Portfolio construction under information asymmetry

THÈSE N^{O} 5606 (2013)

PRÉSENTÉE LE 11 JANVIER 2013

À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS

LABORATOIRE DE COMMUNICATIONS AUDIOVISUELLES

PROGRAMME DOCTORAL EN INFORMATIQUE, COMMUNICATIONS ET INFORMATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Lionel COULOT

acceptée sur proposition du jury:

Prof. R. Urbanke, président du jury

Prof. M. Vetterli, Prof. P. Bossaerts, directeurs de thèse

Prof. P. Embrechts, rapporteur

Prof. L. Mancini, rapporteur

Dr R. Silvotti, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse

2013

A mes parents
A Martial

Abstract

We introduce in this thesis the idea of a variable lookback model, i.e., a model whose predictions are based on a variable portion of the information set. We verify the intuition of this model in the context of experimental finance. We also propose a novel algorithm to estimate it, the variable lookback algorithm, and apply the latter to build investment strategies.

Financial markets under information asymmetry are characterized by the presence of better-informed investors, also called insiders. The literature in finance has so far concentrated on theoretical models describing such markets, in particular on the role played by the price in conveying information from informed to uninformed investors. However, the implications of these theories have not yet been incorporated into processing methods to extract information from past prices and this is the aim of this thesis. More specifically, the presence of a time-varying number of insiders induces a time-varying predictability in the price process, which calls for models that use a variable lookback window. Moreover, although our initial motivation comes from the study of markets under information asymmetry, the problem is more general, as it touches several issues in statistical modeling. The first one concerns the structure of the model. Existing methods use a fixed model structure despite evidences from data, which support an adaptive one. The second one concerns the improper handling of the nonstationarity in data. The stationarity assumption facilitates the mathematical treatment. Hence, existing methods relies on some form of stationarity, for example, by assuming local stationary, as in the windowing approach, or by modeling the underlying switching process, for example, with a Markov chain of order 1. However, these suffer from certain limitations and more advanced methods that take explicitly into account the nonstationariry of the signal are desirable. In summary, there is a need to develop a method that constantly monitors what is the appropriate structure, when a certain model works and when not or when are the underlying assumptions of the model violated.

We verify our initial intuition in the context of experimental finance. In particular, we highlight the diffusion of information in the market. We give a precise definition to the notion of the time of maximally informative price and verify, in line with existing theories, that the time of maximally informative price is inversely proportional to the number of insiders in the market. This supports the idea of a variable lookback model. Then, we develop an estimation algorithm that selects simultaneously the order of the process and the lookback window based on the minimum description length principle. The algorithm maintains a series of estimators, each based on a different order and/or information set. The selection is based on an information theoretic criterion, that ac-

counts for the ability of the model to fit the data, penalized by the model complexity and the amount of switching between models. Finally, we put the algorithm at work and build investment strategies. We devise a method to draw dynamically the trend line for the time-series of log-prices and propose an adaptive version of the well-known momentum strategy. The latter outperforms standard benchmarks, in particular during the 2009 momentum crash. **Key words: nonstationary signal processing, minimum description length, financial modeling, adaptive momentum.**

Résumé

Cette thèse introduit l'idée d'un modèle à horizon variable, c.à.d. un modèle dont les prédictions sont basées sur une portion variable de l'ensemble d'information. Nous vérifions l'intuition qui sous-tend ce modèle dans le contexte de la finance expérimentale. Nous proposons également un nouvel algorithme, l'algorithme à horizon variable, et appliquons ce dernier pour construire des stratégies d'investissement.

Les marchés financiers avec asymétrie de l'information se caractérisent par la présence d'investisseurs mieux informés, aussi appelés initiés. La littérature en finance s'est pour le moment concentrée sur des modèles théoriques décrivant ces marchés, en particulier sur le rôle joué par le prix pour transmettre l'information des investisseurs informés aux non informés. Cependant, les implications de ces théories n'ont à ce jour pas été incorporées dans des méthodes de traitement qui extraient l'information des prix passés et c'est le but de cette thèse. Plus spécifiquement, la présence d'un nombre variable d'initiés induit une variabilité dans la prédictibilité du processus de prix tel qu'elle requiert des modèles à horizon variable. De plus, bien que notre motivation initiale vienne de l'étude des marchés avec asymétrie de l'information, le problème est plus général, car il touche à plusieurs problèmes en modélisation statistique. Le premier concerne la structure du modèle. Les méthodes existantes utilisent une structure fixe en dépit des preuves tirées des données qui appuient l'hypothèse d'une structure adaptable. Le deuxième concerne le traitement inapproprié de la non stationnarité. L'hypothèse de stationnarité facilite le traitement mathématique. Ainsi, les méthodes existantes sont fondées sur une forme ou autre de stationnarité, par exemple, la stationnarité locale dans le traitement par fenêtre ou la modélisation du processus de saut, ex. avec un processus de Markov d'ordre 1. Cependant, ces dernières approches souffrent de certaines limites et des méthodes plus avancées, qui tiennent compte explicitement de la non stationnarité du signal, sont souhaitables. En résumé, il est nécessaire de développer des méthodes qui mesurent en permanence quelle est la structure la plus appropriée, quand un modèle fonctionne et quand il ne fonctionne plus et quand les hypothèses qui sous-tendent le modèle sont violées.

Nous vérifions notre intuition dans le contexte de la finance expérimentale. En particulier, nous mettons en évidence le processus de diffusion de l'information dans le marché. Nous donnons une définition précise de la notion de temps du prix maximalement informatif et vérifions, en accord avec les théories existantes, que le temps du prix maximalement informatif est inversement proportionnel au nombre d'initiés dans le marché. Ce résultat étaie l'idée d'un modèle à horizon variable. Nous développons ensuite un algorithme d'estimation basé sur le principe de description de longueur minimale. L'algorithme calcule une série d'estimateurs, chacun basé sur un

ordre et/ou un ensemble d'information différent. La sélection s'opère selon un critère tiré de la théorie de l'information, qui tient compte de la capacité du modèle à représenter les données, pénalisée par la complexité du modèle et le nombre de sauts entre modèles. Finalement, nous appliquons l'algorithme pour construire des stratégies d'investissement. Nous concevons une méthode pour dessiner dynamiquement la ligne de tendance pour la série temporelle du logarithme du prix et proposons une version adaptable de la célèbre stratégie du moment. Notre stratégie surperforme par rapport aux stratégies de référence standards, en particulier durant le krach de la stratégie du moment en 2009. **Mots-clés : traitement du signal non stationnaire, description de longueur minimale, modélisation financière, stratégie du moment adaptable.**

Acknowledgments

This research was supported by an industrial grant agreement from LGT Group through the LGT & Science program and this is hereby acknowledged here. Moreover all datasets used in this thesis were obtained from LGT/Bloomberg, unless otherwise specified. Special thanks to Gabriel Clerc from the industry liaison office at EPFL for his support in setting up the collaboration contract.

My deepest gratitude goes to my thesis director Prof. Martin Vetterli. Having the chance to work with you was certainly one of my biggest motivations to start a PhD at EPFL. You have been the best advisor I could have hoped for, constantly excited about research, including more mundane finance related topics, and extremely pleasant to work with. Furthermore, I am still amazed how, despite your various assignments, you have managed to find time for our regular meetings. I have really appreciated your guiding style; you have more accompanied this travel by suggesting paths rather than coerced me in any particular directions. Your guidance has also extended far beyond your thesis director's assignment, to encompass important life decisions and career choices and this with my best interest at heart. The best lesson I have learned from you, since you impersonate it, is that one must find his passion in life and turn it into one's daily occupation. This is exactly what I am going to do next. I could not be more thankful for all you have brought me. My gratitude also goes to Prof. Peter Bossaerts, my thesis co-director. We would not have made it without you, your extensive knowledge combining various fields and your genius intuition. Moreover, the index you have built in your brain, which spans more than 50 years of finance research was useful on many occasions. I also would like to thank the members of my thesis committee, in alphabetical order, Prof. Paul Embrechts, Prof. Loriano Mancini and Dr. Roberto Silvotti. You have had the "pleasure" to read my thesis and I should already be extremely grateful just for that. Thanks also for all your comments and feedbacks on the preliminary version of this work, which helped shape its final version.

I would also like to acknowledge the support of my fellow students and LGT colleagues, who contributed to this work. First and foremost, Dr. Marie-Christine Mikl without whom I would not have survived in the sharks tank at LGT. Your advices ideally complemented those of Martin, especially when it came expectation management and company politics. Over time, you have become more than just a colleague, while maintaining the ability to be serious when required. What would have been life at LGT without (a) you reminding me on a daily basis I am balding, skinny (I have the lowest BMI of the whole LGT group) and Swiss Romand (b) your Austro-British sense of humor (c) us showing up in the office with matching outfits (except for

leopard ballerinas) (d) us discussing endlessly on do's and don't in fashion and bitching about the revolting shoes of our colleagues. In brief, you turned my stay at LGT into something considerably more enjoyable. Next comes Dr. Magnus Pirovino, the father of LGT & Science, who made this collaboration possible. Various colleagues also contributed significantly with their feedbacks and comments, in particular, Alexander Zanker, Walter Pfaff, Hans-Peter Oehri, Dr. Matthias Feiler, Stefan Mühlemann, Sudheer Arora and Francesco Valenti. Other colleagues made my lunch in Pfäffikon and train ride more enjoyable, in particular and without particular order, Mario Dal Col, Susan Liu, Michal Dzielinski, Dr. Sven Steude, Monica Maranta, Andrea Ferch, Didier Noverraz and Martin Keller. At EPFL, I would like to thank my fellow LCAV members, particularly Dr. Amina Chebira for her advices and all the good moments of laugh, as well as Andreas Walther, Yann Barbotin, Loïc Baboulaz and Ali Homati for the interesting discussions. Finally, Jacqueline Aeberhard deserves all my consideration for skillfully managing the LCAV's organization and Martin's agenda.

Other people also contributed albeit in a different manner to this thesis. First of all my successive roommates Dinu Niculescu, Dominique Wahl, Ileana Popp and Alexander Bernhart, who had to endure living with me and sometimes had to support the complaints of a PhD student over dinner. My friends, who ensured I stay sane. In first line, Martial de Montmollin, who has always been there for me and made me rediscover the joy of gardening, and Diane Rosier for our endless phone conversations. And of course the following persons, who shall recognize themselves: Miss Béatrice Fraport (or the miracle of Lady Gaga after a night out at HB lamenting about our socially inept bosses), Rudi (Art in town), Bo Rockhard ("Martin Clunes..."), Tiffany Swallow ("One can never be to rich or too thin."), my secret African mistress, Gonzuela Teleñovelas, Posh Bitch, Principessa, Francois and Jonathan. Finally, I would like to thank my family for their love and unconditional support: my sister, who made sure nothing was missing in my apartment in Zürich, my brother who managed to finish his Master before I my PhD. Last but not least my parents. Merci d'avoir toujours été là pour moi, de m'avoir laissé la liberté de choisir et de respecter mes choix. Un grand merci pour tous vos encouragements sans lesquels cette thèse n'aurait jamais été terminée. Cette thèse vous est dediée.

Contents

List of Figures

List of Notations

We adopt the following notational conventions. Indices are denoted in subscript and if i is the canonical index, I represents its maximum value. The superscript $x^{(t)}$ highlights the time dependency of variable x. Vectors $\boldsymbol{x}$ (resp. matrices $\boldsymbol{X}$) are denoted by bold lower (resp. upper) case. Vectors and matrices are typically formed by stacking observations, e.g., $\boldsymbol{x} = (x_1, \ldots, x_i, \ldots, x_I)^T$. $\boldsymbol{x}^T$ denotes the transpose of vector $\boldsymbol{x}$ and it should not be confused with T, the number of observations. $\boldsymbol{x}^{(t_1)>(t_2)}$ represents the concatenation in a sequence of observations of variable x from time t_1 to t_2, $\left(x^{(t_1)}, x^{(t_1+1)}, \ldots, x^{(t_2-1)}, x^{(t_2)}\right)$. The following list contains only notations that appear in different parts of this thesis. A variable is locally adapted by changing its subscript, but it consistently represents the same quantity. For example.

$$p_{i,j}^{(t)}$$

denotes the price at time t in trading session i, which has j insiders, in Section 3.2.2 and

$$p_n^{(t)}$$

the price of stock n at time t in Section 2.1.2.

$-$	Empty sequence
CL	Code length
CL_{past}	Code length of encoding symbols of past periods of piecewise stationarity
CL_{switch}	Code length of encoding switching process
$CNML$	Conditional normalized maximum likelihood
Cov	Covariance, $Cov(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}\left((\boldsymbol{x} - \mathbb{E}(\boldsymbol{x}))^T (\boldsymbol{y} - \mathbb{E}(\boldsymbol{y}))\right)$
H	Entropy
K	Maximum possible order in a family of nested model classes
M	Number of elements in the alphabet
N	Number of stocks in the universe
Q	Quantization function
S	Number of switches
T	Number of observations

U	Utility function
W	Wealth
X	Number of states
α	Intercept in a linear regression
$\bar{r}$	Mean return
β	Regression coefficient (also called loadings or sensitivity)
$\mathbf{0}$	Vector or matrix of all zeros
$\mathbf{1}$	Vector or matrix of all ones
$\boldsymbol{B}$	Matrix of regression coefficients in multivariate linear regression
$\boldsymbol{I}$	Diagonal matrix with ones on the main diagonal and zeros elsewhere
$\boldsymbol{K}$	Kalman gain
$\boldsymbol{\Sigma}$	Covariance matrix
$\boldsymbol{\Theta}$	Set of vector-valued parameters
$\hat{\boldsymbol{\theta}}$	Maximum likelihood estimator of parameters $\boldsymbol{\theta}$
$\boldsymbol{\theta}$	Vector of parameters
δ	Lag
$\digamma$	Dividend at the end of the trading period
$\hat{k}^{(t)}$	Estimate of the order of the process conditional on information at time t
$\hat{t}_s^{(t)}$	Estimate of t_s conditional on information at time t
ι	Private information signal
λ	Proportion of informed investors
$\mathcal{B}$	Probability mass function of a Bernouilli random variable
$\mathcal{I}$	Information set
$\mathcal{M}$	Model
$\mathcal{N}$	Probability distribution function of a normal random variable
$\mathbb{E}$	Expectation operator
$\mathbb{N}$	Set of natural numbers
$\mathbb{P}$	Probability measure
$\mathbb{R}$	Set of real numbers
$\mathbb{Z}$	Set of integers
ω	Portfolio weight
ρ	Coefficient of risk aversion
σ	Volatility, standard deviation

σ^2	Variance
$\wp$	Parameter of Bernouilli random variable, $\wp = \mathbb{P}(y = 0)$
c	Rate of increase of Fisher information
e	Noise process (also called residual process)
f	Factor return
h	Differential entropy
k	Order of the model
l	Dimension of the vector of parameters
m	Minimum number of observations such that the maximum likelihood estimator exists
n	Index of stock
p	Price
q	Quantization level
r	Simple return
$r^{(t)\rightarrow(t')}$	Return between t and t', $t < t'$
r_f	Risk-free rate
$r_{\log}$	Log-return
r_{market}	Return of the market portfolio
s	Index of switch
t,t',t''	Index of time
t_s	Position of switch s (also called last time before which a switch happens)
x	State of the system
y	Observation (also called symbol)
z	Binary representation of the switching process

Acronyms

AIC	Akaike's information criterion.
AR	autoregressive.
BIBO	bounded input bounded output.
BIC	Bayesian information criterion.
BNE	Bayesian-Nash equilibrium.
CAPM	capital asset pricing model.
CARA	constant absolute risk aversion.
CNML	conditional normalized maximum likelihood.
EEG	electroencephalogram.
GARCH	generalized autoregressive conditional heteroskedasticity.
GICS	global industry classification standard.
GLS	generalized least-squares.
GS	Grossman-Stiglitz.
i.i.d.	identically and independently distributed.
IT	information theory.
KF	Kalman filter.
KT	Krichevsky-Trofimov.
LTI	linear time invariant.
MDL	minimum description length.
MSCI	Morgan Stanley Capital International.
MSE	mean squared error.
NML	normalized maximum likelihood.
OLS	ordinary least-squares.

PCA	principal component analysis.
PLS	predictive least-squares.
REE	rational expectations equilibrium.
RHS	right hand side.
S&P	Standards and Poors.
SMI	Swiss market index.
SNR	signal-to-noise ratio.
SP	signal processing.
US	United States.
w.s.s.	wide sense stationarity.

Chapter 1

Introduction

> "Begin at the beginning," the King said gravely, "and go on till you come to the end: then stop."
>
> *Alice's Adventures in Wonderland*
> Lewis Carroll

With the advance of modern computerized systems, the quantity and speed of information availability have increased dramatically in everyday life. These technical advances have been particularly pronounced in financial markets, where financial data terminals with live feeds from the world's exchanges are an imperative for a successful investment manager. These developments have, however, not eliminated the presence of information asymmetry in the market, on the contrary, it has become more relevant than ever. It is with this backdrop that we started our research into the impact of information asymmetry in financial markets. The origin of our research can be traced back to recent contributions in the study of financial markets under information asymmetry, both in the experimental finance and neurofinance field Bruguier et al. [2010], as well as in the more traditional asset pricing and mathematical finance one Biais et al. [2010]. In markets with information asymmetry, certain participants have access to privileged private information. They are referred to as the informed investors, also called insiders, as opposed to the uninformed ones. Biais et al. [2010] have developed a simple investment strategy adopting the point of view of an uninformed investor. It is solely based on price information, hence its name, *price contingent strategy*. Interestingly, it outperforms standard benchmarks. This result proves that an immunization against the adverse selection problem caused by the presence of better-informed investors is not only necessary but also possible. More specifically, the price contingent strategy uses as predictor lagged relative prices. They are defined as the weights of buy-and-hold portfolio with random starting values. Let n be the index of a stock, $n \in \{1, \ldots, N\}$. Let also $p_n^{(t)}$ be the price of stock n at time t and $r_n^{(t)}$ be the simple return of stock n at time t, defined by

$$r_n^{(t)} = \frac{p_n^{(t)} - p_n^{(t-1)}}{p_n^{(t-1)}}. \tag{1.1}$$

The relative price of stock n at time t, $p_{rel,n}^{(t)}$,is then given by

$$p_{rel,n}^{(t)} = \frac{p_{rel,n}^{(t-1)}\left(1 + r_n^{(t)}\right)}{\sum_{n=1}^{N} p_{rel,n}^{(t-1)}\left(1 + r_n^{(t)}\right)}. \tag{1.2}$$

Returns are projected on relative prices lagged by one observation

$$\boldsymbol{r}^{(t)} = \boldsymbol{B}\boldsymbol{p}_{rel}^{(t-1)} + \boldsymbol{e}^{(t)} \tag{1.3}$$

and the estimated regression coefficients $\boldsymbol{B}$ are used to predict the two first moments of stock returns. These are plugged into a portfolio optimizer, so as to obtain a mean-variance optimal portfolio, whose ex ante volatility matches that of the market. The statistical procedure used to estimate the model, namely generalized least-squares (GLS), was judged "primitive", in Peter Bossaerts's own words. This calls for an extension.

Moreover, Peter Bossaerts had a strong intuition on how to extend this: we should use a variable lag of relative price instead of its value lagged by one observation. He then coined the term *variable delay model* to describe such a model. Of course, his intuition came with a justification. I can perfectly picture him explaining it to us, while scribbling on the back of an envelope a graph that is tentatively reproduced in Figure 1.1. The two lines represent the evolution of the price in a market where some private information is distributed to a subset of the market participants, referred to as the insiders, at time 0, and publicly revealed at time T. What distinguishes the two curves is the proportion of insiders. On the one hand, when the insider is a monopolist (blue line), information diffusion is only gradual and the time of maximally informative price happens just before public revelation of the private information Kyle [1985]. On the other hand, when the number of insiders increases (gold line), there is more competition between them, the diffusion of private information is faster and the delay between maximally informative price and public revelation of private information increases. It took us three years to refine this initial intuition, verify it in the context of experimental finance, develop a mathematical solution and its applications in finance. In a nutshell, this is the topic of this thesis. Let us make this idea more definite in the remainder of this chapter.

1.1 Problem formulation and motivation

In the first part of this chapter, we have introduced the idea of a variable delay model, i.e., a model whose predictions are formed using a variable lag of its predictor. Following our work with experimental data, we have refined this initial intuition, leading to the idea of a *variable lookback model*. It is a model that uses a variable portion of its information set to form its prediction. The idea of the variable delay is essentially correct when used in conjunction with relative prices, as they constitute an accumulated measure of past prices. If we use simply past prices instead, we observe that an entire window of past prices should be used as a proxy for private information. Moreover, the size of this window varies over time. This is justified by the presence of a time-varying proportion of insiders, that affects the speed at which

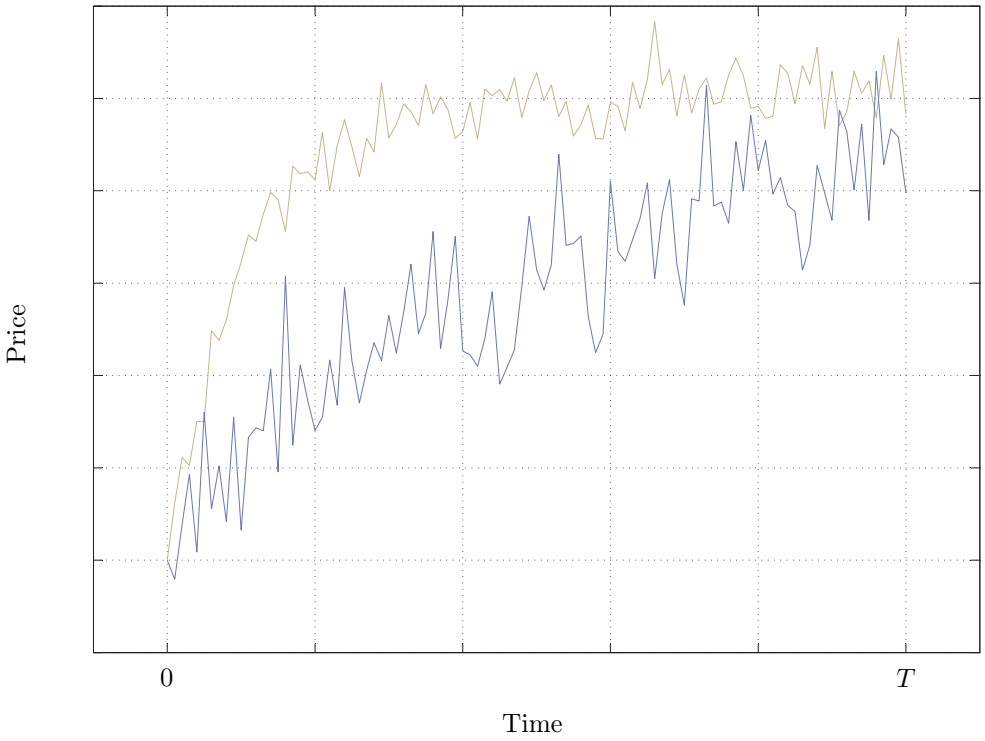


Figure 1.1: *Evolution of the price following the arrival in the market of private information at time 0, that is publicly revealed at time T. The two curves differ by the number of informed investors: one monopolist insider (blue curve) and multiple competing insiders (gold line). The delay between maximally informative price and public revelation of private information at time T is proportional to the number of insiders in the market.*

private information is diffused into the price. This idea is further developed in Section 3.1.3.

Before enunciating the key problems that are at the heart of this thesis, let us clarify the meaning of two terms, , *lookback* and *order*, that are going to constantly come back like leitmotivs. Let us imagine we have to build a statistical model of a financial time-series. Let us also assume this model depends on a l-dimensional vector of parameters

$$\boldsymbol{\theta}_l \in \boldsymbol{\Theta}_l \subset \mathbb{R}^l. \tag{1.4}$$

We call order of a model the number of the parameters in excess of the zero order model. For example, consider a linear factor model describing the return of stock n at time t, reviewed in Section 2.2.1,

$$r_n^{(t)} = \alpha_n + \beta_{n,1} f_1^{(t)} + \ldots + \beta_{n,k} f_k^{(t)} + e_n^{(t)}, \tag{1.5}$$

where $f_k^{(t)}$ is the return on factor k at time t and $e_n^{(t)}$ the residuals at time t. This model is parameterized by $k+2$ parameters, 1 for the intercept α_n, k for the regression

coefficients $\beta_{n,1}, \ldots, \beta_{n,k}$ and 1 for the variance of the residuals σ^2,

$$\boldsymbol{\theta}_{k+2} = \begin{pmatrix} \alpha_n \\ \beta_{n,1} \\ \vdots \\ \beta_{n,k} \\ \sigma^2 \end{pmatrix}. \tag{1.6}$$

The order of the process is simply k, because the zero order model already has two parameters, α_n and σ^2. Moreover, given a set of observations, a standard problem in statistics is to estimate the model, i.e., determine from these observations an estimate of the value of the parameters $\boldsymbol{\theta}_l$. We call *lookback window* the portion of the observations used to estimate the model. In the variable lookback model, the size of this window varies over time.

Remark. The notion of order and lookback are intricate idea, but should not be confused. The lookback window should be larger than the order of the model, in order to be able to compute an estimate of the parameters. For example, a lookback of at least $k + 1$ observations is necessary to obtain a unique value of the maximum likelihood estimator of the coefficients of an autoregressive (AR) process of order k.

1.1.1 Stationarity and the lack thereof

Definitions of stationarity

The problem with the idea of the variable lookback model is that it pushes us outside of the sweet spot preferred of applied mathematicians, statisticians and signal processing experts: *stationarity*. Informally, this term refers to the property of a system that does not change over time. More formal definitions of stationarity are given by Moura [2005]. Let $Y^{(t)}$ be a stochastic process, i.e., a series of random variables indexed by discrete time $t \in \mathbb{Z}$. In the most general terms, a stochastic process is completely characterized by the joint cumulative distribution function of any subset of its observations

$$F(y^{(1)}, \ldots, y^{(T)}) = \mathbb{P}\left(Y^{(1)} \leq y^{(1)}, \ldots, Y^{(T)} \leq y^{(T)}\right). \tag{1.7}$$

Here, the upper case is used to emphasize the random nature of the process, whereas the lower case represents a specific realization of the process. A stochastic process is strictly stationary if the joint cumulative distribution of any subset of observations is invariant by a shift in the observations by δ

$$F(y^{(1+\delta)}, \ldots, y^{(T+\delta)}) = F(y^{(1)}, \ldots, y^{(T)}),\ \forall \delta. \tag{1.8}$$

Very few processes are strictly stationary in practice, e.g., a finite support signal is not strictly stationary. Also, working with this general characterization of a stochastic process is extremely tedious, and one usually resorts to using a more restrictive one. In particular, another approach consists in focusing only on the first two moments of the process, which constitutes a complete characterization if the process is Gaussian.

The corresponding, weaker notion of stationarity, wide sense stationarity (w.s.s.), is given by the two conditions

$$\mathbb{E}\left(Y^{(t)}\right) = \mathbb{E}\left(Y^{(t+\delta)}\right) \tag{1.9}$$

$$Cov(Y^{(t)}, Y^{(t+\delta)}) = \mathbb{E}\left(\left(Y^{(t)} - \mathbb{E}\left(Y^{(t)}\right)\right)^T \left(Y^{(t+\delta)} - \mathbb{E}\left(Y^{(t+\delta)}\right)\right)\right) \tag{1.10}$$

$$= Cov(\delta) \tag{1.11}$$

In plain English, the mean of the process is constant and its autocovariance function is only a function of the lag δ. Furthermore the w.s.s. assumption is extremely convenient in signal processing, because the filtered version of the w.s.s. process by a linear time invariant (LTI), bounded input bounded output (BIBO) stable filter is also w.s.s. and its autocorrelation can be easily computed in the Fourier domain Sbaiz and Ridolfi [2006].

Absence of stationarity in financial time-series

From a finance perspective, there is compelling evidence that the market is not stationary, even in the weaker w.s.s. sense. For example, we will see in Section 2.2.4 that there a stable set of factors describing the cross-section of stock returns does not exist. In particular, when using principal component analysis to extract a series of orthogonal factors, a variable number of factors is necessary to explain a fixed portion of the total variance (Section 2.2.3). Furthermore, a potential justification for the presence of nonstationarity in financial time-series is that of market participants learn Bossaerts and Hillion [1999]. Learning acts as a feedback loop such that profitable arbitrage opportunities are progressively identified by market participants and consequently disappear. Moreover, two types of changes in the system can explain the absence of stationarity, drifts and *jumps*. Drifts on the one hand refer to slow structural changes in the dynamics of the system. Techniques like adaptive filtering Sbaiz and Ridolfi [2006], that assume some form of local stationarity, are well suited to track drifts. On the other hand, jumps correspond to abrupt changes in the dynamics of the system. Consider Figure 1.2 which represents the evolution of the Standards and Poors (S&P)500 index between 2003 and 2010, sampled at daily frequency. Starting in August 2007, the market has entered a phase of acute crisis, that started with the burst of a real estate bubble in the United States (US). It has now evolved into one of the worst global financial and economic crises since the 1930's. Clearly, jumps are better suited to describe the market entering a crisis.

Limitations of existing methods to handle nonstationarity

Despite this evidence, it is surprising to notice that, from a mathematical perspective, everything is done to come back to some form of stationarity when modeling financial time-series. The reason is certainly that this assumption facilitates the mathematical treatment, a role comparable to the Gaussianity assumption. It is also due to the fact that so little is known in the absence of stationarity. One way of bypassing the problem of nonstationarity is to assume some form of local stationarity. This means that the system stays constant over a certain portion of the most recent

Figure 1.2: *Evolution of the log-price of the S&P500 index, between 01.01.2003 and 01.01.2012, sampled at daily frequency.*

observations, which is used for the estimation. This is referred to as local windowing in signal processing. The problem is the choice of the appropriate window size. A short window leads to quick adaptation, but the resulting estimates are noisier. Thus there exists a trade-off between the speed of adaptation and the amount of noise in the solution. However, windowing is intuitively unsuited to handling jumps in the system. Consider again Figure 1.2. Let us assume a 5 year window of monthly data is used to estimate our favorite model to predict the return on the S&P500 index just before the crisis in July 2007. The lookback window is rolled over the observations, i.e., that, for every new observation, we discard the oldest one, add the newest one and run again the estimation procedure. When entering the crisis in August 2007 and in the coming months, the window mostly contains data points which are no longer relevant for the current dynamics of the system. Of course, managers will adapt their window size over time, and this is one of the most often reported changes in quantitative portfolio management by funds in 2009 Shari [2011]. This process is however prone to human bias, and an unbiased and automated solution is highly desirable. Another way to come back to the stationary situation is to model the switching process itself Kim and Nelson [1998]. The example below gives conditions under which an AR process of order 1, which switches between different modes following a Markov chain of order 1 is w.s.s.. This could be generalized to more general classes of processes Timmermann [1999]. We discuss the limitation of this approach in 2.2.4.

Example 1. *Consider the following Markov switching AR process of order 1, which evolves according to*

$$y^{(t)} = \alpha_{x^{(t)}} + \beta_{1,x^{(t)}} \left(y^{(t-1)} - \alpha_{x^{(t-1)}} \right) + \sigma_{x^{(t)}} e^{(t)}, \tag{1.12}$$

where the intercept $\alpha_{x^{(t)}}$, the AR coefficient $\beta_{1,x^{(t)}}$ and the standard deviation $\sigma_{x^{(t)}}$ depends on the state of the system $x^{(t)} \in \{1, \ldots, x\}$. The residuals evolve according to an identically and independently distributed (i.i.d.) Gaussian process with unit variance

$$e^{(t)} \sim \mathcal{N}(0, 1), \tag{1.13}$$

and the switching process $x^{(t)}$, independent of $e^{(t)}$, is governed by a Markov chain of order 1 with X states. This process is parameterized by a matrix of transition probabilities whose elements are given by

$$\Pi_{x,x'} = \mathbb{P}\left(x^{(t)} = x' / x^{(t-1)} = x \right), \tag{1.14}$$

under the constraints

$$0 \leq \Pi_{x,x'} \leq 1 \text{ and } \sum_{x'=1}^{X} \Pi_{x,x'} = 1. \tag{1.15}$$

Let $\boldsymbol{\varpi}$ be the steady state probability of the transition matrix,

$$\boldsymbol{\Pi}^T \boldsymbol{\varpi} = \boldsymbol{\varpi}, \tag{1.16}$$

i.e., the eigenvector of the matrix $\boldsymbol{\Pi}^T$ associated with the eigenvalue 1 and renormalized s.t. the sum of the probabilities equals 1. Provided that $\boldsymbol{\varpi}$ exists and that the system is stable,

$$\sum_{x=1}^{X} \varpi_x |\beta_{1,x}| < 1, \tag{1.17}$$

let us show that $y^{(t)}$ is w.s.s.. The moments are computed in the stationary regime of the switching process. Using backward substitution, we can rewrite (1.12) as

$$y^{(t)} - \alpha_{x^{(t)}} = \sum_{t'=1}^{\infty} \left(\prod_{t''=0}^{t'-1} \beta_{x^{(t-t'')}} \right) \sigma_{x^{(t-t')}} e^{(t-t')} + \sigma_{x^{(t)}} e^{(t)}. \tag{1.18}$$

Taking the expectations on both side and by linearity of the expectation operator

$$\mathbb{E}\left(y^{(t)} - \alpha_{x^{(t)}} \right) = \tag{1.19}$$

$$\textstyle\sum_{t'=1}^{\infty} \mathbb{E}\left(\left(\prod_{t''=0}^{t'-1} \beta_{x^{(t-t'')}} \right) \sigma_{x^{(t-t')}} e^{(t-t')} \right) + \mathbb{E}\left(\sigma_{x^{(t)}} e^{(t)} \right) \overset{(a)}{=} \tag{1.20}$$

$$\textstyle\sum_{t'=1}^{\infty} \mathbb{E}\left(\left(\prod_{t''=0}^{t'-1} \beta_{x^{(t-t'')}} \right) \sigma_{x^{(t-t')}} \right) \mathbb{E}\left(e^{(t-t')} \right) + \mathbb{E}\left(\sigma_{x^{(t)}} \right) \mathbb{E}\left(e^{(t)} \right) = 0, \tag{1.21}$$

where (a) results from the independence between $x^{(t)}$ and $e^{(t)}$. Therefore, using the

law of conditional expectation, the first moment of the process is given by

$$\begin{aligned}\mathbb{E}\left(y^{(t)}\right) &= \mathbb{E}\left(\alpha_{x^{(t)}}\right) && (1.22)\\ &= \sum_{x=1}^{X}\alpha_x\mathbb{P}\left(x^{(t)}=x\right) && (1.23)\\ &= \boldsymbol{\varpi}^T\boldsymbol{\alpha}, && (1.24)\end{aligned}$$

which is independent of t. Furthermore, let us compute

$$\mathbb{E}\left(\left(y^{(t)}-\alpha_{x^{(t)}}\right)^2\right)=\sum_{x=1}^{X}\mathbb{E}\left(\left(y^{(t)}-\alpha_{s(t)}/x^{(t)}=x\right)^2\right)\mathbb{P}\left(x^{(t)}=x\right). \quad (1.25)$$

Let us develop the conditional expectation term in the right hand side (RHS) of the previous equation.

$$\begin{aligned}&\mathbb{E}\left(\left(\beta_{1,x^{(t)}}\left(y^{(t-1)}-\alpha_{x^{(t-1)}}\right)+\sigma_{x^{(t)}}e^{(t)}/x^{(t)}=x\right)^2\right)= && (1.26)\\ &\beta_{1,x}^2\mathbb{E}\left(\left(y^{(t-1)}-\alpha_{x^{(t-1)}}\right)^2/x^{(t)}=x\right)+\sigma_x^2= && (1.27)\\ &\beta_{1,x}^2\textstyle\sum_{x'=1}^{X}\mathbb{E}\left(\left(\ldots\right)^2/x^{(t)}=x,x^{(t-1)}=x'\right)\mathbb{P}\left(x'/x\right)+\sigma_x^2= && (1.28)\\ &\beta_{1,x}^2\textstyle\sum_{x'=1}^{X}\mathbb{E}\left(\left(y^{(t-1)}-\alpha_{x^{(t-1)}}\right)^2/x^{(t-1)}=x'\right)\frac{\Pi_{x',x}\varpi_{x'}}{\varpi_x}+\sigma_x^2. && (1.29)\end{aligned}$$

Therefore, to ensure stationarity, we have

$$\mathbb{E}\left(\left(y^{(t)}-\alpha_{x^{(t)}}\right)^2\right)=\frac{\sum_{x=1}^{X}\sigma_x\varpi_x}{1-\sum_{x,x'=1}^{X}\beta_{1,x}^2\Pi_{x',x}}\varpi_{xi'}=\frac{\boldsymbol{\varpi}^T\boldsymbol{\sigma^2}}{1-\boldsymbol{\varpi}^T\boldsymbol{\Pi}^T\boldsymbol{\beta}_{1,\cdot}^2}. \quad (1.30)$$

Using this result, the variance of the process can be expressed as

$$\begin{aligned}Var\left(y^{(t)}\right) &= \mathbb{E}\left(\left(y^{(t)}-\mathbb{E}\left(y^{(t)}\right)\right)^2\right) && (1.31)\\ &= \mathbb{E}\left(\left(\left(\alpha_{x^{(t)}}-\boldsymbol{\varpi}^T\alpha\right)+\left(y^{(t)}-\alpha_{x^{(t)}}\right)\right)^2\right) && (1.32)\\ &= \mathbb{E}\left(\left(\alpha_{x^{(t)}}-\boldsymbol{\varpi}^T\alpha\right)^2\right)+\mathbb{E}\left(\left(y^{(t)}-\alpha_{x^{(t)}}\right)^2\right) && (1.33)\\ &= \sum_{x=1}^{X}\varpi_x(\alpha_x-\boldsymbol{\varpi}^T\alpha)+\mathbb{E}\left(\left(y^{(t)}-\alpha_{x^{(t)}}\right)^2\right) && (1.34)\\ &= \boldsymbol{\varpi}^T\left(\boldsymbol{\alpha}-\boldsymbol{\varpi}^T\boldsymbol{\alpha}\mathbf{1}\right)+\frac{\boldsymbol{\varpi}^T\boldsymbol{\sigma^2}}{1-\boldsymbol{\varpi}^T\boldsymbol{\Pi}^T\boldsymbol{\beta}_{1,\cdot}^2}. && (1.35)\end{aligned}$$

Therefore, the variance does not depend on time t. Similarly, the autocovariance at

the lag 1 is given by,

$$\mathbb{E}\left(\left(y^{(t)} - \mathbb{E}\left(y^{(t)}\right)\right)\left(y^{(t-1)} - \mathbb{E}\left(y^{(t-1)}\right)\right)\right) = \tag{1.36}$$

$$\mathbb{E}\left(\left(\alpha_{x^{(t)}} - \boldsymbol{\varpi}^T\boldsymbol{\alpha} + \beta_{1,x^{(t)}}\left(y^{(t-1)} - \alpha_{x^{(t-1)}}\right) + \sigma_{x^{(t)}}e^{(t)}\right)\left(y^{(t-1)} - \boldsymbol{\varpi}^T\boldsymbol{\alpha}\right)\right) = \tag{1.37}$$

$$\mathbb{E}\left(\left(\alpha_{x^{(t)}} - \boldsymbol{\varpi}^T\boldsymbol{\alpha}\right)\left(\alpha_{x^{(t-1)}} - \boldsymbol{\varpi}^T\boldsymbol{\alpha}\right)\right) + \mathbb{E}\left(\left(y^{(t-1)} - \alpha_{x^{(t-1)}}\right)^2\right) = \tag{1.38}$$

$$\sum_{x,x'=1}^{X} P_{x'x}\varpi_{x'}(\alpha_x - \boldsymbol{\varpi}^T\boldsymbol{\alpha})(\alpha_{x'} - \boldsymbol{\varpi}^T\boldsymbol{\alpha}) + \mathbb{E}\left(\left(y^{(t-1)} - \alpha_{x^{(t-1)}}\right)^2\right). \tag{1.39}$$

Again, this does not depend on time t. This can be generalized to show that the autocovariance at lag δ is only a function of δ Timmermann [1999].

1.1.2 Key problems

Our discussion on the nonstationarity inherent to financial time-series and the limitations of the current approaches to handle it calls for the development of new signal processing techniques. This is at the heart of this thesis. The goal is to monitor when a model works and when it fails. Moreover, although our initial motivation came from the study of financial markets under information asymmetry, our proposed solution is more general, as it aims to handle nonstationarity in a broader context. The only assumption that we make is that of *piecewise stationarity*. It simply means that there exists a series of switching times $t_0 = 1 < t_1 < \ldots < t_S \leq T$ such that the data are generated by a model of order $k_s \in \{0, \ldots, K\}$ in the interval $[t_{s-1}; t_s)$. The difficulty comes from the fact that we make no assumption on the number of switches S, the positions at which they happen t_s, $s \in \{1, \ldots S\}$ or the switching process itself. See Figure 1.3 for an illustration a piecewise stationary processes. We are willing to put ourselves in the more complicated situation of nonstationarity and abandon any knowledge on the underlying switching process, because we are solely interested in deriving evidence from data. Also, as is the case with all financial signals, we only have access to one realization of the process and we want to draw conclusions just from that. Furthermore, our solution does not rely on the ergodicity assumption.

1.1.3 Key questions

Following our discussion in the previous sections, there are three questions at the heart of this thesis.

Can we verify the intuition of the variable lookback model in experimental finance?

We have introduced the idea of a variable lookback model, initially motivated by the theory of financial markets under information asymmetry. Experimental finance constitutes an ideal setup to verify the intuition underlying the model for various reasons. Firstly, experimental markets represent a simplification of a real world scenario, and it is possible to disentangle competing effects. This is not always possible in field data. Secondly, experiments are run in a controlled setup; each independent replication depends on a set of parameters, and the experimenter can vary the values of the parameters so as to study their impact on the outcome. Finally, when analyzing the result, the experimenter has access to quantities not directly observable in

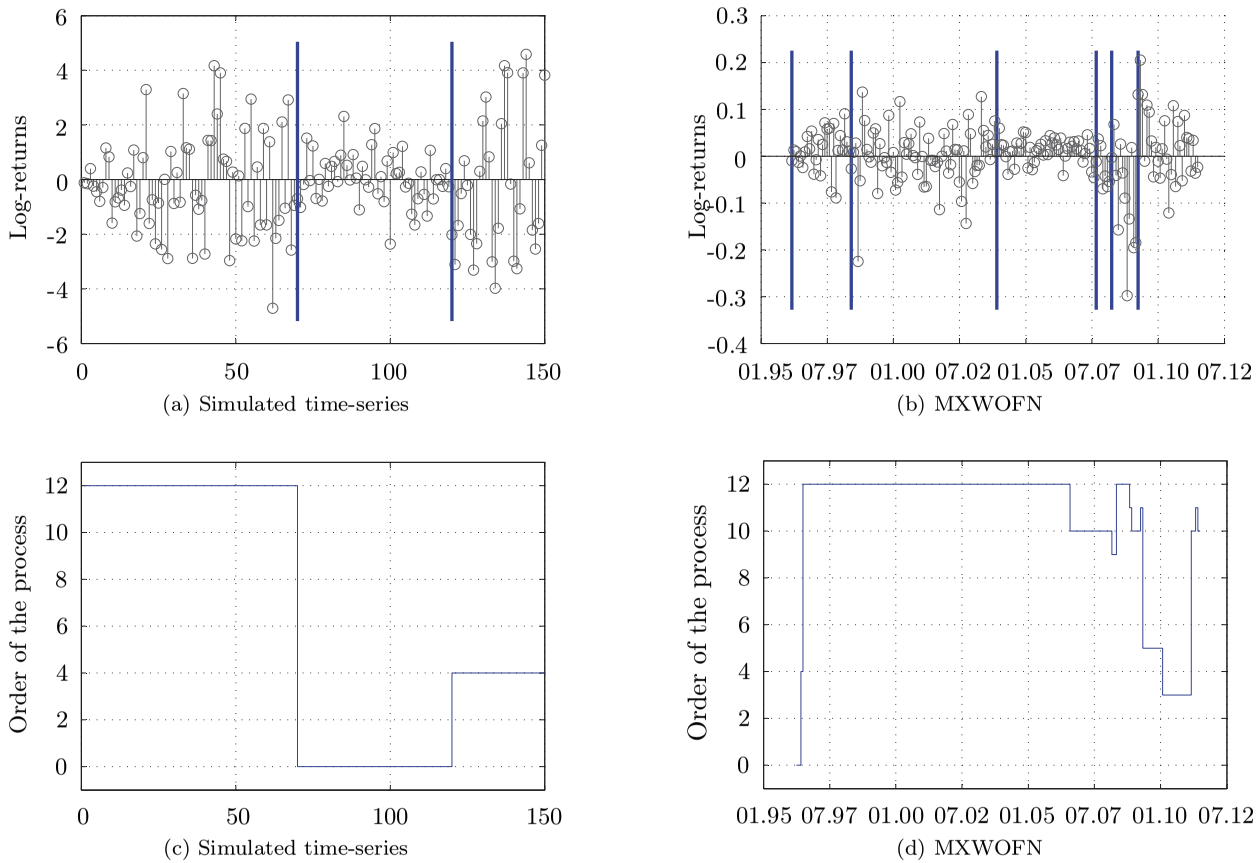


Figure 1.3: *Example of piecewise stationarity processes. (a) Swicthing points (vertical bars) vs. log-returns, simulated time-series. (c) Order of the process, simulated time-series. (b) Switching points identified by the variable lookback algorithm (vertical bars) vs. log-returns, MXWOFN. (d) Order of the process identified by the variable lookback algorithm, MXWOFN.*

field data, that serve as ground truths, for example, the private information signal, or the proportion of insiders. In particular, we need to verify the theoretical prediction that the delay between the time of maximally informative price and full revelation of information is correlated with the number of insiders. The problem of measuring information and maximally informative price also needs to be addressed.

How can we estimate the variable lookback model?

The second question deals with the estimation of the variable lookback model. Although, the model was originally motivated by the study of financial markets under information asymmetry, our solution is more general. Indeed, we are interested in the development of new signal processing algorithm, that are able to handle nonstationarity inherent to financial signals, beyond the existing limited approaches. Furthermore, the solution should be general, independent of the choice of model, so as to avoid turning this thesis into a data mining exercise. Finally, the solution is solely based on the assumption of piecewise stationarity.

How do the resulting strategies perform compared to standard benchmarks?

We are aiming to demonstrate the added value of the proposed algorithm. Adopting a practitioner's perspective, the goal is to put the proposed method to work to build investment strategies. To assess the performance of a strategy, backtest, i.e., replication of investment decisions in the past using historical data, is the method of choice. We review this evaluation methodology in Section 2.1.4. The performance of the proposed strategy is compared with standard benchmarks, for example, the passive indexing strategy or the strategy that is based on a simpler estimation technique.

1.1.4 Unique contributions of this thesis

We have just presented the three key questions at the heart of this thesis, and, logically, unique contributions of this thesis reflect this tripartite structure.

Contributions in the area of experimental finance

The first contribution of this thesis is the novel analysis of experimental market data. Note that we have not conducted the experiments ourselves, but we instead use an existing dataset obtained by Bruguier et al. [2010] and Bossaerts et al. [2010]. In particular, we first develop a method to highlight the diffusion of information into the price using the cross-section of experiments that have the same number of insiders. One may argue that this analysis is already contained in Bossaerts et al. [2010], by "mentally combining" Figure 1 and Figure 4, so as to obtain the time-varying regression coefficients. Our approach has the advantage of making this explicit and visually convincing. This small contribution also immediately rules out the hypothesis that experimental data contain only noise and no useful information. More importantly, we develop a mathematical definition of the intuitive notion of time of maximally informative price and use the Context algorithm Rissanen [2005] to identify it from individual time-series of prices. We stress here that the Context algorithm already

existed, but its use to analyze experimental finance datasets is, to the best of our knowledge, new. Finally, we verify that the time of maximally informative price is inversely proportional to the proportion of insiders in the market, as predicted by both Bayesian-Nash and noisy rational expectations equilibrium theories. We also demonstrate the existence of long-lived informational inefficiencies.

