansen approach: the trace (shown in equation 2.4) and the maximum-eigenvalue statistic (shown in equation 2.5):

$$\lambda_{\text{trace}}(r) = -T \sum_{i=r+1}^{g} \ln(1 - \hat{\lambda}_i) \tag{2.4}$$

$$\lambda_{\max}(r, r+1) = -T \ln(1 - \hat{\lambda_{r+1}}) \tag{2.5}$$

where r is the number of cointegration vectors and $\hat{]}\lambda_i$ is the estimated value for the ith ordered eigenvalue. λ_{trace} null hypothesis is that the number of cointegration vectors is less than or equal to r against that there are more than r. $\lambda_{\max}$ has the null hypothesis that the number of cointegration vectors is r against an alternative of $r+1$.

Cointegration example

If we have two-dimensional process $\mathbf{y}_t$, $t = 1, \ldots, T$ by:

$$\begin{aligned} \mathbf{y}_{1t} &= \sum_{i=1}^{t} \epsilon_{1i} + \epsilon_{2t} \\ \mathbf{y}_{2t} &= a \sum_{i=1}^{t} \epsilon_{1i} + \epsilon_{3t} \end{aligned}$$

FIGURE 2.5. Cointegration example: two I(1) processes (blue and green) and one I(0) process (red).

Since $\mathbf{y}_{1t}$ and $\mathbf{y}_{2t}$ are I(1) processes and there exist a vector $\beta = [a \; -1]$ such that:

$$\beta^\top \mathbf{y}_t = a\mathbf{y}_{1t} - \mathbf{y}_{2t} = a\epsilon_{2t} - \epsilon_{3t} \sim \mathrm{I}(0)$$

then, both processes are said to be cointegrated.

If we add a I(0) process $\mathbf{y}_{3t} = \epsilon_{4t}$ we find that there exists two cointegration vectors now: $\begin{bmatrix} a & -1 & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$ since:

$$\beta^\top \mathbf{y}_t = \begin{bmatrix} a & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{y}_{1t} \\ \mathbf{y}_{2t} \\ \mathbf{y}_{3t} \end{bmatrix} = \begin{bmatrix} a\epsilon_{2t} - \epsilon_{3t} \\ \epsilon_{4t} \end{bmatrix}$$

Figure 2.5 shows the three time series examples.

This example shows how cointegration vectors describes the stable relations between the processes by linear relations that are more stationary than the original process. Cointegration vectors can be obtained using the Johansen method which gives with a certain probability the number of significant vectors. Figures 2.6, 2.7 and 2.8 shows the linear combinations obtained using cointegration vectors given by Johansen method. The method confirm that only two vectors are found and the third one given doesn't result in a I(0) process.

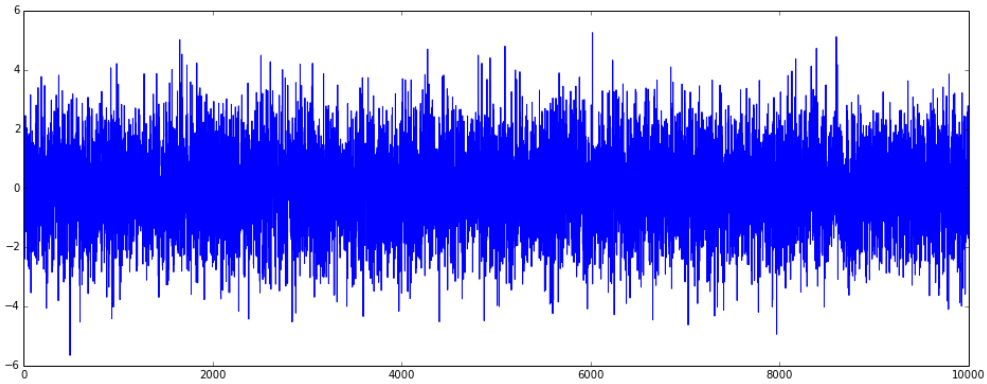


FIGURE 2.6. Time series linear combination using the first cointegration vector

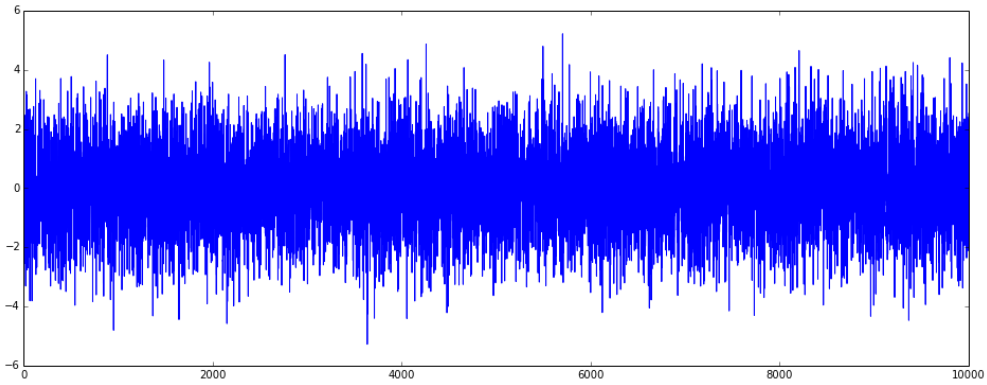


FIGURE 2.7. Time series linear combination using the second cointegration vector

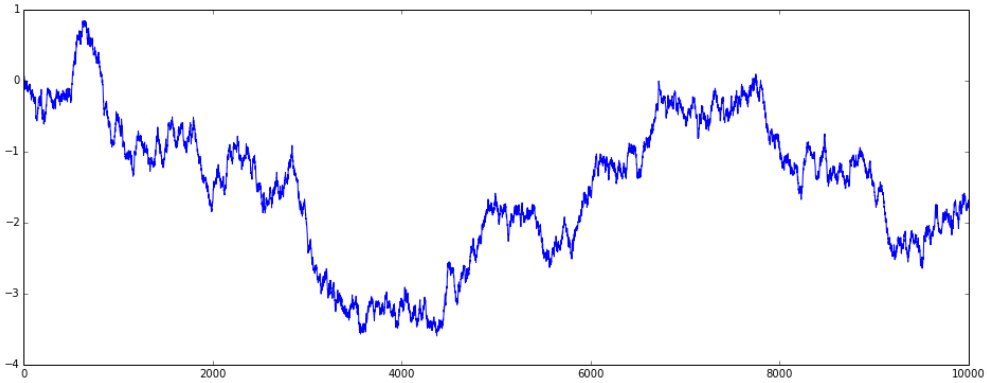


FIGURE 2.8. Time series linear combination using the third cointegration vector

2.2. Unit root tests

Many economic and financial time series exhibit trending behaviour or non-stationary behaviour in the mean or variance. An important econometric task is to determine the type of non-stationarity. For trend stationary I(0) time series a time-trend regression is appropriate. For I(1) time series taking first differences transforms it into a stationary one. For cointegration to determine if time series are I(1) is mandatory to model their long-run relationship. In order to determine if a time series is an a I(1) or I(0) process unit root tests are used. Several tests have been developed but the most widely used is an extension of the Dickey and Fuller test [**?**] called the Augmented Dickey-Fuller (ADF) test developed by Said and Dickey in 1984 [**?**] which tabulated critical values were obtained by [**?**].

The ADF test tests the null hypothesis that a time series y_t is I(1) against the alternative that it is I(0), assuming that the dynamics in the data have an ARMA structure. The ADF test is based on estimating the test regression:

$$y_t = \boldsymbol{\beta}^\top \mathbb{D}_t + \phi y_{t-1} + \sum_{j=1}^{p} \Psi_j \Delta y_{t-j} + \epsilon_t \tag{2.6}$$

where $\mathbb{D}_t$ is a vector of deterministic terms. Lag length p has to be determined when applying the test, Akaike Information Criteria (AIC) is commonly used to determine it. The p-lagged difference terms Δy_{t-j} are used to approximate an ARMA process and the value of p is set so that the error ϵ_t is serially uncorrelated. Under the null hypothesis, y_t is I(1) which implies that $\phi = 1$. The statistic test is the following:

$$\text{ADF}_t = \frac{\hat{\phi} - 1}{SE(\phi)}$$

where SE is the standard error.

The ADF_t statistic is based on the least squares estimates of equation 2.6.

2.3. Volatility in the financial markets

In financial markets, volatility is one of the key elements to model the stochastic dynamic behaviour of financial assets. This is mainly because volatility gives a measure of uncertainty about future returns and it is often viewed as an indicator of the vulnerability of financial markets. It is also important as an input parameter in problems like derivative pricing, hedging and portfolio management. For example, in option pricing we need to known first the volatility of the underlying asset before pricing an option.

Besides, volatility provides important information for studying asset returns because they largely uncorrelated and nonlinearly dependent. However, it has been found that volatility exhibits significant autocorrelation and predictable patterns [**?**] and many models have been proposed to forecast its behaviour.

The volatility of the data is due to a large number of factors affecting the market, with some of them directly measurable such as historical prices, trends, supply and demand. However, others such as monetary policies and news are not directly measurable.

Despite the fact variance and volatility are related, they are not the same concept. Variance is a measure of distribution of returns and is not necessarily bound by any time period. Volatility is a measure of the standard deviation (square root of the variance) over a certain time interval. In finance, variance and volatility both gives you a sense of an asset's risk. Variance gives you a sense of the risk in the asset over its lifetime, while volatility gives you a sense of the movement of the asset in, for example, the past month or the past year. The main underlying difference is in their definition. Variance has a fixed mathematical definition, however volatility does not as such. Volatility is said to be the measure of fluctuations of a process.

Volatility is a subjective term, whereas variance is an objective term i.e. given the data you can definitely find the variance, while you can't find volatility just having the data. Volatility is associated with the process, and not with the data. In order to know the volatility you need to have an idea of the process i.e you need to have an observation of the dispersion of the process. All the different processes will have different methods to compute volatilities based on the underlying assumptions of the process.

2.3.1. Types of volatility. The volatility of a stock is not directly observable [**?, ?**]. For example, daily volatility is not directly observable from only daily returns because there is only one observation in a trading day. If intraday data is available, then volatility could be estimated. However, intraday returns are not the only explanatory variables for volatility and several estimators have been proposed. These estimators are observable variables that are related to the latent variable of interest called volatility proxies [**?**]. Examples of volatility proxies are the following:

2.3.1.1. *Realised volatility.* is also known as historic volatility and it is the actual variance in the price of a stock over time. Realised volatility is measured in terms of the standard deviation using the historical stock prices. It is commonly calculated based

on intraday price returns:

$$r_{t,n} = 100(\ln(p_{t,n}) - \ln(p_{t,n-1})) \tag{2.7}$$

where $p_{t,n}$ is the price observed at day $t = 1, \ldots, T$ and intraday sample $n = 2, \ldots, N$. Realised volatility is defined as:

$$\hat{\sigma}(t) = \sum_{n=1}^{N} r_{t,n}^2 \,, \tag{2.8}$$

where N is the number of intraday samples and T is the number of days. In order to include overnight returns, Hansen and Lunde [**?**] introduced a scaling version of realised volatility using the following definitions:

$$r_t = 100(\ln(p_{t,N}) - \ln(p_{t-1,N})) \tag{2.9}$$

$$\bar{\rho}(t) = \sum_{t=1}^{T} r_t^2 \,. \tag{2.10}$$

where equation (2.9) represents overnight returns and the volatility as equation (2.10), where $p_{t,N}$ is the last intraday sample at day t. The scaled realised volatility $\rho(t)$ is defined as:

$$\rho(t) = \gamma\hat{\rho}(t) \,, \qquad \gamma = \frac{\bar{\rho}(t)}{\sum_{t=1}^{T} \hat{\rho}(t)} \tag{2.11}$$

Realised volatility has also been defined as the absolute value return or as the mean of the sum of intraday squared returns at short intervals of time. The majority of research carried out in the literature obtain the daily volatility as the daily squared returns as is shown in equation (2.10). However, it has been proven that this measurement noise is too high for observing the true volatility process [**?**]. Hansen and Lunde [**?**] stated that the use of a noisy proxy could result in an inferior model being chosen as the best one. The realised volatility, as calculated by the cumulative sum of squared intraday returns and shown in equation (2.8), is less noisy and doesn't lead to choosing an inferior model.

2.3.1.2. *Implied volatility.* not only can be extracted from returns but it can also be derived from option or future pricing models. The volatility obtained corresponds to the market's prediction of future volatility. In finance, an option is a derivative, that is, a contract which gives the owner the right, but not the obligation to buy or sell an underlying asset at a given price called strike price. An option can be executed at any time before an expiration date previously defined no matter what price the underlying asset has. For example, the Black-Scholes model [**?**] determines the fair option value based on stock price, strike price, time to option expiration, the interest rate and volatility. These are known or can be easily obtained from the market, excepting by volatility which must be estimated. However, rather than assuming a volatility a priori and computing option prices from it, the model can be used to estimate volatility at given prices, time to expiration and strike price. This obtained volatility is called the implied volatility of an option. Additionally, some models obtain implied volatility

from futures (other derivative from prices). For instance, the Barone-Adesi and Whaley futures option model [**?**] is also used to determine future volatilities [**?**]. Higher implied volatility is indicative of greater price fluctuation in either direction. Implied volatility is found by determining the value which makes theoretical prices equal to market prices. In this way volatility is "implied" by the current market price of the stock.

In finance, an option is a derivative, that is, a contract which gives the owner the right, but not the obligation to buy or sell an underlying asset at a given price called strike price. An option can be executed at any time before an expiration date previously defined no matter what price the underlying asset has.

The Black-Scholes formula [**?**], developed in the early 1970's, Myron Scholes, Robert Merton and Fisher Black, allows to determine an option value V based on the underlying asset price $S(t)$ at a time t and the following constant parameters:

- σ**:** underlying asset price volatility which measures the standard deviation of the returns
- μ**:** underlying asset drift which is a measure of the average rate of growth of the stock
- E**:** option strike or excersice price
- T**:** option date of expiry
- t**:** current time
- r**:** risk-free interest rate

The Black-Scholes model assume that the underlying price S follows a lognormal random walk:

$$dS = \mu S dt + \sigma S dB \tag{2.12}$$

where B is a Brownian motion. This stochastic differential equation has two components: a deterministic term given by $\mu S dt$ and a random term given by $\sigma S dB$.

A brownian motion B (also called a Wiener process) is a stochastic process characterised by the three following properties:

Continuity: $B(t)$ is a continuos function

Normal increments: $B(t) - B(s)$ has a normal distribution with mean 0 and variance $t - s$.

Independence of increments: for every choice of nonnegative real numbers $0 \leq s_1 < t_1 \leq \cdots \leq s_n < t_n < \infty$, the increment random variables $W_{t_1} - W_{s_1}, \cdots, W_{t_n} - W_{s_n}$ are jointly independent.

An stochastic differential integral has the form:

$$W(T) = \int_0^T f(t) dB(t)\,. \tag{2.13}$$

This equations is also expressed in an abbreviate form:

$$dW = f(t)dB\,. \tag{2.14}$$

Therefore, the integral form of the stock price model shown in equation (2.12) is:

$$S(T) = \int_0^T \mu S(t)dt + \int_0^T \sigma S(t)dB(t) \tag{2.15}$$

Ito's lemma is used to find the differential of a time dependent function of a stochastic process. In option pricing we need to find the option price $V(S(t))$ which depends on a stochastic stock price model $S(t)$. $V(S,t)$ is required to be differentiable function of S and once differentiable function of t.

For trading strategies, the interest is centred in forecasting realised volatility over the life of an option and to take advantage when this volatility differs from the implied volatility. This is called volatility arbitrage. For example, a trader will buy an option and hedge the underlying asset if the implied volatility is under the realised volatility.

2.3.2. Volatility methods. In the existing literature, there are four main classes of asset return volatility models: the general autoregressive conditional heteroskedasticity (GARCH) models, the stochastic volatility (SV) models, the realised volatility models and the machine learning based models. A comparison of the first three models can be found in [**?**].

For many years the most popular methods for estimating financial volatility were the autoregressive conditional heteroskedasticity (ARCH) models [**?**] and the general ARCH (GARCH) models [**?**]. For instance, the GARCH(1,1) defines returns y_t and volatility σ_t as:

$$\begin{aligned} y_t &= \sigma_t \epsilon_t \\ \sigma_t^2 &= \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \end{aligned}$$

where ϵ_t is standard Gaussian white noise, $\alpha_0, \alpha_1, \beta_1 \geq 0$ are required to ensure that the variance will never be negative and $\alpha_1 + \beta_1 < 1$ is needed to guarantee a weakly stationary process [**?**].

Since the introduction of the GARCH models, several extensions have been proposed, but none of them seems to beat the GARCH(1,1) model [**?**]. Despite its popularity, GARCH models have several limitations: firstly, a time series model may be non-linear in mean and/or non-linear in variance, but ARCH and GARCH models are non-linear in variance, but not in mean. Besides, GARCH models often fail to capture highly irregular phenomena, like wild market fluctuations.

SV models explain how volatility varies in a random fashion. These models are useful because they explain why options with different strikes and expirations dates have different Black-Scholes implied volatilities, phenomenon known as the volatility smile. This is useful because the Black-Scholes model assumes that the volatility of

the underlying asset is constant which is not always true. There are several SV models and the most well-known and popular is the Heston model [**?**]. Additional information about SV models can be found in [**?**].

The realised volatility constructed from high frequency intraday returns gave rise to the realised volatility models mainly because the realised volatility series is much more homoskedastic and seems to be a long memory process [**?**]. For realised volatility, the autoregressive fractionally integrated moving average (ARFIMA) process emerged as a standard model [**?**] and many variations have been studied, but all of them produce similar forecasting results to the ARFIMA(1,d,1) model [**?**].

On the other hand, machine learning based models, especially artificial neural networks (ANN) and support vector machines (SVM) have arisen as an alternative to forecast volatility. ANN is a statistical technique inspired by biological neural networks which is capable of changing its structure based on external or internal information during a training phase [**?**]. SVM are supervised learning models for classification analysis which recognize patterns finding a separating hyperplane. An extension for regression analysis is known as support vector regression (SVR).

Since machine learning models and in particular ANN do not require assumptions about the data (gaussianity for example) and allow more explanatory variables than returns to be included, they have become widely used in solving financial problems, specially volatility forecasting [**?, ?**]. There are also many works focused on the using of SVM in volatility forecasting [**?, ?, ?, ?**].

However, just as with ANN, SVMs have scalability problems because their training process is computationally intensive and it is done in batch mode. The scalability problem worsens when new additional training data is available and a re-training process from scratch needs to be done. This problem can be avoided using online machine learning algorithms that allow one instance at a time to be processed with low computationally expensive calculations.

