

Probabilistic data-driven modeling

T Aste

© Tomaso Aste 2019-23
not for distribution
July 25, 2023

© Tomaso Aste 2019-23
not for distribution
July 25, 2023

Dedicated to my daughters

© Tomaso Aste 2019-23
not for distribution
July 25, 2023

© Tomaso Aste 2019-23
not for distribution
July 25, 2023

Contents

<i>List of Symbols</i>	1
<i>Preface</i>	3
Part I Preliminary	7
1 Introduction	9
1.1 Modeling	9
1.2 Models as maps between data representations	10
1.3 Models for real and complex systems	11
1.4 Models as multivariate probabilities	12
2 Fundamentals of probability	14
2.1 Probability	14
2.2 Quantiles	17
2.3 Expected values	19
2.4 Moments of the distribution	21
2.5 Univariate and multivariate probabilities	22
3 Fundamentals of machine learning	24
3.1 Supervised learning	25
3.2 Unsupervised learning	25
3.3 Semi-supervised learning	26
3.4 Reinforcement learning	26
3.5 Training, validating and testing models	27
4 Fundamentals of networks	35
4.1 Networks and graphs	35
4.2 Adjacency matrix, weight matrix, and degree distribution	36
4.3 Paths and walks on networks	39
4.4 Centrality and peripherality	40
4.5 Counting paths through the power of the adjacency matrix	41
4.6 Propagation on networks	43
4.7 Trees forests and higher order networks	45
Part II Foundations of probabilistic modeling	49

5	Univariate probabilities	51
5.1	The normal distribution	51
5.2	Characteristic function	55
5.3	Stable distributions	57
5.4	Tendency towards Lévy-alpha stable distribution	59
5.5	The body and the tails of the distribution	60
5.6	Some common probability density functions	63
5.7	Mixture distributions	71
5.8	Generalized extreme value distribution	72
5.9	Infinitely divisible distributions	74
5.10	Probability distribution space	75
5.11	Data-driven modeling tutorial	75
6	Multivariate probabilities	77
6.1	Joint probabilities	77
6.2	Covariance matrix	78
6.3	Correlation matrix	79
6.4	Multivariate normal distribution	80
6.5	The elliptical distribution family	81
6.6	The Bayes' theorem and conditional probability	85
6.7	Conditional distribution for the elliptical distribution family	88
6.8	Data-driven modeling tutorial	95
7	Entropies	97
7.1	Shannon entropy	97
7.2	The maximum entropy principle	101
7.3	Joint entropy in multivariate systems	102
7.4	Kullback-Leibler divergence and the cross-entropy	104
7.5	Conditional entropy	106
7.6	Other entropies	107
7.7	Data-driven modeling tutorial	108
8	Dependency	109
8.1	Independent variables	109
8.2	Dependency and regression	110
8.3	Linear dependency	112
8.4	Multilinear regression and the covariance matrix	116
8.5	Precision matrix, inverse covariance and partial correlations	118
8.6	A generalized dependency measure	122
8.7	Correlation ratio	124
8.8	Non-linear correlations: rank correlations	126
8.9	Information-theoretic measures of dependency	127
8.10	Lagged correlations	134
8.11	Data-driven modeling tutorial	135
9	Stochastic processes and scaling laws	137
9.1	Stationarity	137
9.2	Scaling laws	139

9.3	The fractal dimension of signals	141
9.4	Random walk processes	144
9.5	Scaling laws of random walk processes	146
9.6	Self-affine, uniscaling processes	148
9.7	Multiscaling processes	150
9.8	Memory and tail effects on the scaling exponents	152
9.9	A stochastic process as a set of Student-t	155
9.10	Data-driven modeling tutorial	160
10	Causality	161
10.1	Cause and effect	161
10.2	Causality and correlations	163
10.3	Wiener-Granger causality	164
10.4	Transfer entropy	166
10.5	Deeper insights into causation	172
10.6	Data-driven modeling tutorial	174
11	Networks as representations of complex systems	176
11.1	Network construction by pruning or joining	176
11.2	Information filtering networks	178
11.3	Higher order network representations	189
11.4	Data-driven modeling tutorial	190
12	Probabilistic modeling with network representations	192
12.1	An information theoretic approach for network learning	192
12.2	Probability decomposition on a clique tree inference structure	198
12.3	Learning clique tree inference structures	200
12.4	Learning network representations for multivariate modeling	206
12.5	Data-driven modeling tutorial	208
Part III Model construction from data		211
13	Nonparametric estimation of univariate probabilities from data	213
13.1	The sample mean	214
13.2	Sample moments	215
13.3	The law of large numbers	216
13.4	Rate of convergence of the sample mean towards the expected value	218
13.5	Estimation of the probability mass function	220
13.6	Estimation of the cumulative probability distribution function	221
13.7	Estimation of the probability density function with histograms	223
13.8	Kernel density estimation (KDE)	227
13.9	Data-driven modeling tutorial	230
14	Parametric estimation of univariate probabilities from data	231
14.1	The method of moments	233
14.2	Maximum likelihood estimation (MLE)	234
14.3	Estimation of the tail exponent in fat-tailed distributions	240

14.4	Body-tail matching	242
14.5	Expectation maximization (EM)	244
14.6	Data-driven modeling tutorial	253
15	Estimation of multivariate probabilities from data	255
15.1	Non-parametric estimation of multivariate probabilities	255
15.2	Non-parametric, non-linear estimation of dependency	257
15.3	Pearson's estimation of the covariance matrix	259
15.4	Sample correlations	260
15.5	Maximum likelihood estimate of the multivariate normal distribution	262
15.6	MLE of multivariate Student-t with EM	263
15.7	The curse of dimensionality	264
15.8	Shrinkage estimation of the covariance matrix	278
15.9	Regularization	280
15.10	Data-driven modeling tutorial	288
16	Time series and probabilistic modeling	290
16.1	Estimation of scaling laws	290
16.2	Estimation of the generalized Hurst exponent	295
16.3	Tests for stationarity	301
16.4	Rolling windows, moving averages and exponential smoothing	303
16.5	Empirical mode decomposition	308
16.6	Time-clustering	311
16.7	Data-driven modeling tutorial	316
17	Construction of network representations from data	318
17.1	Construction of networks from thresholding	318
17.2	Construction of information filtering networks	322
17.3	Information filtering networks for probabilistic modeling	325
17.4	Causal networks construction	330
17.5	Data-driven modeling tutorial	333
18	Assessing the goodness of models	335
18.1	Null models	336
18.2	P-values	336
18.3	Comparing and testing probability estimates	339
18.4	Testing the goodness of regressions	345
18.5	Testing the goodness of classifications	351
18.6	Model evaluation via likelihood	355
18.7	Model selection	362
18.8	Non-parametric validation of models	364
18.9	Subdivision of the dataset in a train, validation, and test subparts	372
18.10	Data-driven modeling tutorial	374
Part IV Closing		377
19	Conclusions	379

19.1	The scientific method	379
19.2	Building models from data	380
19.3	Automated model construction	381
19.4	The end of parsimony	382
19.5	The rise of black boxes	383
19.6	Future of modeling	384
Part V Appendices		385
<i>Appendix A Methods to evaluate implicit equations</i>		387
A.1	Bisection method	387
A.2	Iteration towards a fixed point	387
A.3	Newton–Raphson method	388
A.4	Method of the secants	388
<i>Appendix B Some optimization problems and methods</i>		389
<i>Appendix C Principal components analysis</i>		392
<i>Appendix D Random forest</i>		394
<i>Index</i>		396
<i>References</i>		399

© Tomaso Aste 2019-23
not for distribution
July 25, 2023

© Tomaso Aste 2019-23
not for distribution
July 25, 2023

List of Symbols

\mathbb{R}	real numbers
X	random variable
$\mathbf{X} = (X_1, \dots, X_p)^\top$	column vector of p random variables
x	value of random variable X
\hat{x}	observed value (outcome) of a random variable X
$P(X)$	probability measure
$F(X) = P(X \leq x)$	cumulative distribution function
$\hat{F}(X)$	estimate of the cumulative distribution function
$P(X Y)$	conditional probability
$f(X)$	probability density function
$\hat{f}(X)$	estimate of the probability density function
$\tilde{f}(X)$	model probability density function
\mathcal{L}	likelihood
ℓ	log-likelihood
$\mathbb{E}(X)$	expected value
$Q(\gamma)$	quantile
$\mu = \mathbb{E}(X)$	mean
$\hat{\mu} = 1/q \sum_{k=1}^q \hat{x}_k$	sample mean
$Var(X) = \sigma_X^2 = \mathbb{E}((X - \mu)^2)$	variance
$\hat{\sigma}^2 = 1/q \sum_{k=1}^q (\hat{x}_k - \hat{\mu})^2$	sample variance
σ	standard deviation
$\hat{\sigma}$	sample standard deviation
$m_k = \mathbb{E}(X^k)$	k^{th} moment
$\hat{m}_k = 1/q \sum_{k=1}^q \hat{x}_k^k$	sample k^{th} moment
$\mu_k = \mathbb{E}((X - \mu)^k)$	k^{th} central moment
$\hat{\mu}_k = 1/q \sum_{k=1}^q (\hat{x}_k - \hat{\mu})^k$	sample k^{th} central moment
$\mathcal{G} = (\mathbf{V}, \mathbf{E})$	graph with vertex set \mathbf{V} and edge set \mathbf{E}
$\phi(\omega) = \mathbb{E}(e^{i\omega x})$	characteristic function
$\text{Cov}(X, Y)$	covariance
Σ	covariance matrix
$\hat{\Sigma}$	sample covariance matrix
\mathbf{J}	inverse covariance matrix or precision matrix
$\text{Corr}(X, Y) = \rho_{X,Y}$	correlation
$H(X)$	entropy
$\hat{H}(X)$	entropy estimate
$H(X Y)$	conditional entropy
$I(X; Y)$	Mutual information
$T_{X \rightarrow Y}$	Transfer entropy

Preface

What is this book about and why you might want to read it?

This book introduces and guides the reader through a selection of methodologies and approaches to construct models from data. These data-driven approaches have been originally developed in different fields, from statistics to complexity science. They are general procedures and tools that apply to any domain where models must be built from observational data. This book is also about probabilities and their identification and estimation from data. I indeed approach the general topic of data-driven modeling from the perspective that the entire domain, from linear regressions to deep learning neural networks, can be formulated in terms of the estimation of multivariate probabilities from data. This perspective provides a unifying frame of reference and an interpretation tool for all the topics and methods discussed in this book.

I provide practical tools to characterize, classify and model real systems starting from data. The material I present has become my modeling ‘toolbox’ that I have been using and refining for my research during the last 30 years. I included in this book what – I believe – is a relevant, meaningful, and essential selection of tools. In the case when several methods could be applied, I try to avoid listing all of them; rather, I chose to focus on the one that I believe is the most effective or the one I find simplest or – sometimes – that I like the most. I also give details about some of the often-overlooked aspects in this domain. For instance, how to deal with modeling when the number of observations is small or noisy, or non-stationary. I also discuss the interpretation of statistical validation results from a practical perspective and the complexity of many real-world systems.

The book aims to be readable by anyone with a background in mathematics at the bachelor’s degree level. It is intended to be used both by students as textbook support and by professionals and academics as a reference. I tried to avoid technical jargon and I introduce all new concepts in a way to be as self-contained as possible. Experts in some of the subjects might find some of the introductory parts too basic but, for them, there is no need to read everything from cover to cover. Indeed, I have organized the content in a way that makes it easy for an experienced reader to fast-forward through some basic parts of the book skipping most of the text while retaining the book’s perspective on the topics and then focusing on the more advanced parts.

The book presents my perspective on modeling real complex systems. In do-

ing so it introduces some of the most relevant, established, data-driven modeling tools currently in use, and also some approaches which are still in development. I present statistical tools useful to individuate regularities, discover patterns and laws in complex datasets, and apply them to devise models that help to understand these systems and help to predict their behaviors. Specifically, this book provides the mathematical instruments and the knowledge needed to:

- analyze and characterize complex datasets;
- compute relevant statistical quantities;
- quantify inter-dependency and causality structure between different variables;
- quantify the reliability of data;
- build models for description and prediction of real systems;
- validate hypothesis and models;
- select between alternative models;
- use the outcome of data analytics to develop better tools for description, characterization, and prediction.

Practical issues on data analysis and statistics are covered with specific examples.

I believe that the most important part of this book is Chapter 18 which covers model testing and selection. However, all other chapters are no less important. They are indeed essential to introduce the approach and methodologies and to understand the consequences and the applications. I tried to organize this book as a complete guide through this complex, rich, and fascinating field.

I divided the book into two main parts. Part I, is about foundations and it provides the theoretical basis for probabilistic data-driven modeling. Part II, is instead about the actual construction of models from data, it gives the practical tools and methodologies. Although, unorthodox, a fruitful way to read this book could be starting from Part II and then referring to the definitions and the fundamental concepts in Part I when they become needed. Indeed, I continuously provide cross-references to relevant sections and definitions.

I have taught the content of this book across several universities in Australia and the UK for several years. I had thousands of students who learned how to model real complex systems from this material and I believe I have developed a sense of the content that matters and how to deliver it.

There is a great need to increase data analytics and data-driven modeling capability in the industry. Peoples with data-driven modeling skills are in great and increasing demand. The instruments and tools provided in this book are essential to understanding, modeling, and making practical use of the very large quantity of data that most human activities are currently producing and collecting.

In this book, I adopt the perspective that real, complex, systems should be modeled in terms of the multivariate probability of all variables involved in the system. This would classify my approach and perspective in the realm of Bayesian statistics. However, I am a trained physicist and I have not been educated at statistics schools. Therefore, I can qualify myself neither as a Bayesianist nor as a frequentist. I must say that these classifications mean little to me. In some parts of the book, I report the Bayesian perspective and in other parts instead,

I adopt what is considered to be the frequentist perspective. I do this following what I consider the most intuitive way to present and understand the problem or, what I believe, is the most powerful instrument for the specific problem. I also avoid using more formal approaches to probability such as the Kolmogorov axiomatic formulation. I do not report proofs of theorems. However, I do report mathematically sound demonstrations when I believe they can be useful for a better understanding of the context. Despite the avoidance of jargon and the absence of some formalizations, I try to be precise and mathematically rigorous as much as possible.

Supporting material

The book has a large body of supporting material consisting of data, examples, and Python codes which are provided on the GitHub page:
<https://github.com/FinancialComputingUCL/DataDrivenModeling/>.

This material is organized by chapter number to help the reader to identify the relevant material while reading the book. However, it often mixes topics and methodologies from other chapters. There is also more general and self-containing extra material that covers topics that span across the entire book and focus on specific applications, certain methodologies, and particular narratives. This supporting material is designed to be dynamic and will continuously evolve, be updated, and be enhanced over time.

Acknowledgments

A large number of people have played a vital role in shaping this book, both directly and indirectly. I am grateful to the diverse group of colleagues and friends who generously devoted their time to reading the drafts, offering valuable suggestions, and uncovering numerous mistakes.

To my brilliant colleagues, thank you for your insightful feedback and constructive criticism. Your expertise and attention to detail have significantly improved the quality of this book. I am fortunate to have had such dedicated individuals in my circle.

To my daughters, my family, and my loved ones, thank you for your unwavering support and understanding. Your belief in me and your willingness to listen to my ramblings about scientific concepts have been instrumental in keeping me motivated and avoiding divergences.

This book is the result of a long journey involving the effort of paper coauthors, students, teaching assistants, and many others. I am immensely grateful to each and every person who contributed to its creation. Your involvement has enriched the content and made this scientific endeavor a truly rewarding experience. Thank you all for being part of this journey.

In acknowledging the invaluable contributions to this book, it would be remiss of me to overlook the individuals who have left an indelible mark on its

development. While this list is not exhaustive or intended to be completely comprehensive, I would like to highlight some of the major contributors who have played a significant role. At the forefront, I express my deepest gratitude to Antonio Briola, who, coinciding with the time I dedicated to writing this book, has been my dedicated Ph.D. student. Antonio's involvement went far beyond mere discussions and suggestions for improvements. In fact, he is the major contributor to the GitHub repository, comprising Python codes and datasets that support this book. Additionally, I am immensely grateful to two other exceptional Ph.D. students, Jeremy Turiel and Raymond Wang, who made substantial contributions to shaping the content of this book during that period. I owe a tremendous debt of gratitude to the Financial Computing and Analytics group at UCL, a community of brilliant and talented scientists who have supported me in numerous ways. Their direct involvement in reading the draft, providing insightful feedback, and indirectly inspiring me with their intellect and expertise have been instrumental. It is with great appreciation that I mention the following individuals from this remarkable group: Paolo Barucca, Silvia Bartolucci, Fabio Caccioli, Guido Germano, Denise Gorse, Giacomo Livan, Phelan Carolyn, Geoff Goodell, and Jiahua Xu (...).

Once again, I express my most profound appreciation to all those who have made this scientific endeavor possible. While I was writing I sent the draft version of the book to colleagues who are specialists in their respective fields asking for guidance. I received many very valuable feedbacks, suggestions, and corrections. Let me here mention some of the people toward whom I feel most grateful (...).

Finally, I extend my gratitude to the publishing team who helped transform this manuscript into a tangible book. Your expertise and dedication have been essential in bringing my ideas to life.

Part I

Preliminary

© Tomaso Aste 2019-23
not for distribution
July 25, 2023

© Tomaso Aste 2019-23
not for distribution
July 25, 2023

1

Introduction

We are overwhelmed with data, but information is hard to extract and the consequences of our actions are difficult to predict. Humans have learned to navigate this complexity quite efficiently. Mathematics, statistics, and intelligent machines are still far less capable.

1.1 Modeling

Humans constantly use models to interpret reality. We use models when we try to understand what is happening and when we try to predict what will happen next. We need them to manage, and perhaps reduce, the unpredictability of the world. We need them to take decisions. We need them to survive.

Without models, reality will be an overwhelming amount of data with no information. Indeed, in my view, models are the instruments that transform data into information. Humans construct models in various ways, consciously, or unconsciously, rationally, or instinctively. There is an increasing need to advance knowledge and understanding of the mechanisms and tools for reliable data-driven models. Nowadays, scientists and engineers are developing automated modeling tools. The purpose is to give machines the instruments to select and learn models, developing their ability to interact with the real world and make autonomous decisions. This is broadly called “artificial intelligence” (AI).

There are many different kinds of models; the ones we use, for instance, to choose our food at the market are different from the ones we use to place our satellites into the right orbit. As an example, let’s think about the model we use when we cross a busy road. Let’s make an effort to analyze what we normally, spontaneously, do in this situation. The process is surprisingly complex and sophisticated. Normally, we start looking if there is any vehicle and, if there isn’t, then we cross. In the case there are vehicles, then we gauge their distance and, if they are far enough, we cross. If the vehicles are not too far, then we estimate their speed and predict how long they will take to arrive pondering if it is sufficient for us to cross the road also accounting for the uncertainty of our prediction. There are many other variables that we evaluate. We distinguish cars from trucks and trucks from buses; they have indeed different typical speeds and behaviors. We might consider the vehicle’s trajectory and, in some cases, look at the driver and guess his or her intent to let us pass. We also consider, and normally dismiss, other variables that are in most cases marginal, for instance,

the color of the vehicle, its brand, or the age of the driver. This operation is an example of data-driven modeling. We have learned these operations from observations and, perhaps, from imitation. We don't have a general theory; we do all this spontaneously and instinctively without noticing the amount of data that we are processing and the way we make predictions, we ponder, compare, and, use them. We consider this *simple* but it is a very *complex* operation that we, modern humans, can perform quite easily but that is extremely hard to encode in machines or formulate precisely with our current mathematical and computational tools.

In this book, I'll discuss how to construct models from data.

1.2 Models as maps between data representations

Models can be deterministic, perhaps including some degree of uncertainty due to limited and noisy observations, or they can be probabilistic. Newton's gravitation law and the consequent modeling of the motion of celestial bodies is an example of deterministic modeling which can be written in the form $\mathbf{y} = g(\mathbf{x})$ and for which the true model can be *learned* with arbitrary precision provided a large enough number of observations.¹ The Maxwell-Boltzmann distribution, describing the speed of particles in a gas, is instead an example of probabilistic modeling where the law to be learned is the probability that a particle has a velocity smaller or equal to a given value v : $Probability(V \leq v)$.² This is a different kind of problem than the previous; however, also in this case, one can formulate the problem in the same form $\mathbf{y} = g(\mathbf{x})$ where, this relation represents a form of *probabilistic* dependency between the variables and one must add an uncertainty term, ϵ , to the relation. Some problems cannot be represented in terms of an input and an output and the modeling task becomes the discovery of mutual dependency relations between the variables. The goal becomes to map the structure of the interactions between the system's variables and uncover their similarities, their hierarchies, and their causal relations. In other contexts, one might aim instead to discover emerging behaviors from simple rules, and often for this task simulations are adopted. What I have just described are different forms of modeling, and they all serve the same purpose of helping us to navigate reality by describing what is happening and making a prediction on what will happen. In other terms, models help us to interpret existing observations and make predictions about future ones.

Models are tools to transform data into useful information which can be used for decisions and actions. Data, and observations, can be images, movies, journal articles, chats, financial prices or electric signals, or – indeed – anything that is produced by some process and carries some information. Data can always be represented as points and shapes in a space. The space might not be ordinary; it is often high-dimensional and sometimes non-euclidean. In some cases, such as deep learning, one might use data representations across several spaces. At this level of

¹ Here I use the notation \mathbf{x} and \mathbf{y} to generically indicate two sets of input and output variables and $g(\cdot)$ a function mapping between the two sets.

² Where V is the random variable indicating the velocity of a particle and v is a value.

abstraction, models are maps between points in these spaces. These maps must both guide us as precisely as possible to the positions of the existing observations as well as to help us to find the most likely positions of future observations. The model provides a *description* of the system by locating the observations and their interrelation in these spaces; the model can also provide *prediction* by inferring the presence and location of unobserved points.

Such maps between different data representation spaces are functions that models learn to approximate. Scientists have developed efficient methods to approximate them with so-called universal approximators. There are several universal approximators that can be used for modeling, from polynomials to trees [Tikk et al., 2003]. Certainly, among the most versatile and presently popular, are deep neural networks. Whether deep learning architectures are the best-suited tools for modeling, has still to be proved. However, they are certainly surprisingly good and efficient for many modeling tasks.

There are other kinds of modeling that might not be directly associated with function approximations. In some contexts, a meaningful modeling approach consists in starting from the elementary components of the system and modeling their behaviors and interactions from first principles. This microscopic modeling can produce precise explanations of the system behavior and can help to understand the origin of macroscopic phenomena from fundamental microscopic laws. For instance, this approach is sometimes used to derive the properties of materials from the constituting atoms in so-called *ab initio* simulations. However, in many real and complex systems, there is nothing comparable to the atoms. Indeed, in these systems, the elementary components are complex themselves and their laws of interaction are often unknown. In complex systems, microscopic modeling approaches often end up either being unrealistic oversimplifications or being too complex to be of use to explain the underlying system.

1.3 Models for real and complex systems

Scientists are challenged to handle and solve increasingly complex systems, from markets to self-driving cars. Although data-driven modeling is aiming to ultimately automatize the process of learning new models, human intuition and creativity is still central to this domain. This is why, to successfully build data-driven models, we must learn both the science and the art of modeling, the mathematical rigor and the intuition.

Complex systems are not abstract objects. They are very real, they are everywhere. Humans are complex systems and so are human societies. Animals – even the simplest – are complex systems and many human-created artificial systems – such as financial markets – are complex themselves. The defining characteristic of complexity is that simplification is not possible without losing crucial properties of the system. In complex systems, important properties *emerge* from the combination of many different elements. Although their elements might be known, the emergent property is the result of their combination and prediction is hard to achieve from the analysis of the constituting elements in isolation (see Parisi

[2002], Boccara [2010]). An example is a living organism, even a simple one, which is made of parts with properties and functions that might be well-known and understood. However, even the deepest knowledge of all its parts, will not reveal the most important property of the organism: the fact that it is alive. Being alive is the emergent property of the whole system which is of greater importance than the sum of the parts. Understanding emergent properties is very important in complex systems modeling, it is –literally– a question of life and death, and it is what makes such modeling very challenging.

Henry Louis Mencken – known as the Sage of Baltimore – once noted that '*For every complex problem there is an answer that is clear, simple, and wrong*'. Nonetheless, despite complex systems being challenging, one can devise effective models to describe and predict, at some level of accuracy, the behavior of complex systems. We do it all the time. Indeed, we are complex creatures who live in a complex world. The modeling of complex systems has the same nature and scope as the modeling of any other system. In complex systems, models are used to describe the system and to predict its behavior. However, the task can be more arduous. For instance, often in complex systems not only is the system non-deterministic but also the internal rules change and adapt. Nonetheless, the modeling of these systems is still based on the scientific method's circular approach (see Section 19.1), and the general principles remain the same and models can be formulated and tested using the general tools that are presented in this book.

1.4 Models as multivariate probabilities

Models are used to interpret and understand reality. Some models help us to distinguish between different scenarios such as distinguishing between friends and foes or isolating food from poison. These kinds of models can be defined as recognition and ‘classification’ models. In this setting, one has an input dataset and an output information. For instance, one could have a picture of an animal as input and information about the kind of animal provided by the model as output.

Similarly, one could have a set of observations of a physical system, for instance, the positions of the planets in the heavens over a period of time, and the model might output the mathematical law for the motion of the planets. These problems are often referred to as ‘regression’ and they concern the discovery of relations between two sets of variables one called ‘dependent’ (the output) and the other called ‘independent’ (the input). In this framework modeling consists in solving a so-called ‘inverse problem’: given the observations find the law that generates them.

Models are also used to infer the internal relations between a set of variables. In this case, one has no inputs or outputs and all variables are interdependent. The problem consists in uncovering the structures of similarities, hierarchies, dependencies, and causalities that are in the data structure.

Models must be capable to generalize and predict outcomes that have not

been observed yet. We use them constantly in our everyday life – to survive – for instance, to avoid being crushed by a bus when we cross the road. The accuracy of the model prediction in various circumstances is a measure of the goodness of the model.

Classification, regression, data-structure investigation, and prediction tasks are not necessarily distinct and can be seen as different ways of addressing a problem and interpreting the model. Indeed, the distinction between input and output variables is just a useful convention and the laws that map inputs into outputs coincide with the dependency structure of the dataset. Further, prediction consists in inferring the dependency between variables at different times or across different settings.

In the real and artificial systems that are of interest, the general problem concerns a set of several variables and their relations. In all these cases, the knowledge of the multivariate probability of all the system's variables provides the full information about the system and it is, therefore, the instrument to model it. I indeed argue and demonstrate in this book that the vast majority of what we call modeling can be formulated in terms of multivariate probabilities.

© Tommaso Neri
not for distribution
July 25, 2023

2

Foundamentals of probability

2.1 Probability

Probability is a quantification of the extent to which something is likely to happen. Data-driven probabilistic modeling is about estimating probability from observations. The purpose of this Chapter is to provide a common referential, self-containing background on fundamental concepts and definitions concerning probabilities. In this chapter, I fast-forward through notations, basic definitions, and perspectives that are fundamental to this domain and essential for the rest of the book. Knowledgeable readers could skip this chapter and use it only as a reference for later chapters. Readers interested in further insights on this part can refer to textbooks on probability such as, for instance, Feller [1957] and Ross [2014].

In this book, I consider only real-valued random variables, both continuous or discrete and they are denoted with X . The random variable, X , is in general the outcome of some process and phenomena that results in aleatoric output values. Following the literature, I indicate with upper case X the random variable and with lower case x its value for a given observation. These values can be considered as drawn from a given population, which is the collection of all possible observations and is associated with a probability distribution. Such a probability distribution of the population makes the draw of some observations more likely, others less likely, and some impossible.

Probabilities are real numbers. A value of the probability equal to zero means that the event has no chance to happen. Instead, a value of the probability equal to one represents certainty that the event happens. Probabilities cannot be negative.

Definition 2.1 (Probability). Probability can be formalized in terms of **probability space**, a triplet (Ω, \mathcal{F}, P) where Ω is the **sample space** which is the set of all possible outcomes that are the events in the **event space** \mathcal{F} ; P is the **probability measure**, a function associating each event with a number between 0 and 1.

Definition 2.2 (Random variable). A **random variable** X is a map from the sample space to the real numbers.

2.1.1 Cumulative distribution function

Definition 2.3 (Cumulative distribution function). The **cumulative distribution function** (CDF) is the probability that the random variable $X \in \mathbb{R}$ will take a value smaller than or equal to a value x . I denote it as

$$F(x) = P(X \leq x). \quad (2.1)$$

The cumulative distribution function, $F(x)$, is non-decreasing and it starts from a value equal to zero at the extreme, left, of the support where no smaller observations are possible and it ends to be equal to one at the other extreme, right, of the support where no larger observations are possible. In general, $F(-\infty) = 0$ and $F(+\infty) = 1$.

2.1.2 Complementary cumulative distribution function

Definition 2.4 (Complementary cumulative distribution function). The **complementary cumulative distribution function** $P(X > x)$ is the probability that the random variable $X \in \mathbb{R}$ will take a value larger than x . One has

$$P(X > x) = 1 - F(x). \quad (2.2)$$

Example 2.1. For instance, the likelihood that in a classroom someone picked at random is shorter than 160 cm can be expressed in terms of the cumulative distribution function $F(x) = P(X \leq x)$ with $X = \text{Height}$ and $x = 160$ cm. It is clear that in this example $F(0) = 0$ (nobody can be shorter than 0 cm) and $F(500) = 1$ (no human has ever been reported to have been taller than five meters). Conversely, the probability that someone picked at random is taller than 160 cm is the complementary cumulative distribution function $P(X > x) = 1 - F(x)$.

2.1.3 Probability mass function

If the random variable is discrete as, for example, the number on a face of a dice, then a probability can be associated with each discrete value $p(x) = P(X = x)$. This is called the probability mass function or discrete density function.

Definition 2.5 (Probability mass function). The **probability mass func-**

tion $p(x) = P(X = x)$ is the probability that the discrete random variable $X \in \mathbb{R}$ has a value equal to x . One has $p(x) \geq 0$ and $\sum_{x_k \in \Omega_X} p(x_k) = 1$.

For discrete variables, the cumulative distribution function is given by $F(x) = \sum_{x_k \leq x} p(x_k)$.

Example 2.2. The probability to observe a number, n , between one to six on the face of a cubic dice is $p(X = n) = 1/6$. The probability that by throwing a dice one obtain a number smaller or equal to 4 is instead $P(X \leq 4) = F(4) = 4/6$ and conversely the probability to obtain a number larger than 4 is $P(X > 4) = 1 - F(4) = 2/6$.

2.1.4 Probability density function

When the variable is continuous, the probability must be associated with the likelihood to observe the variable in an interval of values $P(x_1 < X \leq x_2)$. It should be clear from the previous definitions that $P(x_1 < X \leq x_2) = P(x_2 \leq X) - P(x_1 \leq X)$.¹ It is of practical and mathematical convenience to define a function, $f(x)$, called probability density function (PDF), or simply density, such that its integral between two values is equal to the probability.

Definition 2.6 (Probability density function). The **probability density function (PDF)** $f(x)$ is a non-negative Lebesgue-integrable function that satisfies the following conditions:

$$P(x_1 < X \leq x_2) = \int_{x_1}^{x_2} f(x)dx \quad (2.3)$$

with

$$f(x) \geq 0 \quad (2.4)$$

and

$$\int_{-\infty}^{+\infty} f(x)dx = 1. \quad (2.5)$$

The cumulative distribution function is the integral of the probability density function between $-\infty$ and x

$$F(x) = \int_{-\infty}^x f(z)dz . \quad (2.6)$$

¹ Notice that for continuous variables $P(X \leq x) = P(X < x)$ because the likelihood to observe the value of the variable exactly equal to a certain quantity is, in general, infinitesimal.

Equivalently, under some conditions,

$$f(x) = \frac{d}{dx} F(x) . \quad (2.7)$$

Remark 2.1. The probability density function cannot have values smaller than zero but it can have values larger than one under the condition that the integral over the support must be equal to one.

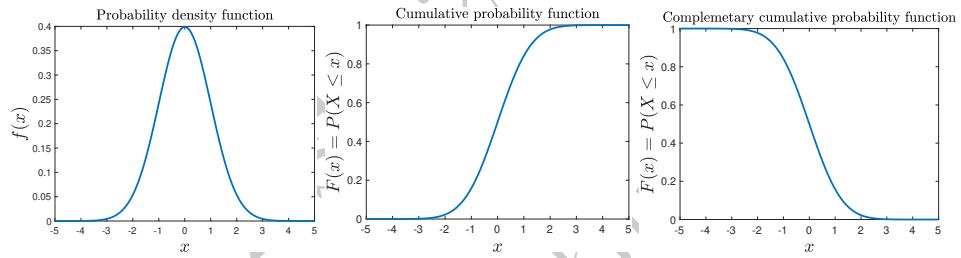


Figure 2.1 (left) An example of a probability density function (a normal distribution), (center) its corresponding cumulative probability function, and (right) its complementary cumulative probability function.

It is evident that a direct consequence of the fact that the PDF must be nonnegative ($f(x) \geq 0$) is that the cumulative density function must be a non-decreasing function with $\frac{d}{dx} F(x) \geq 0$.

Example 2.3 (Probability density function). In Fig.2.1 I report plots for the probability density function, the cumulative distribution function, and the complementary cumulative distribution function for the so-called normal distribution (see Definition 5.1). One can see that the probability density function is in this case a symmetric bell-shaped function; the cumulative distribution is an increasing function that starts from zero and ends to one; while the complementary distribution is a decreasing function starting at one and ending at zero.

2.2 Quantiles

Definition 2.7 (Quantile). The quantile, $Q(\gamma)$, is the value of the variable X at which

$$P(X \leq Q(\gamma)) \leq \gamma \quad (2.8)$$

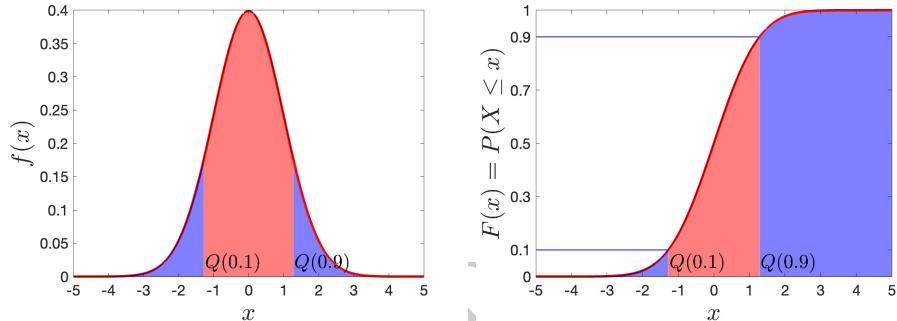


Figure 2.2 The γ -quantile is the value of x at which the cumulative probability function, $F(x)$, is equal to γ (assuming $F(x)$ continuous, differentiable and invertible). The figure reports two examples of 10% and 90% quantiles.

and

$$P(X > Q(\gamma)) \geq 1 - \gamma. \quad (2.9)$$

This is the general expression that simplifies when $F(x)$ is continuous, differentiable, and invertible^a in $x = Q(\gamma)$, because then the quantile is the value at which the cumulative density function equals γ :

$$F(Q(\gamma)) = \gamma. \quad (2.10)$$

^a Unless explicitly mentioned I will only consider continuous, differentiable and invertible $F(x)$.

In Figure 2.2 I provide a graphical representation of the 10% and 90% quantiles. A relevant quantile is the **median** which is the value with half of the density on one side and half on the other side: $F(\text{median}) = 0.5$. If the distribution is symmetric the median coincides with the mean, but in general, they differ.

Remark 2.2. For random variables with finite variances, the difference between the mean μ and the median cannot be larger than one standard deviation (see Eq.2.26).

Example 2.4 (Value at risk). A quantile that is largely used in the evaluation of the risk of financial assets is called **value at risk** (VaR). VaR_α is the α -quantile of the distribution of losses $P(\text{Losses} \leq VaR_\alpha) = \alpha$ where losses are negative returns of an investment. VaR_α is the amount of money at risk of being lost with probability α . Or, in other words, an investor has

a probability α to lose up to VaR_α over a given time horizon. It is indeed literally the ‘value at risk’ given a risk probability α .

For instance, if one has \$100 invested on a certain asset, $VaR_{0.1} = \$15$ means that with 10% probability the investor by selling the asset at a future time t , could lose \$15. Notice that if one has 10% chances to lose \$15 or less, then the chances to lose more than \$15 are 90%. Therefore, the definition of VaR_α that I have just proposed could give a false sense of confidence. This is the reason why often the complementary definition is used instead: the amount one might lose, or more, with a given probability α , which is in previous notation is $VaR_{1-\alpha}$.

Independently on the specific definition, the value at risk is a quantile and to estimate such a risk one must estimate the cumulate probability density function $F(x)$ of the (negative) returns of the asset.

Definition 2.8 (Confidence intervals). Quantiles are strictly related with **confidence intervals**. Indeed, if one wants to establish the interval $[a, b]$ within which a random variable is observed 90% of the cases (i.e. with 90% confidence interval) one must compute the 5% and 95% quantiles: $a = Q(0.05)$ and $b = Q(0.95)$.

2.3 Expected values

The expected value of a continuous random variable X is

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} xf(x)dx \quad (2.11)$$

Remark 2.3. If the variable X is discrete the expected value is

$$\mathbb{E}(X) = \sum_{x_k \in \Omega_X} x_k p(x_k). \quad (2.12)$$

Notice that, the value itself may not be ‘expected’ in common sense, indeed the “expected value” itself may be unlikely or even impossible. This is for instance the case for the expected value for the draw of a cubic dice. The dice has six equivalent probable faces with numbers from 1 to 6, the probability for the draw of each number is $p(x) = 1/6$ for and $x = 1, 2, \dots$ or 6 and the expected value is $\mathbb{E}(X) = \sum_{x_k \in [1, \dots, 6]} x_k p(x_k) = (1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5$, which is not one of the possible values of X .

The notion of expected value can be extended to any function, $g(x)$ of x , as

long as the following integral is defined:

$$\mathbb{E}(g(X)) = \int_{\Omega_X} g(x)f(x)dx , \quad (2.13)$$

where the case $g(X) = X$ is the previous definition for the mean.²

In general, expected values (including the mean) might not be defined in the sense that the integral 2.13 might diverge to plus or minus infinite.

Some useful properties of the expected value (straightforward consequences of the definition) are:

If $c \in \mathbb{R}$ is a constant, then

$$\mathbb{E}(c) = c , \quad (2.14)$$

and

$$\mathbb{E}(cX) = c\mathbb{E}(X) . \quad (2.15)$$

The expected value of the sum of a random variable and a constant is

$$\mathbb{E}(X + c) = \mathbb{E}(X) + c . \quad (2.16)$$

The expected value of the sum of two independent (see Section 8.1) random variables is the sum of the expected values

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y) . \quad (2.17)$$

If two random variables are one smaller than the other $x \leq y$ for all couples of $(x, y) \in \Omega_{X,Y}$, then their expected values follow the same inequality

$$\mathbb{E}(X) \leq \mathbb{E}(Y) . \quad (2.18)$$

In general, the expected value of a multiplication of two random variables is not equal to the multiplication of their expected values. Further, the expected value of a function $g(x)$ is not equal to the function of the expected value. In the case when $g(x)$ is a convex function³ then the following inequality holds

$$\mathbb{E}(g(X)) \leq g(\mathbb{E}(X)) . \quad (2.19)$$

which is known as Jensen's inequality.

Definition 2.9 (Bias). The difference between the expected value of a given quantity and its true value is called **bias**.

² Notice that in Eq.2.11 I integrate over the $(-\infty, +\infty)$ domain while in Eq.2.13 the integration is over the support Ω_X . Indeed, from the definition, $f(x) = 0$ outside the support Ω_X and therefore to the mean in Eq.2.11 contributes only to the region Ω_X . On the other hand, $g(x)$ might not be defined outside Ω_X , and therefore the restriction over the integral must be defined explicitly.

³ A function is convex if for any two values x_1 and x_2 and for any $c \in [0, 1]$ we have $g(cx_1 + (1 - c)x_2) \leq cg(x_1) + (1 - c)g(x_2)$.

Remark 2.4. The cumulative distribution function can be written in terms of the expected value as

$$F(x) = \mathbb{E}(\mathbf{1}_{\{X \leq x\}}) \quad (2.20)$$

where $\mathbf{1}_{\{X \leq x\}}$ is the indicator function which is equal to 1 when $X \leq x$, and zero otherwise.

2.4 Moments of the distribution

The moments of a random variable are quantities that provide important information about the underlying probability distribution.

Definition 2.10 (Moments of a distribution). The **moments** of a distribution are the expected values of integer powers of the random variable.

$$m_k = \mathbb{E}(X^k) . \quad (2.21)$$

with $k \in \mathbb{N}$.

^a Fractional moments can be defined as well and I shall introduce them in Chapter 9. They are $\mathbb{E}(|X|^k)$ with $k \in \mathbb{R}$.

The case $k = 1$ is the **mean** and it is usually denoted with the symbol μ

$$\mu = \mathbb{E}(X) , \quad (2.22)$$

while the k^{th} -moment is instead indicated as

$$m_k = \mathbb{E}(X^k) . \quad (2.23)$$

Deviation from the mean can be quantified from the expected values of the k^{th} -power of the difference between the random variable and the mean. These are called central moments.

Definition 2.11 (Central moments). The **central moments** are the expected value of the variable minus its mean raised to the k^{th} -power.

$$\mu_k = \mathbb{E}((X - \mu)^k) \quad (2.24)$$

with $k \in \mathbb{N}$.

The second ($k = 2$) central moment is called **variance**

$$\sigma^2 = \mathbb{E}((X - \mu)^2) = \mathbb{E}(X) - \mu^2 . \quad (2.25)$$

In this book I denote it either as σ^2 or $Var(X)$ depending on the context. The square root of the variance is called **standard deviation**

$$\sigma = \sqrt{\mathbb{E}((X - \mu)^2)} . \quad (2.26)$$

It is often useful to refer to standardized moments that are k^{th} central moments divided by the standard deviation to the power k . These are measures of the relative deviation of the random variable with respect to the mean in terms of the standard deviations. Two commonly used, and important, standardized moments are:

- Skewness

$$\gamma_3 = \frac{\mu_3}{\sigma^3} \quad (2.27)$$

- Kurtosis

$$\gamma_4 = \frac{\mu_4}{\sigma^4} \quad (2.28)$$

The skewness is associated with the asymmetry of the distribution, a positive skewness indicates that values above the mean are on average more likely than below. Symmetric distributions have zero skewness and negative skewness indicates larger density for values below the mean. The kurtosis is proportional to the fourth central moment and gives large weights to large deviations from the mean. A normally distributed random variable has $\gamma_4 = 3$ therefore one can define the

- Excess kurtosis

$$\gamma_4 - 3 \quad (2.29)$$

which is often used as a measure of deviation from normality. I shall come back to this in later chapters.

2.5 Univariate and multivariate probabilities

Definition 2.12 (Vector of random variables). In this book I denote **vectors of random variables** with the boldfaces symbols: $\mathbf{X} \in \mathbb{R}^{p \times 1}$. Explicitly, they are

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \quad (2.30)$$

or

$$\mathbf{X} = (X_1, \dots, X_p)^\top \quad (2.31)$$

where the components X_i are, unidimensional, scalar random variables.

So far I have only discussed probabilities for one single random variable X . They are called univariate probabilities. However, all the concepts introduced so far are valid for problems in any dimension involving more than one random variable. For these multivariate cases, the probability remains a uni-dimensional

and scalar quantity with the same fundamental properties, as discussed before (it is non-negative and its integral over the whole support is equal to one).

In terms of notation, I shall indicate a set of multivariate random variables with the bold symbol \mathbf{X} representing the vector $\mathbf{X} = (X_1, \dots, X_p)^\top$ and denote multivariate the probability density function as $f(\mathbf{X})$ which is also called joint probability density function.

Essentially nothing fundamental changes when one passes from one variable to several variables. However, the increase in the dimensionality of the problem does pose challenges both practically and conceptually.

Let me mention here for instance the change of perspective for the cumulative distribution function. The concept and definition remain exactly the same: in each dimension, it is the probability to observe values of the variables smaller than or equal to a given value. However, differently from the univariate case, in higher dimensions the space is no longer subdivided into two regions, but instead, there are 2^p combinations to consider where some variables are smaller and others are larger than any given value. It is clear that the exploration and use of this large space of configurations become extremely hard as soon as p is larger than one. This is an instance of what we will later refer to as the ‘curse of dimensionality’.

3

Fundamentals of machine learning

In my view, the purpose of models is to interpret and understand reality: transform the data received from the environment into information that is useful to make decisions and take action. The ambition of machine learning and artificial intelligence is to automate the process of modeling (see also Chapter 19). There are, of course, several different ways to model reality and they are not all applicable to every circumstance. Sometimes the task is to identify a map, $\hat{y} = g(\hat{x})$, between an input dataset \hat{x} and an output dataset \hat{y} . The machine learning community commonly calls this task '**supervised learning**'. When instead is the structure of relations among the variables that must be discovered, then the problem is referred to as '**unsupervised learning**'. This distinction makes good intuitive sense in some contexts, for instance, if one has a dataset of images of dogs and cats (input) and the model must learn how to recognize them (output). The learning from the data is supervised when the model is told which are the dogs and which are the cats (i.e. provides the 'labels'). If images without labels (no \hat{y}) are provided, then the task is to recognize that there are two different kinds of animals by looking at common features and defining differences. In this respect, this learning process is '**unsupervised**'. Beyond, simple examples, the differences between supervised and unsupervised learning can become blurry and in this book, I will mostly avoid this distinction.

In some cases, one does not aim to learn a map or a relation between variables but the goal is instead to learn a procedure, or an algorithm to perform some actions. This is what **reinforcement learning** is addressing where an autonomous 'agent' is learning from data how to optimally perform a task. The agent (or several agents) interacts with the environment and learns from examples by trial and error. This sort of learning process by means of autonomous agents has been used for a very long time in physical sciences, engineering, and even in economics. The idea is that instead of trying to model the entire – complex – behavior of the whole system one models the behavior of the constituents of the system and their interactions. If successful, then the set of rules and behavior learned by the agent (or agents) can be used to automate tasks.

Independently from supervised, unsupervised, or reinforcement learning contexts, when a model is constructed from data, the key element is to identify a convenient quantity that measures how well the model serves a given purpose. This is what is called the **gain function** or the **reward**, that a 'good' model aims to maximize (or conversely **error function** or **loss function** which the

good model aims to minimize). A good model must have small errors and these errors must be small on the observation data as well as on all other possible input data that have not been collected or used yet. This is what is referred as the ‘generalization’ capability of the model. A good model must generalize well and it must still provide reliable outputs also in situations that were not used to train the model.

There is an immense literature on machine learning and the reader interested in further deepening into the subject has a very vast choice, let me here suggest the book by Friedman et al. [2001] and the book by Goodfellow et al. [2016] which I have found very useful and inspiring. Hereafter, I provide only a few fundamental concepts that I’ll then use and recall in the rest of the book.

3.1 Supervised learning

In a supervised learning framework one typically has an input dataset $\hat{\mathbf{x}}$ and an output dataset $\hat{\mathbf{y}}$ and modeling consists in ‘learning’ the function that maps one onto the other:

$$\hat{\mathbf{y}} = g(\hat{\mathbf{x}}). \quad (3.1)$$

Input and output data can be any kind of variable: unidimensional, multidimensional, real, complex, continuous, or discrete. In most real systems this map is not fully deterministic and there is also a noise term that must be added to obtain the following equality:

$$\hat{\mathbf{y}} = g(\hat{\mathbf{x}}) + \epsilon. \quad (3.2)$$

The construction of a model (learning) consists in discovering from observations $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ an approximate function $\hat{g}(\mathbf{X})$ that matches well the map between observations and can also capture the behavior of other, yet unseen observations. This operation requires the solution of what is called an ‘inverse problem’ that indeed consists in searching for the map that connects some observed outputs with some inputs. This operation is, in some contexts, called regression or classification (respectively for continuous or categorical variables).

I shall argue in this book that the general task of supervised learning can be described as estimating the probability for a set of random variables to have values $\mathbf{Y} = \mathbf{x}$ given that another set of random variables has values $\mathbf{X} = \mathbf{x}$.

3.2 Unsupervised learning

In unsupervised learning, there are no ‘labels’ and the goal is to identify the structure and organization of the variables. The learning process consists in estimating from data the structure of similarity and interdependency between the variables. Therefore, ultimately, the general task of both supervised and unsupervised learning concern the estimate of the multivariate probability distribution. This general task is however often practically impossible and the aim is often reduced to computing some particular properties of the multivariate probability

	Supervised learning	Unsupervised learning	Reinforcement learning
Discrete (categorical) variables	classification	clustering	train agents that can perform actions
Continuous variables	regression	dimensionality reduction learning data structures	
Multivariate probability distribution function			

Figure 3.1 Supervised, unsupervised, and reinforcement learning are three common classifications of different approaches to machine learning (see text). Both supervised and unsupervised learning are estimating some properties of the multivariate probability distribution of the system's variables.

distribution such as the conditional expected values (supervised learning) or the dependency structure (unsupervised learning).

The relation between the variables, the structure of their interdependency, and the relative hierarchies are all properties to be ‘learned’ within the unsupervised learning framework. In this respect, network approaches have been shown to be extremely powerful. In these approaches, vertices of the network are the variables and edges represent the interrelations between the variables.

3.3 Semi-supervised learning

Of course, any degree of mixture between labeled and unlabeled data can be imagined and the use of both (normally a small amount of labeled and a larger amount of unlabeled) takes the name of semi-supervised learning. This hybrid approach becomes often necessary because the labeling operation in supervised learning can be very costly (often the labeling is done by humans or through expensive experiments). Furthermore, sometimes some categories are known and others not.

3.4 Reinforcement learning

In reinforcement learning, the basic idea is to model a system by using one or more ‘agents’ that chose their individual actions in a way to maximize some reward. The system evolves in discrete time steps and the agents ‘learn’ from the environment the best actions to maximize reward. Reinforcement learning is related to the dynamic programming methodology introduced by Bellman in the 1950s

[Bellman, 1952, 1954] of which it can be seen as an approximate approach. It is also strictly related to the agent-based-modeling approaches, developed since the 1970s, where systems that present a complex and hard-to-understand behavior are modeled in terms of interactions of simpler ‘agents’ [Helbing, 2012]. Differently from supervised and unsupervised learning, in this case, the learning does not consist in estimating the multivariate probability distribution but rather it consists of establishing a set of rules, a ‘policy’ which produces a maximal reward. However, one could still interpret this approach in terms of probability, specifically both in terms of the probability of a given action and also in terms of changes in the probability of the outcomes with respect to the policy adopted.

In Figure 3.1 I propose an attempt to associate the various methodologies with their outcomes distinguishing between continuous and discrete (or categorical) data.

Remark 3.1. Besides the specific details of all these machine learning methodologies, that are outside the scope of this book, what is important to bear in mind is that these are tools to build models from data and these models are used for specific purposes.

3.5 Training, validating and testing models

A model must describe well the properties of the available observational data. Given a dataset, one wants to train a model that works best on such data minimizing errors and maximizing gains. Furthermore, a model must be able to ‘generalize’ and describe well the properties that might be present in other unseen data. These two requirements are, in general, conflicting. It turns out that models optimized to best perform on a given dataset (the train set) often tend to yield bad performances with new unseen data (the test set). When this happens it is said that the model is ‘overfitting’. This is a consequence of the fact that, by optimizing a model on a given dataset, one might obtain a model that is too specialized for that given set of observations losing however the ability to adjust to new observations. Sometimes, simpler models might fit less well a specific train set but they might work better in generalizing to the overall ensemble of possible datasets. However, there is a point of simplification below in which the model no longer performs well on any dataset. The modeler must find the right compromise between the ability of a model to perform well on existing data and also generalize well on new data. There are several ways to hedge between specialization and generalization. One way is to divide the data into parts and optimize the model on one part while using the other part to quantify how well the model can generalize and fine-tune the process to reach the optimal compromise.

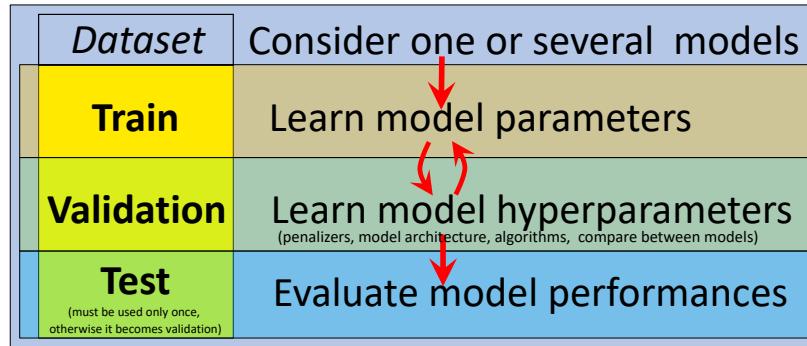


Figure 3.2 In machine learning models the dataset of observations is divided into parts. The model parameters are estimated on a train set, the hyperparameters on a validation set, and the goodness of the model is quantified on a test set. If more than a model is considered, or the model is updated, then the test becomes a further validation set used to select the model. A further test set must be then used to assess the model performances out-of-sample on a dataset that has not been used to construct the model.

3.5.1 Subdivision in train validation and test sets

An overfitting model will work very well on the dataset on which the model has been trained (train data set) but it will work badly on unseen datasets. Instead, if the model works well on a dataset that has not been used for training it is then likely that it would work well also on other datasets yet unseen. Therefore, it is common practice to divide the data in a train set (sometimes called in-sample), where the model is trained, and in a test set (sometimes called out-of-sample), where the model is tested. For some models, there is the need to adjust extra parameters, called ‘hyperparameters’, that are not optimized during training (see Definition 3.1). For this purpose, a third division of the dataset called the validation set, can be introduced. Sometimes, the validation set is needed also to choose between models and/or their architectures.

Definition 3.1. Hyperparameters are special kinds of parameters in a model that are not optimized during training. They must be instead specified a priori.

An example can be the number of clusters in a clustering method.

Summarizing, the data are divided into three, non-overlapping parts.

1. The **train set**, on which the model parameters are optimized (trained) to minimize a loss function or to maximize a gain function.
2. A **validation set**, where other parameters, called hyper-parameters, of the model or of the loss function are optimized. These are parameters or procedures or even alternative models that must be specified before optimization step 1.

3. The goodness of the model is finally assessed by looking at the performances on a third dataset, the **test set** that has not been used for any training or validation purposes.

This procedure is schematically illustrated in Figure 3.2.

Remark 3.2. It is common practice for modelers to use, try, and test different kinds of models modifying procedures and architectures. However, by comparing model performances in the test set one makes the test set become a validation set and results cannot be assumed to be generalizable to other datasets. Another set that has not been seen before by all the models must be used to finally evaluate the best-performing model. Further, for proper statistical validation, more than one test should be performed. This can be done by splitting the test set into subsets.

It must be noticed that the subdivision of the available data into several sets reduces the amount of data for training purposes and therefore can sometimes be not affordable in practice when data are scarce. In this case, it might not be possible to validate a priori assumptions and it might not be possible to test the ability of the model to generalize. When this is the case, the model construction becomes based exclusively on its ability to fit the training dataset perhaps under some restrictive conditions, such as regularizations (see Section 3.5.3), to avoid overfitting.

Cross-validation

The partition of a dataset into the train, validation, and test parts can be repeated over different parts of the dataset. This would produce slightly different results for the estimation of the model parameters and hyperparameters and one could use the average or the most common category. Such a repetition on different partitions and averaging often results in better estimations reducing variability. However, it is not always possible to perform such multi-round cross-validation. In particular, this cannot be easily done when the order of the variables in the dataset is relevant, for instance, this is the case with time series, where past and future data cannot be mixed for forecasting models. Also one must be careful about the way such a cross-validation is defined because, if it includes the test set, then there is a potentially dangerous mix between test data and training data. Overall, it is important to bear in mind that the final assessment of the goodness of a model is reliable only if performed on datasets not used for training or selection purposes.

3.5.2 Bias-variance tradeoff

A good model must certainly fit well the data in the train set. However, one must be aware that, for a different train set, one will learn a different approximate

model. Therefore the ‘goodness’ of a model cannot be exclusively judged from its performance on a train set. Instead, from a general perspective, one aims to judge the quality of a model from its performance across all possible train sets. In a supervised learning regression problem one aims to estimate the relation $Y = g(X) + \epsilon$ (with zero mean and finite variance σ^2) by learning, from a train dataset \mathbf{x}^{train} , a model function $\tilde{g}(X|\mathbf{x}^{train})$. One can quantify the goodness of the model across all possible train datasets from the expected value, computed over all train sets, of the square of the difference between the target variable Y and the model estimate $\tilde{g}(X|\mathbf{x}^{train})$. Such a qualification of the goodness of the model can be decomposed in three terms: a Bias, a Variance, and an irreducible square error (σ^2).

$$\mathbb{E}_{train}((Y - \tilde{g}(X|\mathbf{x}^{train}))^2) = \text{Bias}^2 + \text{Variance} + \sigma^2. \quad (3.3)$$

Where, the term ‘Bias’ is the difference between the expected value of the model and the true function, $g(x)$.

$$\text{Bias} = \mathbb{E}_{train}(g(X) - \tilde{g}(X|\mathbf{x}^{train})). \quad (3.4)$$

Note that $\mathbb{E}_{train}(Y) = \mathbb{E}_{train}(g(X))$ and therefore

$$\text{Bias} = \mathbb{E}_{train}(Y - \tilde{g}(X|\mathbf{x}^{train})).$$

The term ‘Variance’ is instead the variance of the model over the ensemble of train sets

$$\text{Variance} = \mathbb{E}_{train}(\tilde{g}(X|\mathbf{x}^{train})^2) - \mathbb{E}_{train}(\tilde{g}(X|\mathbf{x}^{train}))^2. \quad (3.5)$$

Both the squared Bias and the Variance are nonnegative and they can be minimized to zero if the true model is discovered and $\tilde{g}(X|\mathbf{x}^{train}) = g(X)$. In this case, one is left with the term σ^2 which is the so-called ‘irreducible error’ and it is the variance over the train sets of the noise term

$$\sigma^2 = \mathbb{E}_{train}(\epsilon^2). \quad (3.6)$$

This term is a constant and independent from the model.

The expressions for the three terms, Bias, Variance, and σ^2 in Eq.3.3 can be derived explicitly. Let me first substitute in Eq.3.3 $Y = g(X) + \epsilon$ and

then add and subtract $\mathbb{E}_{train}(\tilde{g}(X|\mathbf{x}^{train}))$, which is:

$$\begin{aligned}
 & \mathbb{E}_{train}((Y - \tilde{g}(X|\mathbf{x}^{train}))^2) = \\
 &= \mathbb{E}_{train}((g(X) + \epsilon - \tilde{g}(X|\mathbf{x}^{train}) - \mathbb{E}_{train}(\tilde{g}(X|\mathbf{x}^{train}))) + \\
 &+ \mathbb{E}_{train}(\tilde{g}(X|\mathbf{x}^{train})))^2) = \\
 &= \mathbb{E}_{train}((g(X) + \epsilon - \mathbb{E}_{train}(\tilde{g}(X|\mathbf{x}^{train}))) - \tilde{g}(X|\mathbf{x}^{train}) + \\
 &+ \mathbb{E}_{train}(\tilde{g}(X|\mathbf{x}^{train})))^2) = \\
 &= \mathbb{E}_{train}((g(X) + \epsilon - \mathbb{E}_{train}(\tilde{g}(X|\mathbf{x}^{train})))^2) - \\
 &- E_{train}((\tilde{g}(X|\mathbf{x}^{train}) - \mathbb{E}_{train}(\tilde{g}(X|\mathbf{x}^{train})))^2) =
 \end{aligned} \tag{3.7}$$

By noticing that the true model function $g(X)$ is independent of the ensemble of train sets and

$$\mathbb{E}_{train}(g(X)) = g(X). \tag{3.8}$$

Furthermore, by also noticing that, over the ensemble of train sets the expected value of the error must be zero

$$\begin{aligned}
 \mathbb{E}_{train}(\epsilon) &= 0, \\
 \mathbb{E}_{train}(\epsilon g(X)) &= 0, \\
 \mathbb{E}_{train}(\epsilon E_{train}(\tilde{g}(X|\mathbf{x}^{train}))) &= 0.
 \end{aligned} \tag{3.9}$$

One obtains

$$\begin{aligned}
 & \mathbb{E}_{train}((Y - \tilde{g}(X|\mathbf{x}^{train}))^2) = \\
 &= \underbrace{(g(X) - \mathbb{E}_{train}(\tilde{g}(X|\mathbf{x}^{train})))^2}_{\text{Bias}^2} + \\
 &+ \underbrace{\mathbb{E}(\tilde{g}(X|\mathbf{x}^{train})^2) - \mathbb{E}_{train}(\tilde{g}(X|\mathbf{x}^{train}))^2}_{\text{Variance}} + \underbrace{\sigma^2}_{\text{Irreducible error}}.
 \end{aligned} \tag{3.10}$$

Models with low Bias across train sets are the ones that adapt by fitting the model to the train data points. These models have lower Bias however they will be more susceptible to changing across different train sets and therefore they will have larger Variance. One particular case is a model that reproduces exactly all data in the train set: $\tilde{g}(\hat{x}) = \hat{y}$, this is called an interpolating model. With such a model one has zero Bias but its Variance is equal to σ^2 . Despite this bias-variance tradeoff will suggest that such overfitting models should generalize worst, it turns out that there are cases when instead they generalize better than more parsimonious models (see Section 3.5.4).

The bias-variance dilemma is the conflict between trying to minimize the Bias by having a model that fits well the data points of the train sets and simultaneously trying to minimize the Variance with models that changes little across different train datasets.

Remark 3.3. The model that minimized $\mathbb{E}_{train}((Y - \tilde{g}(X|\mathbf{x}^{train}))^2)$ is

$$\tilde{g}(X|\mathbf{x}^{train}) = \mathbb{E}(Y|X) = g(X),$$

however to learn this model one would require to be able to compute the exact expected value across all train sets, which demands an infinite amount of data for training.

Natural sciences have mostly proceeded by learning models from $\tilde{g}(X|\mathbf{x}^{train}) \simeq \mathbb{E}(Y|X) = g(X)$. This has been an unquestionably successful strategy but it cannot be applied to all domains. Often, data are limited either because it is too costly to acquire more of them or because there is a finite history of previous events. Furthermore, the underlying systems are often noisy, sometimes with noise component ϵ with diverging variance, $\sigma^2 \rightarrow \infty$, making the convergence of $\tilde{g}(X|\mathbf{x}^{train})$ to the true function $g(X)$ uncertain.

3.5.3 Regularization

Too many, too large, adjustable parameters make models overfit the train set. It is indeed intuitive that with a large enough number of adjustable parameters one can build a model that fits to perfection any input into any target output.

A famous quote, attributed to the Hungarian mathematician John von Neumann (1903 - 1957), recites: “*With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.*”

A way to prevent overfitting is to ‘penalize’ models favoring the ones with smaller and fewer parameters. This is referred to as regularization. The penalization is usually inserted as a cost or penalty by adding a term $R(\tilde{g}(\mathbf{X}))$ to the loss function. Normally such a cost grows with the complexity of the fitting function $\tilde{g}(\mathbf{X})$. In this book, I discuss regularization via penalization and via topological constraints in Section 15.9.

3.5.4 Double descent: cases when overfitting produces better results

Although the paradigm from classical statistical learning theory states that overfitting should provide poorer generalizations and therefore poorer out-of-sample results, in recent years there has been a shred of growing evidence showing that in several cases overfitting can make the system converge faster to a better generalizable solution. This is a relatively novel finding which has become evident from the research in neural networks where models have often an enormous abundance of parameters and, nonetheless, they can train quite efficiently and they are capable to generalize well beyond the training sets. This phenomenon is general and it has been observed in several models from random forests to linear regressions

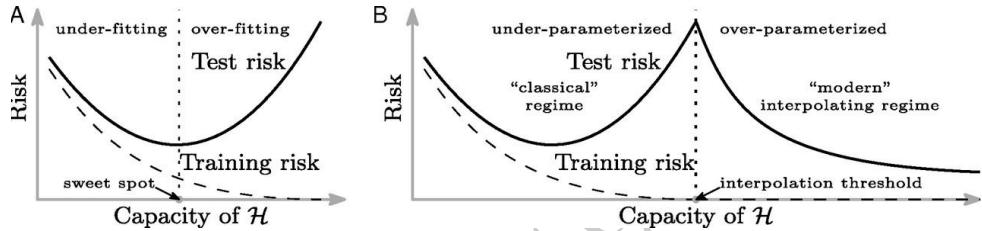


Figure 3.3 Original figure from the Belkin et al. [2019] paper (republished with permission). On the x -axis is model complexity (here called ‘capacity’), and on the y -axis is reported the model error on the test set (here called ‘risk’). On the left, panel A reproduces what is expected within the classical statistical learning theory paradigm: the error decreases with the model complexity on the train set but – after a point – it increases on the test set. On the right, panel B reports instead what is observed in some models where, after an initial behavior in line with classical statistical learning theory, then more complex models start again to produce good results in testing even when the complexity is very large. Intriguingly, the point where the model starts improving again in testing coincides with the ‘interpolation threshold’ which is where the models fit the training set with zero error.

Bartlett et al. [2020]. Indeed, despite the mean-variance statistical learning theoretical foundations (which is a mathematical truth), practitioners started observing that zero-loss training which interpolates the in-sample solutions, returning zero error in training, can give the best test results even when noise and corruption are present in the test data. Furthermore, in some cases, overfitting models can even be trained faster than less complex ones. This phenomenon has been called ‘double descent’ by Belkin et al. [2019]. The term ‘double’ derives from the fact that initially, the system behaves as expected from the mean-variance trade-off and, after an early improvement of the model performances in testing from increasing complexity (the first decent) the system then starts performing worst in testing when further complexity is introduced. However, if one continues to increase the number of parameters, performances in testing start improving again (the second descent). The original pictorial representation of this phenomenon is reproduced in Fig.3.3. There is an increasing amount of research on this phenomenon which has been observed also in some simple linear regression models Hastie et al. [2022] and derived analytically in some cases d’Ascoli et al. [2020]. However, a full understanding is still lacking and some of the results appear to be misleading consequences of specifically designed settings. This is, for instance, the case for the MNIST experiment in Belkin et al. [2019], where the double-descent is a result of a biased experiment setup Chaudhary et al. [2023]. A way to describe overfitting, interpolating, models is considering that instead of having the data described by a model plus an error term, one could see the overfitting model itself containing the error in the parameter sets. One hypothesis is that the excessive abundance of features and parameters can actually help the system to explore the solution space providing an effective regularization that comes from

the overabundance of equivalent solutions. This might help the system to descend in good solutions avoiding getting stuck in local minima or saddles, creating a sort of noisy consensus dynamics.

4

Fundamentals of networks

Networks are extremely powerful tools to represent complex systems. Usually, in a system comprising many interacting variables, one can associate to each variable a vertex and to each related variable an edge in a networked structure.

Networks are advanced and sophisticated mathematical tools. A nice aspect of the mathematics and the science of networks is that most of the terminology coincides with the one used in everyday language. Indeed, overall, the basic definitions and properties of networks mostly follow common language and common sense. Furthermore, networks can be easily visualized into pictures and our brain is particularly well trained at identifying complex network properties. This makes the fundamentals of networks relatively intuitive. However, some care must be paid because sometimes common language, visualization and intuition do not align fully with the mathematical meaning.

4.1 Networks and graphs

A network is made of vertices and edges connecting them. Often in mathematics, the term graph is used instead of the term network. I will use the two terms as synonyms preferring the term graph for the more formal treatment. Edges in a graph can be undirected or they can have a direction from one vertex to the other.

Definition 4.1 (Graph). An **undirected graph** (usually referred to simply as a graph) is a pair of two sets:

$$\mathcal{G} = (\mathbf{V}, \mathbf{E}) \quad (4.1)$$

with $\mathbf{V} = (v_1, \dots, v_p)$ the **vertex set** and \mathbf{E} the set of paired edges (v_i, v_j) , called **edge set**.

A **directed graph** is a graph where the pair of vertices in the edge set is ordered. Conventionally, direction of (v_i, v_j) is from v_i to v_j .

Let me, in this section, provide a minimum set of definitions and concepts that should equip the reader with a sufficient toolset for the rest of the chapter and the book. For a complete reference, I recommend the readers to start from the excellent book by Newman [2018].

Edges in graphs connect distinct couples of vertices but sometimes they can connect a vertex with itself forming a loop. Sometimes vertices can be connected through more than one edge. In some graphs, all vertices are connected to one other. Let me account for this with some further definitions.

Definition 4.2 (Basics on graphs).

- The **order** of a graph is its number of vertices $|\mathbf{V}|$.
- The **size** of the graph is number of edges $|\mathbf{E}|$.
- A **loop** is an edge connecting a vertex to itself.
- Two vertices can be connected with more than one edge in a **multigraph**.
- A **simple graph** has no loops and no multiple edges.
- When all vertices are connected with each other the graph is called **complete**.
- A **k -clique** is a complete graph made of a subset of k vertices all connected to each-other.

A network can be represented with the adjacency matrix \mathbf{A} .

4.2 Adjacency matrix, weight matrix, and degree distribution

Definition 4.3 (Adjacency matrix). The **adjacency matrix** $\mathbf{A} \in \mathbb{N}^{p \times p}$ of a graph is a square matrix with size equal to the number of vertices and elements $A_{i,j}$ equal to the number of edges between vertex i and vertex j . When $A_{i,j} = 0$ it means that there are no edges between the two vertices. In simple graphs $A_{i,j} = 1$ or 0 and $A_{i,i} = 0$. In undirected graphs, the adjacency matrix is symmetric $A_{i,j} = A_{j,i}$. In directed graphs $A_{i,j} = 1$ indicates that there is one edge with direction from i to j (and in the other direction might not be present if $A_{j,i} = 0$).

In a graph, some vertices are more connected than others and some might even be completely disconnected. The degree measures the level of connectedness of each vertex.

Definition 4.4 (Degree). The **degree** k_i of vertex ‘ i ’ in an undirected graph is the number of edges incoming or outgoing from the vertex. It is given by

$$k_i = \sum_{j=1}^p A_{i,j}. \quad (4.2)$$

When the graph is directed one must distinguish between the number of

edges incoming to the vertex, which is called **in-degree**

$$k_i^+ = \sum_{j=1}^p A_{j,i} \quad (4.3)$$

which is the sum of all edges that from neighboring vertices j are incoming into i . Conversely, the number of edges outgoing from the vertex i towards its neighbors is called **out-degree**

$$k_i^- = \sum_{j=1}^p A_{i,j} \quad (4.4)$$

Clearly for undirected graphs $k_i^+ = k_i^- = k_i$.

Vertices in complex networks have, in general, different degrees and one of the simplest and often meaningful quantities that describe the network structure is the distribution of the vertex degrees, called indeed **degree distribution**:

$$p(k) = \frac{\text{Number of vertices with degree } = k}{\text{Total number of vertices}}. \quad (4.5)$$

It turns out, that in many natural systems, such a distribution is ‘fat-tailed’ (see Section 5.5) with power-law tails for large degrees $p(\text{degree} > k) \sim k^{-\gamma}$ for $k \gg 1$ with exponents γ with typical values between 2 and 3 [Barabási, 2009]. Such power-law-tailed distributions are often referred to in the literature as ‘scale-free’ indicating the fact that when the statistics follow such distributions the concept of ‘typical’ scale becomes meaningless. These networks are characterized by a small number of vertices with a very large number of connections (the ‘hubs’) and a large number of vertices with a small number of connections often located at the periphery of the graph (the ‘leafs’).

Edges might carry “weights”, for instance quantifying the relevance of a link, this makes the graph ‘weighted’.

Definition 4.5 (Weighted graphs). In a **weighted graph**, edges have weights. The **weight matrix**, $\mathbf{W} \in \mathbb{R}^{p \times p}$, contains such weights as entries. Asymmetric weight matrices are associated with directed weighted graphs. The sum of the weights of the edges connected to a given vertex is called **strength**.

$$s_i = \sum_j W_{i,j}. \quad (4.6)$$

When \mathbf{W} is asymmetric one can define **in-strength** and **out-strength**

$$s_i^+ = \sum_j W_{j,i} \quad \text{and} \quad s_i^- = \sum_j W_{i,j}. \quad (4.7)$$

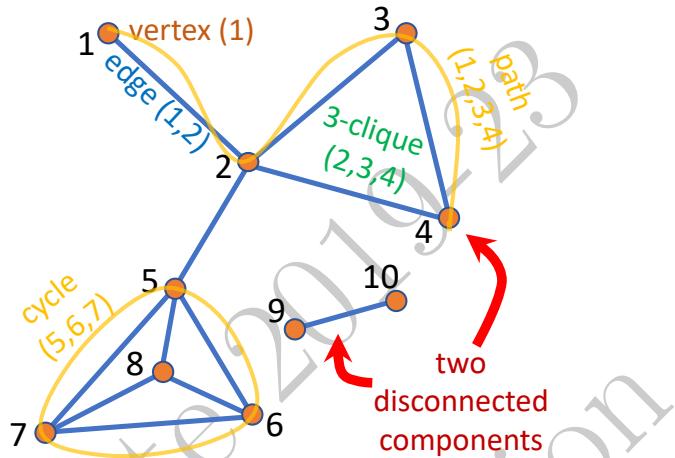


Figure 4.1 Example of a network with ten vertices and twelve edges. The network has one 3-clique $(2,3,4)$, one four-clique $(5,6,7,8)$, two disconnected components $(1,2,3,4,5,6,7,8)$, and $(9,10)$. The dominating set is $(2,5,8)$. It is a chordal graph. An example of a path and an example of a cycle is also illustrated.

Example 4.1 (Graphs). Let me here illustrate a simple example of a graph with ten vertices and twelve edges: $\mathbf{V} = (1, 2, \dots, 10)$, $\mathbf{E} = ((1, 2), (2, 3), (3, 4), (2, 4), (2, 5), (5, 6), (5, 7), (5, 8), (6, 7), (6, 8), (7, 8), (9, 10))$. The pictorial representation is provided in Fig.4.1. Its adjacency matrix is

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \quad (4.8)$$

the degrees are $\mathbf{k} = (1, 4, 2, 2, 4, 3, 3, 3, 1, 1)^\top$, which corresponds to the vector obtained from the sum over the columns of \mathbf{A} (or equivalently its rows).

4.3 Paths and walks on networks

Starting from a vertex one can ‘walk’ through the graph going from vertex to vertex through edges. There are a variety of possible walks and associated definitions.

Definition 4.6 (Walks, paths and distances).

- A **walk** is an alternating sequence of vertices and edges starting and ending with a vertex.
- A **cycle** is a closed walk where the starting and ending vertices are the same.
- A **simple path** is a walk where no vertices (and thus no edges) are repeated.
- The **length** of a path is the number of edges across the path.
- The **shortest path** is the path with minimum length connecting two vertices.
- The **distance** $d_{i,j}$ between two vertices (also called shortest path distance) is the length of the shortest path.
- The **eccentricity** of a vertex is the largest distance between the vertex and any other vertex in the graph.
- The maximum vertex eccentricity is the graph’s **diameter**.
- The minimum vertex eccentricity is the graph’s **radius**.

Example 4.2 (Shortest paths distances). By referring to the network in Fig.4.1 (see Example 4.1) the shortest paths between any couple of vertices is the following matrix

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 2 & 2 & 2 & 3 & 3 & 3 & \infty & \infty \\ 1 & 0 & 1 & 1 & 1 & 2 & 2 & 2 & \infty & \infty \\ 2 & 1 & 0 & 1 & 2 & 3 & 3 & 3 & \infty & \infty \\ 2 & 1 & 1 & 0 & 2 & 3 & 3 & 3 & \infty & \infty \\ 2 & 1 & 2 & 2 & 0 & 1 & 1 & 1 & \infty & \infty \\ 3 & 2 & 3 & 3 & 1 & 0 & 1 & 1 & \infty & \infty \\ 3 & 2 & 3 & 3 & 1 & 1 & 0 & 1 & \infty & \infty \\ 3 & 2 & 3 & 3 & 1 & 1 & 1 & 0 & \infty & \infty \\ \infty & 0 & 1 \\ \infty & 1 & 0 \end{pmatrix} \quad (4.9)$$

The two disconnected components have no paths that connect them and this is represented as infinite distance. In the largest connected component, the largest distance is 3, and this is the diameter of the component, the radius is 2. One might notice that vertex 2 is the only one that has distances

less or equal to 2 to all other vertices in the largest component. In this respect, it is the most central.

Some networks are connected and there is a walk from any vertex to any other vertex. Other networks instead have parts that cannot be reached from other parts.

Definition 4.7 (Connectedness).

- A graph is **connected** if a path exists for any couple of vertices.
- A **connected component** of an undirected graph is a sub-graph in which any two vertices are connected by one or more paths.
- The **dominating set** is a set of vertices whose neighbors, along with themselves, constitute all the vertices in the graph.

Definition 4.8 (Laplacian and degree matrices). The **discrete Laplacian matrix** is a useful matrix representation of the network. For a simple graph, it is defined as

$$\mathbf{L} = \mathbf{D} - \mathbf{A}. \quad (4.10)$$

Where \mathbf{A} is the adjacency matrix and \mathbf{D} is the **degree matrix** which is a diagonal matrix with each diagonal element equal to the degree of the corresponding vertex. The Laplacian is symmetric and semi-positive definite. It can be used to compute several useful properties of the network. The fact that this matrix is called Laplacian, as the divergence of the gradient ∇^2 operator, is not by chance. Indeed, the process of diffusion on a graph of a given quantity $\psi_i(t)$ that at time t is associated with a given vertex i and then it diffuses through the edges making a jump at random at every time step, with a dissipation constant γ , has diffusion equation

$$\frac{d}{dt} \psi(t) + \gamma \mathbf{L} \psi(t) = 0, \quad (4.11)$$

where the time derivative is symbolic for the finite difference $\frac{d}{dt} \psi(t) = \psi(t+1) - \psi(t)$ and $\psi(t)$ is the vector with components $\psi_i(t)$ with $i = 1, \dots, p$. This is indeed analogous to the diffusion equation for a gas but with the Laplace operator ∇^2 replaced by the Laplacian matrix $-\mathbf{L}$.

4.4 Centrality and peripherality

Some vertices are more central, internal, in the middle, of the graph whereas others stay at their periphery. **Centrality** is a very important property for functional properties of the network, with often the more central vertices being also

the most important for processes occurring on the network. There are several possible ways of defining importance and centrality. Possibly the simplest is the **degree centrality** that is indeed the degree with vertices with a larger degree being classified as more central. An intuitive extension of this measure is to account not only for the first neighbor but also the neighbor of the neighbor etc. but with a damping factor γ^d for vertices at the shortest path distance d . This is called **Katz Centrality**. A more sophisticated measure qualifies the number of the shortest path between any two vertices in the graph which pass through a given vertex and uses this as a measure of centrality, this is called **betweenness centrality**. A different perspective is to quantify the importance, c_i , of vertex i with respect to the importance of the first neighbor vertices, such that vertices connected with important vertices are also important: $c = \lambda \sum_{j \sim i} c_j$, which, in vectorial notation is $\mathbf{c} = \lambda \mathbf{A}\mathbf{c}$. One might recognize that this is the equation defining the eigenvectors of \mathbf{A} and indeed the solution for the centrality of each vertex is the eigenvector of the adjacency matrix and it must be the one with all coefficient positive which is the one associated with the largest eigenvalue (there is only one from the Perron–Frobenius theorem¹). This measure of centrality is called *eigenvector centrality*. There are many other measures of importance and centrality that might be relevant in different contexts, the interested reader can refer to Newman [2018].

4.5 Counting paths through the power of the adjacency matrix

The adjacency matrix has non-zero elements between two vertices connected with an edge. Its power \mathbf{A}^k has elements $(\mathbf{A}^k)_{i,j}$ counting the number of (directed or undirected) walks of length k from vertex i to vertex j .

For instance, the diagonal of the adjacency matrix is zero for simple graphs with no loops linking a vertex with itself. Instead, the diagonal of integer powers of the adjacency matrix counts the number of paths of length k which start and end with the vertex (edges can be used several times). Therefore, for undirected simple graphs, \mathbf{A}^2 counts the number of neighbors. Indeed, in two steps one can only move to a neighbor and come back through the same edge. Instead, \mathbf{A}^3 counts twice the number of cycles of length three from a vertex (the number of triangles incident on a vertex) or the number of directed cycles of length 3 in directed graphs. If two vertices are at a distance larger than k then the power k of the adjacency matrix has a zero entry

$$(\mathbf{A}^k)_{i,j} = 0 \quad \text{if } d_{i,j} > k. \quad (4.12)$$

One can for instance verify that, for the adjacency matrix in Eq. 4.8, the two disconnected components 1,...,8 and 9,10 have zero entries for all coefficients belonging to two different components at all orders. Furthermore, for the largest

¹ Notice that the sign of the eigenvector and eigenvalue can be switched arbitrarily. However, the largest eigenvalue in absolute value is the one associated with the component of the eigenvectors all of the same sign and same sign of the eigenvalue. If there are disconnected components then, some eigenvector elements might be zero.

component, one has $(\mathbf{A}^2)_{1,k} = 0$ for $k = 6, 7, 8$ that indeed are at distance 3. However, one can also notice that $(\mathbf{A}^3)_{1,5} = 0$, because although 1, 5 are at distance 2, there are no paths of length 3 that connect them.

The sum of all paths of all lengths can be counted by summing all the powers of the adjacency matrix from zero to infinite, but this, of course, will diverge to infinite unless a damping factor that penalizes proportionally to the length of the path is introduced. For instance, if there is a damping factor γ at each step, then a path of length k has weight γ^k and the sum of the weights of all paths between every couple of vertices is given by the matrix

$$\sum_{k=0}^{\infty} \gamma^k \mathbf{A}^k = (\mathbf{I} - \gamma \mathbf{A})^{-1}, \quad (4.13)$$

where \mathbf{I} is the identity matrix. To converge, the damping factor must be $\gamma < 1/\lambda_1$ with λ_1 the largest eigenvalue of the adjacency matrix. It was proposed by Newman that this matrix of damped paths of all lengths between two nodes could be used as a measure of similarity that he named Katz similarity. Indeed, the Katz centrality is the vector resulting from the sum of the columns of this similarity matrix. This measure has however the defect that tends to assign high similarity to nodes with high degrees simply because more paths come to them through their large number of neighbors. Therefore a sensible variation is to penalize the path by the degree making therefore a walker crossing an edge from vertex i to vertex j dumped by a further factor $1/k_i$, this results in

$$\sum_{k=0}^{\infty} (\gamma \mathbf{D}^{-1} \mathbf{A})^k = (\mathbf{I} - \gamma \mathbf{D}^{-1} \mathbf{A})^{-1}, \quad (4.14)$$

where \mathbf{D} is the matrix with the vertex degrees in the diagonal already defined in Eq.4.10 for the Laplacian. In this case the matrix $\mathbf{L}^{RW} = (\mathbf{I} - \gamma \mathbf{D}^{-1} \mathbf{A})$ is called random walk Laplacian because it is the transition matrix of a random walker on the graph which, at each step, chooses one neighbor at random.

Example 4.3 (The powers of the adjacency matrix). For $\gamma = 0.1$ and the adjacency matrix of the network in Fig.4.1 one obtains

$$(\mathbf{I} - \gamma \mathbf{A})^{-1} = \begin{pmatrix} 1.010 & 0.104 & 0.012 & 0.012 & 0.011 & 0.001 & 0.001 & 0.001 & 0 & 0 \\ 0.104 & 1.045 & 0.116 & 0.116 & 0.109 & 0.014 & 0.014 & 0.014 & 0 & 0 \\ 0.012 & 0.116 & 1.023 & 0.114 & 0.012 & 0.002 & 0.002 & 0.002 & 0 & 0 \\ 0.012 & 0.116 & 0.114 & 1.023 & 0.012 & 0.002 & 0.002 & 0.002 & 0 & 0 \\ 0.011 & 0.109 & 0.012 & 0.012 & 1.050 & 0.131 & 0.131 & 0.131 & 0 & 0 \\ 0.001 & 0.014 & 0.002 & 0.002 & 0.131 & 1.039 & 0.130 & 0.130 & 0 & 0 \\ 0.001 & 0.014 & 0.002 & 0.002 & 0.131 & 0.130 & 1.039 & 0.130 & 0 & 0 \\ 0.001 & 0.014 & 0.002 & 0.002 & 0.131 & 0.130 & 0.130 & 1.039 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.010 & 0.101 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.101 & 1.010 \end{pmatrix} \quad (4.15)$$

and

$$(\mathbf{I} - \gamma \mathbf{D}^{-1} \mathbf{A})^{-1} =$$

$$\begin{pmatrix} 1.003 & 0.101 & 0.003 & 0.003 & 0.003 & 0.000 & 0.000 & 0.000 & 0 & 0 \\ 0.025 & 1.006 & 0.026 & 0.026 & 0.025 & 0.001 & 0.001 & 0.001 & 0 & 0 \\ 0.001 & 0.053 & 1.004 & 0.052 & 0.001 & 0.000 & 0.000 & 0.000 & 0 & 0 \\ 0.001 & 0.053 & 0.052 & 1.004 & 0.001 & 0.000 & 0.000 & 0.000 & 0 & 0 \\ 0.001 & 0.025 & 0.001 & 0.001 & 1.003 & 0.027 & 0.027 & 0.027 & 0 & 0 \\ 0.000 & 0.001 & 0.000 & 0.000 & 0.036 & 1.003 & 0.036 & 0.036 & 0 & 0 \\ 0.000 & 0.001 & 0.000 & 0.000 & 0.036 & 0.036 & 1.003 & 0.036 & 0 & 0 \\ 0.000 & 0.001 & 0.000 & 0.000 & 0.036 & 0.036 & 0.036 & 1.003 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.010 & 0.101 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.101 & 1.010 \end{pmatrix}. \quad (4.16)$$

There are a lot of things to notice in these two matrices. First, Eq.4.15 is symmetric while Eq.4.16 is asymmetric. Both measures identify nodes (5, 6, 7, 8) as a region of larger values. Indeed, this is a 4-clique connected to the rest of the component by an edge only. The other region with larger values is the 3-clique (2, 3, 4) and also the edges (2,5) and (1,2). This indeed, intuitively reflects the vertices that are more often visited by a random walker that moves across the network jumping randomly from vertex to vertex. The asymmetry of Eq.4.16 reveals for instance that (1,2) has a larger weight than (2,1), this is because from vertex 1 the walker can only go to vertex 2 while instead from vertex 2 it has 4 different possibilities.

In terms of centrality measures, one has that the degree is $\mathbf{k} = (1, 4, 2, 2, 4, 3, 3, 3, 1, 1)^\top$ and therefore the degree centrality ranks vertex 2 and 5 as the most central. Betweenness centrality returns 14 shortest paths passing through vertex 2 which is the most central vertex under this measure; vertex 5 is passed by 12 shortest paths, while all other vertices have no shortest paths passing through them and they are therefore all peripheral. The Katz centrality vector retrieved from summing the columns of Eq.4.15 is $\mathbf{K} = (1.15, 1.53, 1.28, 1.28, \mathbf{1.59}, 1.45, 1.45, 1.45, 1.11, 1.11)^\top$, placing therefore again vertex 2 and 5 at the center but this time vertex 5 being more central than 2. Finally, the eigenvector centrality yields to $\mathbf{e} = (0.087, 0.274, 0.128, 0.128, \mathbf{0.519}, 0.453, 0.453, 0.453, 0, 0)^\top$ making, again, vertex 5 most central but also vertices 6, 7, 8 more central than vertex 2.

4.6 Propagation on networks

Consider a generic linear process that describes the evolution of a quantity x_i , associated with each vertex $i = 1, \dots, p$, over a network

$$x_{i,t} = \sum_j x_{j,t-1} M_{j,i} \quad (4.17)$$

The matrix $M_{j,i} \geq 0$ describes the transition of some proportion of the quantities $x_{j,t-1}$ that are on vertices j at time-step $t-1$ into vertex i at time-step t ; it is related with the adjacency matrix or the matrix of weights. In vectorial representation the previous equation is

$$\mathbf{x}_t = \mathbf{M}^\top \mathbf{x}_{t-1}, \quad (4.18)$$

with $\mathbf{x}_t = (\mathbf{x}_{1,t}, \dots, \mathbf{x}_{p,t})^\top$ and $(\mathbf{M})_{i,j} = M_{i,j}$. If the process starts at step 0 with \mathbf{x}_0 , then after t steps it is

$$\mathbf{x}_t = (\mathbf{M}^\top)^t \mathbf{x}_0. \quad (4.19)$$

The power of the weight matrix \mathbf{M}^\top describes the diffusion over the network. This is analogous to the path-counting results in Eqs.4.13 and 4.14, however here one can view it in terms of a diffusion process and study the distribution of \mathbf{x}_t over the network's nodes. For large t , $(\mathbf{T}^\top)^t$ becomes a matrix where all columns are proportional to the eigenvector \mathbf{v}_1 associated with the largest eigenvalue λ_1 (namely $\lambda_1^t \mathbf{v}_1 / \|\mathbf{v}_1\|_1$) and this makes Eq.4.19 returning $\mathbf{x}^* = \mathbf{v}_1 \lambda_1^t \|\mathbf{x}_0\|_1 / \|\mathbf{v}_1\|_1$ for any \mathbf{x}_0 when $t \rightarrow \infty$. Convergence is exponentially fast.

Remark 4.1. By re-writing Eq.4.22 in the Laplacian notation introduced in Eq.4.11 one can identify the Laplacian of the process Eq.4.18 as

$$\mathbf{L} = \frac{1}{\gamma} \mathbf{I} - \mathbf{M}^\top. \quad (4.20)$$

From the previous discussion, it should be clear that the process does not explode only if the dissipation term satisfies $\gamma \leq 1/\lambda_1$. There are some special cases, that I discuss in the following Example, where $\lambda_1 = 1$.

Example 4.4 (Random walk & consensus dynamics on networks). In a **random walk** process the quantity that is updated at each step is the probability to find the walker on a given vertex $x_{i,t} = p_{i,t}$ and \mathbf{M} becomes the transition matrix whose entries are the probabilities that a walker on vertex j jumps on vertex i . To preserve the total number of walkers (or analogously, keep $p_{i,t}$ a probability), \mathbf{M} must be normalized such that $\sum_i M_{j,i} = 1$ or conversely one can re-write the process as

$$p_{i,t} = \sum_j p_{j,t-1} T_{j,i} \quad (4.21)$$

with $T_{j,i} = M_{j,i} / \sum_k M_{j,k}$ the transition matrix.a In vectorial form

$$\mathbf{p}_t = \mathbf{T}^\top \mathbf{p}_{t-1} \quad (4.22)$$

The process is the same as Eq.4.18 but, due to this normalization, the largest eigenvalue is $\lambda_1 = 1$ and $\|\mathbf{p}_t\|_1 = 1$ for all t . For large t the process eventually can reach equilibrium (if the graph is connected and non-

bipartite) and $p_{i,t} = p_{i,t-1} = p_i^*$ for all i and therefore

$$\mathbf{p}^* = \mathbf{T}^\top \mathbf{p}^*, \quad (4.23)$$

with \mathbf{p}^* the column vector with components p_i^* for $i = 1, \dots, p$ and $(\mathbf{T})_{i,j} = T_{i,j}$. This is the eigenvector equation for \mathbf{T}^\top corresponding to the largest eigenvalue $\lambda_1 = 1$. Equivalently, to reach equilibrium one can think of reiterating the process in Eq.4.22 for a very large number of steps until \mathbf{p}^* is reached asymptotically. If one starts from a given \mathbf{p}_0 at step 0, then at step t one has

$$\mathbf{p}_t = (\mathbf{T}^\top)^t \mathbf{p}_0, \quad (4.24)$$

and it is expected that $\mathbf{p}_t \xrightarrow[t \rightarrow \infty]{} \mathbf{p}^*$. Indeed, for large t , $(\mathbf{T}^\top)^t$ becomes a matrix where all columns are the largest eigenvector \mathbf{p}^* and this makes Eq.4.24 returning \mathbf{p}^* for any \mathbf{p}_0 when $t \rightarrow \infty$. Convergence is exponentially fast. When the matrix \mathbf{M} is symmetric, then the limiting probability distribution p_i^* is proportional to the strength of the vertex $s_i = \sum_j M_{i,j}$. From a different perspective, one can regard $x_{i,t}$ as the ‘opinion’ of vertex i at step t and one might want the system to reach an agreed **consensus** by imitating the belief $x_{j,t}$ of neighboring vertices. In the so-called agreement algorithm the value or $x_{i,t}$ is set equal to the average value of the neighbors, which implies to set in Eq.4.17 $M_{j,i} = 1/k_i^+ A_{i,j}$ with $k_i^+ = \sum_j A_{j,i}$ the in-degree of i . Notice that, although similar to the random walk process, this process is instead driven by the transpose of the random walk transfer matrix and does not conserve $\sum_k x_{k,t}$. In this case, the system asymptotically reaches consensus (for connected and non-bipartite graphs) to a uniform vector with $x_i^* = \text{const.}$ across all vertices, which corresponds to the eigenvector of \mathbf{T} with the largest eigenvalue. Generalization to weighted means, $M_{j,i} = 1/s_i^+ W_{i,j}$, and inclusion of the vertex’s own belief in the average are straightforward.

^a This is the probability of the jump $j \rightarrow i$. The matrix \mathbf{M} can be any weight matrix.

4.7 Trees forests and higher order networks

Graphs that have no cycles are called trees or acyclic depending on whether they are undirected or directed.

Definition 4.9 (Trees and forests).

- A **tree** is an undirected graph with no cycles.
- A **spanning tree** is a connected graph with no cycles (a tree connecting all vertices).
- A **forest** is a disconnected graph made of trees.

A spanning tree with p vertices has always $p - 1$ edges independent of its structure. This is the minimum number of edges needed to connect the graph. In a tree, any two vertices are connected through only one shortest path.

Definition 4.10 (Acyclic graphs). Directed graphs with no cycles are called **acyclic**.

Cycles in networks can be of various sizes and some cycles might have shortcuts that can reduce the length of the simple path required to go back to the starting vertex. Such cycles are **reducible**. A special case is when all cycles are reduced to the smallest cycle, which is the triangle.

Definition 4.11 (Chordal graphs). A graph is called **chordal** if all cycles larger than three have a chord, namely an edge between two non-consecutive vertices, which reduces the cycle to a set of triangles.

Definition 4.12 (Simplexes). In graph theory, for undirected graphs, cliques and **simplexes** (or simplices) are strictly related, with the first being topological objects and the second being their corresponding geometrical objects. Specifically, simplexes are d -dimensional geometrical objects: segments, triangles, tetrahedra, and their generalizations in higher dimensions. The cliques associated with such geometrical objects are the network of their edges. A d -dimensional **polytope**, has $d + 1$ vertices all connected to each other. A k -clique is a complete graph with k vertices (often denoted as K_k). Therefore the network of the edges of a d -simplex is a $(d+1)$ -clique. Specifically:

dimension	simplex	polytope	clique	complete graph
0	0-simplex	vertex	1-clique	K_1
1	1-simplex	segment	2-clique	K_2
2	2-simplex	triangle	3-clique	K_3
3	3-simplex	tetrahedron	4-clique	K_4
:	:	:	:	:
d	d -simplex	d -polytope	$(d + 1)$ -clique	K_{d+1}
:	:	:	:	:

Definition 4.13 (Simplicial complexes). A **simplicial complex** is a graph made of simplexes (see Bianconi [2021]). Any undirected graph can be described as a simplicial complex. However, usually, networks are described, through the adjacency matrix, as a set of edges (1-simplexes). The description of a network in terms of its higher-order constituting simplexes introduces a higher-order element that doesn't only describe relations between couples of vertices but also describes relations between groups of fully connected vertices (the simplexes).

A special class of simplicial complexes is the **clique trees** and forests. They are indeed trees made of simplexes (see Definition 11.4). Clique trees are chordal graphs (see Definition 4.11), and vice versa, chordal graphs are clique trees. In terms of homology groups, clique forests are simple objects with only the first Betti number different from zero.^a

Simplicial complexes are used in topological data analysis and in persistent homology to describe some properties of datasets in terms of homology groups properties (see Edelsbrunner et al. [2008], Wasserman [2018]).

^a The Betti number is counting the number of connected components, see Duke [1966].

In this Chapter, I have summarized very essential concepts and definitions from graph theory. I reported only the minimal information that I shall use later on. There is an enormous amount of extra material that one should learn to acquire some knowledge in this domain. However, in this book, I want to focus on the very specific application of graph theory to dependency structure characterization and construction, for this purpose, the previous concepts should be a sufficient reference.

© Tomaso Aste 2019-23
not for distribution
July 25, 2023

Part II

Foundations of probabilistic modeling

© Tomaso Aste 2019²³
not for distribution
July 25, 2023

© Tomaso Aste 2019-23
not for distribution
July 25, 2023

5

Univariate probabilities

There are only two conditions that must be satisfied by a probability density function $f(x)$:

- 1) being a non-negative integrable function, $f(x) \geq 0$;
- 2) having the integral over the whole support equal to one, $\int_{\Omega_X} f(x)dx = 1$.

Therefore there is a very wide range of functions that can be probability density functions. In this Chapter, I mention only a few that are relevant for the study of real, complex, systems and will be used later in this book.

5.1 The normal distribution

The most common probability density function is called normal. The normal distribution is widespread in real systems and it has useful properties which make it appealing for modeling.

Definition 5.1 (Normal distribution). The **normal** probability density function with mean μ and standard deviation $\sigma > 0$ is:

$$\varphi(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \quad (5.1)$$

It has support over the whole domain $x \in (-\infty, +\infty)$. The normal distribution with zero mean, $\mu = 0$, and unitary variance, $\sigma^2 = 1$, is called the ‘standard’ normal distribution.

This probability distribution is known also as Gaussian distribution after Carl Friedrich Gauss (1777-1855) who first described it. A plot of the normal distribution probability density function with zero mean and unitary standard deviation (standard) is reported in Fig.5.1.

The normal distribution is ‘bell’-shaped, it is symmetric around the mean, and it has all moments defined with, in particular: $\mu_2 = \sigma^2$, $\mu_3 = 0$ and $\mu_4 = 3\sigma^4$. Therefore, in this case, the excess kurtosis ($\gamma_4 = \mu_4/\sigma^4 - 3$, see Section 2.4) is zero. This is indeed the reason why γ_4 is called ‘excess kurtosis’ indicating a deviation from the normal distribution case.

The normal distribution is a very important probability distribution; however,

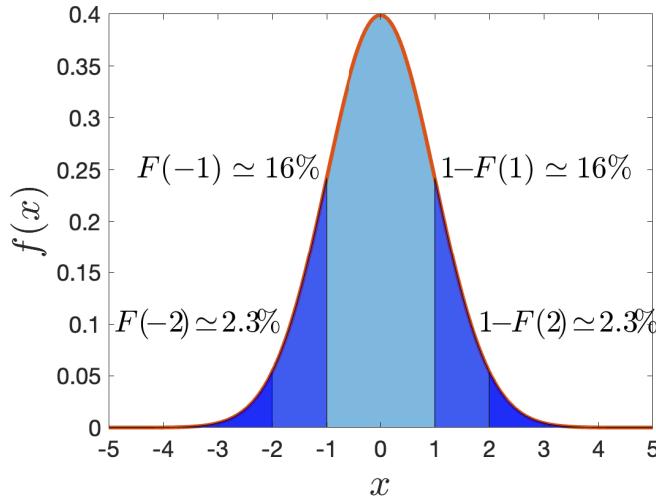


Figure 5.1 Plot of the probability density function $\varphi(x)$ of a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ (standard form). The probability to draw a value larger than one standard deviation from the mean is $P(X - \mu \geq \sigma) \sim 16\%$ and the probability to draw a value larger than two standard deviations from the mean is $P(X - \mu \geq 2\sigma) \sim 2.3\%$. Symmetrically, these are also the probabilities to draw small values at the same distance from the mean.

I will show throughout this book that random variables in real systems, from financial, socio-economics, and complex systems often are not well described by normal distributions. The kurtosis is a simple measure to assess if the probability distribution of a variable is consistent with a normal model. I will introduce in later sections other more sophisticated methods to quantify such differences and to find probability distributions that better describe the statistical properties of these real systems. Indeed, one could argue that a characterizing feature of real and complex systems, such as financial markets or the brain, is their deviation from normal statistics.

5.1.1 Tendency towards the normal distribution: the central limit theorem

Definition 5.2 (Independent and identically distributed variables). A collection of random variables X_1, \dots, X_n is **independent and identically distributed (i.i.d.)** if each random variable, X_k , is drawn from the same population of the others and they are all independent from each other. In-

dependent, means that the value of one variable is not affecting the value of the other (see also Definition 8.1).

There is a very important property that applies to any sum of i.i.d. random variables with finite mean and variance. They converge to only one type of probability density: the normal distribution.

Theorem 5.1 (Central limit theorem). *Consider n i.i.d. random variables X_1, X_2, \dots, X_n with finite mean $\mathbb{E}(X_k) = \mu$ and finite variance $\sigma^2 = \mathbb{E}((X_k - \mu)^2)$. The probability density function of their sum $Y = X_1 + X_2 + \dots + X_n$ tends to converge towards the normal probability density with mean $n\mu$ and variance $n\sigma^2$ when the number of terms, n , tends to infinity:*

$$f(y) = \varphi(y|n\mu, n\sigma^2) = \frac{1}{\sqrt{2\pi n\sigma^2}} \exp\left(-\frac{(y - n\mu)^2}{2n\sigma^2}\right). \quad (5.2)$$

Equivalently, if one considers the mean $Y = (X_1 + X_2 + \dots + X_n)/n$ instead of the sum, the limiting distribution is the Normal distribution with mean equal to μ and variance σ^2/n , which is

$$f(y) = \varphi(y|\mu, \sigma^2/n) = \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left(-\frac{(y - \mu)^2}{2\sigma^2/n}\right), \quad (5.3)$$

indicating that the distribution of the mean becomes increasingly peaked in a narrower region around the mean with the standard deviation that shrinks at a rate $1/\sqrt{n}$.

Example 5.1. Let me consider, for instance, the draw of a six-faced dice. The corresponding random variable, X_1 , returns values between 1 and 6 with uniform probability $1/6$. Therefore, $\mu = \sum_{k=1}^6 k/6 = 3.5$ and $\sigma^2 = \sum_{k=1}^6 k^2/6 - \mu^2 = 35/12 = 2.9167\dots$. By performing an experiment where the dice is drawn 1,000 times, each number on the dice's face will appear a similar number of times. The frequency of occurrence of each number will hence be similar generating a flat histogram (see Section 13.7) as to the one reported in Fig.5.2a. The normal distribution with corresponding mean and standard deviation, is represented on top of the histogram with the red line. One can notice that the two are considerably different. If one instead throws two dice, the distribution of frequency of occurrence of the sum of the number in both faces $Y = X_1 + X_2$, takes a triangular form as shown in Fig.5.2b. This distribution is not well described by a Normal. Passing to four dice, one can see from the figure that the distribution of frequency of occurrence of the sum of the four uniformly distributed random variables $Y = X_1 + X_2 + X_3 + X_4$ starts to resemble the bell-shaped normal distribution reported on top (normal with mean $4 \times 3.5 = 14$ and variance $4 \times 35/12 = 1.66\dots$). Such a resemblance becomes even more evident in the case of ten dice shown in the bottom right Fig.5.2d.

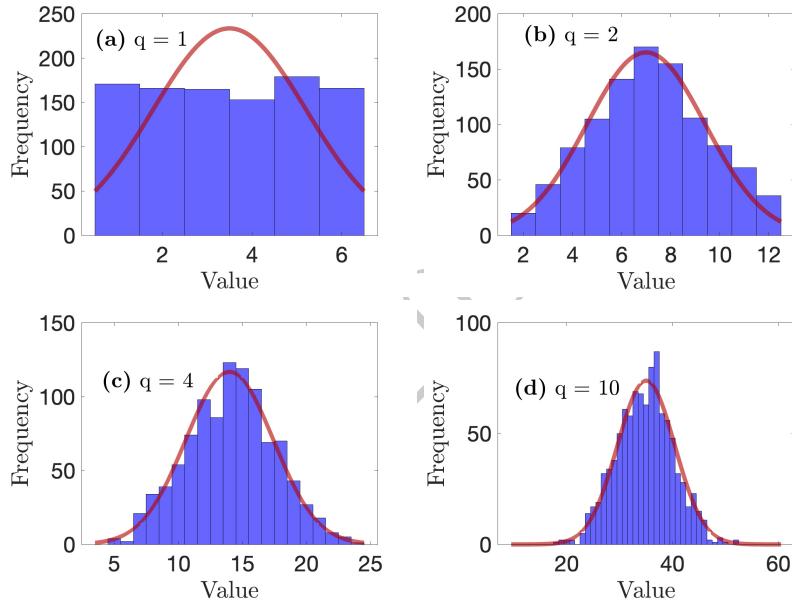


Figure 5.2 Example of convergence towards the normal distribution for the frequencies (blue bars) of the sum of $n = 1, 2, 4$ and 10 dices observed for 1000 draws. The red curve is the probability density function for the normal distribution $\varphi(y|n\mu, n\sigma^2)$ with $\mu = 3.5$ and $\sigma^2 = 35/12$ rescaled by multiplying it by n to match the frequencies.

5.1.2 Rate of convergence towards the normal distribution: Berry-Essen Theorem

Theorem 5.2 (Berry-Essen Theorem). *Given a sequence of n independent identically-distributed random variables X_1, X_2, \dots, X_n each having zero mean $\mathbb{E}(X_k) = 0$, finite positive variance $\mathbb{E}(X_k^2) = \sigma^2 > 0$, and finite third absolute moment $\mathbb{E}(X_k^3) = \gamma$, then the cumulative distribution function $F(Y)$ of the normalized sum*

$$Y = \frac{X_1 + X_2 + \dots + X_n}{\sqrt{n}\sigma} \quad (5.4)$$

differs from the cumulative distribution of a standard normal distribution function $\varphi(Y, 0, 1)$ by no more than a finite number

$$|F(Y) - \varphi(Y, 0, 1)| \leq \frac{C\gamma}{\sigma^3\sqrt{n}} . \quad (5.5)$$

Where C is a positive constant which, for sufficiently large n is between $0.4097 < C < 0.4748$. Notice that this indicates a rate of convergence towards the normal distribution in $1/\sqrt{n}$. (See Berry [1941] and Feller [1957]).

Before adventuring further in discussing other probability density functions, let me introduce some additional, useful, mathematical tools and, in particular, the characteristic function.

5.2 Characteristic function

Given a random variable X with probability density function $f(X)$, its characteristic function is defined as.

Definition 5.3 (Characteristic function).

$$\phi(\omega) = \mathbb{E}(e^{i\omega X}) . \quad (5.6)$$

which is

$$\phi(\omega) = \int_{-\infty}^{+\infty} f(x)e^{i\omega x} dx \quad (5.7)$$

This is the Fourier transform of the probability density function (with the opposite sign at the exponent).

Definition 5.4 (Fourier transform). The Fourier transform of an integrable real function $f(x)$ is

$$\mathcal{F}_f(\omega) = \int_{-\infty}^{+\infty} f(x)e^{-i\omega x} dx \quad (5.8)$$

where $i = \sqrt{-1}$ is the imaginary number. Its inverse is

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \mathcal{F}_f(\omega)e^{i\omega x} d\omega . \quad (5.9)$$

Example 5.2. For instance, the characteristic function of the normal distribution is

$$\phi_N(\omega) = \exp(i\mu\omega - \sigma^2\omega^2/2). \quad (5.10)$$

While, the characteristic function of the uniform distribution ($f(x) = 1/2$ with support $x \in [-1, +1]$) is

$$\phi(\omega) = \frac{\sin(\omega)}{\omega}. \quad (5.11)$$

The characteristic function fully defines the associated probability distribution and vice versa. However, not always both can be written in a simple closed form.

5.2.1 Moment generating function

A function, strictly related to the characteristic function is the moment-generating function that is the characteristic function with argument $-i\omega$

$$M(s) = \mathbb{E}(e^{sx}) . \quad (5.12)$$

This function is extremely useful in practice because all moments can be computed from its derivatives at $s = 0$:

$$m_k = \left. \frac{d^k}{ds^k} M(s) \right|_{s=0} . \quad (5.13)$$

The characteristic function is useful for many purposes particularly when one wants to compute the probability density function of a sum of i.i.d. variables (see Section 5.2.2).

5.2.2 Characteristic function of a sum of i.i.d. variables

Given two independent random variables X_1 and X_2 , both with probability density function $f_X(X)$ with support over the entire real axis, the probability density function of their sum $Y = X_1 + X_2$ is

$$f_Y(y) = \int_{-\infty}^{+\infty} f_X(x) f_X(y - x) dx . \quad (5.14)$$

This integral is called *convolution* and, often, it can be rather elaborate to compute. Conversely, the expression for the characteristic function of Y is much simpler. Indeed, the characteristic function of a linear combination of independent random variables

$$Y = b_1 X_1 + b_2 X_2 + \dots + b_n X_n \quad (5.15)$$

is given by the product of their characteristic functions:

$$\phi_Y(\omega) = \phi_{X_1}(b_1\omega) \cdots \phi_{X_n}(b_n\omega). \quad (5.16)$$

This is why the characteristic function is so useful in the present context: it transforms convolutions into simple products. This is known as the convolution theorem, which was first published in Fourier's book Fourier [1822] but it was already previously proven by Siméon Denis Poisson (see McGillem and Cooper [1991]).

Indeed, one can directly verify that the characteristic function of the probability density function of the random variable $Y = X_1 + X_2$ is:

$$\phi_Y(\omega) = \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} f_X(x) f_X(y - x) dx \right) e^{i\omega y} dy. \quad (5.17)$$

One can call $y - x = z$ and substitute, obtaining

$$\phi_Y(\omega) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_X(x) f_X(z) e^{i\omega(z+x)} dx dz, \quad (5.18)$$

which, indeed, is the product of the two characteristic functions or rather their square, given that in this case they are identical

$$\phi_Y(\omega) = \phi_X(\omega)^2. \quad (5.19)$$

Example 5.3 (Sum of normally distributed variables). For instance, the characteristic function of the sum of two normally distributed variables is

$$\phi_{Y=X_1+X_2}(\omega) = \phi_N(\omega)^2 = \exp(i2\mu\omega - 2\sigma^2\omega^2/2). \quad (5.20)$$

Analogously, the characteristic function of the sum of n normally distributed variables is

$$\phi_{Y=X_1+\dots+X_n}(\omega) = \phi_N(\omega)^n = \exp(in\mu\omega - n\sigma^2\omega^2/2). \quad (5.21)$$

It can be seen that these characteristic functions of the sum of normal variables have still the same original form of the normal distribution with only the parameters changed from μ and σ^2 to $n\mu$ and $n\sigma^2$. In a different notation, given $X_k \sim \mathcal{N}(\mu, \sigma^2)$ then $Y = X_1 + \dots + X_n \sim \mathcal{N}(n\mu, n\sigma^2)$. In this respect, one can say that the normal distribution is *stable* (see Definition 5.5 hereafter) because a sum of normally distributed variables is still a normally distributed variable (with scaled parameters, $\mu \rightarrow n\mu$ and $\sigma^2 \rightarrow n\sigma^2$).

5.3 Stable distributions

A stable distribution has the property that a linear combination of two (or more) independent random variables with stable distribution has also a stable distribution.

Definition 5.5 (Stable random variable). A random variable X is called **stable** if and only if for n independent copies X_i of X , there exist two constants $d_n \in \mathbb{R}$ and $\alpha \in (0, 2]$ such that

$$Y = X_1 + \dots + X_n \stackrel{d}{=} n^{1/\alpha} X + d_n, \quad (5.22)$$

where $\stackrel{d}{=}$ indicates an equality in distribution. In terms of the probability density function, this is

$$f_Y(y) = \frac{1}{n^{1/\alpha}} f_X\left(\frac{y - d_n}{n^{1/\alpha}}\right). \quad (5.23)$$

The coefficient d_n is equal to zero when the variables have zero means $\mathbb{E}(X_k) = \mathbb{E}(Y) = 0$. When $d_n = 0$ the distribution is said to be **strictly stable**.

5.3.1 Lévy-alpha stable distribution

There is an entire class of distributions that are stable. Such a class includes the normal distribution as a special case. Given the great simplification that the characteristic function provides in the computation of the distribution of a sum of variables, (see Section 5.2), it should not be surprising that this class of distributions is defined in terms of their characteristic function. What is however peculiar is that, in general, stable distributions can be only expressed in terms of their characteristic function and do not have a simple closed form for the probability density function. The stable distribution family is often referred to as the Lévy alpha-stable distribution, after Paul Lévy (1886-1971).

Definition 5.6 (Lévy alpha-stable). The **Lévy alpha-stable distribution** has characteristic function

$$\phi(\omega) = \exp \left[i\omega\mu - |c\omega|^\alpha (1 - i\beta \operatorname{sign}(\omega)\Phi) \right]; \quad (5.24)$$

where

$$\Phi = \begin{cases} \tan \frac{\pi\alpha}{2} & \text{if } \alpha \neq 1 \\ -\frac{2}{\pi} \log |c\omega| & \text{if } \alpha = 1 \end{cases}. \quad (5.25)$$

It has support over the whole domain $x, \omega \in (-\infty, \infty)$. It has four parameters: location parameter $\mu \in (-\infty, \infty)$, scale parameter $c \in (0, \infty)$; skewness parameter $\beta \in [-1, 1]$ and stability parameter $\alpha \in (0, 2]$. In this book, I will call α the ‘tail exponent’ because it is the exponent of the power-law tail of this distribution (see Section 5.5).

The α parameter of the Lévy alpha-stable distribution is of particular relevance, especially in the context of this book, because for large values of x (the tail region), the probability density function of the Lévy alpha-stable distribution behaves as a power law $f(x) \underset{|x| \in tails}{\sim} |x|^{-\alpha-1}$ for $\alpha < 1$.

The Lévy alpha-stable distribution has a defined mean equal to μ when $\alpha > 1$. For $\alpha < 2$, the variance and all higher moments are not defined. A special case is when $\alpha = 2$ because the Lévy alpha-stable distribution becomes the normal distribution and therefore the variance and all higher moments become defined. A plot of the Lévy alpha-stable probability density function with zero mean, scale $c = 1$ and $\alpha = 0.75$ is reported in Fig.5.3.

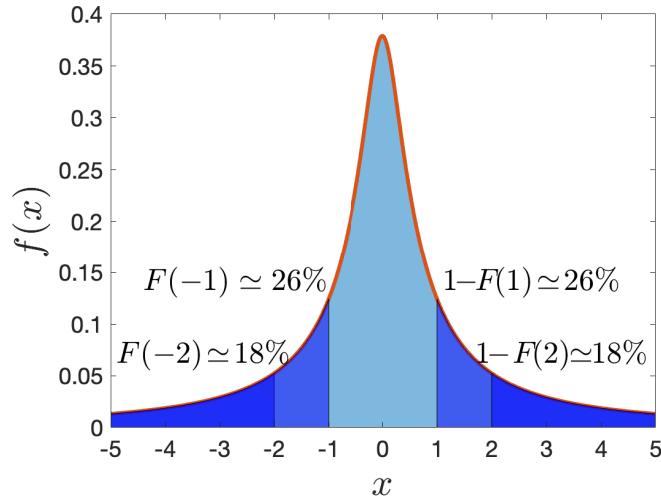


Figure 5.3 Plot of the probability density function $f(x)$ of a Lévy stable distribution with location $\mu = 0$, scale $c = 1$, tail exponent $\alpha = 0.75$, and $\beta = 0$. Notice that in this case, mean, variance, and all higher moments are undefined.

It is straightforward to see from the characteristic function of the Lévy alpha-stable distribution that it is a stable distribution. Indeed, the sum of two variables with Lévy alpha-stable characteristic functions has a characteristic function

$$\phi_{Y=X_1+X_2}(\omega) = \exp \left[i\omega 2\mu - 2|c\omega|^\alpha (1 - i\beta \operatorname{sign}(\omega)\Phi) \right] \quad (5.26)$$

which is a Lévy alpha-stable distribution with location parameter 2μ and scale parameter $2^{1/\alpha}c$. More generally, if one sums n of such variables the location parameter becomes $n\mu$ and the scale parameter becomes $n^{1/\alpha}c$.

5.4 Tendency towards Lévy-alpha stable distribution

The central limit theorem, defined in Theorem 5.1, applies to the sum of any set of i.i.d. random variables with *finite* variance. However, in many practical cases, random variables do not have finite variance and the distribution of their sum no longer converge towards a normal.

An extension of the central limit theorem states that the convergence is towards the stable distribution [Gnedenko et al., 1954].

Theorem 5.3 (Generalized central limit theorem). *Consider n i.i.d. random variables X_1, X_2, \dots, X_n with symmetric distribution around the mean $\mathbb{E}(X_k) = \mu$ and power law tails with tail exponent $\alpha \leq 2$ (e.g. $f(x) \underset{|x| \in \text{tails}}{\sim} x^{-\alpha-1}$). The*

probability density function of their sum $Y = X_1 + X_2 + \dots + X_n$ tends to converge towards a Lévy- α stable probability density with mean $n\mu$ and stability parameter α . If the variables X_k are stable distributions with scale parameter c , then the limiting distribution has scale parameter $n^{1/\alpha}c$.