Contributions in the area of nonstationary signal processing

The second contribution of this thesis is in the area of signal processing for nonstationary signals. We develop an online, universal algorithm for the joint identification of the order of a process and the relevant lookback window. Our approach draws from a large body of literature, but differs from existing approaches. Firstly, our approach is rooted in the information theoretic learning literature, referred to as the minimum description length (MDL). Most model selection criteria developed in this context assume that the underlying sequence of observations is stationary. Also, as studied in van Erven [2006], if one nevertheless applies the criterion on sequences with jumps in their dynamics, the MDL criterion will ultimately adapt and recognize the new dynamics. However, there exists a momentum effect in MDL selection, i.e., a delay between an underlying jump in the system and its detection. This calls for a better method to handle jumps explicitly. Secondly, there are also similarities with the work of Cover and Ordentlich [1996] dealing with universal portfolios. These dynamical portfolios are called universal because they achieve a performance close to the best static portfolio chosen in hindsight within a certain class of portfolios, namely, that of constant rebalanced portfolios. Cover's algorithm is however exponential in the number of assets, which makes it inpractical for concrete applications. Moreover, it is true that the portfolio maximizing the growth rate of wealth is indeed a constant rebalanced portfolio, if investors have power utility functions, in particular log-utility function in their terminal wealth. A constant rebalanced portfolio is however not a target preferred by practitioners. Thus, trying to reach a performance close to it does not seem attractive in the eyes of practitioners, who are typically subject to other investment benchmarks, like the market index. Thirdly, the paper of Willems [1996] is a main source of inspiration, in particular the use of the quadratic tree diagram to maintain in an organized manner competing models of the data. Willems's algorithm is developed to handle binary sequences only, and his approach is Bayesian. That is, he computes a mixture of prediction models over the choice of switching times. The weighting applied to each prediction model is given by the Krichevsky-Trofimov estimator of the probability that this series of switching times occurs Krichevsky and Trofimov [1981]. In comparison, our approach is valid for more general alphabets, including continuous random variables, and we use a selection over the choice of switching times. The main argument in favor of the Bayesian approach is that, if two models are good at predicting a sequence, a combination of the two, for example, by equally weighting each prediction, is also a good predictor. While we completely agree with this idea, it is only valid if the models are good. If some prediction models are particularly bad, this dilutes the predictive power of good models. This is in particular true when the weighting scheme used for combining them does not depend on the predictive power of each model. From a finance perceptive, a Bayesian mixture will result in a strategy that makes only mild bets, and it could not perform

correctly. Also, another reason to select the switching times is that we are interested in the identification of the switching process itself, and would like to relate it to other economic variables. A fourth related work is that of Kozat and Singer [2008], which is developed for sequences taking values in a general alphabet. But their approach is also Bayesian, the algorithm uses a simpler linear tree diagram and the model is more constrained, as it does not allow switches between the order of the process within the same period of piecewise stationarity. Finally, the algorithm is not fully data-dependent, and requires the addition of a parameter; this is further discussed in 4.2.2. Finally, Yamanishi [2007] and Yamanishi and Sakurai [2010] have recently proposed a method for dynamic model selection, that also uses the Krichesky-Trofimov estimator to penalize switches in the model. Their approach is defined for a more constrained structure of switches, as the order of the process is only allowed switching between adjacent values. Also, the first version of their algorithm is offline, and more recent versions perform their operations only in blocks.

Contributions in the area of mathematical finance

Finally, a contribution of this thesis is in the area of mathematical finance. We backtest different strategies resulting from the applications of the proposed algorithm and compare their performance against standard benchmarks, e.g., the market index or the strategy based on a simpler estimation method. Adopting a practitioner's perspective, we present an entire array of indicators, so as to obtain a complete overview of the performance of the strategies. In particular, we develop an adaptive momentum strategy. The momentum strategy is a well known investment strategy, documented by academics' Jegadeesh and Titman [1993] and practitioners' Sanford and Cooper [2006], Koo and Panigirtzoglou [2008] studies alike. Our approach resembles that of time-series momentum of Moskowitz et al. [2012] or Hou and Moskowitz [2005]. In the latter, past prices are also used as a proxy for private information, with the difference that their model uses a fixed lag structure, whereas we allow for a more general model, that switches between phases where (a) momentum performs well, (b) contrarian performs well or (c) neither momentum nor contrarian performs well. Another approach consists of investing in a portfolio of low correlated strategies, instead of trying to time a strategy. This unfortunately does not work in a crisis, where all strategies tend to correlate on the downside. Finally, our approach is also related to Kent [2010]. But his objective is to find a hedge for the momentum strategy, whereas our selection algorithm aims to time the strategy, and enter a contrarian strategy that typically works when momentum fails, and this does so solely based on price information.

1.2 Thesis outline

The remainder of this thesis is organized as follows. In Chapter 2, since we assume that the reader has no prior knowledge in finance, we start by reviewing necessary finance concepts. In particular, we define terms such as return, portfolio, index, etc. We then present two existing applications of signal processing in quantitative finance. Both deal with the problem of factor modeling, the first one is concerned with the estimation of a factor model using the Kalman filter, the second with the extraction of

a set of orthogonal factors using principal component analysis. We aim not to review every existing applications of signal processing in finance. Rather we would like to illustrate the limitations of current approaches, in particular the incorrect handling of nonstationarity.

As already discussed, an important motivation behind this work is the study of financial markets under information asymmetry, and this is the topic of Chapter 3. We first review and compare two existing bodies of literature dealing with markets under information asymmetry, namely the noisy rational expectations theory and the Bayesian Nash equilibrium theory. We also come back to the initial intuition of the variable lookback model. In the second part of the chapter, we present supporting evidence for the variable lookback model in the context of experimental markets. We first explain and comment on the setup of the experiment. We then highlight the process of diffusion of private information into the price, by studying the evolution of the regression coefficient of price on private information. This regression is performed in the cross-section of experiments that have the same number of insiders. We also make more precise the notion of the time of maximally informative price, and use the Context algorithm Rissanen [2005] to detect it. Finally, we verify that the time of maximally informative price is inversely proportional to the proportion of insiders in the market.

In Chapter 4 we present our main solution to estimate the variable lookback model and coin the term *variable lookback algorithm*. As this learning algorithm is based on the *minimum description length* principle, we start by reviewing this large body of literature. Then, starting from a coarse, block diagram view of the algorithm, we further refine its design. Our discussion also encompasses study of the behavior of the algorithm, computational aspects and implementation details. Tests of the algorithm on simulated time-series conclude this chapter. In particular, we test the model selection ability of the algorithm when there is no switch in the system, as well as the detection of a jump in the simplified scenario, when the underlying process switches form a highly correlated AR process of order 1 to an i.i.d. process, and vice versa.

Chapter 5 deals with applications of the variable lookback algorithm in finance. We first present an illustration of the algorithm when used to dynamically draw the trend line for the time-series of log-prices. In that case, there is only one model and the algorithm only decides how to segment the time-series in periods of piecewise stationarity. We then move to a more realistic example, the adaptive momentum strategy. The variable lookback algorithm is in that case used to simultaneously select (a) the lookback window, (b) the type of strategy (momentum, contrarian or none) and (c) the horizon of past returns used as predictors. Detailled evaluation is provided, including a study of the behavior of the strategy during the 2009 crisis.

Chapter 6 contains a summary, further research directions and concluding remarks.

Chapter 2

Signal Processing for Quantitative Finance

> This is the worst signal I have seen in my life.
>
> Martin Vetterli

A signal is a mathematical representation of a real world phenomenon, that varies in time and/or space Prandoni and Vetterli [2008]. For example, a sound, an image or the electric potential measured by a biomedical sensor such as an electroencephalogram (EEG) are all signals. Consequently, signal processing techniques have found applications in a large number of fields: audio, image processing, biomedical signal processing to name a few. Data from the stock markets, such as price, volume and order flow, etc. can also be regarded as signals. It is thus sensible to envision applications of signal processing in finance Cont [2011], Jay et al. [2011], Ganesan [2011]. Existing applications are the subject of this chapter. Why is signal processing in finance of particular relevance today?

(i) Availability of data: over the last 20 years, electronic stock markets have largely replaced physical ones. And high-frequency trading, mostly computer-driven, now accounts for 56% and 38% of equities trades in the US and Europe, respectively Biais and Wolley [2011]. They have lead to the availability of large quantities of data, up to a sampling period of the order of a millisecond.

(ii) Challenging problems: financial signals have unique features such as their nonstationarity, deviation from normality, that pose unique and interesting challenges.

(iii) Lessons from the crisis: starting in August 2007, financial markets have entered a phase of acute crisis, considered to be the worst one since the 1930's. What started as a the burst of a real estate bubble in the US has evolved into a global financial and economic crisis. One of the reasons that have been put forward as origin of this crisis is the increased mathematization of finance, as well as the limitations and failures of the existing methods and models. We can

therefore see the current crisis as an opportunity to suggest improvements. Furthermore, the financial crisis also suggests a change in methodology in finance research; there is a necessity to move away from theoretical evidences, obtained by solving analytically stylized mathematical models, and dwell more on evidences from data. By data, we encompass both field and experimental data. In both cases, signal processing could contribute to the innovative analysis of these datasets.

The remainder of this chapter is organized as follows. To make this thesis as self-contained as possible, we start by reviewing financial concepts necessary for its understanding. We define terms such as return and risk (Section 2.1.2), portfolio and indices (Section 2.1.3) before introducing a special class of investment strategies, quantitative strategies (Section 2.1.4). We then review two existing applications of signal processing in quantitative finance. Both deals with the general problem of factor modeling (Section 2.2.1). First, we use the Kalman filter (KF) to estimate a factor model, which leads to improvements compared to ordinary least-squares (OLS) (Section 2.2.2). Then, we present the principal component analysis (PCA) procedure and apply it to extract a series of orthogonal factors driving the cross-section of stock returns (Section 2.2.3). We do not aim to give an exhaustive treatment of all existing applications of signal processing in finance, especially since signal processing is a large field with blurred boundaries. Rather, we aim to point out the limitations of current applications, in particular the limitations of the windowing approach in dealing with nonstationarity inherent to financial signals (Section 2.2.4).

2.1 Review of financial concepts

2.1.1 Trading in the stock market

A *stock* is a share of ownership of a company Morellec [2008]. It gives its holder certain rights, for example, the entitlement to a fraction of the company's profit, a dividend, provided it is profitable. Moreover, stockholders have the lowest seniority on the assets of a company in the event of a bankruptcy. This means that, in that event, the proceeds of the sales of the company's assets are first attributed to other securities holders, in particular bond holders. The stock market is a place where stocks are traded. An investor takes two types of trading positions in the market, *long* and *short*. A long position allows benefiting from soaring prices. The investor buys the asset now and sells it at some point in the future at some hopefully higher price. On the contrary, a short position allows benefiting from falling prices. The investor sells immediately an asset borrowed from his broker and buys it back later at a hopefully lower price to replace the originally borrowed asset. There are limitations in the short selling activity, which are determined by national or exchange regulations. Furthermore, there is an accounting mechanism involving so-called margin account that limits the position of an individual investor. This ensures he has enough money to buy back the borrowed asset.

In the most general terms, investors trade in the market because they have differences in beliefs, endowments or preferences Grossman and Stiglitz [1980]. The double auction mechanism is now the standard mechanism used by almost all exchanges to conduct trading. An investor can submit two types of order. A limit order is an

order to buy or sell a stock at a certain price SEC [2011a]. This price should be a multiple of the minimum tick size, as determined by the exchange rules. For example, buy stock X (in other words enter a long position on X) when it reaches 100$ is a limit order. If the tick size is 25 cents, the next possible buy limit order is 100.25$. Limit orders are aggregated by a central market maker and displayed in a public order book. In an order book, rows correspond to the price level, and the number in a given row to the number of outstanding orders at this price. See example below. The left column contains all outstanding buy limit orders arranged in decreasing value of their price; the one at the top is called the bid price. Likewise, the right column contains all outstanding sell limit orders, arranged in increasing value of their price; the one at the bottom is called the ask price. The average between the bid and ask price is called the quote mid-price. The difference between the bid and ask price is called the bid-ask spread and corresponds to the reward made by the market maker for his activity. A limit order stays in the order book, unless (a) it is cancelled by its owner, (b) it becomes stale after a certain amount of time depending on exchange rules, or (c) it is matched by a market order. The latter represents another type of order that is executed immediately at the best available market price SEC [2011b]. Observe the difference in associated uncertainty between the two types of order. In a limit order, the price is fixed but the investor does not know when the order will be executed (if at all). In a market order, the order is executed immediately but there is an uncertainty on the price at which this happens.

Example 2. *The following table represents an example of an order book. If an investor submits now a sell market order of 8 units, the market maker first matches this order with the 2 outstanding buy limit orders at 99.50$, the 5 ones at 99.25$ and 1 of the 9 ones at 99.00$. This simple example shows how, by him placing a large order in a rather illiquid market, a single trader can impact the price.*

Number of outstanding orders		
Buy	Sell	Price ($)
		⋮
	10	100.75
	25	100.50
	15	100.25
	–	100.00
–		99.75
2		99.50
5		99.25
9		99.00
		⋮

Table 2.1: *Order book.*

2.1.2 Return and risk

By trading, investors are interested in the resulting rate of *return* (or simply return) of their investments Jondeau [2009], Morellec [2008]. This is because the size of the asset does not matter to them in perfectly competitive markets. Simple return between (discrete) time $t-1$ and t is defined as

$$r^{(t)} = \frac{p^{(t)} - p^{(t-1)}}{p^{(t-1)}}, \tag{2.1}$$

where $p^{(t)}$ is the price of the stock at time t. Furthermore, since the price is a unit root process, taking the first order difference makes the properties of returns easier to handle than that of the price. Given this definition, the wealth of an investor at time $t+1$, $W^{(t+1)}$, that invests at time t all his wealth $W^{(t)}$ in a stock with return $r^{(t+1)}$ is given by

$$W^{(t+1)} = (1 + r^{(t+1)})W^{(t)}. \tag{2.2}$$

Similarly, over T trading periods, we obtain

$$W^{(t+T)} = W^{(t)} \prod_{t'=1}^{T} (1 + r^{(t+t')}). \tag{2.3}$$

Returns are compounded, i.e., the profits of the previous periods are reinvested in subsequent periods. If trading and compounding takes place in continuous time, this leads to the notion of *log-returns*, given by

$$r_{\log}^{(t)} = \log\left(\frac{p^{(t)}}{p^{(t-1)}}\right). \tag{2.4}$$

Observe that simple and log-returns are related by the relation

$$r_{\log}^{(t)} = \log(1 + r^{(t)}). \tag{2.5}$$

Also, when working with log-returns, the return over several periods is equal to the sum of returns over each period, or

$$r_{\log}^{(t-h)\to(t)} \equiv \sum_{t'=0}^{h-1} r_{\log}^{(t-t')}. \tag{2.6}$$

Figure 2.1 represents the distribution of daily log-returns for the S&P500, observed from 01.01.2007 until 01.01.2012, compared to a Gaussian distribution with similar mean and variance. We observe that the distributions of daily log-returns is far from being Gaussian; it is more skewed on the the left, and the tails are fatter than the Gaussian distribution, indicating the presence of so-called extreme events, particularly on the downside. Quite interestingly, this illustrates that the glorified Black-Scholes model Karatzas and Shreve [2004], which assumes the normality of log-returns, is rather ill-suited to describe the distribution of log-returns.

Investors also want to assess the risk of their investments. Since the seminal work of Markowitz [1952], they do this (imperfectly) using the *volatility* of the assets,

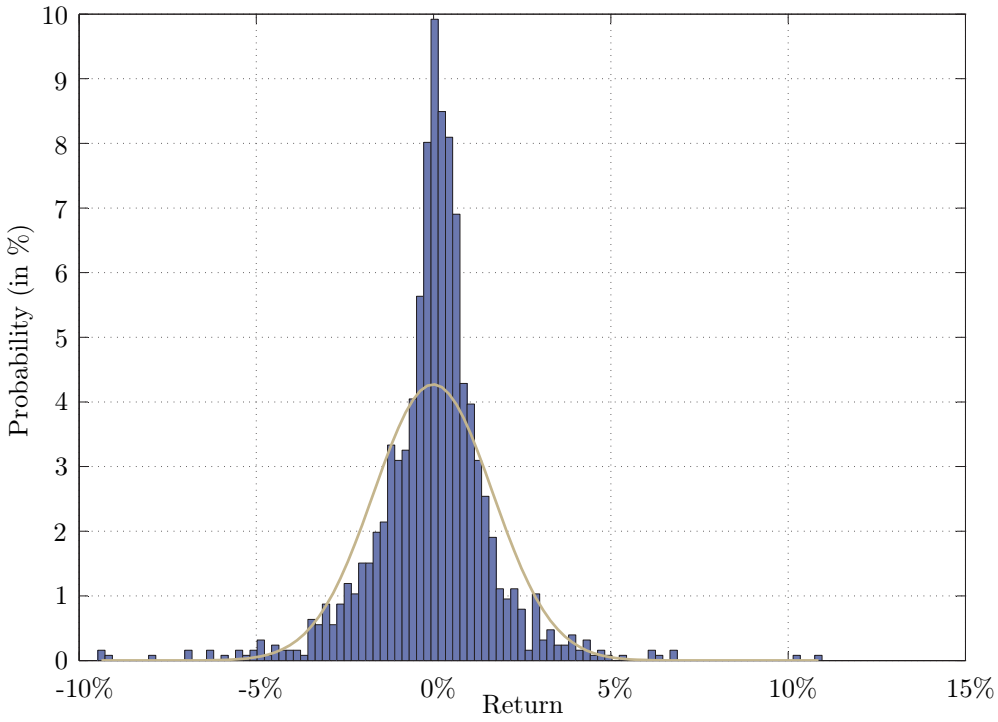


Figure 2.1: *Blue: histogram of daily log-returns of the S&P500 for the period between 01.01.2007 and 01.01.2012. Gold: probability distribution function of the Gaussian distribution fitted on the same observations. We observe that the distribution of log-returns deviates greatly from normality. In particular, the distribution is skewed on the left and the tails are fatter than those of the Gaussian distribution.*

$\sigma^{(t)}$. This is simply the standard deviation of return. This quantity is not directly observable and several estimators based on past observed returns have been proposed Jondeau [2009]. Given a set of T returns observations, the unconditional volatility is given by

$$\sigma = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}\left(r^{(t)} - \bar{r}\right)^2} \tag{2.7}$$

where $\bar{r} = 1/T \sum r^{(t)}$ is the unconditional mean return. Given a window of the latest T observations, the conditional volatility is a local estimate of the volatility,

$$\sigma^{(t)} = \sqrt{\frac{1}{T-1}\sum_{t'=0}^{T-1}\left(r^{(t-t')} - \bar{r}^{(t)}\right)^2}. \tag{2.8}$$

where

$$\bar{r}^{(t)} = \frac{1}{T}\sum_{t=0}^{T-1} r^{(t-t')} \tag{2.9}$$

is the conditional mean. This estimator puts the same weight on all observations. To give more weight to most recent observations, we can use the exponential weighted volatility computed iteratively using the following equations:

$$\bar{r}^{(t)} = \kappa \bar{r}^{(t-1)} + (1-\kappa) r^{(t-1)} \tag{2.10}$$

$$\sigma^{2(t)} = \kappa \sigma^{2(t-1)} + (1-\kappa)\left(r^{(t-1)} - \bar{r}^{(t)}\right)^2 \tag{2.11}$$

$$\sigma^{(t)} = \sqrt{\sigma^{2(t)}}. \tag{2.12}$$

In the equations above, $\bar{r}^{(t)}$ is called the exponential weighted mean return. Note that both exponential weighted mean return and volatility are observable at time $t-1$. κ, the forgetting factor, is set typically to 0.94, as in RiskMetrics. Figure 2.2 represents the exponential weighted volatility of S&P500 from 01.01.2007 to 01.01.2012. We observe the phenomenon of volatility clustering: periods of high volatility are concentrated together and alternate with periods of low volatility. This persistence in the time-series of conditional volatility is well captured by generalized autoregressive conditional heteroskedasticity (GARCH) models Bollerslev [1986].

Figure 2.2: *Exponentially weighted conditional volatility of the daily log-returns of S&P500, the simplest form of GARCH processes, for the period between 01.01.2007 until 01.01.2012. We observe the phenomenon of volatility clustering, i.e., periods of high volatility are concentrated together and alternate with periods of low volatility.*

2.1.3 Portfolio and index

A *portfolio* is a basket of N stocks Morellec [2008]. The return of the portfolio $r_p^{(t)}$that invests a portion $\omega_n^{(t)}$ of its wealth in stock n with return $r_n^{(t)}$ is given by

$$r_p^{(t)} = \sum_{n=1}^{N} \omega_n^{(t)} r_n^{(t)} \tag{2.13}$$

$$= \boldsymbol{\omega}^T \boldsymbol{r}^{(t)}. \tag{2.14}$$

That is, the return of the portfolio is the linear combination of returns of its constituents, each weighted by $\omega_n^{(t)}$. The weights should satisfy for a fully invested portfolio

$$\mathbf{1}^T \boldsymbol{w} = 1, \ \forall t, \tag{2.15}$$

and $w_n < 0$ corresponds to a short position. Note that the formula for the return of a portfolio (2.13) is valid only in conjunction with simple returns, and does not apply with log-returns.

An *index* is a portfolio of stocks that is used to benchmark investments. Stocks are included in an index based on certain characteristics, such as their country or their industry. Also, different weighting schemes are possible. In an equal weighted index, each stock has the same weight. In a market capitalization weighted index, stocks are weighted proportionally to their market capitalization, i.e., their outstanding total market value. The latter scheme introduces a bias towards large companies.

It is very frequent for investors to choose their portfolio weights by performing some sort of optimization. Of great practical interest is the mean-variance portfolio, i.e., the portfolio with the highest expected return for a given level of risk Markowitz [1952]. Mathematically, in its simplest form, this portfolio is obtained by solving the following optimization problem

$$\boldsymbol{\omega^{(t)}} = \underset{\boldsymbol{\omega}}{\operatorname{argmax}}\, \boldsymbol{\omega}^T \mathbb{E}(\boldsymbol{r}^{(t)}) - \frac{\rho}{2} \boldsymbol{\omega}^T Cov(\boldsymbol{r}^{(t)}) \boldsymbol{\omega}, \tag{2.16}$$

where ρ is the coefficient of risk aversion. From the optimization perspective, the problem is extremely well understood and a lot of generalizations of this problem have a clear mathematical solution Boyd and Vandenberghe [2004]. Consider for example the introduction of optimization constraints, such as a maximum stock exposure. The key issue, as already pointed out by Markowitz [1952], is to obtain good estimates of the expected return and covariance matrix of returns. Several methods have been proposed to improve these estimates, for example, shrinkage Ledoit and Wolf [2003] or weighting the observations by an external variable Steude [2011]. The estimation problem is even more acute the portfolio optimization takes into account higher (co-)moments Harvey et al. [2010].

2.1.4 Quantitative strategies

With the problem of portfolio optimization, we have already briefly touched the subject of quantitative strategies, which we develop in this section. Quantitative strategies form a class of investment strategies that are developed by finance practitioners referred to as *quants*. Generally speaking, they have the following characteristics.

(i) Statistical: the strategy is based on an underlying statistical model that captures knowledge about the financial time-series.

(ii) Rule-based: the opening and closing of long/short positions depend on a set of trading rules, that are applied systematically, hence potentially by a computer. This is a natural protection for investors against human investment bias, documented by the theory of behavioral finance Hens and Bachmann [2009].

To evaluate a quantitative strategy, quants perform backtests. These are replications of the investment decisions faced by an investor using historical data. These tests have to be as realistic as possible and great care must be taken to avoid introducing future knowledge in past decisions. Of course, the investment decisions must be based solely on ex ante information. But there are more insidious forms of future knowledge. There is the problem of the choice of strategy itself. Indeed, because of their experience with financial markets, quants tend to select strategies that have performed well historically, but it is not clear whether they would have chosen them ex ante. Bossaerts and Hillion [1999] address this question, using three of the best predictors of stock returns documented in the literature (momentum, value and size). They conclude that an investor would not have been able to choose these strategies ex ante, although the particular choice of predictors introduces a bias towards the alternative hypothesis. Furthermore, another form of future knowledge is introduced when the hyperparameters of a backtest, for example, the estimation window or the order of the model, are determined by maximizing the performance evaluation metric over the entire dataset. Finally, consider the choice of performance metric of a backtest. Risk-adjusted return is the correct performance evaluation metric. Indeed it is misleading to use generated wealth or expected return of the strategy, as different strategies may have incurred different levels of risk. The *Sharpe ratio*, defined as the ratio between the expected return and the volatility of a strategy, is commonly used. However, risk-adjusted return is only a first approximation, and a myriad of other indicators are necessary to fully comprehend a strategy. For example, the performance of the strategy should be decomposed over several periods to test its stability or the amount of trading should be measured, to ensure that the profits of a strategy are not swallowed by its trading costs.

To conclude this review of financial concepts, Table 2.2 summarizes all important notions and relates them to signal processing concepts.

2.2 Review of existing signal processing applications in finance

In this section, we review two existing applications of signal processing in quantitative finance. The first one deals with the application of Kalman filtering to estimate factor models, the second one with the use of PCA to extract a series of orthogonal factors.

2.2.1 Factor models

Both applications we review deal with the general problem of factor modeling. Consider the evolution of the return of a stock, $r_n^{(t)}$. This evolution could be attributed

Name	Definition and related signal processing term
Backtest	Simulation of investment strategy based on historical market data. The goal is to replicate the behavior of an investor in the past as realistically as possible. In particular, all investment decisions must me based on information available before the decision is taken.
Index	Weighted average measure of change of a group of stocks. Stocks are grouped in an index according to a selection criterion such as the country or the sector the stock belongs to. Various weighting scheme are possible, such as weighting by market capitalization or equal weighting. signal processing (SP) term: weighted average.
Leverage	Method to amplify the return of a strategy by borrowing (resp. lending) money at a risk free rate to invest (resp. reduce investment) in a risky asset. SP term: amplifier.
Long	Type of position to bet on rising prices. Open the long position by buying the asset. Close the long position by selling it.
Momentum	Generic name to designate a class of investment strategies that bet on the presence of continuation in the return process. Past winners are the future winners and past loosers are likewise the future loosers.
Portfolio	Basket of stocks
Return	Measure of relative change in the price of a stock. SP term: simple difference.
Sharpe ratio	Measure of the return of an investment strategy per unit of risk. SP term: signal-to-noise ratio (SNR).
Short	Type of position to bet on falling prices. Open the short position by borrowing an asset to ones broker and selling it immediately. Close the short position by buying the asset back to replace the borrowed asset.
Stock	Share of ownership of a company. The holder is entitled a fraction of the profit of the company. Synonym: equities.
Volatility	Standard deviation of returns. The volatility is used as a risk measure. SP term: standard deviation.

Table 2.2: *Glossary of financial terms.*

to two distinct sources of variations: variations that are common to all stocks, called factors, and variations that are proper to a stock, called idiosyncratic returns. For example, it is sensible to assume that stocks belonging to the same industry have common sources of variations, because they evolve in the same economic environment, they have a similar customer base or they face similar costs of commodities, production, transport, etc. But stocks within an industry also evolve differently, for example as the reaction to company specific news. Mathematically this is translated in an stylized manner in a *linear factor model*

$$r_n^{(t)} = \alpha_n + \beta_{n,1} f_1^{(t)} + \ldots + \beta_{n,k} f_k^{(t)} + e_n^{(t)}, \tag{2.17}$$

where $f_k^{(t)}$ is the evolution of factor k at time t, $\beta_{n,k}$, the sensitivity (or loading) of stock n on factor k and $e_n^{(t)}$ the idiosyncratic return of stock n at time t. Stacking the returns of all stocks in a vector, we write in matrix format

$$\boldsymbol{r}^{(t)} = \boldsymbol{B}\boldsymbol{f}^{(t)} + \boldsymbol{e}^{(t)}, \tag{2.18}$$

where $\boldsymbol{f}^{(t)} = \left(1, f_1^{(t)}, \ldots, f_K^{(t)}\right)^T$.

Example 3. *Consider the following example of factor models from the asset pricing literature. In the capital asset pricing model (CAPM) Sharpe [1964], Lintner [1965], the market portfolio $r_m^{(t)}$ is the only source of common variations,*

$$r_n^{(t)} - r_f = \alpha_n + \beta_{n,market}\left(r_{market}^{(t)} - r_f\right) + e_n^{(t)}, \tag{2.19}$$

where r_f is the risk-free rate. $\beta_{n,market}$ is traditionally simply called the beta of a stock. The equation above is equivalent to the standard CAPM formulation if $\alpha_n = 0\ \forall n$. Moreover, the celebrated Fama-French three factors model is an extension of the CAPM Fama and French [1992]. It contains, beside the market factor, two factors called size and value,

$$r_n^{(t)} = \alpha_n + \beta_{n,market} r_{market}^{(t)} + \beta_{n,size} r_{size}^{(t)} + \beta_{n,value} r_{value}^{(t)} + e_n^{(t)}. \tag{2.20}$$

These additional factors are obtained by building the following long-short portfolios. The return of the size factors, $r_n^{(t)}$ is obtained by ranking stocks in the universe at time $t-1$ by their market capitalization, shorting stocks above the top decile and going long stocks below the bottom decile. The return of the value factors is obtained by ranking stocks in the universe at time $t-1$ by their ratio of book to the market value of equities and going long stocks above the top decile and going short stocks below the bottom decile. In both cases, equal weighting between stocks is used.

From a mathematical point of view, factor models allow reducing the dimension of a problem, as the number of stocks is typically large compared to the number of factors, $k \ll N$. Consider the example of the covariance matrix of stock returns. In general, it is parameterized by $N(N-1)/2$ values, as it is a symmetric matrix. Using the factor model given by (2.17), we can write

$$Cov\left(\boldsymbol{r}^{(t)}\right) = \boldsymbol{B}^T Cov\left(\boldsymbol{f}^{(t)}\right)\boldsymbol{B} + Cov\left(\boldsymbol{e}^{(t)}\right). \tag{2.21}$$

Then, the number of parameters drops dramatically. There are $k(k-1)/2$ parameters for the covariance matrix of factors $Cov\left(\boldsymbol{f}^{(t)}\right)$. The idiosyncratic returns are uncorrelated by definition, such that there are only N diagonal elements in the covariance $Cov\left(\boldsymbol{e}^{(t)}\right)$. The factor decomposition allows deriving from a small number of parameters a full rank covariance matrix of returns. Key is the introduction of the diagonal matrix of idiosyncratic variances. Without it, the covariance matrix of returns would be rank deficient, which poses problem, notably in subsequent portfolio optimization.

From an investment perspective, the factor decomposition can be applied in two different manners to build quantitative strategies. On the one hand, a portfolio manager can treat factors as sources of risk for which the holder is rewarded and the idiosyncratic returns, which are not rewarded, are diversified away by construction of a portfolio. The manager then takes a controlled exposure to certain factors, under certain investment constraints, e.g., a limited stock or factor exposure. On the other hand, a portfolio manager can take a long exposure of 1 in the asset and short $\beta_{n,k}$ in each factor k. He is then only exposed to the idiosyncratic component, which indicates whether a stock is currently over- or undervalued compared to its peers. This strategy is called market-neutral, as its evolution is independent of that of the factors. The timing of the exposure is done thanks to a statistical model, for example, a mean-reverting process Avellaneda and Lee [2010].

Of course, the problem of building a factor model remains. This is the topic of the coming sections. In both cases, we work with the Hilbert space of square summable random processes endowed with the scalar product

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \mathbb{E}\left(\left(\boldsymbol{x} - \mathbb{E}(\boldsymbol{x})\right)^T \left(\boldsymbol{y} - \mathbb{E}(\boldsymbol{y})\right) \right). \tag{2.22}$$

2.2.2 Kalman filtering for estimation of factor models

In this section, we review the use of the KF to estimate factor models, i.e., determine the loadings of the model given a set of observations.

Kalman filter

For the sake of completeness, let us start by reviewing the Kalman-Bucy filter in discrete time Moura [2005]. In a nutshell, this filter is used to estimate the state of a linear system from its observations. Its block diagram is represented in Figure 2.3. The system depends on a latent, i.e., not observable vector of states $\boldsymbol{x}^{(t)}$ whose evolution is described by the following linear, possibly time-varying, state equation

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{A}^{(t)}\boldsymbol{x}^{(\boldsymbol{t})} + \boldsymbol{B}^{(t)}\boldsymbol{u}^{(t)} + \boldsymbol{v}^{(\boldsymbol{t})}, \tag{2.23}$$

where $\boldsymbol{u}^{(t)}$ is a vector of exogeneous input and the model noise $\boldsymbol{v}^{(t)}$ is i.i.d., normally distributed with covariance matrix $\boldsymbol{\Sigma}_v$,

$$\boldsymbol{v}^{(t)} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_v\right). \tag{2.24}$$

The state is only observed through the following observation equation

$$\boldsymbol{y}^{(t)} = \boldsymbol{C}^{(t)}\boldsymbol{x}^{(t)} + \boldsymbol{w}^{(t)} \tag{2.25}$$

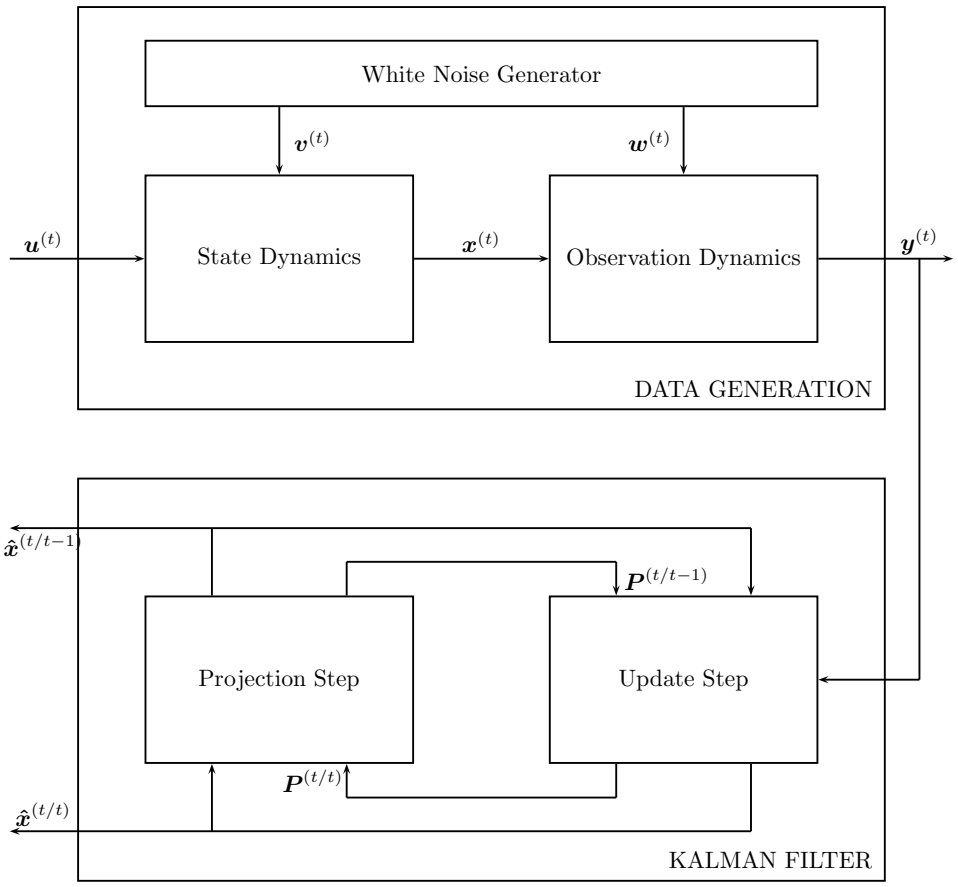


Figure 2.3: *Block diagram of the Kalman-Bucy filter. The data generating process is modeled as a linear state space model. The KF is decomposed in two successive steps: the projection step that projects the estimated state in the future and the update step that updates the state given the current observation.*

where $\boldsymbol{y}^{(t)}$ is the observations, and $\boldsymbol{w}^{(t)}$, the observation noise, is also i.i.d., normally distributed with covariance matrix $\boldsymbol{\Sigma}_w$,

$$\boldsymbol{w}^{(t)} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_w\right). \tag{2.26}$$

The goal is then to form estimates of the states of the system $\boldsymbol{x}^{(t)}$ based on the observations $\boldsymbol{y}^{(t)}$. Suppose that the system matrices $\boldsymbol{A}^{(t)}$, $\boldsymbol{B}^{(t)}$, $\boldsymbol{C}^{(t)}$ as well as the covariance matrices $\boldsymbol{\Sigma}_v$ and $\boldsymbol{\Sigma}_w$ are known. Let us introduce the following notations. Let

$$\hat{\boldsymbol{x}}^{(t/t-1)} = \mathbb{E}\left(\boldsymbol{x}^{(t)} / \boldsymbol{y}^{(t-1)}, \ldots\right) \tag{2.27}$$

be the estimate of the state given information up to time $t-1$. Likewise, let

$$\hat{\boldsymbol{x}}^{(t/t)} = \mathbb{E}\left(\boldsymbol{x}^{(t)}/\boldsymbol{y}^{(t)}, \dots\right) \tag{2.28}$$

be the estimate of the state at time t conditional on information up to time t. Moreover, let

$$\boldsymbol{P}^{(t/t-1)} = Cov\left(\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t/t-1)}/\boldsymbol{y}^{(t-1)}, \dots\right) \tag{2.29}$$

be the covariance of the error at time t conditional on information up to time $t-1$ and

$$\boldsymbol{P}^{(t/t)} = Cov\left(\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t/t)}/\boldsymbol{y}^{(t)}, \dots\right) \tag{2.30}$$

be the covariance of the error at time t conditional on information up to time t.

The KF is represented in Algorithm 2.1. For each observation, the knowledge of the state equation is used to project the state of the system from $t-1$ to t (projection step). Then, given the actual observations, the estimate of the state is updated (update phase). The update is proportional to the prediction error, $\boldsymbol{y}^{(t)} - \boldsymbol{C}^{(t)}\hat{\boldsymbol{x}}^{(t/t-1)}$, as well as a matrix $\boldsymbol{K}^{(t)}$ called Kalman gain. Finally, if the system matrices are unknown, they have to be estimated, for example by maximum likelihood. The likelihood function is computed for a given value of the system matrices from the estimate of the residuals of the filter and it is optimized using a numerical procedure. Details are described in Kim and Nelson [1998].

Algorithm 2.1 Kalman-Bucy filter.

1. Input: observations $\boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(T)}$ and exogeneous inputs $\boldsymbol{u}^{(1)}, \dots, \boldsymbol{u}^{(T)}$
2.
3. % For each observations
4. **for** $t = 1, \dots, T$ **do**
5.
6. % Projection step
7. $\hat{\boldsymbol{x}}^{(t/t-1)} \leftarrow \boldsymbol{A}^{(t)}\hat{\boldsymbol{x}}^{(t-1/t-1)} + \boldsymbol{B}^{(t)}\boldsymbol{u}^{(t)}$
8. $\boldsymbol{P}^{(t/t-1)} \leftarrow \left(\boldsymbol{A}^{(t)}\right)^T \boldsymbol{P}^{(t-1/t-1)}\boldsymbol{A}^{(t)} + \boldsymbol{\Sigma}_v$
9.
10. % Update step
11. $\boldsymbol{K}^{(t)} \leftarrow \boldsymbol{P}^{(t/t-1)}\left(\boldsymbol{C}^{(t)}\right)^T \left(\boldsymbol{C}^{(t)}\boldsymbol{P}^{(t/t-1)}\left(\boldsymbol{C}^{(t)}\right)^T + \boldsymbol{\Sigma}_w\right)^{-1}$
12. $\hat{\boldsymbol{x}}^{(t/t)} \leftarrow \hat{\boldsymbol{x}}^{(t/t-1)} + \boldsymbol{K}^{(t)}\left(\boldsymbol{y}^{(t)} - \boldsymbol{C}^{(t)}\hat{\boldsymbol{x}}^{(t/t-1)}\right)$
13. $\boldsymbol{P}^{(t/t)} \leftarrow \left(\boldsymbol{I} - \boldsymbol{K}^{(t)}\boldsymbol{C}^{(t)}\right)\boldsymbol{P}^{(t/t-1)}$
14. **end for**

Estimation of factor models

Consider again (2.17). One can take as factors observed portfolios or macroeconomic variables suggested by finance theory. The problem is then to estimate the loadings of the model. Of course, is is possible to use OLS regression to find the least-squares estimator. But the KF can be applied to perform this task, with certain

advantages. The central idea is to treat the loadings as the state of the system. Let us stack all loadings of stock n in a vector

$$\boldsymbol{x}_n^{(t)} \equiv \begin{pmatrix} \alpha_n \\ \beta_{n,1} \\ \vdots \\ \beta_{n,k} \end{pmatrix} \tag{2.31}$$

and subsequently loadings for all stocks

$$\boldsymbol{x}^{(t)} \equiv \begin{pmatrix} \boldsymbol{x}_1^{(t)} \\ \vdots \\ \boldsymbol{x}_n^{(t)} \\ \vdots \\ \boldsymbol{x}_N^{(t)} \end{pmatrix}. \tag{2.32}$$

The factor model (2.17) can be rewritten as

$$\underbrace{\begin{pmatrix} r_1^{(t)} \\ \vdots \\ r_n^{(t)} \\ \vdots \\ r_N^{(t)} \end{pmatrix}}_{\boldsymbol{y}^{(t)}} = \underbrace{\begin{pmatrix} \left(\boldsymbol{f}^{(t)}\right)^T & 0 & \dots & & 0 \\ & & \vdots & & \\ 0 & \dots & \left(\boldsymbol{f}^{(t)}\right)^T & \dots & 0 \\ & & \vdots & & \\ 0 & & \dots & 0 & \left(\boldsymbol{f}^{(t)}\right)^T \end{pmatrix}}_{\boldsymbol{C}^{(t)}} \underbrace{\begin{pmatrix} \boldsymbol{x}_1^{(t)} \\ \vdots \\ \boldsymbol{x}_n^{(t)} \\ \vdots \\ \boldsymbol{x}_N^{(t)} \end{pmatrix}}_{\boldsymbol{x}^{(t)}} + \underbrace{\begin{pmatrix} e_1^{(t)} \\ \vdots \\ e_n^{(t)} \\ \vdots \\ e_N^{(t)} \end{pmatrix}}_{\boldsymbol{w}^{(t)}} \tag{2.33}$$

Then, in combination with a state equation describing the evolution of loadings (see examples below), we can use the KF to obtain $\hat{\boldsymbol{x}}^{(t/t-1)}$ and $\hat{\boldsymbol{x}}^{(t/t-1)}$ which correspond, in this case, to the conditional estimates of the loadings based on information up to time $t-1$ and t, respectively.

Example 4. *In this example, the evolution of the loadings is given by the following state equation*

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)}. \tag{2.34}$$

Let us proceed to show that the estimates of the KF correspond to that obtained by GLS. Because of the specific, degenerated form of the state equation

$$\hat{\boldsymbol{x}}^{(t+1/t)} = \hat{\boldsymbol{x}}^{(t/t)} \tag{2.35}$$

and

$$\boldsymbol{P}^{(t+1/t)} = \boldsymbol{P}^{(t/t)}. \tag{2.36}$$

Suppose we have already computed the estimates at time t, $\hat{\boldsymbol{x}}^{(t/t)}$ and $\boldsymbol{P}^{(t/t)}$. Then,

starting from the GLS formula, the covariance of the estimates is given by

$$\boldsymbol{P}^{(t+1/t+1)} = \tag{2.37}$$

$$\left(\sum_{t'=1}^{t+1}\left(\boldsymbol{C}^{(t')}\right)^T\boldsymbol{\Sigma}_w^{-1}\boldsymbol{C}^{(t')}\right)^{-1} = \tag{2.38}$$

$$\left(\sum_{t'=1}^{t}\left(\boldsymbol{C}^{(t')}\right)^T\boldsymbol{\Sigma}_w^{-1}\boldsymbol{C}^{(t')} + \left(\boldsymbol{C}^{(t+1)}\right)^T\boldsymbol{\Sigma}_w^{-1}\boldsymbol{C}^{(t+1)}\right)^{-1} \overset{(a)}{=} \tag{2.39}$$

$$\boldsymbol{P}^{(t/t)} -$$

$$\boldsymbol{P}^{(t/t)}\left(\boldsymbol{C}^{(t+1)}\right)^T\left(\boldsymbol{\Sigma}_w + \boldsymbol{C}^{(t+1)}\boldsymbol{P}^{(t/t)}\left(\boldsymbol{C}^{(t+1)}\right)^T\right)^{-1}\boldsymbol{C}^{(t+1)}\boldsymbol{P}^{(t/t)} \overset{(b)}{=} \tag{2.40}$$

$$\left(\boldsymbol{I} - \boldsymbol{K}^{(t+1)}\boldsymbol{C}^{(t+1)}\right)\boldsymbol{P}^{(t/t)} \overset{(c)}{=} \tag{2.41}$$

$$\left(\boldsymbol{I} - \boldsymbol{K}^{(t+1)}\boldsymbol{C}^{(t+1)}\right)\boldsymbol{P}^{(t+1/t)} \tag{2.42}$$

In the above derivation, (a) results from the application of the matrix inversion lemma Woodbury [1950] and from the definition of the GLS formula for $\boldsymbol{P}^{(t/t)}$. (b) is obtained by identifying the Kalman gain term and (c) simply follows from (2.35). Similarly, we can derive the update rule for the estimates, starting from the GLS formula.

$$\hat{\boldsymbol{x}}^{(t+1/t+1)} = \tag{2.43}$$

$$\boldsymbol{P}^{(t+1/t+1)}\left(\sum_{t'=1}^{t+1}\left(\boldsymbol{C}^{(t')}\right)^T\boldsymbol{\Sigma}_w^{-1}\boldsymbol{y}^{(t')}\right) = \tag{2.44}$$

$$\left(\boldsymbol{I} - \boldsymbol{K}^{(t+1)}\boldsymbol{C}^{(t+1)}\right)\times$$

$$\boldsymbol{P}^{(t/t)}\left(\sum_{t'=1}^{t}\left(\boldsymbol{C}^{(t')}\right)^T\boldsymbol{\Sigma}_w^{-1}\boldsymbol{y}^{(t')} + \left(\boldsymbol{C}^{(t+1)}\right)^T\boldsymbol{\Sigma}_w^{-1}\boldsymbol{y}^{(t+1)}\right) = \tag{2.45}$$

$$\left(\boldsymbol{I} - \boldsymbol{K}^{(t+1)}\boldsymbol{C}^{(t+1)}\right)\hat{\boldsymbol{x}}^{(t/t)} +$$

$$\left(\boldsymbol{I} - \boldsymbol{K}^{(t+1)}\boldsymbol{C}^{(t+1)}\right)\boldsymbol{P}^{(t/t)}\left(\boldsymbol{C}^{(t+1)}\right)^T\boldsymbol{\Sigma}_w^{-1}\boldsymbol{y}^{(t+1)} \tag{2.46}$$

We have simply used (2.41) and recognized the GLS formula for $\hat{\boldsymbol{x}}^{(t/t)}$. Let us manipulate the matrix term of the RHS

$$\left(\boldsymbol{I} - \boldsymbol{K}^{(t+1)}\boldsymbol{C}^{(t+1)}\right)\boldsymbol{P}^{(t/t)}\left(\boldsymbol{C}^{(t+1)}\right)^T = \tag{2.47}$$

$$\boldsymbol{P}^{(t/t)}\left(\boldsymbol{C}^{(t+1)}\right)^T\times$$

$$\left(\boldsymbol{I} - \left(\boldsymbol{C}^{(t+1)}\boldsymbol{P}^{(t/t)}\left(\boldsymbol{C}^{(t+1)}\right)^T + \boldsymbol{\Sigma}_w\right)^{-1}\boldsymbol{C}^{(t+1)}\boldsymbol{P}^{(t/t)}\left(\boldsymbol{C}^{(t+1)}\right)^T\right) = \tag{2.48}$$

$$\boldsymbol{K}^{(t+1)}\left(\boldsymbol{C}^{(t+1)}\boldsymbol{P}^{(t/t)}\left(\boldsymbol{C}^{(t+1)}\right)^T + \boldsymbol{\Sigma}_w - \boldsymbol{C}^{(t+1)}\boldsymbol{P}^{(t/t)}\left(\boldsymbol{C}^{(t+1)}\right)^T\right) = \tag{2.49}$$

$$\boldsymbol{K}^{(t+1)}\boldsymbol{\Sigma}_w \tag{2.50}$$

Therefore,

$$\hat{\boldsymbol{x}}^{(t+1/t+1)} = \hat{\boldsymbol{x}}^{(t/t)} + \boldsymbol{K}^{(t+1)}\left(\boldsymbol{y}^{(t+1)} - \boldsymbol{C}^{(t+1)}\hat{\boldsymbol{x}}^{(t/t)}\right) \tag{2.51}$$

This concludes the proof that the estimates of the KF correspond in this case to the GLS estimates. The KF rule presents the advantage of being an online update rule which does not require to recompute and invert matrix $\boldsymbol{P}^{(t+1/t+1)}$ for every new observation. The use of the KF is therefore computationally more efficient.