2.4. Univariate Time Series modeling

The random walk model, despite its simplicity, is still difficult to outperform for standard econometric forecasting models [**?**]. The random walk model is defined as:

$$\mathbf{y}_t = \mathbf{y}_{t-1} + \epsilon_t \tag{2.16}$$

The naive forecast of the time series difference $\hat{\mathbf{y}}_{t+1}$ for the random walk model is defined as:

$$\hat{\mathbf{y}}_{t+1} = \mathbf{y}_t + \hat{\epsilon}_{t+1} \tag{2.17}$$

where $\hat{\epsilon}_{t+1} = \epsilon_t$.

On the other hand, ARIMA is widely used to forecast returns in finance [**?**]. A process can be modelled as an ARIMA(p,d,q) model if $\mathbf{x}_t = \Delta^d \mathbf{y}_t$, i.e after differencing d times the time series $\mathbf{y}_t$, we get an ARMA(p,q). An ARMA(p,q) model is the following:

$$\mathbf{x}_t = \sum_{i=1}^{p} \phi_i \mathbf{x}_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \tag{2.18}$$

with coefficients $\phi_p \neq 0$, $\theta_q \neq 0$ and $\sigma_\epsilon^2 > 0$.

2.5. Multivariate Time Series modelling

The vector autoregressive VAR(p) model [**?**] is one of the most easy to use, successful and flexible models for the analysis of multivariate time series. It is a natural extension of the univariate autoregressive (AR) model. The VAR model has proven to be useful for describing the dynamic behaviour of economic and financial time series. VAR is a general framework describing the behaviour of a set of l endogenous variables as a linear combination of their last p values, where $l, p \in \mathbb{N}$. In our case, each one of these l variables is a scalar time series $y_{i,t}$, $i = 1, \ldots, l$, and we represent them all together at time t by the the vector time series:

$$\mathbf{y}_t = \begin{bmatrix} y_{1,t} & y_{2,t} & \cdots & y_{l,t} \end{bmatrix}^\top .$$

Notice that the vector $\mathbf{y}_t$ is assumed to be l-dimensional.

The VAR(p) model describes the behaviour of a dependent variable in terms of its own lagged values and the lags of the others variables in the system. The model with p lags is formulated as the system of N:

$$\begin{aligned} \mathbf{y}_t = \mathbf{\Phi}_1 \mathbf{y}_{t-1} + \mathbf{\Phi}_2 \mathbf{y}_{t-2} + \cdots + \mathbf{\Phi}_p \mathbf{y}_{t-p} + \mathbf{c} + \boldsymbol{\epsilon}_t \\ t = p+1, \ldots, N, \end{aligned} \tag{2.19}$$

where $\mathbf{\Phi}_1, \mathbf{\Phi}_2, \ldots, \mathbf{\Phi}_p$ are $l \times l$-matrices of real coefficients, $\boldsymbol{\epsilon}_{p+1}, \boldsymbol{\epsilon}_{p+2}, \ldots, \boldsymbol{\epsilon}_N$ are error terms, $\mathbf{c}$ is a constant vector and N is the total number of samples.

Notice that, regarding our notation of the cointegration section (2.1.6), we have here $\mathbf{y}_t^0 = \mathbf{y}_t$, $\mathbf{y}_t^\nu = \mathbf{y}_{t-\nu}$ and the i-th component of the vector time series $\mathbf{y}_t^\nu$ is the scalar time series $y_{i,t}^\nu$, where $\nu = 1, \ldots, p$ and $i = 1, \ldots, l$. Transposing each equation of the system (2.19) we can write the VAR(p) model in block-matrix form as:

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{E}\,, \tag{2.20}$$

where:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_{p+1}^\top \\ \mathbf{y}_{p+2}^\top \\ \vdots \\ \mathbf{y}_N^\top \end{bmatrix}_{(N-p)\times l} \qquad \mathbf{A} = \left[\begin{array}{cccc|c} \mathbf{y}_p^\top & \mathbf{y}_{p-1}^\top & \cdots & \mathbf{y}_1^\top & 1 \\ \mathbf{y}_{p+1}^\top & \mathbf{y}_p^\top & \cdots & \mathbf{y}_2^\top & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{y}_{N-1}^\top & \mathbf{y}_{N-2}^\top & \cdots & \mathbf{y}_{N-p}^\top & 1 \end{array}\right]_{(N-p)\times(pl+1)}$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{\Phi}_1^\top \\ \mathbf{\Phi}_2^\top \\ \vdots \\ \mathbf{\Phi}_p^\top \\ \mathbf{c}^\top \end{bmatrix}_{(pl+1)\times l} \qquad \mathbf{E} = \begin{bmatrix} \boldsymbol{\epsilon}_{p+1}^\top \\ \boldsymbol{\epsilon}_{p+2}^\top \\ \vdots \\ \boldsymbol{\epsilon}_N^\top \end{bmatrix}_{(N-p)\times l}$$

Taking into account the error term $\mathbf{E}$, equation 2.20 can be solved with respect to $\mathbf{X}$ using the ordinary least squares estimation (see section 2.8).

Conventional regression estimators, including VARs, have good properties when applied to covariance-stationary time series, but encounter difficulties when applied to non stationary or integrated processes.

These difficulties were illustrated by Granger and Newbold in 1974 [**?**] when they introduced the concept of spurious regressions (see section 2.1.4). In 1982, Nelson and Plosser [**?**] showed that unit roots might be present in a wide variety of macroeconomic series in levels or logarithms. This finding made the unit root testing very popular. Aditionally, the implication that variables should be rendered stationary by differencing before they are included in an econometric model was generalised. Further theoretical developments by Granger and Engle in 1987 [**?**] raised the possibility that two or more integrated, non stationary time series might be cointegrated, so that some linear combination of these series could be stationary even though each series were not.

2.6. Vector error correction model

VECM, developed by [**?**], is a linear model for I(1) variables that are cointegrated, see [**?**]. If cointegration exists, variable differences are stationary and they introduce an error correction term which adjusts coefficients to bring the variables back to equilibrium.

In finance, many economic time series are revealed to be stationary when they are differentiated and moreover cointegration restrictions often improve forecasting [**?**]. Therefore, VECM has been widely adopted in finance: [**?**], [**?**], [**?**] and [**?**] to name a few. Pair trading is a very common example of a cointegration application [**?**] but also can also be extended to a larger set of variables [**?**], [**?**].

VECM is obtained re-writing equation 2.19 in terms of the new variable $\Delta\mathbf{y}_t = \mathbf{y}_t - \mathbf{y}_{t-1}$. The VECM model, expressed in terms those differences, takes the form:

$$\Delta\mathbf{y}_t = \mathbf{\Omega}\,\mathbf{y}_{t-1} + \sum_{i=1}^{p-1} \mathbf{\Phi}_i^*\,\Delta\mathbf{y}_{t-i} + \mathbf{c} + \boldsymbol{\epsilon}_t\,, \tag{2.21}$$

where the coefficients matrices $\mathbf{\Phi}_i^*$ and $\mathbf{\Omega}$, expressed in terms of the matrices $\mathbf{\Phi}_i$ of the model VAR shown in equation 2.19, are:

$$\begin{aligned}\mathbf{\Phi}_i^* &:= -\sum_{j=i+1}^{p} \mathbf{\Phi}_j\,,\\ \mathbf{\Omega} &:= -\left(\mathbb{I} - \mathbf{\Phi}_1 - \cdots - \phi_p\right).\end{aligned}$$

The following well known properties of the matrix $\mathbf{\Omega}$ [**?**] will be useful in the sequel:

- If $\mathbf{\Omega} = \mathbf{0}$, there is no cointegration.
- If $rank(\mathbf{\Omega}) = l$, i.e., if $\mathbf{\Omega}$ has full rank, then the time series are not I(1) but stationary.
- If $rank(\mathbf{\Omega}) = r$, $0 < r < l$, then there is cointegration and the matrix $\mathbf{\Omega}$ can be expressed as $\mathbf{\Omega} = \boldsymbol{\alpha}\boldsymbol{\beta}^\top$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $l \times r$ matrices and $\text{rank}(\boldsymbol{\alpha}) = \text{rank}(\boldsymbol{\beta}) = r$.
- The columns of $\boldsymbol{\beta}$ contains the cointegration vectors and the rows of $\boldsymbol{\alpha}$ correspond with the adjusted vectors. $\boldsymbol{\beta}$ is obtained by Johansen procedure [**?**], whereas $\boldsymbol{\alpha}$ has to be determined as a variable in the VECM.

It is worth noticing that the factorisation of the matrix $\mathbf{\Omega}$ is not unique, since for any $r \times r$ nonsingular matrix $\mathbf{H}$, $\boldsymbol{\alpha}^* := \boldsymbol{\alpha}\mathbf{H}$, and $\boldsymbol{\beta}^* = \boldsymbol{\beta}(\mathbf{H}^{-1})^\top$ we have $\boldsymbol{\alpha}\boldsymbol{\beta}^\top = \boldsymbol{\alpha}^*(\boldsymbol{\beta}^*)^\top$. If cointegration exists, then equation (2.21) can be written as follows:

$$\Delta\mathbf{y}_t = \boldsymbol{\alpha}\boldsymbol{\beta}^\top\mathbf{y}_{t-1} + \sum_{i=1}^{p-1} \mathbf{\Phi}_i^*\,\Delta\mathbf{y}_{t-i} + \mathbf{c} + \boldsymbol{\epsilon}_t\,, \tag{2.22}$$

which is a VAR model but for time series differences.

Transposing each equation of the system (2.22) we can write the VECM(p) model in block-matrix form as:

$$\mathbf{B} = \mathbf{A}\mathbf{X} + \mathbf{E}\,, \tag{2.23}$$

where $\mathbf{Y}$ dimension is $((N-p) \times l)$, $\mathbf{A}$ dimension is $((N-p) \times (r + (p-1)l + 1))$, $\mathbf{X}$ dimension is $((r + (p-1)l + 1) \times l)$ and $\mathbf{E}$ dimension is $((N-p) \times l)$:

$$\mathbf{B} = \begin{bmatrix} \Delta\mathbf{y}_{p+1}^\top \\ \Delta\mathbf{y}_{p+2}^\top \\ \vdots \\ \Delta\mathbf{y}_{N}^\top \end{bmatrix} \tag{2.24}$$

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{\alpha}^\top \\ \boldsymbol{\Phi}_1^{*\top} \\ \boldsymbol{\Phi}_2^{*\top} \\ \vdots \\ \boldsymbol{\Phi}_{p-1}^{*\top} \\ \mathbf{c}^\top \end{bmatrix} \tag{2.25}$$

$$\mathbf{E} = \begin{bmatrix} \boldsymbol{\epsilon}_{p+1}^\top \\ \boldsymbol{\epsilon}_{p+2}^\top \\ \vdots \\ \boldsymbol{\epsilon}_N^\top \end{bmatrix} \tag{2.26}$$

and

$$\mathbf{A} = \left[\begin{array}{ccccc:c} \mathbf{y}_p^\top \boldsymbol{\beta} & \Delta\mathbf{y}_p^\top & \Delta\mathbf{y}_{p-1}^\top & \cdots & \Delta\mathbf{y}_2^\top & 1 \\ \mathbf{y}_{p+1}^\top \boldsymbol{\beta} & \Delta\mathbf{y}_{p+1}^\top & \Delta\mathbf{y}_p^\top & \cdots & \Delta\mathbf{y}_3^\top & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{y}_{N-1}^\top \boldsymbol{\beta} & \Delta\mathbf{y}_{N-1}^\top & \Delta\mathbf{y}_{N-2}^\top & \cdots & \Delta\mathbf{y}_{N-p-1}^\top & 1 \end{array}\right]. \tag{2.27}$$

Taking into account the error term $\mathbf{E}$, equation 2.23 can be solved with respect to $\mathbf{X}$ using the ordinary least squares estimation (see section 2.8).

2.7. Cointegration Tests

Johansen method [**?**] allows to find cointegration among several I(1) time series. This test is more general than the Engle-Granger [**?**] test which only allow to find cointegration between two time series. Johansen procedure takes the form of a likelihood-ratio test which determines r cointegration relationships. The method provides two tests based on an eigenvector decomposition of the matrix $\Omega = \alpha\beta^\top$ on equation 2.21 called trace and maximal eigenvalue statistic. Once the number of cointegration number is determined, r more significative eigenvectors provided by Johansen procedure are used to get the cointegration vector matrix $\beta^\top$. Critical values for both tests are limited to 12 and 5 variables in the trace and maximal eigenvalue statistic respectively.

The trace test null hypothesis is $H_0 : r = r_0$ against $H_1 : r > r_0$ and the maximum eigenvalue test null hypothesis is $H_0 : r = r_0$ against $H_1 : r = r_0 + 1$ starting with $r_0 = 0$. In the latest test, if we start with $r_0 = 0$ and H_0 is rejected $H_1 : r = 1$ will be accepted which is not true if there are more than one cointegration relationship, this is why this test is less powerful and the trace test is commonly used.

2.8. Ordinary Least Squares method

When $\mathbf{A}$ is singular, solution to equation (2.23) is given by the ordinary least squares (OLS) method. OLS consists of minimizing the sum of squared errors or equivalently minimizing the following expression:

$$\min_{\mathbf{X}} \quad \|\mathbf{AX} - \mathbf{B}\|_2^2 \tag{2.28}$$

for which the solution $\hat{\mathbf{X}}$ is well-known:

$$\hat{\mathbf{X}} = \mathbf{A}^+ \mathbf{B} \tag{2.29}$$

where $\mathbf{A}^+$ is the Moore-Penrose pseudo-inverse which can be written as follows:

$$\mathbf{A}^+ = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top . \tag{2.30}$$

However, when $\mathbf{A}$ is not full rank, i.e $rank(\mathbf{A}) = k < n \leq m$, $\mathbf{A}^\top \mathbf{A}$ is always singular and equation (2.30) cannot be used. More generally, the pseudo-inverse is best computed using the compact singular value decomposition (SVD) of $\mathbf{A}$:

$$\underset{m\times n}{\mathbf{A}} = \underset{m\times k}{\mathbf{U_1}} \; \underset{k\times k}{\Sigma_1} \; \underset{k\times n}{\mathbf{V}_1^\top}, \tag{2.31}$$

as follows

$$\mathbf{A}^+ = \mathbf{V}_1 \Sigma_1^{-1} \mathbf{U}_1^\top . \tag{2.32}$$

More details can be found in section A.1.

Chapter 3

Machine Learning Models

Machine learning is a scientific discipline focused on the development of algorithms based on learning from examples. This idea has become central to the design of search engines, robots systems and forecasts applications that process large data sets. Usually machines designed to forecast financial time series requires long training period and therefore they are not suitable to be updated with stream data. Online machine learning techniques tackle this problem by updating the model with new data or by computing a new model considering less data so they can give a response in a short period of time.

3.1. Introduction

Machine Learning (ML) studies computer algorithms for learning something such as to complete a task, to make accurate predictions or to behave intelligently. The learning is always based on samples and the objective is about to do better in the future based on the past experiences in an automatic way. ML is a sub-area of artificial intelligence and broadly intersects with other fields such as statistics, mathematics, physics and computer science. There are many examples of machine learning problems: time series forecasting, image processing, face detection, spam filtering, weather prediction, search engines, among many others.

ML models are often more accurate than what can be created through direct programming. The reason for this is that ML models are data driven and are able to examine large amounts of data.

There are three types of ML classified depending on the nature of the learning input or output available to a learning system:

Supervised learning: the input data is a tuple which contains the example input and its desired outputs (also called labels). The list of tuples is called the training set. The concept of supervised learning comes from the supervisor, acting as a teacher in the learning process. The goal is to learn a general rule that maps inputs to outputs optimising a target function. There are two related problem types in supervised learning: classification and regression problems [**?**]. Its two mainstream approaches are: support vector machines (SVMs) [**?**] and ensemble learning [**?**]. Furthermore, supervised learning can be categorised into offline or batch learning and online learning (see section 3.3).

Unsupervised learning: also known as clustering [**?**]. In this type of learning no labels are given and the system has to find a structure on its own, discovering hidden patterns in data.

Reinforcement learning: is the problem faced by an agent that must learn through trial-and-error interactions with a dynamic environment. It is based on programming agents by reward and punishment without the need to specify how the task is to be achieved [**?**].

3.2. Statistical learning theory

All ML problems can be viewed as optimisation problems. The ML core task is to define a learning criterion, i.e the function to be optimised.

Supervised learning is most popular and most commonly used in modelling financial problems and the assessments of this method and their results in practice are fairly good. Therefore, in this thesis we will adhere to this trend and we will use supervised learning.

The basic setting for supervised learning is a *data set* X, and an *outcome set* Y. For our purposes X will be $\mathbb{R}^m$, $m \in \mathbb{N}$, and Y either $\mathbb{R}$ in the case of regression problems, or $\{-1, 1\}$ in the case of classification problems. In addition we have a plurality of *training sets*

$$S = \left\{ (\mathbf{x}_k, y_k) \in X \times Y \mid k = 1, \ldots, n \right\} \subseteq X \times Y \,,$$

where the members of S have been drawn randomly and independently from $X \times Y$ according to an *unknown* joint distribution function $p(\mathbf{x}, y)$ on $X \times Y$. We gather all these training sets in a subset $\mathscr{T}$ of $\mathscr{P}(X \times Y)$.

In addition, there is a *learning algorithm* $\mathcal{A}$ that associates to every data set one and only one learning function $f \in \mathscr{H}$. Thus $\mathcal{A}$ can be considered as a function:

$$\mathcal{A} : \mathscr{T} \to \mathscr{H} \,, \qquad S \mapsto \mathcal{A}(S) = f \quad \forall S \in \mathscr{T} \,.$$

Supervised learning consists then in finding a *learning function* $f : X \to Y$ that minimise the expected error of the loss function $V(f(\mathbf{x}), y) : Y \times Y$ defined as:

$$V(f(\cdot),\cdot) : X \times Y \to \mathbb{R}_0^+, \qquad (\mathbf{x}, y) \mapsto V\big(f(\mathbf{x}), y\big) \tag{3.1}$$

where $V(f(\mathbf{x}), y)$ denote the price paid for mistakes, $f : X = \mathbb{R}^m \to Y$ and $y = f(\mathbf{x}) \in \mathbb{R}$. Therefore, $V(f(\mathbf{x}), y) = 0$ if $f(\mathbf{x}) = y$.

Given a function f a loss function V and a probability distribution p over $X \times Y$, the expected error of f is:

$$E\big[V(f(\cdot),\cdot)\big] = \int_{X\times Y} V\big(f(\mathbf{x}, y)\big)\, dp(\mathbf{x}, y)\,. \tag{3.2}$$

For regression, the most common loss function is square loss or L2 loss function:

$$V(f(\mathbf{x}), y) = \big(f(\mathbf{x}) - y\big)^2 \tag{3.3}$$

another option is the L1 loss

$$V(f(\mathbf{x}), y) = \big|f(\mathbf{x}) - y\big|\,. \tag{3.4}$$

the choice of loss function here gives rise to several well-known learning algorithms such as regularised least squares and support vector machines.