Remark 5.1. I have just shown with Theorem 5.3 that the sum of i.i.d variables converges to the universal attractor class of stable distributions. But, what about the sum of *dependent*, identically distributed variables? Dependence between the variables changes everything. In Section 9.9 I will show that a class of dependent (multivariate), Student-t distributions sums into Student-t distributions (which is not a stable distribution).

5.5 The body and the tails of the distribution

The shape of a probability density function $f(x)$ can be divided into three parts:

1. a central part around the median referred to as the ‘body’ which includes most of the most likely values;
2. the ‘left tail’, where x is much smaller than the median, and includes extreme negative deviations from the median (small quantiles);
3. the ‘right tail’ where x is much larger than the median, and includes extreme positive deviations from the median (large quantiles).

The exact points where the body ends and the tails start are difficult to establish in general because they depend on the kind of distribution. Distributions defined on finite support $x \in [a, b]$ with $a, b \in \mathbb{R}$ and finite, have no tails. All other distributions where the support goes to infinity on one or both sides must satisfy the condition $\int_{-\infty}^{+\infty} f(x)dx = 1$ which implies that the density function, $f(x)$, must go to zero when $|x|$ goes towards infinite, and such a decrease must be at a rate as a power law or faster:

$$f(x) \underset{x \in \text{tails}}{\leq} \frac{1}{|x|^{\alpha+1}}, \quad (5.27)$$

with tail exponent $\alpha > 0$, where the notation $x \in \text{tails}$ indicates extreme values of x , outside the ‘body’ region. The probability distribution can be classified in relation to the tail exponent α .

- For $\alpha \leq 0$ the integral $\int_{-\infty}^{+\infty} f(x)dx$ does not converge and one can only have bounded distributions;
- For $0 < \alpha \leq u$ the distribution is said to be ‘fat tailed’ with all moments m_k with $k \geq u$ not defined;

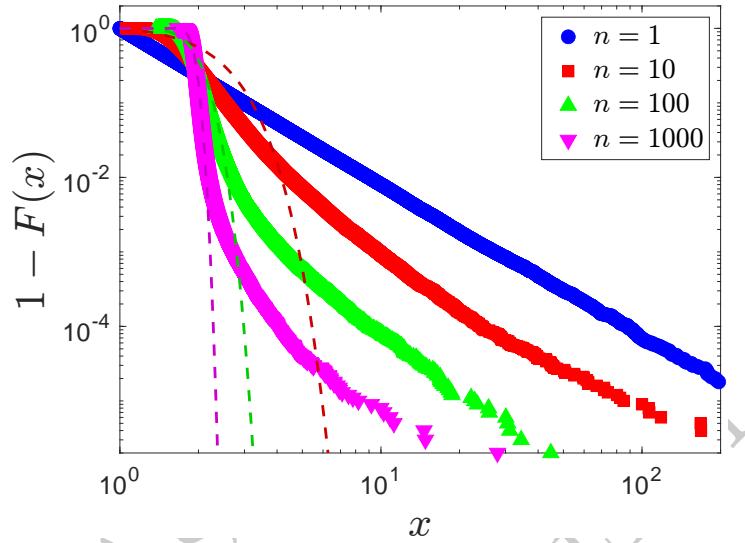


Figure 5.4 Example showing that convergence to the normal distribution might never be achieved in the tails of the distribution of a sum of i.i.d. random variables, even for large aggregation and finite variance. The plots report complementary CDFs for the normalized sums $1/n \sum_{i=1}^n X_i$ of random variables, $X_i > 0$, that are Pareto-distributed with $f(x) \propto x^{-\alpha-1}$ and have exponent $\alpha = 2.1$ (therefore they have finite variance). I plot the CDF for the case $n = 1$ (pure power law) together with the ones for the aggregations of $n = 10$, $n = 100$, and $n = 1,000$ variables. The dashed lines are the CDFs for normal distributions with the same mean and standard deviation of the aggregated variables.

- For $\alpha = \infty$ the distribution is ‘light-tailed’ and all moments are finite and the probability density function tends towards zero at least exponentially fast in the tails.

Remark 5.2. The power law exponent $\alpha = \infty$ indicates a decay rate that is faster than the power law. For instance, the Normal distribution has a rate $f(x) \underset{x \in \text{tails}}{\sim} \exp(-x^2)$.

In many real cases, the right and the left tail of the distribution have different exponents.

The tails of the distribution are very important because they quantify the likelihood of extreme events, such events might be unlikely but when happen they can have strong and disruptive consequences and therefore a good estimation of their likelihood is very important.

Example 5.4 (Sum of power law distributed variables with finite variance: persistence of the tails). Despite the central limit Theorem 5.1 guarantees convergence towards the normal distribution for a sum of random variables with finite variance, the sum of random variables with ‘fat’-tailed distributions with $\alpha \geq 2$, rests fat-tailed with the same exponent α of the individual variables even if their variance is finite. The effect of the sum is to move further away from the mean to the point where the power law tail starts. An example of this, for $\alpha = 2.1$, is provided in Figure 5.4 where the resulting complementary cumulative distribution function of the sum of power-law distributed random variables is reported. One can see that the complementary cumulative distributions display a linear trend in the log-log scale, indicating power law behavior. For $n = 1$ one has a pure power law which is observed to persist in the tails for large aggregations ($n > 1$), even for $n = 1,000$. One can notice that the power law tail starts at larger values of x and at lower probability levels with increasing aggregation.

5.5.1 Chebyshev’s inequality

Observations with very large deviations from the mean are rather unusual when the variance is finite. Chebyshev’s inequality provides the following bounds:

- no more than 1/4 of the values are more than two standard deviations away from the mean;
- no more than 1/9 of the values are more than three standard deviations away from the mean;
- no more than 1/25 of the values are more than five standard deviations away from the mean.
- In general, no more than $1/k^2$ of the values are more than k standard deviations away from the mean.

In a formula, if the random variable X has finite mean μ and standard deviation σ , then

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}. \quad (5.28)$$

Remark 5.3. This inequality is meaningful only for finite variance $\sigma^2 < \infty$. Conversely, unexpected events might become unexpectedly likely when σ^2 diverges. This is not just a hypothetical case, there are indeed several real-world systems well described by statistics with non-defined variance.

Remark 5.4. Chebyshev's inequality provides a rather loose bound in many practical cases. For instance, by comparing with the normal distribution (see Definition 5.1) one can see that the probability of deviations from the mean for the normal distribution is much lower than Chebyshev's bound:

- deviation larger than σ ,
 - normal distribution $P(|X - \mu| > \sigma) \simeq 32\%$,
 - Chebyshev's bound $\leq 100\%$;
- deviation larger than 2σ ,
 - normal distribution $P(|X - \mu| > 2\sigma) \simeq 4.6\%$,
 - Chebyshev's bound $\leq 25\%$;
- deviation larger than 3σ ,
 - normal distribution $P(|X - \mu| > 3\sigma) \simeq 0.27\%$,
 - Chebyshev's bound $\leq 11\%$.

Nonetheless, Chebyshev's inequality is extremely useful because it provides a universal bound that does not depend on the specific distribution.

5.5.2 Cantelli's inequality

In many practical cases, one is more concerned about the fluctuations on one side of the distribution. For instance, for an investor, the likelihood of large gains has very different implications with respect to the likelihood of large losses and one would like to have bounds for each separately. The one-side extension of Chebyshev's inequality is Cantelli's inequality. For the right side of the distribution (X larger than the mean), it is

$$P(X - \mu \geq \lambda) \leq \frac{\sigma^2}{\sigma^2 + \lambda^2}. \quad (5.29)$$

with $\lambda > 0$. While, for the left side of the distribution (X smaller than the mean), it is

$$P(X - \mu \leq -\lambda) \leq \frac{\sigma^2}{\sigma^2 + \lambda^2} \quad (5.30)$$

One might notice that this inequality is very similar to Chebyshev's inequality though if both sides are considered at once, the application of Cantelli's inequality would imply $P(|X - \mu| \geq k\sigma) \leq \frac{2}{1+k^2}$. Instead, Chebyshev's is more restrictive (for $k > 1$).

5.6 Some common probability density functions

Beyond the very important families of normal and stable distributions, there is a vast number of probability density functions that are commonly observed and used for modeling. Let me, therefore, list some of the most relevant ones. Let me

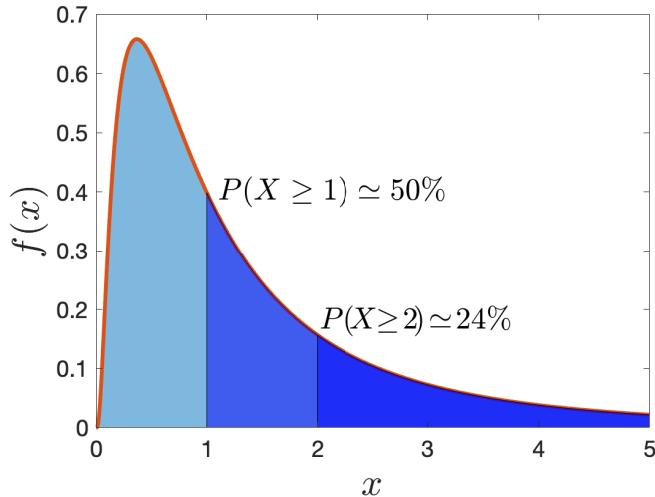


Figure 5.5 Plot of the probability density function $f(x)$ of a log-normal distribution with location $\tilde{\mu} = 0$ and scale $\tilde{\sigma} = 1$. In this case, the mean is at $\mu = \exp(0.5) \simeq 1.65$, while the standard deviation is $\sigma \simeq 2.16$.

however stress that in many real-world cases, random variables do not follow any of these distributions and have instead case-specific properties.

5.6.1 Log-normal distribution

The log-normal distribution describes a random variable whose logarithm is a normal distribution ($\log x \sim \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$):

Definition 5.7 (Log-Normal distribution). The probability density function of the **log-normal distribution** is:

$$f_{\text{logn}}(x|\tilde{\mu}, \tilde{\sigma}^2) = \frac{1}{x\sqrt{2\pi\tilde{\sigma}^2}} \exp\left(-\frac{(\ln x - \tilde{\mu})^2}{2\tilde{\sigma}^2}\right). \quad (5.31)$$

Where the parameter $\tilde{\mu}$ is called location and $\tilde{\sigma} \geq 0$ is called scale. Differently from the normal case, this distribution has support only on the positive domain $x \in (0, +\infty)$.

The log-normal distribution has mean $\mu = \exp(\tilde{\mu} + \tilde{\sigma}^2/2)$ and variance $\sigma^2 = [\exp(\tilde{\sigma}^2) - 1] \exp(2\tilde{\mu} + \tilde{\sigma}^2)$. A plot of the log-normal probability density function with zero location $\tilde{\mu} = 0$ and unitary scale $\tilde{\sigma} = 1$ is reported in Fig.5.5.

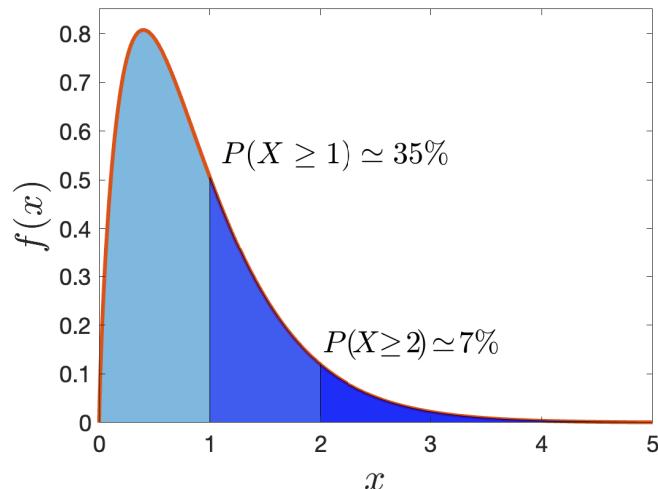


Figure 5.6 Plot of the probability density function of a gamma distribution with shape $\alpha = 1.8$ and rate $\beta = 0.5$. In this case, the mean is $\mu = 0.9$ and the standard deviation is $\sigma \simeq 0.67$.

Remark 5.5. The emergence of log-normal distributions in real systems is a direct consequence of the central limit theorem (see Theorem 5.1). Indeed, when instead of a sum of i.i.d. random variables one considers a product, $Z = X_1 \cdot X_2 \cdots X_n$, then the logarithm of such a product is a sum of i.i.d. random variables $\log Z = \log X_1 + \log X_2 + \dots + \log X_n$ and it must converge towards normal distribution provided that the mean and variance of the logarithm of the random variables $\log X_i$ are finite.

5.6.2 Gamma distribution

The gamma distribution is a common family of distributions which depends on two parameters, α and β , called shape and rate respectively.

Definition 5.8 (Gamma distribution). The probability density function of a **gamma distribution** with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$ is

$$g(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} . \quad (5.32)$$

Sometimes it is expressed in terms of other two parameters: the shape $k = \alpha$ and scale $\theta = 1/\beta$. This distribution is defined only in the positive domain $x \in (0, +\infty)$.

The gamma distribution has mean $\mu = \alpha/\beta$ and variance $\sigma^2 = \alpha/\beta^2$. A plot of the gamma probability density function with shape $\alpha = 1.8$ and rate $\beta = 0.5$ is reported in Fig.5.6.

There are several special cases associated with different α and β which are known under different names. For instance, $X \sim \mathcal{G}(1, 1/\lambda)$ is the **exponential** distribution with rate parameter λ . Another special case is $X \sim \mathcal{G}(k/2, k/2)$ which is the **k-gamma** distribution with k degrees of freedom. This is the probability density function emerging from the sum of k , non-negative, uniformly distributed random variables [Aste and Di Matteo, 2008]. The chi-squared distribution also belongs to the gamma family and it is $X \sim \mathcal{G}(k/2, 1/2)$.

Definition 5.9 (Inverse gamma distribution). The inverse gamma distribution is the probability distribution function of the reciprocal of a gamma-distributed variable.

$$f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{x}\right)^{\alpha+1} e^{-\beta\frac{1}{x}}. \quad (5.33)$$

It is defined over the support $x \in (0, \infty)$, and the parameters $\alpha, \beta > 0$ are respectively called shape and scale.

5.6.3 Chi-squared distribution and F-distribution

Two probability distributions that often arise in the context of statistical validation are the χ^2 distribution and the F-distribution. Indeed, the sum of squares of uncorrelated normally distributed random variables follows a χ^2 distribution and the ratio between two χ^2 -distributed variables follows an F-distribution.

Definition 5.10 (χ^2 distribution). The **Chi-squared distribution** with degrees of freedom k has the probability density function

$$f_{Chi}(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}. \quad (5.34)$$

From Definition 5.8 it should be clear that this distribution is a gamma distribution, $g(x|\alpha, \beta)$, with parameters $\alpha = k/2$ and $\beta = 1/2$.

Definition 5.11 (F-distribution). The **F-distribution** with degrees of freedom d_1 and d_2 has probability density function

$$f_F(x) = c(d_1, d_2) x^{\frac{d_1}{2}-1} \left(1 + \frac{d_1}{d_2} x\right)^{-\frac{d_1+d_2}{2}}, \quad (5.35)$$

with

$$c(d_1, d_2) = \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} \quad (5.36)$$

and $B(\cdot)$ the Beta function.

5.6.4 Student-t distribution

Very often in real, complex systems one observes random variables that are described well with probability density functions that have a compact symmetric body and power law ‘fat’ tails. A distribution with this property which is commonly observed in real systems is the Student-t distribution.

Definition 5.12 (Student-t distribution). The generalized probability density function for the **Student-t distribution** with location parameter μ , scale parameter $\tilde{\sigma}$ and degrees of freedom $\nu \in (0, +\infty)$ is:

$$t(x|\mu, \tilde{\sigma}, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu\tilde{\sigma}^2}} \left(1 + \frac{1}{\nu} \left(\frac{x-\mu}{\tilde{\sigma}}\right)^2\right)^{-\frac{\nu+1}{2}}. \quad (5.37)$$

For $\nu > 1$ the scale parameter μ coincides with the expected value and, for $\nu > 2$ the scale parameter is associated to the variance $\tilde{\sigma}^2 = (\nu - 2)/\nu\sigma^2$. This distribution has support over the whole domain $x \in (-\infty, +\infty)$.

It has the characteristic function

$$\phi(\omega) = \frac{K_{\nu/2}(\sqrt{\nu}|\tilde{\sigma}\omega|) (\sqrt{\nu}|\tilde{\sigma}\omega|)^{\nu/2} e^{i\mu\omega}}{\Gamma(\nu/2)2^{\nu/2-1}} \quad (5.38)$$

with $K_\nu(\cdot)$ a modified Bessel function of the second kind.

Notice that this expression is for the general non-standardized Student-t probability density function. The standardized pdf is simply retrieved from the non-standardized one by imposing $\tilde{\mu} = 0$, $\tilde{\sigma} = 1$. The tail exponent coincides with the degrees of freedom: $\alpha = \nu$. This is a fat-tailed distribution with $t(x|\mu, \tilde{\sigma}, \nu) \underset{|x-\mu|/\tilde{\sigma} \gg 1}{\sim} 1/|x|^{\nu+1}$. A plot of a Student-t probability density function with location $\mu = 0$ and scale $\tilde{\sigma} = 1$ and degrees of freedom $\nu = 1$ is reported in Fig.5.7.

Remark 5.6. The Student-t distribution in its original form has integer degrees of freedom and the general form presented here should be named Pearson Type IV distribution [Pearson, 1894, 1895].

This distribution has its origin in statistics describing the probability distribution of the mean of a normally distributed population when the sample size

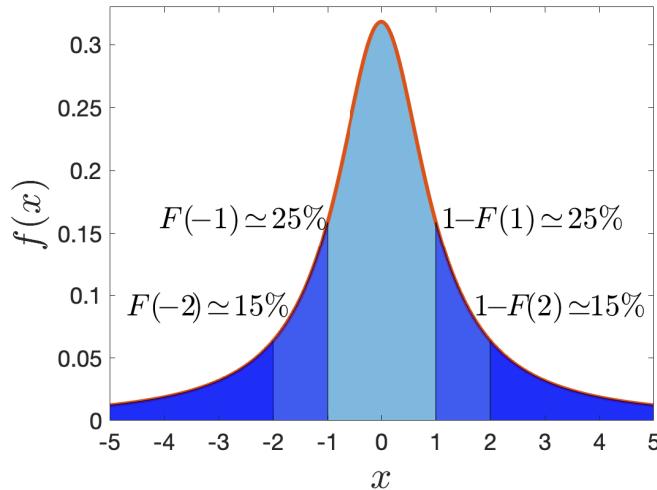


Figure 5.7 Plot of the probability density function of a Student-t distribution with location $\mu = 0$ and scale $\tilde{\sigma} = 1$ and degrees of freedom $\nu = 1$. This distribution has zero mean, undefined variance, and all higher moments.

is small and the standard deviation is unknown Student [1908]. However, it was later observed that it describes very well the statistical properties of a large number of natural and artificial phenomena. It has arisen in the statistical physics literature as the fundamental distribution function associated with non-extensive entropies and, in that context, takes the name of q-Gaussian or q-Tsallis distribution [Tsallis, 1988, 2009b, Umarov et al., 2008].

Cauchy distribution

A special case of the Student-t distribution with $\nu = 1$ is the Cauchy distribution. This is a case where the mean and variance and all the higher moments are undefined.

Definition 5.13 (Cauchy distribution). The probability density function for the **Cauchy distribution** is

$$f_{Cauchy}(x|x_0, \gamma) = \frac{1}{\pi\gamma} \left[\frac{\gamma^2}{(x - x_0)^2 + \gamma^2} \right]. \quad (5.39)$$

In this case, the location parameter conventionally takes the symbol x_0 and the scale parameter is $\gamma > 0$. It has support over the whole domain $x \in (-\infty, +\infty)$.

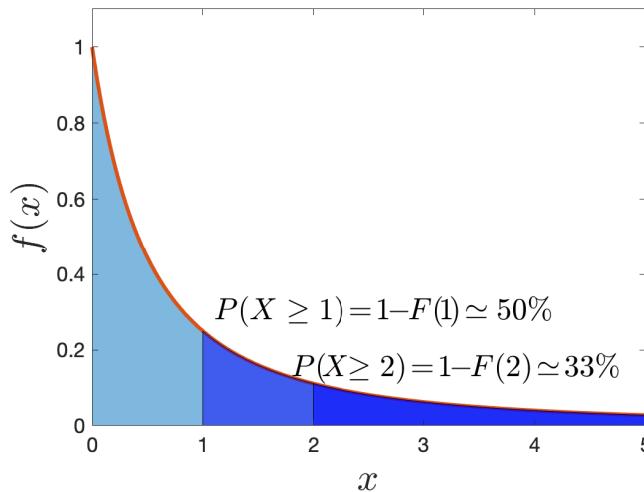


Figure 5.8 Plot of the probability density function of a Pareto distribution with exponent $\alpha = 1$ and $x_m = 1$. This distribution has an undefined mean, variance, and all higher moments.

5.6.5 Pareto distribution

Vilfredo Pareto, an Italian engineer, economist sociologist, and political scientist, observed that the distribution of wealth in Europe at the end of the 1800' followed a power-law pattern. Since then, it has been established that such a power law, the Pareto distribution, is indeed a good model for individual wealth distribution but also for a nation's GDP, income, and other measures of wealth across the world. Power laws have two very important properties: the first is that they are fat-tailed meaning that there is a sizable fraction of the population with values much larger than the mean; the second is that they are peaked on the left side of the support, meaning that the most likely value in the population is the minimum. A plot of a Pareto distribution probability density function with exponent $\alpha = 1$ and $x_m = 1$ is reported in Fig.5.8. In terms of modeling mechanism, power laws distributions naturally emerge when additional wealth is created at a rate proportional to the owning. This is the 'rich get richer' mechanism which apparently reproduces well how fortunes are made.

Definition 5.14 (Pareto distribution). The probability density function of the **Pareto distribution** is

$$f_{\text{Pareto}}(x|x_m, \alpha) = \alpha \frac{x_m^\alpha}{x^{\alpha+1}}, \quad (5.40)$$

with $\alpha > 0$ the shape parameter, and $x_m > 0$ the scale parameter. It has support $x \in [x_m, +\infty)$.

It is easy to see that the Pareto distribution can be defined as well with the support in the negative domain $x \in (-\infty, x_m]$ with $x_m < 0$ by writing $f_{Pareto}(x) = \frac{\alpha}{|x_m|} \left(\frac{x_m}{|x|}\right)^{\alpha+1}$. The Pareto distribution is defined only for $\alpha > 0$ because, otherwise, its integral over the whole support diverges. Similarly, it has defined mean only for $\alpha > 1$, defined variance only for $\alpha > 2$ and it has, in general, defined k^{th} moment m_k and central moment μ_k only for $\alpha > k$.

The Pareto distribution is often used in combination with other distributions (e.g. the normal or the Student-t) to fit the tails of an observed distribution. I will come back to this with an example in Section 14.4.

5.6.6 Generalized Pareto distribution

Another distribution often used to model the tail region is the Generalized Pareto distribution.

Definition 5.15 (Generalized Pareto). The probability density function of a **generalized Pareto distribution** with location parameter $\tilde{\mu} \in (-\infty, \infty)$ and scale parameter $\tilde{\sigma} \in (0, \infty)$ is

$$f_{GenPar}(x|\tilde{\mu}, \tilde{\sigma}, \xi) = \frac{1}{\tilde{\sigma}} \left(1 + \xi \frac{x - \tilde{\mu}}{\tilde{\sigma}}\right)^{-(1/\xi+1)} \quad (5.41)$$

its support is $x \geq \tilde{\mu}$ for $\xi > 0$ or $\tilde{\mu} \leq x \leq \tilde{\mu} - \tilde{\sigma}/\xi$ for $\xi < 0$.

This probability distribution has defined mean $\mu = \tilde{\mu} + \frac{\tilde{\sigma}}{1-\xi}$ for $\xi < 1$ and defined variance $\sigma^2 = \frac{\tilde{\sigma}^2}{(1-\xi)^2(1-2\xi)}$ for $\xi < 1/2$. It is fat-tailed with tail exponent $1/\xi$.

5.6.7 Location-scale family distributions

The normal distribution, the Student-t, and many other probability distribution functions mentioned in this Chapter belong to the location-scale family.

Definition 5.16 (Location-scale family distributions). For the **location-scale family** the probability density function can be expressed as

$$f_{LS}(x) = f\left(\frac{x - \tilde{\mu}}{\tilde{\sigma}}\right), \quad (5.42)$$

where the two parameters $\tilde{\mu}$ and $\tilde{\sigma}$ are respectively the location and scale parameters. If X belongs to the location-scale family distribution, then the linear transformation $Y = a + bX$ is equal in distribution to X and the cumulative distribution function has the property: $F_Y(y) = F_X((y - a)/b)$.

The location-scale family depends on two parameters: location and scale. Therefore all moments must be a function of them. Specifically, if the second moment, μ_2 , is defined, then all higher-order moments must depend on it. Indeed, the characteristic function (see Definition 5.3) of a location-scale distribution is

$$\phi(\omega) = e^{i\tilde{\mu}\omega} \psi(\tilde{\sigma}\omega). \quad (5.43)$$

With $\psi(\cdot)$ independent from $\tilde{\mu}$ or $\tilde{\sigma}$. By using the moment generating formula Eq.5.13, if the first moment, m_1 , is zero, then all odd moments vanish and the even moments are [Berkane and Bentler, 1986]

$$\mu_{2m} = c_{2m} \mu_2^m, \quad (5.44)$$

with

$$c_{2m} = \frac{(2m)!}{(2^m m!) \psi^{(1)}(0)} \frac{\psi^{(m)}(0)}{\psi^{(1)}(0)}. \quad (5.45)$$

Where $\psi^{(m)}(0)$ indicated the m^{th} derivative of $\psi(z)$ computed at $z = 0$.

5.7 Mixture distributions

Probability density functions with various shapes and properties can be constructed by mixing together other probability distribution functions with various weights.

In the discrete case, the mixture of a number of distributions, $f_z(x, \boldsymbol{\theta}(z))$ with $z = 1, \dots, n$, has the form:

$$f(x|\boldsymbol{\theta}) = \sum_{z=1}^n p_z f(x|\boldsymbol{\theta}(z)) \quad (5.46)$$

where the sum of the weights p_z must add to one and, if they are all positive, then $f(x|\boldsymbol{\theta})$ is guaranteed to be a probability distribution function and the weights are a probability mass function. This form of mixture is used for instance in the kernel density estimator method (see Section 13.8) which is a tool to estimate the probability density directly from data.

A continuous spectrum of weights can be equivalently used:

$$f(x|\boldsymbol{\theta}) = \int p(z) f(x|\boldsymbol{\theta}(z)) dz. \quad (5.47)$$

Normally the mixture weights distribution $p(z)$ is a probability density function and, in this case, $f(x|\boldsymbol{\theta})$ is guaranteed to be a probability density function.

A common construction uses normal distributions for $f(x|\boldsymbol{\theta}(z))$ and the mixture takes the name of Gaussian mixture. Though the principle is general and mixtures can be done with any distribution and even with a mixture of distributions.

5.7.1 Student-t distribution as a Gaussian mixture

The Student-t distribution can be retrieved as a mixture of a normal distribution with a gamma distribution. I shall show later in this book that this can be convenient in some circumstances, for instance, to estimate the Student-t parameters from data. It also brings a different perspective on the way the Student-t distribution should be regarded and how it may emerge as the underlying distribution in some circumstances.

Specifically, one has that the Gaussian mixture

$$t(x|\mu, \beta\tilde{\sigma}/\alpha, 2\alpha) = \int_0^\infty g(\tau|\alpha, \beta)\varphi(x|\mu, \tilde{\sigma}^2/\tau)d\tau \quad (5.48)$$

is Student-t with location μ , scale $\beta\tilde{\sigma}/\alpha$ and degrees of freedom 2α . Note that the usual form

$$t(x|\mu, \tilde{\sigma}, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu\tilde{\sigma}^2}} \left(1 + \frac{1}{\nu} \left(\frac{x-\mu}{\tilde{\sigma}}\right)^2\right)^{-\frac{\nu+1}{2}} \quad (5.49)$$

is retrieved for $\tau \sim \mathcal{G}(\nu/2, \nu/2)$ which is the k-gamma distribution (with $k = \nu$).

Remark 5.7. The k-gamma distribution $\mathcal{G}(k/2, k/2)$, see also Section 5.6.2, was originally retrieved in the study of volume fluctuations in disordered materials. It emerged as the probability density function of volumes resulting from the sum of k random volumes with uniformly distributed probabilities Aste and Di Matteo [2008], Coniglio et al. [2017]. It turns out that a large number of natural systems have structures with k-gamma distributions.

Therefore, a process that follows normal statistics but takes place within such a k-gamma distributed structure will eventually display Student-t distributions. This could be a fascinating explanation for the emergence of Student-t distributions in complex systems, which are indeed systems where several components are sharing common resources (in analogy with volumes) and have heterogeneous dynamics that have both local and global dynamics.

5.8 Generalized extreme value distribution and Fréchet, Weibull and Gumbel sub-families

The central limit theorem provides a formidable understanding of the convergence of the sum of n random variables demonstrating that they all converge to one class of distributions. There is another kind of convergence that has very important for practical applications: the distribution of the *maximum* among a set of n random variables. For instance, one might be interested in the probability distribution

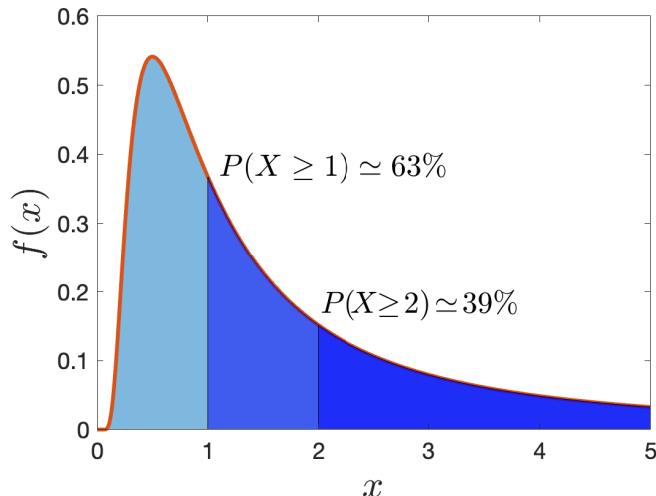


Figure 5.9 Plot of the probability density function of a Generalized extreme value distribution with location $\mu = 0$, scale $\sigma = 1$, shape $\xi = 1$. This distribution has an undefined mean, undefined variance, and all higher moments.

for the largest damage in the most catastrophic hurricane over a series of n hurricanes.

Let me consider a set of i.i.d. observations x_1, \dots, x_n drawn from a population with probability density function $f(x)$. I search for the distribution of the maximum value $M_n = \max\{x_1, \dots, x_n\}$ for a given number n of occurrences.

Theorem 5.4 (Fisher–Tippett–Gnedenko extreme value theorem). *Given a sequence of i.i.d. random variables X_1, X_2, \dots, X_n and let $M_n = \max\{X_1, \dots, X_n\}$. If exists a sequence of pairs of real numbers (a_n, b_n) such that each $a_n > 0$ and*

$$\lim_{n \rightarrow \infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = F(x)$$

then the limit distribution function $F(x)$ belongs to either the Gumbel, Fréchet or Weibull location-scale family distributions that can be grouped into the generalized extreme value distribution.

Definition 5.17 (Generalized extreme value distribution). Given a random variable x , defining

$$z = \frac{(x - \tilde{\mu})}{\tilde{\sigma}} \quad (5.50)$$

with $\tilde{\mu} \in \mathbb{R}$ location parameter and $\tilde{\sigma} > 0$ scale parameter, the **generalized**

extreme value cumulative distribution function is defined as

$$F_{GE}(z|\xi) = \begin{cases} \exp(-(1+\xi z)^{-1/\xi}) & \xi \neq 0 \\ \exp(-\exp(-z)) & \xi = 0 \end{cases} \quad (5.51)$$

where $\xi \in \mathbb{R}$ is the shape parameter. For $\xi > 0$, the distribution has support in $z > -1/\xi$, while for $\xi < 0$ it has support for $z < -1/\xi$, and for $\xi < 0$ it has support over the entire real axis $z \in (-\infty, +\infty)$. The sub-families, defined by $\xi > 0$ and $\xi < 0$ correspond, respectively, to the **Fréchet** and **Weibull** families, whereas the **Gumbel** is associated with $\xi = 0$.

The Gumbel distribution is associated with the distribution of the maxima of a set of normally distributed variables and it is light-tailed. For $\xi > 0$, the generalized extreme value distribution is fat-tailed. In the tail region, it can be fitted well with $F_{GE}(z) \underset{z \text{ large}}{\sim} z^{-\alpha}$ and consequently $f_{GE}(z) \underset{z \text{ large}}{\sim} z^{-\alpha-1}$ where the tail exponent is given by the inverse of the shape parameter $\alpha = 1/\xi$. A plot of the probability density function of a generalized extreme value distribution with location $\mu = 0$, scale $\sigma = 1$, shape $\xi = 1$, is reported in Fig.5.9.

5.9 Infinitely divisible distributions

Probability density functions of random variables which are equivalent to the sum of i.i.d. random variables are called infinitely divisible.

Definition 5.18 (Infinitely divisible). A random variable is **infinitely divisible** if, for every natural number $n \in \mathbb{N}$, it can be represented as a sum of n i.i.d. random variables $X_1 + \dots + X_n$. A probability density function is infinitely divisible if and only if its characteristic function is, for every natural number $n \in \mathbb{N}$, the n^{th} power of some characteristic function:

$$\phi(\omega) = [\psi(\omega)]^n \quad (5.52)$$

Remark 5.8. Notice that infinitely divisible random variables do not have to be the sum of i.i.d. variables, it suffices that their distribution is equivalent to one of a sum of i.i.d. variables.

Infinitely divisible distributions include stable distributions but they are a broader and more general class. Indeed, the i.i.d. variables do not necessarily have the same probability distribution as the sum.

An example of a distribution that is infinitely divisible but is not stable is, for instance, the Poisson distribution which has the probability density function $f(x) = \lambda^x e^{-\lambda} / x!$ and characteristic function $\phi(\omega) = \exp(\lambda(e^{i\omega} - 1))$. Another example is the Gamma distribution (see Definition 5.8) that has characteristic

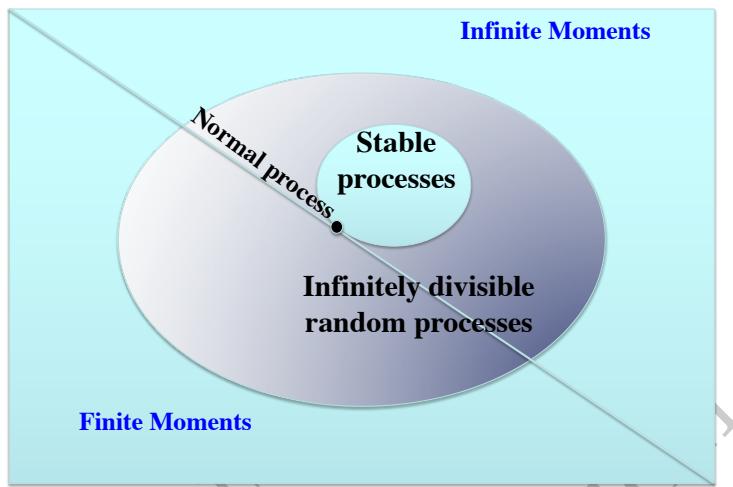


Figure 5.10 Representation of the various categories of probability distributions.

function $\phi(\omega) = (1 - \theta i\omega)^{-k}$. The above distributions are infinitely divisible because their characteristic function can be written as $\phi(\omega) = [\psi(\omega)]^n$ but they are not stable because the scaling $f_Y(y) = \frac{1}{n^{1/\alpha}} f_X(\frac{Y-d}{n^{1/\alpha}})$ (see Definition 5.5) does not hold. Indeed their attractor distribution is the normal distribution (they have finite variance).

5.10 Probability distribution space

Summarizing, one can classify *any* probability distribution into four general classes. The first distinction is between distributions with all moments finite and distributions with only some moments finite. To the first category belong all distributions whose tails decrease exponentially or faster (light-tailed distributions) or bounded distributions defined on finite support $x \in [a, b]$. To the second category belong all distributions that decrease in the tails as power laws $f(x) \underset{|x| \text{ large}}{\sim} |x|^{-\alpha-1}$ that have defined moments m_k only for $k < \alpha$. Infinitely divisible distributions can belong to both these categories, whereas stable distributions only to the second category with an exception for the normal distribution which is the only probability density function that is both stable and has finite moments. Figure 5.10 pictorially represents these classes.

5.11 Data-driven modeling tutorial

<https://github.com/FinancialComputingUCL/DataDrivenModeling/Ch5>

The tutorial for this Chapter covers various topics on univariate probabilities including: the convergence towards the normal distribution (Example 5.1); the convergence toward the Lévy-alpha stable distribution; the persistence of the tails in the case of aggregation of random variables with finite variance but power-law tails (Example 5.4). It also discusses comparison between the quantiles for various random variables with the bound from Chebyshev's inequality.

Exercises

- Using by parts integration formula, demonstrate that $\mathbb{E}(X) = \int_0^\infty (1 - F(x))dx - \int_{-\infty}^0 F(x)dx$.
- Compare Chebyshev's inequality bounds with the quantiles of Student-t distributions with degrees of freedom $\nu = 2.01$ and $\nu = 5$.
- Demonstrate that $\mathbb{E}(X^k) = k \int_0^\infty x^{k-1} (1 - F(x))dx - \int_{-\infty}^0 x^{k-1} F(x)dx$.
- By using Eqs. 5.38, 5.44 and 5.45, derive the expression for the kurtosis of a Student-t distribution.
- Derive the expression for the fourth central moment of a sum of n i.i.d. random variables.
- Discuss why a random variable X from a Pareto's distribution (see Definition 5.14) with $\alpha = 3$, has finite variance but undefined kurtosis.
- Discuss how a sum of i.i.d. random variables X_i from a Pareto's distribution with $\alpha = 3$ can converge towards the normal distribution which has zero skewness and finite kurtosis.
- What is the shape parameter for the distribution of the maxima of a set of normally distributed variables?

6

Multivariate probabilities

6.1 Joint probabilities

When more than one random variable is involved, the probability is called ‘multivariate’. All the fundamental concepts about probabilities rest the same as in the multivariate case. However, when there is more than one variable involved then some extra concepts arise. For instance, for two random variables, X and Y , one must distinguish between the probabilities to observe them in conjunction or independently. The probability to observe them in conjunction is called joint probability, while the probability to observe one variable independently of the value of the other is called marginal probability.

Explicitly, for discrete variables, the joint probability mass function to observe both $X = x$ and $Y = y$ is denoted with

$$p_{XY}(x, y), \quad (6.1)$$

the marginal probability of $X = x$ independently of the value of Y is instead denoted as $p_X(x)$ and it is equal to the sum of the joint probabilities over all values of Y :

$$p_X(x) = \sum_{y \in \Omega_Y} p_{XY}(x, y), \quad (6.2)$$

and vice versa, the marginal probability of $Y = y$ independently of the value of X is

$$p_Y(y) = \sum_{x \in \Omega_X} p_{XY}(x, y). \quad (6.3)$$

Example 6.1 (Joint probability). For instance, if the two discrete variables are the age of an individual and the color of her/his hair. $p(\text{Age} = 25, \text{Hair-Color} = \text{blond})$ is the probability that, across a population, a person selected at random has blond hair and is 25 years old. In a population of adults in their 20’s, these two variables are most likely unrelated: individuals are blond or dark-haired independently of their age. However, the hairs of older individuals eventually turn white and there is a larger probability of blond hair in very young populations of European descent. Therefore, age

and hair color are not independent variables even if they might be if one restricts the age range of the investigation.

The joint cumulative distribution function is the probability that the random variable X takes values smaller or equal than x and jointly the random variable Y takes values smaller or equal than y :

$$F_{XY}(x, y) = P(X \leq x, Y \leq y), \quad (6.4)$$

this definition holds for both discrete and continuous variables.

Remark 6.1. One can notice that, for each couple of values ($X = x, Y = y$) the two-dimensional plane (X, Y) can be divided into four quadrants respectively associated with combinations of regions with X and Y larger or smaller than the two values x and y . Considering more than two variables, with p variables the space gets subdivided into 2^p regions. One can therefore realize that the concept of multivariate cumulative distribution becomes rapidly complex because it involves a very large number of possibilities.

For two continuous random variables, the joint probability density function is

$$f_{XY}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{XY}(x, y) \quad (6.5)$$

if $F_{XY}(x, y)$ is differentiable in (x, y) . Conversely one has

$$F_{XY}(x, y) = \int_{-\infty}^x \int_{-\infty}^y f_{XY}(x, y) dx dy, \quad (6.6)$$

if $f_{XY}(x, y)$ is integrable in $(-\infty, x)$ and $(-\infty, y)$.

The marginal probability density functions are

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dy; \quad f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dx. \quad (6.7)$$

In this case, they are univariate probabilities but if one considers the case with more than two variables, the marginals are multivariate probabilities of a subset of the variables.

The extension to multivariate cases with more than two variables is straightforward and, to handle it, I shall adopt the boldface notation to indicate vectors (see Definition 2.12). Consistently the joint probability density function takes the form $\mathbf{f}_{\mathbf{X}}(\mathbf{x})$, indicating the value of the density for $\mathbf{X} = \mathbf{x}$ (i.e. $X_1 = x_1, \dots, X_p = x_p$).

6.2 Covariance matrix

The expected values for multivariate probabilities involve more than one variable as well. A very important case is the covariance which is the expected value of the product of the deviation from the mean of two variables

Definition 6.1 (Covariance). The **covariance** between two random variables X and Y with expected values μ_X and μ_Y , is defined as:

$$\text{Cov}(X, Y) = \mathbb{E}((X - \mu_X)(Y - \mu_Y)). \quad (6.8)$$

The covariance of a variable with itself is the variance:

$$\text{Cov}(X, X) = \sigma_X^2. \quad (6.9)$$

When there are more than two variables, say $\mathbf{X} = (X_1, \dots, X_p)^\top$, a covariance can be defined for any couple of variables (X_i, X_j) . This results in a $p \times p$ matrix that is named covariance matrix.

Definition 6.2 (Covariance matrix). The covariance matrix, $\Sigma_{\mathbf{XX}}$, for a set of random variables $\mathbf{X} = (X_1, \dots, X_p)^\top$, is a $p \times p$ matrix with elements:

$$(\Sigma_{\mathbf{XX}})_{i,j} = \Sigma_{i,j} = \text{Cov}(X_i, X_j). \quad (6.10)$$

The covariance matrix is symmetric ($\Sigma_{i,j} = \Sigma_{j,i}$) and has the variances on the diagonal ($\Sigma_{i,i} = \sigma_i^2$).

6.3 Correlation matrix

A quantity strictly related to covariance is the correlation which simply is the covariance normalized by the product of the standard deviations.

Definition 6.3 (Correlation). The **correlation** between two random variables X and Y is:

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}((X - \mu_X)(Y - \mu_Y))}{\sqrt{\mathbb{E}((X - \mu_X)^2)\mathbb{E}((Y - \mu_Y)^2)}}. \quad (6.11)$$

The correlation is often referred to as the **Pearson's correlation**.

Remark 6.2. The correlation between a variable and itself is equal to one: $\text{Corr}(X, Y) = 1$. Correlations have the nice property of being defined between $[-1, +1]$ and quantify the amount of linear dependency between the variables, this will be discussed in detail in Chapter 8, specifically Section 8.3.2.

The bounds $[-1, +1]$ for the correlation coefficient is a direct consequence of the Cauchy-Schwarz inequality (for two random variables A and B ,

$\mathbb{E}(AB)^2 \leq \mathbb{E}(A^2)\mathbb{E}(B^2)$) which imposes

$$\text{Cov}(X, Y)^2 \leq \text{Var}(X)\text{Var}(Y). \quad (6.12)$$

and therefore, from Eq.6.11,

$$|\text{Corr}(X, Y)| \leq 1. \quad (6.13)$$

Analogously with the covariance matrix, the correlation matrix is defined as a $p \times p$ matrix whose coefficients are the correlations between the couples of variables.

Definition 6.4 (Correlation matrix). The correlation matrix is a $p \times p$ matrix with elements:

$$(\mathbf{C}_{\mathbf{XX}})_{i,j} = \rho_{i,j} = \text{Corr}(X_i, X_j). \quad (6.14)$$

The correlation matrix is symmetric, $\rho_{i,j} = \rho_{j,i}$, and has ones on the diagonal, $\rho_{i,i} = 1$.

We shall see in the rest of this book, and especially in Chapter 8, that the covariance and correlations matrices are very important and of great practical use for the modeling of multivariate systems.

6.4 Multivariate normal distribution

I have already discussed in the univariate case that the normal distribution plays a very important role. This is also the case in the multivariate realm. Let's start with two variables X_1 and X_2 . The bivariate normal distribution is

$$\varphi(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{1,2}^2}} \exp\left(-\frac{1}{2}d_{12}^2\right) \quad (6.15)$$

where

$$d_{12}^2 = \frac{1}{(1-\rho_{1,2}^2)} \left(\frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} - \frac{2\rho_{1,2}(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} \right), \quad (6.16)$$

with $\rho_{1,2}$ the correlation coefficient between the two variables and σ_1^2 , σ_2^2 the respective variances of the two variables.

By using vector and matrix notations the bivariate normal and in general any multivariate normal can be written in a more compact and convenient form.

Definition 6.5 (Multivariate normal). The probability density function of

the multivariate normal distribution of p random variables, $\mathbf{X} \in \mathbb{R}^{p \times 1}$ is:

$$\varphi(\mathbf{X} = \mathbf{x}) = \frac{1}{\sqrt{(2\pi)^p |\Sigma_{\mathbf{XX}}|}} \exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}})^T \Sigma_{\mathbf{XX}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}}) \right]. \quad (6.17)$$

It has support on the entire \mathbb{R}^p space. The term $\Sigma_{\mathbf{XX}}$ is the $p \times p$ covariance matrix which must be invertible, the term $\Sigma_{\mathbf{XX}}^{-1}$ is the inverse of the covariance matrix and $|\Sigma_{\mathbf{XX}}|$ indicates the determinant of the covariance matrix.

One can verify that the bivariate case is retrieved with $\mathbf{x} = (x_1, x_2)^T$, $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ and

$$\Sigma_{\mathbf{XX}} = \begin{pmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 \\ \rho_{1,2}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}. \quad (6.18)$$

One can see from the definition of the multivariate normal distribution that the coefficients are actually not the elements of the covariance matrix but of its inverse $\mathbf{J}_{\mathbf{XX}} = \Sigma_{\mathbf{XX}}^{-1}$ (notice that $|\Sigma_{\mathbf{XX}}| = 1/|\Sigma_{\mathbf{XX}}^{-1}|$). Indeed, the inverse covariance plays a very important role not only in the multivariate normal statistics but also in multilinear regression and in general in the multivariate statistics of the elliptical distribution family.

Definition 6.6 (Precision matrix). The inverse covariance

$$\mathbf{J}_{\mathbf{XX}} = \Sigma_{\mathbf{XX}}^{-1}, \quad (6.19)$$

is also known as concentration matrix or **precision matrix**. This name comes from the term ‘precision’ that in statistics is the reciprocal of the variance. We shall see in Chapter 8 that the elements of $\mathbf{J}_{\mathbf{XX}}$ are associated to conditional linear dependence. Couples of conditionally independent variables X_i, X_j have $(\mathbf{J}_{\mathbf{XX}})_{i,j} = 0$.

6.5 The elliptical distribution family

The multivariate normal distribution belongs to a large family of multivariate probability density functions known as the elliptical distribution family [Fang, 2018].

Definition 6.7 (Elliptical family distribution). The probability density function of the **elliptical family** (when defined) can be written as:

$$f(\mathbf{X} = \mathbf{x}) = k_p |\boldsymbol{\Omega}|^{-\frac{1}{2}} g(\tilde{d}_{\mathbf{X}}^2), \quad (6.20)$$

where, k_p is a constant, $g(\cdot)$ is a scalar function, which is non-negative Lebesgue measurable on $[0, \infty)$ such that $\int_0^\infty t^{p/2-1} g(t) dt < \infty$, and it is called density generator function. The term $\tilde{d}_{\mathbf{x}}^2$ is a non-negative scalar which is a generalization of the (squared) Mahalanobis distance. Ω is a positively defined matrix named scale matrix.

Definition 6.8 (Mahalanobis distance). For a multivariate set of random variables $\mathbf{X} \in \mathbb{R}^{p \times 1}$, the **Mahalanobis distance** is defined as [Mahalanobis, 1936]:

$$d_{\mathbf{x}} = \sqrt{(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^T \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})}. \quad (6.21)$$

It is a measure of the distance of an observation, \mathbf{x} , from the centroid position $\boldsymbol{\mu}_{\mathbf{x}}$ scaled by the metric defined by the covariance of the variables. Extreme, ‘atypical’ observations have larger Mahalanobis distances than ‘typical’ observations which are nearer to the centroid position. The covariance scaling makes the $d_{\mathbf{x}} = 1$ surface being an ellipsoid in p -dimensions.

Definition 6.9 (Generalized Mahalanobis distance). The generalization of the Mahalanobis distances is

$$\tilde{d}_{\mathbf{x}} = \sqrt{(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^T \boldsymbol{\Omega}_{\mathbf{xx}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})}. \quad (6.22)$$

Where Ω is the scale matrix. When the covariance is defined, the scale matrix is proportional to it. For the multivariate normal case, the scale matrix coincides with the covariance matrix and $\tilde{d}_{\mathbf{x}} = d_{\mathbf{x}}$. For the elliptic family distribution, ellipsoidal surfaces with equal Generalized Mahalanobis distance, $\tilde{d}_{\mathbf{x}} = \text{const.}$, have equal density.

The multivariate normal is the most used member of the elliptical family and it has $g(d_{\mathbf{x}}^2) = \exp(-d_{\mathbf{x}}^2/2)$ (notice $\tilde{d}_{\mathbf{x}} = d_{\mathbf{x}}$ in this case). Another noticeable member of the elliptical family is the multivariate Student-t distribution which has $g(\tilde{d}_{\mathbf{x}}^2) = [1 + \frac{\tilde{d}_{\mathbf{x}}^2}{\nu}]^{-\frac{\nu+p}{2}}$. In this case, when $\nu > 2$, the matrix $\boldsymbol{\Omega}_{\mathbf{xx}}$ in Eq.6.22 is proportional to the covariance matrix: $\boldsymbol{\Omega}_{\mathbf{xx}} = \frac{\nu-2}{\nu} \boldsymbol{\Sigma}_{\mathbf{xx}}$.

Definition 6.10 (Multivariate Student-t). The probability density function of the **multivariate Student-t** is

$$t(\mathbf{X} = \mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Omega}_{\mathbf{xx}}, \nu) = \frac{\Gamma[(\nu + p)/2]}{\Gamma(\nu/2) \sqrt{\nu^p \pi^p |\boldsymbol{\Omega}_{\mathbf{xx}}|}} \left[1 + \frac{1}{\nu} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^T \boldsymbol{\Omega}_{\mathbf{xx}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}) \right]^{-\frac{\nu+p}{2}}, \quad (6.23)$$

where $\Omega_{\mathbf{XX}}$ is the scale matrix, p is the dimension of $\mathbf{X} \in \mathbb{R}^{p \times 1}$ and $\nu > 0$ is a parameter known as ‘degrees of freedom’. For $\nu > 1$, the term $\mu_{\mathbf{X}}$ is the vector of expected values; for $\nu > 2$ the covariance matrix is $\Sigma_{\mathbf{XX}} = \nu/(\nu - 2)\Omega_{\mathbf{XX}}$ otherwise they are undefined. The case $\nu = 1$ is the multivariate Cauchy distribution.

Another sub-family of the elliptical distribution family is the power exponential family which corresponds to $g(\tilde{d}_{\mathbf{X}}^2) = \exp(-(\tilde{d}_{\mathbf{X}}^2)^{\beta})$ for $\beta \in (0, \infty)$. Clearly, the multivariate normal is the case $\beta = 1$ (note however the absence of the factor $1/2$, which means $\Omega_{\mathbf{XX}} = \frac{1}{2}\Sigma_{\mathbf{XX}}$ in the power exponential family notation). The Laplace distribution has instead $g(\tilde{d}_{\mathbf{X}}^2) = \exp(-|\tilde{d}_{\mathbf{X}}^2|)$ and it is the case with $\beta = 1/2$.

In some cases, in which the probability density function cannot be written in a simple explicit form, the elliptical distribution can be conveniently expressed in terms of the characteristic function, which is of the form

$$\varphi_{\mathbf{X}}(\omega) = e^{i\mu^T \omega} \psi\left(-\frac{1}{2}\omega^T \Omega_{\mathbf{XX}} \omega\right). \quad (6.24)$$

where $\psi(\cdot)$ is a scalar function. For the multivariate normal case, $\psi(\cdot) = \exp(\cdot)$.

Remark 6.3. The elliptical distribution family is invariant under affine transformations. In other words, if \mathbf{X} belongs to an elliptical distribution then the distribution of

$$\mathbf{Y} = \mathbf{a} + \mathbf{B}^T \mathbf{X} \quad (6.25)$$

with $\mathbf{B} \in \mathbb{R}^{p_X \times p_Y}$, $\mathbf{a} \in \mathbb{R}^{p_Y \times 1}$, $p_Y \leq p_X$ and $\text{rank}(\mathbf{B}) = p_Y$, is also elliptically distributed. Specifically, it has,

$$\mu_{\mathbf{Y}} = \mathbf{B}^T \mu_{\mathbf{X}} + \mathbf{a}, \quad (6.26)$$

and

$$\Omega_{\mathbf{YY}} = \mathbf{B}^T \Omega_{\mathbf{XX}} \mathbf{B}, \quad (6.27)$$

and for $p_Y < p_X$

$$g_Y(t) = \int_0^\infty s^{(p_X - p_Y)/2 - 1} g_X(t + s) ds, \quad (6.28)$$

and instead

$$g_Y(t) = g_X(t), \quad (6.29)$$

for $p_Y = p_X$.

Therefore, in general, the distribution rests elliptical under affine transformations but the density generator changes. However, for some functional forms of $g(\cdot)$ the density generator rests also invariant (keeps the same functional form), and therefore the distribution itself is invariant. This is

the case for the normal distribution, where $g_Y(t) = g_X(t) = \exp(-t/2)$. It is also the case for the Student-t, where by applying Eq.6.28 to $g_X(t) = [1 + \frac{t}{\nu}]^{-\frac{\nu+p_X}{2}}$ one retrieves $g_Y(t) = [1 + \frac{t}{\nu}]^{-\frac{\nu+p_Y}{2}}$.

Example 6.2 (Distribution of a linear combination of multivariate elliptical variables). A case of practical relevance is the resulting distribution from a sum (or any linear combination) of variables (X_1, X_2, \dots, X_p) which are multivariate elliptically distributed. Here, I am interested in the distribution of a linear combination of these variables, $Y = \beta_1 X_1 + \dots + \beta_p X_p$ or, in vectorial notation:

$$Y = \mathbf{B}^\top \mathbf{X}. \quad (6.30)$$

This is an affine transformation of the same kind of the one in the previous Remark 6.3. One has therefore that Y must follow as well a (univariate) elliptical distribution that one can write in terms of its characteristic function as:

$$\varphi_Y(\omega) = e^{i\mu_Y \omega} \psi\left(-\frac{1}{2}\Omega_{YY}\omega^2\right). \quad (6.31)$$

where $\mu_Y = \mathbf{B}^\top \boldsymbol{\mu}_X$ and $\Omega_{YY} = \mathbf{B}^\top \boldsymbol{\Omega}_{XX} \mathbf{B} = \sigma_Y^2$ are scalars. One can notice that independently on the kind of elliptical multivariate form of the distribution of \mathbf{X} , the distribution of Y depends only on two parameters: the location μ_Y and the scale σ_Y . Indeed, univariate elliptical distributions are a sub-family of the location-scale family (see Definition 5.16).

Remark 6.4. In the previous Example (6.2) I have shown that any linear combination of multivariate Student-t distributed variables has a Student-t distribution. However, it is known that the Student-t distribution is not a stable distribution (see Section 5.3) and therefore a linear combination of i.i.d. Student-t distributed variables is not Student-t distributed and converges towards a stable distribution. This apparent contradiction is a consequence of the fact that this invariance is for a linear combination of multivariate Student-t distributed variables, not i.i.d. Student-t distributed variables. Indeed, multivariate Student-t distributed variables are not independent even when $\boldsymbol{\Omega}$ is a diagonal matrix.

6.5.1 Marginal distribution for the elliptical family

A consequence of the invariance under affine transformations is that the marginal distributions of elliptical family distributions are elliptical as well. In particular, consider the elliptically distributed random variable $\mathbf{Z}^\top = (\mathbf{X}^\top, \mathbf{Y}^\top)$ with pa-

rameters μ_Z , Ω_{ZZ} and density generator functional parameter $g_Z(\cdot)$. To obtain the marginal for \mathbf{Y} one can use the affine transformation $\mathbf{Y} = \mathbf{a} + \mathbf{B}^\top \mathbf{Z}$ with $\mathbf{B}^\top = (\mathbf{0}_{p_X}^\top, \mathbf{1}_{p_Y}^\top)$ where $\mathbf{0}_{p_X}$ is a $p_X \times 1$ vector of zeros and $\mathbf{1}_{p_Y}$ is a $p_Y \times 1$ vector of ones. From the invariance for affine transformation (see Remark 6.3) one has that the marginal must also be elliptical. Specifically, the marginal of the variable \mathbf{Y} has expected values μ_Y and scale matrix Ω_{YY} , which is a block element of the joint shape matrix:

$$\Omega_{ZZ} = \begin{pmatrix} \Omega_{XX} & \Omega_{XY} \\ \Omega_{YX} & \Omega_{YY} \end{pmatrix}. \quad (6.32)$$

The functional parameter $g_Y(\cdot)$ is instead

$$g_Y(t) = \int_0^\infty s^{\frac{p_X}{2}-1} g_Z(t+s) ds. \quad (6.33)$$

It is evident from this formula that the functional form of $g_Y(\cdot)$ can coincide with the functional form of $g_Z(\cdot)$ only in some cases. Such special cases, however, notably include the normal and the Student-t distributions.

Remark 6.5. In the case of Student-t distribution, the degrees of freedom are the same for the joint and the marginal probability density functions.

6.6 The Bayes' theorem and conditional probability

The probability to observe one variable while the value of another variable is constrained to a certain value is called conditional probability. For two discrete variables the conditional probability to draw $X = x$ when $Y = y$ is denoted as

$$p_{X|Y}(x|y) = P(X = x|Y = y). \quad (6.34)$$

The conditional probability is related to the joint and marginal probabilities through the Bayes' formula

$$p_{X|Y}(x|y) = \frac{p_{XY}(x,y)}{p_Y(y)} \quad \text{for } p_Y(y) \neq 0. \quad (6.35)$$

For continuous variables, one can write this expression in terms of the density functions, and it takes the form

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} \quad \text{for } f_Y(y) \neq 0. \quad (6.36)$$

The Bayes' formula is the basis of Bayesian statistics which is a specific way to describe and interpret probability and statistics where probability is considered to express a degree of belief in an outcome. Such a belief can derive from observations and it can then be refined through the acquisition

of new data. This process can indeed be expressed through the formula:

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{p_Y(y)}. \quad (6.37)$$

Which is a different way of writing the Bayes' formula Eq.6.35, with $p_{XY}(x,y)$ written as $p_{Y|X}(y|x)p_X(x)$ (by using the Bayes' formula itself). The interpretation of the above formula in Bayesian statistics is that y represents new evidence that is used to update the probability

$$p_X(x) \text{ which is the } \mathbf{prior \ probability} \text{ (of } x) \quad (6.38)$$

expressing the belief about x before the new evidence is considered. Instead, the quantity

$$p_{Y|X}(y|x) \text{ is the } \mathbf{likelihood} \text{ (of } y \text{ given } x) \quad (6.39)$$

which is the probability of y for a given x . Finally, the quantity

$$p_{X|Y}(x|y) \text{ is the } \mathbf{posterior \ probability} \text{ (of } x \text{ given } y) \quad (6.40)$$

expressing the new belief about x after the new evidence y has been considered. The unconditional probability $p_Y(y)$ is often hard to compute but in many cases, this is not necessary because the aim is to maximize the posterior probability, $p_{X|Y}(x|y)$, for a given evidence y , in this case, $p_Y(y)$ is a constant coefficient of proportionality and one has to maximize only $p_{Y|X}(y|x)p_X(x)$.

Remark 6.6. Bayes was a reverend who lived in the 18th century. He is considered to be the first who formulated the concept of conditional probability and proposed a methodology to compute the conditional probability in his '*An Essay towards solving a Problem in the Doctrine of Chances*' published in 1763. It was not a theorem and not even a formula. Indeed, at that time probabilities were not formulated yet in the modern mathematical form and the domain was studied mostly for understanding chances in gambling. A mathematical formulation came much later, in 1812, within the '*Théorie analytique des probabilités*' by Pierre-Simon Laplace.

Given the 'holy' roots of the initiator, it might not be too much of a surprise that Bayesian probability principles have become a sort of cult among statisticians.

Example 6.3 (Bayesian interpretation for the Student-t distribution as the marginal of a Gaussian mixture). I have already discussed in Section 5.7.1 that the univariate Student-t distribution can be retrieved as a Gaussian mixture. This can be reformulated in terms of multivariate probabilities: the

Student-t distribution is the marginal distribution of a bivariate probability density function given by the product of normal and gamma distributions:

$$f(X, \tau) = g(\tau|\alpha, \beta)\varphi(x|\mu, \tilde{\sigma}^2/\tau) \quad (6.41)$$

which is sometimes referred to as normal-gamma distribution. Specifically, let the conditional distribution of X given τ be normal (see Definition 5.1) with mean μ and variance $\tilde{\sigma}^2/\tau$

$$X|\tau \sim \mathcal{N}(\mu, \tilde{\sigma}^2/\tau) \quad (6.42)$$

and let τ be a gamma-distributed random variable (see Definition 5.8), independent from X , with shape parameter α and rate parameter β

$$\tau \sim \mathcal{G}(\alpha, \beta). \quad (6.43)$$

Then, the marginal distribution of X is Student-t distributed (see Definition 5.12), with location μ , scale $\frac{\beta}{\alpha}\tilde{\sigma}$ and degrees of freedom 2α :

$$X \sim \mathcal{T}(\mu, \frac{\beta}{\alpha}\tilde{\sigma}, 2\alpha). \quad (6.44)$$

In Bayesian terms, one can say that the Student-t is the marginalized posterior distribution of a normal distribution with unknown variance with conjugate prior that follows an inverse gamma distribution (a gamma distribution for the conjugate, i.e. $1/\tau$, see Definition 5.9).

The multivariate version is identical except for the use of bold symbols and the covariance:

$$\mathbf{X}|\tau \sim \mathcal{N}(\boldsymbol{\mu}, \tilde{\boldsymbol{\Sigma}}/\tau). \quad (6.45)$$

One can observe that in this representation as a Gaussian mixture the multivariate Student-t distribution is very peculiar because its fat-tailed variability is the consequence of the gamma-distributed term that induces variations in the covariance of the multivariate normal. Therefore the covariance structure is determined by the multivariate Gaussian.

Definition 6.11 (Likelihood and log-likelihood). In this section, I introduced a quantity named **likelihood** (see, Eq.6.39). Within the Bayesian perspective, the likelihood is the conditional probability of $Y = y$ given $X = x$. Throughout this book, we shall see the likelihood mentioned in the context of modeling as the probability that a given model assigns to an observation. This is also a conditional probability but the conditioning is with respect to the set of the model parameters. It is quite intuitive that a meaningful model constructed from data must return large probabilities for the data from which it has been constructed. In other words, it must return large likelihoods for the observation set, and often the likelihood is

the ‘gain’ function that models aim to maximize (see, for instance, Sections 14.2 and 18.6). A quantity often used is the **log-likelihood**, which is simply the logarithm of the likelihood. We shall see that the log-likelihood is often more practical to handle.

6.7 Conditional distribution for the elliptical distribution family

Let me conclude this chapter by discussing the conditional probability density function for the elliptical distribution family, for which the joint density function is in the form $f(\mathbf{X}, \mathbf{Y}) = k|\Omega_{\mathbf{XX}}|^{-\frac{1}{2}}g(\tilde{d}^2)$ (when defined, see Eq.6.20). The multivariate conditional probability of $\mathbf{Y} \in \mathbb{R}^{p_Y \times 1}$ given $\mathbf{X} \in \mathbb{R}^{p_X \times 1}$ is defined through the Bayes formula $f(\mathbf{Y}|\mathbf{X} = \mathbf{x}) = f(\mathbf{X} = \mathbf{x}, \mathbf{Y})/f(\mathbf{X} = \mathbf{x})$. When conditioning, one variable is fixed to some values, $\mathbf{X} = \mathbf{x}$, therefore $f(\mathbf{X} = \mathbf{x})$ in the previous formula is a constant and consequently

$$f(\mathbf{Y}|\mathbf{X} = \mathbf{x}) \propto f(\mathbf{X} = \mathbf{x}, \mathbf{Y}) \propto g(\tilde{d}^2). \quad (6.46)$$

The term \tilde{d}^2 in the argument can be re-written as

$$\tilde{d}^2 = \tilde{d}_{\mathbf{Y}|\mathbf{x}}^2 + \tilde{d}_{\mathbf{x}}^2 \quad (6.47)$$

with

$$\tilde{d}_{\mathbf{x}}^2 = (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^\top \Omega_{\mathbf{XX}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}), \quad (6.48)$$

which is a constant for any given $\mathbf{X} = \mathbf{x}$. The other term is instead

$$\tilde{d}_{\mathbf{Y}|\mathbf{x}}^2 = (\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}|\mathbf{x}})^\top \mathbf{J}_{\mathbf{YY}|\mathbf{x}} (\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}|\mathbf{x}}) \quad (6.49)$$

with

$$\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{x}} = \boldsymbol{\mu}_{\mathbf{Y}} + \Omega_{\mathbf{YX}} \Omega_{\mathbf{XX}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}), \quad (6.50)$$

and

$$\mathbf{J}_{\mathbf{YY}|\mathbf{x}} = (\Omega^{-1})_{\mathbf{YY}} = (\Omega_{\mathbf{YY}} - \Omega_{\mathbf{YX}} \Omega_{\mathbf{XX}}^{-1} \Omega_{\mathbf{XY}})^{-1}, \quad (6.51)$$

(notice that this is not $\Omega_{\mathbf{YY}}^{-1}$). The terms $\Omega_{\mathbf{XX}}$, $\Omega_{\mathbf{XY}}$, $\Omega_{\mathbf{YY}}$ and $\Omega_{\mathbf{YX}}$ are the block elements of Ω (as in Eq.6.32) and the matrix $\Omega_{\mathbf{XX}}$ is assumed invertible.

Therefore, the conditional probability density function (when is defined) for any probability of the elliptical family is of the form:

$$f(\mathbf{Y}|\mathbf{X} = \mathbf{x}) = k_{p_Y} |\mathbf{J}_{\mathbf{YY}|\mathbf{x}}|^{1/2} g(\tilde{d}_{\mathbf{Y}|\mathbf{x}}^2 + \tilde{d}_{\mathbf{x}}^2). \quad (6.52)$$

The functional form of the conditional probability is, therefore, very similar to the form of the joint probability density function beside the additive constant $\tilde{d}_{\mathbf{x}}^2$ in the argument.

Let me provide an intuitive geometric perspective of a multivariate distribution from the elliptical family. The value of the probability density at any given point in space is determined by the generalized Mahalanobis distance (see Definition 6.9): two observations at the same generalized Mahalanobis distance have

the same probability density value. In particular, geometrically, the ellipsoidal surfaces $\tilde{d}_{\mathbf{X}}^2 = \text{const.}$ describes multivariate configurations with equal probabilities. Small fluctuations from the centroids are contained in the distribution's body within the surface $\tilde{d}_{\mathbf{X}}^2 = 1$ while atypical large fluctuations are positioned outside this surface. From Eqs.6.49-6.51 one observes that, from a geometrical perspective, conditioning to $\mathbf{X} = \mathbf{x}$ can provoke three effects:

- a shift of the barycentre of the ellipsoid;
- a rotation of the ellipsoid;
- a change of sizes of the principal axis of the ellipsoid.

Figure 6.1 depicts such effects in a schematic way.

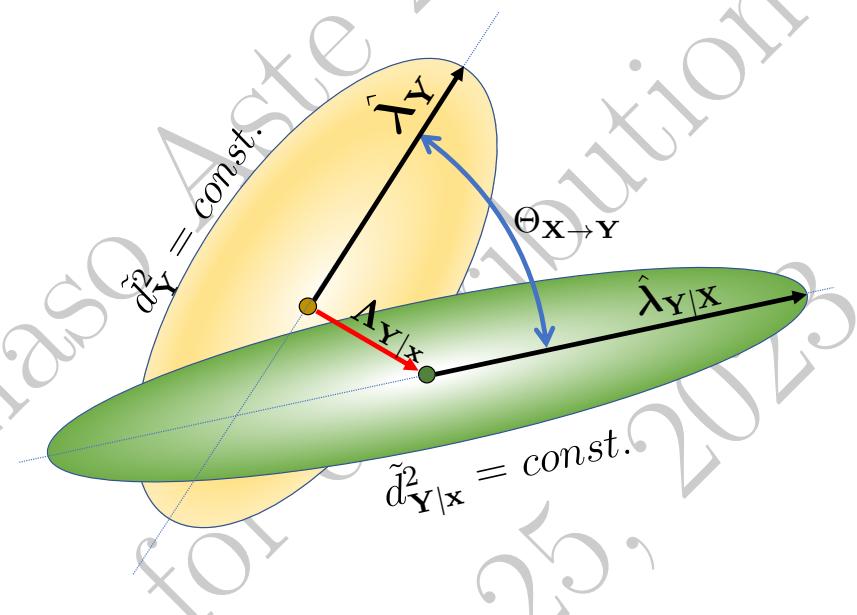


Figure 6.1 A pictorial representation of the effect of conditioning on the equiprobability surface of a multivariate elliptical distribution. Such surfaces are ellipsoids in a $p_{\mathbf{Y}}$ -dimensional space. They are respectively described by the equations $\tilde{d}_{\mathbf{Y}}^2 = \text{const.}$ and $\tilde{d}_{\mathbf{Y}|\mathbf{x}}^2 = \text{const.}$ (see Definition 6.9 and Eq.6.49). Conditioning to $\mathbf{X} = \mathbf{x}$ shifts the barycentre of the ellipsoid $\tilde{d}_{\mathbf{Y}}^2 = \text{const.}$ by the vector $\Delta_{\mathbf{Y}|\mathbf{x}}$, it rotates the most elongated axis by an angle $\Theta_{\mathbf{X} \rightarrow \mathbf{Y}}$ and it changes the length of such axis from $\hat{\lambda}_{\mathbf{Y}}^{1/2}$ to $\hat{\lambda}_{\mathbf{Y}|\mathbf{x}}^{1/2}$.

The shift in the position of the centroid of the probability distribution of the \mathbf{Y} variables caused by conditioning to \mathbf{X} is

$$\Delta_{\mathbf{Y}|\mathbf{x}} = \mu_{\mathbf{Y}|\mathbf{x}} - \mu_{\mathbf{Y}} = \Omega_{\mathbf{Y}\mathbf{X}}\Omega_{\mathbf{X}\mathbf{X}}^{-1}(\mathbf{x} - \mu_{\mathbf{X}}). \quad (6.53)$$

It must be noticed that this shift coincides with the change in the expected value from the multilinear regression of \mathbf{Y} with respect to $\mathbf{X} = \mathbf{x}$. This effect of conditioning is normally the most relevant for what concerns stress propagation.

Indeed, I shall show in the next chapter that entropy and consequently uncertainty is always reduced by conditioning.

The rotation of the ellipsoid and the changes in sizes of the principal axis are more subtle and sophisticated measures. They quantify changes in the likelihood of relative variations within the stressed subset of variables. Let me first notice that the principal axis of the ellipsoid are directed along the eigenvectors of the scale matrix (or the covariance matrix, when defined) and their sizes are the square roots of the eigenvalues. The eigenvalues and eigenvectors of the conditional and unconditional scale matrices do not coincide either in size or in direction.

The most elongated axis of the ellipsoid points in the direction of the eigenvector associated with the largest eigenvalue of the scale matrix. This is the linear combination of variables that carries the largest variance. It is therefore an important quantity because it carries the greatest risk. Let me denote the two largest eigenvalues with $\hat{\lambda}_Y^{1/2}$ and $\hat{\lambda}_{Y|X}^{1/2}$ respectively for the unconditioned and conditioned cases, while I denote the associated eigenvectors respectively with \hat{u}_Y and $\hat{u}_{Y|X}$. These eigenvectors are both unitary in the module and therefore the change is a rotation by the angle:

$$\Theta_{X \rightarrow Y} = \arccos(\hat{u}_Y^\top \hat{u}_{Y|X}). \quad (6.54)$$

This rotation of the axis of the largest elongation of the equiprobability ellipsoidal surface provides information on the relative increase or decrease of risk associated with each variable. Under conditioning, some variables might take larger weights than in unconditioned situations highlighting systemic fragility. The larger the rotation the larger is the disruptive effect of conditioning on the variables. An application of conditioned probability to stress testing was explored in the paper Aste [2021] where this reasoning and these measures were applied to financial systems.

For all these conditional distributions, for the entire elliptical family, the expected values are

$$\mathbb{E}[Y|X = x] = \mu_{Y|x}. \quad (6.55)$$

Instead, the conditional covariance (when defined) depends on the distribution. Hereafter, I provide some more details specific to the multivariate normal and the multivariate Student-t distributions.

6.7.1 Conditional distribution for the multivariate normal

In the multivariate normal case, the density generator function is

$$g(\tilde{d}^2) = \exp(-\tilde{d}^2/2). \quad (6.56)$$

This form is identical to the standard monovariate normal distribution with the squared Mahalanobis distance in place of the univariate variable. This implies that the cumulative probability also coincides. For instance, as depicted in Fig.5.1 the probability of observations further than Mahalanobis distance $d^2 = 1$ is about

32% (equivalent to one standard deviation) while $d^2 = 4$ (equivalent to two standard deviations) is about 4.6%.

For what concerns that conditional distribution, this normal case is particularly simple because the exponential form of the density generator function makes the additive constant term in $d^2 = d_{\mathbf{Y}|\mathbf{x}}^2 + d_{\mathbf{x}}^2$ become an irrelevant multiplicative constant, giving

$$f(\mathbf{Y}|\mathbf{X} = \mathbf{x}) = k_{p_Y} |\Sigma_{\mathbf{YY}|\mathbf{x}}|^{-1/2} \exp(-d_{\mathbf{Y}|\mathbf{x}}^2/2) \quad (6.57)$$

which is a multivariate normal with expected values

$$\mu_{\mathbf{Y}|\mathbf{x}} = \mu_{\mathbf{Y}} + \Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1} (\mathbf{x} - \mu_{\mathbf{x}}), \quad (6.58)$$

and covariance

$$\Sigma_{\mathbf{YY}|\mathbf{x}} = \Sigma_{\mathbf{YY}} - \Sigma_{\mathbf{YX}} \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{XY}}. \quad (6.59)$$

In another notation:

$$\mathbf{Y}|\mathbf{x} \sim \mathcal{N}(\mu_{\mathbf{Y}|\mathbf{x}}, \Sigma_{\mathbf{YY}|\mathbf{x}}). \quad (6.60)$$

Notably, the conditional covariance is different from the unconditional, it is shaped by the correlations between the conditioning and conditional variables but it does not depend on the value of the conditioning variable $\mathbf{X} = \mathbf{x}$.

Example 6.4 (Conditional probability density function for the bivariate normal). By using the previous formulas one can verify that the conditional $f(Y = y|X = x)$ probability density function for the **bivariate normal** is a univariate normal distribution with mean

$$\mu_{Y|x} = \mu_Y + \rho_{XY} \frac{\sigma_Y}{\sigma_X} (x - \mu_X)$$

and

$$\sigma_{Y|x}^2 = \sigma_Y^2 (1 - \rho_{XY}^2).$$

6.7.2 Conditional distribution for the multivariate Student-t

For the multivariate Student-t case, one has

$$g(\tilde{d}^2) = (1 + \tilde{d}^2/\nu)^{-\frac{\nu+p}{2}}. \quad (6.61)$$

Also in this case, as for the multivariate normal, the form of the density is identical to the univariate Student-t with \tilde{d}^2 taking the place of x^2 . For instance, as depicted in Fig.5.7, when $\nu = 1$, observations outside the Mahalanobis ellipsoidal surface with $\tilde{d}^2 = 1$ have a probability of 50%. However, this depends on the degrees of freedom and, for instance, for $\nu = 2$, this probability reduces to about 42%, while when $\nu \rightarrow \infty$ this probability becomes the same as for the normal distribution at about 32%.

Conditioning transforms the multivariate Student-t is another Student-t. Indeed, the additive constant term in $\tilde{d}^2 = \tilde{d}_{\mathbf{Y}|\mathbf{x}}^2 + \tilde{d}_{\mathbf{x}}^2$ can be handled so to keep the formula in the same form, obtaining

$$f(\mathbf{Y}|\mathbf{X} = \mathbf{x}) = k_{p_Y} |\tilde{\Omega}_{\mathbf{YY}|\mathbf{x}}|^{-1/2} \left(1 + \frac{\hat{d}_{\mathbf{Y}|\mathbf{x}}^2}{\nu_{\mathbf{Y}|\mathbf{x}}} - \frac{\nu_{\mathbf{Y}|\mathbf{x}} + p_Y}{2}\right)^{-\frac{\nu_{\mathbf{Y}|\mathbf{x}} + p_Y}{2}}. \quad (6.62)$$

with

$$\nu_{\mathbf{Y}|\mathbf{x}} = \nu + p_{\mathbf{x}}, \quad (6.63)$$

and

$$\hat{d}_{\mathbf{Y}|\mathbf{x}}^2 = \frac{\nu + p_{\mathbf{x}}}{\nu + \tilde{d}_{\mathbf{x}}^2} \tilde{d}_{\mathbf{Y}|\mathbf{x}}^2. \quad (6.64)$$

Which, when the covariance is defined ($\nu > 2$) is equal to $\hat{d}_{\mathbf{Y}|\mathbf{x}}^2 = \frac{\nu + p_{\mathbf{x}}}{\nu + \tilde{d}_{\mathbf{x}}^2} \frac{\nu - 2}{\nu} d_{\mathbf{Y}|\mathbf{x}}^2$.

Note that, this expression for $\hat{d}_{\mathbf{Y}|\mathbf{x}}^2$ comes from the fact that the two terms

$$\bar{k}(\nu + \tilde{d}_{\mathbf{Y}|\mathbf{x}}^2 + \tilde{d}_{\mathbf{x}}^2)^{-\frac{\nu+p}{2}}, \quad (6.65)$$

and

$$\tilde{k}(\nu_{\mathbf{Y}|\mathbf{x}} + \hat{d}_{\mathbf{Y}|\mathbf{x}}^2)^{-\frac{\nu_{\mathbf{Y}|\mathbf{x}} + p_Y}{2}}, \quad (6.66)$$

must be identical. Consequently, by comparing the exponents, follows

$$\nu_{\mathbf{Y}|\mathbf{x}} = \nu + p - p_Y = \nu + p_{\mathbf{x}} \quad (6.67)$$

while by comparing the terms in the parenthesis one must have $\nu + \tilde{d}_{\mathbf{Y}|\mathbf{x}}^2 + \tilde{d}_{\mathbf{x}}^2 \propto \nu_{\mathbf{Y}|\mathbf{x}} + \hat{d}_{\mathbf{Y}|\mathbf{x}}^2$, which can be obtained by imposing

$$\hat{d}_{\mathbf{Y}|\mathbf{x}}^2 = \frac{\nu + p_{\mathbf{x}}}{\nu + \tilde{d}_{\mathbf{x}}^2} d_{\mathbf{Y}|\mathbf{x}}^2. \quad (6.68)$$

The previous expression for the conditional density function $f(\mathbf{Y}|\mathbf{X} = \mathbf{x})$ is clearly still a multivariate Student-t. The expectation values are $\mu_{\mathbf{Y}|\mathbf{x}}$ (see Eq.6.50) and this coincides with the shift for the normal case and for the whole elliptical family distribution. The scale matrix also changes similarly to the normal term (see Eq.6.51) but it also acquires a multiplicative term that is dependent on the values of the conditioning variable:

$$\tilde{\Omega}_{\mathbf{YY}|\mathbf{x}} = \frac{\nu + \tilde{d}_{\mathbf{x}}^2}{\nu + p_{\mathbf{x}}} (\Omega_{\mathbf{YY}} - \Omega_{\mathbf{YX}} \Omega_{\mathbf{XX}}^{-1} \Omega_{\mathbf{XY}}). \quad (6.69)$$

The degrees of freedom also become change increasing by the number of the conditioning variables ($p_{\mathbf{x}}$). Notice that therefore when conditioning to several variables (i.e. $p_Y \gg 1$) the conditioned distribution becomes indistinguishable from a normal distribution.

When defined, the conditional covariance is

$$\tilde{\Sigma}_{\mathbf{YY}|\mathbf{x}} = \frac{\nu + \tilde{d}_x^2}{\nu + p_x - 2} (\Omega_{\mathbf{YY}} - \Omega_{\mathbf{YX}} \Omega_{\mathbf{XX}}^{-1} \Omega_{\mathbf{XY}}). \quad (6.70)$$

This is proportional to the expression for the conditional covariance for the multivariate normal distribution, however, the term \tilde{d}_x^2 in the additional coefficient in front is now a function of the values of the conditioning variables \mathbf{x} .

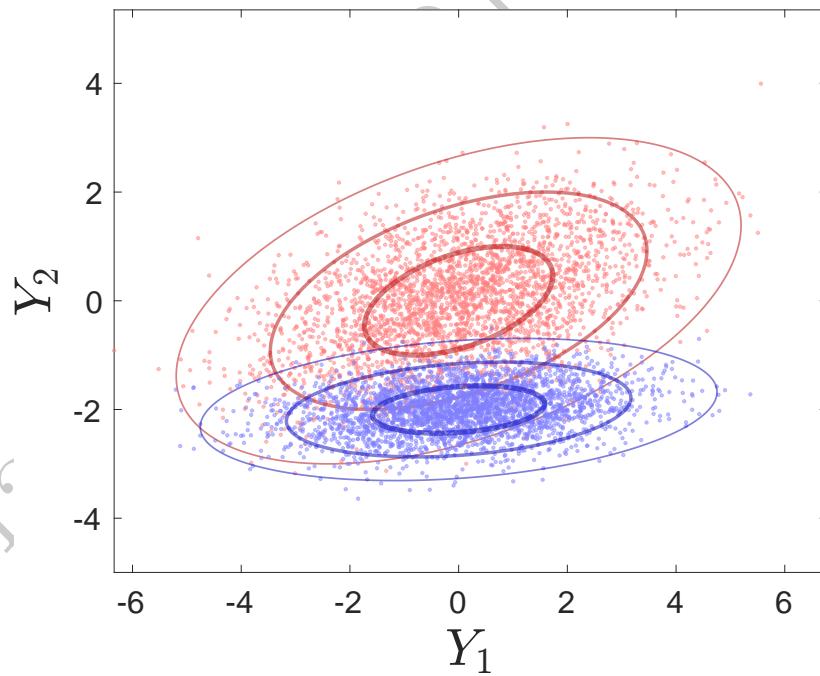


Figure 6.2 Exemplification of the effect of conditioning for an elliptical multivariate distribution of three variables X, Y_1, Y_2 (multivariate normal). (Red dots) scatter plot of 3,000 observations drawn from the unconditioned marginal distribution of Y_1, Y_2 . (Blue dots) scatter plot of 3,000 observations drawn from a conditioned distribution of Y_1, Y_2 when $X = -2$. One can notice that the centroids have shifted by $\Lambda_{\mathbf{Y}|\mathbf{x}} = (0, -2)$, the ellipsoid have rotated by an angle $\Theta_{\mathbf{X} \rightarrow \mathbf{Y}} = 15.16$ degrees in the clockwise direction, and it shrunk with the largest eigenvalue passing from 3.28 to the value of 2.52. The lines are equiprobability regions.

Example 6.5 (Stress testing with multivariate conditional elliptical probabilities.). Systemic risk, in a complex system with several interrelated variables, such as a financial market, is quantifiable via the conditional multivariate probability distribution describing the reciprocal influence between the system's variables. The effect of stress on the system is reflected by the

change in the conditional probability with respect to the unconditioned. The conditioning variable set, \mathbf{X} , is the set of ‘stressing’ variables while the conditioned variable set, \mathbf{Y} , are the ‘stressed’ variables. The conditional probability distribution function $f(\mathbf{Y}|\mathbf{X} = \mathbf{x})$ can provide a full quantification of the propagation of stress induced by stressing the variables \mathbf{X} to extreme values. The conditional probability distribution provides indeed the complete knowledge about the effects of the stressing variables $\mathbf{X} = \mathbf{x}$ on the statistics of the stressed variables \mathbf{Y} .

Suppose, for instance, that \mathbf{X} are the returns of sets of assets in a given industry sector, say the changes in the prices of oil and gas ($\mathbf{X} \in \text{Oil \& Gas}$). Through $f(\mathbf{Y}|\mathbf{X} = \mathbf{x})$, one can compute how a strong negative change in the prices of such commodities ($\mathbf{X} = \mathbf{x} < 0$) impacts on the changes in values in another sector, such as, for instance, consumer goods ($\mathbf{Y} \in \text{Consumer Goods}$).

I argue that for the purpose of induced stress quantification, the most important effect is the centroid’s shift, $\Delta_{\mathbf{Y}|\mathbf{x}}$ (see Eq.6.53). For instance, in the example I mentioned about the effect of losses in oil and gas on returns in consumer goods, this shift quantifies the average losses in the consumer goods sector caused by losses in the oil and gas sector. In most cases, this shift is the largest effect because conditioning always reduces uncertainty. Rotation of the equiprobability ellipsoid and the changes in the sizes of the principal axis are other important effects that impact the relative propagation of the risk in the subset of stressed variables.

The whole family of elliptical distributions shares the same shape matrices that might differ only by a scalar factor. Therefore, these measures are general measures of risk independent from modeling details.

Let me demonstrate the effect of conditioning on the probability density function with a practical example for three multivariate normal variables X and $\mathbf{Y} = (Y_1, Y_2)$ with zero means $\boldsymbol{\mu} = (0, 0, 0)^\top$ and covariance

$$\Sigma_{\mathbf{Z}\mathbf{Z}} = \begin{pmatrix} 1.0 & 0.7 & 0.9 \\ 0.7 & 3.0 & 0.8 \\ 0.9 & 0.8 & 1.0 \end{pmatrix}. \quad (6.71)$$

where I denoted $\mathbf{Z} = (X, \mathbf{Y})$. The unconditioned covariance for \mathbf{Y} is

$$\Sigma_{\mathbf{YY}} = \begin{pmatrix} 3.0 & 0.8 \\ 0.8 & 1.0 \end{pmatrix} \quad (6.72)$$

while, using Eq.6.51 (see also Eq.6.59 for the normal case) the conditioned one is equal to

$$\Sigma_{\mathbf{YY}|\mathbf{x}} = \begin{pmatrix} 2.51 & 0.17 \\ 0.17 & 0.19 \end{pmatrix} \quad (6.73)$$

Fig. 6.2 reports with red dots the scatter plot of 3,000 observations drawn from the marginal distribution $f_{\mathbf{Y}}(\mathbf{Y})$. The figure reports also with red

lines the equiprobability ellipsoids corresponding to Mahalanobis distances $d_{\mathbf{Y}} = 1, 2$ and 3 . The draw of 3,000 observations from the conditional distribution $f_{\mathbf{Y}|x}(\mathbf{Y}|X = -2)$ is reported in the same figure with blue dots and blue lines. One can visually see that the blue point cloud has shifted, rotated, and compactified with respect to the unconditional red one. The centroids' shift is: $\Lambda_{\mathbf{Y}|x} = (0, -2)$. The rotation angle is $\Theta_{\mathbf{x} \rightarrow \mathbf{y}} = 15.16$ degrees in the clockwise direction while the largest eigenvalue passes from a value of 3.28 for the unconditioned covariance to the value of 2.52 for the conditioned covariance. Therefore the riskiest unconditioned portfolio has a variance 3.28 but by conditioning, even under stress, the largest variance is 2.52. Indeed, the losses in the stress scenario are all carried by the shift of -2 in the expected values. Notably, all these considerations are equally valid for any multivariate elliptical probability and the Student-t in particular. However, for non-normal models, the values of the probabilities change. For instance, assuming a Student-t model with the same covariance one would have the same shift of the centroids as for the normal case but the Mahalanobis distance terms have extra factors in front, namely $\frac{\nu-2}{\nu}$ for the unconditional and instead $\frac{\nu+p_{\mathbf{x}}}{\nu+d_{\mathbf{x}}^2} \frac{\nu-2}{\nu}$ for the conditional. The term $\tilde{d}_{\mathbf{x}}^2$ has expected value $\mathbb{E}(\tilde{d}_{\mathbf{x}}^2) = p_{\mathbf{x}}$ and it can vary in the range $[0, \infty)$; it can therefore have a very large impact on the conditional probability distribution. In the present example, the conditioning variable has zero mean and value $x = -2$, while $\Sigma_{\mathbf{xx}} = 1$ and therefore $d_{\mathbf{x}}^2 = 4$.

An application of this stress testing qualification to a financial system is discussed in the paper Aste [2021].

6.8 Data-driven modeling tutorial

<https://github.com/FinancialComputingUCL/DataDrivenModeling/Ch6>

The tutorial for this Chapter covers various topics on multivariate probabilities including: the relationship between the covariance matrix and the scale matrix; comparison between marginal and conditional probability distribution functions; and examples of the use of elliptical distribution family for stress testing (Example 6.5).

Exercises

- Derive the correlation coefficient of two random variables X and Y which are related by $Y = -0.1X + 3$.
- Derive the expression for the correlation coefficient of two random variables

related by $Y = \beta X + \epsilon$ with ϵ a random noise with zero mean and unitary standard deviation.

- Compute the correlation coefficient between the random variables X and $|X|$ assuming X is defined in the hole domain $(-\infty, +\infty)$, it has a symmetric probability density function with mean $\mu_X = 0$.
- Derive the expression for the probability density function for the sum of two Student-t multivariate random variables X_1 and X_2 with means μ_X, μ_Y , degrees of freedom ν and scale matrix

$$\Omega_{\mathbf{xx}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \quad (6.74)$$

- By using the discussion in Remark 6.3 derive the degrees of freedom of a sum of p multivariate Student-t variables with degrees of freedom ν .
- By using Bayes' formula, demonstrate that the marginal probability is the expected value of the conditional probability computed over the conditioning variable

$$p_X(x) = \mathbb{E}(p_{XY}(x|y))_Y. \quad (6.75)$$

- Discuss if a conditioned variable, $Y|\mathbf{X} = \mathbf{x}$, can have larger variance $\sigma_{Y|\mathbf{x}}^2$ than the unconditioned counterpart σ_Y^2 .

Entropies

The concept of entropy emerged at the time of steam engines when scientists were engaged in finding the relation between heat, work, and energy. They soon realized that not all the thermal energy in a system can be transformed into mechanical work. Indeed, there is an extra term, proportional to the temperature, a quantity that Clausius named entropy [Clausius, 1879]. Later, the microscopical interpretation of the laws of thermal systems provided by Boltzmann's statistical mechanics [Boltzmann, 1868] made it clear that the entropy is associated with the statistical properties of the heated matter and measures the degree of disorder or randomness in the system. Finally, it emerged that structural disorder and information are two perspectives on the same subject and entropy quantifies to the amount of uncertainty on the microscopic states associated with a macrostate of a thermal system.

From a probabilistic modeling perspective, entropy quantifies uncertainty: the less we know about the possible value of a random variable the larger its entropy is. Entropy is a very important concept in physics but it is also central in information theory. In these two fields, entropy was introduced by following very different paths, however, in both fields, the entropy is a measure of uncertainty. In physics, in the statistical mechanics' definition, the entropy is proportional to the logarithm of the number of microscopic configurations that a system can access

$$S = k_B \log W \quad (7.1)$$

where W is the volume of the accessible configuration space and k_B is the Boltzmann's constant. The larger W is, the greater is the uncertainty to find the system in a given configuration.

7.1 Shannon entropy

In information theory, the concept of entropy was introduced by Shannon while studying the 'lost' information in telephone communications [Shannon, 1948]. Shannon is indeed considered the father of information theory.

Definition 7.1 (Shannon entropy). For a discrete probability distribution,

$P(X)$, positively defined over the support Ω_X , the Shannon entropy is:

$$H(X) = -\mathbb{E}(\log P(X)) = - \sum_{x \in \Omega_X} P(x) \log(P(x)). \quad (7.2)$$

Analogously, for continuous variables with probability density function $f(x) > 0$ in Ω_X , the Shannon entropy is:

$$H(X) = -\mathbb{E}(\log f(X)) = - \int_{\Omega_X} f(x) \log f(x) dx. \quad (7.3)$$

Remark 7.1. Note that the definition requires the probabilities to be positive. This is always possible as far as the support, Ω_X is restricted to non-zero values of $P(X)$ or $f(X)$.

Example 7.1. For the normal distribution (see Definition 5.1) the Shannon entropy is the logarithm of the standard deviation plus a constant:

$$H(X) = -\mathbb{E}(\log f(X)) = \log \sigma + \frac{1}{2} + \frac{1}{2} \log(2\pi) . \quad (7.4)$$

Showing that, indeed, the entropy increases when the distribution is broader and the uncertainty on the possible value of a draw of the random variable is larger.

The Student-t (see Definition 5.12) has

$$H(X) = \log \tilde{\sigma} + \frac{\nu+1}{2} \left(\psi \left(\frac{\nu+1}{2} \right) - \psi \left(\frac{\nu}{2} \right) \right) + \log \left(\sqrt{\nu} B \left(\frac{\nu}{2}, \frac{1}{2} \right) \right) \quad (7.5)$$

with $\psi(\cdot)$ the digamma function and $B(\cdot)$ the beta function.

For the log-normal distribution (see Definition 5.7) it is

$$H(X) = \log \sigma + \log \left(e^{\tilde{\mu} + \frac{1}{2}} \sqrt{2\pi} \right) . \quad (7.6)$$

For a continuous uniform distribution ($f(x) = 1/(b-a)$ for $x \in [a, b]$ and zero elsewhere), the Shannon entropy is the logarithm of the width of the interval.

$$H(X) = \log(b-a) . \quad (7.7)$$

One can see that the Shannon entropy in general increases as the logarithm of the width of the distribution. It is a measure of the broadness of the uncertainty associated with random draws from a population with a given probability distribution.

In general, the Shannon entropy is a measure of the broadness of the probability distribution. When the base two logarithm is used, then the entropy, $H(X)$,

quantifies the information content in terms of the number of bits, and it is referred to as Shannon entropy. For discrete variables, given a sequence of q values, $\hat{x}_1, \dots, \hat{x}_q$, of the random variable X , the Shannon entropy is the number of alternative yes/no questions per unit length that are necessary to define the values of the variables in the sequence. Again we see that $H(X)$ is a measure of uncertainty: the more unpredictable is the sequence, the larger is the entropy.

Remark 7.2. We have so far two definitions for entropy: Boltzmann's definition

$$S = k_B \log W \quad (7.8)$$

and Shannon's definition:

$$H(X) = -\mathbb{E}(\log P(X)). \quad (7.9)$$

These two definitions coincide (apart for the constant k_B) because in statistical mechanics the microstates are assumed to be equiprobable and therefore $P(\text{system state} = k) = 1/W$ and $H(X) = -\mathbb{E}(\log P(\text{system state} = k)) = \log W$. Indeed, Shannon's entropy was introduced in physics by Gibbs several years before Shannon.

The Shannon entropy was introduced as a quantification of information that can be transmitted by a signal. One can figure this out by considering a time series of T observations of a variable X that can have a number, A , of discrete values. Or – analogously – let's consider a “message” of length T written with an “alphabet” containing a number A of characters.

There are T positions where the first letter can be placed, $T - 1$ positions where the second letter can go, etc. until the last, T^{th} , a letter which has only one position left. This is a number of combinations equal to:

$$T(T - 1)(T - 2)\dots 1 = T!. \quad (7.10)$$

However, in a ‘message’ some ‘characters’ are repeated several times and the exchange of two identical characters in different places does not change the sentence, therefore for each character, \hat{x}_i ($i = 1..A$) repeated n_i times we must consider that $n_i!$ combinations are identical and do not carry any extra information. Consequently, when we have repetitions, we must divide the total number of combinations by $n_i!$ for each i . It results that the total number of different possible combinations in a signal of length T containing a number n_1 of character \hat{x}_1 , a number n_2 of character \hat{x}_2, \dots and a number n_A of character \hat{x}_A is:

$$W = \frac{T!}{n_1! n_2! \dots n_A!}. \quad (7.11)$$

One can, for instance, see that by using only one kind of character \hat{x}_i

(i.e. $n_i = T$ and $n_k = 0$ for all $k \neq i$), then this number is $W = 1$ (no information). Conversely, the maximum is achieved when all characters in the signal are different and all $n_i = 1$.

If we measure the information in terms of bits, G , we have

$$W = 2^G \quad (7.12)$$

which means

$$G = \log_2 \frac{T!}{n_1! n_2! \cdots n_A!}. \quad (7.13)$$

We can now use Stirling's approximation for the log-factorial, $\log_2 n! \simeq n \log_2 n - n$ (which is quite accurate for large enough n), obtaining:

$$G \simeq - \sum_{k=1}^A n_k \log_2 \frac{n_k}{T}. \quad (7.14)$$

One might recognize that n_k/T is the relative frequency of occurrence of the character \hat{x}_k in the signal, and it tends to the probability $P(\hat{x}_k)$ in the limit of an infinitely long signal (law of large numbers, see Section 13.3). Therefore, for large T , the number of bits G in a signal of length T is

$$G \simeq -T \sum_{k=1}^A P(\hat{x}_k) \log_2 P(\hat{x}_k). \quad (7.15)$$

which is the Shannon entropy multiplied by T . This conversely implies that the Shannon entropy (when computed in base-2 logarithm) is the number of bits per unit length of a signal:

$$H(X) = \frac{G}{T}. \quad (7.16)$$

This quantity is the (maximum) amount of information that can be encoded into a signal with a given probability distribution, $P(\hat{x}_k)$ of occurrence of characters \hat{x}_k for $k = 1, \dots, A$. Notice that, the more the signal is random, the larger the amount of information that it can carry. Indeed, the maximum of $-P(\hat{x}_k) \sum_k \log P(\hat{x}_k)$ is achieved by the uniform distribution with $P(\hat{x}_k) = 1/T$ for all k .

Entropy is a measure of information: it accounts for the maximum number of bits per unit length that a signal X with probability distribution $f(X)$ can transmit.

Entropy is a measure of uncertainty: the larger the entropy, the smaller the possibility to predict the next output in a data series.

Entropy is maximal when the distribution is uniform $f(X) = p_0$ (unless other constraints are imposed, see Section 7.2). This indeed corresponds to maximum uncertainty: when all values of X are equally likely to be observed, the prediction of the next value in a sequence is the hardest.

Example 7.2. Let me discuss a bit more with this example what it means that entropy is a measure of uncertainty. Let's suppose that I think of a number, any number, between 1 and 100. The entropy is the logarithm of the number of questions someone must ask me before finding the right number. The larger the entropy, the harder is to guess the number.

Let me discuss this in detail. There are several ways to ask questions, for instance, if one tries to guess the number directly by listing the number one by one in a sequence such as 1, 2, 3,... the discovery process might require up to 99 questions. However, by asking the right questions it would take at most 7 questions to discover the right number. Indeed, if the choice is between n numbers, one can proceed by dividing first the set into two subsets and ask in which of the two subsets the number is (i.e. smaller or equal to 50?) and then continue by bisecting the range in which the number is until all possibilities are explored. If H questions are asked, the number of possibilities explored is 2^H . Consequently, if one had to guess between n possibilities, the number of steps needed is $H = \lceil \log_2 n \rceil$ which, in the case of $n = 100$, is $H = 7$. This number is strictly related with the Shannon entropy which, indeed, it is $H = -\mathbb{E}(\log p) = -\log p$, where, in this case, p is uniform and equal to $p = 1/n$ because all numbers between 1 and n have equal probability and therefore $H = \log n$. The base-2 logarithm makes the units measured of the Entropy in bits.

The result $H = 7$ has been obtained under the assumption that the probability distribution of the number to guess is uniform. If instead, the numbers between 1 and 100 were not equiprobable, for instance, if the number to guess is the age of an individual in a classroom, that in principle can be between 0 and 100 but in practice, in a classroom, is highly likely to be peaked around one age value only, then the entropy will be smaller because there is less uncertainty and it is easier to guess the answer with a smaller number of questions.

7.2 The maximum entropy principle

The maximum entropy principle is a tool from statistical physics that can be very useful, in some cases, to discover the functional form of the probability distribution. It states that the best probability distribution is the one with the largest entropy under some conditions provided by the knowledge we have of the system. The conditions typically are on some expectation values (e.g. distribution with a given mean and variance), some symmetry properties, and the support interval. One can see this principle as advocating ‘maximal ignorance’, in other words, one adopts the distribution that maximizes the uncertainty about what it is not known. Within the Bayesian perspective, the principle of maximum entropy is often advocated to set the uniform prior probability.

Example 7.3. Let me provide a few examples of distributions emerging from the maximum entropy principle for some given conditions. I restrict the examples to continuous variables. Therefore, I search for the function $f(X)$ that maximizes $H = -\mathbb{E}(\log f(X))$ under the desired conditions. Two conditions are common to all, namely positivity $f(x) > 0$ and normalization $\int_{\Omega_X} f(x)dx = 1$.

- By imposing the condition of being defined in a **finite support** $x \in [a, b]$ the **uniform** distribution maximizes entropy:

$$f(x) = \frac{1}{b-a}. \quad (7.17)$$

- By conditioning the **mean** $\mathbb{E}(X) = \mu$ and the **support** $x \geq 0$ one obtains that entropy is maximized by the **exponential** density function:

$$f(x) = \frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right). \quad (7.18)$$

- By conditioning the **mean** $\mathbb{E}(X) = \mu$ and **variance** $\mathbb{E}((X - \mu)^2) = \sigma^2$ one obtains that entropy is maximized by the **normal** density function:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \quad (7.19)$$

Note that all distributions can be described as maximum entropy solutions under some conditions. For instance, the Student-t maximizes entropy under the condition $\mathbb{E}(\ln(\nu + X^2)) = constant$; however, the meaning of conditions like this one is hard to find.

7.3 Joint entropy in multivariate systems

The concept of entropy, introduced in the previous sessions for the univariate case can be directly extended to any dimension. The Shannon entropy, in particular, becomes

$$H(\mathbf{X}) = -\mathbb{E}_{\mathbf{X}}(\log f(\mathbf{X})), \quad (7.20)$$

where $\mathbf{X} \in \mathbb{R}^{p \times 1}$ is a p-dimensional multivariate random variable. Entropy is a measure of how broad the distribution is and, in multidimensional spaces, this can be quantified by the volume of the accessible configuration space. Which is indeed the original Boltzmann's definition of entropy (see Remark 7.2). The Shannon entropy of several variables quantifies the global uncertainty associated with the set of variables and it is called joint entropy.

7.3.1 Entropy of the multivariate normal

One can compute directly from Eq.7.20 the Shannon entropy for the multivariate normal obtaining:

$$H(\mathbf{X}) = \frac{1}{2} \log(|\Sigma_{\mathbf{XX}}|) + \frac{p}{2} + \frac{p}{2} \log(2\pi) \quad (7.21)$$

where $|\Sigma_{\mathbf{XX}}|$ is the determinant of the covariance matrix. It is worth pointing out that the determinant of a $p \times p$ covariance matrix can be indeed interpreted in terms of a volume of a simplex in p dimensions (see Remark 7.3 below).

Remark 7.3. A simplex is a generalization to arbitrary dimensions of the triangle in two dimensions and the tetrahedron in three dimensions (see Definition 4.12). A k -simplex is a k -dimensional object with $k+1$ vertices that are all connected through edges. The volume of a k -simplex with vertices $\mathbf{v}_0, \mathbf{v}_1, \dots, \mathbf{v}_k$ is:

$$\text{k-Simplex Volume} = \frac{|\mathbf{D}|}{k!} \quad (7.22)$$

where, $\mathbf{D} \in \mathbb{R}^{k \times k}$ is a matrix whose elements $(\mathbf{D})_{i,j} = v_{i,j}$ the components of the vectors: $v_{i,j}$ component j of vector \mathbf{v}_i . For the entropy of multivariate normal case, the term $|\Sigma_{\mathbf{XX}}|$ is the volume of a $(p-1)$ -simplex with vertices $\mathbf{v}_i = (\text{Cov}(X_i, X_1), \text{Cov}(X_i, X_2), \dots, \text{Cov}(X_i, X_p))$.

7.3.2 Entropy of the multivariate elliptical family

A very similar expression to the one for the entropy of the multivariate normal distribution (Eq.7.21) holds the broader family of elliptical distribution:

$$H(\mathbf{X}) = \frac{1}{2} \log(|\Omega_{\mathbf{XX}}|) + H_{\mathbf{X}_0} \quad (7.23)$$

where $\Omega_{\mathbf{XX}}$ is the scale matrix and $H_{\mathbf{X}_0}$ is the entropy associated with a vector of standardized random variables with zero expected values and the identity as shape matrix. This term is therefore independent of the location vectors and the shape matrices of \mathbf{X} and depends only on the number of variables and kind of distribution.

For instance, for the multivariate Student-t distribution for p variables and with ν degrees of freedom one has

$$H_{\mathbf{X}_0} = -\log \frac{\Gamma(\frac{\nu+p}{2})}{\Gamma(\frac{\nu}{2})(\nu\pi)^{\frac{p}{2}}} + \left(\frac{\nu+p}{2} \right) \left(\psi\left(\frac{\nu+p}{2}\right) - \psi\left(\frac{\nu}{2}\right) \right) \quad (7.24)$$

with $\psi(\cdot)$ the digamma function. One can see that in this case, $H_{\mathbf{X}_0}$ depends on p and ν .

Example 7.4 (Determinants). Let me write explicitly the determinants for low dimensional cases.

- For only one variable $\mathbf{X} = X_1$ then

$$|\boldsymbol{\Omega}_{\mathbf{XX}}| = \sigma_1^2, \quad (7.25)$$

with σ_i^2 the variance of X_i ($i = 1$).

- For two variables $\mathbf{X} = (X_1, X_2)$ then

$$|\boldsymbol{\Omega}_{\mathbf{XX}}| = \sigma_1^2 \sigma_2^2 (1 - \rho_{1,2}^2), \quad (7.26)$$

with $\rho_{i,j}$ the correlation coefficient between X_i and X_j ($i, j = 1, 2$).

- For three variables $\mathbf{X} = (X_1, X_2, X_3)$ then

$$|\boldsymbol{\Omega}_{\mathbf{XX}}| = \sigma_1^2 \sigma_2^2 \sigma_3^2 (1 - \rho_{1,2}^2 - \rho_{1,3}^2 - \rho_{2,3}^2 + 2\rho_{1,2}\rho_{1,3}\rho_{2,3}). \quad (7.27)$$

These expressions will turn handy in later examples.

7.4 Kullback-Leibler divergence and the cross-entropy

Often modelers want to quantify how ‘distant’ a probabilistic model is from another. A measure of such distance is the Kullback-Leibler divergence.

Definition 7.2 (Kullback-Leibler divergence). For discrete variables the **Kullback-Leibler divergence (KLD)**, between two probability mass functions $P(x) > 0$ and $Q(x) > 0$, both defined over the support Ω_X , is

$$D_{\text{KL}}(P \parallel Q) = \sum_{x \in \Omega_X} P(x) \log \frac{P(x)}{Q(x)}. \quad (7.28)$$

Notice that, if $Q(x) = 0$ the Kullback-Leibler divergence can be still defined if $P(x) = 0$.

Analogously, for continuous variables, the KLD between two probability density functions $f(x) > 0$ and $q(x) > 0$, both defined over the support Ω_X , is

$$D_{\text{KL}}(f \parallel q) = \int_{\Omega_X} f(x) \log \frac{f(x)}{q(x)} dx. \quad (7.29)$$

The KLD is always non-negative and it is equal to zero only when the probabilities are identical (i.e. $P(x) = Q(x)$ or $f(x) = q(x)$). The KLD is asymmetric in the two probability distributions.

The fact that KLD is non-negative for any couple of models $f(x)$ and $q(x)$ is easy to retrieve from the Gibbs’ inequality (i.e. by using $\log h(x) \leq h(x) - 1$,

for $h(x) > 0$, which implies

$$\begin{aligned} D_{\text{KL}}(f \parallel q) &= \int_{\Omega_X} f(x) \log \frac{f(x)}{q(x)} dx = - \int_{\Omega_X} f(x) \log \frac{q(x)}{f(x)} dx \\ &\geq - \int_{\Omega_X} \left(f(x) \frac{q(x)}{f(x)} - f(x) \right) dx = \int_{\Omega_X} f(x) dx - \int_{\Omega_X} q(x) dx = 0. \end{aligned} \quad (7.30)$$

Let me consider a case when $q(x) = \tilde{f}(x)$ is a model for the probability density function of a random variable X which has true probability density function $f(x)$ (the true distribution). In this case, the KLD measures the amount of information lost when the model $\tilde{f}(x)$ is used to approximate the true distribution $f(x)$. Good models must lose little information while bad models will lose large information. One might notice that the KLD is the difference between $\mathbb{E}_f(\log f(x))$ and $\mathbb{E}_f(\log \tilde{f}(x))$ (where I explicitly indicate that the expected value is over $f(x)$). In the Bayesian inference perspective, $\tilde{f}(x)$ is interpreted as the prior distribution and $f(x)$ the posterior, and the KLD is referred to as a relative entropy: the entropy of the prior $H_{\text{prior}} = -\mathbb{E}_f(\log \tilde{f}(x))$ minus the entropy of the posterior $H_{\text{posterior}} = -\mathbb{E}_f(\log f(x))$. However, one must note that $-\mathbb{E}_f(\log \tilde{f}(x))$ is not a Shannon entropy because the expectation value is taken over the reference distribution $f(x)$. This is indeed the so-called cross entropy (see Definition 7.4).

Definition 7.3 (*J*-divergence). There is a ‘symmetrized’ version of the KLD, which is called ***J*-divergence** $J(f, q) = D_{\text{KL}}(f \parallel q) + D_{\text{KL}}(q \parallel f)$ and it can be sometimes useful.

Example 7.5 (KLD for normal distributions). When the two distributions are normal respectively with expected values μ_f and μ_g and variances σ_f^2 and σ_g^2 , the KLD is

$$D_{\text{KL}}(f \parallel q) = \frac{1}{2} \left(\frac{\sigma_f^2}{\sigma_g^2} - \log \frac{\sigma_f^2}{\sigma_g^2} + \frac{(\mu_f - \mu_g)^2}{\sigma_g^2} - 1 \right). \quad (7.31)$$

One can see that both the means and the variances contribute to the KLD. It is also clear that $D_{\text{KL}}(f \parallel q) \geq 0$ (because $x - \log x \geq 1$) and the equality holds only when the two distributions are identical.

7.4.1 Kullback-Leibler divergence for multivariate systems

For multivariate systems, the concept stays identical to the univariate case and Kullback-Leibler divergence between the multivariate distributions $f(\mathbf{Z})$ and $q(\mathbf{Z})$ is

$$D_{\text{KL}}(f \parallel q) = \mathbb{E}_f(\log f(\mathbf{Z})) - \mathbb{E}_f(\log q(\mathbf{Z})), \quad (7.32)$$

Example 7.6 (KLD for multivariate normal distributions). When the two distributions are p -dimensional multivariate normals $f = \mathcal{N}(\boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f)$ and $q = \mathcal{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$, KLD is:

$$D_{\text{KL}}(f \| q) = \frac{1}{2} \left(\text{tr} (\boldsymbol{\Sigma}_q^{-1} \boldsymbol{\Sigma}_f) - \log \frac{|\boldsymbol{\Sigma}_f|}{|\boldsymbol{\Sigma}_q|} + (\boldsymbol{\mu}_q - \boldsymbol{\mu}_f)^\top \boldsymbol{\Sigma}_q^{-1} (\boldsymbol{\mu}_q - \boldsymbol{\mu}_f) - p \right). \quad (7.33)$$

As for the unidimensional case, both the means and the covariances contribute towards the KLD which is asymmetric, non-negative, and equal to zero only when the two distributions are identical.

Definition 7.4 (Cross entropy). The term

$$H(\mathbf{Z}|q) = -\mathbb{E}_f(\log q(\mathbf{Z})), \quad (7.34)$$

can be seen as an estimate of the Shannon entropy $H(\mathbf{Z}) = -\mathbb{E}_f(\log f(\mathbf{Z}))$ by using the model $q(\mathbf{Z})$. It is called **cross-entropy**.

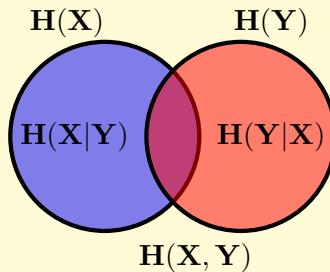
7.5 Conditional entropy

The entropy of a set of variables \mathbf{Y} conditioned to another set of variables, \mathbf{X} , is instead called conditional entropy. It quantifies the uncertainty on \mathbf{Y} remaining when the other variables \mathbf{X} are fully known. It is defined as the difference between the joint entropy $H(\mathbf{X}, \mathbf{Y})$ and the marginal entropy $H(\mathbf{X})$.

Definition 7.5. The **conditional entropy** is:

$$H(\mathbf{Y}|\mathbf{X}) = H(\mathbf{X}, \mathbf{Y}) - H(\mathbf{X}). \quad (7.35)$$

A pictorial representation is



It does not depend on the values of $\mathbf{X} = \mathbf{x}$ because the conditioning with respect to any value of \mathbf{X} is considered. The Shannon conditional entropy is:

$$H(\mathbf{X}|\mathbf{Y}) = -\mathbb{E}_{\mathbf{XY}}(\log f_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}|\mathbf{X})) \quad (7.36)$$

which, for two continuous variables is

$$H(X|Y) = - \int_{\Omega_X} \int_{\Omega_Y} f_{XY}(x, y) \log f_{Y|X}(y|x) dx dy. \quad (7.37)$$

Remark 7.4. Sometimes, the relation for the conditional entropy, or rather its equivalent expression

$$H(\mathbf{X}, \mathbf{Y}) = H(\mathbf{Y}|\mathbf{X}) + H(\mathbf{X}), \quad (7.38)$$

is called “chain rule”.

For the elliptical family probability distribution the Shannon conditional entropy is therefore:

$$H(\mathbf{X}|\mathbf{Y}) = \frac{1}{2} \log \frac{|\boldsymbol{\Omega}_{\mathbf{Z}\mathbf{Z}}|}{|\boldsymbol{\Omega}_{\mathbf{XX}}|} + H_{Z_0} - H_{X_0} \quad (7.39)$$

where $\mathbf{Z} = \mathbf{X} \cup \mathbf{Y}$.

7.6 Other entropies

Although the Shannon entropy is by far the most common and most used entropy, there are other measures of uncertainty that are also called entropies. In a general form, for an arbitrary discrete probability distribution with probabilities P_1, \dots, P_n , with $P_k > 0$ and $\sum_{k=1}^n P_k = 1$, entropy can be defined as a functional of the probabilities: $H(P_1, \dots, P_n)$ [Khinchin, 2013]. This is however a very generic definition and clearly some properties must be satisfied by the functional to be a ‘proper’ entropy.

Definition 7.6 (Shannon-Khinchin axioms). Shannon-Khinchin listed four conditions as requirements for an entropy functional:

1. $H_n(P_1, \dots, P_n)$ is continuous with respect to all its arguments;
2. $H_n(P_1, \dots, P_n)$ takes its largest value for the uniform distribution $P_k = 1/n$;
3. $H_{n+1}(P_1, \dots, P_n, 0) = H_n(P_1, \dots, P_n)$ the entropy does not change if a state with zero probability is added;
4. $H(A, B) = H(B|A) + H(A)$ (composition rule) the entropy of a system split into two subsystems A and B equals the entropy of A plus the entropy of B , conditional on A .

The Shannon entropy is the only functional that satisfies the four Shannon-Khinchin axioms. These axioms are very reasonable and the first three are considered desirable for any entropy measure. On the other hand, the universality of

the fourth axiom (the composition rule) is questionable and, if dropped, it opens the space to other entropies. One is the Rényi entropy [Rényi et al., 1961]

$$H_\alpha = \frac{1}{1-\alpha} \log \sum_{x \in \mathcal{X}} P(x)^\alpha \quad (7.40)$$

that can be seen as a generalization of the Shannon entropy which is retrieved in the limit $\alpha \rightarrow 1$.

Another generalization is the Tsallis' entropy [Tsallis 1988]

$$S_q = \frac{k}{q-1} \left(1 - \sum_{x \in \mathcal{X}} P(x)^q \right) \quad (7.41)$$

which again retrieves Shannon entropy in the limit $q \rightarrow 1$.