Example 5. *Consider also the following example taken from Jay et al. [2011]. The eight factors model of Fung and Ksieh [2002] is used to track the return of a hedge fund. The loadings evolve according to the following state equation*

$$\boldsymbol{x}^{(t)} = \boldsymbol{x}^{(t-1)} + \boldsymbol{v}^{(t)} \tag{2.52}$$

where the noise $\boldsymbol{v}^{(t)}$ *is uncorrelated and identical across stocks*

$$\boldsymbol{v}^{(t)} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \boldsymbol{\Sigma} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \boldsymbol{\Sigma} \end{pmatrix}\right) \tag{2.53}$$

Figure 2.4 compares the performance of the KF and the OLS method by measuring the relative error , i.e., the absolute difference between the observed and estimated returns. The improvement is by two orders of magnitude.

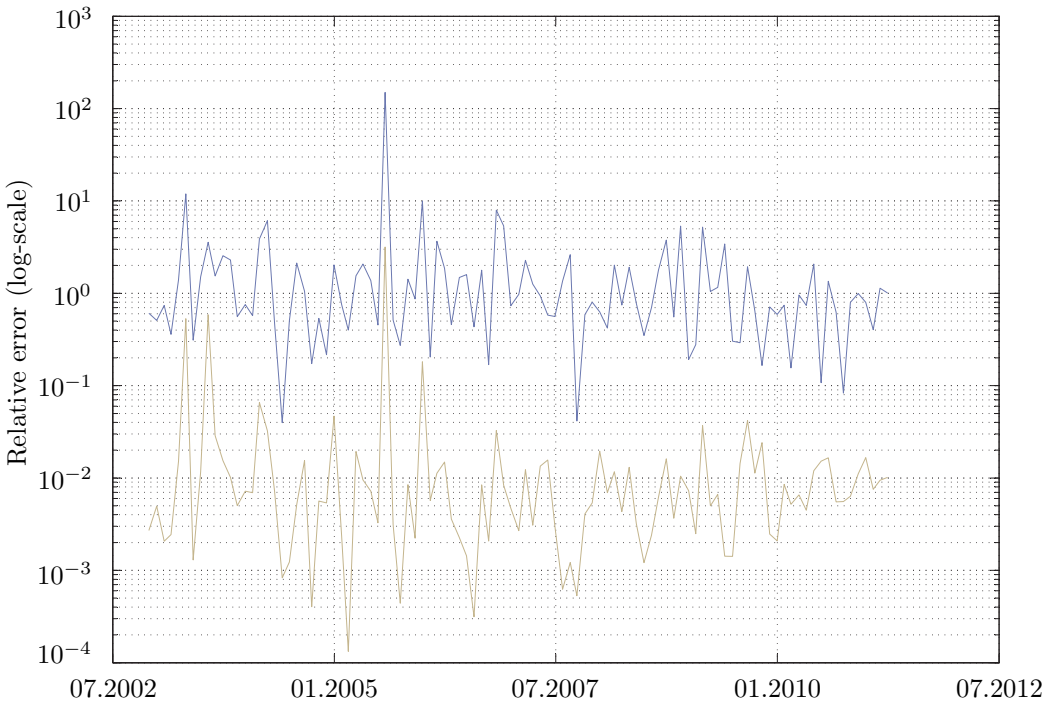


Figure 2.4: *Comparaison of the relative error, i.e., the absolute difference between observed and estimated returns, using OLS (blue) and the KF (gold). The improvement is by two orders of magnitude.*

2.2.3 PCA for extraction of orthogonal factors

In the previous section, the factors were observed quantities, either observed portfolios or macroeconomic variables. The problem was then to estimate the models,

i.e., determining the value of the loadings given a set of observations. In this section, we explore another approach called PCA which allows simultaneously finding a set of orthogonal factors and their associated loadings.

PCA procedure

We start by reviewing the PCA procedure also called Karuhen-Loewe transform in signal processing Sbaiz and Ridolfi [2006]. It builds a series of orthogonal factors such that successive factors have a decreasing explanatory power. Orthogonal means uncorrelated in this case, due to the definition of the relevant scalar product in (2.22). The details of the procedure are presented in Algorithm 2.2. Given a set of observations, PCA starts by computing an estimate of the correlation matrix. Working with correlation instead of covariance allows working with assets that have volatility of different scales. Also, other estimates than the sample correlation can be used. Then, the eigenvalue decomposition is used to find an orthogonal basis whose elements corresponds to the columns of matrix $\boldsymbol{U}$. The notation reflects the unitary property of the matrix

$$\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}, \tag{2.54}$$

since the correlation matrix is positive semi-definite and symmetric. Moreover, Figure 2.5 represents the first eigenvalues of the correlation matrix, ordered in descending order. If the system were well conditioned, there would exist a sharp drop between significant positive eigenvalues and zero nonsignificant eigenvalues. Because of the presence of noise, the boundary between significant and nonsignificant eigenvalues is blurred and the system is said to be ill-conditioned. It is not entirely clear then how many factors we should retain. The quantity

$$\frac{D_{n,n}}{\sum_{n=1}^{N} D_{n,n}} \tag{2.55}$$

can be interpreted as the percentage of total variance explained by factor k. One choice consists in using a fixed number of factors K. Figure 2.6a represents the evolution over time of the variance explained by a set of 15 factors for the US equities market. We observe that a fixed number of factors explain a variable portion of the total variance. Another choice, described in Algorithm 2.2, consists in choosing a number of factors so as to explain a fixed proportion of the total variance. Figure 2.6b gives the evolution over time of number of factors necessary to explain a 75% of the total variance of the US equities market. Finally, factors are obtained by projecting returns onto this set of k orthogonal vectors.

Economic interpretation of factors extracted by PCA

PCA allows a simultaneous identification of the factors and their loadings. However, it comes with the disadvantage that the resulting factors are sometimes hard to interpret. However, for financial markets data and the first factors, the following results are well established Avellaneda and Lee [2010]. Firstly, Figure 2.7 represents the loadings of all US stocks on the first factor. They are ordered in decreasing orders and stocks are labeled with their industry code. We observe that all loadings are positive. Also, Figure 2.8 compares the wealth generated by investing in the first factor

Algorithm 2.2 Principal Component Analysis.

1. Input: $\boldsymbol{r}^{(t)}, t = 1, \ldots, T$ as series of observations of returns, ς, the target explained variance
2. % Computation of the sample correlation matrix
3. **for** $n = 1, \ldots, N$ **do**
4. $\quad \bar{r}_n \leftarrow \frac{1}{N} \sum_{t=1}^{T} r_n^{(t)}$
5. $\quad \sigma_n \leftarrow \sqrt{\frac{1}{N-1} \sum_{t=1}^{T} \left(r_n^{(t)} - \mu_n \right)^2}$
6. $\quad \widetilde{r}_n^{(t)} \leftarrow \frac{r_n^{(t)} - \bar{r}_n}{\sigma_n}, \; t = 1, \ldots, T$
7. **end for**
8. $\boldsymbol{\Sigma} \leftarrow \frac{1}{T-1} \left(\widetilde{\boldsymbol{r}}^{(t)} \right)^T \widetilde{\boldsymbol{r}}^{(t)}$
9.
10. % Eigenvalue decomposition
11. $[\boldsymbol{U}, \boldsymbol{D}] \leftarrow eig(\boldsymbol{\Sigma}), \; s.t. \; \boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{\boldsymbol{T}}$
12.
13. % Choice of eigenvalue
14. Choose k as the minimum such that $\sum_{n=1}^{k} \frac{D_{n,n}}{\sum_{n=1}^{N} D_{n,n}} \geq \varsigma$
15. % Computations of factor
16. $f_k^{(t)} \leftarrow \boldsymbol{U}_{.,k}^T \widetilde{\boldsymbol{r}}^{(t)}$

and the market capitalization index. The two curves are extremely correlated and the minor difference is explained by a difference in volatility: the market capitalization index has a lower volatility, as it is biased towards large stocks. Thus, we associate the first factor with the market portfolio, like the CAPM or Fama-French's three factors model. Figure 2.9 and 2.10 represents the loadings of stocks on the second and third factors, respectively. Similarly, they are displayed in decreasing order and stocks are labeled with their industry code. Note that certain loadings have to be negative to satisfy the orthogonality condition. Moreover, we observe that stocks that load a lot on the same factor tend to belong to the same industry. This phenomenon, called coherence, corresponds to the intuition that stocks in the same industry have a similar evolution. Coherence disappears progressively as we observe loadings associated with subsequent, smaller eigenvalues.

2.2.4 Shortcomings of existing applications of SP to finance

We have just reviewed two applications of signal processing in quantitative finance, namely the estimation of factor models using the KF and the extraction of orthogonal factors using PCA. Using these examples, but also other references, we proceed in this section to illustrate the limitations of existing approaches.

Simple toolbox applications

In most existing applications, signal processing is regarded as a set of methods and algorithms that are applied to solve problems in quantitative finance. This is legitimate, given the success that some of these techniques have had when applied

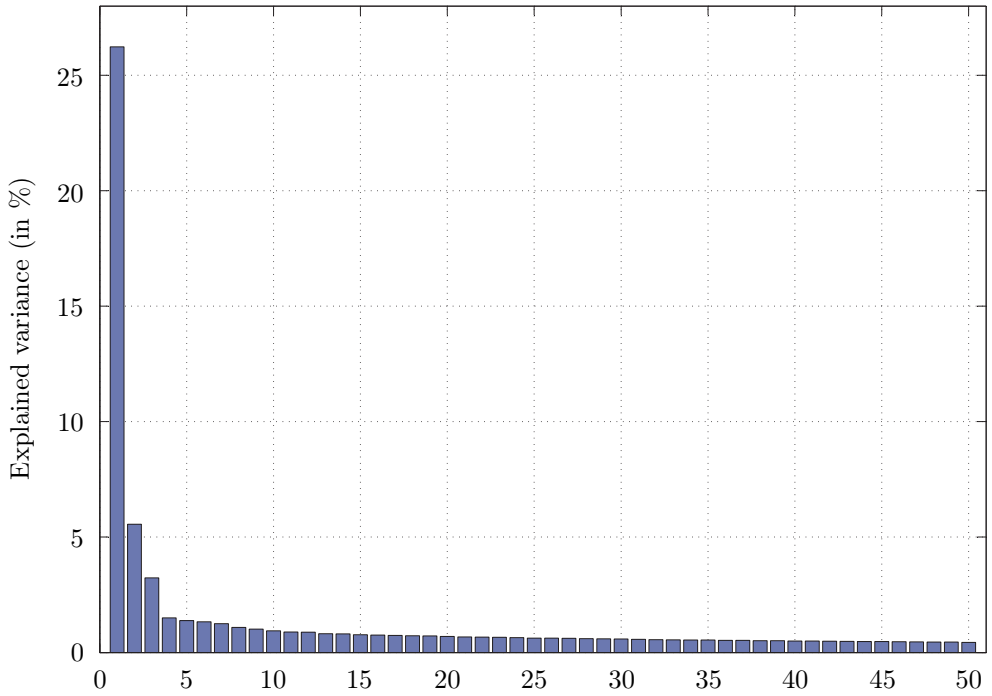


Figure 2.5: *Eigenvalue of the correlation matrix expressed as percentage of explained variance. Correlation is estimated from daily returns of US stocks, from 01.05.2006 until 01.05.2007.*

in other fields Oppenheim et al. [1999]. However, certain of these techniques are not adapted to the specificity of the financial signals and the constraints of investment management. Consider the example of adaptive filtering Sayed [2003]. This technique was developed in the context of audio applications where computational power is limited, for example due to the limited power available in the hardware implementation. Thus, the solution of the normal equation is not obtained directly by matrix inversion but approximated by an iteration of the steepest descent algorithm, one iteration per observation. And to increase the adaptation speed, the signal processing expert increases the sampling frequency of the process. In financial applications with a relatively low frequency of trading, computational power is not anymore the limiting factor as significantly more time is available to compute an update per sample. But correlations in the filter inputs, and characteristics of the noise process such as its heteroskedasticity and SNR pose other challenges. This simple example illustrates why we need to move beyond toolbox applications of signal processing techniques, as they have been developed with different sets of assumptions and design constraints in mind. This has already been done in the history of signal processing. Think for example of specific transforms and lossy compression scheme that takes into account the properties of the human visual system to compress images Vetterli and Kovacevic [1995].

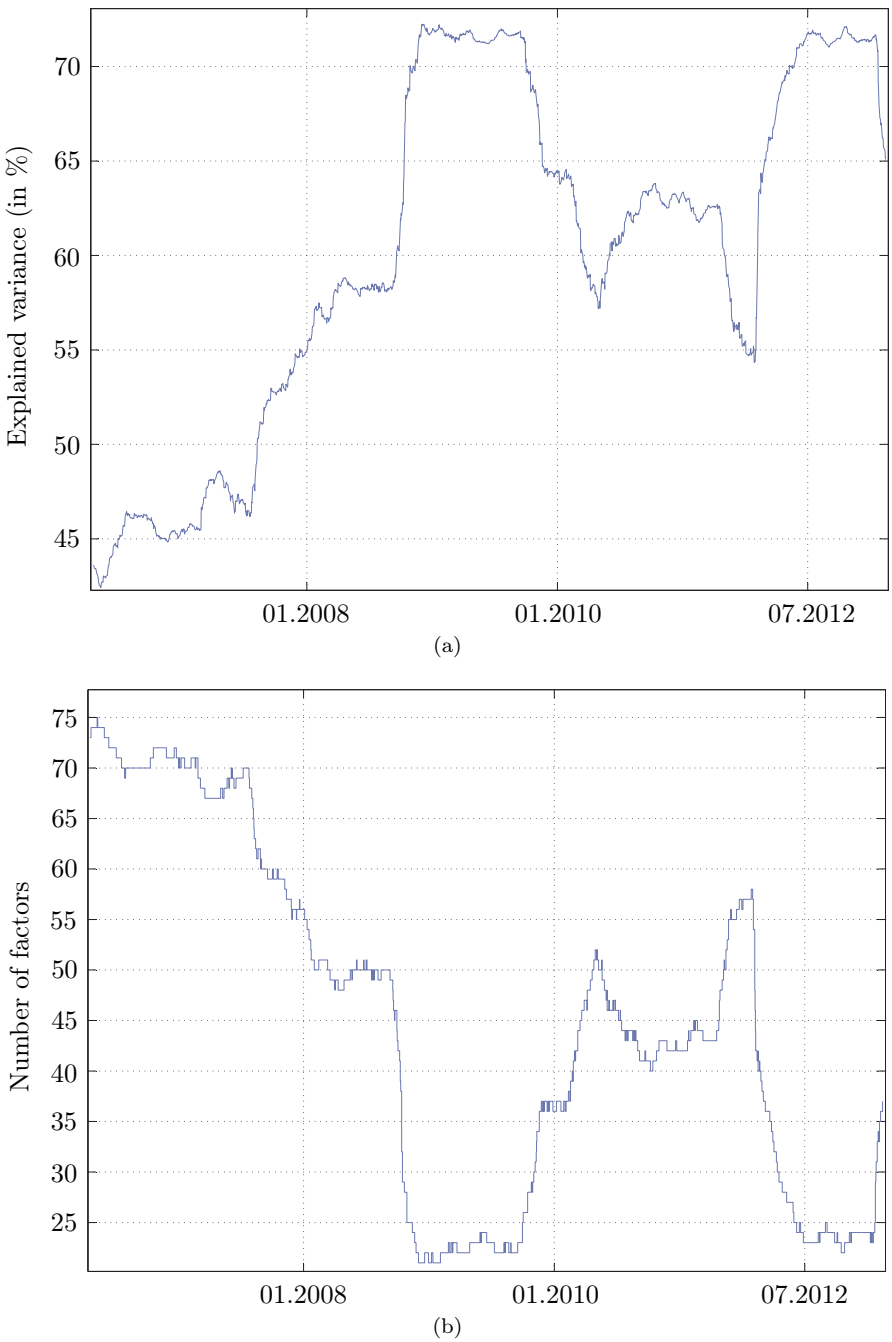


Figure 2.6: *(a) Percentage of explained variance using 15 factors. (b) Number of factors necessary to explain 75% of the total variance. Correlation in both cases is estimated from daily returns of US stocks, using a 1-year lookback window.*

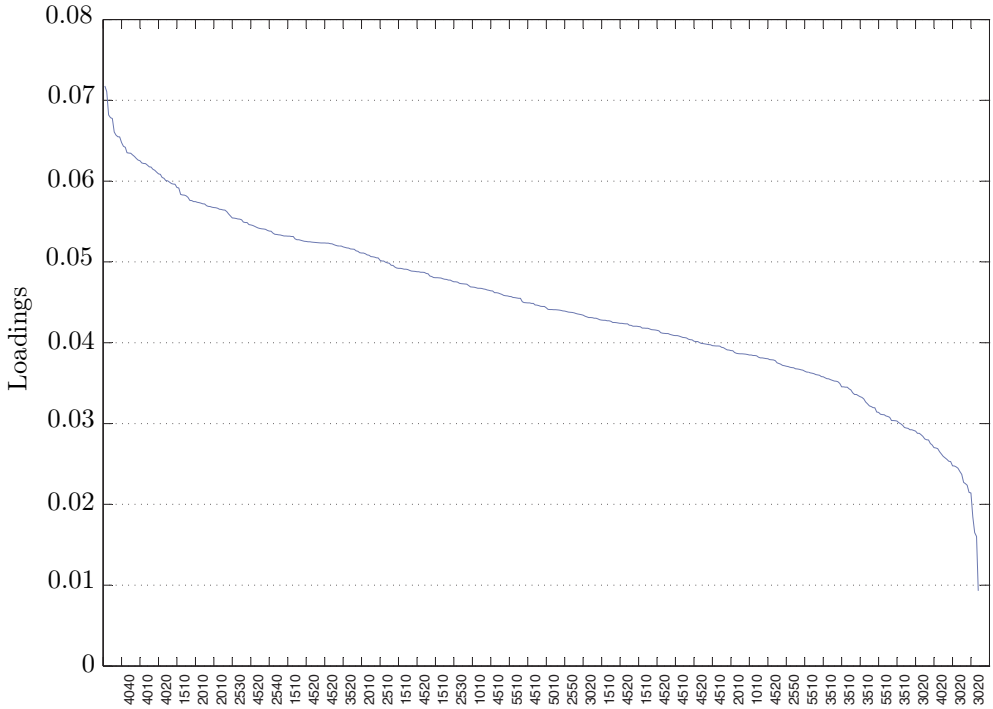


Figure 2.7: *Loadings on the first factor ordered in decreasing order, compared with the industry the stock belongs to. Correlation is estimated from daily returns of US stocks, from 01.05.2006 until 01.05.2007.*

Belief in a unique model as an absolute truth

We have seen that a quantitative strategy relies on an underlying statistical model that represents a view on the time-series of interest, typically some function of the returns. For example, quants use a factor model with a fixed set of factors to reduce the dimension of the cross-section of returns, as seen in Section 2.2.1. Momentum type strategy is another example where a fixed portion of past returns is used as predictors for future returns Jegadeesh and Titman [1993]. More specifically, this strategy consists in going long (resp. short) on the assets that have performed well (resp. poorly) over the last 12 months of data, ignoring the last monthś observation. There are no reasons to believe this specific model should hold and despite the lack of economic understanding of why this strategy delivers outperformance, it is widely used in practice Novy-Marx [2012]. There are two problems with this approach.

The first one is the fixed structure of the model. There are overwhelming evidences that a fixed set of factors or predictors does not capture well the cross-section of stock returns. For example, we have already seen in PCA that a fixed number of orthogonal factors explains a variable portion of the total variance. Moreover, Cooper et al. [2005] review a set 48 in-sample predictors of stock returns documented in the literature. They apply to each of them the same out-of-sample test with the same dataset consisting of more than 30 years of monthly observations. They conclude on the absence of

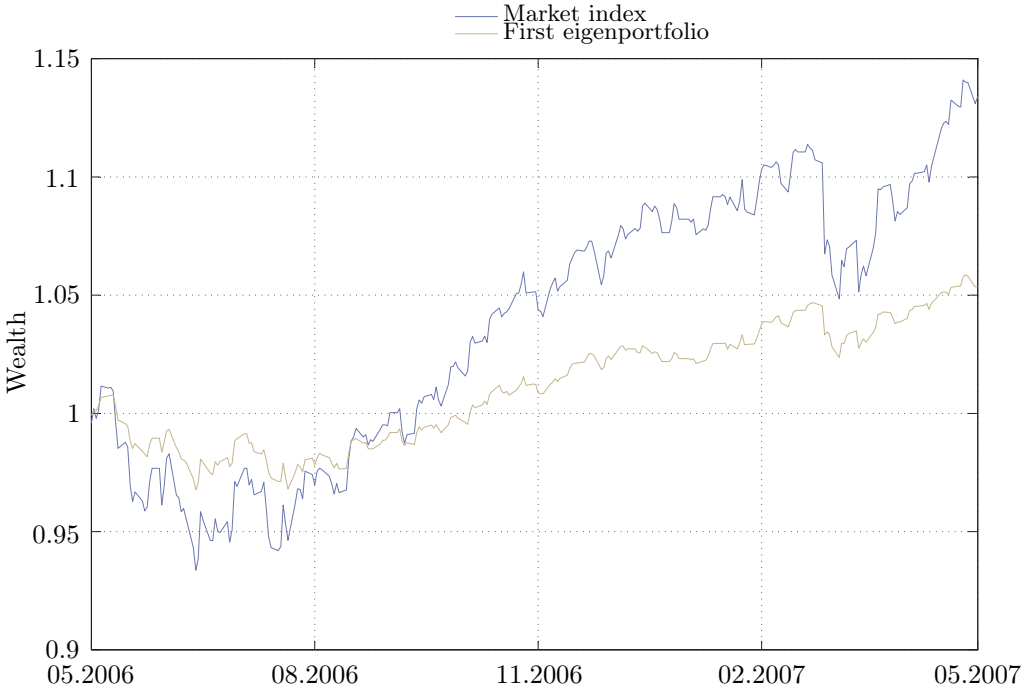


Figure 2.8: *Evolution of the wealth generated when investing in the first factor compared with the evolution of the market capitalization weighted index.*

out-of-sample predictability for all 48 predictors, even though each was documented to have significant in-sample predictability. Likewise, Bossaerts and Hillion [1999] conduct an out-of-sample study, where model selection criteria are used to select the set of factors among the three best predictors of equities premium (momentum, size and value). Their strategy fails to deliver any outperformance compared to indexing. In both cases, this failure is the symptom of the nonstationarity present in the data. The latter is interpreted as the consequence of learning in financial markets, which act as a feedback loop in the system and modifies it permanently Bossaerts and Hillion [1999]. In summary, we should not be using a single fixed model, but allows for certain adaptation among different models, including the model of order 0.

The second problem resides in the way we treat the model under consideration. This is almost a philosophical question, but it is important since it serves as foundation for our approach. Quants typically assume the existence of a true distribution according to which data are distributed. And a model is a more or less accurate simplification of this true distribution. This approach is certainty influenced by physics, where it is sensible to assume the existence of laws, albeit complex, governing the system. In finance, this is unlikely to be the case and we should therefore not base our approach on this assumption.

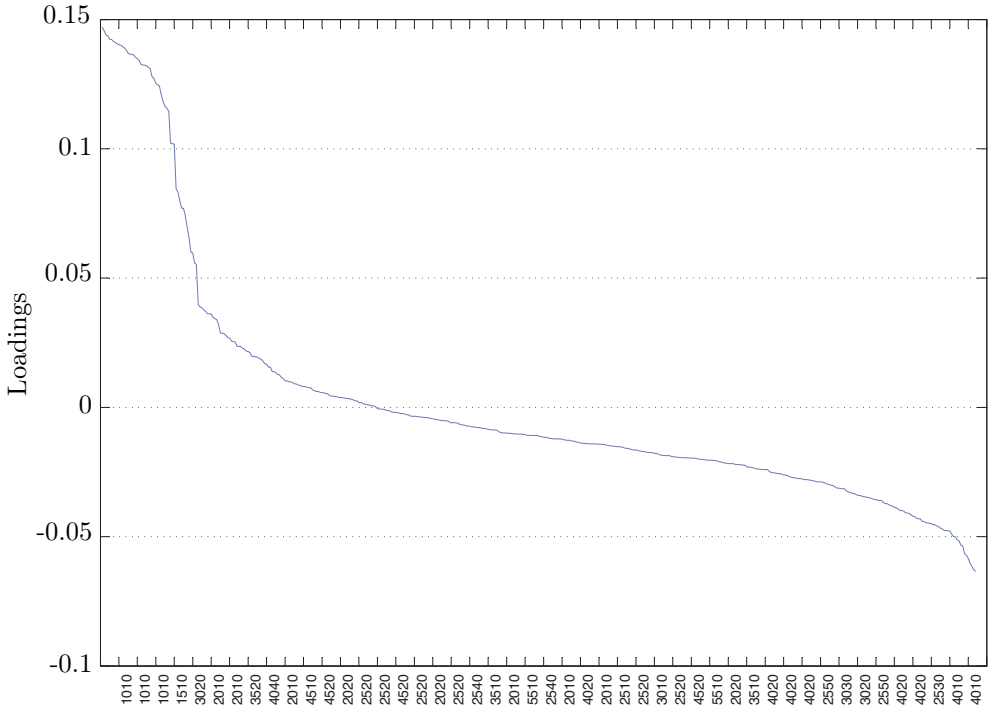


Figure 2.9: *Loadings on the second factor ordered in decreasing order, compared with the industry the stock belongs to. Correlation is estimated from daily returns of US stocks, from 01.05.2006 until 01.05.2007.*

Dimensionality and complexity of the model

Another problem recurrent in finance is the question of model dimensionality and complexity. For example, consider the problem of choosing the number of factors to include in a linear factor model (2.17). In PCA we have presented a procedure to select the number of factors so as to explain a fixed proportion of the total variance. But with this procedure, we run the risk of overfitting, by including factors corresponding to noise, which goes against the idea of building a factor model. Somehow we would like to penalize the additional complexity that represents the addition of another factor and accept it only if the improvement in terms of explained variance matches at least the additional complexity. Existing model selection criteria, such as the well-known Bayesian information criterion (BIC) and Akaike's information criterion (AIC), offer a partial answer, but, as we will see in the coming chapters, suffer from certain limitations. Quite interestingly, in both BIC and AIC, the complexity is proportional to the dimensionality of the model. But, as we will see with other model selection criteria, this should not be the case and two models with the same number of parameters may have a different complexity.

The problem of dimensionality and the so-called curse of dimensionality is particularly apparent in certain studies. For example, Jondeau [2010] develops an extension of the GARCH model to better capture the asymmetric behavior in the tail of the

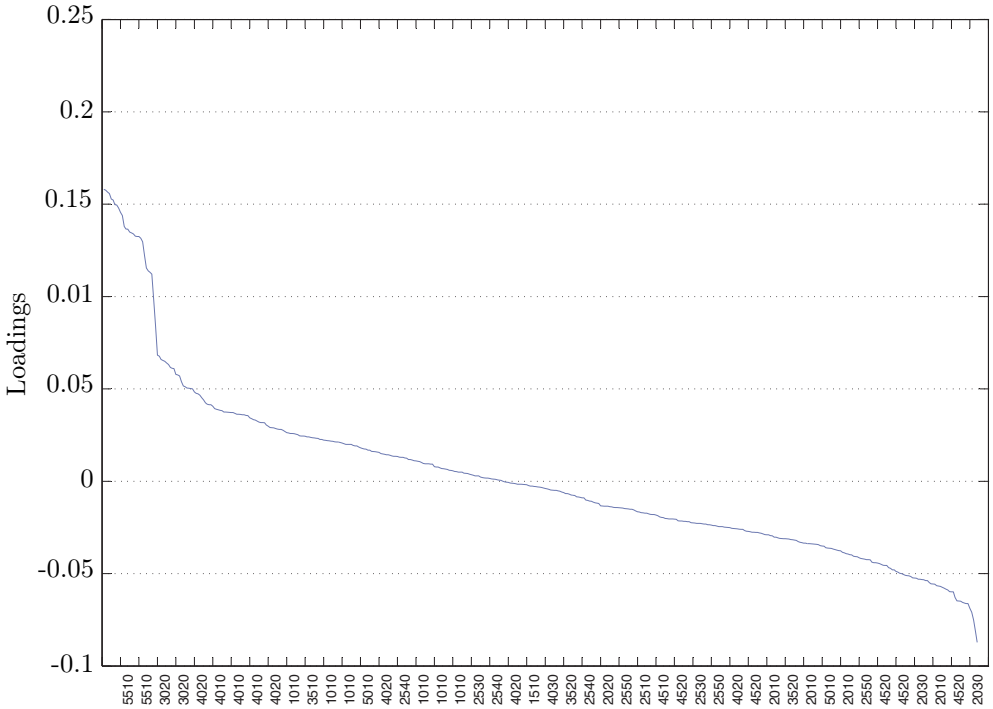


Figure 2.10: *Loadings on third factor ordered in decreasing order, compared with the industry the stock belongs to. Correlation is estimated from daily returns of US stocks, from 01.05.2006 until 01.05.2007.*

return distribution. But this comes at the price of introducing additional parameters controlling the behavior of the tail. Of course, the general principle of maximum likelihood could (and is) applied to infer from a set of observations the value of these parameters. But the larger the dimension of the parameters space, the larger the set of observations necessary to estimate it, which goes against the idea of handling non-stationarity. In summary, the curse of dimensionality limits greatly the applicability of these models..

Incorrect handling of stationarity

The problem of nonstationarity is pervasive in all financial modeling problems. But this is typically bypassed in one way or another, as we detail below. This is certainly because most mathematical and signal processing techniques rely on some form of stationarity in the data and little is known otherwise.

One way of bypassing the nonstationarity problem is to assume some form of local stationarity. That is, the system stays constant over a certain portion of the most recent observations and is estimated using these data. This is the so-called *windowing* approach. We have seen several examples across this chapter. The conditional volatility estimates is the simplest one. In applications of the Kalman filter, the parameters of the system have to be estimated and this is done by maximization of the likelihood

function over a window of observations. In PCA, local stationarity is implicitly assumed when computing the correlation matrix over a fixed window of observations. The problem with local stationarity is the choice of the appropriate window of data. A short window leads to quick adaptation but the resulting estimates are more sensitive to noise. There is thus a natural trade-off between the speed of adaptation and the noise in the solution. More worryingly, one common practice is to set the window size as a hyperparameter of the backtest and fine-tuned it based on observed data. This corresponds to the introduction of future knowledge in the backtest. Also, this is not suitable to handle abrupt changes in the dynamics of the system. During a stable market phase, for example the bull market between 2003 and 2007, this procedure will tend to use larger window size, for example, five years of monthly data. Then, when the market jumps to a crisis mode, the estimation of the model still use a large portion of observations that are not anymore relevant for the new market conditions. Of course, managers will adapt their window size over time, and this is one of the most often reported change in quantitative portfolio management by funds in 2009 after the 2008 financial crisis Shari [2011]. But this process is prone to human bias, and an unbiased and automated adaptive solution is highly desirable.

Another way to bypass the problem of nonstationarity is to model the underlying switches between different regimes of the system. For example, in state space model with Markov switching, a Markov process governs the switching between state space equations with different system matrices Kim and Nelson [1998]. With this approach, we come back to the convenient stationarity situation and the maximum likelihood principle is used to find the parameters of the models. This comes at the price of introducing an additional number of parameters, such that we face again the curse of dimensionality problem. For example with a Markov process of order 1 that switches between X state, we have already at least X parameters for each state and $X \times (X-1)$ parameters of the matrix of transition probabilities of the Markov switching process. This grows quickly to unpractical numbers for models with more than three states. Beyond the problem of dimensionality, our experience with Markov processes shows that their dynamics is not suitable to represent processes involving humans. The latter have usually a longer memory than just the last observation, as in Markov process of order 1. As a result, the maximum likelihood estimation procedure can select a system that switches often, whereas a system with less switches and longer periods between switches better reflects reality. We somehow would like to introduce a penalty for these excessive switches, but the theory of Markov switching processes gives little guidance on how to achieve this, even in an ad hoc manner.

Summary

In this chapter, we have first reviewed financial concepts necessary for the understanding of this thesis. (Log-)return is the key quantity of interest for investors. The risk of an investment is imperfectly measured by the volatility, i.e., the standard deviation of return, and the Sharpe ratio, defined as the ratio between expected return and volatility, is used as a key performance metric to compare investment strategies. We have also related financial concepts to SP ones. In particular, the Sharpe ratio corresponds to the notion of SNR, a standard performance metric in SP. We have also reviewed two applications of SP in quantitative finance. The first one deals with the estimation of a factor model using the KF. Through two concrete examples, we have illustrated that the KF offers not only better estimates of the loadings, compared to standard OLS, but also a computationally more efficient online update rule. The second application deals with the use of PCA to extract a series of uncorrelated factors. The first extracted factor is associated to the market index, a standard factor already included in factor models from the asset pricing literature, such as the CAPM. The next factors are associated industry indices, as suggested by the phenomenon of coherence. Finally, we have also illustrated the limitations of the current approaches. Two problems are particularly important for the remainder of this thesis. The first one concerns the fixed structure of the model that is used to model the data, although a certain number of evidences support more adaptive model structures. The second one is the improper handling of the nonstationarity inherent to financial signals, and, more specifically, the limitations of local windowing or Markov switching processes. Solutions to these problems are the topic of the coming chapters.

Chapter 3

Financial Markets Under Information Asymmetry

> "Private information is practically the source of every large modern fortune."
>
> *An Ideal Husband*
> Oscar Wilde

This chapter is concerned with the general topic of information asymmetry in financial markets. Under this assumption, certain market participants have access to privileged, private information. We then distinguish two groups of investors, the informed - also called *insiders* - in opposition to the uninformed ones. In practice, there exists an important difference between insider and informed trading. Typically, the latter refers to the legal collection and exploitation of private information, for example, analysts covering a firm. In contrast, insider trading refers to the illegal exploitation of informational advantages and cases are unfortunately regularly reported in the press Simonian [2011], Scannel [2011]. However, we use the two terms interchangeably in the remainder of this thesis. This is because we adopt in our investment strategies the point of view of an uninformed investor facing informed ones. We are therefore not concerned with the legality of insiders' investments.

Historically, the literature in finance has first developed around the opposite assumption of homogeneous information. According to it, all market participants have access to the same level of information. Note that it was clear for its authors that this assumption was not realistic Lintner [1965], but necessary for their mathematical treatment. Combining the optimality of the mean-variance efficient portfolio Markowitz [1952], i.e., the portfolio with the highest expected return for a given level of risk, with this assumption, the authors conclude that the portfolio of all invested wealth, the market portfolio, is mean-variance efficient. Thus indexing, i.e., holding the market portfolio is optimal in equilibrium. If the market also contains a risk-free asset, all investors hold the same portfolio of risky securities but the specific combination of this portfolio and the risk-free asset depends on each investor's risk appetite.

There is thus a dichotomization of the investment decision, as it is divided into two independent tasks. This is called the fund separation property. Moreover, as a direct implication in terms of asset pricing, the CAPM Sharpe [1964], Lintner [1965] gives the appropriate rate of return of a risky security n, $r_n^{(t)}$, as a sole function of the return of the market portfolio $r_{market}^{(t)}$,

$$\mathbb{E}(r_n^{(t)}) - r_f = \beta_{n,market} \left(\mathbb{E}(r_{market}^{(t)}) - r_f \right), \tag{3.1}$$

where r_f is the risk-free rate.

Departing from the homogeneous information assumption, we review and compare two large bodies of literature in finance dealing with financial markets under information asymmetry, namely noisy rational expectations equilibrium (REE) (Section 3.1.1) and Bayesian-Nash equilibrium (BNE) (Section 3.1.2). In noisy REE, the equilibrium price emerges from the interaction of price-taking agents, whereas in BNE agents strategically take into account the impact of their actions on the price. We also establish a link between the presence of insiders and the necessity to introduce a variable lookback in prediction model (Section 3.1.3), a potential justification for the coming chapters. We then test and confirm our initial intuition in the context of experimental markets, where the experimenter controls private information signal and the number of insiders. We first highlight the information diffusion process (Section 3.2.2) before developing a method to detect the *time of maximally informative price* (Section 3.2.3). The latter corresponds to the first point in time where all information diffusion has taken place.

3.1 Theory of markets under information asymmetry

Two large bodies of literature in finance study the problem of information asymmetry, noisy REE and BNE. Both theories deal with the following questions.

(i) Role of the price in conveying information: this role of the price, beyond that of clearing the market, is of great importance. And although a form of the efficient market hypothesis corresponds to the situation where prices reflect all information, both public and private Fama [1970], little is understood on how this is obtained. Intuitively, the information diffusion takes place because informed agents reveal (at least partly) their private information when trading and uninformed agents (imperfectly) infer the level of private information from observed prices. This inference capacity is an assumption in noisy REE. Furthermore, a recent neurofinance study provides a foundation for this inference mechanism, called Theory of Mind Bruguier et al. [2010]. Interestingly, it is related to the social skills of market participants, such as their ability to detect others' intentions by observing their behavior, rather than traditionally sought after quantitative skills.

(ii) Implications of information asymmetry for portfolio choice and asset pricing: the questions are: does the fund separation property still hold under information asymmetry? Is there an equivalent CAPM relationship to price risky securities? How should uninformed agents tilt their portfolio to cope with the *winner's curse problem*? Indeed, uninformed agents serve as counter-parties

to informed ones. The market then clears, but the price at which uninformed agents buy (resp. sell) securities is too high (resp. too low)) compared to the level implied by private information. Winner's curse is a term that emanates from the theory of common value auctions and is also referred to as adverse selection in the literature.

3.1.1 Noisy rational expectations theory

In noisy REE theory, the price emerges from the interaction of price-taking agents, studied under the rational expectations assumption. From historical perspective, it was developed from the pioneer work of Grossman and Stiglitz [1980]. In their paper, they introduce the notion of a noisy rational expectations equilibrium price, as a price that conveys private information. But it does so only imperfectly such that informed traders are rewarded for collecting and exploiting costly private information. The presence of noise resolves the so-called Grossman-Stiglitz (GS) paradox, i.e., the impossibility of having an equilibrium in the market if information is costly and price fully revealing. Then, Grossman and Stiglitz make seven conjectures regarding the properties this equilibrium price possesses. Of particular relevance for us is the following one.

Hypothesis 1. *"The more individuals are informed, the more informative is the price system." Grossman and Stiglitz [1980]*

Grossman and Stiglitz then proceed to prove all their conjectures in the case of a simple analytical model, where the market only contains one risky and one risk-free security. Moreover, interactions between informed and uninformed market participants take place during a single trading period, with consumption at the end of trading. Their model was later on extended to the case of multiple risky securities economies by Admati [1985]. Admati's model not only better reflects reality. It also has a richer structure, such as effects that are not possible in the single security case. This is due to the general correlation structure in her model. For example, in one of her simulations with only two risky securities, she is able to find assets whose price decreases while their final payoff increases or Giffen goods, i.e., assets whose demand increases when their price increases. Her model was later extended to the case of multiple trading period by Brennan and Cao [1997], but with only one consumption period at the end of trading. Finally, the most recent overlapping generations models Biais et al. [2010] constitute the ultimate refinement. The market contains multiple risky securities that are traded over multiple periods of trading with intermediate consumption.

Model and notations

Before we proceed, let us introduce the following model and notations. Each trading period extends from time t to time $t+1$. There is a continuum of agents $a \in [0;1]$ that have constant absolute risk aversion (CARA) utility function in their wealth at the end of the trading period, $W_a^{(t+1)}$,

$$U(W_a^{(t+1)}) = \exp\left(-\rho W_a^{(t+1)}\right), \tag{3.2}$$

where ρ is the coefficient of absolute risk aversion. We consider a market with continuously indexed agent so as to make the resulting model analytically tractable, in particular by using the central limit theorem. Agents trade in a market that consists of N risky securities that pay a random dividend at the end of the trading period, $\boldsymbol{F}^{(t+1)}$, and a risk-free asset with risk-free rate r_f. There is also a portion λ of agents that are informed, which could either be an exogenous or endogenous variable of the model. In the latter case, market participants decide to become informed at a certain cost before trading starts. Then, λ is determined in equilibrium under the condition that the ex ante expected utility of informed and uninformed investors are equal. Before the start of trading, insiders receive a private signal that consists of a noisy version of the vector of final dividends

$$\boldsymbol{\iota}_a^{(t)} = \boldsymbol{F}^{(t+1)} + \boldsymbol{\epsilon}_a^{(t)}. \tag{3.3}$$

Each insider a receives a different signal. This combined with the multiple number of correlated risky assets makes the resulting price informative beyond the information already contained in an agent's private signal. The noise $\boldsymbol{\epsilon}_a^{(t)}$ is such that it is uncorrelated between agents and has zero mean in the cross-section of agents

$$\int_0^1 \boldsymbol{\epsilon}_a^{(t)} da = 0. \tag{3.4}$$

Thus at the aggregate level, i.e., over all participants,

$$\int_0^1 \boldsymbol{\iota}_a^{(t)} da = \boldsymbol{F}^{(t+1)}. \tag{3.5}$$

which means that the market overall knows perfectly the terminal dividend. The covariance matrix of the noise, $\boldsymbol{\Sigma}_\epsilon$, controls for the quality of the private information signal. As noted by Grossman and Stiglitz, the better the quality of the private signals, i.e., the lower variance Σ_ϵ, the more informative the price is. In the absence of noise, i.e., $\boldsymbol{\Sigma}_\epsilon = \boldsymbol{0}$, the price becomes fully informative and the GS paradox is not resolved.

(Noisy) rational expectations assumptions

The equilibrium price is obtained under the rational expectations assumption. Under this assumption, market participants know the probabilistic structure of the model. For example, in Grossman and Stiglitz [1980], uninformed agents know the joint distribution between price and final dividend. Because of their probabilistic knowledge, agents can optimally extract information from their information set $\mathcal{I}_a^{(t)}$ in the sense that the estimates they make are not systematically biased. Furthermore, one may wonder where this probabilistic knowledge that characterize the rational expectations assumptions comes from. Suppose that market participants have already played the game repeatedly. During this initial phase, learning takes place and terminates eventually after a certain number of iterations. Then, when agents continue playing the game, they use the same distribution they have learned to compute their demand, such that this distribution persists. In summary, analyzing a model under the rational expectations assumption means analyzing it in the steady phase of the

learning process. There is however an issue with the pure rational expectations assumptions. The resulting price becomes fully informative such that all participants choose to be uninformed. But if nobody is informed, it pays some to become informed. Thus there cannot exist an equilibrium and the GS paradox is not resolved.

There is an elegant way to resolve the GS paradox by the addition of a source of noise. Generally speaking, it takes the form of an uncertainty in the total supply or the demand of the securities. And it could be attributed to very different sources in practice such as noise traders, liquidity traders, life-cycle investors or traders that have an imperfect knowledge of the structure of the problem. For example, in Grossman and Stiglitz [1980] the aggregate supply of the risky asset is a random variable and the price is a linear function of it and the private information signal. As a result, the uninformed agents, who only observe the price, are unable to disentangle the evolution of the price driven by a change in private information from that driven by a change in the aggregate supply. In Biais et al. [2010], the aggregate supply is deterministic. The noise is introduced by the addition of a random endowment shock, i.e., an additional random source of income, received only by insiders at the end of the trading period. The component of this endowment shock spanned by the final payoffs of securities at the end of the trading period, $\boldsymbol{F}^{(t+1)} + \boldsymbol{p}^{(t+1)}$, enters in the portfolio choice problem of insiders. In the remainder of this chapter, we loosely call endowment shock this component of the endowment shock. This reflects the fact that investors should diversify the risk of their income in their portfolio. For example, the employee of a bank, whose income depends on his employer and more generally on the financial industry, should reduce his exposure to this sector in his portfolio. Similar to Grossman and Stiglitz [1980], the uninformed participants are unable to distinguish variations in the price that are caused by a change in the aggregate endowment shock from that by a change in private information. In summary, with the noisy rational expectation assumption, market participants still make correct inferences conditional on their information set, but they do this only imperfectly such that the price does not fully convey private information. Therefore, the resulting equilibrium price resolves the GS paradox.

Remark. How can we interpret the rational expectations assumption, in particular in the case where the interaction takes only place over a single period? Imagine there exists a series of parallel economies where agents play the same game. All probability distributions and expectations are defined in the cross-section of parallel economies. Then, in this cross-section the probabilistic knowledge is correct, although single realization might differ from the mean of the distribution. This is somehow comparable with the repeated sampling framework underlying the parametric estimation theory. In that case, the properties of an estimator are defined in the cross-section of repeated experiments, but the performance of the estimator on an individual realization might be quite poor.

Key results

We now explain the general method employed to find the analytical solution of the model. Firstly, the existence of a unique, linear equilibrium price function is assumed. The equilibrium price is, generally speaking, measurable in past prices, the vector of aggregate private information signal (the final cash flow because of (3.5))

and aggregate supply (possibly augmented with the aggregate endowment shock). The existence and uniqueness of this equilibrium remains to be proven later. Then, the optimal demand of agents is computed. It corresponds to the vector of portfolio weights as a function of the price. It is obtained by maximizing the expected utility of agents, conditional on their information set $\mathcal{I}_a^{(t)}$. In the case of informed agents, the latter consists of the private information signal and past prices. Indeed, in multivariate setting, price contains information beyond what is already contained in an insider's private signal. In Grossman and Stiglitz [1980], all insiders receive the same information signal, such that they do not use price when computing their demand. But this is particular to the one risky asset model. For uninformed agents, the information set consists only of past prices. Note that because agents have CARA utility function their demand does not depends on their wealth. Also, the fund separation property does not hold anymore as informed and uninformed agents hold different portfolios of risky securities, reflecting their different information sets. Finally, the equilibrium price is obtained by imposing the market clearing condition, i.e., aggregate demand equals to aggregate supply.