Since the true distribution is unknown and only training samples are available, the objective is to estimate a function $\widehat{f}$ through empirical risk (training error) minimisation (ERM):

$$\widehat{f} = \arg\min_{f \in \mathscr{H}} R_{\text{emp}}[f]$$

where,

$$R_{\text{emp}}[f] = \frac{1}{n} \sum_{i=1}^{n} V(f(x_i), y_i) \tag{3.5}$$

The ERM is a Riemann sum approximation of the expected error of f shown in equation 3.2. By the law of large numbers it is known that:

$$\frac{1}{n} \sum_{i=1}^{n} V\big(f(\mathbf{x}_i, y_i)\big) \overset{n\to\infty}{\longrightarrow} E\big[V(f(\cdot),\cdot)\big]\,. \tag{3.6}$$

3.2.1. Learning algorithms. In all learning algorithms that are trained from example data, there is a tradeoff between three factors [**?**]:

Complexity or capacity of the hypothesis class: is defined in terms of the Vapnik-Chervonenkis (VC) dimension, which corresponds to the number of training points that can be classified exactly by the hypothesis space. More formally, the VC dimension of the hypothesis space $\mathscr{H}$ defined over instance space X is the size of the largest finite subset of X shattered by $\mathscr{H}$. The dataset S is shattered by hypothesis space $\mathscr{H}$ if and only if for every dichotomy of S

there exists some hypothesis in $\mathscr{H}$ consistent with this dichotomy. A dichotomy of a set S is a partition of S into two disjoint subsets [**?**].

For example, the VC dimension of the lines in the plane is explained in figure 3.1:

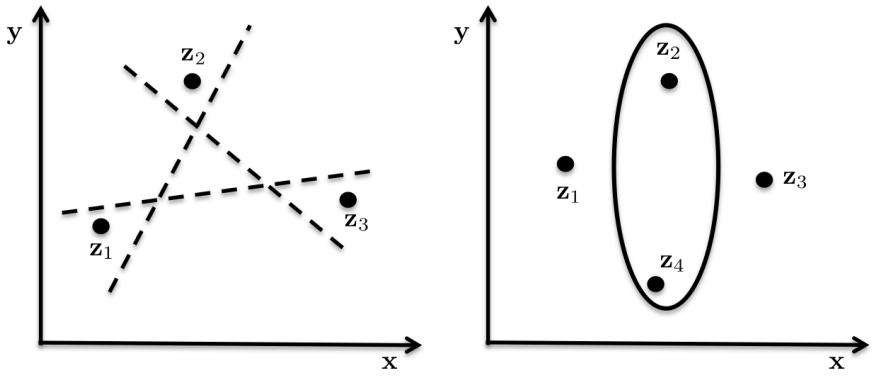


FIGURE 3.1. VC dimension example from [**?**]. The example shows 2-dimensional vectors $z_k = (x_k, y_k)$, $k = 1, 2, 3, 4$. The left figure shows that the 3 dichotomies can be shattered by a line in a plane. However, if we add an extra vector z_4, it is impossible to shatter them with a line. Therefore, linear classifiers has VC dimension equals to $d + 1$ where d is the dimension of the instance space X.

A key point is Radon's theorem in discrete geometry [**?**] (co(A) denotes the convex hull of A):

THEOREM 1 (J. Radon). *Let $S \subset \mathbb{R}^d$ be a subset containing at least $d + 2$ elements. Then there exist* **disjoint** *subsets $A \subset S$ and $B \subset S$ such that*

$$co(A) \cap co(B) \neq \varnothing$$

Radon's theorem implies that in $\mathbb{R}^d$ any set of $d + 2$ or more vectors has at least one dichotomy which cannot be separated (shattered) by any plane in $\mathbb{R}^d$. Thus the VC-dimension in $\mathbb{R}^d$ cannot be greater than $d + 1$.

Generalisation accuracy on new examples: is the capability of the algorithm to generate the right output for an input instance outside the training set. It is measured by the expected risk (equation 3.2).

The amount of training data: in most cases, generalisation accuracy increases as the amount of training data increases.

The relationship between generalisation error and the amount of training or input data is intuitive: the more data is given to the learning model f, the more evidence the algorithm has about the problem. In the limit, the input data will contain every possible example, so the algorithm will generalize perfectly. However, for some $f \in \mathscr{H}$, the best hypothesis $\hat{f}$ might be far away to the best hypothesis class but it is the best that all

the learners can hope. Therefore, to choose the right hypothesis class $\mathscr{H}$ is crucial. The error of this best hypothesis is formalised as the *approximation or generalisation error* [**?**].

Additionally, we do not have infinite data but only some finite random sample set to find an hypothesis, the additional error caused by finiteness of the data, is called *estimation error*. The amount of data needed to ensure a small *estimation error* is referred to as *sample complexity* of the problem.

The relationship between the generalisation error and the hypothesis complexity is less intuitive. If the class $\mathscr{H}$ complexity (VC dimension) increases, then to mantain an *estimation error*, the *sample complexity* increases. Therefore, if we have a low complex model we will reduce the *sample complexity* but it will increase *generalisation error*. The hypothesis has to be sufficiently complex to capture the characteristics of the data.

The figure 3.2 illustrates how generalisation accuracy depends on the complexity of the model and the amount of training data.

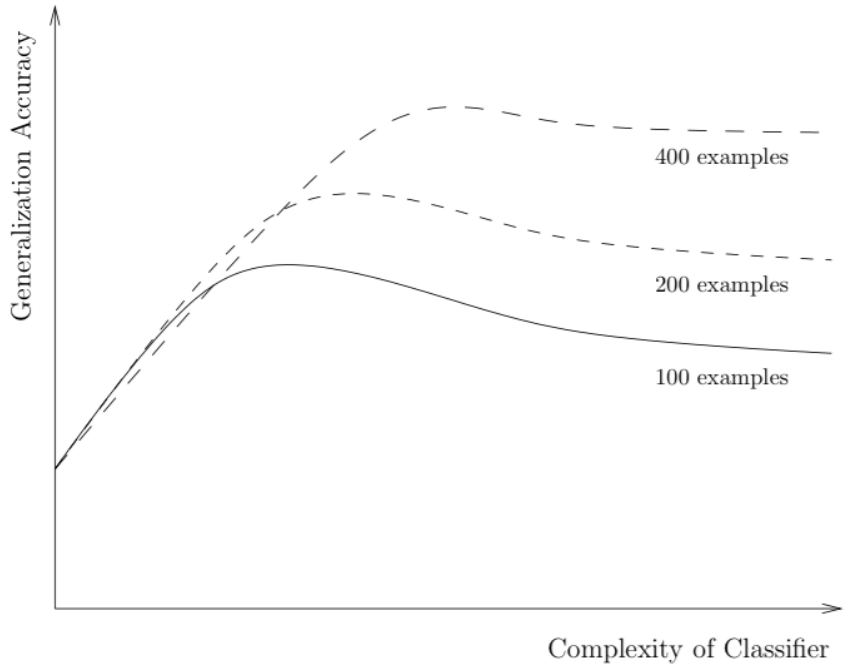


FIGURE 3.2. Tradeoff in empirical learning from [**?**]. The generalization accuracy is not affected by the amount of examples for low complex classifiers. However, for some more complex classifiers, the number of examples improves the generalization accuracy. The key is to find the best generalization accuracy and the lowest complex classifier with the available training data.

3.2.2. The bias-variance tradeoff. A very complex models will fit the training data better (low bias) but it could overfit it and generalise poorly (high variance), this is also called the bias-variance tradeoff.

Figure 3.3 shows *prediction or estimation error* for training and testing sample. In general, in the training sample prediction error is decreasing with more complex models but it overfit the data and it is not a good model for the test data.

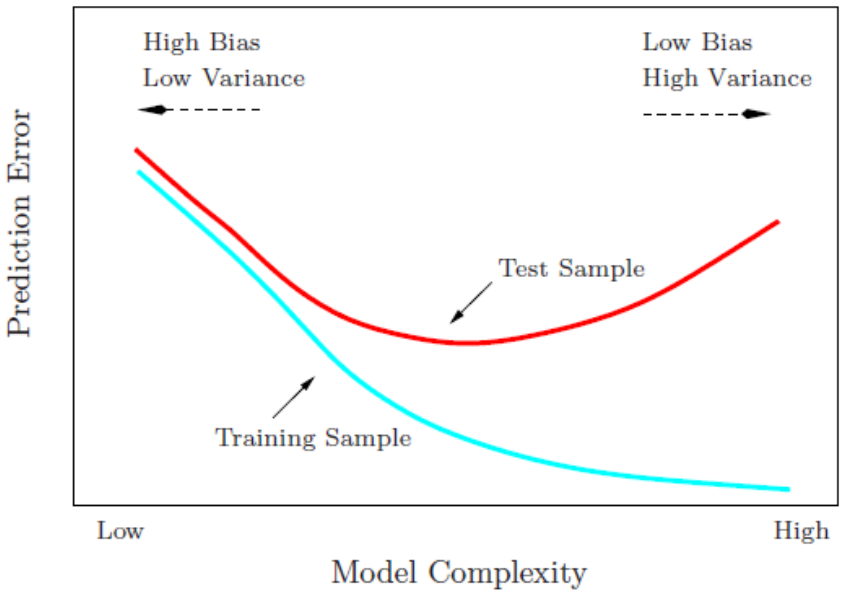


FIGURE 3.3. Training and test error. Image source is [**?**].The blue curve represents the training error which decreases with more complex models. However, the model starts to overfit the data reducing the generalisation capacity of the model,this occurs when the prediction error in the test sample starts to increase (red curve).

Models learning error can be split into two main components: error due to bias and error due to variance. Bias measures the prediction error of the model $\hat{f}$ from the real value, see equation 3.8. Variance is taken as the variability of a model prediction for a given data point or its sensitivity to small fluctuations in the input data, see equation 3.9.

If we have a learning model $\hat{f}$, its expected generalisation error on a testing sample x can be decomposed as shown in equation 3.7 (proof in [**?**]):

$$\mathrm{E}\left[\left(y - \hat{f}(\mathbf{x})\right)^2\right] = \mathrm{Bias}(\hat{f}(\mathbf{x}))^2 + \mathrm{Var}(\hat{f}(\mathbf{x})) + \sigma^2 \tag{3.7}$$

where:

$$\text{Bias}(\hat{f}(\mathbf{x})) = \mathrm{E}\big[\hat{f}(\mathbf{x})\big] - f(\mathbf{x}) \tag{3.8}$$

$$\text{Var}(\hat{f}(\mathbf{x})) = \mathrm{E}\Big[\big(\hat{f}(\mathbf{x}) - \mathrm{E}[\hat{f}(\mathbf{x})]\big)^2\Big] \tag{3.9}$$

Figure 3.4 shows how testing error can be decomposed into bias and variance components. The bias shown in equation 3.8 can be obtain through resampling of the input data x, therefore an average of the predicted values can be calculated. If these prediction values are substantially different to the true value, the bias will be high.

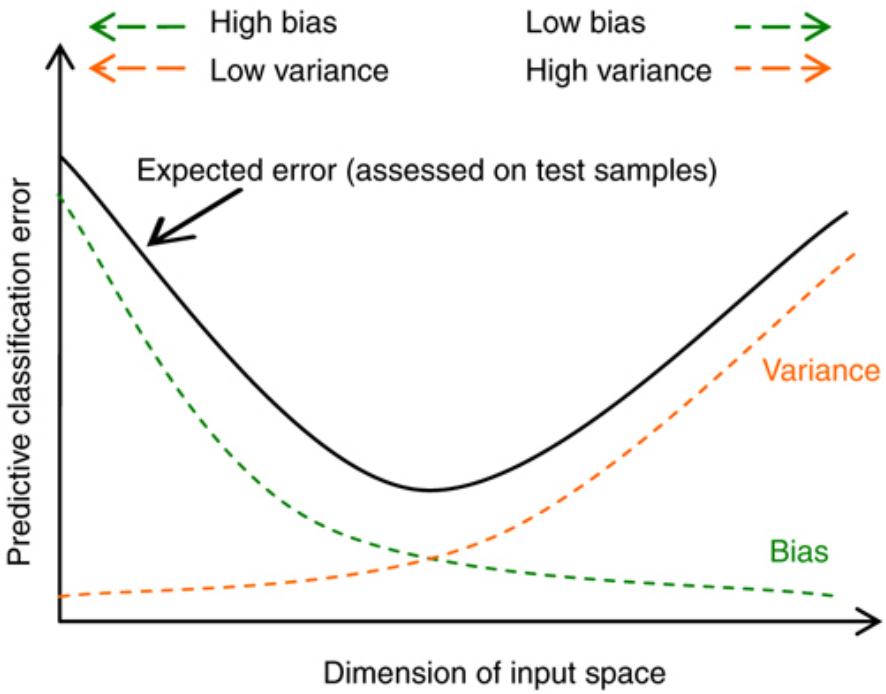


FIGURE 3.4. Bias variance tradeoff

The error due to variance is the amount by which the prediction, over one training set, differs from the expected predicted value, over all the training sets. Variance measures how inconsistent are the predictions from one another, over different training sets, not whether they are accurate or not [**?**].

3.2.3. Generalised Cross Validation. This bias-variance tradeoff is the reason why the data is usually divided in three subsets: training, validation and testing set. The training set is used to construct the model, the validation set to estimate the generalisation error and finally the testing set to estimate the accuracy of the model. Usually the partition is 50% for training, 25% for validation and 25% for testing purposes. When the amount of data is limited, this procedure can be extended to a Cross Validation (CV) approach [**?**] . Various splitting strategies lead to various CV estimates. In K-fold

cross-validation [**?**] the training data is divided randomly into K distinct subsets, then the model is trained using K-1 subsets, and tested on the remaining subset. The process of training and testing is then repeated for each of the K possible choices of the subset omitted from the training. The average performance on the K omitted subsets is the estimate of the generalisation performance.

3.2.4. Regularization. Regularization can be understood using two contexts: learning theory (probabilistic) and inverse problems (deterministic). In the context of learning, regularization refers to techniques allowing to avoid over-fitting and the desire property of the selected estimator is to perform well on new data (to generalise), e.g. regularized least squares. In the context of inverse problems, regularization objective is to stabilise, with respect to noise, a possibly ill-conditioned matrix inversion problem e.g spectral cut-off and Tikhonov regularization or ridge regression (RR) [**?**]. More details in section A.6.

In particular, it is well known that regularization schemes such as RR or Tikhonov regularization can be effectively used in the context of learning [**?**].

Tikhonov introduced a regularization which ensures well-posedness and generalisation of ERM, i.e prevents overfit, by constraining the hypothesis space $\mathscr{H}$ usually called regularised ERM or Tikhonov regularisation.

Tikhonov regularisation considers a functional of the form:

$$R_{\text{emp}}[f] = \frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i) + \lambda \mathcal{R}(f) \tag{3.10}$$

where $X \subseteq \mathbb{R}^m$ and $Y = \mathbb{R}$, V is the loss function defined in the equation 3.1 and $\mathcal{R}(f)$ is the regulariser, a penalisation on f. One example of regularisation in linear models is Ridge Regression (RR), which is a regularised least squares method.

3.2.5. Ridge Regression. Ridge regression (RR) was independently proposed by Tikhonov [**?**] and Phillips [**?**].

RR corresponds to a regularised least squares method. The least squares (LS) method is a well known way to solve a regression problem.

LS method consists of minimising the sum of squared errors:

$$\begin{aligned} J(\mathbf{X}) &= \sum_{t=1}^{n} (f(\mathbf{x}_t) - y_t)^2 \\ &= \sum_{t=1}^{n} (\mathbf{A}^\top \mathbf{x}_t - y_t)^2 \\ &= \|\mathbf{AX} - \mathbf{Y}\|_2^2 \end{aligned}$$

where

(3.11)

$$\mathbf{X} = \begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_n \\ | & | & & | \end{bmatrix} \quad , \quad \mathbf{A} = \begin{bmatrix} — & \mathbf{a}_1^\top & — \\ — & \mathbf{a}_2^\top & — \\ & \vdots & \\ — & \mathbf{a}_m^\top & — \end{bmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{bmatrix} | & | & & | \\ \mathbf{y}_1 & \mathbf{y}_2 & \dots & \mathbf{y}_n \\ | & | & & | \end{bmatrix}$$

and $f(\mathbf{x}_t) = \mathbf{A}^\top \mathbf{x}_t$.

To solve equation 3.11 is equivalent to find a solution of:

$$\underset{m\times n}{\mathbf{A}} \ \underset{n\times l}{\mathbf{X}} = \underset{m\times l}{\mathbf{Y}} \tag{3.12}$$

The optimal solution for $\mathbf{X}$ is:

$$\mathbf{X} = \mathbf{A}^+ \mathbf{Y} \tag{3.13}$$

where $\mathbf{A}^+$ is the Moore-Penrose pseudo-inverse which can be written as follows:

$$\mathbf{A}^+ = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top . \tag{3.14}$$

However, when $\mathbf{A}$ is not full rank, i.e $rank(\mathbf{A}) = k < n \leq m$, $\mathbf{A}^\top \mathbf{A}$ is always singular and equation (3.14) cannot be used. More generally, the pseudo-inverse is best computed using the compact singular value decomposition (SVD) of $\mathbf{A}$:

$$\underset{m\times n}{\mathbf{A}} = \underset{m\times k}{\mathbf{U_1}} \ \underset{k\times k}{\mathbf{\Sigma_1}} \ \underset{k\times n}{\mathbf{V_1}^\top} \tag{3.15}$$

as follows

$$\mathbf{A}^+ = \mathbf{V_1} \mathbf{\Sigma_1^{-1}} \mathbf{U_1^\top} \tag{3.16}$$

Proof can be found in the appendix section A.2.

In order to avoid the singularity of the matrix $\mathbf{A}^\top \mathbf{A}$, a regularisation term λ is introduced:

$$J(\mathbf{X}) = \|\mathbf{AX} - \mathbf{Y}\|_2^2 + \lambda \|\mathbf{X}\|^2 \tag{3.17}$$

which optimal solution $\mathbf{X}_*$ is well known:

$$\mathbf{X}_* = (\mathbf{A}^\top \mathbf{A} + \lambda \mathbb{I})^{-1} \mathbf{A}^\top y \, , \tag{3.18}$$

Proof can be found in section A.3.

The method is called ridge regression because the term $\lambda\mathbb{I}$ adds positive entries along the diagonal (ridge) to avoid the singularity of the covariance matrix $\mathbf{A}^\top \mathbf{A}$. This addition ensures that all of the covariance matrix eigenvalues will be strictly greater than 0, i.e the solution becomes unique.

3.2.6. The Lambda parameter. The additional term $\lambda\|\mathbf{X}\|_2^2$ in the optimisation problem shown in equation (3.17) has two effects on the solution: shrinks the coefficients towards zero and improves the conditioning of the problem.