The application of the principle of maximum entropy to these alternative entropies provides a natural way to describe the emergence of non-normal distributions. For instance, Tsallis' entropy is associated with non-extensive formulations of the statistical mechanics and it is related to the emergence of Student-t distributions with tail exponents $\alpha = q$ (named Tallis distribution or q -gaussian in this context, see Tsallis [2009a]).

7.7 Data-driven modeling tutorial

<https://github.com/FinancialComputingUCL/DataDrivenModeling/Ch7>

The tutorial for this Chapter covers various topics on entropies including: comparisons between the values of Shannon entropies for various kinds of distributions (Example 7.1); comparisons between entropies and conditional entropies and their relation with mutual information and correlations.

Exercises

- Demonstrate that the uniform distribution maximizes entropy in the absence of other constraints.
- Compute the maximum entropy solution for the probability density function under the conditions: $E(\ln(\nu + (X - \mu)^2 / \sigma^2)) = c$, $f(x) > 0$ and $\int_{-\infty}^{+\infty} f(x) dx = 1$.
- Derive the expression (E.7.6) for the KLD of two multivariate normal distributions.
- Discuss if a conditioned set of variables, $\mathbf{Y}|\mathbf{X} = \mathbf{x}$, can have larger entropy than the unconditioned counterpart \mathbf{Y}

8

Dependency

Our world is interconnected and events do not happen in isolation. For modeling purposes, it is essential to understand and quantify the kind, the strength and the significance of interdependency between variables. From a probabilistic perspective, it is rather intuitive that the mutual dependence between two random variables X and Y should be measurable from the ‘difference’ between the probability to observe them together and the probability to observe them separately. This is indeed at the basis of all measures of dependency that I shall discuss hereafter. Another, and equivalent, perspective is to look at dependency as the ability of a random variable (X) to describe another variable (Y). This is called regression, and the task is to find a function $g(\cdot)$ such that $Y = g(X) + \epsilon$. If the variability of the error, ϵ , is small with respect to the variability of Y , then this implies that X is able to describe a significant part of the variability of Y and therefore the two variables are dependent. Alternatively to the regression approach (and equivalently to it), one can look directly at information measures. If the conditional entropy, $H(Y|X)$, is small with respect to the entropy of Y ($H(Y)$) then it means that the knowledge of variable X reduces uncertainty on the variable Y and therefore the two variables are dependent. These three perspectives, 1. probabilistic; 2. regression and; 3. information theoretic; are equivalent and are indeed different ways to address the same problem. In this chapter I am discussing these three perspectives and their connections.

8.1 Independent variables

In order to discuss the dependency between variables let me start by defining when they are not-dependent. Variable Y is not dependent on variable X if and only if the conditional probability, $P(Y|X)$, is unaffected by the values of X and therefore it is equal to the marginal probability $P(Y)$. Bayes’ formula (Eq.6.35) implies that, equivalently, when two variables are independent the joint probability must be equal to the product of the marginals over the whole support.

Definition 8.1 (Independent variables). Two discrete random variables X and Y are **independent** if and only if:

$$P_{X,Y}(X, Y) = P_X(X)P_Y(Y). \quad (8.1)$$

Analogously two random continuous variables X and Y are independent if and only if:

$$f_{X,Y}(x,y) = f_X(x)f_Y(y). \quad (8.2)$$

This can also be expressed in terms of cumulative distributions:

$$F_{X,Y}(x,y) = F_X(x)F_Y(y). \quad (8.3)$$

In standard notation, two independent random variables X and Y are indicated as

$$X \perp\!\!\!\perp Y. \quad (8.4)$$

These identities for independence are far reaching because the biconditional connection (the if and only if) states that the identities are both criteria to establish independence and a consequence of independence. In this Chapter I will discuss several tools to establish and quantify different kinds of dependency between random variables. However, despite the variety of methodologies, all these methods are different applications or consequences of the previous identities.

Remark 8.1. All three identities in Definition 8.1 are equivalent and dependent variables violate these identities. The case with cumulative distributions can be interpreted in terms of the so called '**quadrant dependency**' which has an intuitive meaning. Specifically, we have two main scenarios:

- 1) Positive quadrant dependency, $F_{X,Y}(x,y) > F_X(x)F_Y(y)$ which tells us that "when X is large Y is also likely to be large";
- 2) Negative quadrant dependency, $F_{X,Y}(x,y) < F_X(x)F_Y(y)$ which instead tells us that "when X is large Y is likely to be small".

Remark 8.2. Extension of Eqs.8.1-8.3 to more than two variables is somehow straightforward and, in the notation of this book, one can simply write the previous identities using the vectorial representations changing X and Y into \mathbf{X} and \mathbf{Y} . There are however some subtle details to be taken into account. In the multivariate context, the concept of independence is extended between sets of variables. For instance, when two sets are independent then all subsets of the two sets are independent as well: if $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}$ which implies that for any $\mathbf{X}_i \subseteq \mathbf{X}$ and $\mathbf{Y}_j \subseteq \mathbf{Y}$ one has $\mathbf{X}_i \perp\!\!\!\perp \mathbf{Y}_j$.

8.2 Dependency and regression

Measuring dependency is equivalent to finding the map $g(X)$ of random variable X into random variable $Y \simeq g(X)$. This is called **regression** and I have intro-

duced it already in Chapter 3 where I mentioned the supervised learning ideas. Indeed, the task of the modeler is often to ‘learn’ the function $g(X)$.

Definition 8.2 (Regression). **Regression** analysis aims to uncover the functional relation between two random variables X and Y :

$$Y = g(X) + \epsilon. \quad (8.5)$$

Here, $g(X)$ is a function mapping X into Y and ϵ is a random variable called error.

The task of the regression is to find the function $g(X)$ which minimizes the error $\epsilon = Y - g(X)$.¹ What property of the error one aims to minimize depends on the system and on the purpose of the regression. A common quantity to minimize is the mean square error, which is the variance of ϵ . However, in some circumstances, such as in the presence of fat tails, the variance might not be defined or it can be hard to estimate with precision due to outliers and therefore other quantities such as the mean absolute error might be used instead. In other contexts, the relevant quantile of the error can be minimized as in the so-called quantile regression. It should be quite clear from the previous chapter that another quantity that one might aim to minimize is the entropy of the error $H(\epsilon)$ which is the uncertainty on Y not explained by X .²

Definition 8.3 (Entropy of the regression error). For two dependent random variables X and Y with $Y = g(X) + \epsilon$, the entropy of the error ϵ is the conditional entropy which was introduced in Definition 7.5:

$$H(Y|X) = H(\epsilon) = H(Y - g(X)). \quad (8.6)$$

Indeed, when X is fixed to a given value, then $Y|X = \epsilon + const..$

Performances of the models and their computability often depend on the property of the error that it is minimized.

Remark 8.3. Often, in the literature, for the regression of Y with respect to X , the random variable Y is referred to as the ‘dependent’ variable and the random variable X as ‘independent’. This is however quite misleading because the dependency is always between both variables and one can always equivalently write the inverse relation $X = \tilde{g}(Y) + \tilde{\epsilon}$ reversing therefore the roles of the variables. Indeed, when dependency is concerned, there is no directionality. Both variables carry the same information about each-other.

¹ The regression error is the difference between the true value of Y and $g(X)$. Instead the difference between the observed value \hat{Y} and its estimate $\hat{g}(\hat{X})$ is called ‘residual’.

² Notice that, for normal modeling the minimization of the entropy of the error, $H(\epsilon)$, coincides with the minimization of the error’s variance.

In other contexts, Y is named ‘response’ variable and X is named ‘predictor’ or ‘explanatory variable’ or, in machine learning, ‘feature’.

8.3 Linear dependency

When the relation between the two variables is linear, then the dependency is called ‘linear’ and the regression is called linear regression.

Definition 8.4 (Linear dependency). When the regressor that minimize the error ϵ is a linear equation, $g(X) = \beta_0 + \beta_1 X$, and therefore the dependency relation between X and Y is given by

$$Y = \beta_0 + \beta_1 X + \epsilon \quad (8.7)$$

with $\beta_1 \neq 0$ and ϵ an error independent on X , then the variables are said to be **linearly dependent**.

Linear dependency is simple and very important. The vast majority of studies of dependency between variables concern linear dependency. We shall see shortly that linear dependency is directly related with commonly used quantities such as correlation and covariance. Linear dependency is also directly related with multivariate normal modeling. It is therefore important to understand well the fundamentals of linear dependency before adventuring into the non-linear domain.

8.3.1 Covariance and dependency

To establish dependence between two variables it is necessary to measure the difference between the joint probability and the product of the marginal probabilities (see Eqs.8.1, 8.2 and 8.3). For this purpose, one could, for instance, measure the integral of the deviation from the equality in Eq.8.3 over the entire support:

$$\text{Cov}(X, Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (F_{X,Y}(x, y) - F_X(x)F_Y(y)) dx dy. \quad (8.8)$$

This quantity is called covariance. This is however a very unusual expression for it. Indeed, the common expression for the covariance is defined in terms of expected values

$$\text{Cov}(X, Y) = \mathbb{E}((X - \mu_X)(Y - \mu_Y)) \quad (8.9)$$

(see Definition 6.1), which can be proved to be identical to the previous through the Hoeffding lemma [Hoeffding, 1940].

When two variables are independent, the difference $F_{X,Y}(x, y) - F_X(x)F_Y(y)$ is always equal to zero and therefore their covariance is equal to zero. Conversely,

the expression in Eq.8.8 reveals that the covariance is an integral over the support of such a difference and therefore there can be instances when dependent variables have both positive and negative deviations from the equality in Eq.8.3 and this could eventually result into zero covariance even if the variables are not independent. Therefore, independent variables have zero covariance but not vice-versa: there can be dependent variables with zero covariance.

Example 8.1 (Zero covariance for dependent variables). An example of two variables which are dependent but have zero covariance is $Y = X^2$ when X is assumed to have a symmetric distribution around zero. Indeed, symmetry implies $\mathbb{E}(X) = 0$ and $\mathbb{E}(X^3) = 0$, consequently we have $\text{Cov}(X, Y) = \mathbb{E}((X - \mu_X)(Y - \mu_Y)) = \mathbb{E}(X(X^2 - \mu_Y)) = \mathbb{E}(X^3) - \mu_Y\mathbb{E}(X) = 0$.

Remark 8.4. When the dependency between the two variables is linear, then zero covariance implies independence. Indeed, for linearly dependent variables ($Y = \beta_0 + \beta_1 X + \epsilon$):

$$\text{Cov}(X, Y) = \beta_1 \mathbb{E}((X - \mu_X)^2) = \beta_1 \sigma_X^2. \quad (8.10)$$

The variables X, Y are **linearly independent** if and only if $\beta_1 = 0$. Eq.8.10 incidentally also reveals that the linear regression coefficient β_1 is directly proportional to the covariance.

Positive covariance is retrieved when two variables vary together in the same direction: when one variable is large also the other variable is likely to be large and vice versa when one variable is small also the other is likely to be small. Instead, negative covariance is associated to a couple of variables that vary in opposite directions: when one is large the other is likely to be small.

8.3.2 Correlation and the strength of linear dependency

The value of the covariance depends on the scale of measure of the variables. For example, the value of the covariance between two variables expressed in millimetres is one million time larger than the covariance between the same two variables expressed in meters. Therefore, the value itself is meaningless. If one instead scales the covariance by dividing it by the product of the standard deviations, then the measure becomes independent on scale and unit of measure

$$\text{Corr}(X, Y) = \rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}. \quad (8.11)$$

This quantity is called correlation and it has been already defined in Definition 6.3.

This measure is scale-invariant, $\text{Corr}(aX, Y) = \text{Corr}(X, Y)$; it is symmetrical

$\text{Corr}(X, Y) = \text{Corr}(Y, X)$; its largest value is +1 and its smallest is -1 with zero indicating linear independence.

Correlation quantifies linear dependency between two variables. Indeed, when two variables are linearly related, $Y = \beta_0 + \beta_1 X + \epsilon$, then the correlation coefficient and the coefficients in the linear relation are related. Specifically:

$$\text{Corr}(X, Y) = \frac{\beta_1}{\sqrt{\beta_1^2 + \frac{\text{Var}(\epsilon)}{\text{Var}(X)}}}. \quad (8.12)$$

One can observe that when $\text{Var}(\epsilon) \rightarrow 0$ then $\text{Corr}(X, Y) = \beta_1/|\beta_1| = 1$ for $\beta_1 > 0$ or = -1 for $\beta_1 < 0$. Conversely, when the variance of the error, $\text{Var}(\epsilon)$, increases, then the correlation between the variables decreases and eventually it goes towards zero for $\text{Var}(\epsilon) \gg \text{Var}(X)$ even if $\beta_1 \neq 0$.

Example 8.2 (Least squares solution of linear regression and correlation coefficients). In the linear regression analysis, given two random variables X and Y , one wishes to compute the coefficients that best reproduce their linear relation. Specifically, one searches for the coefficients β_0 and β_1 that minimize $\text{Var}(\epsilon)$ in the expression

$$Y = \beta_0 + \beta_1 X + \epsilon. \quad (8.13)$$

This is the so-called least squares method, because minimizing the loss function $L = \text{Var}(\epsilon) = \mathbb{E}(\epsilon^2)$ is indeed equivalent to minimize the expected value of the square of the difference between the variable Y and its linear model $\beta_0 + \beta_1 X$: $L = \mathbb{E}((Y - (\beta_0 + \beta_1 X))^2)$. One therefore searches for the parameters β_1 and β_0 that minimize L .

This can be done analytically, indeed

$$\text{Var}(\epsilon) = \text{Var}(Y - \beta_1 X) = \text{Var}(Y) + \beta_1^2 \text{Var}(X) - 2\beta_1 \text{Cov}(X, Y) \quad (8.14)$$

and by differentiating with respect to β_1 and equalling to zero one has

$$\beta_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \rho_{XY} \frac{\sigma_Y}{\sigma_X}. \quad (8.15)$$

The coefficient β_0 can be then obtained from the expected value

$$\mathbb{E}(Y) = \beta_1 \mathbb{E}(X) + \beta_0 \quad (8.16)$$

by substituting the expression for β_1 :

$$\beta_0 = \mathbb{E}(Y) - \frac{\text{Cov}(XY)}{\text{Var}(X)} \mathbb{E}(X). \quad (8.17)$$

One might notice that by substituting these coefficients, the variance of the error is: $\text{Var}(\epsilon) = \text{Var}(Y)(1 - \frac{\text{Cov}(X,Y)^2}{\text{Var}(X)\text{Var}(Y)})$ or

$$\frac{\text{Var}(\epsilon)}{\text{Var}(Y)} = (1 - \text{Corr}(X, Y)^2). \quad (8.18)$$

This reveals that the square of the correlation coefficient $R^2 = \text{Corr}(X, Y)^2$ is a measure of the goodness of the linear model. Large R^2 , close to 1, indicate that the error is small and the linear model describes well the dependency between the two variables. Whereas when $R^2 \rightarrow 0$ the linear model is not describing well the relation between the variables. The R^2 is called coefficient of determination (see Definition 8.5) and it is indeed used to quantify the goodness of the linear regression (see Section 18.4.1).

Definition 8.5 (Coefficient of determination). The **coefficient of determination**, R^2 , is the square of the correlation coefficient: $R^2 = \text{Corr}(X, Y)^2$. It is a measure of the goodness of the linear dependency model fit. $R^2 = 1$ means that the variable Y is perfectly described by $\beta_0 + \beta_1 X$. Conversely, $R^2 = 0$ means that the linear relation $\beta_0 + \beta_1 X$ does not capture any feature of the variable Y . For a given coefficient of determination R^2 , one can say that “the linear equation $\beta_0 + \beta_1 Y$ describes $R^2\%$ of the variance of variable Y ” (see previous Example 8.2)

Remark 8.5. Note that the regressions

$$Y = \beta_0 + \beta_1 X + \epsilon \quad (8.19)$$

or

$$X = \beta'_0 +' \beta'_1 Y + \epsilon' \quad (8.20)$$

have different minimal errors $\text{Var}(\epsilon) \neq \text{Var}(\epsilon')$ but they are equivalent in terms of the description of the linear relation between the variables. Indeed, in both regressions the coefficient of determination is the same and given by:

$$R^2 = \beta_1 \beta'_1. \quad (8.21)$$

Remark 8.6. The error ϵ are uncorrelated with X when β_1 have the optimal solution value which minimizes $\text{Var}(\epsilon)$. Indeed, we have

$$\text{Cov}(X, \epsilon) = \text{Cov}(X, Y - \beta_0 - \beta_1 X) = \text{Cov}(X, Y) - \beta_1 \text{Cov}(X, X) \quad (8.22)$$

and by substituting the solution

$$\beta_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \quad (8.23)$$

one gets

$$\text{Cov}(X, \epsilon) = \text{Cov}(X, Y) - \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \text{Cov}(X, X) = 0. \quad (8.24)$$

Note that instead variable Y is correlated with the error and

$$\text{Cov}(Y, \epsilon) = \text{Var}(Y)(1 - \text{Corr}(X, Y)^2), \quad (8.25)$$

which is the same as $\text{Cov}(\epsilon, \epsilon) = \text{Var}(\epsilon)$ and it is the quantity that is minimized in the regression. From Eq.8.25 one has for the correlation

$$\text{Corr}(Y, \epsilon) = \sqrt{\frac{\text{Var}(Y)}{\text{Var}(\epsilon)}}(1 - \text{Corr}(X, Y)^2) = \sqrt{1 - R^2} = \sqrt{1 - \text{Corr}(X, Y)^2}. \quad (8.26)$$

8.4 Multilinear regression and the covariance matrix

Covariances and correlations are related also to the coefficients in multilinear regressions, which are regressions of a variable Y with respect to several variables X_i with $i = 1 \dots p_X$:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{p_X} X_{p_X} + \epsilon. \quad (8.27)$$

By using a vector notation one can write the previous expression as

$$Y = \boldsymbol{\beta}^\top \mathbf{X} + \epsilon. \quad (8.28)$$

with $\boldsymbol{\beta}^\top = (\beta_0, \beta_1, \beta_2, \dots, \beta_{p_X})$ and $\mathbf{X} = (1, X_1, X_2, \dots, X_{p_X})^\top$. In this notation, the least squares solution follows very similar steps to the linear regression case with two variable only. The loss function is

$$\begin{aligned} L &= \text{Var}(\epsilon) = \text{Var}(Y - \boldsymbol{\beta}^\top \mathbf{X}) \\ &= \text{Var}(Y) + \boldsymbol{\beta}^\top \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})^\top] \boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(Y - \mu_Y)] \end{aligned} \quad (8.29)$$

and the minimum is obtained by differentiating with respect to each of the coefficients β_i and equalling to zero

$$\boldsymbol{\beta}^* = \boldsymbol{\Sigma}_{\mathbf{XX}}^{-1} \boldsymbol{\Sigma}_{\mathbf{XY}} \quad (8.30)$$

where

$$\boldsymbol{\Sigma}_{\mathbf{XX}} = \mathbb{E}((\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})^\top) \quad (8.31)$$

is the covariance matrix of variables \mathbf{X} . Whereas,

$$\boldsymbol{\Sigma}_{\mathbf{XY}} = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(Y - \mu_Y)] \quad (8.32)$$

is the vector of covariances between the components of the variable \mathbf{X} and the variable Y .

Substituting the expression for β^* , the variance of the error is

$$\text{Var}(\epsilon) = \text{Var}(Y) - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}. \quad (8.33)$$

Remark 8.7. By writing explicitly Eq.8.30 in terms of the matrices elements, each regression coefficient β_i^* is given by

$$\beta_i^* = \sum_{j=1}^{p_X} (\Sigma_{XX}^{-1})_{i,j} (\Sigma_{YX})_j \quad (8.34)$$

with

$$(\Sigma_{YX})_j = \sigma_Y \sigma_{X_j} \rho_{YX_j} \quad (8.35)$$

and where σ_Y and σ_{X_j} are respectively the standard deviations of Y and X_j ; ρ_{YX_j} is the correlation between Y and X_j ; $(\Sigma_{XX}^{-1})_{j,i}$ is the element i,j of the inverse of the covariance matrix Σ_{XX} .

A generalization to the case when Y is a set of p_Y variables $\mathbf{Y} = (Y_1, \dots, Y_{p_Y})^\top$, is obtainable by just repeating the previous procedure for each component of \mathbf{Y} and its expression is identical to the previous expressions with the bold symbols use instead:

$$\mathbf{Y} = \mathbf{B}^\top \mathbf{X} + \boldsymbol{\epsilon}. \quad (8.36)$$

Where also the error have dimension p_Y , $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_{p_Y})^\top$ and β has become the $(p_X + 1) \times p_Y$ matrix \mathbf{B} . The variables \mathbf{Y} are in general dependent and their covariance is directly related to the coefficients \mathbf{B} :

$$\text{Cov}(\mathbf{Y}, \mathbf{Y}) = \mathbf{B}^\top \text{Cov}(\mathbf{X}, \mathbf{X}) \mathbf{B} + \text{Cov}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}). \quad (8.37)$$

Note the use of notation $\text{Cov}(\mathbf{Y}, \mathbf{Y}) = \Sigma_{YY}$ for the matrix of covariances between the components of \mathbf{Y} (and analogously for $\text{Cov}(\mathbf{X}, \mathbf{X})$ and $\text{Cov}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})$).

The solution for the matrix of coefficients \mathbf{B} that minimize the sum of the variances of the errors,

$$\sum_{k=1}^n \text{Var}(\epsilon_k) = \|\mathbf{Y} - \mathbf{B}^\top \mathbf{X}\|^2, \quad (8.38)$$

sometimes called total variance, is essentially the same as outlined before because one can solve the equations by solving for each Y_k separately. The overall solution is therefore

$$\mathbf{B}^* = \Sigma_{XX}^{-1} \Sigma_{YX} \quad (8.39)$$

with $\Sigma_{YX} \in \mathbb{R}^{p_Y \times p_X+1}$ and $\Sigma_{XX} \in \mathbb{R}^{p_X+1 \times p_X+1}$.

It was noted by Wilks [Wilks, 1934] that the same solution minimizes also the determinant of the covariance of the errors $|\text{Cov}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})|$.

The covariance of the errors $\text{Cov}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})$ is the ‘conditional covariance’, as defined for the multivariate normal distribution in Section 6.7.1. Indeed, it is what is left of the covariance of \mathbf{Y} when it is linearly conditioned to \mathbf{X} :

$$\text{Cov}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}) = \text{Cov}(\mathbf{Y} - \mathbf{B}^\top \mathbf{X}, \mathbf{Y} - \mathbf{B}^\top \mathbf{X}) \quad (8.40)$$

By substituting the expression $\mathbf{B}^* = \boldsymbol{\Sigma}_{\mathbf{XX}}^{-1} \boldsymbol{\Sigma}_{\mathbf{XY}}$ one gets.

$$\text{Cov}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}) = \boldsymbol{\Sigma}_{\mathbf{YY}} - \boldsymbol{\Sigma}_{\mathbf{YX}} \boldsymbol{\Sigma}_{\mathbf{XX}}^{-1} \boldsymbol{\Sigma}_{\mathbf{XY}}. \quad (8.41)$$

Which is indeed identical to the one for the covariance matrix of the conditional multivariate normal distribution (see expression for $\boldsymbol{\Sigma}_{\mathbf{YY}|\mathbf{X}}$ in equation 6.59). Therefore, from the Bayes’ formula, this is the covariance associated with the ratio between two multivariate normals $\varphi(\mathbf{Z} = \mathbf{z})/\varphi(\mathbf{Y} = \mathbf{y})$ with $\mathbf{Z} = \mathbf{X} \cup \mathbf{Y}$ and has as determinant the ratio of the determinants:

$$|\text{Cov}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})| = |\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}\boldsymbol{\epsilon}}| = \frac{|\boldsymbol{\Sigma}_{\mathbf{ZZ}}|}{|\boldsymbol{\Sigma}_{\mathbf{XX}}|}. \quad (8.42)$$

Remark 8.8. In analogy with the univariate case (see Remark 8.6), also in this multivariate case the errors $\boldsymbol{\epsilon}$ are uncorrelated with \mathbf{X} when \mathbf{B} has the optimal solution values. Indeed, one has

$$\begin{aligned} \text{Cov}(\mathbf{X}, \boldsymbol{\epsilon}) &= \text{Cov}(\mathbf{X}, \mathbf{Y} - \mathbf{B}^\top \mathbf{X}) \\ &= \text{Cov}(\mathbf{X}, \mathbf{Y}) - \text{Cov}(\mathbf{X}, \mathbf{X})\mathbf{B} \end{aligned} \quad (8.43)$$

and substituting the solution

$$\mathbf{B}^* = \boldsymbol{\Sigma}_{\mathbf{XX}}^{-1} \boldsymbol{\Sigma}_{\mathbf{XY}} \quad (8.44)$$

one gets

$$\text{Cov}(\mathbf{X}, \boldsymbol{\epsilon}) = \boldsymbol{\Sigma}_{\mathbf{XY}} - \boldsymbol{\Sigma}_{\mathbf{XX}} \boldsymbol{\Sigma}_{\mathbf{XX}}^{-1} \boldsymbol{\Sigma}_{\mathbf{XY}} = 0 \quad (8.45)$$

8.5 Precision matrix, inverse covariance and partial correlations

The inverse of the covariance matrix, $\boldsymbol{\Sigma}^{-1}$, also called precision matrix (see Definition 6.6), appears in all previous expressions for the multilinear regressions as well as in the expression for several multivariate probability distributions. It is indeed a very important quantity and its estimation is crucial for the modeling of several multivariate systems. In the multivariate normal case, zero elements in the precision matrix correspond to couples of variables that are conditionally independent. The off-diagonal coefficients of the precision matrix normalized by the corresponding diagonal elements are called partial correlations.

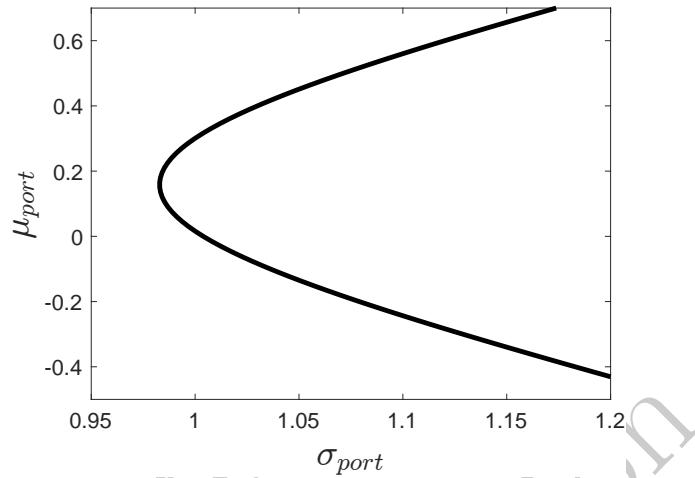


Figure 8.1 Markowitz efficient frontier obtained for a system of three variables with the covariance in Eq.8.59 and expected values $\mu = (0.1, -0.2, 0.3)^\top$.

Definition 8.6 (Partial correlations). For a set of random variables \mathbf{X} the **partial correlation** between two variables $X_i, X_j \in \mathbf{X}$, conditioned to all others $\mathbf{Z} = \mathbf{X} \setminus (X_i, X_j)$ (i.e. all \mathbf{X} variables except for X_i or X_j) is defined as:

$$\rho_{X_i, X_j \cdot \mathbf{Z}} = -\frac{(\boldsymbol{\Sigma}^{-1})_{i,j}}{\sqrt{(\boldsymbol{\Sigma}^{-1})_{i,i}(\boldsymbol{\Sigma}^{-1})_{j,j}}}. \quad (8.46)$$

Partial correlations are Pearson's correlations (see Definition 6.3) but between the errors of the multilinear regressions of X_i and X_j with respect to all other variables $\mathbf{Z} = \mathbf{X} \setminus (X_i, X_j)$:

$$\rho_{X_i X_j \cdot \mathbf{Z}} = \text{Corr}(X_i - \boldsymbol{\beta}_{X_i}^{*\top} \mathbf{Z}, X_j - \boldsymbol{\beta}_{X_j}^{*\top} \mathbf{Z}). \quad (8.47)$$

where $\boldsymbol{\beta}_{X_i}^* = \boldsymbol{\Sigma}_{\mathbf{Z}\mathbf{Z}}^{-1} \boldsymbol{\Sigma}_{\mathbf{Z}X_i}$ and $\boldsymbol{\beta}_{X_j}^* = \boldsymbol{\Sigma}_{\mathbf{Z}\mathbf{Z}}^{-1} \boldsymbol{\Sigma}_{\mathbf{Z}X_j}$ (see Eq.8.30).

In the case when one has only three variables: X , Y and Z , then the partial correlation between X and Y can be expressed in terms of the correlations as

$$\rho_{XY \cdot Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1 - \rho_{XZ}^2}\sqrt{1 - \rho_{YZ}^2}}. \quad (8.48)$$

Example 8.3 (Maximum mean and minimum variance of a linear combination of random variables: the Markowitz portfolio theory). As a relevant and practical example where the covariance and its inverse play a very prominent role, let me introduce a classical portfolio optimization problem. Let's suppose there are p assets in a portfolio with returns described by p

random variables X_1, \dots, X_p . Let's call w_1, \dots, w_p the weights for each asset in the portfolio, with the condition $\sum_{k=1}^p w_k = 1$, so that the portfolio return is given by

$$R_{port} = w_1 X_1 + \dots + w_p X_p. \quad (8.49)$$

The portfolio optimization problem consists in finding the weights w_i which are simultaneously maximizing portfolio's returns and minimizing investment's risk.

The Markowitz portfolio construction chooses the weights such that the expected returns of the portfolio are at a given rate and the variance is minimal [Markowitz, 1952, 1968]. Minimization of the variance is a simple way to reduce risk. Indeed, it constrains the width of the distribution and a smaller variance corresponds to a smaller risk of large price variations. For some families of distributions, such as the normal distribution, but also the elliptical family, all moments are proportional to the variance (when defined) and therefore the minimization of the variance is a quite general way to minimize risk.

The expected value of the portfolio return is

$$\mu_{port} = \mathbb{E}\left(\sum_{k,j=1}^p w_k X_k\right) = \sum_{k,j=1}^p w_k \mathbb{E}(X_k) = w_1 \mu_1 + \dots + w_p \mu_p \quad (8.50)$$

with $\mu_k = E(X_k)$. The portfolio variance is instead

$$\sigma_{port}^2 = \mathbb{E}((R_{port} - \mu_{port})^2) = \mathbb{E}\left(\sum_{k,j=1}^p w_k (X_k - \mu_k)(X_j - \mu_j) w_j\right). \quad (8.51)$$

The expected value can be moved inside the sum and one can notice that $\mathbb{E}((X_k - \mu_k)(X_j - \mu_j)) = \Sigma_{k,j}$ are the elements of the covariance matrix Σ and therefore the portfolio variance can be written as

$$\sigma_{port}^2 = \sum_{k,j=1}^p w_k \Sigma_{k,j} w_j = \mathbf{w}^\top \Sigma \mathbf{w}. \quad (8.52)$$

With $\mathbf{w} = (w_1, \dots, w_p)^\top$, the vector of weights.

Within the Markowitz's approach one wants to minimize σ_{port} while keeping μ_{port} at a given value and also while imposing that the sum of the weights must be equal to one. For this purpose, one can use the Lagrange multiplier method (see Bertsekas [1982] and the note after this example) where instead of minimizing directly a quantity one aims to minimize plus the constraints multiplied by a constant (the so called Lagrange multipliers). In this case, the quantity to minimize is σ_{port} and the constraints are

$\sum_{k=1}^p w_k \mu_k = \mu_{port}$ and $\sum_{k=1}^p w_k = 1$. The Lagrangian function is therefore:

$$\mathcal{L} = \sum_{k,j=1}^p w_k \Sigma_{k,j} w_j - \lambda \sum_{k=1}^p w_k \mu_k - \gamma \sum_{k=1}^p w_k. \quad (8.53)$$

or, in matrix notation,

$$\mathcal{L} = \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w} - \lambda \mathbf{w}^\top \boldsymbol{\mu} - \gamma \mathbf{w}^\top \mathbf{1}. \quad (8.54)$$

Where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)^\top$ and $\mathbf{1} = (1, \dots, 1)^\top$. The extreme is searched by solving the system of equations

$$\frac{\partial}{\partial w_k} \mathcal{L} = 0 \text{ with } k = 1 \dots p. \quad (8.55)$$

That has solution

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1} (\lambda \boldsymbol{\mu} + \gamma \mathbf{1}). \quad (8.56)$$

Note that the portfolio's weights, \mathbf{w} are determined by the inverse covariance $\boldsymbol{\Sigma}^{-1}$. Therefore, a good estimation of the 'true' covariance is the essential ingredient in the Markowitz portfolio construction. For a given portfolio return, μ_p , the Lagrange multipliers are:

$$\begin{aligned} \lambda &= \frac{\mu_p - b/a}{\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} - b^2/a} \\ \gamma &= \frac{1 - \lambda b}{a} \end{aligned} \quad (8.57)$$

with

$$\begin{aligned} a &= \mathbf{1}^\top \boldsymbol{\Sigma}^{-1} \mathbf{1} \\ b &= \mathbf{1}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}. \end{aligned} \quad (8.58)$$

Normally λ is kept as a 'free parameter' with $\lambda = 0$ corresponding to the case when the mean portfolio returns, μ_{port} , are not constrained and the portfolio is the 'minimum variance' one. By varying λ one explores the so-called 'efficient frontier' where the mean returns μ_{port} and the portfolio variance σ_{port}^2 change simultaneously with the value of λ .

In practice, $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ must be estimated from data and this makes the Markowitz weighting in Equation 8.56 becoming sub-optimal especially when the estimate is from small observation sets. This will be examined further in Section 15.7.3 where methods to estimate these parameters are discussed. In particular, the interested reader might want to see Example 15.7, where the effect of the estimation of the covariance on the portfolio performances is discussed.

Let me show the Markowitz efficient frontier for a simple example with

three random variables with covariance matrix

$$\Sigma = \begin{pmatrix} 1.0 & 0.7 & 0.9 \\ 0.7 & 3.0 & 0.8 \\ 0.9 & 0.8 & 1.0 \end{pmatrix}. \quad (8.59)$$

(notice that this is the same as in Eq.6.71). I compute the weights using Eq.8.56 fixing the parameters λ and γ with Eqs.8.57 and 8.58 for a range of expected portfolio returns, μ_{port} between -0.5 and 0.7. In Fig.8.1 I report the curve for μ_{port} vs. σ_{port} which is called the efficient frontier.

The **Lagrange multiplier** is a method used to find extrema of a function $g(x)$ subject to some constraints, i.e. $g_0(x) = 0$. The method consists in finding the critical points of the Lagrange function

$$\mathcal{L}(x, \lambda) = g(x) - \lambda g_0(x), \quad (8.60)$$

instead of $g(x)$. The Lagrange multiplier theorem guarantees that there exists a unique Lagrange multiplier λ^* for which, at $x = x^*$, $d\mathcal{L}(x, \lambda^*)/dx|_{x=x^*} = 0$ and both $g(x^*)$ is extreme (or a stationary point with zero derivative) while the constraint is satisfied $g_0(x^*) = 0$. The method can be directly extended to problems with several variables and several constraints. (See [Bertsekas, 1982] for further reading.)

8.6 A generalized dependency measure

The use of the correlation coefficient to measure dependence between variables is very common and widespread. However, this measure can be very problematic and it might sometimes lead to serious faults. Indeed, as discussed in the previous part of this chapter, the correlation coefficient quantifies linear dependency between two variables indicating how well the relation $Y = \beta_0 + \beta_1 X + \epsilon$ is describing the dependency between X and Y (see Eq.8.12). However, dependency between variables can be described by any functional form $Y = g(X) + \epsilon$ (see Section 8.2) and sometimes, especially when $g(X)$ is not monotonic, correlations will not provide any valuable quantification of the true dependency between the variables (see Example 8.1). Other problems might arise with non-normally distributed variables. Indeed, we already noticed that the standard deviation is not defined for random variables with fat-tailed power law distributions with tail exponent smaller or equal than 2. This implies that for these variables the correlation coefficient is not defined as well. Moreover, when the tail index α (see 5.5) belongs to the interval $(2, 4]$, the correlation coefficient exists but its sample estimator is highly unreliable because its distribution has undefined second moments and therefore it can have large variations. Despite all these shortcomings, the corre-

lation coefficient has many nice properties that one would like to keep also in a more general dependency measure. Specifically:

1. the correlation coefficient between two variables is symmetric, $\rho_{X,Y} = \rho_{Y,X}$;
2. the correlation coefficient has values in the defined interval $\rho_{X,Y} \in [-1, +1]$ with the extremes corresponding to complete linear dependence and complete linear anti-dependence;
3. zero correlation coefficient, $\rho_{X,Y} = 0$, implies linear independence between variable X and variable Y .

It would be excellent to have an equivalent measure which applies to any kind of dependency, not only to the linear one. This was the idea of Rényi that indeed in his 1959 paper proposed the following five requirements for a ‘good’ measure of dependency [Rényi, 1959b]:

A measure of dependency $C(X, Y)$ between two random variables X and Y must:

- 1) $C(X, Y) = C(Y, X)$, be symmetric
- 2) $0 \leq C(X, Y) \leq 1$, have values within $[0, 1]$;
- 3) $C(X, Y) = 1$, if there is a strict dependence between X and Y (i.e. if $X = g_1(Y)$ or $Y = g_2(X)$, with $g_1(\cdot), g_2(\cdot)$ Borel-measurable functions, see Halmos [1950]);
- 4) $C(g_1(X), g_2(Y)) = C(X, Y)$, be invariant for transformations with $g_1(\cdot), g_2(\cdot)$ Borel-measurable functions;
- 5) $C(X, Y) = |\text{Corr}(X, Y)|$, if X, Y are linearly dependent.

A proper definition of **Borel measurable** functions goes outside the style of this book. For the purpose of this criteria for dependency, one needs functions that map finite domains into finite codomains. Interested readers can refer to the vast literature on measure theory, perhaps starting from Halmos [1950].

The proposal of Rényi was therefore very simple: quantifying dependency in terms of the absolute value of correlations between functions of the variables. If there exist two functions of the variables X and Y such that their correlation is significantly different from zero, then the two variables are dependent. Given that the sign of such a correlation becomes meaningless, he proposed to use the absolute value.

The quantity

$$C(X, Y) = \sup_{g_1(\cdot), g_2(\cdot)} |\text{Corr}(g_1(X), g_2(Y))|, \quad (8.61)$$

is a dependency measure that can detect and quantify any kind of non-linear dependency. However, for practical purposes, it must be noticed that the space of all measurable functions is infinitely large containing all sorts of weird functions. Nonetheless, this is a very clean and powerful generalization of a measure

of dependency beyond the linear case. Furthermore, all considerations about correlations and linear regressions discussed in Section 8.2 are still applicable to this case. The search for the supremum in Eq.8.61 coincides with the search for the regression:

$$g_1(Y) = g_2(X) + \epsilon, \quad (8.62)$$

that minimizes the error ϵ . Indeed, as I pointed out already in Section 8.2: the dependency problem always maps into a regression problem.

8.7 Correlation ratio

To measure dependency between two random variables X and Y one can look for the function $g(X)$ that best fits Y and then measure the goodness of the fit (goodness of the regression). If the fit is good and variable Y is described well with $g(X)$, then the two variables must be dependent. A quantification of the goodness of the fit is given by the ratio between the variance of the error and the variance of the regressed variable. The best fitting function is the one providing the supremum of such a ratio:

$$\inf_{g(\cdot)} \frac{\text{Var}(Y - g(X))}{\text{Var}(Y)} = 1 - \eta_{XY}^2, \quad (8.63)$$

where inf stands for infimum, i.e. the greatest lower bound. Indeed, if $g(X)$ fully describes Y this quantity is zero, instead if $g(X)$ is unable to describe anything of the variability of Y than this ratio is one.³ The quantity η_{XY}^2 is named **correlation ratio**. Let me note that in the case of linear regression the correlation ratio η_{XY}^2 coincides with the square of the correlation coefficient ρ_{XY}^2 and therefore it can be seen as a generalization of the coefficient of determination (see Definition 8.5). Indeed, especially when it comes to its empirical estimation, η_{XY}^2 is often called R^2 also in the non linear case (see Section 18.4.1).

If there are no other conditions, the function that minimizes the mean square error of the regression $Y = g(x) + \epsilon$ is the conditional expected value $g(x) = \mathbb{E}(Y|X = x)$.

Indeed, it is immediate to prove directly that the quantity $\mathbb{E}((Y - c)^2)$ is minimized by $c = \mathbb{E}(Y)$. Therefore, in the conditioned context $\mathbb{E}((Y - g(X))^2|X = x)$ is minimal with

$$g(x) = \mathbb{E}(Y|X = x) \quad (8.64)$$

for any value of $X = x$.

By using this conditional expected value then one can also use of the equality $\text{Var}(Y - \mathbb{E}(Y|X)) = \text{Var}(Y) - \text{Var}(\mathbb{E}(Y|X))$ (see note below) which implies

³ The supremum cannot do worst than one because, one is always achievable with $g(X) = 0$.

therefore that the correlation ratio can be written as

$$\eta_{XY}^2 = \frac{\text{Var}(\mathbb{E}(Y|X))}{\text{Var}(Y)} \quad (8.65)$$

Note that $\text{Var}(Y - \mathbb{E}(Y|X)) = \text{Var}(Y) - \text{Var}(\mathbb{E}(Y|X))$. Indeed:

$$\begin{aligned} \text{Var}(Y - \mathbb{E}(Y|X)) &= \\ &= \mathbb{E}(Y^2) + \mathbb{E}(\mathbb{E}(Y|X)^2) - 2\mathbb{E}(Y\mathbb{E}(Y|X)) \\ &= \mathbb{E}(\mathbb{E}(Y|X)^2) - \mathbb{E}(Y^2) \end{aligned} \quad (8.66)$$

Where I used $\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y)$ (law of total expectation) and $\mathbb{E}(Y\mathbb{E}(Y|X)) = \mathbb{E}(Y^2)$. Similarly,

$$\begin{aligned} \text{Var}(\mathbb{E}(Y|X)) &= \\ &= \mathbb{E}((\mathbb{E}(Y|X) - \mathbb{E}(Y))^2) \\ &= \mathbb{E}(\mathbb{E}(Y|X)^2) + \mathbb{E}(Y)^2 - 2\mathbb{E}(Y)^2 \\ &= \mathbb{E}(\mathbb{E}(Y|X)^2) - \mathbb{E}(Y)^2 \end{aligned} \quad (8.67)$$

By comparing the two expressions one obtains the above expression.

This is related to the law of total variance, which states that

$$\text{Var}(Y) = \mathbb{E}(\text{Var}(Y|X)) + \text{Var}(\mathbb{E}(Y|X)). \quad (8.68)$$

One can see that the correlation ratio η_{XY}^2 , when defined, satisfies all criteria listed in Section 8.6 for a general measure of dependency. When expressed in terms of ratio between the variance of the error of a regression and the variance of the variable (i.e. Eq.8.63), it is the supremum of the generalization to non-linear cases of the R^2 . It is therefore a general measure of goodness of any kind of regression. A practical problem for the computation of the correlation ratio η_{XY}^2 from data is that the conditional expectation $\mathbb{E}(Y|X)$ is not in general known and it is often hard to estimate from observations. This is indeed still the original regression problem: finding the function $g(X) = \mathbb{E}(Y|X)$ which minimizes the variance of the error. For such a regression, one can use many tools. Notably, support vector machines and forward neural networks are both, in theory, able to discover any functional relation between two variables (they are universal approximators, see Hammer and Gersmann [2003] and Hornik et al. [1989]). However, often the search for dependency through regression with a universal approximator is complicated by the fact that solutions tend to be overfitting. In order to reduce the number of parameters, it is often more practical to consider correlations between classes of pre-defined functions.

8.8 Non-linear correlations: rank correlations

There is a vast number of possible measures for non-linear dependency. Indeed, we have seen in the previous Section that any correlation between functions of the variables is a possible way to quantify non-linear dependency (Eq.8.61).

Here I mention a few simple measures that are often used in the literature and belong to a class known as ‘rank correlations’. They are the Pearson’s correlations between functions of the ‘ranks’ of the variables.

Definition 8.7 (Rank). The **rank** is the position in a ordered list. For examples the list of numbers 3.0, 1.3, 4.5, 2.1, 1.5 can be sorted in ascending order obtaining 1.3, 1.5, 2.1, 3.0, 4.5 and the **ordinal ranking** is: $\text{Rank}(3.0) = 4$, $\text{Rank}(1.3) = 1$, $\text{Rank}(4.5) = 5$, $\text{Rank}(2.1) = 3$, $\text{Rank}(1.5) = 2$.

When the same entry is repeated more than once in the list (equal tier) there are different ways to assign ranks. In this book I shall use a method called **fractional ranking** where to the equal tiers is associated a rank equal to the average position of that entry in the sorted list. For instance, the set 3.0, 1.3, 4.5, 2.1, 1.5, 3.0, 2.1 has fractional ranking 1, 2, 3.5, 3.5, 4, 5.5, 5.5, 7.

Example 8.4 (Fractional ranking). The list 3.0, 1.3, 4.5, 2.1, 2.1 can be ordered in ascending order obtaining 1.3, 2.1, 2.1, 3.0, 4.5. One can notice that it has two equal tier numbers respectively in position 2 and 3, therefore their fractional ranking is 2.5 for both: $\text{Rank}(3.0) = 4$, $\text{Rank}(1.3) = 1$, $\text{Rank}(4.5) = 5$, $\text{Rank}(2.1) = 2.5$.

Other rankings are possible but fractional ranking is particularly good for the kind of problems addressed in this book.

The ranked elements do not have to be numbers, for instance, a typical application concerns words or letters. For example the list D, G, A, D, E, F can be sorted in ascending alphabetic order, A, D, D, E, F, G and their rank is $\text{Rank}(D) = 2.5$, $\text{Rank}(G) = 6$, $\text{Rank}(A) = 1$, $\text{Rank}(E) = 4$, $\text{Rank}(F) = 5$.

By using rank statistics one automatically considers non-linear relations. In rank statistics, outlying observations from fat-tails become less impactful because outlying large values of the variables are simply ranked at the extreme, one step away from the others. However, in rank statistics only the relative order and not the actual value is taken into account and therefore risk might be underestimated.

8.8.1 Spearman- ρ

The Spearman- ρ correlation is the Pearson correlation between the ranks of the variables.

$$r_{XY} = \text{Corr}(\text{Rank}(X), \text{Rank}(Y)). \quad (8.69)$$

This rank correlation measures the strength of monotonic dependence between the variables. It is a good non-linear measure, robust also in presence of outliers. However, if the dependence is not monotonic then the Spearman- ρ correlation can be highly inaccurate.

In the case of bivariate normally distributed variables, Spearman correlation and Pearson correlations are related:

$$r_{XY} = \frac{6}{\pi} \arcsin\left(\frac{\rho_{XY}}{2}\right). \quad (8.70)$$

8.8.2 Kendall- τ

The Kendall- τ correlation is the Pearson correlations between the signs of the rank differences. For two sets of observations $(\hat{x}_1, \dots, \hat{x}_q)$ and $(\hat{y}_1, \dots, \hat{y}_q)$ the Kendall- τ is

$$\tau_{XY} = 2 \frac{\sum_{i < j} \text{sign}(\hat{x}_i - \hat{x}_j) \text{sign}(\hat{y}_i - \hat{y}_j)}{n(n-1)}, \quad (8.71)$$

where $\text{sign}(\Delta) = +1$ if $\Delta > 0$ and $\text{sign}(\Delta) = -1$ if $\Delta < 0$ and $\text{sign}(\Delta) = 0$ if $\Delta = 0$. Notice that the sum at the nominator counts the number of concordant pairs, i.e. pairs of values of X and Y variables that both increase or both decrease in the same observation interval. Instead, the quantity $n(n-1)/2$ is the number of all possible pairs. If all pairs simultaneously increase or decrease, it means that the two signals always go in the same directions and in this case the Kendall- τ correlation is equal to one. Conversely, minus one is obtained when all pairs go in opposite directions. The Kendall- τ correlation captures well non-linear monotonic dependency and tends to be more robust than the Spearman- ρ in many practical situations. However, it becomes computationally expensive when the number of observations is large.

In the case of bivariate normally distributed variables, Kendall- τ correlation and Pearson- ρ correlations are related:

$$\tau_{XY} = \frac{2}{\pi} \arcsin(\rho_{XY}). \quad (8.72)$$

8.9 Information-theoretic measures of dependency

I started this chapter by introducing, in Section 8.1, the concept of dependency in terms of the difference between the joint probability and the product of the marginal probabilities (Eq.8.1). From a general probabilistic perspective, to quantify dependency one could therefore measure the distance between the joint probability distribution function $f_{X,Y}(X, Y)$ and the product of the two marginal distributions $f_X(X)$ and $f_Y(Y)$. The two variables are independent if and only if such a distance is zero.

A general measure of distance between two probability distributions functions is the Kullback–Leibler divergence $D_{KL}(P \parallel Q)$ (Definition 7.2). In this case, one must choose $P = f_{X,Y}(X, Y)$ and $Q = f_X(X)f_Y(Y)$ obtaining a measure of

dependence that is equal to zero if and only if the variables are independent and increases in value more the variables are dependent. In information theory such a special KLD takes the name of mutual information.

8.9.1 Mutual information

Definition 8.8 (Mutual information for Shannon entropy). The **mutual information** for Shannon entropy is the Kullback-Leibler divergence between the joint distribution and the product of the marginals:

$$D_{\text{KL}}(f_{X,Y}(x,y) \parallel f_X(x)f_Y(y)) = \int_{\Omega_X} \int_{\Omega_Y} f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)} dx dy. \quad (8.73)$$

and it is denoted with $I(X;Y)$.

One can re-write Eq.8.73 as

$$\begin{aligned} I(X;Y) &= \int_{\Omega_X} \int_{\Omega_Y} f_{X,Y}(x,y) \log f_{X,Y}(x,y) dx dy \\ &\quad - \int_{\Omega_X} \int_{\Omega_Y} f_{X,Y}(x,y) \log f_X(x) dx dy - \int_{\Omega_X} \int_{\Omega_Y} f_{X,Y}(x,y) \log f_Y(y) dx dy \\ &\quad - \int_{\Omega_X} f_X(x) \log f_X(x) dx - \int_{\Omega_Y} f_Y(y) \log f_Y(y) dy. \end{aligned} \quad (8.74)$$

Which is the difference between the Shannon entropies:

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \quad (8.75)$$

The mutual information can be defined between any two sets of variables **X** and **Y**. It is an information-theoretic measure because it quantifies the amount of information that the set of variables **X** have about the set of variables **Y**. Indeed, dependency between variables ultimately means that a set of variables carries some information about the other set. The mutual information, $I(X;Y)$ in Definition 8.8, quantifies the reduction in uncertainty about **Y** produced by the knowledge of **X** (and vice-versa). Analogously, it is the amount of uncertainty about **Y** left by the full knowledge of **X**. It is a measure of dependency with larger values corresponding to larger dependency.

Definition 8.9 (Mutual information general definition). The **mutual information** is the difference between the full entropy of variable **Y** and the

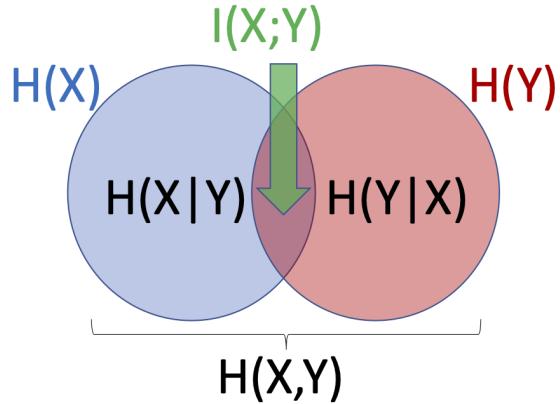


Figure 8.2 Pictorial representation of the mutual information and its relation with joint, marginal and conditional entropies.

entropy of the conditioned variable $\mathbf{Y}|\mathbf{X}$

$$I(\mathbf{X}; \mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}), \quad (8.76)$$

or analogously by substituting the expression Eq.7.37 for the conditional entropy

$$I(\mathbf{X}; \mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y}) - H(\mathbf{X}, \mathbf{Y}). \quad (8.77)$$

Indeed, if two sets of variables are independent, the conditioning has no effect and $H(\mathbf{Y}) = H(\mathbf{Y}|\mathbf{X})$, and $H(\mathbf{X}, \mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y})$ resulting into zero mutual information. Conversely, the larger is the dependence, the smaller will be the conditional entropy with respect to the unconditioned one resulting therefore in large mutual informations. A pictorial representation is provided in Figure 8.2.

Remark 8.9. The Shannon conditional entropy $H(\mathbf{Y}|\mathbf{X}) = H(\mathbf{X}, \mathbf{Y}) - H(\mathbf{X})$ (see Definitions 7.5 and 8.3) is given by $-\mathbb{E}_{\mathbf{XY}}[\log f_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y})]$. Note that it is not dependent on the values of $\mathbf{X} = \mathbf{x}$ and it is not equal to $-\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\log f_{\mathbf{Y}|\mathbf{x}}(\mathbf{Y})]$ which is instead a function of \mathbf{x} .

Example 8.5 (Mutual Information for the bivariate normal). When two variables are linearly dependent and their joint distribution is a bivariate normal, the mutual information directly maps into the correlation coefficient ρ_{XY} between the two variables. Indeed, from Eq.8.73 and the expres-

sion for the determinant in Eq.7.26, one has

$$I(X; Y) = -\frac{1}{2} \log(1 - \rho_{XY}^2). \quad (8.78)$$

Example 8.6 (Mutual Information for elliptical family linear case). Also in the multivariate case of two sets of variables, \mathbf{X} , \mathbf{Y} one has a simple relation that can be directly derived from the definition of multivariate entropies for the elliptical family (see Equation 7.23):

$$I(\mathbf{X}; \mathbf{Y}) = \frac{1}{2} \log \frac{|\Omega_{\mathbf{XX}}||\Omega_{\mathbf{YY}}|}{|\Omega_{\mathbf{ZZ}}|}. \quad (8.79)$$

with $\mathbf{Z} = (\mathbf{X}^\top, \mathbf{Y}^\top)^\top$. This is a generalization of Eq.8.78 which coincides with it in the case of two variables.

8.9.2 Higher order information share: interaction information

The mutual information between any two set of variables, \mathbf{X}_1 and \mathbf{X}_2 , is defined from Definition 8.9. It represents the information in common between the two sets; it is given by the difference between the entropies of the variables when considered separately with respect to the entropy of the joint variables, and it is a non-negative quantity. This concept can be generalized to more than two sets of variables by quantifying the difference between the entropies of different aggregations of variables. However, the interpretation of this quantity in terms of shared information becomes less intuitive. Indeed, for instance, it can be negative.

For n sets of variables $\mathbf{X}_1, \dots, \mathbf{X}_n$, a higher order version of the mutual information, sometimes called interaction information, or cross-information, is defined as [Rosas et al., 2019]

$$I_n(\mathbf{X}_1; \mathbf{X}_2; \dots; \mathbf{X}_n) = \sum_{k=1}^n (-1)^{k+1} \sum_{i_1 < \dots < i_k} H(\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_k}) \quad (8.80)$$

Where in the sum, the indexes $i_1 \dots i_k$ take values between 1 and n producing all combinations of terms without repetitions. One can see that for $n = 2$ the expression for the mutual information is retrieved:

$$I_2(\mathbf{X}_1; \mathbf{X}_2) = I(\mathbf{X}_1; \mathbf{X}_2) = H(\mathbf{X}_1) + H(\mathbf{X}_2) - H(\mathbf{X}_1, \mathbf{X}_2). \quad (8.81)$$

For three groups of variables \mathbf{X}_1 , \mathbf{X}_2 and \mathbf{X}_3 one has instead

$$\begin{aligned} I_3(\mathbf{X}_1; \mathbf{X}_2; \mathbf{X}_3) &= H(\mathbf{X}_1) + H(\mathbf{X}_2) + H(\mathbf{X}_3) \\ &\quad - H(\mathbf{X}_1, \mathbf{X}_2) - H(\mathbf{X}_1, \mathbf{X}_3) - H(\mathbf{X}_2, \mathbf{X}_3) \\ &\quad + H(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3) \end{aligned} \quad (8.82)$$

This interaction information obeys the inductive relation

$$I_n(\mathbf{X}_1; \mathbf{X}_2; \dots; \mathbf{X}_n) = I_{n-1}(\mathbf{X}_1; \mathbf{X}_2; \dots; \mathbf{X}_{n-1}) - I_{n-1}(\mathbf{X}_1; \mathbf{X}_2; \dots; \mathbf{X}_{n-1} | \mathbf{X}_n). \quad (8.83)$$

Indeed, $H(\mathbf{X}_1, \dots, \mathbf{X}_{n-1} | \mathbf{X}_n) = H(\mathbf{X}_1, \dots, \mathbf{X}_n) - H(\mathbf{X}_n)$, which substitute into Eq.8.80 verifies Eq.8.83.

By using this interaction information one can operate an exact expansion of the multivariate entropy

$$H(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{n-1}, \mathbf{X}_n) = \sum_i H(\mathbf{X}_i) - \sum_{i < j} I(\mathbf{X}_i; \mathbf{X}_j) + \sum_{i < j < k} I_3(\mathbf{X}_i; \mathbf{X}_j; \mathbf{X}_k) - \dots \quad (8.84)$$

note the alternate signs. This expansion could be quite insightful because it can be seen as an approximation where every term brings in the residual contribution of higher order mutual information between n sets of variables. These are higher order interactions that eventually became zero when the combination of groups of variables provides no extra information than the uncombined groups. From Eq.8.83 one can see that this is the case when conditioning to an extra variable makes no difference. The alterante signs makes interpretation complicated because it appears as if positive interaction information terms either increase or decrease the approximate entropy depending on the order.

Notice that Eq.8.84 must be considered with attention; indeed, if one considers all terms of the expansion, one might notice that the contribution at each order cancel the contribution at the previous order and adds the terms for that order only. In this respect, it appears to be nothing more than a nice notation trick. However, in Section 12.1.2 I'll discuss an interpretation in terms of sparse network representation, where not all terms are included, which can make this expansion insightful.

In general terms, the combination of variables can provide more or less information than the sum of the parts considered in isolation. The case when the combination provides overall less information is associated to the existence of a common information shared among the group of variables. This situation is referred as **redundancy** and produces positive contributions to the interaction information. The opposite case occurs when the information about a subset of variables can enhance the dependency among the others. This results in a negative contributions to the interaction information and is referred as **synergy**. For instance, in a system of three variable sets \mathbf{X}_1 , \mathbf{X}_2 and \mathbf{X}_3 , one has positive interaction information, $I_3(\mathbf{X}_1; \mathbf{X}_2; \mathbf{X}_3) > 0$, when the information of a third variable explains some of the dependency between the other two. This is the redundancy case. Conversely, a negative interaction information, $I_3(\mathbf{X}_1; \mathbf{X}_2; \mathbf{X}_3) < 0$, indicates that taking into account the information of a third variable increases the dependency between the other two, and this is the synergy case. An extreme example of redundancy is three equal variables $X_1 = X_2 = X_3$ where therefore one variable explains the other but if the two are constrained to the third then there is

no extra information provided. This implies $I(X_1; X_2|X_3) = 0$ but $I(X_1; X_2) > 0$ and therefore $I_3(X_1; X_2; X_3) > 0$. An example of synergy is the XOR function between two independent binary variables: $X_3 = X_1 \oplus X_2$. The three variables are pairwise independent, however the knowledge about two of them in combination provides the full information about the third. This implies $I_3(X_1; X_2|X_3) > 0$ but $I(X_1; X_2) = 0$ and therefore $I_3(X_1; X_2; X_3) < 0$.

Example 8.7 (Redundancy and synergy for three multivariate elliptical variables). Let me explore redundancy and synergy among three variables for the case of three multivariate elliptical variables. In this case, the expression for the higher order mutual information can be written explicitly in terms of correlation coefficients, namely

$$I(X_1; X_2) = -\frac{1}{2} \log(1 - \rho_{1,2}^2); \quad (8.85)$$

while

$$I_3(X_1; X_2; X_3) = +\frac{1}{2} \log \left(\frac{1 - \rho_{1,2}^2 - \rho_{1,3}^2 - \rho_{2,3}^2 + 2\rho_{1,2}\rho_{1,3}\rho_{2,3}}{(1 - \rho_{1,2}^2)(1 - \rho_{1,3}^2)(1 - \rho_{2,3}^2)} \right). \quad (8.86)$$

Where I made use of the following identities for multivariate normals (for other elliptical distributions they differ only by a constant that cancels out):

$$\begin{aligned} H(X_i) &= \frac{1}{2} \log(\sigma_i^2) + \frac{p_i}{2} + \frac{p_i}{2} \log(2\pi) \\ H(X_i, X_j) &= \frac{1}{2} \log(\sigma_i^2 \sigma_j^2 (1 - \rho_{i,j}^2)) + \frac{p_i + p_j}{2} + \frac{p_i + p_j}{2} \log(2\pi) \\ H(X_i, X_j, X_k) &= \frac{1}{2} \log(\sigma_i^2 \sigma_j^2 \sigma_k^2 (1 - \rho_{i,j}^2 - \rho_{i,k}^2 - \rho_{j,k}^2 + 2\rho_{i,j}\rho_{i,k}\rho_{j,k})) \\ &\quad + \frac{p_i + p_j + p_k}{2} + \frac{p_i + p_j + p_k}{2} \log(2\pi) \end{aligned} \quad (8.87)$$

- Examples of **redundancy** can be obtained for $\rho_{1,2}, \rho_{1,3}, \rho_{2,3}$ all equal and positive with larger positive values of $I_3(X_1; X_2; X_3)$. For instance for $\rho_{1,2} = \rho_{1,3} = \rho_{2,3} = \rho$ one has $I_3(X_1; X_2; X_3) = 0.006$ for $\rho = 0.2$ and instead $I_3(X_1; X_2; X_3) = 0.24$ for $\rho = 0.7$.
- In this ternary system, examples of **synergy** are always observed when one of the correlations is zero. For instance, for $\rho_{1,2} = 0$ and $\rho_{1,3} = \rho_{2,3} = 0.7$ gives $I_3(X_1; X_2; X_3) = -1.3$. But also other combinations with all correlations different from zero can return negative values.

Notice that, while $I(X_1; X_2)$ is independent from the sign of the correlation, instead $I_3(X_1; X_2; X_3)$ depends on their signs. For instance, for $\rho_{1,2} = \rho_{1,3} = 0.2$ and $\rho_{2,3} = -0.2$, one has $I_3(X_1; X_2; X_3) = -1.2 \cdot 10^{-2}$ (instead of $6 \cdot 10^{-3}$ when the three correlations are all positive), in this case, the change of sign of one of the correlations changes both value and sign of the third-order mutual information.

Example 8.8 (Redundancy and synergy for more than three variables).

Stepping up the order of the mutual information, one retrieves similar results. Essentially, cliques of similarly correlated variables tend to have positive interaction information, they are ‘redundant cliques’, while ‘synergic cliques’ with negative interaction information are typical of heterogeneous interactions with a mix of large and low or negative correlations.

For instance, for four variables all equally correlated with correlation coefficients $\rho = 0.2$ one has $I_4(X_1; X_2; X_3; X_4) = 2.5 \cdot 10^{-3}$ while if one of the correlations is set to -0.2 then $I_4(X_1; X_2; X_3; X_4) = -1.5 \cdot 10^{-3}$. Similarly, for five variables all equally correlated with correlation coefficients $\rho = 0.2$ one has $I_5(X_1; X_2; X_3; X_4; X_5) = 1.2 \cdot 10^{-3}$ while if one of the correlations is set to -0.2 then $I_4(X_1; X_2; X_3; X_4) = -4.0 \cdot 10^{-5}$.

8.9.3 Total correlation, residual entropy and O-information

For more than three variables the formulation of redundancy and synergy becomes increasingly complicated. Let me here briefly sketch an approach by Rosas et al. [2019] which I find very insightful. First of all, one wants to quantify the difference between a system of variables, $\mathbf{X} = (X_1, \dots, X_n)^\top$, represented as all independent variables with respect to the actual interacting system of variables. This is quantifiable from the difference between the sum of the entropies of the variables and the joint entropy, which is called **total correlation**

$$C_n(\mathbf{X}) = \sum_{k=1}^n H(X_k) - H(\mathbf{X}), \quad (8.88)$$

For two variables ($n = 2$), this quantity is the mutual information. In general, it is a non-negative quantity measuring the strength of the overall dependency between the variables. It is said that this is a measure of the strength of ‘collective constraints’.

Another important quantity is the amount of independent information carried by each variable that can be quantified by the entropy of a variable conditioned to all others: $H(X_i|\mathbf{X}_{\mathbf{k}\setminus i})$ where $\mathbf{k} = (1, 2, 3, \dots, n)^\top$ is the index set and $\mathbf{X}_{\mathbf{k}\setminus i} = (X_1, X_2, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$. The sum of these terms has been named residual entropy or erasure entropy [Abdallah and Plumley, 2012, Verdu and Weissman, 2008] and the difference between such a sum and the joint entropy

$$D_n(\mathbf{X}) = H(\mathbf{X}) - \sum_{i \in \mathbf{k}} H(X_i|\mathbf{X}_{\mathbf{k}\setminus i}), \quad (8.89)$$

is sometimes called **excess entropy** [Olbrich et al., 2008]. Intriguingly, for two variables this quantity is also equal to the mutual information. In general, above the two variables it quantifies the independent contributions of the variables to the joint entropy, it is a measure of ‘shared randomness’.

The total correlation $C_n(\mathbf{X})$ and the excess entropy $D_n(\mathbf{X})$ are quantifying two

different aspects of collective information. One can see the first more as a global dependency from shared information and therefore a global measure contributing to redundancy while the other as a measure contributing to synergy. In Rosas et al. [2019] the authors proposed to use the difference between $C_n(\mathbf{X})$ and $D_n(\mathbf{X})$ to quantify the redundancy/synergy imbalance in a system and they called such a difference **O-information**. For two variables this quantity is zero, for three variables it coincides with $I_3(X_i, X_k, X_j)$. In general, when the O-information is positive they say that the system is redundancy dominated, while when the O-information is negative they say it is synergy dominated. The O-information is different from the interaction information $I_n(\mathbf{X})$.

Examples of prevalently redundant systems of variables with positive O-information are factor models where a set of random variables depend on a certain number of common factors. Instead, random systems of correlated variables tend to have negative O-information.

In Sections 12.1.2 and 17.2-17.4 I will return to this topic when I address the issue of the representation of multivariate models via networks and the construction of such networks from data. For some further readings on this topic I suggest: Ting [1962], Yeung [1991], Matsuda [2000], Bell [2003], Bettencourt et al. [2008], Williams and Beer [2010, 2011].

Remark 8.10. The papers by Williams and Beer [2010, 2011], provide an interesting definition of redundancy and synergy as two non-negative terms whose difference is the interaction information. In general, both effects are simultaneously present and therefore zero interaction information does not necessarily imply absence of interaction. Furthermore, for this reason, Venn diagram representations, such as the one for the pairwise mutual information in Fig.8.2, cannot be directly extended to three or more variables.

8.10 Lagged correlations

In processes that evolve with time $\mathbf{X}_t = (X_{1,t}, X_{2,t}, \dots, X_{p,t})^\top$ with $t = 1..T$ (i.e. stochastic processes, see Doob and Doob [1953] and Chapter 9) some signals at a given instant in time, t , might have information about the signal at a following instant in time $t + h$. This can be due to some underlying causality between the two (as we shall explore in Chapter 10) or by other factors such as delay in the reception/processing of the signal from a common factor. In general, dependency between variables in such processes might need to be established not only between simultaneous variables (i.e. $\text{Corr}(X_{i,t}, X_{j,t})$, sometimes called cross-correlations), but also between variables at different times. In terms of correlations this is: $\text{Corr}(X_{i,t'}, X_{j,t})$, which are the so-called ‘lagged correlations’ or lagger cross-correlations where the ‘lag’ is $h = t - t'$.

Definition 8.10 (Lagged correlations). The cross-correlation between two random variables shifted in time by a given time ‘lag’ h is called **lagged correlation**:

$$\rho_{X_i X_j}(h) = \text{Corr}(X_{i,t-h}, X_{j,t}). \quad (8.90)$$

One might note that when the two variables are the same, then $\rho_{X_i X_i}(h)$ is the autocorrelation. Indeed sometimes lagged correlations are called cross-autocorrelations. One might also notice that at lag $h = 0$ the lagged correlation coincides with the correlation $\rho_{X_i X_j}(h = 0) = \rho_{i,j}$. In a stationary process, one must have $\text{Corr}(X_{i,t-h}, X_{j,t}) = \text{Corr}(X_{i,t}, X_{j,t+h})$ and, equivalently, $\text{Corr}(X_{i,t}, X_{j,t'}) = \rho_{X_i X_j}(t' - t) = \rho_{X_i X_j}(t - t')$.

This definition can be directly extended to all other measures of dependency that I have introduced so far just by considering measures between lagged variables.

If one process anticipates the other at a given lag h one should measure a significant dependency between $X_{i,t-h}$ and $X_{j,t}$. In this case one can say that X_i is leading X_j at lag h . The presence of lagged dependency is not a demonstration of causality (see Chapter 10 for a full discussion). Indeed, there can be a very large number of factors that make a signal anticipate another signal without any causal relation. However, any temporal causality must be related with lagged dependency and the absence of such lagged dependency is a strong indication of absence of temporal causality.

8.11 Data-driven modeling tutorial

<https://github.com/FinancialComputingUCL/DataDrivenModeling/Ch8>

The tutorial for this Chapter covers various topics on dependency including: a discussion of the relation between Pearson correlations and linear regression (Example 8.2); a comparison between correlations and partial correlations; a comparison between Pearson correlations, rank correlations and mutual information; the use of inverse covariance in the Markowitz portfolio theory (Example 8.3).

Exercises

- Retrieve the identity:

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (F_{X,Y}(x,y) - F_X(x)F_Y(y)) dx dy = \mathbb{E}((X - \mu_X)(Y - \mu_Y)).$$
- Considering the regression $Y = \beta_0 + \beta_1 X + \epsilon$, demonstrate that if $\text{Cov}(Y, \epsilon) = 0$ then $\text{Corr}(Y, X) = 1$.

- By using fractional ranking compute Spearman- ρ correlation for $\hat{\mathbf{x}} = (3.50, 11.51, 1.02, 9.44, 6.24, 11.51, 0.49, 2.78, 3.50)$ and $\hat{\mathbf{y}} = (13.35, 33.71, 2.56, 34.26, 24.35, 40.19, 4.15, 4.15, 13.35)$.
- Considering a multilinear regression, between two sets of random variables belonging to the elliptical family, derive the link between the generalized variance of the error, $|\text{Cov}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})|$, and the conditional entropy $H(\mathbf{X}|\mathbf{Y})$.
- Derive the Markowitz portfolio weights (see Example 8.3) for the case of uncorrelated assets.
- Show that $\min_c (\mathbb{E}((X - c)^2)) = \mathbb{E}(X)$.
- Show that $I(X; Y)$ between two bivariate normal random variables, \mathbf{X} and \mathbf{Y} can be approximated as $I(X; Y) \simeq 1/2 \rho_{X,Y}^2$. Discuss the possible generalization of this approximation to non-normally distributed variables.
- Discuss the link between the mutual information $I(\mathbf{X}; \mathbf{Y})$ between two sets of random variables, \mathbf{X} and \mathbf{Y} , belonging to the elliptical family and the generalized variance of the error of the multilinear regression between \mathbf{X} and \mathbf{Y} .

Stochastic processes and scaling laws

Definition 9.1 (Stochastic process). A **stochastic process** is a collection of random variables, defined on a common probability space, called the **state space**, and indexed by a set of numbers, $\mathbf{t} = (t_1, \dots, t_T)$ which is called the **index set**

$$\mathbf{X}_{\mathbf{t}} = (X_{t_1}, \dots, X_{t_T}). \quad (9.1)$$

The indices set t_1, \dots, t_T represent points at different stages of ‘evolution’ of the process. When the variables are multivariate, $\mathbf{X}_{t_s} \in \mathbb{R}^p$ then the process is denoted as $\mathbf{X}_{\mathbf{t}} \in \mathbb{R}^{p \times T}$.

A stochastic process is therefore a multivariate system of $p \times T$ random variables. The variables follow a non-commutable sequence and are not independent. When the index represents time-steps, the process can be seen as the ‘evolution’ of a random variable through time. In this chapter, I will refer only to processes evolving in time. However, the approaches and results that I present are universal, independent on the meaning of the index set. Normally T is large, often one seeks results for $T \rightarrow \infty$, and this is called asymptotic limit. We are therefore dealing with a very high-dimensional system of dependent variables. The dependency in the time-direction is often referred to as autocorrelation or, more generally, autodependency. This is a special kind of multivariate system, quite different in nature and meaning from the dependency between different random variables that I have discussed so far in this book. Therefore, it needs to be studied with specifically devised tools.

9.1 Stationarity

Let me start from basic. A process is stationary when its probability distribution does not change with time.

Definition 9.2 (Strict stationarity). Given a univariate stochastic process $\{X_{\mathbf{t}}\}$ with cumulative distribution function $F(X_{t_1}, \dots, X_{t_T})$, one says it is

strictly stationary if

$$F(X_{t_1}, \dots, X_{t_T}) = F(X_{t_1+\tau}, \dots, X_{t_T+\tau}) \quad (9.2)$$

for all $T \in \mathbb{N}$, $\tau \in \mathbb{N}$. This definition applies equivalently for a multivariate system $\mathbf{X}_t = (X_{1,t}, \dots, X_{p,t})$.

Sometimes lower forms of stationarity are used such as requiring that the moments of the distribution do not change when the observation window is shifted in time.

It must be noticed that, real complex systems typically evolve and adapt and consequently the statistical properties of real complex processes change continuously in time. Therefore real process are often non-stationary. However, in some cases, non-stationary processes, can be made stationary by operating some transformations on the variables. For instance, if a random variable has a linear trend that makes its mean increasing constantly in time, it could be sufficient to eliminate such a linear trend to make the process stationary. In the same way, more complex trends or regular changes, can be eliminated by subtracting polynomial functions or periodic functions that capture seasonalities. Another common way to make a series of variables, \mathbf{X}_t , stationary is by considering the differences between consecutive variables, which are called ‘returns’.

Definition 9.3 (Returns). The difference between two stochastic variables at times t and $t - s$ are called **returns**

$$\mathbf{R}_t = \mathbf{X}_t - \mathbf{X}_{t-s}. \quad (9.3)$$

The interval size s is often called ‘lag’ or ‘horizon’. In this Chapter I consider only integer values of s .

For instance, in finance, the price of an asset (i.e. a stock price) is not a stationary variable, nor is its logarithm. However, the difference between the logarithm of the price at a given time t and the logarithm of the price at a previous time $t - \tau$ (called log-returns, see Definition 9.4) can be often considered stationary for many practical purposes.

Definition 9.4 (Log-returns). The **log-returns** are the differences between logarithms of prices at two different times t and $t - s$:

$$r_t = \log Price_t - \log Price_{t-s}. \quad (9.4)$$

They represent rates of changes of the prices during the interval $s > 0$ and can be directly compared with interest rates. The interval size s is often called ‘lag’ or ‘horizon’.

For the rest of this chapter, unless otherwise mentioned, I’ll assume we are deal-

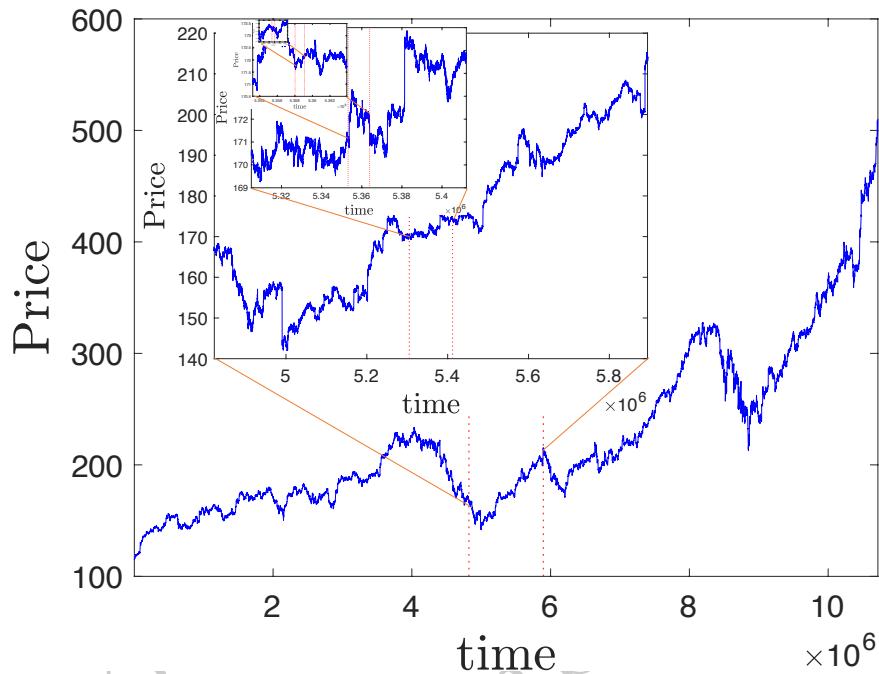


Figure 9.1 Example of ‘zooming’ into a stochastic proces. The figure reports AAPL stock prices in the period 03/01/2017 to 24/08/2020 with observations registered every second (see Section 16.1). The main figure reports the whole period; its inset is a period 10 times smaller (indicated with dotted vertical lines); the inset inside the inset is a period 100 times smaller; and so on to the last inset that is spanning a period 10,000 time smaller.

ing with stationary variables. Statistical tests to assess stationarity are described in Section 16.3.

9.2 Scaling laws

Another aspect, specific to stochastic processes evolving in time, concerns the change of the statistical properties by aggregation of the data over time-windows of different sizes. This is like ‘zooming’ in or out of a signal and look at details at different time-scales (see an example for the AAPL stock price in Fig.9.1. It turns out that often signals have similarities or affinities at different time-scales and there are laws that map the statistical properties of stochastic signals across time-scales.

Definition 9.5 (Scaling). In general terms, I indicate with the term **scaling** changes in the system’s properties with size. For instance, the area of a circle

with diameter Δ scales as Δ^2 . Similarly, the volume of a sphere of diameter Δ scales as Δ^3 .

From a probabilistic modelling perspective, in this book I use the term scaling to indicate how the statistical properties of a stochastic process change when observed at different levels of granularity (i.e. at different *scales*). Specifically, scaling laws describe how the statistical properties of the variables at given times t are shaping the statistical properties of their sums (or their means) over time-intervals $(t-s, t]$. Indeed, the aggregation over a time-interval of size s is changing the time-horizon or the ‘scale’ of the problem. For instance, if one has a signal at frequency of seconds, its mean over 60 second will provide a measure of the ‘scaled’ signal over minutes.

I have already discussed in previous chapters that the aggregation of random variables changes their statistical properties. For instance, the sum of two variables has different statistical properties than the variables themselves (see Example 5.1). The law of change for some properties is very simple. For example, the mean of a sum of random variables must be equal to the sum of the means. However, beyond the mean, things become less simple.

Example 9.1. The ability to infer the statistical properties of a sum of variables from the properties of individual variables can be extremely powerful in many practical cases. For instance, the price of a financial product fluctuates continuously depending on the flow of ask, bids and the volumes involved. If one looks at the relative variation of the present price with respect to the price one minute earlier (the 1 min return) one is likely to observe small changes; if one instead looks at the variation from yesterday’s price (the 1 day return) the variation is likely to be larger; and, similarly, the 1 week return, the 1 month return and the 1 year return will have increasingly larger likelihood of large changes. Understanding how the probability of a given price variation changes with the time-horizon is very important in the estimation of the risk of an investment. In this case, one wants to see if there is a pattern describing the ‘scaling’ of the probability distribution function of the returns at the different time-scales. If that pattern exists, then one could infer the risk of investments over horizons of years simply by estimating the statistics over a few minutes of observation. I shall show in this chapter that, although it is not straightforwardly simple, such scaling laws exists and can be used for this purpose.

In general, in a stochastic process the signal X_t is not the variable of interest, one rather looks at the returns over a given time-horizon $s > 0$, $R_{t,s} = X_t - X_{t-s}$ (see Definition 9.3). The general aim is to learn the probability density function of the signal at all time-scales:

$$f_{t,s}(R_{t,s}) \quad (9.5)$$

where I indicated explicitly that the distribution depends on the time-scale s and, in general, in absence of stationarity for $R_{t,s}$, it might even be dependent on t . The scaling laws are the laws that govern the transformation of $f_{t,s}(R_{t,s})$ with s . They are the maps between the probability distributions at different time horizons. An example of such changes of statistical properties with the scale, is reported, for log-returns in Fig.16.1 in Chapter 16, where how to measure and quantify these scaling laws is discussed.

The return at a given time-horizon s is aggregating the changes of the signal over the time window $(t-s, t]$, indeed, in general, it is a sum of returns at a shorter scale:

$$R_{t,s} = X_t - X_{t-s} = X_t - X_{t-1} + X_{t-1} - X_{t-2} + \dots + X_{t-s} = \sum_{u=t-s+1}^t R_{u,1}. \quad (9.6)$$

Consequently, the study of the scaling laws corresponds to the study of the changes in the probability density function of the random variable $R_{t,s}$ as a consequence of aggregation. The general problem is to find the transformation laws for the expression for the probability density function at different levels of aggregation.

9.3 The fractal dimension of signals

A fractal is an object where sub-parts at different scales have the same statistical character.

Remark 9.1. The concept of **fractals** comes from Benoit B. Mandelbrot [Mandelbrot and Mandelbrot, 1982] who first coined the word fractal to define signals or objects that scales with non-integer exponents.

By looking at stochastic processes one might notice that some have similarities across different time-scales. They can indeed be considered examples of fractals. We are probably more accustomed to examples of fractals such as the Sierpinksky gaskets or the cauliflower where if one looks at parts of smaller or larger dimensions one observe similar shapes and geometries and they are indeed called self-similar. The fractal nature of signals can be also discovered by observing that by looking at two plots of the same process (i.e. X_t vs. t), the first over a very large time window and the other over a smaller sub-part, the two look similar with ups and downs through a rugged curve. However, they are not self-similar, but rather they are self-affine. Indeed, differently from the previous geometric examples, the x -direction (time) and the y -direction do not scale in the same way.

For a stochastic signal evolving in time, the overall variations over a small interval are expected to be smaller than the variations observed over a larger interval. Namely, one expects

$$\mathbb{E}(|X_t - X_{t-s}|) > \mathbb{E}(|X_t - X_{t-s'}|) \quad \text{for } s > s'. \quad (9.7)$$

However, by ‘scaling’ them appropriately, they can be compared:

$$\mathbb{E}(|X_t - X_{t-s}|) \simeq \left(\frac{s}{s'}\right)^H \mathbb{E}(|X_t - X_{t-s'}|). \quad (9.8)$$

I shall show shortly that the scaling exponent H is directly related with the fractal nature of the signal.

From a geometrical perspective one can see that a straight line has dimension one and any smoothly curved lines have also dimension one. However, one can figure out that a rough curve covers a larger portion of the plane than a smooth line and consequently its dimension must be larger than one (if its rough nature continues at all scales). The upper limit for the dimensionality of a line on the plane is associated to an extremely rough curve that homogeneously covers the entire plane. Such a curve would have dimension two, but this limit is not reachable. Indeed, the fractal dimension for a continuous curve can have values in the interval $D_f \in [1, 2]$. If the curve is not continuous then the dimension can go below one to zero.

A way to quantify the fractal dimension, D_f , is the so-called ‘box counting’ approach where one can imagine to divide a set into a quantity of smaller subparts (the ‘boxes’). For any given scale (the box-size) one needs a larger number of boxes to cover an object which has larger fractal dimension.

Definition 9.6 (Box counting fractal dimension). Consider a set of size Δ (for instance a signal over a time-interval Δ). Suppose to cover the set with N_s boxes of size s . The number of boxes needed to cover the set must increase when the size decreases. The **box counting fractal dimension**, D_f , is related with the way this number changes with the size:

$$D_f = \lim_{s \rightarrow 0} \frac{\log N_s}{\log(\Delta/s)}. \quad (9.9)$$

Example 9.2. One might notice that, for non-fractal continuous objects, the previous definition of the fractal dimension in terms counting boxes, coincides with the common definition of dimension. For instance, if one divides a line of length Δ into segments of size s their number will be $N_s = \Delta/s$ and Eq.9.12 yields to the fractal dimension $D_f = 1$. Analogously, for a continuous two dimensional set of size Δ one has $N_s = (\Delta/s)^2$ and $D_f = 2$. In general, for continuous non-fractal objects in D dimensions the number of boxes scales as $N_s = (\Delta/s)^D$. But this procedure can be applied to any other set with self-affine properties.

The box counting fractal dimension is known also as Minkowski–Bouligand dimension. There are other ways to define dimensions in fractals and they coincide only under some conditions Mandelbrot [1977]. In this book I will refer only to the box-counting dimension.

Remark 9.2. There are several definitions of fractal dimensions which are relevant or useful in different contexts and they do not always coincide Metzler and Klafter [2000]. One definition which is particularly meaningful in the context of this book is the **information dimension**, introduced in Balatoni and Rényi [1956] and Rényi [1959a] as rate of increase of the Shannon entropy of a random variable with respect to the discretization scale

$$D_I = \lim_{n \rightarrow \infty} \frac{\mathbf{H}_0(\xi_n)}{\log n}. \quad (9.10)$$

Where n is a discretization factor which can be related with the inverse of the scale with $n = \Delta/s$, while ξ_n is the discretization, at scale s , of a continuous variable x . Conceptually, this dimension is very similar to the box counting one because the Shannon entropy is the logarithm of the number of configurations at a given scale s . The entropy increases when s decreases because, the smaller the s the finer is the description of the signal and therefore the larger its information content.

The fractal dimension tells us how the typical range of values of a random variable scale with the characteristic size used to observe them. If one analyzes a process at different scales, the changes in the signal increases by increasing the scale. It turns out that, in many general and practical cases, these changes can be described as a power law:

$$\mathbb{E}(|X_t - X_{t-s}|) \propto s^H, \quad (9.11)$$

with

$$H = 2 - D_f. \quad (9.12)$$

This relation between the exponent H and the fractal dimension D_f can be directly inferred by using the box counting reasoning. Indeed, by looking at the plot of a signal over a period equal to the size s of a ‘box’, the curve must occupy a portion of the plane which has an interval equal to s on the x -axis, while on the y -axis, from Eq.9.11, one expects in average an interval proportional to s^H . So the area occupied by the signal is $s \times s^H$. Over the whole period Δ one has Δ/s of such intervals of size s and the area occupied by the signal is $s \times s^H (\Delta/s)$. The boxes have area s^2 and therefore number of boxes required to cover the signal over the period Δ is proportional to $N_s \propto s \times s^H (\Delta/s)/s^2$. When substituted into the box-counting definition of fractal dimension (Definition 9.6) it yields indeed to: $D_f = 2 - H$.

9.4 Random walk processes

Possibly the simplest kind of stochastic process is the one resulting from a cumulative sum of i.i.d. variables, which is called random walk.

The term “random walk” was first used by Karl Pearson in 1905. In a letter to Nature (consisting only of a few lines) Pearson [1905]. In the letter, he asked to the community which is the distance travelled by a man who walks a step, then turns a random angle and walks another step, repeating the process n times; the “drunkard’s walk”. A related question was originally posed to Pearson by Sir Ronald Ross (Noble prize in 1902 for identifying the mosquito transmission of malaria) who, in 1904, asked Pearson to calculate the diffusion speed of mosquitos in a forest. The Pearson’s question was answered by Lord Rayleigh, who had already solved the problem in a more general form for vibrations with unit amplitude and arbitrary phase. In the same year, Albert Einstein published his paper on Brownian motion [Einstein, 1905] which he modeled as a random walk driven by collisions with gas molecules. Einstein did not seem to be aware of the related works of Rayleigh and Pearson. The next year, Smoluchowski [Smoluchowski, 1906] also published very similar ideas, again independently.

Actually, Pearson, Rayleigh, Einstein and Smoluchowski were all unaware that five years earlier an economics student, Louis Bachelier, introduced in his thesis a random walk model to describe the variation of stock prices [Bachelier, 1900]. He, correctly, pointed out relations with the theory of errors and the normal distribution. However, Bachelier’s work was not immediately recognized as important and it took several decades before his contribution to the theory of stochastic processes started to become known and appreciated. He is now considered a pioneer and the forefather of financial mathematics.

The mathematical formulation of a random walk process is very simple: the value of a variable at time-step t is equal to its value at time-step $t - 1$ plus a random term

$$X_{t+1} = X_t + \eta_t , \quad (9.13)$$

with $X_{t=0} = X_0$ and $t = 1, \dots, T$. In its original formulation the ‘step’ or ‘random noise’ terms η_t were discrete $+1$ or -1 moves. However, it is straightforward to generalize it using random noise terms η_t that are i.i.d. random variables belonging to some probability distribution. The noise terms are normally assumed to have zero mean $\mathbb{E}(\eta_t) = 0$ and their independence implies $\mathbb{E}(\eta_t \eta_{t'}) = \mathbb{E}(\eta_t) \mathbb{E}(\eta_{t'}) = 0$ for all $t \neq t'$. The variable X_t described by the random walk’s Equation 9.13 is a stochastic process.

Remark 9.3. The random walk process in Equation 9.13 is directly related with the so called **Wiener process** which is its continuous-time limit. In its original formulation the noise term is normally distributed. Extensions to other distributions of the noise term have been studied and the general class of this processes take the name of **Lévy processes**. It can be formalized

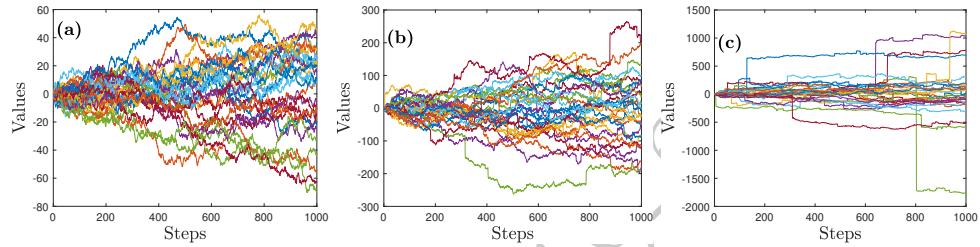


Figure 9.2 Plots of random walk processes starting from $X_0 = 0$ and proceeding for $T = 1,000$ steps with three different noise terms. (a) normal noise with zero mean and unitary standard deviation. (b) symmetric power law noise with exponent $\alpha = 2$. (c) symmetric power law noise with exponent $\alpha = 1.5$. Each panel reports 30 realizations.

as:

$$dX_t = a dt + b dW_t \quad (9.14)$$

When a and b also depend on X and t it takes the name of **Ito process**. It must be noticed that the limit to the continuum of the random walk process has some subtle issues that must be carefully accounted for.

Random walks are not stationary, because the statistical properties of the variable X_t change with time t . Indeed, the longer is the walking time the more likely is to find the walker at a larger distance from the starting point. In terms of probability distribution of the variable X_t , this means that the probability distribution of the walker position gets broader and broader with time.

The random walk process has the appeal to be a very simple model for a stochastic process. We shall see shortly that such a simplicity allows to retrieve the scaling laws analytically. However, it is well established that several real stochastic processes, such as the variation of stock prices, have a more complex dynamics than random walks.

The random walk model in Equation 9.13 assumes independent steps, discrete times and it also considers equally spaced time steps. One can extend the model to consider (auto)correlated steps and/or continuous times or/and non uniform spacing between time-steps. However, with a few noticeable exceptions, these extensions make the model no longer analytically treatable.

Example 9.3 (Random walks). Let me here discuss three random walk processes generated with Eq.9.13 with three different noise terms.

1. Normal noise with zero mean and unitary standard deviation $\eta_t \sim \mathcal{N}(0, 1)$. Thirty realizations of this process are reported in Fig.9.2(a);
2. Symmetric power law noise, $f(\eta_t) \propto |\eta_t|^{-\alpha}$, with exponent $\alpha = 2$. Thirty realizations are reported in Fig.9.2(b);

3. Symmetric power law noise, $f(\eta_t) \propto |\eta_t|^{-\alpha}$, with exponent $\alpha = 1.5$. Thirty realizations are reported in Fig.9.2(c).

All examples start from $X_0 = 0$ and are reported for the same number of steps $T = 1,000$. One might note that the power law distributions make the process to have large isolated jumps that are larger in size when for the lower exponent ($\alpha = 1.5$, Fig.9.2(c)). This is indeed a visible consequence of fat tails that make larger fluctuations more likely.

9.5 Scaling laws of random walk processes

I have already introduced in Chapter 2 the central limit theorem which states that the sum of independent identically-distributed variables will tend to be distributed accordingly with a limiting distribution that is the normal distribution if the variance is finite (see Section 5.1) or the Lévy-alpha stable distribution if the variance is not defined (see Section 5.3).

By re-writing the random walk Equation 9.13 one can observe that

$$\begin{aligned} X_t &= X_{t-1} + \eta_{t-1} \\ &= X_{t-2} + \eta_{t-2} + \eta_{t-1} \\ &\vdots \\ &= X_1 + \eta_1 + \eta_2 + \cdots + \eta_{t-1}. \end{aligned}$$

Putting in evidence that, in a random walk process, the resulting value of X_t is the sum of t , i.i.d., random variables. Analogously, any difference between the value of the variable at a given time t and the value at a previous time $t-s$, $R_{t,s} = X_t - X_{t-s}$, which is called ‘return’ at time-horizon s , is the sum of s , i.i.d., random variables:

$$R_{t,s} = X_t - X_{t-s} = \sum_{k=t-s}^{t-1} \eta_k. \quad (9.15)$$

If the noise terms, η_t , are i.i.d. random variables from a strictly stable distribution with tail exponent $\alpha \leq 2$ and zero mean (a Lévy α -stable, see Definition 5.5 and Equation 5.24), then, for any s , the distribution of $R_{t,s} = X_t - X_{t-s}$ is also a strictly stable distribution and it must scale with s as:

$$f_s(r) = \frac{1}{s^{1/\alpha}} f_1\left(\frac{r}{s^{1/\alpha}}\right). \quad (9.16)$$

Essentially what happens is that the distribution of $R_{t,s}$ become broader with growing s and such a broadness increases as $s^{1/\alpha}$ (i.e. the scaling exponent in Eq.9.11 is $H = 1/\alpha$, for $\alpha \leq 2$). Equation 9.16 indicates that by rescaling the random variable $R_{t,s}$ by a factor $s^{-1/\alpha}$, all distributions, at any time scale s , will collapse on the same distribution that is the one for η_t . From this scaling law, it is straightforward to compute the scaling properties, which are the ones for

the parameters of the Lévy α -stable distribution. Specifically, in this process the location is zero, whereas the scale parameter changes with the time horizon s as

$$c_s = s^{1/\alpha} c_1 \quad (\alpha \leq 2) \quad (9.17)$$

and the other parameters rest unchanged. Analogously, the quantiles scale as

$$Q_s(\gamma) = s^{1/\alpha} Q_1(\gamma). \quad (9.18)$$

And similarly the moments scale accordingly. The expected values of any power k (with $k \in \mathbb{R}$, not necessarily integer) of the returns will scale with the time-horizon as:

$$\mathbb{E}(|R_{t,s}|^k) = s^{k/\alpha} \mathbb{E}(|\eta|^k) \quad (9.19)$$

and diverge for $k \geq \alpha$. However, one must note that this holds only for the case when the noise terms are random variables from a strictly stable distribution with $\alpha \leq 2$ and $k < \alpha$.

By comparing Eq.9.19 with Eq.9.11 and Eq.9.12 it is clear that, for random walk processes with i.i.d. noise term η_t belonging to a strictly stable distribution with tail exponent $\alpha \leq 2$, the fractal dimension is:

$$D_f = 2 - \frac{1}{\alpha}. \quad (9.20)$$

Thus for the case of normally distributed noise ($\alpha = 2$) one has $D_f = 3/2$. When fat-tailed Lévy alpha-stable distributions are involved (see Definition 5.6), and $\alpha \in [1/2, 2]$, one can span the range of fractal dimensions from $3/2$ to zero. For $\alpha < 1/2$ the fractal dimension cannot be any longer defined with the box-counting method.

When the noise terms η_t do not follow a stable distribution, then the scaling law Eq.9.16 will hold only asymptotically at large horizons, when the aggregated distribution converges towards a stable distribution. In this case, accordingly with the central limit theorem and its generalization (see Theorems 5.1 and 5.3), the convergence will be either to a Lévy α -stable or to a normal distribution depending whether the variance is undefined or finite. The overall process of broadening of the distribution of X_t with t is similar to the one just described for the case with a noise term belonging to a stable distribution. However, during the transition towards the asymptotic stable distribution the scaling of the distribution will no longer be captured by a single scaling factor of the variable. Typically, when the noise is fat-tailed but with $\alpha \geq 2$, and therefore the variance is finite and the asymptotical convergence is towards the normal distribution, the body of the distribution of the aggregated variables converges quite rapidly towards the normal but the tails persist affecting the scaling. Different parts of the distribution will scale with different factors, with the body scaling faster than the tails. I have already presented an example (see Example 5.4) of this different scaling in Section 5.2.2 where I discussed the sum of i.i.d. variables and its convergence towards the normal distribution. I shall further discuss this dishomogeneous scaling in Chapter 16, where the estimation of the scaling laws is addressed.

9.6 Self-affine, uniscaling processes

Random walks with noise terms that have strictly stable distributions are simple stochastic processes and they belong to a category called uniscaling, or self-affine.

Definition 9.7 (Uniscaling or Self affine). A process X_t is **uniscaling**, or self affine, if there exist a coefficient $H > 0$ such that for all $s > s' > 0$

$$R_{t,s} \stackrel{d}{=} \left(\frac{s}{s'}\right)^H R_{t,s'}, \quad (9.21)$$

with $\stackrel{d}{=}$ indicating **equality in distribution**. The coefficient H is known as self-affine index or scaling exponent.

This exponent is often named Hurst exponent and, in the present case of uniscaling processes, it also coincide with the Hölder exponent (see Section 9.7.1). It was noticed by Mandelbrot that it is “Serendipitously, the names of Hurst and Hölder shared the same initial letter” [Mandelbrot, 2013].

In terms of the probability density function, Equation 9.21 can be written as

$$f_s(r) = \left(\frac{s}{s'}\right)^{-H} f_{s'}\left(\left(\frac{s}{s'}\right)^{-H} r\right) \quad (9.22)$$

In uniscaling processes a single scaling factor, $(s/s')^{-H}$, is sufficient to rescale all distributions at all scales onto one only. For random walks with strictly stable noise terms, I have already identified this scaling law of the distribution (Equation 9.16) which is indeed identical to the above definition with

$$H = \frac{1}{\alpha} \quad (9.23)$$

for tail exponent $\alpha \leq 2$.

I have already observed (see Eq.9.19) that this scaling law implies as a consequence that the k th central moments of the distribution must scale as

$$\mu(k, s) = \mathbb{E}(|R_{t,s}|^k) = \left(\frac{s}{s'}\right)^{kH} \mathbb{E}(|R_{t,s'}|^k), \quad (9.24)$$

which therefore implies that

$$\mu(k, s) = \left(\frac{s}{s'}\right)^{kH} \mu(k, s'). \quad (9.25)$$

Despite Eq.9.24 being identical to Eq.9.19 it is, in this case, applicable to any uniscaling process and not exclusively to random walks that are one particular instance of uniscaling process.

Random walks and uniscaling processes are nice and simple models but they do not capture most of the essential properties of real signals. For instance, the independence between successive steps is very unusual in real processes where there are several good reasons that often make the direction and the size of the

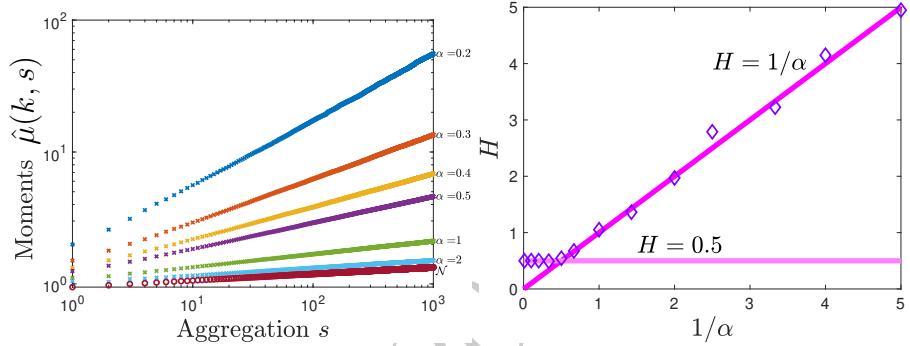


Figure 9.3 Scaling laws for artificial random walks with symmetric power law noise. On the left panel I report the estimate for the k -moments $\mathbb{E}(|R_{t,s}|^k)$ at various levels of aggregation s . The straight line behaviours in log-log scale indicate that they are proportional to a power of s (i.e. s^γ with $\gamma > 0$). This is in agreement with Eq.9.25 where $\gamma = kH$. On the right panel I show that the estimated Hurst exponent, \hat{H} (diamond symbols), is well described by $H = 1/\alpha$ for $\alpha \leq 2$ (the diagonal line) and then it becomes $H = 1/2$ when $\alpha > 2$ (the horizontal line).

next steps related to what happened in the previous steps. Furthermore, the assumption of discrete, equally sized, time-steps is also quite a strong reduction of the possible behaviors in real systems. There are several models to generate stochastic processes with more complex properties than random walks. In particular, the financial mathematics discipline has developed several stochastic models, from fractional Brownian motion to rough paths (see for instance Nourdin [2012] and Friz and Hairer [2020]), that capture some aspects of the complexity of real signals. In this book I do not discuss these generative models and I only look at the properties of stochastic signals. I discuss instruments to characterize and quantify their scaling properties independently from the details of the model, or the real system, that originates them.

Example 9.4 (Scaling laws for random walks processes). Let me here illustrate the scaling laws in random walk processes by investigating them with artificial data. Analogously to Example 9.3, I generate random walk processes with symmetric power law noise with a range of exponents α from 0.2 to 100. I let a large number (10,000) of independent random walks processes evolve and I first estimate the moment of the distribution $\mathbb{E}(|R_{t,s}|^k)$ with $k = 0.1$. Note that I use a small value of k because the moments are undefined when $k > \alpha$. The scaling law of the exponent is depicted in Fig.9.3 (left). One can see from the linear trends in the log-log scale that indeed, accordingly with Eq.9.24, $\mathbb{E}(|R_{t,s}|^k) \propto s^\gamma$ with $\gamma > 0$ a parameter that decreases with the exponent α . In the same figure, I also report the result for a random walk process with normal noise (labelled with \mathcal{N}). Pro-

cesses with $\alpha > 2$ are not reported in this figure because their scaling law become overlapped with the process with normal noise.

In order to verify Eq.9.24, and validate the link between the scaling of the moments, the Hurst exponent and the tail exponents (Eq.9.24) I have also estimated the moments $\mathbb{E}(|R_{t,s}|^k)$ for a range of values of k between 0.1 to 5. Specifically, for each k , if $k < \alpha$, I estimated the values of γ by computing the linear regression coefficient for $\log(\mathbb{E}(|R_{t,s}|^k))$ with respect to $\log s$. I then verified that the linear relation $\gamma = kH$ is well followed for $k < \alpha$ and I estimated the Hurst exponent H by regressing γ with respect to k (only for $k < \alpha$). Fig.9.3 (right) reports the resulting estimated values of the Hurst exponent H vs. α . One can observe that indeed, for $\alpha \leq 2$, the estimated Hurst exponent, \hat{H} (diamond symbols), is well described by $1/\alpha$. Then, when $\alpha > 2$ the Hurst exponent becomes $H = 0.5$. Note that I have used the ‘hat’ to indicate the estimated values of a quantity. This notation will be used through the book and in particular in Part II.

9.7 Multiscaling processes

Most stochastic processes, real or artificial, have complex scaling laws and cannot be described by a simple uniscaling law. These processes have self-affine properties and scaling laws but they have more than one characteristic scaling parameter depending on the property in exam. They have more than one scaling, and therefore they are called multi-scaling, or equivalently, multifractal.

In this book, I discuss multiscaling from two different perspectives. The first, concerns heterogeneity in time, where different parts of the signal have different roughness profiles, and this is quantified through the Hölder exponent. The second, is instead associated with the changes of the shape of the distribution with the time-horizon, where different moments scale with different characteristic exponents, this multiscaling is quantified through a generalization of the Hurst exponent. The two perspectives are not independent; they are actually strictly related.

Remark 9.4. The term **multifractals** comes from a 1980 paper by Firsh and Parisi on fully developed turbulence [Frisch and Parisi, 1980]. However, the idea was already formulated earlier in works by Mandelbrot himself and by Novikov and others [Novikov, 1971, Mandelbrot, 1972, 1974, 2013].

9.7.1 Hölder exponent

Many processes are not homogeneous and real signals have some parts with higher roughness profile and other parts that are smoother making the scale of self-affinity different in different parts of the process. This can be expressed in a

similar way to the scaling relation in Equation 9.21 where the scaling exponent H is however replaced with a local one h_t :

$$R_{t,s} \stackrel{d}{=} s^{h_t} R_{t,1}. \quad (9.26)$$

The exponent h_t is known as singularity exponent or Hölder exponent. This equation describes a process that scales with s similarly to the uniscaling one but such a scaling is local and depends on the point in time t . Different parts of the signal have different local scaling and overall the signal is characterized by a distribution of local exponents with probability distribution $f(h)$. This distribution is called multifractal spectrum or singularity spectrum.

Remark 9.5. The multifractal spectrum $f(h)$ is the probability distribution of the Hölder exponent for a given process. It is a fractal dimension itself or, more generally, it is the spectrum of the fractal dimensions of the process. From a ‘box counting’ perspective, the number of boxes, $N_s(h)$, with singularity strength between h and $h + dh$ is scaling as $N_s(h) \sim s^{-f(h)}$ (see Halsey et al. [1986].) As pointed out by Mandelbrot, a multifractal process requires the use of ‘multiboxes’.

9.7.2 Generalized Hurst exponent

I have introduced with Eq.9.25 the scaling of the moments for uniscaling processes where a single exponent H can describe the scaling law of all exponents. In a multiscaling process the moments of the returns over a time-horizon, s , scale instead with different exponents:

$$\mu(k, s) = \mathbb{E}(|R_{t,s}|^k) = s^{kH(k)} \mathbb{E}(|R_{t,1}|^k), \quad (9.27)$$

where $k \geq 0$, and $H(k) > 0$ is called generalized Hurst exponent. Notice that this expression is almost identical to the one for a uniscaling process (Eq.9.25, here I assume $s' = 1$ for simplicity). However, now the scaling exponent is not a constant but it depends on the order k of the moment.

The concept of multiscaling has its origin in the field of turbulence. A quantity that is often used in the turbulence literature is the scaling function or mass exponent function, which is directly related with the generalized Hurst exponent via:

$$\tau(k) = kH(k) - 1 \quad (9.28)$$

The Legendre transform of $\tau(k)$ gives the singularity strength h

$$h = \frac{d}{dk} \tau(k). \quad (9.29)$$

The singularity spectrum $f(h)$ is instead given by:

$$f(h) = kh - \tau(k). \quad (9.30)$$

In the studies of turbulence, the term ‘singularity’ refers to the point at which the flow velocity is discontinuous. In this book I am discussing scaling properties from a temporal signal perspective and therefore such a language remains a bit abstract. However, it is important to keep coherence with established literature definitions.

9.8 Memory and tail effects on the scaling exponents

Scaling of the probability distribution of signals is often observed in real systems. The origin of such scaling laws are in general complex and mostly unknown. I have just shown in the previous sections that for simple uniscaling processes, such as the random walk with strictly stationary noise term, the scaling of the returns at scale s coincides with the scaling of the probability distribution and the scaling exponent is directly related with the tail exponent through $H = 1/\alpha$. In random walk processes there is no memory and the scaling is entirely originated by the aggregation law of the probability distribution. In random walk models with stable distributions one has therefore that the exponent is, in general, $H \geq 1/2$ with $H = 1/2$ for the normal case.

Real processes are not random walks and one of the main features is that changes of the signal at different time steps are not independent. The correlation between two instances of a stochastic process at different times is named autocorrelation.

Definition 9.8 (Autocorrelation). The **autocorrelation** in a stochastic process X_1, \dots, X_t is the correlation between two instances of the process at two different times:

$$\text{Corr}(X_t, X_{t'}) = \frac{\mathbb{E}((X_t - \mu_X)(X_{t'} - \mu_X))}{\sqrt{\mathbb{E}((X_t - \mu_X)^2(X_{t'} - \mu_X)^2)}}.$$

For stationary processes, $\text{Corr}(X_t, X_{t'})$ must be dependent only on the absolute value of the difference $\Delta = |t - t'|$. The autocorrelation has its largest value at $\Delta = 0$ where it is equal to one.

In terms of the returns, by assuming $\mathbb{E}(R_{t,s}) = 0$ and stationarity, the expression for the autocorrelation takes the form:

$$\rho(\Delta) = \frac{\mathbb{E}(R_{t,s} R_{t-\Delta,s})}{\mathbb{E}(R_{t,s}^2)}. \quad (9.31)$$

The autocorrelation describes the ‘memory’ of the signal. The scaling properties are related with the memory effects.

There is a special class of stochastic processes that is called fractional Brownian

motion which is a generalization of the Brownian motion but with correlated increments. Specifically, for this process, the correlation between the returns at $t - \Delta/2$ and the returns at $t + \Delta/2$ is:

$$\rho(\Delta) = 2^{2H-1} - 1; \quad (9.32)$$

which does not depend on Δ revealing that this process has infinite memory with ‘persistent’ positive autocorrelation for $1/2 < H < 1$ and ‘anti-persistent’ negative autocorrelation for $0 < H < 1/2$.

Remark 9.6. The fractional Brownian motion is a non-stationary process. The autocorrelation must therefore be defined carefully because the interval of measure affects the result. This is why in Eq.9.32 I have taken the interval centred around the time t .

The infinite memory of the fractional Brownian motion is highly unrealistic, and, although fractional Brownian motion has proved to be an effective modeling tool, one expects, in real processes, memory to decrease with the time-lag Δ and eventually go to zero when $\Delta \rightarrow \infty$.

Example 9.5. For financial time series, the changes in price have very short memory with typically an anti-persistent behavior at short time scales below the second and then an absence of significant correlations for larger horizons. This is because the presence of autocorrelations in the returns would allow to predict the future change of price and therefore to make money out of it (the so-called arbitrage opportunity). The short-time anti-persistence is within a range of times and price variations that makes extremely hard, if not impossible, to directly profit from this property. On the other hand, if one observes the autocorrelation of the absolute value of the returns, a significant positive autocorrelation is revealed. Such persistence lasts even for months with a slow decay. This is related to the known fact that when there is a large variation of a price it is more likely that afterwards other large variations occur. The sign of the variation is random and therefore it is difficult to speculate from this information. This phenomenon is often referred to as ‘volatility clustering’. Interestingly, very similar clusterings of fluctuations are observed in earthquakes where after a primary shock a series of secondary shocks of decreasing intensity are often likely to occur.

In the literature, processes with Hurst exponent $H > 1/2$ are often referred as persistent and instead anti-persistent the ones with $H < 1/2$. However, beside the special case for fractional Brownian motion, and the special case discussed in the next section, the relation between the scaling exponent and the autocorrelation is in general non-trivial. Nonetheless, it is intuitive from the fractal dimension and box counting arguments (see Section 9.3) to figure out that an highly auto-correlated signal with strong persistence must be ‘smoother’ than

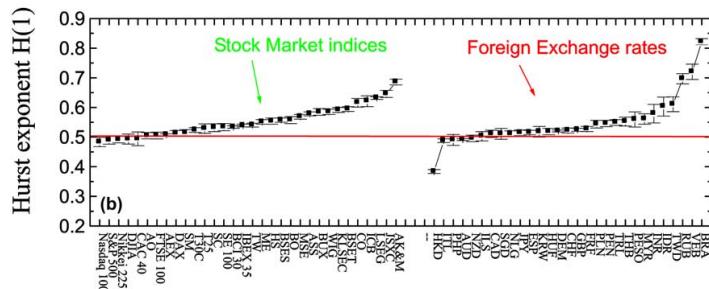


Figure 9.4 Generalized Hurst exponent $H(1)$ for the Stock Market indices and Foreign Exchange rates. (Figure reproduced with permission from Di Matteo et al. [2005]).

an anti-persistent signal. Therefore, similarly to the fractional brownian motion case, one expects smaller fractal dimensions in the first case and larger fractal dimensions in the second. Small fractal dimensions are associated with large Hurst exponents H and vice versa, therefore one still expects large H begin associated with persistent behaviors and small H with anti-persistent behaviors with $H = 1/2$ corresponding to the uncorrelated Brownian motion case.

Processes where the autocorrelation goes to zero exponentially fast are associated with ‘short-range’ persistence and, if no other factors (such as fat-tails) are involved, one expects a scaling exponent $H = 1/2$. In other processes, memory is instead lost at a much slower rate with autocorrelations that decrease towards zero as power laws. In this case, there are long-range memories the scaling exponent is expected to differ from $H = 1/2$. Persistent signals, with positive autocorrelations, are expected to have scaling exponents larger than 0.5 while anti-persistent signals, with negative autocorrelations, results in scaling exponents smaller than 0.5.

Example 9.6 (Hurst exponent for different financial markets). It is quite intuitive that in a developed, liquid, market the variation of prices must be harder to predict than in a less developed, less liquid market. In particular, if the log-returns of prices in a market are strongly persistent, then they can be easily predicted allowing for arbitrage. In Di Matteo et al. [2005] it was indeed shown that the deviation of the generalised Hurst exponent from the value of 0.5 is related with the stage of development of the market. It was found that the less-developed, less-liquid markets have Hurst exponents above 0.5 (persistent behavior) while instead liquid and well-developed markets have Hurst exponents below or around 0.5. Figure 9.4 summarizes the these findings. The figure refers to computation of the generalized Hurst exponent (see Eq.9.27) from time series of daily prices for the 1990’ decade. Although these are historic data from a time when markets were very different, still hurst exponents differentiate across markets. For

instance, by using daily prices of the MOEX Russia's index for the period Sept 2017-Sept 2022, one gets a generalized Hurst exponent $H(1) = 0.54$ while for the NASDAQ Composite (IXIC) index during the same period one retrieves $H(1) = 0.47$.^a

^a Data for the closing daily prices of such indexes can be obtained from yahoo finance: finance.yahoo.com/. Last accessed July 2022.

9.9 A stochastic process as a set of Student-t random variables multivariate across time

From a general probabilistic perspective, a stochastic process can be seen as a multivariate set of returns $(R_{t-k+1,1}, \dots, R_{t,1})$ indexed over a set of points in time. Where $R_{t,1} = X_t - X_{t-1}$ is the return at time t for the time-horizon $s = 1$. If such a system of random variables has joint probability density function

$$f(R_{t-k+1,1}, \dots, R_{t,1}), \quad (9.33)$$

a general problem is to find the scaling law which links this multivariate density function of the returns at horizon $s = 1$ with the density function, $f_{t,s}(R_{t,s})$, of returns at horizon s .

This is in general a very challenging problem. However, in some cases, it simplifies to tractable expressions. This is for instance the case when $f(R_{t-s+1,1}, \dots, R_{t,1})$ belongs to the elliptical distribution family with $\mu_i = 0$ and scale matrix $\Omega_{\mathbf{RR}}$. In this case, the aggregate probability density function, $f_{t,s}(R_{t,s})$, of the returns at horizon s , $R_{t,s} = R_{t-s+1,1} + \dots + R_{t,1}$, belongs to the location scale family with zero mean and scale

$$\sigma(s) = c \sqrt{\sum_{i,j=1}^s (\Omega_{\mathbf{RR}})_{i,j}}. \quad (9.34)$$

Where c is a proportionality constant independent from s . This expression is a consequence of the invariance of elliptical distribution over linear combination, and uses the fact that the return at horizon s is the sum of s returns at horizon 1, namely $R_{t,s} = \sum_{i=1}^s R_{t-i+1,1}$. Eq.9.34 was explicitly derived in Example 6.2. The functional form of $f_{t,s}(\cdot)$ can in general depend on s (see Eq.6.28 in Remark 6.3), but not in the cases of multivariate normal and multivariate Student-t where instead it keeps the same form.