Numerical simulations highlight important features of the model. As already mentioned, the equilibrium price is a linear function of several random vectors of the model. Let us call $\boldsymbol{B}$ be the matrix of the multivariate regression coefficients of price on private information. Figure 3.1 represents the evolution of the diagonal element of the matrix $B_{n,n}$ as a function of the proportion of insiders λ for asset n. This is the sensitivity of the price on its private information. It thus represents a measure of price informativeness. When $\lambda = 0$, the coefficient $B_{n,n}$ is equals to zero. If there is no insider, the price cannot depend on private information. When λ increases, $B_{n,n}$ increases as well. Thus the larger the proportion of informed investors, the more the price depends on private information, the more informative the price system is. This confirms Hypothesis 1.

3.1.2 Bayesian-Nash equilibrium theory

Let us now turn our attention to the alternative theory of markets under information asymmetry, namely BNE theory Holden and Subrahmanyam [1992]. In this theory, the equilibrium price emerges from the interaction in a repeated game of strategic, noncooperative and rational investors. Each term characterizing an investor in BNE is important; let us come back to each individually.

(i) Strategic: this means that market participants decide on their holdings by taking into account the influence of their trades on the price process. They are also aware of the presence of other strategic, potentially informed investors. As a consequence, in BNE, a single agent can impact the price.

(ii) Noncooperative: this means that agents cannot communicate in order to coordinate their actions. They would be better off by coordinating and colluding against the central market maker, but this is not possible because they only interact in an anonymous market.

(iii) Rational: this refers to the assumption that investors decide on their holdings by maximizing their expected utility function, conditional on their information set. Moreover, a key ingredient in the model is the introduction of an additional source of noise, which results in a game of imperfect information. Noise traders,

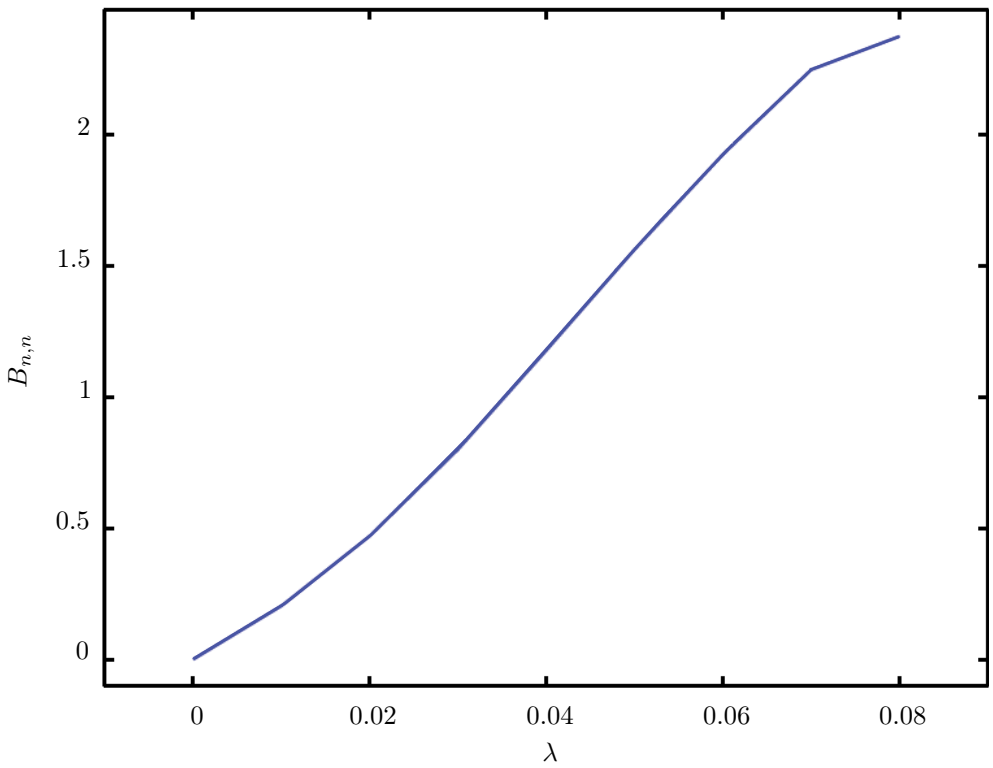


Figure 3.1: *Multivariate regression coefficient of price on private information for asset n, $B_{n,n}$, as a function of the proportion of informed investors λ. The larger this proportion is, the more the price is sensitive to changes in private information, thus informative. Biais et al. [2010]. Reproduced by permission of Oxford University Press.*

who also submit order to the central market maker, constitute this source of noise. Because of their demands, an informed agent cannot infer the order of other informed agents by observing the price. In the absence of noise, a similar result is obtained if agents are risk averse. In that case, they hedge their positions in risky assets and this prevent the price to become fully revealing.

The model is analyzed so as to obtain its unique linear Bayesian-Nash equilibrium. The model is solved by backward induction, starting at the last period. The Bayesian-Nash equilibrium of the subgame formed by the last trading period is computed under the assumptions that agents set their demand so as to maximize their utility function and the price clears the market, i.e., total supply equals to total demand. Then, the Bayesian-Nash equilibrium of the ultimate period is computed similarly, but conditional on the solution at the last period, etc. Figure 3.2 illustrates the central results of the BNE theory. It represents the variance of the final dividend conditional on price for market with different number of insiders. Intuitively, the smaller this variance is, the more informative the price is. We first note that, this variance decreases over time, stated differently, there is diffusion of private information into the price

process. However it does so quite differently for markets with different number of insiders. In the case of a monopolistic insider analyzed by Kyle [1985] the unique insider takes into account the price impact of his trades so as to optimally hide his action. The diffusion of private information takes place at a constant, almost linear rate. With more insiders, the decrease in variance is considerably faster. Thus, the speed of information diffusion increases with the number of insiders, again a confirmation of Hypothesis 1. Moreover, unlike predicted by noisy REE, all informational inefficiencies disappear very quickly. Even in the case of two insiders, the variance of the final dividend drops after a couple of trading periods. This is justified by the aggressivity of traders, who want to benefit first from their informational advantages while not being able to coordinate their actions.

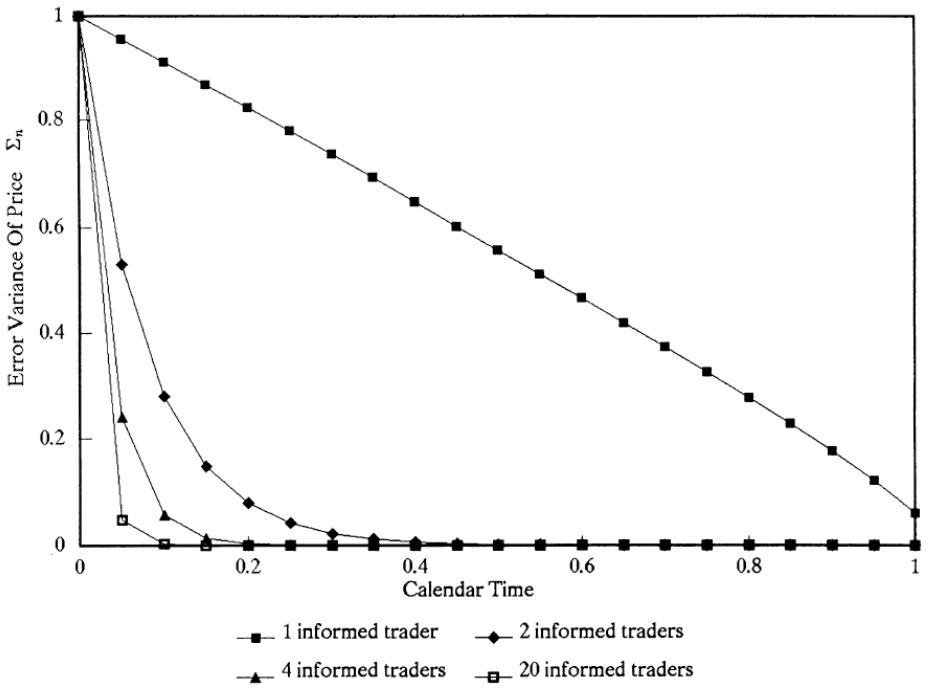


Figure 3.2: *Conditional variance of the final dividend in calendar time, for markets with different number of informed investors. The conditional variance decreases as a function of time, but more abruptly for market with more insiders. When there is only one insider, the variance decreases slowly, at an almost linear rate. Even in markets with two insiders, all informational inefficiencies disappear rapidly. Holden and Subrahmanyam [1992]. Reproduced by permission of John Wiley & Sons, Inc.*

3.1.3 Our view

Comparaison between noisy REE and BNE

We have just reviewed two existing theories of financial markets under information asymmetry, noisy REE and BNE. Table 3.1 summarizes our discussion and compares their frameworks, assumptions and implications. The biggest difference lies in the presence, on the one hand, of price-taking continuously indexed agents, where no one can impact the price, compared, on the other hand, to strategic agents, which can individually impact the price. Moreover, whereas both theories agrees on the fact that the larger proportion of informed investors leads to a more informative price (Hypothesis 1), they disagree on the existence of long-lived informational inefficiencies. BNE supports their absence, because agents trade aggressively in order to be the first one to benefit from their informational advantages. Also, the nature of the equilibrium that both theories approaches is different. In BNE, the equilibrium price is defined in the sense of a form of the efficient market hypothesis, i.e., prices are such that they reflect all information, both public and private Fama [1970]. In noisy REE, equilibrium price is not fully revealing of private information. To anticipate the results of the coming sections, experimental finance confirms that the nature of the resulting equilibrium is better described by noisy REE but BNE gives a mechanism (strategic interaction) explaining the emergence of this equilibrium.

Information asymmetry and variable lookback

Let us now establish a link between the presence of informed investors and the idea of a variable lookback model. This is only a potential justification for it. Uninformed agents should make their predictions using past prices as a proxy for private information. But how they do this depends on the relative proportion of insiders in the market. In the extreme case of no informed agent (resp. all informed agents), there is no (resp. immediate) diffusion of private information into the price process, thus no predictability. From the point of view of an uninformed agent, there is no point to try to make predictions. In the case of a unique informed agent Kyle [1985], he can optimally hide his action when trading, because he is taking into account the impact of his trades on the price process. The information diffusion is very slow, predictability is small and a large portion of past observations is relevant when forming predictions. On the contrary, when the number of informed agents increases, they face a dilemma. They want to hide their actions, while being the first to benefit from their knowledge of private information. The information diffusion is considerably faster, the price process more predictable and a shorter proportion of the past observations is relevant when forming predictions. Now, suppose information arrives in the market as a point process with both private information and public information that resolves private information. There is consequently a time-varying random number of insiders in the market that induces a time-varying predictability of the price. More importantly, uninformed agent should form their prediction using a time-varying portion of past prices. This is the central idea behind our variable lookback model. Before we study the estimation of this model in Chapter 4, let us first confirm this intuition in the context of experimental finance.

	Noisy REE	BNE
Framework of analysis		
	Equilibrium under the rational expectations assumption	Bayesian-Nash equilibrium of repeated imperfect information game
Assumptions		
Investors' Rationality	✓	✓
Strategic	✗	✓
	Price takings agents	Noise traders are price insensitive
Cooperative	✗	✗
Correct estimates and conjectures	✓	✓
Market clears	✓	✓
Results		
Nature of the equilibrium	Price partially reveals private information	Price reveals all public and private information
Price informativeness proportional to the number of insiders	✓	✓
Existence of long-lived informational inefficiency	✓	✗

Table 3.1: *Comparaison of noisy REE and BNE.*

3.2 Confirmation from experimental finance

Experiments have a long tradition in sciences, in particular physics, as a way to test theories. But, it is only relatively recently that they have found their way in finance research. See for example the pioneer work of Plott and Sunder [1988], who conducted at that time his experiments openly in a classroom. This delay is possibly attributable to the large quantities of field data already available. However, unlike field data, their experimental counterparts present several advantages. In the first place, the experiment represents a simplification of a real world scenario. For example, participants interact in a complete market made of a limited number of securities. Obviously, the level of complexity has to be balanced. On the one hand, it should be complex enough so as to rule out trivial solutions. On the other hand, it should be simple enough, in particular, to be able to disentangle competing effects, something not always possible in field data. Experiments retain as a central element of their complexity the human being. This distinguish experimental data from simulations, which, despite their complexity, remain only stylized mathematical models. In summary, an experiment is skillfully designed as to retain the essence of the problem. In the second place, experiments are run in a controlled setup. It depends on a set of parameters that forms its experimental condition. The experimenter then replicates the experiment using various parameters' value so as to study their impact on the outcome. For example, in experimental markets with insiders, the experimenter controls the number of informed participants or the quality of the private information signal. Finally, the experimenter has access in his analysis to quantities not directly observable in practice, for example, private information signals received by insiders.

3.2.1 Description of the experiment

In this section, we describe the setup of the experiment used to generate data analyzed in the coming sections Bossaerts et al. [2010], Bruguier et al. [2010]. Twenty subjects are recruited to participate in an experiment run using a computerized system. They are either finance professionals or graduate students in finance. Before the start of the experiment, subjects receive the following instructions. The market in which they trade contains only two securities, called X and Z, that pay a random, complementary dividend between 0 and 50 cents at the end of each trading period. That means that if X pays F cents of dividend, then Z pays $50 - F$ cents. Only trading of X is permitted and short selling is allowed, in the limit that a subject does not become bankrupt. Overall, at the market level, there is the same amount of X and Z. Thus the total amount of money distributed by the experimenter is fixed and the unconditional price of X is 25 cents as there is no aggregate risk. The experimenter makes sure the subjects understand the instructions and answers possible questions. Then, the experiment itself starts and consists of a series of 13 independent trading sessions. Before each session, participants receive an initial portfolio of securities X and Z, different for each participant. Also, there is a subgroup of subjects, the insiders, that receive a private signal related to the realization of the final dividend F. More specifically, the signal gives a 10 cents bound within which the final dividend is comprised. For example, if the private signal is 40 cents, insiders know that the final dividend lies between 35 and 45 cents. Like in noisy REE, the signal consists

of a noisy version of the final dividend in order to prevent the price from being fully revealing. Then, the trading session starts and last 5 minutes. Trading takes place in a double auction market, described in Section 2.1.1. Subjects can submit buy and sell limit orders, that accumulate in a publicly visible order book, as well as buy and sell market orders. At the end of the trading period, the value of dividend is publicly revealed and subjects receive an amount of money that corresponds to their final portfolio. On average, subjects make 55$ per experiment. There are in total of 5 experiments.

A certain number of comments regarding the design of the experiment are in order. Firstly, by design, subjects have a clear incentive to trade in the market. Indeed, as suggested by finance theory, there exist three reasons for trading, namely differences in endowments, beliefs and preferences, and all of them are present in the experiment. The difference in endowments comes from the different initial portfolio of securities X and Z received by subjects before the start of trading. The difference in beliefs results from the introduction of a private signal for insiders. And the difference in preferences results from the difference in risk aversion between subjects, a reasonable assumption in practice. Risk averse subjects want to balance their portfolio so as to equate the quantity of X and Z and reduce the risk of their holdings. Secondly, one may argue that the subjects do not comprehend the game they are playing. As a result, they play simple, stupid strategies (even if they are harmful for them) and experimental data contain only spurious noise. Several ingredients in the design prevent this. First of all the anonymity of the market and the relatively large number of participants ensure that these noise effects average out. If one subject plays one random strategy, we can assume an other one plays another random yet uncorrelated strategy and their combined effect do not affect the outcome. The anonymity in the market is ensured by the special design of the room where the experiments take place: subjects seat in a cubicle and are prevented to see the screen and the keyboard of their peers. Moreover, subjects are extensively trained before the start of the experiment. Beside the instructions they receive, as already mentioned, they trade during a couple of sessions to familiarize themselves with the rules of the experiments and the double auction mechanism. Results are only registered after this initial training phase. Finally, certain critics concern the choice of the double auction as market mechanism, in particular. that it is too complex. The use of a double auction market mechanism produces an efficient allocation and is able to reproduce stylized facts predicted by the theory. Moreover, the complexity of this market mechanism plays in his favor, as it is extremely hard to strategize against it given the large set of available actions (choice between limit and market orders, level of the limit order). Also, this is the market mechanism used practically by every stock exchanges, which makes the experiments closer to reality.

The experimenter registers the entire evolution over time of the order book, from which he can compute certain quantities, such as the quote mid-price, etc. The evolution of the price for the 13 trading sessions of experiment 1 is represented in Figure 3.3. In general, we observe that the price tends to approach the private information signal.

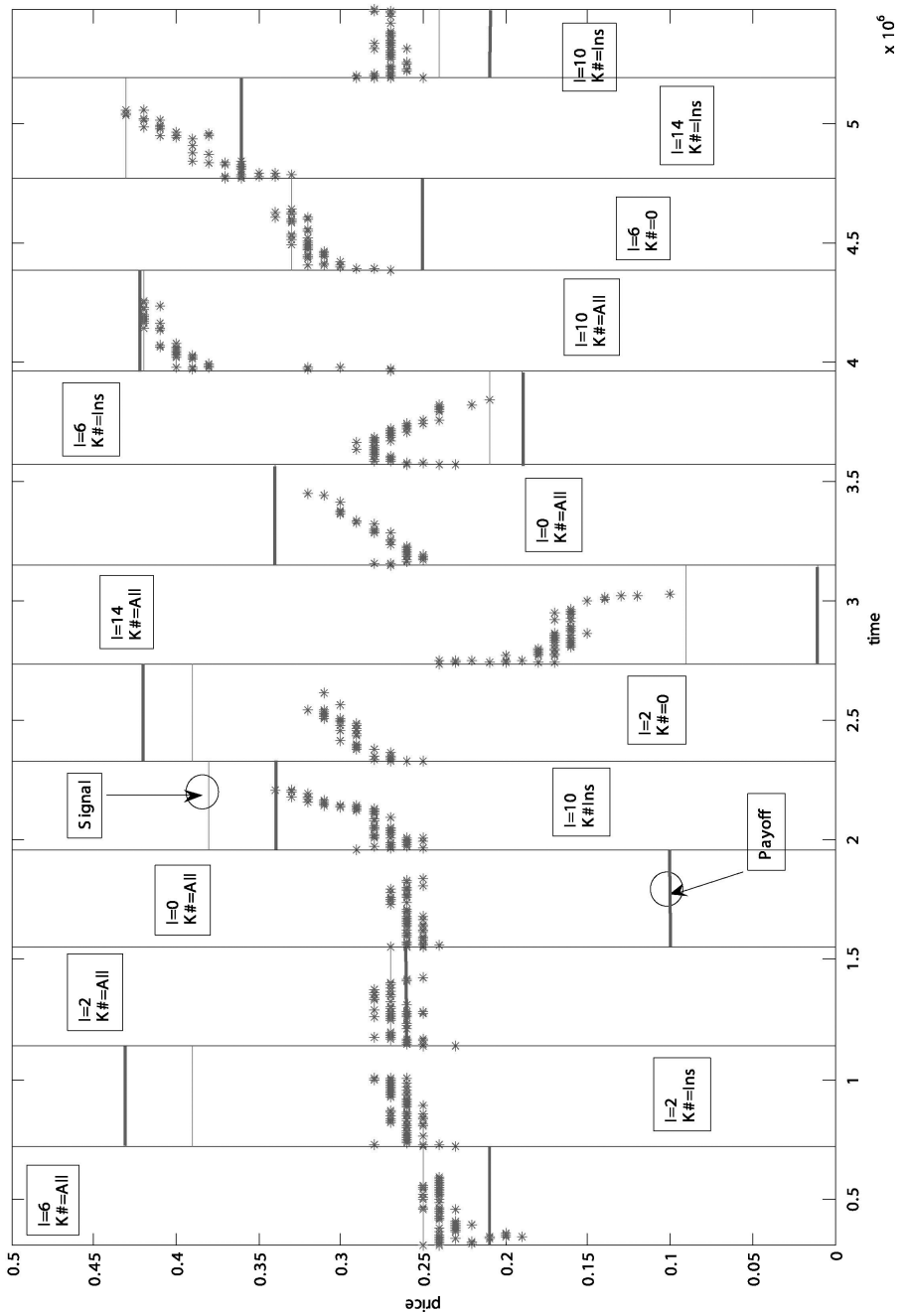


Figure 3.3: *Evolution of the quote mid-point price for the 13 trading sessions of experiment 1. Price approaches the private signal, in general. Bruguier et al. [2010]. Reproduced by permission of John Wiley & Sons, Inc.*

3.2.2 Information diffusion

Analysis method

In a first analysis, we aim to highlight the diffusion process of private information into the price, as well as study the influence of the number of insiders on the speed of this diffusion. We run the cross-sectional regression of price on private information, where time-series are grouped per experimental condition (number of insiders). More precisely, let us define

$$p_{i,j}^{(t)} \tag{3.6}$$

to be the price of security X at time t, in trading session i which has j insiders. Then, let us stack in a vector $\boldsymbol{p}_{\cdot,j}^{(t)}$ all prices that have the same number of insiders j. Likewise, let $\iota_{i,j}$ be the private information signal received by insiders in trading session i which has j insiders. By default, we set $\iota_{i,0} = 0$, i.e., there is no private information in markets with no insider. Then, we define the vector $\boldsymbol{\iota}_{\cdot,j}$ by stacking all private information signals of the trading sessions with j insiders. Then, the equation describing the linear cross-sectional regression is then given by

$$\boldsymbol{p}_{\cdot,j}^{(t)} = \alpha_j^{(t)} + \beta_j^{(t)} \boldsymbol{\iota}_{\cdot,j} + \boldsymbol{e}_{\cdot,j}^{(t)}, \tag{3.7}$$

where $\boldsymbol{e}_{\cdot,j}^{(t)}$ represents the (vector of) residuals. The estimates of the time-varying parameters $\alpha_j^{(t)}$ and $\beta_j^{(t)}$ are obtained by OLS.

Results

The first quantity of interest is $\beta_j^{(t)}$. It reflects the amount of private information reflected in the price at time t in a market with j insiders. Figure 3.4 represents the evolution over time of this regression coefficient, for different number of insiders. The case with 0 insiders is a sanity check. There is no private information to be reflected in the price and thus the regression is zero all the time. Except for this case, the coefficient increases in time thus the price becomes more informative over time, albeit differently depending on the number of informed investors. On the one hand, when the number of insiders is large (10 for example), the coefficient increases and stabilizes at a plateau quickly. Information is rapidly impeded in the price and the price is almost perfectly correlated with the private information signal, as the coefficient reaches a value close to one. On the other hand, with a small number of insiders (6 for example), the regression coefficient increases at a slower rate and the plateau level is lower. Information diffusion is slower and the final price is less informative. In the case with two insiders, plateau level is not reached at the end of the 5 minutes of trading, indicating that the diffusion process is still taking place. In conclusion, the speed at which information is reflected in the price increases with the relative proportion of informed investors, as predicted by both noisy REE and BNE theory. Moreover, as illustrated by the case with two insiders, all information inefficiencies do not disappear quickly, supporting the noisy REE theory.

Another quantity of interest is the mean squared error (MSE), defined by

$$MSE_j^{(t)} = \frac{1}{N} \sum_{i=1}^{n} (e_{i,j}^{(t)})^2. \tag{3.8}$$

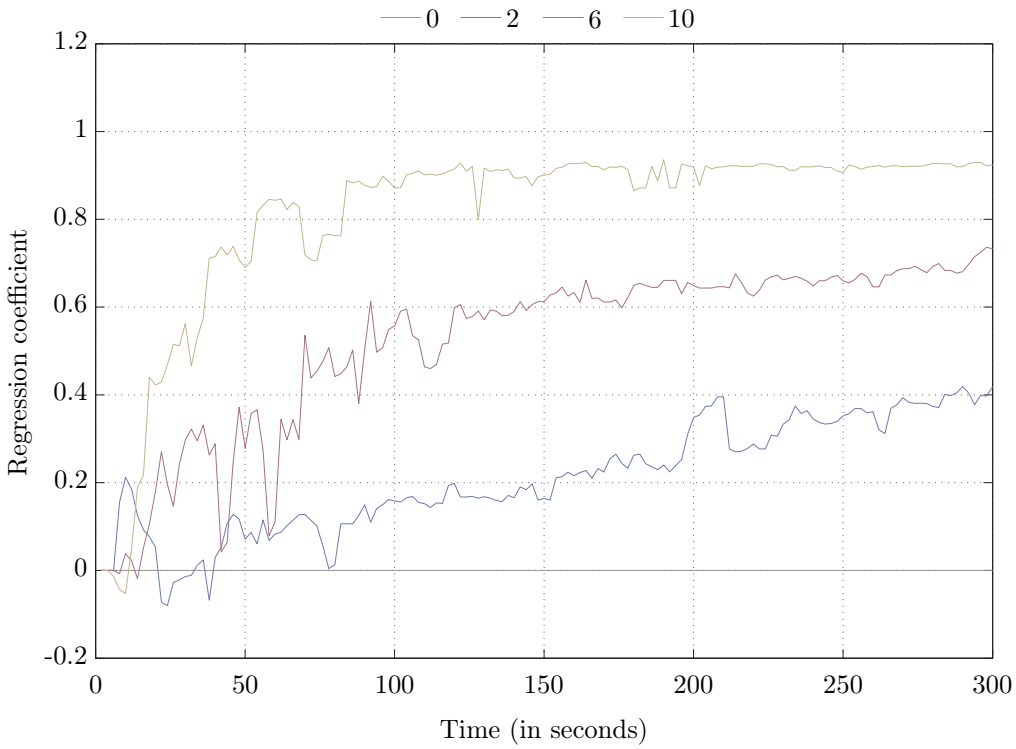


Figure 3.4: *Evolution over time of the cross-sectional regression coefficient of the price on private information, as a function of the number of insiders.*

A finance practionner calls the square root of the mean squared error volatility and there are two reasons to focus on this quantity. Patterns in the volatility, in particular GARCH-like features such as the clustering of volatility, were the only significant patterns detected in the time-series of price that could explain how uninformed agents detect the presence of insiders Bruguier et al. [2010]. Our results is in line with those of Bruguier et al. [2010]. Moreover, the arrival of information in the market is associated with an increase of volatility, followed by a subsequent decrease due to incorporation of information Lamoureux and Lastrapes [1990]. Figure 3.5 represents the evolution over time of the MSE of the residuals as a function of the number of insiders. The picture looks somehow less clear than Figure 3.4 but we can still make the following observations. At the beginning of the period, corresponding to the arrival of private information, the volatility increases and tends to decrease over time. That is, information is progressively impeded in the price. Interestingly, in the case of two insiders, the MSE remains high at the end of the period. This reflects again that the diffusion of private information into the price is still taking place at the end of the 5 minutes of trading.

This analysis highlights clearly the diffusion process of the private information into the price process. However, it could not be performed in field data, because we have adopted the point of view of an omniscient experimenter. Indeed, both the number

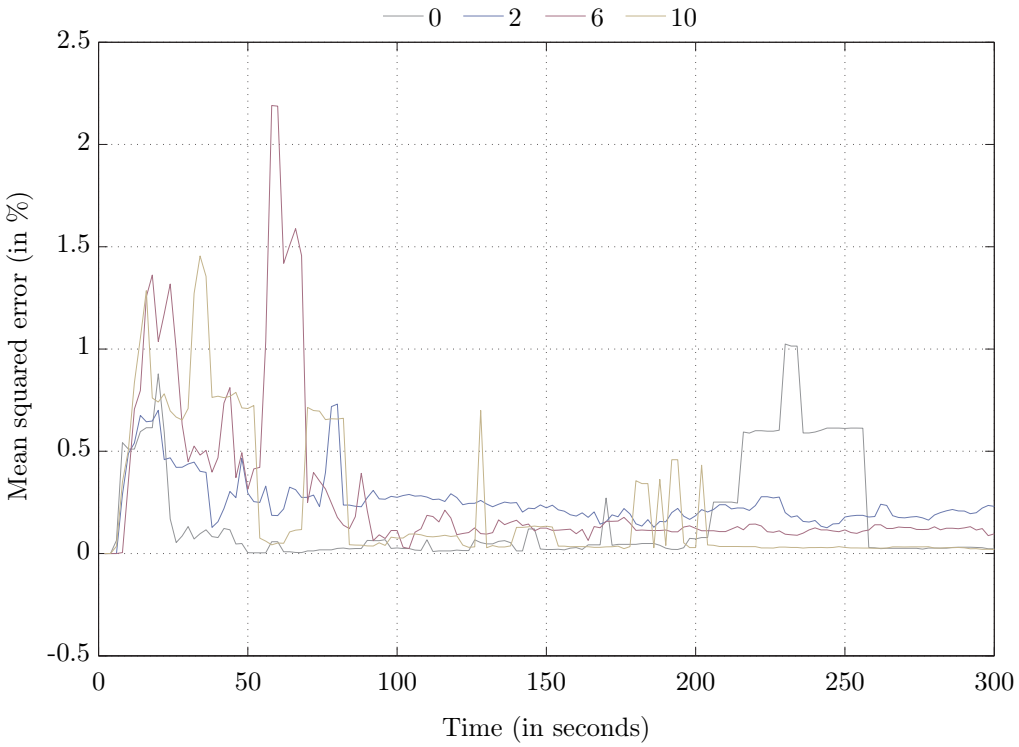


Figure 3.5: *Evolution over time of the mean squared error, as a function of the number of insiders.*

of insiders as well as the private signals are necessary for our analysis, but they are not observable in practice.

3.2.3 Detection of the time of maximally informative price

Analysis methodology

In a second analysis, we adopt the point of view of an uninformed investor that has only access to a single realization of the time-series of prices. Assuming information arrives in the market at $t = 0$, our objective is to detect the time of maximally informative price and relates it to the proportion of informed insiders in the market. Let us start by clarifying this notion of time of maximally informative price. Intuitively, this is the first point in time where all information diffusion has taken place. Consider again Figure 3.4. We distinguish two distinct phases. The regression coefficient $\beta_j^{(t)}$ first monotonically increases when the information gets impounded in the price and then reaches a plateau when all information diffusion has taken place. In the presence of risk neutral market participants, this happens but, at the same time, the price is a martingale and there is no predictability in the price process. Conversely, in the presence of risk averse market participants, there is predictability in the price process during the phase of information diffusion. Indeed, during this phase, uninformed in-

vestors perceive more risk due to the adverse selection effect. Consequently, the price drops initially to compensate the risk averse agents for holding the risk of the asset. On average, the expected return should be higher than in the absence of insiders. As the information is impounded in the price, the intensity of the adverse selection effect decreases, the risk premium decreases as well and prices increase. The predictable change in price lasts until no more information is being incorporated in the price. Then, the time of maximally informative price corresponds to the first point in time characterized by the absence of predictability in the price process. Moreover, this phase of initial predictability is shorter in market with more insiders, as the information is incorporated in the price more quickly. Stated differently, the time of maximally informative price should be smaller in markets with more insiders.

Let us now give a more formal treatment of this notion. To represent the dependencies of the current price on past k prices, we could use a Markov process of order k,

$$\mathbb{P}\left(y^{(t)}/y^{(t-1)}, y^{(t-2)}, \ldots\right) = \mathbb{P}\left(y^{(t)}/y^{(t-1)}, y^{(t-2)}, \ldots, y^{(t-k)}\right) \tag{3.9}$$

$$= \mathbb{P}\left(y^{(t)}/\boldsymbol{y}^{(t-k)>(t-1)}\right) \tag{3.10}$$

where $y^{(t)}$ are the observations of the process, typically some function of the price. Recall also that $\boldsymbol{y}^{(t')>(t'')}$ denotes the concatenation in a sequence of the observations from t' to t''. In a Markov process of order k, the emission probability of the next symbol is conditioned on a fixed portion of k past symbols. Markov processes are extremely versatile, but suffer from the curse of dimensionality. If $y^{(t)}$ can takes M values, the Markov process is parameterized by a matrix of $M^k \times (M-1)$ transition probabilities. To resolve this dimensionality problem, let us consider instead tree machine Rissanen [2005]. In a tree machine, the probability of an observation depends on a set of past observations of different length, called *context* $\boldsymbol{\xi}$. We can represent the data generating process as an M-ary tree. Each node corresponds to a context, in particular the root node to the empty context $-$. Furthermore, if edges are labeled with the M values taken by $y^{(t)}$, the context of a node corresponds to the path from the root leading to that node. Moreover, each leaf node, i.e., a node without children stores a conditional distribution of the next symbol given its context $\mathbb{P}\left(y^{(t)}/\boldsymbol{\xi}\right)$. Note that a Markov process of order k corresponds to the complete M-ary context tree with k levels. Also, to sample a random process from a context tree, the following method is applied. Suppose we have already generated a set of symbols $y^{(1)}, \ldots, y^{(t-1)}$. Then, we climb the tree following past symbols $y^{(t-1)}, y^{(t-2)}$, etc. until we reach a leaf node. The next symbol is then generated by drawing at random from the conditional distribution stored at that node $\mathbb{P}\left(y^{(t)}/\boldsymbol{\xi}\right)$.

Example 6. *Suppose observations takes binary values, i.e.,* $y^{(t)} \in \{0, 1\}$. *Figure 3.6 represents an example of a binary context tree. By convention, edges from a parent node to its left (resp. right) child is labeled with 0 (resp. 1). The second node on the second row has context* 01 *which corresponds to the path of going from the root to this node.*

The Context algorithm Rissanen [2005], described in Algorithm 3.1, allows building a context tree given a sequence of observations $y^{(1)}, \ldots, y^{(T)}$. To simplify the

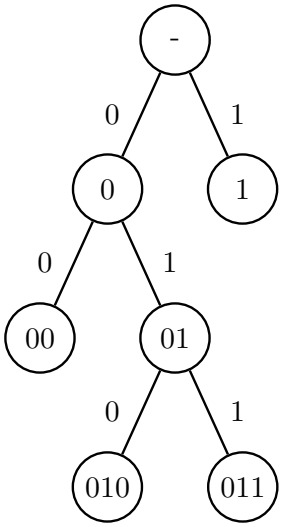


Figure 3.6: *Example of context tree.*

treatment, suppose the observations take binary values. We later generalize the algorithm to the case of a general, possibly continuous alphabet. Each node in the tree is indexed by its context $\boldsymbol{\xi}$. It also stores the number of 0 and 1 observed so far, $count_0$, $count_1$ and an estimate of the distribution of the next symbol conditional on its context $\mathbb{P}(y/\boldsymbol{\xi})$. Each node also contains a code length CL, that represents the capacity of a node to encode observations efficiently. The link between data compression and statistical learning will be made clear in the next chapter, with the introduction of the MDL principle. For the time being, let us just assume that a short code length is good. The algorithm starts with a root node, labeled with the empty context $-$, with count $count_0(-) = count_1(-) = 1$ and code length $CL(-) = 0$. Then for every observation, the algorithm first encodes it (encoding or pruning step), and then updates the knowledge stored in the tree (update and growth step). During the encoding step, starting at the root, the algorithm climbs the tree by reading past symbols in reverse order $y^{(t-1)}, y^{(t-2)}, \ldots$. It stops when (a) it reaches a leaf node or (b) it reaches a node whose children are not more efficient than itself. This happens when the code length of the parent node is smaller than the sum of code lengths of its children. Because we are interested in the length of the context used for encoding, the algorithm returns it. During the growth and update phase, the algorithm, restarting at the root, climb the tree following past symbols in reverse order. For every node visited, the algorithm increases their count of symbols, code length and update the estimate of the probability distribution. The algorithm stops either (a) when it reaches a leaf node or (b) when an internal node's count becomes 2. This last condition allows the tree growing only at repeated occurrences of a context. Finally, if the algorithm has reached a leaf node, it is extended by its two children with count equals to 1 and with code length from their parent node.

Algorithm 3.1 Context algorithm.

Input: $y^{(t)}, t = 1, \ldots, T$

Classdef $\{count_0, count_1, CL, \mathbb{P}(\ldots)\}$ as Node

%Initialization
$Node(-).count_0 \leftarrow 1$
$Node(-).count_1 \leftarrow 1$
$Node(-).CL \leftarrow 0$

%For each observation
for $t = 1, \ldots, T$ **do**

  %Choice of encoding node
  Start at root node
  repeat
    Climb the tree following past observations $y^{(t-1)}, y^{(t-2)}, \ldots$
  until $Node(\boldsymbol{\xi})$ is a leaf $\vee\, Node(\boldsymbol{\xi}, 0).CL + Node(\boldsymbol{\xi}, 1).CL > Node(\boldsymbol{\xi}).CL$
  return Length of $\boldsymbol{\xi}$

  %Update phase
  Start at root node
  repeat
    Climb the tree following following past observations $y^{(t-1)}, y^{(t-2)}, \ldots$
    $Node(\boldsymbol{\xi}).count_{y^{(t)}} \leftarrow Node(\boldsymbol{\xi}).count_{y^{(t)}} + 1$
    $Node(\boldsymbol{\xi}).CL \leftarrow Node(\boldsymbol{\xi}).CL - \log_2 \mathbb{P}\left(y^{(t)}/\boldsymbol{\xi}\right)$
    Update model $Node(\boldsymbol{\xi}).\mathbb{P}(\ldots)$ given $y^{(t)}$
  until $Node(\boldsymbol{\xi})$ is a leaf $\vee\, Node(\boldsymbol{\xi}).count_{y^{(t)}} = 2$

  %Growth phase
  if $Node(\boldsymbol{\xi})$ is a leaf **then**
    $Node(\boldsymbol{\xi}, 0).count_0 \leftarrow 1$
    $Node(\boldsymbol{\xi}, 0).count_1 \leftarrow 1$
    $Node(\boldsymbol{\xi}, 0).CL \leftarrow Node(\boldsymbol{\xi}).CL$
    $Node(\boldsymbol{\xi}, 1).count_0 \leftarrow 1$
    $Node(\boldsymbol{\xi}, 1).count_1 \leftarrow 1$
    $Node(\boldsymbol{\xi}, 1).CL \leftarrow Node(\boldsymbol{\xi}).CL$
  end if
end for

Let us now generalize the Context algorithm to the case where the observations take value in a nonbinary alphabet Rissanen [2007]. The tree is grown using a quantized version of the observed symbols. For example, consider the following quantization scheme

$$Q\left(y^{(t)}\right) = \begin{cases} 1 & \text{if } y^{(t)} \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{3.11}$$

where $Q\left(y^{(t)}\right)$ corresponds to the quantized version of observation $y^{(t)}$. Moreover, the model describing the data conditional on the context is defined at the same precision as the original alphabet. For example, with continuous random variable and a context $\boldsymbol{\xi}$ of length k we can use an autoregressive processes of order k

$$y^{(t)} = \alpha + \beta_1 y^{(t-1)} + \ldots + \beta_k y^{t-k} + e^{(t)}. \tag{3.12}$$

The model is estimated using observations whose past symbols match context $\boldsymbol{\xi}$. Finally, the coding distribution corresponds to a universal model for that class of processes.

Let us come back to the problem of detecting the time of maximally informative price. When the information diffusion has taken place and no more predictability of price exists, the Context algorithm should encode subsequent symbols using the root node, i.e., a model of order zero. And we equate time of maximally informative price with the first time the algorithm comes back to a model of order zero. If this does not happen, the algorithm does not conclude and does not output a value. More precisely, we apply the Context algorithm on the time-series of price changes

$$y^{(t)} = p^{(t)} - p^{(t-1)}. \tag{3.13}$$

A ternary tree is grown using the following quantization

$$Q\left(y^{(t)}\right) = \begin{cases} 1 & \text{if } y^{(t)} > 0 \\ 0 & \text{if } y^{(t)} = 0 \\ -1 & \text{if } y^{(t)} < 0. \end{cases} \tag{3.14}$$

Each node use as model an autoregressive process, whose order equals to the length of its context. It is estimated using observations where this context occurs. Finally we use as universal model the probability distribution corresponding to an existing model selection criterion, namely the normalized maximum likelihood (NML) Rissanen [2007].

Results

Figure 3.7 represents the time of maximally informative price identified by the Context algorithm. Only the first set of 13 trading sessions of experiment 1 are used. However, there are not 13 points one the figure because the algorithm does not always conclude. Identified times of maximally informative price are plotted against the number of insiders and the curve represents the OLS regression of the time of maximally informative price on the inverse number of insiders. The regression coefficient is statistically significant. This again confirms Hypothesis 1.

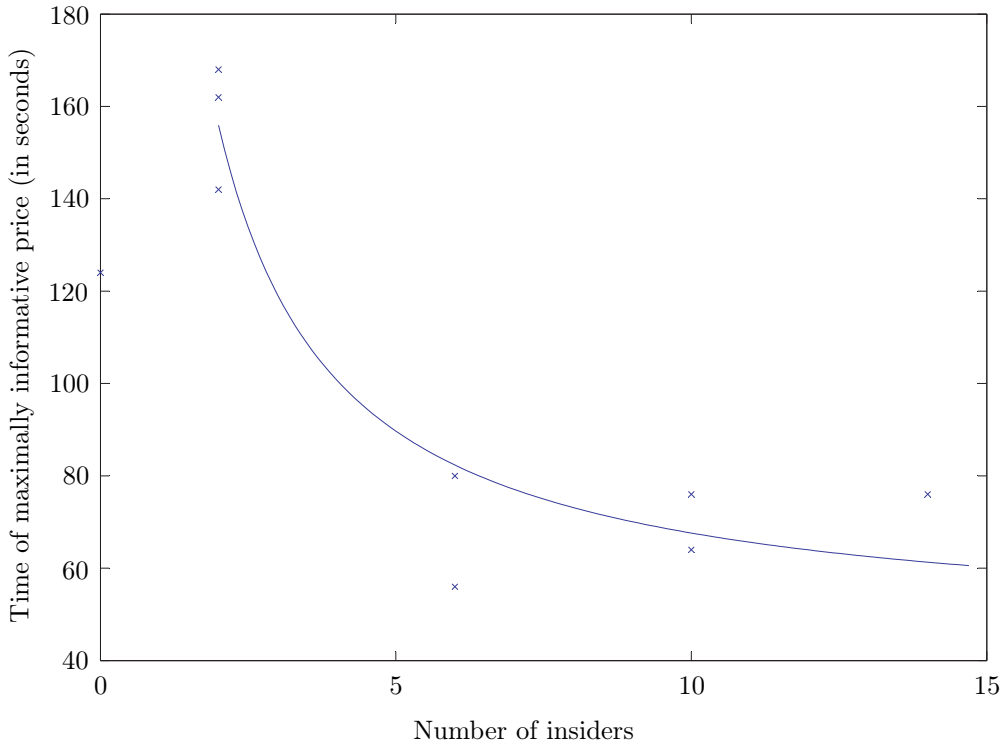


Figure 3.7: *Estimated times of maximally informative price displayed as a function of the number of insiders. Only the time-series from the first experiment are used. The curve corresponds to the OLS regression of the time of maximally informative price on the inverse number of insiders.*

Figure 3.8 represents the same result for all 5 experiments, each with 13 trading sessions. Each color represents a different experiment number. Similarly, we display the estimated times of maximally informative price as a function of the number of insiders grouped per number of insiders and plot the OLS regression of the time of maximally informative price on the inverse number of informed investors. The black curve represents the random coefficient regression Poi [2003]. This regression adds to each experiment a common constant, suppose to capture the difference between groups of participants involved in different experiments. For example, think of the difference of aggressivity in trading of a group of participants which makes a market with only 2 insiders in one case resembles one with 6 insiders in another case. Again, the random coefficient regression is significant. This supports Hypothesis 1.

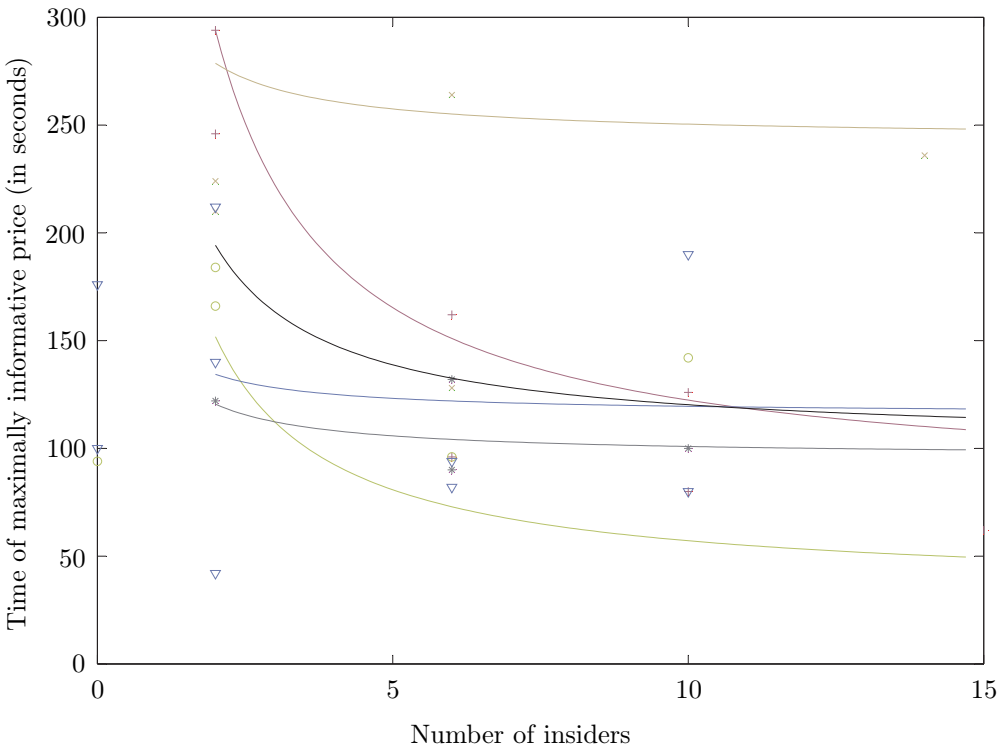


Figure 3.8: *Estimated times of maximally informative price, as a function of the number of insiders. Time-series from all 5 experiments are represented by a different color. The colored curve corresponds to the OLS regression of the time of maximally informative price on the inverse number of insiders and the black curve to the Swany's random coefficient regression.*

Summary

In this chapter, we have reviewed and compared two theories of financial markets under information asymmetry, the noisy REE and the BNE. Whereas in REE price emerges as the equilibrium between price-taking agents under the rational expectations assumption, it results from the strategic interaction of investors in BNE. In both theories, price plays an articulated role in conveying information from informed to uninformed market participants. Both theories agrees that the informativeness of the price is proportional to the relative proportion of informed investors in the market. However, in noisy REE, the price does not fully reveals all private information, whereas all informational inefficiencies disappear very quickly in BNE. We have then tested both theories in the context of experimental finance. We were able to highlight the information diffusion process as well as determine the price of maximally informative price using the Context algorithm. Interestingly, our view reunites both theories, in the sense that the type of equilibrium the price system converges to is better described by noisy REE but BNE explains how this equilibrium emerges. The existence of long-lived informational inefficiencies is also a good news from the uninformed agent point of view: there is money left on the table and strategies that optimally extract information from past prices are profitable. The aim is not to recover the private information signal of course, but rather to characterize the diffusion of this information so as to distinguish if the evolution of price is driven by information or not. This is the topic of the coming chapters.

Chapter 4

Variable Lookback Algorithm

> When I look back on my life, it's not that I don't want to see things exactly as they happened, it's just that I prefer to remember them in an artistic way.
>
> Lady Gaga

In the introduction of this thesis, we have proposed the idea of a variable lookback model, i.e., a statistical model where a time-varying portion of the past observations is relevant for the future. We have also seen in Chapter 3 that a potential justification is given by the theory of financial markets under information asymmetry; the relative proportion of informed investors controls the speed of diffusion of private information into the price, and since this proportion varies over time, a time-varying portion of past prices is relevant for the future. We have also verified this intuition in the context of experimental finance. In particular, we have seen that the time of maximally informative price, i.e., the first point in time where all information diffusion has taken place is inversely proportional to the number of informed investors. Expressed in terms of lookback window, this leads to the statement above. The variable lookback model is more generally motivated by the need to develop learning methods able to handle nonstationarity inherent to financial signals, whether or not it is the consequence of the presence of information asymmetry. Moreover, as we have also seen in Chapter 2, current solutions, in particular local windowing, suffer from certain shortcomings.