When $\mathbf{A}$ is orthonormal then $\mathbf{A}^\top\mathbf{A} = \mathbb{I}$ and there is a simple relation between the ridge estimator and the OLS estimator:

$$\begin{aligned}\mathbf{X}_*(\lambda) &= (\mathbf{A}^\top\mathbf{A} + \lambda\mathbb{I})^{-1}\mathbf{A}^\top\mathbf{Y} \\ &= (\mathbb{I} + \lambda\mathbb{I})^{-1}\mathbf{A}^\top\mathbf{Y} \\ &= (1+\lambda)^{-1}\mathbb{I}\mathbf{A}^\top\mathbf{Y} \\ &= (1+\lambda)^{-1}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{Y} \\ &= (1+\lambda)^{-1}\mathbf{X}\end{aligned}$$

Figure 3.5 shows a visual example of the shrinking of the coefficients:

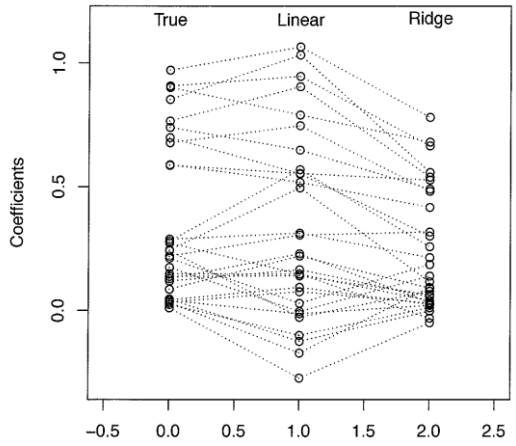


FIGURE 3.5. Shrink of regression coefficients

On the other hand, the effect of adding the term $\lambda\mathbb{I}$ to the matrix $\mathbf{A}^\top\mathbf{A}$ (equation (3.18)) improves its condition number since it increases its diagonal values when $\lambda > 0$. The matrix $\mathbf{A}^\top\mathbf{A}$ is symmetrical ($(\mathbf{A}^\top\mathbf{A})^\top = \mathbf{A}^\top\mathbf{A}$) and therefore diagonalizable. If we know the eigenvalue decomposition of $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, then:

$$\begin{aligned}\mathbf{A}^\top\mathbf{A} + \lambda\mathbb{I} &= \mathbf{V}\mathbf{\Sigma}^\mathbf{2}\mathbf{V}^\top + \lambda\mathbf{V}\mathbf{V}^\top \\ &= \mathbf{V}(\mathbf{\Sigma}^2 + \lambda\mathbb{I})\mathbf{V}^\top,\end{aligned}$$

where

$$\mathbf{\Sigma}^2 + \lambda\mathbb{I} = \begin{bmatrix}\sigma_1^2+\lambda & & & \\ & \sigma_2^2+\lambda & & \\ & & \ddots & \\ & & & \sigma_n^2+\lambda\,.\end{bmatrix}$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$.

Since the condition number of a matrix $\mathbf{A}$ is defined as:

$$\kappa = \|\mathbf{A}\|\|\mathbf{A}^{-1}\|$$

If matrix $\mathbf{A}$ is non-singular, its condition number can be expressed in terms of its singular values. The effect of adding the regularization term affects the condition number as follows:

$$\begin{aligned} \kappa_{ols} &= \|\mathbf{A}\|\|\mathbf{A}^{-1}\| = \frac{\sigma_1}{\sigma_n} \\ \kappa_{ridge} &= \|\mathbf{A}^\top\mathbf{A} + \lambda\mathbb{I}\|\|(\mathbf{A}^\top\mathbf{A} + \lambda\mathbb{I})^{-1}\| = \frac{\sigma_1 + \lambda}{\sigma_n + \lambda} \end{aligned}$$

It is easy to see that the term λ improves the condition number:

$$\frac{\sigma_1 + \lambda}{\sigma_n + \lambda} < \frac{\sigma_1}{\sigma_n} \qquad \forall \quad \lambda > 0$$

However, λ cannot be too large. Tipically λ is small and its magnitude depends on the matrix $\mathbf{A}$.

For rank deficient matrices we know that $\det(\mathbf{A}\mathbf{A}^\top) = 0$, adding the term $\lambda\mathbb{I}$ we have that $\det(\mathbf{A}\mathbf{A}^\top + \lambda\mathbb{I}) = p(\lambda)$ where $p(\lambda)$ is a polynomial of degree n ($\mathbf{A}$ is $m \times n$). The zeros of $p(\lambda)$ are discretes, so it can be represented as:

$$p(\lambda) = \lambda(\lambda - \lambda_1)^{n_1}(\lambda - \lambda_2)^{n_2} \ldots (\lambda - \lambda_s)^{n_s}$$

where $n_1 + n_2 + \cdots + n_s = n$.

This means that λ must be small in order to ensure that $p(\lambda)$ does not vanish.

3.2.7. Selection of Lambda. One of the ways to determine parameter λ is using the bias-variance tradeoff (see section 3.2.2). This parameter is crucial for ridge regression since it could reduce the expected prediction error by reducing variance, considering a biased estimator. It is known that the prediction error can be express as a decomposition between bias and variance.

The solution of OLS is well known $\hat{\mathbf{X}} = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{Y}$, and its bias ($Bias(\hat{f}(\hat{\mathbf{X}}))$ is 0 (Proof in section A.4).

The bias of ridge regression when $\mathbf{A}\mathbf{A}^\top$ is non-singular can be obtained expressing ridge regression solution λ in terms of OLS solution $\hat{\mathbf{X}}$:

$$Bias(\mathbf{X}(\lambda)) = \mathbf{W}\mathbf{X} - \mathbf{X} \neq 0 \tag{3.19}$$

where $\mathbf{W} = (\mathbb{I} + \lambda(\mathbf{A}^\top\mathbf{A})^{-1})^{-1}$

The variance of OLS is:

$$Var(\hat{\mathbf{X}}) = \sigma^2(\mathbf{A}^\top\mathbf{A})^{-1}$$

and the variance of ridge regression is:

$$Var(\mathbf{X}(\lambda)) = \sigma^2 \mathbf{W}(\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{W}^\top$$

Proof in section A.4.

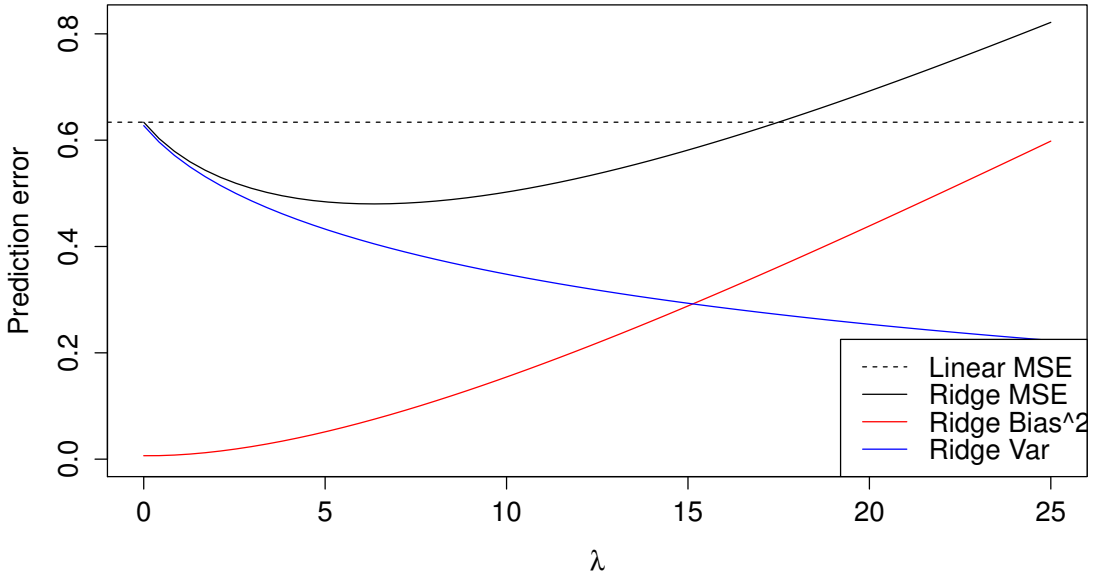


FIGURE 3.6. Bias-variance tradeoff depending on lambda. Dotted line corresponds to OLS prediction error (invariant with lambda). Black line corresponds to RR prediction error which can be decomposed in $Bias^2$ (in red) + Variance (in blue).

The figure 3.6 shows the bias-variance tradeoff given by equation (3.7).

Despite OLS has zero bias, its variance is greater than ridge for small values of λ. It can be shown that, in terms of prediction error, ridge (black line) is lower than OLS (dotted line) [**?**]. Ridge regression shows an increasing squared bias and a decreasing variance. See proof in section A.5.

3.2.8. Machine learning algorithms. In recent years many successful machine learning applications have been developed. Artificial neural networks (ANN) and Support Vector Machines (SVM) have been some of the most popularly used machine learning algorithm [**?**]. Historically, SVMs emerged after ANN.

The main characteristics of neural networks are that they have the ability to learn complex nonlinear input-output relationships, they use sequential training procedures they adapt themselves to the data.

ANN is inspired by biological learning systems organised into layers and have unidirectional connections between the layers, feed-forward networks (FFN) are the most used. FFN consist of a series of layers. The first layer has a connection from the network

input. Each subsequent layer has a connection from the previous layer. The final layer produces the network's output. Figure 3.7 shows a FFN with its different layers.

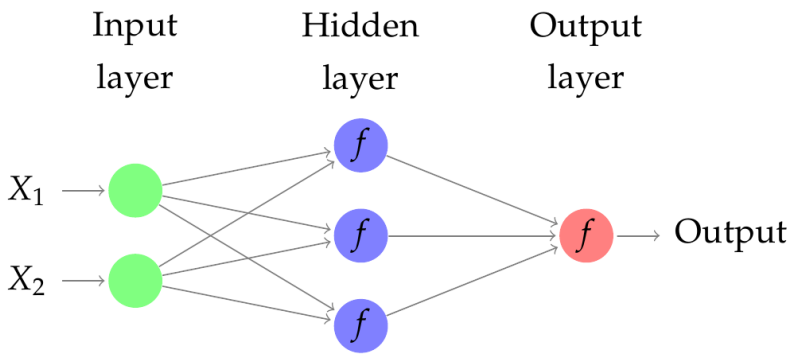


FIGURE 3.7. Feed-forward neural network (FFN). FNN consists of an input layer, an output layer and at least one hidden layer between the input and output layer.

Support Vector Machine (SVM) was introduced in 1992 by Vapnik and his coworkers [**?**]. In its original form, SVM is a training algorithm for linear classification. Only later it was used for regression, principal component analysis, novelty detection and also for non-linear case. However, unlike ANN, it is very well based on theory in statistical learning [**?**]. SVM tunes the capacity of maximising the margin between the training patterns and the decision boundary. The solution is expressed as a linear combination of supporting patterns, which are the training patterns close to the decision boundary, called the support vectors.

The major difference between SVM and ANN is the error optimisation. In ANN, the aim of learning is to obtain a set of weight values which minimise the training error while training adjust the capacity of the machine.

For nonlinear case, SVM mapped the data sets of input space into a higher dimensional feature space, which is linear and the large margin learning algorithm is then applied. However, the mapping can be implicitly done by kernel functions. In the high dimensional feature space, simpler and linear hyper plane classifiers that have maximal margin between the classes can be obtained.

3.3. Online learning

Online learning is a supervised machine learning framework that is useful when we have sequential access to a sample only once. This differs from the classical batch learning, where there is an entire data set available and the learner can build the internal model without any limits in accessing the data. In batch learning, there is time enough to carefully analyse the dataset, build large predictive models and combine them in a sophisticated way. However, when execution time is critical and new data is required to be included in the model, online learning algorithms are a better option.

In batch learning, the training phase is commonly a computationally expensive process. Therefore, when new data arrives it can't be included easily into the model. Moreover, it could happen that we won't have enough time to process the new data before more data arrives. In online learning algorithms there is no training phase, the model is updated and evaluated at every time step. This model updating is computationally less expensive than a training phase.

Online algorithms allow incremental learning by processing one instance at a time. This is done updating the current model instead of building the model from scratch.

Online learning is also useful when some past data may be irrelevant or we want to improve computational efficiency. However, the use of less historical data could affect accuracy.

The goal of online learning is the same as batch learning, predicting targets as accurate as possible. For example, stock market prediction can be seen as online learning. The algorithm makes a prediction of the stock, little time after the real stock price is available, this information can be incorporated to the learner to further improve the prediction accuracy. In general, there is too much data available in an online learning setup, the data set grows continuously. Offline learning has equal or superior accuracy compared to online learning when the same amount of data is used.

Classic statistical theory of sequential prediction enforces strong assumptions on the statistical properties of the input sequence (for example, stationary stochastic process). However, these assumptions can be unknown or change over time. In online learning there is no previous assumption about the data and the sequence is allowed to be deterministic, stochastic or even adaptive.

Moreover, in case we receive data streams, ANN or SVM cannot introduce new information into the model without a re-training process, so we will have to use the same non-updated model until we decide to compute another one if it is possible. Online learning algorithms allow one example at a time to be introduced into an existing model incrementally [**?**]. This is extremely important when the problem has large data streams and real-time forecasting must be done. This is the most common scenario when we want to forecast a wide range of data such as stock prices and volatilities, electricity power, intrusion detection, web-mining, server load, etc. Besides, many problems of

high interest in machine learning can be treated as online ones and they can also use these types of algorithms.

The online learning framework was first introduced in the perceptron algorithm [**?**]. There are several other widely used online methods such as passive-aggressive [**?**], stochastic gradient descent [**?**], aggregating algorithm [**?**] and the second order perceptron [**?**]. In [**?**] an in-depth analysis of online learning is provided. Applications in finance has been widely used: study presented by [**?**] applied ridge regression in an online context and more recently, time series forecasting using online learning has been presented [**?**].

The motivation for online learning is to obtain computational efficiency and tackle the shifting problem, i.e. that the distribution of the data is unknown or changes over time. Online learning algorithms can deal with this problem because they have a tracking ability which is a strategy based on retaining weak dependence on past examples by using two types of models:

a) **memory boundedness:** consists of limiting the number of support vectors in order to improve computational efficiency. One example of this is the budget perceptron [**?**] which reduces the number of examples used for prediction. Alternatively, in the forgetron algorithm [**?**] the damage caused by removing old examples is discussed, which can be avoided by removing samples with small influences. Other examples are the sliding window kernel (RLS) [**?**], which only considers a sliding window of the most recent data, and in [**?**] is shown a variant of aggregating algorithm for regression [**?**] considering only a sliding window of the most recent data, optimising also common matrix operations.

b) **weight decay:** one example of this is the shifting perceptron algorithm which implements an exponential decaying scheme for the examples [**?**]. Performance of an online learning algorithm is measured by the cumulative loss it suffers along its run on a sequence of examples. In order to minimise this loss, the learner may update the hypothesis after each round so as to be more accurate in later rounds.

There is sometimes confusion about online and incremental learning concepts. Incremental learning refers to any online learning process that learns the same model as would be learnt by a batch learning algorithm.

Incremental learning is useful when the input to a learning process is stream data, with the need or desire to be able to use the result of learning at any point in time, based on the input observations received so far.

Incremental learning is very useful when there is no need to record fundamental data and only a summary needs to be retained. Due to this, incremental algorithms are often characterised as memoryless, because no memory of past data is required. The algorithm is online but not incremental if it doesn't produce the same result for all observations that the corresponding batch algorithm would for these same observations.

Algorithm 1 shows the online learning algorithm structure:
where l is some loss function. Performance is later measured after T trials as:

Algorithm 1 Structure of a Learning System

1: Receives input $\mathbf{x}_t$
2: Makes prediction $\hat{\mathbf{y}}_t$
3: Receives response $\mathbf{y}_t$
4: Incurs loss $l_t(\mathbf{y}_t, \hat{\mathbf{y}}_t)$

$$L_T = \sum_{t=1}^{T} l_t(\mathbf{y}_t, \hat{\mathbf{y}}_t)$$

The objective is to minimise this loss function for all instances. The quality of online learning algorithms is measured by a quantity known as regret which is the difference between the performance of the online algorithm and its optimal predictor $E^* \in \Theta$ given by:

$$L_T^* = \min_{E \in \Theta} L_T^E \,,$$

where $L_T^E = \sum_{t=1}^{T} l_t(\mathbf{y}_t, \mathbf{y}_t^E)$ and $\mathbf{y}_t^E$ is the expert estimation.

Therefore regret is defined as:

$$R_T = L_T - L_T^*$$

3.3.1. Online Ridge Regression. Online RR is the online formulation of the regularised Least Squares method and is based on the following equivalent formulation of the RR optimal solution.

Since

$$\mathbf{A} = \begin{bmatrix} — & \mathbf{a}_1^\top & — \\ — & \mathbf{a}_2^\top & — \\ & \vdots & \\ — & \mathbf{a}_m^\top & — \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} — & \mathbf{b}_1 & — \\ — & \mathbf{b}_2 & — \\ & \vdots & \\ — & \mathbf{b}_m & — \end{bmatrix} \tag{3.20}$$

equation (3.18) can also be written as:

$$\begin{aligned} \mathbf{X}_{\text{ridge}} &= (\mathbf{A}^\top \mathbf{A} + \lambda \mathbb{I})^{-1} \mathbf{A}^\top \mathbf{Y} \\ &= \Big(\sum_{t=1}^{m} \mathbf{a}_t \mathbf{a}_t^\top + \lambda \mathbb{I} \Big)^{-1} \sum_{t=1}^{m} \mathbf{a}_t \mathbf{y}_t \,. \end{aligned}$$

Lets define $\mathbf{S} = \sum_{t=1}^{m} \mathbf{a}_t \mathbf{a}_t^\top + \lambda \mathbb{I}$ and $\mathbf{W} = \sum_{t=1}^{m} \mathbf{a}_t \mathbf{y}_t$, so the algorithm 2 shows the iterative formulation:

Algorithm 2 Online Ridge Regression

Require:

$\{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$: m input vectors

$\{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$: m target vectors

λ: regularization parameter

Ensure:

$\{f(\mathbf{a}_1), \ldots, f(\mathbf{a}_m)\}$: model predictions

Initialize $\mathbf{S} = \lambda\mathbb{I}$ and $\mathbf{W} = 0$
for $t = 1$ to m **do**
  read new $\mathbf{a}_t$
  $\mathbf{X} = \mathbf{S}^{-1}\mathbf{W}$
  output prediction $f(\mathbf{a}_t) = \mathbf{X}^\top \mathbf{a}_t$
  $\mathbf{S} = \mathbf{S} + \mathbf{a}_t\mathbf{a}_t^\top$
  Read new y_t
  $\mathbf{W} = \mathbf{W} + \mathbf{a}_t\mathbf{y}_t$
end for

3.3.2. The Aggregating Algorithm for Regression. The AAR, proposed by [**?**] (also known as Vovk-Azoury-Warmuth predictor [**?**]), is an application of the aggregating algorithm to the problem of regression. The idea is introduce the new input vector $\mathbf{x}_{m+1}$ to solve the model parameters:

$$\mathbf{X}_{aar} = \Big(\sum_{t=1}^{m+1} \mathbf{a}_t\mathbf{a}_t^\top + \gamma\mathbb{I} \Big)^{-1} \sum_{t=1}^{m} \mathbf{a}_t\mathbf{y}_t \,. \tag{3.21}$$

If we define $\mathbf{S} = \sum_{t=1}^{m+1} \mathbf{a}_t\mathbf{a}_t^\top + \gamma\mathbb{I}$ and $\mathbf{W} = \sum_{t=1}^{m} \mathbf{a}_t\mathbf{y}_t$, the algorithm 3 is slightly different to the algorithm 2, which updated matrix $\mathbf{S}$ before making the prediction.