In this case, of multivariate elliptically distributed variables, the Hurst exponent does not depend on the moments $H(k) = H$, indeed in the location-scale family all moments scale with the same scale parameter H (see Example 6.2). However, H can have different values depending on the form of the scale matrix. It must be noticed that in this formulation of stochastic processes as random variables multivariate across time, the elements of the scale matrix are proportional to the autocorrelation between the returns at two times that are $\Delta = |i-j|$

apart: $(\Omega_{RR})_{i,j} \propto \rho(\Delta)$. Therefore, the Hurst exponent is directly related with the autocorrelation through

$$\sum_{i,j=1}^s \rho(|i-j|) \underset{s \rightarrow \infty}{\sim} s^{2H} \quad (9.35)$$

which is a way to rewrite Eq.9.34 making explicit the relation between autocorrelation and the Hurst exponent. In a process with only the diagonal elements of the scale matrix different from zero ($\rho(0) = 1$, $\rho(\Delta) = 0$ for any $\Delta > 0$) the sum in Eq.9.34 scales linearly with s yielding to a Hurst exponent $H = 1/2$. This is consistent with the discussions in the early sessions of this Chapter. Indeed, a diagonal scale matrix implies zero autocorrelations. However, this is a delicate case because, despite having zero autocorrelations, the returns of a multivariate elliptically distributed process with a diagonal scale matrix are not independent, except for the normal case. Such a linear scaling is also observed for any process with short-range positive autocorrelations. When instead the autocorrelation is long-ranged and positive, then the Hurst exponents becomes larger than 1/2. For instance, for

$$\rho(\Delta) \propto \Delta^{\beta-1} \quad (9.36)$$

with $0 \leq \beta < 1$, one has

$$\sum_{i,j=1}^s \rho(|i-j|) \sim s^{\beta+1}, \quad (9.37)$$

and therefore

$$H = 1/2 + \beta/2. \quad (9.38)$$

This yields to Hurst exponents $H \in [1/2, 1]$. Hurst exponents smaller than 1/2 can only be obtained when negative autocorrelations are present (see also discussion in Example 9.7). In this case, depending on the balance between positive and negative correlations, and their range, all values of Hurst exponents, all the way down to $H = 0$ can be obtained. One therefore retrieves, within this framework, the relations between the Hurst exponent $H > 1/2$ and the ‘persistent’ process (positive autocorrelations) and $H < 1/2$ with ‘anti-persistent’ process (negative autocorrelations) that I described in the previous section for the fractional Brownian motion. However, in the present framework, the relation between the form of the autocorrelation and the scaling is straightforward and explicit through Eqs.9.34 and 9.35.

In summary, one has:

- $H = 1/2$, when $|\rho(\Delta)|$ is decreasing towards zero faster than Δ^{-1} ;
- $1/2 < H < 1$, when $\rho(\Delta) > 0$ decreasing towards zero slower than Δ^{-1} ;
- $H = 1$ for constant positive autocorrelation $\rho(\Delta) = \gamma > 0$;
- $0 < H < 1/2$ for specific combinations of positive and negative autocorrelations.

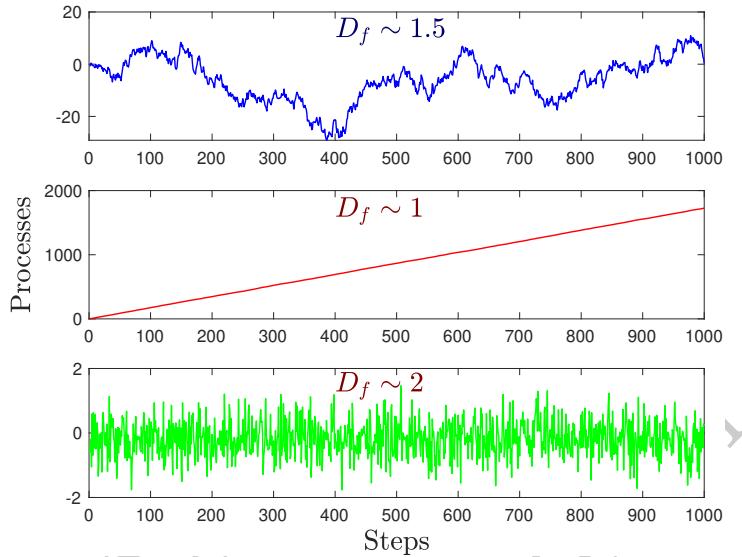


Figure 9.5 Examples of three processes generated using Student-t variables multivariate across time. The process on top has a diagonal (autocorrelation) scale matrix (Eq.9.40), it is therefore linearly uncorrelated but still the returns at each time step are not independent. The process on the middle has a full constant (autocorrelation) scale matrix (Eq.9.42), and it results essentially in a straight line because the returns at each time step are fully correlated and therefore all steps have similar values (a close look reveals some small deviations from a straight line). The process on the bottom has scale matrix (autocorrelation) with negative values adding to a constant (Eq.9.46), it is therefore an anti-correlated mean-reverting process. The three process have respectively Hurst exponents $H \sim 0.5$, 1, and 0 and consequently, from $D_f = 2 - H$, the fractal dimension is respectively $D_f \sim 1.5$, 1, and 2.

Example 9.7 provides a few practical and insightful cases where it is easy to link the Hurst exponent and the autocorrelation.

Example 9.7. From Eq.9.35 it is clear that the Hurst exponent is directly and explicitly related with the way the sum of the autocorrelation over a window of size s increases with the window size. Let me investigate this scaling with a few concrete examples.

a. Process with zero autocorrelations

Consider, the case of absence of autocorrelations, which is

$$\begin{aligned}\rho(0) &= 1 \\ \rho(\Delta) &= 0 \text{ for all } \Delta \neq 0.\end{aligned}\tag{9.39}$$

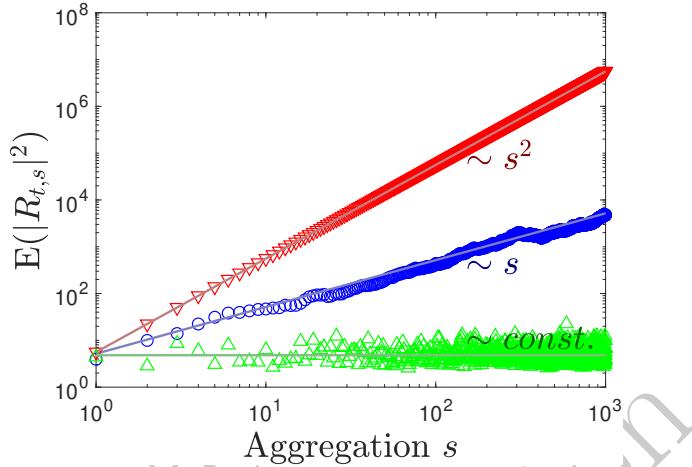


Figure 9.6 Scaling of the second moment, $\mathbb{E}(|R_{t,s}|^2)$, for Student-t processes multivariate across time with three different scale (autocorrelation) matrices (same processes as in Fig.9.5). The red lower triangles are associated with a process with a full constant (autocorrelation) scale matrix. It has a scaling that follows well $\mathbb{E}(|R_{t,s}|^2) \propto s^2$. The blue circles are associated with a process with a diagonal (autocorrelation) scale matrix. It has a scaling that follows well $\mathbb{E}(|R_{t,s}|^2) \propto s$. The green upper triangles are associated with a process with a scale matrix (autocorrelation) with negative values adding to a constant. It has a scaling that follows well $\mathbb{E}(|R_{t,s}|^2) \propto \text{const.}$. Consequently, from the relation $\mathbb{E}(|R_{t,s}|^2) \propto s^{2H}$, one has respectively $H \sim 1$, 0.5 , and 0 for the three processes.

The sum is

$$\sum_{i,j=1}^s \rho(|i - j|) = s. \quad (9.40)$$

Yielding, through Eq.9.35, to $H = 1/2$. As already noticed, $\rho(|i - j|) = 0$ (for $|i - j| > 0$) does not imply independence of the returns, but rather only linear independence. An example of this kind of process is reported in Fig.9.5 top panel, while the corresponding scaling law for the second moment, $\mathbb{E}(|R_{t,s}|^2)$ is reported in Fig.9.6 with blue circles. One can verify that indeed this process scales with $H = 1/2$ which corresponds to $D_f = 1.5$.

b. Process with constant positive autocorrelation

Consider the autocorrelation

$$\begin{aligned} \rho(0) &= 1 \\ \rho(\Delta) &= \gamma \text{ for all } \Delta \neq 0. \end{aligned} \quad (9.41)$$

The sum is

$$\sum_{i,j=1}^s \rho(|i-j|) = s + \gamma \frac{s(s-1)}{2} \underset{s \rightarrow \infty}{\sim} s^2 \quad (9.42)$$

Yielding, through Eq.9.35, to $H = 1$. An example of this kind of process is reported in Fig.9.5 middle panel, while the corresponding scaling law for the second moment, $\mathbb{E}(|R_{t,s}|^2)$ is reported in Fig.9.6 with lower red triangles.

c. Process with positive autocorrelation decreasing as power law with Δ

Consider the autocorrelation

$$\begin{aligned} \rho(0) &= 1 \\ \rho(\Delta) &= \gamma \Delta^{\beta-1} \quad \text{for all } \Delta \neq 0. \end{aligned} \quad (9.43)$$

with $0 \leq \beta \leq 1$, which is referred to as “power law banded” [Mirlin et al., 1996]. The scaling of the sum can be computed by first noticing that $\sum_{i,j=1}^s |i-j|^{\beta-1} \underset{s \rightarrow \infty}{\sim} s \sum_{\Delta=1}^s \Delta^{\beta-1}$ and then by using the continuous limit to compute the sum as an integral. The result is

$$\sum_{i,j=1}^s \rho(|i-j|) \underset{s \rightarrow \infty}{\sim} \begin{cases} s & \text{for } \beta \leq 0 \\ s^{1+\beta} & \text{for } \beta > 0 \end{cases}. \quad (9.44)$$

Yielding, through Eq.9.35, to Hurst exponents $H = 1/2 + \beta/2$. Therefore, for $\beta \in [0, 1]$, one has $H \in [0.5, 1]$. An example of this kind of process is reported in Fig.9.5 lower panel, while the corresponding scaling law for the second moment, $\mathbb{E}(|R_{t,s}|^2)$ is reported in Fig.9.6 with upper green triangles . Once can verify that indeed this process scales with $H = 0$ which corresponds to $D_f = 2$.

d. Process with positive and negative autocorrelations

Consider the autocorrelation

$$\begin{aligned} \rho(0) &= 1 \\ \rho(1) &= -0.5 \\ \rho(\Delta) &= 0 \quad \text{for all } \Delta > 1 \end{aligned} \quad (9.45)$$

In this case, the sum is not scaling with s , rather it is always

$$\sum_{i,j=1}^s \rho(|i-j|) = 1. \quad (9.46)$$

Yielding, through Eq.9.35, to Hurst exponent $H = 0$.

Larger Hurst exponents in the range $H \in (0, 0.5]$, can be, for instance, obtained by adding a banding with $\rho(\Delta) = \gamma \Delta^{\beta-1}$ for all $\Delta > 1$.

Remark 9.7. This case of a multivariate elliptically distributed set of returns, $f(R_{t-s+1,1}, \dots, R_{t,1})$ which I have just discussed, is quite peculiar because I have just shown that the Hurst exponent is independent on the degree of the moment. This would indicate a unscaling process, however the form of the distribution might vary with aggregation (except for normal and Student-t). This is therefore an instance of a process that is not unscaling but has constant Hurst exponent across moments. Furthermore, for the Student-t case, the Hurst exponent is related to the autocorrelation but not to the tail exponent.

9.10 Data-driven modeling tutorial

<https://github.com/FinancialComputingUCL/DataDrivenModeling/Ch9>

The tutorial for this Chapter covers various topics on scaling laws including: random walk processes (Example 9.3); scaling of the moments in random walk processes (Example 9.4); Student-t autocorrelated processes (Example 9.7).

Exercises

- Using the arguments for the estimation of the fractal dimension with the box-counting method, estimate the fractal dimension of a circle.
- Comment whether the fractal dimensions of triangles, squares or pentagons should be different.
- What is the Hurst exponent of a random walk process with Lévy-stable distribution with scaling exponent equal to 1.5?
- Can a random walk process have Hurst exponent smaller than 0.5?
- Provide an example of a process with Hurst exponent smaller than 0.5.
- A self-affine process has variance of the returns, σ_s^2 , at time horizon $s = 1$ equal to 3. Knowing that the Hurst exponent of the process is 1.5 estimate the variance at time horizon $s = 100$.

10

Causality

10.1 Cause and effect

Whenever we witness something happening, we often wonder about what might have caused it, and time and time again we find ourselves short of an answer. Causality is hard to establish. Nonetheless, finding the cause of events and phenomena is the core of scientific investigation and the basis of what we define as knowledge and understanding of the world around us.

In our personal experience, we have learned that causality conclusions are often subject to personal perspectives, opinions, and beliefs. This is in general a broad philosophical issue, which I will not address in this book. Rather, in this chapter, I address this problem from a perspective that is narrow but it provides a clean way to quantify causality from data. I shall refer to this as statistical temporal causality. The idea was first formulated by Wiener [1956] which proposed as a signature of a causal relation between two variables the fact that a variable provides extra information for the forecasting of the other. Wiener formulation was first implemented, within a linear regression framework for financial time-series, by Granger [1969] and I shall refer to this approach as Wiener-Granger causality (see Section 10.3). Such an approach to causality is limited and can be misleading; however, it has the advantage of being well-defined and quantifiable in many practical situations. In this chapter, I shall present an information-theoretic generalization of the Wiener and Granger approach, where I quantify causality in terms of the amount of extra information the past of a variable carries about the future of another variable. Or, in other terms, how much a change in one variable leads to a change in another variable while discounting for factors that might derive from the rest of the system. It should be clear therefore that this causality measure requires the use of conditional probability and the ordering of observations in lead and lag terms. I shall indeed show in Section 10.4 that such a statistical temporal causality is quantifiable as a conditional lagged dependency. At the end of this Chapter, in Section 10.5, I dedicate a small section to provide a broader view of the problem of causation.

The Wiener-Granger approach to causality requires a temporal notion because it assumes that the cause must precede the effect. Indeed, when things happen simultaneously or their temporal sequence cannot be established, it becomes impossible to attribute to one the role of cause and to the other the role of effect, from observations only, without extra knowledge about the wider context. For

instance, when a car turns at a road crossing we know that it is the driver that acts on the steering wheel and makes the car turn. Therefore the driver is causing the car to turn. If, however, one has only access to the observations of the car movement and of the steering wheel rotation, we will not be able to establish if it is the steering wheel that causes the car to turn or vice versa is the turning of the car that causes the steering wheel to rotate. The two phenomena occur simultaneously, we can certainly say that there is a dependency (the knowledge of one phenomenon gives us information about the other), but we cannot say that there is a causal relation.

When there is a temporal sequence of events, it can become easier to establish causal relations. This is because, as a rule of thumb, the cause precedes the effect. However, even this very simple, largely intuitive, and often true, rule can lead to false results. For instance, going back to the example of a turning car at a road crossing, the observer could notice that the turn of the car is preceded by the activation of the directional lights; a straightforward conclusion would therefore be that directional lights cause the car to turn. This is obviously wrong, but the reason why it is so obvious to us is that we know that there is a conscious driver, and therefore we know that the turn follows the driver's decision. From the sole observation of the data, the existence of a 'hidden variable' (the driver) that is causing the occurrence of events in the observed sequence might not be discoverable.

In many domains, the sequence of events over time is not the central focus in the narrative of causality studies. For instance, in biology or medicine, one wants to assess the causal effect of a given intervention (a drug or a treatment). Of course, the intervention must have taken place before the analysis of its effects, but the crucial point is to compare samples where the intervention took place with samples where no intervention was done (the so-called 'control' group). In this chapter, and for the rest of this book, I'll always refer to the temporal sequence of the variables assuming that an action (or intervention) in the past has taken place so as to modify the future properties of a variable. At a fundamental level, referring to information transfer between two variables in time or instead talking specifically about intervention is pertinent to the narrative of the particular domain. The substance is the quantification of the effect of the information from a variable (i.e. intervention/no-intervention) on the probability distribution of another variable.

The quantification of causality from the linear regressions was proposed by Clive Granger (1934-2009) in his celebrated paper [Granger, 1969]. In 2003 he, and Robert F. Engle, were awarded the Nobel Memorial Prize in Economic Sciences in recognition of their contributions to the analysis of time series data. The general idea must be however attributed to Norbert Wiener (1894-1964) who formulated it in his inspiring paper "Theory of prediction" [Wiener, 1956]. Wiener was a visionary and talented scientist

who laid down the basis of what we call artificial intelligence (AI) and is the father of cybernetics. He is best known for the introduction of the concept of feedback and control systems which indeed led him to develop cybernetics, which in the title of his book [Wiener, 2019] he defines as “the study of control and communication in the animal and the machine”.

10.2 Causality and correlations

It is often said that “correlation does not imply causation”. This is certainly true, one must not confuse causality with correlation; correlations quantify linear dependency, whereas causality (in the Wiener-Granger formulation) must quantify how much the values of a random variable determine the values of another random variable discounting for all other factors. To quantify causality, one is therefore looking for the ‘extra’ contribution that the knowledge of a variable provides to the knowledge of another variable. However, correlations, dependency, and causality are very close concepts, and they are hard to disentangle. The presence of correlations between two variables does not imply any causality; nonetheless, phenomena that are causally related are often characterized by the presence of correlations, especially lagged correlations (see Definition 8.10). In the linear case, one can demonstrate (see Remark 10.2) that lagged correlations are a necessary condition for causality. One can in general say that correlation does not imply causation, but causation does require non-simultaneous conditional dependency.

The question to ask is whether events in the past may be related to events in the future and how. Let me consider two stochastic processes X_t and Y_t (see Definition 9.1), if there is a significant dependence between the variable $X_{t-\lambda}$ at time $t - \lambda$ and the other variable Y_t at time t , then I can use the information about X at previous times to predict what might happen to Y at a later time. In terms of dependency, we say that X has a leading lag dependency with Y at lag λ . This however does not imply causality, because the variable Y could have such a leading lag-dependency with itself (e.g. autocorrelations), and therefore the dependency with X might not be of extra relevance to predict its behavior.

Definition 10.1 (Lags). The concept of ‘lags’ refers to values of a stochastic variable at times previous to a specific point in time t . For instance, X_{t-1} is the lagged variable associated to X_t with lag $\lambda = 1$. Lags can be also a vector of different values, for instance, the vector of lags $\boldsymbol{\tau} = (\tau_1, \dots, \tau_m)^\top$ is associated with the lagged time series

$$X_{t,\boldsymbol{\tau}}^- = (X_{t-\tau_1}, \dots, X_{t-\tau_m})^\top, \quad (10.1)$$

and, consistently with this notation, for any other variable, Y_t , and lag

vector $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)^\top$, one equivalently has

$$Y_{t,\lambda}^- = (Y_{t-\lambda_1}, \dots, Y_{t-\lambda_n})^\top. \quad (10.2)$$

10.3 Wiener-Granger causality

Following the original idea of Wiener and Granger, considering two stationary stochastic processes X_t and Y_t , one aims to establish if the value y_{t+h} of the stochastic variable Y_s at time $s = t + h$ is exclusively the consequence of randomness and the previous history, $y_{t,\lambda}^- = (y_{t-\lambda_1}, \dots, y_{t-\lambda_n})$, of the variable Y itself, or it is also influenced by the history, $x_{t,\tau}^- = (x_{t-\tau_1}, \dots, x_{t-\tau_m})$, of another variable X . This can be established by using two regressions: if by using the past of variable X one can better predict the future of variable Y with respect to using only the past of Y itself, then it implies that the past of variable X has some Wiener-Granger causal influence on variable Y .

Remark 10.1. In technical jargon, the variable at $t+h$ is referred to as ‘at h steps ahead’. Whereas the variables at $t-\lambda_1, \dots, t-\lambda_n$ and $t-\tau_1, \dots, t-\tau_m$ are referred to as ‘lagged’. Here, I always assume $h \geq 0$ and $\lambda_i, \tau_i > 0$. It is common to use the same lags for the two variables, $\boldsymbol{\lambda} = \boldsymbol{\tau}$, and often one single value for the lag ($n = m = 1$) is used. The choice of the lags in causal models is a feature selection problem. As such the choice of the lags involves choosing the most relevant and informative variables (features) to be used in the model. Ultimately, the goal is to strike a balance between including enough lagged variables to capture relevant patterns and avoiding overfitting and curse of dimension by including too many lags that may not add significant value to the model’s predictive capabilities.

Granger’s original proposal [Granger, 1969] was to use linear regressions. Specifically, he proposed to compare the regression of Y_{t+h} with respect to its past $Y_{t,\lambda}^-$ with another regression that includes also the past of the other variable $X_{t,\tau}^-$. Explicitly:

$$Y_{t+h} = \beta_0 + \sum_{s=1}^n \beta_{1,s} Y_{t-\lambda_s} + \epsilon_1 \quad (10.3)$$

and

$$Y_{t+h} = \beta_2 + \sum_{s=1}^n \beta_{3,s} Y_{t-\lambda_s} + \sum_{s=1}^m \beta_{4,s} X_{t-\tau_s} + \epsilon_2. \quad (10.4)$$

In this notation the ‘lags’ $\lambda_1, \dots, \lambda_n$ and τ_1, \dots, τ_m are not necessarily coincident. Often the setting is with $h = 0$ and a single common lag $\lambda > 0$, where therefore the ‘present’ value Y_t is inferred from the ‘past’ lagged values $Y_{t-\lambda}$ and $X_{t-\lambda}$.

If the second regression (Eq.10.4) is significantly better than the first (Eq.10.3), then it means that the past of X is influencing the future of Y beyond the contribution already provided by the past of Y itself. How ‘better’ the second regression estimates the future of Y can be quantified by comparing the variances of the two errors (see also discussion on the goodness of regressions in Section 18.4)

$$F_{X \rightarrow Y} = \log \frac{\text{Var}(\epsilon_1)}{\text{Var}(\epsilon_2)}. \quad (10.5)$$

If the past of X contributes significantly to the regression of Y_{t+h} , then $\text{Var}(\epsilon_2) \ll \text{Var}(\epsilon_1)$ and $F_{X \rightarrow Y}$ is large. Conversely, if the past of X does not contribute significantly to the regression of Y_{t+h} , then $\text{Var}(\epsilon_2) \approx \text{Var}(\epsilon_1)$ and $F_{X \rightarrow Y}$ is small or zero. This quantifies the linear dependence between Y_{t+h} and $(X_{t-\tau_1}, \dots, X_{t-\tau_m})$ conditioned on $(Y_{t-\lambda_1}, \dots, Y_{t-\lambda_n})$.

For uncorrelated, normally distributed errors, the quantity

$$f = \frac{\text{Var}(\epsilon_1) d_2}{\text{Var}(\epsilon_2) d_1}, \quad (10.6)$$

with $d_1 = m$ and $d_2 = q - n - 1$, where q is the sample size, follows the F-distribution (see Section 5.6.3) with degrees of freedom d_1 and d_2 .

Notice that, as far as these regression errors are estimated in-sample, $\text{Var}(\epsilon_2)$ is always smaller than $\text{Var}(\epsilon_1)$ because – even in the absence of any conditional dependency – the regression with a larger number of parameters must always provide lower in-sample errors. If instead one computes the residuals out-of-sample, it is then possible that the model with a larger number of parameters performs worse than the simpler one, this is commonly referred to as ‘overfitting’ (see Section 3.5.4).

Remark 10.2. The regression coefficients depend on the correlations between the variables (see Remark 8.7). If all lagged correlations between Y_{t+h} and $X_{t-\tau_s}$ are zero, then the coefficients $\beta_{4,s} = \frac{\sigma_Y}{\sigma_X} \rho_{X_{t-\tau_s}, Y_t} = \frac{\sigma_Y}{\sigma_X} a_{X,Y}(\tau_s)$ are zero as well. Having non-zero lagged correlations is a necessary condition for linear Granger causality.

Remark 10.3. The Granger causality from the regressions in Eqs.10.3 and 10.4 uses past values of the random variable X to predict the random variable Y . If one includes in the regression also the present value of X (i.e.

$X_t)$, then the Granger causality is called “instantaneous”. However, this poses interpretability problems because one cannot associate a direction to the instantaneous correlation between X_t and Y_t .

10.3.1 Non-linear Wiener-Granger causality

Granger causality in its original formulation detects linear causality. It is indeed a comparison between the errors of the two linear regressions Eqs.10.3 and 10.4. However, it is straightforward to generalize Granger’s implementation to any kind of non-linear regression. Specifically, a generalized, non-linear Wiener-Granger causality approach compares the following two general regressions:

$$\begin{aligned} Y_{t+h} &= g_1(Y_{t-\lambda}^-) + \epsilon_1 \\ Y_{t+h} &= g_2(Y_{t-\lambda}^-, X_{t-\tau}^-) + \epsilon_2. \end{aligned} \quad (10.7)$$

As before, the comparison can be performed in terms of the variance of the errors:

$$F_{X \rightarrow Y} = \log \frac{\text{Var}(\epsilon_1)}{\text{Var}(\epsilon_2)}. \quad (10.8)$$

However, in this non-linear case, the statistics of $F_{X \rightarrow Y}$ is in general unknown and the validation of the causal relation might need to adopt non-parametric criteria, such as the z-score (see Section 18.8.1). The comparison of the variances of the two errors is a fair measure to assess the relative goodness of fit of the two models; however, also other measures of uncertainty can be used. For instance, the entropy. I have indeed already discussed that dependency can be quantified in terms of reduction in entropy as a consequence of conditioning (see Definition 8.9). It should be therefore clear that $H(\epsilon_1) - H(\epsilon_2)$ is a measure of causality. It is called transfer entropy.

10.4 Transfer entropy

The fundamental idea underneath Wiener-Granger causality is to associate causality with the fact that the past of a variable ($X_{t-\tau}^-$) has extra information about the future of another variable (Y_{t+h}) discounting for the information provided by the past of the variable itself ($Y_{t-\lambda}^-$). In Section 7.1, I have introduced the entropy (Shannon entropy) as a measure of information, therefore the amount of information provided by $X_{t-\tau}^-$ and $Y_{t-\lambda}^-$ can be expressed in terms of two conditional entropies $H(Y_{t+h}|Y_{t-\lambda}^-)$ and $H(Y_{t+h}|X_{t-\tau}^-, Y_{t-\lambda}^-)$ (see Remark 8.9). The first quantifies the uncertainty left in the variable Y_{t+h} when its past ($Y_{t-\lambda}^-$) is known. The second quantifies the uncertainty left in the variable Y_{t+h} when both its past ($Y_{t-\lambda}^-$) and the other variable past ($X_{t-\tau}^-$) are known. Their difference is therefore the extra contribution of $X_{t-\tau}^-$ to Y_{t+h} , and it is a qualification of Wiener-Granger

causality that extends to non-linear, non-normal cases. It takes the name of transfer entropy [Schreiber, 2000]:

$$T_{\mathbf{X} \rightarrow \mathbf{Y}} = H(\mathbf{Y}_{t+h} | \mathbf{Y}_{t,\lambda}^-) - H(\mathbf{Y}_{t+h} | \mathbf{X}_{t,\tau}^-, \mathbf{Y}_{t,\lambda}^-), \quad (10.9)$$

which can also be written as the difference of the entropies of the errors of the regressions, Eqs.10.7.

$$T_{\mathbf{X} \rightarrow \mathbf{Y}} = H(\epsilon_1) - H(\epsilon_2). \quad (10.10)$$

When the variables are multivariate normally distributed then $T_{\mathbf{X} \rightarrow \mathbf{Y}} = \frac{1}{2} F_{\mathbf{X} \rightarrow \mathbf{Y}}$ (see Example 10.2).

Remark 10.4. From a general probabilistic perspective, one wants to establish if the two conditional distributions

$$P(Y_{t+h} | X_{t,\tau}^-, Y_{t,\lambda}^-) \quad (10.11)$$

and

$$P(Y_{t+h} | Y_{t,\lambda}^-) \quad (10.12)$$

are equivalent. Applying Bayes' formula to both, one can write the first as $P(Y_{t+h}, X_{t,\tau}^-, Y_{t,\lambda}^-) / P(X_{t,\tau}^-, Y_{t,\lambda}^-)$ and the second as $P(Y_{t+h}, Y_{t,\lambda}^-) / P(Y_{t,\lambda}^-)$. Multiplying both terms by $P(X_{t,\tau}^-, Y_{t,\lambda}^-)$ one can see that the original equivalence is the same as the equivalence between $P(Y_{t+h}, X_{t,\tau}^-, Y_{t,\lambda}^-)$ and $P(Y_{t+h}, Y_{t,\lambda}^-)P(X_{t,\tau}^-, Y_{t,\lambda}^-) / P(Y_{t,\lambda}^-)$. I discussed in Section 7.4 that a way to quantify such an equivalence is via the Kullback-Leibler divergence. One can indeed verify directly (see Definition 7.2) that the transfer entropy, when computed in terms of Shannon entropy, is a Kullback-Leibler divergence:

$$T_{X \rightarrow Y} = D_{KL}(P, Q), \quad (10.13)$$

between

$$P = P(Y_{t+h}, X_{t,\tau}^-, Y_{t,\lambda}^-), \quad (10.14)$$

and

$$Q = \frac{P(Y_{t+h}, Y_{t,\lambda}^-)P(X_{t,\tau}^-, Y_{t,\lambda}^-)}{P(Y_{t,\lambda}^-)}. \quad (10.15)$$

One might notice that the expression for the transfer entropy Eq.10.9 resembles the one for the mutual information in Section 8.9. Indeed, the transfer entropy is a conditional mutual information:

$$T_{\mathbf{X} \rightarrow \mathbf{Y}} = I(\mathbf{Y}_{t+h}; \mathbf{X}_{t,\tau}^- | \mathbf{Y}_{t,\lambda}^-). \quad (10.16)$$

It is therefore a measure of **conditional, lagged dependency**. The conditioning to the past values of the stochastic variable Y itself is crucial to distinguish between causality and lagged dependency. To assess causation one must establish

the dependence between the future of Y and the past of X , conditioned to the past of Y .

10.4.1 Conditional transfer entropy

Exactly in the same way I defined transfer entropy in Eq.10.9 and Eq.10.16, one can define a conditional form of transfer entropy simply by imposing a further condition on other variables \mathbf{Z} (not containing \mathbf{Y}_{t+h} and $\mathbf{X}_{t,\tau}^-$)

$$T_{\mathbf{X} \rightarrow \mathbf{Y} | \mathbf{Z}} = I(\mathbf{Y}_{t+h}; \mathbf{X}_{t,\tau}^- | \mathbf{Y}_{t,\lambda}^-, \mathbf{Z}), \quad (10.17)$$

These other conditioning variables \mathbf{Z} must be in the past because, as for $\mathbf{Y}_{t,\lambda}^-$, one is discounting the knowledge of them when quantifying the effect of $\mathbf{X}_{t,\tau}^-$ on the reduction of uncertainty on \mathbf{Y}_{t+h} . Therefore, the notation $\mathbf{Z}_{t,\gamma}^-$ should be applied, and it is often used in the literature. However, the formulation is completely general and one might be in some cases interested to discount information on other variables, not necessarily in the past.

Remark 10.5. It must be noticed that the estimation of the transfer entropy involves the estimation of the entropy of at least three variables: the future variable Y_{t+h} and two lagged variables $X_{t-\tau}$ and $Y_{t-\lambda}$. Typically, as a common practice, to keep dimensionality low the transfer entropy is computed for a single lag $\tau = \lambda$, performing various measurements for different lag values. Then the largest or/and the most significant value at a given lag λ^* is adopted. Conditioning to other variables increases the dimensionality of the problem.

10.4.2 Transfer entropy for the multivariate elliptical family

The entropy of multivariate elliptical variables is given by $H(\mathbf{Z}) = \frac{1}{2} \log(|\Omega_{\mathbf{Z}\mathbf{Z}}|) + H_{Z_0}$ (see Eq.7.23 in Section 7.3). Recalling that the conditional entropy is $H(\mathbf{A}|\mathbf{B}) = H(\mathbf{A}, \mathbf{B}) - H(\mathbf{B})$ (see Definition 8.9), by substituting directly into the expression in Eq.10.9, using the Shannon entropy, one has

$$H(\mathbf{Y}_{t+h} | \mathbf{X}_{t,\tau}^-, \mathbf{Y}_{t,\lambda}^-) = \frac{1}{2} \log \left(\frac{|\text{Cov}(\{\mathbf{Y}_{t+h}, \mathbf{X}_{t,\tau}^-, \mathbf{Y}_{t,\lambda}^-\})|}{|\text{Cov}(\{\mathbf{X}_{t,\tau}^-, \mathbf{Y}_{t,\lambda}^-\})|} \right) + H_{Y_0}. \quad (10.18)$$

One also has

$$H(\mathbf{Y}_{t+h} | \mathbf{Y}_{t,\lambda}^-) = \frac{1}{2} \log \left(\frac{|\text{Cov}(\{\mathbf{Y}_{t+h}, \mathbf{Y}_{t,\lambda}^-\})|}{|\text{Cov}(\{\mathbf{Y}_{t,\lambda}^-\})|} \right) + H_{Y_0}. \quad (10.19)$$

Combined, through Equation 10.9 they lead to

$$T_{\mathbf{X} \rightarrow \mathbf{Y}} = \frac{1}{2} \log \left(\frac{|\text{Cov}(\{\mathbf{Y}_{t+h}, \mathbf{Y}_{t,\lambda}^-\})| \cdot |\text{Cov}(\{\mathbf{X}_{t,\tau}^-, \mathbf{Y}_{t,\lambda}^-\})|}{|\text{Cov}(\{\mathbf{Y}_{t+h}, \mathbf{X}_{t,\tau}^-, \mathbf{Y}_{t,\lambda}^-\})| \cdot |\text{Cov}(\{\mathbf{Y}_{t,\lambda}^-\})|} \right). \quad (10.20)$$

Here I used the notation, $\text{Cov}(\{\mathbf{A}\}) = \boldsymbol{\Sigma}_{\mathbf{AA}}$ to indicate the covariance matrix of the set of random variables \mathbf{A} and $|\cdot|$ is the determinant. The brackets $\{\mathbf{A}, \mathbf{B}\}$ indicate the joint system of variables or, in other notation, $\text{Cov}(\{\mathbf{A}, \mathbf{B}\}) = \boldsymbol{\Sigma}_{\mathbf{ZZ}}$ with $\mathbf{Z}^\top = (\mathbf{A}^\top, \mathbf{B}^\top)$.

Remark 10.6. Notice that in this notation, when the set of random variables \mathbf{A} is composed of only one random variable, A , then the covariance coincides with the variance:

$$\text{Cov}(A) = \sigma_A^2, \quad (10.21)$$

and the determinant is the variance as well

$$|\text{Cov}(A)| = \sigma_A^2. \quad (10.22)$$

Example 10.1 (Transfer entropy between two variables at one lag). In order to better understand expression Eq.10.20 it can be useful to derive it explicitly in terms of correlations for the simplest case two variables and one lag. From Eq.10.20 one has

$$|\text{Cov}(\{\mathbf{Y}_{t+h}, \mathbf{Y}_{t,\lambda}^-\})| = \det \begin{bmatrix} \Sigma_{Y,Y} & \Sigma_{Y,Y^-} \\ \Sigma_{Y,Y^-} & \Sigma_{Y,Y} \end{bmatrix} = \Sigma_{Y,Y}^2 - \Sigma_{Y,Y^-}^2; \quad (10.23)$$

$$|\text{Cov}(\{\mathbf{X}_{t,\tau}^-, \mathbf{Y}_{t,\lambda}^-\})| = \det \begin{bmatrix} \Sigma_{X,X} & \Sigma_{X,Y} \\ \Sigma_{X,Y} & \Sigma_{Y,Y} \end{bmatrix} = \Sigma_{X,X} \Sigma_{Y,Y} - \Sigma_{X,Y}^2. \quad (10.24)$$

$$\begin{aligned} |\text{Cov}(\{\mathbf{Y}_{t+h}, \mathbf{X}_{t,\tau}^-, \mathbf{Y}_{t,\lambda}^-\})| &= \det \begin{bmatrix} \Sigma_{Y,Y} & \Sigma_{Y,X^-} & \Sigma_{Y,Y^-} \\ \Sigma_{Y,X^-} & \Sigma_{X,X} & \Sigma_{X,Y} \\ \Sigma_{Y,Y^-} & \Sigma_{X,Y} & \Sigma_{Y,Y} \end{bmatrix} \\ &= \Sigma_{Y,Y} (\Sigma_{X,X} \Sigma_{Y,Y} - \Sigma_{X,Y} \Sigma_{X,Y}) \\ &\quad - \Sigma_{Y,X^-} (\Sigma_{Y,Y} \Sigma_{Y,X^-} - \Sigma_{X,Y} \Sigma_{Y,Y^-}) \\ &\quad + \Sigma_{Y,Y^-} (\Sigma_{X,Y} \Sigma_{Y,X^-} - \Sigma_{X,X} \Sigma_{Y,Y^-}); \end{aligned} \quad (10.25)$$

$$|\text{Cov}(\{\mathbf{Y}_{t,\lambda}^-\})| = \Sigma_{Y,Y}; \quad (10.26)$$

Where I indicated the coefficients of the covariance matrices by using the simplified symbols using Y for \mathbf{Y}_{t+h} and analogously $\mathbf{X}_{t,\tau}^-$ for X^- and $\mathbf{Y}_{t,\lambda}^-$ for \mathbf{Y}^- . I also used the fact that stationarity imposes that the covariance $\text{Cov}(X^-, X^-) = \text{Cov}(X, X)$ and similarly $\text{Cov}(X^-, Y^-) = \text{Cov}(X, Y)$ and $\text{Cov}(Y^-, Y^-) = \text{Cov}(Y, Y)$. By noticing that $\Sigma_{X,X} = \sigma_X^2$, $\Sigma_{X,Y} = \sigma_X \sigma_Y \rho_{X,Y}$, $\Sigma_{Y,Y} = \sigma_Y^2$, $\Sigma_{Y,X^-} = \sigma_X \sigma_Y \rho_{Y,X^-}$ and $\Sigma_{Y,Y^-} = \sigma_Y^2 \rho_{Y,Y^-}$, one has

$$|\text{Cov}(\{\mathbf{Y}_{t+h}, \mathbf{Y}_{t,\lambda}^-\})| = \sigma_Y^4 (1 - \rho_{Y,Y^-}^2); \quad (10.27)$$

$$|\text{Cov}(\{\mathbf{X}_{t,\tau}^-, \mathbf{Y}_{t,\lambda}^-\})| = \sigma_X^2 \sigma_Y^2 (1 - \rho_{X,Y}^2). \quad (10.28)$$

$$\begin{aligned}
|\text{Cov}(\{\mathbf{Y}_{t+h}, \mathbf{X}_{t,\tau}^-, \mathbf{Y}_{t,\lambda}^-\})| &= \sigma_X^2 \sigma_Y^4 (1 - \rho_{X,Y}^2) \\
&\quad - \sigma_X^2 \sigma_Y^4 \rho_{Y,X^-} (\rho_{Y,X^-} - \rho_{X,Y} \rho_{Y,Y^-}) \\
&\quad + \sigma_X^2 \sigma_Y^4 \rho_{Y,Y^-} (\rho_{X,Y} \rho_{Y,X^-} - \rho_{Y,Y^-}) \\
&= \sigma_X^2 \sigma_Y^4 (1 - \rho_{X,Y}^2 - \rho_{Y,X^-}^2 + 2\rho_{Y,X^-} \rho_{X,Y} \rho_{Y,Y^-} - \rho_{Y,Y^-}^2);
\end{aligned} \tag{10.29}$$

$$|\text{Cov}(\{\mathbf{Y}_{t,\lambda}^-\})| = \sigma_Y^2; \tag{10.30}$$

Overall it results

$$T_{\mathbf{X} \rightarrow \mathbf{Y}} = \frac{1}{2} \log \frac{(1 - \rho_{Y,Y^-}^2)(1 - \rho_{X,Y}^2)}{1 - \rho_{X,Y}^2 - \rho_{Y,X^-}^2 - \rho_{Y,Y^-}^2 + 2\rho_{Y,X^-} \rho_{X,Y} \rho_{Y,Y^-}}. \tag{10.31}$$

One can notice that $T_{\mathbf{X} \rightarrow \mathbf{Y}} = 0$ when $\rho_{Y,X^-} = 0$ and either ρ_{Y,Y^-} or $\rho_{X,Y}$ or both are zero. Indeed, $Y \perp X^-$ implies that Y must be independent from one or both Y^- and X , or conversely $Y \not\perp Y^-$ and $Y \not\perp X$ implies $Y \not\perp X^-$.

The expression for the conditional transfer entropy $T_{\mathbf{X} \rightarrow \mathbf{Y} | \mathbf{Z}}$ can be obtained following the same reasoning just adding an extra conditioning set, \mathbf{Z} , of conditioning variables:

$$T_{\mathbf{X} \rightarrow \mathbf{Y} | \mathbf{Z}} = \frac{1}{2} \log \left(\frac{|\text{Cov}(\{\mathbf{Y}_{t+h}, \mathbf{Y}_{t,\lambda}^-, \mathbf{Z}\})| \cdot |\text{Cov}(\{\mathbf{X}_{t,\tau}^-, \mathbf{Y}_{t,\lambda}^-, \mathbf{Z}\})|}{|\text{Cov}(\{\mathbf{Y}_{t+h}, \mathbf{X}_{t,\tau}^-, \mathbf{Y}_{t,\lambda}^-, \mathbf{Z}\})| \cdot |\text{Cov}(\{\mathbf{Y}_{t,\lambda}^-, \mathbf{Z}\})|} \right). \tag{10.32}$$

The simplicity of computation and generality of these expressions for the transfer entropy is quite remarkable. The quantification of causality from observations using this formula only depends on the estimate of the covariances or rather their determinants. Despite the simplicity of this, their evaluation can be challenging in high-dimensional cases, such as when several lags are considered simultaneously, or when the conditioning on a large set of variables is considered. In general, one needs at least a number of observations larger than the number of variables to obtain non-zero determinants. The issue of estimation of covariances from data is addressed in Chapter 15.

Example 10.2 (Wiener-Granger causality and transfer entropy for multivariate normal variables). Let me briefly discuss here the equivalence between linear Granger causality and transfer entropy for multivariate normal variables. As I pointed out in Remark 8.9, the conditional entropy $H(A|B)$ between two random variables A and B is the entropy of the error of the regression of A with respect to B . In the linear case, this error is given by the linear regression, $A|B = A - \beta B = \epsilon$, and its entropy is

$$H(A|B) = H(\epsilon) = \frac{1}{2} \log \text{Var}(\epsilon). \tag{10.33}$$

For the transfer entropy, I have therefore

$$H(Y_{t+h} | \mathbf{X}_{t,\tau}^-, \mathbf{Y}_{t,\lambda}^-) = \frac{1}{2} \log \text{Var}(\epsilon_2), \quad (10.34)$$

where ϵ_2 is the error of $Y_{t+h} | \mathbf{X}_{t,\tau}^-, \mathbf{Y}_{t,\lambda}^-$, which is given by the regression Eq.10.4, and

$$H(Y_{t+h} | \mathbf{Y}_{t,\lambda}^-) = \frac{1}{2} \log \text{Var}(\epsilon_1), \quad (10.35)$$

where ϵ_1 is the error of $Y_{t+h} | \mathbf{Y}_{t,\lambda}^-$, which is given by the regression Eq.10.3. Therefore

$$T_{\mathbf{X} \rightarrow \mathbf{Y}} = \frac{1}{2} \log \frac{\text{Var}(\epsilon_1)}{\text{Var}(\epsilon_2)}. \quad (10.36)$$

Consequently, $T_{\mathbf{X} \rightarrow \mathbf{Y}} = \frac{1}{2} F_{\mathbf{X} \rightarrow \mathbf{Y}}$ and this demonstrates that the original Granger causality is the linear bivariate case of the transfer entropy approach. From the discussion in Section 10.3.1 it should be also clear that the generalized non-linear Wiener-Granger causality expressed in terms of the regressions Eq.10.7 is equivalent to transfer entropy also for the non-linear general case. Even for entropies different from the Shannon one.

Note that Eq. 10.20 indicates that to compute Granger causality there is no need to actually compute the regressions in Eqs. 10.3 and 10.4. Rather, the estimate of the determinants of the covariance matrices is sufficient:

$$F_{\mathbf{X} \rightarrow \mathbf{Y}} = \log \left(\frac{|\text{Cov}(\{Y_{t+h}, \mathbf{Y}_{t,\lambda}^-\})| \cdot |\text{Cov}(\{\mathbf{X}_{t,\tau}^-, \mathbf{Y}_{t,\lambda}^-\})|}{|\text{Cov}(\{Y_{t+h}, \mathbf{X}_{t,\tau}^-, \mathbf{Y}_{t,\lambda}^-\})| \cdot |\text{Cov}(\{\mathbf{Y}_{t,\lambda}^-\})|} \right). \quad (10.37)$$

Which is the same as Eq.10.20 except for the 1/2 factor in front. Contrary to the original expressions for the linear regressions, these expressions do not require the computation of the inverse covariance but instead the estimation of the determinants. Nonetheless, one can notice that the result is undefined if the determinants are zero and therefore in the cases when the matrices are not invertible.

Remark 10.7. It has been pointed out by James et al. [2016] that transfer entropy does not directly quantify the flow of information between variables, it is rather a more subtle concept associated with the reduction of uncertainty. It seems to me that the name ‘transfer entropy’ captures well this subtlety.

10.4.3 Enhancement and inhibition of causation

Extension of transfer entropy to more than two groups of variables is not trivial. Indeed, one faces the same problems described in Section 8.9.2: beyond pair-

wise interactions, there are synergic and redundant terms that contribute with opposite signs.

Let me restrict the attention to the case of three variables, or three sets of variables **X**, **Y**, and **Z**. In this case, one might ask which effect has the set of variables **Z** over the causal relationship between the sets **X**, **Y**? This is a question of practical relevance, indeed in many situations, the presence of other variables can affect the observed causation and this effect can be both enhancement or suppression. What one wants to quantify is, therefore, the difference between the transfer entropy between two sets of variables **X**, **Y** with respect to the transfer entropy between the same two sets but conditioned to a third set **Z**, which is

$$T_{\mathbf{X} \rightarrow \mathbf{Y}} = T_{\mathbf{X} \rightarrow \mathbf{Y}} - T_{\mathbf{X} \rightarrow \mathbf{Y} | \mathbf{Z}} \quad (10.38)$$

where the conditioning should be to the past of **Z**. A negative value of $T_{\mathbf{X} \rightarrow \mathbf{Y}}$ corresponds to an enhancement induced by **Z** to the causation of **X** on **Y**. Contrarily, a positive value of $T_{\mathbf{X} \rightarrow \mathbf{Y}}$ corresponds to an inhibition induced by **Z** to the causation of **X** on **Y**. This is an asymmetric measure where the roles of the three sets of variables are very distinct.

Remark 10.8. The conditional transfer entropy (Section 10.4.1) is the transfer entropy between two conditioned sets of variables $\mathbf{X}|\mathbf{Z}$ and $\mathbf{Y}|\mathbf{Z}$ and it coincides with the transfer entropy between the unconditioned **X** and conditioned $\mathbf{Y}|\mathbf{Z}$. Instead the transfer entropy between conditioned $\mathbf{X}|\mathbf{Z}$ and unconditioned **Y** has a different expression.

10.5 Deeper insights into causation

The Wiener-Granger approach provides an extremely useful and computationally simple way to quantify causality. However, this approach alone might not be sufficient to establish the true cause of a phenomenon.

Philosophers have been arguing for millennia about causation. It is indeed the fundamental element for the construction of knowledge. I find the first Chapter of the book by Hulswit [2002] to be a clear accounting of philosophers' elaborations around the idea of causation and I recommend to refer to that publication for a detailed accounting. Let me here provide a short and simplified accounting of the evolution of these ideas. Since, at least Aristotle it was clear that causality is related to phenomena that happen in succession but it was also understood that the observation of such a combination of phenomena is not sufficient to establish causation. Descartes proposed that cause must be "efficient" and provoke directly the effect and he argued that there must be a logical "necessity" between cause and effect. This was further elaborated by Spinoza and Leibniz. The need for an action to provoke an effect was argued by Locke and, later, Newton introduced the idea that there must be an active power, or a force, that modifies the state of

another body producing the observed effect. Hume proposed three factors that must be present in causal relation: (1) contiguity (in space and time) of cause and effect; (2) priority in time of cause to effect; and (3) a necessary connection between cause and effect. Kant argued that the effect not only should follow the cause but it must be a necessity after the cause following a “universal” rule. Mill summarized the concept of necessity in two statements: 1. ‘*A* is the cause of *B*’ means ‘*A* is the initiator of a change in *B*’; 2. ‘*A* is the cause of *B*’ means ‘Given the occurrence of *B*, *A* must necessarily have occurred’.

The requirement of necessity puts into question the possibility of measuring causality from observations only. Indeed, through observations one can only estimate the likelihood of the occurrence of an event but one cannot establish a mechanical necessity. From these philosophers’ refinements on the idea of causation it should be clear that, while the Wiener-Granger approach does accounts for some of the philosophers’ prescriptions, it falls short on others. Wiener’s concept that the past of *A* provides information about the future of *B* is in line with the idea that some “efficient” action must be done by *A* to cause *B*. Further, as far as the steps ahead and lags are small, the process is contiguous and has the correct priority satisfying the first two of Hume’s conditions. Instead ‘necessity’ and ‘universality’ cannot be directly accounted for with the Wiener-Granger approach. They require a proper investigation of the phenomenon where the regressions or information transfer methodologies can be used as instruments. In other words, indications for causality can be quantified with statistical methods, but other actions, beyond data analytics, must be undertaken to establish causation in a definite way.

The impossibility to establish causality from the sole statistical analysis and the need to include actions in the search for causation has been clearly argued by Pearl [Pearl et al., 2000, Pearl, 2019] who proposed a causal hierarchy organized into three layers: level 1. is Associational, and concerns the observation of co-occurring phenomena; level 2. is Interventional, and one investigates what happens if an intervention is done; level 3. is Counterfactual, and one investigates what would have happened if an intervention was performed, given that something else, in fact, occurred. These levels are hierarchical because to be able to answer the third, one must be capable to handle the second, which requires the first for its construction from observations. Pearl argues in Pearl [2019], Pearl and Mackenzie [2018] that, from pure observational statistical analysis, the top level cannot be learned. This, in general, requires a model of the system that goes beyond probability. Other kinds of modeling, such as agent-based modeling or ab initio simulations, could be more adequate for this task. The Wiener-Granger approach to causality is limited to the first level and can be used as an investigation tool for the two other levels by devising appropriate experiments.

Example 10.3 (Establish causation from observations and experiments).
Let me reconsider the example of a turning car at a road crossing that I

have used at the beginning of this chapter. I argued that the observations of directional lights switching on preceding the turning of the car is an indication of dependency between the two phenomena however, it cannot be an indication of causation. In this example, observation alone cannot establish causal relation. Nonetheless, specifically designed observations and experiments can bring us closer to establishing the actual cause of the phenomena.

For instance, one could observe that, while at the crossing the directional lights and the car turning are related with large likelihood, there are other instances, in other locations, where the car turns without directional lights. This is indeed the case when the road turns without crossings. Therefore, the necessity of directional lights for cars to turn can be excluded.

Furthermore, now that direct causation has been excluded but dependency has been established, one can argue that there must be a third variable that first causes the lights to switch on and then makes the car turn: the driver. If we cannot see inside the car and cannot analyze its mechanics, the existence of a driver remains a hypothesis, a hidden variable. This would be a working hypothesis, which could be reinforced by observing other phenomena that require the intervention of another factor to happen.

10.6 Data-driven modeling tutorial

<https://github.com/FinancialComputingUCL/DataDrivenModeling/Ch10>

The tutorial for this Chapter covers various topics on causality including: lagged correlations; linear Granger causality; non-linear Granger causality and Transfer entropy.

Exercises

- Given the time series X_s and Y_s generated from:

$$\begin{aligned} Y_t &= -aY_{t-1} + bX_{t-2} + \epsilon_t \\ X_t &= -cX_{t-1} + \eta_t \end{aligned} \tag{10.39}$$

with ϵ_t and η_t random variables with zero mean and unitary variance and a, b, c positive constants.

- Discuss whether Y is caused by X or vice versa.
- Demonstrate that some lagged correlations are not zero.
- Discuss which lags must have a causality signal.

- iv. Compute the expression for $F_{\mathbf{X} \rightarrow \mathbf{Y}}$ and $T_{\mathbf{X} \rightarrow \mathbf{Y}}$ assuming ϵ_t and η_t are univariate normal with zero means, unitary standard deviation and are uncorrelated.
 - v. By using the F-distribution, compute the p -value for $a = b = c = 1$.
- A process has $Y_t|Y_{t-1} \sim \mathcal{N}(0, 1)$ and $Y_t|Y_{t-1}, X_{t-1} \sim \mathcal{N}(0, 1/2)$, quantify causality using information theoretic measures.

11

Networks as representations of complex systems

We live in a small world where every individual, every firm, and every place are connected to every other through a network of intricate links. This interconnection provides strength to our society and great potential opportunities for individuals. However, this is also the vehicle through which risk can spread. This is true for all complex systems where the web of relations between the variables makes the system both resilient and vulnerable. The modeling of complex systems requires modeling this network of interactions and the processes that occur within.

A network model of a complex system is an abstract object representing the relevant interactions in the system. I shall discuss in the next chapter that it is also a tool to produce a model of the system in terms of multivariate probability. In this chapter, I'll guide the reader through a few methodologies to build representation networks for the modeling of complex systems.

Networks are rather powerful tools because, despite their simplicity and intuitiveness, they represent complex objects that would otherwise be hard to describe and handle. Networks have the added advantage that our brain is trained to visualize patterns in complex, interconnected systems and therefore the use of networks for visualization and investigation of complex structures can be very effective.

11.1 Network construction by pruning or joining

Networks might be constructed at random simply by picking disconnected vertices and joining them with a given probability. Such networks are called ‘random graphs’, but often they are called Erdős-Rényi model after the Hungarian mathematicians, Paul Erdős and Alfréd Rényi, who first described this random construction and computed the properties of such random graphs [Erdős, 1959] (see also Newman [2018]).

Example 11.1 (Random networks). The Erdős-Rényi approach is the simplest way to generate a random graph. It requires only setting the probability to join two vertices with an edge or equivalently the number of edges to insert. The construction proceeds by choosing at random a couple of ver-

tices not selected before and connecting them with an edge until the desired density is reached. One can build variations around this elementary model, for instance forbidding some edges or enforcing some structure, or making the probability of an edge dependent on some properties of the vertices.

Some networks are constructed with a mix of randomness and constraints. For instance, one might want to construct a network with a given degree distribution but where edges are added at random. This is called the ‘configurational model’.

Example 11.2 (Configurational model). To generate a random graph which, despite being random, has a defined degree distribution one can start from vertices already connected with the right number of ‘half edges’, called ‘stubs’. Then one proceeds by connecting at random these stubs obtaining a randomly connected network with the desired degree distribution. Of course, in order to join all the stubs, one needs an even number of them, therefore the desired degree sequence is achieved exactly only if the sum of the degrees is an even number (see Newman [2018]).

There are several possible constraints or biases that yield various kinds of models. In general, these are used to test how random real networks are and therefore produce null models (see Section 18.1). The idea is to compare a real network with random networks and establish if the properties of the real network are significantly different from the ones of a random network and therefore if there is extra information contained in the structure of the real one.

Another procedure to construct networks is to join vertices in order to maximize some gain function. When the gain function is the product of the properties of the vertices, then several properties of the network can be computed. This is the so-called ‘fitness model’.

Example 11.3 (Fitness model). The fitness model focuses on vertices and their associate ‘fitness’ ξ_i . Two vertices i, j , are joint depending on the fitnesses of both. In this model, the probability to join the vertices is, therefore, a function of the fitnesses ξ_i : $p_{i,j} = g(\xi_i, \xi_j)$ Caldarelli et al. [2002]. A version of the fitness model is the preferential attachment where iteratively vertices attach to each other depending on their existing degree Bianconi and Barabási [2011]. In such a model, a vertex with a larger degree has more chances to be chosen in the so-called ‘rich get richer’ mechanism. This leads to a power law degree distribution with exponent $\gamma = 3$.

From a very general perspective, networks can be constructed either by adding edges to a sparser graph (normally starting from a disconnected graph with no edges) or by pruning edges from a denser network (normally starting from the

complete graph where all couple of edges are connected). These two procedures might be combined in some adaptive methodologies. The networks produced by adding or pruning can result to be the same and often the choice regarding whether to prune or add is dictated by computational convenience. Indeed, sparse networks require less adding than pruning.

Besides edges, higher-order elements, such as triangles or tetrahedra, can be used for the construction of networks. Indeed, I shall present in Section 17.2 an approach to building clique forests by joining simplexes. Another example is the construction of complex networks by combining ‘simple’ Platonic solids Song et al. [2012a].

Example 11.4 (Thresholding). Given a weighted graph with scalar measure $w_{i,j}$ associated with each couple of vertices, a simple way of pruning is by thresholding, where one prunes edges that carry weights that are lower than a given threshold and keeps instead the ones above such a threshold (or vice versa). The sparsity of the graph depends on the thresholding value. A reciprocal procedure that leads to the same graph is to start from a set of disconnected vertices and add edges that carry the largest weight ending the procedure when all edges associated with weights above a given threshold value have been inserted. The weight measure $w_{i,j}$ that is thresholded can be anything. For instance, it can be a function of the fitnesses of the vertices, retrieving, therefore, the fitness model. There is a rather vast literature that studies networks constructed by thresholding the correlation matrix $w_{i,j} = \rho_{i,j}$. These networks are called ‘correlation networks’ (see Example 17.1).

11.2 Information filtering networks

It is quite general and intuitive that the network representation of a complex system with several variables and many interactions between them must retain the most relevant links while discarding the spurious ones. This is for instance the reasoning for the networks generated by thresholding on edge weights discussed in Example 11.4. However, thresholding unavoidably results in over-representing strong but redundant relations, typically making parts of the network very dense, while discarding weak but important relations, often resulting in a disconnected overall structure. Connectedness is a global topological property, which is often desirable. The main idea underlying information filtering network construction is to generate a sparse graph that retains the most relevant links while also preserving some global properties. For instance, one often aims to prune the network as much as possible but also retain it connected, this forces the network into a tree structure. Indeed, trees are the connected graphs with the minimum number of links (they have $p - 1$ for p vertices). Beyond connectedness and trees, there are several other desirable topological constraints. For instance, planarity or

cordality can be imposed generating other kinds of information-filtering networks. Let me start with the trees, which are the simplest, and then engage with more complex structures.

11.2.1 The minimum spanning tree (MST)

In 1926 a Czech scientist Otakar Boruvka, conceived a way to most efficiently design the power grid coverage of Moravia [Boruvka, 1926, Nešetřil et al., 2001]. His idea was to avoid cycles and connect each point that needed power from the powered point at the shortest distance. Such a network is a tree (it has no cycles), it connects all vertices (it is ‘spanning’) and it is the connected graph with minimum total edge distance. It is called the minimum spanning tree.

Definition 11.1 (Minimum Spanning Tree (MST)). The **Minimum Spanning Tree (MST)** is a connected, undirected, graph with positive edge weights with the minimum total sum of edge weights. Analogously one can define the **Maximum Spanning Tree** as the spanning tree with maximum total edge weight sum. MSTs have no cycles, all vertices are connected by a walk and have $|E| = p - 1$ edges.

In most practical applications the maximum spanning tree is the structure of interest, however, for historical reasons most algorithms are devised for the minimum spanning tree problem. Clearly, the two problems can be mapped into each other. For instance, if the set of edge weights is $w_{i,j} > 0$ with $i, j \in [1, \dots, p]$, by searching for the minimum spanning tree of $1/w_{i,j}$ one obtains the maximum spanning tree for $w_{i,j}$.

Remark 11.1. It is actually a very bad idea to use a tree structure for a power grid network. Indeed, such a structure is extremely vulnerable to failure: the removal of any edge will result in disconnecting a part of the network. In general, systems must be able to keep functioning despite some local failures, for this a power grid needs cycles that provide alternative paths. Optimization tends to produce fragile systems. Some redundancy is good to prevent systems from systemic failures. One might notice that redundancy, is always naturally integrated in the design of biological systems.

There are several possible algorithms that can be used to build an MST. The two best-known, and most intuitive, are named after two mathematicians who introduced them, Prim and Kruskal (even if the former algorithm was discovered by others before Prim). They are both greedy algorithms and they both start from a set, \mathbf{S} , of edge weights and generate an MST with computational complexity $|\mathbf{S}| \log |\mathbf{V}|$ (see Boruvka [1926], Prim [1957], Dijkstra et al. [1959], Kruskal [1956]). There are algorithms that can perform faster than Prim’s and Kruskal’s, down to

almost linear time $\mathcal{O}(|\mathbf{S}|)$ [Chazelle, 2000]. Note that in the contexts in which I will apply these approaches, I start from dense networks with $\mathcal{O}(|\mathbf{S}|) = \mathcal{O}(|\mathbf{V}|^2)$.

Algorithm 11.1: Prim's algorithm for MST

input A set \mathbf{S} of edges (i, j) with positive weights, $w_{i,j} > 0$.
initialize Start from an empty graph, $\mathbf{E} = \emptyset$, $\mathbf{V} = \emptyset$.
 Choose a vertex i at random and include it in the MST, $\mathbf{V} \leftarrow i$.
while There are still vertices not included in the MST, $|\mathbf{V}| < p$, **do**

- Find the edge with the smallest weight connecting a vertex in the MST with another vertex not yet included in the MST:
 $v = \min_{j \notin \mathbf{V}} (w_{k,j} | k \in \mathbf{V})$.
- Include the vertex v in the vertex set: $\mathbf{V} \leftarrow v$.
- Include the edge (v, k) in the edge set: $\mathbf{E} \leftarrow (v, k)$.

output The MST: $\mathcal{G} = (\mathbf{V}, \mathbf{E})$.

Algorithm 11.2: Kurskal's algorithm for MST

input A set \mathbf{S} of edges (i, j) with positive weights, $w_{i,j} > 0$.
initialize Set $\mathbf{E} = \emptyset$, $\mathbf{V} = (1, \dots, p)$.
initialize Create a forest where every vertex is a separate tree: $\mathcal{T} \leftarrow \mathbf{V}$.
while There are still edges in \mathbf{S} to add **do**

- Find the edge in \mathbf{S} with smallest weight connecting two different trees: $(u, v) = \min_{k,j \in \mathbf{S}} (w_{k,j} | k \in \mathbf{t}_a, j \in \mathbf{t}_b, \mathbf{t}_a, \mathbf{t}_b \in \mathcal{T}, \mathbf{t}_a \neq \mathbf{t}_b)$.
- Join the trees \mathbf{t}_a and \mathbf{t}_b through the new edge (u, v) creating a single tree \mathbf{t}_c .
- Remove \mathbf{t}_a and \mathbf{t}_b from \mathcal{T} and add \mathbf{t}_c .
- Remove the edge (u, v) from \mathbf{S} and include it to the forest's edge set: $\mathbf{E} \leftarrow (u, v)$.

output The MST or forest: $\mathcal{G} = (\mathbf{V}, \mathbf{E})$.

Remark 11.2. The MST has two nice properties: first, if there is no degeneracy and each edge has a distinct weight, then there is only one, unique minimum spanning tree; second, the complexity of the problem can be solved in polynomial time, as indeed the two greedy Prim's and Kurskal's algorithms demonstrate. Interestingly, a small variation on the problem,

such as finding the three with minimum weight which spans only a subset of vertices (Steiner tree problem) is instead NP-complete.

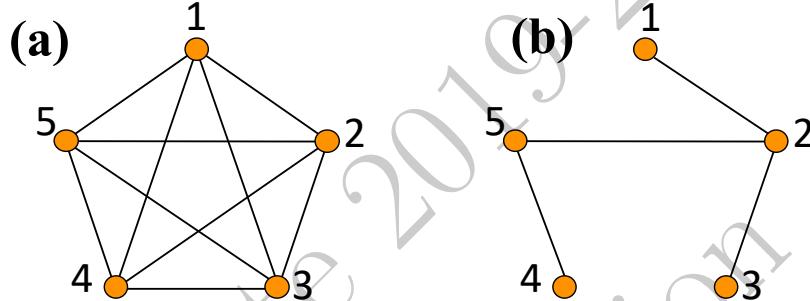


Figure 11.1 (a) The complete graph, K_5 , is made of 5 vertices all connected to one other through 10 edges. (b) A spanning three sub-graph; it is a connected graph with 4 edges and no cycles.

MSTs are used in many different domains, indeed it is often very useful to extract a connected network with minimum total weight. They have the nice characteristic of being optimally connected using the minimum possible number of edges ($p - 1$). In particular, the maximum spanning tree can be regarded as a backbone structure representing the essential interactions in the system where only the most important connections are retained.

Remark 11.3. Maximum spanning tree construction. Often the network representation that one aims to construct for modeling purposes must maximize the sum of the weights and not minimize it. This is the case of the ‘Maximum Spanning Tree’ defined earlier (see Definition 11.1). The two constructions are complementary and can be straightforwardly interchanged. Namely for the Maximum Spanning Tree, in Prim’s and Kurscal’s algorithms, one must seek to connect the edge with the largest weight instead of the one carrying the smallest one.

11.2.2 Planar maximally filtered graphs (PMFG)

The MST problem could be seen as a minimization (or maximization) problem with a topological constraint: find the graph with minimum (maximum) weight constrained to being connected and having no cycles. From this perspective, one could conceive several other minimization (maximization) problems on networks under some topological constraint. Topologically, the simplest structures beyond trees are planar graphs.

Definition 11.2 (Planar graphs and maximal planar graphs). A **planar graph** is a graph that can be embedded on a sphere without edge intersections.

A planar graph to which no edges can be added without violating planarity is called **maximal planar graph**. A maximal planar graph cannot have cycles larger than three and they are therefore **triangulations** of the sphere. Any triangulation of the sphere is maximal planar and it has $3p - 6$ edges and $2p - 4$ surface triangles.^a

^a Notice that the total number of 3-cliques can be larger because there can be collar cycles of three edges that are not surface triangles.

Analogously to the MST problem, one might want to search for the planar graph with maximal edge weight.

Definition 11.3 (Maximum weight planar graph problem (MWPG)). The problem of finding the planar graph with maximum edge weight is called **maximum-weight planar graph** problem. Sometimes called maximum-weight planar subgraph problem because one searches for the planar subgraph of the complete graph K_p with weights $w_{i,j} > 0$ with $i, j \in (1,..p)$.

However, differently to the MST, where the solution is discoverable in polynomial time, in this case, the problem is NP-hard. Nonetheless, there are algorithms that reach sub-optimal solutions in polynomial time.

Planar maximally filtered graphs (PMFG)

As for all hard problems one can devise approximate solutions that provide good results in polynomial time. One of these solutions was proposed by Tumminello et al. [2005], where the Authors introduced a Greedy algorithm, analogous to Kurscal's algorithm for the MST (see Algorithm 11.2). Such a network was named Planar Maximally Filtered Graph (PMFG).

Algorithm 11.3: PMFG construction for MWPG

```

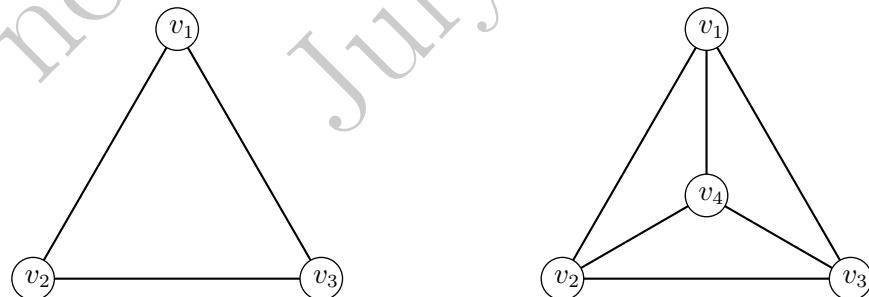
input A  $p \times p$  matrix of edges with positive weights,  $w_{i,j} > 0$ .
initialize Create an ordered set of edges in descending weight rank:
 $S_k = (v_k, u_k)$  s.t.  $w_{v_{k+1}, u_{k+1}} < w_{v_k, u_k}$ .
initialize Start from an empty graph  $\mathbf{E} = \emptyset$ ,  $\mathbf{V} = \emptyset$ ,  $k \leftarrow 1$ .
while There are still edges to include in the PMFG,  $|\mathbf{E}| < 3p - 6$ , do
    if including edge  $S_k = (v_k, u_k)$  from the ordered set in the
    graph does not violate planarity then
        • Include the vertices  $v_k, u_k$  in the vertex set, if not already in-
          cluded:  $\mathbf{V} \leftarrow v_k, u_k$ .
        • Include the edge  $(v_k, u_k)$  in the edge set:  $\mathbf{E} \leftarrow (v_k, u_k)$ .
         $k \leftarrow k + 1$ .
output The PMFG:  $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ .

```

It was proved in Tumminello et al. [2005] that the PMFG structure always includes the maximum spanning tree as a sub-structure. This algorithm is $\mathcal{O}(p^3)$ with the slowest part being the verification of planarity (a $\mathcal{O}(p^2)$ operation) which must be done at each stage of the process when a new edge is included in the PMFG ($\mathcal{O}(p)$ times). Such validation can be eliminated by adopting a different perspective and building maximal planar graphs using construction moves that automatically preserve planarity.

11.2.3 Triangulated maximally filtered graphs (TMFG)

Among the large range of possibilities for the construction of meaningful planar network representations, there exists a very simple graph-construction move that preserves embedding. It consists in adding a vertex inside a triangle that is lying on the surface and connecting it to the tree vertices of the triangle forming in this way three new surface triangles as displayed in the following drawing where vertex v_4 is inserted inside the v_1, v_2, v_3 triangle:



It is clear that the new structure with four vertices (a tetrahedron) can be drawn on the same portion of the plane of the original triangle and does not affect the planarity of the graph. This procedure to construct maximal planar graphs

with the largest edge weights was introduced in Massara et al. [2017] and named Triangulated Maximally Filtered Graph (TMFG).

Algorithm 11.4: TMFG construction for MWPG

input A $p \times p$ matrix of edges with positive weights, $w_{i,j} > 0$.
initialize Start from the triangle (u_1, u_2, u_3) with largest edge weight.
initialize $\mathbf{V} \leftarrow u_1, u_2, u_3$, $\mathbf{E} \leftarrow (u_1, u_2), (u_2, u_3), (u_3, u_1)$,
 $\mathbf{T} \leftarrow (u_1, u_2, u_3)$.
while There are still vertices not included in the TMFG, $|\mathbf{V}| < p$, **do**

- Find the vertex v that yields to maximum weight with existing triangles in \mathbf{T} : $v = \max_{k \notin \mathbf{V}} (w_{k,j_1} + w_{k,j_2} + w_{k,j_3} | (j_1, j_2, j_3) \in \mathbf{T})$.
- Add three new triangles to $\mathbf{T} \leftarrow (v, j_1, j_2), (v, j_2, j_3), (v, j_3, j_1)$.
- Remove triangle (j_1, j_2, j_3) from \mathbf{T} .
- Update $\mathbf{V} \leftarrow v, \mathbf{E} \leftarrow (v, j_1), (v, j_2), (v, j_3)$

output The TMFG: $\mathcal{G} = (\mathbf{V}, \mathbf{E})$.

This algorithm has complexity $\mathcal{O}(p^2)$ and it can be made efficient by keeping a gain table that associates vertices, not included in the TMFG, with triangular faces of the TMFG (see Algorithm 1 in Massara et al. [2017]). The resulting network is a clique-tree (see Definition 11.4) made of 4-cliques separated by 3-cliques, it has $3p - 6$ edges which is the maximum number of edges for a planar graph. The TMFG graph is therefore maximally planar (see Definition 11.2). An example of a simple TMFG is illustrated in Fig.11.2.

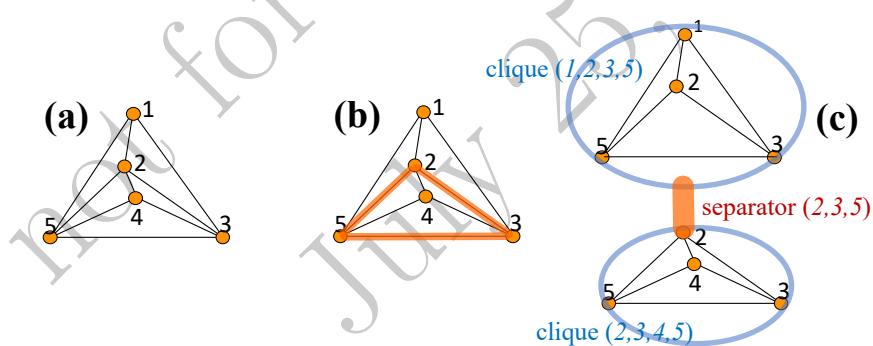


Figure 11.2 Example of a simple TMFG made of 2 cliques (the tetrahedra $(1, 2, 3, 5)$ and $(2, 3, 4, 5)$) and one separator (the triangle $(2, 3, 5)$). (a) reports the TMFG network. (b) highlights the separator $(2, 3, 5)$. (c) represents the network as a clique tree made of 2 cliques connected by one separator.

Definition 11.4 (Clique trees and forests). A **clique tree** is a graph made of a set of cliques that are connected in a tree structure by means of a set of separators. A **separator** is a clique that, if removed from a connected graph, makes the resulting graph disconnected. The number of components resulting from the removal of a separator is one plus the **multiplicity of the separator**. A **clique forest** is a disconnected set of clique trees.

Definition 11.5 (Clique tree width). The size of the largest clique in a clique tree minus one is the graph **treewidth**. Many optimization problems in graphs (including some NP-hard and NP-complete) can be solved efficiently in polynomial time in graphs with small treewidth.

Example 11.5 (The spanning tree as a clique tree). A simple example of a clique tree is a spanning tree. In this case, the cliques have size two and are the edges. The separators have size one and are the subset of vertices with degrees larger than one. The multiplicity of a separator is the degree of the corresponding vertex minus one.

Example 11.6 (The TMFG as a clique-tree). The TMFG is a clique tree made of tetrahedra separated by triangles. In the construction in Algorithm 11.4 the clique and separators sets \mathcal{C}, \mathcal{S} can be constructed step-by-step by adding, within the while loop, the tetrahedron $\mathcal{C} \leftarrow (v, j_1, j_2, j_3)$ to the clique set and adding the 3-clique $\mathcal{S} \leftarrow (j_1, j_2, j_3)$ to the separator set. In this case, separators have multiplicity one they are 3-cliques joining two tetrahedra. Note that the PMFG might instead not be a clique tree. It was named ‘bubble three’ in Song et al. [2012b].

Clique trees involving cliques with more than 4 vertices are no longer planar graphs.

11.2.4 Graph embedding on surfaces of arbitrary genus

Beyond the sphere, one might want to explore networks that are drawn, without edge intersections, on more complex surfaces. In topology, a surface is a two-dimensional space that is locally homomorphic to the Euclidean plane. It is locally a two dimensional object that can be navigated with a two Euclidean coordinates system. Topological graph theory studies how graphs can be drawn on a surface without edge intersections. This is referred to as embedding a graph in a surface. The sphere is a very simple surface¹ on which one can embed only a restricted class

¹ I will only consider connected, non-empty, non-orientable topological surfaces with no boundaries.

of graphs (the planar graphs, indeed). More complex graphs can be embedded in more complex surfaces. A measure of the complexity of a surface is its **genus**.

Definition 11.6 (Surface genus and graph genus). The **genus** of a surface accounts for the number of ‘holes’ (or equivalently ‘handles’) in a surface: the sphere has no holes ($g = 0$), the torus has one hole ($g = 1$), etc.. It is also the maximum number of cuttings one could perform without disconnecting the surface.

The **genus of a graph** is the minimal genus of an embedding surface such that the graph can be drawn on it without edge intersections.

It was demonstrated that the complete graph with p vertices (K_p) can be embedded on a non-orientable surface with genus

$$\hat{g}_p = \left\lceil \frac{(p-3)(p-4)}{12} \right\rceil \text{ for } p > 2. \quad (11.1)$$

This is the inverse of the solution of the Map Color Theorem proved by Ringel and Youngs in 1968 [Ringel, 1974]. Given that any graph is a subgraph of K_p , then any graph can be embedded on a surface with a genus smaller or equal to \hat{g}_p . However, the problem of finding the embedding of a given graph on a surface is NP-hard.

Example 11.7 (Graph embedding and surface genus). It is evident that on a surface of a sphere ($g = 0$) one can always embed an isolated edge, but also any tree. A single cycle can be also always embedded because it is homomorphic to a circle. But when the graph becomes more connected eventually the graph cannot any longer be embedded on a sphere and a more complex embedding is needed. One can for instance verify that 3 and 4-cliques can be drawn on a sphere without edge intersections but 5-cliques cannot. Indeed, by using Eq.11.1 one can observe that: $\hat{g}_3, \hat{g}_4 = 0$ but $\hat{g}_5 = 1$. Therefore any graph containing K_5 as a subgraph cannot be embedded on the surface of the sphere. This is indeed a core part of the **Kuratowski theorem** which states that *a graph is planar if and only if it does not contain a subgraph that can be constructed as a subdivision of K_5 or a complete bipartite graph with 3 vertices ($K_{3,3}$)*.

Clique trees are a natural extension of the concept of trees beyond simple structures. They have a very important property: clique trees are chordal (see Definition 4.11) and also any chordal graph is a clique tree or forest. This makes this class of graph extremely useful for probabilistic modeling because, as I shall discuss in the next Chapter (Section 12.1), on top of chordal structures one can construct exact inference models. Let me, therefore, extend the TMFG idea beyond planarity and generalize it to the construction of chordal graphs.

11.2.5 Maximally filtered clique forests (MFCF)

One can build a clique tree piece by piece by joining couples of cliques through subparts of them which become separators. One can consider a general problem where weights – or gains – can be associated with joining cliques and their separators rather than simply the edges that connect them. This opens a completely new scenario.

An algorithm for this construction can be designed in a similar way to Kurscal's algorithm for MSTs. However, the complexity of the problem scales up to $\mathcal{O}(p^k \log p)$ for arbitrary gain functions, where p is the number of vertices and k is the largest clique size and it is assumed $p \gg k$.

The computational complexity of the construction of a clique forest by joining together cliques can be estimated considering that, in an undirected graph with p vertices, there are $p(p-1)/2$ possible edges, $p(p-1)(p-2)/6$ possible triangles and in general $|\mathcal{C}_k| = \binom{p}{k}$ possible k -cliques. Any algorithm which explores the best way to join a pair of cliques would involve at least two stages: (i) the construction of a list with all k -cliques; (ii) the search for connections with other k -cliques sharing a separator. Without optimization, stage (i) creates a list of $\mathcal{O}(p^k)$ elements each of whom can attach to $\mathcal{O}(p^{k-s})$ other elements and the graph-construction process can require up to $\mathcal{O}(p^k)$ steps. This is a large computational burden accounting up to $\mathcal{O}(p^k \times p^{k-s} \times p^k) = \mathcal{O}(p^{3k-s})$ if not optimized. However, we have seen earlier that both Prim's and Kurscal's algorithms are in $\mathcal{O}(p^2 \log p)$ despite $k = 2$, $s = 1$, and therefore $3k - s = 5$. Indeed, searches can be reduced to $\mathcal{O}(\log p)$ on ordered lists.

Therefore, an optimized algorithm can find solutions in $\mathcal{O}(p^k \log p)$. However, still, if one wants to explore exhaustively all configurations, then $k = \mathcal{O}(p)$ and with current computational tools this method becomes uncomputable for systems of sizes larger than the order of ten vertices.

In analogy with the TMFG construction, one can simplify the complexity of the procedure by starting from a seed structure and attaching one vertex at a time instead of linking two full cliques. This produces an approximate solution which, given the seed, is constructible in $\mathcal{O}(p)$. Finding, the seed can be reasonably done in $\mathcal{O}(p^2)$ in most practical cases.

Algorithm 11.5: MFCF construction for maximal chordal weighted graph problem

input A gain function, $G(\cdot, \cdot)$.
input the minimum clique size, Min_Cl.
input the maximum clique size, Max_Cl.
input the number of times separators can be used (multiplicity), Max_Mult.
initialize Start from a seed clique \mathbf{c}_0 with vertices \mathbf{v}_0 and edges \mathbf{e}_0 :
 $\mathbf{V} \leftarrow \mathbf{v}_0$, $\mathbf{E} \leftarrow \mathbf{e}_0$, $\mathcal{C} \leftarrow \mathbf{c}_0$.
while There are still vertices not included in the MFCF, $|\mathbf{V}| < p$ **do**

- Find a vertex, $v \notin \mathbf{V}$, and a sub-clique, \mathbf{s} , with size larger than $\text{Min_C} - 1$ but smaller than Max_Cl , that return the largest gain: $v = \max_{k \notin \mathbf{V}} (G(k, \mathbf{s}) | \mathbf{s} \subseteq \mathbf{c} \in \mathcal{C} \text{ and } \text{Min_C} - 1 \leq |\mathbf{s}| < \text{Max_Cl})$.
- Create a new clique $\mathbf{c}' = \mathbf{s} \cup v$
- Add the new clique \mathbf{c}' to the clique set $\mathcal{C} \leftarrow \mathcal{C} \cup \mathbf{c}'$.
- If $\mathbf{s} = \mathbf{c}$ remove the clique \mathbf{c} from \mathcal{C} .
- Update $\mathbf{V} \leftarrow \mathbf{V} \cup v$, $\mathbf{E} \leftarrow (\text{edges between } v \text{ and } \mathbf{c})$.

output The MFCF($\text{Min_Cl}, \text{Max_Cl}, \text{Max_Mult}$): $\mathcal{G} = (\mathbf{V}, \mathbf{E})$.

By comparing the two Algorithms 11.5 and 11.4 one can notice that the MFCF construction is a generalization to arbitrary gain function, arbitrary clique size, arbitrary separator size and arbitrary separator multiplicity of the TMFG algorithm.² Accordingly with Algorithm 11.5 I shall denote MFCF network with minimum clique size Min_Cl, maximum clique size Max_Cl and maximum multiplicity Max_Mult, as:

$$\text{MFCF}(\text{Min_Cl}, \text{Max_Cl}, \text{Max_Mult}). \quad (11.2)$$

The MFCF network family includes the MST and the TMFG. In particular, the MST is an MFCF with both maximum clique and minimum clique equal to 2 and where a separator (a vertex in this case) can be used multiple times.

$$\text{MST} = \text{MFCF}(2, 2, p - 1). \quad (11.3)$$

Similarly,

$$\text{TMFG} = \text{MFCF}(4, 4, 1), \quad (11.4)$$

the TMFG is an MFCF with both maximum clique and minimum clique equal to 4 and where a separator (a triangle in this case) can be used only once. There are some extra details, for instance, the choice of the seed clique \mathbf{c}_0 is very important

² Arbitrariness is within the constraints that the clique sizes must be smaller or equal than the number of vertices in the graph, the separator size must be smaller or equal than the size of the smallest clique it joins and its multiplicity cannot be larger than the number of cliques.

and can significantly change the results. Furthermore, a threshold on the minimum gain could be added leaving therefore some vertices unattached. Variations of the algorithm can be devised to allow the construction of forests instead of trees. However, these details are not essential in the context of this book and I direct the interested reader to the reference papers [Massara and Aste, 2019, Massara et al., 2017] and the GitHub project Aste et al. for codes and practical inputs.

11.3 Higher order network representations

I have, so far, described networks in terms of sets of vertices (\mathbf{V}) and edges (\mathbf{E}) (i.e. $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, see Definition 4.1). From a geometrical perspective, a vertex is a zero-dimensional object (a point) while an edge is a segment, therefore a one-dimensional object. I have already discussed that networks can include elements with dimensions larger than one, such as triangles, tetrahedra, etc. In general, these are n -dimensional simplexes (see Definition 4.12) that, in many practical applications, might be crucial for the functional properties of the system that the network is representing, and they must be accounted for explicitly. The simplest representation of a network which includes higher-dimensional sub-structures can be produced by adding triplets (triangles), quadruplets (tetrahedra), etc. to the sets in \mathcal{G}

$$\mathcal{G} = (\mathbf{S}_0, \mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3, \dots) \quad (11.5)$$

with $\mathbf{S}_0 = \mathbf{V}$ the set of zero-dimensional simplexes, $\mathbf{S}_1 = \mathbf{E}$ the set of one-dimensional simplexes and, in general, \mathbf{S}_n the set of n -dimensional simplexes. It must be noted that in this representation cliques and simplexes might be different objects. A clique is a skeleton of edges of a complete sub-graph, while the corresponding simplex is a geometrical object in a n -dimensional space.

Let me also describe an alternative representation for higher-order networks, which is a native network approach and therefore can be more appealing in some contexts. This is a layered representation that explicitly takes into account the higher-order sub-structures and their interconnections in the network. This representation is particularly suitable for chordal graphs (see Definition 4.11). Extensions to more general classes of networks, which include different motifs other than simplexes, can be made, however, it is more complex and one might lose some uniqueness properties and therefore it is not considered in this book.

A visual example of a higher order chordal network (clique tree) is provided in Fig.11.3(a). In the figure, the maximal cliques (largest fully-connected subgraphs) are highlighted and plotted, in panel (b), as clique-tree nodes. Such vertices are connected to each other through separators (see Definition 11.4). One can observe that, in the figure, the separator constituted by the vertex '4' has multiplicity 1, while the separator constituted of edge '4-6' has multiplicity 2 and indeed it appears twice.

A layered higher-order representation can be constructed by representing the d -dimensional simplexes with vertices in layer d . The structure starts with the

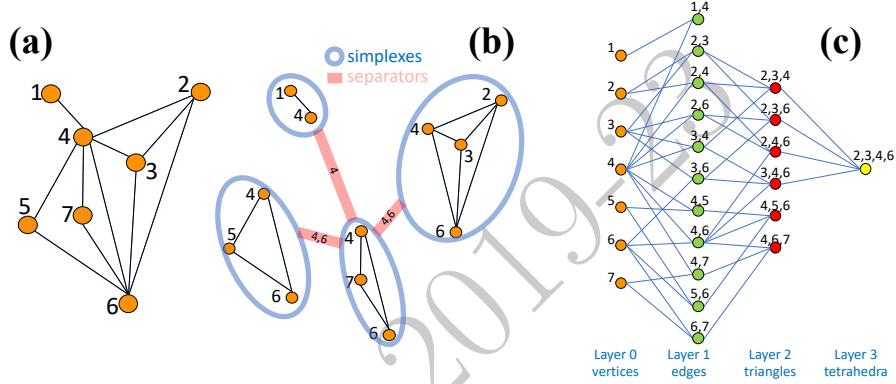


Figure 11.3 (a) Visual example of a higher order network made of 7 vertices, 11 edges, 6 triangles and 1 tetrahedron. (b) This higher-order network is a clique tree made of four cliques (maximal cliques highlighted with blue ellipses) connected through three separators (the tick red edges). (c) Layered representation of the same network where vertices in each layer L_d are associated with the d -dimensional simplexes in the structure.

vertices in layer 0; then a couple of vertices connect to edges represented with the nodes in layer 1; edges connect to triangles in layer 2; triangles connect into tetrahedra in layer 3, and so on. This is illustrated in Figure 11.3(c). One can verify that links between nodes in layers d and $d + 1$ are the connections between d to $d + 1$ simplexes in the network. The degree on the left of nodes in L_d is always equal to d . The degree on the right of nodes in L_d is the multiplicity of the simplex. The d -dimensional simplexes with no connections towards $d + 1$ are the maximal cliques in the network (i.e. the vertices in the clique tree in Fig.11.3(b)). Such representation has a one-to-one correspondence with the original network but shows explicitly the simplexes and sub-simplexes and their interconnection in the structure. All information about the network, at all dimensions, is explicitly encoded in this representation including elements such as maximal cliques, separators, and their multiplicity (see caption of Fig.11.3 for further details).

Clearly, this layered representation can be directly used for the characterization of the structure and for a detailed description of its components and the relations between them.

One might notice the resemblance between this layered structure and the layered architecture of deep neural networks. Indeed, these representations can be directly used for computational purposes [Wang and Aste, 2022].

11.4 Data-driven modeling tutorial

<https://github.com/FinancialComputingUCL/DataDrivenModeling/Ch11>

The tutorial for this Chapter covers various topics on network representations including: construction of networks by pruning; construction of the TMFG graph; construction of MFCC graphs.

Exercises

- Prove that a connected graph with p vertices must have at least $p - 1$ edges and that a connected graph with such a minimum number of edges must be a tree.
- Compare the computational complexity of a model network construction where the structure is generated by joining vertices, edges, or triangles.
- Construct higher-order network representation described in Section 11.3 for the network of Fig.4.1.

12

Probabilistic modeling with network representations

From a probabilistic perspective, the properties of a multivariate probability distribution that models a system under examination are determined by the network representing the interconnections between the variables in the system. Multivariate probability models can be directly constructed using this network structure that I shall call ‘*representative network*’. In some instances, this is the network of inferences, which is the structure representing the conditional dependencies or causalities between the variables. But, overall, this representative network is more general: it is an abstract object representing the relevant interactions in the system. It is also a modeling tool that helps to estimate the underlying multivariate probability. In this chapter, I will guide the reader through a few methodologies to construct multivariate probabilistic modeling based on network representations.

12.1 An information theoretic approach for network learning

The purpose of the network construction is to represent the structure of a multivariate probability and to provide guidance for the construction of models that best represent the data. The general problem of finding the network representation would require two ingredients: 1. have models that depend on a network representation \mathcal{M}_G ; 2. generate all possible networks and find the one that best describes the properties of the multivariate system of variables. This is a known NP-hard problem.

From an information theoretic perspective the problem consists in finding the multivariate probability density function with representation structure \mathcal{G} , $\tilde{f}(\mathbf{X}|\mathcal{G})$, that best describes the true underlying distribution $f(\mathbf{X})$ (which is unknown). I intentionally call \mathcal{G} the ‘representation structure’ instead of the more common term ‘inference structure’ because the second implies assumptions of conditional independence that are not necessary at this stage. Within this perspective, differently from graphical modeling (see Definition 12.1), the network does not necessarily coincide with the conditional independence structure. It is rather a representation of the multivariate probability structure. Ultimately, the goodness of the multivariate model represented by the network \mathcal{G} is tested from the gain function that typically, in this context, is the likelihood estimated on an out-of-sample dataset (i.e. not the train set, see Chapter 18).

In general, to quantify the distance between a model, $f(\mathbf{X}|\mathcal{G})$, and the true distribution, $f(\mathbf{X})$, one can use the Kullback-Leibler divergence (see Section 7.4)

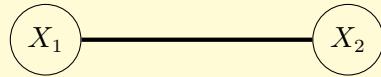
$$D_{\text{KL}}(f \parallel \tilde{f}) = \mathbb{E}(\log f(\mathbf{X})) - \mathbb{E}(\log \tilde{f}(\mathbf{X}|\mathcal{G})), \quad (12.1)$$

which must be minimized. The first term is independent of the model and therefore its value is irrelevant to the purpose of discovering the representation network. The second term, $-\mathbb{E}(\log \tilde{f}(\mathbf{X}|\mathcal{G}))$ (note the minus), instead depends on \mathcal{G} and must be minimized. This term is the estimate of the entropy of the multivariate system of variables \mathbf{X} by using of the model $\tilde{f}(\mathbf{X}|\mathcal{G})$:

$$H(\mathbf{X}|\mathcal{G}) = -\mathbb{E}(\log \tilde{f}(\mathbf{X}|\mathcal{G})); \quad (12.2)$$

it is the cross entropy (see Definition 7.4). Given that the true underlying distribution is unknown, the expectation value cannot be computed exactly, however, it can be estimated with arbitrary precision from sampling. Indeed, as we shall see in the next chapter, the sample mean converges to the expected value (this is the law of large numbers, Section 13.3). The sample estimation of $\mathbb{E}(\log \tilde{f}(\mathbf{X}|\mathcal{G}))$ approximates the expected value of the log-likelihood of the model $\tilde{f}(\mathbf{X}|\mathcal{G})$ (see Definition 14.1). Therefore, the construction of the representation network must aim to maximize the likelihood of the model.

Definition 12.1 (Graphical modeling). There is a graphical representation of conditional independence which can be extremely useful. In this representation, two variables directly connected with an edge are dependent:



$$P(x_1, x_2) \neq P(x_1)P(x_2) \text{ (} X_1, X_2 \text{ dependent).}$$

Two variables belonging to two disconnected components of the graph are independent:



$$P(x_1, x_4) = P(x_1)P(x_4) \text{ (} X_1, X_4 \text{ independent)}$$

$$P(x_2, x_3) = P(x_2)P(x_3) \text{ (} X_2, X_3 \text{ independent)}$$

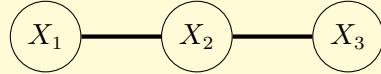
$$P(x_2, x_4) = P(x_2)P(x_4) \text{ (} X_2, X_4 \text{ independent)}$$

$$P(x_1, x_3) = P(x_1)P(x_3) \text{ (} X_1, X_3 \text{ independent)}$$

$$P(x_1, x_2) \neq P(x_1)P(x_2) \text{ (} X_1, X_2 \text{ dependent)}$$

$$P(x_3, x_4) \neq P(x_3)P(x_4) \text{ (} X_3, X_4 \text{ dependent).}$$

Two variables connected through a path but not directly with an edge are conditionally independent:



$$P(x_1, x_3|x_2) = P(x_1|x_2)P(x_3|x_2) \text{ (} X_1, X_3 \text{ conditionally independent);}$$

$$\text{but } P(x_1, x_3) \neq P(x_1)P(x_3) \text{ (} X_1, X_3 \text{ dependent);}$$

This way of associating conditional independence to a graph structure is called **graphical modeling**.

12.1.1 Pairwise interactions

Let me first address the network learning problem in terms of pairwise interactions. This does not mean simply two variables, but rather two groups of variables \mathbf{X}_1 and \mathbf{X}_2 . In the perspective of this book an edge between two vertices in the network would represent the direct inclusion in the model of some pairwise relation between the associated variables. If the association is between sets of variables, then edges are added between all couples between the two sets.

I have discussed in Chapter 8 that dependency between two sets of variables implies that a variable can be in part described in terms of the other, in other words, they share some information. This shared information can be accounted for by the mutual information $I(\mathbf{X}_1, \mathbf{X}_2)$. Indeed, this mutual information is quantifying how much the models gain by including explicitly the links between \mathbf{X}_1 and \mathbf{X}_2 in the representation. I shall discuss in this Chapter that this gain is exactly equal to $I(\mathbf{X}_1, \mathbf{X}_2)$ for graphical models, and it is approximated by this mutual information in general. Therefore, a meaningful network representation must connect sets of variables that share large information. Networks that aim to maximize such pairwise information can be constructed by means of one of the information filtering methods described in Section 11.2 by using the mutual information as a gain function.

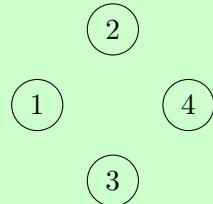
12.1.2 Higher order interactions

Accounting for pairwise interactions is not a trivial task, but accounting for higher-order interactions is harder. Let me recall the multivariate entropy expansion introduced in Section 8.9.2, Eq.8.84. One can associate each element of a network representation of a multivariate system of variables with its contribution towards the estimate of the system's entropy and therefore to the contribution to the model's likelihood. Namely, the first terms of the expansion are associated with vertices, which contribute with $H(X_i)$ and represent independent variables. The second contribution comes from the edges and it is (minus) the mutual information $-I(X_i, X_j)$ between couples of variables. These terms are negative and reduce the entropy associated to the first terms. The third terms are the triangles that contribute with $+I_3(X_i, X_j, X_k)$. If we restrict to clique-tree representations then, the next contribution comes from tetrahedra, namely the tetrahedron with vertices (c_1, c_2, c_3, c_4) , contributes with $-I_4(X_{c_1}, X_{c_2}, X_{c_3}, X_{c_4})$. In general, at all orders, k -cliques contribute with $(-1)^{k+1} I_k(X_{c_1}, \dots, X_{c_k})$.

The aim is to obtain a model with minimum cross entropy which in turn corresponds to the largest likelihood. By looking at each contribution separately, it appears that edges contribute with $-I(X_i, X_j)$ and always reduce cross en-

tropy (indeed $I(X_i, X_j) \geq 0$) while triangles contribute with $+I_3(X_i, X_j, X_k)$ and reduce cross entropy only if they are synergic ($I(X_i, X_j, X_k) < 0$). It must be however noticed that most of the contributions from these terms cancel each other and therefore this analysis can be misleading. Let me provide here below a few examples of network representations and their associated entropy approximators that can clarify these accountings.

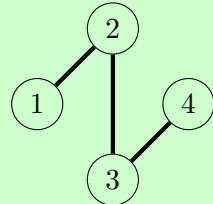
Example 12.1 (Multivariate entropy approximation for a given representation). The simplest representation has four isolated nodes



which contribute to the cross entropy as:

$$H(X_1, X_2, X_3, X_4 | \mathcal{G}) \simeq h_1 = H(X_1) + H(X_2) + H(X_3) + H(X_4). \quad (12.3)$$

By adding three edges (1, 2), (2, 3) and (3, 4), one obtains a connected line graph.



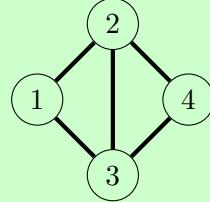
The contributions to the cross entropy of these new edges are respectively $-I(X_1, X_2)$, $-I(X_2, X_3)$ and $-I(X_3, X_4)$ which must be added to the previous contribution from the vertices, obtaining:

$$\begin{aligned} H(X_1, X_2, X_3, X_4 | \mathcal{G}) &\simeq \\ h_2 &= h_1 - I(X_1, X_2) - I(X_2, X_3) - I(X_3, X_4) = \\ &= H(X_1) + H(X_2) + H(X_3) + H(X_4) - \\ &\quad - H(X_1) - H(X_2) + H(X_1, X_2) - \\ &\quad - H(X_2) - H(X_3) + H(X_2, X_3) - \\ &\quad - H(X_3) - H(X_4) + H(X_3, X_4) = \\ &= H(X_1, X_2) + H(X_2, X_3) + H(X_3, X_4) - H(X_2) - H(X_3) \end{aligned} \quad (12.4)$$

Notice that, as for the Eq.8.84 expansion, in the previous expression most

terms cancel with each other leaving a few higher order terms and some terms associated with elements that are double counted (i.e. vertex 2 is counted in both in $H(X_1, X_2)$ and $H(X_2, X_3)$ and therefore it is discounted once, same for vertex 3).

If I add now two new edges, $(1, 3)$ and $(2, 4)$ I obtain two triangles joined by one edge

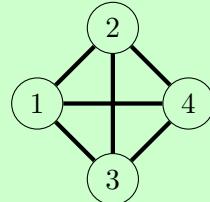


The edges contribute $-I(X_1, X_3)$, $-I(X_2, X_4)$ while the two newly formed triangles, $(1, 2, 3)$, $(2, 3, 4)$, contribute with $I_3(X_1, X_2, X_3)$ and $I_3(X_2, X_3, X_4)$ respectively. This results in the following approximator:

$$\begin{aligned}
 H(X_1, X_2, X_3, X_4 | \mathcal{G}) &\simeq \\
 h_3 = h_2 - I(X_1, X_3) - I(X_2, X_4) + I_3(X_1, X_2, X_3) + I_3(X_2, X_3, X_4) &= \\
 = H(X_1, X_2) + H(X_2, X_3) + H(X_3, X_4) - H(X_2) - H(X_3) - \\
 - H(X_1) - H(X_3) + H(X_1, X_3) - \\
 - H(X_2) - H(X_4) + H(X_2, X_4) + \\
 + H(X_1) + H(X_2) + H(X_3) - H(X_1, X_2) - H(X_2, X_3) - H(X_1, X_3) + \\
 + H(X_1, X_2, X_3) + \\
 + H(X_2) + H(X_3) + H(X_4) - H(X_2, X_3) - H(X_3, X_4) - H(X_2, X_4) + \\
 + H(X_2, X_3, X_4) &= H(X_1, X_2, X_3) + H(X_2, X_3, X_4) - H(X_2, X_3). \quad (12.5)
 \end{aligned}$$

Again one can notice that the lower order terms have been canceled and only the contribution from the two triangles and the (double counted) separating edge were left.

By adding an extra edge between $(1, 4)$ I then close the structure into a tetrahedron.



The added edge contributes with $-I(X_1, X_4)$, but one must notice that the structure has also gained two new triangles $(1, 2, 4)$ and $(1, 3, 4)$ and –

of course – the tetrahedron $(1, 2, 3, 4)$. These new elements must add their contribution to the previous approximator, resulting in:

$$\begin{aligned}
 H(X_1, X_2, X_3, X_4 | \mathcal{G}) &\simeq \\
 h_4 &= h_3 - I(X_1, X_4) + I_3(X_1, X_2, X_4) + I_3(X_1, X_3, X_4) - I_4(X_1, X_2, X_3, X_4) = \\
 &= H(X_1, X_2, X_3) + H(X_2, X_3, X_4) - H(X_2, X_3) - \\
 &- H(X_1) - H(X_4) + H(X_1, X_4) + \\
 &+ H(X_1) + H(X_2) + H(X_4) - H(X_1, X_2) - H(X_2, X_4) - H(X_1, X_4) + \\
 &+ H(X_1, X_2, X_4) + \\
 &+ H(X_1) + H(X_3) + H(X_4) - H(X_1, X_3) - H(X_3, X_4) - H(X_1, X_4) + \\
 &+ H(X_1, X_3, X_4) - H(X_1) - H(X_2) - H(X_3) - H(X_4) + \\
 &+ H(X_1, X_2) + H(X_1, X_3) + H(X_1, X_4) + H(X_2, X_3) + H(X_2, X_4) + \\
 &+ H(X_3, X_4) - H(X_1, X_2, X_3) - H(X_1, X_2, X_4) - H(X_1, X_3, X_4) - \\
 &- H(X_2, X_3, X_4) + H(X_1, X_2, X_3, X_4) = H(X_1, X_2, X_3, X_4). \tag{12.6}
 \end{aligned}$$

One can notice in these examples that the contributions are always the sum of the contribution from the cliques minus the contribution of the separators (that are double counted). These network representations are all indeed instances of clique-tree representations of graphical models for which the result is always in this form.

12.1.3 Higher order cross entropy contributions with clique-tree representations

The accounting of higher order contribution to the cross entropy by the network representation is simplified for clique-tree structures. Indeed, given a clique-tree representation with clique-set \mathcal{C} , each k -clique contributes with its cross entropy $H(X_{c_1}, \dots, X_{c_k})$. There are however several double counting associated with the elements that belong to more than one clique. These are the separators and their double contribution must be subtracted. Overall, for a given clique-tree representation \mathcal{G} with clique-set \mathcal{C} and separator-set \mathcal{S} the approximator of the system's entropy is

$$H(X_1, \dots, X_p | \mathcal{G}) \simeq \sum_{\mathbf{c} \in \mathcal{C}} H(\mathbf{X}_{\mathbf{c}}) - \sum_{\mathbf{s} \in \mathcal{S}} H(\mathbf{X}_{\mathbf{s}}). \tag{12.7}$$

The problem of learning high-dimensional models that maximize likelihood is reduced to a set of low-dimensional models over the cliques and separators substructures. This is a general result, which, as I shall discuss in the next session, is exact when the clique tree representation is the inference structure (graphical models) and it is an approximation in the general case.

12.2 Probability decomposition on a clique tree inference structure

The time complexity of exact inference is NP-hard and therefore not feasible from any finite set of observations [Birnbaum, 1962, Ernst, 2004]. In practice, one can tackle this problem either by first estimating the joint probability distribution and then extracting a set of the most relevant dependencies or by first inferring the most relevant dependencies and then estimating the joint probability distribution. Either way, the approach will lead to an approximate solution.

Let me first assume that the inference structure is known and it has the structure of a clique tree. This leads to important simplifications. Then I shall proceed the opposite way and discuss how a clique tree inference structure representation can be learned from data (see Sections 12.3 and 12.4) .

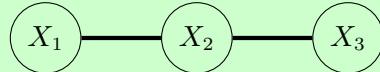
When the inference structure is known and it is chordal, then it is made of cliques attached through separators in a clique-tree structure. On such a structure Bayes' formula can be applied to the cliques and separators' conditional probabilities obtaining the following decomposition for the joint probability density function (see Lauritzen [1996]):

$$f(\mathbf{x}) = \frac{\prod_{c \in \mathcal{C}} f_c(\mathbf{x}_c)}{\prod_{s \in \mathcal{S}} f_s(\mathbf{x}_s)}. \quad (12.8)$$

Where \mathcal{C} is the set of cliques and \mathbf{X}_c is the set of variables associated with the vertices in $c \in \mathcal{C}$ and similarly \mathcal{S} is the set of separators and \mathbf{X}_s is the set of variables associated with the vertices in $s \in \mathcal{S}$. Separators with multiplicity k are counted $k - 1$ times.

Example 12.2 (Probability decomposition for two cliques and one separator). Let me provide an example for the simplest case: two cliques c_1 and c_2 with one separator s between them.

First, let me do this for three variables: X_1 , X_2 , and X_3 .

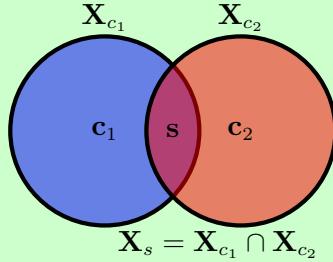


The variables associated with the cliques are $\mathbf{X}_{c_1} = (X_1, X_2)$ and $\mathbf{X}_{c_2} = (X_2, X_3)$, and the variable $\mathbf{X}_s = X_2$ is the separator.

Applying Bayes' formula for the joint probability density function, one has $f(x_1, x_2, x_3) = f(x_1, x_3|x_2)f(x_2)$, but $X_1|X_2$ is independent from $X_3|X_2$ and therefore $f(x_1, x_3|x_2) = f(x_1|x_2)f(x_3|x_2)$ and consequently $f(x_1, x_2, x_3) = f(x_1|x_2)f(x_3|x_2)f(x_2)$. One can apply again Bayes' formula to write the conditional probabilities in terms of the joint: $f(x_1|x_2) = f(x_1, x_2)/f(x_2)$ and $f(x_3|x_2) = f(x_3, x_2)/f(x_2)$. Obtaining therefore overall the following expression:

$$f(x_1, x_2, x_3) = \frac{f(x_1, x_2)f(x_3, x_2)}{f(x_2)}. \quad (12.9)$$

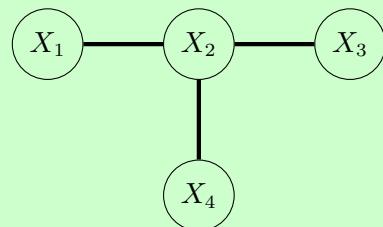
I have therefore decomposed the joint probability for the tree variables in terms of the probabilities of the cliques and separators in the inference graph structure. This procedure is completely general and can be extended to any set of variables $\mathbf{X} = \mathbf{X}_{c_1} \cup \mathbf{X}_{c_2}$ and $\mathbf{X}_s = \mathbf{X}_{c_1} \cap \mathbf{X}_{c_2}$ in an analogous structure of cliques and separators:



This inference structure is indicating that the two subsets of variables \mathbf{X}_{c_1} and \mathbf{X}_{c_2} are conditionally independent given the subset of variables \mathbf{X}_s , the Bayes' formula tells us that $f(\mathbf{x}) = f_{c_1}(\mathbf{x}_{c_1}|\mathbf{x}_s)f_{c_2}(\mathbf{x}_{c_2}|\mathbf{x}_s)f_s(\mathbf{x}_s)$ or, by rewriting the conditional probabilities as $f_{c_1}(\mathbf{x}_{c_1}|\mathbf{x}_s) = f_{c_1}(\mathbf{x}_{c_1})/f_s(\mathbf{x}_s)$ and $f_{c_2}(\mathbf{x}_{c_2}|\mathbf{x}_s) = f_{c_2}(\mathbf{x}_{c_2})/f_s(\mathbf{x}_s)$ one has

$$f(\mathbf{x}) = \frac{f_{c_1}(\mathbf{x}_{c_1})f_{c_2}(\mathbf{x}_{c_2})}{f_s(\mathbf{x}_s)} . \quad (12.10)$$

Extension and generalization to any clique-forest inference graph structure is straightforward with the only consideration that separators with multiplicity larger than two must be accounted for more than once. For instance the structure:



has

$$f(x_1, x_2, x_3, x_4) = \frac{f(x_1, x_2)f(x_3, x_2)f(x_2, x_4)}{f(x_2)^2} . \quad (12.11)$$

When the number of variables is large and the size of the cliques is small $p \gg \max(|c| \in \mathcal{C})$ (small treewidth) Eq.12.8 is an enormous reduction in the dimensionality of the joint probability distribution problem. This simplifies both the models and their calibration. This probability decomposition in Eq.12.8 provides a link between local probabilistic properties (the probabilities of variables in the cliques) and the global properties of the whole system (the joint proba-

bility of all variables) The problem passes from the whole system level to the local levels of each clique, for this reason, it was named LoGo (Local-Global) by Barfuss et al. [2016].

12.2.1 Multivariate normal case

For the case when the probability density function is a multivariate normal a straightforward consequence of the decomposition in Eq.12.8 is that the entries i, j of the inverse covariance matrix can be written as:

$$(\boldsymbol{\Sigma}^{-1})_{i,j} = (\mathbf{J})_{i,j} = \sum_{c \in \mathcal{C}} (\boldsymbol{\Sigma}_c^{-1})_{i,j} - \sum_{s \in \mathcal{S}} (\boldsymbol{\Sigma}_s^{-1})_{i,j}, \quad (12.12)$$

when they belong to a clique ($(i, j) \in \mathcal{C}$), otherwise $(\mathbf{J})_{i,j} = 0$ ($(i, j) \notin \mathcal{C}$). This implies that for a multivariate system of variables with a clique tree inference structure the inverse covariance is sparse. Sparse inverse covariances are valuable tools for modeling because they involve a smaller number of parameters to be estimated. In other words, sparsity reduces the dimensionality of the problem. In the case when the maximum clique size is a number independent from the total number of variables (i.e. for the MST it is 3 and for the TMFG it is 4) then the total number of the model's parameters passes from $\mathcal{O}(p^2)$ to $\mathcal{O}(p)$. From the expression of the multivariate normal decomposed with Eq.12.8, it is straightforward to demonstrate that the determinants decompose as well:

$$|\mathbf{J}| = \frac{\prod_{c \in \mathcal{C}} |\mathbf{J}_c|}{\prod_{s \in \mathcal{S}} |\mathbf{J}_s|} \text{ or } |\boldsymbol{\Sigma}| = \frac{\prod_{c \in \mathcal{C}} |\boldsymbol{\Sigma}_c|}{\prod_{s \in \mathcal{S}} |\boldsymbol{\Sigma}_s|}. \quad (12.13)$$

12.3 Learning clique tree inference structures

If the inference network is a clique tree, then the decomposition of the probability $\tilde{f}(\mathbf{X}|\mathcal{G})$ (Eq.12.8) provides an excellent tool to learn a good structure for \mathcal{G} . Indeed, through that decomposition, the Shannon cross entropy $H(\mathbf{X}|\mathcal{G})$ can be written as:

$$H(\mathbf{X}|\mathcal{G}) = \sum_{c \in \mathcal{C}} H(\mathbf{X}_c) - \sum_{s \in \mathcal{S}} H(\mathbf{X}_s). \quad (12.14)$$

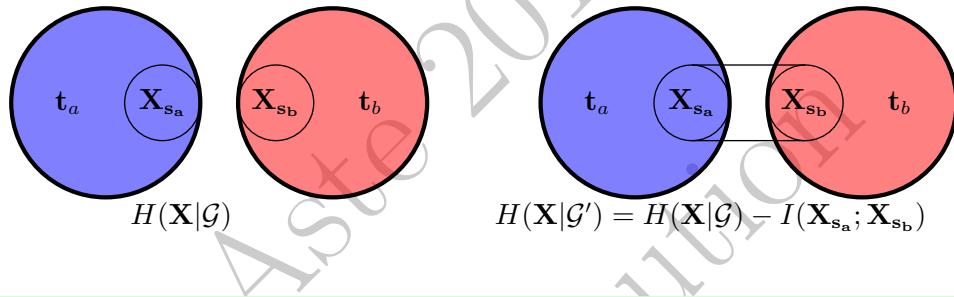
It is therefore a simple sum of local terms associated with cliques and separators. In any construction of the inference network made by combining together parts, the change in the cross entropy will be only associated with the parts of the inference network that join together. By joining together parts in the inference network one is assuming direct dependency between two sets of variables. This operation always lower cross entropy and the task of learning the inference structure is to find the connections that lower cross entropy most significantly.

When two sub-cliques \mathbf{s}_a and \mathbf{s}_b of two disjoined clique trees are joined together to form a larger clique tree, they generate a new clique $\mathbf{s}_a \cup \mathbf{s}_b$ and two new separators \mathbf{s}_a and \mathbf{s}_b . If one is constructing an inference model where the network

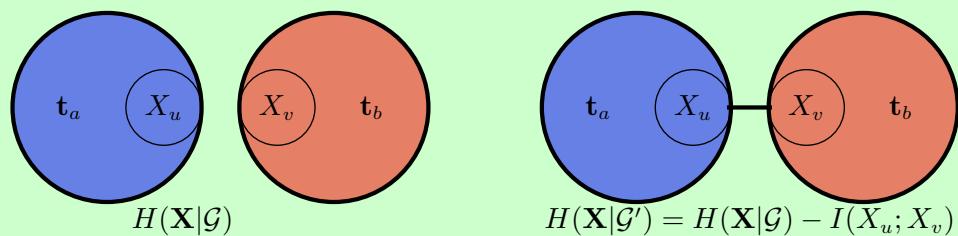
represents the conditional independence structure, then from Eq.12.14 the change in cross-entropy is

$$\Delta H = H(\mathbf{X}_{\mathbf{s}_a}, \mathbf{X}_{\mathbf{s}_b}) - H(\mathbf{X}_{\mathbf{s}_a}) - H(\mathbf{X}_{\mathbf{s}_b}) = -I(\mathbf{X}_{\mathbf{s}_a}; \mathbf{X}_{\mathbf{s}_b}), \quad (12.15)$$

this is minus the mutual information between the joined sets of variables. This is intuitive: the reduction in cross-entropy associated with the explicit inclusion into the model of the dependency between variables is the mutual information (see Section 8.9). In a schematic drawing, this is:



Example 12.3 (Cross entropy changes for a MST inference network construction). Let me consider the construction of the inference network \mathcal{G} as a spanning tree by using Kurscal's Algorithm 11.2, which joins at each step two separated trees \mathbf{t}_a and \mathbf{t}_b with one edge (u, v) between two vertices of the trees ($u \in \mathbf{t}_a$ and $v \in \mathbf{t}_b$). Here I address the problem for the maximum spanning tree, because I aim to maximize the mutual information captured by the network representation. Clearly the problem is complementary to the minimum spanning tree for which the Kurscal's algorithm was originally designed.



Joining two trees \mathbf{t}_a and \mathbf{t}_b by an edge (u, v) adds the clique (u, v) , it also might add separators depending on the properties of the two trees. Let me list all cases:

1. if both u and v are isolated vertices, then no separators are added;
2. if u is part of the \mathbf{t}_a tree and v is instead isolated, then the separator v is added (vice versa for the case u isolated and v attached to a tree);
3. if both u and v are part of two disjoined trees ($u \in \mathbf{t}_a$ and $v \in \mathbf{t}_b$), then two separators u and v are added.

Despite these different mechanisms, in all these three scenarios the cross entropy changes by the amount of mutual information $I(X_u; X_v) = H(X_u) + H(X_v) - H(X_u, X_v)$. Indeed, for these three scenarios:

1. the cross entropy $H(\mathbf{X}|\mathcal{G}) = H(X_u) + H(X_v)$ becomes $H(\mathbf{X}|\mathcal{G}) = H(X_u, X_v)$, therefore the difference is $\Delta H = I(X_u; X_v)$;
2. the cross entropy $H(\mathbf{X}|\mathcal{G}) = H(X_{t_a}) + H(X_v)$ becomes $H(\mathbf{X}|\mathcal{G}) = H(X_{t_a}) + H(X_u, X_v) - H(X_u)$, therefore the difference is $\Delta H = I(X_u; X_v)$ (and the same when vertex u is part of a tree instead);
3. the cross entropy $H(\mathbf{X}|\mathcal{G}) = H(X_{t_a}) + H(X_{t_b})$ becomes $H(\mathbf{X}|\mathcal{G}) = H(X_{t_a}) + H(X_{t_b}) + H(X_u, X_v) - H(X_u) - H(X_v)$, therefore again the difference is $\Delta H = I(X_u; X_v)$.

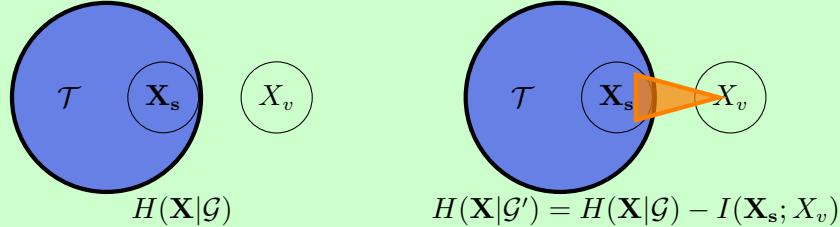
By constraining the inference structure to be a spanning tree, the dimensionality of the problem is reduced to a set of two-dimensional maximizations and it can be solved exactly. This is indeed the solution proposed by Chow and Liu [1968] to estimate p -dimensional (discrete) probability distributions with a product of second-order distributions.