We are now ready to introduce our new solution, a learning algorithm that selects simultaneously the order of the process and the lookback window based on the MDL principle. We therefore coin the term variable lookback algorithm for it. The solution is algorithmic, which means that we describe a general procedure. It is furthermore independent of the choice of models, so as to prevent turning this thesis into a mere data mining exercise. Moreover, the proposed solution is universal, which informally means that it is somehow close to the best solution in hindsight, i.e., with full knowledge of the whole history of the process. We will give in this chapter a more precise definition of this notion of universality. How close we are to the best solution in hindsight represents the cost of learning progressively the model in an online manner.

Of course, if we had a strong idea on the data generating process and the underlying model, this is the penalty we would incur to use a universal model. But since we work with financial time-series, this is unlikely to be the case. Finally, our solution is only based on the idea of piecewise stationarity. This means there exists a set of times $t_0 = 1 < t_1 < \ldots < t_S \leq T$ such that the data are generated by a model of order $k - S \in \{0, \ldots, K\}$ in the interval $[t_{s-1}; t_s)$. We make no assumption on the distribution of switching times, their numbers S or the positions at which they occur $\{t_s\}_{s=1}^{S}$. Furthermore, the solution does not rely on the ergodicity of the time-series.

The remainder of this chapter is organized as follows. We start by reviewing necessary notions of information theory (IT) (Section 4.1.1). Of particular importance, we see how Kraft's inequality and the information inequality establish two links between the notion of a probability distribution and that of a code length function. We also review the existing theory of MDL as it underlies our proposed solution (Section 4.1.2). Starting with the origin of MDL rooted in Kolmogorov's algorithmic complexity theory, we give a first crude version of MDL, before enunciating a modern version based on the idea of a universal model. Later, we present our algorithm (Section 4.2.1 and 4.2.2). Towards the end of the chapter, we present the first tests of the algorithm on simulated time-series, whereas applications on real financial data is the topic of Chapter 5. We first test the model selection criterion in the absence of switch in the system (Section 4.3.1). Then, we study the behavior of the algorithm in the idealized scenario when the underlying system switches between a highly correlated model to and i.i.d. one, and vice versa (Section 4.3.2).

4.1 Background material

4.1.1 Basics of information theory

IT is a branch of communication theory that was developed from the work of Shannon [1948]. We start by reviewing the basics of this theory, especially coding theory, as it underlies the MDL principle. Of crucial importance are the links between probability distributions and code length functions.

Code and code length function

Coding theory is a branch of IT that is concerned with the design of codes, i.e., binary representations of sequences of observations. From a communication perspective, a sender and a receiver exchange messages over a communication channel. The sender encodes its original message into a bitstream, which the receiver has to decode to recover the original message. Intuitively, the goal is to achieve a short description length, while ensuring that the code is decodable and that no information is lost. At first, this problem might appear quite remote from our concerns, but we will see that there is a link between coding and learning theory.

Let us now give a more formal treatment of this notion of code Telatar [2006]. Suppose that the message to encode is a sequence of observations, whose elements, or *symbols*, $y^{(t)}$ take value in a countable set $\mathcal{Y}$, called *alphabet*. A code is a mapping

of the symbol to a sequence of binary elements or *bits*,

$$\begin{array}{rcccc} C & : & \mathcal{Y} & \to & \{0,1\}^* \\ & & y^{(t)} & \to & C(y^{(t)}). \end{array} \tag{4.1}$$

The binary representation of symbol $y^{(t)}$, $C(y^{(t)})$, is called its code word. A code has to be invertible such that the decoder can recover the original symbol from its code word. Also, a code defined over symbols is naturally extended to encode an entire sequence of symbols by encoding each symbol of the sequence individually,

$$\begin{array}{rcccc} C^T & : & \mathcal{Y}^T & \to & \{0,1\}^* \\ & & \boldsymbol{y}^{(1)>(T)} & \to & C\left(\boldsymbol{y}^{(1)>(T)}\right) = C\left(y^{(1)}\right) \ldots C\left(y^{(T)}\right). \end{array} \tag{4.2}$$

A code whose natural extension is invertible is called uniquely decodable. Moreover, uniquely decodable code are called *prefix-free* if no code word is a prefix of another code word. Prefix-free codes have the desirable properties of being instantaneously decodable. The decoder does not need to wait until it receives subsequent binary elements to disentangle which binary symbols correspond to which code words. This holds without the introduction of a special symbol representing the limit between different code words. We focus in the remainder solely on prefix-free codes.

As already mentioned, the resulting length of the code measured in bits is the key quantity of interest when designing a code. Hence, we can associate to a code its *code length function* that maps the original sequence to a natural number,

$$\begin{array}{rcccc} CL & : & \mathcal{Y} & \to & \mathbb{N} \\ & & \boldsymbol{y}^{(1)>(T)} & \to & CL\left(C(\boldsymbol{y}^{(1)>(T)})\right). \end{array} \tag{4.3}$$

The most obvious choice of code consists in describing each element of $\mathcal{Y}$ using $\log_2 M$ bits, where M is the cardinality of the alphabet. The resulting code length for a sequence of T observations, $T \log_2 M$, is an upper bound on the achievable code length. To achieve a better code length, we can exploit restrictions in the original message. If certain observations are more likely than others, we can encode them using a shorter code word such that the expected code length is reduced. Moreover, we can encode entire sequences of symbols at once, which leads potentially to more efficient solutions than the natural extension of a code defined on individual symbols. For example, suppose we encode an English text. The frequency of letters is not uniform and the vocabulary and syntax defines constraints on admissible sequences of letters. Finally, in the remainder of this thesis, we drop the integer constraint of the code length function, because this facilitates the mathematical treatment and the resulting code we design is within one bit of the constrained scheme Gruenwald [2007].

Links between code length functions and probability distributions

We now review two important links between probability distributions and code length functions, that are central for the understanding of the MDL theory Gruenwald [2005], Gruenwald [2007]. Intuitively, when designing prefix-free codes over a given alphabet $\mathcal{Y}$, we observe that we can attribute a short code length CL to a limited number of symbols or sequences. For example, observe that there can be at most one symbol with a corresponding code word of less than 1 bit, 3 symbols with a

corresponding code word of less than 2 bits, etc. Likewise, when defining probability distributions over a given alphabet $\mathcal{Y}$, we cannot associate a large probability to large number of symbols, because the probability has to sum to 1,

$$\sum_{y \in \mathcal{Y}} \mathbb{P}(y) = 1. \tag{4.4}$$

Again, observe that there can be only 1 symbol with probability greater than 1/2, 3 symbols with probability greater than 1/4, etc. This intuition can be made precise thanks to Kraft's inequality

$$\sum_{y \in \mathcal{Y}} 2^{-CL(y)} \leq 1. \tag{4.5}$$

Suppose there exists a probability distribution $\mathbb{P}$ defined on $\mathcal{Y}$. Let us define the code length function CL as

$$CL(y) = -\log_2 \mathbb{P}(y), \quad \forall y \in \mathcal{Y}. \tag{4.6}$$

Being a proper probability distribution, $\mathbb{P}$ satisfies

$$\sum_{y \in \mathcal{Y}} \mathbb{P}(y) = 1 \Rightarrow \sum_{y \in \mathcal{Y}} 2^{-CL(y)} = 1. \tag{4.7}$$

Thus the code length function satisfies Kraft's inequality and by the converse of Theorem 5.2.1 in Cover and Thomas [1991], there exists a prefix-free code whose code length is given by CL. Conversely, suppose there is a prefix-free code with code length function CL. Moreover, let us define the probability $\mathbb{P}$ as in (4.6). Then, by Theorem 2.5.1 in Cover and Thomas [1991], the code length function satisfies Kraft's inequality

$$\sum_{y \in \mathcal{Y}} 2^{-CL(y)} \leq 1 \Rightarrow \sum_{y \in \mathcal{Y}} \mathbb{P}(y) \leq 1. \tag{4.8}$$

Which means that $\mathbb{P}$ is a proper, albeit possibly degenerated, probability distribution. In summary, the use of Shannon-Fano coding Cover and Thomas [1991] allows establishing a first link between probability distributions and code length functions given by (4.6).

There exists another important link. In the previous section, we have treated probability distributions as mathematical objects, without assuming that data come from this distribution. If data y are distributed according to probability distribution $\mathbb{P}$, taking the prefix-free code, whose code length given by (4.6), is optimal in some sense. More specifically, as a consequence of the information inequality Cover and Thomas [1991],

$$\begin{aligned} \mathbb{E}_{\mathbb{P}}(CL(y)) &= \mathbb{E}_{\mathbb{P}}(-\log_2 \mathbb{P}(y)) && (4.9) \\ &\leq \mathbb{E}_{\mathbb{P}}(-\log_2 \mathbb{P}'(y)) && (4.10) \end{aligned}$$

where $\mathbb{P}'$ is any another probability distribution defined on $\mathcal{Y}$, Shannon-Fano coding minimizes the expected code length, if data are distributed according to $\mathbb{P}$.

Entropy, a measure of information

We review in this section key quantities pervasive in IT and give their coding interpretation Telatar [2006], Gruenwald [2007]. Suppose the observation y is modeled by a random variable, that takes value in a discrete alphabet $\mathcal{Y}$ and has probability mass function $\mathbb{P}(y) = \psi(y)$.The *entropy* of y is defined by

$$\begin{aligned} H(y) &= \mathbb{E}\left(-\log_2 \psi\left(y\right)\right) && (4.11)\\ &= -\sum_{y\in\mathcal{Y}} \psi\left(y\right)\log_2 \psi\left(y\right). && (4.12)\end{aligned}$$

Note that the entropy is a function of the distribution ψ and not of the random variable y itself. Given the link between probability distributions and code length functions given by (4.6), the entropy represents the expected code length when encoding y using the prefix-free code whose code length is given by $-\log_2 p(y)$. The more the model p is restrictive about y, the smaller the entropy is. Moreover, we can define the relative entropy or Kullback-Leibler distance between two probability mass functions p and p' as

$$\begin{aligned} D\left(\psi,\psi'\right) &= \mathbb{E}\left(-\log_2 \psi'\left(y\right)\right) - \mathbb{E}\left(-\log_2 \psi\left(y\right)\right) && (4.13)\\ &= -\sum_{y\in\mathcal{Y}} \psi\left(y\right)\log_2 \psi'\left(y\right) + \sum_{y\in\mathcal{Y}} \psi\left(y\right)\log_2 \psi\left(y\right) && (4.14)\\ &= -\sum_{y\in\mathcal{Y}} \psi(y)\log_2 \frac{\psi'\left(y\right)}{\psi\left(y\right)}. && (4.15)\end{aligned}$$

Again we can give the following coding interpretation. Suppose that the data are effectively generated by drawing from distribution ψ. The differential entropy then measures the additional number of bits need to encode the data using distribution ψ' instead of the optimal ψ.

If y is a continuous random variable, $\psi(y)$ defines a probability distribution function and, provided it exists, we can define the differential entropy

$$\begin{aligned} h(y) &= \mathbb{E}\left(-\log_2 \psi\left(y\right)\right) && (4.16)\\ &= -\int_y p(y)\log_2 \psi(y)dy. && (4.17)\end{aligned}$$

Unlike the standard entropy, the differential entropy can be negative. Also, there exists an important link between the differential entropy of a continuous random variable and the discrete random variable obtained by quantizing y at precision q Cover and Thomas [1991]

$$H\left(Q(y)\right) = h(y) - \log_2(q). \tag{4.18}$$

where Q denotes the quantization function

$$Q(y) = y' \text{ if } y \in [y' - q/2;\, y' + q/2). \tag{4.19}$$

Therefore, the differential entropy has a similar coding interpretation as the entropy.

4.1.2 Model selection and MDL principle

We review in this section the MDL principle, a general principle for model selection rooted in the theory of coding. Central for its understanding are the links between code length functions and probability distributions we have reviewed in the previous section. Generally speaking, model selection is concerned with the choice of the "best" explanation of observed data, among a set of competing ones. For example, if we model a set of observations with an autoregressive process, the question is how many lags should we include in the model. Similarly, in a linear regression model, the question is which subset of regressors should we use. From a finance practitioner point of view, the main goal is certainly not to recover the correct order of a process. Indeed, as seen in Chapter 2 the latter most likely does not exist. And even if we assumed the data were distributed according to some true complex distribution, it would be better to use a simpler model, especially if the complex model is incidental, i.e., applies to few observations. The aim is rather to avoid overfitting and select models with good out-of-sample predictability.

Origin of MDL

Let us start by reviewing the origin of MDL Rissanen [2005]. The traditional approach in parametric estimation theory starts with the assumption of the existence of an underlying true distribution, according to which observed data are distributed. This distribution depends of a l-dimensional vector of parameters $\boldsymbol{\theta}_l$, which are inferred from the observations either by maximization of the (log-)likelihood function or by the minimization of the norm of the residuals. For example, the least-squares (resp. least absolute deviation) method minimizes the l_2 (resp. l_1) norm of the residuals. The typical problem with this approach is that, when we increase the number of parameters, the resulting model fits the data better and one runs in the disastrous conclusion that the best model is also the most complex one.

From an historical perspective, this problem was recognized early. But it was only addressed using ad hoc methods, for example, in linear regression by choosing the subset of regressors maximizing the standardized coefficient of determination Harvey [1991]. Intuitively, the general idea consists in introducing a penalty for the additional parameters, which compensates for the improvement in fit. Akaike [1974] was the first to address the problem formally. Akaike's analysis is based on the assumption of the existence of a true distribution, which generates data. The goal is to choose among a collection of models the best one, under the assumption that the true model lies outside this collection. Akaike's key idea it to measure the quality of the model by the information theoretic Kullback-Leibler distance between the model and the true distribution. This distance has to be estimated from the data. It turns out to be difficult, but, quite surprisingly, there exists a very elegant and simple answer for the asymptotic mean Kullback-Leibler distance, where the mean is taken over all models with the same number of parameters. The resulting AIC is given by

$$AIC(l) = 2l - 2LLF(\boldsymbol{y}^{(1)>(T)}/\boldsymbol{\theta}_l) \tag{4.20}$$

where l is the number of parameters and $LLF(\ldots)$ the log-likelihood function. Unfortunately, AIC suffers from several limitations. First, it is only an asymptotic result and nothing can be said about its finite sample properties. AIC is also based on the

assumption of the existence of a true distribution. More worryingly, AIC is not consistent when the true model lies inside the collection of models under consideration. Consistency pertains to the desirable feature of a statistical method that reaches the correct answer given a sufficiently large number of observations. Similarly to AIC, BIC is given by Schwartz [1978]

$$BIC(l) = l \log(T) - 2LLF(\boldsymbol{y}^{(1)>(T)}/\boldsymbol{\theta}_l) \tag{4.21}$$

where T is the number of observations. Again, this is only an asymptotic result. Also, note that in both AIC and BIC, the complexity is directly proportional to the dimensionality of the vector of parameters L. This is not the case in general and we are going to see in MDL that models with the same number of parameters may have a different complexity.

The ultimate answer to the problem of model selection is given by Kolmogorov's and Solomonoff's algorithmic complexity theory Cover and Thomas [1991]. In this theory, computer programs, that write data and stop, serve as code word. They are described using a general purpose programming language. Also, since a computer program cannot start by itself, a program that stops before another one is not its prefix. Therefore, computer programs define a prefix-free code and their associated code lengths a complexity measure. Moreover, the choice of programming language does not affect the model selection criterion because if two programming languages are used to describe the same data, the length of their programs asymptotically differs only by a constant, albeit potentially large. This invariance principle removes the arbitrariness of the choice of programming language. There is however two problems with Kolmogorov's complexity theory. First, the invariance principle is only valid asymptotically but does not apply on small samples. More importantly, Kolmogorov's complexity is noncomputable, i.e., there exists no automated procedure to find the shortest computer program given data or even its length. Algorithmic complexity theory establishes a link between learning and data compression, but it also put a serious limit on what could be achieved by a model selection scheme. There is no automated procedure to find the best model when the latter is expressed in the most general terms using computer programs. The next best thing to do is to restrict the choice of coding schemes and define a model selection criterion valid for this restricted class. This is the idea behind Rissanen's MDL principle.

Crude version of MDL

MDL is a general framework to address the problem of model selection, first proposed by Rissanen [1978], Rissanen [2005], Gruenwald [2005]. The central idea of Rissanen's theory is inspired by Kolmogorov's algorithmic complexity as it equates the problem of learning with data compression. Intuitively, learning means finding regularities in observed data and these regularities are used by a coding scheme to compress the sequence of observations. Also, more regularities means more compression, and purely random sequences are the least compressible. In algorithmic complexity, computer programs are used to devise a prefix-free code and a notion of complexity is chosen to be the length of the shortest computer program. In order not to fall back into the noncomputability problem, computer programs are replaced by a less general class of probability distributions or equivalently statistical models.

The MDL principle, in a first approximation, advocates to pick the model with the smallest combined description length of the model $\mathcal{M}$ and the data using the model, informally

$$\underset{\mathcal{M}}{\operatorname{argmin}} \left\{ CL(\mathcal{M}) + CL(\boldsymbol{y}^{(\mathbf{1})>(\boldsymbol{T})}/\mathcal{M}) \right\}. \tag{4.22}$$

Given the link between code length functions and probability distributions reviewed in Section 4.1.1, the second term is given by

$$CL(\boldsymbol{y}^{(\mathbf{1})>(\boldsymbol{T})}/\mathcal{M}) = -\log_2 \mathbb{P}(\boldsymbol{y}^{(\mathbf{1})>(\boldsymbol{T})}), \tag{4.23}$$

where $\mathbb{P}$ is the probability distribution corresponding to the model $\mathcal{M}$. We typically work in the framework of parametric estimation theory, where the model depends on a l-dimensional vector of parameters

$$\boldsymbol{\theta}_l \in \boldsymbol{\Theta}_l \subset \mathbb{R}^l. \tag{4.24}$$

Let us denote $\hat{\boldsymbol{\theta}}_l$ the maximum likelihood estimates of $\boldsymbol{\theta}_l$,

$$\hat{\boldsymbol{\theta}}_l = \underset{\boldsymbol{\theta}_l}{\operatorname{argmax}} \mathbb{P}\left(\boldsymbol{y}^{(1)>(T)}/\boldsymbol{\theta}_l\right). \tag{4.25}$$

Therefore, the term that minimizes the code length function given the model (4.23) corresponds to the opposite of the maximum log-likelihood function (4.25), and we recognize a standard measure of fit for a model.

With MDL, there is no such things as a true distribution or not. A model is just used to express properties about data, like the vocabulary and the syntax of a language. Of course, this could be done at different precision levels. Certain models allow expressing only coarse properties, while others allow for a more precise description, up to the point of capturing "noise". Consider the following example to illustrate this point. Suppose that we have to encode the very long sequence of alternating 0 and 1, 01010101010…. On the one hand, we could use a Bernouilli model, i.e., a family of i.i.d. Bernouilli random variable parameterized by $\wp = \mathbb{P}\left(y^{(t)} = 0\right)$. Then,

$$\begin{aligned} \mathbb{P}\left(\boldsymbol{y}^{(1)>(T)}\right) &= \prod_{t=1}^{T} \mathbb{P}\left(y^{(t)}\right) && (4.26) \\ &= \wp^{\sharp \text{ of 0's}} (1-\wp)^{\sharp \text{ of 1's}} && (4.27) \end{aligned}$$

The achieved compression is small because the sequence is almost random. On the other hand, we could use a Markov model of order 1 with

$$\mathbb{P}\left(y^{(t)} = 0/y^{(t-1)} = 1\right) = \mathbb{P}\left(y^{(t)} = 0/y^{(t-1)} = 1\right) = 1, \tag{4.28}$$

The model perfectly describes data, a high compression is achieved and there is no noise. The ultimate difficulty is of course to come up with a good choice of models and model classes, but this is left to humans, based on their experience or knowledge they have on the process at hand. This is anyway the best thing to do, since no computable automated procedure can be devised for model selection.

Remark. Note that we use in the remainder of this thesis the term model in an information theoretic sense, which differs from its statistical meaning. A model

corresponds to a unique probability distribution, and in a parametric estimation theory framework to a single value of parameters θ_l. Statisticians call an information theoretic model a point hypothesis. A model class is a set of models with a similar functional form, for example the autoregressive process of order k, which corresponds to a joint hypothesis or a model in a statistical sense.

Universal model

In our first crude version of the MDL principle, we have surreptitiously ignored the problem of encoding the code length of the model $CL(\mathcal{M})$. In modern version of MDL, this is addressed with the use of *universal models*, which we review now Cover and Thomas [1991], Gruenwald [2005], Gruenwald [2007]. The universal model for a class of models is defined as the unique representative model that is able to represent well any models in that class. The term universal model is quite misleading, as it corresponds to a unique probability distribution and it is only universal with respect to a certain class of models. Let us make this idea more precise. Suppose the model depends on a l-dimensional vector of parameters. Also, suppose the maximum likelihood estimator of the parameters $\hat{\boldsymbol{\theta}}_l$, a function of the sequence of observations $\boldsymbol{y}^{(1)>(T)}$, exists and is unique. Our definition of the universal model translates into the following mathematical condition. For any realization $\boldsymbol{y}^{(1)>(T)}$, if the code length of the best model in hindsight

$$-\log_2 \mathbb{P}\left(\boldsymbol{y}^{(1)>(T)}/\hat{\boldsymbol{\theta}}_l\left(\boldsymbol{y}^{(1)>(T)}\right)\right) \tag{4.29}$$

is small, in other words if the model fits the data well, the code length of the universal model CL_u is small as well. The difficulty comes from the fact that the choice of universal model should happen before observing the actual sequence $\boldsymbol{y}^{(1)>(T)}$. Otherwise, this does not define a coding scheme, as the decoder is not able to perform the same optimization as the encoder, which necessitates the knowledge of the sequence.

Is it possible to achieve a code whose code length is as good as the best in hindsight code for all realizationa of the sequence? The answer is no. Indeed suppose such a code exists. Its code length function CL_u, with associated probability $\mathbb{P}_u$, satisfies

$$\begin{aligned} CL_u\left(\boldsymbol{y}^{(1)>(T)}\right) &= -\log_2 \mathbb{P}_u\left(\boldsymbol{y}^{(1)>(T)}\right) && (4.30)\\ &\leq -\log_2 \mathbb{P}\left(\boldsymbol{y}^{(1)>(T)}/\hat{\boldsymbol{\theta}}_l\left(\boldsymbol{y}^{(1)>(T)}\right)\right). && (4.31) \end{aligned}$$

Then

$$\begin{aligned} \sum_{\boldsymbol{y}^{(1)>(T)}} \mathbb{P}_u\left(\boldsymbol{y}^{(1)>(T)}\right) &\geq && (4.32)\\ \sum_{\boldsymbol{y}^{(1)>(T)}} \mathbb{P}\left(\boldsymbol{y}^{(1)>(T)}/\hat{\boldsymbol{\theta}}_l\left(\boldsymbol{y}^{(1)>(T)}\right)\right) &= && (4.33)\\ \sum_{\boldsymbol{y}^{(1)>(T)}} \max_{\boldsymbol{\theta}_l} \mathbb{P}\left(\boldsymbol{y}^{(1)>(T)}/\boldsymbol{\theta}_l\right) &> 1 && (4.34)\\ & && (4.35) \end{aligned}$$

Thus, as the corresponding $\mathbb{P}_u$ does not satisfy the property of a distribution, its associated code length CL_u cannot be the code length of a prefix-free code. The next best thing we can ask for is to achieve a performance close to the best in hindsight. This leads to the definition of a universal model in the individual sequence sense as a model whose code length is within a constant, albeit large, of the best solution in hindsight. Mathematically,

$$CL_u\left(\boldsymbol{y}^{(1)>(T)}\right) = -\log_2 \mathbb{P}\left(\boldsymbol{y}^{(1)>(T)}/\hat{\boldsymbol{\theta}}_l\left(\boldsymbol{y}^{(1)>(T)}\right)\right) + cst \tag{4.36}$$

where the constant *cst* is required to grow sublinearly in the number of observations T.

Example 7. *In Bayesian statistics, the parameters $\boldsymbol{\theta}_l$ is assumed to be distributed according to a prior distribution $\mathbb{P}_{prior}$.The distribution of the observations is then given by*

$$\mathbb{P}_{Bayes}\left(\boldsymbol{y}^{(1)>(T)}\right) = \sum_{\boldsymbol{\theta}_l} \mathbb{P}\left(\boldsymbol{y}^{(1)>(T)}/\boldsymbol{\theta}_l\right)\mathbb{P}_{prior}\left(\boldsymbol{\theta}_l\right) \tag{4.37}$$

Thus the corresponding code length satisfies $\forall \boldsymbol{\theta}_l$

$$\begin{aligned} CL_{Bayes} &= -\log_2 \mathbb{P}_{Bayes}\left(\boldsymbol{y}^{(1)>(T)}\right) && (4.38)\\ &= -\log_2 \sum_{\boldsymbol{\theta}_l} \mathbb{P}\left(\boldsymbol{y}^{(1)>(T)}/\boldsymbol{\theta}_l\right)\mathbb{P}_{prior}\left(\boldsymbol{\theta}_l\right) && (4.39)\\ &\overset{(a)}{\leq} -\log_2 \mathbb{P}\left(\boldsymbol{y}^{(1)>(T)}/\boldsymbol{\theta}_l\right)\mathbb{P}_{prior}\left(\boldsymbol{\theta}_l\right) && (4.40)\\ &\overset{(b)}{=} -\log_2 \mathbb{P}\left(\boldsymbol{y}^{(1)>(T)}/\boldsymbol{\theta}_l\right) - \log_2 \mathbb{P}_{prior}\left(\boldsymbol{\theta}_l\right). && (4.41) \end{aligned}$$

(a) holds because a sum of positive terms is at least as large as one of its terms and the negative logarithmic $-\log_2$ is monotonically decreasing and (b) is true for all values of the parameters, in particular for the maximum likelihood estimator $\boldsymbol{\theta}_l = \hat{\boldsymbol{\theta}}_l$. Therefore, the Bayesian model is a universal model.

Refined version of MDL

In the refined version of MDL, a universal model for the class of models under consideration is used to encode simultaneously the model and the data given the model Gruenwald [2007], Rissanen [2005]. We have just seen one example of universal model, but there exist others. Different universal models for a given class lead to different MDL-based model selection criteria. One way of building universal models is by solving an optimization problem. For example, let us define the regret of a universal model $\mathbb{P}_u$ as

$$R = -\log_2 \mathbb{P}_u\left(\boldsymbol{y}^{(1)>(T)}\right) + \log_2 \mathbb{P}\left(\boldsymbol{y}^{(1)>(T)}/\hat{\boldsymbol{\theta}}_l\left(\boldsymbol{y}^{(1)>(T)}\right)\right). \tag{4.42}$$

The regret represents the extra number of bits necessary to encode the sequence compared to the best fitting solution in hindsight Rissanen [2005]. Let us find the universal model that solves the following min max optimization

$$\min_{\mathbb{P}_u} \max_{\boldsymbol{y}^{(1)>(T)}} R. \tag{4.43}$$

This can be interpreted as a game against Nature, where the experimenter chooses $\mathbb{P}_u$ and Nature chooses the sequence such that the experimenter incurs the largest possible regret. Goal is to minimize the worst case regret, which is the best solution in the individual sequence sense. The solution to this optimization, known as the NML or Shtarkov distribution, is given by

$$\mathbb{P}_{NML} = \frac{\mathbb{P}\left(\boldsymbol{y}^{(1)>(T)}/\hat{\boldsymbol{\theta}}_l\left(\boldsymbol{y}^{(1)>(T)}\right)\right)}{\sum_{\boldsymbol{y}^{(1)>(T)}}\mathbb{P}\left(\boldsymbol{y}^{(1)>(T)}/\hat{\boldsymbol{\theta}}_l\left(\boldsymbol{y}^{(1)>(T)}\right)\right)}. \tag{4.44}$$

The denominator is called the parametric complexity. NML achieves constant maximum regret, i.e., for all sequences $\boldsymbol{y}^{(1)>(T)}$ the maximum regret is equal to the parametric complexity

$$\max_{\boldsymbol{y}^{(1)>(T)}} R = \sum_{\boldsymbol{y}^{(1)>(T)}} \mathbb{P}\left(\boldsymbol{y}^{(1)>(T)}/\hat{\boldsymbol{\theta}}_l\left(\boldsymbol{y}^{(1)>(T)}\right)\right). \tag{4.45}$$

Despite its attractiveness, NML suffers from certain limitations. In particular, the sum or equivalently integral for continuous random variables in the denominator might not be finite, in which case the NML does not exist, and another criterion should be used. This happens for large classes of models, e.g., for the class of linear regression models with Gaussian disturbance.

Using MDL for model selection presents several advantages over other approaches. First of all, MDL is based on sound philosophical foundations, as it does not assume the existence of a true distribution according to which the data are distributed. This is of particular interest in finance where the existence of an underlying true distribution is rather unlikely to exist. In MDL, there is no such thing as a correct or a wrong model. A model or a model class is simply used to express properties about data and MDL offers a fair mean of comparison between them. Also, since the complexity is defined as the code length function associated with a probability distribution, the unit is bits. This allows comparing models and model classes that have different structure or number of parameters. Furthermore, the only principle underlying MDL is Occam's razor, a very general principle found throughout sciences and engineering that advocates to use the simplest best explanation. Secondly, MDL selection criteria are developed in the individual sequence sense. This is again very important in finance where we have access to a single realization of the process and do not want to make the assumption of ergodicity of financial markets. Moreover, in some cases, the formula for the constant appearing in the universal model definition is exact on small samples and we can be confident about the selection decision on small samples as well. Finally, MDL is related to the Bayesian approach but it avoids some of its interpretation problems, in particular when the experimenter has no knowledge on the underlying prior distribution.

MDL in finance

As already noted, MDL-based model selection criteria present several advantages. It is surprising to note they have not been widely adopted by the finance research community. A notable exception is the predictive least-squares (PLS) criterion, that

is used, e.g., in the paper of Bossaerts and Hillion [1999]. This paper studies various statistical model selection criteria to select the best out-of-sample set of predictors of stock returns among the three best predictors of stock returns. PLS belongs to the more general family of plug-in models. Consider the class of models of order k parameterized by the l-dimensional vector of parameters $\boldsymbol{\theta}_l$. Let m be the minimum number of observations such that the maximum likelihood $\hat{\boldsymbol{\theta}}_l$ exists. Then, a universal model for that class is given by the plug-in model

$$\mathbb{P}_{PLS}\left(\boldsymbol{y}^{(1)>(T)}\right) = \mathbb{P}_0\left(y^{(1)>(m)}\right)\prod_{t=m+1}^{T}\mathbb{P}\left(y^{(t)}/\hat{\boldsymbol{\theta}}_l\left(\boldsymbol{y}^{(1)>(t-1)}\right)\right), \tag{4.46}$$

This is called a plug-in model because we are plugging in as value of the parameters the maximum likelihood estimator based on observations up to time $t-1$. Also, PLS is a prequential scheme, a term coined by Dawid [1984] to describe a model which is sequentially formed using information up to time $t-1$ and evaluated against observation at time t, $\forall t = m+1, \dots T$. The PLS criterion corresponds to the accumulated log-loss of the prequential model where the loss function is measured by the density function

$$\mathbb{P}\left(y^{(t)}/\hat{y}^{(t)}\right). \tag{4.47}$$

Intuitively, if the prediction of the model is good (resp. bad), the distribution $\mathbb{P}(y^{(t)}/\hat{y}^{(t)})$ is large (resp. small) and the corresponding code length small (resp. large). The idea is also related to cross-validation, where a portion of the information is used to estimate the model and another one to test the model. But unlike cross-validation, the prequential approach works sequentially and a point is predicted only once. Note also that another estimator than the maximum likelihood one can be used to define a plug-in distribution. Moreover, for the class of linear regression models with Gaussian disturbance, the plug-in model defines a universal model, called PLS

$$PLS(k) = \sum_{t=m+1}^{T}\left(y^{(t)} - \left(\boldsymbol{f}^{(t)}\right)^T\hat{\boldsymbol{\theta}}_l^{(t-1)}\right)^2, \tag{4.48}$$

where $\boldsymbol{f}^{(t)}$ is the vector of regressors. It is well known that PLS has a poor performance in practice Wei [1992], Rissanen et al. [2010]. But this is not a general feature of MDL-based model selection criteria. Indeed, we can build different model selection criteria, each corresponding to a different universal model for the class of models under consideration, with different associated performance.

We have introduced MDL with the aim of finding models with good out-of-sample performance in terms of prediction. We have phrased the problem using models, their corresponding probability distributions and associated code length functions. Somehow this seems quite remote from the concerns of a finance practitioner, whose ultimate goal is to design strategies with good out-of-sample performance. We see that this is not the case. Firstly, it is tempting to use standard measure of performance of the strategy, like the Sharpe ratio, to select the best strategy. But there is no theoretically sound approach to take into account the complexity of the model underlying the strategy and the resulting method could only be ad hoc. Secondly, the

strategy of a good investor should inspire a good data compressor, and conversely. On the one hand, Cover and Thomas [1991] develops a coding scheme where the log-wealth of the strategy betting on a sequence of outcome is used as codeword for that sequence, such that a good gambler is also a good data compressor. On the other hand, Kelly's gambling establishes another link between data compression and gambling Kelly [1956], Cover and Thomas [1991]. In gambling, the goal is to maximize the growth rate of the investment, given by the expected value of the terminal log-wealth at time T. This is achieved by Kelly's gambling, that bets on each outcome proportionally to the conditional distribution on that outcome conditional on side information, for example past observations. Hence, Kelly's gambling is also called proportional gambling. Therefore, a good predictor, in other words a good compressor, is also a good gambler. Finally, observe that the growth rate of the investment, the measure being maximized by Kelly's gambling, corresponds to the opposite of the code length of the distribution, the measure being minimized in MDL.

4.2 Variable lookback algorithm

Following our review of MDL in the previous section, which underlies our proposed solution, we are now ready to present the variable lookback algorithm. Remember that it is only based on the assumption of piecewise stationarity, that replaces the stationarity assumption. This means there exist a series of switching times $t_0 = 1 < t_1 < \ldots < t_S \leq T$ such that the data are represented by a model of order $k_s \in \{0, \ldots, K\}$ in the interval $[t_{s-1}; t_s)$. Recall that there are no assumption on the switching process, the number of switching times S or their positions t_s.

4.2.1 Basics of the algorithm

In this section, we describe a coarse version of the algorithm, whereas additional details are treated in the next section.

Coding of observations

We consider first the problem of selecting the order of the process $k_s \in \{0, \ldots, K\}$ within a period of piecewise stationarity $t \in [t_s, t_{s+1})$. Based on our review of MDL, we have to design a universal code for the class of processes of order up to K to encode the sequence of observations within that period. Also, the code length of this scheme serves as a measure of complexity of the model and the MDL principle advocates to pick the model with the shortest code length. We choose to use an existing universal model, the recently proposed conditional normalized maximum likelihood (CNML) Rissanen et al. [2010]. We describe it briefly. Consider the class of parametric models of order k parameterized by a l-dimensional vector of parameters $\boldsymbol{\theta}_l$. Let m be the smallest number of observations $t - t_s + 1$ such that the maximum likelihood $\hat{\boldsymbol{\theta}}_l \left(y^{(t_s)>(t-t_s+1)} \right)$ can be computed uniquely. Let us also define the two sequences $\boldsymbol{y}_0 = y^{(t_s)>(t_s+m-1)}$ and $\boldsymbol{y}_1 = y^{(t_s+m)>(t_s+t-1)}$. Conditional on $\boldsymbol{y}_0$, the

CNML criterion is given by

$$\mathbb{P}_{CNML}(\boldsymbol{y}_1/\boldsymbol{y}_0,k)=\frac{\mathbb{P}\left(\boldsymbol{y}_0,\boldsymbol{y}_1/\hat{\boldsymbol{\theta}}_l(\boldsymbol{y}_0,\boldsymbol{y}_1)\right)}{\sum_{\boldsymbol{y}_1}\mathbb{P}\left(\boldsymbol{y}_0,\boldsymbol{y}_1/\hat{\boldsymbol{\theta}}_l(\boldsymbol{y}_0,\boldsymbol{y}_1)\right)}. \tag{4.49}$$

The CNML was obtained as a solution of an optimization problem, namely the minimization of the maximum conditional regret, mathematically

$$\min_{\bar{\mathbb{P}}}\max_{\boldsymbol{y}_1}\left\{\ -\log_2\left(\bar{\mathbb{P}}\left(\boldsymbol{y}_1/\boldsymbol{y}_0\right)\right)+\ \log_2\left(\mathbb{P}_{\hat{\boldsymbol{\theta}}_l(\boldsymbol{y}_0,\boldsymbol{y}_1)}\left(\boldsymbol{y}_0,\boldsymbol{y}_1\right)\right)\ \right\} \tag{4.50}$$

The term inside the curly brackets, called conditional regret, measures the difference in bits between coding the sequence $\boldsymbol{y}_1$ using probability $\bar{\mathbb{P}}$ and coding the same sequence using the best fitting model in hindsight, corresponding to the maximum likelihood estimator of $\hat{\boldsymbol{\theta}}_l$. Several reasons motivate the choice of the CNML criterion. First, the criterion minimizes the maximum conditional regret and is thus a universal model in the individual sequence sense. Second, it is a very practical MDL-based model selection criterion. It not only avoids the problem of infinite complexity of the NML criterion, but it is considerably easier to evaluate. Indeed, the integral term in the denominator of (4.50) can be evaluated analytically for large classes of processes, in particular, for the class of linear regression models with Gaussian or Laplace residuals Rissanen et al. [2010]. Note also that, unlike NML, the CNML criterion defines a prequential coding scheme. This can be verified by showing that the probability distribution associated with the CNML criterion satisfies

$$\begin{cases}\mathbb{P}_{CNML}\left(-/\boldsymbol{y}_0,k\right)=1\\ \mathbb{P}_{CNML}\left(\boldsymbol{y}_1/\boldsymbol{y}_0,k\right)>0,\ \forall\boldsymbol{y}_1\\ \sum_{y^{(t_s+t)}}\mathbb{P}_{CNML}\left(\boldsymbol{y_1},y^{(t_s+t)}/\boldsymbol{y}_0,k\right)=\mathbb{P}_{CNML}\left(\boldsymbol{y_1}/\boldsymbol{y}_0,k\right).\end{cases} \tag{4.51}$$

Let us verify the last consistency condition.

$$\sum_{y^{(t_s+t)}}\mathbb{P}_{CNML}\left(\boldsymbol{y_1},y^{(t_s+t)}/\boldsymbol{y}_0,k\right)\overset{(a)}{=} \tag{4.52}$$

$$\sum_{y^{(t_s+t)}}\mathbb{P}_{CNML}\left(y^{(t_s+t)}/\boldsymbol{y}_0,\boldsymbol{y_1},k\right)\mathbb{P}_{CNML}\left(\boldsymbol{y}_1/\boldsymbol{y}_0,k\right)\overset{(b)}{=} \tag{4.53}$$

$$\mathbb{P}_{CNML}\left(\boldsymbol{y}_1/\boldsymbol{y}_0,k\right)\sum_{y^{(t_s+t)}}\mathbb{P}_{CNML}\left(y^{(t_s+t)}/\boldsymbol{y}_0,\boldsymbol{y_1},k\right)\overset{(c)}{=} \tag{4.54}$$

$$\mathbb{P}_{CNML}\left(\boldsymbol{y}_1/\boldsymbol{y}_0,\boldsymbol{y_1},k\right)\frac{\sum_{y^{(t_s+t)}}\mathbb{P}\left(\boldsymbol{y}_0,\boldsymbol{y}_1,y^{(t_s+t)}/\hat{\theta}\left(\boldsymbol{y}_0,\boldsymbol{y}_1,y^{(t_s+t)}\right)\right)}{\sum_{y^{(t_s+t)}}\mathbb{P}\left(\boldsymbol{y}_0,\boldsymbol{y}_1,y^{(t_s+t)}/\hat{\theta}\left(\boldsymbol{y}_0,\boldsymbol{y}_1,y^{(t_s+t)}\right)\right)}= \tag{4.55}$$

$$\mathbb{P}_{CNML}\left(\boldsymbol{y}_1/\boldsymbol{y}_0,k\right). \tag{4.56}$$

(a) corresponds to the definition of conditional expectation $\mathbb{P}(A/B)=\mathbb{P}(A,B)/\mathbb{P}(B)$. (b) holds because $\mathbb{P}_{CNML}\left(\boldsymbol{y}_1/\boldsymbol{y}_0,k\right)$ does not depend on $y^{(t+t_s)}$. (c) is obtained by using the definition of the CNML criterion (4.50).

Remark. Consider two classes of models with a similar functional form, one of order k_1, the other of order k_2. Let m_1 and m_2 be the respective minimum number of observations such that the maximum likelihood can be computed. It is important to compare these two model classes using the same number of observations, thus $m=\max_{i=1,2}\{m_i\}$. Otherwise the model selection criterion does not correspond to the code length of a prefix-free code Gruenwald [2007].

Coding of switches

Let us now introduce switches between periods of piecewise stationarity. Let us call pattern of switches the collection of switching times $t_0 = 1 < t_1 < t_2 < \ldots < t_s \leq T$. The central idea behind the variable lookback algorithm is to use a two-part code to jointly encode the observations and the pattern of switches. Given a certain pattern of switches, we have just seen that the CNML criterion is used to encode observations over each period of piecewise stationarity. Let us now describe how we encode the pattern of switches. Remember that we make neither assumption on the number of switches in the system nor on the positions at which they happen. To each pattern of switches, we can associate a binary sequence that contains a 1 where a switch occurs and a 0 otherwise. Mathematically,

$$z^{(t)} = \begin{cases} 1 & \text{if} \quad t \in \{t_0, \ldots, t_s\} \\ 0 & \text{otherwise.} \end{cases} \tag{4.57}$$

One candidate for modeling a binary sequence is the Bernouilli process, i.e., a sequence of i.i.d. Bernouilli distributed random variable parameterized by the probability $\wp$ that a zero occurs

$$\mathbb{P}\left(z^{(t)} = 0\right) \sim i.i.d\, \mathcal{B}(\wp). \tag{4.58}$$

In this case $\theta = \wp$ and the maximum likelihood estimator $\hat{\theta}$ is given by

$$\hat{\theta}\left(z^{(1)>(t)}\right) = \frac{\sharp \text{ of 0's in } \boldsymbol{z}^{(1)>(t)}}{t}. \tag{4.59}$$

which is not defined at $t = 0$. We need to associate to a pattern of switches a measure of complexity, and the MDL theory suggests to use the code length of a universal model for that class of processes. We choose to use the Krichevsky-Trofimov (KT) Krichevsky and Trofimov [1981] probability distribution

$$\mathbb{P}_{KT}\left(z^{(t+1)} = 0/\boldsymbol{z}^{(1)>(t)}\right) = \frac{\sharp \text{ of 0's in } \boldsymbol{z}^{(1)>(t)} + 1/2}{t+1}, \tag{4.60}$$

which is a universal model for the class of Bernouilli process. Observe that the KT estimator is a prequential plug-in model, where the maximum likelihood estimator is replaced by the maximum likelihood estimator of the original dataset augmented by half a 0 observation and half a 1 observation. This avoids the problem of infinity of the standard maximum likelihood estimator. Finally, note that the fact that we use a universal model for the class of Bernouilli model does not imply that we assume that $z^{(t)}$ follows a Bernouilli process.

We can associate to each pattern of a node in a quadratic tree diagram as in Willems [1996]. This allows organizing in a structured manner the different pattern of switches in the variable lookback algorithm. A node in the corresponding quadratic tree is parameterized by three variables:

(i) $\boldsymbol{t}$ **:** the time of the current observation,

(ii) $\boldsymbol{s}$ **:** the number of switches so far,

(iii) $\boldsymbol{t_s}$ **:** the time of the first observation in the current period of piecewise stationarity.

A portion of the quadratic tree is represented in Figure 4.1 for the first 4 iterations of the algorithm. By default, when we move up in the tree, we introduce a switch and when we move down along a branch, there is no switch. Each node in the tree is labeled with a visual representation of the corresponding segmentation of the sequence in periods of piecewise stationarity. Each square represents a different observation and the shade of grey the model used to encode the observation. For example, consider node $(2, 1, 2)$, that has only two observations. The first observation is associated with one period of piecewise stationarity, the second one with another period of piecewise stationarity. Similarly, consider node $(4, 2, 3)$, that has 4 observations. The first one corresponds to one period of piecewise stationarity, the second one to another one and the last two to yet another one. On the one hand, the number of models grows linearly in the number of observation t. This can be checked visually for example at the fourth iteration where there are only four different shades of grey representing all model conditional on information from time 1, 2, 3, and 4, respectively, up to time 4. On the other hand, the number of nodes grows in the the square of the number of observations. To be precise, the number of nodes at time t is given by

$$\frac{t(t-1)}{2} + 1 \tag{4.61}$$

This is significantly smaller than 2^t, because several paths in the tree lead to the same node. See, for example node $(4, 2, 4)$ in Figure 4.1.

Block diagram

We are now ready to present our algorithm, whose block diagram is depicted in Figure 4.2. The top block represents an idealized data generating process. The signal $y^{(t)}$ is obtained by filtering a white noise process, to which an additive white noise $e^{(t)}$ is added. The two white noise sources are uncorrelated. A switch controls the choice of the filter such that only one is active at every time t. Note that we make no assumption on the process controlling the switch (e.g., exponential switching process) so as to obtain a nonstationary output. The bottom block represents the estimation and model selection procedure. The series of inverse lag operators allow aligning the sequence so as to obtain models depending on different portions of the information set. For example the first one on the top one selects all information from time 1 to t, the second one from 2 to t, etc. The estimation is conducted by a series of $T - 1$ adaptive filters. Each is based on a different information set; in particular, filter i is based on information from time i to t. One of the output of the adaptive filter, the estimated residuals of the model, controls its adaptation, which is depicted by the arrow from the output of the filter and crossing the filter box diagonally. The outputs of the adaptive filters serve as inputs to a dynamic programming algorithm, the Viterbi algorithm, which determines an estimate of the order of the process $\hat{k}^{(t)}$ and the last time before which a switch happens $\hat{t}_s^{(t)}$, conditional on information at time t.