3.3.3. Competitive analysis. Competitive analysis was designed for analysing the performance of an online algorithm compared with its optimal offline algorithm. An optimal offline algorithm can view the sequences of requests in advance. The effectiveness of an online algorithm [**?**] may be measured by its competitive ratio which is the worst-case ratio of the online algorithm and the optimal offline algorithm.

Algorithm 3 *The aggregating algorithm for regression*

Require:

$\{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$: m input vectors

$\{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$: m target vectors

λ: regularisation parameter

Ensure:

$\{f(\mathbf{a}_1), \ldots, f(\mathbf{a}_m)\}$: model predictions

Initialize $\mathbf{S} = \lambda\mathbb{I}$ and $\mathbf{W} = 0$

for $t = 1$ to m **do**

  read new $\mathbf{a}_t$

  $\mathbf{S} = \mathbf{S} + \mathbf{a}_t\mathbf{a}_t^\top$

  $\mathbf{X} = \mathbf{S}^{-1}\mathbf{W}$

  output prediction $f(\mathbf{a}_t) = \mathbf{X}^\top\mathbf{a}_t$

  Read new $\mathbf{y}_t$

  $\mathbf{W} = \mathbf{W} + \mathbf{a}_t\mathbf{y}_t$

end for

3.4. Evaluation methods

Forecast performance is evaluated using different methods. We have chosen three measures usually used:

MAPE: Mean Average Percent Error which presents forecast errors as a percentage.

$$\text{MAPE} = \frac{1}{N}\sum_{t=1}^{N}\frac{|\mathbf{y}_t - \hat{\mathbf{y}}_t|}{|\mathbf{y}_t|} \times 100 \tag{3.22}$$

MAE: Mean Average Error which measures the distance between forecasts to the true value.

$$\text{MAE} = \frac{1}{N}\sum_{t=1}^{N}|\mathbf{y}_t - \hat{\mathbf{y}}_t| \tag{3.23}$$

MSE: Mean Square Error measures the distance between forecasts and the true values and large deviations from the true value have a large impact due to squaring forecast error.

$$\text{MSE} = \frac{\sum_{t=1}^{N}(\mathbf{y}_t - \hat{\mathbf{y}}_t)^2}{N} \tag{3.24}$$

RMSE: Root Mean Square Error also measures the distance between forecasts to the true values but, unlike MAE, large deviations from the true value have

a large impact on RMSE due to squaring forecast error.

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{N}(\mathbf{y}_t - \hat{\mathbf{y}}_t)^2}{N}} \tag{3.25}$$

***U*-statistic:** the Theil's U-statistic, presented by [**?**], is a unit free measure obtained as the ratio between the root MSE (RMSE) of a model and the RMSE of the naive random walk model.

3.5. Model selection

Akaike Information Criterion (AIC) is often used in model selection where AIC with smaller values are preferred since they represent a trade-off between bias and variance. AIC is obtained as follows:

$$AIC = \underbrace{-\frac{2l}{N}}_{\text{bias}} + \underbrace{\frac{2k}{N}}_{\text{variance}} \tag{3.26}$$

where

l: is the loglikelihood function
k: number of estimated parameters
N: number of observations

Loglikelihood function is obtained from the Residual Sum of Squares (RSS):

$$l = -\frac{N}{2}\left(1 + ln(2\pi) + ln\left(\frac{RSS}{N}\right)\right) \tag{3.27}$$

Chapter 4

Fast and adaptive cointegration based model for forecasting high frequency financial time series

---

Cointegration is a long-run property of some non-stationary time series where a linear combination of those time series is stationary. This behaviour has been studied in finance because cointegration restrictions often improve forecasting. The Vector Error Correction Model (VECM) is a well-known econometric technique that characterises short-run variations of a set of cointegrated time series incorporating long-run relationships as an error correction term. VECM has been broadly used with low frequency time series. We aimed to adapt VECM to be used in finance with high frequency stream data.

Cointegration relations change in time and therefore VECM parameters must be updated when new data is available. We studied how forecasting performance is affected when VECM parameters and the length of historical data used change in time. We observed that the number of cointegration relationships varies with the length of historical data used. Moreover, parameters that increased these relationships in time led to better forecasting performance. Our proposal, called an Adaptive VECM (AVECM) is to make a parameters grid search that maximises the number of cointegration relationships in the near past. To ensure the search can be executed fast enough, we used a distributed environment .

The methodology was tested using four 10-second frequency time series of the Foreign Exchange market. We compared our proposal with ARIMA and the naive forecast of the random walk model. Numerical experiments showed that on average AVECM

performed better than ARIMA and random walk. Additionally, AVECM significantly improved execution times with respect to its serial version.

This work is published in [**?**].

4.1. Introduction

In finance, it is common to find variables with long-run equilibrium relationships. This is termed cointegration and reflects the idea that some sets of variables cannot wander too far from each other. Cointegration means that one or more linear combinations of these variables are stationary even though individually they are not. Some models, such as the Vector Error Correction (VECM), see [**?**], take advantage of this property and describe the joint behaviour of several cointegrated variables. VECM introduces this long-run relationship among a set of cointegrated time series as an error correction term. These time series must be integrated of order 1, denoted I(1), i.e. they become stationary at their first differences. In finance, I(1) time series are very common and to introduce cointegration restrictions in models often improves forecasting, see [**?**]. Therefore, VECM has been widely adopted in financial applications, among others: [**?**], [**?**], [**?**] and [**?**]. VECM has also been used in pair trading, see [**?**], or models with more than two variables, see for example [**?**] and [**?**].

Cointegration relationships can be found in low and high frequency data of two or more assets. While cointegration in low frequency data is motivated by a long-run equilibrium relationship between economic forces, cointegration in high frequency data has its foundation in statistical arbitrage theory which is very helpful to detect mean-reverting trades. Information about cointegrated assets in high frequency data could be used as an input for high frequency trading strategies, using it as a signal to capture small profits in short term trades, see [**?**]. [**?**] addressed the benefits of using higher frequency data to analize cointegration. [**?**] also found that cointegration may differ with different data frequencies and could appear whith increased data frequency.

The use of VECM with high frequency data is mainly limited by computationally expensive routines. Firstly, VECM parameters are obtained using the ordinary least squares (OLS) method, developed by [**?**]. Since OLS involves many calculations, the parameter estimation is computationally expensive when the number of lagged values and data increases. Secondly, obtaining cointegration vectors is also an expensive routine because the Johansen method is required, which is of order $O(n^3)$ ([**?**]). [**?**] addressed the advantage of distributed processing over conventional rolling window processing. Therefore, our aim was to study if a parallel version of VECM can be used with high frequency stream data.

Our approach was to determine, adaptively, the number of observations and lags of VECM which maximise cointegration relations in the past in a distributed environment. We called our proposal Adaptive Vector Error Correction (AVECM). AVECM parallelises this search of parameters in order to update them before new data arrives. Model effectiveness is focused on out-of-sample forecast rather than in-sample fitting. This criterion allows AVECM prediction capability to be expressed rather than just explaining data history. The forecast capability of our method was measured using MSE and the Theil's U-statistic, see [**?**], widely used in economic forecast. Tests were run using four currency rates: Euro (EUR) to United States Dollar (USD) (EURUSD), British Pound (GBP) to USD (GBPUSD), USD to Swiss Franc (CHF) (USDCHF) and USD to Japanese Yen (JPY) (USDJPY) with a 10-second frequency.

The AVECM algorithm is presented in section 4.2. In section 4.3 we describe the tests carried on to assess the accuracy and the execution time of AVECM. This section also includes a description of the test data. Section 4.4 contains the conclusions and a discussion of future research.

4.2. Methodology

Cointegration vectors can be found applying the Johansen method which uses a sample of the last historical data. However, VECM assumes cointegration vectors do not change in time. In fact, [**?**] addresses that the long-run relationships between the time series might change due to several economic factors that can lead to structural breaks in the cointegration relationship. In order to show that the number of cointegration vectors depends on the amount L of historical data and the number of lags p in the VECM, we used a grid search. We arbitrarily defined a grid of possible values for L and p. L goes throughout $[2, 14]$ hours (1 *hour* = 360 data points) with a step size of 4 hours and p throughout $[1, 5]$ with a step size of 1. The idea was to show the variability of the number of cointegration vectors when we changed these two parameters. We used four forex rates: EURUSD, GBPUSD, USDCHF and USDJPY with 10-second frequency. Data started at 13:00 GMT of the 13th of August 2014, when the New York and London financial markets opened.

Figure 4.1 shows the distribution of the number of cointegration vectors given by the Johansen method for different values of $L = [2, 6, 10, 14]$ hours and $p = 1$. This procedure was carried out by a sliding window of historical data moving 1000 times. We observed that the distribution of cointegration vectors changed with different values of L. When $L = 2$ hours, there was no cointegration in more than 60% of the iterations. Cointegration increased when $L = 6$ hours and was maximum when $L = 10$ hours where one cointegration vectors was found for all 1000 iterations. The ocurrence of cointegration started to decrease at $L = 14$ hours.

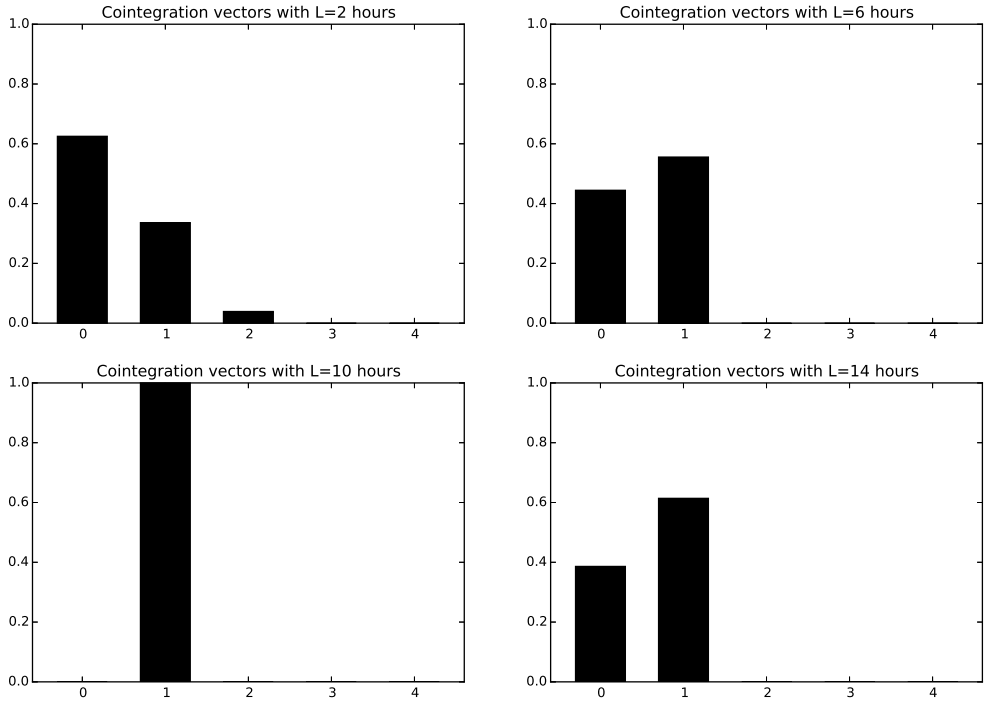


FIGURE 4.1. Distribution of the number of cointegration vectors using $p = 1$ lags. Four possible values for windows size L are shown (2, 6, 10 and 14) hours (1 hour = 360 data points).

From section 2.6 we know that $r = 0$ means no cointegration and $r = l$ (we are using four rates, so $l = 4$) reveals that no process is I(1) but stationary. The interesting cases of cointegration are those where r lies strictly between 0 and 4, i.e. $0 < r < 4$.

In order to measure the extent of cointegration, we introduce a *percentage of cointegration* as following:

$$PC = \frac{\#\{it \mid it \text{ has } r \text{ c.v. with } 0 < r < l\}}{\#it} \times 100 \tag{4.1}$$

where c.v. stands for cointegration vectors and it is the number of iterations.

The goal of our next experiment was to find a relationship between the ratio PC and the performance measure MSE (see equation 3.24). L was defined between $[2, 14]$ hours, that corresponded to $[720, 5040]$ data points, and p took values between $[1, 5]$.

Figure 4.2 shows the relationship between MSE and PC and L. We found that higher cointegration percentage leads to improved performance accuracy in terms of lower MSE. Also, increasing the size of the sliding window L doesn't necessarily help to reduce MSE.

Therefore we proposed to choose L and p in order to maximise the percentage of cointegration PC in the near past. This process was done every time that new data

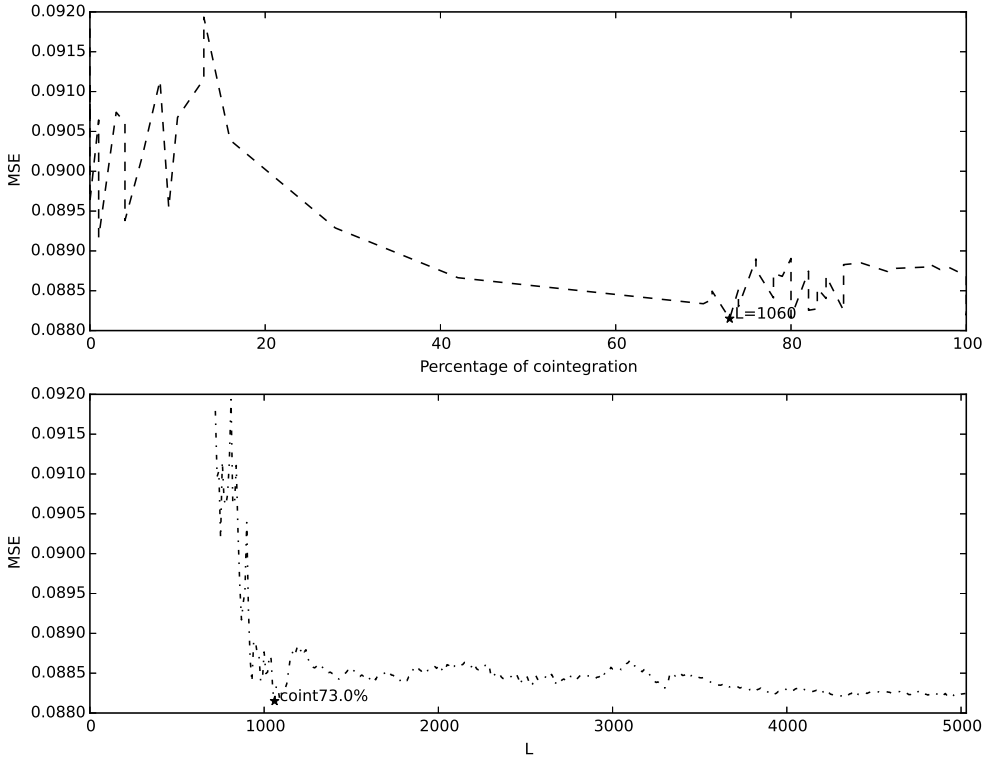


FIGURE 4.2. MSE versus the percentage of cointegration considering 1000 iterations. Optimum windows size L found was 1060. Below MSE versus L shows a rapidly decreasing behaviour, founding minimum at PC = 73%

was processed. However, this search can be slow if we try different values for L and p and therefore we distributed this calculation in order to reduce searching time.

Our proposal is then a modified version of VECM, with parameter p and the amount of historical data used L obtained at every step. Only the search of L and p was done in a distributed environment, since this is the most expensive routine. Our proposal is called Adaptive Vector Error Correction model (AVECM). AVECM is detailed in the algorithm 4 which summarises our proposal.

The input of AVECM is time series prices which are cointegrated. The starting point of testing is j and the total number of iterations is it. We need to ensure that j is at least the maximum value of Ls. Ls and ps are the possible values for L and p. The function **get_best_params** makes this grid search on the two vector lists Ls and ps and returns the parameters L and p which maximise the percentage of cointegration PC (see equation 4.1) for a pre-defined number of iterations m. This function is implemented in a distributed environment, thus ensuring a response before new data is available. After L and p parameters are found, VECM is built and used to forecast the next data point.

Algorithm 4 AVECM: Adaptive VECM.

Require:

$\mathbf{y}$: matrix with N input vectors and l time series
j: Starting point of testing
it: Ending point of testing
ps: list of p values
Ls: list of L values ($L < N$)
m: Iterations to determine parameters ($m < N - L$)

Ensure:

$\{\hat{\mathbf{y}}[1], \ldots, \hat{\mathbf{y}}[it]\}$: prediction vectors

1: **for** $i = j$ to it **do**
2: $\mathbf{Y} \leftarrow \mathbf{y}[:, i-1]$
3: $L, p \leftarrow$ `get_best_params`$(Ls, ps, m, \mathbf{Y})$
4: $model = \text{VECM}(\mathbf{Y}, L, p)$
5: $\hat{\mathbf{y}}[i-j] = model.predict()$
6: **end for**

4.2.1. Model comparison. We compared our proposal, in terms of performance, with the naive forecast of the random walk model and ARIMA. It is still difficult to outperform the random walk model for standard econometric forecasting models despite its simplicity, see [**?**]. The random walk model is defined as:

$$\mathbf{y}_t = \mathbf{y}_{t-1} + \epsilon_t \tag{4.2}$$

The naive forecast of the time series difference $\hat{\mathbf{y}}_{t+1}$ for the random walk model is defined as:

$$\hat{\mathbf{y}}_{t+1} = \mathbf{y}_t + \hat{\epsilon}_{t+1} \tag{4.3}$$

where $\hat{\epsilon}_{t+1} = \epsilon_t$.

On the other hand, ARIMA is widely used to forecast returns in finance, see [**?**]. A process can be modelled as an ARIMA(p, d, q) model if $\mathbf{x}_t = \Delta^d \mathbf{y}_t$, i.e after differencing d times the time series $\mathbf{y}_t$, we get an ARMA(p, q). Since we are modelling returns, we used $d = 1$.

4.3. Experimental results

4.3.1. Data. All the experiments and AVECM tests were carried out using four foreign exchange rates all related to USD: EURUSD, GBPUSD, USDCHF and USDJPY. We chose the most traded rates related to USD so they were likely to be cointegrated.

This data was collected from [**?**], a free database which gives access to the Swiss Foreign Exchange marketplace.