In the special case of modeling with normal distributions, then $I(X_u; X_v) = -\frac{1}{2} \log(1 - \rho_{u,v}^2)$ (see Example 8.5) and the problem becomes the one of maximization of the square correlation (i.e. at each step, join the two vertices with the largest square correlation).

Remark 12.1. In the linear dependency case the maximization of the mutual information between two variables corresponds to the maximization of the square correlation coefficient. This links this topic with the vast literature on correlation networks (see Example 17.1). However, the present approach is much broader because: first, I consider the general non-linear case; and second, I do not limit the problem to the relation between two variables but rather to the dependency between sets of variables bringing the topics to the emerging field of the so-called “higher-order interactions” in ecological [Mayfield and Stouffer, 2017] and physical complex systems [Battiston et al., 2021].

Example 12.4 (Cross entropy change for an MFCF inference network construction). Let me now consider, instead of the spanning tree (where vertices, 0-dimensional simplexes, are joined together through an edge, 1-dimensional simplex), a higher order graph where higher dimensional simplexes are joined through higher dimensional separators. Specifically, let me follow the construction of the MFCF (see Algorithm 11.5) where a n -simplex is joined to a disconnected vertex at every step. Schematically, the

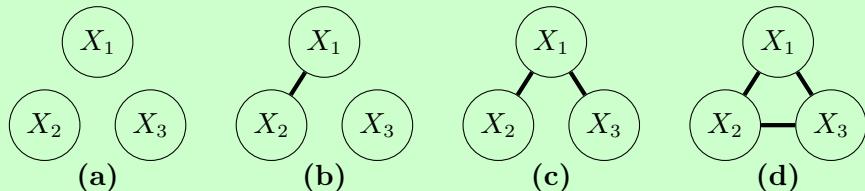
process is similar to the construction of the MST described in the previous example, although in this case, the analogy is with Prim's algorithm (see Algorithm 11.1) where, at each step, an isolated vertex is added to the tree, which in this case is the clique tree \mathcal{T} . In a picture, this is



Where the orange triangle indicates that a set of edges forming a clique (not just one edge) is inserted to connect the sub-clique s with vertex v . In this case the change in cross entropy for the whole system is due to the addition to the clique tree of a new clique $c' = s \cup v$, the addition of a new separator s and the elimination of an isolated vertex that becomes part of the clique three. This accounts for: $\Delta H = H(\mathbf{X}_s \cup X_v) - H(\mathbf{X}_s) - H(X_v)$ which is minus the mutual information $I(\mathbf{X}_s; X_v)$. This is very similar to the result in the previous example. However, in this case, \mathbf{X}_s is a set of $|s|$ variables and not only one variable. The mutual information $I(\mathbf{X}_s; X_v)$ is a multidimensional object, it accounts for the contribution from the combination of a vertex, v , with a simplex s and, in general, it is not equal to the sum of the contribution of the added edges.

For the special case of multivariate normal models, this contribution is given by Eq.8.79.

Example 12.5 (Higher order contributions). Let me consider a very simple case of three variables associated with the following network representations:



They respectively have:

- a** $H(X_1, X_2, X_3 | \mathcal{G}_a) = H(X_1) + H(X_2) + H(X_3);$
- b** $H(X_1, X_2, X_3 | \mathcal{G}_b) = H(X_1, X_2) + H(X_3);$
- c** $H(X_1, X_2, X_3 | \mathcal{G}_c) = H(X_1, X_2) + H(X_1, X_3) - H(X_1);$
- d** $H(X_1, X_2, X_3 | \mathcal{G}_d) = H(X_1, X_2, X_3).$

Let me consider a simple network construction where the gain function is the pairwise mutual information between the couples of vertices that are joined at each step. This is:

$$\begin{aligned}\mathbf{a} \rightarrow \mathbf{b} \quad G_{a \rightarrow b} &= I(X_1; X_2); \\ \mathbf{b} \rightarrow \mathbf{c} \quad G_{b \rightarrow c} &= I(X_1; X_3); \\ \mathbf{c} \rightarrow \mathbf{d} \quad G_{c \rightarrow d} &= I(X_2; X_3).\end{aligned}$$

One can notice that while

$$H(X_1, X_2, X_3 | \mathcal{G}_a) - G_{a \rightarrow b} = H(X_1, X_2, X_3 | \mathcal{G}_b) \quad (12.16)$$

and

$$H(X_1, X_2, X_3 | \mathcal{G}_b) - G_{b \rightarrow c} = H(X_1, X_2, X_3 | \mathcal{G}_c). \quad (12.17)$$

Conversely,

$$H(X_1, X_2, X_3 | \mathcal{G}_c) - G_{c \rightarrow d} \neq H(X_1, X_2, X_3 | \mathcal{G}_d). \quad (12.18)$$

Indeed, the network (d) has a higher dimensional structure (the triangle) which is not accounted for in the pairwise gain function. Not incidentally, the missing term is $I_3(X_1, X_2, X_3)$ (see Eq.8.84) which is the higher order mutual information term associated with the three variables.

I have just demonstrated in the previous Example 12.4 that the clique expansion move (attachment of an isolated node to an existing clique of sub-clique in the network) accounts correctly for the cross entropy reduction. In this case, the representation (d) must be constructed directly from (b) by joining the clique (X_1, X_2) with the isolated vertex X_3 . This is associated with gain

$$\mathbf{b} \rightarrow \mathbf{d} \quad G_{b \rightarrow d} = I((X_1, X_2); X_3) = H(X_1, X_2) + H(X_3) - H(X_1, X_2, X_3).$$

It can be directly verified that, with this construction indeed

$$H(X_1, X_2, X_3 | \mathcal{G}_b) - G_{b \rightarrow d} = H(X_1, X_2, X_3 | \mathcal{G}_d). \quad (12.19)$$

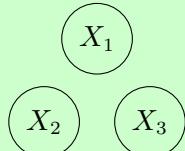
Example 12.6 (Quantification of network representations contributions). Let me use the previous example to quantify the network contributions in a simple example for a multivariate normal model with the following covariance matrix

$$\Sigma_{\mathbf{Z}\mathbf{Z}} = \begin{pmatrix} 1.0 & 0.7 & 0.9 \\ 0.7 & 3.0 & 0.8 \\ 0.9 & 0.8 & 1.0 \end{pmatrix}. \quad (12.20)$$

which is the same as in Example 6.5. By computing directly the entropy

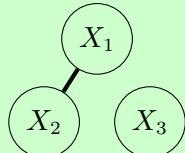
of each model for the multivariate normal case (see Eq.7.21) I obtain the following.

Model (a)



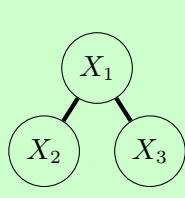
$$\begin{aligned} H(X_1, X_2, X_3 | \mathcal{G}_a) &= H(X_1) + H(X_2) + H(X_3) = \\ &= \frac{3}{2} + \frac{3}{2} \log(2\pi) + \log \sigma_1 + \log \sigma_2 + \log \sigma_3 = \\ &= \frac{3}{2} + \frac{3}{2} \log(2\pi) + \log 1.0 + \log \sqrt{3.0} + \log 1.0 = 4.806 \end{aligned}$$

Model (b)



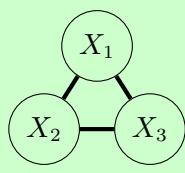
$$\begin{aligned} H(X_1, X_2, X_3 | \mathcal{G}_b) &= H(X_1, X_2) + H(X_3) \\ &= \frac{3}{2} + \frac{3}{2} \log(2\pi) + \frac{1}{2} \log |\Sigma_{12}| + \log \sigma_3 = \\ &= \frac{3}{2} + \frac{3}{2} \log(2\pi) + \frac{1}{2} \log \begin{vmatrix} 1.0 & 0.7 \\ 0.7 & 3.0 \end{vmatrix} + \log 1.0 = 4.717 \end{aligned}$$

Model (c)



$$\begin{aligned} H(X_1, X_2, X_3 | \mathcal{G}_c) &= H(X_1, X_2) + H(X_1, X_3) - H(X_1) \\ &= \frac{3}{2} + \frac{3}{2} \log(2\pi) + \frac{1}{2} \log |\Sigma_{12}| + \frac{1}{2} \log |\Sigma_{13}| - \log \sigma_1 = \\ &= \frac{3}{2} + \frac{3}{2} \log(2\pi) + \frac{1}{2} \log \begin{vmatrix} 1.0 & 0.7 \\ 0.7 & 3.0 \end{vmatrix} + \\ &\quad + \frac{1}{2} \log \begin{vmatrix} 1.0 & 0.9 \\ 0.9 & 1.0 \end{vmatrix} - \log 1.0 = 3.886 \end{aligned}$$

Model (d)



$$\begin{aligned} H(X_1, X_2, X_3 | \mathcal{G}_d) &= H(X_1, X_2, X_3) \\ &= \frac{3}{2} + \frac{3}{2} \log(2\pi) + \frac{1}{2} \log |\Sigma_{123}| = \\ &= \frac{3}{2} + \frac{3}{2} \log(2\pi) + \frac{1}{2} \log \begin{vmatrix} 1.0 & 0.7 & 0.9 \\ 0.7 & 3.0 & 0.8 \\ 0.9 & 0.8 & 1.0 \end{vmatrix} = 3.855 \end{aligned}$$

This example shows that more connected network representations yield probabilistic models with lower entropy and therefore larger expected likelihoods.

An equivalent approach, with identical results, consists in using the sparse inverse covariance formula Eq.12.12 and

$$H(\mathbf{X} | \mathcal{G}) = \frac{p}{2} + \frac{p}{2} \log(2\pi) - \frac{1}{2} \log |\mathbf{J}_{\mathcal{G}}| \quad (12.21)$$

which is direct consequence of Eq.7.21 for the entropy of p multivariate normal variables with covariance $\Sigma = \mathbf{J}_{\mathcal{G}}^{-1}$. Specifically one has:

Model	$\mathbf{J}_{\mathcal{G}}$	$ \mathbf{J}_{\mathcal{G}} $	$H(\mathbf{X} \mathcal{G})$
(a)	$\begin{pmatrix} 1.000 & 0 & 0 \\ 0 & 0.333 & 0 \\ 0 & 0 & 1.000 \end{pmatrix}$	0.333	4.806
(b)	$\begin{pmatrix} 1.195 & -0.279 & 0 \\ -0.279 & 0.398 & 0 \\ 0 & 0 & 1.000 \end{pmatrix}$	0.398	4.717
(c)	$\begin{pmatrix} 5.458 & -0.279 & -4.737 \\ -0.279 & 0.398 & 0 \\ -4.737 & 0 & 5.263 \end{pmatrix}$	2.097	3.887
(d)	$\begin{pmatrix} 5.268 & 0.045 & -4.777 \\ 0.045 & 0.424 & -0.379 \\ -4.777 & -0.379 & 5.603 \end{pmatrix}$	2.232	3.855

12.4 Learning network representations for multivariate modeling

Graphical modeling is a powerful instrument that allows decomposing of high-dimensional, complex, multivariate models into conditionally independent low-dimensional interconnected parts. However, often, the true joint probability structure is not made of meaningful conditionally independent terms and, while the simplification of the model provided by a sparse representation can be extremely useful, the decomposition into conditionally independent parts is not necessary or convenient. For modeling purposes, one might desire the simplification provided by a network representation of a multivariate model without the need for the decomposition into conditionally independent parts. In this case, the network \mathcal{G} becomes a representation of the model structure in terms of lower dimensional elements. In other words, the network is a sparsification of the model that reduces dimensionality but does not necessarily represent conditional independence. By dropping the conditional independence of graphical models one greatly expands the applicability of network representations to multivariate probabilistic modeling. However, the contribution of the network to cross entropy and, ultimately, the model-likelihood, becomes less direct and harder to compute.

Example 12.7 (Sparse representation). An example of a sparse network representation of a multivariate model, that does not impose or imply conditional independence, can be obtained by using the sparse inverse covariance from Eq.12.12 as inverse scale matrix, $\Omega^{-1} = \mathbf{J}$, in a multivariate joint probability model from the elliptical family (see Section 6.5). Indeed, with an exception for the multivariate normal case, zero entries in the inverse scale matrix do not correspond to conditional independence. Nonetheless, through \mathbf{J} , the model depends on the network structure that determines

the set of non-zero parameters. This is a form of zero-norm regularization and it is discussed further in Section 15.9.4.

The learning of such network representation should follow the same general principle described for the inference structure learning; the model representation, \mathcal{G}^* , must be learned in such a way that uncertainty (i.e. cross entropy) is minimized (and consequently model-likelihood is maximized):

$$\mathcal{G}^* = \operatorname{argmin}_{\mathcal{G}} (H(\mathbf{X}|\mathcal{G})). \quad (12.22)$$

For this general case, when the network is a representation and not an inference structure, it is rather intuitive that a model would benefit most by joining together in the network representation subsets of variables that share the largest information. However, the gain in likelihood ultimately depends on the application of the network representation in model construction. Therefore, ultimately, the gain achieved by modifying the network from \mathcal{G} to \mathcal{G}' by joining subsets of variables is

$$\Delta H = H(\mathbf{X}|\mathcal{G}') - H(\mathbf{X}|\mathcal{G}) \quad (12.23)$$

and it might be not reducible to local terms. Still, this gain can be estimated from the model log-likelihood (see Eq.12.2 and discussion there) and used as guidance for the network construction algorithm. I shall show in Chapter 18 that such a log-likelihood difference is a standard tool for model comparison (see Section 18.6.1).

A schematic, pictorial, representation of the procedure that constructs network representations for multivariate modeling is reported in Fig.12.1.

Remark 12.2. An alternative network representation of conditional independence is the **Bayesian Networks** which represent, with directed acyclic graphs, Markov chains of conditional probabilities. In a nutshell, these networks are representing the conditional probability relation $p(x_1, x_2) = p(x_2|x_1)p(x_1)$; which is a way to write the Bayes' formula. Such a relation is represented in the Bayesian network with a directed edge $X_1 \rightarrow X_2$. The vertex X_1 takes the name of ‘parent’, while X_2 is named ‘child’. When X_1 has his own parent, the graph $X_0 \rightarrow X_1 \rightarrow X_2$ represents $p(x_0, x_1, x_2) = p(x_2|x_1)p(x_1|x_0)p(x_0)$ which is a Markov chain where one assumed conditional independence between $X_0|X_1$ and $X_2|X_1$ (which is – indeed – the Markov property). In this case, X_0 is named ‘ancestor’. The generalization of this construction to a complexity of parents ancestors and children in an acyclic network representation is the essence of Bayesian networks, where children are dependent on their parents but conditionally independent from their ancestors. Note that the direction $X_1 \rightarrow X_2$ does not imply that X_1 is conditionally independent of X_2 . On the contrary, two dependent variables are always dependent in both directions

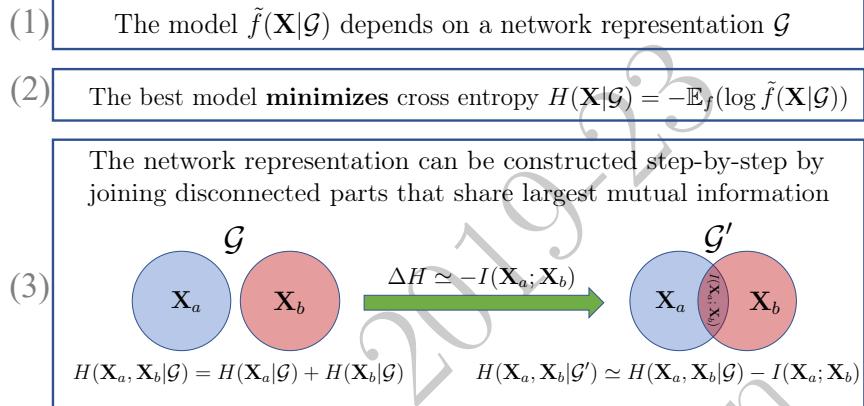


Figure 12.1 A schematic representation of the network construction procedure. 1) Consider a model $\tilde{f}(\mathbf{X}|\mathcal{G})$ which depends on a network representation \mathcal{G} . 2) The best model is the one that **minimizes** the cross-entropy $H(\mathbf{X}|\mathcal{G}) = -\mathbb{E}(\log \tilde{f}(\mathbf{X}|\mathcal{G}))$ (i.e. the one that **maximizes** the expected value of the **log-likelihood** of model $\tilde{f}(\mathbf{X}|\mathcal{G})$). 3) The network representation of the model can be constructed step-by-step joining disconnected parts that share the largest mutual information, decreasing at each step the model cross-entropy of an amount equal to or comparable with the mutual information $\Delta H \approx -I(\mathbf{X}_a; \mathbf{X}_b)$.

(i.e. $p(x_2|x_1) \neq p(x_2)$ implies $p(x_1|x_2) \neq p(x_1)$ and $p(x_1, x_2) \neq p(x_1)p(x_2)$). Therefore parents are conditionally dependent on their own children as well. I will not address any further this topic in this book. Interested, readers can deepen the investigation starting from Lauritzen [1996] (specifically, Chapter 3, section 3.2.2), and Pearl et al. [2000].

12.5 Data-driven modeling tutorial

<https://github.com/FinancialComputingUCL/DataDrivenModeling/Ch12>

The tutorial for this Chapter covers various topics on probabilistic modeling with network representations, including: the construction of the sparse inverse covariance matrix; and the comparison between the likelihoods of various probabilistic models associated with different network representations (Example 12.6).

Exercises

- Demonstrate that the covariance defined by the decomposition in Eq.12.12 is positively definite if the covariances of the cliques and separators are positively definite.
- Discuss why a denser model has always a larger likelihood than a sparser one.
- Compare the cross entropy expression associated with a model with three isolated vertices with an inference model where the three vertices form a linear chain and a model where they form a triangular clique.
- Discuss the application of the probability decomposition in Eq.12.8 to a model construction that uses bidimensional histograms.

© Tomaso Aste 2019-23
not for distribution
July 25, 2023

Part III

Model construction from data

© Tomaso Aste 2019₂₃
not for distribution
July 25, 2023

© Tomaso Aste 2019-23
not for distribution
July 25, 2023

13

Nonparametric estimation of univariate probabilities from data

In this chapter and in the next chapter I will address a fundamental issue that has also a practical relevance: given a set of observations $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, \dots, \hat{x}_q)$ of a random variable X how one can estimate the underlying probability distribution?¹

When it comes to estimating a probability distribution there are two possible major approaches: parametric and nonparametric.

In the nonparametric approach, one does not assume a priori the functional form of the underlying probability distribution. I shall show in Sections 13.5 to 13.8 that there are some choices to make about methodology and there actually are some parameters to choose (called hyperparameters). However, in its essence, the nonparametric approach builds the distribution directly from the data. From a data-driven modeling perspective, this is conceptually what one should aim to achieve: obtain a model (the probability distribution) using only the data themselves. The weakness of this approach is that normally a large number of observations is required to obtain good estimates and, instead, in many practical cases only a limited number of observations are available.

The problem is greatly simplified if one knows a priori the kind of distribution of the population. This information is in general not known, however some natural phenomena and some artificial systems follow known probability distributions and this has been theoretically or empirically established. If one knows – or assumes – that the data are drawn from a given type of probability distribution, then one has only to estimate from data the parameters of such distribution and not its entire functional form. This is the so-called parametric approach that I shall illustrate in the next Chapter. Let me note that assuming the ‘wrong’ distribution can result in extremely bad models with a very poor estimation of the likelihood of some events. This is particularly critical when distributions are ‘fat-tailed’ and there are sizeable probabilities of extreme events which are hard to estimate.

¹ Notice that I use the ‘hat’ symbol, \hat{x} , to indicate the observed, measured, value of variable X . In the statistical literature this is often referred as the ‘outcome’ of random variable X when a particular event or trial occurs. I shall use the hat symbol also for the sample measures (e.g. sample mean, Eq.13.1).

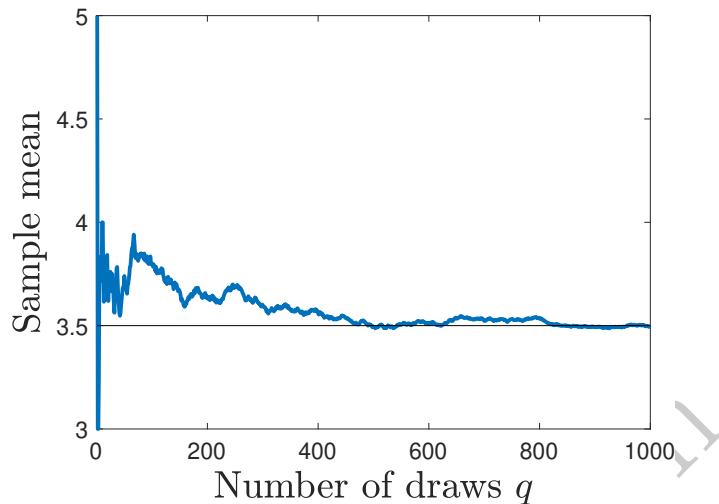


Figure 13.1 Example of sample mean convergence towards the mean for the random draw of a dice, with sample means computed over an increasing number of observations from $q = 1$ to $q = 1,000$.

13.1 The sample mean

For both parametric and nonparametric approaches, to estimate the properties of a random variable, one must start from the simplest ones, such as the moments of the distribution.

If one has q independent observations $(\hat{x}_1, \hat{x}_2, \dots, \hat{x}_q)$ of the random variable X , the sample estimate of the first moment (sample mean) is:

$$\hat{\mu} = \frac{1}{q}(\hat{x}_1 + \hat{x}_2 + \dots + \hat{x}_q) = \frac{1}{q} \sum_{i=1}^q \hat{x}_i. \quad (13.1)$$

This is a very simple and fundamental formula that can be used to estimate expected values from observations. We shall see in the next few sections how such a sample mean converges towards the population mean and how this estimate can be applied to the estimation of all expected values.

Example 13.1 (Convergence of the sample mean towards the expected value). Consider a six-faced dice that is thrown repeatedly. The draws are random with each face appearing with equal probability. For instance, the first ten draws might be the set of observations: $\hat{x} = (6, 1, 5, 5, 6, 1, 3, 2, 5, 3)$. By using Eq.13.1, the sample mean for this set of observations is $\hat{\mu} = 3.7$ which is a little bit higher than the population mean which is $\mu = 3.5$. If one repeats throwing the dice and adding observations to the estimate of the sample mean, eventually it will converge towards the population mean value $\hat{\mu} \xrightarrow{q \rightarrow \infty} 3.5$. This is shown in Figure13.1 where the results for the

sample mean, computed up to 1,000 observations, are reported. One can see that the convergence towards the mean value of 3.5 is clear. However, one can also notice that, even after 1,000 observations, there are still some differences. I shall indeed show, in Section 13.3, that convergence is in $1/\sqrt{q}$ and therefore 1,000 samples have an expected precision of a few percent.

13.2 Sample moments

All moments of the distribution can be estimated with the same procedure as the sample mean (Definition 13.1). This is actually more general, indeed, the expected value of any function $g(X)$ of the variable X , if finite, can be estimated from observational data, by computing the sample mean of the function:

$$\mathbb{E}(g(x)) \simeq \frac{1}{q} \sum_{i=1}^q g(\hat{x}_i), \quad (13.2)$$

if the function $g(X)$ is defined over the points $(\hat{x}_1, \dots, \hat{x}_q)$.

This is a very general and simple estimation tool that can be used to estimate all the moments.

Definition 13.1 (Sample moments). The **sample moments** are:

$$\hat{m}_k = \frac{1}{q} \sum_{i=1}^q \hat{x}_i^k. \quad (13.3)$$

Definition 13.2 (Sample central moments). The **sample central moments** are:

$$\hat{\mu}_k = \frac{1}{q} \sum_{i=1}^q (\hat{x}_i - \hat{m}_1)^k \quad (\text{for } k > 1). \quad (13.4)$$

Remark 13.1. Note that I use conventional notation for the sample mean, which is $\hat{\mu}$, and then $\hat{\mu}_k$ for the sample central moments, but only for $k > 1$.

For instance, the sample variance is the sample central second moment and it explicitly reads

$$\hat{\mu}_2 = \hat{\sigma}^2 = \frac{1}{q} \sum_{i=1}^q (\hat{x}_i - \hat{m}_1)^2 = \frac{1}{q} \sum_{i=1}^q (\hat{x}_i - \hat{\mu})^2. \quad (13.5)$$

A question that naturally arises is whether these sample means do estimate well the true expectation values and, more precisely, one wants to know, first,

if they converge asymptotically (for $q \rightarrow \infty$) to the population's expected value and, second, which is the rate of convergence with q .

13.3 The law of large numbers

The law of large numbers tells us that, when the number of observations increases and under some conditions, the sample mean of a variable eventually converges toward the expected value of the population.

$$\hat{\mu} = \frac{1}{q} \sum_{i=1}^q \hat{x}_i \xrightarrow[q \rightarrow \infty]{} \mu. \quad (13.6)$$

There are several formal ways to prove the convergence of the sample mean towards the population mean. This is a very important fact because it gives a precise meaning to the relevance of observations and provides instruments to evaluate how well one can estimate the statistical properties of a variable from a set of observations. Indeed, this is the law that links observations with the probability distribution of the population.

13.3.1 Mean-square law of large numbers

Definition 13.3 (Mean square convergence). A sequence of random variables Y_1, \dots, Y_q, \dots converges in the **mean square** sense to a random variable Y if

$$\lim_{q \rightarrow \infty} \mathbb{E}(|Y_q - Y|^2) = 0 , \quad (13.7)$$

and then one writes $Y \stackrel{m.s.}{=} \lim_{q \rightarrow \infty} Y_q$.

The law of convergence of the sample mean to the population mean must be established in general for all possible sets of observations, not just for a given set of observations $\hat{x}_1, \hat{x}_2, \dots$. Therefore, I consider each of the observations, $\hat{x}_1, \hat{x}_2, \dots$ as drawn from a set of independent and identically distributed random variables X_i with $i = 1, 2, \dots$. The law of large numbers tells us that the mean over such a set of variables

$$\bar{X}_q = \frac{1}{q} \sum_{i=1}^q X_i, \quad (13.8)$$

converges in the mean square sense to the population mean

$$\lim_{q \rightarrow \infty} \bar{X}_q \stackrel{m.s.}{=} \mu. \quad (13.9)$$

Notice that, the mean \bar{X}_q is a random variable itself and any draw of \bar{X}_q is an observed sample mean $\hat{\mu}$. Convergence of the sample mean towards the population mean is obtained when the distribution of \bar{X}_q is narrowly peaked around μ . If

the variance of such distribution shrinks to zero, then convergence is in the mean square sense.

This convergence can be derived with a few simple passages. Indeed, one has:

$$\begin{aligned}
 \mathbb{E}((\bar{X}_q - \mu)^2) &= \mathbb{E}\left(\left(\frac{1}{q} \sum_{i=1}^q X_i - \mu\right)^2\right) \\
 &= \frac{1}{q^2} \sum_{i=1}^q \mathbb{E}(X_i^2) + \frac{1}{q^2} \sum_{\substack{i,j=1 \\ i \neq j}}^q \mathbb{E}(X_i)\mathbb{E}(X_j) + \mu^2 - 2\mu \frac{1}{q} \sum_{i=1}^q \mathbb{E}(X_i^2) \\
 &= \frac{1}{q^2} \left(\sum_{i=1}^q (\mathbb{E}((X_i - \mu)^2) + \mu^2) \right) + \frac{q^2 - q}{q^2} \mu^2 - \mu^2 \\
 &= \frac{\sigma^2}{q}
 \end{aligned} \tag{13.10}$$

and therefore

$$\lim_{q \rightarrow \infty} \mathbb{E}((\bar{X}_q - \mu)^2) = 0, \tag{13.11}$$

which means that the sample mean converges to the population's expected value in the mean square sense.

This derivation also shows that the standard deviation of the mean \bar{X}_q shrinks towards zero as $1/\sqrt{q}$.

There are different kinds of convergence of random variables and the mean square one is one of various possibilities. Another useful convergence is defined by the probability itself on which the following variation of the law of large numbers is based.

13.3.2 Weak law of large numbers

Definition 13.4 (Convergence in probability). A sequence of random variables Y_1, \dots, Y_q, \dots **converges in probability** to a random variable Y if for any $\varepsilon > 0$ one has

$$\lim_{q \rightarrow \infty} P(|Y_q - Y| > \varepsilon) = 0, \tag{13.12}$$

and one writes

$$\lim_{q \rightarrow \infty} Y_q \xrightarrow{P} Y. \tag{13.13}$$

For a set of uncorrelated random variables X_1, X_2, \dots with the same mean μ

and with

$$\lim_{q \rightarrow \infty} \frac{1}{q^2} \sum_{i=1}^q \mathbb{E}((X_i - \mu)^2) = 0 \quad (13.14)$$

one has that the mean \bar{X}_q converges in probability (see definition 13.4) to the population mean:

$$\lim_{q \rightarrow \infty} \bar{X}_q \xrightarrow{P} \mu. \quad (13.15)$$

It is straightforward to verify this convergence by using Chebyshev's inequality (Section 5.5.1). Specifically, one must demonstrate that for any $\varepsilon > 0$

$$\lim_{q \rightarrow \infty} P(|\bar{X}_q - \mu| \geq \varepsilon) = 0. \quad (13.16)$$

This can be obtained by noticing that the variance of \bar{X}_q is σ^2/q (with σ the variance of X_i , assumed defined for simplicity) and using the Chebyshev's inequality:

$$P(|\bar{X}_q - \mu| \geq \frac{k}{\sqrt{q}}\sigma) \leq \frac{1}{k^2}. \quad (13.17)$$

Using a value of k such that $\varepsilon \leq \frac{k}{\sqrt{q}}\sigma$ one has

$$P(|\bar{X}_q - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{q\varepsilon}. \quad (13.18)$$

that indeed, for any finite ε , goes to zero when $\sigma^2/q \xrightarrow{q \rightarrow \infty} 0$.

One can notice that the distribution of \bar{X}_q is narrowing around μ with a convergence law of the standard deviation in $1/\sqrt{q}$.

Notice that, despite the above derivation is for finite variance σ , in its general formulation, this law does not assume finite variance for X_i . Indeed, the condition in Eq.13.14 can be satisfied by diverging second central moments.

13.4 Rate of convergence of the sample mean towards the expected value

The sample moments converge *almost surely* (strong convergence) to the corresponding moments of the population if these are finite and if the observations are independent and identically distributed (i.i.d.). Asymptotically (i.e. when q is large and goes towards ∞), if the mean and variances are defined, then the rate of convergence of the sample mean $\hat{\mu}$ to the population mean μ is in $1/\sqrt{q}$, implying that for any positive η exists q^* such that, for $q > q^*$

$$|\hat{\mu} - \mu| < \frac{\eta}{\sqrt{q}}, \quad (13.19)$$

Therefore, for instance, to halven the likelihood of $\hat{\mu}$ to be observed above a given distance from the mean one must quadruple the number of observations.

When the variance is not defined, the weak law of large numbers still guarantees convergence toward the mean. This is of particular interest when fat-tailed distributions with undefined second moments are involved. A result by Marcinkiewicz–Zygmund (see Brillinger [1962]) demonstrates that for fat-tailed independent random variables with tail exponents α , for any positive η exists q^* such that, for $q > q^*$

$$|\hat{\mu} - \mu| < \frac{\eta}{q^{1-1/\alpha}} \quad \text{for } 1 < \alpha < 2 \quad (13.20)$$

while

$$|\hat{\mu} - \mu| < \frac{\eta}{q^{1/2}} \quad \text{for } \alpha \geq 2. \quad (13.21)$$

Remark 13.2. Note that this law of convergence holds true for any sample estimation of the expected value of any real function of the random variable $\mathbb{E}(g(X))$. Indeed, any function of a random variable is a random variable. This, for instance, applies for the function, $g(X) = \mathbf{1}_{a \leq X < b}$, that returns 1 if the value of the variable $X = x$ is in the interval $x \in [a, b)$. This particular case is of interest because the sample mean of this function is the relative frequency of the observations in the interval $[a, b)$. This means that the relative frequencies converge toward the probability. I will make use of this in Sections 13.5, 13.6, and 13.7 where I discuss nonparametric estimation of the probability distribution function from relative frequencies.

Remark 13.3. When a moment of a random variable is undefined (i.e. it is divergent to plus or minus infinite), the sample moment (being a sum of finite numbers) will always return finite numbers. However, such numbers will increase, in absolute value, with the number of observations. This is a very simple way to identify, from observations, that a moment of a random is undefined.

Remark 13.4. Besides the convergence rate, the probability distribution of the sample mean around the expected value is known in some cases. For instance, the central limit theorem (see Theorem 5.1) states that if

$$X_k \sim \mathcal{N}(\mu, \sigma^2) \quad (13.22)$$

then the mean is also normal with

$$\bar{X}_q \sim \mathcal{N}(\mu, \sigma^2/q). \quad (13.23)$$

Indicating, indeed, that the distribution of \bar{X}_q narrows around the expected value with a standard deviation that decreases with $1/\sqrt{q}$.



Example 13.2 (Bias of the sample variance). The expected value of the mean is the population mean

$$\mathbb{E}(\bar{X}_q) = \mathbb{E}\left(\frac{1}{q} \sum_{i=1}^q X_i\right) = \frac{1}{q} \sum_{i=1}^q \mathbb{E}(X_i) = \mu. \quad (13.24)$$

In this respect, the sample mean is an unbiased estimator (see Definition 2.9). Conversely, this is not true for the sample variance. Indeed,

$$\begin{aligned} \mathbb{E}\left(\frac{1}{q} \sum_{i=1}^q (X_i - \bar{X}_q)^2\right) &= \frac{1}{q} \sum_{i=1}^q \mathbb{E}\left((X_i - \frac{1}{q} \sum_{j=1}^q X_j)^2\right) \\ &= \frac{1}{q} \sum_{i=1}^q \left(\mathbb{E}(X_i^2) + \frac{1}{q^2} \sum_{j,k=1}^q \mathbb{E}(X_j X_k) - 2 \frac{1}{q} \sum_{j=1}^q \mathbb{E}(X_i X_j) \right) \end{aligned}$$

Assuming the observations are i.i.d., one has

$$\begin{aligned} \frac{1}{q^2} \sum_{j,k=1}^q \mathbb{E}(X_j X_k) &= \frac{q}{q^2} \mathbb{E}(X^2) + \frac{q(q-1)}{q^2} \mathbb{E}(X)^2 \\ \frac{1}{q} \sum_{j=1}^q \mathbb{E}(X_i X_j) &= \frac{1}{q} \mathbb{E}(X^2) + \frac{q-1}{q} \mathbb{E}(X)^2 \end{aligned}$$

where $\mathbb{E}(X) = \mu$ and $\mathbb{E}(X^2) = \sigma^2 + \mu^2$

Therefore:

$$\mathbb{E}\left(\frac{1}{q} \sum_{i=1}^q (X_i - \bar{X}_q)^2\right) = \frac{q-1}{q} \mu_2, \quad (13.25)$$

indicating that there is a systematic error in the sample estimation of the variance. An unbiased estimator is

$$\hat{\sigma}_{unbiased}^2 = \frac{q}{q-1} \hat{\mu}_2. \quad (13.26)$$

All sample estimates of all central moments are biased.

13.5 Estimation of the probability mass function

For modeling purposes, one wants to estimate the entire probability distribution not only its moments. Let me first consider the simpler case of a discrete random variable X of which one has q observations $\hat{x}_1, \dots, \hat{x}_q$. If $n(x)$ is the number of times the value $X = x$ is observed (frequency), then the weak law of large numbers tells

us that the relative frequency of its occurrence in the observation set approximates well the probability and, for large q , will converge to it in the limit $q \rightarrow \infty$:

$$P(X = x) \stackrel{P}{\lim}_{q \rightarrow \infty} \frac{n(x)}{q}. \quad (13.27)$$

From the discussion in Section 13.4, it should be clear that the convergence rate is in $1/\sqrt{q}$ when the variance is defined.

For continuous variables, the same is true for the frequency of observation within a given interval. Let me start with the interval $[-\infty, x)$ that identifies the cumulative probability distribution function.

13.6 Estimation of the cumulative probability distribution function

For both continuous or discrete variables, one can count the number of observations in the observation set $\hat{\mathbf{x}} = (\hat{x}_1, \dots, \hat{x}_q)$ that have values smaller or equal to a given value x

$$n(\hat{\mathbf{x}} \leq x) = \sum_{k=1}^q \mathbf{1}_{\hat{x}_k \leq x} \quad (13.28)$$

which is called ‘order statistics’. Given a set of q observations $\hat{x}_1, \dots, \hat{x}_q$ of a random variable X , the indicator function $\mathbf{1}_{\hat{x}_i \leq x}$ returns 1 when $\hat{x}_i \leq x$ and zero otherwise. Therefore, indeed, $n(\hat{\mathbf{x}} \leq x) = \sum_{i=1}^q \mathbf{1}_{\hat{x}_i \leq x}$ counts the number of observations that have values smaller or equal to x . This counting corresponds to the ordinal ranking of the variable i.e. the position of the variable in a sorted list (see Definition 8.7). One can use the same argument on the relative frequencies, discussed in the previous section, to verify that the frequency estimate of the cumulative distribution function must converge to the population’s cdf:

$$F(x) = P(X \leq x) \stackrel{P}{\lim}_{q \rightarrow \infty} \frac{n(\hat{\mathbf{x}} \leq x)}{q}. \quad (13.29)$$

This provides a practical tool to estimate the cumulative probability distribution from data with strong convergence provided by the weak law of large numbers (see Section 13.3.2).

The estimator $n(\hat{\mathbf{x}} \leq x)/q$ is a good unbiased estimator of the cumulative probability, however, for a small number of observations, a better sample estimate of the cumulative probability is

$$\hat{F}(x) = \frac{n(\hat{\mathbf{x}} \leq x)}{q+1}. \quad (13.30)$$

The argument in favor of this second estimator is that by using the formula with q at the denominator, given a set of observations $\hat{x}_1, \dots, \hat{x}_q$ with the largest of them having value \hat{x}_{max} , then the sample estimate of the cumulative probability of \hat{x}_{max} would be equal to one. This is consistent with

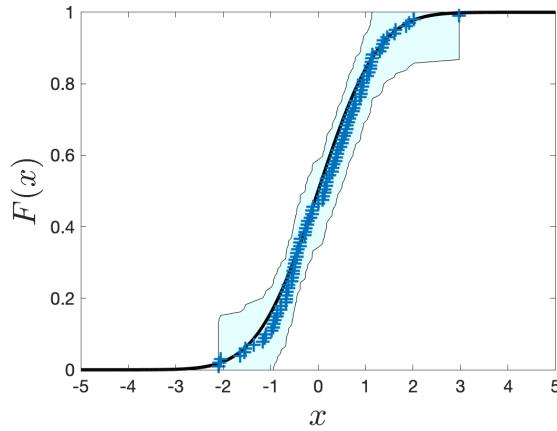


Figure 13.2 Rank-frequency plot (see Remark 13.5) representation of the estimation of a cumulative probability function with $\hat{F}(x) = \frac{n(\hat{x} \leq x)}{q+1}$ for a set of $q = 100$ observations (symbols) from a normally distributed population (black smooth line). The cyan band is the 90% confidence interval from Dvoretzky–Kiefer–Wolfowitz inequality.

the given set of observations but unlikely because, in general, unless \hat{x}_{max} is the true maximum, if one increases the number of observations there is a finite likelihood to have some of them with values larger than \hat{x}_{max} . The estimate with $q + 1$ at the denominator takes into account this possibility making the sample estimate of the cumulative probability of \hat{x}_{max} smaller than one: $\hat{F}(\hat{x}_{max}) = \frac{q}{q+1} < 1$. Of course, the two coincide in the asymptotic limit. This choice is especially useful when quantiles are estimated. Indeed, by using the $1/q$ factor we imply that the 100% quantile is the largest value in the sample, instead, the theoretical limit is infinite (or the largest value in the support interval). Sometimes, for quantile purposes, the estimator $\hat{F}(X) = (n(\hat{x} \leq x) - 0.5)/q$ is used instead.

Remark 13.5. The **rank frequency plot** is a way to plot the sample estimate of the cumulative probability function $\hat{F}(x) = \frac{n(\hat{x} \leq x)}{q+1}$ from a set of q observations. On the y-axis one reports the rank of the observation (from small to large) divided by $q + 1$ (see the previous justification for $q + 1$) and on the x-axis the values. See Fig.13.2 for the use of this plot in a practical example.

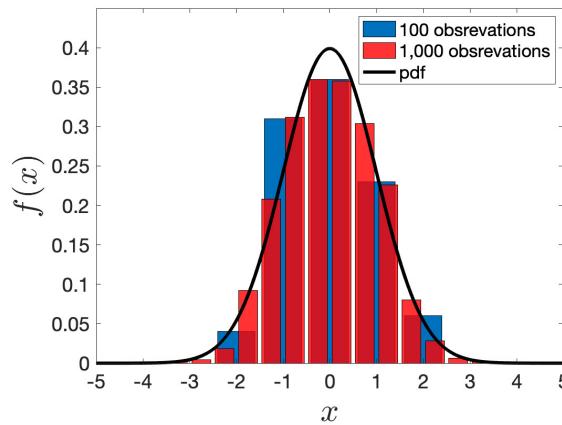


Figure 13.3 Estimation of a probability density function from relative frequencies using respectively 100 (blue bars) and 1,000 (red bars) observations drawn from a normal population with zero mean and unitary standard deviation. The histograms have equal bins respectively of size $h = 1$ and $h = 0.5$. The black curve is the probability density function of the population.

13.6.1 Convergence towards the population's distribution

The cumulative relative frequency $n(\hat{\mathbf{x}} \leq x)/q = \frac{1}{q} \sum_{i=1}^q \mathbf{1}_{\hat{x}_i \leq x}$ is the sample mean of the indicator function $\mathbf{1}_{\hat{x}_i \leq x}$. Therefore, the law of large numbers ensures that the convergence towards the expected value is in $1/\sqrt{q}$. There are several results for the rate of convergence and probably the most intuitive and of practical utility is the Dvoretzky-Kiefer-Wolfowitz inequality [Dvoretzky et al., 1956] which provides an interval of values containing the true CDF, with probability $1 - \gamma$:

$$\frac{n(\hat{\mathbf{x}} \leq x)}{q} - \frac{c_\gamma}{\sqrt{q}} \leq F(x) \leq \frac{n(\hat{\mathbf{x}} \leq x)}{q} + \frac{c_\gamma}{\sqrt{q}} \quad (13.31)$$

with $c_\gamma = \sqrt{\log(2/\gamma)/2}$.

Example 13.3. For example, for a confidence level of $1 - \gamma = 90\%$, one has $c_{0.1} = 0.12$, which means that for 100 observations one will have an expectation range around the estimate of ± 0.12 (see also Fig.13.2).

13.7 Estimation of the probability density function with histograms

The sample estimate of the cumulative probability distribution function from the ranking of the data is precise, useful, and uses all available information. However, it yields to a non-continuous non-differentiable function and therefore

the probability density function cannot be directly estimated from it by differentiation. One can, however, estimate such a derivative by taking discrete intervals of size h : $\hat{f}(x) \simeq (\hat{F}(x - h/2) - \hat{F}(x + h/2))/h$, which is the relative number of observations with values between $x - h/2$ and $x + h/2$ (i.e. $n(x + h/2 < X \leq x - h/2) = n(\hat{x} \leq x + h/2) - n(X \leq x - h/2)$) divided by h :

$$\hat{f}(x) = \frac{n(x - h/2 < X \leq x + h/2)}{qh}. \quad (13.32)$$

This is the equal-bin histogram estimate of the probability density function with a bin size equal to h .

The bin size h is a free parameter and its choice can strongly affect results. For a given dataset, the number of bins depends on h via $B = (\max(\hat{x}_k) - \min(\hat{x}_k))/h$. If h is too small, there are too many bins and too few observations counting inside each bin. Vice versa with h too large, too much information is lost by aggregating the data. There have been several proposals to identify the optimal value of the bin size h . However, they all have shortcomings and depend strongly on the data. The most common choices for the number of bins are the square root of the number of observations $B = \lceil \sqrt{q} \rceil$ and the Sturges' formula $B = \lceil \log_2 q \rceil + 1$ [Sturges, 1926]. In general, it is important to visualize the result to assess qualitatively its meaningfulness, and quantitatively the optimal parameter h can be discovered by minimizing an error over a validation dataset.

Example 13.4 (Entropy estimate with histograms with equal bin-width). One of the quantities that one might aim to compute from the estimate of the probability density function is the Shannon entropy. Consistently with the procedure for the histogram estimate, one replaces the integral in the Shannon entropy expression (Eq.7.3) with the following sum:

$$\hat{H}(X) \simeq - \sum_{b=1}^B \hat{f}(x_b) \log \hat{f}(x_b) h, \quad (13.33)$$

where h is the bin size and B is the number of bins and $\hat{f}(x_b)$ is the estimated value (via Eq.13.32) of the probability density function in each bin. It turns out that this estimate is quite accurate in most practical applications.

Let me here report a few estimates for the case $X \sim \mathcal{N}(0, 1)$ which has Shannon entropy $H(X) = \frac{1}{2} \log(2\pi) + \frac{1}{2} = 1.4189\dots$. Summary results from sample estimates of $\hat{H}(X)$ using histograms are in the table below for various sample sizes q (rows) and a number of bins B (columns) between 2 and 100. The values are the means over 100 randomly generated datasets and the \pm indicate the standard deviation.

$q \setminus B$	2	10	50	100
50	1.748 ± 0.275	1.291 ± 0.116	-	-
100	1.937 ± 0.229	1.369 ± 0.071	1.144 ± 0.092	-
1,000	2.019 ± 0.081	1.434 ± 0.026	1.390 ± 0.026	1.364 ± 0.025
10,000	2.098 ± 0.167	1.448 ± 0.008	1.417 ± 0.006	1.413 ± 0.007
100,000	2.350 ± 0.103	1.457 ± 0.005	1.420 ± 0.002	1.419 ± 0.002
1,000,000	2.347 ± 0.074	1.466 ± 0.004	1.420 ± 0.001	1.419 ± 0.001

samples with sizes smaller than or equal to the number of bins are not computed. Results for larger bin sizes between 200 to 5,000 are in the table below:

$q \setminus B$	200	500	1,000	5,000
1,000	1.314 ± 0.023	1.181 ± 0.030	-	-
10,000	1.408 ± 0.008	1.395 ± 0.008	1.374 ± 0.007	1.217 ± 0.014
100,000	1.417 ± 0.002	1.417 ± 0.002	1.415 ± 0.002	1.397 ± 0.002
1,000,000	1.419 ± 0.001	1.419 ± 0.001	1.418 ± 0.001	1.417 ± 0.001

One might notice that the results are all consistent and highly reproducible. Notice that, when the number of bins is small (much smaller than the sample size) the estimate of the entropy is biased towards larger values than the true one. Conversely, with a large number of bins, the bias is in the opposite direction with the entropy being slightly underestimated, the threshold however depends on the sample size.

13.7.1 Histograms with variable bin-width

In some cases, it is convenient to gather the observations into bins of different widths. This is motivated by the aim to populate bins more evenly and therefore have equivalent statistical significance across the different bins. The probability density function can be directly estimated by using bin width covering the finite interval $(a, b]$ by using

$$\hat{f}(x) = \frac{n(\hat{\mathbf{x}} < b) - n(\hat{\mathbf{x}} \leq a)}{q(b-a)} \quad \text{for } x \in (a, b]. \quad (13.34)$$

The use of variable bin-width can be very convenient especially when distributions have ‘fat tails’ and therefore there are a few observations that are far away from the mean that would be badly captured by an equal binning which would result either with very large bins that do not capture the actual shape of the distribution or several bins with zero observations and a few bins in the tails with only a few observations. In these cases, logarithmic binning is often used, where instead of equal-sized intervals in the variable one takes equal-sized intervals in the logarithm of the variable (i.e. $[0.01, 0.1]$, $(0.1, 1]$, $(1, 10]$, $(10, 100]$, ...). Another possible binning in these cases is the quantile binning where the intervals are

chosen such that the number of observations inside each bin is the same (or similar). This method has the advantage to provide bins with equal statistical weight although it could result in too large bins in the tails. An extreme of this binning is when only one observation per bin is included, in this case, for observations sorted in increasing order with no repetition $\hat{x}_1 < \hat{x}_2 \dots < \hat{x}_q$, the estimate for the probability density function becomes $\hat{f}(x) = 1/(\hat{x}_{i+1} - \hat{x}_i)$ for $x \in (\hat{x}_i, \hat{x}_{i+1}]$, which is indeed the density corresponding to the distribution of the observations. With this binning there is no loss of information, however, the result is often too noisy and the use of singly populated bins is not recommendable in general. What is important here to retain is that Equation 13.34 is completely general and applies to any kind of chosen binning.

Example 13.5 (Entropy estimate with histograms with variable bin-width). Let me repeat the identical experiment as in the previous example but estimate entropy using variable bin width adopting a quantile partition where for any number of bins I take bins such that they are equally populated (plus or minus one). In this case, therefore, the counting of observations inside each bin is always the same $\hat{f}(x_b) = q/B$ (if q multiple of B otherwise the number might vary by one unit), the points x_b are the quantiles each corresponding to $Q(x_b) = \hat{F}(x_b)$, while the size of each bin is $h_b = Q(x_b) - Q(x_{b-1})$ with x_0 the minimum value in the sample while x_B is the maximum. The estimate of entropy is

$$\hat{H}(X) \simeq - \sum_{b=1}^B \hat{f}(x_b) \log(\hat{f}(x_b)/h_b), \quad (13.35)$$

This estimate is often more accurate than the one with equal bins especially when fat-tailed distributions are concerned. Let me report the results for the same number of bins and sample sizes as in the previous example

$q \setminus B$	2	10	50	100
50	1.441 ± 0.139	1.284 ± 0.121	1.098 ± 0.123	—
100	1.596 ± 0.106	1.360 ± 0.072	1.191 ± 0.086	1.115 ± 0.079
1,000	1.852 ± 0.064	1.496 ± 0.028	1.400 ± 0.026	1.368 ± 0.024
10,000	2.037 ± 0.049	1.563 ± 0.019	1.441 ± 0.008	1.423 ± 0.007
100,000	2.170 ± 0.042	1.599 ± 0.013	1.453 ± 0.004	1.435 ± 0.003
1,000,000	2.281 ± 0.033	1.627 ± 0.010	1.462 ± 0.002	1.440 ± 0.002

and

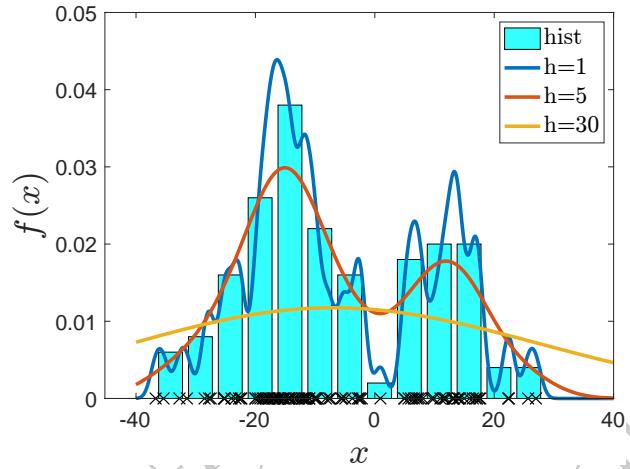


Figure 13.4 Kernel density estimation of a distribution from a set of 100 observations. The figure reports the data points on the x axis. I report a histogram estimate (bars) obtained using equal bins of size equal to 5. The lines are instead the three KDEs respectively with $h = 1, 5, 30$.

$q \setminus B$	200	500	1,000	5,000
1,000	1.327 ± 0.023	1.227 ± 0.027	1.146 ± 0.027	—
10,000	1.413 ± 0.008	1.396 ± 0.007	1.370 ± 0.007	1.229 ± 0.007
100,000	1.426 ± 0.002	1.419 ± 0.002	1.415 ± 0.002	1.395 ± 0.002
1,000,000	1.429 ± 0.001	1.422 ± 0.001	1.420 ± 0.001	1.417 ± 0.001

As for the previous example, the values are the means over 100 randomly generated datasets, and the \pm indicates the region of uncertainty within one standard deviation. Overall results tend to be in line with the previous ones with marginally better estimates obtained when the number of bins is comparable to the sample size.

13.8 Kernel density estimation (KDE)

Histograms can provide excellent approximations for the probability density function, however, they are non-continuous estimators and aggregation of the data into bins reduces the available information in the dataset. An alternative method is kernel density estimation (KDE) which is indeed a way to generate a continuous probability density function from observations without discretizing. KDE is a sum of functions each of them associated with one observation. Formally the KDE estimator $\hat{f}(x)$ can be written as:

$$\hat{f}(x) = \frac{1}{qh} \sum_{k=1}^q K\left(\frac{x - \hat{x}_k}{h}\right). \quad (13.36)$$

Where is $h > 0$ is a scalar called ‘bandwidth’ and $K(\cdot)$ is the ‘kernel function’. Such a function can, in principle, have any form under the conditions of being, integrable, non-negative, and with its integral over the whole support equal to one

$$\begin{aligned} K(x) &\geq 0 \\ \int_{-\infty}^{+\infty} K(x) dx &= 1. \end{aligned} \quad (13.37)$$

Let me first notice that these conditions are the same as the ones for a probability density function. Therefore, the kernels $K(\cdot)$ are probability density functions, and the KDE method approximates a probability density function with a mean over a set of probability density functions with location on each observation and of with scale proportional to h . This method can be seen as a way to construct histograms with smooth densities with width h around the data points instead of just counting the data points inside each bin. The law of large numbers applies also to this particular case of the sample mean and the convergence rate towards the population’s distribution must go as $1/\sqrt{q}$.

The two conditions on the kernels in Eqs.13.37 are necessary and sufficient to guarantee that $\hat{f}(x)$ is a probability density function. Indeed, the first implies $\hat{f}(x) \geq 0$ and from the second $\int_{-\infty}^{+\infty} \hat{f}(x) dx = \frac{1}{q} \sum_{k=1}^q \frac{1}{h} \int_{-\infty}^{+\infty} K\left(\frac{x-\hat{x}_k}{h}\right) dx = 1$. We can also see that the expression in Eq.13.36 for KDE can be generalized to different sets of kernels and bandwidths for each datapoint

$$\hat{f}(x) = \frac{1}{q} \sum_{k=1}^q \frac{1}{h_k} K_k \left(\frac{x - \hat{x}_k}{h_k} \right). \quad (13.38)$$

Remark 13.6. Several kernel functions can be adopted. Normally, they are restricted to be symmetric. Common choices are:

- **Uniform** - which uses the *uniform distribution*, $K(s) = 1/2$ for $s \in [-1, 1]$ and 0 otherwise;
- **Gaussian** - which uses the *normal pdf*, with $\mu = 0$ and $\sigma^2 = 1$, $K(s) = \varphi(s)$.

Let me note that despite both the histogram and the kernel methods being nonparametric, nonetheless they do have parameters and use assumptions such as the bin sizes, the bandwidth, and the choice of kernels.

Example 13.6 (Histograms and KDE). In Figure 13.4 I report the histogram and the KDE for the probability density function associated with a dataset of $q = 100$ observations gendered by randomly mixing data from two normal distributions with 70% of the observations from a normal centered on $\mu_1 = -15$ and with $\sigma_1 = 10$ and the other 30% of the observations from another normal distribution centered on $\mu_2 = 15$ and with $\sigma_2 = 5$. The histogram is constructed using 12 equal bins of size $h = 5$. I also report three KDEs constructed with a Gaussian kernel and bandwidths respectively equal to $h = 1, 5$, and 30 . One can clearly see that $h = 1$ is too small and the resulting estimate of the probability density function has large fluctuations. The choice of $h = 15$ is instead quite good showing well the two underlying Gaussians present in the population. Finally, $h = 30$ is too large and the resulting probability density function is too smooth. In Example 14.6 I'll show how to estimate the best bandwidth using the maximum likelihood method.

The bandwidth h is a free parameter. In this method, the choice of the right kernel with the right bandwidth is crucial. Too small bandwidths produce pdfs with oscillating values with local maxima around the observations. Whereas, too large bandwidths flatten down the distribution losing information. The effect of bandwidth on the shape of the kernel density is made evident in Fig.13.4 where one can notice that for $h = 1$ the function has peaked in correspondence with each datapoint, whereas for $h = 30$ it is excessively smooth capturing only a mean behavior and losing information on the two major peaks.

In general, the bandwidth must decrease with the number of observations and increase with the width of the underlying distribution. There is a formula by Silverman [Silverman, 2018] that applies when the data are from a normal distribution with variance σ^2 :

$$h \simeq \sigma q^{-1/5}. \quad (13.39)$$

For instance, applied to the dataset in Example 13.6, Fig.13.4, this would lead to h in a range between 2 and 4 respectively adopting $\sigma = 5$ or 10 whereas one gets $h = 6.6$ by using the sample standard deviation of the dataset, which is $\hat{\sigma} = 16.5$. Qualitatively, a good practice is to see visually in a plot the resulting pdf and look for the smallest values of h that produce pdfs which are smooth but not ‘too flat’. Quantitatively, let me note that the bandwidth h is a hyperparameter, and one must identify this parameter by minimizing the error over a validation dataset.

In the next chapter, I'll introduce other methodologies to infer the right parameters. Specifically, in Example 14.6 I will discuss how to find h by maximizing likelihood with cross-validation.

Remark 13.7. Nonparametric **estimation of the entropy** requires in general the nonparametric estimation of the population probability distribution function. In principle, therefore the problem coincides with what I discussed in this chapter. For instance, for the Shannon entropy (see Section 7.1) it is required the estimation of the expected value minus the logarithm of the probability. However, when the true distribution is unknown, one estimates the cross-entropy (i.e. the expected value minus the logarithm of the model probability). This coincides with (minus) the model likelihood (see Definition 14.1).

13.9 Data-driven modeling tutorial

<https://github.com/FinancialComputingUCL/DataDrivenModeling/Ch13>

The tutorial for this Chapter covers various topics on the nonparametric estimation of univariate probabilities from data, including: the convergence of the sample means towards the expected value (Example 13.1), the estimation of the cumulative distribution function, estimation of the probability density function with histograms and Kernel density methods (Example 13.6).

Exercises

- From $\hat{\mathbf{x}} = (3.6, 8.0, 5.7, 4.9, 5.7, 4.7, 4.8, 6.4, 6.4, 6.4)$ compute the sample mean and the sample estimates of the first four central moments.
- Assuming that the values of $\hat{\mathbf{x}}$ from the previous question are from a normal distribution, verify Eq.13.31 is followed.
- By using Chebyshev's inequality (Section 5.5.1) estimate how likely is to draw an observation with $\hat{x} \leq 0$ from the same population of $\hat{\mathbf{x}}$.
- Derive the bias on $\hat{\mu}_3$.
- Derive the Kernel density estimation that best fits $\hat{\mathbf{x}}$ and compute the likelihood of the ‘test’ $\hat{\mathbf{x}}^{test} = (5.5, 9.1, 4.3, 3.8, 6.1, 3.7, 3.6, 5.3, 5.2, 5.1)$.

Parametric estimation of univariate probabilities from data

In the parametric approach for the estimation of probabilities one assumes a population distribution which depends on a set of parameters $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_d\}$ and then estimates the distribution's parameters which best represent the observation set.

Normally, parametric estimations of the probability distribution function are simpler to implement and to compute with respect to the non-parametric counterparts and they often require a smaller number of observations to reach equivalent levels of accuracy. However, they require the notion of the underlying distribution of the population. Unfortunately, in general, the underlying distribution of the population is unknown. However, one can assume an a-priori distribution and then verify, a-posteriori, if it represents well the observations. Be mindful that wrong choices of the distribution can result in very bad estimates of the probability of events. For instance, I have already discussed that the normal distribution is very common and represents well the properties of random variables in many natural and artificial systems. However, there are systems – such as the financial markets and most complex systems – that do not follow normal statistics. For the modeling of such systems, the assumption of normal distribution cause a great underestimation of the likelihood of extreme events and ultimately can have disastrous consequences.

Example 14.1 (Estimation of financial risk with normal models). Let me exemplify the problem concerning the assumption of an a-priori population which arises in the parametric estimation of probability. Let me consider the risk of large losses for investments in the financial market. For this purpose, I use data for the NASDAQ Composite daily adjusted closing prices (from Yahoo Finance) for the period February 1971 to September 2022 for a total of $q = 13,013$ observations. From the daily log-returns,

$$\hat{r}(t) = \log Price(t) - \log Price(t-1), \quad (14.1)$$

I observe that there are three days in the period with losses larger than 10%, they are the 19th of October 1987 ('black Friday' crash), the 14th of April 2004 (dot-com bubble burst) and the 16th of March 2020 (COVID crash). From a non-parametric perspective, I can estimate that, historically, the

likelihood for this market to loose 10% or more is $3/13,013 \simeq 0.23\%$ (see Eq.13.29). Conversely, from a parametric perspective, if I model the log-returns as normally distributed, with mean and standard deviation equal to the sample ones ($\hat{\mu} = 0.000368$ and $\hat{\sigma} = 0.0127$) I retrieve $P(R \leq -0.1) = 1.01 \cdot 10^{-15}$ that would erroneously suggest that these events are very unlikely to happen even once in the entire life of the universe. Figure 14.1(a) reports a comparison between the non-parametric histogram estimation of the probability density functions for the NASDAQ log-returns with the parametric normal model. It is evident that there are great differences and normal modeling should not be used in this domain.

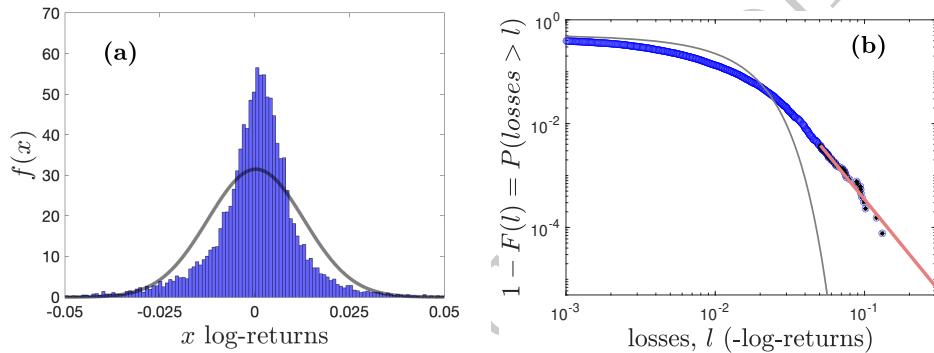


Figure 14.1 (a) Probability density function estimates for the log-returns of NASDAQ Composite daily adjusted closing prices in the period February 1971 to September 2022 ($q = 13,013$ observations). The bars are the histogram-non parametric estimate while the black line is the normal distribution calibrated with moments equal to the sample moments. (b) Complementary cumulative distribution function sample estimate for losses (blue circles) reported in log-log scale. The black line is the the normal distribution while the red straight line is the fit of the tail with a power-law $\propto l^{-\alpha}$ (where $l = -x$ are the losses). The datapoints used for the fit of the tail are marked with a black star inside the blue circle. Notice that $F(0) \simeq 0.5$ because the distribution is centered around zero and the losses are about half of the observations.

Wrong parametric models can be badly misleading, however, appropriate parametric estimates can be very good and they tend to preform better than the non-parametric ones in the tail region of the distribution. Indeed, this is the region of rare events, where the observations provide poor statistics and where often one must estimate the probability of events that never happened before and are not part of the observation set. Therefore they cannot be measured directly from the data. This is however also a region where the chosen parametric form of the distribution can be very different from the population distribution and therefore grossly miss-estimate the probability. The lack of observational data in this tail region might make hard to detect such a misalignment.

Example 14.2 (Parametric tail-modeling). In this example I use the same data for NASDAQ log-returns in the previous example but I focus on the modeling of the tail region of large losses. Extreme losses in financial systems are quite well modeled with power-law distributions. This is indeed shown in Figure 14.1(b) where it is reported the sample estimate of the conjugate cumulative probability

$$P(\text{losses} > l) \simeq 1 - \hat{F}(l) = \frac{n(\text{losses} > l)}{q}, \quad (14.2)$$

together with a parametric estimate of the tail region ($\text{losses} > 0.05$) which is modeled as a power law

$$1 - \tilde{F}(l) \propto l^{-\alpha} \quad (14.3)$$

with tail exponent $\alpha = 3.5$ (see Section 14.3 and Example 14.7 for the computation of this exponent). One can observe that the power law does fit well observations for losses larger than 5%. For instance, by using this parametric estimate the likelihood for a return smaller than -0.1 (i.e. a loss larger than 0.1) is 0.35 % which is in line with observations. This parametric estimation could provide a good indication of the likelihood for even larger losses, that have not been observed yet. For instance, it would imply that the chance of occurrence of a ‘very bad day’ with 50% or more losses is about one in 35 years.

14.1 The method of moments

One of the oldest, simplest and most intuitive method to derive estimates for population parameters, dating back at least to Karl Pearson (1857-1936), is the method of moments. It is intuitive, easy to implement and it often leads to estimators which are good and easy to validate statistically.

The method is based on the fact that the central moments μ_k are functions of the parameters of the distribution $\theta_1, \dots, \theta_d$, and sometimes they coincide with them.

Example 14.3 (Method of moments for the normal distribution). Assuming normal distribution, then there are two parameters ($d = 2$), which are the mean $\theta_1 = \mu$ and the variance $\theta_2 = \sigma^2$. In this case, the method of moments simply uses the sample moments as parameters $\hat{\theta}_1 = \hat{\mu}$ and $\hat{\theta}_2 = \hat{\sigma}$.

In general, the method of moments consists in solving the system of d equations

to estimate the parameters from the sample estimate of the central moments:

$$\begin{aligned}\mu_1(\theta_1, \dots, \theta_d) &= \hat{\mu}_1 \\ \mu_2(\theta_1, \dots, \theta_d) &= \hat{\mu}_2 \\ &\vdots \\ \mu_k(\theta_1, \dots, \theta_d) &= \hat{\mu}_k.\end{aligned}\tag{14.4}$$

Where the solution, $\boldsymbol{\theta}^* = \{\theta_1^*, \theta_2^*, \dots, \theta_d^*\}$, is the set of parameters of a probability distribution function whose central moments are equal to the sample central moments.

Example 14.4 (Method of moments for the Student-t distribution). Assuming Student-t distribution, then $d = 3$ and the parameters are the location $\theta_1 = \mu$, the scale $\theta_2 = \hat{\sigma}$ and the degrees of freedom $\theta_3 = \nu$. One could therefore use the first three non-zero central moments, which are the mean $\mu_1 = \mu$, the variance $\mu_2 = \nu/(\nu - 2)\sigma^2$ (for $\nu > 2$) and the kurtosis $\gamma_4 = \mu_4/\mu_2^2 = 3 + 6/(\nu - 4)$ (for $\nu > 4$). In this case, the method of moments will resolve $\hat{\theta}_1 = \hat{\mu}$, $\hat{\theta}_2 = \hat{\theta}_3/(\hat{\theta}_3 - 2)\hat{\sigma}$ and $\hat{\theta}_3 = 4 + 6/(\hat{\gamma}_4 - 3)$. However, if we have $\nu \leq 4$ (as in most practical cases) this solution cannot be applied because the kurtosis is undefined. Furthermore, even when $\nu > 4$ the estimation of the degrees of freedom from the kurtosis is a poor because large moments are harder to estimate with precision. In this case, the estimation of the degrees of freedom by using the maximum likelihood approach or by fitting the tail exponent are preferable (see discussion in Section 14.3) .

14.2 Maximum likelihood estimation (MLE)

One wants to find the parameters $\theta_1, \dots, \theta_d$ of the assumed parametric distribution such that the empirical observations $\hat{x}_1, \dots, \hat{x}_q$ are most ‘likely’ to be observed. The likelihood is indeed the probability to draw a set of observations with values $\hat{x}_1, \dots, \hat{x}_q$ assuming a given probability distribution function.

Definition 14.1 (Likelihood). The **likelihood function** is defined as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}|\hat{\mathbf{x}}) &= P(\hat{\mathbf{x}}|\boldsymbol{\theta}), \text{ for discrete variables} \\ \mathcal{L}(\boldsymbol{\theta}|\hat{\mathbf{x}}) &= f(\hat{\mathbf{x}}|\boldsymbol{\theta}), \text{ for continuous variables.}\end{aligned}\tag{14.5}$$

With $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, \dots, \hat{x}_q)$.

Given a probability distribution function and an observation set $\hat{\mathbf{x}}$, one wants to find the parameters which maximize the likelihood. It is often more convenient to search for the maximum of the logarithm of this function that is called **log-**

likelihood

$$\ell(\hat{\boldsymbol{\theta}}|\hat{\mathbf{x}}) = \log \mathcal{L}(\boldsymbol{\theta}|\hat{\mathbf{x}}). \quad (14.6)$$

Note that, being the logarithm a monotonically increasing function, the maximum of the logarithm of a positive function coincides with the maximum of the function itself.

We therefore search for the set of parameters which maximize the log-likelihood $\ell(\hat{\boldsymbol{\theta}}|\hat{\mathbf{x}})$ of the model $\tilde{f}(\hat{\mathbf{x}}|\boldsymbol{\theta})$:

$$\boldsymbol{\theta}^* = \operatorname{argmax}_{\boldsymbol{\theta}} \left(\log \tilde{f}(\hat{\mathbf{x}}|\boldsymbol{\theta}) \right). \quad (14.7)$$

the solution, $\boldsymbol{\theta}^* = \{\theta_1^*, \theta_2^*, \dots, \theta_d^*\}$, is the set of parameters that maximizes the probability density function computed at the observed values $\hat{\mathbf{x}} = \{\hat{x}_1, \dots, \hat{x}_q\}$.

Remark 14.1. If the observations are i.i.d. (see Definition 8.1) then the joint probability density function factorizes:

$$\tilde{f}(\mathbf{x}|\boldsymbol{\theta}) = \tilde{f}(x_1|\boldsymbol{\theta}) \dots \tilde{f}(x_q|\boldsymbol{\theta}) \quad (14.8)$$

and therefore the log-likelihood can be written as a sum

$$\ell(\hat{\boldsymbol{\theta}}|\hat{\mathbf{x}}) = \sum_{s=1}^q \log \tilde{f}(\hat{x}_s|\boldsymbol{\theta}), \quad (14.9)$$

this often simplifies computations and it is the main reason why log-likelihood is often used instead of the likelihood in this context.

This is a multi-dimensional optimization problem and in general the maximum is not easy to find. Furthermore, the solution might be optimal for the given observation set $\hat{\mathbf{x}}$ but with other observation sets it can be sub-optimal. The challenge is to find the parameters $\boldsymbol{\theta}$ that provide high likelihoods across datasets with the hope they will generalize well also for data not observed yet. These ‘good’ solutions are not necessarily maxima on any of the available datasets. As I mentioned already in Chapter 3, this is a central issue in data science and machine learning.

Remark 14.2. If the **log-likelihood** $\ell(\hat{\boldsymbol{\theta}}|\hat{\mathbf{x}})$ is continuous and differentiable on the parameter space and if the maximum is an interior point, then the

maximum likelihood estimate is a root of the system of derivatives

$$\begin{aligned}\frac{\partial}{\partial \theta_1} \ell(\boldsymbol{\theta}|\hat{\mathbf{x}}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} &= 0 \\ &\vdots \\ \frac{\partial}{\partial \theta_k} \ell(\boldsymbol{\theta}|\hat{\mathbf{x}}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} &= 0 \\ &\vdots \\ \frac{\partial}{\partial \theta_d} \ell(\boldsymbol{\theta}|\hat{\mathbf{x}}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} &= 0\end{aligned}$$

As for any study of functions one must also verify that the solution is a global maximum and not a local extremum or a minimum.

Within a Bayesian interpretation, the maximum likelihood estimator can be described as the maximum a-posteriori estimate of the parameter $\boldsymbol{\theta}$ given a uniform prior distribution of the parameters. From the Bayes' formula (see Section 6.6)

$$P(\boldsymbol{\theta}|\hat{\mathbf{x}}) = \frac{P(\hat{\mathbf{x}}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\hat{\mathbf{x}})} . \quad (14.10)$$

By assuming a uniform prior $P(\boldsymbol{\theta}) = \text{const}$ and given that $P(\hat{\mathbf{x}})$ is independent from $\boldsymbol{\theta}$, the maximization of the Bayes' posterior probability $P(\boldsymbol{\theta}|\hat{\mathbf{x}})$ is reduced to the maximization of $P(\hat{\mathbf{x}}|\boldsymbol{\theta})$ which is indeed the likelihood (see Definition 14.1).

14.2.1 Relative frequencies are maximum likelihood estimators of the probability

In order to quantify how well a probability density function is estimated by using the observed frequencies one can compute its likelihood. Let me consider the case of discrete variables that is simpler; the case of continuous variables can be handled in the same way by using intervals.

Given a set of i.i.d. observations $\hat{x}_1, \dots, \hat{x}_q$ of a discrete random variable X . The log-likelihood associated with the probability mass function $P(x)$ is

$$\ell = \sum_{s=1}^q \log P(\hat{x}_s) . \quad (14.11)$$

The value of some observations might be repeated several times and this is accounted by the frequency $n(\hat{x})$. Using the frequencies, the sum over the events in

the previous equation can be transformed into a sum over the observed values

$$\ell = \sum_{\hat{x} \in \hat{\mathbf{X}}} n(\hat{x}) \log P(\hat{x}) . \quad (14.12)$$

This formula is a generic expression for the log-likelihood, that depends on the choice of the probability mass function $P(\hat{x})$. It can be proved (see inset afterwards) that this likelihood is maximized when the probability mass function is the relative frequency $\hat{P}(\hat{x}) = n(\hat{x})/q$.

To prove this I use the Kullback-Leibler divergence (KLD, see Definition 7.2) as measure of distance between the relative frequency $\frac{n(x)}{q}$ and the population ‘true’ probability $P(x)$. The proof is a straightforward consequence of the fact that KLD is non-negative, which means

$$D_{KL} \left(\frac{n(x)}{q} \parallel P(x) \right) = \sum_{\hat{x} \in \hat{\mathbf{X}}} \frac{n(\hat{x})}{q} \log \frac{n(\hat{x})/q}{P(\hat{x})} \geq 0 \quad (14.13)$$

which implies

$$\sum_{\hat{x} \in \hat{\mathbf{X}}} \frac{n(\hat{x})}{q} \log \frac{n(\hat{x})}{q} \geq \sum_{\hat{x} \in \hat{\mathbf{X}}} \frac{n(\hat{x})}{q} \log P(\hat{x}) . \quad (14.14)$$

The log likelihood for the set of observations $\hat{x} \in \hat{\mathbf{X}}$, assumed i.i.d., is: $\ell = \sum_{\hat{x} \in \hat{\mathbf{X}}} n(\hat{x}) \log P(\hat{x})$. Therefore, the expression Eq.14.14 implies that ℓ is maximized for $\hat{P}(\hat{x}) = n(\hat{x})/q$.

The same result is obtainable for continuous variables, by following the same steps.

In other words, the relative frequencies $n(x)/q$ are the MLE solution for the probability mass function.

Remark 14.3. There is an apparent contradiction between the maximum likelihood method and maximum entropy principle (see Section 7.2). Indeed, the normalized log-likelihood is a sample estimate of (minus) the Shannon entropy. Therefore, the maximum likelihood method yields solutions that minimize the sample estimate of the entropy. Conversely, the maximum entropy principle seeks for solutions that maximize entropy. Therefore, the two methods seemingly deliver opposite solutions. But this is not the case, in fact the two are consistent: one method seeks for a model that minimizes uncertainty on the observation points, where the solution is known, while the other method maximizes model uncertainty elsewhere.

Let me conclude this discussion on the maximum likelihood estimation with two examples of application: one for the the estimation of the parameters of the

normal distribution and the other for the estimation of the bandwidth in the kernel estimation method. Another application of MLE is provided in Section 14.3.1 for the tail exponent of fat-tailed distributions.

Example 14.5 (MLE for the normal distribution). Assuming i.i.d. observations, the log likelihood function for a normal distribution with mean μ and variance σ^2 is

$$\ell(\hat{\boldsymbol{\theta}}|\hat{\mathbf{x}}) = \sum_{s=1}^q \left(-\frac{1}{2} \log(2\pi\sigma^2) - \frac{(\hat{x}_s - \mu)^2}{2\sigma^2} \right) \quad (14.15)$$

Accordingly with MLE, I now estimate the optimal parameters by searching for the roots of the first derivatives

$$\begin{aligned} \frac{\partial}{\partial \mu} \ell(\boldsymbol{\theta}|\hat{\mathbf{x}}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} &= \sum_{s=1}^q \frac{\hat{x}_s - \hat{\mu}}{\hat{\sigma}^2} = 0 \\ \frac{\partial}{\partial \sigma} \ell(\boldsymbol{\theta}|\hat{\mathbf{x}}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} &= -\frac{q}{\hat{\sigma}} + \sum_{s=1}^q \frac{(\hat{x}_s - \hat{\mu})^2}{\hat{\sigma}^3} = 0 \end{aligned}$$

that lead to the solutions

$$\mu^* = \hat{\mu} = \frac{1}{q} \sum_{s=1}^q \hat{x}_s \quad (14.16)$$

and

$$(\sigma^*)^2 = \hat{\sigma}^2 = \frac{1}{q} \sum_{s=1}^q (\hat{x}_s - \hat{\mu})^2, \quad (14.17)$$

that are indeed the sample mean and variance.

In this case of the normal distribution the maximum likelihood estimation of the parameters is rather straightforward and can be obtained in closed form. This is however not the case for the Student-t distribution where instead the likelihood maximization requires a more elaborate procedure, named expectation maximization, that shall be discussed in Section 14.5.2.

Example 14.6 (MLE with cross-validation for the kernel density estimation). If one tries to use MLE to find the bandwidth h in the Kernel density estimation method (see Section 13.8 and Example 13.6), one would obtain $h \rightarrow 0$. This is a typical example of overfitting. Indeed, distributions of zero width centered around the observations describe perfectly the observation data. However, they will fail to generalize to new data. MLE solutions have the tendency to overfit unless the number of parameters or their values are adequately regularized. A way to cure overfitting is to search for the model

parameters $\boldsymbol{\theta}$ using only a part of the dataset (the ‘train set’ $\hat{\mathbf{x}}^{tra}$) and then estimate the goodness of the model by using another part of the dataset (the ‘validation set’ $\hat{\mathbf{x}}^{val}$).

In the present MLE for KDE context this means to use $\hat{\mathbf{x}}^{tra}$ for the centroids and use instead $\hat{\mathbf{x}}^{val}$ to find the maximum likelihood bandwidth h . Let me do it explicitly for Gaussian kernels:

$$\tilde{f}(x) = \frac{1}{hq^{tra}} \sum_{s=1}^{q^{tra}} \varphi\left(\frac{x - \hat{x}_s^{tra}}{h}\right). \quad (14.18)$$

and

$$h^* = \operatorname{argmax}_h \left(\sum_{j=1}^{q^{val}} \log \frac{1}{hq^{tra}} \sum_{s=1}^{q^{tra}} \varphi\left(\frac{\hat{x}_j^{val} - \hat{x}_s^{tra}}{h}\right) \right). \quad (14.19)$$

The maximum is computable by equaling to zero the derivative with respect to h and this lead to a similar expression as the one for σ^2 in the previous example

$$\frac{\partial}{\partial h} \ell(h|\hat{\mathbf{x}}^{val}) \Big|_{h=h^*} = \sum_{j=1}^{q^{val}} \frac{\frac{1}{hq^{tra}} \sum_{s=1}^{q^{tra}} \left(-\frac{1}{h} + \frac{(\hat{x}_j^{val} - \hat{x}_s^{tra})^2}{h^3}\right) \varphi\left(\frac{\hat{x}_j^{val} - \hat{x}_s^{tra}}{h}\right)}{\tilde{f}(\hat{x}_j^{val})} \Big|_{h=h^*} = 0 \quad (14.20)$$

leading to

$$h^* = \sqrt{\frac{\sum_{j=1}^{q^{val}} \sum_{s=1}^{q^{tra}} w_{j,s}(h) (\hat{x}_j^{val} - \hat{x}_s^{tra})^2}{\sum_{j=1}^{q^{val}} \sum_{s=1}^{q^{tra}} w_{j,s}(h)}} \quad (14.21)$$

with

$$w_{j,s}(h) = \frac{\varphi\left(\frac{\hat{x}_j^{val} - \hat{x}_s^{tra}}{h}\right)}{\tilde{f}(\hat{x}_j^{val})} \quad (14.22)$$

Equation 14.21 is implicit but the solution, h^* , can be found iteratively as a fixed point (see Appendix A.2).

Application of this MLE cross-validation solution for Kernel density estimation for the same dataset of 100 points as in Figure 13.4, returns a median value of the bandwidth of $h^* = 3.53$. Specifically, the cross-validation procedure (see Section 3.5.1) was iterated over ten divisions in training and validating subsets respectively composed of 70 and 30 elements chosen at random, the optimal estimate of the bandwidth was taken as the median over these ten cross-validations.

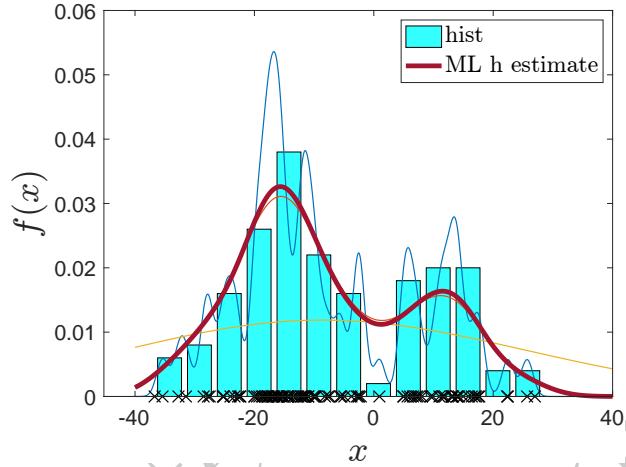


Figure 14.2 Kernel density estimation for the same dataset as in Figure 13.4 with the bandwidth h estimated through maximum likelihood with cross-validation. In this example the value of the optimal bandwidth results in $h^* = 3.53$. The other three tinnier lines correspond to $h = 1, 5, 30$ (same as in Fig.13.4) and are here reported as reference. Note that the line for $h = 5$ is indistinguishable from the one for $h^* = 3.53$.

14.3 Estimation of the tail exponent in fat-tailed distributions

In real, complex systems the probability distribution often reveals a power law behaviors in the ‘tail’ region. Such a behavior can be directly observed by looking at the conjugate cumulative distribution in the tails by plotting the complementary cumulative distribution $P(X > x) = 1 - F(x)$ (right tail) and the cumulative distribution $P(-X > x) = F(x)$ (left tail) in log-log scale. Indeed, a power law behavior implies $P(X > x) \propto x^{-\alpha}$ and therefore $\log P(X > x) \simeq -\alpha \log x + const$, that is a linear decreasing trend in $\log P(X > x)$ vs. $\log x$ (a straight line in log-log scale, see Fig.14.1(b)).

14.3.1 Maximum likelihood estimation of the tail exponent

Assuming power law tails, one can write the log-likelihood function for the tail part of the distribution $x \geq \check{x} > 0$ (or, for the left tail, $x < \check{x} < 0$) as

$$\tilde{f}(x) = c \frac{\alpha}{|\check{x}|} \left(\frac{\check{x}}{x} \right)^{\alpha+1} \quad (14.23)$$

and maximize it to obtain the maximum likelihood estimation of the exponent α . From the sub-set of observations $\{\hat{x}_1, \dots, \hat{x}_q\}$ selected in the tail of the distribution $|\hat{x}_s| \geq |\check{x}|$, the log-likelihood associated with the tail is $\ell_{tail} = \sum_{s=1}^q \log \tilde{f}(\hat{x}_s)$, which by substituting the power law form of $\tilde{f}(\hat{x}_s)$ becomes :

$$\ell_{tail} = \sum_{s=1}^q \log \left(\frac{\alpha}{|\check{x}|} \left(\frac{\check{x}}{\hat{x}_s} \right)^{\alpha+1} \right) + const. \quad (14.24)$$

Its maxima is computable by imposing to zero the derivative with respect to α .

The partial derivative with respect to α is:

$$\frac{\partial}{\partial \alpha} \ell_{tail} = \sum_{s=1}^q \left(\frac{1}{\alpha} + \log \frac{\check{x}}{\hat{x}_s} \right) \quad (14.25)$$

the equation for the root is

$$\sum_{s=1}^q \left(\frac{1}{\alpha} + \log \frac{\check{x}}{\hat{x}_s} \right) \Big|_{\alpha=\alpha^*} = 0. \quad (14.26)$$

The solution for the maximum likelihood estimation of the tail exponent is:

$$\alpha^* = \frac{1}{\frac{1}{q} \sum_{s=1}^q \log \frac{\hat{x}_s}{\check{x}}}. \quad (14.27)$$

Note that in this estimate, \check{x} is an important parameter because it defines the region where the ‘tail’ begins. However, for this parameter, the maximum likelihood method cannot be applied because the maximum is not internal but rather always on the boundary of the domain.

One observes that the partial derivative of ℓ_{tail} with respect to \check{x} is:

$$\frac{\partial}{\partial \check{x}} \ell_{tail} = q \frac{\alpha}{\check{x}}, \quad (14.28)$$

which is a quantity different from zero (for $\alpha \neq 0$ or $q \neq 0$) and it has the sign of \check{x} (for $\alpha > 0$). This means that the extrema is on the boundary and specifically it is the largest (absolute) value in the domain.

14.3.2 Log-log fitting of the tail of the complementary cumulative distribution

The MLE estimator in Eq.14.27 is known as Hill tail index estimator [Hill, 1975] and it is highly popular. There are however a large number of other estimators for the tail index α and, in my opinion, in most cases, the best approach is to estimate the exponent of the tail by best fitting the tail part of the complementary cumulative distribution with a power law.

In the tail region, where the probability density function is well described by a power law, the logarithm of the complementary cumulative distribution has a linear decreasing trend with $\log x$

$$\log P(X > x > \check{x}) \simeq -\alpha \log \frac{x}{\check{x}} + const. \quad (14.29)$$

Therefore, an estimation of the tail exponent can be derived by best-fitting $\log \hat{P}(X > \hat{x})$ vs. $\log \hat{x}$ in the tail region.

The complementary cumulative distribution can be estimated from the ratio (see Section 13.6)

$$1 - \hat{F}(x) = 1 - \frac{N(x)}{q+1}. \quad (14.30)$$

When the tail decays as a power law, one expects $1 - \hat{F}(x) \propto (\frac{\check{x}}{x})^\alpha$ and the linear regression coefficient that minimizes mean square error between $\log(1 - \hat{F}(\hat{x}_s))$ and $-\alpha \log \frac{\hat{x}_s}{\check{x}} + const.$ is

$$\alpha^* = \frac{\sum_{s=1}^q (\log \frac{\hat{x}_s}{\check{x}} - \hat{\mu}_a)(\log(1 - \hat{F}(\hat{x}_s)) - \hat{\mu}_b)}{\sum_{s=1}^q (\log \frac{\hat{x}_s}{\check{x}} - \hat{\mu}_a)^2} \quad (14.31)$$

where

$$\hat{\mu}_a = \frac{1}{q} \sum_{s=1}^q \log \frac{\hat{x}_s}{\check{x}} \quad (14.32)$$

$$\hat{\mu}_b = \frac{1}{q} \sum_{s=1}^q \log(1 - \hat{F}(\hat{x}_s)) \quad (14.33)$$

Notice that, differently from the Hill estimator (see Eq.14.27), in this case α^* does not depend explicitly on \check{x} because the contribute in $\hat{\mu}_a$ cancels with the other term in the parenthesis. Nonetheless, \check{x} is important because it defines the region where the ‘tail’ begins and therefore the choice of the observations to consider.

14.4 Body-tail matching

The ‘tails’ of probability distributions are the extreme parts, far away from the mean, where only a few observations are found. The central part is instead the region near to the mean where most of the observations are found. Often, in many real systems, it is convenient to approximate the probability distribution by using two different models for the body and the tails. In some cases, when the risk of rare and large events are of interest for the modeler, the body can be modeled with the sample cumulative distribution and the tail is instead modeled parametrically. Independently on the modeling of the body and tail parts, the issue is how to connect the two models.

Body and tails must be ‘matched’ at the point \check{x} where the body ends and the tail begins. The only constraint is that the overall probability density function must be non-negative and must integrate to one. One of the criteria for such a matching is making the cumulative distribution function continuous, which consists in setting the parameters such that

$$\tilde{F}_{body}(\check{x}) = \tilde{F}_{tail}(\check{x}), \quad (14.34)$$

which needs to be applied separately for both the left and the right tails. This

criterion is intuitive, however does not correspond to maximum likelihood or best fitting. If the purpose is to exclusively model the tails, then there is no need to constraint the match with the body and the modeling of the tail might be more accurate.

It is not straightforward to establish the point, \check{x} , where the body ends and the tail begins. Indeed, it depends on the distribution of the population and also on the goals of the model (i.e. whether one aims to better model the tails or the body). As for parametric models the point where tail begins, \check{x} , can be considered as a parameter to adjust using one of the criteria such as maximum likelihood estimation (see Example 14.9)). As for the other parametric optimizations, an error function must be defined (i.e. negative likelihood) and it must be minimized. Minimization on the validation set can reduce overfitting.

Example 14.7 (Modeling tails with power laws). For risk estimation and management purposes, the tails of the distributions are the parts that must be modeled and for these extreme parts power laws distributions are often particularly suitable. Let me first focus on only one of the tails, the right tail with $r = x$ when $x \geq 0$, I'll then repeat identically the operation for the left tail just, flipping the distribution and proceed using the variable $l = -x$ when $x < 0$. Formally,

$$\tilde{f}_r(r) = \begin{cases} \hat{f}(r) & \text{for } 0 \leq r \leq \check{r} \\ c_r \frac{\alpha_r}{r^{\alpha_r+1}} & \text{for } r > \check{r}, \end{cases}; \quad \tilde{f}_l(l) = \begin{cases} \hat{f}(l) & \text{for } 0 \leq l \leq \check{l} \\ c_l \frac{\alpha_l}{l^{\alpha_l+1}} & \text{for } l > \check{l}, \end{cases} \quad (14.35)$$

and equivalently for the cumulative

$$\tilde{F}_r(x) = \begin{cases} \hat{F}(r) & \text{for } 0 \leq r \leq \check{r} \\ c_1 \frac{1}{r^{\alpha_r}} & \text{for } r > \check{r}, \end{cases}; \quad \tilde{F}_l(l) = \begin{cases} \hat{F}(l) & \text{for } 0 \leq l \leq \check{l} \\ c_l \frac{1}{l^{\alpha_l}} & \text{for } l > \check{l}. \end{cases} \quad (14.36)$$

Let me discuss a practical case for APPLE stocks daily log returns. I collected daily adjusted closing prices for APPLE from Yahoo Finance [Finance] for the period 12 December 1980 to 09 September 2022 for a total of 10,525 observations. The complementary cumulative distributions for right side of the distribution, which in this cases are gains (positive log-returns) and for the left side of the distribution, which are losses (negative log-returns) are reported in Fig.14.5 with upward and downward triangular symbols. I identify as ‘tails’ the observations above the 95% quantile. In the present case, \check{r} , \check{l} cannot be identified by maximizing likelihood because the sample estimate always returns better in-sample likelihoods and it would lead to no tails to fit. The identification \check{r} , \check{l} by maximizing likelihood is meaningful when both body and tails are modeled parametrically. I will show this in Example 14.9. Alternatively, one could do an out-of-sample likelihood maximization. Here I instead use Eq.14.31 to compute the best-fitting tail exponent. I obtain respectively $\alpha_r = 3.7$ for the gains

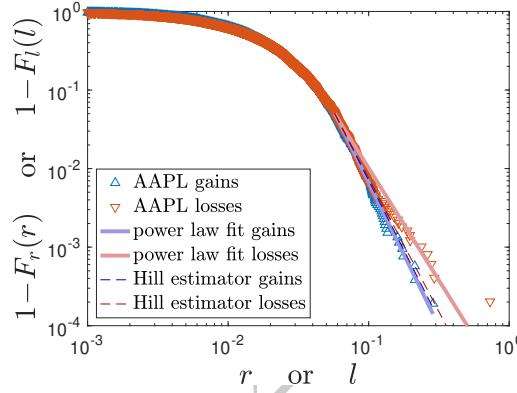


Figure 14.3 Complementary cumulative distributions for gains (positive log-returns) and for the losses (negative log-returns) for the APPLE inc. for the period 12 December 1980 to 09 September 2022 (10,525 observations) plotted in log-log scale. The last 95% quantile of the tails are fitted with power law distributions (straight lines in log-log scale). Notice that, differently from Fig.14.1(b) here the complementary cumulative distributions start from the value of 1 on the left side. Indeed, here I consider the right and the left sides as independent distributions (returns and losses) and therefore the probability of negative values is zero.

and $\alpha_l = 2.9$ for the losses. By using the Hill estimator formula, Eq.14.27, I obtain instead $\alpha_r = 3.5$ for the gains and $\alpha_l = 3.4$ for the losses. As one can note the Hill estimator fails to capture the fatter tails with smaller exponent in the losses. The resulting left- and right-tails distributions for both the best-fitting and Hill estimators are shown in Fig.14.5. For these tail models, the body-tail matching was done by making the cumulative continuous through Eq.14.34.

14.5 Expectation maximization (EM)

The expectation maximization (EM) is a method to find the maximum likelihood solution for the parameter set, $\boldsymbol{\theta}$, of a probability density function $\tilde{f}(x|\boldsymbol{\theta})$, given a set of observations $\hat{\mathbf{x}} = \{\hat{x}_1, \dots, \hat{x}_q\}$ of the continuous random variable X . The approach assumes that the complete model depends also on a variable, z that is not observed and it is called ‘latent’. The complete model is therefore $\tilde{f}(z, x|\boldsymbol{\theta})$ and what one wants to maximize is the likelihood of its marginal distribution:

$$\tilde{f}(x|\boldsymbol{\theta}) = \int_{z \in \Omega_z} \tilde{f}(z, x|\boldsymbol{\theta}) dz , \quad (14.37)$$

which is indeed the probability density function whose likelihood must be maximized. The likelihood of the complete model $\tilde{f}(z, x|\boldsymbol{\theta})$ cannot be maximized directly because indeed there are no observations for the latent variable z , however

one can maximize instead the expected value of $\log \tilde{f}(z, \hat{\mathbf{x}}|\boldsymbol{\theta})$ over the variable z given the observations $\hat{\mathbf{x}}$ while assuming a given set of parameters $\boldsymbol{\theta}$. There are many real systems where the observable variables are only a part of the variables that determine the behavior of the system and this assumption of the existence of a ‘latent’ variable can be indeed very realistic. Furthermore, in many problems it turns out the the maximization of the expected value of $\log \tilde{f}(z, \hat{\mathbf{x}}|\boldsymbol{\theta})$ can be simpler than the maximization of $\tilde{f}(x|\boldsymbol{\theta})$. The EM method can be therefore seen either as a description of a real system with real latent variables or as a useful mathematical trick.

Specifically, the EM works in two steps:

- **E step.** Given a set of parameters $\boldsymbol{\theta}^t$ compute the function

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = \int_{z \in \Omega_z} \tilde{f}(z|\hat{\mathbf{x}}, \boldsymbol{\theta}^t) \log \tilde{f}(z, \hat{\mathbf{x}}|\boldsymbol{\theta}) dz. \quad (14.38)$$

- **M step.** Maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$ with respect to $\boldsymbol{\theta}$ in order to find the new set of estimators $\boldsymbol{\theta}^{t+1}$:

$$\boldsymbol{\theta}^{t+1} = \operatorname{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t). \quad (14.39)$$

It turns out that the parameters $\boldsymbol{\theta}^{t+1}$ that are improving the estimation of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$ are improving at least as much the log-likelihood associated with $\tilde{f}(x|\boldsymbol{\theta})$.

To show that the values of $\boldsymbol{\theta}$ that are improving the value of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$ increase also the log-likelihood $\log \tilde{f}(\hat{\mathbf{x}}|\boldsymbol{\theta}^t)$ we can use the Baye’s formula:

$$\tilde{f}(z, \hat{\mathbf{x}}|\boldsymbol{\theta}) = \tilde{f}(z|\hat{\mathbf{x}}, \boldsymbol{\theta}) \tilde{f}(\hat{\mathbf{x}}|\boldsymbol{\theta}), \quad (14.40)$$

which implies

$$\log \tilde{f}(\hat{\mathbf{x}}|\boldsymbol{\theta}) = \log \tilde{f}(z, \hat{\mathbf{x}}|\boldsymbol{\theta}) - \log \tilde{f}(z|\hat{\mathbf{x}}, \boldsymbol{\theta}). \quad (14.41)$$

Multiplying both sides by $\tilde{f}(z|\hat{\mathbf{x}}, \boldsymbol{\theta}^t)$ and integrating over z gives

$$\log \tilde{f}(\hat{\mathbf{x}}|\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) + D(\boldsymbol{\theta}^t|\boldsymbol{\theta}). \quad (14.42)$$

With

$$D(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = - \int \tilde{f}(z|\hat{\mathbf{x}}, \boldsymbol{\theta}^t) \log \tilde{f}(z|\hat{\mathbf{x}}, \boldsymbol{\theta}) dz. \quad (14.43)$$

The likelihood difference is

$$\log \tilde{f}(\hat{\mathbf{x}}|\boldsymbol{\theta}) - \log \tilde{f}(\hat{\mathbf{x}}|\boldsymbol{\theta}^t) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) - Q(\boldsymbol{\theta}^t|\boldsymbol{\theta}^t) + D(\boldsymbol{\theta}^t|\boldsymbol{\theta}) - D(\boldsymbol{\theta}^t|\boldsymbol{\theta}^t). \quad (14.44)$$

The last two terms are the Kullback-Leibler divergence between $\tilde{f}(z|\hat{\mathbf{x}}, \boldsymbol{\theta}^t)$ and $\tilde{f}(z|\hat{\mathbf{x}}, \boldsymbol{\theta})$ and, as such, this term is non-negative (see Section 7.4)

$$D(\boldsymbol{\theta}^t|\boldsymbol{\theta}) - D(\boldsymbol{\theta}^t|\boldsymbol{\theta}^t) = D_{KL}(\tilde{f}(z|\hat{\mathbf{x}}, \boldsymbol{\theta}^t) \| \tilde{f}(z|\hat{\mathbf{x}}, \boldsymbol{\theta})) \geq 0. \quad (14.45)$$

Consequently, the log-likelihood improves as $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$ or more.

$$\log \tilde{f}(\hat{\mathbf{x}}|\boldsymbol{\theta}) - \log \tilde{f}(\hat{\mathbf{x}}|\boldsymbol{\theta}^t) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) - Q(\boldsymbol{\theta}^t|\boldsymbol{\theta}^t).$$

Therefore any $\boldsymbol{\theta}$ that increases $Q(\boldsymbol{\theta}^t | \boldsymbol{\theta}^t)$ increases also the log-likelihood $\log \tilde{f}(\hat{\mathbf{x}} | \boldsymbol{\theta}^t)$. This is the reason why this method is used interactively. However, convergence to maximum likelihood is not guaranteed.

14.5.1 EM for Gaussian mixtures

The most common application of EM is when data are modeled as the result of two or more processes but when, for each observation, one does not have the information about the process from which the observation is drawn. Consider, for instance, observations drawn with probability p from a normal distribution $\varphi(x|\mu_1, \sigma_1^2)$ or with probability $1-p$ from another normal distribution $\varphi(x|\mu_2, \sigma_2^2)$. If the dataset tells us from which of the two models each observation is drawn from, then the maximum likelihood can be applied to each of the two normal distributions using the relative observations retrieving the estimates of the parameters μ_z and σ_z^2 ($z = 1, 2$). Furthermore, in this case, the maximum likelihood for p is the relative frequency of draws from model ‘1’. However, if one does not know the model from which each observation is drawn, then the likelihood of the mixture must be maximized under the assumption that there is a latent variable ‘ p_z ’.

$$\tilde{f}(x|\boldsymbol{\theta}) = \sum_{z=1,2} p_z \varphi(x|\mu_z, \sigma_z^2). \quad (14.46)$$

with $p_1 = p$ and $p_2 = 1 - p$ and $\boldsymbol{\theta} = \{p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2\}$. In the EM framework, the joint probability is

$$\tilde{f}(x, z|\boldsymbol{\theta}) = p_z \varphi(x|\mu_z, \sigma_z^2). \quad (14.47)$$

Given q independent observations $\hat{\mathbf{x}} = \{\hat{x}_1, \dots, \hat{x}_q\}$ one has

$$\tilde{f}(\mathbf{z}, \hat{\mathbf{x}}|\boldsymbol{\theta}) = \prod_{s=1..q} p_{z_s} \varphi(\hat{x}_s|\mu_{z_s}, \sigma_{z_s}^2). \quad (14.48)$$

EM requires first to compute the expectation

- **E step.**

From Eq.14.38 one has

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = \mathbb{E}(\log \tilde{f}(z, \hat{\mathbf{x}}|\boldsymbol{\theta})|\hat{\mathbf{x}}, \boldsymbol{\theta}^t). \quad (14.49)$$

For q i.i.d. observations $\hat{\mathbf{x}} = \{\hat{x}_1, \dots, \hat{x}_q\}$ it becomes

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = \sum_{s=1}^q \mathbb{E}(\log \tilde{f}(z, \hat{x}_s|\boldsymbol{\theta})|\hat{x}_s, \boldsymbol{\theta}^t) \quad (14.50)$$

Since the observations are i.i.d., consequently the condition in the expected value does not have to depend on all observations but only the one in each term:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = \sum_{s=1}^q \mathbb{E}(\log \tilde{f}(z, \hat{x}_s|\boldsymbol{\theta})|\hat{x}_s, \boldsymbol{\theta}^t). \quad (14.51)$$

Explicitly we have

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = \sum_{s=1}^q \sum_{z_s=1,2} \tilde{f}(z_s|\hat{x}_s, \boldsymbol{\theta}^t) \log(p_{z_s} \varphi(\hat{x}_s|\mu_{z_s}, \sigma_{z_s}^2)). \quad (14.52)$$

Notice that, from the Bayes' formula

$$\tilde{f}(z|x, \boldsymbol{\theta}) = \frac{\tilde{f}(x, z|\boldsymbol{\theta})}{\tilde{f}(x|\boldsymbol{\theta})}. \quad (14.53)$$

and therefore

$$\tilde{f}(z|x, \boldsymbol{\theta}) = \frac{p_z \varphi(x|\mu_z, \sigma_z^2)}{\sum_{z'=1,2} p_{z'} \varphi(x|\mu_{z'}, \sigma_{z'}^2)}. \quad (14.54)$$

- **M step.**

To find the maximum of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$ one equals to zero its partial derivatives with respect with each of the parameters. The derivative with respect to p is:

$$\frac{\partial}{\partial p} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = \sum_{s=1}^q \tilde{f}(z=1|\hat{x}_s, \boldsymbol{\theta}^t) \frac{1}{p} - \tilde{f}(z=2|\hat{x}_s, \boldsymbol{\theta}^t) \frac{1}{1-p} \Big|_{p=p^{t+1}} = 0 \quad (14.55)$$

that gives

$$p^{t+1} = \frac{\sum_{s=1}^q \tilde{f}(z=1|\hat{x}_s, \boldsymbol{\theta}^t)}{\sum_{z=1,2} \sum_{j=1}^q \tilde{f}(z|\hat{x}_j, \boldsymbol{\theta}^t)}. \quad (14.56)$$

Note that $\sum_{z=1,2} \tilde{f}(z|\hat{x}_j, \boldsymbol{\theta}^t) = 1$ and the sum at the denominator is equal to q and therefore

$$p_z^{t+1} = \frac{1}{q} \sum_{s=1}^q w_s^t(z) \quad (14.57)$$

with

$$w_s^t(z) = \tilde{f}(z|\hat{x}_s, \boldsymbol{\theta}^t). \quad (14.58)$$

Explicitly, from Eq.14.54

$$w_s^t(z) = \tilde{f}(z|\hat{x}_s, \boldsymbol{\theta}^t) = \frac{p_z \varphi(\hat{x}_s|\mu_z^t, \sigma_z^{t2})}{\sum_{z'=1,2} p_{z'}^t \varphi(\hat{x}_s|\mu_{z'}^t, \sigma_{z'}^{t2})}. \quad (14.59)$$

with $p_{z=1} = p$ and $p_{z=2} = 1 - p$. To estimate the parameters μ_z (with $z = 1$ or 2) one follows the same steps as in the previous Example 14.5

$$\frac{\partial}{\partial \mu_z} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = \sum_{s=1}^q \tilde{f}(z|\hat{x}_s, \boldsymbol{\theta}^t) \frac{\hat{x}_s - \mu_z}{2\sigma_z^2} \Big|_{\mu_z=\mu_z^{t+1}} = 0 \quad (14.60)$$

resulting in

$$\mu_z^{t+1} = \frac{\sum_{s=1}^q w_s^t(z) \hat{x}_s}{\sum_{s=1}^q w_s^t(z)}. \quad (14.61)$$

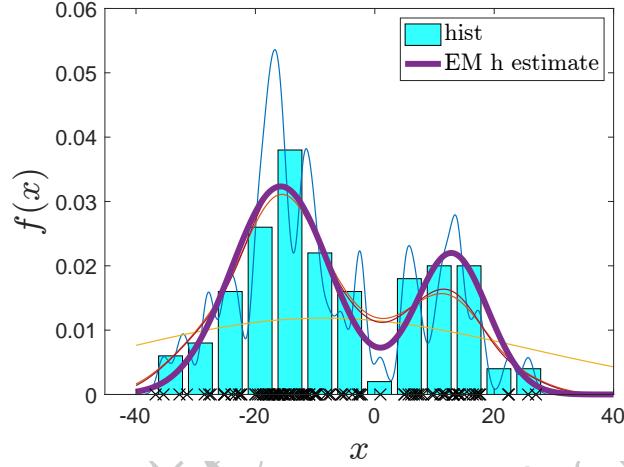


Figure 14.4 Estimation of the density using EM for a mixture of two normal distributions (tick line). The dataset is the same as in Figures 13.4 and 14.2. The tinnier lines correspond to the other models discussed previously and reported also in the mentioned figures.

Analogously for σ_z ,

$$\frac{\partial}{\partial \sigma_z} Q(\boldsymbol{\theta} | \boldsymbol{\theta}^t) = \sum_{s=1}^q \tilde{f}(z | \hat{x}_s, \boldsymbol{\theta}^t) \left(\frac{(\hat{x}_s - \mu_z^{t+1})^2}{\sigma_z^3} - \frac{1}{\sigma_z} \right) \Bigg|_{\sigma_z = \sigma_z^{t+1}} = 0$$

resulting in

$$(\sigma_z^{t+1})^2 = \frac{\sum_{s=1}^q w_s^t(z) (\hat{x}_s - \mu_z^{t+1})^2}{\sum_{s=1}^q w_s^t(z)}. \quad (14.62)$$

Example 14.8 (Expectation maximization method for a Gaussian mixture). Let me apply the EM method to model the dataset already used in Figures 13.4 and 14.2. I start the process assuming the process is a Gaussian mixture with two terms $p = 2$ and I use Eqs. 14.61 and 14.62 to estimate iteratively the parameters. The result is reported in Figure 14.4 where one can see that the resulting model describes very well observations. Indeed, that dataset was generated from a population made by a mixture of two normal distributions respectively with $\mu_1 = -15$, $\sigma_1 = 10$, $\mu_2 = 15$, $\sigma_2 = 5$, $p_1 = 0.7$ and $p_2 = 0.3$. The values retrieved with EM are respectively $\mu_1^* = -15.8$, $\sigma_1^* = 8.25$, $\mu_2^* = 12.9$, $\sigma_2^* = 6.04$, $p_1^* = 0.67$ and $p_2^* = 0.33$. In this example, convergence of the iterative process for the coefficients is reached after 100 to 200 steps depending on the starting conditions.

I have shown so far how to treat the sum of two normal distributions but this approach can be easily generalized to the sum of any number of normal distributions just by letting the variable z to have more than two values. It can

also be directly generalized to the case when z is a continuous variable with its own probability density function $f_Z(z)$.

14.5.2 EM for the Student-t distribution

Definition 14.2 (Log-likelihood of Student-t distribution). From the Definition 5.12 of the probability density function of a Student-t, it directly follows that the log-likelihood for a set of i.i.d. observations $\hat{\mathbf{x}} = (\hat{x}_1, \dots, \hat{x}_q)^\top$ is:

$$\ell(\mu, \hat{\sigma}, \nu | \hat{x}) = q \log \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2}) \sqrt{\pi \nu \hat{\sigma}^2}} - \frac{\nu+1}{2} \sum_{s=1}^q \log \left(1 + \frac{1}{\nu} \left(\frac{\hat{x}_s - \mu}{\hat{\sigma}} \right)^2 \right). \quad (14.63)$$

The Student-t distribution can be constructed as a Gaussian mixture with a gamma distribution:

$$t(x|\boldsymbol{\theta}) = \int_0^{+\infty} g(z | \frac{\nu}{2}, \frac{\nu}{2}) \varphi(x | \mu, \hat{\sigma}^2 / z) dz \quad (14.64)$$

with $\varphi(x | \mu, \hat{\sigma}^2 / z)$ the probability density function of a normal distribution with mean μ and variance $\hat{\sigma}^2 / z$ and

$$g(z | \frac{\nu}{2}, \frac{\nu}{2}) = \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} z^{\frac{\nu}{2}-1} e^{-\frac{\nu}{2}z} \quad (14.65)$$

the probability density function of a gamma distribution with both scale and rate parameters equal to $\frac{\nu}{2}$, sometimes called k-gamma distribution (see Section 5.6.2). To estimate the parameters using EM one can follow a similar path as for the previous example for the mixture of two gaussians. By identifying

$$\tilde{f}(x, z | \boldsymbol{\theta}) = g(z | \frac{\nu}{2}, \frac{\nu}{2}) \varphi(x | \mu, \hat{\sigma}^2 / z). \quad (14.66)$$

One has

$$\tilde{f}(z|x, \boldsymbol{\theta}) = \frac{\tilde{f}(x, z | \boldsymbol{\theta})}{\tilde{f}(x | \boldsymbol{\theta})} \quad (14.67)$$

and

$$\tilde{f}(z | \hat{x}_s, \boldsymbol{\theta}) \propto \tilde{f}(\hat{x}_s, z | \boldsymbol{\theta}^t) = g(z | \frac{\nu}{2}, \frac{\nu}{2}) \varphi(\hat{x}_s | \mu, \hat{\sigma}^2 / z) = \quad (14.68)$$

which is

$$\tilde{f}(z | \hat{x}_s, \boldsymbol{\theta}) \propto g(z | \frac{\nu+1}{2}, \frac{\nu}{2} + \frac{(\hat{x}_s - \mu)^2}{2\hat{\sigma}^2}). \quad (14.69)$$

• **E step.**

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = \sum_{s=1}^q \int_0^\infty \tilde{f}(z|\hat{x}_s, \boldsymbol{\theta}^t) \log \tilde{f}(\hat{x}_s, z|\boldsymbol{\theta}) dz. \quad (14.70)$$

From the expression for $\tilde{f}(z|\hat{x}_s, \boldsymbol{\theta}^t)$ (see Equation 14.69) it follows

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = \sum_{s=1}^q \int_0^\infty \tilde{f}(z|\hat{x}_s, \boldsymbol{\theta}^t) \left(\log g(z|\frac{\nu}{2}, \frac{\nu}{2}) + \log \varphi(\hat{x}_s|\mu, \frac{\hat{\sigma}^2}{z}) \right) dz. \quad (14.71)$$

- **M step.** From the Definition 5.1, the log-likelihood $\log \varphi(\hat{x}_s|\mu, \frac{\hat{\sigma}^2}{z})$ is

$$\log \varphi(x_s|\mu, \frac{\hat{\sigma}^2}{z}) = -\frac{1}{2} \log \frac{\hat{\sigma}^2}{z} - \frac{z(x_s - \mu)^2}{2\hat{\sigma}^2} - \frac{1}{2} \log(2\pi). \quad (14.72)$$

Substituting Eq.14.72 in Eqs.14.71, yields to

$$\frac{\partial}{\partial \mu} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = \sum_{s=1}^q \int_0^\infty z \tilde{f}(z|\hat{x}_s, \boldsymbol{\theta}^t) \frac{\hat{x}_s - \mu}{\hat{\sigma}^2} dz \Big|_{\mu=\mu^{t+1}} = 0,$$

resulting in

$$\mu^{t+1} = \frac{\sum_{s=1}^q w_s^t \hat{x}_s}{\sum_{s=1}^q w_s^t}, \quad (14.73)$$

with

$$w_s^t = \frac{\nu^t + 1}{\nu^t + \frac{(\hat{x}_s - \mu^t)^2}{(\hat{\sigma}^t)^2}}. \quad (14.74)$$

Indeed,

$$w_s^t = \int_0^\infty z \tilde{f}(z|\hat{x}_s, \boldsymbol{\theta}^t) dz. \quad (14.75)$$

From Equations 14.75 and 14.69

$$w_s^t \propto \int_0^\infty z g(z|\frac{\nu^t + 1}{2}, \frac{\nu^t}{2} + \frac{(\hat{x}_s - \mu^t)^2}{2\hat{\sigma}^{t2}}) dz,$$

that is the expected value for a gamma distribution with $\alpha = \frac{\nu^t + 1}{2}$ and $\beta = \frac{\nu^t}{2} + \frac{(\hat{x}_s - \mu^t)^2}{2\hat{\sigma}^{t2}}$ (see Session 5.6.2), and therefore one retrieves

$$w_s^t = \frac{\nu^t + 1}{\nu^t + \frac{(\hat{x}_s - \mu^t)^2}{(\hat{\sigma}^t)^2}}. \quad (14.76)$$

Analogously for $\hat{\sigma}$

$$\frac{\partial}{\partial \hat{\sigma}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = \sum_{s=1}^q \int_0^\infty \tilde{f}(z|\hat{x}_s, \boldsymbol{\theta}^t) \left(-\frac{1}{\hat{\sigma}} + \frac{z(\hat{x}_s - \mu^{t+1})^2}{\hat{\sigma}^3} \right) dz \Big|_{\hat{\sigma}=\hat{\sigma}^{t+1}} = 0,$$

resulting in

$$(\hat{\sigma}^{t+1})^2 = \frac{1}{q} \sum_{s=1}^q w_s^t (\hat{x}_s - \mu^{t+1})^2. \quad (14.77)$$

Finally one can do the same for ν

$$\frac{\partial}{\partial \nu} Q(\boldsymbol{\theta} | \boldsymbol{\theta}^t) \Big|_{\nu=\nu^{t+1}} = 0. \quad (14.78)$$

From the explicit expression,

$$\frac{\partial}{\partial \nu} Q(\boldsymbol{\theta} | \boldsymbol{\theta}^t) = \sum_{s=1}^q \int_0^\infty \tilde{f}(z | \hat{x}_s, \boldsymbol{\theta}^t) \frac{\partial}{\partial \nu} \log g(z | \frac{\nu}{2}, \frac{\nu}{2}) dz, \quad (14.79)$$

one has

$$\frac{\partial}{\partial \nu} \log g(z | \frac{\nu}{2}, \frac{\nu}{2}) = -\frac{1}{2}\psi(\frac{\nu}{2}) + \frac{1}{2} \log \frac{\nu}{2} + \frac{1}{2} + \frac{1}{2} \log z - \frac{1}{2}z, \quad (14.80)$$

where $\psi(\cdot)$ is the digamma function.

Obtaining

$$\psi\left(\frac{\nu^{t+1}}{2}\right) - \log \frac{\nu^{t+1}}{2} = 1 + \sum_{s=1}^q \mathbb{E}(\log z - z | \hat{x}_s, \boldsymbol{\theta}^t). \quad (14.81)$$

Where

$$\mathbb{E}(h(z) | \hat{x}_s, \boldsymbol{\theta}^t) = \int_0^\infty h(z) \tilde{f}(z | \hat{x}_s, \boldsymbol{\theta}^t) dz. \quad (14.82)$$

However, this solution for ν , is not in closed form and must be solved numerically (see Appendix A).