Let us now describe the operations of the dynamic programming algorithm Viterbi [1967]. The algorithm is best described using a trellis diagram, which, in this case,

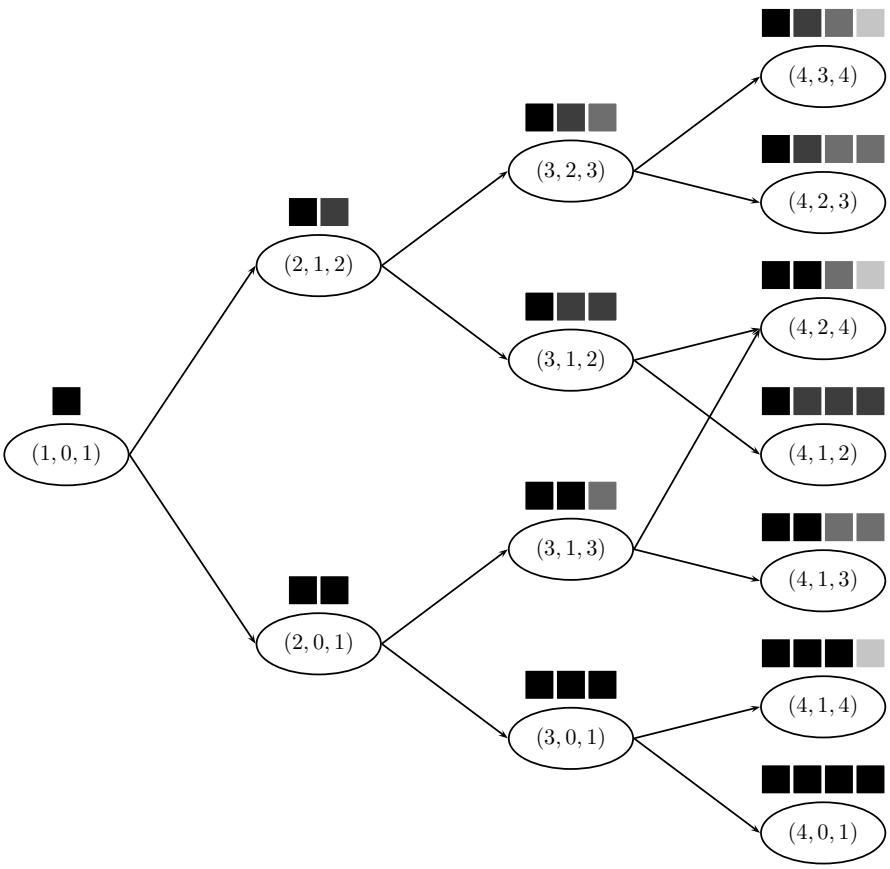


Figure 4.1: *Portion of the quadratic tree diagram for the first 4 iterations of the algorithm. Each node in the tree is parameterized by three variables: t, the time of the current observation, s, the number of switches so far, and t_s, the last time before which a switch happens. Nodes of the quadratic tree are used to maintain in an organized manner competing explanations of data.*

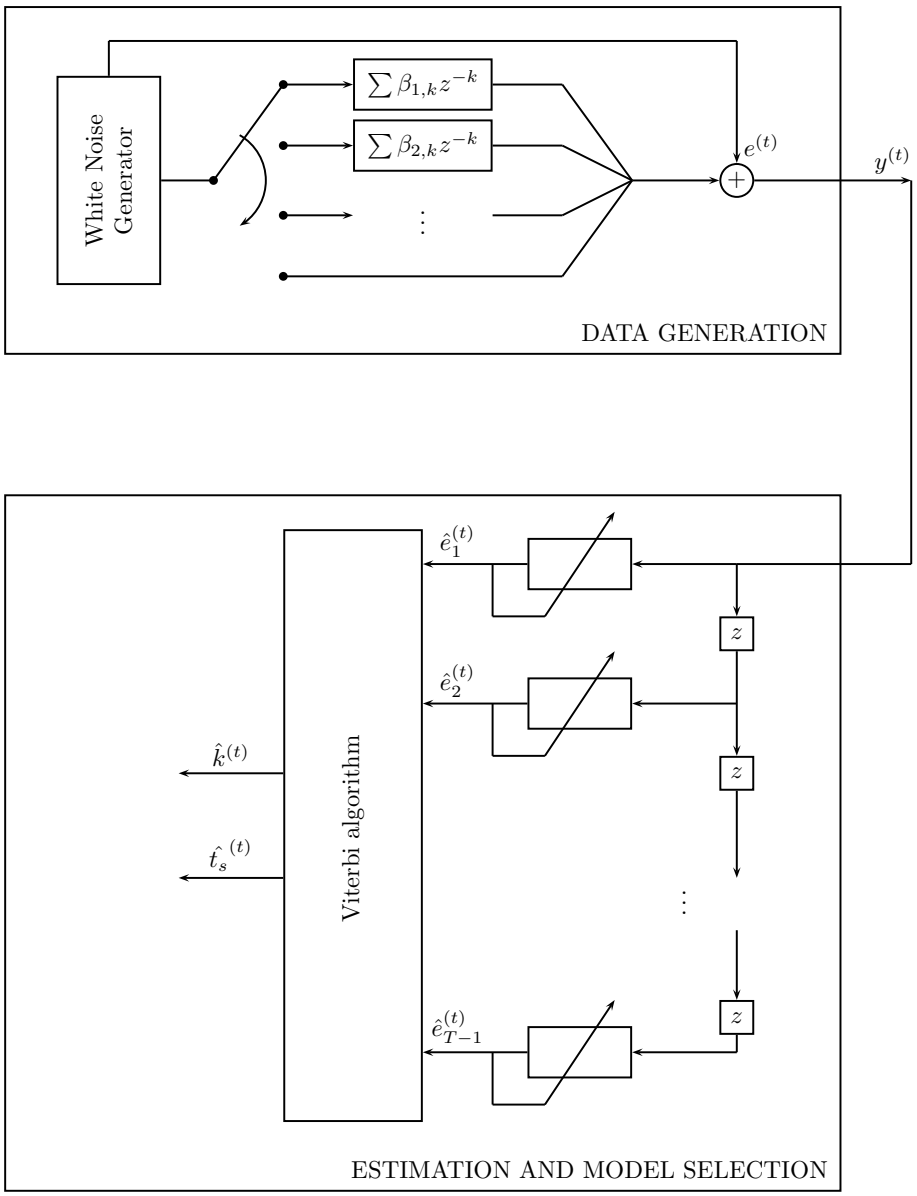


Figure 4.2: *Block diagram of the variable lookback algorithm: idealized data generating process and corresponding estimation and model selection procedure.*

takes the form of the quadratic tree. Consider Figure 4.3 which represents the first 4 iterations of the algorithm. It looks similar to Figure 4.1, but this time, we would like to focus on the operations of the dynamic programming algorithm. On the one hand, each node is labeled with the code length corresponding to encoding the current observation $y^{(t)}$ using the CNML distribution estimated using observations in the current period of piecewise stationarity up to time $t-1$. On the other hand, edges are labeled with the code length of encoding the binary representation of the pattern of switches $z^{(t)}$ using the KT probability. When moving along a path in the tree, we cumulate the cost of the nodes (code length of coding observations) and the cost of the edges (code length of coding switches). Finally, the dynamic programming algorithm is a shortest path algorithm that selects the path in the tree with the smallest accumulated cost.

4.2.2 Details of the algorithm

Pseudo Code

We now give in this section a detailed version of the variable lookback algorithm. As suggested by the block diagram, we need to (a) maintain a series of estimators of nested model classes of order $k = 0, \ldots, K$, each depending on a different portion of the information set and (b) perform a dynamic programming task to find the path of shortest cost in the quadratic tree. Our implementation of the Viterbi algorithm corresponds to a full expansion of the trellis diagram, similar to Djikstra's shortest path algorithm Dijkstra [1959]. Consider the pseudocode of the variable lookback algorithm given in Algorithm 4.1. We use the formalism of object-oriented programming. There are two classes of objects, $Node$ and $Model$. The object $Node$ describes a node in the quadratic tree diagram. Each node stores a pointer to its parent, t_{s-1}, used for backtracking. Each node also maintains and updates up to $K+1$ model classes of order $0, \ldots, K$ in the form of an array of $Model$ objects. Furthermore, a node contains also the code length corresponding to the encoding of symbols from previous periods of piecewise stationarity, CL_{past} and the code length associated with the KT probability of reaching that node in the quadratic tree, CL_{switch}. The object $Model$ stores the current estimate of the model parameters $\boldsymbol{\theta}_l$ and the CNML criterion for all model classes of order up to $k = 0, \ldots, K$. The object $Model$ also instantiates two functions, a constructor and a function to update the parameters and the CNML criterion given a new observation, updateModel(...). Given these two objects, the algorithm works as follows. The algorithm starts at node $(1, 0, 1)$. Given a new observation, the algorithm crosses every existing node and splits it in two nodes, one corresponding to a restarted estimation procedure, one corresponding to a continuation of the estimation procedure. The estimate of the parameters of the model, $\boldsymbol{\theta}_l$, the CNML criterion and the code length associated with the KT probability are updated accordingly. Finally, the total score associated with each node is the sum of the code length of encoding symbols from past and current periods of piecewise stationarity and CL_{switch}. Then, the MDL principle advocates to pick the model that minimizes this quantity.

Remark. Consider line 41 of Algorithm 4.1, where, following the MDL principle, we select the best node and order of the process. The minimization is also over the choice of order k for the current period of piecewise stationarity. The term $CNML$ in the pseudo code corresponds to the accumulated cost of coding symbols in the current

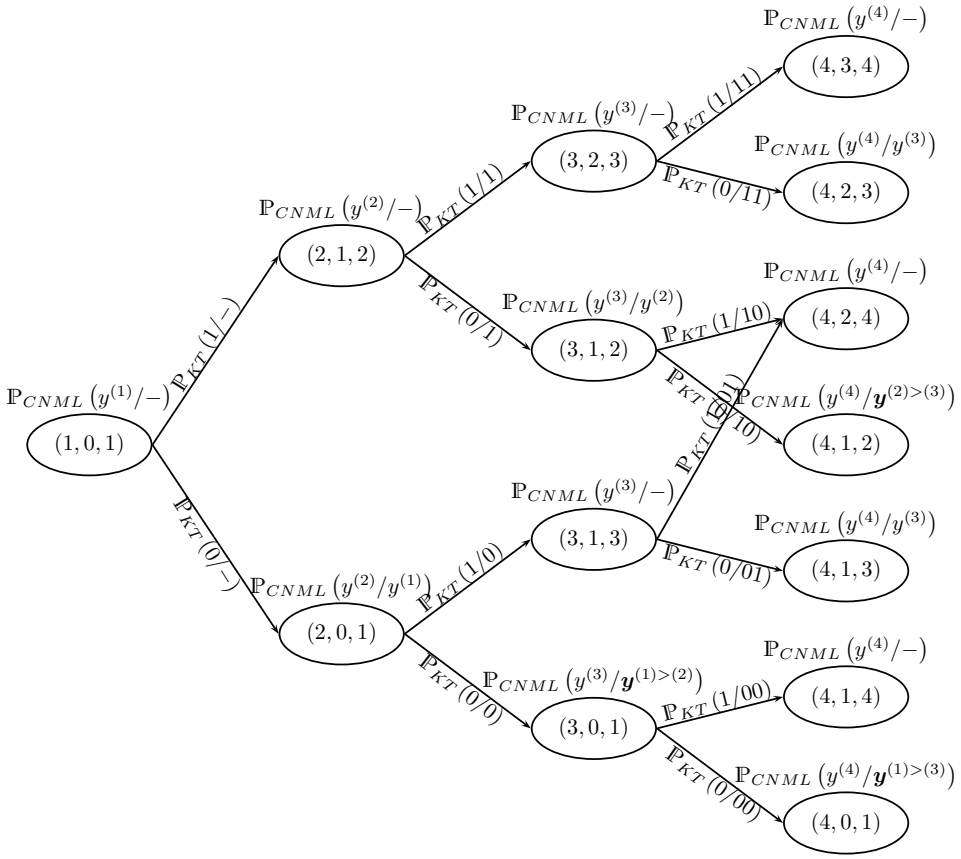


Figure 4.3: *Trellis diagram of the Viterbi algorithm for the first 4 iterations. Labels of the nodes correspond to the code length of coding the current observation using the CNML criterion estimated using data from the current period of piecewise stationarity up to time $t-1$. Labels on the edges correspond to the code length of coding the binary representation of the pattern of switches using the KT probability. The Viterbi algorithm finds the path of shortest accumulated cost.*

period of piecewise stationarity. Suppose that at some point the best order is k_1, i.e.,

$$k_1 = \underset{k}{\operatorname{argmin}}\, Node(t, s, t_s).Model(k).CNML \tag{4.62}$$

At a later point, another order k_2 minimizes this quantity. Compared to our description with the trellis diagram, this corresponds to a relabeling of the cost of the node for the current period of piecewise stationary, without affecting the cost reported from previous period of piecewise stationarity, CL_{past}. This illustrates that the trellis diagram does not fully reflects all subtleties of the variable lookback algorithm.

Coding until the CNML is available

There is an important detail we have ignored so far in the description of the variable lookback algorithm. When we restart the estimation procedure, i.e., introduce a switch in the process, the CNML criterion is not immediately available. Typically $m = K_{max} + 1$ is the minimum number of observations such that the maximum likelihood estimator can be computed for all model classes up to order K. Then, for node (t, s, t_s), the first CNML criterion can be computed for t such that $t - t_s + 1 \geq m + 2$. Furthermore, observe that, even if we knew the position of the switches t_s, there would exist a tradeoff between the accuracy in estimation and the accuracy in coding. Indeed, a model immediately available for coding is estimated using data from the previous period of piecewise stationarity, whereas a model available later is estimated using the correct data, but use the former model to code observations in the meantime. Given that we do not know the position of the switching times t_s, we can think of different solutions to handle this problem. We have chosen the one where the algorithm in the meantime uses the model of the preceding period of piecewise stationarity to encode observations until the CNML of the new period can be computed. Alternatively, we could have chosen the one where the model of the current period is initialized using data from the previous period, which makes the criterion immediately available. However, the two solutions are equivalent, as they simply correspond to a different parametrization of the index of the nodes in the quadratic tree. Indeed, with the second solution, a node at time t that is just restarted, (t, s, t) is estimated using data at time $t - 1, \ldots, t - m - 2$. This is equivalent, with the first solution, to node (t, s, t_s) where $t_s = t - m - 2$. These two nodes have the same code length of coding symbols. Also, since the cost of switching only depends on the number of observations t and the number of switches so far s, but not on the actual position of switches t_s, the two nodes have the same total code length. It is up to the algorithm, in a completely data-dependent fashion, to decide whether it chooses the accuracy in coding or that in estimation. Actually, there exists a third solution to this problem used by Kozat and Singer [2008]. The solution consists in replacing the maximum likelihood estimator by a ridge regressor estimator, also called penalized least-squares estimator Gruenwald [2007]. It is obtained by minimizing the standard least-squares function penalized by the Mahalanobis distance $\mu_0^T \Sigma_0^{-1} \mu_0$. This is equivalent to add to the original set of observations a set of artificial observations with mean μ_0 and variance Σ_0. This makes the criterion available immediately, at the cost of introducing an extra parameter, $\mu_0^T \Sigma_0^{-1} \mu_0$. We choose our approach because it is fully data-dependent.

Algorithm 4.1 Variable lookback algorithm.

Input: observations $\{y^{(t)}\}_{t=1,\ldots,T}$, regressors of order k $\{\boldsymbol{F}_k^{(t)}\}_{t=1,\ldots,T,k=0,\ldots,K}$
%Class definition
Classdef $\{\ \boldsymbol{\theta}, CNML\ \}$ as $Model$
Classdef $\{\ t_{s-1}, CL_{\text{past}}, CL_{\text{switch}}, Model()\ \}$ as $Node$

%Initialization
$Node(1,0,1).CL_{\text{past}} \leftarrow 0$
$Node(1,0,1).CL_{\text{switch}} \leftarrow 0$
$Node(1,0,1).Model(0) \leftarrow$ new $Model\left((y^{(1)}, \boldsymbol{F}_0^{(1)}\right)$

%For each observation
for $t = 2, \ldots, T$ **do**

  %For each existing node in the tree at $t-1$
  for $s = 0, \ldots, t-2$ **do**
    for $t_s = s+1, \ldots, t-1$ **do**

      %Restarted node
      $Node(t, s{+}1, t).t_{s-1} \leftarrow (Node(t, s{+}1, t).t_{s-1};\quad Node(t{-}1, s, t_s).t_s)^T$
      $Node(t, s+1, t).CL_{\text{switch}} \leftarrow Node(t-1, s, t_s).CL_{\text{switch}} - \log_2\left(\frac{s+1/2}{t+1}\right)$
      $Node(t, s+1, t).Model(0). \leftarrow$ new $Model\left((y^{(t)}, \boldsymbol{X}_0^{(t)}\right)$

      %Continued node
      $Node(t, s, t_s).CL_{\text{switch}} \leftarrow Node(t-1, s, t_s).CL_{\text{switch}} - \log_2\left(\frac{t-s-1/2}{t+1}\right)$
      for $k = 0, \ldots, t-t_s-1$ **do**
        $Node(t, s, t_s).Model(k) \leftarrow Node(t-1, s, t_s).Model(k)$.updateModel $(y^{(t)}, \boldsymbol{F}_k^{(t)})$
      end for
      if $t - t_s + 1 \leq m$ **then**
        $Node(t, s, t_s).Model(t-t_s) \leftarrow \text{new}Model(y^{(t_s \wedge t)}, \boldsymbol{F}_{t-t_s}^{(t_s > t)})$
      end if

    end for
  end for

  %Update code length of symbols of past periods of piecewise stationarity
  for all $Node(t, s, t_s) : t - t_s = m$ **do**
    $[Node(t, s, t_s).CL_{\text{past}},\ Node(t, s, t_s).t_{s-1}] \leftarrow \min_{t' \in Node(t,s,t_s).t_{s-1}, k} Node(t, s, t').CL_{\text{past}} + Node(t, s, t').Model(k).CNML$
  end for

  %Model selection by MDL principle
  return $\text{argmin}_{t,s,t_s,k}\, Node(t, s, t_s).CL_{\text{past}} + Node(t, s, t_s).Model(k).CNML + Node(t, s, t_s).CL_{\text{switch}}$
end for

The report of the cost of coding symbols from the previous period of piecewise stationarity is described in line 37 of Algorithm 4.1. It takes place at time $t = t_s + m + 1$, i.e., at the iteration preceding the availability of the CNML criterion. Observe that a selection is performed among competing path in the tree leading to that node. Whatever switching or symbols appear in the future, all these paths have a similar future evolution. Following the MDL principle, it is valid to select the best of them, i.e., the smallest, and discard the other strictly dominated solutions.

Remark. Both CNML and KT are prequential schemes, and their combination is prequential as well. However, it seems that, because we perform a selection among past paths, the consistency condition (4.51) is not respected. This issue is resolved by noting that a code word is implicitly reserved for suboptimal solutions, but not explicitly stored in the quadratic tree.

Cost of switching

We have seen that the variable lookback algorithm attributes to each node in the quadratic tree an information theoretic score, or cost function, given by

$$\begin{aligned} &\min_k \{Node(t,s,t_s).Model(k).CNML\} + \\ &Node(t,s,t_s).CL_{\text{past}} + Node(t,s,t_s).CL_{\text{switch}}. \end{aligned} \tag{4.63}$$

It is the sum of code length resulting from the encoding of nodes in the past and current periods of piecewise stationarity and the code length of encoding the switching pattern using the KT distribution. Furthermore, the MDL principle advocates to find the node that minimizes this quantity. Let us first understand how the algorithm changes from selecting one node to another. Let $(t, s-1, t_{s-1})$ be the index of a node in the quadratic tree diagram and (t, s, t_s) be the index of the restarted node on the same path, where the restart happens at time t_s. Figure 4.4 represents the evolution of the code length of these two nodes, as of time t_s. The total code length is decomposed, following (4.63), into the code length of the past and current periods of piecewise stationarity, and the code length of switching. Up to time $t = t_s$, the two nodes are identical and have the same cost functions. At time $t = t_s$, the restarted node (t_s, s, t_s) has a higher cost of switching, because it contains one more switch compared to $(t_s, s-1, t_{s-1})$. Indeed, the encoder needs to transmit the information to the decoder that there is a switch in the system. Also, since the CNML criterion of the restarted node is not available, the cost of coding symbols is the same for the two nodes until $t_s + m + 2$ and the algorithm never selects the restarted node until then. At time $t = t_s + m$, we can compute for the restarted node the first estimate of the parameters of all model classes up to order K and at time $t = t_s + m + 1$, the code length of the previous periods, CL_{past} is updated

$$\min_{t' \in t_{s-1}, k} Node(t, s-1, t').CL_{past} + Node(t, s-1, t').Model(k).CNML \tag{4.64}$$

At time $t = t_s + m + 2$, the CNML criterion is now available and the restarted node is fully operational. As of this point, a competition between the two nodes begins. If the model of the restarted node represents the data well, its cost of coding symbols increases at a lower rate than that of the continued node. Ultimately, at some point t', the improvement in coding symbols is such that it compensates the additional cost

of switching and the total code length of the restarted node becomes smaller. The algorithm selects it then. This clearly corresponds to our intuition we phrased in Section 2.2.4 where the improvement in the fit of the model should be penalized by a certain complexity measure. But, unlike in other ad hoc method, the value of the threshold is given by a theoretically sound quantity and not by an arbitrary parameter value.

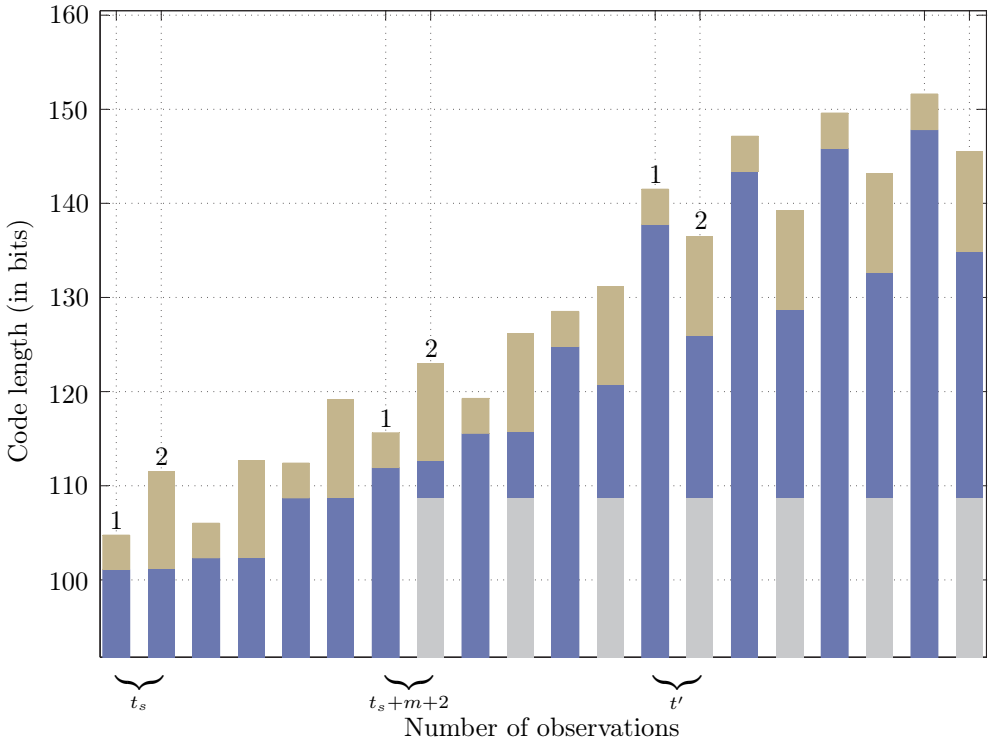


Figure 4.4: *Evolution of the cost function of continued (1) and restarted (2) model. The underlying process switches at $t = t_s$ between an i.i.d. process to a highly correlated autoregressive process of order 1. Total code length is decomposed into code length of symbols from past periods of piecewise stationarity (grey), code length of symbols from current period of stationarity (blue) and code length of switching (gold). At time $t = t_s + m + 2$, the restarted model is operational and at time $t = t'$, the restarted model is selected over the continued one, as it has a smallest total code length.*

We will see, when we move to applications of the variable lookback algorithm in Chapter 5, that the first term in the CNML criterion resembles the entropy of the process. Also, since we are interested in the time-series of returns, the random variable $y^{(t)}$ are continuous and the associated differential entropy might be negative. Also, if we scale the initial sequence by a factor a, the differential entropy is consequently scaled by factor $\log_2(a)$ Cover and Thomas [1991]. Then, it seems we are not defining a consistent model selection criterion. In particular, the CL_{switch} does not appear to be a consistent penalty for switches in the structure of the system. This is however

not the case. Indeed, in practice continuous variable are quantized at precision level q. And the differential entropy of the random variable X, $h(X)$, is related to the entropy of the quantized version of the variable, $H(X)$ by the relation Cover and Thomas [1991]

$$H(X) = h(X) - \log(q). \tag{4.65}$$

Suppose we fix a desired precision level. If we scale the original sequence by a, we can adapt the precision level so as to represent the same quantity of information. Also, since accounting for the quantization implies adding a term $-t \log_2(q)$ to each node in the tree, we do not include it explicitly in the score of our algorithm as it does not impact the final selection. Thus, the cost of switching given by the code length corresponding to the KT probability is a valid penalty for the switches in the system.

Computational aspects

The pseudo-code of the variable lookback algorithm given by Algorithm 4.1 maintains a quadratic number of nodes in the quadratic tree diagram. In particular, the number of nodes at time t is equal to

$$\frac{t(t-1)}{2} + 1 = O(t^2). \tag{4.66}$$

There exist alternatives to the full expansion of the trellis diagram, which amount to use a different shortest path algorithm than Dijkstra's algorithm. For example, we have developed an implementation of the variable lookback algorithm using A* algorithm Hart et al. [1968]. In A*, a heuristic function is used to estimate of a lower bound on the remaining cost up to time T. Then, the algorithm expands first the nodes that have the smallest actual incurred cost augmented by this heuristic function. The choice of the heuristic function is crucial for the performance of A*. In our case, there exists no good candidate. We have used as heuristic function the cost of encoding with the KT distribution the $T - t$ sequence of zeros, given the current path in the tree

$$-\log_2 \mathbb{P}_{KT}\left(0 \ldots 0 / z^{(t)}\right) \tag{4.67}$$

which constitutes a lower bound on the cost of coding swicthes until time T. Our experience shows that, in conjunction with this heuristic function, A* is not computationally more efficient than Dijkstra's algorithm, because we are expanding almost all nodes and the overhead of keeping track of which nodes are expanded makes the algorithm performs poorly. Also, from a memory usage perspective, the full expansion requires to store only an array of nodes at time t, whereas we have to store a tree of nodes with A*.

Moreover, in the pseudo code of Algorithm 4.1, each node is storing a model and it seems that the number of models grows also in $O(t^2)$. This contrasts with the block diagram where the number of models grows linearly. As can be seen in Figure 4.1, several nodes use the same model for their current period of piecewise stationary. For example, node $(4, 1, 4)$ and $(4, 2, 4)$ both use model restarted at time 4. Therefore, unlike in the pseudo code of Algorithm 4.1, we can maintain a linear number of models and store in a node only a pointer to the relevant classes of models. The latter is actually simply given by t_s, the last time before which a switch happens. This remark concludes our description of the variable lookback algorithm.

4.3 Test on simulated data

We aim to ultimately apply the variable lookback algorithm to problems in finance, and this is the topic of Chapter 5. It is however of great importance to first test the behavior of the algorithm on simulated time-series. First, when using simulation, there exists an underlying ground truth against which the output of the algorithm can be compared. Think for example of the order of the process generating the data or the actual position of a switch. This comparison ensures that the algorithm is not systematically flawed. Secondly, like in experimental finance, a simulated time-series corresponds to an idealized, simplified and controlled scenario. There is no problem of model misspecification and it is possible to disentangle competing effects, unlike in field data. Also, by varying the values of the parameters, typically the variance of the observation noise, the experimenter can study how the performance of the algorithm degrades in the presence of noise.

4.3.1 Test of the CNML criterion

Description of the experiment

This first test aims to assess the ability of the CNML criterion to identify the correct order of the process $k \in \{1, \ldots, 10\}$. The experiment consists of a series of independent iterations. At each iteration, a time-series of $T = 100$ observations is obtained by sampling an AR process of order k

$$y^{(t)} = \beta_1 y^{(t-1)} + \beta_2 y^{(t-2)} + \ldots + \beta_k y^{(t-k)} + e^{(t)}, \tag{4.68}$$

where the residuals $e^{(t)}$ are i.i.d., standard normal

$$e^{(t)} \sim \mathcal{N}(0, 1). \tag{4.69}$$

The coefficients of the filter are obtained by sampling uniformly at random the k roots of the polynomials in z^{-1}

$$1 - \beta_1 z^{-1} - \beta_2 z^{-2} - \ldots - \beta_k z^{-k} \tag{4.70}$$

such that (a) all roots lie strictly inside the unit circle, so as to obtain a strictly stable system, and (b) complex roots come in complex conjugate pairs, so as to obtain real coefficients $\beta_1, \ldots, \beta_k$. Then, considering the nested model classes of AR process of order up to $K = 10$, the CNML criterion is computed sequentially for the simulated time-series. Following the MDL principle, an estimate of the order is chosen, so as to minimize at every time t the CNML criterion,

$$\hat{k}_i^{(t)} = \underset{k}{\operatorname{argmin}}\, CNML(k, t), \tag{4.71}$$

where $\hat{k}_i^{(t)}$ denotes the estimates of the order of the AR process at time t for the i^{th} iteration of the experiment. There are $I = 100$ iterations and we compute an estimate of the probability that the CNML criterion correctly infers the order, given by

$$\mathbb{P}\left(\hat{k}^{(t)} = k\right) = \frac{\sharp \text{ of times } k_i^{(t)} = k}{I}. \tag{4.72}$$

Similarly, we can compute an estimate of the probability that this criterion over- and underestimates the correct order k.

Results

Figure 4.5 represents the estimate of the probability that the CNML criterion underestimates (light blue), correctly estimates (medium blue) and overestimates (dark blue) the correct order of the process k as a function of the number of observations $t = 20, 30 \dots 100$. Different plots correspond to different orders $k = 0, \dots, 5$. The results are in line with those of Rissanen et al. [2010]. We observe that the probability of correctly estimating the model increases with the number of observations; this indicates some form of consistency of the criterion which, given a sufficient amount of data, finds the correct order of the process. But, comparing across plots processes with different orders k, this increase occurs at a slower rate for processes of higher order k. We also observe that the light blue area tends to be larger than the dark blue one for almost all number of observations and orders. Thus, the CNML criterion has a tendency to underestimate the true order of the process. Stated differently, the CNML criterion is rather conservative in its selection and selects a higher order model only when it is fairly confident. This is good for financial applications, where overfitting is a serious concern.

Remark. We have observed that the probability to identify the correct order of the process increases with the number of observations of the process. Intuitively, this probability also depends on the intensity of the effect, i.e., how correlated the process is, and small effects take more time to detect. To illustrate this, consider the two classes of AR models of order $k = 1$, one where the AR coefficient is constrained to be small,

$$|\beta_1| \in (0; 0.1], \tag{4.73}$$

and one where the AR coefficient is constrained to be large,

$$|\beta_1| \in [0.9; 1). \tag{4.74}$$

Figure 4.6 represents the result of the same experiment when the time-series is generated by sampling from these two model classes. This confirms our intuition: the correct order is identified more quickly for highly correlated processes. By sampling the roots uniformly at random in the unit circle, Figure 4.5(b) averages out qualitatively different results.

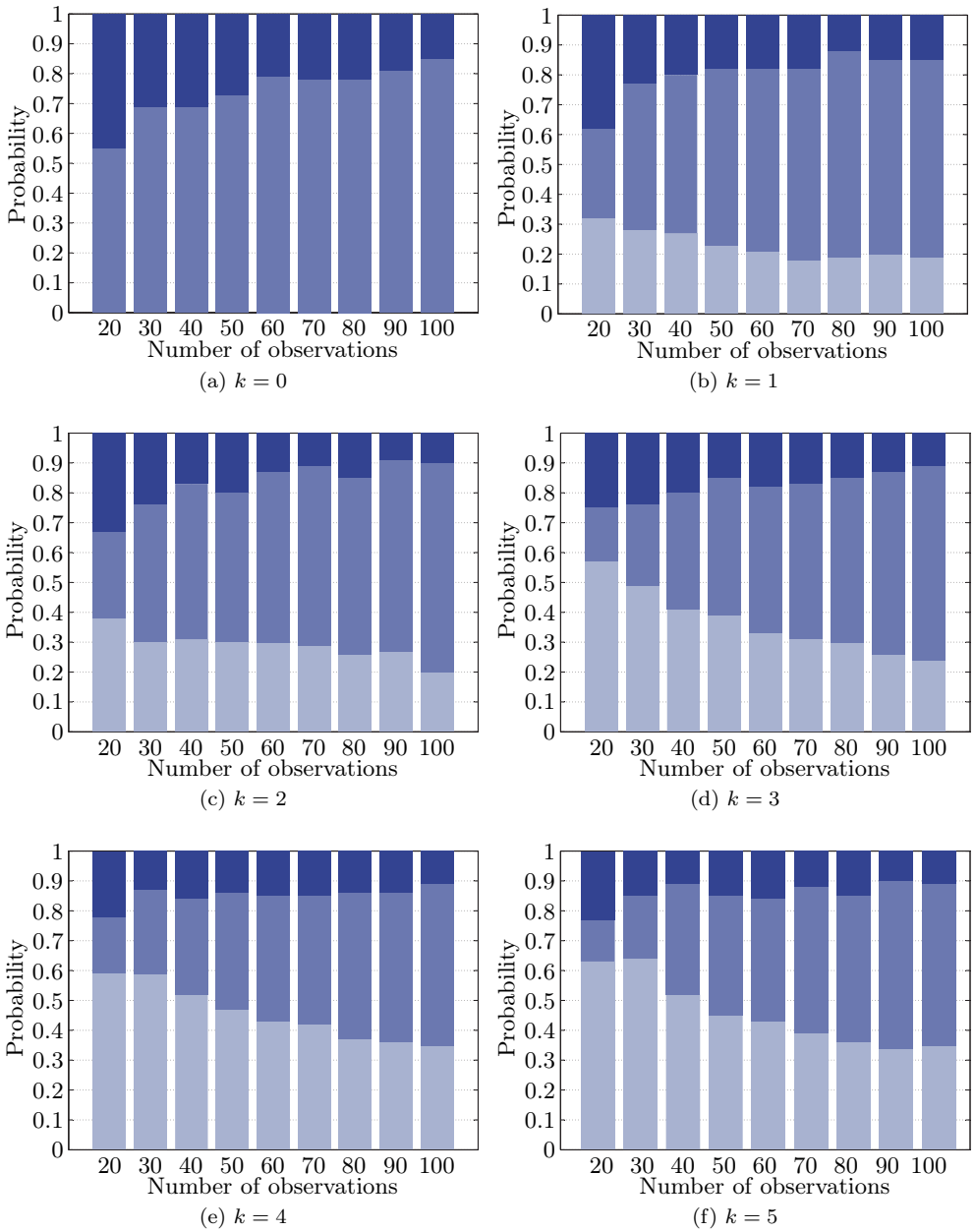
4.3.2 Test of switching point detection

Description of the experiment

The second test aims to assess the ability of the variable lookback algorithm to detect a switch in the underlying dynamics of the system. The experiment consists of a series of independent iterations. At each iteration, a time-series of $T = 100$ observations is generated by sampling to an AR process, that switches after 50 observations between an i.i.d. process to a highly correlated AR process of order 1, or vice versa. See example in Figure 4.3.2. The regression coefficient is constrained to be in the interval

$$|\beta_1| \in [0.9; 1), \tag{4.75}$$

so as to obtain a strongly correlated process. Let us justify this choice. The quality of the estimation has a direct influence on the selection of the variable lookback algorithm

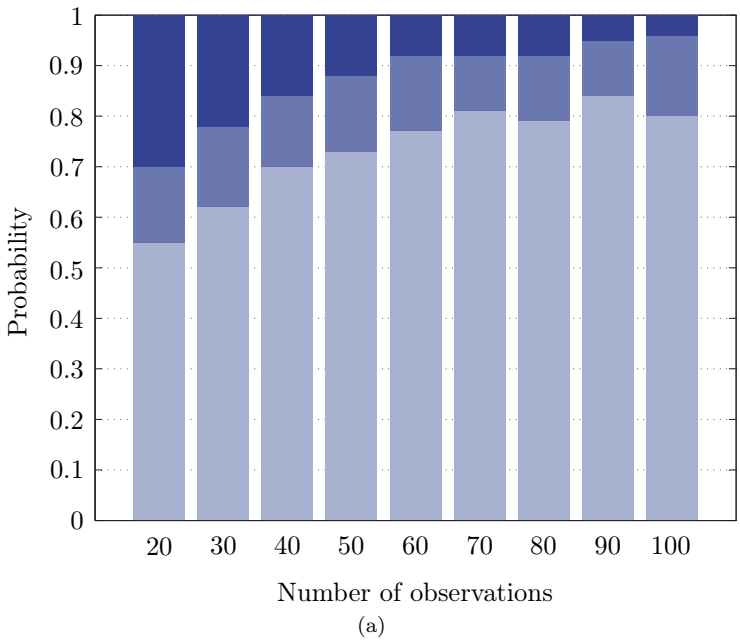


(a) $k = 0$ (b) $k = 1$

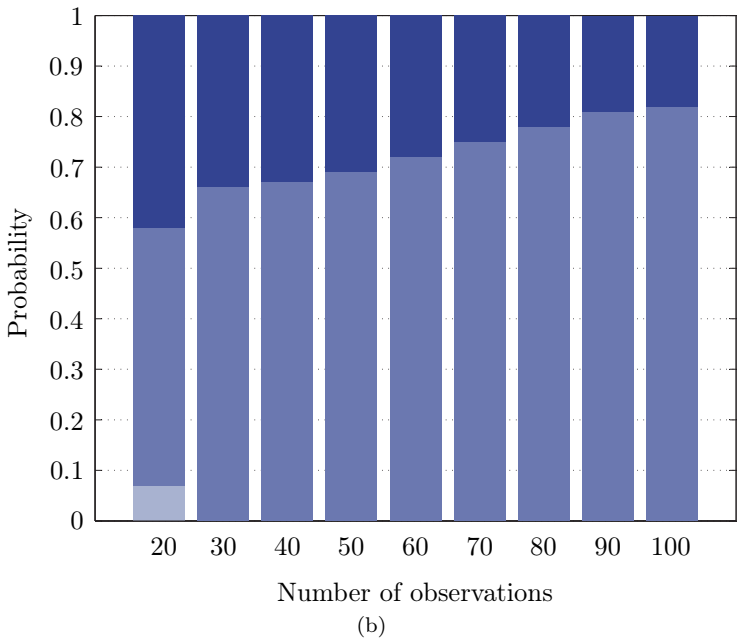
(c) $k = 2$ (d) $k = 3$

(e) $k = 4$ (f) $k = 5$

Figure 4.5: *Estimate of the probability that the CNML criterion underestimates (light blue), correctly estimates (medium blue) and overestimates (dark blue) the correct order of the process, k, after t observations. The underlying process is a stable, autoregressive process with real coefficients of order $k = 0$ in (a), $k = 1$ in (b), $k = 2$ in (c), $k = 3$ in (d), $k = 4$ in (e) and $k = 5$ in (f). The probability to detect the correct order increases with the number of observations, but at a slower rate for higher order processes.*



(a)



(b)

Figure 4.6: *Estimate of the probability that the CNML criterion underestimates (light blue), correctly estimates (medium blue) and overestimates (dark blue) the correct order of the process, k, after t observations. The process is a stable, autoregressive process with real coefficients of order $k = 1$ with small (a) and large (b) correlation. Stronger effect are detected more quickly.*

and it might be hard to disentangle whether a switch is not detected because there is a flaw in the variable lookback algorithm or in the underlying estimation procedure. With such a strong correlation effect, it is extremely likely going to be picked by the estimation algorithm and we can focus on the ability of the variable lookback algorithm to detect a switch. Note also that it is not sufficient to simply append two independent time-series, the second one has to be initialized using data from the first one. Then, considering the nested model classes of autoregressive process of order $k = 0, 1$, we use the variable lookback algorithm to compute an estimate of the switching time $t_{s,i}^{(t)}$ at each iteration i. There are $I = 100$ iterations and we can compute in the cross-section of iterations an empirical histogram of the estimated $t_s^{(t)}$

$$\mathbb{P}\left(t_s^{(t)} = t'\right) = \frac{\sharp \text{ of times } t_{s,i}^{(t)} = t'}{I}, \tag{4.76}$$

which is a function of the number of observations.

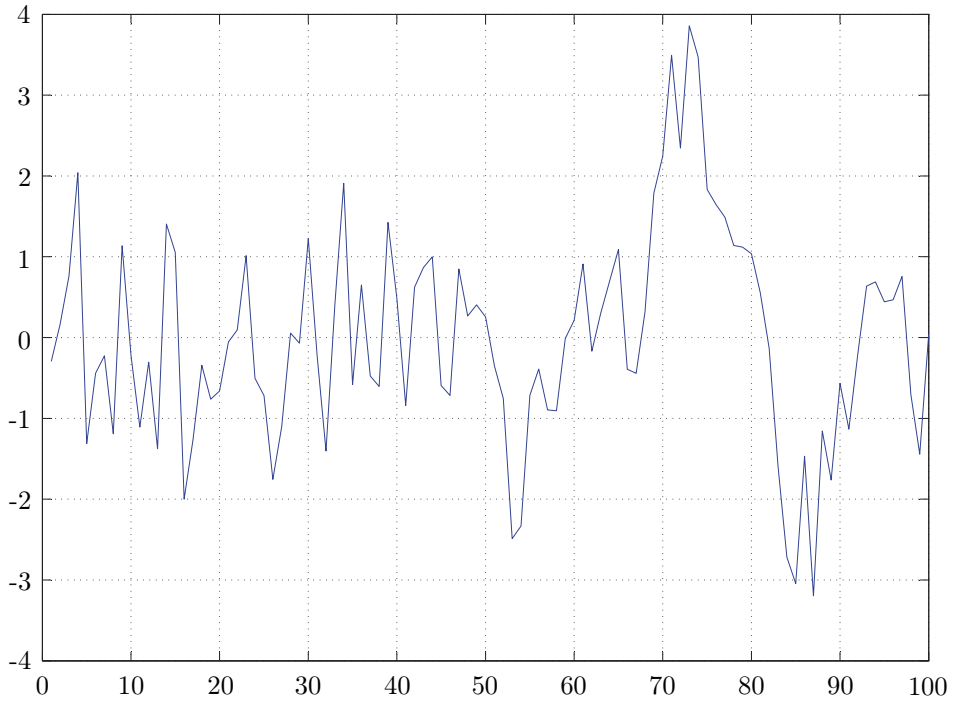


Figure 4.7: *Sample path of a process that switches at time $t = 51$ between an i.i.d. process to a highly correlated AR process of order 1, $\beta_1 = 0.92$.*

Results

Consider first the case where the system switches from an i.i.d. to a highly correlated AR(1) process. Figure 4.8 represents the histogram of the estimated switching time t_s as a function of the number of observation. The histogram is seen by taking a slice of the figure for a fixed number of observations. Ideally, if the algorithm

perfectly and immediately identifies the correct t_s, the histogram has a single peak, at $t_s = 0$ for $t < 51$ and $t_s = 51$ otherwise. Let us compare this ideal case with what we observe in practice. Up to time $t = 50$, we see that the histogram is almost only concentrated on $t_s = 0$. There is no switch in the system and the system does not detect one. In technical terms, the number of false positive is negligible. After the switch has happened at $t = 51$, the system starts detecting a switch. Then, the histogram of t_s is more dispersed compared to the situation before the switch, but it is centered approximately on the value of 50. Moreover, as explained in Section 4.2.2, the detection of a switch is not immediate, and, e.g., there is with 70 observations still approximately 30% of the iterations that do not detect a switch. The number of false positive is large, but decreases as we use more observations of the process, again a form of consistency of the algorithm. The same comments applies for Figure 4.9.

There is a form of duality between the two scenarii, and it is legitimate to wonder whether this duality is reflected in the output of the algorithm. This is indeed the case. Remember that, when the system chooses the position of a switching time, it also chooses in a completely data-dependent manner between the accuracy in the estimation and the accuracy in the coding. On the one hand, when the system switches from an i.i.d. to a strongly correlated AR(1) process, the algorithm tends to select 51 as the position of the switching time. In that case, the AR(1) process is estimated using correct data, and an i.i.d. process is used to encode the process in the meantime (accuracy in the estimation). On the other hand, when the system switches between an AR(1) to an i.i.d. process, the algorithm tends to select 47 as the position of the switching time. In that case, the i.i.d. process is estimated using data from the previous period, but it is immediately ready for coding observations at time 51 (accuracy in coding). In both case, the length of the period where an i.i.d. process is used is inflated. This indicates again a certain form of conservatism in the decision of the algorithm, which selects a higher order process only when it is fairly confident.

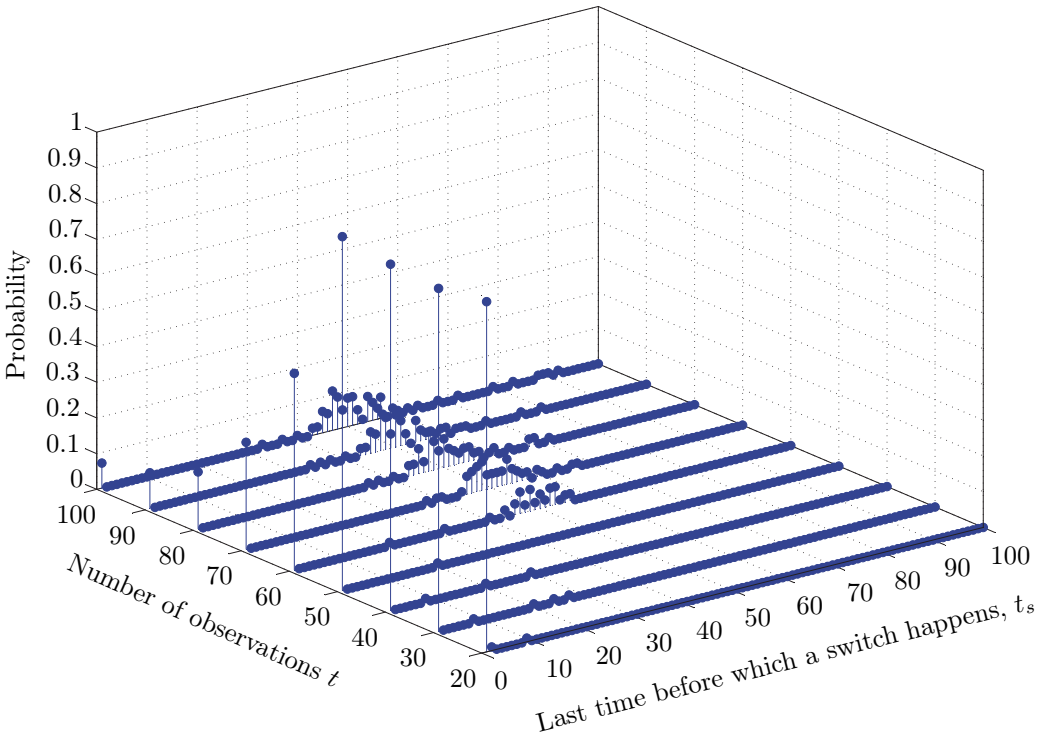


Figure 4.8: *Histogram of the identified switching time t_s as a function of the number of observations t. The underlying process switches at $t = 51$ from an i.i.d. process to a highly correlated AR(1) process.*

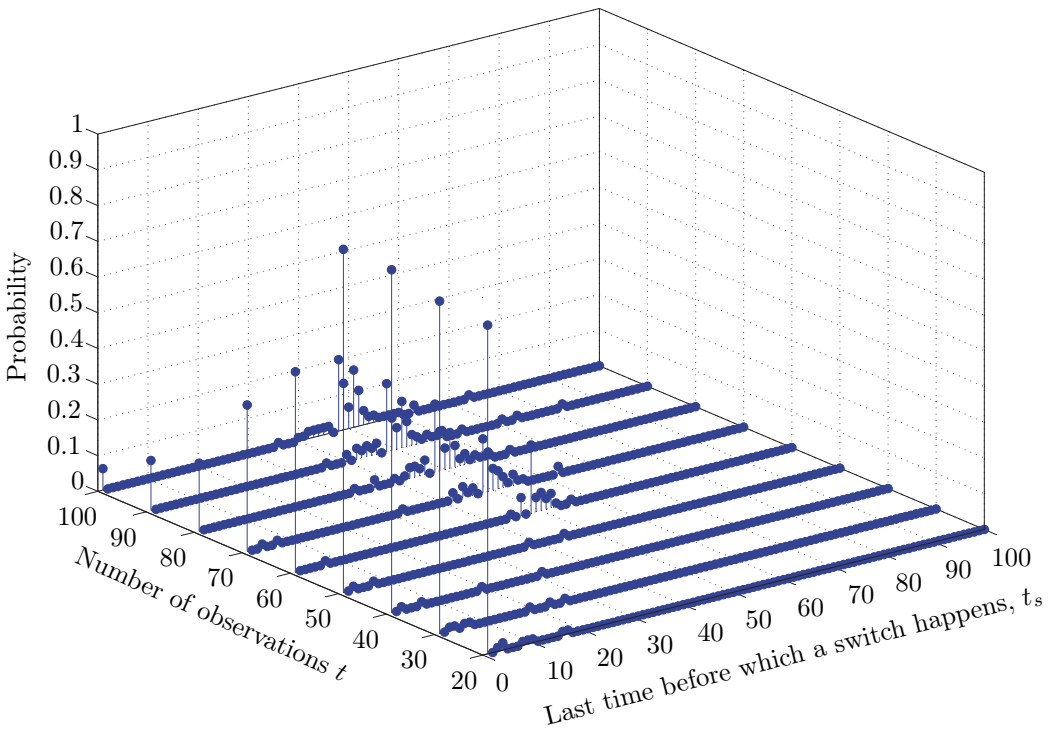


Figure 4.9: *Histogram of the identified switching time t_s as a function of the number of observations t. The underlying process switches at $t = 51$ from a highly correlated AR(1) process to an i.i.d. process.*

Summary

In this chapter, we have introduced our solution to estimate the variable lookback model based on the MDL principle. Since this principle underlies our solution, we have started by reviewing it. MDL equates the problem of learning with data compression: a statistical model aims to capture the regularities used by a data compression engine to compress a sequence of observations. The measure of complexity is given by the code length of a universal model for a class of processes, i.e., the unique probability distribution that is able to represent well any model in that class. Finally MDL advocates to choose the model class minimizing this measure. We have then described the variable lookback algorithm, which is based on the joint coding of the sequence of observations and the underlying binary representation of the pattern of switches. The algorithm combines a series of estimators with a dynamic programming algorithm to select the order and the lookback window. Finally, we have tested the algorithm on simulated time-series, to check its ability to select the correct order of the process and to detect the presence of a switch in the underlying dynamics of the system. Applying the algorithm to financial problems is the topic of the coming chapter.