The tests were done using 10-second frequency from ask prices from the 11th to the 15th of August 2014. Since one day corresponds to 8640 data points and we used 5 days of data, we have 43,200 data points in total.

4.3.2. Unit root tests. Before running the tests, we firstly checked whether the time series were I(1) using the Augmented Dickey Fuller (ADF) test with lags $p = 1, 2, 3, 4, 5$. [**?**] presented critical values for rejection of hypothesis of a unit root: -2.62 (1%), 1.94 (5%) and 1.62 (10%). Table 1 shows that all currency rates cannot reject the unit root test in their level form considering different lags, but they rejected it with their first differences. This means that all of them are I(1) time series and we are allowed to use VECM and therefore our proposed AVECM.

Variable	ADF(1)	ADF(2)	ADF(3)	ADF(4)	ADF(5)
EURUSD	-0.052	-0.054	-0.054	-0.054	-0.054
GBPUSD	-0.744	-0.784	-0.805	-0.837	-0.846
USDCHF	-0.476	-0.493	-0.493	-0.495	-0.502
USDJPY	0.357	0.360	0.360	0.367	0.367
Δ EURUSD	-128.4*	-128.4*	-96.85*	-89.12*	-89.12*
Δ GBPUSD	-131.4*	-112.7*	-102.5*	-92.86*	-88.29*
Δ USDCHF	-127.8*	-127.8*	-96.94*	-88.82*	-80.79*
Δ USDJPY	-135.1*	-135.1*	-101.2*	-101.2*	-101.2*

* Indicates significance at 1% level
** Indicates significance at 5% level
*** Indicates significance at 10% level
MacKinnon critical values for rejection of hypothesis of a unit root are: -2.62 (1%), -1.94 (5%) and -1.62 (10%)
ADF(d) Augmented Dickey-Fuller test with lag d

TABLE 1. Unit roots tests for EURUSD, GBPUSD, USDCHF and USDJPY at 10-second frequency.

4.3.3. Performance accuracy. Algorithms AVECM, ARIMA and the naive random walk were tested using four days of data (from the 12th to the 15th of August 2014). For AVECM we considered different number of iterations (parameter m in algorithm 4): 10, 50 and 100. We tried 12, 24 and 47 different pair of combinations for L and p. Possible values for L were in the interval $[2, 14]$ hours and p can have values in between $[1, 5]$. Best AVECM performance was compared against ARIMA and the random walk model. Table 2 shows the out-of-sample performance measures: MSE and U-statistic for AVECM and ARIMA. In terms of both measures we found that AVECM

is superior to ARIMA and the naive random walk model. We also included the p-value that proves that the difference in the MSE is significant at the 99% significance level in three of the four currency rates and at 90% in the case of GBPUSD. The U-statistic shows that AVECM and ARIMA are superior to the naive random walk model and that our proposal is also superior to ARIMA.

TABLE 2. AVECM performance

	MSE			U-statistic	
	AVECM	ARIMA	p-value	AVECM	ARIMA
EURUSD	1.0702 e-09	1.1481 e-09	9.2509 e-12	0.6863	0.7108
GBPUSD	1.6630 e-09	1.7408 e-09	6.9519 e-02	0.6866	0.7025
USDCHF	5.8503 e-10	6.3545 e-10	2.8999 e-14	0.6803	0.7091
USDJPY	6.3483 e-06	6.5194 e-06	6.8536 e-05	0.6964	0.7057

4.3.4. Parallel implementation. To determine L and p based on maximising the percentage of cointegration requires use of the Johansen method which is a computationally expensive routine. This procedure is done by the function get_best_params shown in algorithm 4. In order to improve the execution time of this search, our proposal included a parallel search of VECM parameters using high performance computing. The main objective was to obtain a response before a new data arrived in the following 10 seconds.

The Johansen method is already programmed in the Python Statsmodels library, see [**?**], and the parallel implementation was developed using MPI in Python. We chose MPI because it allows large-scale parallel applications with wide portability to be built, being able to run in large clusters or on local computers. We tested our proposal in a cluster with 2 servers Xeon E5-2667 (2.90GHz) of 24 cores each (48 cores in total) and 24GB RAM. In order to compare serial and parallel execution times in AVECM, we set parameter it = 100 in algorithm 4.

The L parameter was always chosen between 2 and 14 hours and p always took values between 1 and 5. Parameter $nparams$ represents the number of pairs (L,p) used to maximise the percentage of cointegration.

Execution time depends directly on L, p and $nparams$ used, since they determine the size of matrix $\mathbf{A}$ (see equation 2.27) and therefore affect the OLS function execution time. Therefore, if we try more combinations of L and p (increasing $nparams$) the serial algorithm will take longer. For this financial time series we are interested in execution times below 10 seconds (the time series frequency).

The superior figure in 4.3 shows that best performance accuracy measurements are achieved in times near or below 10 seconds in the parallel version. Contrarily, serial times are higher, above 10 seconds in most cases. Figure 4.3 also shows the speed-up

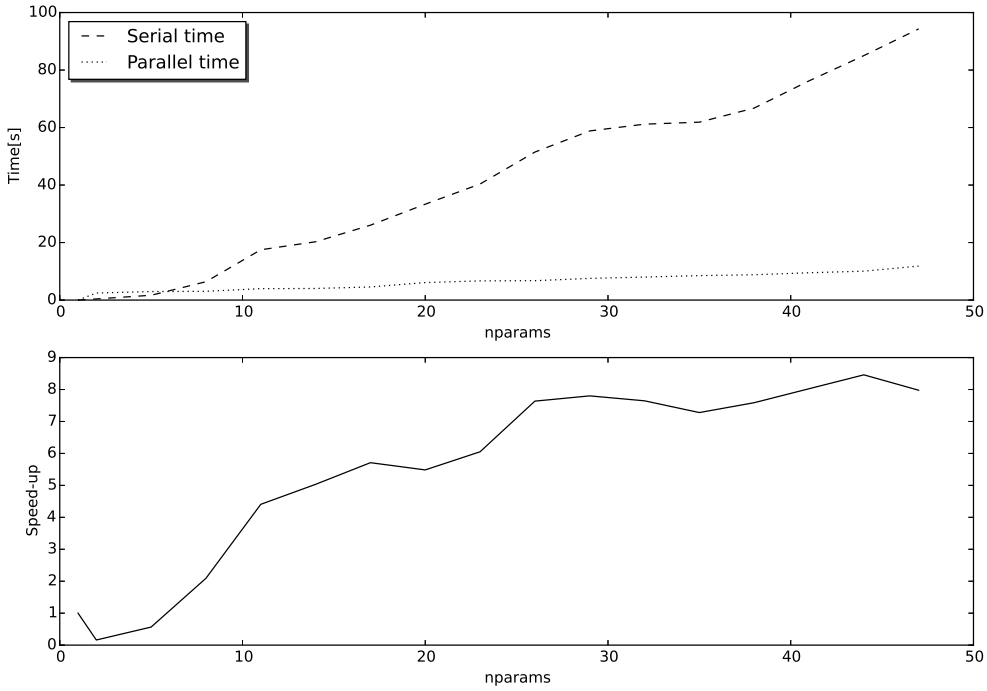


FIGURE 4.3. Computing time of sequential and parallel algorithm is shown in the upper figure. Speed-up is shown below.

in computing time for AVECM which is near 9X if we use more than 25 parameters (*nparams*).

Execution times do not consider the loading data time, just that of finding best parameters and matrix operations. The time MPI spends transferring data and synchronising processes is about two seconds independently of the number of processes considered. Execution times were measured using the Time python library.

4.4. Conclusions

Cointegration in financial time series has been largely studied and the Johansen method is commonly used to obtain cointegration relationships. In practice, it has been found that cointegration relations change with time. However, model-based cointegration such as VECM assumes that cointegration remains unchanged in time. We empirically showed that the Johansen method is sensitive to the number of lags but also to the amount of data considered.

Moreover, we introduced the notion of *percentage of cointegration* and found that out-of-sample forecast performance MSE is related to the value of this figure in the last samples. We used this information to set the model parameters. Our proposal AVECM

consists of an adaptive algorithm to update VECM parameters every time that new data is available. These parameters are found by maximising the percentage of cointegration of the last samples or iterations.

Despite the fact that high frequency Forex data can be spurious, the model performance can be less reliable (and more spurious) relative to the lower frequencies (such as 1 minute or 5 minute intervals) adopted by some other studies. However, the deficiency is offset by gain in accuracy from parallel processing which is capable of searching or examining a much larger state space given the same computational time.

Determining VECM parameters was the most expensive routine and it was run using parallel processes using MPI which allowed a grid search within a range of values for L and p to be made. Tests were done using real currency rates data.

Results showed that our proposed AVECM improves performance measures by finding parameters of L and p maximising the percentage of cointegration.

The parallel implementation allowed the execution times to be reduced more than 9 times and therefore a response time was obtain before 10 seconds. Since we used 10-second frequencies we can say that our proposal is suitable for use in an online context for real applications because response times were less than this frequency. Cointegration information can now easily be used as an integration tool to detect arbitrage opportunities or risk control.

For future study, it would be interesting to explore the relationship between cointegration and performance in order to propose new criteria for improving VECM parameters. It would also be interesting to include more explaining variables such as bid-ask spread and change in volume.

Chapter 5

An Online Vector Error Correction Model for Exchange Rates Forecasting

The Vector Error Correction Model (VECM) is an econometric model which characterises the joint dynamic behaviour of a set of cointegrated variables in terms of forces pulling towards equilibrium. In the previous proposal, the Adaptive VECM (AVECM) showed that to be able to use VECM at every time step can be used for forecasting purposes. However, to avoid extensive calculations, a parallel approach was implemented.

In this study, an Online VEC model (OVECM) is proposed, which optimises how model parameters are obtained using a sliding window of the most recent data. OVECM, unlike AVECM, updates the parameters at each step, instead of obtaining new ones. This proposal also takes advantage of the long-run relationship between the time series in order to obtain improved execution times.

Our proposed method is tested using four foreign exchange rates with a frequency of 1-minute, all related to the USD currency base.

OVECM is compared with VECM and ARIMA models in terms of forecasting accuracy and execution times. We show that OVECM outperforms ARIMA forecasting and enables execution time to be reduced considerably while maintaining good accuracy levels compared with VECM.

This work is published in [**?**].

5.1. The problem

Both VECM and VAR model parameters are obtained using ordinary least squares (OLS) method. Since OLS involves many calculations, the parameter estimation method is computationally expensive when the number of past values and observations increases. Moreover, obtaining cointegration vectors is also an expensive routine.

Recently, online learning algorithms have been proposed to solve problems with large data sets because of their simplicity and their ability to update the model when new data is available. The study presented by [**?**] applied this idea using ridge regression.

There are several popular online methods such as perceptron [**?**], passive-aggressive [**?**], stochastic gradient descent [**?**], aggregating algorithm [**?**] and the second order perceptron [**?**]. In [**?**], an in-deph analysis of online learning is provided.

In this proposal, an online formulation of the VECM called Online VECM (OVECM) is presented. OVECM is a lighter version of VECM which considers only a sliding window of the most recent data and introduces matrix optimizations in order to reduce the number of operations and therefore execution times. OVECM also takes into account the fact that cointegration vector space doesn't experience large changes with small changes in the input data.

OVECM is later compared against VECM and ARIMA models using four currency rates from the foreign exchange market with 1-minute frequency. VECM and ARIMA models were used in an iterative way in order to allow fair comparison. Execution times and forecast performance measures MAPE, MAE and RMSE were used to compare all methods.

Model effectiveness is focused on out-of-sample forecast rather than in-sample fitting. This criteria allows the OVECM prediction capability to be expressed rather than just explaining data history.

The next sections are organized as follows: the OVECM algorithm proposed is presented in section 5.2. Section 5.3 gives a description of the data used and the tests carried on to show accuracy and time comparison of our proposal against the traditional VECM and section 5.4 includes conclusions and a proposal for future study.

5.2. Methodology

Since VECM parameter estimation is computationally expensive, we propose an online version of VECM (OVECM). OVECM considers only a sliding window of the most recent data. Moreover, since cointegration vectors represent long-run relationships which vary little in time, OVECM determines firstly if they require calculation.

OVECM also implements matrix optimisations in order to reduce execution time, such as updating matrices with new data, removing old data and introducing new cointegration vectors.

Algorithm 5 shows our OVECM proposal.

Algorithm 5 OVECM: Online VECM

Require:

$\mathbf{y}$: matrix with N input vectors and l time series
p: number of past values
L: sliding window size ($L < N$)
mean_error: MAPE threshold
n: interpolation points to obtain MAPE

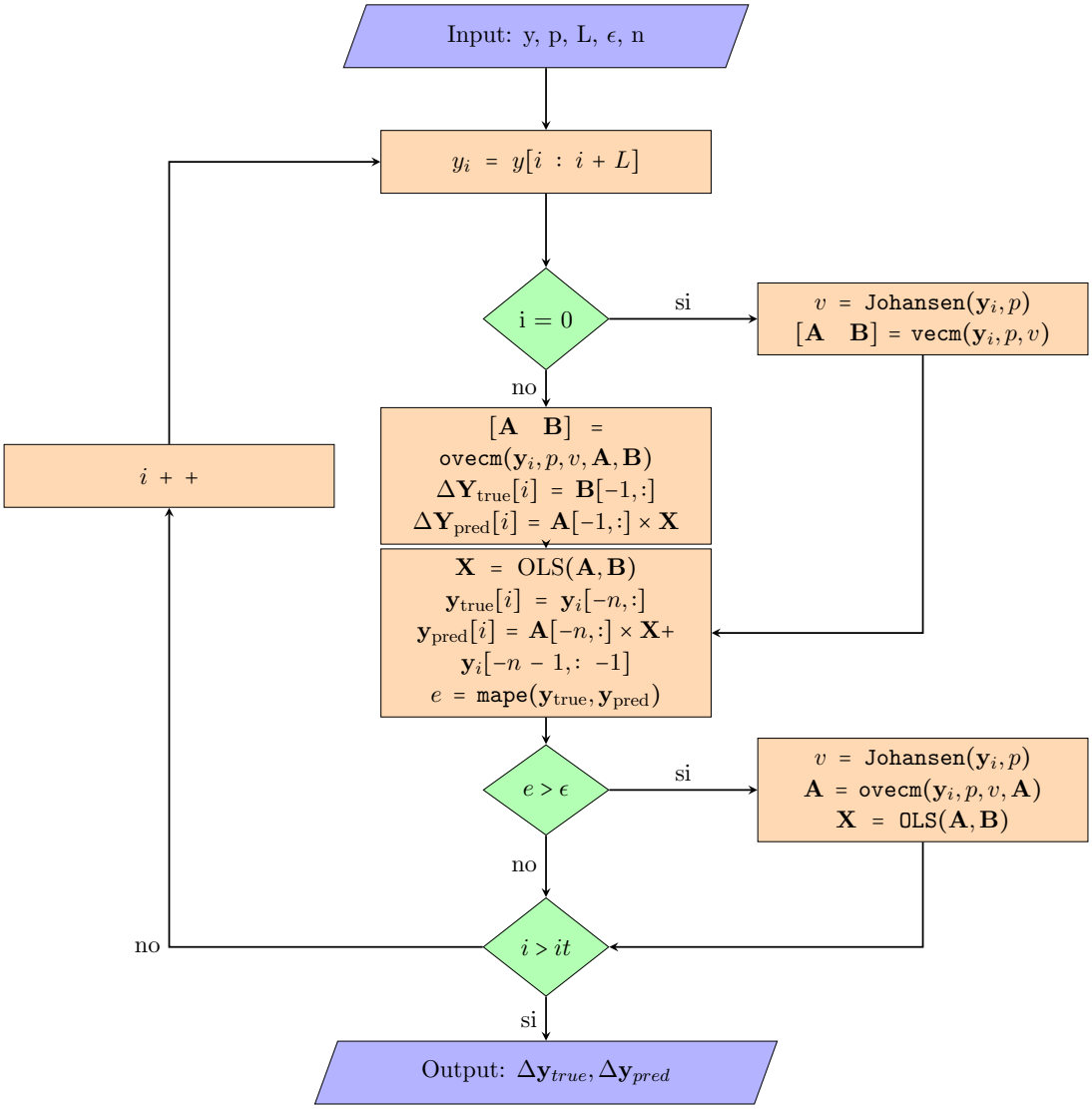
Ensure:

$\{\mathbf{y}_{\text{pred}}[L+1], \ldots, \mathbf{y}_{\text{pred}}[N]\}$: model predictions

1: **for** $i = 0$ to $N - L$ **do**
2: $\mathbf{y}_i \leftarrow \mathbf{y}[i : i + L]$
3: **if** i $= 0$ **then**
4: $v \leftarrow \texttt{getJohansen}(\mathbf{y}_i, p)$
5: $[\mathbf{A} \quad \mathbf{Y}] \leftarrow \texttt{vecMatrix}(\mathbf{y}_i, p, v)$
6: **else**
7: $[\mathbf{A} \quad \mathbf{Y}] \leftarrow \texttt{vecMatrixOnline}(\mathbf{y}_i, p, v, \mathbf{A}, \mathbf{Y})$
8: $\Delta\mathbf{Y}_{\text{pred}}[i] \leftarrow \mathbf{A}[-1, :] \times \mathbf{X}$
9: **end if**
10: $\mathbf{X} \leftarrow \texttt{OLS}(\mathbf{A}, \mathbf{Y})$
11: $e \leftarrow \texttt{mape}(\mathbf{Y}[-n, :], \mathbf{A}[-n, :] \times \mathbf{X})$
12: **if** $\texttt{mean}(e) >$ mean_error **then**
13: $v \leftarrow \texttt{getJohansen}(\mathbf{y}_i, p)$
14: $\mathbf{A} \leftarrow \texttt{vecMatrixUpdate}(\mathbf{y}_i, p, v, \mathbf{A})$
15: $\mathbf{X} \leftarrow \texttt{OLS}(\mathbf{A}, \mathbf{Y})$
16: **end if**
17: **end for**
18: $\mathbf{Y}_{\text{true}} \leftarrow \mathbf{Y}[L + 1 : N]$
19: $\mathbf{Y}_{\text{pred}} \leftarrow \mathbf{Y}[L : N - 1] + \Delta\mathbf{Y}_{\text{pred}}$

- The function `getJohansen` returns cointegration vectors given by the Johansen method considering the trace statistic test at 95% significance level.
- The function `vecMatrix` returns the matrices (2.27) and (2.24) that allows VECM to be solved.
- The function `vecMatrixOnline` returns the matrices (2.27) and (2.24) aggregating new data and removing the old one, avoiding calculation of the matrix $\mathbf{A}$ from scratch.
- Out-of-sample outputs are saved in the variables $\mathbf{Y}_{\text{true}}$ and $\mathbf{Y}_{\text{pred}}$.