Example 14.9 (Modeling with Student-t). Let me use the same data for APPLE as in Example 14.7. The aim is to estimate of the probability density function. Let me start with the sample estimate with equal-bin histograms, which I report in Fig.14.5(a). I then estimate the best parameters for the Student-t by using the expectation maximization solutions (see Section 14.5.2) and I report the resulting probability density function with the black line in the figure. It is quite evident that this is a good fit for the distribution. The empirical complementary cumulative distributions for the left and right tails are reported in Fig.14.5(b) in log-log scale. The upward triangles are the gains (positive log-returns) while the downwards are the losses (negative log-returns). In this figure it is also reported the EM solutions for the Student-t which still demonstrates a very good fit but also shows that, in the tail region, it cannot capture the asymmetry in the tail

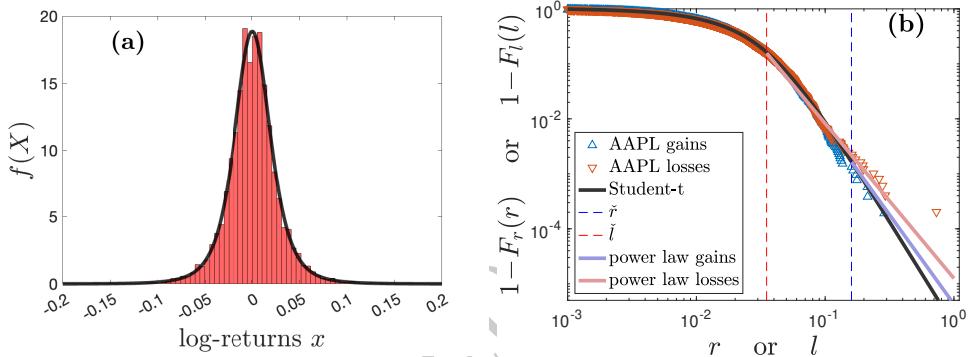


Figure 14.5 (a) Sample probability density function (histogram) and maximum likelihood EM solution for the parametric Student-t distribution (line). (b) Sample complementary cumulative distributions for the positive (gains, upward triangles) and negative (losses, downward triangles) log-returns. The black line is the Student-t EM solution, while the blue and red lines are power law tails best-fitting with exponents and maximum likelihood \bar{x} .

exponents. Indeed, the Student-t is symmetric with tail exponent α equal to the degrees of freedom ν for both left (losses) and right tail (returns). For this dataset, the EM solution returns an exponent, $\nu = \alpha = 3.78$. I however, want to construct a model that contains the observed asymmetry in the tails. To this end I adopt an hybrid model that uses EM Student-t in the for the body and then two power laws with different exponents in the tails. For the tail models, I proceed as described in Example 14.7, however this time I search for the best \check{r} , \check{l} recursively maximizing the likelihood of the left and right part of the distribution separately. I do this by simply computing the likelihood of the hybrid model for a set of different \check{x} and then I pick the best (this is called ‘grid’ method). In this I obtain, for the returns $\check{r} = 0.16$, while for the losses $\check{l} = 0.035$. They are reported as dashed vertical lines in Fig.14.5(b). One can notice, visually, that these points are very good estimates of the region where indeed the tail starts to behave linearly in log-log scale (i.e. as power law). From Eq.14.31, I then retrieve the best-fitting exponents $\alpha_r = 3.25$ and $\alpha_l = 2.81$. This hybrid model has a relative gain in likelihood of 3% with respect to the EM Student-t model. Let me notice the remarkable asymmetry in the model of the left (losses) and right tails (gains). While the gains are well modeled with the Student-t distribution and the power law tail only marginally improves the model (less than 0.1% improvement in likelihood) and only in the very last part of the tail (above the 99% quantile). Conversely, the modeling of extreme losses is highly modified by the power law tail which improves the likelihood for this part by 5.8% and affects 15% of the data (from 85% quantile). Both models

find fatter tails than the Student-t indicating that extreme fluctuations are under-estimated with Student-t modeling. However, the deviation of the left tail (losses) is much more prominent with $\alpha_l = 2.81$ which would even imply an undefined skewness.

Improvements to the modeling could be introduced by amending the condition of continuity of the cumulative distribution (Eq.14.34) letting some more freedom to the two models to adjust. However, this is not the purpose of the present example. Let me notice that continuity is not necessary but $F(x)$ must be a non-decreasing function. In terms of risk this difference in the modeling of the tails is quite dramatic. For instance, from these tail models, chances that in a single day the price will double are less than once in 400 years. Conversely, the chances that in a single day, the price reduces to half are more than once in 32 years.

14.6 Data-driven modeling tutorial

<https://github.com/FinancialComputingUCL/DataDrivenModeling/Ch14>

The tutorial for this Chapter covers various topics on the parametric estimation of univariate probabilities from data, including: the modeling with normal distributions (Example 14.1), the modeling of fat-tailed datasets with power laws (Example 14.7), the use of the expectation maximization method for a Gaussian mixture (Example 14.8), and the modeling with Student-t distributions (Example 14.9).

Exercises

- Consider the set of ten i.i.d. observations $\mathbf{x} = (-0.35, 1.64, -0.46, -0.55, 0.97, 2.03, -4.18, 0.03, -1.02, 1.26)$. By using the method of moments compute the normal distribution model and calculated its log-likelihood.
- By using the data for \mathbf{x} and the Hill formula Eq.14.27
 - estimate the left and right tail exponents and compute likelihood;
 - discuss the criterion of choice of the left and right values of \check{x} and compute likelihood;
 - discuss dependence of the results by \check{x} .
- Repeat the previous exercise by best fitting the tails with a power law (see formula Eq.14.31) instead of the Hill estimator.
- Assuming the population distribution is a Student-t use the expectation-maximization method to estimate the parameters and compute the likelihood.

- Discuss implications of all the above modeling approaches on the estimation of risk.

© Tommaso Aste 2019-23
not for distribution
July 25, 2023

Estimation of multivariate probabilities from data

From a fundamental perspective, the estimation of multivariate probabilities from observations has no differences from the univariate case. However, from the practical and technical perspectives, the increase in the problem's dimensionality makes the estimation task harder, more complicated, more data-greedy, and computationally intensive.

As for the univariate case, there are two main approaches to estimating multivariate probability distributions from data. In the parametric approach, one assumes observations from a process with a known probability distribution and then uses the data to estimate the parameters of such a distribution. Conversely, in the non-parametric approach, the distribution is not assumed as a prior, and its properties are directly derived from the data. Both these estimation processes are very similar to the univariate case.

Let me briefly discuss the non-parametric estimate concerning only the histogram method in the first two Sections of this Chapter, then for the rest of the Chapter, I'll discuss parametric estimations and focus on the estimation of the covariance matrix. Indeed the covariance matrix is a crucial parameter set for a large class of probability distributions, namely the elliptical family. I will also discuss the so-called ‘curse of dimensionality’ and ways to tackle it via shrinkage and regularization methods.

15.1 Non-parametric estimation of multivariate probabilities

The non-parametric approach is conceptually identical to the one described in Chapter 13 for the uni-dimensional case. For instance, the histogram method proceeds in the same way as the approach in one dimension, however, the bins pass from being segments to being squares in two dimensions, cubes in three dimensions, and hypercubes afterward. For instance, analogously to Eq.13.32, in two dimensions the estimate of the bivariate probability density function is given by the frequency of the number of observations within the rectangular region of a two-dimensional bin divided by the area of the bin

$$\hat{f}(x, y) = \frac{n(x - h_X/2 < X \leq x + h_X/2, y - h_Y/2 < Y \leq y + h_Y/2)}{qh_Xh_Y}. \quad (15.1)$$

Also in the multidimensional case, one can use bins of variable sizes. The problem of the partition of the multidimensional space can become much more complex.

plex and rich because it can be divided into cells that are not rectangles, cuboids, or hyper-rectangles in general.

For p variables, the problem is p -dimensional and, if each dimension splits into b bins, then the entire space becomes subdivided in b^p bins. One can see that to populate with a sizable number of observations in each of the bins one needs either a large number of observations $q > b^p$ or a small number of bins per dimension $b < q^{1/p}$. For example, if one has three variables and $q = 1,000$ observations the number of bins per dimension, b , must be smaller than 10 in order to have on average one observation per bin.

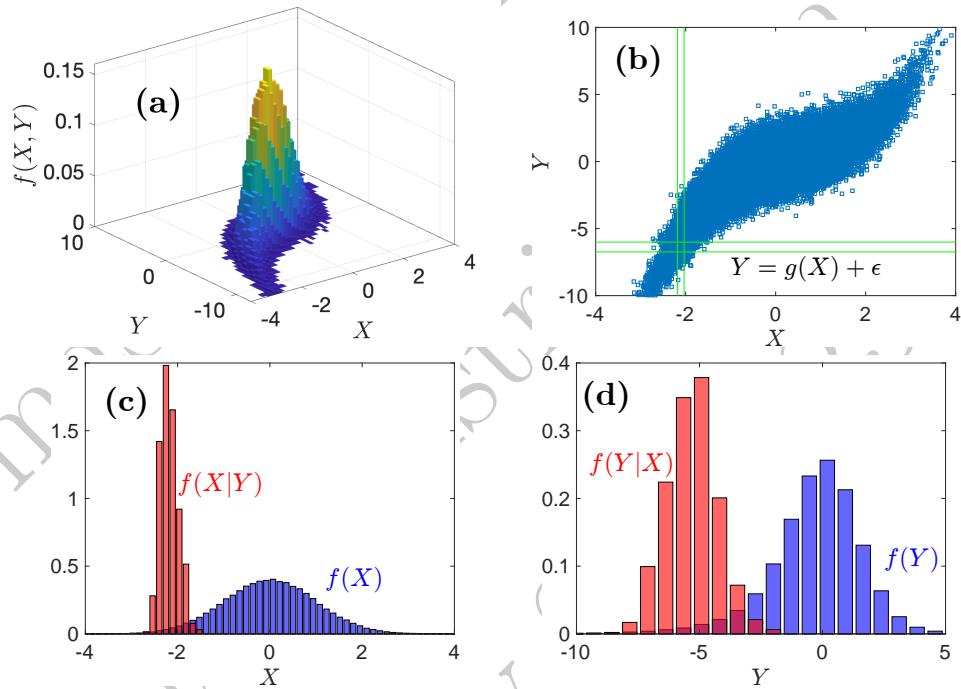


Figure 15.1 (a) Two-dimensional histogram for two coupled random variables X and Y . (b) Scatter plot of Y vs. X , which illustrates the dependency between the two variables. (c) Comparison between the marginal probability for X and the conditional probability for $X|Y$ using histogram estimates (the conditioning is with respect to the values of $Y \in (-2.180, -2.035]$, the horizontal lines in (b)). (d) Comparison between the marginal probability for Y and conditional probability $Y|X$ histogram estimates (the conditioning is with respect to the values of $X \in (-2.180, -2.035]$, the vertical lines in Fig.(b)).

Let me proceed with the following example to illustrate the use of multidimensional histograms in practice.

Example 15.1 (Modeling bi-variate distribution with histograms). Let me deepen a bit into the multivariate probabilities and their non-parametric representations, with a simple practical example. I consider two dependent random variables. Specifically

$$X \sim \mathcal{N}(0, 1) \quad (15.2)$$

and

$$Y = c_1 x + c_2 X^2 + c_3 X^3 + \epsilon \quad (15.3)$$

with $\epsilon \sim \mathcal{N}(0, 1)$. I generated synthetic data with $q = 100,000$ observations and $c_1 = 1$, $c_2 = -0.3$, $c_3 = 0.2$. The histogram estimate of the joint probability density function computed from these data using 60 equally spaced bins is reported in Fig.15.1(a). The other panels in the Figure report respectively: (b) the scatter plot of variable Y vs. variable X ; (c) the marginal probability density function for variable X and the conditional probability density function for variable X given variable Y in the range $(-6.74 - 6.01]$ (values between the horizontal lines in Fig.15.1(b)); equivalently (d) reports the marginal probability density function for variable Y and the conditional probability density function for variable Y given variable X in the range $(-2.180, -2.035]$ (which are the values between the vertical lines in Fig.15.1(b)).

From this example, the dependency between X and Y is very evident both from Fig.15.1(b) which shows a cloud of points with clearly non-circular shapes and Figs. 15.1(c,d), which reveals a great difference between the marginals and the conditional probability density functions. Furthermore, Fig.15.1(b) also clearly reveals that the dependency must be non-linear because the cloud of points is not simply elongated in one direction but rather has a defined shape.

Kernel density estimation methods can also be directly extended to multiple dimensions just by using multivariate kernels. However, also in this case, the increase in dimensionality makes the process harder, requiring a density of points that is increasing geometrically with the power of p or, equivalently, the size of the bandwidth must also increasing with the power of p or faster.

15.2 Non-parametric, non-linear estimation of dependency

To estimate dependency one must quantify the difference between the joint and the marginal probability density functions (see Section 8.1). Therefore, the same tools for the estimation of multivariate probabilities can also be used to estimate dependency. For instance, the effect of dependency is very clear from Figs.15.1(c,d) where it is evident that the conditional distributions are very different from the marginals. The correlation coefficient is the main quantity used to estimate dependency and I shall devote the next Section to discuss its estimation.

However, correlations measure linear dependency and in this section, I want to address the issue from more general non-linear and non-parametric perspectives. For this purpose, I have discussed in the first part of this book that mutual information and its higher-order extensions are the best-suited measures (see Section 8.9). The quantification of mutual information, and related quantities, require the estimation of the Shannon entropy, which can be estimated non-parametrically, for instance by using histogram approximations.

Let me, in the following Example, showcase the non-parametric estimation of mutual information and estimation of dependency for the dataset used in Example 15.1 and Fig.15.1.

Example 15.2 (Estimating non-parametric, non-linear measures of dependency). In this example, I use the same dataset as in the previous Example 15.1.

First of all, let me notice that clearly the two variables are dependent and this is visible from Fig.15.1(b) where a clear regression map $Y \simeq g(X)$ emerges. Such a dependency is also observable by looking at the conditional probabilities that strongly differ from the marginal ones. It is evident that the dependency is not linear, indeed the scatter plot of the two variables does not cloud around a straight line. Nonetheless, in terms of correlations, I retrieve a Person's correlation of $\hat{\rho}_{XY} = 0.801$ (see Eq.15.11) which is a very clear indication of strong (monotonic) dependency.

In terms of entropies, by using the histogram method (and Eqs.13.32, 13.33, 15.1) one has for the joint entropy

$$\begin{aligned} H(X, Y) &= - \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) \log f(x, y) dx dy \\ &\simeq - \sum_{b_x=1}^{B_X} \sum_{b_y=1}^{B_Y} \hat{f}(x_{b_x}, y_{b_y}) \log \hat{f}(x_{b_x}, y_{b_y}) h_X h_Y \end{aligned} \quad (15.4)$$

where B_X and B_Y are the number of bins in each of the two dimensions and h_X, h_Y are the bin sizes while $\hat{f}(x_{b_x}, y_{b_y})$ are the relative frequencies of observations with values within the bins located at (x_{b_x}, y_{b_y}) (see also Example 13.4). Extensions to any number of variables are straightforward by approximating the integral into sums, however, precision becomes extremely problematic, and appropriate numerical integration tools must be adopted Press et al. [2007]. From the synthetic dataset of $q = 100,000$ observations of bivariate non-linearly dependent variables from Exemple 15.1, by using $B_X = B_Y = 60$, I obtain the following estimates:

$$\begin{aligned} \hat{H}(X, Y) &\simeq 2.86; \\ \hat{H}(X) &\simeq 1.42; \\ \hat{H}(Y) &\simeq 2.02. \end{aligned} \quad (15.5)$$

From these estimates, I can then deduce the presence of dependency by quantifying mutual information

$$\hat{I}(X; Y) = \hat{H}(X) + \hat{H}(Y) - \hat{H}(X, Y) \simeq 0.59, \quad (15.6)$$

this value should be compared with the one from the expression for the mutual information for linearly dependent variables, which yields $-\frac{1}{2} \log(1 - \hat{\rho}_{XY}^2) = 0.51$.

Another measure of dependency is the conditional entropy, which I can again estimate from the previous quantities, retrieving

$$\hat{H}(Y|X) = \hat{H}(X, Y) - \hat{H}(X) \simeq 1.438. \quad (15.7)$$

In this example, the variables X and Y are generated by the artificial signal described by Eq.15.3, with $\epsilon \sim \mathcal{N}(0, 1)$.

The conditional entropy, $H(Y|X)$, must be compared with $H(Y)$ qualifying the reduction in uncertainty on Y provided by the knowledge about X . Clearly, the difference between $H(Y)$ and $H(\epsilon) = H(Y|X)$ is – again – the mutual information.

I'll return to this example in Section 18.4.2 where I discuss using $H(\epsilon)$ as a generalized measure of dependency.

15.3 Pearson's estimation of the covariance matrix

Let me for the rest of this Chapter focus on the parametric estimation problems and specifically on the estimation of the covariance matrix. The central role of the covariance matrix in multivariate probability distributions has been discussed at large in Chapter 6. It should therefore be clear that its precise estimation is crucial for the parametric estimate of several multivariate distributions.

Pearson's estimation consists in evaluating each coefficient of the covariance matrix by using the sample mean. Given a set of multidimensional observations of the random variables $\mathbf{X} = (X_1, \dots, X_p)^\top$:

$$\begin{aligned} \hat{\mathbf{x}}_1 &= (\hat{x}_{1,1}, \dots, \hat{x}_{p,1})^\top \\ \hat{\mathbf{x}}_2 &= (\hat{x}_{1,2}, \dots, \hat{x}_{p,2})^\top \\ &\vdots \\ \hat{\mathbf{x}}_q &= (\hat{x}_{1,q}, \dots, \hat{x}_{p,q})^\top, \end{aligned} \quad (15.8)$$

the Pearson's estimate of the covariance $\text{Cov}(X_i, X_j) = (\hat{\Sigma})_{i,j}$, is:

$$(\hat{\Sigma})_{i,j} = \frac{1}{q} \sum_{s=1}^q (\hat{x}_{i,s} - \hat{\mu}_i)(\hat{x}_{j,s} - \hat{\mu}_j). \quad (15.9)$$

where $(\hat{\Sigma})_{i,j}$ is the entry i, j of the sample covariance matrix $\hat{\Sigma}$.

This sample estimate turns out to be the maximum likelihood estimate for

the multivariate normal distribution (see Session 15.5). However, this quantity is biased and the unbiased version has $1/q$ replaced with $1/(q - 1)$.

Remark 15.1. In several instances it is necessary to compute the logarithm of the determinant of the covariance matrix $\log |\hat{\Sigma}|$. For large matrices, the determinant can become extremely large often outside computational capability. However, its logarithm is normally a relatively small number. A way to compute the logarithm of the determinant without directly computing the value of the determinant is by making use of the eigenvalues λ_i of the covariance matrix $\hat{\Sigma}$. Indeed, the determinant is given by the product of the eigenvalues and therefore its logarithm is the sum of the sum of the logarithms of the eigenvalues

$$\log |\hat{\Sigma}| = \sum_{i=1}^p \log \lambda_i. \quad (15.10)$$

15.4 Sample correlations

The estimation of the correlation coefficients follows directly from Pearson's estimate of the covariance coefficients normalizing them by the sample standard deviations:

$$\hat{\rho}_{i,j} = \frac{(\hat{\Sigma})_{i,j}}{\hat{\sigma}_i \hat{\sigma}_j}. \quad (15.11)$$

This estimate returns numbers in the interval $[-1, +1]$. Positive values of $\hat{\rho}_{i,j}$ are an indication of linearly correlated variables, while negative coefficients indicate linearly anti-correlated variables. The extreme values ± 1 indicate perfect linear relation between the two sets of data $\hat{\mathbf{x}}_i = b\hat{\mathbf{x}}_j + c$, with the sign of $\hat{\rho}_{i,j}$ equal to the sign of b .

A small absolute value of the sample correlation coefficient might indicate that there is no linear relation between the two variables. However, it could also correspond to a small, but significantly different from zero, linear correlation. In order to establish if there is a linear dependency between two variables, it is important to establish how far from zero the value of the sample correlation coefficient must be, to be considered significantly different from zero.

15.4.1 Significance of the sample correlation coefficient

Consider q observations from a pair of uncorrelated, normally distributed variables. Even if the two variables are uncorrelated and therefore their true correlation is zero, their sample correlation coefficient will result in a finite value

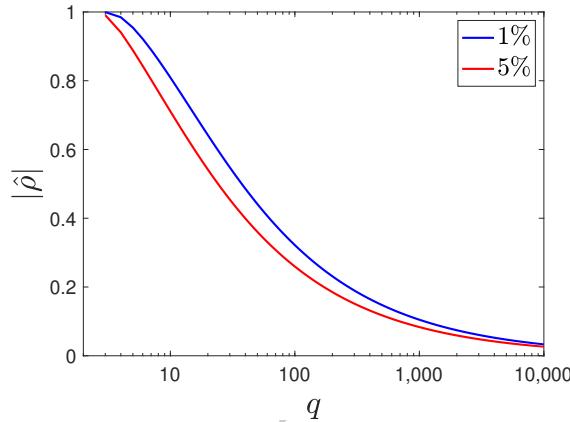


Figure 15.2 The two lines indicate the boundaries of the regions where sample correlation coefficients with absolute values $|\hat{\rho}|$ larger than the bounds are significant at 5% and 1% significance levels respectively. The significance region increases with the sample size.

different from zero which will eventually tend to zero only asymptotically when $q \rightarrow \infty$.¹

For uncorrelated normally distributed variables, the sample correlation coefficient from a sample of size q has values that vary around zero, and the quantity

$$t = \frac{\hat{\rho}_{i,j}}{\sqrt{1 - \hat{\rho}_{i,j}^2}} \sqrt{q - 2} \quad (15.12)$$

is following a Student-t distribution with $q - 2$ degrees of freedom, $\mu = 0$ and $\sigma = 1$. This result can be extended to some cases of non-normally distributed variables, where, however, sometimes (when the probability has fat-tails) the degrees of freedom must be reduced.

The fact that t follows a Student-t distribution can be used to discard the hypothesis that the observed value of the sample correlation is only incidental and therefore this distribution can be used to associate significance to the measure. Indeed, if the probability of t is small, it means that it is unlikely that uncorrelated variables can result in such a value, meaning that $\hat{\rho}_{i,j}$ is significantly different from zero. I shall discuss in more detail statistical validation in Chapter 18 where I shall introduce also non-parametric methods that do not require formulating hypotheses about the population statistics.

Example 15.3 (Significance levels for sample correlations). Let me visualize Equation 15.12 by plotting, in Figure 15.2, the minimum absolute values

¹ From the discussion in Section 13.4 it should be clear that the convergence rate of the sample correlation coefficient towards the true population correlation coefficient is in $1/\sqrt{q}$ but only if the distribution is not fat-tailed with $\alpha \leq 4$.

of the sample correlation coefficient above which it is unlikely to retrieve values for uncorrelated variables at a given confidence level. Specifically, the 5% (1%) line in the Figure is the lower boundary of the region where sample correlation coefficients with values larger than that boundary have a likelihood smaller than 5% (1%) to be produced by chance from uncorrelated variables. This is the so-called p-value, which is discussed in Section 18.2. One can note that for a small number of observations, $q \leq 10$, any sample correlation, $\hat{\rho}$, with absolute values below 0.8 is likely to be obtained by chance from uncorrelated variables and therefore cannot be considered significant. Instead, for a number of observations $q = 100$ or above, sample correlations with absolute values above 0.2 are likely to be significant. While to achieve some significance for values of $|\hat{\rho}|$ below 0.1 one needs 300 or more observations.

Notice that for real-non normal processes, it is likely that a larger number of observations are required to reach the same levels of confidence. Specifically, in many real data, such as financial returns, there are two main factors that make the estimate poorer. First, the variables follow fat-tailed distributions with tail exponents $\alpha < 4$. This implies that the correlations (who are second-order moments) might be definite but their variances (fourth-order moments) are not, and therefore very large fluctuations in the observed values of the correlations are likely to occur. Second, real variables are often auto-correlated and this modifies the statistics as if the effective sample size is smaller.

15.5 Maximum likelihood estimate of the multivariate normal distribution

Definition 15.1 (Multivariate normal log-likelihood). The **log-likelihood of the multivariate normal distribution** for a set of multidimensional observations is:

$$\ell = \frac{q}{2} \log |\mathbf{J}| - \frac{1}{2} \sum_{s=1}^q (\hat{\mathbf{x}}_s - \boldsymbol{\mu}_s)^\top \mathbf{J} (\hat{\mathbf{x}}_s - \boldsymbol{\mu}_s) \quad (15.13)$$

where $\mathbf{J} = \boldsymbol{\Sigma}^{-1}$ is the inverse covariance matrix. It can be written as:

$$\ell = \frac{q}{2} \log |\mathbf{J}| - \frac{q}{2} \text{Tr}(\mathbf{J} \hat{\boldsymbol{\Sigma}}) - \frac{pq}{2} \log(2\pi) \quad (15.14)$$

with $\hat{\boldsymbol{\Sigma}}$ the sample estimate of the covariance (see Eq. 15.9).

Note that often, the constant term $-\frac{pq}{2} \log(2\pi)$ is omitted. Also, normally, the likelihood per unit sample size ℓ/q is used.

The maximum likelihood parameters for the multivariate normal distribution are obtained by searching for the zeros of the partial derivatives of the log-

likelihood with respect to $\boldsymbol{\mu}$ and \mathbf{J}

$$\frac{\partial \ell}{\partial \mu_i} = \sum_{s=1}^q \sum_{j=1}^p J_{i,j} (\hat{x}_{j,s} - \mu_j), \quad (15.15)$$

$$\frac{\partial \ell}{\partial \mathbf{J}} = \frac{q}{2} \boldsymbol{\Sigma} - \frac{1}{2} \sum_{s=1}^q (\hat{\mathbf{x}}_s - \boldsymbol{\mu})(\hat{\mathbf{x}}_s - \boldsymbol{\mu})^\top, \quad (15.16)$$

(the Jacobi's formula, $\partial \log |\mathbf{J}| / \partial \mathbf{J} = \mathbf{J}^{-1} = \boldsymbol{\Sigma}$, was used).

For the means, by setting the first derivative equal to zero, one obtains:

$$\mu_i^* = \frac{1}{q} \sum_{s=1}^q \hat{x}_{i,s}. \quad (15.17)$$

that is the sample mean.

For the covariance, one has:

$$\boldsymbol{\Sigma}^* = \frac{1}{q} \sum_{s=1}^q (\hat{\mathbf{x}}_s - \boldsymbol{\mu})(\hat{\mathbf{x}}_s - \boldsymbol{\mu})^\top = \hat{\boldsymbol{\Sigma}} \quad (15.18)$$

which is Pearson's sample covariance estimate (Eq. 15.9).

15.6 Maximum likelihood estimate of the multivariate Student-t with the expectation-maximization method

Definition 15.2 (Multivariate Student-t log-likelihood). Given a set of q multidimensional observations $\{\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_q\}$, with $\hat{\mathbf{x}}_1 \in \mathbb{R}^p$, the log-likelihood of the multivariate Student-t is

$$\ell = q \log \left(\frac{\Gamma(\frac{\nu+p}{2})}{\nu^{p/2} \pi^{p/2} \Gamma(\frac{\nu}{2})} \right) + \frac{q}{2} \log |\mathbf{J}| - \frac{\nu + p}{2} \sum_{s=1}^q \log \left(1 + \frac{1}{\nu} (\hat{\mathbf{x}}_s - \boldsymbol{\mu})^\top \mathbf{J} (\hat{\mathbf{x}}_s - \boldsymbol{\mu}) \right). \quad (15.19)$$

As for the univariate case, a closed form for the maximum likelihood parameters of the multivariate Student-t does not exist and, as for the univariate case, the expectation-maximization approach must be used instead.

The expectation maximization approach for the multivariate estimation of the parameters of the Student-t distribution follows the same steps as for the univariate case described in Section 14.5.2. Indeed, the multivariate Student-t can be represented as a Gaussian mixture of multivariate normal distributions:

$$t(\mathbf{x}|\boldsymbol{\theta}) = \int_0^{+\infty} g(z|\frac{\nu}{2}, \frac{\nu}{2}) \phi(\mathbf{x}|\boldsymbol{\mu}, \tilde{\boldsymbol{\Sigma}}/z) dz \quad (15.20)$$

with $\phi(\mathbf{x}|\boldsymbol{\mu}, \tilde{\boldsymbol{\Sigma}}/z)$ the probability density function of a multivariate normal distribution with means $\boldsymbol{\mu}$ and covariance matrix $\tilde{\boldsymbol{\Sigma}}/z$, and where $g(z|\frac{\nu}{2}, \frac{\nu}{2})$ is the

probability density function of a gamma distribution with both scale and rate parameters equal to $\frac{\nu}{2}$.

The EM procedure for the mean is identical to the univariate case with the only difference being the scalar parameters replaced by their bold symbols indicating vectors in p dimensions. One obtains:

$$\boldsymbol{\mu}^{t+1} = \frac{\sum_{s=1}^q w_s^t \hat{\mathbf{x}}_s}{\sum_{s=1}^q w_s^t}. \quad (15.21)$$

with

$$w_s^t = \int_0^\infty z g(z | \frac{\nu}{2}, \frac{\nu}{2}) \phi(\hat{\mathbf{x}}_s | \boldsymbol{\mu}^t, \tilde{\boldsymbol{\Sigma}}^t / z) dz$$

which can be computed explicitly in the same way as for Equation 14.76 leading to

$$w_s^t = \frac{\nu^t + p}{\nu^t + (\hat{\mathbf{x}}_s - \boldsymbol{\mu}^t)^\top \tilde{\mathbf{J}}^t (\hat{\mathbf{x}}_s - \boldsymbol{\mu}^t)} \quad (15.22)$$

where $\tilde{\mathbf{J}}^t = (\tilde{\boldsymbol{\Sigma}}^t)^{-1}$. For the shape matrix, we can follow similar steps as for the univariate case resulting in the following expression for the matrix elements:

$$(\tilde{\boldsymbol{\Sigma}}^{t+1})_{i,j} = \frac{1}{q} \sum_{s=1}^q w_s^t (\hat{x}_{i,s} - \hat{\mu}_i^{t+1})(\hat{x}_{j,s} - \hat{\mu}_j^{t+1}). \quad (15.23)$$

The updating expression for the parameter ν^t remains unchanged with respect to the univariate case (see Section 14.5.2) with the scalar \hat{x}_s substituted with the vector $\hat{\mathbf{x}}_s$. However, as for the univariate case, it must be noted that the EM estimate is not the best approach to estimate this parameter. Models that better fit the data are obtained by setting the parameter ν from the fits of the tails of the marginal distributions (they all have the same ν) or by the maximum likelihood method for the tails of the marginals (see Sections 14.3.1 and 14.3.2).

15.7 The curse of dimensionality

The information contained in a set of q observations of p variables scales as $q \times p$ whereas the number of parameters scales as p^2 (i.e. the covariance matrix) or faster. Therefore, for a given number of observations, q , by increasing the number of variables, p eventually the problem becomes under-constrained. This is sometimes counter-intuitive because it means that gathering more information about a system by taking into account of another variable, might make the problem become less definite and eventually unsolvable.

Example 15.4 (The more you look the less you see.). Let me go back to the problem of a pedestrian crossing a road mentioned in the introduction of this book (see Chapter 1). When one crosses a busy road, to evaluate the chances to cross safely one usually ‘models’ the system by estimating the

speed and trajectories of the vehicles. There are many other factors that could be taken into account, some clearly important, such as visibility, but others less clearly relevant such as the color of the vehicle, its brand, the name of the driver, the plate number, etc.. Most of the time, the pedestrian accounts for the most important factors and ignores the less important ones with an intuitive process. Somehow, we learned that these extra factors are irrelevant and taking into account such information will not increase our chances to cross the road safely. It turns out that from a mathematical perspective, by taking into account these extra variables one might end up reducing the accuracy of the model, to the point to make it unsolvable.

For the purpose of quantifying the accuracy of the estimation of matrices, it is useful to use norms, which are defined as follows.

Definition 15.3 (Norm). The **norm** $\|\mathbf{v}\|$ of a vector $\mathbf{v} \in \mathbb{R}^p$ is a scalar measure of distance. It must be non-negative and satisfy three conditions:

1. $\|\mathbf{v}\| = 0$ iff $\mathbf{v} = \mathbf{0}$, being equal to zero at the origin;
2. $\|a\mathbf{v}\| = |a|\|\mathbf{v}\|$, being absolutely scalable;
3. $\|\mathbf{v} + \mathbf{u}\| \leq \|\mathbf{v}\| + \|\mathbf{u}\|$, satisfying the triangular inequality.

The norm of a matrix $\mathbf{A} \in \mathbb{R}^{p_1 \times p_2}$ is also a scalar with the above properties with, in addition, the sub-multiplicative property, which holds for two square matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p \times p}$

4. $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|$.

An often used norm for a $p_1 \times p_2$ matrix, \mathbf{A} , is the **Frobenius norm** which is a scalar equal to the square root of the sum of the squares of the matrix elements:

$$\|\mathbf{A}\|_2 = \sqrt{\sum_{i=1}^{p_1} \sum_{j=1}^{p_2} a_{i,j}^2}. \quad (15.24)$$

Generalization to other norms, known as n -norm, is

$$\|\mathbf{A}\|_n = \left(\sum_{i=1}^{p_1} \sum_{j=1}^{p_2} |a_{i,j}|^n \right)^{1/n}. \quad (15.25)$$

For vectors, they often take the name of L_n -norm. The L_2 -norm is the vector's length in p -dimensional euclidean geometry

$$\|\mathbf{v}\|_2 = (v_1^2 + v_2^2 + \dots + v_p^2)^{1/2}. \quad (15.26)$$

The L_1 -norm is the sum of the absolute values of the components

$$\|\mathbf{v}\|_1 = (|v_1| + |v_2| + \dots + |v_p|). \quad (15.27)$$

While the L_0 -norm is the number of components different from zero. Let me notice that this is actually not a norm because it does not satisfy condition 2. Nonetheless, it is commonly called norm.

15.7.1 Covariance estimation and condition number

When models are using elliptical multivariate distributions, the covariance matrix (when defined) is the main part of the parameter set. It has $p(p+1)/2$ distinct elements, and therefore the number of parameters in this model increases as $\mathcal{O}(p^2)$. If one has q observations the amount of data points is instead $q \times p$ and consequently, it increases linearly with the dimension p . No matter how many observations (q) one can gather, there is always a point at which, by increasing the dimensionality (p), the estimation of the whole covariance matrix becomes too uncertain and eventually unsolvable. It must be noted that this is not trivial and might not be so evident at first glance. For instance, when each element of the covariance is estimated through Pearson's method, one is using observations from two variables only. This is therefore a low dimensional problem and a precise quantification of each coefficient needs a relatively small number of data points. Indeed, the sample covariance between two variables will converge towards the population covariance at a rate $1/\sqrt{q}$ and it was, for instance, shown in Example 15.3 and Figure 15.2 that, with 100 observations, correlations above 0.2 are very unlikely to occur by chance. Therefore, why does the entire covariance become ill-definite when p increases while its coefficients stay at the same level of significance? What causes this dimensionality 'curse'?

A way to better understand this point is by looking at the condition number.

Definition 15.4 (Condition number). The condition number, κ , measures how much, in the worst case, the output of a function is affected by a change in the input argument. For $\kappa = 1$ the noise in the input is not amplified, for $\kappa > 1$ there is amplification in the worst case.

For a linear equation

$$\mathbf{Y} = \mathbf{B}\mathbf{X}, \quad (15.28)$$

with \mathbf{B} a square matrix, the **condition number of the matrix \mathbf{B}** is

$$\kappa = \|\mathbf{B}\|_2 \|\mathbf{B}^{-1}\|_2, \quad (15.29)$$

where $\|\cdot\|_2$ is the Frobenius' norm. The value of κ provides a bound on how accurate can be the solution. Specifically, in this linear case, the relative error in the solution can be up to κ times the relative error in the variable.

This condition number can be also computed from:

$$\kappa = \frac{|\lambda_{\max}(\mathbf{B})|}{|\lambda_{\min}(\mathbf{B})|}, \quad (15.30)$$

where $\lambda_{\max}(\mathbf{B})$ and $\lambda_{\min}(\mathbf{B})$ are the maximum and minimum eigenvalues of \mathbf{B} . One can view κ as the precision that can be lost in the operation. For instance, $\kappa \simeq 10^5$ means that one loses five significant digits in precision. In general, a double-precision floating point number carries fifteen significant digits, therefore when $\kappa > 10^{15}$ the operation can wipe out all information from the data and the result is likely to be completely random. See Edelman [1988] and Myers and Myers [1990] for further readings.

In the case of covariances computed from uncorrelated random variables, the eigenvector spectrum is known from the Marčenko-Pastur formula [Marčenko and Pastur, 1967] and therefore the condition number can be computed explicitly.

Definition 15.5 (Marčenko-Pastur formula). The probability density function, $f(\lambda)$, for the eigenvalues of a covariance matrix computed from q observations from a set, \mathbf{X} , of p i.i.d. uncorrelated random variables with variance $\sigma^2 < \infty$ follows the **Marčenko-Pastur** formula. In the limit $p, q \rightarrow \infty$, but p/q finite, one has:

$$f(\lambda) = \frac{q}{2\pi\sigma^2 p} \frac{\sqrt{(\lambda_{\max} - \lambda)(\lambda - \lambda_{\min})}}{\lambda} \quad (15.31)$$

where

$$\lambda_{\min} = \sigma^2 \left(1 - \sqrt{p/q}\right)^2, \quad \lambda_{\max} = \sigma^2 \left(1 + \sqrt{p/q}\right)^2, \quad (15.32)$$

are, respectively, the largest and smallest eigenvalues.

This distribution applies to uncorrelated random variables. An approximated solution when the variables are all correlated with the same correlation coefficient $\rho > 0$ is:

$$\begin{aligned} \lambda_{\min}(\rho) &\simeq \sigma^2 \left(1 - \sqrt{p/q}\right)^2 (1 - \rho) \\ \lambda_{\max}(\rho) &\simeq \sigma^2 \left(1 + \sqrt{p/q}\right)^2 (1 - \rho) + \sigma^2 p \rho. \end{aligned} \quad (15.33)$$

This indicates that correlations are further reducing the value of the smallest eigenvalue while instead, for large p , the largest eigenvalue can increase. I shall show in the next example that this has important consequences on the condition number and therefore on the effect of dimension on model sensitiveness on input uncertainty. The interested reader can refer to the

original paper by Marčenko and Pastur [1967] and the book by Livan et al. [2018] for details and further readings.

By combining the condition number formula, Eq.15.30, with the Marčenko-Pastur formula for the largest and smallest eigenvalues, Eq.15.32, one retrieve:

$$\kappa_{MP} = \frac{(1 + \sqrt{p/q})^2}{(1 - \sqrt{p/q})^2} \text{ for } q > p. \quad (15.34)$$

Which turns out to be indeed very well followed by the condition number for the covariance matrix of uncorrelated variables. I demonstrate this in the following example.

Example 15.5 (Condition number for covariance of uncorrelated variables). I compute the condition number of covariance matrices from samples of $p = 100$ random, non-correlated normally distributed, variables with zero mean and unitary variance, and sizes from $q = 100$ to $10,000$. In Fig.15.3(a) I compare this condition numbers with the formula for κ_{MP} (Eq.15.34). The agreement is excellent. One can note that, for $q/p \simeq 2$, κ is in the interval between 10^1 and 10^2 implying that one loses no more than two digits in precision for the linear transformation $\Sigma^{-1}\mathbf{X}$, which is the main term in linear regressions (see Example 15.9 for an in-depth discussion of the error in multilinear regression).

For correlated variables, the spectrum of a covariance matrix from a correlated process broadens considerably outside the Marčenko-Pastur spectrum (as Eq.15.33 describes). Therefore, the condition number can become considerably larger. For instance, for a set of p variables all coupled with constant correlation ρ , for large p , the coefficient of determination, computed combining Eq.15.30 with Eq.15.33, becomes well approximated by

$$\kappa_{MP}(\rho) = \kappa_{MP} + \frac{p}{(1 - \sqrt{p/q})^2} \frac{\rho}{1 - \rho}. \quad (15.35)$$

Let me note that, for finite correlations, $\kappa_{MP}(\rho)$ increases linearly with dimension p and in the limit $p/q \rightarrow 0$ it converges towards $p\rho/(1 - \rho)$ instead of 1. Implying that, even when the number of observations is large – even in the asymptotic limit when it goes to infinite – in a correlated system, uncertainty still increases with the dimensionality of the problem. The coefficient of determination diverges both linearly with p , and also, for finite p , when $\rho \rightarrow 1$.

Example 15.6 (Condition number for covariance of correlated variables). I compute the condition number of covariance matrices from samples of

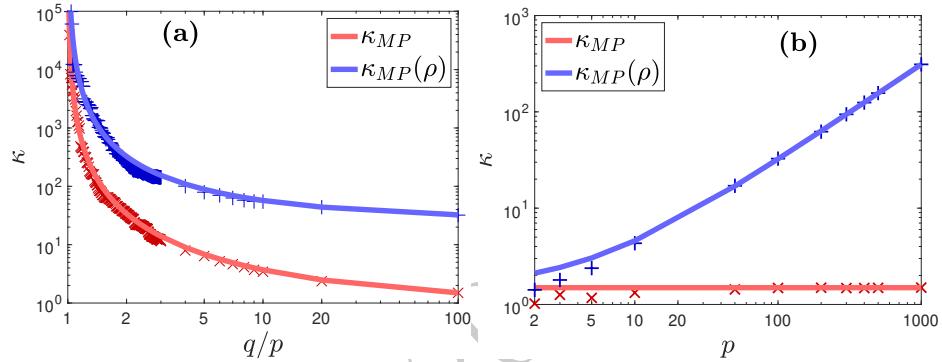


Figure 15.3 Symbols: condition numbers for covariance matrices from $p = 100$ random variables both uncorrelated and correlated ($\rho = 0.2$) computed from samples of different sizes from $q = 100$ to $q = 10,000$ (+ uncorrelated and \times correlated). Lines: theoretical approximations for the condition numbers from Eqs. 15.34 and 15.35. Panel (a) reports the condition numbers vs. p/q with $p = 100$, while panel (b) reports the condition numbers vs. p with $q = 100p$.

$p = 100$ random, correlated multivariate normal variables with zero means, unitary variance, equal correlation $\rho = 0.2$, and sizes from $q = 100$ to 10,000. In Fig. 15.3(a) I compare this condition numbers with the formula for κ_{MP} (Eq. 15.35). One can note the considerable difference in the results for the condition number with respect to the uncorrelated case, and also the excellent agreement with the theoretical results.

15.7.2 Dimensionality reduction

An intuitive solution for the curse of dimensionality would be to reduce the number of variables in the observation set to the minimum possible number selecting only the most informative ones. However, a priori, it is not possible to know which are the informative variables and which are the ones that can be discarded. Testing all possible combinations of variables is often unfeasible because the number of such tests scales as $\mathcal{O}(p!)$, whereas, attempting to reduce dimensionality a-posteriori incurs the same curse of dimensionality as the original problem. The literature categorizes dimensionality reduction procedures into two main approaches named as **feature selection** and **feature extraction**. The first approach selects a subset of the original features, while the second operates on a transformation of the original feature set producing a new feature set (of lower dimension). A selection of features can be done in many different ways, the literature refers to this as subset selection, and approaches are called wrappers, filters, and embedded methods (please refer to the comprehensive book James et al. [2013] for further reading).

Feature extraction can be performed by using an even wider number of methodologies. For instance, a very popular methodology is the principal component analysis (PCA, see Appendix C). Another valid way to explore the feature space is the random forest method (see Appendix D). I shall also introduce, later in this chapter in Section 15.9.4, a topological regularization approach that can drastically reduce dimensionality through a network-based construction (see also Section 17.3). The interested reader can start deepening the topic from James et al. [2013].

Remark 15.2. It is important to notice that the curse of dimensionality is not automatically solved by dimensionality reduction. Specifically, it can be healed by reducing the dimensionality before processing the data (i.e. by using feature selection) but if instead the data are transformed contextually with the dimensionality reduction, then whether the curse of dimensionality is healed or not depends on the process. For instance, PCA is not a solution to the curse of dimensionality problem. Indeed, the PCA method requires the estimation of the inverse covariance and therefore it can be used only when the covariance is well conditioned.

15.7.3 Estimation of the precision matrix

One of the most evident manifestations of the curse of dimensionality concerns the estimation of the precision matrix (inverse covariance matrix). Indeed, as I have illustrated in the previous Section, the covariance matrix becomes badly conditioned when p becomes large (see Figure 15.3). Let me here illustrate the consequences of such bad conditioning with two examples. In the first example, I quantify the distance between the estimated precision matrix and the true precision matrix by using the Frobenius norm.

Example 15.7 (Effect of sample size on covariance estimate: relative Frobenius norm measure). Let me define the following relative Frobenius norm

$$r_1 = \frac{\|\hat{\Sigma} - \Sigma\|_2}{\|\Sigma\|_2}. \quad (15.36)$$

This is a quantity that is equal to zero if the estimate covariance matrix, $\hat{\Sigma}$ coincides with the true covariance matrix Σ and, instead, it becomes large when their coefficients differ. A value of r_1 around one indicates that on average the difference between the coefficients in the two covariance matrices is of the same order of the values of the coefficients of the true matrix themselves. Analogously, for the inverse I introduce the following

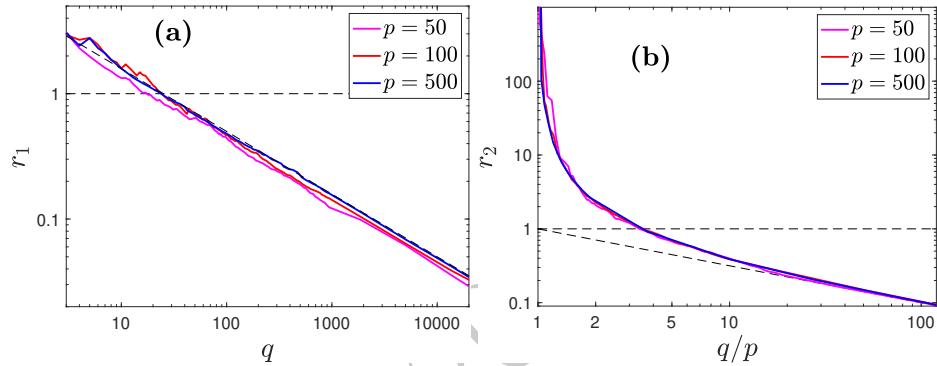


Figure 15.4 (a) r_1 : relative Frobenius norm of the difference between the true covariance and its sample estimate vs. sample size q . One can observe that, in this example, all results follow well the law $r_1 \sim c/\sqrt{q}$, with $c \simeq 5$. (b) r_2 : relative Frobenius norm of the difference between the true precision matrix and the inverse of the sample covariance estimate vs. relative sample size q/p . One can observe that, in this example, asymptotically, all results converge to the law $r_2 \sim \sqrt{p/q}$. (Data: artificially generated, correlated ($\rho = 0.2$), multivariate normally distributed variables with various p and q , see Example 15.7).

relative Frobenius norm

$$r_2 = \frac{\|\hat{\Sigma}^{-1} - \Sigma^{-1}\|_2}{\|\Sigma^{-1}\|_2}. \quad (15.37)$$

These two measures are, of course, not independent, and it is known (see Wilkinson [1971, 1994]) that the relative distances of the inverse and of the direct are related by

$$r_2 \leq \kappa r_1, \quad (15.38)$$

with κ the condition number (see Definition 15.4). In Fig. 15.4 I report r_1 and r_2 computed from $p = 50, 100$ and 500 random, correlated normally distributed, variables with zero mean, unitary variance, and sample sizes from $q = 100$ to $10,000$. The variables are all correlated with equal correlation $\rho = 0.2$ (as in Section 15.7.1). Let me first observe from Fig. 15.4(a) that the results for r_1 are independent of the dimension p . This indeed, follows intuition: from the law of large number, we know that each element of the covariance matrix will converge towards the true value for large sample sizes. Fig. 15.4(a) indeed shows that this convergence is well followed by the entire norm. One observes, for this example, that the relative Frobenius norm of the difference between the sample and true covariance matrices follows very well the relation

$$\hat{r}_1 \simeq \frac{c}{\sqrt{q}} \quad (15.39)$$

with $c \simeq 5$. One can also observe that $r_1 > 1$ until $q \sim 10$ and then it becomes significantly smaller than one for $q > 100$. This is indeed consistent with the discussion in Example 15.3 and Fig.15.2 on the significance of the correlation coefficients.

For the inverse case, I find instead that r_2 strongly depends on the dimension p and it scales with q/p . This is illustrated in Fig.15.4(b). Let me stress that these results for the precision matrix, r_2 , are very different with respect to the ones for the covariance, r_1 . First, r_2 diverges when $q/p \rightarrow 1$. This divergence at $q = p$ is a consequence of the fact that the sample covariance becomes not invertible when $q < p$ and $\|\hat{\Sigma}^{-1}\|_2 \rightarrow \infty$ when $q \rightarrow p^+$. Second, one can observe that r_2 decreases with q/p reaching an asymptotic regime well described by

$$r_2 \simeq \sqrt{\frac{p}{q}}. \quad (15.40)$$

One can observe that r_2 starts having values below 1 after $q/p \sim 4$; meaning that, for this example, until the number of observations is larger than four times the number of variables the estimate of the inverse covariance has, on average, errors that are larger in amplitude than the true values of the coefficients themselves. Results are similar for a range of values of the correlation and sample size with correlation reducing the errors while sample size incensing them.

The convergence of the coefficients of the sample covariance to the true covariance is an unavoidable consequence of the law of large numbers (see Section 13.3.2). The central limit theorem (see Section 5.1) also guarantees that, if the fourth moment is defined, deviations must converge to zero as $1/\sqrt{q}$. This is confirmed with this Example, where indeed, in Figure 15.4 we observe that both the measures r_1 and r_2 follow the $1/\sqrt{q}$ convergence law. Note that, for this law of convergence, in order to use the central limit theorem, one needs a definite fourth moment because here one is looking at the statistics of the second moments. Often, in many practical cases, when fat-tails with $2 < \alpha < 4$ are present, the fourth moments are not defined and such a convergence law is in $q^{1-1/\alpha}$ (see Section 13.4).

The quantification of the distance between two matrices is a non-trivial matter and the various available methods, including the Frobenius norm, just explored in Example 15.7, have their shortcomings. Let me, therefore, look at the same problem as in the previous example but use as a distance measure the Kullback-Leibler divergence between the population and the estimated distributions. This is a measure of the goodness of the estimate. Indeed, if the estimate is exact (i.e. $\hat{\Sigma} = \Sigma$), then the Kullback-Leibler divergence between the true and the estimated distribution must be zero, while, instead, the distance becomes larger

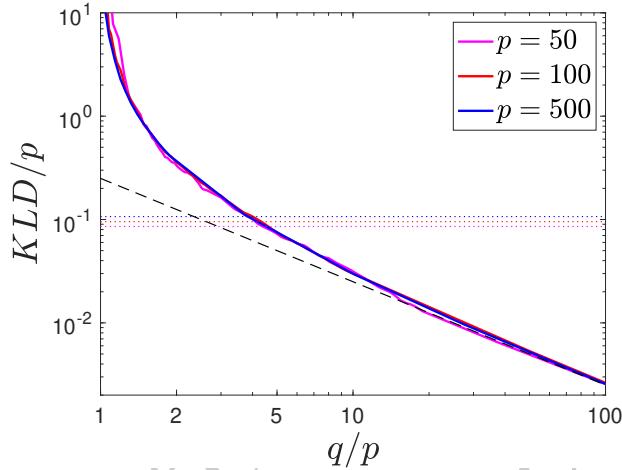


Figure 15.5 Kullback–Leibler divergence per unit of dimension, between the true probability density function and the probability density function estimated from the sample covariance and sample mean. The horizontal dotted lines are the results of a model which assumes uncorrelated variables. The dashed line is instead the law $0.251/q$ which appears to reproduce well the asymptotic behavior of KDL/p . (Data: artificially generated, correlated ($\rho = 0.2$), multivariate normally distributed variables with various p and q , same as for Fig.15.4).

for poorer estimates. This measure of distance between the true and the estimated covariance is meaningful and interpretable only for multivariate normal modeling. However, it can anyway be used in all cases when the covariances are defined and invertible.

Example 15.8 (Effect of sample size on covariance estimate: Kullback–Leibler divergence). The expression for the Kullback–Leibler divergence between the population and the estimated distributions can be computed from Eq.7.33 (see Section 7.4.1).

$$KLD = \frac{1}{2} \left(Tr(\hat{\Sigma}^{-1} \Sigma) - \log(|\hat{\Sigma}^{-1}| |\Sigma|) + (\hat{\mu} - \mu)^\top \hat{\Sigma}^{-1} (\hat{\mu} - \mu) - p \right). \quad (15.41)$$

(Notice that $|\hat{\Sigma}^{-1}| = 1/|\hat{\Sigma}|$.) It should be quite evident from this equation, that one expects this quantity to increase linearly with dimension $KLD \propto p$.

I have computed this measure for the same dataset as in Example 15.7. In Figure 15.5 I report KLD/p for various q and three different values of p . One can observe that indeed the Kullback–Leibler divergence scales with dimension p and decreases with the sample size q . From the figure one

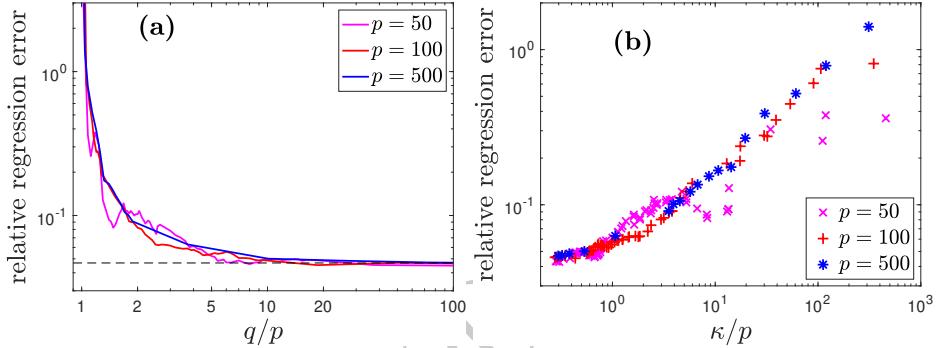


Figure 15.6 (a) Out-of-sample (test set) relative regression error $\text{Var}(\hat{y}^{\text{test}} - \tilde{y}^{\text{test}})/\text{Var}(Y)$ vs. relative sample size q/p for the model $Y = \beta^\top \mathbf{X} + \epsilon$, (parameters are computed on the train set and error on the test set). (b) The same out-of-sample relative regression error plotted versus the condition number per unit of dimension, κ/p , of $\Sigma_{\mathbf{XX}}$. (Data: artificially generated, correlated ($\rho = 0.2$), multivariate normally distributed variables with various p and q , same as for Fig.15.4).

notices that asymptotically results are well described by the law

$$KDL \simeq 0.25 \frac{p}{q}. \quad (15.42)$$

When compared with the Kullback–Leibler divergence between the true distribution and a distribution of uncorrelated variables ($\hat{\Sigma}$ being replaced with a matrix with all off-diagonal elements equal to zero), one observes that the distance from the sample estimates becomes lower than the uncorrelated one in the region of q/p larger than 5. This indicates, that for this example, one needs an observation set of the size of five times or more the number of variables to start getting estimations of the correlation structure of the probability density function that is more significant than assuming an uncorrelated system. This observation is fully consistent with the results in Fig.15.4 for the relative Frobenius distance of the inverse covariance (the r_2 measure) which indeed is reaching error levels below the signal only after $q/p \sim 4$.

Let me conclude this example by noticing that, the KLD between the true model and its estimate is the expected value of the negative log-likelihood plus a constant, $KLD = \bar{\ell}_\infty - \mathbb{E}(\ell)$, where – not incidentally – the constant $\bar{\ell}_\infty$ offsets the KLD value to zero when the estimated model coincides with the true one (e.g. in the asymptotic limit when $q \rightarrow \infty$ and $\hat{\Sigma} \rightarrow \Sigma$).

As a measure of the goodness of estimation, more than a generic distance between the true and the estimated covariances (or inverse covariances), one would like to quantify the effect of the estimate on the performances of a practically

relevant problem, such as linear regression. This is illustrated in the following example.

Example 15.9 (Effect of the covariance estimate on multilinear regression). Let me consider the multilinear regression of a random variable Y with respect to a set of p random variables \mathbf{X}

$$Y = \boldsymbol{\beta}^\top \mathbf{X} + \epsilon. \quad (15.43)$$

Quantifying the error in this operation which is a consequence of the uncertainty on the estimation of the covariance is, of course, of great practical relevance. I discussed in Section 8.4 that, the parameters that minimize the variance of the error ϵ are

$$\boldsymbol{\beta}^\top = \boldsymbol{\Sigma}_{Y\mathbf{X}} \boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}. \quad (15.44)$$

Supposing, I estimate $\boldsymbol{\beta}$ from the sample means over a set of q observations, I want to quantify the effect of the uncertainty on the estimation of the covariances on the precision of the regression results.

To this purpose, from the dataset used in the previous two examples, I generate a variable $Y = \mathbf{b}^\top \mathbf{X} + \eta$ with \mathbf{b}^\top a random vector of coefficients with $\|\mathbf{b}^\top\|_2 = 1$ and η a normal noise term with variance 0.1 (recall that X_i have unitary variance, therefore, the noise term contributes about 10% to the variance of the signal). I produce multivariate artificial datasets $\hat{\mathbf{x}}$ of various sizes and compute the associated variables $\hat{y}\mathbf{b}^\top \hat{\mathbf{x}} + \eta$. I divide each dataset $(\hat{y}, \hat{\mathbf{x}})$ into an in-sample training set, and an out-of-sample testing part both of size $q.a$ Using the training part $(\hat{y}^{train}, \hat{\mathbf{x}}^{train})$, I estimate the best-fit regression coefficients from the sample covariances

$$\hat{\boldsymbol{\beta}}^\top = \hat{\boldsymbol{\Sigma}}_{Y\mathbf{X}} \hat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{-1}. \quad (15.45)$$

I then use these parameters to perform the linear regression for the test data

$$\tilde{y}^{test} = \hat{\boldsymbol{\beta}}^\top \hat{\mathbf{x}}^{test}, \quad (15.46)$$

and I measure the difference between the true values of \hat{y}^{test} and their estimate from regression \tilde{y}^{test} . This is a measure of the effect of uncertainty of the sample estimations of the covariances. Indeed, if I could use the exact regressions coefficients $\boldsymbol{\beta}$ from the true covariances, then $Var(\hat{y}^{test} - \tilde{y}^{test}) = Var(\eta)$, instead if I use the estimate coefficients $\hat{\boldsymbol{\beta}}$ I expect $Var(\hat{y}^{test} - \tilde{y}^{test}) > Var(\eta)$.

Figure 15.6(a) reports $Var(\hat{y}^{test} - \tilde{y}^{test})/Var(Y)$ as a function of relative size q/p . One can notice that when the number of observations q becomes considerably larger the number of variables p the relative error becomes near to the minimum expected, which is $Var(\eta)/Var(Y)$.

Previously in this section, I have discussed the relation between the condition number and the goodness of the estimate. Intuitively, one would expect

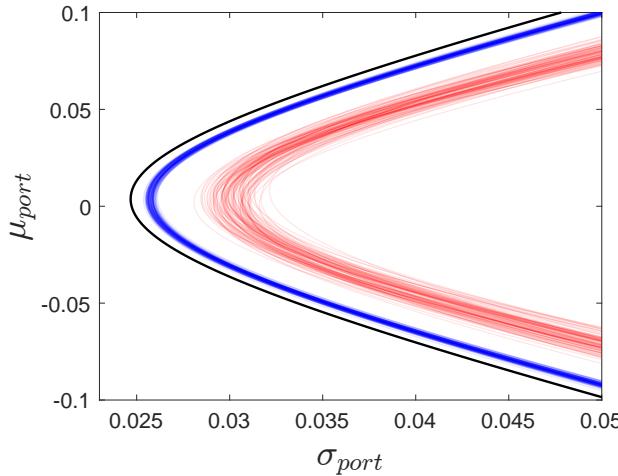


Figure 15.7 Markowitz efficient frontiers were obtained using two different estimates of the covariance and mean for a portfolio with $p = 100$. The continuous blue lines are Markowitz's efficient frontiers for 100 datasets with $q = 1, 200$; while the red lines refer to the case for $q = 300$. The fact that the red lines are on the right side of the blue ones indicates that the estimate of the covariance with a lower number of observations is less efficient. The black line is the theoretical efficient frontier obtainable with the exact parameters. (Data: set of $p = 100$ artificially generated, correlated, multivariate normal variables, see Example 15.10).)

that larger condition numbers must correspond to poorer regression performances. This is indeed exemplified in Figure 15.6(b) where it shows that the relative regression error increases with the condition number scaling as κ/p .

^a While the size of the train dataset is crucial because the precision of the estimate of sample covariances increases with $1/q$, the size of the test dataset is essentially irrelevant to our analysis. I however keep the two of the same size for simplicity.

Let me conclude this excursus around the sample estimate of the covariance matrix by looking at a specific application. One of the cornerstones in finance is the Markowitz portfolio construction that I have introduced in Example 8.3. In the Markowitz approach, the optimal weights (Eq.8.56), are directly obtained from the inverse covariance. Therefore, this is a practical instance where the estimate of the inverse covariance matrix has direct consequences on performances. Let me dive into this with the following example.

Example 15.10 (Effect of the covariance estimate on portfolio performances). In the Markowitz portfolio construction the optimal weights are

given by:

$$\hat{\mathbf{w}} = \hat{\mathbf{J}} \left(\lambda \hat{\boldsymbol{\mu}} + \frac{1 - \lambda \mathbf{1}^\top \hat{\mathbf{J}} \hat{\boldsymbol{\mu}}}{\mathbf{1}^\top \hat{\mathbf{J}} \mathbf{1}} \mathbf{1} \right), \quad (15.47)$$

(see Equation 8.56). Where I denoted with $\hat{\boldsymbol{\mu}}$ the estimate of the mean returns and with $\hat{\mathbf{J}} = \hat{\boldsymbol{\Sigma}}^{-1}$ the inverse of the covariance estimate (i.e. the estimate of the precision matrix). Note that the covariance is estimated from the observation set. This is – unavoidably – an ‘in-sample’ estimate. As discussed in Example 8.3, the parameter λ is a Lagrange multiplier associated with the mean portfolio return. The portfolio’s expected return is:

$$\mu_{port} = \hat{\mathbf{w}}^\top \boldsymbol{\mu}. \quad (15.48)$$

And the portfolio variance is:

$$\sigma_{port}^2 = \hat{\mathbf{w}}^\top \boldsymbol{\Sigma} \hat{\mathbf{w}}. \quad (15.49)$$

the parameter λ in Eq.15.47 can be treated as a free parameter: by varying $\lambda \in (-\infty, +\infty)$ one explores the constrained space of all best μ_{port} and σ_{port} which is the so-called ‘efficient frontier’. The case $\lambda = 0$ is the unconstrained case and corresponds to the ‘minimum variance’ portfolio.

To quantify the effect of uncertainty on the covariance estimation on Markowitz’s efficient frontier, I generated 100 correlated multivariate normal datasets with unitary standard deviation, with means distributed around zero and random correlation matrices, and with average absolute values of the correlation coefficients around 0.075. In Figure 15.7 I report the plots for the efficient frontiers (i.e. μ_{port} vs. σ_{port} from Eqs.15.48 and 15.49 for a range of λ) from two different estimates of the parameters ($\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$) for $p = 100$ and respectively with $q = 300$ and $q = 1,200$. It is evident from the figure that estimates with a smaller number of samples ($q = 300$, red lines) are further at the right side of the efficient frontier, corresponding to portfolios with lower average returns and higher average standard deviations. The best-performing portfolio (the efficient frontier reported on the left with the tick black line) is the one where the estimated means and covariance coincide with the population ones $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}$ and $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}$. Such a model is by definition also the one that maximizes likelihood.

Therefore, accordingly, with Markowitz portfolio construction, the best weights must be the ones obtained with the parameters that maximize the likelihood. Note that this is completely general. Indeed, both the Markowitz maximization procedure and the maximum likelihood argument do not use any assumption or hypothesis except that the first and second moments must be definite and both $\hat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}$ must be invertible. Such a generality is remarkable but also quite intuitive: the best model yields the best performances.

Figure 15.8 demonstrates that indeed models with larger likelihoods return

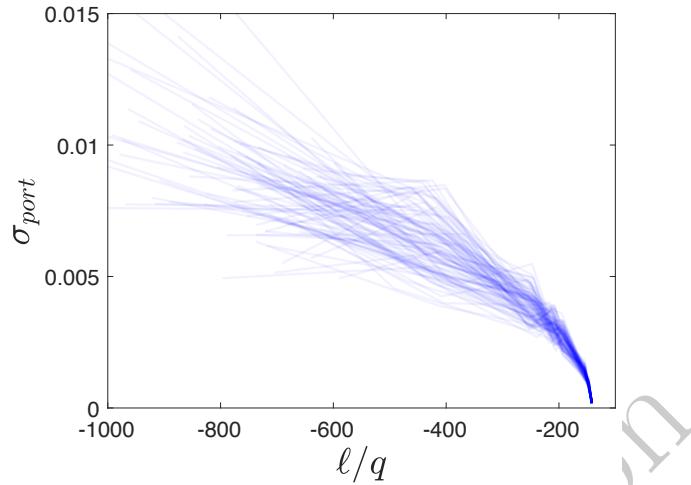


Figure 15.8 Portfolio standard deviation σ_{port} obtained from Markowitz's minimum variance portfolio ($\lambda = 0$) versus model log-likelihood per observation, ℓ/q . The blue lines report results over 100 different datasets. One sees that larger likelihoods are associated with better-performing portfolios which have smaller standard deviations. (Data: $p = 100$ artificially generated, correlated, multivariate normal variables – same as in Fig.15.7.)

better solutions for portfolio optimization. The example is for minimum variance portfolio ($\lambda = 0$), using multivariate normal likelihood (see Definition 15.1). The computation is, as before (see Fig.15.7), over 100 artificial correlated datasets with different sizes from $q = 150$ to $q = 20,000$ for a model with $p = 100$.

15.8 Shrinkage estimation of the covariance matrix

I have discussed in the previous Section that Pearson's covariance estimate can produce poor estimates when the number of observations, q , is small and the number of variables, p , is large. The computation of the inverse covariance becomes even impossible when $q \leq p$. It has been proposed that instead of using the sample covariance $\hat{\Sigma}$ in many cases it is more convenient to use the following covariance

$$\Sigma_{Shrink} = (1 - \delta)\hat{\Sigma} + \delta\mathbf{T}, \quad (15.50)$$

where \mathbf{T} is a ‘target’ covariance matrix and $\delta \in [0, 1]$ is a scalar parameter. This method was called shrinkage because the sample matrix is ‘shrunk’ towards the target and, in doing so, some values of the matrix elements that are too high or too low with respect to the true values are shrunk towards less extreme target values.

The target matrix must have the property of being a positive definite, invertible

covariance matrix. For $\delta = 0$ the shrunk matrix coincides with the sample matrix whereas for $\delta = 1$ it coincides with the target. Being the sum of two covariance matrices Σ_{Shrink} is also a covariance matrix and, given that the target matrix is invertible, then it is also always invertible for any $\delta > 0$, even when the sample covariance is not (i.e. it is a squared positively definite matrix).

Remark 15.3. When $q > p$ and there are no variables that are linear combinations of the others, then the sample covariance matrix is positively definite, which means that its eigenvalues are all larger than zero and the covariance matrix is invertible. However, when $q < p$ then $p - q$ eigenvalues become equal to zero and the matrix is no longer invertible. The addition of a vector of positive values to the diagonal makes the matrix become again positive-definite and invertible. Indeed, these values directly add to the sample covariance matrix eigenvalues. More generally, any sample covariance matrix can be inverted if added to a positive definite matrix.

Several possible target matrices have been proposed. Let me list here below a few target matrix types that have been used and are among the most relevant:

1. The simplest choice is the identity matrix with ones on the diagonal and zeros elsewhere. This makes the target an uncorrelated system of variables with unitary variance. This however can be inappropriate when variables have different variances. Indeed, for a given δ the shrinkage will be proportionally larger for variables with lower variance.
2. A matrix with the sample variances on the diagonal and zeros elsewhere. This corresponds to an uncorrelated system of variables with variances equal to the sample ones.
3. A matrix with ones on the diagonal and constant covariance off-diagonal.
4. A matrix with the sample variances on the diagonal and products of the sample standard deviations elsewhere.
5. Any positively-definite square $p \times p$ matrix that contains information about known structural features of the system.

The choice of the shrinkage constant δ is very important. Ledoit and Wolf in Ledoit and Wolf [2003, 2004] derived a formula for the optimal choice of δ . However, this is a hyper-parameter and it should be chosen from the out-of-sample performances using cross-validation as we shall see shortly in Chapter 18.

It has been shown by many authors that the shrunk estimate of the covariance, with the right shrinkage constant and the right target matrix, performs better than the sample covariance in a large range of practical applications and, in many cases, any shrinkage with small δ improves results.

15.9 Regularization

If a model is complex enough, it can adapt to the dataset to any degree of perfection providing therefore a perfect (in sample) ‘fit’ of the data. This is called an interpolating model. However, when such a model is tested on a different dataset (out-of-sample) it often tends to perform less well because it ‘overfits’ the data, modeling the noise instead of the signal (see Section 3.5 for further considerations about model training and overfitting). One way to reduce overfitting is to keep the model simpler and to reduce the range of variation of its parameters making it less able to adapt to the train, in-sample, dataset. Regularization is one of the main tools to reduce overfitting.

In most instances, regularization is achieved through penalization by adding to the loss function a regularization function that penalizes models that are too complex or/and have too large parameter values. In a formula, for a model \mathcal{M} with parameters $\boldsymbol{\theta}$, given a dataset $\hat{\mathbf{x}}$, one adds to the loss function $L(\mathcal{M}, \hat{\mathbf{x}}, \boldsymbol{\theta})$ a penalizing term $R(\mathcal{M}, \hat{\mathbf{x}}, \boldsymbol{\theta})^2$ ²

$$\operatorname{argmin}_{\boldsymbol{\theta}} (L(\mathcal{M}, \hat{\mathbf{x}}, \boldsymbol{\theta}) + \lambda R(\mathcal{M}, \boldsymbol{\theta})), \quad (15.51)$$

where $R(\mathcal{M}, \boldsymbol{\theta})$ is generically a function of the model \mathcal{M} and its parameters $\boldsymbol{\theta}$; λ is a hyper-parameter that can be used to tune the amount of regularization. The regularizer hyper-parameter λ can be a vector and can have different values for regularization for the different coefficients or groups of coefficients. This procedure is quite general and it is referred to as the penalty method for constrained optimization Bertsekas [1982].

When the model is the joint probability distribution function for the system’s variables, $\tilde{f}(\hat{\mathbf{x}}, \boldsymbol{\theta})$, then the regularization is a penalization term which is added to the negative log-likelihood

$$\operatorname{argmin}_{\boldsymbol{\theta}} \left(-\log \mathcal{L}(\tilde{f}(\hat{\mathbf{x}}, \boldsymbol{\theta})) + \lambda R(\tilde{f}, \boldsymbol{\theta}) \right) \quad (15.52)$$

where the minus sign in front of the likelihood term is because the likelihood must be maximized.

In this case, one could adopt a Bayesian perspective deriving the regularization term from the prior probability for the model and its parameters. Indeed, one aims to maximize the probability of the model and its coefficients for a given observation set: $\operatorname{argmax}_{\boldsymbol{\theta}} P(\tilde{f}, \boldsymbol{\theta} | \hat{\mathbf{x}})$. From Bayes’ formula, this probability is proportional to the product between the likelihood, $P(\hat{\mathbf{x}} | \tilde{f}, \boldsymbol{\theta})$, and the prior, $P(\tilde{f}, \boldsymbol{\theta})$,

$$P(\tilde{f}, \boldsymbol{\theta} | \hat{\mathbf{x}}) = P(\hat{\mathbf{x}} | \tilde{f}, \boldsymbol{\theta}) P(\tilde{f}, \boldsymbol{\theta}). \quad (15.53)$$

Taking the logarithm, this leads to the sum of the log-likelihoods of each

² The model \mathcal{M} can be either a probability distribution function $\tilde{f}(\hat{\mathbf{x}}, \boldsymbol{\theta})$ or a regression/classification function $g(\hat{\mathbf{x}}, \boldsymbol{\theta})$.

observation, $\log \mathcal{L}(\tilde{f}(\hat{\mathbf{x}}, \boldsymbol{\theta}))$, with the log of the prior probability that is, therefore, the regularizer term:

$$\lambda R(\tilde{f}, \boldsymbol{\theta}) = -\log P(\tilde{f}, \boldsymbol{\theta}). \quad (15.54)$$

This point is expanded in Section 18.7 where criteria to select between models are discussed.

It is clear from this derivation that by knowing the exact form of the regularizer and its coefficient, λ , one would be able to retrieve the exact model. However, priors are in general unknown and so are the regularizers. One therefore normally proceeds by guessing, assuming some standard regularizer's form. For instance, a prior with the normal distribution of the coefficients leads to a penalization with the sum of the squares, which is known as L_2 -norm regularization that I shall discuss in the next session.

When the model is a regression or a classification problem one can adopt the same perspective. Indeed, regression/classification and dependency in multivariate probabilistic modeling are intimately related problems and the search for the best regressor/classification function $Y = g(\mathbf{X}, \boldsymbol{\theta})$ is directly related to the search for the best joint probability distribution function for the system's variables, $\tilde{f}(Y, \mathbf{X}, \boldsymbol{\theta})$. For instance, the minimization of the squared error is achieved by $g(\mathbf{X}) = \mathbb{E}(Y|\mathbf{X})$.

15.9.1 L_2 -norm regularization: ridge regression

The L_2 -norm regularization consists in finding the coefficient that minimizes the loss function, $L(\mathcal{M}, \hat{\mathbf{x}}, \boldsymbol{\theta})$, while penalizing it with the sum of the square of the coefficients:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}} (L(\mathcal{M}, \hat{\mathbf{x}}, \boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_2^2), \quad (15.55)$$

imposing therefore $R(\mathcal{M}, \boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2$.

The effect of the L_2 -norm regularization is to reduce the size of the coefficients by ‘shrinking’ them, reducing in this way overfitting. The L_2 -norm is differentiable and therefore sometimes the minimization can be found analytically and otherwise, it can be discovered by gradient descent with numerical methods.

Historically, the L_2 -norm regularization was first proposed by Andrey Niko layevich Tikhonov et al. [1995] who proposed it as a regularizer for model functions written in the form:

$$g(\mathbf{X}, \boldsymbol{\theta}) = g(\mathbf{W}\mathbf{X}, \boldsymbol{\theta}). \quad (15.56)$$

This is a general and meaningful way to write the model function because it emphasizes the role of the coefficients in relation to the observations and it has a direct interpretation in linear regression.³ Indeed, by writing the model function

³ $\mathbf{W}\mathbf{X}$ can be interpreted as a linear feature extraction term. In artificial neural networks, this is a

in terms of the product \mathbf{WX} , one represents explicitly that larger parameters $(\mathbf{W})_{i,j}$ provide a relatively larger importance to the relationship between some couples variables (i, j) with respect to other couples. The regularizer is

$$R(\mathcal{M}, \boldsymbol{\theta}) = \|\mathbf{W}\|_2^2. \quad (15.57)$$

Note that there can be other parameters in the model that are not coupled with the variables and they are not included in this regularization.

Remark 15.4. A generalization of L_2 regularization can be performed for any function by taking the norm of the function:

$$R(g) = \|g\|_{\mathcal{H}}^2, \quad (15.58)$$

which is the inner product $\langle g, g \rangle_{\mathcal{H}}^{1/2}$ in the reproducing kernel Hilbert space \mathcal{H} Young [1988].

Maximum likelihood with L_2 -norm regularization for multi-normal distribution

Let me deepen into the L_2 -norm Tikhonov regularization by applying it to modeling with the multi-normal distribution. To this purpose, it is convenient to write the log-likelihood of the multi-normal model in terms of \mathbf{WX} . Let me, indeed, assume the inverse covariance \mathbf{J} being expressed as the product

$$\mathbf{J} = \mathbf{W}^\top \mathbf{W} \quad (15.59)$$

with \mathbf{W} a real $p \times p$ matrix.

Notice that this decomposition of $\mathbf{J} = \boldsymbol{\Sigma}^{-1}$ is always possible as far as the covariance is positive definite. Indeed, \mathbf{W} can be constructed with columns (n) containing the coefficients of the eigenvectors $\mathbf{w}^{(n)}$ of $\boldsymbol{\Sigma}$ a

$$(\mathbf{J})_{i,j} = (\mathbf{W}\mathbf{W}^\top)_{i,j} = \sum_{n=1}^p w_i^{(n)} w_j^{(n)} / e_n. \quad (15.60)$$

From this expression, one might notice, on passing, that the smallest eigenvalues give larger weights to the elements of the inverse covariance \mathbf{J} .

^a The eigenvectors $\mathbf{w}^{(n)}$ satisfy the equation $\boldsymbol{\Sigma}\mathbf{w}^{(n)} = e_n \mathbf{w}^{(n)}$, when $\boldsymbol{\Sigma}$ is positive definite and therefore invertible. Note that the inverse matrix $(\mathbf{J} = \boldsymbol{\Sigma}^{-1})$ has the same eigenvectors with reciprocal eigenvalues divided by the square root of the associated eigenvalue e_n : $W_{i,n} = w_i^{(n)} / \sqrt{e_n}$.

The L_2 -norm regularizer takes the form

$$R(\tilde{f}, \boldsymbol{\theta}) = \|\mathbf{W}\|_2^2 = \text{Tr}(\mathbf{W}\mathbf{W}^\top), \quad (15.61)$$

convolutional layer. The model function $g(\mathbf{X}, \boldsymbol{\theta})$ is normally a regression/convolution, however, it can also be a probability density function.

and Tikhonov's regularized log-likelihood for the multivariate normal distribution is

$$\tilde{\ell}(\boldsymbol{\theta}|\hat{\mathbf{x}}) = \frac{q}{2} \log |\mathbf{W}\mathbf{W}^\top| - \frac{1}{2} \sum_{s=1}^q (\hat{\mathbf{x}}_s - \boldsymbol{\mu})^\top \mathbf{W}\mathbf{W}^\top (\hat{\mathbf{x}}_s - \boldsymbol{\mu}) - \lambda \text{Tr}(\mathbf{W}\mathbf{W}^\top). \quad (15.62)$$

The minus sign in front of the regularization parameter λ accounts for the fact that the likelihood is maximized and therefore the penalizations must reduce it. Substituting back $\mathbf{J} = \mathbf{W}^\top \mathbf{W}$ it reads

$$\tilde{\ell}(\boldsymbol{\theta}|\hat{\mathbf{x}}) = \frac{q}{2} \log |\mathbf{J}| - \frac{1}{2} \sum_{s=1}^q (\hat{\mathbf{x}}_s - \boldsymbol{\mu})^\top \mathbf{J} (\hat{\mathbf{x}}_s - \boldsymbol{\mu}) - \lambda \text{Tr}(\mathbf{J}). \quad (15.63)$$

Notice that this regularization is the L_2 -norm of Tikhonov's log-likelihood maximized in terms of \mathbf{W} . The regularizer term is $\text{Tr}(\mathbf{J})$ and not $\|\mathbf{J}\|_2^2$, which would give a different solution.

By differentiating with respect to \mathbf{J} , and using the identities

$$\frac{\partial \log |\mathbf{J}|}{\partial \mathbf{J}} = \mathbf{J}^{-1} = \boldsymbol{\Sigma}, \quad (15.64)$$

and

$$\frac{\partial \text{Tr} \mathbf{J}}{\partial \mathbf{J}} = \mathbf{I}, \quad (15.65)$$

with \mathbf{I} the $p \times p$ identity matrix, I obtain

$$\frac{\partial \ell(\boldsymbol{\theta}|\hat{\mathbf{x}})}{\partial \mathbf{J}} = \frac{q}{2} \boldsymbol{\Sigma} - \frac{q}{2} \hat{\boldsymbol{\Sigma}} - \lambda \mathbf{I}, \quad (15.66)$$

Equalling to zero I obtain the solution

$$\boldsymbol{\Sigma} = \hat{\boldsymbol{\Sigma}} + \lambda \mathbf{I}. \quad (15.67)$$

One can verify that for any $\lambda > 0$ the estimated covariance matrix $\boldsymbol{\Sigma}$ is positive definite and therefore the original assumption of it being invertible is satisfied.

One can also recognize that this is the same expression for the estimate of the covariance proposed by the shrinkage method except for a missing term $1 - \lambda$ in front of $\hat{\boldsymbol{\Sigma}}$. The effect is the same, namely by adding λ to the diagonal the sample covariance matrix is shrunk towards a target identity matrix. Other target matrices can be equivalently obtained by using local regularizer parameters. Namely penalizing with $\boldsymbol{\Lambda} \mathbf{W}$ with $(\boldsymbol{\Lambda})_{i,j} = \lambda_{i,j}$ one can shrink towards any target matrix $\boldsymbol{\Lambda}^\top \boldsymbol{\Lambda}$.

Example 15.11 (Maximum likelihood with L_2 -norm regularization for multi-normal distribution). As an example, I applied this regularization to the multivariate normal modeling discussed in Example 15.8 and Fig.15.5. The changes in the estimate of the Kullback–Leibler divergence between the

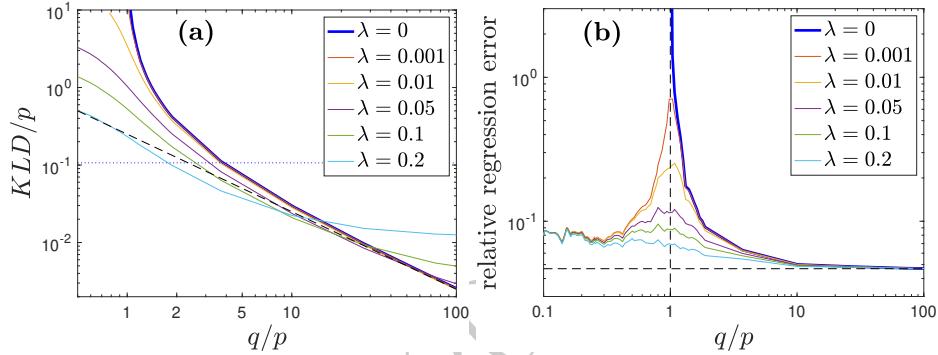


Figure 15.9 Comparison between performances of regularized models ($\lambda > 0$) with respect to non-regularized ($\lambda = 0$). The tick blue line corresponds to $\lambda = 0$ and it is the same as in Fig.15.5 for the case $p = 500$. The thinner lines are the solution with L_2 -norm regularization with $\lambda = 0.001, 0.01, 0.05, 0.1, 0.2$. **(a)** Kullback–Leibler divergence per variable between the true and the estimated distribution vs. relative sample size q/p . The horizontal dotted line is the KLD/p for an uncorrelated model with a diagonal covariance matrix. the dashed line is $0.251/q$ as in Fig.15.5. **(b)** Relative regression error vs. relative sample size q/p . The horizontal dashed line is the minimum achievable relative regression error. (Data: artificially generated, correlated, multivariate normal variables – same as Fig.15.5 – for $p = 500$.)

true and the estimated distribution are reported in Fig15.9(a) for the case $p = 500$ and various values of the regularizer parameter λ . One can notice that, differently from Fig.15.5, now there is no longer a divergence at $q \rightarrow p^+$. Indeed, the L_2 -norm regularized covariance (Eq.15.67) is now invertible also when $q < p$. However, one can notice that below $q \sim 2p$, the KLD is larger than the one for the uncorrelated model (horizontal dotted line in Fig.15.9(a)). One can observe that while the larger value of λ produces models with marginally lower KLD in the region of small q , then they produce models with larger KLD in the region of large q . Indeed, eventually, in the asymptotic limit, $q \rightarrow \infty$ the non-regularized estimate ($\lambda = 0$) will tend to the true covariance and must outperform the regularized one. It must be remarked that, in this procedure, I only used the construction $\mathbf{J} = \mathbf{W}^\top \mathbf{W}$ together with the L_2 -norm regularization. This implies that any model function that depends on $(\hat{\mathbf{x}}_k - \boldsymbol{\mu})^\top \mathbf{J} (\hat{\mathbf{x}}_k - \boldsymbol{\mu})$ can be regularized in this way. This automatically extends this procedure to the entire family of elliptical distributions when the shape matrix is positive definite.

Multivariate normal modeling and linear regression are strictly related and indeed the previous results for the L_2 -norm regularization for multi-normal distribution can be directly mapped into the regularization of multi-linear regression. This takes the name of Ridge regression.

Example 15.12 (Ridge regression). The multi-linear regression of a variable Y with respect to a set of p variables $\mathbf{X} = (X_1, \dots, X_p)^\top$ has the form

$$Y = \boldsymbol{\beta}^\top \mathbf{X} + \epsilon \quad (15.68)$$

In this context, L_2 -norm regularization implies minimization of the mean square loss (see also Section 8.4) subject to a penalizer proportional to the sum of the square of the components of $\boldsymbol{\beta}$

$$\operatorname{argmin}_{\boldsymbol{\beta}} ((Y - \boldsymbol{\beta}^\top \mathbf{X})^2 + \lambda \|\boldsymbol{\beta}\|_2^2). \quad (15.69)$$

The minima can be computed analytically by differentiating with respect to $\boldsymbol{\beta}$ and equalling to zero obtaining

$$\boldsymbol{\beta} = (\hat{\Sigma}_{\mathbf{XX}} + \lambda \mathbf{I})^{-1} \Sigma_{\mathbf{XY}}. \quad (15.70)$$

That is the same solution as in the non-regularized case (see Equation 8.30) but where however the estimation of the covariance is no longer the sample covariance but instead the shrunk (or –indeed – the L_2 -norm regularized) one:

$$\hat{\Sigma}_{\mathbf{XX}} + \lambda \mathbf{I}, \quad (15.71)$$

which is indeed identical to Eq.15.67.

Recalling Example 15.9 one can verify that by using the above L_2 -norm penalized estimate of the covariance one gets better results in terms of the relative regression error and also one can still perform the regression even when $q < p$. This is shown in Figure 15.9(b), where the relative regression error $Var(\hat{y}^{test} - \tilde{y}^{test})/Var(Y)$ (see Example 15.9) computed on the out-of-sample test set is reported.

One can notice that for finite λ this regression error (sometimes called risk) is not monotonic. It is large for small training samples and then it decreases with training set size, however, it then increases (even diverges for $\lambda \rightarrow 0$) when q approaches p . This is the ‘interpolation threshold’ where the regression error in the train set goes to zero. Afterward, the out-of-sample error decreases again and asymptotically, for $q \rightarrow \infty$, reaches the theoretical minima. It was pointed out by Nakkiran [2019] that this is an instance of ‘double descent’ (see Section 3.5.4) showing that ‘More Data Can Hurt’. This kind of double descent has been named ‘sample-wise double descent’.

15.9.2 L_1 -norm regularization: LASSO

A logical extension of the L_2 -norm regularization is using the L_1 norm instead. Formally:

$$\operatorname{argmin}_{\boldsymbol{\theta}} (L(\mathcal{M}, \hat{\mathbf{x}}, \boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_1). \quad (15.72)$$

This regularization is often called LASSO (Least Absolute Shrinkage and Selection Operator) Santosa and Symes [1986], Tibshirani [1996].

Similarly to L_2 -norm, also L_1 -norm regularization has the effect of ‘shrinking’ the coefficients. However, in this case, the penalization with the sum of absolute values has the further effect to force the value of some coefficients to zero, effectively eliminating them from the problem generating in this way sparser models. This is referred to, in some literature, as ‘soft thresholding’ distinguishing it from ‘hard thresholding’ where some coefficients are put to zero while the others are left unchanged. The fact that some coefficients go to zero, can turn out to be extremely useful in many cases simplifying the model and making it easier to interpret. The L_1 -norm has the problem that it is not differentiable and this makes the discovery of the minimum harder. Nonetheless, there are efficient numerical strategies that can accomplish this task. One must however be mindful that the result is, in general, not deterministic yielding to different solutions every run.

The L_1 -norm regularization of the multilinear regression problem is

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} ((Y - \boldsymbol{\beta}^\top \mathbf{X})^2 + \lambda \|\boldsymbol{\beta}\|_1). \quad (15.73)$$

The effect of this regularization is to produce a vector of coefficients $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ with some entries equal to zero. This implies that some of the $\mathbf{X} = (X_1, \dots, X_p)^\top$ variables do not contribute to the regression reducing the dimensionality of the model and simplifying the interpretation of the results.

15.9.3 GLASSO: sparse inverse covariance with L_1 -norm regularization

For the multivariate normal problem, the L_1 -norm regularization can be formulated as:

$$\underset{\mathbf{J}}{\operatorname{argmin}} \left(-\log |\mathbf{J}| + \frac{1}{q} \sum_{s=1}^q (\hat{\mathbf{x}}_s - \boldsymbol{\mu})^\top \mathbf{J} (\hat{\mathbf{x}}_s - \boldsymbol{\mu}) + \lambda \|\mathbf{J}\|_1 \right). \quad (15.74)$$

One might note that $\frac{1}{q} \sum_{s=1}^q (\hat{\mathbf{x}}_s - \boldsymbol{\mu})^\top (\hat{\mathbf{x}}_s - \boldsymbol{\mu}) = \hat{\Sigma}$ and therefore the previous expression can be equivalently written as:

$$\underset{\mathbf{J}}{\operatorname{argmin}} \left(-\log |\mathbf{J}| + \hat{\Sigma} \mathbf{J} + \lambda \|\mathbf{J}\|_1 \right). \quad (15.75)$$

The effect of the L_1 -norm regularization in the multivariate normal case is to produce sparse precision matrices \mathbf{J} . This can provide insights into the structure of interdependence between the variables (see Section 8.5). This L_1 norm penalized maximization is called Graphical-LASSO or GLASSO Friedman et al. [2008].

The estimation of the sparse inverse GLASSO covariance can be obtained by the same procedure as for the L_1 -norm LASSO regression solution. Indeed, the two problems are strictly related.

A combination of L_1 and L_2 -norm regularization is often referred to as the elastic net.

15.9.4 L_0 -norm regularization: best subset selection

The LASSO L_1 -norm regularization is often used because it reduces problem dimensionality and it produces sparse problems providing easier interpretability. However, if the goal is to produce models with some of the parameters forced to zero, then the appropriate norm is the L_0 -norm that simply counts the number of non-zero elements. Formally:

$$\operatorname{argmin}_{\boldsymbol{\theta}} (L(\mathcal{M}, \hat{\mathbf{x}}, \boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_0). \quad (15.76)$$

This regularization is often called best subset selection, it penalizes models with more parameters while keeping instead the optimized value of the non-zero parameters unchanged. In this respect, it is not a ‘shrinkage’ procedure.

For very simple models with only a few parameters, this L_0 -norm regularization can be performed exhaustively by putting to zero a subset of the parameters at a time, exploring all configurations combinatorially. This however becomes rapidly uncomputable when the dimensionality of the problem becomes large (indeed for $p \sim 10$ the number of combinations is already several millions and when $p \sim 100$ it becomes much larger than the number of atoms in the universe). Furthermore, not all models can be kept consistent by putting an arbitrary selection of coefficients to zero. For instance, while the multilinear regression could be explored in this way (for small p); conversely, in the directly related problem of modeling with multivariate normal distributions, one cannot arbitrarily put to zero coefficients of the matrix \mathbf{J} because the precision matrix must be positive definite. For this purpose, the topological regularization procedure, presented in the next section, has been devised.

15.9.5 Topological regularization: sparse inverse covariance with LoGo

A regularization that does not use penalization, but rather constraints on the selection of non-zero parameters, is the topological regularization. I mentioned in the previous session that GLASSO provides a tool to sparsify the inverse covariance, however, it also introduces a shrinkage element. Furthermore, GLASSO can become computationally demanding, and in some circumstances, especially for very sparse models, numerical solutions become hard to obtain.

An alternative method to sparsify the inverse covariance without shrinkage consists in retaining different from zero only the coefficients of the inverse covariance that are associated with a given network representation. In Bayesian terms, one defines a prior network representation and then maximizes likelihood under such network representation constraint. I have shown in Chapter 4 how

these networks, named information filtering networks, can be conveniently constructed. For the purpose of regularization, it is sufficient to assume that this network is given and it has the property to be chordal (see Definition 4.11).

Such a sparsification is computationally fast and can be run in parallel. A combination of this regularization with L_2 regularization or a shrinkage towards some structured target matrix, such as the constant correlation matrix, can be implemented straightforwardly on the local estimates of the clique's and separators' covariances. It can be applied equivalently to any shape matrix for the elliptical distribution family. Furthermore, the expectation-maximization procedure can be applied to this local structure. For further details the interested reader can refer to Massara and Aste [2019], Aste [2020].

15.9.6 Early stopping

In machine learning, a form of regularization consists in ending the procedure for loss minimization (or gain/likelihood maximization) early, before the actual minimum is reached. This yields to worst results in the in-sample training set but it often results in better performances in the out-of-sample testing set.

How ‘early’ one must stop, is a form of hyper-parameter to be discovered with cross-validation.

The nature of these sub-optimal early solutions depends on the kind of process addressed. Generally speaking, if the optimization is a gradient-descent operation, one can argue from very general geometrical arguments that the optimization process spends almost all its evolution time on the boundaries of the basin of attraction of solutions. In non-linear cases when there is more than one solution this method could efficiently provide a way to explore the phase space at the boundaries between equivalent solutions providing hybrid models that are less overfitting and could be better performing in generalizing to datasets unseen before.

15.10 Data-driven modeling tutorial

<https://github.com/FinancialComputingUCL/DataDrivenModeling/Ch15>

The tutorial for this Chapter covers various topics on the estimation of multivariate probabilities from data, including: The estimation of correlations and covariance matrices from data, the effect of sample size on covariance estimate, the condition number (Examples 15.5, 15.6, and 15.7), the Ridge regression (Example 15.12) and regularizations.

Exercises

- Consider the two sets of observations $\hat{\mathbf{x}} = (1.01, -0.91, 1.31, 0.14, 1.46, 1.69, -0.34, 1.71, 2.00, 1.73)^\top$ and $\hat{\mathbf{x}} = (0.59, 0.80, -0.79, 2.28, -0.66, 2.24, 0.50, 1.79, 3.56, 2.40)^\top$.
 - i. Compute the sample covariance $\text{Cov}(X, Y)$.
 - ii. Compute the sample correlation $\text{Corr}(X, Y) = \hat{\rho}_{XY}$.
 - iii. Use the t-statistic test to estimate if the correlation is significantly different from zero.
 - iv. Discuss the consistency of this result with what is reported in Figure 15.2.
- Compute the condition number κ of the matrix:

$$\mathbf{B} = \begin{pmatrix} 1.89 & 0.59 & -0.45 & -0.04 \\ 0.59 & 0.95 & 0.17 & -0.26 \\ -0.45 & 0.17 & 0.41 & -0.10 \\ -0.04 & -0.26 & -0.10 & 0.57 \end{pmatrix}. \quad (15.77)$$

- Assuming $\mathbf{B} = \Sigma$ is the sample covariance matrix from $q = 10$ observations, compare the result for the condition number κ with the estimation from the Marčenko-Pastur formula for the max and min eigenvalues.
- Compute the weight for Markowitz's minimum variance optimal portfolio, using $\Sigma = \mathbf{B}$.
- Verify that $\tilde{\mathbf{B}} = \mathbf{B} + \lambda \mathbf{I}$ has smaller condition numbers for \mathbf{B} for any value of $\lambda > 0$.
- Using the LoGo topological regularization method, compute the sparse inverse of \mathbf{B} associated with a network with the adjacency matrix

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}. \quad (15.78)$$

Verify that it is positive definite.

16

Time series and probabilistic modeling of evolving processes at different time-scales

In this Chapter, I discuss how to quantify the properties of stochastic processes which evolve over time. In this case, the observation of the process during a given lapse of time produces a set of values that are recorded one after the other in an ordered sequence and take the name of ‘time series’.

Definition 16.1 (Time series). Consider a stochastic process X_t with index set $\mathbf{t} = (t_1, \dots, t_T)$ (see Definition 9.1); each observation of such multivariate set of random variables, $X_{\mathbf{t}} = (X_{t_1}, \dots, X_{t_T})$, is a time-ordered series of numbers

$$(\hat{x}_s)_{s \in \mathbf{t}} = (\hat{x}_{t_1}, \hat{x}_{t_2}, \dots, \hat{x}_{t_T}) \quad (16.1)$$

which is called **time series**. The variables at each point in time $t \in (t_1, \dots, t_T)$ can be multidimensional $\mathbf{X}_t \in \mathbb{R}^p$ and the process has therefore $p \times T$ dimensions, $\mathbf{X}_{\mathbf{t}} \in \mathbb{R}^{p \times T}$. The time observation $t \in (t_1, \dots, t_T)$ of variable $i \in (1, \dots, p)$ is denoted with $\hat{x}_{i,t}$.

When, as often is the case, the times are equally spaced, with unitary intervals, between two values $t_{min} < t_{max}$, then the time series can be also denoted as:

$$(\hat{\mathbf{x}}_s)_{s=t_{min}}^{t_{max}} = (\hat{\mathbf{x}}_{t_{min}}, \hat{\mathbf{x}}_{t_{min}+1}, \dots, \hat{\mathbf{x}}_{t_{max}}). \quad (16.2)$$

^a Notice the difference between bold \mathbf{t} that is a sequence and t that is an element of this set $t \in \mathbf{t}$

I'll first discuss how to quantify scaling properties, and then, from Section 16.4, how to quantify properties that evolve with time.

16.1 Estimation of scaling laws

Complex systems, such as financial markets, have dynamics that span several orders of magnitude. For instance, orders in the markets are executed at speeds that often reach the microsecond, but there are investors that plan investments for decades in the future. I have already discussed in Chapter 9 that the statistical properties of a system change with the scale and, in some cases, these changes follow known ‘scaling laws’. From Section 9.2 it should be clear that the general challenge is to estimate the laws of change in the probability density function of a

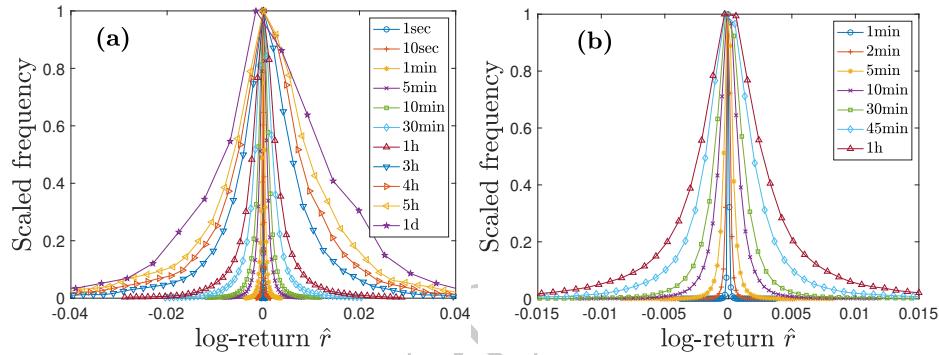


Figure 16.1 Relative frequencies for the distributions of log-returns ($\hat{r}_{t,s} = \log Price_t - \log Price_{t-s}$) at different horizons s , for the AAPL prices observed between 03/01/2017 and 24/08/2020 (see details in Example 16.1). The frequencies are divided by their maximum value in order to have all maxima at the value of one and make them all visible. (a) Time horizons between 1 second and 1 day (five orders of magnitude). (b) Detail for the smaller range of horizons between 1 minute and 1 hour.

random variable (the returns $R_{t,s} = X_t - X_{t-s}$) associated to a process (X_t) when it is observed at different time-scales (s). In other words, one seeks to identify the changes of the probability density function $f_{t,s}(R_{t,s})$ with the time-horizon s . The plots in Fig. 16.1 illustrates such changes for a practical example of log-returns of AAPL price. This is illustrated in the following Example.

Example 16.1 (Scaling of AAPL log-returns). I investigate the scaling laws for AAPL prices in the period between 09:30 of 03/01/2017 and 14:00 of 24/08/2020 with observations registered every second (a total of over 18 million observations). I computed the log-returns $\hat{r}_{t,s} = \log Price_t - \log Price_{t-s}$ at various time horizons s between one second to one day. Fig. 16.1 shows the relative frequencies for the AAPL log returns at various time-horizon s . By looking at this Figure it is clear that the log-return distribution broadens with s indicating that for larger horizons it is more likely to observe larger price variations and therefore larger log-returns. This is intuitive, indeed the longer the time-lapse is and the more the price can change. The most common measure for the width of a distribution is the standard deviation. Figure 16.2 reports, in log-log scale the increase with s of the standard deviation $\hat{\sigma}_s$ of the log returns $\hat{r}_{t,s}$ at horizon s . In this figure, the linear trend of the log-log plot indicates that these standard deviations are following the scaling law $\hat{\sigma}_s \propto \hat{\sigma}_1 s^{\hat{H}}$ which, notably, holds over a range of time horizons spanning more than five orders of magnitude. In this case, the exponent value, estimated by minimizing the mean square error of the linear fit $\log \hat{\sigma}_s = \hat{H} \log \hat{\sigma}_1 s + c$, is $\hat{H} = 0.497$. This is an estimate of the Hurst exponent (see Section 9.7.2) and, specifically of $H(2)$, the Hurst

exponent associated with the second moment of the distribution. Note that this value is very close to $H = 0.5$ which is expected for the Brownian motion.

If the distribution of $\hat{r}_{t,s}$ is broadening with the time horizon as $s^{\hat{H}}$, then the scaled signal, $s^{-\hat{H}} \hat{r}_{t,s}$, must have a similar distribution at all time-horizons. This is tested in Fig.16.3 where one can see that indeed the plots of Fig.16.1 now collapse into similar behaviors when scaled. This is particularly noticeable for time horizons in a range between 1 and 60 minutes where the plots are almost indistinguishable, as shown in Fig.16.3(b). This is an empirical validation of the scaling law Eq.9.21. However, despite the noticeable collapse of the plots, one can also observe that there are clear differences between the distributions (also in the 1-60 min range) and indeed two-sample Kolmogorov-Smirnov tests (see Massey Jr [1951] and Section 18.3.3) return as unlikely the hypotheses that these scaled signals belong to the same probability distribution. Indeed, the scaling law Eq.9.21 holds for self-affine, uniscaling processes. It is however known that financial log-returns are not uniscaling self-affine processes (see next Example 16.2) and that their distribution scale differently for the various moments and with the time horizons as well. Furthermore, it is quite evident Fig.16.3(a) that there are significant differences in the distributions both at small horizons below the minute and at horizons above one hour. This, together with the effects from autocorrelations and slow convergence towards stable distribution, reflects the different kinds of actors in the market that operate with different frequencies, depending on their own objectives.

From a data-driven perspective, the discovery, validation, and quantification of scaling laws is a very hard task. The main problem is typically data scarcity. Indeed, to estimate properly scaling properties one must estimate the probability distributions of a system over a broad range of time scales. Often, over these time scales non-stationarity and seasonality (cyclical changes) play very important roles and the datasets might not be directly usable for modeling. Another source of problems is that the scaling property, such as the Hurst exponent (H , see Sections 9.6 and 9.7), often depends on more than one factor. For instance, H is affected by both the autocorrelations and the tail exponent. Disentangling these factors is not trivial. A further problem arises from aggregation. I have argued in Section 9.5 that in uniscaling processes, such as the random walk, the scaling law is the law governing the scaling of stable probability distributions. However, this is true only asymptotically or in the special case when the noise term is a stable distribution. In systems such as financial markets typically one has probability distributions with finite second moments but with power law tails (with exponents $\alpha > 2$). These are not stable distributions because they tend to converge towards a normal distribution. However, in practice, the asymptotic limit is never reachable because the tails persist over all orders of aggregation (see Example 5.4).

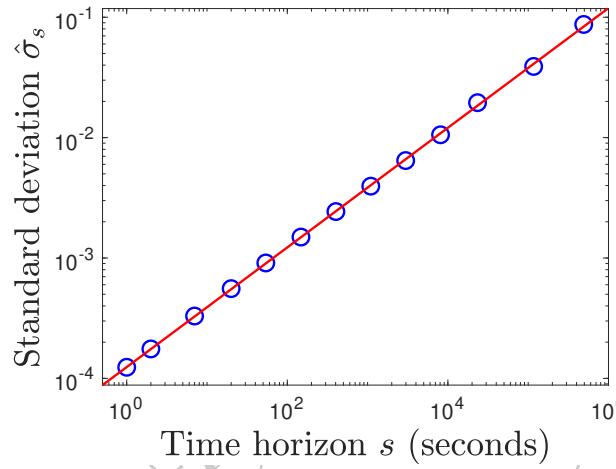


Figure 16.2 Standard deviations of the log returns vs. the time horizon s for the same data as in Fig.16.1. The red line is the best fit for the scaling law $\hat{\sigma}_s = cs^{\hat{H}}$ which gives exponent value $\hat{H} = 0.497$. Note that the linear behavior is across more than five orders of magnitude in time horizons and three orders of magnitude in standard deviation values. See Example 16.1 for further details.

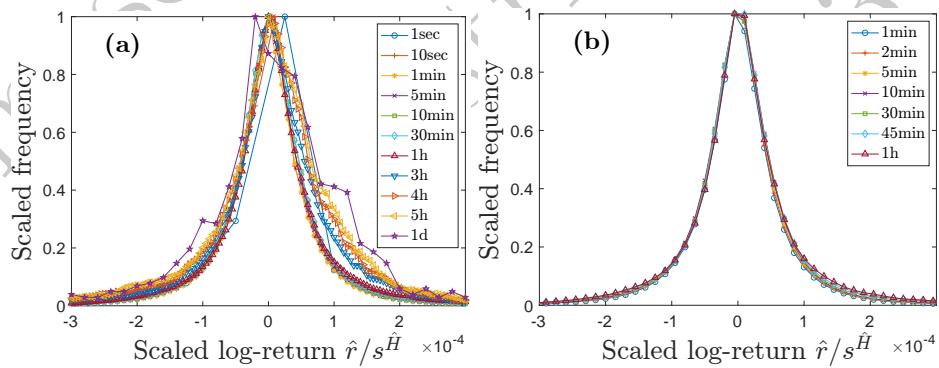


Figure 16.3 Same distributions as in Fig.16.1 but with the log-returns scaled by $s^{\hat{H}}$ with $\hat{H} = 0.497$ (best-fit parameter for the scaling of the standard deviation, see Fig.16.2). One can notice the collapse of the curves into a more similar behavior, especially for the data in the time-horizon range between 1 and 60 minutes (b).

There is a general modeling issue here. On one hand, random walk processes are simple but unrealistic, on the other hand, realistic processes are hard to construct and harder to solve making the whole exercise useless. There are three main, related, factors that make these real processes complex:

- memory effects in the returns;
- non-uniform distribution of the times intervals between successive changes;

- non-normal or stable distribution of the returns.

For instance, the log-return series at one second horizon reveals a small but significant autocorrelation. Let me in this example hereafter highlight some properties of real processes that make them more complex than simple random walks.

Example 16.2 (Real financial data are not random walk processes). I analyze log-returns for the AAPL dataset already introduced in Example 16.1. I observe that the log-return series at one second horizon reveals a small but significant autocorrelation. For the period observed, I find that, by sampling every second, price changes on average every 1.8 seconds and, if the previous change was negative, then there is a tendency (2% more likely) to have the next change positive, indicating that during this period, losses have a tendency to be followed by gains. This is a sign of memory in the returns which in this instance reveals a **mean reverting** tendency. Conversely, if the previous change was positive, then there is a tendency (1% more likely) to have the next change positive as well, indicating a **momentum** dynamic, with a persistent tendency to have a sequence of positive gains. Autocorrelations, and non-uniformity in times of price changes are related. These time series, sampled at a second frequency, have a number of successive observations of zero returns. The probability to observe a zero return after another zero return is over 20% larger than the probability that a zero return is instead followed by a positive or negative change. The average number of seconds between two changes of price is 1.8 seconds but, while 67% of changes happen within the sampling time of one second, other 33% of changes take over two seconds, with 1% taking above 11 seconds. This persistence of the price is also a memory effect that has consequences on the dynamics of the process. As discussed, in Chapter 9, returns at various time horizons are the aggregation of returns at smaller horizons (i.e. the returns at one minute are the sum of the 60 returns at one second present in the minute). In the case of financial price log-returns, there is a broad consensus that at a given horizon, the distribution is fat-tailed, with tail exponent α larger than two and characterized by a distribution that is not normal or Levy-stable. In the present case one can verify that, while the sample standard deviation of the log-returns returns is finite with value $\hat{\sigma}_1 \sim 0.00012$, the sample kurtosis has instead a value around 10^5 indicating asymptotic divergence of the fourth moment. Consistently, with other observations, one can conclude that there is the presence of fat tails with a tail exponent which must be larger than 2 (because the variance is finite) and smaller than 3 (because the kurtosis diverges). Indeed, an estimation over the left and right 1% quantiles, using Eq.14.27, gives $\alpha_{left} \simeq 2.77$ and $\alpha_{right} \simeq 2.82$, for the left and right tails respectively. The fact that the tail-exponent is larger than two ensures finite variance and asymptotic convergence toward the normal distribution. However, at any

aggregation stage, one would observe a mixture with the body of the distribution becoming increasingly similar to a normal distribution but with fat tails persisting in the extreme parts of the distribution. Such a changing shape of the aggregate distribution is affecting the scaling laws.

Despite modeling shortcomings and analytical difficulties, there are several techniques that allow to estimate scaling laws and study the evolution of the probability distribution of signals at different time horizons. The literature is vast and non-trivial and in this book, I will present only the generalized Hurst exponent technique which has been proven to be reliable to estimate the Hurst exponent in many practical cases. I will also introduce, the empirical mode decomposition, which is a way to decompose the signal into components at different time scales. Such a decomposition is directly related to scaling laws and it is an extremely useful tool to study the properties of signals at different scales and also to detrend signals in a non-parametric manner.

16.2 Estimation of the generalized Hurst exponent

I have argued in the previous Session that, while there is good evidence of scaling behavior in the observation of the log-return distributions at different time scales (see Example 16.1), there is also good evidence that the uniscaling hypothesis (see Section 9.5) is over-simplifying the complex nature of these signals (see Example 16.2). A tool to capture the complexity of the change in distribution with aggregation is the generalized Hurst exponent that characterizes the scaling of a set of moments of the distribution (see Section 9.7.2).

The computation of the generalized Hurst exponent is rather straightforward from the estimation of the sample moments at various horizons s , which are

$$\hat{\mu}(k, s) = \frac{1}{T-s} \sum_{t=s+1}^T |\hat{r}_{t,s} - \hat{\mu}(s)|^k , \quad (16.3)$$

where

$$\hat{r}_{t,s} = \hat{x}_t - \hat{x}_{t-s} \quad (16.4)$$

is the return at time t over the horizon s , and

$$\hat{\mu}(s) = \frac{1}{T-s} \sum_{t=s+1}^T \hat{r}_{t,s} , \quad (16.5)$$

is the sample mean of such a return, and $k \in \mathbb{R}$ is a scalar number (not necessarily integer) and must have values within a reasonable interval, normally within the range -1 to 4. Often in the literature, the sample mean is omitted and put to zero. Indeed, the expected value of the return must be zero in a stationary process, and in most practical cases, the sample mean return $\hat{\mu}(k, s)$ is infinitesimal. However, especially at high frequencies, this term can have a sizable effect and it can be viewed as a local adaptive detrending of the signal.

Remark 16.1. Let me note that $\hat{\mu}(k, s)$ is an average moment from a set of non-overlapping sample moments. Indeed, given a set of observations $\hat{x}_1, \dots, \hat{x}_T$ for a given horizon $1 \leq s < T$, one can compute the first return as $\hat{x}_{s+1} - \hat{x}_1$, the second as $\hat{x}_{2s+1} - \hat{x}_{s+1}$, the third as $\hat{x}_{3s+1} - \hat{x}_{2s+1}$, and so on. There are $\lfloor(T-1)/s\rfloor$ of such returns for a period with T observations. These returns are non-overlapping changes of the signal over a horizon s that start at $j = 1$. If one starts at any other $1 < j < s$ instead of $j = 1$, then one obtains a different, but equivalent, sequence of returns. I denote the horizon s return at $t = ns + j$ with

$$\hat{r}_n(s)_j = \hat{x}_{ns+j} - \hat{x}_{(n-1)s+j} \quad (16.6)$$

where $1 < n \in \mathbb{N}$ is an integer counter with $ns + j \leq T$.^a For a given j , the k sample moment at horizon s is

$$\hat{\mu}(k, s)_j = \frac{1}{\tilde{N}} \sum_{n=1}^{\tilde{N}} |\hat{r}_n(s)_j - \hat{\mu}(s)_j|^k , \quad (16.7)$$

where $\tilde{N} = \lfloor(T-j)/s\rfloor$ is the number of returns at horizon s over the period $[j, \dots, T]$ and

$$\hat{\mu}(s)_j = \frac{1}{\tilde{N}} \sum_{n=1}^{\tilde{N}} \hat{r}_n(s)_j , \quad (16.8)$$

is the sample mean of the return at horizon s with seed j . The analysis of the statistical properties of the returns at horizon s can be equivalently performed for each value of $j = 1, \dots, s$.

The dependence on the seed j enriches the statistics but also makes it more complicated to handle. There are a different number of seeds for each horizon and the size of the return's series depends on j . For simplicity, one can consider the average over the seeds j of the k -moment $\hat{\mu}(k, s)_j$, which is indeed Eq.16.3. However, one could make use of these richer seed-dependant statistics to estimate the effect of outliers and compute confidence intervals on the estimation of the scaling quantities. This is illustrated in Example 16.3.

^a Notice that to ensure $ns + j \leq T$ then one must have $j \leq T - ns$. When n has the largest value, then $j \leq T - s\lfloor(T-1)/s\rfloor$ which can force j to be smaller than s even reducing it to only $j = 1$ when $T - 1$ is multiple of s .

Coming back to the estimation of the scaling properties, from Eq.9.27 one expects these k -moments to scale as:

$$\hat{\mu}(k, s) \propto s^{k\hat{H}(k)} . \quad (16.9)$$

Such a scaling law must be first verified empirically from the data by testing if the logarithm of $\hat{\mu}(k, s)_j$ follows a linear behavior as a function of the logarithm

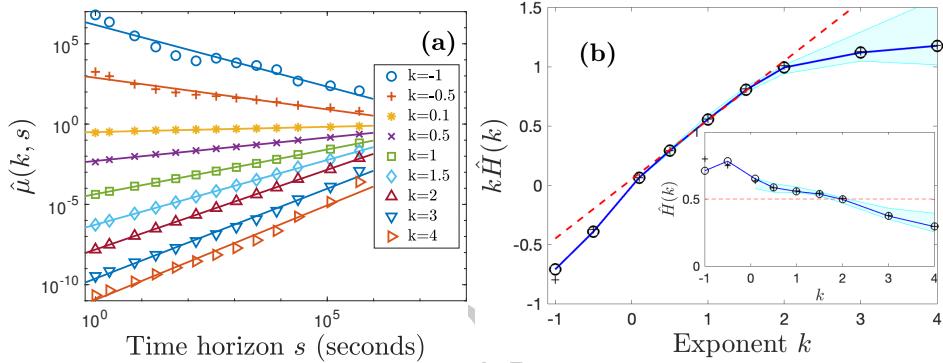


Figure 16.4 Illustration of the scaling of the sample central moments with time-horizon. Plots in (a) report $\hat{\mu}(k, s)$ vs. s showing that the linear scaling law in Eq.16.10 is well followed, especially for intermediate values of the exponents (i.e. $k \sim 1$). The plot in (b) reports $k\hat{H}(k)$, while the inset reports $\hat{H}(k)$. The horizontal dashed line is $\hat{\mu}(k, s) = H = 0.5$. The band is between the 1% and 99% quantiles computed from the scaling of $\hat{\mu}(k, s)_j$ for various values of seeds j (see Example 16.3 and Eqs.16.12, 16.13). The dashed line corresponds to $\hat{\mu}(k, s) = H = 0.5$, the value expected for Brownian motion.

of s .

$$\log \hat{\mu}(k, s) = k\hat{H}(k) \log s + \text{const.} \quad (16.10)$$

Such a scaling law is tested in Fig.16.4 and discussed in Example 16.3.

Remark 16.2. In the literature, instead of the scaling of the sample moment $\hat{\mu}(k, s)_j$ it has been considered a normalized version of it that is called **structure function**:

$$K(k, s) = \frac{\frac{1}{T-s} \sum_{t=s+1}^T |\hat{r}_{t,s}|^k}{\frac{1}{T} \sum_{t=1}^T |\hat{x}_t|^k} . \quad (16.11)$$

with $\hat{r}_{t,s} = \hat{x}_t - \hat{x}_{t-s}$. One might note that the term at the denominator of this equation does not depend on s and therefore it plays no role in the scaling property of $K(k, s)$. One might also notice that in the expression for $K(k, s)$ the sample mean is not subtracted. This is often unimportant for stationary series because, in practice, the mean of the returns is infinitesimally small and can be considered zero.

Example 16.3 (Generalized Hurst exponent for AAPL log-returns). By using the same dataset as in Examples 16.2 and 16.1 I test the scaling law of Eq.16.10 by computing the sample moments $\hat{\mu}(k, s)_j \propto s^{k\hat{H}(k)}$ and estimat-

ing the linear regression coefficient for $\log \hat{\mu}(k, s)$ vs. $\log s$, which is, indeed, $k\hat{H}(k)$. Fig.16.4(a) reports in log-log scale the plots for $\hat{\mu}(k, s)$ with k in the range between -1 to 4. The linear trends in the log-log plots are very evident indicating therefore that Eq.16.10(a) is well followed, especially in the range of exponents around unity, $k \sim 1$. The best fitting linear regression coefficients $k\hat{H}(k)$ are reported in Fig.16.4(b) and $\hat{H}(k)$ are reported in the inset. One can notice that the exponent $\hat{H}(k)$ depends on k and it has an overall decreasing trend around the value 0.5 which is crossed near $k = 2$. This indicates, deviations from pure Brownian motion. The origin of such deviations has to be searched in the interplay between autocorrelations, ‘fat-tails’, and also the complex nature of the process that, at different scales, is influenced by different factors and actors.