Chapter 5

Applications to Finance

> Noémie, please meet Lionel. He works in finance. One day he will be rich [or not].
>
> Martin Vetterli

In the preceding chapters, we have introduced the idea of a variable lookback model, a model where a variable portion of the information set is relevant for future predictions. We have also proposed an algorithm based on the MDL principle to estimate it, the variable lookback algorithm. We have described this algorithm in the most general terms, in particular, independently of the choice of models and model classes. Furthermore, our solution is solely based on the assumption of piecewise stationarity. Moreover, it works as follows. It maintains various model classes, each based on a different portion of the information set, whose number grows linearly in the number of observations. This set of estimators is combined with a quadratic dynamic programming algorithm resembling the Viterbi algorithm, so as to select simultaneously the order of the process and the relevant lookback window by minimizing an information theoretic score. The latter measures the ability of the model to fit data penalized by the model complexity and the complexity of the underlying switching process between different regimes. From a learning perspective, the algorithm decides sequentially either to keep and update the current model based on the latest observation (continued estimation) or to restart the learning procedure and forget all properties learned so far (restarted estimation). The restarted model is chosen over the continued one, when it is better by a certain threshold, whose value is not a parameter of the algorithm, which can later be fine-tuned to fit a certain dataset. Rather, it corresponds to the complexity of encoding an additional switch in the system associated with a chosen description method.

In this chapter, we would like to demonstrate the added value of the proposed algorithm by applying it to concrete problems in finance. In the most general terms, any applications of the algorithm should start with the specification of model classes to describe the data at hand. Remember that, as seen in our review of MDL in Section 4.1.2, the choice of models and model classes is left to humans. Indeed there are no automated procedures to find the shortest description when the latter is expressed

using a computer program, that prints data and stops. Moreover, it is of great interest to reinterpret the output of the algorithm, the selected order and lookback window, in the context of each application. Finally, observe that the difficulty in the applications of the algorithm comes from the fact that it does not depend on hyperparameters, which can be fine-tuned by the experimenter. At the same time, this is the main interest of this chapter. This is not a mere data mining exercise. The proposed strategies are fully out-of-sample and as such could have been identified and applied by an investor ex ante.

The remainder of this chapter is organized as follows. We start by describing an application of the algorithm to dynamically draw the trend line for the time-series of log-prices (Section 5.1.1 and 5.1.2). This aims to be more an illustration of the algorithm and of its underlying concepts rather than the basis for investment strategies. Therefore, our analysis does not contain a formal evaluation of the resulting predictions. We then consider a more realistic example and apply the variable lookback model to make the well known momentum strategy adaptive. We describe the methodology (Section 5.2.1) and relate it with other existing approaches (Section 5.2.2), since momentum is a well documented investment strategy. We also evaluate the output of the algorithm using a backtest. We build a strategy that leverages investments in the risky asset based on the sign of the prediction of next period's return (Section 5.2.3). We also highlight the behavior of the strategy in 2009, which like the 1930's, is known to be one of the historical periods where momentum strategies have dramatically failed.

5.1 Dynamic trend line

5.1.1 Description of the methodology

The first application of the variable lookback algorithm aims at drawing dynamically the trend line for the time-series of log-prices. The trend line is a tool commonly used by investment professionals, which makes the output of the algorithm comprehensible by a large audience of them. We aim to mimic the behavior of a portfolio manager observing a time-series on Bloomberg in a completely automated and data-dependent manner. The class of models that we consider to describe the evolution of the log-prices over a period of piecewise stationarity is simply given by

$$\log p^{(t)} = \alpha + \beta t + e^{(t)}, \tag{5.1}$$

where the residuals $e^{(t)}$ are assumed to be i.i.d. Gaussian with zero mean and constant variance σ^2. Hence the vector of parameters is given by

$$\boldsymbol{\theta} = \begin{pmatrix} \alpha \\ \beta \\ \sigma^2 \end{pmatrix}. \tag{5.2}$$

We should not talk about piecewise stationarity, as the trend line model is non-stationary over $[t_{s-1}; t_s)$, but we do so in a loose way. Over a period of piecewise stationarity, the value of the parameters that corresponds to the best fitting line in the least-squares sense is computed. Note that it also corresponds to the maximum

likelihood estimator of the parameters $\boldsymbol{\theta}$, because of the Gaussianity assumption on the residual process. Furthermore, let us reinterpret the output of the algorithm in the context of this experiment. In that particular case, there is no choice of order, since there is only one model class. Also, the selection of the lookback window corresponds to a segmentation of the original time-series in periods of piecewise stationarity, where a different trend line is fitted on each period. The role of the algorithm is then to decide, given a new observation, whether to carry on using the current trend line and update its slope and intercept, or to completely forget the past and restart a new trend line. This happens if the improvement in the fit of the new line compensates the added complexity of introducing a switch in the system.

Let us also briefly compare our approach with other automated trend line procedures. One alternative consists in finding the best fitting line over a certain window of the latest observations. Another one, current in technical analysis and also referred to as support trend line, consists in joining the lowest low with the next highest high over a certain lookback window, provided no significance crossing has already happened Schwager [1995]. In both cases, the window size is a parameter of the procedure and its choice is problematic, since the optimal window size appears to change over time. On the contrary, our approach does not suffer from this drawback, as it is fully data-dependent and no parameters can be fine-tuned to control the output.

The algorithm is applied on the 10 Morgan Stanley Capital International (MSCI) World global industry classification standard (GICS) sector total return indices, sampled at monthly frequency from January 1995 until July 2011. Each time-series represents the weighted average returns of a large basket of stocks belonging to a certain sector. The grouping of stocks in a sector follows the GICS classification and the weight of each stock depends on its market capitalization.

5.1.2 Results

Figure 5.1 and 5.2 represent the output of the algorithm for the MSCI World Financials index (MXWOFN) after July 2007 and March 2009, respectively. These two dates are obviously not chosen at random. The first one corresponds to the last observation before the start of the subprime crisis in the US, the second one to the lowest observation in the market, preceding an important rebound particularly acute for all financials stocks. Each plot contains the evolution of the log-price of the index (solid blue line), as well as various trend lines (gold line) conditional on information up to time t. Observe the difference between two types of trend lines. The trend lines of the previous periods of piecewise stationarity are represented using a dashed line, whereas the trend line of the current period of piecewise stationarity is represented using a solid line. This emphasizes the fact that only the trend line in the current period of piecewise stationary is active: its slope and intercept are updated given a new observation. Firstly, we observe that the algorithm is extremely responsive. For example, in March 2009 the index reaches its lowest point and only after two observations in April and May 2009 does the algorithm introduce a switch and a new positive trend line. It is then well positioned to benefit from the rebound of the market as of June 2009. This should be compared against the output of a windowing algorihm, which uses a fixed rolling lookback window. Of course, this other algorithm will progressively adapt, and update the slope of the trend line until it reaches a positive

value, as the lookback window contains more and more points corresponding to the rebound. But this is never as rapid as our proposed solution. Also, observe that the algorithm can introduce a switch and later on revert its decision. For example, this happens in Figure 5.1 between February 2008 and May 2008. Given the information up to February 2008, the algorithm has chosen to introduce a new trend line, whose improvement in fit compensates the additional cost of encoding an additional switch. At a later point in May, this does not hold anymore and the algorithm reverts its decision. With this example, our aim was mainly to illustrate the variable lookback algorithm and its underlying concepts. Therefore, we do not perform any further evaluation of the output of the variable lookback algorithm beyond this visual inspection, in particular the backtest of the strategy based on its predictions.

5.2 Adaptive momentum strategy

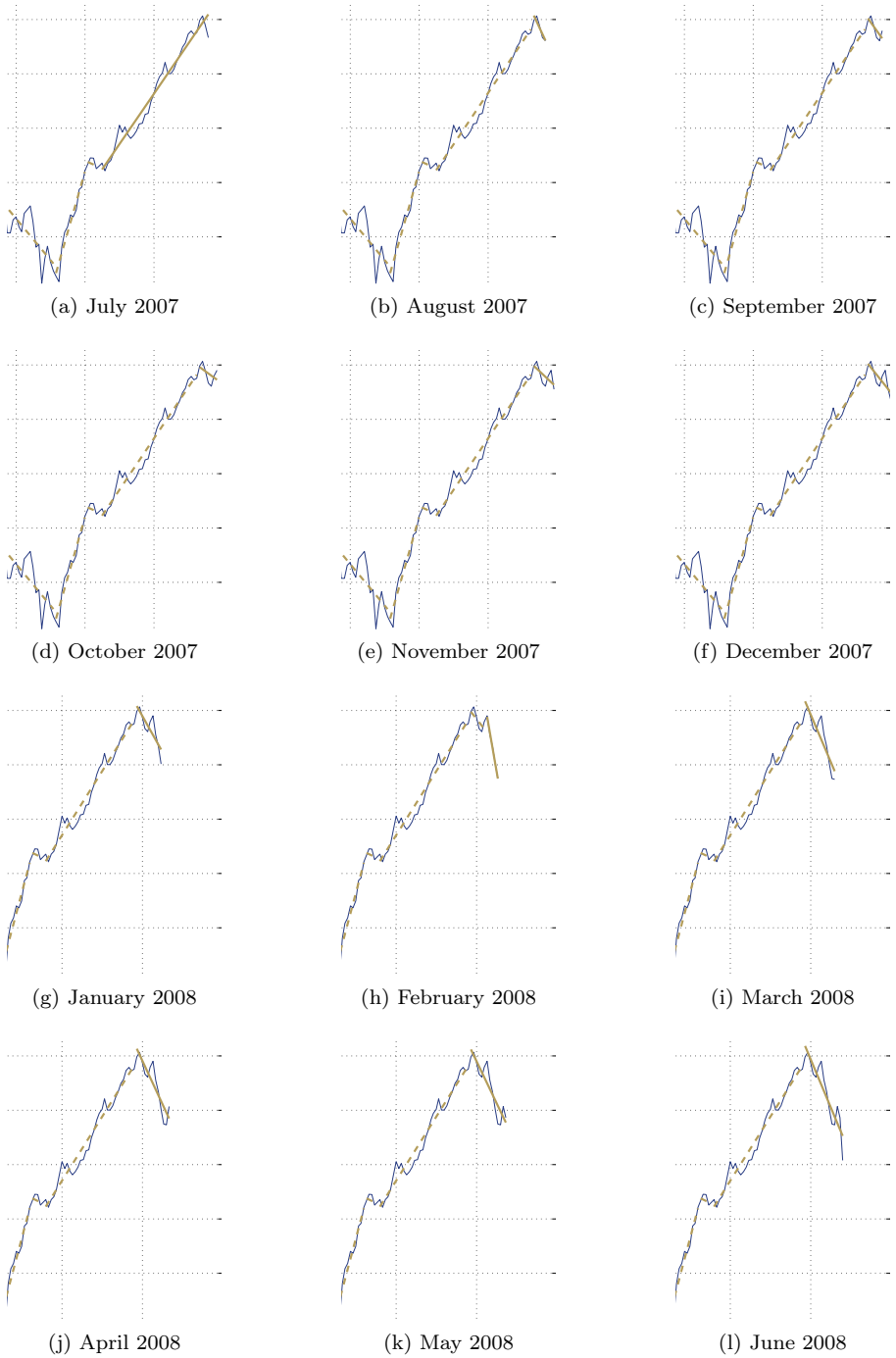
The second application of the variable lookback algorithm aims to develop an adaptive momentum strategy. Momentum generally refers to an investment strategy that bets on continuation in the return process. Informally, past winners are the future winners, and past losers the future losers. We also use the term momentum-type strategies to encompass more general forms of this strategy, in particular the contrarian strategy: past winners are the future losers, and past losers the future winners. Despite its widespread use in practice, it is quite disconcerting to note that there is only a rather limited economic understanding of this strategy Kelsey et al. [2011]. The most convincing explanation of momentum so far is given by Hong and Stein [1999]. Their study develops a model where only two types of market participants interact in the market: news watchers, that observe only information, and momentum traders, that trade solely based on price information. Because news watchers are prevented to observe the price, there is an underreaction of the price to information. Therefore, it is economically profitable for momentum traders to start chasing trends and trade on them. The difficulty for a momentum trader is to distinguish price changes due to a change in information from those caused by other momentum traders. Hence, "early momentum traders impose a negative externality on future momentum traders". Because of this, the presence of momentum traders induces a subsequent overreaction of the price to information.

5.2.1 Description of the experiment

Let us start by describing the setup of the experiment. As before, we first need to define model classes to describe data over a period of piecewise stationarity. As the goal is to capture persistence in the time-series of returns, we consider the class of autoregressive processes of order $k = 0, \dots, K$ given by

$$r_{\log}^{(t)} = \alpha + \sum_{k=1}^{K} \beta_k r_{\log}^{(t-k)} + e^{(t)}. \tag{5.3}$$

that form a collection of $K+1$ nested model classes. The residuals $e^{(t)}$ follow an i.i.d. Gaussian distribution with zero mean and constant variance σ^2. The Gaussianity assumption is not necessary to compute an estimate of the parameters but used in



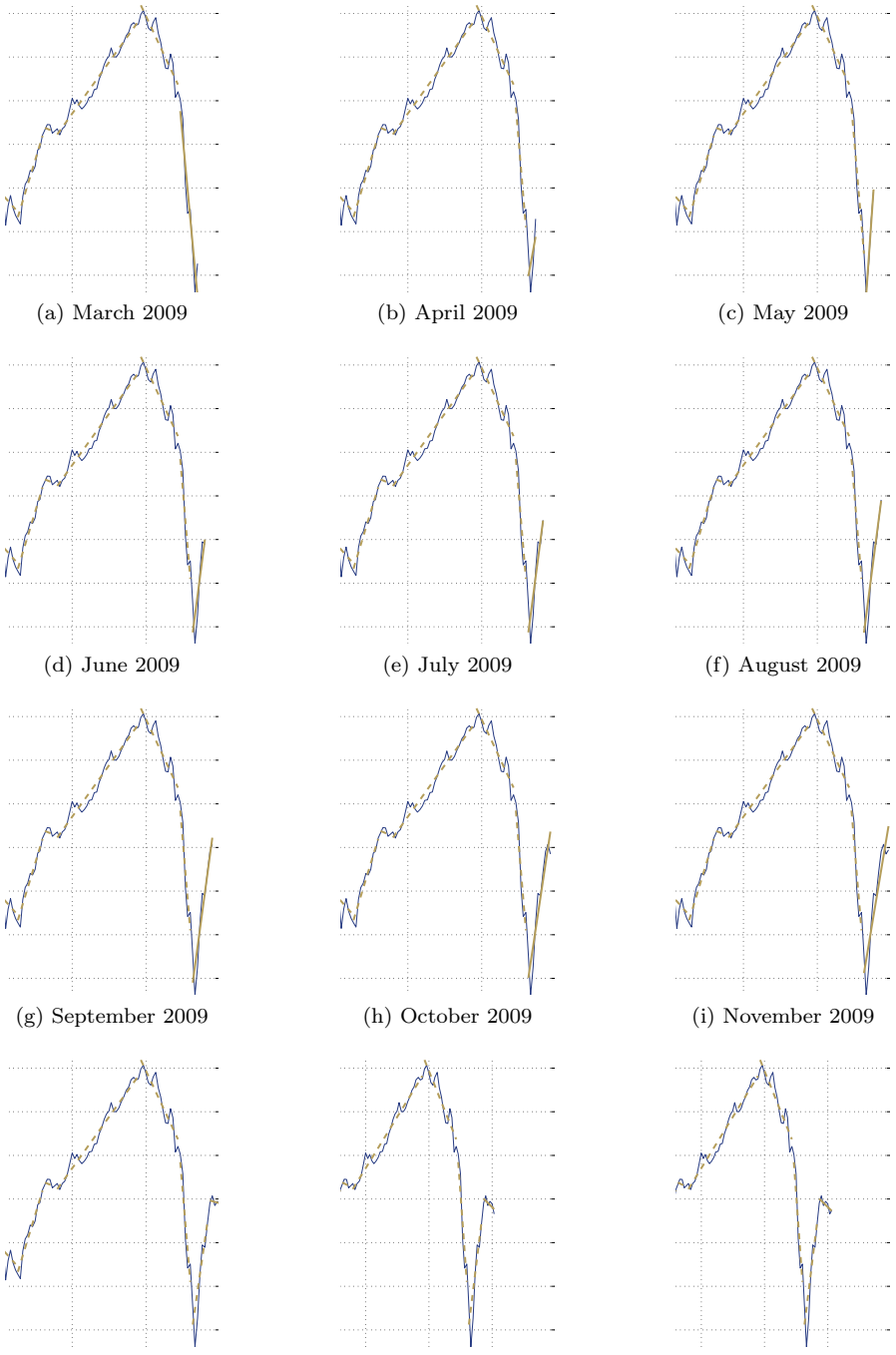
(a) July 2007 (b) August 2007 (c) September 2007

(d) October 2007 (e) November 2007 (f) December 2007

(g) January 2008 (h) February 2008 (i) March 2008

(j) April 2008 (k) May 2008 (l) June 2008

Figure 5.1: *Dynamic trend line, sequential plot from July 2007 until July 2008, MSCI World Financials index.*



(a) March 2009 (b) April 2009 (c) May 2009

(d) June 2009 (e) July 2009 (f) August 2009

(g) September 2009 (h) October 2009 (i) November 2009

(j) December 2009 (k) January 2010 (l) February 2010

Figure 5.2: *Dynamic trend line, sequential plot from March 2009 until February 2010, MSCI World Financials index.*

the derivation of the CNML criterion. We restrict attention to a simpler version of this model class that is obtained by constraining all regression coefficients to be equal.

$$\beta_k = \beta_1 = \beta, \ \forall k. \tag{5.4}$$

In that case,

$$\begin{aligned} r_{\log}^{(t)} &= \alpha + \sum_{k=1}^{K} \beta r_{\log}^{(t-k)} + e^{(t)} && (5.5) \\ &= \alpha + \beta r_{\log}^{(t-k)\to(t-1)} + e^{(t)}, && (5.6) \end{aligned}$$

where we have used the property of log-returns reviewed in Section 2.1.2 according to which the sum of returns over a given period is equal to the return over this period. Then, the vector of parameters is given by

$$\boldsymbol{\theta} = \begin{pmatrix} \alpha \\ \beta \\ \sigma^2 \end{pmatrix}. \tag{5.7}$$

These very simple model classes correspond to the one commonly used in momentum-type strategies where returns are projected on past returns over a given *horizon*, typically over the last 12 months. When the coefficient $\beta > 0$ (resp. $\beta < 0$), we obtain a momentum (resp. contrarian) strategy. Hence, the basic building block of our algorithm corresponds to a well-established strategy understood and applied by a large number of investment professionals. The output of the variable lookback algorithm is thus easily comprehensible and does not suffer from the black box problem of other quantitative strategies. Moreover, the role of our proposed algorithm is to select optimally and simultaneously the order of the process and the lookback window. The order in this case corresponds to the horizon of past returns over which returns are projected. The choice of the lookback window corresponds to a segmentation of the sequence into segments where different "modes" can be identified, namely: (i) momentum strategy works (ii) contrarian strategy works (iii) neither momentum nor contrarian strategy work. It is safe to include this latter option in our model selection scheme. If the system detects this mode, an investor enters a neutral position and avoids taking risky bets.

For the considered classes of models, the CNML criterion is readily available and we report here the equations for the sake of completeness. Details can be found in Rissanen et al. [2010]. Define the vector of regressors

$$\boldsymbol{f}^{(t)} = \begin{pmatrix} 1 \\ r_{\log}^{(t-k)\to(t-1)} \end{pmatrix}. \tag{5.8}$$

The following equations are used to update the estimator of the parameters of the model sequentially

$$\hat{\boldsymbol{\theta}}^{(t)} = \hat{\boldsymbol{\theta}}^{(t-1)} + \hat{e}^{(t)} \boldsymbol{V}^{(t-1)} \boldsymbol{x}^{(t)} \tag{5.9}$$

$$\hat{e}^{(t)} = \left(r^{(t)} - \left(\boldsymbol{x}^{(t)} \right)^T \hat{\boldsymbol{\theta}}^{(t-1)} \right) / \left(1 + c^{(t)} \right) \tag{5.10}$$

$$\boldsymbol{V}^{(t)} = \left(\boldsymbol{I} - \frac{\boldsymbol{V}^{(t-1)} \boldsymbol{f}^{(t)} \left(\boldsymbol{f}^{(t)} \right)^T}{\left(1 + c^{(t)} \right)} \right) \boldsymbol{V}^{(t-1)} \tag{5.11}$$

$$c^{(t)} = \left(\boldsymbol{f}^{(t)} \right)^T \boldsymbol{V}^{(t-1)} \boldsymbol{f}^{(t)}, \tag{5.12}$$

where $\hat{e}^{(t)}$ represents the prediction error estimates, $\boldsymbol{V}^{(t)}$ the inverse of the covariance matrix of the regressors and $1 + c^{(t)}$ the relative increase of (Fisher) information induced by the observation at time t. (5.9) and (5.11) define an online update rule for the estimation of the ordinary least-squares estimator of the parameters of the model. The latter is given by

$$\hat{\boldsymbol{\theta}}^{(t)} = \left(\left(\boldsymbol{F}^{(t)} \right)^T \boldsymbol{F}^{(t)} \right)^{-1} \left(\boldsymbol{F}^{(t)} \right)^T \boldsymbol{y}^{(t)}. \tag{5.13}$$

where

$$\boldsymbol{F}^{(t)} = \left(\left(f^{(1)} \right)^T, \ldots, \left(f^{(t)} \right)^T \right)^T. \tag{5.14}$$

Also, since the residuals are Gaussian, the least-squares estimator corresponds to the maximum likelihood estimator of the parameters. The online rule is derived from this definition following the same methodology as Example 4. That is, the regression coefficients are treated as latent state variables of a system whose state equation is given by

$$\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)} \tag{5.15}$$

and observation equation by

$$y^{(t)} = \left(\boldsymbol{f}^{(t)} \right)^T \boldsymbol{\theta}^{(t)} + e^{(t)}. \tag{5.16}$$

(5.9) and (5.11) result from the application of Kalman filter equations to the above state space model. Moreover, the CNML criterion is in this case given, for $t \geq m+2$, by

$$\begin{aligned} CNML(t,k) &= \frac{t-m}{2} \log_2 \tau^{(t)} - \log_2 \hat{e}^{(m+1)} - \log_2 \frac{\Gamma(\frac{t-m}{2})}{\Gamma(\frac{1}{2})} \\ &+ \sum_{t'=m+2}^{t} \log_2 \left(\sqrt{\pi}(1 + c^{(t')}) \right), \end{aligned} \tag{5.17}$$

where Γ is the Gamma function and $\tau^{(t)}$ is the sum of squared prediction error estimates

$$\tau^{(t)} = \sum_{t'=m+1}^{t} \left(\hat{e}^{(t)} \right)^2. \tag{5.18}$$

Note that (5.17) criterion is not an approximation. We can obtain one by applying Stirling's formula for the Gamma function Whittaker and Watson [1990]

$$\log_2 \Gamma(t) \approx \frac{1}{2} \log_2 \frac{2\pi}{t} + t \log_2 \frac{t}{\mathrm{e}}. \tag{5.19}$$

Then, since $\Gamma(1/2) = \sqrt{\pi}$, the CNML criterion can be rewritten as

$$\begin{aligned} CNML(t,k) &= \frac{t-m}{2}\log_2(2\pi \mathrm{e}\frac{\tau^{(t)}}{t-m}) + \sum_{t'=m+1}^{t}\log_2\left(1+c^{(t')}\right) \\ &\quad + \frac{1}{2}\log_2 t + O(1), \end{aligned} \tag{5.20}$$

where all constants are grouped in the generic $O(1)$ term. We recognize the first term in the glsrhs that corresponds to the entropy of a Gaussian process with variance $\tau^{(t)}/(t-m)$. The first term also resembles the cusum of squares, a test statistic used to identify a change in regression relationships Brown et al. [1975]. But our approach is purely data-dependent and does not contain any parameter, in particular, an arbitrarily chosen test size.

5.2.2 Comparison with other momentum-type strategies

Momentum-type strategies are not only documented by numerous academic studies but also applied by a large number of investment professionals. It is thus of great interest to compare our approach with existing ones. This is the topic of this section.

Firstly, we distinguish two large classes of momentum-type strategies, namely cross-sectional and time-series momentum. On the one hand, cross-sectional momentum, documented by Jegadeesh and Titman [1993], uses the relative performance of stocks to identify past winners and losers. The strategy is obtained by building a zero investment portfolio as follows. At every time t, stocks in the universe are sorted based on their last 12 months performance and a portfolio going long (resp. short) stocks in the top (resp. bottom) decile is constructed. This strategy generates a significant positive performance. The performance is even increased when measuring past performance ignoring last month's observation. Furthermore, this performance is interpreted as a reward for a distinct source of risk, and it is thus included in many factor models. On the other hand, time-series momentum, reviewed by Moskowitz et al. [2012] for future contracts on various assets and by Sanford and Cooper [2006] or Koo and Panigirtzoglou [2008] for European fixed income instruments, identify past winners based on their absolute performance. Our approach is closely related to time-series momentum, but differs in various aspects. Our model is richer as it allows more complex dynamics: our algorithm identifies based on the significance of β and its sign, which is the most appropriate strategy to play: momentum strategy ($\beta > 0$), contrarian strategy ($\beta < 0$) or none (β not significant). Moreover, the horizon of our strategy is not a parameter, potentially fined-tuned on a certain dataset, but selected by our algorithm.

Secondly, the idea of using past returns as proxy for private information is already found in Hou and Moskowitz [2005], but their model uses a fixed structure with the last 5 lagged monthly returns. Our approach is more adaptive and decides on the appropriate structure in a completely data-dependent manner.

Thirdly, our strategy is also related to an older strand of finance literature that studies the autocorrelation in the time-series of returns Lo and MacKinlay [1988]. This study concludes on the presence of a small but significant effect up to high lag. There are two shortcomings in this approach. Firstly, the test of autocorrelation allows

only testing static effect: a fixed lag of past return is correlated with returns over a given investment horizon. Secondly, the analysis technique implicitly assumes the underlying time-series is stationary and tends to average out effects. For example, if the underlying time-series switches between two modes both characterized by a high autocorrelation but of opposite sign, an analysis over the entire time-series is not going to detect any effect, despite its significance over each period. By allowing a flexible structure and departing from the stationarity assumption, our approach does not suffer from these limitations.

Finally, it is well recognized that momentum, like any other quantitative strategies, does not work consistently over time and various studies try to address this issue. For example, Kent [2010] studies methods to hedge the momentum strategy, so as to reduce losses when momentum fails. Also, Kelsey et al. [2011] study the influence of external variables on the performance of momentum. One of these variables is a measure of volatility of the market and the authors conclude that when the volatility increases, the performance of momentum decreases. It has proven wrong in 2009, when the level of volatility was very high and momentum extremely profitable. Our approach is a contribution to this effort, but it bases its decision solely on price information and past performance of the momentum strategy.

5.2.3 Results

We discuss in this section the results of the experiment and evaluate the output of the variable lookback algorithm by backtesting the resulting investment strategy.

Illustration of algorithm output

Figure 5.3 illustrates the output of the algorithm for one of the time-series of log-returns, the MSCI World Information Technology Index (MXWOIT). On the one hand, the plots on the left are pertaining to the choice of lookback window. The top left plot 5.3(a) represents the segmentation of the time-series in periods of piecewise stationarity. The segmentation points are indicated by the vertical lines whose thickness is proportional to the number of times a switching point is selected. This is compared against the evolution of the log-price and the exponential weighted volatility estimate. We observe that the algorithm segments periods not according to the general directions of the index (alternating growth and decrease of the log-price), but according to the volatility of the index (alternating high and low volatility). This makes sense since the estimation procedure is driven by the second order characteristics of the process. Also, crisis periods are typically characterized by an increase in the volatility. The bottom left plot 5.3(c) represents a complementary image of the top left one. It represents the evolution over time of the selected t_s value, the last time before which a switch happens. Compared to the top left plot, this one not only indicates where the switching point is detected but also when this happens.

On the other hand, the plots on the right are pertaining to the choice of model for the current period of piecewise stationarity. The top right plot 5.3(b) represents the type of strategy employed when making prediction of returns at time t, green for momentum and red for contrarian, depending on the sign of the coefficient $\hat{beta}$ conditional on information at time $t-1$. The type of strategy is again compared against the log-price and exponential weighted volatility of the index. Note that

the figure does not permit to conclude whether the strategy outperforms or not the index. Indeed, the experimenter could be tempted to check visually whether red areas correspond to falling prices and green areas to soaring ones. However, the experimenter tends to select areas in the graph confirming his preconceived idea and ignoring the others. This well documented bias is known as the confirmation biais Taleb [2008]. Going back to the figure, the bottom right plot 5.3(d) represents the evolution of the selected order of the process. In this case, an order of 12 means that return is projected on the return over the last 12 months. Although in our model the order is supposed to remain fixed over periods of piecewise stationarity, we see that in this plot, the order varies within these periods. This is because we expect the estimate of the order to be imprecise, especially just after a switch has happened. Moreover, we observe that periods of decrease of the log-price are associated with a drop of the order of the process. This is interesting since these extreme drop of prices are typically driven by investors' panic or liquidity shocks hence not by fundamental information. The algorithm correctly concludes that there is no information that can be extracted from these prices, hence the reduction in the order of the process. The effects we have observed are not particular to the information technology sector and apply more generally to other ones. For example, Figure 5.4 illustrates the output of the algorithm for MSCI World Financials index.

Backtest

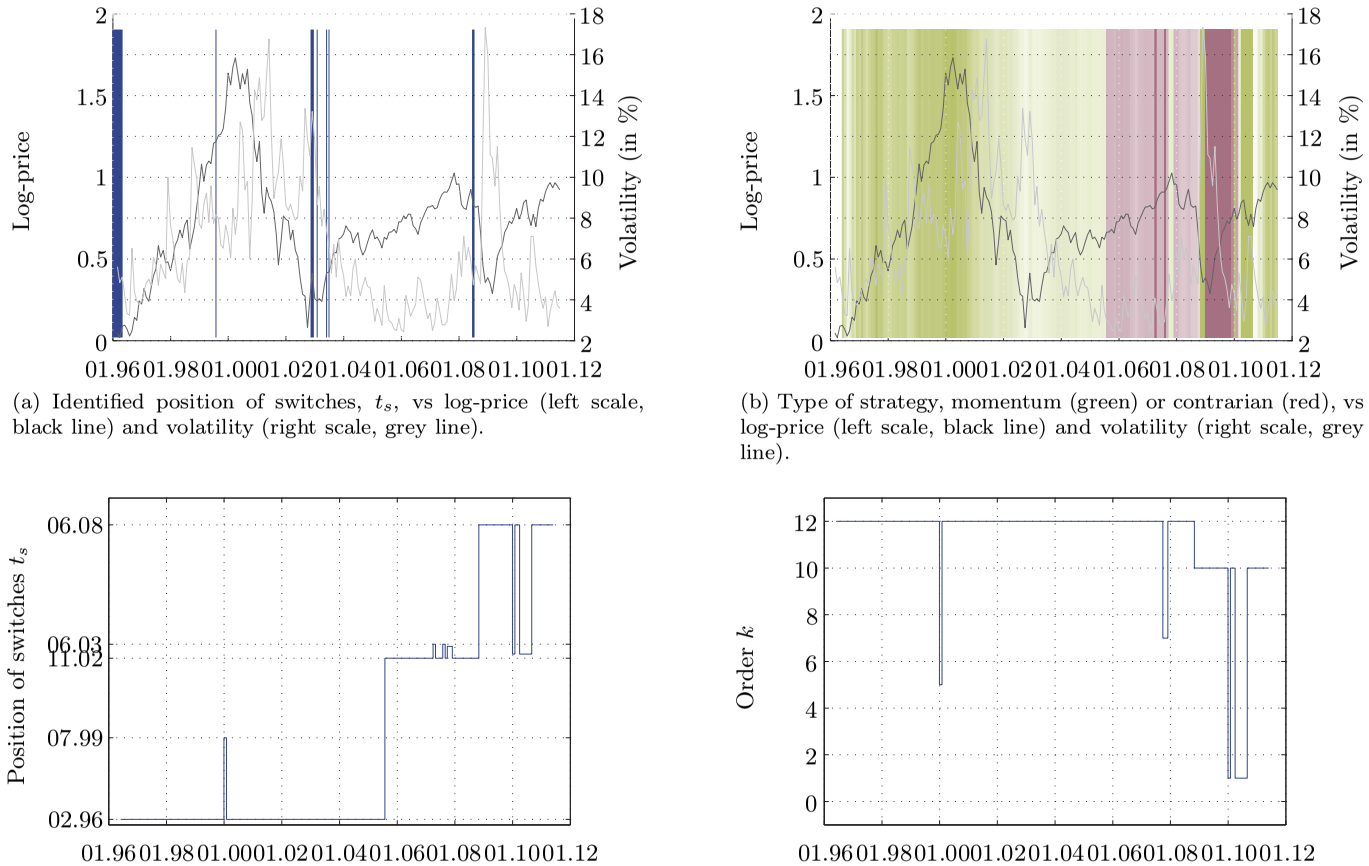
Let us now describe the method used to test formally the output of the algorithm. At every time t, the algorithm selects the relevant lookback, or alternatively the last time before which a switch happens $\hat{t}_s^{(t)}$, and the order of the process $\hat{k}^{(t)}$. The resulting prediction of the next period return is then given by

$$\hat{r}_{\log}^{(t+1)} = \hat{\beta}\left(\hat{k}^{(t)}, \hat{t}_s^{(t)}\right) r_{\log}^{(t-\hat{k}^{(t)})>(t)}, \tag{5.21}$$

which is observable at time t. In order to build a strategy, the following trading rule is applied, where the leverage depends on the sign of the prediction at time t. We go long 130% (resp. 70%) of the asset at time t when our prediction is positive (resp. negative). Also, we go long 100% of the asset when no prediction is made, i.e., when a model of order 0 is selected. Mathematically, the weight of the portfolio at time t, $w^{(t)}$, is given by

$$\omega^{(t)} = \begin{cases} 1.3 & \text{if} \quad \hat{k}^{(t)} > 0, \hat{y}^{(t+1)} > 0 \\ 0.7 & \text{if} \quad \hat{k}^{(t)} > 0, \hat{y}^{(t+1)} < 0 \\ 1 & \text{if} \quad \hat{k}^{(t)} = 0. \end{cases} \tag{5.22}$$

This change of leverage is implemented in practice by borrowing or lending money at a risk-free rate, $r_f = 0$ for the sake of simplicity. Also, the value of the potential leverage is fixed and is not scaled by an estimate of the volatility. Scaling the leverage by the volatility is a well-known practitioners' trick to inflate the Sharpe ratio of a strategy. Here we would like to test solely predictions and disentangle whether the improvement in Sharpe ratio is caused by our algorithm or by a scaling of the leverage. The adaptive momentum strategy is compared against two standard benchmarks, namely the passive indexing strategy and the simple 12 months momentum. In the simple passive strategy, the exposure in the asset is maintained at 100% for all time.



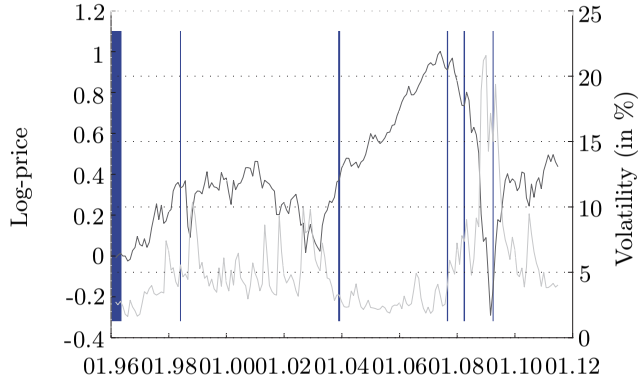
(a) Identified position of switches, t_s, vs log-price (left scale, black line) and volatility (right scale, grey line).

(b) Type of strategy, momentum (green) or contrarian (red), vs log-price (left scale, black line) and volatility (right scale, grey line).

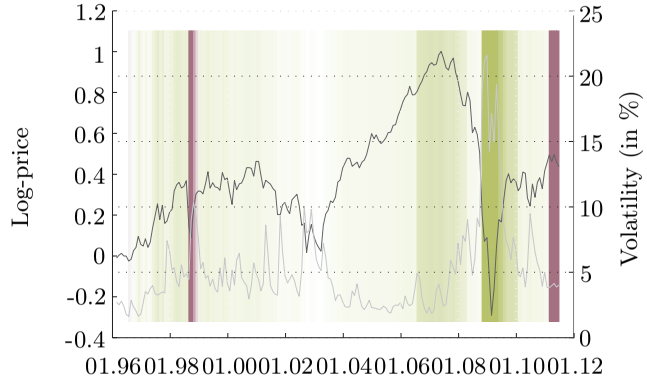
(c) Evolution over time of the position of switches, t_s, conditional at time t.

(d) Evolution over time of the identified order of the process, k, conditional at time t.

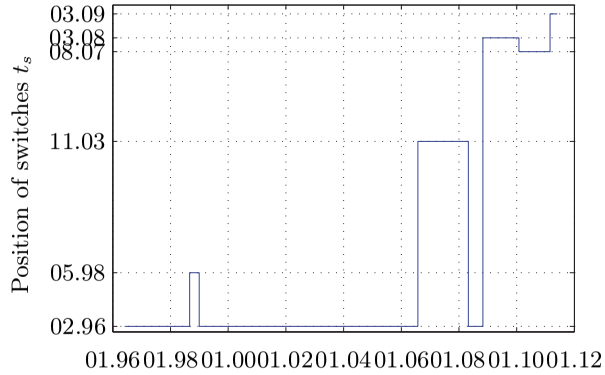
Figure 5.3: *lllustration of the output of the variable lookback algorithm. The latter is used to estimate model (5.6) on the time-series of log-returns of the MSCI World Information Technology index (MXWOIT) from 01.01.1995 until 01.07.2011.*



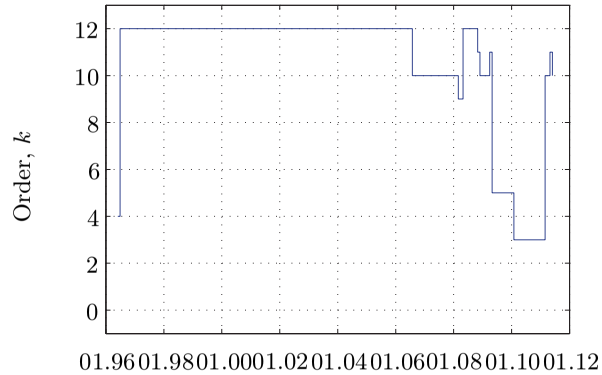
(a) Identified position of switches, t_s, vs log-price (left scale, black line) and volatility (right scale, grey line).



(b) Type of strategy, momentum (green) or contrarian (red), vs log-price (left scale, black line) and volatility (right scale, grey line).



(c) Evolution over time of the identified position of switches, t_s, conditional at time t.



(d) Evolution over time of the identified order of the process, k, conditional at time t.

Figure 5.4: *lllustration of the output of the variable lookback algorithm. The latter is used to estimate model (5.6) on the time-series of log-returns of the MSCI World Financials index (MXWOFN) from 01.01.1995 until 01.07.2011.*

This strategy bets on the absence of momentum effect. In the simple 12 months momentum strategy, we go long 130% (resp. 70%) when the return over the last 12 months is positive (resp. negative). In other terms, we assume $\hat{\beta} > 0, \forall t$ and this strategy bets on the omnipresence of the momentum effect. We run the backtest on the time-series of log-returns of the 10 MSCI World total return sector indices sampled at a monthly frequency from January 1995 until July 2011. Each time-series represents the weighted average returns of a large basket of stocks belonging to a certain sector. The grouping of stocks in a sector follows the GICS classification and the weight of each stock depends on its market capitalization.

Backtest results

Figure 5.5 represents the result of the backtest for MSCI World Financials (MXWOFN) for all three strategies, namely passive indexing (grey), simple 12 months momentum (blue) and adaptive momentum (gold). We present a whole array of indicators, so as to build a complete overview of the performance of the strategy. Let us detail each now. The top left plot 5.5(a) represents the evolution of the log-wealth for all strategies. A strategy outperforms another one when the slope of their log-wealth significantly diverges, because this reflects a significant difference in their growth rate. For example, we observe that both standard and adaptive momentum strategies outperform the passive strategy between the end of 2005 and July 2007. Likewise, adaptive momentum outperforms simple momentum between the first quarter of 2010 and the first quarter of 2011. The top right plot 5.5(b) decomposes the return of the strategy on an annual basis. Ideally, a strategy should outperforms the market in both directions, i.e., have a higher positive return during positive years and a lower negative returns during negative years.. The aim of this plot is also to analyze the stability of the outperformance of a strategy against its benchmarks. The question is whether the outperformance is stable over time, or concentrated on a couple of extreme values. The latter are likely to bias summary statistics, which are reported in the legend. In particular, we give the average annual return, the average annualized volatility and the average annualized Sharpe ratio. The latter is defined as the ratio between the average of monthly return and the standard deviation of monthly return, scaled by $\sqrt{12}$ so as to obtain an annualized measure. We observe that the adaptive momentum strategy outperforms the simple indexing in almost all years and even outperforms the simple 12 months momentum during the present crisis. What is particularly striking is the ability of the adaptive momentum strategy to outperform indexing during the negative year of 2008, like the simple momentum strategy, but to also outperform both strategies during the consecutive positive years in 2009. We come back to the analysis of the behavior of the strategy in 2009 in the next section. Moreover, the bottom left plot 5.5(c)represents the evolution of the portfolio weight for simple and adaptive momentum strategies. The turnover, defined as the sum of the absolute difference between consecutive portfolio weights,

$$TO = \sum_{t=2}^{T} |\omega^{(t)} - \omega^{(t-1)}|, \tag{5.23}$$

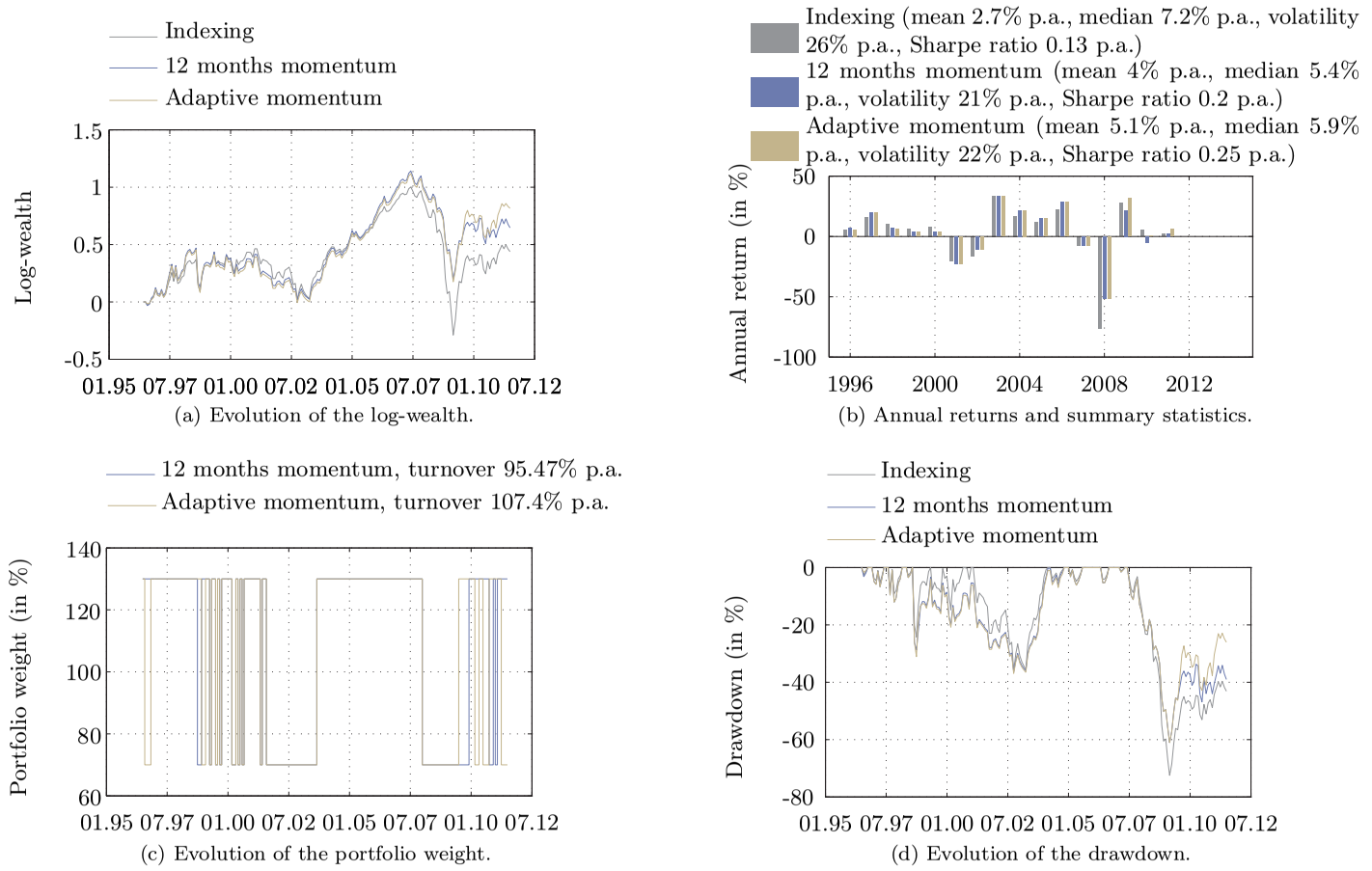
measures the amount of trading incurred by a strategy. We report this number, since the profits generated by a strategy might be consumed by trading costs, which are not

included in our analysis. Note that the turnover of the adaptive momentum strategy is approximately equal to that of simple momentum. Finally, the bottom right plot 5.5(d) represents the drawdown of the strategy. This is defined as the difference between the current maximum value of the wealth and the current value,

$$DD^{(t)} = max_{t' \in \{1,\dots,t\}} \left\{ W^{(t)} \right\} - W^{(t)}, \tag{5.24}$$

and is expressed in percentage term. It is commonly used to assess the risk of a strategy. The drawdown of the strategy is similar to the drawdown of indexing and simple momentum before the 2008 crisis and is significantly better afterwards. Again, the comments we have made are not particular to the financials sector and we report as well the complete performance analysis for the Information Technology index in Figure 5.6.