- The model is solved using OLS.
- In-sample outputs are saved in the variables $\Delta\mathbf{y}_{\text{true}}$ and $\Delta\mathbf{y}_{\text{pred}}$.
- The function `mape` gets the in-sample MAPE for the l time series.
- Cointegration vectors are obtained at the beginning and when they are required to be updated. This updating is decided based on the in-sample MAPE of the last n inputs. The average of all MAPEs are stored in the variable e. If the average of MAPEs (`mean`(e)) is above a certain error given by the mean_error threshold, cointegration vectors are updated.
- If new cointegration vectors are required, the function `vecMatrixUpdate` only updates the corresponding columns of matrix $\mathbf{A}$ affected by those vectors (see equation 2.27).



A pseudocode of the algorithm 5 is detailed in the figure5.2.

Our proposal was compared against VECM and ARIMA. Both algorithms were adapted to an online context in order to get a reasonable comparison with our proposal (see algorithms 6 and 7). VECM and ARIMA are called with a sliding window of the most recent data, whereby the models are updated at every time step.

Algorithm 6 SLVECM: Sliding window VECM

Require:

$\mathbf{y}$: matrix with N input vectors and l time series
p: number of past values
L: sliding window size ($L < N$)

Ensure:

$\{\mathbf{y}_{\text{pred}}[L+1], \ldots, \mathbf{y}_{\text{pred}}[N]\}$: model predictions

1: **for** $i = 0$ to $N - L$ **do**
2: $\quad \mathbf{y}_i \leftarrow \mathbf{y}[i : i + L + 1]$
3: $\quad model = VECM(\mathbf{y}_i, p)$
4: $\quad \mathbf{Y}_{\text{pred}}[i] = model.predict(\mathbf{y}[i + L])$
5: **end for**
6: $\mathbf{Y}_{\text{true}} = \mathbf{y}[i + L + 1 : N]$

Since we know out time series are I(1) SLARIMA is called with $d = 1$. ARIMA is executed for every time series.

Algorithm 7 SLARIMA: Sliding window ARIMA

Require:

$\mathbf{y}$: matrix with N input vectors and l time series
p: autoregressive order
d: integrated order
q: moving average order
L: sliding window size ($L < N$)

Ensure:

$\{\mathbf{y}_{\text{pred}}[L+1], \ldots, \mathbf{y}_{\text{pred}}[N]\}$: model predictions

1: **for** $i = 0$ to $N - L$ **do**
2: $\quad$ **for** $j = 0$ to $l - 1$ **do**
3: $\quad\quad \mathbf{y}_i \leftarrow \mathbf{y}[i : i + L + 1, j]$
4: $\quad\quad model = ARIMA(\mathbf{y}_i, (p, d, q))$
5: $\quad\quad \mathbf{Y}_{\text{pred}}[i, j] = model.predict(\mathbf{y}[i + L, j])$
6: $\quad$ **end for**
7: **end for**
8: $\mathbf{Y}_{\text{true}} = \mathbf{y}[i + L + 1 : N, :]$

Both OVECM and SLVECM time complexity is dominated by Johansen method which is $O(n^3)$. Thus, both algorithms order is $O(Cn^3)$ where C is the number of iterations.

5.3. Experimental results

5.3.1. Data. Tests of SLVECM, SLARIMA and our proposal OVECM were carried out using foreign four exchange data rates: EURUSD, GBPUSD, USDCHF and USDJPY. This data was collected from the free database Dukascopy which gives access to the Swiss Foreign Exchange Marketplace [**?**].

The reciprocal of the last two rates (CHFUSD, JPYUSD) were used in order to obtain the same base currency for all rates. The tests were done using 1-minute frequency from ask prices which corresponded to 1.440 data points per day from the 11th to the 15th of August 2014.

5.3.2. Unit root tests. Before running the tests, we firstly checked if they were I(1) time series using the Augmented Dickey Fuller (ADF) test.

Table 1. Unit roots tests

	Statistic	**Critical value**	**Result**
EURUSD	-0.64	-1.94	True
ΔEURUSD	-70.45	-1.94	False
GBPUSD	-0.63	-1.94	True
ΔGBPUSD	-54.53	-1.94	False
CHFUSD	-0.88	-1.94	True
ΔCHFUSD	-98.98	-1.94	False
JPYUSD	-0.65	-1.94	True
ΔJPYUSD	-85.78	-1.94	False

Table 1 shows that all currency rates cannot reject the unit root test but they rejected it with their first differences. This means that all of them are I(1) time series and we are allowed to use VECM and therefore OVECM.

5.3.3. Parameter selection. In order to set OVECM parameters: windows size L and lag order p, we propose to use several window sizes: $L = 100, 400, 700, 1000$. For every window size L we chose lag order with minimum AIC.

ARIMA parameters were also obtained using AIC. Parameters optimisation is presented in table 2:

TABLE 2. Parameters optimisation. VECM order and ARIMA parameters were selected using AIC.

Windows size L L	VECM order (p)	ARIMA order (p, d, q)
100	2	p=2,d=1,q=1
400	5	p=1,d=1,q=1
700	3	p=2,d=1,q=1
1000	3	p=2,d=1,q=1

In OVECM we also define a mean_error variable, which was defined based on the in-sample MAPEs. Figure 5.1 shows how MAPE moves and how mean_error variable help us to decide whether new cointegration vectors are needed.

5.3.4. Execution times. We ran OVECM and SLVECM 400 iterations. SLARIMA execution time is excluded because its is not comparable with OVECM and SLVECM. SLARIMA was created based on statsmodels library routine ARIMA.

The execution times are shown in the table 3.

TABLE 3. Execution times

	L	order	e	Time[s]
OVECM	100	p=2	0	2.492
OVECM	100	p=2	0.0026	1.606
SLVECM	100	p=2	–	2.100
OVECM	400	p=5	0	3.513
OVECM	400	p=5	0.0041	2.569
SLVECM	400	p=5	–	3.222
OVECM	700	p=3	0	3.296
OVECM	700	p=3	0.0032	2.856
SLVECM	700	p=3	–	3.581
OVECM	1000	p=3	0	4.387
OVECM	1000	p=3	0.0022	2.408
SLVECM	1000	p=3	–	3.609

It is clear that execution time depends directly on L and p since they determine the size of matrix $\mathbf{A}$ and therefore affect the OLS function execution time. It is worthy of note that execution time also depends on mean_error because it determines how many times OVECM will recalculate cointegration vectors which is an expensive routine.

Figure 5.1 shows an example of the in-sample MAPE for 50 iterations. When the average of the in-sample MAPEs is above mean_error new cointegration vectors

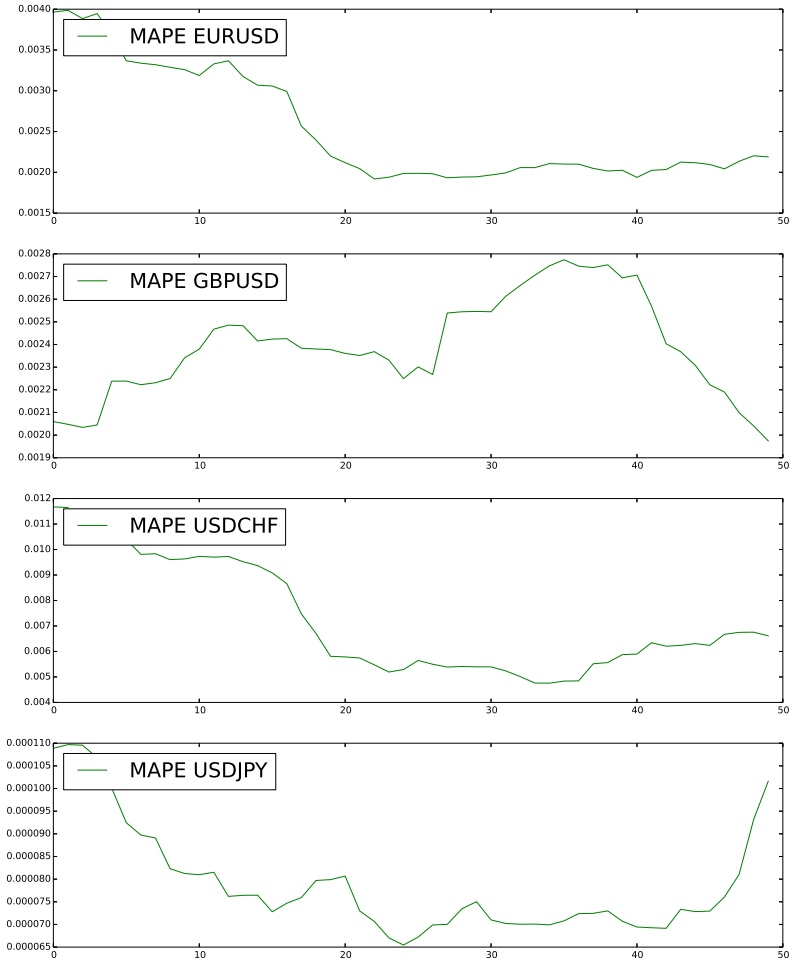


FIGURE 5.1. In-sample MAPEs example for 50 minutes. The average of them is considered to obtain new cointegration vectors.

are obtained. In consequence, OVECM performance increases when mean_error increases. However, this could affect accuracy, but table 4 shows that using an appropriate mean_error doesn't affect accuracy considerable.

5.3.5. Performance accuracy. Table 4 shows in-sample and out-of-sample performance measures: MAPE, MAE and RMSE for OVECM, SLVECM and SLARIMA. Test were done using the parameters defined in table 2. We can see that OVECM has very similar performance than SLVECM and this support the theory that cointegration vectors vary little in time. Moreover, OVECM also out performed SLARIMA based on these three performance measures.

We can also notice that in-sample performance in OVECM and SLVECM is related with the out-of-sample performance. This differs with SLARIMA which models with

TABLE 4. Model measures

Model				In-sample			Out-of-sample		
Method	L	Parameters	e	MAPE	MAE	RMSE	MAPE	MAE	RMSE
OVECM	100	P=2	0.0026	0.00263	0.00085	0.00114	0.00309	0.00094	0.00131
OVECM	400	P=5	0.0041	0.00378	0.00095	0.00127	0.00419	0.00103	0.00143
OVECM	700	P=3	0.0032	0.00323	0.00099	0.00130	0.00322	0.00097	0.00132
OVECM	1000	P=3	0.0022	**0.00175**	**0.00062**	**0.00087**	**0.00172**	**0.00061**	**0.00090**
SLVECM	100	P=2	-	0.00262	0.00085	0.00113	0.00310	0.00095	0.00132
SLVECM	400	P=5	-	0.00375	0.00095	0.00126	0.00419	0.00103	0.00143
SLVECM	700	P=3	-	0.00324	0.00099	0.00130	0.00322	0.00098	0.00132
SLVECM	1000	P=3	-	**0.00174**	**0.00061**	**0.00087**	**0.00172**	**0.00061**	**0.00090**
SLARIMA	100	p,d,q=2,1,1	-	0.00285	0.00110	0.00308	0.00312	0.00098	0.00144
SLARIMA	400	p,d,q=1,1,1	-	0.00377	0.00101	0.00128	0.00418	0.00106	0.00145
SLARIMA	700	p,d,q=2,1,1	-	0.00329	0.00102	0.00136	0.00324	0.00097	0.00133
SLARIMA	1000	p,d,q=2,1,1	-	**0.00281**	**0.00074**	**0.00105**	**0.00177**	**0.00063**	**0.00092**

good in-sample performance are not necessarily good out-of-sample models. Moreover OVECM outperformed SLARIMA using the same window size.

Figure 5.2 shows the out-of-sample forecasts made by our proposal OVECM with the best parameters found based on table 4 which follows the time series very well.

5.4. Conclusions

A new online vector error correction method was presented. We have shown that our proposed OVECM considerably reduces execution times without compromising solution accuracy. OVECM was compared with VECM and ARIMA with the same sliding window sizes and OVECM outperformed both in terms of execution time. Traditional VECM slightly outperformed our proposal but the OVECM execution time is lower. This reduction of execution time is mainly because OVECM avoids the cointegration vector calculation using the Johansen method. The condition for getting new vectors is given by the mean_error variable which controls how many times the Johansen method is called. Additionally, OVECM introduces matrix optimisation in order to get the new model in an iterative way. We could see that our algorithm took much less than a minute at every step. This means that it could also be used with higher frequency data and would still provide responses before new data arrives.

For future study, it would be interesting to improve the out-of-sample forecast by considering more explicative variables, to increase window sizes or trying new conditions to obtain new cointegration vectors.

Since OVECM is an online algorithm which optimises processing time, it could be used by investors as an input for strategy robots. Moreover, some technical analysis methods could be based on its output.

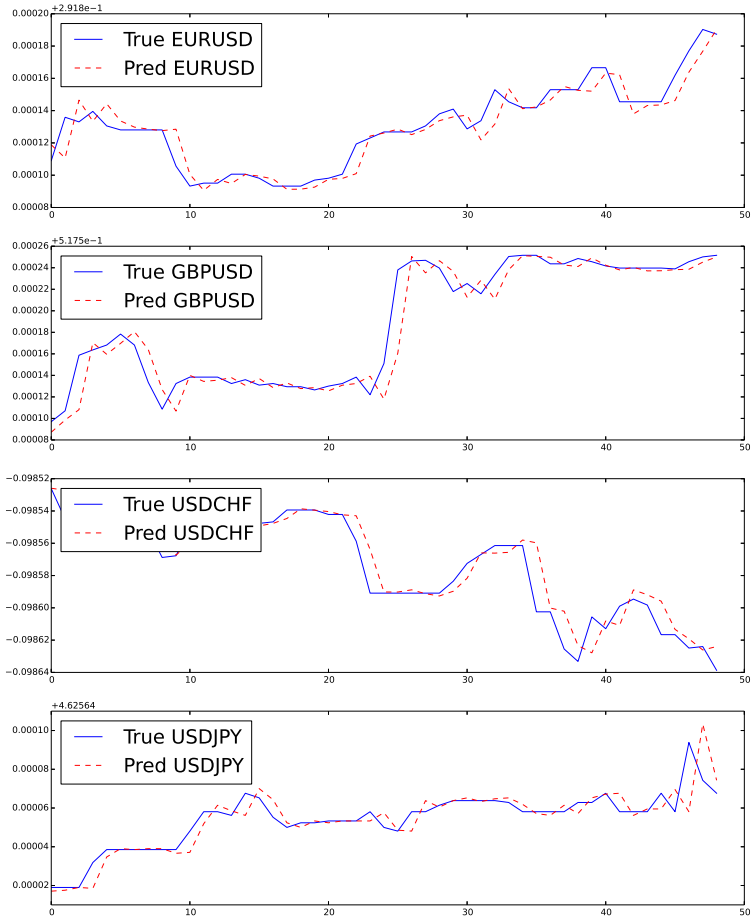


FIGURE 5.2. OVECM forecasting accuracy example for 50 minutes using $L = 1000$ and $p = 3$

Chapter 6

Conclusions and Future Work

Main results and contributions are presented in this chapter along with some suggested future work.

6.1. Conclusions of this thesis

This thesis is a multidisciplinary work that involves knowledge of Finance and Economics, Time Series, Machine Learning, Parallel Computing and Scientific Computing. We specifically addressed the joint dynamic behaviour of financial time series that are said to be cointegrated. A broadly used approach, called VECM (Vector Error Correction Model), characterises this behaviour in terms of forces pulling towards equilibrium called cointegration. However, the use of VECM has been limited to low frequency time series processed in batch mode. In this thesis we explored the use of VECM with high frequency data and we found that the main limitation was computational. The focus of our work was developing a model to improve the performance forecasting of financial time series maintaining good execution times in order to use it with high frequency data. VECM parameter estimation uses two computationally expensive routines: the Johansen method, to obtain cointegration vectors, and the ordinary least squares method to solve the system. In this thesis, different ways to explore cointegration in high frequency time series were proposed, always considering the computational limitations of the VECM.

The study of cointegration in high frequency data was done using two different approaches presented in Chapters 4 and 5, called AVECM and OVECM:

AVECM: AVECM is an adaptive version of VECM which includes a new method to choose VECM parameters based on the maximisation of the percentage of

cointegration. AVECM was implemented using MPI to search on a grid of possible values. We evaluated our results in terms of performance and execution times:

Performance: Results showed that AVECM improves performance measures by finding parameters of L and p maximising the percentage of cointegration. AVECM performance was compared against ARIMA and the random walk model. We showed that the out-of-sample performance for AVECM is superior to ARIMA and the naive random walk model in terms of the MSE and U-statistic. This result is very important since it is still difficult to outperform ARIMA and the random walk model for standard econometric forecasting models despite their simplicity.

Execution times: execution times were reduced more than 9 times ensuring a response time before the processing of the next data point. Despite the fact that high frequency Forex data can be spurious, the model performance can be less reliable (and more spurious) relative to the lower frequencies (such as 1 minute or 5 minute intervals) adopted by some other studies. However, the deficiency is offset by gain in accuracy from parallel processing which is capable of searching or examining a much larger state space given the same computational time.

OVECM: We proposed an online version of VECM (OVECM). OVECM optimises how model parameters are obtained using a sliding window of the most recent data. OVECM, unlike AVECM, updates the parameters at each step, instead of obtaining new ones. This proposal takes advantage of the long-run relationship between the time series in order to obtain improved execution times. OVECM also introduces matrix optimisation in order to obtain the new model in an iterative way.

Performance: OVECM was compared with the traditional VECM and ARIMA. Despite the fact that traditional VECM slightly outperformed our proposal, the OVECM execution time is lower. This result can be explained because OVECM avoids the calculation of cointegration vectors at each time step which improves execution times but can affect forecast performance.

Execution times: OVECM was compared with the traditional VECM and ARIMA with the same sliding window sizes. As a result, OVECM outperformed both in terms of execution time. This reduction of execution time is mainly because OVECM avoids the cointegration vector calculation using the Johansen method. Aditionally, OVECM took much less than a minute at every step making it possible to use with higher frequency data.

Finally, cointegration information can now easily be used as an integration tool to detect arbitrage opportunities or risk control in financial time series.

6.2. Contributions of this thesis

This research helped to integrate machine learning techniques and parallel computing with econometric models such as VECM so they can be used with high frequency data. We also observed that cointegration relationships are sensitive to the choice of the amount of data and VECM parameters and this can be used to improve forecasting performance. We proposed two variations of VECM, OVECM and AVECM which could be extended to other econometric models.

Regarding the initial objectives defined at the beginning of this thesis we can say:

- $\mathcal{O}_1$: *A review of the literature on time series analysis models including machine learning techniques.*
  The literature review was focused in finance, time series concepts and models and machine learning including theory and applications.
- $\mathcal{O}_2$: *Development of a set of known features of the studied time series and the application to improve forecasting.*
  We included specific knowledge about forex rates, which were those used in our experiments, such as trade best trading times, stylised facts, percentage of cointegration, time series frequency, among others.
- $\mathcal{O}_3$: *Development of parallel and efficient algorithms to ensure quick response times .*
  We showed that high performance computing will allow the use of computationally expensive methods to forecast high frequency data ensuring quick response times.
- $\mathcal{O}_4$: *Deep mathematical analysis of the proposal and financial concepts involved.*
  The two proposals were presented with a strong mathematical foundation, we added definitions and proofs of all the concepts involved.
- $\mathcal{O}_5$: *Design and implement representative set of experiments in order to show when and why the proposal performs better.* All the experiments were designed to give a general view of the method performance. We clearly showed the use and limitations of our proposals.