The bands in Fig.16.4(b) (and its inset) are the 1% and 99% quantiles of the values obtained by computing the exponents for the seed-dependent moments $\hat{\mu}(k, s)_j$ separately for each seed value j (see Remark 16.1 and Eq.16.7). Specifically, in analogy with the scaling relation, Eq.16.9, for each j I expect the following scaling law

$$\hat{\mu}(k, s)_j \propto s^{k\hat{H}(k)_j} , \quad (16.12)$$

and consequently

$$\log \hat{\mu}(k, s)_j = k\hat{H}(k)_j \log s + \text{const.} . \quad (16.13)$$

Therefore, in this case, the Hurst exponent depends on the seed value j and for a given time series one generates several equivalent values of the exponent for each j . In this example, the number of observations is 18 million and the statistics on j is very rich of data. For the range of values of the horizons between one second to 2.5 hours, I compute over 9,000 equivalent values of the seed-dependent hurst exponent with seeds values also between one second to 2.5 hours. Over this set of measures I then compute the quantiles reported in Fig.16.4(b). One might note that, for this seed-dependent estimate, I do not report values for negative k . Indeed, at high frequency, prices often return to their values and therefore there is a very large abundance of zero returns that make the negative k exponents diverge. This effect is more severe in this j dependent case with respect to the averaged one. Notice that, in this case, the contribution from the sample mean return $\hat{\mu}(k, s)_j$ in Eq.16.7 is very important and its removal (as often done in the literature) changes significantly the results.

By looking at Fig.16.4 one can notice that, while the scaling of all moments is well-followed with good linear trends in the log-log scale (panel (a)), there are however clear deviations from what is expected from pure – uniscaling – Brownian motion (panel (b)). This is a further indication that the real signal is more complex and cannot be modeled as Brownian motion. The origin of these deviations, their quantification, and their interpretation are not simple. There

are numerical and practical challenges that derive from the fact that, to obtain a good estimate of the slope, one needs a large range of values for the horizon s spanning across several orders of magnitude. However, the aggregation in s is highly data-intensive. Indeed, if one has T observations \hat{x}_t , at time horizon 1 one has $T - 1$ returns but they will be reduced to $T - s$ returns at time horizon s . These are, however, overlapping returns that carry similar information, if one considers non-overlapping returns then their number reduces to $\lfloor (T - 1)/s \rfloor$. The effect of this can be observed in Fig.16.1 where the frequencies start to become noisy above hourly horizons because statistics start to become insufficient. Other challenges are associated with the fact that the differentiation between uniscaling and multiscaling processes, requires the exploration of a range of k which must be as broad as possible to properly assess the dependence of $k\hat{H}(k)$ on k . However, the range of k cannot be arbitrarily large. In particular, if the process has returns with fat tails distributions, then the moments will become undefined for k larger or equal to the tail exponent α . Therefore, k must be smaller than α . In practice, the estimation of moments with large k becomes extremely dependent on the presence of outliers in the observation set, it is therefore recommendable to limit the computation to relatively small moments, typically not larger than $k = 3$. Negative values of k can be, in principle, explored, however, if the process can have zero returns then the computations for $k < 0$ might return errors.

In order to distinguish between the effects of the tails and the persistence of memory effects, deriving from autocorrelations, one can compare results from the empirical returns with null-hypothesis results obtained by shuffling the returns and therefore eliminating the temporal memory. For reference, this shuffling method is discussed in detail in Section 18.8.1 together with other model testing techniques. However, let me briefly explain it here briefly because it is straightforward. The approach simply consists in re-ordering the time entries of the returns (see Example 16.4) eliminating in this way memory effects. The resulting Hurst exponents associated with these null models can be considered as the sole contribution from the probability distribution aggregation without memory effects. This operation can eventually be done by allowing sampling repetition (see resampling, see Section 18.8.2) to generate intervals of confidence. Conversely, one can estimate the contribution from temporal memory disentangled from the aggregation part by generating a series of returns that preserves the original time order but have normal statistics. This can be done by taking ranks of the returns in the series and transforming the value of the returns into the associated quantiles for a normal distribution. In this way the aggregation statistic is normal and it is expected to produce $H = 1/2$ for uncorrelated Brownian motion. Deviations from this value are therefore exclusively associated with memory effects. In this case, block-bootstrapping can be used to generate an interval of confidence (see Section 18.8.2).

Example 16.4 (Effects of autocorrelations and tails on the AAPL log-returns scaling). For the same dataset in Examples 16.2, 16.1 and 16.3, I tested the effect of autocorrelations by randomizing the order of log returns eliminating the presence of memory in the signal (shuffling). The hypothesis tested here is that, if the observed deviations from Brownian motion are exclusively due to memory effects, then their removal should produce unscaling results.

I then tested the effect of the fat tails by substituting the values of the one-second non-zero returns with the relative quantiles of a normal distribution, obtaining therefore a signal of returns that are normally distributed but keep memory with the same sequence of large and low values as the original signal. The same scaling analysis performed on this normalized signal returns results for a process with the same memory structure as the original but without the effects of the fat tails.

I have run the scaling analysis for a range of exponents values between $k \in (0, 4]$ for the real data (Real) and four generated signals: 1. Brownian motion (BM); 2. shuffled real data (Real Sh.); 3. normalized real data (Real N.); 4. shuffled and normalized real data (Real Sh. N.). Results for $\hat{H}(k)$ are reported in the following Table.

k	$H(k)$				
	Real	BM	Real Sh.	Real N.	Real Sh. N.
0.1	0.650	0.501	0.635	0.857	0.891
0.5	0.585	0.501	0.583	0.578	0.567
1.0	0.557	0.501	0.560	0.532	0.523
1.5	0.537	0.501	0.544	0.521	0.507
2.0	0.497	0.500	0.506	0.519	0.500
3.0	0.374	0.500	0.380	0.526	0.491
4.0	0.294	0.500	0.299	0.539	0.486

One can first notice that only the Brownian motion (BM) returns consistently Hurst exponents very close to 0.5 across all values of k . It appears that the elimination of autocorrelations via shuffling (Real Sh.) tends to produce an increase in the estimated $\hat{H}(k)$ for $k = 1, 1.5$ and 2. Instead, in the same range of k , normalization (Real N.) seems to reduce the variability of $\hat{H}(k)$ with a decrease for $k = 1$ and 1.5 and an increase for $k = 2$. For larger k , the normalized (Real N.) case has quite different $\hat{H}(k)$ with respect to the original (Real) and the shuffled ones (Real Sh.). Indeed, while the non-normalized signals return a drop well below 0.5 in the values of $\hat{H}(k)$, instead the normalized is retrieving values above 0.5. This effect is likely to be a consequence of fat-tails which makes large moments diverge and affect severely the estimation of the Hurst exponent in the non-normalized series. It must be noticed that the shuffled and normalized process (Real Sh. N.), is not a Brownian motion and has more complex scaling laws. Indeed, despite

shuffling and normalization, there is still a non-normal component in the return distributions due to the presence of a large number of zero returns that gives the distribution a non-normal profile with a spike at zero. This must be the origin of the deviations from 0.5 observed also for the Real Sh. N. signal.

16.3 Tests for stationarity

In everything I have presented so far, I have implicitly assumed that the process under study is stationary. Such stationarity assumption must be validated. To assess strong stationarity one must verify that the whole probability distribution does not change with time while weak stationarity requires only to verify some of the statistical properties (see Definition 9.2).

There are several tests to reject/accept the stationarity hypothesis. The general concept is to verify that the statistical properties do not change with shift in time. Two common implementations are the Dickey–Fuller test and the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests. Both these tests assume that a time series y_t , $t = 1 \dots T$ can be written as a sum of three terms: a stationary stochastic component x_t , a deterministic component D_t and a (stationary) residual ϵ_t :

$$y_t = x_t + D_t + \epsilon_t. \quad (16.14)$$

The tests first take off the deterministic part, normally by subtracting a linear trend, and then assess if the statistics of x_t does not change with time.

Definition 16.2 (Characteristic equation). Let's consider a stochastic process x_t and let's assume it can be constructed as

$$x_t = a_1 x_{t-1} + a_2 x_{t-2} + \dots + a_s x_{t-s} + \epsilon_t. \quad (16.15)$$

The **characteristic equation** of this process is:

$$z^s - a_1 z^{s-1} - a_2 z^{s-2} - \dots - a_{s-1} z - a_s = 0. \quad (16.16)$$

The roots are the solutions of the characteristic equation.

16.3.1 Dickey–Fuller test

The Dickey–Fuller test assumes that the process can be written in terms of the following relation (called autoregressive):

$$x_t = a_1 x_{t-1} + \epsilon_t. \quad (16.17)$$

This autoregressive process is not stationary, however, the difference $x_t - x_{t-1}$ can be stationary but only if $a_1 = 1$ because, in this case, it reduces to $x_t - x_{t-1} = \epsilon_t$. A

general way to test for $a_1 = 1$ is testing for the roots of the characteristic equation (see Definition 16.2), if the absolute value of the root of the characteristic equation is smaller than one, then the process is considered stationary (this is called unit root test).

Definition 16.3 (Random walk). The **random walk** (see Section 9.4) is a very simple stochastic process defined by

$$y_t = y_{t-1} + \epsilon_t \quad (16.18)$$

where ϵ_t are i.i.d. random variables (see definition 5.2). It is therefore a special case, with $a_1 = 1$, of the autoregressive process Eq.16.17.

Remark 16.3. A random walk is not stationary, indeed its characteristic equation is

$$z - 1 = 0. \quad (16.19)$$

which means it has one unit root. Its first difference is instead stationary (indeed it is ϵ_t).

16.3.2 Augmented Dickey-Fuller test

The Dickey–Fuller test can be extended to more complex autoregressive processes and it is named the augmented Dickey–Fuller test (ADF). Specifically, it assumes that the process is expressible as

$$x_t = a_1 x_{t-1} + a_2 x_{t-2} + \dots + a_s x_{t-s} + \epsilon_t. \quad (16.20)$$

Again, if the absolute values of the roots of the characteristic equation are all smaller than one then the process is considered stationary. Instead, if the absolute value of one root is equal to one and all others are smaller than one, then the first difference of the process is stationary. It is said that it is a unit root process. If the absolute values of m roots are equal to one and all the others are smaller than one, then the m^{th} difference of the process is stationary. If any root is larger than one in absolute value, then the process is explosive. Note that the unit root test applies on the stochastic part, x_t , of the signal and assesses stationarity around a deterministic ‘trend’ D_t (see Eq.16.14).

16.3.3 Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test

The KPSS test consists instead in falsifying the null hypothesis that the process is in the form

$$y_t = r_t + c t + \epsilon_t \quad (16.21)$$

where r_t is a random walk: $r_t = r_{t-1} + \gamma_t$ with the noise term γ_t being an arbitrary i.i.d. random variable. Clearly, if the process is described by the expression above then it is not stationary because the random walk is not a stationary process (see Remark 16.3).

Note the stationarity tests presented in this section assess stationarity around a deterministic trend. Therefore, despite the tests might give a positive result for stationarity the signal might be non-stationary but when the deterministic part is detrended, the residual is stationary.

16.4 Rolling windows, moving averages and exponential smoothing

During their evolution, stochastic processes might change their statistical properties (i.e. they might be non-stationary). In this case, local estimates of the statistical properties, from the analysis of observations around a given time t , might be necessary to model the dynamic evolution of the system. One common approach is to use a temporal ‘window’ of Δ observations between a time $t - \Delta + 1$ and a time t and then ‘roll’ this window with a given step size h (i.e. the successive window being between $t - \Delta + 1 + h$ and $t + h$). When $h \geq \Delta$, the sequence of windows have no common observations and they are referred to as non-overlapping; otherwise, they share some observations and are called overlapping. On these rolling windows, one can build dynamical models that evolve in time. For instance, one might compute the sample mean over the window between $t - \Delta + 1$ and t .

Definition 16.4 (Moving average). The sample mean computed over a rolling window is called **moving average** or **rolling mean**. The sample mean associated with the window between $t - \Delta + 1$ and t is

$$\hat{\mu}_{t,\Delta} = \frac{1}{\Delta} \sum_{s=t-\Delta+1}^t \hat{x}_s. \quad (16.22)$$

In this definition, I associate the results with the time at the right end side of the window (latest observation time, t , in the window). In some contexts, it is preferred to associate instead the middle point in the window $t - \Delta/2$. Be mindful that, when adopting this second definition, in forecasting problems one must be very careful of linkage of future data in the forecast.

Analogously, any other moment and any other statistical property, including the entire probability distribution, can be estimated over rolling windows and they become time-dependent dynamical properties. This rolling window approach is straightforward and all tools and approaches presented so far can be directly applied to the window subsamples as much as to the whole set. There are however some challenges related to the choice of the window size Δ . The optimal choice for the window size depends on the problem, the properties of the system and

the quantities one is investigating. There isn't any simple and unique recipe for this because there are two conflicting goals:

1. for statistical accuracy purposes, one would like to have Δ as large as possible;
2. for modeling purposes, one would like to have local estimations that capture well the evolution of the system in time and therefore one would like to have Δ as small as possible.

16.4.1 Weighted moving averages and exponential smoothing

One compromise is to use weighted averages with weights such that more recent events give larger contributions to the statistical measure than older events, making the rolling average a better representation of the statistical properties at a given point in time. This reduces the propagation of the effect of outlying events that otherwise, in the unweighted case, keep affecting the results until they get outside the rolling window (see Example 16.5). This is what, for instance, it is done with weighted moving averages and with exponential smoothing.

Definition 16.5 (Weighted moving average). By using a mask of weights, $\mathbf{w} = (w_1, \dots, w_\Delta)$, one can compute averages that associate different relative importance to observations from different positions in the rolling window. For instance, one wants to give a larger weight to recent events (i.e. closer to t) and give a smaller weight to old events (i.e. closer to $t - \Delta + 1$). For any choice of weightings, w_1, \dots, w_Δ , the **weighted average** can be computed from

$$\hat{\mu}_{t,\Delta} = \frac{1}{\sum_{k=1}^{\Delta} w_k} \sum_{s=t-\Delta+1}^t w_{s-t+\Delta} \hat{x}_s. \quad (16.23)$$

A common choice of weighing is an exponentially increasing weight from 1 to Δ as

$$w_k = \exp\left(\frac{k}{\theta}\right), \quad (16.24)$$

with $\theta > 0$. The coefficient θ is a characteristic time indicating the scale of the time-lapse of contributing observations prior to the time t . An empirical rule of thumb, that has been shown to provide good output [Pozzi et al., 2012], is to set $\theta = \Delta/3$. This exponential weighting can be effective in capturing local changes in signal properties, this is illustrated in Example 16.5.

Definition 16.6 (Exponential smoothing). Similar results to weighted rolling averages with exponential weights are obtained with **exponential smoothing**. However, in this case, there is no window and the signal is recursively weighted in a way to assign larger weights to the last observation and smaller weights to the previous ones. The exponentially smoothed

average is defined as

$$\begin{aligned}\hat{\mu}_{1,\alpha} &= \hat{x}_1; \\ \hat{\mu}_{s,\alpha} &= \alpha \hat{x}_s + (1 - \alpha) \hat{\mu}_{s-1,\Delta} \quad \text{for } s > 1;\end{aligned}\tag{16.25}$$

with $\alpha \in (0, 1)$. One can notice that older observations are weighted with smaller weights and, eventually, become negligible. Indeed, the contribution of \hat{x}_{t-k} to $\hat{\mu}_{t,\alpha}$ is weighted with $w_k = \alpha(1-\alpha)^k$ which tends to zero when $k \rightarrow \infty$. This corresponds to an exponential weighting with $\theta = -1/\ln(1-\alpha)$ over an expanding rolling window stretching between 1 and t .

The exponential smoothing yields very similar results to the exponential weighting. However, while the rolling window produces its first results only after Δ observations, the exponential smoothing has outputs from the first observation. This can be an advantage of this method that, however, must be pondered with the fact that the statistics of the exponential smoothing are computed over a changing number of observations. A comparison between the rolling average, the weighted rolling average, and the exponential smoothing is described in Example 16.5, and illustrated in Fig.16.5.

In the same way, as I have just illustrated for the sample averages, one can compute the weighted average of the exponential smoothed average for any function of the variable. This is also illustrated in the Example 16.5, Fig.16.5. Before diving into this example let me define the annualized volatility, which I will indeed compute as a useful practical quantity in the example.

Definition 16.7 (Annualized volatility). In finance the variability of a price, $Price(t)$, of an asset is often quantified in terms of the standard deviation of its log-returns ($r(t) = \ln Price(t) - \ln Price(t-1)$) normalized on annual basis multiplying the sample standard deviation by the square root of the number of observations in a year. This takes the name of **annualized volatility**.

$$\text{Annualized Volatility} = \sqrt{T_{\text{year}}} \sqrt{\frac{1}{T} \sum_{t=1}^T (r(t) - \hat{\mu})^2} \tag{16.26}$$

with T the total number of observations, T_{year} the number of observations in a year (which is customarily set at 252 working days), and with $\hat{\mu}$ the sample mean

$$\hat{\mu} = \frac{1}{T} \sum_{t=1}^T r(t). \tag{16.27}$$

One might notice that accordingly to the discussion about scaling (see Chapter 9 and Section 16.1), the correct scaling factor to compute the

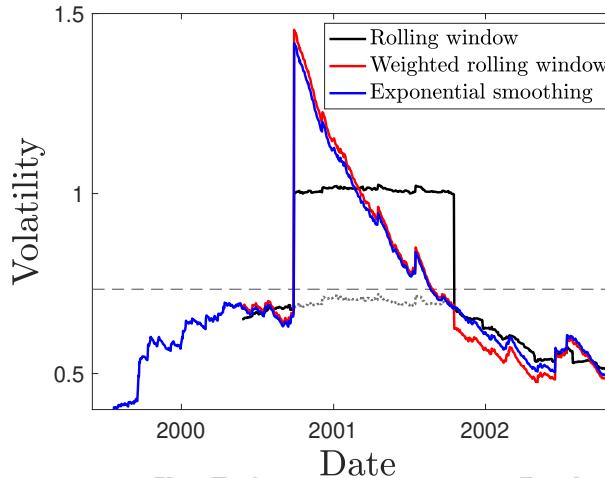


Figure 16.5 Comparison between dynamical estimation for annualized volatilities for Apple shares in the period between June 1999 and December 2002. The black curve is the rolling annualized volatility (Eq.16.28) over a window of $\Delta = 252$ working days. The dotted grey curve is the same rolling annualized volatility but excluding the outlying datapoint for Friday 29/09/2000. The red curve is the exponentially weighted rolling volatility (Eq.16.29) with $\theta = 84$. The blue curve is the exponential smoothing estimate (Eq.16.30) with $\alpha = 1 - \exp(-1/84)$. The dashed horizontal line is the average annualized volatility over the whole period.

annualized volatility should not be the square root (i.e. $\sqrt{T_{\text{year}}} = T_{\text{year}}^H$ with $H = 0.5$) but rather the generalized Hurst exponent $H(2)$.

Example 16.5 (Moving averages and exponential smoothing). Let me here compare different dynamical estimators: rolling averages, exponentially weighted averages and exponential smoothed averages. I apply these dynamical estimators to the computation of annualized volatilities for Apple share prices (see Definition 16.7). I consider Apple shares between June 1999 and December 2002 ($T = 865$ daily observations, same dataset as in Example 14.7). The standard deviation in this period, computed from daily log returns, was 0.046 and the annualized volatility was therefore $\text{Volatility} = \sqrt{252} \times 0.046 = 0.73$. This is very large annualized volatility; although Apple is a share characterized by large variations, there is also one very particular event that strongly influences this volatility estimate. Indeed, the analyzed period includes Friday 29 September 2000, a day when Apple's share fell 0.52. This is an unusually high price movement and it had a strong effect over the whole period. For instance, by excluding that day the volatility for the whole period gets down to 0.62. The persistent

effect of this outlying day on the volatility can be studied by looking at the volatility computed over a rolling window. Figure 16.5 reports with the black line this rolling volatility computed over a window of $\Delta = 252$ days (one year).

$$V_1(t) = \sqrt{252} \sqrt{\frac{1}{\Delta} \sum_{s=t-\Delta+1}^t (r(s) - \hat{\mu}_{t,\Delta})^2} \quad (16.28)$$

with $\hat{\mu}_{t,\Delta}$ the rolling mean from Eq.16.22. One can note a sharp increase in the rolling yearly volatility in correspondence to the day 29/09/2000. Notably, such a surge rests at high levels on a plateau for about a year, until the outlaying day becomes no longer included in the rolling window statistics. The line in dotted grey in Fig. 16.5 reports, for comparison, the rolling yearly volatility computed excluding the returns of the day 29/09/2000. One can see that the plateau is not present and it is exclusively an effect due to the outlaying negative returns on that Friday.

Weighting the window with exponential weights $w_k = \exp(-k/\theta)$ one computes the weighted moving average

$$V_2(t) = \sqrt{252} \sqrt{\frac{1}{\sum_{k=1}^{\Delta} w_k} \sum_{s=t-\Delta+1}^t w_{s-t+\Delta} (r(s) - \hat{\mu}_{t,\Delta})^2} \quad (16.29)$$

with $\theta = 250/3 = 84$, as suggested in Pozzi et al. [2012]. This yields to the result reported with the red line in Fig. 16.5. One can notice that weighting does eliminate the plateau with the large abrupt descent at the end of the window size observed for the unweighted rolling average. However, there is still a strong persisting effect that is ‘smoothed’ exponentially by the weighting.

The exponentially smoothed average is computed as

$$V_3(t) = \sqrt{252} \sqrt{v(t)}; \quad (16.30)$$

with $v(t)$ from the following recursive procedure

$$\begin{aligned} v(1) &= 0; \\ v(s) &= \alpha(\hat{x}_s - \hat{\mu}_{t,\Delta})^2 + (1 - \alpha)V(s-1) \quad \text{for } s > 1; \end{aligned} \quad (16.31)$$

with $\hat{\mu}_{t,\Delta}$ the exponentially smoother average from Eq.16.25, $\alpha = 1 - \exp(-1/\theta)$, $\theta = 84$ and $\Delta = 252$, as before. Results are reported with the blue line in Fig. 16.5. One can observe that the resulting estimates are very similar to the ones obtained with the exponentially weighted window. However, while the red plot shows a residual discontinuity Δ points after 29/09/2000 when this outlying day exits from the window, the exponential smoothing does not have this spurious effect.

From this example, it is evident that both weighted rolling averages and

exponential smoothing can provide a time-varying estimate of dynamical statistical properties. Both can ‘cure’ the effect of outlying data by reducing their effects exponentially. The weighted rolling average has the advantage of being flexible in the choice of weights (though usually they are chosen exponential) and having the statistic computed on the same number of observations, making, therefore, results easily comparable. Exponential smoothing has the advantage to return results from the beginning of the sample set. Conversely, the unweighted rolling average can be highly misleading because it is highly sensitive to outlying events that happened in the past. However, the equal weighting of the observations in the window makes it simpler to implement some statistical tests such as bootstrapping and shuffling (see Sections 18.8). Overall, the various methods are equivalent and the criteria for selection of one method over the other depends strongly on the purpose of the analysis.

16.5 Empirical mode decomposition

Empirical mode decomposition (EMD) is a numerical methodology to decompose a signal into components at different characteristic time scales. Usually, this methodology is not mentioned among the techniques to study the scaling of signals. Indeed, the components are not the signal at different time scales as in the self-affine picture, rather they are components that, when summed together, reconstruct the original signal. Nonetheless, these components have different characteristic time scales, and their scaling is indeed directly related to the Hurst exponent Nava et al. [2016a].

The empirical mode decomposition is a fully data-driven decomposition that can be applied to non-stationary and non-linear data Huang et al. [1998]. The purpose of the method is to identify a finite set of oscillations with a scale defined by the local maxima and minima of the data itself. Each oscillation is empirically derived from data and it is called intrinsic mode function (IMF). An IMF must satisfy two criteria:

- 1 The number of extrema and the number of zero crossings must either be equal or differ at most by one.
- 2 At any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

The IMFs are obtained through a process that makes use of local extrema to separate oscillations starting with the highest frequency. Hence, given a time series $(\hat{x}_1, \hat{x}_2, \dots, \hat{x}_T)$, the process decomposes it into a finite number of intrinsic mode functions denoted as $IMF_{k,t}$, with $k = 1, \dots, n$ and a residue $r_{n,t}$. The residue is the non-oscillating drift of the data. At the end of the decomposition process, the original time series can be reconstructed as:

$$\hat{x}_t = \sum_{k=1}^n IMF_{k,t} + r_{n,t}. \quad (16.32)$$

From a practical perspective, the EMD methodology is quite straightforward to implement. It is done in steps in a recursive way. By following Nava et al. [2016a] notation, one can describe the EMD procedure with the following steps

Algorithm 16.1: Empirical Mode Decompositions (EMD)

```

input A time series  $\hat{x}_t$  with  $t = 1, \dots, T$ .
initialize For every  $t$  set  $\hat{r}_{0,t} = \hat{x}_t$ ;
initialize For every  $t$  set  $h_{0,t} = \hat{x}_t$ ;
initialize Set the IMF index  $k = 0$ ;
while  $r_{k,t}$  is neither a constant nor a monotonic slope or contains only
one extremum do
    - Increase  $k \leftarrow k + 1$ .
    - Set the iteration counter  $i = 1$ .
    while  $h_{i-1,t}$  does not satisfy the IMF's conditions do
        • find the set of local minima of  $h_{i-1,t}$ ;
        • create the lower envelope  $E_{l,t}$  by interpolating between the
          minima;
        • find the set of local maxima of  $h_{i-1,t}$ ;
        • create the upper envelope  $E_{u,t}$  by interpolating between the
          maxima;
        • calculate the mean of both envelopes as  $m_{i-1,t} = (E_{u,t} +
          E_{l,t})/2$ ;
        • subtracts the envelope mean from the input time series, ob-
          taining  $h_{i,t} = h_{i-1,t} - m_{i-1,t}$ ;
        • increase  $i \leftarrow i + 1$ .
    - Set  $IMF_{k,t} = h_i$ ;
    - Set  $r_{k,t} = r_{k-1,t} - IMF_{k,t}$ .
output The Intrinsic Mode Functions  $IMF_{u,t}$  with  $u = 1, \dots, k$  and a
residual  $r_{k,t}$ .
```

The EMD does not require stationarity. An example of EMD for the SP500 index during for the period 05/05/2014–05/11/2014 (from Nava et al. [2016a]) is reported in Fig.16.6.

The time scale, of the IFM components is their average period τ_k , which is given by the total time period divided by the number of zero crossings plus one. For instance, the IFM reported in Fig.16.6 have periods ranging from 1.6 minutes for the first component to 11.6 days for the 17th component.

The sum of the residual and all IFM components reconstructs the signal. The residue is the overall trend of the signal and each component (starting from

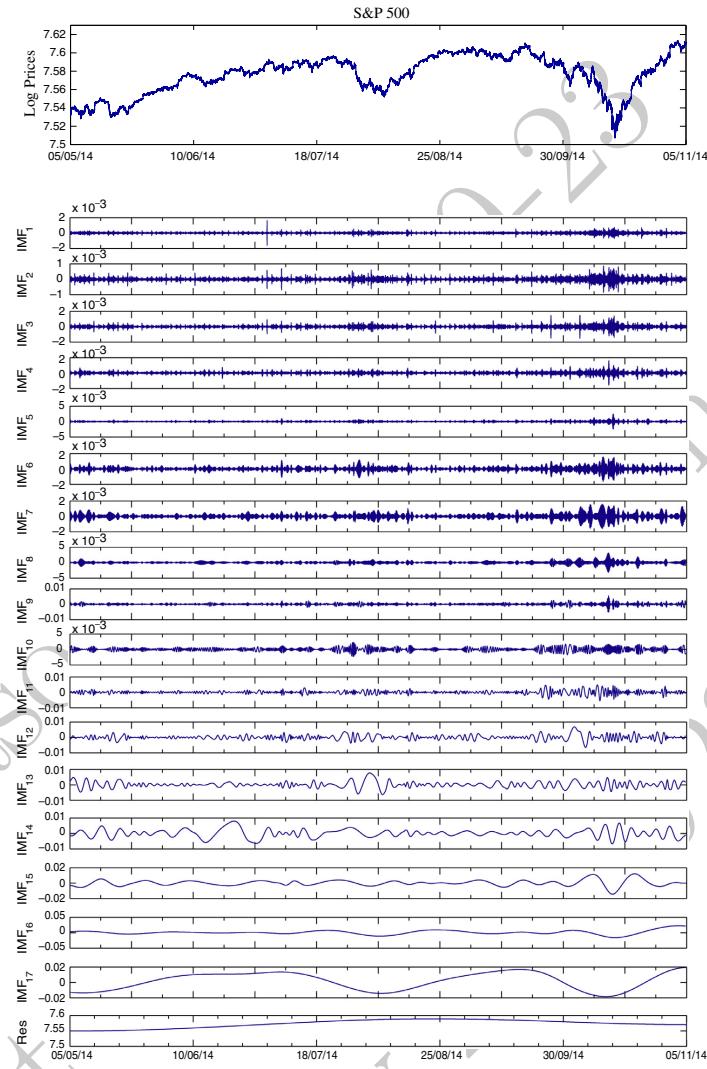


Figure 16.6 S&P500 index during the period 05/05/2014–05/11/2014 (top) and the corresponding EMD decomposition in Intrinsic Mode Functions (IFM). There are 17 IFM components and one residual. Reproduced, with permission, from Nava et al. [2016a].

the lowest frequency component) are an additional oscillating trend on shorter time scales. From this perspective, the EMD construction can be seen as a very effective de-trending and smoothing tool. This is illustrated in Fig.16.7 where a trend for the S&P500 signal is obtained by adding the residual; and the last IFM component and it is reported together with the original signal.

Each IFM and the residual contribute in different amounts to the signal. Quantifying the contribution of each IFM to the total variance of the signal is a way to

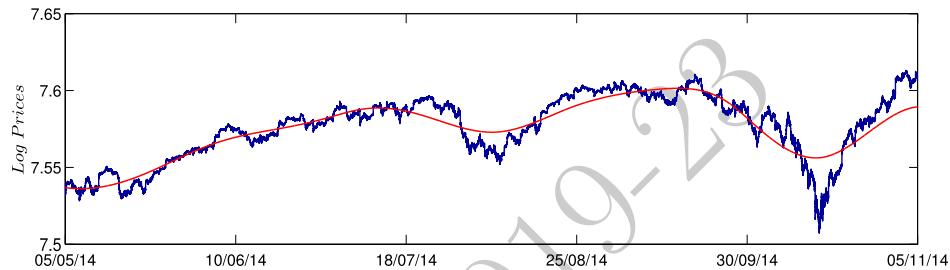


Figure 16.7 An illustration that EMD can be used as a detrending tool. The overall global trend is given by the residual and each component, from low frequency to high frequency (last to first), adds oscillating trends. The figure report in red the trend from the residual plus the last IFM component. In blue is instead reported the original signal for comparison. Reproduced, with permission, from Nava et al. [2016a].

measure the contribution to the signal from different time scales. Therefore, it is an alternative way to quantify scaling which was indeed explored in two papers: Nava et al. [2016a,b]. Briefly, the scaling law is retrieved by the law of change of the variance of the IFM with the period, τ_i . One can verify that the IFM's variance adheres to the following law

$$\text{Var}(IFN_i) \sim \tau_i^{2H}. \quad (16.33)$$

This is consistent with the scaling laws introduced in Chapter 9 and can be seen as a further, empirical, way to estimate the scaling exponent H .

16.6 Time-clustering

I have already mentioned that, in many practical cases, signals are non-stationary and their statistical properties depend on time. By detrending or differencing, one might generate stationary observations from signals that originally are non-stationary. There are however cases where the non-stationarity is an important characteristic of the signal and one might want to exploit it. For instance, there are systems that change their state, and the statistics of different states are different and must be treated separately. A simple example could be the statistics of temperature variations during nice sunny summer days or during stormy winter nights. The system is in two distinct states (the extension to more than two states is straightforward) and its modelling must be different in the case of nice-sunny-summer-days (state 1) with respect to the case of stormy-winter-nights (state 2). One can make use of such differences in the statistics for identifying the two states. Indeed, if the two states have different statistical properties, it means that the model for state 1 (\mathcal{M}_1) would better describe the observations during state 1 while the model for state 2 (\mathcal{M}_2) would better describe the observations during state 2. If one has ‘labelled’ data (and enough for each category) with the

two states explicitly distinguished (i.e. nice-sunny-summer-day or stormy-winter-night), then the solution is to build the two models from data and use them for the relative categories. However, in general, one has ‘unlabelled’ observations and, a priori, it is not known to which state each observation belongs. From a data-driven probabilistic perspective, as a first step, one can start by assigning the data into states randomly and build from these two randomly labelled datasets two models \mathcal{M}_1^1 and \mathcal{M}_2^1 . Then, at the following step, one can reconsider each observation and change its labelling if it is better described by the model for the other set and with these two newly labelled sets one can re-build the two models (\mathcal{M}_1^2 and \mathcal{M}_2^2). By iterating this procedure one eventually reach a point, t , after which re-labelling does not improve results and the models \mathcal{M}_1^t and \mathcal{M}_2^t converge to two attractors. The goodness of each model to describe the observations is the likelihood, in this context often one considers ‘costs’ to minimize and therefore the negative log-likelihood might be used as in Hallac et al. [2017]. When the models are normal distributions, then the minimization of the negative log-likelihood coincides with the minimization of the Euclidean distance between the points and the centroids of the assigned clusters. In this case, this minimization procedure coincides with the very popular k-means clustering.

16.6.1 Time clustering via k-means

With the k-means clustering one seeks to assign q observations $\hat{\mathbf{x}}_i \in \mathbb{R}^p$ ($i = 1 \dots q$) into K sets (clusters) $\mathcal{S} = (\mathbf{S}_1, \dots, \mathbf{S}_K)$, where each observation, i , is assigned to the set, \mathbf{S}_k with nearest centroid $\hat{\boldsymbol{\mu}}_k = 1/q_k \sum_{i \in \mathbf{S}_k} \hat{\mathbf{x}}_i$, with q_k the number of elements in \mathbf{S}_k . The optimal assignment must minimize the sum of the distances between the points and the cluster centroids

$$\operatorname{argmin}_{\mathcal{S}} \left(\sum_{k=1}^K \sum_{\hat{\mathbf{x}} \in \mathbf{S}_k} \|\hat{\mathbf{x}} - \hat{\boldsymbol{\mu}}_k\|_2^2 \right). \quad (16.34)$$

The general problem is NP-hard. However, when the number of clusters, K , is given, the problem can be solved exactly in $\mathcal{O}(q^{pK+1})$.

Good approximate solutions can be found by using a simple iterative algorithm which is called **k-means**. The algorithm starts with an initial set of means $\hat{\boldsymbol{\mu}}_k$ (usually computed from randomly assigned sets) and then proceeds iterating the following two steps.

- i) **assign** each observation to the cluster with nearest centroid: $i \in \mathbf{S}_k = \operatorname{argmin}_{\mathcal{S}} (d_{i,k}^2)$
with $d_{i,k}^2 = \|\hat{\mathbf{x}}_i - \hat{\boldsymbol{\mu}}_k\|_2^2$ the square of the Euclidean distance between the observation i and the centroid of cluster k .
- ii) **update** the centroids accordingly with the new assignment.

The iteration ends when the assignments no longer change. The time complexity becomes $\mathcal{O}(qpK)$ per iteration. The result of this (naive) k-means algorithm is not deterministic (the solution tends to be different for different starting configurations and different choices of the sequence of reassignments), and there is

no guarantee of convergence to an optimum solution. However, in most cases, the convergence is quite fast and hits similar configurations that are close to the optimal solution.

Common variations on the methodology use different kinds of distances, $d_{i,k}$, with respect to the Euclidean distance. A similar clustering is k-medoids where the main difference consists in using a set of data points in place of the centroids. See Rokach and Maimon [2005] for a more complete account of clustering methodologies.

16.6.2 Penalized time clustering with Viterbi path

In stochastic processes, time-clustering into states representing different aspects of the process' evolution could result in an excessive and unrealistic splitting of the states through time. On one hand, there is the need to separate the states as much as possible but, on the other hand, one would like to keep some temporal coherence and prevent continuous switching between states. A simple way to reduce switching is to penalize it by modifying the distance measure of an observation at time t , $d_{t,k}$, adding a penalization term when the observation switches between two different clusters from $t-1$ to t .

$$\tilde{d}_{t,k_t}^2 = d_{t,k_t}^2 + \gamma[k_{t-1} \neq k_t] \quad (16.35)$$

where $[k_{t-1} \neq k_t]$ is the Iverson bracket which returns 1 if the statement is true and 0 otherwise (in this case, 0 if the state rests the same or 1 if it is switched). This additional term, although simple, makes the solution to the problem harder. Indeed, the system is now interacting between adjacent times and the optimal solution must explore all possible paths. A clever way to assign elements to clusters when there is a transition cost was proposed by Viterbi [1967]. The Viterbi algorithm finds the optimal path in $\mathcal{O}(TK^2)$ and it is widely used in several application domains from deep-space communication to speech recognition. The present case is a bit simpler than the general Viterbi problem because the transition cost is simply penalizing for cluster switch and this can reduce the algorithm to $\mathcal{O}(TK)$. A compact algorithm for this simplified problem was proposed by Hallac et al. [2017]. An equivalent algorithm is sketched hereafter (see Procacci and Aste [2019]).

Algorithm 16.2: Find the least-cost path with switch penalty.

This algorithm proceeds into two loops. The first is a backward loop which populates the matrix $\text{PathCost}(t, k)$ with the minimum cost of a forward path from cluster k at t to T . The second is a forward loop which starts from the minimum cost state at $t = 1$ and then follows such a minimum cost path populating the sequence Path .

Input $\text{Cost}_{t,k} = \text{cost}$ for assigning observation at t to cluster k

Input $\gamma = \text{time consistency penalizer parameter}$

Initialize $\text{PathCost} \leftarrow T \times K$ matrix of zeros

for each observation time $t = (T - 1), \dots, 1$ **do**

for each cluster $k = 1, \dots, K$ **do**

$\text{PathCost}(t, k) =$

$\min_{k' \in [1 \dots K]} (\text{PathCost}(t + 1, k') + \text{Cost}(t + 1, k') + \gamma[k \neq k'])$

$\text{Path}(1) = \operatorname{argmin}_{k' \in [1 \dots K]} (\text{PathCost}(1, k') + \text{Cost}(1, k'))$

for each observation time $t = 2, \dots, T$ **do**

$\text{Path}(t) =$

$\operatorname{argmin}_{k' \in [1 \dots K]} (\text{PathCost}(t, k') + \text{Cost}(t, k') + \gamma[\text{Path}(t - 1) \neq k'])$

Output Path : sequence of T cluster indices associated with minimum cost

An application of this kind of temporal clustering is discussed in the following Example where the $\text{Cost}(t, k)$ is the square Euclidean distance $d_{t,k}^2$ (see Definition 16.6.1).

Example 16.6 (Time clustering via Viterbi path). I consider the set of $T = 4,148$ observations of daily closing prices of the Dow Jones Industrial Average index in the period 03/01/2006–27/06/2022. I want to assign observation days to two temporal clusters ($K = 2$) in such a way to minimize the penalized distance $\tilde{d}_{t,k_t}^2 = d_{t,k_t}^2 + \gamma[k_{t-1} \neq k_t]$. With the distance $d_{t,k_t}^2 = \|\hat{\mathbf{x}}_t - \hat{\boldsymbol{\mu}}_{k_t}\|_2^2$ and \mathbf{x}_t the log-return of the price at time t . For $\gamma > 0$, I use the least-cost-path search algorithm described in Algorithm 16.2. The resulting temporal clustering is reported in Fig.16.8. The background colors indicate the assignment of each time to one of the clusters (cluster 1 in blue and cluster 2 in green), whereas the plot is the price of the index. The penalization parameter was fixed at $\gamma = 5 \cdot 10^{-5}$. This number was chosen by using a grid search and picking the largest value that still

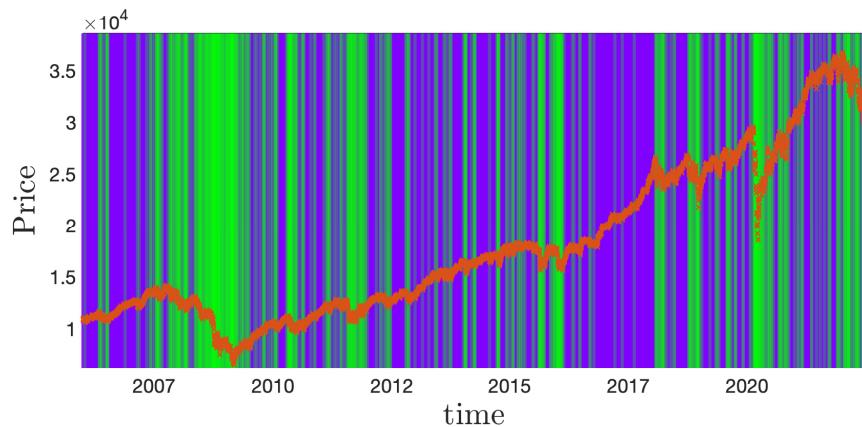


Figure 16.8 Time clustering of the Dow Jones Industrial Average index from daily prices during the period 03/01/2006–27/06/2022.

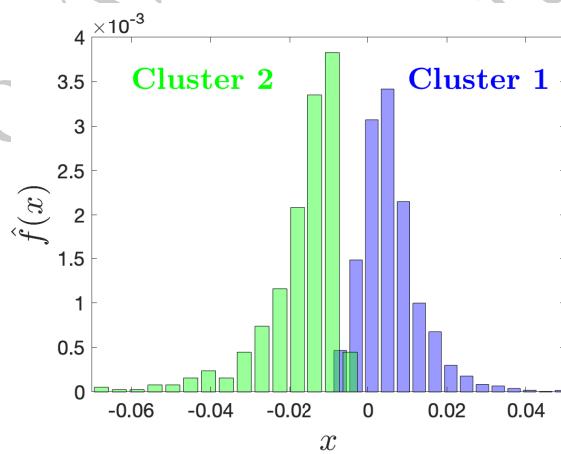


Figure 16.9 Log-return distributions for each cluster from the time clustering of the Dow Jones Industrial Average index (same data used for Fig.16.8).

produces two populated clusters with one populated with at least 10% of the observations. Indeed, for large penalizations, the solution becomes only one cluster because switching between two clusters becomes too costly. In the reported case, cluster 1 has 3,654 observations (about 88%) and cluster 2 has 494 observations (about 12%). Cluster 1 lasts on average 9.9 days while Cluster 2 lasts on average 1.3 days only. One can notice from the figure that one cluster tends to be associated with upward market trends and the other with downward ones. This is indeed confirmed by the values

of the two means that are respectively $\hat{\mu}_1 = 0.31\%$ (cluster 1, in blue) and $\hat{\mu}_2 = -2.04\%$ (cluster 2, in green). Fig.16.9 reports the distributions of log returns for the two clusters. The distinction between the kind of processes in the clusters is very evident with only small percentages of observations overlapping in the two distributions. One can see that cluster 2 is associated with very strong negative market movements. Indeed, the days in this cluster include the 2007-09 financial crisis, the 2020 COVID downturn, and the market stress surrounding the 2022 Russian invasion of Ukraine.

Here I have used the Euclidean distance as a cost measure, this coincides with the minimization of the variance (σ^2) in the clusters and, in turn, this corresponds to the minimization of the negative log-likelihood for normal modeling (indeed, $-\ell_{normal} = \frac{q}{2}(\ln \sigma^2 - \ln(2\pi))$). However, other cost measures could be more significant. Alternative modeling might seek the use of other distributions, such as, for instance, the Student-t. It must be noted that Fig.16.9 reports distributions inside the clusters that have a peculiar nature, which, for the example reported, do not resemble normal or Student-t. Therefore, other distances/costs/loss measures with practical relevance could be better suited to model this temporal clustering.

16.7 Data-driven modeling tutorial

<https://github.com/FinancialComputingUCL/DataDrivenModeling/Ch16>

The tutorial for this Chapter covers various topics on data-driven modeling of time series, including: stationarity tests, rolling averages and exponential smoothing (Example 16.5), scaling laws (Examples 16.1 and 16.3), and empirical mode decomposition.

Exercises

- An exponentially smoothed average $\hat{\mu}_{s,\alpha}$ is computed using $\alpha = 0.01$. Calculate the weight associated with x_t over the value of the average at $\hat{\mu}_{t+100,\alpha}$.
- Calculate after how many steps an outlying entry \hat{x}_t contributes to $\hat{\mu}_{s,\alpha}$ for less than 1%.
- Given a financial asset with standard deviation computed from daily log-returns equal to 0.01, compute the yearly volatility. Explicitly list all assumptions that this estimate of the yearly volatility implies.
- Consider a stochastic signal with Hurst exponent $\hat{H} = 0.4$. Assuming that the process is unscaling, estimate the value of the standard deviation of at returns at aggregation $s = 252$ given that the standard deviation at $s = 1$ is $\hat{\sigma}_1 = 0.01$.

- Two uniscaling processes both have standard deviation at $s = 1$ is $\hat{\sigma}_1 = 0.01$ but Hurst exponents $\hat{H}_1 = 0.4$ and $\hat{H}_1 = 0.6$, respectively. Compute the difference in the estimate of the standard deviation at $s = 252$ between the two processes and comment implications in terms of risk estimate.
- By applying EMD to a stochastic signal, one obtains four intrinsic mode functions respectively with average periods $\tau_1 = 1.1$ second, $\tau_2 = 33$ seconds, $\tau_3 = 1.2$ minutes and $\tau_4 = 32$ minutes. If the standard deviations of these components are respectively $\hat{\sigma}_1 = 1.1$, $\hat{\sigma}_2 = 9.0$, $\hat{\sigma}_3 = 14.8$ and $\hat{\sigma}_4 = 117.1$, estimate the Hurst exponent.

Construction of network representations from data

The generation of networks from data has become increasingly popular in data-driven modeling with applications to many different research domains. In most cases, the system under exam is a complex system made of several elements and characterized by several variables. Typically, the network vertices represent the variables and the edges represent their relations. The purpose of such a network is to characterize the interrelations between these variables and use the network for quantitative modeling. In terms of probabilistic modeling, such network representations can be either the inference structure of conditional independence or, more generally, a sparse representation describing the essential interactions from which the probabilistic model is constructed.

In this chapter, I first introduce a few practical and useful methodologies to construct networks by using various measures. These methods can be all classified under the name of ‘information filtering networks’ (see Section 11.2), which include both inference networks and networks that do not strictly represent the inference structure, but rather provide a model representation into an essential, sparse, backbone structure of interrelations. I then explain how to make use of these networks to learn, from data, multivariate probabilistic models.

Networks are extremely powerful tools to characterize complex relations between components and often discoveries can be made from the uncovering of such a structure.

17.1 Construction of networks from thresholding

At the simplest level, in the kind of problems, I am addressing here, one starts from a measure, $w_{i,j}$, defined for each couple of vertices with $i \in (1\dots p)$ and $j \in (1\dots p)$. Such a measure, $w_{i,j}$, can be one of many different quantities depending on the specific problem, in this section I will provide some examples when $w_{i,j}$ is the correlation coefficient or its statistical significance level. The assumption that $w_{i,j}$ is defined for each couple of edges (i, j) is not necessary and one can adapt the methodologies described below forbidding the presence of some edges or imposing the presence of others.

The simplest measure which describes a dependency relation between variables is the correlation and, not surprisingly, a vast body of the literature has been focused on the construction of networks from correlations (see for instance, Bonanno et al. [2004], Tumminello et al. [2007], Aste et al. [2010], Marti et al.

[2021]). Besides correlations, other dependency measures, such as mutual information, have been used Fiedor [2014]. Furthermore, directed graphs associated with measures of causality have been also explored Aste [2019].

The network representing the full set of $w_{i,j}$ for all couples of vertices, is the complete graph and, although it contains the full information it is dense and its structure is not conveying any useful information. Therefore one wants to apply pruning to transform this complete graph into a sparse structure that, while retaining most of the information of $w_{i,j}$, it also has a meaningful topological structure. Let me start with the simplest method, which is also the most used, and then move towards more sophisticated approaches.

The simplest way to construct a sparse network from a set of $w_{i,j}$ measures is by thresholding, retaining in the network only the edges with values larger than the threshold. When the measure, $w_{i,j}$, on which the thresholding is made is a statistical validation quantity such as the z-score (see Sections 15.4.1 and 18.8.1), then the network becomes a ‘statistical validated’ network and its structure represents the significant structure of interrelations at a given confidence level.

Remark 17.1. While the validation by thresholding over the z-score is similar to the one performed directly on the correlation; conversely, the validation on the p-value (in this case retaining only the edges with a p-value below a given threshold) is more delicate because, typically, p-values have very steep changes from value very close to zero to large values within a narrow correlation range. This makes the tuning of the method more problematic. For instance, in the Example 17.1 all p-values have values below the numerical precision (i.e. in this case below 10^{-14}) and therefore for this example thresholding on p-values would not be practically applicable.

Let me illustrate the construction of a network by thresholding with the following example.

Example 17.1 (Correlation network via thresholding). Let me consider the construction of a network from a correlation measure $w_{i,j} = \rho_{i,j}$ using thresholding. The network representation of $w_{i,j}$ is a complete graph and one can sparsify it by keeping all edges with the absolute value of the correlation larger than a given threshold value θ . The adjacency matrix (see Definition 4.3) is therefore

$$A_{i,j} = \begin{cases} 1 & \text{if } |\rho_{i,j}| \geq \theta \\ 0 & \text{if } |\rho_{i,j}| < \theta. \end{cases} \quad (17.1)$$

Clearly for $\theta = 0$ one obtains the complete graph, whereas for $\theta > 1$ one has a set of disconnected vertices.

Let me apply this simple construction method to a real case. I have analyzed

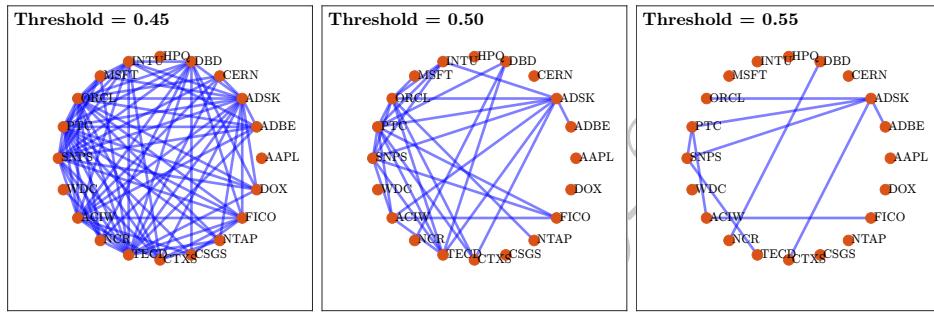


Figure 17.1 Correlation networks are obtained by keeping only edges with weights (correlations) larger than a given threshold value. One can note that the networks become gradually sparser by increasing the value of the threshold.

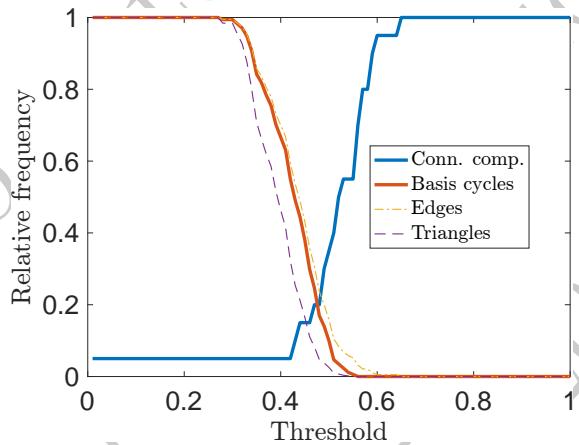


Figure 17.2 Fraction of connected components with respect to the number of vertices; the normalized number of basis cycles; the number of edges and triangles in the thresholded correlation network as a function of the threshold value. The normalization of each plotted measure is with respect to the respective value in the complete graph. One can notice that by decreasing the threshold the network changes from the complete graph (threshold at one) where every vertex is connected with every other to a set of isolated vertices with no cycles or edges (threshold zero).

the daily closing prices of 20 NASDAQ stocks in the technology sector (software and computer hardware) for the period of five years between 01 Jan 2010 to 31 Dec 2014 (1,303 trading days). I have computed the Pearson's correlation coefficients from the daily log-returns (see Definition 9.4), obtaining correlation coefficients between a couple of stocks in a range between 0.27 and 0.64, with median 0.44. These are indeed highly correlated variables from stocks that are belonging to the same industry sector and

therefore tend to be positively correlated. Let me note that, given the high correlation values and the size of the observation sample all correlations are significant with the least significant p-value still below 10^{-14} and the smallest z-score above 9.

Fig.17.1 shows on the left the network corresponding to a threshold value $\theta = 0.45$. One can observe that the network is quite dense. By moving the threshold from low to high values one observes that the network passes from complete ($\theta = 0$) to fully disconnected ($\theta = 1$) and the complexity of the structure evolves. This is shown in Fig.17.2 where I report the relative number of connected components with respect to the number of vertices, the normalized number of basis cycles, the normalized number of edges and triangles present in the network with normalization with respect to the corresponding numbers in the complete graph. It is clear that, while the network is gradually including edges, the number of irreducible cycles increases making the structure not only more connected but also more complex. Note that in correlation networks there is a strong tendency to have all triangular basis cycles. The kind of networks that one obtains at various threshold levels is plotted in Fig.17.1 where networks corresponding to thresholds $\theta = 0.45, 0.50$, and 0.55 are reported. One can notice that by decreasing the threshold (i.e. including edges with smaller weights) the network becomes locally densely connected however, there are still vertices (i.e. HPQ and AAPL) that rest disconnected. This is typical of this kind of network construction.

If instead of the absolute value of the correlation the threshold is taken over other measures, such as retaining only correlations with a statistical significance above some value (e.g. by using the t-test as described in Section 15.4.1, or the z-score in Section 18.8.1) one obtains similar kinds of results. However, in this case, these networks are referred to as ‘statistically validated’ networks Tumminello et al. [2011].

Remark 17.2. Extending further the approach illustrated in the previous example, one enters into the territory of the so-called **topological data analysis** Wasserman [2018] and the related **persistent homology** Edelsbrunner et al. [2008] approaches where these networks are analyzed at various threshold levels in term of their simplicial complex structure. However, these correlation networks have, in general, little structure from the homology perspective. Indeed, in most cases, only the 0th Betti number (i.e. number of connected components) is different from zero. For instance, in the Example 17.1 and relative Figs.17.1 and 17.2 the 1st Betti number (the genus or the number of holes) is larger than zero only in a few instances at thresholding values around 0.4. No other larger Betti numbers are observed. Consistently, the number of simplexes of dimension larger than 0

(i.e. edges, triangles, tetrahedra, ...) decreases with increasing threshold all in a rather similar way, as one can see from Fig.17.2. Overall, it appears that the persistent homology approach is not particularly informative for the study of these kinds of datasets.

17.2 Construction of information filtering networks

Graphs generated by thresholding on a scalar measure $w_{i,j}$ are simple to construct. However, as discussed in Example 17.1 often returns networks that poorly represent the data structure. Of course, the usefulness of a method depends on the dataset and on the kind of information one wants to extract. Specifically, in complex systems, where many variables are interacting in several complex ways, one wants to retrieve information that is representative of such a complex interaction structure. For instance, I have discussed in the previous section that a typical feature of networks generated by thresholding is that, for a given threshold, some parts of the networks are very dense while others are still sparse or disconnected. From a representation perspective, this means that some parts retain more edges than necessary while the interactions with other parts are ignored. It is typical of complex systems to have interactions operating over a broad scale of values and, often, rare ‘weak links’ can be more significant and representative than frequent and redundant ‘strong links’. For example, a link that connects two disjoint parts of the network is more representative than an extra link in one densely connected part. Therefore, in the literature, methodologies that generate connected networks with similar sparsity levels in all their parts have been proposed. These networks, which maximize a given gain function while taking into account some topological constraint, have been named information filtering networks [Tumminello et al., 2005]. A connected structure that is similarly sparse and similarly simple in all its parts and which retains the largest weights is the maximum spanning tree (MST), which indeed has been the first of these kinds of networks ever used. The MST solves this problem with a topological constraint: the structure must be a tree that retains the largest weights. A natural extension of such a topological constraint is to impose that the network is embedded on a surface. When the surface is planar, homological to the surface of the sphere, the problem becomes to maximize the retained weights under the constraint of being planar. This is the reasoning which motivated the introduction of the PMFG and, later, the TMFG (see Sections 11.2.2 and 11.2.3, where these kinds of networks have been introduced). Another kind of topological constraint consists in imposing that the structure is a clique tree. For cliques of size 2, one can re-obtain the MST, for cliques of size 4 one can obtain the TMFG, but for cliques of other sizes, one obtains an entire family of networks that we named MFCF (see Section 17.2).

In the construction described in Algorithm 11.5, the MFCF algorithm requires three input arguments, Min_Cl, Max_Cl, and Max_Mult (i.e. minimum clique size, maximum clique size, and separators’ multiplicity). Results also depend on the

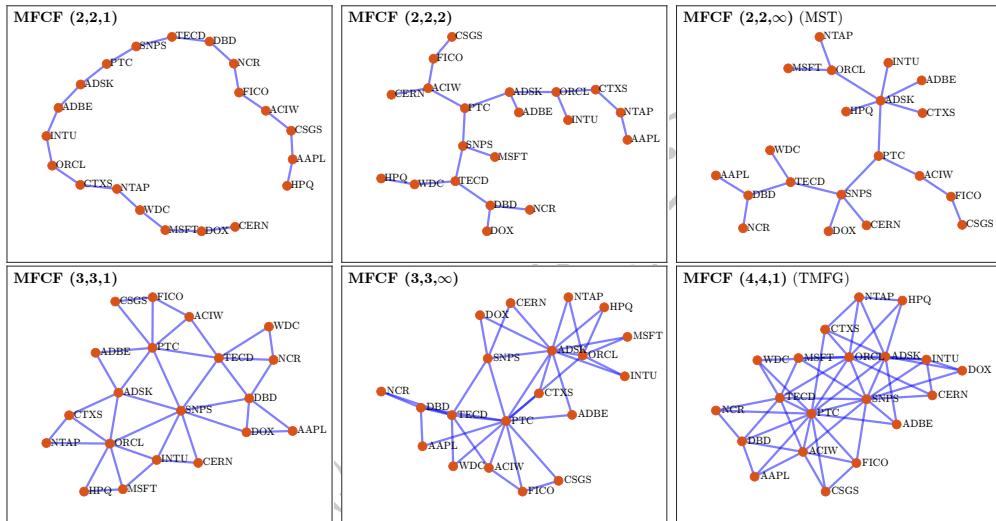


Figure 17.3 MFCF networks constructed from correlations using the same dataset used for Fig.17.1. The three networks on the top are trees, while the three on the bottom are planar graphs.

choice of gain function and the seed clique (see the reference papers [Massara and Aste, 2019, Massara et al., 2017] and the GitHub project Aste et al.). Let me showcase the construction details and the effect of the various parameters with the following example for $\text{MFCF}(\text{Min_Cl}, \text{Max_Cl}$ and Max_Mult) construction.

Example 17.2 ($\text{MFCF}(\text{Min_Cl}, \text{Max_Cl}$ and Max_Mult)). In this example, I use the same dataset of 20 assets on NASDAQ as in Example 17.1. I construct MFCF graphs using as a gain function the sum of the square correlation coefficients of the edges in each clique. This is a simple choice that allows us to compare directly with the correlation networks constructed via thresholding in the previous example. This is also a fair approximator for the maximum likelihood for various models. As a seed clique, I use the clique with minimum clique size (Min_Cl) that has the largest gain weight. In Figure 17.3 I report the resulting MFCF networks for various combinations of the parameters. The notation is $\text{MFCF}(\text{Min_Cl}, \text{Max_Cl}, \text{Max_Mult})$.

- MFCF(2,2,1), has max and min clique sizes equal to 2 (an edge) and the separators (vertices) can be used only once. The result can only be a line. It is quite clear that this graph is an oversimplified representation. Not allowing branches make the structure too constrained. Nonetheless, this simple line structure has already some features that will emerge in more complex representations, note, for instance, the relative centrality of PCT, ADSK and SNPS.

- MFCF(2,2,2), allows only edges as cliques but lets the vertices have coordination up to three (i.e. separators can be used twice), letting the structure branch out in a tree structure. With respect to the line graph MFCF(2,2,1), this is a richer structure.
- MFCF(2,2, ∞), is the maximum spanning tree (MST), where there are no cliques larger than 2, but separators can have any degree of multiplicity. With respect to the previous tree, one can see that some vertices have larger degrees. Specifically, the vertex associated with ADSK has degree 6, while SNPS has degree 4. Let me note that this and all previous graphs are spanning trees, they have the same number of edges ($p - 1 = 19$) and therefore the same sparsity. The extra structural information derives only from the constraint on the maximum multiplicity.
- MFCF(3,3,1), introduces triangular cliques. The network density almost doubles with respect to the previous three structures, passing to $2p - 3 = 37$ edges. This is a tree made of triangles. The constraint on the separator multiplicity limits the branching of the clique tree to a maximum of other three triangles. This is an outer planar and chordal graph. It is a very simple and easy to visualize graph, yet it is rich in structural information. Quite remarkably, to the best of my knowledge, these graphs have never been studied before.
- MFCF(3,3, ∞), allows the separators (edges) to have any multiplicity. This is a planar and chordal graph with $2p - 3 = 37$ edges. With respect to the previous structure, one can see that the removal of the constraint on separator's multiplicity has made the structure of this network more complex.
- MFCF(4,4,1), is the TMFG. This network is a clique tree made of tetrahedra. This is a planar and chordal graph with $3p - 6 = 54$ edges, which is the maximum number of edges that a planar graph can have and it is therefore called maximally planar. One can see that the structure maintains some of the properties identified in the previous sparser graphs but the denser structure makes it richer and more informative.

I shall come back on the use of these networks for probabilistic modeling in Examples 17.3 and 17.4.

The previous example should have made clear that information filtering networks, and the MFCF networks in particular, are a very vast family of networks with a sparse but information-rich structure. The fact that MFCF operates directly with simplexes (cliques), makes these networks belong to the ‘higher order’ graph realm. A visual comparison, between the networks obtained by thresholding (Fig.17.1) and by MFCF (Fig.17.3) demonstrates the strong differences in their structural properties and their informative power as representations of the underlying complex system. The MFCF, clique tree family provides not only a useful visual and descriptive representation of the system, but these are also

chordal graphs and can be directly used in data-driven probabilistic modeling in two different ways:

1. as inference structure for the construction of multivariate probabilistic models by making use of the joint probability decomposition into conditionally independent clique subparts (see Section 12.2);
2. as topological regularization instruments for dimensionality reduction in multivariate modeling without imposing conditional independence (see Section 15.9.5).

Let me address these two points in detail in the next section.

17.3 Information filtering networks for data-driven multivariate probabilistic modeling

The clique tree information filtering networks, such as the MFCF family, can be directly used to estimate multivariate probabilities. There are two main approaches that I shall illustrate in the following two subsections.

17.3.1 Multivariate models from probability decomposition

The first approach assumes that the network is representing the inference structure and the probability distribution function is computed using the decomposition in Eq.12.8:

$$p(\mathbf{x}) = \frac{\prod_{c \in C} p_c(\mathbf{x}_c)}{\prod_{s \in S} p_s(\mathbf{x}_s)}, \quad (17.2)$$

which I have rewritten above, for discrete variables, in terms of the probabilities, $p(\cdot)$, instead of the densities. This is a very convenient decomposition because it reduces the high-dimensional problem of estimating the whole multivariate probability $p(\mathbf{x})$ to a set of low-dimensional problems concerning the estimation of $p_c(\mathbf{x}_c)$. The lowest dimension is as low as two when the network structure is a forest. This applies to both parametric and non-parametric estimation procedures. With this construction, one explicitly imposes conditional independence between sets of variables that are not directly connected through a separator in the network. Let me illustrate the application of this decomposition in the following example.

Example 17.3 (Multivariate models from probability decomposition). Let me use again the multivariate dataset of 20 stocks log-returns that I first introduced in Example 17.1. Also, let me use the MFCF network representations that I have constructed in Example 17.2 and are shown in Fig.17.3. In that case, I used as gain function the sum of the square-correlation, this was in order to compare with results from correlation networks obtained via

thresholding. In this example I also report results for a MFCF construction but with the mutual information as a gain function.

I use multivariate normal modeling and compute the average log-likelihoods per observation of the six examples in Fig.17.3 plus two other examples and two reference structures. As a measure for the goodness of the model, I used the mean log-likelihoods per observation ($\bar{\ell} = \ell/q$) which I have computed both in-sample (over the train set period 01/01/2010-31/12/2014) and out-of-sample (over the test set period 01/01/2015-31/12/2019).

The following table reports a summary of relevant metrics for these MFCF models. Specifically: the total value of the sum of square correlations retained by the networks, the in-sample and out-of-sample mean log-likelihoods per observation ($\bar{\ell} = \ell/q$).

Model	Gain function	$ E $	sum ρ^2	train ℓ	test ℓ
Empty	—	0	0	50.080	47.216
MFCF(2,2,1)	sum ρ^2	19	9.30	55.191	51.799
MFCF(2,2,2)	sum ρ^2	19	9.78	55.474	51.301
MFCF(2,2, ∞)	sum ρ^2	19	10.04	55.617	51.374
MFCF(3,3,1)	sum ρ^2	37	18.58	56.312	52.015
MFCF(3,3,2)	sum ρ^2	37	18.75	56.326	52.056
MFCF(3,3, ∞)	sum ρ^2	37	18.84	56.334	52.064
MFCF(4,4,1)	sum ρ^2	54	26.76	56.666	52.617
MFCF(4,4,1) MI	Mut. Inf.	54	26.61	56.667	52.592
Complete	—	190	82.61	57.160	53.049

The values of the average log-likelihoods per observation have no direct, absolute, meaning. They are significant only when log-likelihoods from different models are compared. In this case, the two models that should be the reference are the one where the network is empty and there are only isolated vertices, and the complete graph where the network is full and everything is connected with everything else. The values corresponding to these two reference structures are reported in the first and last row of the table.

Network complexity increases from top to bottom of the table. One can note that, regardless of the number of edges ($|E|$), by increasing complexity, all networks tend to increase the sum of the square of the edges' correlations and the in-sample (train set) and out-of-sample (test set) log-likelihoods. Model MFCF(4,4,1) MI has been trained to maximize the mutual information between the joining vertex and any sub-clique of the clique tree. This model indeed improves on in-sample (train set) likelihood with respect to MFCF(4,4,1) (row above) but only very marginally and it loses on the out-of-sample (test set) likelihood and on the sum of squares. It is important to stress that both these models are TMFG networks, although constructed using two different gain functions.

Overall the result in the table demonstrates that sparse model decomposition via MFCF networks produces models with comparable likelihood to the full models. The question that springs to mind is whether sparse models do better than the full one. A hint should come from the discussion in Chapter 15 and in particular from Fig.15.9 where it is shown that sparse models overperform full models in the regions where the number of observations is comparable or smaller than the number of variables. In the example used for the table, one has $p = 20$ variables and $q = 1,303$ observations, therefore $q/p \simeq 65$, which is a region where one expects the full model to perform relatively better than the sparse ones.

I have therefore performed a set of other experiments, using the same dataset, but with train sets taking observation windows of various sizes from $q = 5$ to $q = 2,500$ extending backward from the 1st of October 2019. For each of these experiments, I have computed the out-of-sample likelihood over the same test set of 3 months ranging from October 3rd, 2019 to 31st December 2019 for a total of 64 observations. For each of these different train sets, I repeated the computations of the table for all MFCF models, the empty graph model, and the complete graph model. The results are reported in Fig.17.4 on the left panel where each plot corresponds to a probabilistic model associated with a specific MFCF network. First notice that the value of the likelihoods are consistent with the results in the above table. Second it is evident from the figure that, while the model associated with the complete graph is performing very well for large q/p ratios, at smaller values of this ratio (q/p below 10) the sparse models outperform the full ones with all analyzed sparse models outperforming the full model for $q/p < 5$. Let me finally notice that likelihood appears slightly decreasing for large train sets above $q/p \sim 50$, which is a period of above 4 years of daily observations. This is, most likely, the consequence of non-stationarity: the system changes with time and old observations become less useful for model construction.

17.3.2 Multivariate models from sparse inverse covariance estimate

The second approach for the estimation of the multivariate probability distribution with information filtering networks uses the network structure to better estimate the parameters of a multivariate model distribution. Specifically, for the elliptical family probability distribution, the network is used to compute the non-zero elements of the sparse inverse covariance (or shape) matrix. This is a form of topological regularization (see also Section 15.9.5 and Aste [2020]) where the number of parameters of the model (i.e. the coefficients of the inverse covariance) is reduced from $\mathcal{O}(p^2)$ to a fewer number of the relevant ones which correspond to the edges in the information filtering network, typically $\mathcal{O}(p)$. In the case of multivariate normal distributions, this approach and the one described in the pre-

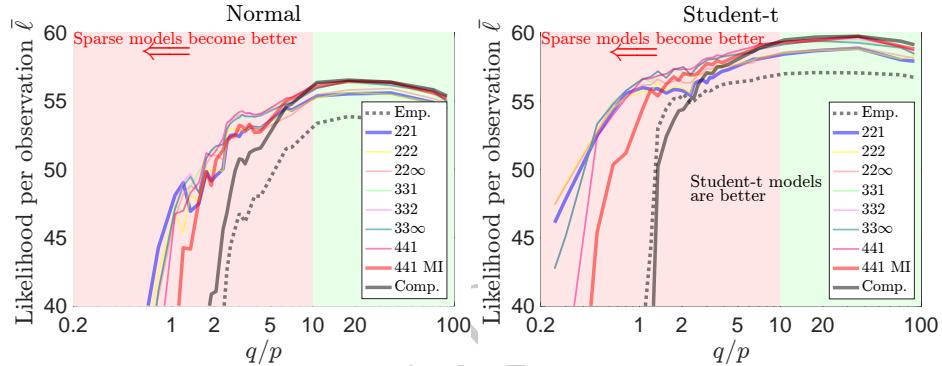


Figure 17.4 Test set likelihoods per observation, $\bar{\ell}$, for various models vs. the relative train sample size q/p . On the left are reported results for multivariate normal models while on the right are results for multivariate Student-t models. The various plots are associated with different network representations and follow the same notation used in Examples 17.3 and 17.4.

vious section, coincide. However, in general, for non-normal models, they don't, and this second method does not impose conditional independence between disconnected variables, and therefore the probability decomposition in cliques and separators does not hold any longer. In this respect, the local gain in likelihood is achieved by joining two cliques, described in Section 12.1, should not be applied in this case.¹ Other gain functions that consider the overall change in likelihood must be used instead for the construction of these models.

In practice, this approach is quite simple, one must compute a chordal information filtering network, for instance, the MFCF, and then use sparse inverse covariance instead of the full in the associated probabilistic model (see Section 15.9.4). This procedure is best illustrated with the following example.

Example 17.4 (Multivariate models from sparse inverse covariance estimate). In this example, I use the same data and the same network representations as in the previous example but instead of using the probability decomposition, I use the sparse inverse covariance approach. As stated earlier, for multivariate normal models the two approaches are identical, and therefore all results reported in the table in Example 17.3 hold for both approaches. I have therefore to illustrate this procedure with a non-normal model as an example. Let me use the multivariate Student-t model. The multivariate Student-t distribution depends on the following parameters (see Definition 6.10): the means, the degrees of freedom, and the inverse covariance (precisely the inverse scale, or shape, matrix). For the estimate of the means, I use the sample estimate as for the previous example. For the

¹ However, it is still a good approximation.

estimate of the degrees of freedom, I use the maximum likelihood method (see Eq.14.27) applied over both tails of all marginals and averaged. For the estimate of the scale matrix, I use the sparse inverse covariance obtained over the various information-filtering network structures. The results are as follows.

Model	Gain function	$ E $	train ℓ^{St}	test ℓ^{St}
Empty	–	0	55.850	55.354
MFCF(2,2,1)	sum ρ^2	19	57.877	56.956
MFCF(2,2,2)	sum ρ^2	19	58.120	56.897
MFCF(2,2, ∞)	sum ρ^2	19	58.258	57.037
MFCF(3,3,1)	sum ρ^2	37	58.952	57.691
MFCF(3,3,2)	sum ρ^2	37	58.965	57.677
MFCF(3,3, ∞)	sum ρ^2	37	59.018	57.665
MFCF(4,4,1)	sum ρ^2	54	59.322	58.113
MFCF(4,4,1) MI	Mut. Inf.	54	59.332	58.113
Complete	–	190	59.741	58.546

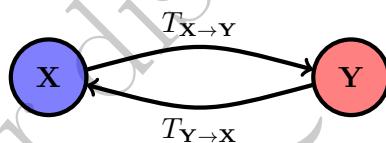
The overall results are very similar to what was reported for multivariate normal models in the previous example with sparse models associated with the MFCF networks producing excellent results with a likelihood comparable to the full model even for networks with large sparsity. In Fig.17.4 on the right panel I report the results for the likelihood associated with all MFCF network probabilistic models for the multivariate Student-t case vs. relative sample size. Results are consistent with what was observed for the multivariate normal models the main difference being an overall likelihood increase between 5% to 15% for Student-t models with respect to normal models indicating therefore that multivariate Student-t is a better-suited model than multivariate normal for this dataset. This amount of likelihood increase is quite remarkable considering that within the normal or Student-t modeling the range of likelihood values from the best to worst models stays within 6% and 12% respectively. Let me notice that further improvement in the likelihood performances could be obtained by applying the expectation maximization estimate of the parameters (see section 15.6) on the local sparse network representation Aste [2020].

A quite remarkable result that one can observe from Fig.17.4 is that models constructed from only $q = 5$ observations (extreme left-side) can still yield relatively high values of the likelihood but they must be sparse. For instance the model associated with network MFCF(2,2, ∞), has $\bar{\ell}^{St} = 47.40$ and it is marginally better than MFCF(2,2,1) with $\bar{\ell}^{St} = 46.1$. Furthermore, already from $q = 10$ sparse models such as MFCF(2,2, ∞) have already $\bar{\ell}^{St}$ above 53, which is within 12% from the absolute maximum reached by the full and TMFG IM models both reaching $\bar{\ell}^{St} = 59.75$ but constructed by using a much larger set with $q = 713$ observations.

17.4 Causal networks construction

Causality is a directional quantity and therefore causal networks are usually depicted with directed edges. Within the probabilistic data-driven modeling perspective, the problem of quantification of causality concerns the estimation of a multivariate joint probability between lagged variables. I have indeed pointed out in Section 10.4 that the transfer entropy, which is one of the most relevant measures for statistical causality, is a conditional mutual information between lagged variables (see Eq.10.16). Therefore, ultimately, we are facing the same kind of estimation problem as discussed in the previous part of this chapter. However, there is a further dimensionality increase due to the use of the lagged variables and a further increase in the complexity of the interpretation of the model's outcomes. The topic is vast, and quite subtle, with several approaches and several perspectives. I suggest the interested reader to start the investigation of alternative approaches from the classical book of Pearl et al. [2000].

I introduced the concept of transfer entropy $T_{\mathbf{X} \rightarrow \mathbf{Y}}$ between two sets of variables in Section 10.4. This quantity measures the reduction in uncertainty about the values of variables \mathbf{Y} at future times provided by the knowledge of past values of variables \mathbf{X} , discounting for the knowledge from the past of \mathbf{Y} . It is a non-negative scalar number. The larger this number is, the stronger the information held by the past of \mathbf{X} about the future of \mathbf{Y} . Often, both directions $\mathbf{X} \rightarrow \mathbf{Y}$ and $\mathbf{Y} \rightarrow \mathbf{X}$ have significant transfer entropy values. In a network representation, this bi-directional causal relation can be depicted as below:



Bi-directionality of causation is not a contradiction since the information in one direction is not the same as the one in the other. This makes the present perspective fundamentally divergent from the field of structure learning, where causality cycles are not admitted. Sometimes, the difference between the transfer entropies in the two directions is used and some authors call it **information flow**. In this case, the direction is assigned to the (positive) flow corresponding to the direction of the largest transfer entropy.

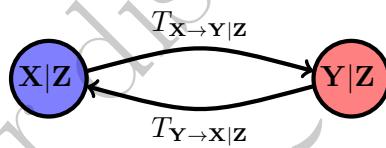
Transfer entropy requires the estimation of mutual information between two sets of variables (the past of \mathbf{X} and the future of \mathbf{Y}) conditioned to, at least, a third set (the past of \mathbf{Y} , see Section 10.4). Therefore, its computation is involving several variables and the curse of dimensionality is affecting precision or even computability.

For large sets of variables, one might want to construct a causality network where the most relevant pairwise causal relations are represented. In principle, one, naively, would like to associate directed edges only to unique relations of the causal influence of one set of variables onto the other set, discounting the contribution of all other variables. However, I have already commented in Section

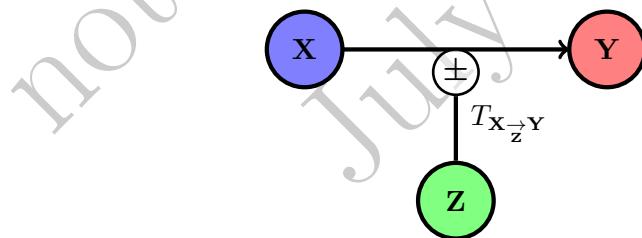
8.9.2 that, for more than two sets of variables, there are both redundant and synergic contributions (see Definition 8.10) which either can decrease or increase the pairwise mutual information. The interpretation of these positive and negative contributions in the context of causality is not straightforward and can produce misleading conclusions, as pointed out by James et al. [2016].

A set of pairwise transfer entropy relations between several groups of variables can be represented with a directed network which, sometimes, is named transfer entropy network. The network can be made sparse by retaining only the most relevant entries either by thresholding or validating. This network however does not represent the transfer entropy structure, it is just a representation of the collection of pairwise interactions.

The conditional pairwise transfer entropy $T_{X \rightarrow Y|Z}$, is more significantly related to the network representation, in this case, conditioning should be with respect to all other variables or the nearest neighbor of the X , Y pair. For multivariate systems with a large number of variables, conditioning to the neighbors only reduces issues with the curse of dimensionality. However, both the pairwise transfer entropy $T_{X \rightarrow Y}$, which ignores the possible contribution from other variables, Z , and the conditional pairwise transfer entropy $T_{X \rightarrow Y|Z}$, which instead eliminates the contribution from the set Z by conditioning (see sub-Section 10.4.1), do not represent correctly the true structure of causal relations, which are ultimately very hard to uncover from statistical analysis only. The network representation of this bi-directional conditional causal relation can be depicted as below:



Instead, the three variable extension $T_{X \rightarrow Y|Z}$ introduced in Section 10.4.3 can be visually represented as:



This is an instance of a higher-order network because the vertical edge is not acting between two vertices but rather between a vertex and an edge. Transfer entropy network representations are, intrinsically, higher order networks because one does not simply join two variables with a directed edge but rather one connects several sub-sets of variables. Overall, these representations are hard to display in graphical form and they are difficult to interpret from a modeling

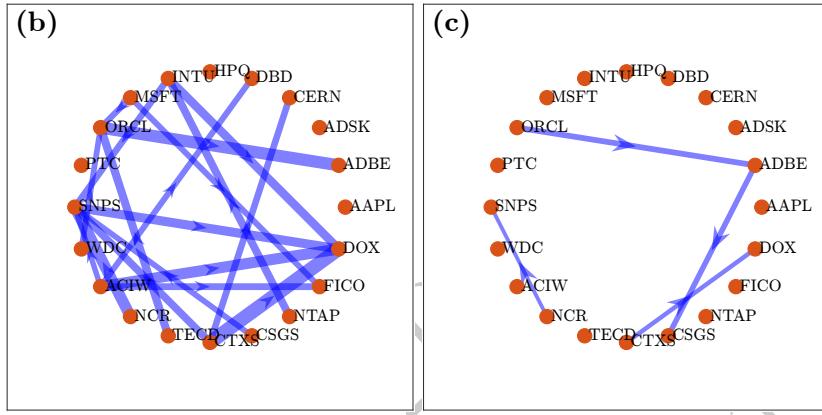


Figure 17.5 Three examples of transfer entropy networks on a system of the 20 NASDAQ technology stocks from Fig.17.1 where edges correspond to significant transfer entropies with z-scores above 5. (a) reports the transfer entropy links corresponding to a 3-day lag, which turns out to be the most significant; (b) reports the conditional transfer entropy links corresponding to a 3-day lag.

perspective. Let me conclude this section with the following example where the previously mentioned measures of causality are quantified.

Example 17.5 (Transfer entropy network). By using the same dataset as in Example 17.2 (daily closing prices of 20 NASDAQ stocks in the technology sector for the period of five years between 01 Jan 2010 to 31 Dec 2014) I compute the transfer entropies and the conditional transfer entropies between couples of stocks using linear elliptical modeling (i.e. using Eqs.10.20 and 10.32). In Fig.17.5 I represent with directed edges only (a) the transfer entropies and (b) the conditional transfer entropies validated with a z-score (see Section 18.8.1) larger than 5. For each couple of variables (i, j) , the conditioning is with respect to the past all other variables. For both the transfer entropies and the conditional ones I use a three-day lag because this lag produces the largest number of validated edges (densest graph) for this dataset.

One might notice that the strongest link in Fig.17.5(a) is between ORCL and ADBE for which I obtain $TE_{ORCL \rightarrow ADBE} = 6.5 \cdot 10^{-3}$ and a z-score of 18.5. The opposite direction has instead $TE_{ADBE \rightarrow ORCL} = 1.8 \cdot 10^{-3}$ and a z-score of 2.4. Conditioning to all other variables keeps a residual conditional transfer entropy at $TE_{ORCL \rightarrow ADBE|z} = 4.3 \cdot 10^{-3}$ and a z-score of 7.2. Not surprisingly, in this case, conditioning to all other variables reduces the transfer entropy value. By using Section 10.4.3 notation with $\mathbf{X} = ORCL$,

$\mathbf{Y} = \text{ADBE}$, and $\mathbf{Z} = \text{all other variables (at lag 3-days)}$, I have:

$$T_{\mathbf{X} \rightarrow \mathbf{Y}} = +2.2, \quad (17.3)$$

This is a typical inhibition instance where some of the information that was transferred from ORCL to ADBE was shared with the rest of the system and therefore conditioning reduces causation. More surprisingly, is instead the appearance of a new link between ADBE and CSGS in the conditioned case Fig.17.5(b). For this case we have $TE_{\text{ADBE} \rightarrow \text{CSGS}|\mathbf{z}} = 4.1 \cdot 10^{-3}$ with z-score 8.9 while the unconditioned case had $TE_{\text{ADBE} \rightarrow \text{CSGS}} = 2.2 \cdot 10^{-3}$ with z-score 2.6. This is a case of enhancement of causal relation between two variables induced by the set of the other variables. By using Section 10.4.3 notation, with $\mathbf{X} = \text{ADBE}$, $\mathbf{Y} = \text{CSGS}$, and $\mathbf{Z} = \text{all other variables (at lag 3-days)}$, I can write:

$$T_{\mathbf{X} \rightarrow \mathbf{Y}} = -1.9 \cdot 10^{-3}. \quad (17.4)$$

This enhancement of causality by conditioning could be, for instance, the consequence of reduced noise from the rest of the system.

17.5 Data-driven modeling tutorial

<https://github.com/FinancialComputingUCL/DataDrivenModeling/Ch17>

The tutorial for this Chapter covers various topics on the construction of useful network representations, including: correlation networks via thresholding (Example 17.1), TMFG and MCFC methodologies (Example 17.2), and multivariate modeling with sparse inverse covariance estimate (Example 17.4)

Exercises

- Construct a threshold correlation network from the following correlation matrix:

$$C = \begin{pmatrix} 1 & 0.58 & -0.54 & -0.51 & 0.16 \\ 0.58 & 1 & -0.25 & -0.22 & -0.17 \\ -0.54 & -0.25 & 1 & 0.83 & 0.46 \\ -0.51 & -0.22 & 0.83 & 1 & 0.66 \\ 0.16 & -0.17 & 0.46 & 0.66 & 1 \end{pmatrix}. \quad (17.5)$$

- Assuming that the previous correlation matrix has been obtained from 20 multivariate observations, construct a validated network where only correlations with p-values below 1% are reported. (You might use Eq.15.12).

- Construct the maximum spanning tree using the measure $w_{i,j} = C_{i,j}^2$.
- Construct MCFC(3,3,1) using the measure $w_{i,j} = C_{i,j}^2$.
- Compute the sparse inverse correlation associated with the maximum spanning tree.
- Consider the observation: $\mathbf{X}_t = (-0.34, 0.24, -0.48, -0.35, -0.11)$. Assuming, the variables have zero means and are from a multivariate normal distribution. Using C as a correlation matrix and the results from the previous exercises, compare the likelihoods using the full covariance, the MST sparse inverse covariance, and the MCFC(3,3,1) sparse inverse covariance.
- Compare the previous results for the log-likelihoods with the ones obtained from Student-t models with degrees of freedom $\nu = 3.8$.

Assessing the goodness of models

All models are wrong. However, some models are useful and can help to understand and navigate reality. Some models provide a precise, quantitative, representation of the system under study and can be used for description or prediction with good accuracy. Other models might be less precise but simpler or easier to interpret, or faster to compute and the loss in precision could be compensated by these other features. Eventually, some models do not provide any extra information about the system and their use is misleading; they must be avoided. Being able to quantify the ‘goodness’ of models and assess which one is more adequate for a given purpose is as important as the model itself. As a matter of fact, in data-driven modeling, the two operations are not distinct because the model is constructed, learned, and trained in a way to maximize its goodness for a given purpose.

The goodness of models is somehow subjective and task-dependent: a model is good if it performs well the task that it must perform. Therefore, the quantification and qualification of the model’s goodness cannot be universal but it depends on the model and its purpose. Furthermore, assessing the goodness of the model has a different meaning depending on the part of the dataset under exam (see Section 3.5.1). Indeed, the goodness of the model in training quantifies how well the model fits the dataset and it is used to learn the model parameters. Differently, in the validation set, the goodness measure provides an instrument to compare models and their architectural choices, it is used to learn the hyperparameters from data and select a model. Finally, in testing, the model goodness is the instrument to evaluate model performances, even on data not seen before. Often different measures of goodness are used for training, validation, and testing purposes.

There are many kinds of models and for each kind of model, there are different goodness criteria. In this chapter, I shall review some of the most common methods which can be used to assess the goodness of models and select the best performing among different competing models. This topic is vast. This chapter aims at providing my perspective on some of what I consider essential instruments. I also try to give a unified vision from an information theoretic perspective, showing how some of the commonly used methods are related.

These days data scientists and data-driven-modelers commonly adopt models from software packages and statistical validation tools are routinely part of the package outputs. This has made the use of different models and their comparison

much simpler and more approachable even for people with little knowledge about the underlying theoretical foundations of the model. For this reason, it has become even more important to know and understand the measures used for quantifying the goodness of models, enabling modelers to judge which are the best and most appropriate models to use in every situation.

18.1 Null models

It has been argued that it is impossible to prove from observations that a hypothesis is correct [Popper, 1934]. Indeed, one would need to verify that the model based on such a hypothesis correctly provides the right answers across all possible instances. Conversely, one needs only one observation that is in contradiction with the hypothesis to demonstrate that there are circumstances when the model is not right. In other words, to prove that sometimes a model does not work is possible, whereas to demonstrate that a model always work is impossible.¹

This is the logic behind the use of ‘null models’: to understand which is the ‘best’ model for a given system one can start excluding some models that one can prove to be faulty; these are the so-called ‘null hypotheses’ that one aims to prove wrong. These null hypotheses do not have to be trivial models, they can be any kind of model that we wish to test and compare against other models. For instance, within the probabilistic perspective of this book, a typical null model is to hypothesize that the observations follow a normal distribution, if one finds that such a hypothesis is unlikely, then it can be excluded that the population is normally distributed and one can attempt to describe our problem using another kind of distribution such as, for instance, the Student-t. Another instance where p-values are often used is to reject the null hypothesis that a model is not providing any information and therefore that it is not a good model. Also in this case a small p-value will reveal that it is unlikely that the model under examination is not informative. This excludes some level of confidence that the model is bad; however, this does not imply that the model is actually good. In some cases, the p-value can be used to distinguish between two models. If the null hypothesis that the two models are performing equivalently is rejected, this implies that the two models might not be equivalent and one model might therefore be significantly better than the other.

18.2 P-values

The probability, under a given hypothesis, to observe a value of a random variable X that is more extreme than the one actually observed, \hat{x} , is the p-value. In a formula:

$$p_{value} = P(X \leq \hat{x} | Hypothesis), \quad (18.1)$$

¹ Note that I am not talking about the falsification of a model. The fact that a model does not work perfectly does not ‘falsify’ it and, vice versa, in the real world there are no exact and absolute models. For instance, the general relativity theory explains phenomena more precisely than Newton’s dynamics but one cannot say that it falsifies it and we still use and teach both.

(or, vice versa, $P(X > \hat{x}|\text{Hypothesis})$ if the other extreme is of relevance). One can recognize, that this is the cumulative distribution of the likelihood. Indeed, the likelihood function, $P(X = \hat{x}|\text{Hypothesis})$ (see Definition 14.1), is the probability to observe a given value, \hat{x} , under some hypothesis (i.e. assuming a known population probability distribution). Note that, here I use $P(\cdot)$ to denote both probability or density depending on the variables being discrete or continuous.

If the hypothesis in question is a ‘null model’, one wants to retrieve a small p_{value} , which indicates that, under such hypothesis, the observation of values equal to \hat{x} or smaller (larger) is highly unlikely and therefore one can discard the hypothesis that the null model is true. In general, one looks for small p-values to discard the null hypothesis. How small this value should be is conventional and normally p-values below 5% or 1% are considered low enough to discard a hypothesis.

Despite the fact that a small p-value is a good indication that a model is likely not to be good and could be discarded, conversely a non-small p-value is not at all a measure of the goodness of a model. Indeed, observations that are drawn from a population that coincides with the model under test (the ‘true’ model) return p-values that are uniformly distributed between 0 and 1.

Let me show that, when the hypothesis under the p-value test is actually the ‘true model’ that has generated the data, then the p-values are uniformly distributed between zero and one. Indeed, given that the cumulative distribution is a non-decreasing function of the variable, then the condition $X \leq \hat{x}$ implies $F(X) \leq F(\hat{x})$. Consequently, the p-value can be written as

$$p_{value} = P(F_{true}(X) \leq F_{true}(\hat{x})|\text{Hypothesis}) \quad (18.2)$$

where $F_{true}(\hat{x})$ is the cumulative distribution of the ‘true’ model. If the population distribution is indeed the one of the null hypothesis, then $F_{true}(\hat{x}) = P(X \leq \hat{x}|\text{Hypothesis})$ and consequently

$$p_{value} = P(F_{true}(X) \leq p_{value}|\text{Hypothesis} = \text{True}). \quad (18.3)$$

The cumulative distribution of the p_{value} coincides with the p-value. This can occur only if the distribution of the p-values is uniform ($f(p_{value}) = 1$ over the support $p_{value} \in [0, 1]$). Indeed,

$$P(F_{true}(X) \leq p_{value}|\text{Hypothesis}) = \int_0^{p_{value}} 1 dp = p_{value}. \quad (18.4)$$

Consequently, by construction, p-values cannot be used to support a hypothesis but only to reject it (see also Goodman [2008]).

The fact that the true model has uniform p-values distribution means that, although a p-value of 1% is a good indication that the model under test is not good, there is still one chance per 100 to obtain such a small value from the true model. The same applies to very large p-values, which are close to 1. They

do not indicate, at all, that the model is good. On the contrary, when the true model is applied, large p-values are as unlikely to happen as small p-values. Consequently, large p-values are likely to indicate that the hypothesis under test must be reverted and it is the likelihood to find the variable larger or equal to X which is very small and the (conjugate) hypothesis should be discarded.

The rejection of the null hypothesis through the p-value is a formidable investigation instrument that greatly helps in the identification and construction of models. However, it has been often misused and abused, to the point that the American statistical association felt the need to warn the researcher's community about the interpretation and use of p-values.

Remark 18.1. “P values, the ‘gold standard’ of statistical validity, are not as reliable as many scientists assume.”

In 2016 the American statistical association (ASA) released a statement on statistical significance and p-values that took many by surprise. This was the first time that ASA took a position in any statistical practice. Clearly, they felt the strong need to “shed light on an aspect of our field that is too often misunderstood and misused in the broader research community”.

They said:

1. *P-values can indicate how incompatible the data are with a specified statistical model.*
2. *P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*
3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*
4. *Proper inference requires full reporting and transparency.*
5. *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*
6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*

This caution note does not mean that p-values are not useful or that they must be dismissed. It simply indicates that they must be used properly. As a good practice, when testing hypotheses with p-values, it is recommended to produce a set of p-values by testing the hypothesis over several independent observation sets. Indeed, a repeated number of small p-values is a strong indication that the null hypothesis is poor and should be rejected, on the other hand, a large and uniform range of p-values does indicate that the null hypothesis cannot be rejected. In cases when further independent observations cannot be obtained and the dataset cannot be split, distributions of p-values can still be obtained through bootstrapping, Jackknife, or holdout methods (see later in this Chapter in Sections 18.8.2, 18.8.2 and 18.8.2).

Remark 18.2. I have shown (see Eqs. 18.1-18.3) that the p-value is not useful to identify the ‘true model’ because it returns values from a uniform distribution under the true hypothesis. One can see from the derivation that this is the direct consequence of the fact that the p-value is a cumulative likelihood. On the contrary, the probability

$$P(X = \hat{x} | Hypothesis) \quad (18.5)$$

is instead useful to identify the ‘true model’. Indeed, it is the likelihood of \hat{x} and we shall see in Section 18.7 that this is the main quantity that is used to quantify the goodness of models in most of the model selection criteria.

18.2.1 Testing multiple hypotheses: Bonferroni correction

If one tests a large number of hypotheses (i.e. try several models), then the probability that, eventually, one hypothesis among many is statistically rejected – by pure chance – becomes large. Indeed, just from statistical fluctuations one of the p-values could end up being smaller than the threshold.

To take this into account, Bonferroni [1936] proposed to divide the p-value level by the number of hypotheses. For instance, if 10 different hypotheses are tested at 1% p-value, then the actual p-value for each hypothesis must be ten times smaller at 0.1%. The correction is a consequence of the fact that in a test of m hypotheses, if each hypothesis has a p-value smaller than γ/m then the ensemble of the hypotheses has a p-value smaller than γ .

It could be noticed that the world’s research community as a whole is continuously testing a very large number of hypotheses on any given subject. Often they do so by using the same dataset. The probability that a team obtains, by chance, small p-values increases with the number of researches performed. Therefore, p-values should be always highly penalized. On the other hand, we have seen that not-small-enough p-values are not proof that the hypothesis is valid, and also small p-values are not conclusive proof that an hypothesis is false. A better way to test a hypothesis is to test its adherence with datasets not used to formulate the hypothesis itself (out-of-sample data) and perform multiple tests on independent datasets.

18.3 Comparing and testing probability estimates

Probabilistic modeling consists in estimating the probability distribution from observations. In this context, the quantification of the goodness of the model concerns the estimation of how well the probability distribution, learned from the data, describes the observations and it is therefore similar to the true distribution of the population. Such quantification is, in general, provided by the likelihood, as I shall discuss in Section 18.7. However, there are several other non-likelihood-based methodologies that can be useful in specific contexts. Let me in this section

list a few simple and fundamentally relevant methods to assess the goodness of univariate probability models.

18.3.1 Q-Q plot

The quantile-quantile plot is a graphical tool that is useful to compare a model probability distribution against empirical data. It consists in generating a plot with their quantiles respectively in the x- and y-axis. It is particularly useful to spot inconsistencies, especially in the ‘tail’ region of the distribution.

In practice, one starts with an observation set $(\hat{x}_1, \dots, \hat{x}_q)$ and a model cumulative distribution function $\tilde{F}(x)$ with known inverse $Q(p) = \tilde{F}^{-1}(p)$ (i.e. the expression for the quantile, see Section 2.2). The Q-Q plot is then constructed by plotting on the y-axis the observations \hat{x}_k and on the x-axis the corresponding quantiles $\tilde{F}^{-1}(\hat{F}(\hat{x}_k))$. Where $\hat{F}(x)$ is the empirical estimate of the cumulative distribution function computed from the order statistics $N(\hat{x}_k)$ of the observations as $\hat{F}(\hat{x}_k) = N(\hat{x}_k)/q$ (or $\hat{F}(\hat{x}_k) = N(\hat{x}_k)/(q + 1)$, see Section 13.6). If the model exactly describes the dataset then we must have $\tilde{F}^{-1}(\hat{F}(\hat{x}_k)) = \hat{x}_k$ and the plot shall result in a straight diagonal line $y = x$. Deviations from such a line are indicating inconsistencies between the model and the empirical observations.

The Q-Q plot can also be used with two empirical quantile estimations to compare the statistical consistency of two samples of data or it can be as well used with two model distributions to compare the differences between them.

Q-Q plots are commonly used to compare datasets with the normal and other common probability distribution functions where the inverse of the cumulative is known. However, in some cases, finding the explicit expression for the inverse of a cumulative distribution can be challenging and therefore there are instances where this method cannot be used.

18.3.2 P-P plot

Instead of comparing quantiles, one can directly compare cumulative distribution functions, avoiding in this way the need to compute the inverse cumulative distribution function of the model, which in some instances can be problematic. The P-P plot reports on the x-axis the theoretical cumulative distribution from the model $(\tilde{F}(\hat{x}_k))$ and on the y-axis the empirical estimate $\hat{F}(\hat{x}_k)$ (see Section 13.6).

The P-P plot is a graphical representation of the probability integral transformation (see for instance Casella and Berger [2021]). Given a random variable, X with continuous cumulative distribution function $F(X)$, the random variable $Y = F(X)$ has uniform distribution in the interval $[0, 1]$. Indeed,

$$P(Y < y) = P(F(X) < y) = P(X < F^{-1}(y)) = y \quad (18.6)$$

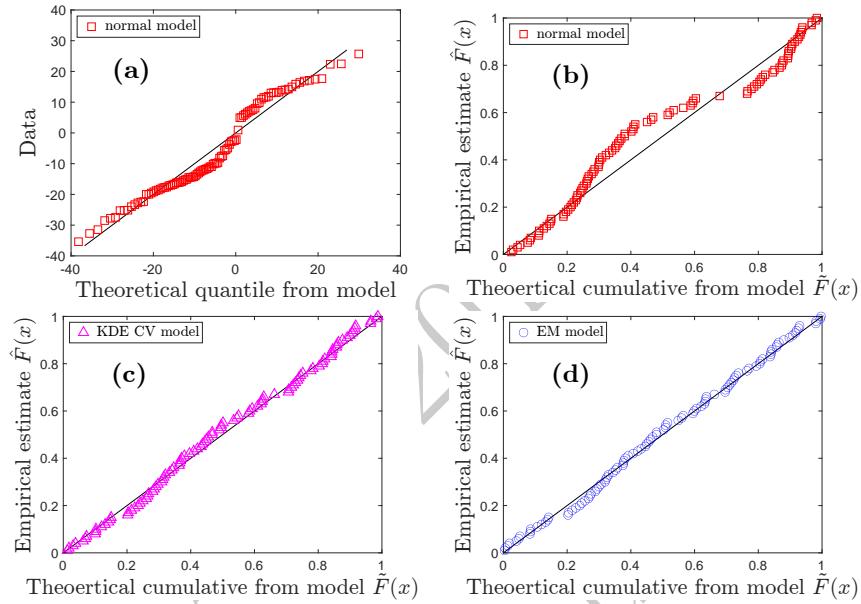


Figure 18.1 Use of Q-Q and P-P plots to test different models for the dataset introduced in Example 13.6. (a) Q-Q plot for a normal distribution model with mean and standard deviation given by the sample measures (method of moments); (b) P-P plot for the same normal distribution model as in (a); (c) P-P plot for the KDE model with bandwidth $h^* = 3.53$ estimated from cross-validation (Example 14.19 and Fig.13.4); (d) P-P plot for the EM model reported in Figure 14.4 (see Example 14.8).

which can be written as

$$P(X < F^{-1}(y)) = F(F^{-1}(y)) = y. \quad (18.7)$$

Therefore $F(y) = y$ and $f(y) = \text{const}$. Which means indeed that the distribution is uniform. Notice that this is the same fact we observed for the p-value. This becomes a very useful tool to test whether a random variable is drawn from a given known distribution, indeed, for this purpose, it is sufficient to test for the uniformity of $F(X)$.

Example 18.1 (Q-Q and P-P plots). In order to illustrate the Q-Q and the P-P plots let me return to the dataset introduced in the Example 13.6 and already used for Figures 13.4, 14.2 and 14.4. The Q-Q plot for a normal model with mean and standard deviation equal to the sample measures (method of moments) is reported in Figure 18.1(a) while the P-P plot is reported in Figure 18.1(b). Note that the information conveyed by the Q-Q and P-P plots are quite similar. However, the P-P plot tends to be less sensi-

tive than the Q-Q plot for what concerns deviations in the tails (deviations that are very evident from Figs.18.2 (c),(d) from Example 18.2, where a fat-tailed distribution of real financial returns is investigated). Figure 18.1 (c), (d) report P-P plots for the kernel density estimation with maximum likelihood bandwidth value from the Example 14.6 and the Gaussian mixture model from Example 14.8. As mentioned, this is one of the main advantages of the P-P plot that does not require to compute the inverse of the model's cumulative distribution and therefore can be computed more easily for special kinds of models. By comparing Fig.18.1(b) with Fig.18.1(c,d) one observes clear visual evidence that the kernel density and the Gaussian mixture models are better describing the data resulting in a narrow spread of points very close to the diagonal straight line. Quantitative evidences through validation tests are discussed in the following sub-Sections 18.3.3 and 18.3.4.

18.3.3 Kolmogorov-Smirnov test

The Q-Q plot and the P-P plot methods, discussed in the previous two subsections, look at the difference between the theoretical probability distribution and the empirical estimates. These methods are essentially visual and qualitative.

More quantitatively, one might want to measure the distance between the empirical $\hat{F}(x)$ and the model $\tilde{F}(x)$ cumulative distribution functions. For this purpose, one might observe that the two simplest distance measures between these functions are the absolute value and the square of the difference. These are indeed the distances that two of the main statistical validation tests use: the Kolmogorov-Smirnov test and the Anderson-Darling test.

The Kolmogorov-Smirnov (KS) test [Massey Jr, 1951] computes the largest absolute value of the difference between $\hat{F}(x)$ and $\tilde{F}(x)$:

$$D_q = \sup_{x_k \in \{\hat{x}_1, \dots, \hat{x}_q\}} |\hat{F}(\hat{x}_k) - \tilde{F}(\hat{x}_k)|. \quad (18.8)$$

Under the null hypothesis, for continuous $\tilde{F}(x)$, the quantity $\sqrt{q}D_q$ converges, asymptotically, to the Kolmogorov distribution, which does not depend on $\tilde{F}(x)$.

Definition 18.1 (Kolmogorov distribution). The **Kolmogorov distribution** quantifies the probability that a stochastic process, $X(t)$, of duration q and such that it starts and ends at values equal to zero, $X(0) = X(q) = 0$ (a process called Brownian bridge), reaches an absolute maximum value $K = \max |X(t)|$ larger than a given number. It is

$$P(K > x) = 2 \sum_{s=1}^{\infty} (-1)^{s-1} e^{-2s^2 x^2}. \quad (18.9)$$

Asymptotic convergence is quite slow, however, there are good estimates of the distribution of $\sqrt{q}D_q$ for finite q (see Vrbik [2018]). The KS test returns a p -value: a probability under the null hypothesis to obtain a value of D_q that is larger or equal to the measured one.

KS test is commonly used to reject the null hypothesis that data are drawn from a normal distribution. However, the KS test can be performed also to compare consistency between two samples computing the distance as the supremum of the absolute difference between the two empirical estimates of the cumulative distribution function. This is called the two-sample Kolmogorov–Smirnov test.

The KS test has two main weaknesses:

1. it relies on a supremum, which means that it is therefore extremely sensitive to outliers;
2. the difference between the cumulative distributions $\tilde{F}(x)$ and $\hat{F}(x)$ might be large in the body of the distribution where both $\tilde{F}(x)$ and $\hat{F}(x)$ have values around 0.5, but they will be minimal in the tail regions where both $\tilde{F}(x)$ and $\hat{F}(x)$ have values close to 0 (left tail) or 1 (right tail). Therefore KS test is not able to discriminate well between models when they differ in the tail regions.

The Anderson-Darling test addresses both of these issues.

18.3.4 Anderson-Darling test

The distance between cumulative density functions, used for the KS test, could be misleading because, in the tail regions, the differences tend to zero while in these regions small differences can have dramatic consequences. The Anderson-Darling (AD) test [Anderson and Darling, 1952] computes instead a weighted average distance between the cumulative density functions

$$A^2 = q\mathbb{E} \left[\frac{(\hat{F}(x) - \tilde{F}(x))^2}{\tilde{F}(x)(1 - \tilde{F}(x))} \right]. \quad (18.10)$$

The idea is based on the fact that, if the data come from the model distribution $\tilde{F}(x)$, then the quantity $(\hat{F}(x) - \tilde{F}(x))^2/(\tilde{F}(x)(1 - \tilde{F}(x)))$ is a random variable with uniform distribution. The AD test measures the degree of uniformity of such a distribution. Given a continuous cumulative distribution $F(x)$ and a dataset $(\hat{x}_1, \dots, \hat{x}_q)$ which is ordered in a non-decreasing order $(\hat{x}_{k_1} \leq \hat{x}_{k_2} \dots \leq \hat{x}_{k_q})$ the quantity, A^2 , can be estimated from

$$A^2 = \sum_{s=1}^q \frac{1-2s}{q} [\ln(F(\hat{x}_{k_s})) + \ln(1 - F(\hat{x}_{q+1-k_s}))] - q. \quad (18.11)$$

Under the null hypothesis (that the data are from the distribution $F(x)$), the probability distribution of A^2 , is the Anderson-Darling distribution. This has a complicated form, however, a closed form for $q \rightarrow \infty$ was computed by Marsaglia and Marsaglia [2004] and values for small q are provided by approximate expressions.

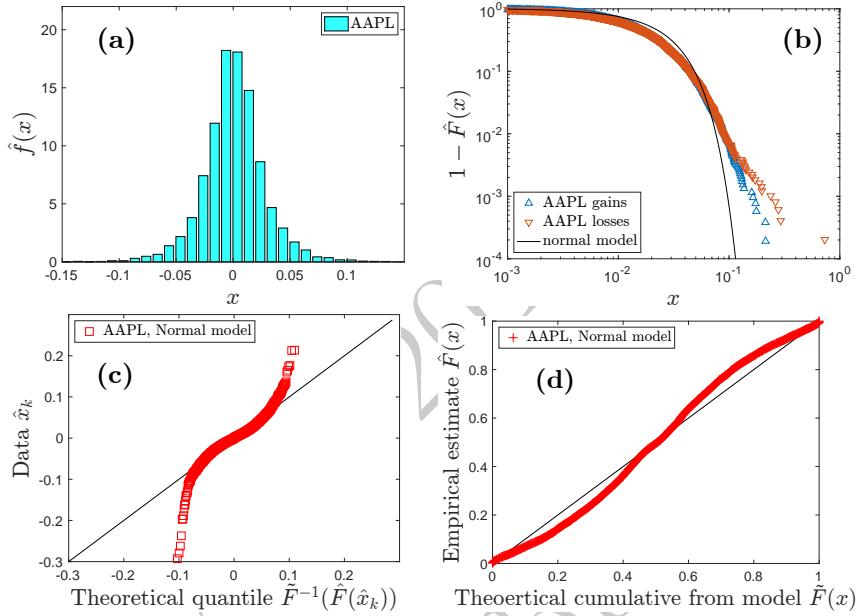


Figure 18.2 (a) Histogram estimate of the PDF for the log-returns of Apple daily log-price returns during the period December 1985 to December 2019; (b) Complementary cumulative distributions for the positive and negative returns plotted in log-log scale; (c) Q-Q plot comparison with best fitting (maximum likelihood) with normal distribution; (d) P-P plot for the log returns of Apple stock prices compared with the normal distribution. One can notice that, differently from the Q-Q plot, the P-P plot does not reveal differences in the tails.

Example 18.2 (Distribution of financial return data). Let me here consider a practical example for the distribution of the variation of prices of a stock equity. Specifically, I look at the log-returns (see Definition 9.4) for the Apple (AAPL) stock price. Data are daily adjusted closing prices from the period December 1985 to December 2019 where there are a total of 9478 trading days.

Figure 18.2 reports the (a) histogram estimate of the PDF and (b) the empirical estimates of the complementary cumulative distributions computed separately for the positive and negative returns and plotted in a log-log scale. The Q-Q plot in Fig. 18.2 (c) reveals that the empirical distribution is very different from a best fitting, maximum likelihood, normal distribution especially in the tail regions and more remarkably for the negative returns (i.e. the losses, left tail).

The P-P plot is instead shown in figure 18.2 (d). Let me notice that in this plot the large deviations in the tail regions observed in the Q-Q plot are no longer visible.

Both Kolmogorov-Smirnov and Anderson-Darling tests reject the null hypothesis that the data are drawn from a normal distribution with $p_{value} < 10^{-15}$ (using a normal model with sample mean and standard deviation, respectively $\hat{\mu} = 7 \cdot 10^{-4}$ and $\hat{\sigma} = 2.9 \cdot 10^{-2}$).

Notice that the observed largest negative return has a value of -0.73, which is over 24.9 standard deviations away from the mean. The probability of such a value for a normal distribution model will be below 10^{-100} . Considering that the age of the universe is less than 10^{13} days it means that, under a normal model, it would have never had a chance to occur, ever. Instead, on Friday 29 September 2000, Apple's stock price fell down 51.9 percent. This was a consequence of an announcement that the fourth-quarter profit was lower than expected. It is worth noting that Chebyshev's inequality (see Section 5.5.1) returns a probability smaller than 0.2% for this extreme event, assessing it therefore likely every 10 years. However, Chebyshev's inequality provides very loose bonds.

The AD test rejects the hypotheses of stable distribution with $p_{value} \simeq 5 \cdot 10^{-7}$ (using best-fit estimate $\hat{\mu} = 7 \cdot 10^{-4}$, $\hat{C} = 2 \cdot 10^{-2}$ and $\hat{\alpha} = 1.7$ and $\hat{\beta} = 2 \cdot 10^{-2}$, computed using Matlab 'fitdist'). The stable distribution hypothesis could also be rejected at 5% by the KS test that returns $p_{value} \simeq 2 \cdot 10^{-2}$. The AD test also rejects the hypotheses of Student-t distribution with $p_{value} \simeq 7 \cdot 10^{-3}$ (using best-fit estimate $\hat{\mu} = 8 \cdot 10^{-4}$, $\hat{\sigma} = 2 \cdot 10^{-2}$ and $\hat{\nu} = 3.8$, computed using Matlab 'fitdist'). Conversely, KS test does not reject the hypothesis of Student-t distribution returning $p_{value} \simeq 0.12$. This shows that the AD test is more restrictive than KS test especially when deviations in the tails are concerned.

18.4 Testing the goodness of regressions

Given a set of coupled observations²

$$\begin{aligned} & (\hat{\mathbf{x}}_1, \hat{\mathbf{y}}_1) \\ & (\hat{\mathbf{x}}_2, \hat{\mathbf{y}}_2) \\ & \vdots \\ & (\hat{\mathbf{x}}_q, \hat{\mathbf{y}}_q), \end{aligned} \tag{18.12}$$

one wants to 'learn' the model $g(\cdot)$ that best explains the relation

$$\mathbf{Y} = g(\mathbf{X}) + \epsilon. \tag{18.13}$$

between the two sets of random variables \mathbf{X} and \mathbf{Y} from which the observations have been drawn. Depending on the context, this task is called 'supervised learning', or 'regression', or 'prediction'. (When the variable set \mathbf{Y} is discrete – or

² I use bold symbols indicating that in general, both variables are multivariate with $\hat{\mathbf{x}}_s = (\hat{x}_{1,s}, \hat{x}_{2,s}, \dots, \hat{x}_{p_X,s})^\top$ and $\hat{\mathbf{y}}_s = (\hat{y}_{1,s}, \hat{y}_{2,s}, \dots, \hat{y}_{p_Y,s})^\top$, with $s = 1, \dots, q$.

equivalently categorical – the task is instead called ‘classification’ but I’ll discuss this in the next section.)

Normally, the true model $g(\mathbf{X})$ is unknown and the aim is to ‘learn’ from data a model $\tilde{g}(\mathbf{X})$ which minimizes the residual

$$\tilde{\epsilon} = \mathbf{Y} - \tilde{g}(\mathbf{X}). \quad (18.14)$$

The quantity $\tilde{\epsilon}$ is a set of random variables itself and defining the measure to choose for such minimization is the first task.³ A possible measure is the entropy of the residual, $H(\tilde{\epsilon})$, which is indeed a quantification of uncertainty. However, entropy is quite a sophisticated measure that sometimes can be challenging to quantify. The simplest and most common measure is the variance or the sum of the squares, this is what the R-square approach uses.

Let me notice that for a broad range of statistics, minimizing entropy corresponds to minimizing the variance $\text{Var}(\tilde{\epsilon})$ or the determinant of the covariance of the error’s components $|\Sigma_{\tilde{\epsilon}\tilde{\epsilon}}|$ respectively in the univariate or multivariate cases.

18.4.1 R-square

One of the most commonly used quantities to estimate the goodness of a regression is the sum of the squared values of the residual which is called indeed Residual Sum of Squares (RSS). Let me consider the case when Y is univariate⁴

$$\text{RSS} = \sum_{i=1}^q (\hat{y}_i - \tilde{g}(\hat{\mathbf{x}}_i))^2 = \sum_{i=1}^q \tilde{\epsilon}_i^2. \quad (18.15)$$

One can see that, if $\mu_{\tilde{\epsilon}} = 0$ (which is true for an unbiased model), this quantity coincides with q times the sample estimation of the variance of the residual, $\hat{\sigma}(\tilde{\epsilon})$ (see Eq.13.5). For a given observation set, smaller values of RSS imply that the model is a better fit of the dataset. However, the value of RSS has little meaning if not compared to a reference value. It is also a dimensional quantity, with values related to the unit of measure of Y . A natural choice is to compare the residual sum of squares with the Total Sum of Squares (TSS)

$$\text{TSS} = \sum_{i=1}^q (\hat{y}_i - \hat{\mu}_Y)^2, \quad (18.16)$$

with $\hat{\mu}_Y$ the sample mean of Y . The ratio between RSS and TSS is a good estimate of how well the model explains the variability of Y . In particular, the

³ Notice that I used the ‘tilde’ on the symbol for the residual, $\tilde{\epsilon}$, to distinguish it from the regression error ϵ . Later I will add the ‘hat’, $\hat{\epsilon}$, to indicate its values from observations.

⁴ This univariate case is actually quite general because any multivariate case can be reduced to a set of univariate instances by computing the R^2 for each component of the multivariate \mathbf{Y} and summing them

relative difference, $(\text{TSS}-\text{RSS})/\text{TSS}$, is known as the coefficient of determination or R-square:

$$\hat{R}^2 = 1 - \frac{\text{RSS}}{\text{TSS}} \quad (18.17)$$

Large values of \hat{R}^2 indicate that a large fraction of the variance Y is explained by the variables \mathbf{X} through the model $\tilde{g}(\mathbf{X})$.

Remark 18.3. In this part of the literature, another notation is often used with the value estimated by the regression ($\tilde{g}(\hat{\mathbf{x}}_i)$) denoted with \hat{y}_i . However, in this Book, I have already used the ‘hat’ symbol to denote observations and in my notation, \hat{y}_i is the observed output corresponding to the input $\hat{\mathbf{x}}_i$. Therefore, I adopted a different notation for the model estimate. Also, the sample mean of Y (which in my notation is $\hat{\mu}_Y$) is instead often denoted with \bar{y} in the literature.

Definition 18.2. In this Section, I have introduced two new symbols RSS and TSS. In the literature, there are several other symbols that are commonly used for the same or strictly related quantities. Let me try to account for them and their relations.

TSS is the total sum of squares also known as the sum of square total (SST), it is equal to the sample variance of Y times the size of the dataset $\text{TSS} = \text{SST} = q\hat{\sigma}_Y^2$.

RSS is the residual sum of squares also known as the sum of square error (SSE). This is strictly related to the mean square error (MSE), which is RSS/q .

SSR is the sum of square regression, or explained sum of squares (ESS), and it is

$$\text{SSR} = \text{ESS} = \sum_{i=1}^q (\tilde{g}(\hat{\mathbf{x}}_i) - \hat{\mu}_Y)^2. \quad (18.18)$$

Generally, an optimal regression model, which minimizes the sample variance of $\tilde{\epsilon}$, has consequently $\hat{\mu}_Y = \hat{\mu}_{\tilde{g}(\mathbf{X})}$ and $\sum_{i=1}^q (\tilde{g}(\hat{\mathbf{x}}_i) - \hat{\mu}_Y) = 0$ (i.e. it is unbiased with residual uncorrelated with predictor). This, in turn, implies $\text{TSS} = \text{RSS} + \text{SSR}$. Notice that this is valid only for the optimal regression model, sub-optimal models do not satisfy this equality.