The annualized Sharpe ratio is our key performance measure when comparing strategies. Furthermore, recall that leverage does not affect the Sharpe ratio, as reviewed in Section 2.1.4. This is similar to the situation in signal processing where the amplification of a signal does not modify its SNR ratio. Therefore, a strategy that constantly leverages its investments has the same Sharpe ratio as passive indexing. Also, a comment regarding the use of Sharpe ratio in conjunction with log-returns is in order. In a static context, the use of log-returns implicitly correct for risk by assuming log-utility. Indeed, if an investor has log-utility, he maximizes the logarithm of his terminal wealth, which is equal to the sum of log-returns. Thus there is no need to correct for risk and expected log-return is the correct performance evaluation metric. In an intertemporal context, where one wants to adjust for risk period-by-period, and probably also for changes in the investment opportunity set, Rubinstein's CAPM Rubinstein [1976] offers a framework within which to think about this. An investor with a log-utility maximizes the sum of the logarithm of its intermediate consumption. Assuming joint log-normality of the asset or portfolio return and the market portfolio, there exists a relationship resembling Sharpe's CAPM, where the beta needs to be doubled (and corrected by a factor proportional to the variance term that is likely going to be small). Again, the use of log-returns correct for risk. However, the widespread use of Sharpe ratio in conjunction with log-returns in practice explains why we use it as our key performance measure. Figure 5.7 represents the annualized Sharpe ratio of the three strategies for all sectors. Adaptive momentum outperforms its two benchmarks in almost all assets. Figure 5.8 represents also the annualized Sharpe ratio for all sectors, when returns are projected on past cumulated returns ignoring last month's observation. In this case, the outperformance of adaptive momentum compared to its two benchmarks is even more evident. This is somehow related to the already mentioned method of ignoring the last month's observation in cross-sectional momentum strategy. This is also related to the recent paper of Novy-Marx [2012] that advocates that in the simple 12 months strategy, the first 6 months portion explains most of the predictive power of the strategy. It seems to be a stable feature of data, but there is yet only partial economic justification for it. With daily data, market microstructure effects have been put forward to explain it, but they are unlikely to persist at a monthly level Jegadeesh and Titman [1993]. This has also been interpreted as a consequence of the short-term contrarian effect caused by market overreaction de Bondt and Thaler [1985].



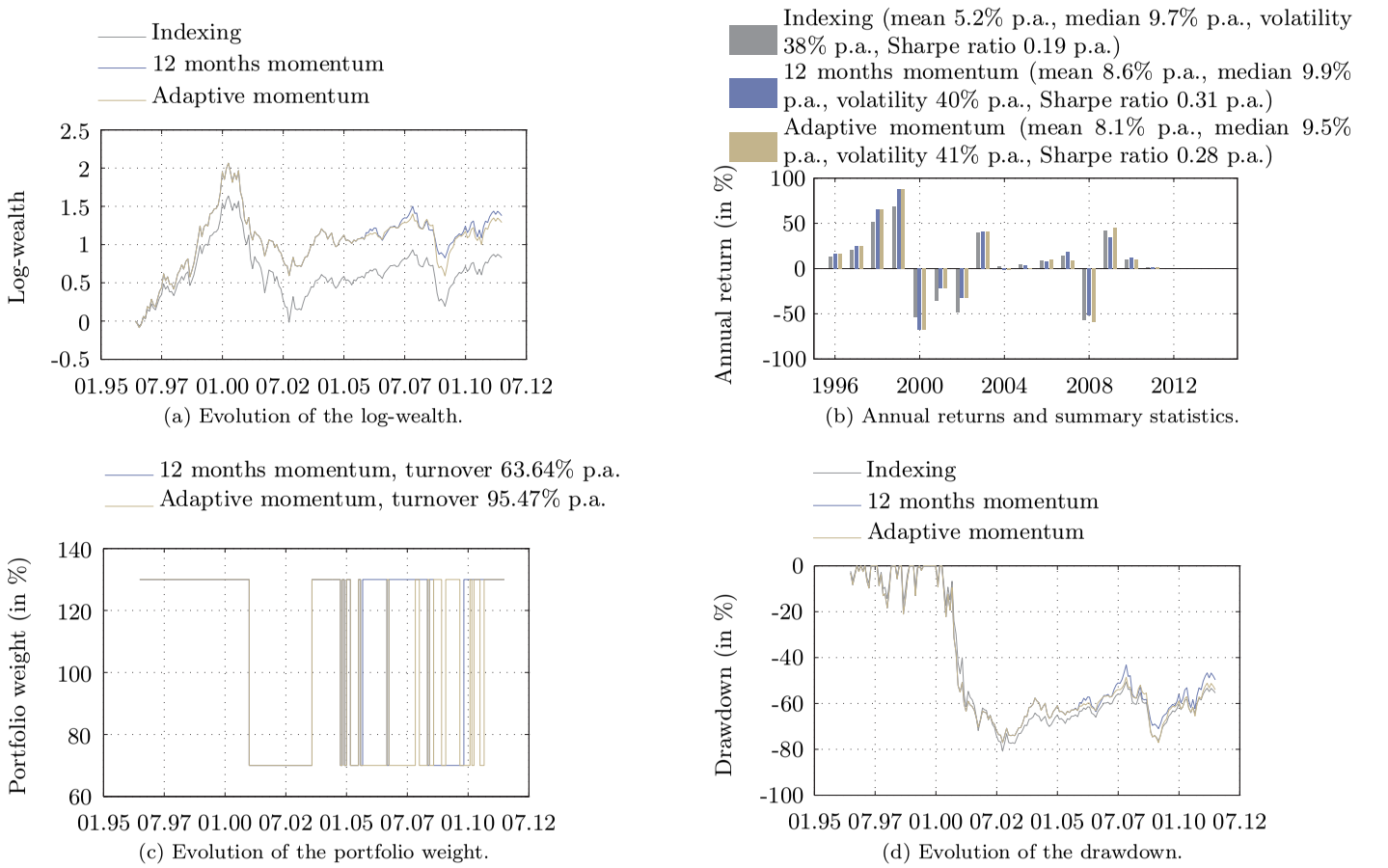
(a) Evolution of the log-wealth.

(b) Annual returns and summary statistics.

(c) Evolution of the portfolio weight.

(d) Evolution of the drawdown.

Figure 5.5: *Backtest of the indexing (grey), simple 12 months momentum (blue) and adaptive momentum strategy (gold) for the MSCI World Financials index (MXWOFN) from 01.01.1995 until 01.07.2011.*



(a) Evolution of the log-wealth.

(b) Annual returns and summary statistics.

(c) Evolution of the portfolio weight.

(d) Evolution of the drawdown.

Figure 5.6: *Backtest of the indexing (grey), simple 12 months momentum (blue) and adaptive momentum strategy (gold) for the MSCI World Information Technology index (MXWOIT) from 01.01.1995 until 01.07.2011.*

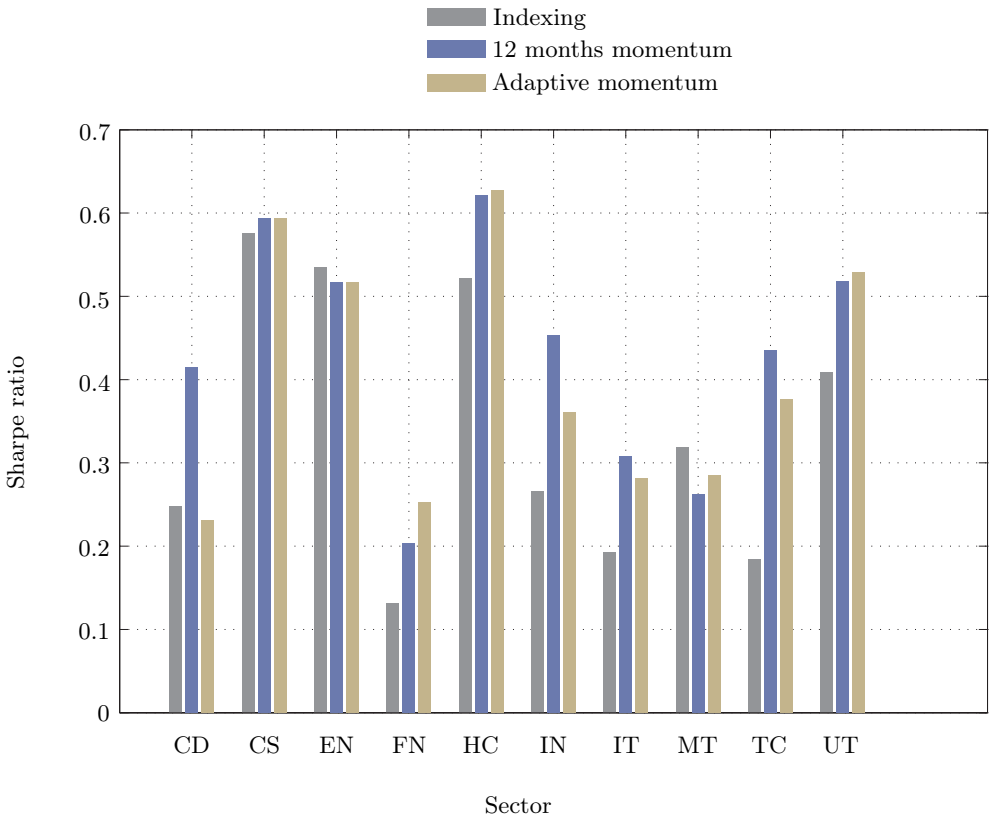


Figure 5.7: *Sharpe ratio of the indexing (grey), simple 12 months momentum (blue) and adaptive momentum strategy (gold) for all MSCI World GICS level 1 sectors. The adaptive momentum outperforms the two benchmarks.*

Behavior during the 2009 momentum crash

2009 was certainly a challenging year for investment managers. The market started with a sharp decline in the first quarter, followed by an important rebound. For example, in Switzerland, the Swiss market index (SMI) index has doubled between March and October 2009. This period was particularly difficult for quantitative strategies, in particular momentum strategies. Together with the 1930's, it is one of the historical periods where momentum has failed dramatically. In this section, we highlight here the behavior of our adaptive momentum strategy during this particular phase.

Figure 5.9 represents our analysis of the strategy between January 2007 and January 2010 for MSCI World Financials index (MXWOFN). The top left plot 5.9(a) represents the type of strategy used, green for momentum, red for contrarian, against the log-wealth of the strategy and its two benchmarks. The shade of the color is proportional to the significance of the effect. The top right plot 5.9(b) represents the sign of the predicted return, green for positive, red for negative values, against the log-wealth of the strategy and its two benchmarks. Again the shade of the color is proportional to the absolute value of the predicted return. The bottom left plot 5.9(c)

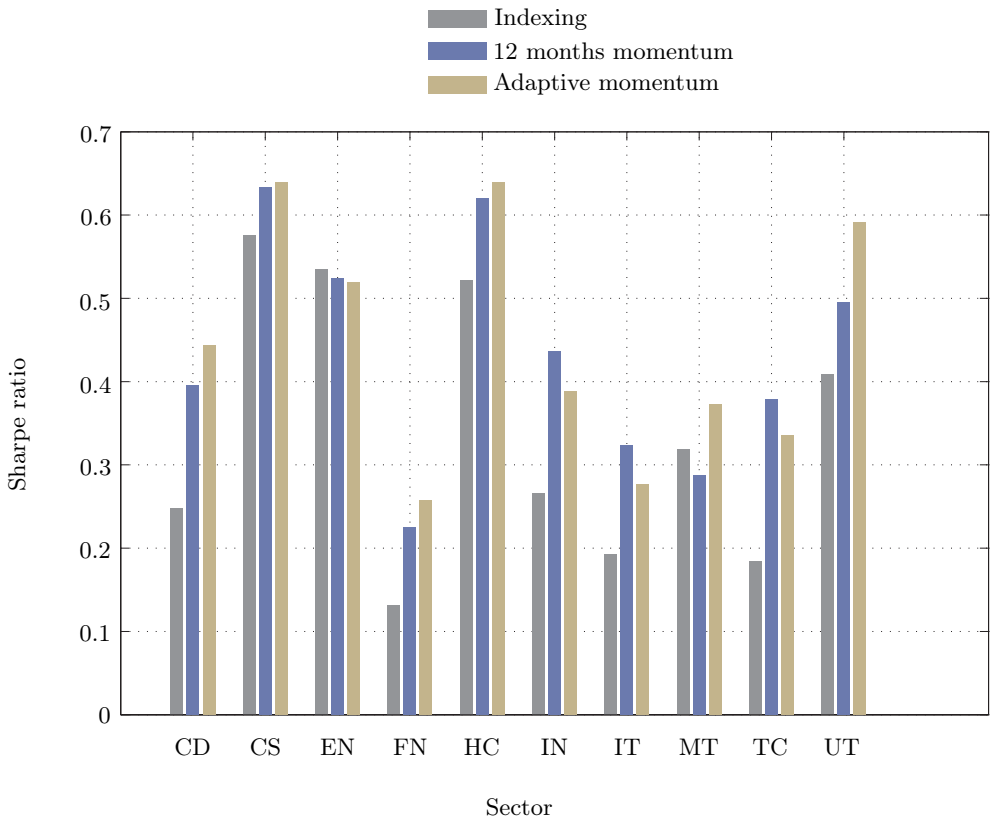
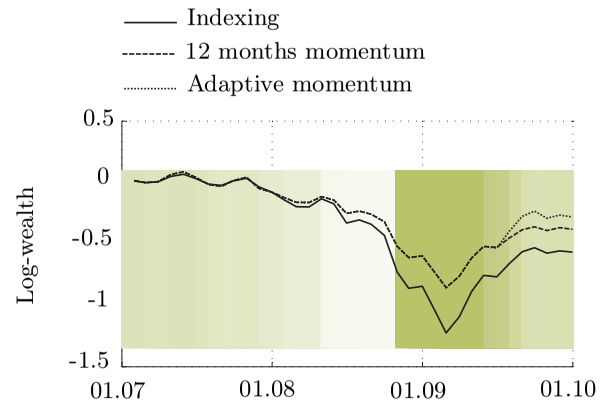


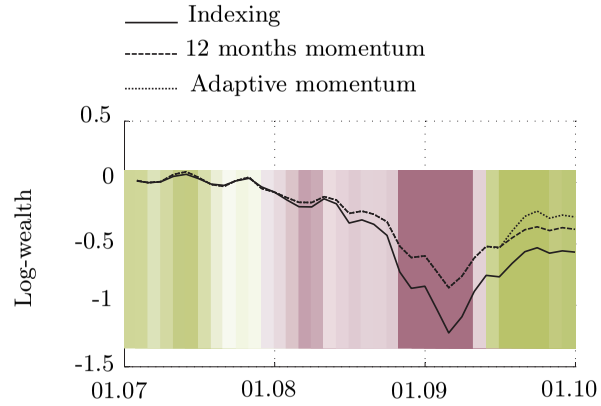
Figure 5.8: *Sharpe ratio of the indexing (grey), simple 12 months momentum (blue) and adaptive momentum strategy (gold) for all MSCI World GICS level 1 sectors. The outperformance is even stronger when removing the last month observation.*

pertains to the choice of lookback window conditional on information up to time t, whereas the bottom right one 5.9(d) to the choice of order of the process conditional on information up to time t. In the case of MXWOFN, we see that the algorithm decides to use a momentum strategy during the whole period. Generally speaking, playing a momentum strategy is a good idea when there is a clear trend in the market, either positive or negative, but not during turning points. At the beginning of 2007, the algorithm is already using a momentum strategy with an order of 12, in other words, with an horizon of 12 months. It is well positioned to benefit from the current rising market condition. When the market enters a crisis in August 2007, past cumulated returns become quickly negative. The algorithm is again well positioned with a momentum strategy, as the market carries on moving downward. In February 2009, the algorithm decides to change the lookback window and consider that only observations as of March 2008 are relevant for the current environment and simultaneously reduces the horizon of momentum to three months. This decision is solely based on data and cannot be further justified. When the market starts to rebound in March 2009, the adaptive momentum uses a shorter horizon than the simple 12 months

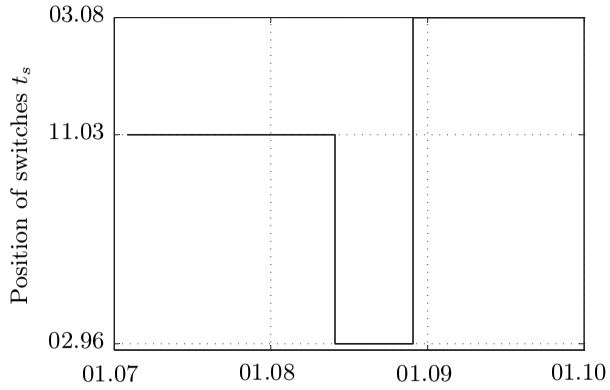
momentum, past cumulated returns becomes quickly positive, and the algorithm is correctly positioned to benefit from the rebound as of May 2009. On the contrary, simple momentum is lagging behind, and takes more time to adapt. This change of order explains why adaptive momentum outperforms simple momentum as of this point. Figure 5.10 represents the same analysis for the Utilities sector (MXWOUT). The outperformance is not explained in this case by a change of order, that more or less stays constant during the whole analysis period, but by a change of lookback window. This change happens in May 2009, and the type of strategy moves from being a momentum strategy to a contrarian one. As already noted, the cumulated returns over the past 12 months is negative, such that the algorithm predicts positive returns and is correctly positioned for catching the rebound. This decision lasts for three months, and is then reverted to a momentum strategy, which is again correct since the past cumulated 12 months return has become positive and the algorithm still bets on a positive return value. Simultaneously, the lookback window goes back to its original value. In summary, these two figures illustrate two mechanisms, namely change of order and change of lookback, used by the algorithm to adapt to current market conditions.



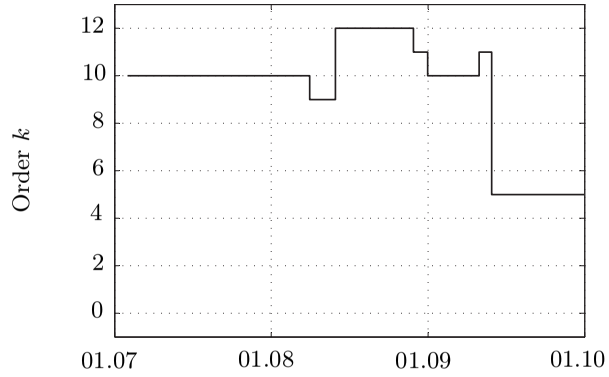
(a) Log-wealth vs. type of strategy: momentum (green) and contrarian (red).



(b) Log-wealth vs sign of prediction: positive (green) vs negative (red) values.

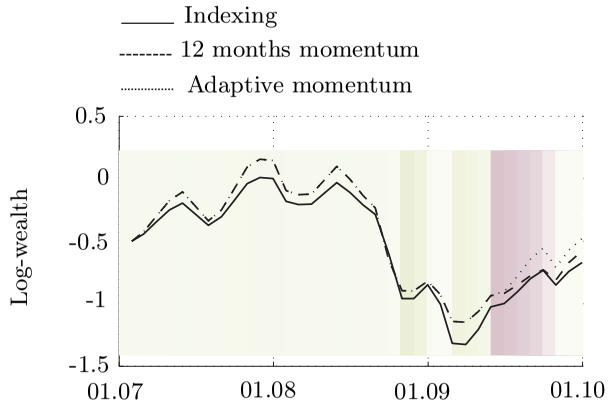


(c) Identified position of switches, t_s, conditional at time t.

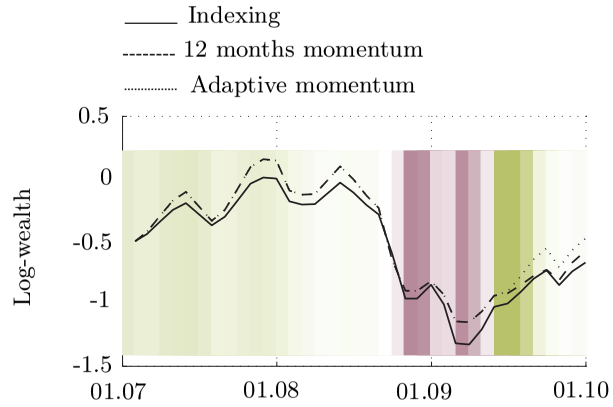


(d) Identified order of the process, k, conditional at time t.

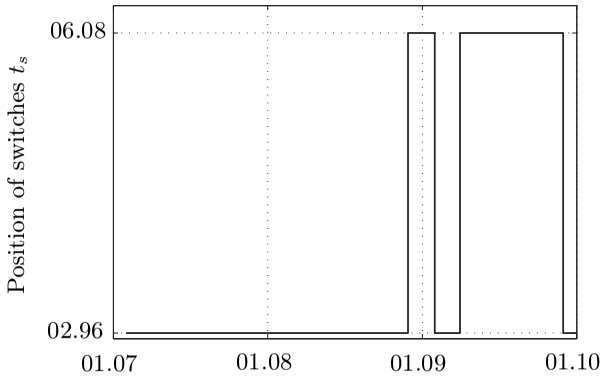
Figure 5.9: *Behavior of the adaptive momentum strategy during the 2009 momentum crash, MSCI World Financials (MXWOFN).*



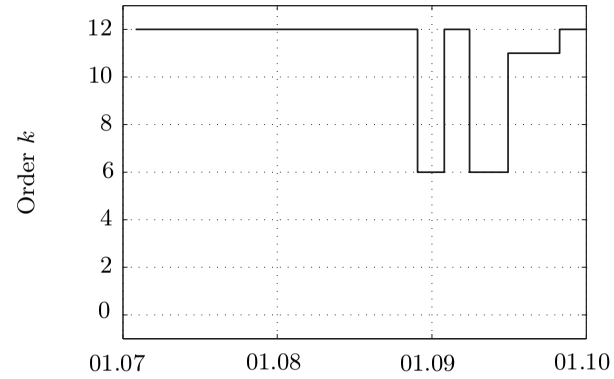
(a) Log-wealth vs. type of strategy: momentum (green) and contrarian (red).



(b) Log-wealth vs sign of prediction: positive (green) vs negative (red) values.



(c) Identified position of switches, t_s, conditional at time t.



(d) Identified order of the process, k, conditional at time t

Figure 5.10: *Behavior of the adaptive momentum strategy during the 2009 momentum crash, MSCI World Utilities (MXWOUT).*

Summary

In this chapter, we have reviewed two applications of the variable lookback algorithm to finance. In the most general terms, the algorithm is used to sequentially and simultaneously identify the various modes of the market by segmenting the original time-series in different periods of piecewise stationarity, as well as the appropriate structure of the model over each period. More specifically, the first application results in an automated and data-dependent algorithm to draw the trend line for the time-series of log-prices. This is mainly an illustration of the algorithm and its underlying concepts, rather than the basis for a realistic investment strategy. The second application results in an adaptive version of the momentum strategy. We backtest the strategy that levers the portfolio depending on the significance of the identified effect and the sign of predicted returns. The strategy outperforms standard benchmarks, passive indexing and simple 12 months momentum strategy. Moreover, because of its ability to vary the lookback window and the order the process, the strategy performs particularly well during the 2009 momentum crash.

Chapter 6

Conclusion

> I have come to the conclusion, after many years of sometimes sad experience, that you cannot come to any conclusion at all.
>
> *In Your Garden Again*
> Vita Sackville-West

6.1 Summary

We have started this thesis with the introduction of the idea of a variable lookback model, a statistical model where a time-varying portion of the information set is relevant when forming predictions. The introduction of a variable lookback is an implication of information asymmetry in the market; the presence of a time-varying number of informed investors, which controls for the speed of diffusion of private information into the price, justifies a time-varying lookback. We have also seen that the model is more general, as it aims to handle nonstationarity inherent to financial time-series. More specifically, our aim was to derive a general procedure that constantly monitors when a model works and when not, what are the appropriate model structure and lookback, and this in a purely data-dependent manner.

We have further motivated our approach in two different manners. On the one hand, we have reviewed existing applications of signal processing in quantitative finance, with the intention to illustrate the shortcomings of existing approaches rather than provide an exhaustive treatment. Both applications we have reviewed are concerned with the general problem of factor modeling; the first one uses Kalman filtering to improve the estimation of the model, the second one principal component analysis to automatically extract a series of orthogonal factors and their associated loadings. The first shortcoming pertains to the structure of the model. The questions are what is the appropriate order of the model? is the effect significant? When the model is estimated by maximum likelihood, increasing the number of parameters increases the fit of the model. As a consequence, the conclusion that the most complex model is also the best one is unavoidable. This calls for the development of theoretically sound

complexity measure; a model of higher order is chosen over one of lower order if the improvement in fit at least compensates the additional complexity. Another approach consists in introducing new parameters in the model, so as to better capture stylized facts present in the time-series of interest. But the curse of dimensionality limits the practical impact of this approach. The second shortcoming pertains to the nonstationarity of the signals. The current methods bypass the problem, by either assuming some form of local stationarity, as in local windowing, or by modeling the switching process itself. With local windowing, the choice of window size is problematic, as the optimal window size changes over time. And modeling the switching process suffers quickly from the curse of dimensionality. Furthermore, Markov switching processes not only quickly suffer from the curse of dimensionality but are also not well suited to capture long term memory effects characteristic of processes where humans beings are involved. On the other hand, since our initial motivation came from studies of markets under information asymmetry, we have reviewed and compared two larges bodies of literature, the noisy REE and BNE. Whereas in REE price emerges as the equilibrium between price-taking agents under the rational expectations assumption, it results from the strategic interaction of investors in BNE. In both theories, price plays an articulated role in conveying information from informed to uninformed market participants. Both theories agree that the informativeness of the price is proportional to the relative proportion of informed investors. However, in noisy REE, the price does not fully reveals all private information, whereas all informational inefficiencies disappear very quickly in BNE. Interestingly, our view reunite both theories, in the sense that the type of equilibrium the price system converges to is better described by noisy REE but strategic interaction explains how this equilibrium emerges. We have also tested our initial intuition in the context of experimental finance. We have highlighted the process of diffusion of information into the price. We have also defined the notion of time of maximally informative price and used the Context algorithm to estimate it. We have verified that the time of maximally informative price is inversely proportional to the number of informed investors, a supportive evidence for the idea of a variable lookback model.

We have then developed a general algorithm to estimate the variable lookback model and coined the term variable lookback algorithm for it. It is solely based on the assumption of piecewise stationarity. This means that the system switches between different regimes at unknown point in time and there are no assumptions on the number of switches or their positions. The algorithm is rooted in the MDL approach, that equates learning with data compression. The code length of the universal model with respect to a class of models defines a complexity measure. Furthermore, the choice of models and model classes is left to humans, since Kolmogorov's complexity, which is based on a more general description method using computer programs, is noncomputable. Moreover, the variable lookback algorithm works as follows. It maintains various model classes, each based on a different portion of the information set, whose number grows linearly in the number of observations. This set of estimators is combined with a quadratic dynamic programming algorithm resembling the Viterbi algorithm, so as to select simultaneously the order of the process and the relevant lookback window by minimizing an information theoretic score. The latter measures the ability of the model to fit data, penalized by the model complexity and the complexity of the underlying switching process between different regimes. From

a learning perspective, the algorithm decides sequentially either to keep and update the current model based on the latest observation (continued estimation) or to restart the learning procedure and forget all properties learned so far (restarted estimation). The restarted model is chosen over the continued one, when it is better by a certain threshold, whose value is not a parameter, which can later be fine-tuned to fit a certain dataset. Rather, it corresponds to the complexity of encoding an additional switch in the system associated with a chosen description method. Furthermore, our solution is characterized by a certain number of features

(i) Purely data-dependent: we aim to only derive evidence from data, there are no parameters that can be fine-tuned on a specific dataset.

(ii) Individual sequence sense: we have access to only one realization of the process and conclude from there.

(iii) Sequential learning: the algorithm forms sequential predictions that are compared against realizations; we learn properties from the data on the fly.

(iv) Model diversification: we consider in parallel several nested model classes, the appropriate one is chosen in a completely data-dependent manner. We constantly monitor when a model works and when not. The latter case allows entering a neutral position in our investments.

(v) Complexity and dimensionality: our approach explicitly takes into account the model complexity in a theoretically sound manner.

Finally, we have conducted various tests of the algorithm, first on simulated time-series then on field data. The first test was designed to assess the ability of the CNML criterion to detect the correct order of the process. The second test was assessing the ability of the algorithm to detect a switch when the underlying model switches from a highly correlated autoregressive process of order 1 to an i.i.d. process and vice versa. We have then presented two applications of the variable lookback algorithm in finance. In the most general terms, the algorithm is used to sequentially and simultaneously identify the various modes of the market by segmenting the original time-series in different periods of piecewise stationarity, as well as the appropriate structure of the model over each period. More specifically, the first application results in an automated and data-dependent algorithm to draw the trend line for the time-series of log-prices. This is mainly an illustration of the algorithm and its underlying concepts, rather than the basis for a realistic investment strategy. The second application results in an adaptive version of the momentum strategy. We have backtested the proposed strategy that levers the portfolio depending on the significance of the identified effect and the sign of the predicted returns. The strategy outperforms standard benchmarks, passive indexing and simple 12 months momentum strategy. Moreover, because of its flexibility in varying the lookback window and the order the process, the strategy performs particularly well during the 2009 momentum crash.

6.2 Future research

Our work like any research projects calls for various extensions, which, mainly for the lack of time, could not further be explored. This is the topic of this section. They are ordered by fields of contribution.

Future research in experimental finance and neurofinance

Certain extensions of our research lie in the area of experimental finance and neurofinance. A first extension of our research is the design of experimental markets, whose information process is more realistic and the analysis of the resulting datasets. We have seen in the description of the experiments in Section 3.2.1 that the information arrival process used is extremely simplistic; private information is revealed to the insiders before trading starts and is resolved by the revelation of the final dividend at the end of the trading period. On the one hand, this is inspired by the theoretical model of Brennan and Cao [1997] or Holden and Subrahmanyam [1992]. Thus there exists a detailed theoretical analysis whose implications can be compared against the results of the experiment. On the other hand, this is too simplistic to reflect the complexity of the information process in financial markets. Usually, information arrives as a point process, where subsequent pieces of information either subsume or complement existing ones. Therefore, there is a need to design experiments whose information process better reflects this complexity. Towards this goal, we can imagine an experiment where subsequent pieces of private information are revealed to the insiders, such that the precision of the private information signal, i.e., the inverse of the variance of the noise of this signal, is increased over time in a controlled manner.

Experimental finance is also an ideal setup to address the question of how humans beings perform model selection tasks. We have seen in this thesis that, from a mathematical perspective, the minimum description length approach presents several advantages. The principle is rooted on sound epistemological foundations, in particular, as it does not assume the existence of a "true" distribution according to which observed data are distributed. The only principle underlying MDL is Occam's razor, a general principle in sciences and engineering that advocates to pick the simplest best explanation. MDL is also related to the Bayesian approach, but it avoids some of its interpretation difficulties. From a neuroscience perspective, there is a consensus around the idea that humans estimate statistical models by reinforcement learning. Intuitively, this means that humans perform the estimation task by trial an error. Starting from an initial guess, the parameters are progressively updated at an adaptation speed proportional to the prediction error. This is similar to the method of Kalman filtering. Whereas the estimation question has been addressed, the question remains on how do human beings choose among competing explanations of data. In particular, is there a neuroscience foundation for the theory of MDL? More specifically, this could be addressed by designing an experiment where subjects are asked to make sequential predictions. Their predictions can then be compared against the predictions made by automated model selection procedures, in particular Bayesian or MDL-based ones. Intuitively, predictions that match closely one approach support it. There are various challenges that should be addressed. First, it is better to ask subjects to elicit a prediction, rather than a model order, as this notion is more intuitive. But the prediction has to be linked unambiguously with the underlying order of the model. Moreover, the experiment should be designed in such a way that its is possible to disentangle various automated model selection criteria. Observe, for example, that both MDL and Bayesian criterion explore the entire set of predictors, but differ in the way they are combined. Bayesian approach uses a mixture of various predictors, whereas MDL selects a single one.

Future research in signal processing

Our work also calls for various extensions in the area of signal processing. Firstly, the version of the variable lookback algorithm we have developed is univariate, as it processes each time-series individually. Of course, it is possible to combine them to form a portfolio of univariate strategies. But, because a great deal of effects is present in the cross-section of stocks, the development of a multivariate version of the algorithm is desirable. Certainly, the extension of the model selection criterion used to encode observations over a period of piecewise stationarity to a multivariate setting is not problematic. The problem comes from the rapidly impractical number of possibilities, when both the order of each process and their lookback window is allowed varying simultaneously. One obvious solution is to allow only the order of each univariate process varying. At a monthly frequency and at the index level, this is motivated by the fact that the univariate algorithm usually picks dates that corresponds to major events that affect the cross-section of stocks. But this is hard to reconcile this approach with our initial justification for the variable lookback model, where the lookback window is a proxy for the relative speed of information diffusion in the market, created by a time-varying number of informed investors.

Secondly, it is of great interest to develop a measure of the confidence in the selection performed by the algorithm. Of course, the compression factor, i.e., the ratio of code length between two models, offers a mean of pairwise comparison between two models. Intuitively, the larger this factor is, the more confidence we can have in the selected model. But a theoretical rooted measure, similar to the one found in Thomas et al. [1995], is desirable.

A third generalization is the development of new MDL-based model selection criteria, in particular the conditional normalized maximum likelihood for more general classes of models. So far, all the models we have used can be casted as a linear regression model, for which the CNML criterion can be evaluated analytically. However, a considerable number of models in finance, for example the celebrated GARCH model Bollerslev [1986] or the autoregressive conditional duration model for high frequency data Engle and Russel [1998] cannot be framed as a linear regression model. Also, an estimate of the parameters of these models are obtained by (quasi-)maximum likelihood methods, where the optimization is performed numerically. Therefore, the evaluation of the denominator in the definition of the CNML criterion (4.49) is also numerical, and efficient scheme must be developed to estimate this integral, for example, by cleverly sampling the sample space.

Future research in mathematical finance

Other extensions lie in the area of mathematical finance. Firstly, we can extend the backtest we have documented in different ways. First, the backtest we have presented scale the leverage only based on the sign of the prediction, and more elaborate trading rules could be considered. We can also perform the same test on other asset classes, sampled at different frequencies. Gold for example is particularly challenging for current methods, therefore of potential interest. We shall as well consider other equities indices, for example based on countries or regions, as they exhibit different statistical features.

Secondly, our experience shows that the quality of the estimator has an impact on

the selection performed by the variable lookback algorithm. So far, we have used the ordinary least-squares estimator, which is the best linear unbiased estimator if the error is zero-mean, uncorrelated and homoskedastic. Since these properties do not hold for financial time-series, it is of interest to test other estimators. One approach consists in renormalizing the original time-series, so as to compute a weighted maximum likelihood estimator Steude [2011]. For example, we can give more weight to the most recent observations and this can be generalized to include other variables that proxy for the arrival of information in the market. Another approach consists in using the feasible weighted least-squares estimator. This is a two-pass procedure. The first pass corresponds to the OLS estimator and is used to compute an estimate of the variance of the residuals. The latter is used as renormalizing factor in the second pass. Various heteroskedastic consistent covariance estimators that have been proposed in the literature White [1980] should be compared.

Bibliography

Admati, A. A. (1985). A noisy rational expectations equilibrium for multi-asset securities markets. *Journal of Finance*, 53(3):629–657.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transacions on Automatic Control*, 19(6):716–723.

Avellaneda, M. and Lee, J.-H. (2010). Statistical arbitrage in the US equities market. *Quantitative Finance*, 10(7):761–782.

Biais, B., Bossaerts, P., and Spatt, C. (2010). Equilibrium asset pricing and portfolio choice under asymmetric information. *Review of Financial Studies*, 23(4):1503–1543.

Biais, B. and Wolley, P. (2011). High frequency trading. Working paper.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

Bossaerts, P., Frydman, C., and Ledyard, J. (2010). The speed of information revelation and eventual price quality in markets with insiders : comparing two theories. *Review of Finance*. Revise and Resubmit.

Bossaerts, P. and Hillion, P. (1999). Implementing statistical criteria to select return forecasting models: what do we learn? *Review of Financial Studies*, 12(2):405–428.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Brennan, M. and Cao, H. (1997). International portfolio investment flows. *Journal of Finance*, 52(5):1851–1880.

Brown, R. L., Durbin, J., and Evans, J. M. (1975). Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society. Series B (Methodological)*, 37(2):149–192.

Bruguier, A., Quartz, S., and Bossaerts, P. (2010). Exploring the nature of "trader intuition". *Journal of Finance*, 65(5):1703–1723.

Cont, R. (2011). Statistical modeling of high-frequency financial data. *IEEE Signal Processing Magazine*, 28(5):16–25.

Cooper, M., Gutierrez, R. C., and Marcum, B. (2005). On the predictability of stock returns in real time. *Journal of Business*, 78(2):469–500.

Cover, T. and Ordentlich, E. (1996). Universal portfolios with side information. *IEEE Transactions on Information Theory*, 42(2):348–363.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. Wiley.

Dawid, P. (1984). Present position and potential developments: some personal views. *Journal of the Royal Statistical Society A*, 147(2):278–292.

de Bondt, W. and Thaler, R. (1985). Does the stock market overreacts. *Journal of Finance*, 40(3):793–805.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271.

Engle, R. F. and Russel, J. R. (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica*, 66(5):1127–1162.

Fama, E. (1970). Efficient capital markets: a review of theory and empirical work. *Journal of Finance*, 25(2):383–417.

Fama, E. F. and French, K. R. (1992). The cross-section of expected stock returns. *Journal of Finance*, 47(2):427–465.

Fung, W. and Ksieh, D. (2002). Asset-based style factors for hedge funds. *Financial Analysts Journal*, 8(5):16–27.

Ganesan, G. (2011). A subspace approach to portfolio analysis. *IEEE Signal Processing Magazine*, 28(5):49–60.

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

Gruenwald, P. (2005). A tutorial introduction to the minimum description length principle.

Gruenwald, P. (2007). *The Minimum Description Length Principle*. MIT Press.

Hart, P. E., Nilsson, N. J., and Raphael, B. (1968). A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107.

Harvey, A. C. (1991). *Forecasting structural time-series model and the Kalman filter*. Cambridge University Press.

Harvey, C. R., Liechty, J. C., Liechty, M. W., and Mueller, P. (2010). Portfolio selection with higher moments. *Quantitative Finance*, 10(5):469–485.

Hens, T. and Bachmann, K. (2009). *Behavioral Finance for Private Banking*. Wiley Finance.

Holden, C. and Subrahmanyam, A. (1992). Long-lived private information and imperfect competition. *Journal of Finance*, 47(1):247–270.

Hong, H. and Stein, J. S. (1999). A unified theory of underreaction, momentum trading, and overreaction in asset markets. *Journal of Finance*, 54(6):2143–2184.

Hou, K. and Moskowitz, T. J. (2005). Market frictions, price delay, and the cross-section of expected returns. *Review of Financial Studies*, 18(3):981–1020.

Jay, E., Duvaut, P., Darolles, S., and Chrétien, A. (2011). Multifactor models. *IEEE Signal Processing Magazine*, 28(5):37–48.

Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers. *Journal of Finance*, 48(1):65–91.

Jondeau, E. (2009). Econometrics. Unpublished lecture notes.

Jondeau, E. (2010). Asymmetry in tail dependence of equity portfolios. Working paper.

Karatzas, I. and Shreve, S. E. (2004). *Stochastic Calculus for Finance*. Springer Verlag.

Kelly, J. L. (1956). A new interpretation of information rate. *Bell System Technical Journal*, 35:917–926.

Kelsey, D., Kozhan, R., and Pang, W. (2011). Asymmetric momentum effects under uncertainty. *Review of Finance*, 15(3):603–631.

Kent, D. (2010). Momentum crashes.

Kim, C.-J. and Nelson, C. (1998). *State-Space Models with Regime-Switching: Classical and Gibbs-Sampling Approaches with Applications*. MIT Press.

Koo, G. and Panigirtzoglou, N. (2008). Global bond momentum.

Kozat, S. and Singer, A. (2008). Universal switching linear least-squares prediction. *IEEE Transactions on Signal Processing*, 56(1):189–204.

Krichevsky, R. and Trofimov, V. (1981). The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–207.

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

Lamoureux, C. and Lastrapes, W. (1990). Heteroskedasticity in stock return data: volume versus GARCH effects. *Journal of Finance*, 45(1):221–229.

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621.

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics*, 47(1):13–37.

Lo, A. and MacKinlay, C. (1988). Stock prices do not follow random walks: evidence from a simple specification test. *Review of Financial Studies*, 1:41–66.

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

Morellec, E. (2008). Introduction to finance. Unpublished lecture notes.

Moskowitz, T. J., Ooi, Y. H., and Pedersen, L. H. (2012). Time-series momentum. *Journal of Financial Economics*, 104(2):228–250.

Moura, J. M. F. (2005). Applied stochastic processes. Unpublished lecture notes.

Novy-Marx, R. (2012). Is momentum really momentum? *Journal of Financial Economics*, 103(3):429–453.

Oppenheim, A. V., Schafer, R. W., and Buck, J. R. (1999). *Discrete-Time Signal Processing*. Prentice Hall.

Plott, C. and Sunder, S. (1988). Rational expectations and the aggregation of diverse information in laboratory security markets. *Econometrica*, 56(5):1085–1118.

Poi, B. (2003). Swamy's random coefficient model. *The Stata Journal*, 3(3):302–308.

Prandoni, P. and Vetterli, M. (2008). *Signal Processing for Communications*. EPFL Press.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5):465–471.

Rissanen, J. (2005). Lectures on statistical modeling theory.

Rissanen, J. (2007). *Information and Complexity in Statistical Modeling*. Springer.

Rissanen, J., Roos, T., and Myllymäki, P. (2010). Model selection by sequentially normalized least-squares. *Journal of Multivariate Analysis*, 101(4):839 – 849.

Rubinstein, M. (1976). The valuation of uncertain income streams and the pricing of options. *The Bell Journal of Economics*, 7(2):407–425.

Sanford, G. and Cooper, G. (2006). Euro fixed income momentum strategy.

Sayed, A. H. (2003). *Fundamentals of Adaptive Filtering*. John Wiley & Sons.

Sbaiz, L. and Ridolfi, A. (2006). Statistical signal processing. Unpublished lecture notes.

Scannel, K. (2011). Rajaratnam guilty of insider trading. `http://www.ft.com/intl/cms/s/0/a22a994a-71b0-11e0-9adf-00144feabdc0.html`.

Schwager, J. D. (1995). *Technical Analysis*. John Wiley and Sons.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

SEC (2011a). Limit order. `http://www.sec.gov/answers/limit.htm`.

SEC (2011b). Market order. `http://www.sec.gov/answers/mktord.htm`.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423,623–656.

Shari, M. (2011). Quant shops innovate to survive. `http://www.ft.com/intl/cms/s/0/ce7548a8-46ac-11e0-967a-00144feab49a.html`.

Sharpe, W. F. (1964). Capital asset prices: a theory of market equilibrium under conditions of risks. *Journal of Finance*, 19(3):425–442.

Simonian, H. (2011). Three quit after Sonova insider trading probe. `http://www.ft.com/intl/cms/s/0/e46d57a6-5a96-11e0-8900-00144feab49a.html`.

Steude, S. (2011). Weighted maximum likelihood for risk prediction. Technical report, University of Zürich.

Taleb, N. N. (2008). *The Black Swam: the Impact of the Highly Improbable*. Penguin.

Telatar, E. (2006). Information theory and coding. Unpublished lecture notes.

Thomas, J. K., Scharf, L. L., and Tufts, D. W. (1995). The probability of subspace swap in the SVD. *IEEE Transactions on Signal Processing*, 43(3):730–736.

Timmermann, A. (1999). Moments of Markov switching models. *Journal of Econometrics*, 96(1):75–111.

van Erven, T. (2006). Momentum effect in MDL and Bayesian prediction.

Vetterli, M. and Kovacevic, J. (1995). *Wavelets and Subband Coding*. Prentice Hall.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269.

Wei, C. Z. (1992). On predictive least squares. *Annals of Statistics*, 20(1):1–42.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.

Whittaker, E. T. and Watson, G. N. (1990). *A Course in Modern Analysis*. Cambridge University Press.

Willems, F. (1996). Coding for a binary independent piecewise-identically-distributed source. *IEEE Transactions on Information Theory*, 42(6):2210 – 2217.

Woodbury, M. A. (1950). Inverting modified matrices. Technical report, Princeton University.

Yamanishi, K. (2007). Dynamic model selection with its applications to novelty detection. *IEEE Transactions on Information Theory*, 42(6):2180 – 2189.

Yamanishi, K. and Sakurai, E. (2010). Extensions and probabilistic analysis of dynamic model selection. In *Proc. of the 2010 Workshop on Information Theoretic Methods for Science and Engineering, Tampere, Finland.*

Curiculum Vitæ

Lionel Coulot

Audiovisual Communications Laboratory (LCAV)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
1015 Lausanne, Switzerland

Personal

Date of birth:	June 28, 1984
Nationality:	Swiss
Civil status:	Single

Education

2008–2012	PhD candidate at the School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
2002–2007	BSc and MSc in Communication Systems, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
2004-2005	Exchange Student, Carnegie Mellon University, Pittsburgh, PA, USA

Professional experience

09/2008–12/2012	**Research assistant**, Audiovisual Communications Laboratory (LCAV), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
01/2008–12/2012	**Quantitative Analyst Trainee**, LGT Capital Management AG, Pfäffikon SZ, Switzerland
03/2007–10/2007	**Visiting Student**, Microsoft Research Asia, Beijing, People's Republic of China
08/2006–08/2006	**Intern**, Lombard Odier, Geneva, Switzerland

Publications

Journal papers

1. L. Coulot, M. Vetterli and P. Bossaerts. MDL-based variable lookback algorithm and application to finance. *IEEE Transactions on Signal Processing*. Submitted
2. T. Blu, P. L. Dragotti, M. Vetterli, P. Marziliano and L. Coulot. Sparse sampling of signal innovation: algorithms and performance bound. *IEEE Signal Processing Magazine, Special Issue on Compressed Sampling*. March 2008

Conference papers

1. J. Kovacčević, G. Srinivasa, A. Chebira, L. Coulot, H. Kirschner, J.M.F. Moura, E.G. Osuna and R.F. Murphy. Topology Preserving STACS Segmentation of Protein Subcellular Location Images. *International Congress of the International Society for Analytical Cytology*. May 2006
2. L. Coulot, H. Kirschner, A. Chebira, J. M. F. Moura, J. Kovacevic, E. G. Osuna and R. Murphy. Topology Preserving STACS Segmentation of Protein Subcellular Location Images. *IEEE International Symposium on Biomedical Imaging*. April 2006

Awards and honors

2009-2012	LGT & Science full PhD scholarship
2008	Logitech award for Master thesis
2005-2007	Scholarship of excellence for Master studies

Languages

French (mother tongue), English (fluent) and German (fluent).