6.3. Future Work

For future study, it would be interesting to explore the relationship between cointegration and performance for more assets, including not only forex rates but also stocks. It would also be interesting to include more explaining variables such as bid-ask spread and change in volume.

The online approach for other econometrics models it will be also worthy of study. Many of them could be adapted and used with higher frequency data. In this thesis, the online version of OLS was studied, but the online version of ridge regression could also be applied to obtain a solution with better generalisation capabilities.

Finally, it will be also interesting to improve the out-of-sample forecast by considering more explicative variables, to increase window sizes or try different conditions to obtain new cointegration vectors.

Appendix A

Proofs

Some useful proofs for the understanding of this thesis.

A.1. Pseudo-inverse computed using the compact SVD

Proof

Since $\mathbf{A}$ is singular the problem shown in equation (2.28) has not solution, the minimum norm given by equation (3.13) is obtained by solving the equivalent problem:

$$\mathbf{A}\hat{\mathbf{X}} = \mathbf{PB}$$

where $\mathbf{P} = \mathbf{U_1}\mathbf{U_1}^\top$ is the projection onto the Col($\mathbf{A}$).

Since $\mathbf{V} = \underset{(n\times k)}{[\ \mathbf{V_1}\ }\mid \underset{(n\times k)}{\mathbf{V_2}\ ]}$ and $\mathbf{V_1}^\top\mathbf{V_2} = \mathbf{0}$ we can express $\hat{\mathbf{X}} = \mathbf{V_1}\mathbf{x_1} + \mathbf{V_2}\mathbf{x_2}$ with $\mathbf{x_2} = \mathbf{0}$ because $\hat{\mathbf{X}}$ lives in the Row($\mathbf{A}$) given by $\mathbf{V_1}$, so we have:

$$\begin{aligned}
\mathbf{A}\hat{\mathbf{X}} &= \mathbf{PB} \\
\mathbf{U_1}\mathbf{\Sigma_1}\mathbf{V_1}^\top\hat{\mathbf{X}} &= \mathbf{U_1}\mathbf{U_1}^\top\mathbf{B} \\
\mathbf{V_1}^\top\hat{\mathbf{X}} &= \mathbf{\Sigma_1^{-1}}\mathbf{U_1}^\top\mathbf{B} \\
\mathbf{V_1}^\top\mathbf{V_1}\mathbf{x_1} &= \mathbf{\Sigma_1^{-1}}\mathbf{U_1}^\top\mathbf{B} \\
\mathbf{x_1} &= \mathbf{\Sigma_1^{-1}}\mathbf{U_1}^\top\mathbf{B}
\end{aligned}$$

from this result we can obtain $\hat{\mathbf{X}}$ and therefore the pseudo-inverse expression:

$$\begin{aligned}\hat{\mathbf{X}} &= \mathbf{V_1 x_1} \\ &= \mathbf{V_1 \Sigma_1^{-1} U_1^\top B} \\ \mathbf{A}^+ &= \mathbf{V_1 \Sigma_1^{-1} U_1^\top}.\end{aligned}$$

■

A.2. The pseudo-inverse computed using the compact singular value decomposition (SVD)

$$\mathbf{A}^+ = \mathbf{V_1 \Sigma_1^{-1} U_1^\top} \tag{A.1}$$

Proof

In the case matrix $\mathbf{A}$ is non singular, its solution is straight forward:

$$\mathbf{X} = \mathbf{A}^{-1}\mathbf{Y} \tag{A.2}$$

Since the problem shown in equation (3.12) has not solution, the minimum norm given by equation (A.2) is obtained by solving the equivalent problem:

$$\mathbf{A}\hat{\mathbf{X}} = \mathbf{PY}$$

where $\mathbf{P} = \mathbf{U_1 U_1^\top}$ is the projection onto the Col($\mathbf{A}$).

Since $\mathbf{V} = \underset{(n\times k)}{[\ \mathbf{V_1}\ |} \underset{(n\times k)}{\mathbf{V_2}\ ]}$ and $\mathbf{V_1^\top V_2} = \mathbf{0}$ we can express $\hat{\mathbf{X}} = \mathbf{V_1 X_1} + \mathbf{V_2 X_2}$ with $\mathbf{X_2} = \mathbf{0}$ because $\hat{\mathbf{X}}$ lives in the Row($\mathbf{A}$) given by $\mathbf{V_1}$, so we have:

$$\begin{aligned}\mathbf{A}\hat{\mathbf{X}} &= \mathbf{PY} \\ \mathbf{U_1 \Sigma_1 V_1^\top}\hat{\mathbf{X}} &= \mathbf{U_1 U_1^\top Y} \\ \mathbf{V_1^\top}\hat{\mathbf{X}} &= \mathbf{\Sigma_1^{-1} U_1^\top Y} \\ \mathbf{V_1^\top V_1 X_1} &= \mathbf{\Sigma_1^{-1} U_1^\top Y} \\ \mathbf{X_1} &= \mathbf{\Sigma_1^{-1} U_1^\top Y}\end{aligned}$$

from this result we can obtain $\hat{\mathbf{X}}$ and therefore the pseudo-inverse expression:

$$\begin{aligned}\hat{\mathbf{X}} &= \mathbf{V_1 X_1} \\ &= \mathbf{V_1 \Sigma_1^{-1} U_1^\top Y} \\ \mathbf{A}^+ &= \mathbf{V_1 \Sigma_1^{-1} U_1^\top}.\end{aligned}$$

■

A.3. Ridge regression optimal solution

$$\mathbf{X}_* = (\mathbf{A}^\top \mathbf{A} + \lambda \mathbb{I})^{-1} \mathbf{A}^\top y\,,$$

Proof

$$\begin{aligned} J(\mathbf{X}) &= \|\mathbf{A}\mathbf{X} - \mathbf{Y}\|_2^2 + \lambda \|\mathbf{X}\|^2 \\ &= \sum_{t=1}^{N} (\mathbf{X}^\top \mathbf{a}_t - y_t)^2 + \lambda \sum_{i=1}^{p} \mathbf{X}_i^2 \\ &= (\mathbf{X}^\top a_1 - y_1)^2 + \cdots + (\mathbf{X}^\top a_N - y_N)^2 + \lambda(\mathbf{X}_1^2 + \cdots + \mathbf{X}_p^2) \end{aligned}$$

taking derivatives

$$\begin{aligned} \frac{\partial J(\mathbf{X})}{\partial \mathbf{X}_1} &= 2(\mathbf{X}^\top \mathbf{a}_1 - y_1)\mathbf{a}_{11} + \cdots + 2(\mathbf{X}^\top \mathbf{a}_N - y_N)\mathbf{a}_{N1} + 2\lambda \mathbf{X}_1 \\ &= 2\mathbf{a}_1^\top (\mathbf{A}\mathbf{X} - \mathbf{Y}) + 2\lambda \mathbf{X}_1 \\ &\vdots \\ \frac{\partial J(\mathbf{X})}{\partial \mathbf{X}_p} &= 2\mathbf{a}_p^\top (\mathbf{A}\mathbf{X} - \mathbf{Y}) + 2\lambda \mathbf{X}_p \end{aligned}$$

Then we have that:

$$\frac{\partial J(\mathbf{X})}{\partial \mathbf{X}} = 2\mathbf{A}^\top (\mathbf{A}\mathbf{X} - \mathbf{Y}) + 2\lambda \mathbf{X}$$

Since $\frac{\partial J(\mathbf{X})}{\partial \mathbf{X}} = 0$ we have:

$$\begin{aligned} 2\mathbf{A}^\top (\mathbf{A}\mathbf{X} - \mathbf{Y}) + 2\lambda \mathbf{X} &= 0 \\ \mathbf{A}^\top \mathbf{A}\mathbf{X} - \mathbf{A}^\top \mathbf{Y} + \lambda \mathbf{X} &= 0 \\ (\mathbf{A}^\top \mathbf{A} + \lambda \mathbb{I})\mathbf{X} &= \mathbf{A}^\top \mathbf{Y} \\ \mathbf{X} &= (\mathbf{A}^\top \mathbf{A} + \lambda \mathbb{I})^{-1} \mathbf{A}^\top \mathbf{Y} \end{aligned}$$

■

A.4. Bias and Variance

The OLS bias can be obtained as:

$$\begin{aligned} Bias(\hat{f}(\hat{\mathbf{X}})) &= E[\hat{\mathbf{X}}] - \mathbf{X} \\ &= E[(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{Y}] - \mathbf{X} \\ &= E[(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top(\mathbf{A}\mathbf{X})] - \mathbf{X} \\ &= \mathbf{X} - \mathbf{X} \\ &= 0 \end{aligned}$$

The Ridge Regression bias can be obtained as:

$$\begin{aligned} \mathbf{X}(\lambda) &= (\mathbf{A}^\top\mathbf{A} + \lambda\mathbb{I})^{-1}\mathbf{A}^\top\mathbf{Y} \\ &= (\mathbb{I} + \lambda(\mathbf{A}^\top\mathbf{A})^{-1})^{-1}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{Y} \\ &= (\mathbb{I} + \lambda(\mathbf{A}^\top\mathbf{A})^{-1})^{-1}\hat{\mathbf{X}} \\ &= \mathbf{W}\hat{\mathbf{X}} \end{aligned}$$

where $\mathbf{W} = (\mathbb{I} + \lambda(\mathbf{A}^\top\mathbf{A})^{-1})^{-1}$ it is defined for simplicity. Ridge regression bias is then obtained as:

$$\begin{aligned} Bias(\mathbf{X}(\lambda)) &= E[\mathbf{X}(\lambda)] - \mathbf{X} \\ &= E[\mathbf{W}\hat{\mathbf{X}}] - \mathbf{X} \\ &= \mathbf{W}\mathbf{X} - \mathbf{X} \neq 0 \end{aligned}$$

The variance of OLS is:

$$Var(\hat{\mathbf{X}}) = \sigma^2(\mathbf{A}^\top\mathbf{A})^{-1}$$

and the variance of ridge regression is:

$$\begin{aligned} Var(\mathbf{X}(\lambda)) &= Var(\mathbf{W}\hat{\mathbf{X}}) \\ &= E[(\mathbf{W}\hat{\mathbf{X}} - E[\mathbf{W}\hat{\mathbf{X}}])(\mathbf{W}\hat{\mathbf{X}} - E[\mathbf{W}\hat{\mathbf{X}}])^\top] \\ &= \mathbf{W}E[(\hat{\mathbf{X}} - E[\hat{\mathbf{X}}])(\hat{\mathbf{X}} - E[\hat{\mathbf{X}}])^\top]\mathbf{W}^\top \\ &= \mathbf{W}Var(\hat{\mathbf{X}})\mathbf{W}^\top \\ &= \sigma^2\mathbf{W}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{W}^\top \end{aligned}$$

A.5. Ridge regression shows an increasing squared bias and a decreasing variance

Proof
Since $Bias(\mathbf{X}(\lambda)) \neq 0$ this imply that

$$Bias(\mathbf{X}(\lambda))^2 > 0$$

we know that $\mathbf{A}^\top\mathbf{A}$ has an eigenvalue decomposition $\mathbf{A}^\top\mathbf{A} = \mathbf{V}\Sigma\mathbf{V}^{-1}$.

$$\begin{aligned}
Var(\mathbf{X}(\lambda)) &= \sigma^2\mathbf{W}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{W}^\top \\
Var(\mathbf{X}(\lambda)) &= \sigma^2(\mathbb{I}+\lambda(\mathbf{A}^\top\mathbf{A})^{-1})^{-1}(\mathbf{A}^\top\mathbf{A})^{-1}((\mathbb{I}+\lambda(\mathbf{A}^\top\mathbf{A})^{-1})^{-1})^\top \\
&= \sigma^2(\mathbb{I}+\lambda(\mathbf{V}\Sigma\mathbf{V}^{-1})^{-1}(\mathbf{V}\Sigma\mathbf{V}^{-1})^{-1}((\mathbb{I}+\lambda(\mathbf{V}\Sigma\mathbf{V}^{-1})^{-1})^{-1})^\top \\
&= \sigma^2\mathbf{V}(\mathbb{I}+\lambda\Sigma^{-1})^{-1}\Sigma^{-1}(\mathbb{I}+\lambda\Sigma^{-1})^{-1}\mathbf{V}^{-1} \\
&= \sigma^2\mathbf{V}\mathbf{D}^{-1}\mathbf{V}^{-1}
\end{aligned}$$

where $\mathbf{D}^{-1} = (\mathbb{I}+\lambda\Sigma^{-1})^{-1}\Sigma^{-1}(\mathbb{I}+\lambda\Sigma^{-1})^{-1}$ is a diagonal matrix with coefficients:

$$d_i = \frac{\sigma_i^{-1}}{(1+\lambda\sigma_i^{-1})^2}$$

■

A.6. Efficient computation

In regression problems we always require getting a matrix inverse which is computationally expensive. However, in stream data problems there is a way to obtain a matrix inverse approximation using the Sherman-Morrison-Woodbury. This formula allows to get the inverse of a matrix $\mathbf{A}+\mathbf{uv}^\top$ if we previously calculated the inverse of $\mathbf{A}$ as follows:

$$(\mathbf{A}+\mathbf{uv}^\top)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A^{-1}uv}^\top\mathbf{A^{-1}}}{\mathbf{1}+\mathbf{v}^\top\mathbf{A^{-1}u}} \tag{A.3}$$

Alternatively, Coleman and Sun [**?**] presented an iterative algorithm which uses $\mathbf{X}(\lambda)$ to approximate $\mathbf{X}(0)$.

Using the compact SVD (shown in equation (3.15)) $\mathbf{X}(\lambda)$ is expressed as follows:

$$\begin{aligned}
\mathbf{X}(\lambda) &= (\mathbf{A}^\top\mathbf{A}+\lambda\mathbb{I})^{-1}\mathbf{A}^\top\mathbf{Y} \\
&= \mathbf{V}_1(\Sigma_1^2+\lambda\mathbb{I})^{-1}\mathbf{\Sigma_1}\mathbf{U_1^\top}\mathbf{Y} \qquad \text{(A.4)}
\end{aligned}$$

where is easy to see that $\mathbf{X}(\lambda) \to \hat{\mathbf{X}} = \mathbf{V}_1\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^\top\mathbf{Y}$ as $\lambda \to 0$.

The method consists in obtaining $\mathbf{X}(\lambda)$ and then refine by adding more terms of its Taylor expansion to approximate $\mathbf{X}(0) = \hat{\mathbf{X}}$. The Taylor expansion about λ_0 is:

$$\mathbf{X}(\lambda) = \mathbf{X}(\lambda_0) + \sum_{k=1}^{\infty}\mathbf{s}_k(\lambda-\lambda_0)^k \tag{A.5}$$

where $\mathbf{s}_k = \frac{1}{k!}\mathbf{X}(\lambda)^{(k)}$ and $\mathbf{X}(\lambda)^{(k)}$ is obtained by taking differences $\frac{\partial}{\partial\lambda}$ of:

$$\begin{aligned}
(\mathbf{A}^\top\mathbf{A} + \lambda\mathbb{I})\mathbf{X}(\lambda) &= \mathbf{A}^\top\mathbf{Y} \\
(\mathbf{A}^\top\mathbf{A} + \lambda\mathbb{I})\mathbf{X}(\lambda)^{(1)} + \mathbf{X}(\lambda) &= 0 \\
\mathbf{X}(\lambda)^{(1)} &= -(\mathbf{A}^\top\mathbf{A} + \lambda\mathbb{I})^{-1}\mathbf{X}(\lambda) \\
\mathbf{X}(\lambda)^{(2)} &= -2(\mathbf{A}^\top\mathbf{A} + \lambda\mathbb{I})^{-1}\mathbf{X}(\lambda)^{(1)} \\
&\vdots \\
\mathbf{X}(\lambda)^{(k)} &= -k(\mathbf{A}^\top\mathbf{A} + \lambda\mathbb{I})^{-1}\mathbf{X}(\lambda)^{(k-1)}
\end{aligned}$$

for $\lambda = 0$ we have:

$$\mathbf{X}(0) = \mathbf{X}(\lambda_0) + \sum_{k=1}^{\infty}(-1)^k\mathbf{s}_k\lambda_0^k \tag{A.6}$$

therefore in order to ensure convergence, we can see that λ_0 cannot be large.

The algorithm for computing $\mathbf{X}(0)$ is the following:

Algorithm 8 Algorithm for handling rank deficient matrices

Require:

$\mathbf{A}$: design matrix

$\mathbf{Y}$: response matrix

λ: rank deficient parameter

Ensure:

$\mathbf{X}$: parameters

1: $\mathbf{M} = \mathbf{A}^\top\mathbf{A}$
2: Initialize $\mathbf{QR} = \mathbf{M}$
3: $\mathbf{X} = \mathbf{R}^{-1}\mathbf{Q}^\top\mathbf{Y}$
4: $\mathbf{s} = \mathbf{X}$
5: **for** $i = 1, 2, 3, \ldots$ **do**
6: $\quad\mathbf{s} = -(\mathbf{M} + \lambda\mathbb{I})^{-1}\mathbf{s}$
7: $\quad\mathbf{X} = \mathbf{X} + (-1)^i\mathbf{s}\lambda^i$
8: **end for**

The algorithm 8 solves equation (A.6). However, this version is unstable since typically $\|\mathbf{s}\|$ is very large and λ^i is very small (λ is small).

The following algorithm shows a more stable version of algorithm 8.

Algorithm 9 Algorithm for handling rank deficient matrices improved

Require:

$\mathbf{A}$: design matrix

$\mathbf{Y}$: response matrix

λ: rank deficient parameter

Ensure:

$\mathbf{X}$: parameters

1: $\mathbf{M} = \mathbf{A}^\top \mathbf{A}$
2: Initialize $\mathbf{QR} = \mathbf{M}$
3: $\mathbf{X} = \mathbf{R}^{-1}\mathbf{Q}^\top \mathbf{Y}$
4: $\mathbf{t} = \mathbf{X}$
5: **for** $i = 1, 2, 3, \ldots$ **do**
6: $\quad \mathbf{t} = \lambda \mathbf{t}$
7: $\quad \mathbf{t} = -(\mathbf{M} + \lambda \mathbb{I})^{-1}\mathbf{t}$
8: $\quad \mathbf{X} = \mathbf{X} + \mathbf{t}$
9: **end for**

Both algorithms are equivalent, but algorithm 9 is more stable and converges in typically less than 10 steps. It is important to notice that the QR factorisation is computed only once and it is computationally less expensive than the SVD.