Remark 18.4. The bias-variance tradeoff (see Section 3.5.2) tells us that the ‘goodness’ of a model cannot be judged from its performance on one train set but rather it must be judged from its performance across all possible

train sets. In the bias-variance jargon, models with small RSS have a small ‘bias’.

I have already introduced the coefficient of determination in Chapter 8 (see Definition 8.5) where I have indeed discussed that, for the linear regression model, R^2 coincides with the product of the regression coefficients $R^2 = \beta_1\beta'_1$ (see Eq.8.21). Indeed, in the case of linear regression, \hat{R}^2 from Eq.18.17 is indeed the square of the sample estimate of the correlation coefficient between variable X and variable Y (i.e. $\hat{R}^2 = \hat{\rho}_{XY}^2$) but this is not the case, in general, for non-linear regressions where the \hat{R}^2 takes a broader meaning.

Remark 18.5. In Section 8.7 the non-linear version of the R^2 was called correlation ratio, η_{XY}^2 , to distinguish it from the linear case. This is the common notation in the literature for what concerns the general definition (i.e. Eq.8.63), however when it comes to its estimation most of the literature refers to it generically as R-square also for the non-linear case. Nonetheless, the empirical estimator \hat{R}^2 introduced in this section is actually the empirical estimate of the correlation ratio η_{XY}^2 . I will however maintain the \hat{R}^2 notation for coherence with the majority of the literature.

Remark 18.6. From Eq.18.17 one can derive that, for unbiased models, the \hat{R}^2 is directly proportional to Pearson’s correlation coefficient, $\hat{\rho}_{Y,\tilde{g}(X)}$, between Y and $\tilde{g}(X)$, namely

$$\hat{R}^2 = \frac{2\hat{\sigma}_Y \hat{\sigma}_{\tilde{g}(X)} \hat{\rho}_{Y,\tilde{g}(X)} - \hat{\sigma}_{\tilde{g}(X)}}{\hat{\sigma}_Y^2}. \quad (18.19)$$

This correlation coefficient, $\hat{\rho}_{Y,\tilde{g}(X)}$, is actually in itself a good alternative measure of goodness of modeling and it is used in some cases (see MCC in Section 18.5.1). This correlation coefficient $\hat{\rho}_{Y,\tilde{g}(X)}$ could be seen as an instance of the generalized dependency measure discussed in Section 8.6. It measures how well model $\tilde{g}(X)$ captures the underlying (non-linear) dependency between X and Y .

Definition 18.3. Sometimes, in processes with extreme events, when statistics are fat-tailed, the variance might be ill-defined and the **mean of absolute error** is preferable to the mean square error. This quantity is

$$\text{MAE} = \frac{1}{q} \sum_{i=1}^q |\hat{\epsilon}_i|. \quad (18.20)$$

Clearly better models have smaller residuals and correspond to smaller MAE.

There are many other measures of deviation. Normally they measure the deviation of a random variable from its mean or median. A closely related quantity to MAE is the **mean absolute deviation** (MAD) which takes the sample mean of the variable minus its mean or median. It often coincides with MAE because, for unbiased models $\hat{\mu}_\epsilon = 0$, and often the residual is symmetric therefore it has also zero median. These are dimensional quantities that depend on the unit of measure of Y and a normalization is necessary to use them for comparison with other measures. A scale-independent quantity is the **mean absolute percentage error** (MAPE), where the sample mean is taken on relative deviation, for the residual it is $100\% / q \sum_{i=1}^q |\hat{\epsilon}_i / \hat{y}_i|$.

Adjusted R-square

Models with greater complexity and a larger number of parameters are likely to better fit the training data due to overfitting. Indeed, in the case when the model is optimized by ordinary least square regression it can be proved that \hat{R}^2 calculated in-sample is a non-decreasing function of the number of parameters. There are therefore good reasons to introduce an adjusted version of \hat{R}^2 that takes into account the number of parameters in the model. One option is

$$\text{Adjusted } \hat{R}^2 = 1 - \frac{\text{RSS}}{\text{TSS}} \frac{(q-1)}{(q-n-1)} \quad (18.21)$$

where n is the number of parameters in the model and q is the number of observations. One can directly notice that for $n > 1$ the factor $(q-1)/(q-1-n)$ is larger than one and penalizes models with a larger number of parameters. One can also notice that this factor can become quite large and even diverge at $n = q-1$. It yields instead to negative Adjusted \hat{R}^2 coefficients when $n \geq q$. Vice versa, in the limit $q \rightarrow \infty$ the coefficient tends to 1 and the adjusted measure tends to \hat{R}^2 .

18.4.2 Information theoretic generalization of R-square

The concept underneath R-square is to measure the reduction in uncertainty about \mathbf{Y} provided by the knowledge of \mathbf{X} through a model $\tilde{g}(\mathbf{X})$. A good measure of uncertainty is the entropy. In analogy with the R-square, one could use directly the entropy of the residual, $H(\mathbf{Y} - \tilde{g}(\mathbf{X})) = H(\tilde{\epsilon})$ to quantify this uncertainty. However, this can often be problematic. For instance, for continuous variables, the value of the entropy has no absolute meaning, indeed it depends on the unit of measure of the variable (e.g. the Shannon entropy of a continuous variable expressed in meters is different from the entropy of the same variable expressed in millimeters). Therefore, in some contexts, the quantification of mutual information is preferable. Specifically, one might want to compare the residual

uncertainty on $\tilde{\epsilon}$ with respect to the original uncertainty on \mathbf{Y} . The goodness of the model can be quantified in terms of reduction in uncertainty on \mathbf{Y} provided by the model $\tilde{g}(\mathbf{X})$, which is the difference

$$H(\mathbf{Y}) - H(\tilde{\epsilon}) \geq H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}) = I(\mathbf{X}; \mathbf{Y}). \quad (18.22)$$

An information-theoretic generalization of the R-square is, therefore,

$$\hat{R}_I^2 = 1 - \exp(-2\hat{I}(\mathbf{X}; \mathbf{Y})). \quad (18.23)$$

where $\hat{I}(\mathbf{X}; \mathbf{Y})$ is the sample estimate of $I(\mathbf{X}; \mathbf{Y})$ obtained by using model $\tilde{g}(\mathbf{X})$). One can directly verify that for the linear regression model, it returns indeed $\hat{R}_I^2 = \hat{R}^2 = 1 - \hat{\rho}_{Y,\epsilon}^2 = \hat{\rho}_{XY}^2$ (see also Remark 8.6). It is also clear that when there is no measurable shared information between the variables $\hat{I}(\mathbf{X}; \mathbf{Y}) = 0$ and $\hat{R}_I^2 = 0$. This is a very simple and useful generalization that is directly applicable to multidimensional and non-linear cases. An advantage of \hat{R}_I^2 with respect to the direct use of the information-theoretic measure $\hat{I}(\mathbf{X}; \mathbf{Y})$ is that this measure is defined in the interval $[0, 1]$ and fulfills the Renyi's criteria for a generalized dependency measure (see Section 8.6).

Remark 18.7. In the multivariate normal case and in the broader class of elliptical distributions, R_I^2 can be estimated from the sample covariances:

$$\hat{R}_I^2 = \frac{|\hat{\Sigma}_{ZZ}|}{|\hat{\Sigma}_{\tilde{\epsilon}\tilde{\epsilon}}||\hat{\Sigma}_{YY}|} \quad (18.24)$$

with $|\hat{\Sigma}_{\tilde{\epsilon}\tilde{\epsilon}}|$, $|\hat{\Sigma}_{YY}|$ and $|\hat{\Sigma}_{ZZ}|$ respectively the determinants of the sample covariance matrices of $\tilde{\epsilon}$, \mathbf{Y} and $\mathbf{Z} = (\tilde{\epsilon}^\top, \mathbf{Y}^\top)^\top$.

Notice that, in the simplest case of only two bivariate elliptical variables, by substituting $\mathbf{Z} = (Y - \tilde{g}(X), Y)$, one retrieves Eq.18.19.

Example 18.3 (Non-linear measure of dependency). The estimate of the mutual information which is used in the dependency measure proposed in Eq.18.23 has been already computed in Example 15.2 for a non-linear bivariate synthetic dataset. Let me, therefore, use the same data and the same histogram estimates for the entropies to compute \hat{R}_I^2 :

$$\hat{R}_I^2 \simeq 1 - \exp(-2 \cdot 0.59) = 0.69. \quad (18.25)$$

Notice that, for this dataset, the linear correlation coefficient is $\hat{\rho} = 0.801$, and indeed the previous estimate gives $\sqrt{\hat{R}_I^2} = 0.832$, which, consistently, is a little larger than $\hat{\rho}$ capturing some extra contribution from nonlinearity.

18.4.3 Statistical validation of nested regression models

Definition 18.4 (Netsed models). Two models \mathcal{M}_1 and \mathcal{M}_0 are called **nested** when model \mathcal{M}_1 can be transformed into model \mathcal{M}_0 by imposing some constraints on \mathcal{M}_1 . These constraints consist typically in forcing some parameters to zero and therefore reducing the parameter set Θ_0 of model \mathcal{M}_0 to a subset of the parameter set Θ_1 of \mathcal{M}_1 .

Under the assumption that the variables are normally distributed, and for nested models, the R-square can be statistically validated by using the F-distribution (see Definition 5.11). Indeed, given two models \mathcal{M}_0 and \mathcal{M}_1 with model 1 nested within model 2. Assuming that \mathcal{M}_0 has a number of parameters p_1 and \mathcal{M}_1 as instead $p_2 > p_1$ parameters. For a set of q observations, the residual sums of squares (i.e. $\text{RSS}_k = \sum_{i=1}^q (\tilde{g}_k(\hat{\mathbf{x}}_i) - \hat{y}_i)^2$ with $\tilde{g}_k(\hat{\mathbf{x}}_i)$ the predictor of model k , with $k = 1, 2$) follow chi-square distributions and their ratio an F-distribution (see Section 5.6.3). Specifically,

$$\left(\frac{\text{RSS}_1 - \text{RSS}_2}{p_2 - p_1} \right) / \left(\frac{\text{RSS}_2}{q - p_2} \right) \sim \mathcal{F}(p_2 - p_1, q - p_2). \quad (18.26)$$

This is the core of the analysis of variance (ANOVA) test. In order to apply this result to the statistics of the R-square, one must consider model 1 as the sample mean of Y ($\tilde{g}_1(\hat{\mathbf{x}}_i) = \hat{\mu}_Y$, and therefore $p_1 = 1$), while model 2 is instead a model with n parameters ($\tilde{g}_1(\hat{\mathbf{x}}_i) = \tilde{g}(\hat{\mathbf{x}}_i)$, using previous notation, therefore $p_2 = n + 1$). In this case $\text{RSS}_1 = \text{TSS}$ and $\text{RSS}_2 = \text{RSS}$ and the formula becomes

$$\left(\frac{\text{TSS} - \text{RSS}}{n} \right) / \left(\frac{\text{RSS}}{q - n - 1} \right) = \frac{R^2}{1 - R^2} \frac{q - n - 1}{n} \sim \mathcal{F}(n, q - n - 1). \quad (18.27)$$

This validation is usually applied to linear regression models but it is more general as far as the models are nested and the variables are all normal. However, both these assumptions are rather unrealistic when non-linear modeling of real systems is concerned.

18.5 Testing the goodness of classifications

When variable Y is discrete or categorical, then the regression problem $Y = g(\mathbf{X})$ becomes a classification problem.

18.5.1 Counting right and wrong classifications

In a binary classification problem, one has two output categories (i.e. cats and dogs), and the model either predicts the right output or misclassifies it. One might associate the values 0 and 1 with the two categories (i.e. cats=0 and dogs=1).⁵

⁵ If there are more than two categories, then the classification is called multi-label or multi-class.

Some of the approaches here discussed can be directly extended to multi-class problems. Others

The counting of the number of times the output $\hat{y}_i = 1$ is guessed correctly by the model (i.e. $\tilde{g}(\hat{\mathbf{x}}_i) = 1$), is called the ‘number of true positive’ (TP). Instead, the number of times the output $\hat{y}_j = 0$ is guessed correctly (i.e. $\tilde{g}(\hat{\mathbf{x}}_j) = 0$) is called true negative (TN). The wrong guesses are instead respectively called false positive (FP) if the model guessed $\tilde{g}(\hat{\mathbf{x}}_j) = 1$ but the real value was $\hat{y}_j = 0$ and false negative (FN) if the model guessed $\tilde{g}(\hat{\mathbf{x}}_i) = 0$ but the real value was $\hat{y}_i = 1$. These are the elements of the so-called confusion matrix.

Definition 18.5 (Confusion Matrix). The **confusion matrix** summarizes the performances of a classification algorithm in terms of instances of actual classes (columns) and instances of predicted classes (rows).

		True/Actual Class		Total Predicted
Predicted Class	Positive	Positive	Negative	TP+FP
	Negative	FN	TN	
Total True		TP+FN	FP+TN	

Good models must have large TP and TN (diagonal elements) and small FP and FN (off-diagonal elements). However, the qualification of ‘good’ depends on the problem and on the balance between the number of 0 and 1 in the sample (the class imbalance). Indeed, if the model has a large imbalance (i.e. a large difference in the numbers of 0 and 1) it becomes easier to guess the most abundant class. There are several measures of goodness of the classification based on these counting. Let me here mention the few that are most used.

- The fraction of true positives (TP) relative to the total number of positives in the sample (TP+FN) is called **sensitivity**

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}}. \quad (18.28)$$

- The fraction of true negative (TN) relative to the total number of negatives in the sample (TN+FP) is called **specificity**

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}}. \quad (18.29)$$

- The number of correct guesses (TP+TN) divided by the total sample (TP+FP+TN+FN) is called **accuracy**

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}}. \quad (18.30)$$

- The fraction of true positive (TP) with respect to the total number of positive

require some adjustments. Typically one can use binary classifications either for one class vs. rest or between two classes.

guesses ($TP + FP$) is called **precision**

$$\text{Precision} = \frac{TP}{TP+FP}. \quad (18.31)$$

- The harmonic mean of precision and sensitivity is the **F1 score**

$$F1 = \frac{2TP}{2TP+FP+FN}. \quad (18.32)$$

All these measures return scores in the interval $[0, 1]$ with low scores associated with bad classifications and high scores associated with better classifications. However, they return different scores for the same classifier and the judgment on the goodness of the classification depends on the purpose of the classification itself. Indeed, there are cases where it is very important to capture the largest number of correct positives (TP large) and the miss classification of the negatives is less important, for this case, Sensitiveness is a meaningful score. However, in other contexts, it is equally important to correctly classify both positives and negatives and in such a case the Accuracy is more adequate. In all cases, class imbalance (i.e. relative number of positives and negatives in the sample) strongly affects results. For instance, in a sample with a very large abundance of positives with respect to negatives, Precision is always close to 1 independent of the goodness of the classification. There are therefore weighted, corrected, counterparts of these measures, normally referred to as ‘Adjusted’, that take into account the imbalance.

A very useful measure, which is less prone to imbalance and interpretation issues is the **Matthews correlation coefficient** (MCC) which is the Pearson correlation between the output guessed by the model and the true output:

$$MCC = \text{Corr}(\hat{y}, \tilde{g}(\hat{x})) \quad (18.33)$$

which can be expressed in terms of the previous measures

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}. \quad (18.34)$$

The MCC is recognized to be one of the best measures of being reliable also when the class imbalance is high. The MCC measure was previously known under the name of Pearson’s phi coefficient and Yule phi coefficient before Brian W. Matthews attached his name to it for the machine learning community.

Remark 18.8. For a random classifier, the probability to obtain TP successes (true positives) over q draws with replacement in a sample of size N that contains P elements of a the desired category, follows the **hypergeometric distribution**

$$p_{\text{random}}(TP) = \frac{\binom{P}{TP} \binom{N-P}{q-TP}}{\binom{N}{q}}. \quad (18.35)$$

Drawing with replacement is not the most typical practice in classification, however, this probability can be used to quantify the over-representation or under-representation of some categories in a sample and this can be very useful in clustering models.

18.5.2 Cross-entropy

Often models produce the probability of the category for a given input. For example, if the classification concerns distinguishing between images of dogs and cats, the input $\hat{\mathbf{x}}$ is an image, and the output is the probabilities $\tilde{p}(Y = \text{dog}|\hat{\mathbf{x}})$ and $\tilde{p}(Y = \text{cat}|\hat{\mathbf{x}})$. The previous binary output is obtained from the model's conditional probability by thresholding with a scalar $\theta \in [0, 1]$, normally set at 0.5

$$\tilde{g}(\hat{\mathbf{x}}_i) = \begin{cases} 1 & \text{if } \tilde{p}(Y = 1|\hat{\mathbf{x}}_i) \geq \theta \\ 0 & \text{otherwise} \end{cases}. \quad (18.36)$$

To quantify the goodness of the classification one can consider minimizing a distance between the true model $p_{\text{true}}(\mathbf{y}|\hat{\mathbf{x}})$ and the approximated model $\tilde{p}(\mathbf{y}|\hat{\mathbf{x}})$. Such a distance can be quantified with the Kullback-Leibler divergence (see Section 7.4)

$$\hat{D}_{KL}(p_{\text{true}}\| \tilde{p}) = \sum_{\mathbf{y} \in \Omega_Y} p_{\text{true}}(\mathbf{y}|\hat{\mathbf{x}}) \log \frac{p_{\text{true}}(\mathbf{y}|\hat{\mathbf{x}})}{\tilde{p}(\mathbf{y}|\hat{\mathbf{x}})}, \quad (18.37)$$

which is equal to zero when the model coincides with the true one and becomes larger than zero when the two deviate. Measuring the goodness of models in terms of Kullback-Leibler divergence with respect to the true model is a standard and logical way to proceed, not only in the context of classification. Indeed, it is easy to verify that this measure is equal to the negative log-likelihood plus a constant.

The true model is in general unknown, however, the expected value can be estimated by using the sample mean and the estimate of Eq.18.37 is

$$\hat{D}_{KL}(p_{\text{true}}\| \tilde{p}) = \sum_{k=1}^q \log \frac{p_{\text{true}}(\hat{\mathbf{y}}_k|\hat{\mathbf{x}}_k)}{\tilde{p}(\hat{\mathbf{y}}_k|\hat{\mathbf{x}}_k)} = \sum_{k=1}^q \log p_{\text{true}}(\hat{\mathbf{y}}_k|\hat{\mathbf{x}}_k) - \sum_{k=1}^q \log \tilde{p}(\hat{\mathbf{y}}_k|\hat{\mathbf{x}}_k). \quad (18.38)$$

The first term is independent from the model, while the second term

$$\hat{H}(p_{\text{true}}, \tilde{p}) = \sum_{k=1}^q \log \tilde{p}(\hat{\mathbf{y}}_k|\hat{\mathbf{x}}_k), \quad (18.39)$$

is the one that the model aims to maximize. This is the sample estimate of (minus) the cross entropy (see Definition 7.4) and it coincides with the log-likelihood of the conditional probability. Therefore, models that minimize the cross entropy are maximizing the likelihood.

Definition 18.6 (ROC curve and AUC). It is important to recall that the classification is, for most of the modeling, the output is a thresholding operation of the estimated output probability ($\tilde{p}(y|\hat{\mathbf{x}})$), and therefore it depends on the threshold value θ . In order to assess such a dependency, one can run several classifications with different threshold values between 0 and 1 and then plot a graph with on the y-axis the sensitivity and on the x-axis the false positive rate ($FP/(TN+FP)$), that is given by 1-Specificity. This is called the **ROC curve** (acronym for Receiver Operating Characteristic). If the probability $\tilde{p}(y|\hat{\mathbf{x}})$ is random and the model has, therefore, no capability of predicting the correct category of Y given $\hat{\mathbf{x}}$, then the ROC curve is a diagonal straight line. Conversely, if the model truly reflects the conditional likelihood, then all points must concentrate on the upper left corner where Sensitivity = 1 and the false positive rate is zero. The deviation above the diagonal line is considered an indication of the goodness of the classifier.

The acronym ROC stands for ‘receiver operating characteristic’ and derives from its first use for operators of military radar receivers. It must be noticed that in this computation only the threshold changes while the other model parameters are left unchanged.

The **area under the curve (AUC)** is a related measure that summarizes the performance of the classifier across all threshold values. It consists of the estimation of the area under the ROC curve. The true model, with perfect classification, must have AUC equal to 1, a random classification has instead $AUC = 1/2$, while a terrible classification could even reach $AUC = 0$.

18.6 Model evaluation via likelihood

From a probabilistic perspective, a regression consists in estimating the conditional expected value $\tilde{g}(\hat{\mathbf{x}}) = \mathbb{E}(Y|X = \hat{\mathbf{x}})$,⁶ while a classification is a thresholding of the estimate of the conditional probability $P(Y|X = \hat{\mathbf{x}})$. Independently on the task, good modeling demands a good estimation of the probability distribution function of the set of variables.

Following intuition, for a given set of observations, $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, \dots, \hat{x}_q)^\top$, if a hypothesis returns a larger probability than another for that observation (i.e. a larger likelihood $P(\hat{\mathbf{x}}|Hypothesis)$), then it describes better the observed data, and it must be therefore considered preferable.

Before proceeding any further let me identify more precisely what I mean with ‘*Hypothesis*’. This is a rather complex and rich concept. For parametric probabilistic modeling, the hypothesis comprises two parts: a model \mathcal{M} and its parameters $\boldsymbol{\theta}$. For non-parametric probabilistic modeling, the model \mathcal{M} could also comprise a procedure and the parameter set $\boldsymbol{\theta}$ might include hyperparameters’

⁶ When the variance of the regression error is minimized.

choices (i.e. the bin size for histograms). Hereafter, I'll write the likelihood of a set of observations $\hat{\mathbf{x}}$ described by a model \mathcal{M} with parameter set $\boldsymbol{\theta}$ with $P(\hat{\mathbf{x}}|\mathcal{M}, \boldsymbol{\theta})$ for both discrete and continuous variables and for both parametric and non-parametric modeling.

The ultimate purpose of modeling is to discover the true distribution $f_{true}(\mathbf{x})$ which generated the observed data. The data-driven-modeling process consists in finding a good approximation $\tilde{f}(\mathbf{x}|\mathcal{M}, \boldsymbol{\theta})$ which is as close as possible to the true distribution. A measure of distance between two probability distributions is the Kullback-Leibler divergence (see Definition 7.2)

$$D_{KL}(f_{true} \parallel \tilde{f}) = \int_{\Omega_X} f_{true}(\mathbf{x}) \log \frac{f_{true}(\mathbf{x})}{\tilde{f}(\mathbf{x}|\mathcal{M}, \boldsymbol{\theta})} d\mathbf{x} = \mathbb{E}(\log f_{true}(\mathbf{x})) - \mathbb{E}(\log \tilde{f}(\mathbf{x}|\mathcal{M}, \boldsymbol{\theta})). \quad (18.40)$$

In this expression for the Kullback-Leibler divergence the first term is independent of the model, therefore the model \mathcal{M}^* and parameters $\boldsymbol{\theta}^*$ that minimize such divergence are the ones that maximize the second term

$$(\mathcal{M}^*, \boldsymbol{\theta}^*) = \sup_{\mathcal{M}, \boldsymbol{\theta}} \mathbb{E}(\log \tilde{f}(\mathbf{x}|\mathcal{M}, \boldsymbol{\theta})). \quad (18.41)$$

In other terms, good models maximize the expected value of the log-likelihood.

The true distribution is in general unknown, therefore the expected value cannot be computed exactly. However, it can be estimated from observations. Indeed, the expected value of the log-likelihood can be approximated with arbitrary precision by using the sample mean (see the law of large numbers Section 13.3).

Given a set of observations, $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, \dots, \hat{x}_q)^\top$, if they are i.i.d. the joint probability of the set of observations becomes equal to the product of the marginal probabilities of each observation, and therefore

$$\log P(\hat{\mathbf{x}}|\mathcal{M}, \boldsymbol{\theta}) = \sum_{k=1}^q \log P(\hat{x}_k|\mathcal{M}, \boldsymbol{\theta}). \quad (18.42)$$

Over an (infinitely) large set of observations, the sum converges towards q times the expected value of the likelihood of each observation, which is in turn equal to the expected value of the joint log-likelihood itself

$$\log P(\hat{\mathbf{x}}|\mathcal{M}, \boldsymbol{\theta}) \underset{q \gg 1}{\sim} q \mathbb{E}(\log P(x|\mathcal{M}, \boldsymbol{\theta})) = \mathbb{E}(\log P(\mathbf{x}|\mathcal{M}, \boldsymbol{\theta})). \quad (18.43)$$

Therefore, in the asymptotic limit, when the observation set size becomes infinite and for i.i.d. observations, maximization of the likelihood for the entire set of observations coincides with the maximization of the expected value of the log-likelihood. For non-i.i.d. observations, for instance, in the case of a non-stationary process, there are memory effects and path dependency, and normally the problem can be addressed by estimating expected values over sets of observations.

The maximization of the likelihood over any finite set of data (the train set) leads unavoidably to models that perform on average better on the data used for the maximization than over a dataset not seen by the model before. To assess the

general goodness of a model, its performance must be instead evaluated over all possible observations. Indeed, it is the expected value that must be maximized. To better estimate the expected value of the likelihood it is good practice to divide the available data into train, validation, and test sets (see Section 3.5.1 and also, later in this Chapter, Section 18.9). The train set $\hat{\mathbf{x}}_{train}$ is used to learn the optimal parameter set $\boldsymbol{\theta}^*$ (which is however optimal for this set and not in general). Hyperparameters, model architectures, and the choice between competing models (which are forms of hyper-parameters) can be established on the validation set $\hat{\mathbf{x}}_{valid}$. The best model, \mathcal{M}^* , can be assessed from the performances on the test set $\hat{\mathbf{x}}_{test}$. The likelihood can be used to assess model goodness both in-sample (train set $\hat{\mathbf{x}}_{train}$) and out-of-sample (validation $\hat{\mathbf{x}}_{valid}$ and test on $\hat{\mathbf{x}}_{test}$ sets). One expects that a well-learned, well-selected model must perform out-of-sample in a comparable way as in-sample, and models that have statistically significant better performances on the train set should also outperform the other competing models on the validation and test sets. Such out-of-sample performances are the ultimate test for model evaluation and selection.

The value of the likelihood in itself has little meaning, especially in the case of continuous variables. Indeed, while for discrete variables the likelihood is a probability, for continuous variables it is instead a density and it is therefore a quantity that depends on the unit of measure of the variable. What is meaningful is instead to compare, for the same set of observations $\hat{\mathbf{x}}$, two values of likelihoods associated with two different models or associated with the same model with different parameters. In particular, their ratio or, equivalently, the difference of their logarithms are meaningful, dimensionless quantities.

18.6.1 Likelihood ratio

Consider two hypotheses for a parametric model \mathcal{M} where one is formulated over a restricted subset of parameters $\boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$. These two models are called nested (see Definition 18.4).

To quantify the difference between the two hypotheses one can use the quantity

$$\lambda_{LR} = -2 \ln \left(\frac{\sup_{\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0} P(\hat{\mathbf{x}} | \mathcal{M}, \boldsymbol{\theta}_0)}{\sup_{\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}} P(\hat{\mathbf{x}} | \mathcal{M}, \boldsymbol{\theta}_1)} \right). \quad (18.44)$$

The restricted model, the null hypothesis, cannot have a larger (in-sample) supremum of the likelihood than the full model and therefore λ_{LR} is non-negative. Under some restrictive conditions, one can associate a probability to the likelihood ratio λ_{LR} . Indeed, Wilks' theorem [Wilks, 1934] guarantees asymptotical convergence of λ_{LR} to a χ^2 distribution with degrees of freedom given by the difference in the dimensionalities of $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}_0$. However, the two models must be nested and the supremum values of the estimated best parameters $\boldsymbol{\theta}_0^*$ and $\boldsymbol{\theta}^*$ must be discoverable and lie within the interior of the parameter space and not be extrema. Under these conditions, in the asymptotic limit, the likelihood

ratio can be quantified in terms of probability yielding confidence intervals and p-values for the null hypothesis. This is called the **likelihood ratio test**. When the best values of the parameters are known, the Likelihood ratio is proved to be the most powerful among all other possible statistical tests [Neyman and Pearson, 1933]. However, the supremum for the parameter set is not always discoverable, indeed many of the optimization methodologies do not guarantee convergence. Therefore, in practice, there is often the risk that a model appears better than another only because the other has been poorly estimated.

18.6.2 Relative Likelihood

Although the likelihood ratio test and Wilks' theorem are of relatively limited applicability, it rests nonetheless true that models with larger likelihoods are better at describing the data and they are therefore in general preferable.

Let me consider, more generally, two models \mathcal{M}_0 and \mathcal{M}_1 with two parameter sets $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$. For a given set of observations $\hat{\mathbf{x}}$, the two values of the likelihoods $P(\hat{\mathbf{x}}|\mathcal{M}_0, \boldsymbol{\theta}_0)$ and $P(\hat{\mathbf{x}}|\mathcal{M}_1, \boldsymbol{\theta}_1)$ quantify how well the models and their parameters describe the observations. In order to compare two likelihoods, one can take their ratio:

$$\lambda = \frac{P(\hat{\mathbf{x}}|\mathcal{M}_0, \boldsymbol{\theta}_0)}{P(\hat{\mathbf{x}}|\mathcal{M}_1, \boldsymbol{\theta}_1)}. \quad (18.45)$$

A model with a larger likelihood describes better the dataset than a model with a smaller one. Therefore, this ratio, or rather its logarithm, which is the difference between the two log-likelihoods, can be used to quantify relative model performances. However, for model selection, there are other elements that must be taken into account and one of them is the complexity of the model.

18.6.3 Akaike's Information Criterion (AIC)

If two models have similar likelihoods but the one with a larger likelihood is more complex and has more parameters, one might be tempted to adopt the simpler model even if it has a smaller likelihood. Akaike proposed a simple information-theoretic criteria to take into account model complexity. If a model has n parameters then the AIC value is

$$\text{AIC} = 2n - 2 \ln(\sup_{\boldsymbol{\theta} \in \Theta} P(\hat{\mathbf{x}}|\mathcal{M}, \boldsymbol{\theta})). \quad (18.46)$$

Therefore AIC is proportional to the negative log-likelihood penalized by the number of parameters in the model.⁷ A model with a smaller AIC is preferred to a model with a larger AIC. Therefore, AIC selects models with larger values of the likelihood but it penalizes them for their number of parameters.

⁷ The reasoning beyond this criterion is that the log-likelihood is an estimate of the reduction of uncertainty provided by the model. In terms of bits, a model with more parameters is provided with extra information which can be accounted for as one bit per parameter.

Remark 18.9. If one has two models and therefore two values of AICs with $AIC_0 < AIC_1$ the quantity

$$\exp \frac{AIC_0 - AIC_1}{2} \propto \lambda \quad (18.47)$$

is the relative likelihood introduced in Eq.18.45 when the two models have the same number of parameters, otherwise it is proportional to it.

Remark 18.10. The value of the log-likelihood increases proportionally to the observation size and therefore for large sample sizes the penalization with the number of parameters becomes irrelevant and $AIC \xrightarrow[q \rightarrow \infty]{} -2 \ln(\sup_{\theta \in \Theta} P(\hat{\mathbf{x}}|\mathcal{M}, \boldsymbol{\theta}))$.

When the two models are nested and the number of observations tends to infinity then the difference $AIC_0 - AIC_1$ is the likelihood ratio.

Adjusted AIC for small samples: AICc

For small sample sizes (q of the same order of n), when computed in-sample, AIC tends to select models with more parameters because these models are overfitting and therefore have larger in-sample likelihoods. Therefore, a correction to penalize further the number of parameters for small sample sizes has been proposed:

$$AICc = 2 \frac{qn}{q - n - 1} - 2 \ln(\sup_{\theta \in \Theta} P(\hat{\mathbf{x}}|\mathcal{M}, \boldsymbol{\theta})). \quad (18.48)$$

Which can be rewritten as $AICc = AIC + 2(n^2 + n)/(q - n - 1)$. The penalization diverges when q reaches $n + 1$ which is the point where overfitting models can interpolate the data returning the largest in-sample likelihoods. In the opposite limit, for large sample sizes, $q \rightarrow \infty$, this criteria coincide with AIC.

18.6.4 Bayesian information criterion (BIC)

From a Bayesian perspective, one aims to find the best model irrespective of the parameters' choice and this can be achieved by maximizing the marginal likelihood

$$P(\hat{\mathbf{x}}|\mathcal{M}) = \int_{\theta \in \Theta} P(\hat{\mathbf{x}}|\mathcal{M}, \boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta}. \quad (18.49)$$

However, the probability distribution of the parameters (the prior $P(\boldsymbol{\theta}|\mathcal{M})$) is in general unknown. One can nonetheless seek an approximate expression that holds for a class of prior distributions. It was noticed by Schwarz [1978] that, in the large sample limit, the main contribution to the marginal log-likelihood, $\log P(\hat{\mathbf{x}}|\mathcal{M})$, is the maximum log-likelihood $\log P(\hat{\mathbf{x}}|\mathcal{M}, \boldsymbol{\theta}^*)$ and, under some not too restrictive assumptions, for large q , the extra contribution is approximated by $-\frac{1}{2}n \ln q$ (with

n the number of parameters in \mathcal{M}). Therefore, another quantification of the goodness of a model, which takes into account the tradeoff between likelihood and model dimension, is the following Bayesian information criterion

$$\text{BIC} = n \ln q - 2 \ln(\sup_{\theta \in \Theta} P(\hat{\mathbf{x}}|\mathcal{M}, \theta)). \quad (18.50)$$

The coefficient for the penalization for the number of model parameters is here an increasing function of the size of the sample ($\ln q$). Interestingly, this is opposite to the AICc criterion where the penalization is instead a decreasing function of q . However, the increment of the penalization is in $\log q$ while the expected change of the log-likelihood is in q (precisely, from Eqs.18.42 and 18.43, it is $q\mathbb{E}(\ln P(\hat{\mathbf{x}}|\mathcal{M}, \theta))$). Therefore, in the asymptotic limit, one expects the unrestricted maximum likelihood solution to prevail for both criteria.

The underlying quantity that is gauged by both criteria is the model likelihood, but clearly AIC and BIC criteria are different and they are appropriate for different purposes. It is often argued that BIC is more appropriate in situations where the sample size is limited, while AIC is better suited for cases where the sample size is large. However, there is no strict rule for when to use one criterion over the other, and sometimes it may be beneficial to consider both BIC and AIC to gain a more comprehensive understanding. Additionally, other factors, such as the specific problem at hand and the theoretical background, should also be taken into account.

Another important approach for model selection consists in estimating the model's likelihood out-of-sample. This changes the perspective because over-complicated-overfitting models might result in large in-sample likelihoods but they get automatically penalized by the out-of-sample likelihood estimation. In this case a ‘Bayesian’ approach might consist in assessing the likelihood for a range of parameters and hyper-parameters, possibly with some prior ansatz on their range and distribution. This is shown and discussed in the following example.

Example 18.4 (Comparison between likelihoods from different models). In this example, I compare likelihoods from different models, parametric and non parametric and both in-sample and out-of-sample.

Let me first consider the model described in Example 13.6 (see also Fig.13.4) where I use non-parametric modeling tools, namely the histogram and the kernel density estimation (KDE), to model a synthetic dataset. In that example, I compare models with different KDE’s bandwidth h , showing its effect on the distribution. The qualitative outcome from the observation of the plots indicated that $h \sim 5$ is a good parameter choice for KDE, while histogram bin size was directly fixed at 5 (note that these two quantities are comparable but not the same). A cross-validation maximum likelihood estimation approach to establish the KDE bandwidth parameter was introduced in Example 14.6 obtaining $h^* = 3.53$ via Eq.14.21 (see also Fig.14.2),

while an Expectation Maximization approach, described in Example 14.8 and Fig.14.4 returned the best estimation Gaussian mixture model.

Let me first compare directly for all these models the likelihoods per observation computed in-sample on the training dataset $\hat{\mathbf{x}}^{tra}$. Results are reported in the following table.

Model	Parameters	log-likelihood (in-sample)
Histogram	bin size = 2	-3.84
	bin size = 3	-3.88
	bin size = 5	-3.96
	bin size = 10	-4.01
KDE cross validation	$h^* = 3.53$	-3.99
Expectation Maximization	$\mu_1^* = -15.8, \sigma_1^* = 8.25$	
	$\mu_2^* = 12.9, \sigma_2^* = 6.04$	
	$p_1^* = 0.67, p_2^* = 0.33$	-4.01
True model	$\mu_1 = -15, \sigma_1 = 10$ $\mu_2 = +15, \sigma_2 = 5$ $p_1 = 0.70, p_2 = 0.30$	-4.04

It can be noticed that all log-likelihood values are very similar. As it should be expected, in this in-sample estimation, the histogram method with the smallest bin sizes is the best performing, while the true model is the worst performing. This might sound counter-intuitive but it is instead what is normally observed. Indeed, as a consequence of overfitting, models with a larger number of parameters and more flexibility to adapt to observations (i.e. the histogram with a small bin size, in this example) return the largest likelihood on training data over-performing the true model. In this example, we can directly observe that, for non-nested models, the overfitting is related to the adaptability of the model to the data rather than the number of parameters. Indeed, for instance, the expectation maximization has the largest number of parameters but it is the worst performing among the estimated solutions.

Let me then compute the likelihoods per observation on an out-of-sample validation dataset, $\hat{\mathbf{x}}^{val}$, never seen by the models for training purposes. Let me consider $q_{val} = 10,000$ observations in order to have a good statistics.

An adjustment must be made to the previous equal bin histogram estimate because, out of sample, it would almost surely return $-\infty$ as a consequence of data in $\hat{\mathbf{x}}^{val}$ that lie outside the bins set by $\hat{\mathbf{x}}^{tra}$. To tackle this, one can use non-equal bins and in particular add a padding making the two extreme bins larger. However, the size of such a padding is arbitrary and, for the same observation set, the log-likelihood decreases by increasing padding size. For this example, I use the smallest padding size that includes all out-of-sample data. This is however an important information leakage and

results can only be interpreted as upper limits for this histogram method. Out-of-sample log-likelihood results are reported in the following table.

Model	Parameters	log-likelihood (out-of-sample)
Histogram	bin size = 2	$-\infty$
	bin size = 3	≤ -4.176
	bin size = 5	≤ -4.150
	bin size = 10	≤ -4.154
KDE cross-validation	$h^* = 3.53$	-4.099
Expectation Maximization	$\mu_1^* = -15.8, \sigma_1^* = 8.25$	
	$\mu_2^* = 12.9, \sigma_2^* = 6.04$	-4.092
	$p_1^* = 0.67, p_2^* = 0.33$	
True model	$\mu_1 = -15, \sigma_1 = 10$	
	$\mu_2 = +15, \sigma_2 = 5$	-4.066
	$p_1 = 0.70, p_2 = 0.30$	

Let me notice that the ranking of the results is now reverted. The true model has the largest out-of-sample likelihood (as it must), followed by the Expectation Maximization, then the KDE cross-validation, and finally the histogram models. Also within the histogram models ranking is changed with the small bin sizes becoming the worst performing and with the largest likelihood achieved at bin size = 5. A grid-search KDE's bandwidth associated with the largest out-of-sample likelihood returns $h^* = 4.20$ which slightly improves the likelihood of the cross-validation estimate yielding to -4.097.

Let me point out that the likelihood values are very similar across all models and performances must be compared on the third decimal. For this example, AIC, AICc, and BIC criteria do not change significantly results. From these data and models, although the relative differences in log-likelihood are quite small, one can figure out that EM does better than KDE and that histograms must have bin sizes around 5. This is exactly the same conclusion reached qualitatively by observing the behavior of these models in Fig.14.2 (see also Example 14.6), but now this conclusion is supported by a quantitative tool.

^a This changes also the estimation of the in-sample likelihood that becomes -3.89, -3.97, -4.03, -4.04 respectively for bin sizes 2, 3, 5, 10.

18.7 Model selection

The likelihood

$$P(\hat{\mathbf{x}}|Hypothesis), \quad (18.51)$$

quantifies the probability to observe the data $\hat{\mathbf{x}} = (\hat{x}_1, \dots, \hat{x}_q)^\top$ under a given *Hypothesis* (e.g. a model \mathcal{M} and its parameters $\boldsymbol{\theta}$). However, to select the best model, one would rather directly quantify the probability of the hypothesis under the observation evidence:

$$P(\text{Hypothesis}|\hat{\mathbf{x}}). \quad (18.52)$$

These two probabilities are strictly related through the Bayes' formula

$$P(\text{Hypothesis}|\hat{\mathbf{x}}) = \frac{P(\hat{\mathbf{x}}|\text{Hypothesis})P(\text{Hypothesis})}{P(\hat{\mathbf{x}})}. \quad (18.53)$$

In Bayesian terminology

$$P(\text{Hypothesis}|\hat{\mathbf{x}}) \quad \text{is the } \mathbf{posterior} \text{ probability, while the} \quad (18.54)$$

$$P(\hat{\mathbf{x}}|\text{Hypothesis}) \quad \text{is the } \mathbf{likelihood}, \text{ and} \quad (18.55)$$

$$P(\text{Hypothesis}) \quad \text{is the } \mathbf{prior} \text{ probability.} \quad (18.56)$$

In Bayesian reasoning, the prior probability $P(\text{Hypothesis})$ describes the uncertainty about the model, and the posterior $P(\text{Hypothesis}|\hat{\mathbf{x}})$ describes instead the remaining uncertainty after the observations have been considered. This is the relevant quantity for model selection. However, priors are in general unknown and in most cases the only quantity in Eq.18.53, related to the dataset that can be computed and maximized is the likelihood $P(\hat{\mathbf{x}}|\text{Hypothesis})$. Indeed, $P(\hat{\mathbf{x}})$ is a scalar independent from the hypothesis while $P(\text{Hypothesis})$ is a subjective term. The *Hypothesis* term comprises two parts, \mathcal{M} and $\boldsymbol{\theta}$, and in some circumstances, one wants to assess the goodness of a model irrespective of the parameter's choice. While the maximization of $P(\mathcal{M}, \boldsymbol{\theta}|\hat{\mathbf{x}})$ maps directly into the likelihood-based approaches (see previous Section); instead, Bayesian approaches focus on the maximization of the marginal

$$P(\mathcal{M}|\hat{\mathbf{x}}) = \int_{\boldsymbol{\theta} \in \Theta} P(\mathcal{M}, \boldsymbol{\theta}|\hat{\mathbf{x}}) d\boldsymbol{\theta}. \quad (18.57)$$

Given a set of competing models \mathcal{M}_k , from the Bayesian model selection perspective, one would choose the model that maximizes the marginal posterior probability

$$\sup_{\mathcal{M}_k} P(\mathcal{M}_k|\hat{\mathbf{x}}), \quad (18.58)$$

given the set of evidences $\hat{\mathbf{x}}$.

The comparison between two competing models, \mathcal{M}_1 and \mathcal{M}_2 , can be quantified by comparing the two probabilities taking the ratio

$$\frac{P(\mathcal{M}_1|\hat{\mathbf{x}})}{P(\mathcal{M}_2|\hat{\mathbf{x}})} \quad (18.59)$$

Using (twice) the Bayes' formula one has

$$P(\mathcal{M}_k|\hat{\mathbf{x}}) = P(\mathcal{M}_k, \hat{\mathbf{x}})P(\hat{\mathbf{x}}) = P(\hat{\mathbf{x}}|\mathcal{M}_k)P(\mathcal{M}_k)P(\hat{\mathbf{x}}) \quad (18.60)$$

and the ratio becomes

$$\frac{P(\mathcal{M}_1|\hat{\mathbf{x}})}{P(\mathcal{M}_2|\hat{\mathbf{x}})} = \frac{P(\hat{\mathbf{x}}|\mathcal{M}_1)}{P(\hat{\mathbf{x}}|\mathcal{M}_2)} \times \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)}. \quad (18.61)$$

Where the marginal likelihood is given by Eq.18.49. In Eq.18.61, the first term, containing the ratio between the two likelihoods given the models, is called the Bayes factor

$$K = \frac{P(\hat{\mathbf{x}}|\mathcal{M}_1)}{P(\hat{\mathbf{x}}|\mathcal{M}_2)}. \quad (18.62)$$

The second term in Eq.18.61 is instead the ratio between the prior probability of the models, ‘the model’s prior odds’. Therefore, in Bayesian’s model selection terminology, one has that the ‘the model’s posterior odds’ are retrieved from the ‘the model’s prior odds’ by multiplying by the Bayes factor K . Starting from some prior beliefs about the model, through the average observation likelihoods in the Bayes factor, one obtains a refined estimate of the relative model probability given the observations.

Usually, this Bayesian approach to model selection requires the knowledge of priors for both the models (to compute the Bayes factor, Eq.18.62) and the priors for the parameters (to compute $P(\hat{\mathbf{x}}|\mathcal{M}_k)$, see Eq.18.49). These are, in general, unknown and results depend on their choices. In data-driven modeling, one would rather avoid the reliance on subjective choices and use instead so-called non-informative priors. Indeed, in the absence of extra information, one can assign the same probability to all models (uniform distribution). With uniform priors probabilities, the posterior is proportional to the marginal likelihood, and the comparisons between the two models (see Eq.18.61) reduce to the Bayes factor. For parameter priors, uniform distribution cannot be used, unless a range of accessible parameters is known. In the alternative, a common choice is a prior parameter distribution with a peak at the maximum likelihood solution and then with an arbitrary variance. Whoever, allowing too much prior dispersion can result in smaller estimates of the posterior model probability $P(\mathcal{M}_k|\hat{\mathbf{x}})$ and can affect model selection choices.

18.8 Non-parametric validation of models

There is a profusion of statistical tests devised to associate a probability to measures of goodness of models and therefore to provide a p-value or a confidence interval. In this Chapter, I have already mentioned a few. However, all these tests rely on assumptions that are not always valid and they are often unverifiable. All these significance measures come nowadays together with software packages and I leave it to the readers to make wise use of them being mindful of the issues with p-values and statistical tests, mentioned earlier in this Chapter.

Let me here instead focus on non-parametric approaches which have the great advantage to rely less on assumptions and therefore have broader applicability. They however are more computationally intensive. The validation procedures I

shall introduce in this section are very general and they can be applied to any goodness measure or any parameter estimation.

In general, there are two main ways to validate models non-parametrically. The first approach uses observational data to generate a set of null models and create therefore null-model statistics that can be compared with the outcomes of the model under test. The second approach uses instead the model itself to and uses different sampling of the observation dataset to quantify intervals of confidence for the model outcomes and its parameters.

18.8.1 Shuffling data to generate null-models statistics

A nonparametric way to create null hypothesis for testing a given model is to compare the outcomes of the model on the observational data with the outcomes on a dataset that is derived from these observations but randomizes relations and maps between the variables. This is an extremely simple procedure and it can be achieved by randomly and independently shuffling (see Example 18.5) the observation order of the variables. This creates a new dataset that has the same entries and the same marginal statistics for each variable but the relations between the variables have been eliminated. This shuffling procedure can be applied to any multivariate dataset but also to any set for regression or classification. If the model task is a regression or a classification (supervised learning, i.e. learning the model $\mathbf{Y} = g(\mathbf{X})$ from a set of examples $(\hat{\mathbf{x}}_s, \hat{\mathbf{y}}_s)$ with $s = 1, \dots, q$), such randomization could shuffle the outputs $\hat{\mathbf{y}}_s$ by substituting the indices s with random indices breaking the correspondence with the inputs $\hat{\mathbf{x}}_s$. If the model instead concerns the inference structure of a multivariate system (unsupervised learning) then the shuffling is between the indices i, j of each couple of variables $\hat{\mathbf{x}}_i$ and $\hat{\mathbf{x}}_j$. Some constrained randomization that eliminates some relations but maintains others can be devised. One of the advantages of this method is that, while the dependency relations between variables are eliminated, all other properties of the variables are maintained. Indeed, data are not changed only their relative order is randomized. There is a combinatorially large number of different shuffling for each variable, therefore one can create rich statistics by computing null-hypothesis model outcomes on many different reshuffling.

Example 18.5 (Shuffling). Let me illustrate the shuffling procedure with an example with two variables X_1 and X_2 and $q = 3$ observations. Namely, $\hat{\mathbf{x}}_1 = (\hat{x}_{1,1}, \hat{x}_{2,1}, \hat{x}_{3,1})$ and $\hat{\mathbf{x}}_2 = (\hat{x}_{1,2}, \hat{x}_{2,2}, \hat{x}_{3,2})$. Shuffling consists in creating a sequence where the observation order is randomized independently for the two variables. For instance $\hat{\mathbf{x}}_1^{shuffled} = (\hat{x}_{3,1}, \hat{x}_{1,1}, \hat{x}_{2,1})$ and $\hat{\mathbf{x}}_2^{shuffled} = (\hat{x}_{2,2}, \hat{x}_{3,2}, \hat{x}_{1,2})$. In this simple example, there are 6 combinations of these shuffling that do not reproduce the original match between observations.

This shuffling procedure can be applied to assess the reliability of any measure, S , associated with any model. The statistical distribution of the shuffled values

$S^{shuffle}$ can reveal if the measured outcome, S , is likely to be the outcome of a random chance or instead, it is specific to a special kind of relationship between the variables, a pattern, that the model is therefore identifying. The measure can be any relevant quantity such as the likelihood, or the R^2 , the AIC, the precision, the transfer entropy, or any other estimate of a model's parameter such as the correlation coefficient. By convention, hereafter I shall consider that larger values of the measure S are more significant (the conjugate measure can be used otherwise).

Standard score or z-score

Suppose one has created $N^{shuffle}$ samples and the corresponding set of null-model measures $S_i^{shuffle}$ with $i = 1, \dots, N^{shuffle}$. If $\hat{\mu}_S^{shuffle}$ is the estimated sample average and $\hat{\sigma}_S^{shuffle}$ is the estimated standard deviation of the shuffled measures (assuming they are defined). The relative significance of the model measure S with respect to the measures outgoing from the null-shuffled models can be quantified with:

$$z = \frac{S - \hat{\mu}_S^{shuffle}}{\hat{\sigma}_S^{shuffle}}. \quad (18.63)$$

It is clear that if the measure S is significantly larger than $\hat{\mu}_S^{shuffle}$, then this means that it is unlikely it is the result of a random chance and the null hypothesis can be excluded. Dividing by $\hat{\sigma}_S^{shuffle}$ makes the deviation in terms of the standard deviation, that is an estimate of the interval of variability of $S_i^{shuffle}$, and it has the advantage of making z non-dimensional and scale-invariant. Large values of z are unlikely outcomes of the null model and therefore large values of z (typically $z > 3$) indicate that the model under consideration is likely to better describe the observations than the null model. Indeed, we have seen in Section 5.5.1 that large deviations from the mean are unlikely for any kind of statistics with defined mean and variance. Chebyshev's inequality establishes a bound indicating that the likelihood must be smaller than $1/z^2$. For instance, $z > 3$ has a likelihood to happen under the null hypothesis which must be below $1/9$ and usually, it is much smaller than that. In the cases where the variance $\sigma_S^{shuffle}$ is not defined, one can use equivalent measures of deviations such as replacing $\hat{\sigma}_S^{shuffle}$ in Eq.18.63 with the mean absolute value of the deviation. In unusual cases where the mean is not defined, this methodology cannot be used directly and variables must be transformed.

Sample p-values

A quantitative estimate of the likelihood of a model measure S with respect to the null hypothesis can be obtained by estimating empirically the conjugate cumulative distribution $P(S^{shuffled} > S)$.

$$\hat{p}_{value} = \frac{\sum_{i=1}^{N^{shuffle}} \mathbf{1}_{S \leq S_i^{shuffle}}}{N^{shuffle}}, \quad (18.64)$$

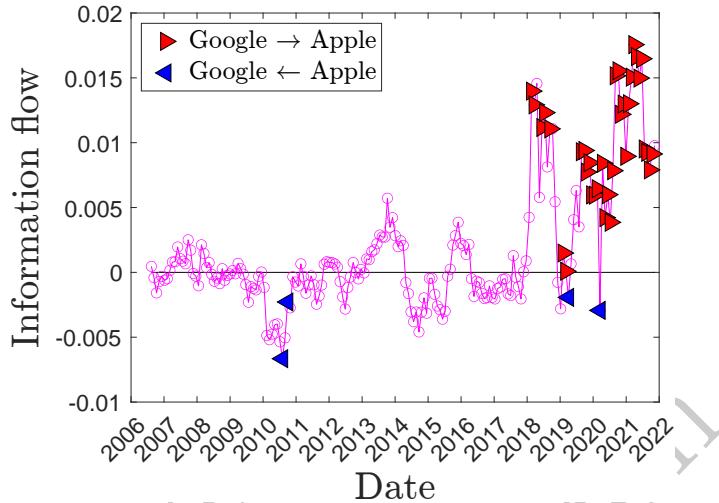


Figure 18.3 Information flow ($T_{\text{Google} \rightarrow \text{Apple}} - T_{\text{Google} \leftarrow \text{Apple}}$) computed from daily log-returns of Google (Alphabet Inc.) and Apple (Apple Inc.) stock prices from Nasdaq with a rolling window of 500 days in the period 2004-2021. The filled triangles correspond to periods where causality is detected at significance levels $z > 4$ and sample p-values are below 0.1%. Right pointing triangles correspond to Google causing Apple, instead left pointing triangles indicated Apple causing Google. Transfer entropies are computed at lags 1 to 7 days and the largest are chosen.

which is, indeed, the sample estimate of the p-value for the null hypothesis that the shuffled measure returns a larger score than S (the sum over $\mathbf{1}_{S_i^{shuffled} > S}$ counts the number of times $S_i^{shuffled} > S$). As a rule of thumb, for a meaningful estimation of a p-value at a given p_v level, one needs at least $2/p_v$ samples. From Section 13.6 it should be clear that convergence towards the true p-value distribution is in $(\text{number of shuffles})^{-1/2}$. As discussed in Section 18.2 the estimation of a set of p-values over independent observations is recommendable in order to properly evaluate model rejection or consistency.

Example 18.6 (Non parametric estimation of significance of dependency and causality). Let me provide an example of the use of z-score to test the significance of correlation coefficients. I consider two coupled sets of observations $\hat{\mathbf{z}} = ((\hat{x}_1, \hat{y}_1), (\hat{x}_2, \hat{y}_2), \dots, (\hat{x}_q, \hat{y}_q))^\top$ from which the Pearson's correlation $\hat{\rho}$ can be computed (see Section 15.4). To generate the shuffled set one must randomize the mixing of the couples by randomly shuffling the order of the \hat{y} entries changing $\hat{\mathbf{z}}$, for example, to $\hat{\mathbf{z}}_i^{shuffled} = ((\hat{x}_1, \hat{y}_5), (\hat{x}_2, \hat{y}_1), \dots, (\hat{x}_q, \hat{y}_k))^\top$. On each shuffled dataset an estimate of the shuffled correlation coefficient is computed $\hat{\rho}_i^{shuffled}$ and the operation can

be repeated several times obtaining $\hat{\mu}_\rho^{shuffle}$, $\hat{\sigma}_\rho^{shuffle}$ and the z -score via Eq.18.63.

As a real example, let me look at the correlation between the log returns of Apple Inc. and Alphabet Inc (Google) stock prices on the Nasdaq in a period of common trade between 19 August 2004 and 22 November 2021 (4,347 trading days). The Person's correlation over the whole period is $\hat{\rho} \simeq 0.50$, which is a considerably large value. The z -score computed over 1,000 shuffled series is $z = 7.9$ and $\hat{p}_{value}^{shuffle} = 0$, indicating, therefore, a very strong dependency between these two stocks which is unlikely to be a spurious outcome by random chance. If I look instead at a sub-sample of the dataset containing only the last two months (the last 42 trading days), correlation is still strong at $\hat{\rho} \simeq 0.46$ however the significance starts to be less strong with $z = 3.1$ and $\hat{p}_{value}^{shuffle} = 0.001$. This is mainly a consequence of the poorer statistical significance of smaller samples. A parametric validation using the t-test statistics, introduced in Section 15.4.1, provides $p_{value} < 10^{-15}$ and $p_{value} = 0.00098$ respectively for the measure over the whole period and the one over the last 42 trading days.

One might have noticed that the sample p-value and the parametric ones essentially coincide, this is a very good proof of consistency but also questions the need for the non-parametric testing. Let me point out that, first the parametric test relies on assumptions that sometimes are not fulfilled and that are often hard to verify. Second, there are measures for which there are no theoretical parametric ways to assess their significance.

Let me deepen the analysis of the interrelations between Apple and Google stock prices by looking at their causality ties measured in terms of the sample estimate of the transfer entropy (see Section 10.4). For the full dataset, I obtain small transfer entropies in either directions $Google \rightarrow Apple$ or $Google \leftarrow Apple$ with a larger value of the transfer entropy in the direction $Google \rightarrow Apple$ where $T_{Google \rightarrow Apple} \simeq 4 \cdot 10^{-4}$ and $z_{Google \rightarrow Apple} \simeq 2$. Looking at sub-periods of 2 years windows (500 trading days) I find again the prevalence of transfer entropy in the direction $Google \rightarrow Apple$ with periods with significant causality relations, with $z > 3$ and sample p-values below 1% in both directions with the prevalence of Google causing Apple in the last three years of the dataset (after 2018), and conversely prevalence of Apple causing Google in the previous years. I observe that signal decreases with lag size except for lags of 6 or 7 days that have a large signal, similar to the one for lag 1 day. These results are shown in Fig.18.3 where the difference $T_{Google \rightarrow Apple} - T_{Google \leftarrow Apple}$, known as **information flow**, is reported for a 500 days (about two years) rolling window. The periods with significant transfer entropies and their directions are highlighted with filled symbols. This corresponds to measure when in the prevalent direction, the z -score is larger than 4 and the sample p-value is smaller than 0.1%. The transfer entropies are computed at lags between one to seven days and

the maximum transfer entropy across the lags is chosen. In most cases, it is the one-day lag that contributes but in a few cases, five, six, or seven days lags have larger transfer entropies.

18.8.2 Resampling to generate model statistics

In order to assess the significance of a given measure one would like to figure out if such a measure is typical or just incidental. In other words, one would like to have not only one but several measures and then analyze their statistical distribution.

If one has a very large observation set and therefore can access a large enough number of sub-samples that are themselves with sufficient sizes to achieve the desired model precision, then one can generate several independent outcomes for any model and use them to generate interval of confidence for the measure and to compare and discriminate between different models \mathcal{M}_1 and \mathcal{M}_0 . The problem is however that often one has a limited amount of data. In this context, resampling methodologies can become handy.

Bootstrapping

Bootstrapping is a general resampling methodology that allows to estimate level of confidence in empirical measures. It can be used for any measure. The methodology consists of re-sampling of an original observation set $\hat{\mathbf{x}}_s = (\hat{x}_{1,s}, \dots, \hat{x}_{p,s})^\top$ with $s = 1, \dots, q$ by picking randomly q observations from the original set, with repetition.

Definition 18.7 (Sampling with repetition). Random **sampling with repetition** means that in a sequence some elements will be randomly picked more than once and others not picked at all. This produces a new sample that is statistically consistent with the original one but does not contain all the same elements.

The bootstrapping process produces from one original observation set $(\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_q)$ several ‘bootstrapped’ sets $(\hat{\mathbf{x}}_1^b, \dots, \hat{\mathbf{x}}_q^b)$ with $b = 1, \dots, N_{\text{bootstrap}}$ that can be used to estimate statistical properties of any measure that is a function of the variables. In particular, quantities that one wants to estimate are the interval of confidence, which are the quantiles of the measures computed over the bootstrapped sets.

Remark 18.11. For continuous variables there are $q!$ possible different resampling and, therefore, for q large enough, statistics can be in principle performed on a very large number of resampling. This is, of course, numerically costly and unnecessary. Therefore, normally one limits this number to a number larger than the reciprocal of the precision one aims to estimate

in the confidence intervals. For instance, for 1%-99% quantiles one aims to compute statistics from no more than $N_{bootstrap} = 1,000$ resampling which is one order of magnitude above the desired precision level.

When the random variables are time-ordered and with ‘memory’ (i.e. auto-correlations), then one usually wants to generate bootstrapped series that still preserves some of the memory. In this case, one can sample in ‘blocks’ of a given size h of consecutive observations. This is called ‘block bootstrapping’.

Jackknife resampling

Jackknife is another resampling methodology. Differently from bootstrapping, in this procedure samples contain all observations except one that is left out. The resampling proceeds leaving out a different observation at each step. For q observations, there are q different resampling of size $q - 1$ which leave out one observation. The procedure is particularly useful to handle datasets with a few outlying values because there will be considerable statistical differences in the samples where the outliers are not present in the resampled set and therefore the presence and effect of outliers can be measured.

Cross-validation, holdout set

Similarly to Jackknife, one can hold out a subset containing more than one observation. Without overlaps, the resampling operation consists in dividing the sample in k random parts of size q/k (assuming this is an integer number) and then producing k different samples each having one of the k sub-parts held out. The process can be repeated by repartitioning the sample in another set of k parts. Indeed, for an observation set of size q , given k , there are $q!/((q - k)!k!)$ possible distinct partitions (this is the binomial coefficient). This procedure can result particularly useful for variance and bias estimations.

18.8.3 Undersampling, oversampling and synthetic data

In classification problems, class imbalance is a major problem causing poor training. Indeed, intuitively, the classifier has more examples from the most abundant class and tries to optimize results on this class eventually resulting in poor performances on the less represented classes.

In order to make the classes equally populated there are two options: 1. undersampling; 2. oversampling. Undersampling consists in downsizing the classes with a larger number of examples and removing observations until the dataset is balanced with all classes with the same number of examples. Conversely, oversampling consists in artificially increasing the size of the classes with a smaller number of examples adding observations. There are various ways to do this, the simplest being just duplicating examples at random, this takes the name of Random Over-Sampling (ROS). A more sophisticated approach is to create new examples by merging existing ones into new hybrid examples, this is for instance

what the Synthetic Minority Oversampling Technique (SMOTE) does, where a sample from the minority class is selected at random and then merged with a similar sample generating a new synthetic example similar to the two [Chawla et al., 2002]. There are various ways to do the merging between existing examples and generate new synthetic ones, a common choice is the Adaptive Synthetic Sampling Approach (ADASYN) [He et al., 2008] which, similarly to SMOOTE, operates a merge between two similar samples that are however not chosen at random but accordingly with some weight where one generates more synthetic data similar to examples that are harder to learn.

18.8.4 Data augmentation via generation of synthetic data

Oversampling is a way to generate synthetic data. Alternatively, producing a generative model that outputs data that are similar to the real observations is both a way to better train models on larger datasets or, in itself, a way to understand the mechanism that has generated the data. Indeed, historically, in science, simulations and generative models have been used extensively to describe some of the important characteristics of real systems and investigate the inner mechanisms of systems. One common approach is to generate data with known characteristics and use these synthetic data to test the ability of a model to retrieve the characteristics. In this context, synthetic data have been proven to be extremely useful to design models and evaluating their performances before they are used on real data. Another approach is instead to construct generative models using simple elements and rules. If the properties of the generated data are similar to the ones of the system of interest, then these generative models can be used to train descriptive and predictive models. This approach can also help to shed light on hidden mechanisms that produce the observed properties of the system. This topic touches an enormous literature that the interested reader can deepen starting from: Heermann [1990], Sanchez and Lucas [2002], Hafner [2008]. Sometimes, the generative part and the predictive part are integrated within the same modeling tool. This is, for instance, the approach underlying deep learning methods such as encoder-decoder and Generative adversarial network architectures (see Goodfellow et al. [2016], Goodfellow et al. [2014] and references generated from these pioneering papers).

From the perspective of this book, where modeling has been defined as ‘learning the joint probability of the system’, it must be clear to the reader that, indeed, if one acquires a good knowledge of such a joint probability, then the task of generating any realistic synthetic dataset becomes very easy. However, a good knowledge of the joint probability must be the final goal of modeling and not the starting point. Variational autoencoders and GANs, can provide a tool to generate synthetic data that follow well the conditional dependency structure between the system’s variables and therefore can help modelers to characterize such a structure. In most cases, more direct approaches to the learning of the joint probability of the system are more advantageous than the generation of synthetic signals. There are cases where synthetic data can be necessary, this might be

for instance the case when data are sensitive and the original data cannot be released. In these cases, producing ‘deepfake’ datasets might become the only way to access data for data-driven modeling [Little, 1993]. One must be aware that, despite synthetic data could turn out to be excellent tools for modeling in some circumstances, the history of scientific research has repeatedly shown that relying on hypotheses instead of direct observations can produce misleading results.

18.9 Subdivision of the dataset in a train, validation, and test subparts

I mentioned already in Section 3.5.1 that learning a model and testing it are distinct operations that must be performed on distinct datasets. As a general approach, one divides the observation set into three parts: 1. a train set; 2. a validation set; 3. a test set. The train set is where the parameters of the model are optimized. The validation set is where the hyperparameters are chosen (see Definition 3.1) and also where different models are compared. The test set is where the actual out-of-sample performances of the model must be quantified. Parameters are optimized for best performances on the training set, therefore performances are highest in this set. Performances on the validation set are used to set the model’s hyper-parameters and to select between models. Hyper-parameters are often regularizers or penalizer that make the performances of the model over the train set worst but aim to enhance the out-of-sample performances on the test set. Finally, the ability of the selected model to generalize for unseen data is measured by the performances in the test set. The performances in the train set are often referred to as in-sample performances while performances on the test set are often referred to as out-of-sample performances. Performances in the validation set are also ‘in-sample’ but they are usually used to infer the out-of-sample ones. The literature is plentiful of recipes for the division of the observation set into the training, validation, and test subsets. The rule is very simple: the three sets must be large enough to guarantee the desired statistical accuracy of the quantified measures. In most cases, statistical significance requires less data than optimization, therefore the larger part of the dataset is reserved for training (often 70%) then another (20%) for validation, and finally the rest for testing. However, one must make sure that the size of the dataset left for the test is sufficiently large to guarantee the accuracy of the results. Unfortunately, the relation between sample size and the accuracy of the sample estimates of a given measure depends on the measure, the model, and the population distribution. Therefore, a unique rule cannot be established. However, this book and specifically Chapters 13, 14 and 15 should be of valuable guidance for this purpose. Let me finally notice that when data are ordered temporarily, as in time series, the test set should normally follow the train set because one aims to predict events in the future. Sometimes, the validation set is chosen before the train set to avoid a gap between the train and the test.

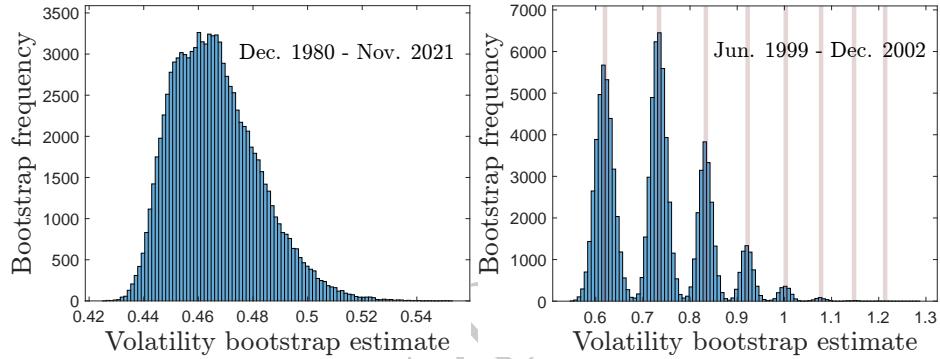


Figure 18.4 Frequencies of annualized volatility values for log-returns of Apple Inc. obtained from bootstrapping 100,000 times. The left panel refers to a period of over 20 years between Dec. 1980 and Nov. 2021. The right panel instead refers to a shorter period of about 3 years between Jun. 1999 and Dec. 2002 which includes a strong outlying event when a loss of over 50% was recorded on Friday 29 September 2000. The peaks in the left panel correspond to samples where the exceptionally large loss has been included in the statistics a different number of times. Data are the same as in Examples 16.5, 14.7, 14.9 and 18.2.

Example 18.7 (Bootstrapping, Jackknife and intervals of confidence). Let me consider the dataset for Apple Inc. that I have already used in Example 16.5 (and also in Examples 14.7, 14.9, 18.2 and Figure.14.5).

I first consider the whole period between 12 Dec. 1980 and 11 Nov. 2021 for a total of $q = 9,956$ trading days. I compute the daily prices log-returns and, from them, the annualized volatility using Definition 16.7 retrieving the value 0.47. I perform the bootstrapping on this dataset by taking randomly with repetition q observations and then repeating the sampling $B = 100,000$ times. For each random sampling, I compute the annualized volatility obtaining therefore B different values. The histogram reporting the frequencies of obtained bootstrapping volatility values is reported in the left panel of Fig.18.4. One can notice a relatively broad and skewed distribution. The median over the B bootstrapping values is 0.46 while the confidence interval given by the 1% and 99% quantiles is [0.44, 0.51]. This analysis is over a very long period of time. As I mentioned in Example 16.5, during this period there was a special event on Friday 29 September 2000 that is worth studying in further detail. Indeed, on that Friday, the price of Apple stock slumped by over 50%, a loss of more than 25 standard deviations.

I, therefore, repeated the bootstrap analysis by considering data only in the period between June 1999 and December 2002 ($q = 865$ daily observations). The annualized volatility over this period is $Volatility = \sqrt{252}\hat{\sigma} = 0.73$. As one can immediately notice this is larger than the previous value computed

over the whole period. The outlying value of Friday 29 September 2000 is having a strong effect on the volatility value, indeed by removing this single day the volatility becomes 0.62. Let me here investigate how this special outlying datapoint is affecting the overall statistics by using the bootstrapping method. As before, I take at random, with repetition, q observations and then I repeat the sampling $B = 100,000$ times. The histogram reporting the frequencies of obtained bootstrapping volatility values is shown in the right panel of Fig.18.4. One can notice several frequency peaks that were not present in the bootstrap statistics over the whole period. They are the effect of the outlying Friday 29 September 2000 loss that, in the first peak from left is not included (by chance) in the statistics, in the second is included once, in the third twice, etc.. The vertical brown lines are indeed the estimates for the average volatility obtained by excluding the outlaying point (0.62), or by including it once (0.73), or twice (0.83), or three times (0.92), etc.. This unusual repetition of peaks is a very clear indication that the extreme event Friday 29 September 2000 is special and not representative of the usual behavior of the log-returns. Indeed, it must be considered that I intentionally selected a shot period around this exceptional event. From the bootstrapped samples I can compute intervals of confidence by estimating the qualities of the bootstrapped volatilities. It results that the 1% quantile is 0.58 while the 99% is 1.00. This is a considerable upward shift and expansion of the interval of confidence. It must be noticed that together with the effect of the outlying point, this period was characterized by a large volatility in the stock markets especially in the technology sector. The bootstrapping intervals of confidence indicate that the hypothesis that the statistical properties of the signal during the sub-period June 1999 and December 2002 are comparable with the overall properties of the signal over the full period is below 1%. This is evidence of non-stationarity. By applying the Jackknife on this shorter 3-year dataset, one no longer observes the spurious repetition of peaks for sampling with or without the large Friday 29 September 2000 loss. Indeed, in the Jackknife there is only one sample, over $q = 865$ without that day, this produces only one outlying point at 0.62 with the rest spread in a narrow peak around 0.7343 with 1% and 99% quantiles respectively at 0.7319 and 0.7347. By removing the outlaying day from the original sample and performing the Jackknife one gets instead an interval of confidence between 0.617 and 0.62.

18.10 Data-driven modeling tutorial

<https://github.com/FinancialComputingUCL/DataDrivenModeling/Ch18>

The tutorial for this Chapter covers various topics on the goodness of models, including: Q-Q and P-P plots, Kolmogorov-Smirnov, and Anderson-Darling tests (Example 18.2), the use of the confusion matrix to assess goodness of classification, nonparametric measures (Example 18.6), bootstrapping, and Jackknife (Example 18.7).

Exercises

1. In September 2000 Apple's stock price lost nearly 52%. Knowing that the sample mean of the price log-returns in the two previous years was $\hat{\mu} \simeq 2 \cdot 10^{-3}$ and the sample standard deviation was $\hat{\sigma} \simeq 4 \cdot 10^{-2}$. Compute the p-values associated with this log return assuming that the probability distribution of the log returns is:
 - i. normal;
 - ii. Student-t with degrees of freedom $\nu = 3.1$;
 - iii. symmetric Levy-stable distribution with tail exponent $\alpha = 1.5$.

Compare the results and discuss them, also taking into account multiple hypotheses testing.

2. Use the information provided with the previous example to compute the likelihood for the three models (normal, Student-t, Levy-stable).
3. Discuss how the Adjusted AIC and the Bayesian information criterion should be applied in the case of the previous example.
4. Estimate the z-score and the sample p-value for the Person's correlation coefficient between $\hat{\mathbf{x}} = (0.61, -1.02, -1.59, 0.78, 0.83, -1.14, -1.43, 0.38, -0.48, 0.63)$ $\hat{\mathbf{y}} = (-0.50, -0.09, -3.24, -0.44, 2.39, -0.74, -2.17, -1.48, 0.73, 0.40)$. Discuss statistical significance also comparing with the t-test.
5. Use the $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ in the previous example to compute 10% and 90% intervals confidence from bootstrapping.

© Tomaso Aste 2019-23
not for distribution
July 25, 2023

Part IV

Closing

© Tomaso Aste 2019-23
not for distribution
July 25, 2023

© Tomaso Aste 2019-23
not for distribution
July 25, 2023

Conclusions

The field of modeling is undergoing rapid transformation, with new innovations constantly altering perspectives. However, despite the changing landscape, the concepts outlined in this book are based on solid mathematical principles and will remain foundational to data-driven modeling. Let me conclude this book by providing my broader perspective on what modeling might become in the future. And, for this purpose, let me start from the past.

19.1 The scientific method

Since, at least, the sixteenth century, science has developed its way to build and validate models that have been given the name of *scientific method* [Popper, 1934; Gower, 1997]. The scientific method is a circular approach based on observation of the system, formulation of a model, testing the goodness of such model through further observations, and comparison of the results with the ones obtained with alternative models, all under the principle of parsimony (see Figure 19.1).

Building models that can produce accurate predictions are at the core of scientific research. Being able to make predictions is essential to formulate hypotheses and theories that can be tested with further observations and can be iteratively refined or discarded and changed to obtain better predictions. Prediction, in this context, has a vast meaning embracing all inferences, relations, or gatherings around common classifications not provided directly by the observations or not provided precisely by them. Model complexity is another key ingredient; the principle of parsimony recommends that a simpler model with fewer parameters and fewer assumptions should be preferable with respect to a more complex model if both produce comparable predictions. This is an application of the *Occam's razor* principle introduced, in the thirteenth century, by the English Franciscan friar, William of Ockham (pictured on the bottom right of Figure 19.1).

The success of the scientific method is unquestionable. One can say with confidence that it has been the very basis of most of the knowledge about natural and artificial systems that humankind has acquired in the last five hundred years. However, epistemologists have repeatedly shown that the scientific method has not been always rigorously followed by scientists [Feyerabend, 1975]. There are several situations where the scientific method is hard to apply, for instance, when experiments cannot be repeated because the system under examination is unique and evolves with time. Nonetheless, its circular approach, which starts and ends

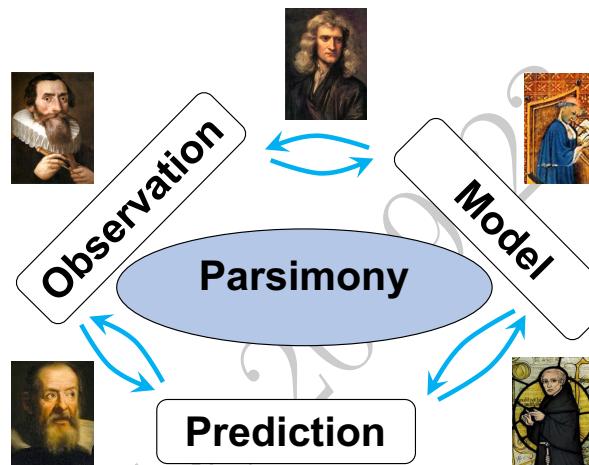


Figure 19.1 The scientific method is a circular process: from observations (data), models capable of predictions are constructed, and their predictions are tested against further observations, all within the principle of parsimony, where simpler models are preferred. The people represented in the images are, from the top right in clockwise order: Nicholas Oresme (1320-1382); William of Ockham (1285-1347); Galileo Galilei (1564-1642); Johannes Kepler (1571-1630); Isaac Newton (1643-1727). These are philosophers and natural philosophers who have greatly contributed to science and to the elaboration of the scientific method.

with observations passing through modeling and prediction, is a very powerful methodology for building models from observations.

19.2 Building models from data

Historically, models were generated starting from human intuition, they were built upon experience and then they were tested and validated, or eventually falsified, with experiments. This is a *top down* approach where mechanisms and interactions are assumed *a priori* and then data are used to calibrate and validate the model.

There is no need to argue or demonstrate that one of the novel elements of this epoch is the abundance and availability of data. This, combined with unprecedented access to powerful computational capabilities, is radically changing the way we are constructing and operating our models. Rather than following scientists' genius intuitions, increasingly models are constructed directly from data. The appealing aspect of this data-driven model construction process is that it can be automated. The challenge becomes to test the validity and goodness of the model, which must perform not only on the dataset from which it has been constructed but it must also generalize to other circumstances, perhaps even to situations that never happened before. This is a *bottom up* approach that starts from the analytics of observation data and constructs the model directly from the

data. Data-driven modeling is not only a new opportunity but rather a necessity because the complexity of most of the problems that we are addressing nowadays leaves less space for the top-down approach. It is indeed very unlikely that, someone, one day, will come up with a genius idea and a clean equation to solve the problem of recurring financial crises or the prediction of earthquakes. These problems are much more complex than modeling the motion of the heavenly bodies and the top-down approach, based on assumptions from human intuition is less effective and of narrower applicability.

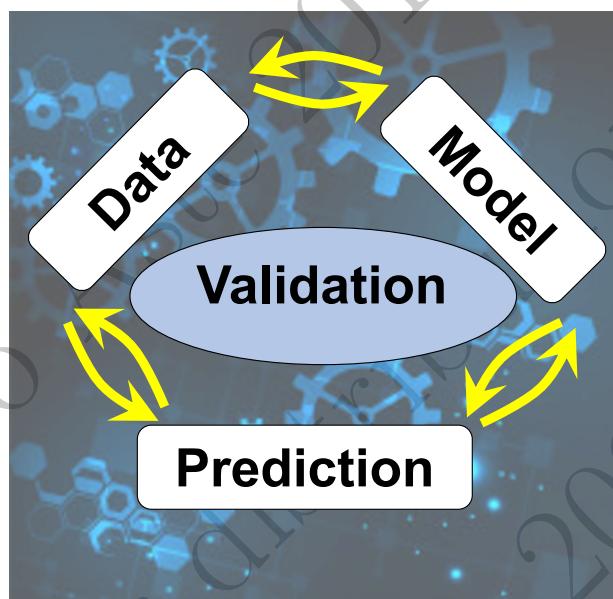


Figure 19.2 Machine learning modeling also adopts a circular process similar to the one at the basis of the scientific method (Fig.19.1). It starts from observations (in this case, usually referred to as data), and produces a model which yields predictions that are validated on the data. The main differences are that the circular process is automated, a large number of (non-linear) models are produced and the validation part becomes crucial to select the most useful model. The parsimony principle is less relevant and, sometimes, it is ignored.

19.3 Automated model construction

The data-driven modeling methodologies that I have presented in this book can be used – and it is used – for the automation of the modeling procedure.

Automated modeling is the unfolding ‘revolution’ at the core of present-day artificial intelligence. In this context, the word ‘intelligence’ is often abused because, for the vast majority of the endeavors in the artificial intelligence field, the aim is not to produce intelligent machines but rather to automate the modeling process and make machines able to operate in complex environments taking actions

and decisions that – so far – only ‘intelligent’ humans were able to take. Nonetheless, this continuously challenges our definition and understanding of intelligence, sometimes producing outputs that we thought could have been generated only by creative beings with good cognitive skills.

In my perspective, machine learning and artificial intelligence are still rooted in the circular modeling approach I have described for the scientific method. However, such a process is automated and does not need assistance from human ‘genius’ scientists. With respect to the original formulation of the scientific method, state-of-the-art machine learning, and data-science protocols are expanding the observation and model generation parts digging and manipulating vast datasets, and generating a large number of alternative models. Indeed, automation of the bottom-up modeling approach expands enormously the number of models that can be proposed.

Interestingly, one can view, present day, forward neural networks machines as incorporating directly in their architectures the circular learning methodology which starts from data and produces predictions that are used to refine the model via iterative repetitions (see Fig.19.3).

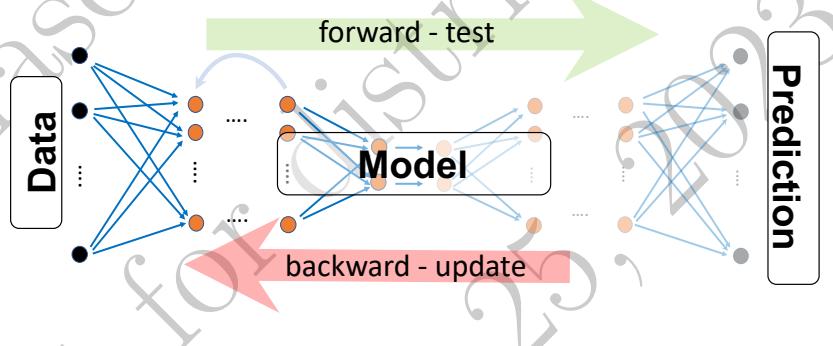


Figure 19.3 In a general scheme of a deep learning machine, one can identify three common general elements: the forward propagation of the signal, the backward propagation of the error, and the recurrent construction of the model. The process of learning is circular, starting from the data, producing a model, comparing the model prediction with the data, and retraining the model to achieve better prediction.

19.4 The end of parsimony

The *parsimony* principle has been at the center of the scientific method. Generations of students have been thought that models must be rational, understandable, simple, and with a minimum number of adjustable parameters. However, things have changed and this paradigm is now drifting away. One striking difference between the traditional human-made models and machine-generated models

is the very large number of adjustable parameters of the latter. At the center of these modeling tools, there is no longer the principle of parsimony but rather methodologies for model selection and validation. The scheme for modeling accordingly with the scientific method stays similar in its cyclical structure that starts and ends on data (observations) but, at the center, validation takes the place of parsimony, as visualized in Figure 19.2.

19.5 The rise of black boxes

Models are used to describe systems, explain their inner mechanisms, and predict their behavior. Machine-generated models and neural networks, in particular, tend to produce models that can be very good for prediction, but they are often useless for description. This modeling approach is putting into question the very meaning of modeling and, I would say, the actual meaning of knowledge. Without a theoretical underpinning, that humans can understand and convey to each other, on what basis can one be confident that the model will continue to provide accurate and reliable predictions in the future? What do we really know about a system in which we can predict its behavior but cannot explain why? And what does explaining really mean? These are philosophical issues, hard to resolve. However, it is important to be conscious that we are witnessing a radical change in the way systems are analyzed and modeled and that this radical change could have profound implications for what we presently call scientific knowledge.

Most of the scientific problems solved in the last few centuries are associated with finding a linear map $y = b_0 + b_1x_1 + b_2x_2 + \dots$ between two subsets of variables.¹ However, a very large number of problems cannot be addressed with linear modeling. When linear models are concerned, approximate solutions with similar error levels are similar; they have coefficients b_i which are at a similar distance from the coefficients of the exact solution. When non-linear models are concerned instead, the complexity of the problem increases considerably. Indeed, there exists a combinatorially large number of different models that provide approximate solutions with equivalent error levels but that are totally different from each other. Therefore, how to interpret results becomes blurry and the meaning of ‘description’ and ‘interpretation’ fades away because if the same outcome can be reached by completely different models, then the details of the model have little overall meaning. Each model can produce good predictions but each model structure and property does not provide useful information about the system under examination. These models are ‘black boxes’ where their internal structure is not only hard to analyze, it is actually meaningless.

Interpretability of artificial intelligence models is an actively debated topic that sometimes comes under the name of *explainable artificial intelligence*. Model interpretability touches on several fundamental issues and questions the essence of what we define as knowledge. It has also several practical consequences from

¹ Or some function of the variables $g_0(y) = b_0 + b_1g_1(x_1) + b_2g_2(x_2) + \dots$, where the b_i are parameters (independent from x_k and y).

ethics to reliability. In this domain, most questions rest unanswered. What I want to stress here is that the problem of interpretability is not an exclusive problem of some specific artificial intelligence approaches, such as neural networks; it is the unavoidable consequence of the shift of present-day modeling towards data-driven non-linear tools.

To explain artificial intelligence, blackbox-machines can be opened up and their functioning mechanisms might be decrypted in full detail. However, the non-uniqueness of the solution, and the existence of a combinatorially large number of very different machines which provide similar prediction power, make the interpretation of each solution meaningless. Data-driven modeling with non-linear tools provides us with millions of completely different ‘theories’ which might work at the same level of accuracy and predictive power. However, differently from what the top-down scientific approaches have done in the past, these competing theories are hard to select or combine in a unified paradigm that can advance what we become accustomed to calling scientific knowledge. I believe that, with the evolution of artificial intelligence, the issue of interpretability will move into the domain of human-machine interaction where blackboxes’ behavior will be judged more or less ‘interpretable’ in terms of the machine’s ability to explain to humans its decisions.

19.6 Future of modeling

Artificial intelligence is proving to be a powerful tool in the modeling of both natural and artificial systems. Automated modeling tools are helping us to better navigate complex systems in ways so far unexplored. Artificial intelligence has the potential to bring about major advancements in a wide range of fields, from science and engineering to economics and social sciences.

At the beginning of this book, I argued that models are the instruments to transform data into information. We have the chance to witness the unfolding of a major revolution in the way information is created. This will strongly modify science and society offering new tools to navigate our complex world.

Part V

Appendices

© Tomaso Aste 2019-23
not for distribution
July 25, 2023

© Tomaso Aste 2019-23
not for distribution
July 25, 2023

Appendix A

Methods to evaluate implicit equations

Often we are faced with the problem of finding the solution (root) of an equation like this:

$$g(x) = 0.$$

sometimes such an equation can be made explicit but often such a solution is not achievable analytically. In this case we can resort to numerical methods to estimate the root. Here I list three of the most common and practical ones that are used in the book.

A.1 Bisection method

If $g(x)$ is continuous in an interval (a_1, a_2) which contains one root then $g(a_1)$ and $g(a_2)$ must have opposite signs. A very good way to find the starting interval is to plot the function and take two points around the zero. Let $a_3 = (a_1 + a_2)/2$ the middle of the original interval, then take the point $a_4 = a_1$ or $a_4 = a_2$ where $g(a_4)$ has opposite sign to $g(a_3)$ iterate the process using the interval between a_3 and a_4 . The iteration can be stopped at $x^* = a_n$ if the process reaches a point where $|g(a_n)| < \epsilon$.

A.2 Iteration towards a fixed point

Often in optimization problems the implicit equation is in the form

$$x = h(x)$$

One of the simplest ways to find its solution is starting from a guess value for $x = x_0$ and then iteratively substituting into the equation

$$x_{n+1} = h(x_n).$$

The operation can be terminated when a fixed point with $x_{n+1} = x_n = x^*$ is reached or when the difference is small $|x_{n+1} - x_n| < \epsilon$. The method converges to the right root x^* if the function is continuous and derivable and if the initial guess is not too far from the root itself and if the derivative of $h(x)$ in the neighbours of x^* is smaller than one in absolute value.

A.3 Newton–Raphson method

A more effective convergence can be obtained with

$$x_{n+1} = x_n - \frac{h(x_n)}{h'(x_n)}.$$

with $h'(x)$ the first derivative in x of $h(x)$. However, sometimes this derivative could be hard to compute.

A.4 Method of the secants

This method uses the approximation $h'(x_n) \simeq (h(x_n) - h(x_{n-1}))/(\bar{x}_n - \bar{x}_{n-1})$ in the Newton–Raphson method:

$$x_{n+1} = x_n - (\bar{x}_n - \bar{x}_{n-1}) \frac{h(x_n)}{h(x_n) - h(x_{n-1})}.$$

Appendix B

Some optimization problems and methods

B.1 Linear programming

Linear programming (sometimes called linear optimization) is an optimization method of a linear objective function,

$$\text{minimize } g(\mathbf{X}) = \mathbf{c}^\top \mathbf{X}, \quad (\text{B.1})$$

subject to linear constraints.

A linear inequality constraint $\mathbf{a}^\top \mathbf{X} \geq 0$ (or ≤ 0) divides the space into two half. Solutions that satisfy all constraints might, or might not exist. Several of such inequality constraints might create a bounded polytope when there are at least $p+1$ of them. Depending on the intersections of the half-spaces existing solutions might lay inside a bounded polytope or outside of it. The solution, $\mathbf{c}^\top \mathbf{X} = c^*$ will lay on a plane, and therefore if an optimal solution exists, there will be in general a region of equivalent solutions.

A linear equality constraint $\mathbf{a}^\top \mathbf{X} = 0$ is instead a plane in the variable's space which is a $p - 1$ dimensional space. If there is more than one linear equality constraint solutions will be in the intersections of such planes. Each linear equality constraint reduces by one the dimensionality of the solution space and one needs p , linearly independent, linear equality constraints (non-parallel planes) to have a single point intersection (a unique solution).

If a minimum solution that satisfies all constraints exists, the linear programming problem is called feasible. Otherwise, it is called infeasible.

Depending on the domain of study relevant problems might have a large number of constraints or only a few.

Remark B.1. The use of the term “Programming” is historical, referring to a ‘procedure’ for solving the problem. It is an unfortunate term because it creates confusion with computer programming. The term ‘optimization’ would be more appropriate, but, for some reason, it is less used.

B.2 Quadratic programming

In quadratic programming the objective function is a multivariate quadratic function.

$$\text{minimize } g(\mathbf{X}) = \frac{1}{2} \mathbf{X}^\top \mathbf{Q} \mathbf{X} + \mathbf{c}^\top \mathbf{X}, \quad (\text{B.2})$$

subject to linear constraints.

The enforcement of linear constraints, either equalities or inequalities, makes the considerations in the previous section still valid for this problem. Indeed, for $\mathbf{Q} = 0$ this becomes the previous problem. The difference is that, for a positive defined \mathbf{Q} , the objective function is bounded from below and a unique solution for the quadratic objective function exists. However, it could be outside the feasible space of solutions imposed by the constraints and therefore the feasible solution might be laying on a boundary of the feasible space. For \mathbf{Q} with negative eigenvalues, the quadratic problem is unbounded and the feasible minimum can also be at a boundary of the feasible space or at infinite. Quadratic, constraints can be treated within the same framework.

B.2.1 Least squares

A special case of positive defined \mathbf{Q} is the least squares problem which is

$$\text{minimize } g(\mathbf{X}) = \|\mathbf{Y} - \mathbf{B}^\top \mathbf{X}\|^2, \quad (\text{B.3})$$

subject to linear constraints. Which is the previous problem with $\mathbf{Q} = \mathbf{B}\mathbf{B}^\top$ and $\mathbf{c} = -\mathbf{Y}^\top \mathbf{B}$. I have discussed the solution of the dual of this problem (i.e. the problem for \mathbf{B} not \mathbf{X}) in Section 8.4 (see Eqs.8.38 and 8.39) for the case when there are no constraints or the constraints are equalities and they are taken into account with the method of Lagrange multipliers. In the case of linear inequality constraints, the problem must be solved numerically.

B.3 Non-linear programming

When the objective function, or the constraints, or both are non-linear (and non-quadratic) the problem is called non-linear programming and in general, it must be solved numerically. The problem of constraint's feasibility and the solution's bound (convex objective function, in this case) rest conceptually the same but it can become considerably less intuitive. There is a large and increasing provision of methodologies to address these problems.

When the problem is convex (or concave in case of maximization) and differentiable then there are necessary and sufficient conditions for a solution to be optimal. Essentially, the minimum is where the derivatives in all directions (gradient) of an associated Lagrangian function are equal to zero, $\nabla g(\mathbf{X}^*) = 0$. The solution is unique and optimal if there are no other local minima. Normally, in non-linear problems, there are a very large number of local minima and the search is for one local minimum in a region of the feasibility space. Searching for

a minimum can be done by starting from some arbitrary point and descending towards the minima in a sequence of steps. This is the basics of gradient descent methodologies, nowadays predominant in optimization problems especially for neural networks where gradients can be computed efficiently. It is also the basis of methodologies such as sequential quadratic programming, where the objective function is locally approximated with a quadratic form and constraints are linearly approximated. Other approaches explore the landscape through random or weighted local moves accepting the ones that produce improvements in Monte Carlo dynamics often referred to, in these contexts, as stochastic optimization.

For general reference, I suggest to start from: Nocedal and Wright [2006], Gill et al. [2019].

Appendix C

Principal components analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique, used for feature extraction, that aims at finding a new set of uncorrelated variables, called principal components, by linearly transforming the original variables X_1, \dots, X_p .

The algorithm starts by standardizing the data by subtracting the mean and dividing by the standard deviation

$$\tilde{X}_i = \frac{X_i - \mu_i}{\sigma_i}. \quad (\text{C.1})$$

The algorithm is built from the correlation matrix \mathbf{C} (or the covariance matrix of standardized variables, which is the same thing). It then performs and eigendecomposition to obtain eigenvectors \mathbf{u}_k and eigenvalues λ_k ($i = 1, \dots, p$).

$$\mathbf{C}\mathbf{u}_k = \lambda_k \mathbf{u}_k \quad (\text{C.2})$$

Eigenvectors \mathbf{u}_k represent the directions or axes in the original feature space, while eigenvalues λ_k are associated with their importance. The eigenvectors are sorted based on their eigenvalues, representing the most important axes in descending order. The principal components are the components of the original standardized variables along the eigenvector directions

$$V_k = \tilde{\mathbf{X}} \mathbf{u}_k \quad (\text{C.3})$$

The dimensionality of the original data is then reduced by projecting onto a new feature space defined by the principal components with the largest eigenvalues.

Each principal component contributes to the total variance in the dataset in proportion to the corresponding eigenvalue:

$$\text{Contribution to Variance of Component } k = \frac{\lambda_k}{\sum_i \lambda_i}. \quad (\text{C.4})$$

Principal components with larger eigenvalues explain a larger portion of the variance and are considered more significant in capturing the underlying patterns or structure of the data.

While the PCA is a very effective method for dimensionality reduction it is not a solution for the curse of dimensionality. Indeed, when p is large and the number of observations is small, the correlation matrix \mathbf{C} cannot be estimated

precisely from data and the principal component becomes a noisy representation of the dataset.

PCA has been used extensively in statistics, data science, and machine learning and there is a huge literature. The interested reader could start from the papers by Pearson [1901] and Hotelling [1933] that introduced the topic. Then, among the several contributions, I suggest Jolliffe [2002].

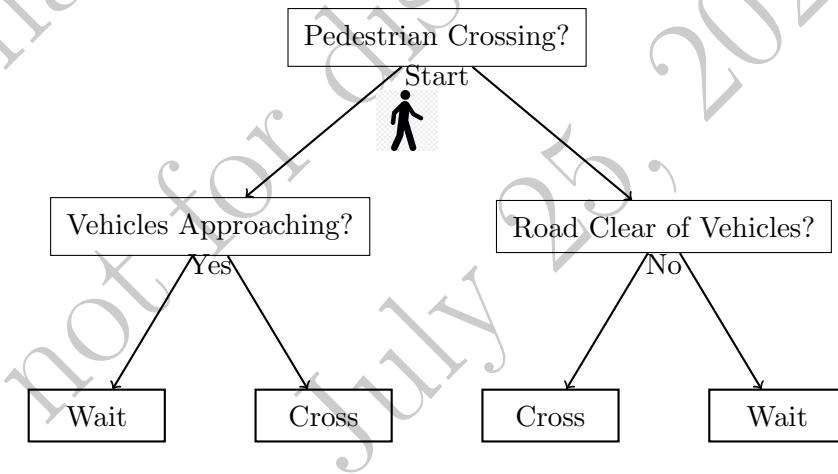
Appendix D

Random forest

D.1 Decision trees

Decision trees are used for both classification and regression tasks. A decision tree is a non-parametric supervised learning algorithm that has a hierarchical, tree structure. The tree starts with a root node, which represents the entire dataset. The root node is divided into subsequent nodes called internal nodes, each corresponding to a feature or attribute. The internal nodes are connected by branches, which represent the possible values or outcomes of the corresponding feature. The leaf nodes at the end of the branches represent the final predictions or decisions.

A very rudimentary decision tree for the problem of a pedestrian crossing a road (See Chapter 10) is illustrated below.



In a binary decision tree (as the one above), each internal node has two branches representing two possible outcomes of a binary decision on categorical or numerical features. While binary decisions (e.g., yes/no) are the most common, it is possible to have more than two branches at a node, allowing for multiway splits. In a multiway decision tree, each internal node can have more than two branches, corresponding to multiple possible outcomes. The tree partitions the feature space into regions corresponding to different outcomes or predictions. By recursively

splitting the decision space based on different features and their values, the decision tree creates a hierarchical structure that guides the decision-making process. The root node represents the entire decision space or feature space. Each internal node represents a partition into subsets of the decision space. The splitting condition determines how the feature space is divided into separate regions. Branches represent the different paths or regions that the decision space can follow based on the values of the features. Leaf nodes are the terminal nodes, the endpoints of the decision process. They are the final predictions or decisions pointing to a specific region of the decision space.

For literature reference on decision trees one can refer to Kingsford and Salzberg [2008], Song and Ying [2015].

D.2 Random forest

The random forest algorithm consists of the construction of a “forest” of multiple, individual, decision trees. It is an ensemble method where, typically, each tree in the forest is built using a random subset of the training data and a random subset of the features. The final prediction is determined by aggregating the predictions of all the trees. This can be done either by majority voting (for classification) or averaging (for regression). There are several variations and refinements of the algorithm which explore different ways of constructing the ensemble and the aggregation of the individual trees’ decisions. The idea was first formulated by citeho1995random who also coined the name. However, curiously, a trademark “Random Forests” was registered over ten years later by Leo Breiman and Adele Cutler.

D.3 Gradient boosting

Similarly with Random Forest, Gradient Boosting [Friedman, 2001] is another ensemble learning algorithm that builds an ensemble of decision trees. However, unlike Random Forest, Gradient Boosting builds the ensemble sequentially rather than independently. Each new tree is built to improve the previous trees, focusing on the samples with larger residuals or errors. The predictions of the individual trees are combined using a weighted sum, where the weights are determined by the gradient descent optimization process. A noticeable variation is XGBoost (Extreme Gradient Boosting) which introduces additional enhancements, such as column subsampling and an overall more efficient implementation, making it a popular choice. In general, Gradient Boosting algorithms improve considerably performance and computational efficiency with respect to the original random forest approach. For references on gradient boosting I suggest starting from the original paper by Friedman [2001] and then read Friedman [2002], Natekin and Knoll [2013], Chen et al. [2015].

Index

- Accuracy, 352
Adjacency matrix, 36, 41, 319
Akaike's Information Criterion (AIC), 358, 360, 366
Anderson-Darling (AD) test, 343, 345
Annualized volatility, 305
Autocorrelation, 135, 137, 152, 157, 292, 300
Bandwidth, 228, 238, 257, 341
Bayes theorem, 85, 101, 109, 118, 167, 198, 236, 280, 359, 363
Bayesian information criterion (BIC), 359
Bias, 20, 29, 260, 347, 370
Bias variance tradeoff, 29, 347
Black box, 383
Body and tail of the distribution (see fat-tail), 240, 242, 244
Bonferroni correction, 339
Bootstrapping, 299, 338, 369
Causal network, 330
Causality, 134, 161, 163, 172, 192, 330, 367
Central limit theorem, 53, 146, 219, 272
Central moments, 21, 148, 215
Centrality and peripherality, 40
Characteristic equation, 301
Characteristic function, 55, 83
Chebyshev's inequality, 62, 218, 366
Chordal graph, 46, 186, 198, 324
Complementary cumulative distribution, 15, 241, 344
Conditional entropy, 106, 166
Conditional expected value, 26, 124
Conditional independence, 192, 318, 325
Conditional probability, 85, 90, 109, 161, 207, 354, 355
Conditional transfer entropy, 168, 170, 172, 332
Confidence interval, 19, 296, 358, 364
Configurational model, 177
Configurational model], 177
Confusion Matrix, 352
Correlation, 79
Correlation matrix, 80, 392
Correlation ratio, 124, 348
Covariance, 79
Covariance matrix, 79, 103, 116, 169, 200, 255, 259, 278
Cross-entropy, 104, 201, 230, 354
Cross-validation, 29, 229, 238, 288, 360, 370
Cumulative distribution function, 15, 54, 78, 137, 221, 242, 340, 342, 343
Cumulative distribution function estimation of, 221
Curse of dimensionality, 23, 255, 264, 330, 392
Data augmentation, 371
Degree, 36, 185, 190, 324
Degree distribution, 36, 37, 177
Degrees of freedom, 66, 83, 85, 103, 165, 261, 357
Dickey-Fuller test, 301
Double descent, 32, 285
Early stopping, 288
Efficient frontier (see also Markowitz's portfolio theory), 121, 277
Elliptical distribution, 81, 88, 103, 155, 350
Empirical mode decomposition (EMD), 308
Entropy, 97, 101, 111, 193, 194, 205, 224, 346
Excess kurtosis, 22, 51
Expectation maximization (EM), 244, 263, 288
Expected value, 19, 147, 214, 215
Exponential smoothing, 303
F1 score, 353
Fat-tail, 37, 62, 67, 69, 87, 122, 147, 213, 238, 240, 242, 244, 262
Feature extraction, 269, 392
Feature selection, 164, 269
Fitness model, 177, 178
Fractal and fractal dimension, 141, 142, 147, 154
Gamma distribution, 65, 66, 72, 249
Generalized central limit theorem, 59
Generalized dependency measure, 123, 348, 350
Generalized extreme value distribution, 73
Generalized Hurst exponent, 151, 295
GLASSO, 286
Goodness of models, 335
Goodness of regression, 345

- Granger causality (see Wiener-Granger causality), 164
Graph (see Network), 35
Graphical models, 193, 197
Hölder exponent, 150
Higher order interactions, 130, 189, 194, 197
Higher order network, 45, 189, 203
Histogram, 223–225, 255, 360
Hurst exponent, 148, 292, 295
Hyperparameter, 28, 356, 372
IFN probability decomposition, 198, 325
Independent and identically distributed (i.i.d.), 52, 144, 218, 267
Independent variables, 20, 52, 109, 112, 127
Information filtering networks, 178, 322
Interaction information, 130
Inverse covariance, 81, 118, 200, 278
Jackknife, 338, 370
Joint probability, 23, 77, 109, 112, 127, 155, 198, 280, 330, 371
k-gamma distribution, 66, 72, 249
Kendall- τ correlation, 127
Kernel density estimator (KDE), 227, 238, 257
Kernel function, 228
Kolmogorov-Smirnov (KS) test, 342, 345
Kolmogorov-Smirnov test, 292
Kullback-Leibler divergence, 104, 128, 167, 193, 237, 272, 354
Kruscal's algorithm, 180
Kurtosis, 22
Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, 302
 L_0 , L_1 , L_2 , and L_n -norms, 265, 266, 281, 282, 285
Lévy-alpha stable distribution, 58, 146
Lagged correlation, 134, 163
Lagrange multiplier, 122, 277, 390
Laplacian, 40
Law of large numbers, 100, 193, 216, 356
Likelihood, 86, 87, 192, 194, 234, 355
Likelihood ratio, 357
Linear dependency, 79, 112, 165, 202, 260
Location-scale family distributions, 70, 155
Log-likelihood, 87, 193, 235
Log-return, 138
LoGo, 200, 287
Mahalanobis distance, 82
Marčenko-Pastur, 267
Marginal probability, 77, 109, 112, 127
Markowitz portfolio theory, 119, 276
Matthews correlation coefficient (MCC), 353
Maximally filtered clique forests (MFCF), 187, 202, 322
Maximum entropy, 101, 237
Maximum likelihood estimation (MLE), 234, 263
Maximum spanning tree, 179, 201, 322
Method of moments, 233
MFCF, 187, 202, 323
Minimum spanning tree (MST), 179, 322
Models, 9
Moments of a distribution, 21
Moving average, 303
MST, 179, 201, 322
Multifractal, 150
Multilinear regression, 89, 116, 275, 286
Multiscaling, 150, 299
Multivariate normal, 80
Multivariate Student-t, 82
Mutual information, 128
Nested models, 351
Network, 35, 176
Network representation, 131, 192, 287, 318
Nonparametric estimation, 213
Normal distribution, 51
Null model, 336
O-information, 133
P-P plot, 340
P-value, 336, 366
Parametric estimation, 231
Pareto distribution, 69
Parsimony, 382
Partial correlation, 118
Pearson's correlation, 79
Pearson's correlation estimator, 260
Pearson's covariance estimator, 259
Planar maximally filtered graphs (PMFG), 181
PMFG, 182
Posterior probability, 86, 105, 236, 363
Precision, 353
Precision matrix, 81, 118
Prim's algorithm, 180
Prior probability, 86, 101, 105, 236, 255, 281, 363
Probability, 14
Probability decomposition, 198, 325
Probability density function, 16
Probability mass function, 15
Q-Q plot, 340
Quantiles, 17
R-square, 346
Random variable, 14
Random walk, 44, 144, 302
Rank correlations, 126
Rank frequency plot, 222
Ranking, 126
Redundancy, 131, 133, 331
Regression, 25, 109–111, 164, 345
Regularization, 32, 280, 325

- Reinforcement learning, 26
Relative Likelihood, 358
Residual entropy, 133
Resampling, 369
Return, 138
ROC curve, 355
Rolling windows, 303
Sample central moments, 215
Sample correlation, 260
Sample mean, 214
Sample moments, 215
Scaling & scaling law, 139, 290
Scientific method, 379
Self-affine process, 148
Sensitivity, 352
Shannon entropy, 97, 128, 166, 224, 237, 258
Shannon-Khinchin axioms, 107
Shrinkage, 278
Shuffling, 299, 365
Simplex, 46, 103, 189, 202, 324
Simplicial complex, 321
Simplicial complexes, 47
Skewness, 22
Sparse inverse covariance, 200, 286, 327
Spearman- ρ correlation, 126
Specificity, 352
Stable distribution, 57, 146
Standard deviation, 21
Standard score (or z-score), 366
Stationarity, 137, 301
Stirling approximation, 100
Stochastic process, 137
Student-t distribution, 67, 72
Supervised learning, 25
Synergy, 131, 133, 331
Tail exponent, 58–60, 67, 74, 122, 147, 148, 152, 233, 240, 299
Tail exponent log-log fitting estimator, 241
Tail exponent ML estimator, 241
Tail of the distribution, 60
Test set, 28
Thresholding, 178, 318
Time clustering, 311
Time series, 290
TMFG, 183, 184, 200, 322
Topological regularization, 287, 325
Total correlation, 133
Train set, 28
Transfer entropy, 166, 330, 367
Tree and forests, 45
Triangulated maximally filtered graphs (TMFG), 183
Unscaling, 148, 292, 298, 299
Unsupervised learning, 25
Validation set, 28
Value at risk, 18
Variance, 21

References

- S. A. Abdallah and M. D. Plumley. A measure of statistical complexity based on predictive information with application to finite spin systems. *Physics Letters A*, 376(4):275–281, 2012.
- T. W. Anderson and D. A. Darling. Asymptotic theory of certain “goodness of fit” criteria based on stochastic processes. *The annals of mathematical statistics*, pages 193–212, 1952.
- T. Aste. Cryptocurrency market structure: connecting emotions and economics. *Digital Finance*, 1(1):5–21, 2019.
- T. Aste. Topological regularization with information filtering networks. *arXiv preprint arXiv:2005.04692*, 2020.
- T. Aste. Stress testing and systemic risk measures using elliptical conditional multivariate probabilities. *Journal of Risk and Financial Management*, 14(5):213, 2021.
- T. Aste and T. Di Matteo. Emergence of gamma distributions in granular materials and packing models. *Physical Review E*, 77(2):021309, 2008.
- T. Aste, G. Massara, A. Briola, and R. Wang. Financial computing & analytics group ucl. URL <https://github.com/FinancialComputingUCL>.
- T. Aste, W. Shaw, and T. Di Matteo. Correlation structure and dynamics in volatile markets. *New Journal of Physics*, 12(8):085009, 2010.
- L. Bachelier. Théorie de la spéculation. In *Annales scientifiques de l’École normale supérieure*, volume 17, pages 21–86, 1900.
- J. Balatoni and A. Rényi. Remarks on entropy. *Publ. Math. Inst. Hung. Acad. Sci*, 1:9–40, 1956.
- A.-L. Barabási. Scale-free networks: a decade and beyond. *science*, 325(5939):412–413, 2009.
- W. Barfuss, G. P. Massara, T. Di Matteo, and T. Aste. Parsimonious modeling with information filtering networks. *Physical Review E*, 94(6):062306, 2016.
- P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.
- F. Battiston, E. Amico, A. Barrat, G. Bianconi, G. Ferraz de Arruda, B. Franceschiello, I. Iacopini, S. Kéfi, V. Latora, Y. Moreno, M. M. Murray, T. P. Peixoto, F. Vaccarino, and G. Petri. The physics of higher-order interactions in complex systems. *Nature Physics*, 17(10):1093–1098, 2021. doi: 10.1038/s41567-021-01371-4. URL <https://doi.org/10.1038/s41567-021-01371-4>.
- M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.
- A. J. Bell. The co-information lattice. In *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation: ICA*, volume 2003. Citeseer, 2003.
- R. Bellman. On the theory of dynamic programming. *Proceedings of the National Academy of Sciences of the United States of America*, 38(8):716, 1952.
- R. Bellman. The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6):503–515, 1954.
- M. Berkane and P. Bentler. Moments of elliptically distributed random variates. *Statistics & Probability Letters*, 4(6):333–335, 1986.

- A. C. Berry. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society*, 49(1):122–136, 1941.
- D. P. Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, New York, 1982.
- L. M. Bettencourt, V. Gintautas, and M. I. Ham. Identification of functional information subgraphs in complex networks. *Physical review letters*, 100(23):238701, 2008.
- G. Bianconi. *Higher Order Networks: An Introduction to Simplicial Complexes*. Cambridge University Press, 2021.
- G. Bianconi and A.-L. Barabási. Competition and multiscaling in evolving networks. In *The Structure and Dynamics of Networks*, pages 361–367. Princeton University Press, 2011.
- A. Birnbaum. On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306, 1962.
- N. Boccara. *Modeling complex systems*. Springer Science & Business Media, 2010.
- L. Boltzmann. Studien über das gleichgewicht der lebenden kraft. *Wissenschaftliche Abhandlungen*, 1:49–96, 1868.
- G. Bonanno, G. Caldarelli, F. Lillo, S. Micciche, N. Vandewalle, and R. N. Mantegna. Networks of equities in financial markets. *The European Physical Journal B*, 38(2):363–371, 2004.
- C. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.
- O. Boruvka. O jistém problému minimálním. 1926.
- D. R. Brillinger. A note on the rate of convergence of a mean. *Biometrika*, 49(3/4):574–576, 1962.
- G. Caldarelli, A. Capocci, P. De Los Rios, and M. A. Munoz. Scale-free networks from varying vertex intrinsic fitness. *Physical review letters*, 89(25):258702, 2002.
- G. Casella and R. L. Berger. *Statistical inference*. Cengage Learning, 2021.
- S. Chaudhary, J. Down, N. D’Souza, Y. Gu, and H. Yan. Parameter-wise double descent - a unified model or not? UCL - COMP0031 student project, March 2023.
- N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
- B. Chazelle. A minimum spanning tree algorithm with inverse-ackermann type complexity. *Journal of the ACM (JACM)*, 47(6):1028–1047, 2000.
- T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.
- C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968.
- R. Clausius. *The mechanical theory of heat*. Macmillan, 1879.
- A. Coniglio, M. P. Ciamarra, and T. Aste. Universal behaviour of the glass and the jamming transitions in finite dimensions for hard spheres. *Soft matter*, 13(46):8766–8771, 2017.
- S. d’Ascoli, M. Refinetti, G. Birolì, and F. Krzakala. Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*, pages 2280–2290. PMLR, 2020.
- T. Di Matteo, T. Aste, and M. M. Dacorogna. Long-term memories of developed and emerging markets: Using the scaling analysis to characterize their stage of development. *Journal of banking & finance*, 29(4):827–851, 2005.
- E. W. Dijkstra et al. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959.
- J. L. Doob and J. L. Doob. *Stochastic processes*, volume 7. Wiley New York, 1953.
- R. A. Duke. The genus, regional number, and betti number of a graph. *Canadian Journal of Mathematics*, 18:817–822, 1966.
- A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

- A. Edelman. Eigenvalues and condition numbers of random matrices. *SIAM journal on matrix analysis and applications*, 9(4):543–560, 1988.
- H. Edelsbrunner, J. Harer, et al. Persistent homology-a survey. *Contemporary mathematics*, 453:257–282, 2008.
- A. Einstein. Über die von der molekularkinetischen theorie der wärme geforderte bewegung von in ruhenden flüssigkeiten suspendierten teilchen. *Annalen der physik*, 4, 1905.
- P. Erdős. Rényi, a.:” on random graphs. *I*”. *Publicationes Mathematicae (Debre*, 1959.
- M. D. Ernst. Permutation methods: a basis for exact inference. *Statistical Science*, pages 676–685, 2004.
- K. W. Fang. *Symmetric multivariate and related distributions*. CRC Press, 2018.
- W. Feller. *An introduction to probability theory and its applications*. Wailey & Sons, 1957.
- P. Feyerabend. *Against Method: Outline of an Anarchistic Theory of Knowledge*. New Left Books, 1975.
- P. Fiedor. Networks in financial markets based on the mutual information rate. *Physical Review E*, 89(5):052801, 2014.
- Y. Finance. Yahoo finance historic prices. URL <https://finance.yahoo.com/>.
- J. B. J. Fourier. *Théorie analytique de la chaleur*. Firmin Didot, 1822.
- J. Friedman, T. Hastie, R. Tibshirani, et al. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.
- J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.
- J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.
- J. H. Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378, 2002.
- U. Frisch and G. Parisi. Fully developed turbulence and intermittency. *New York Academy of Sciences, Annals*, 357:359–367, 1980.
- P. K. Friz and M. Hairer. *A course on rough paths*. Springer, 2020.
- P. E. Gill, W. Murray, and M. H. Wright. *Practical optimization*. SIAM, 2019.
- B. Gnedenko, A. Kolmogorov, B. Gnedenko, and A. Kolmogorov. Limit distributions for sums of independent. *Am. J. Math*, 105, 1954.
- I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.
- S. Goodman. A dirty dozen: twelve p-value misconceptions. In *Seminars in hematology*, volume 45, pages 135–140. Elsevier, 2008.
- B. Gower. *Scientific method: An historical and philosophical introduction*. Psychology Press, 1997.
- C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.
- J. Hafner. Ab-initio simulations of materials using vaspm: Density-functional theory and beyond. *Journal of computational chemistry*, 29(13):2044–2078, 2008.
- D. Hallac, S. Vare, S. Boyd, and J. Leskovec. Toeplitz inverse covariance-based clustering of multivariate time series data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 215–223, 2017.
- P. R. Halmos. *Measure theory*, volume 18. Springer-Science, New York, 1950.
- T. C. Halsey, M. H. Jensen, L. P. Kadanoff, I. Procaccia, and B. I. Shraiman. Fractal measures and their singularities: The characterization of strange sets. *Physical review A*, 33(2):1141, 1986.
- B. Hammer and K. Gersmann. A note on the universal approximation capability of support vector machines. *neural processing letters*, 17(1):43–53, 2003.

- T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.
- H. He, Y. Bai, E. A. Garcia, and S. Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. IEEE, 2008.
- D. W. Heermann. Computer-simulation methods. In *Computer Simulation Methods in Theoretical Physics*, pages 8–12. Springer, 1990.
- D. Helbing. Agent-based modeling. In *Social self-organization*, pages 25–70. Springer, 2012.
- B. M. Hill. A simple general approach to inference about the tail of a distribution. *The annals of statistics*, pages 1163–1174, 1975.
- W. Hoeffding. Masstabvariante korrelationstheorie. *Schriften des Mathematischen Instituts und Instituts für Angewandte Mathematik der Universität Berlin*, 5:181–233, 1940.
- K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.
- H. Hotelling. Analysis of a complex of statistical variables with principal components. *J. Educ. Psy.*, 24:498–520, 1933.
- N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*, 454(1971):903–995, 1998.
- M. Hulswit. *From cause to causation: A Peircean perspective*, volume 90. Springer Science & Business Media, 2002.
- G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.
- R. G. James, N. Barnett, and J. P. Crutchfield. Information flows? a critique of transfer entropies. *Physical review letters*, 116(23):238701, 2016.
- I. T. Jolliffe. *Principal component analysis for special types of data*. Springer, 2002.
- A. Y. Khinchin. *Mathematical foundations of information theory*. Courier Corporation, 2013.
- C. Kingsford and S. L. Salzberg. What are decision trees? *Nature biotechnology*, 26(9):1011–1013, 2008.
- J. B. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, 7(1):48–50, 1956.
- S. L. Lauritzen. *Graphical Models*. Oxford:Clarendon, 1996.
- O. Ledoit and M. Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, 10(5):603–621, 2003.
- O. Ledoit and M. Wolf. Honey, i shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4):110–119, 2004.
- R. J. Little. Statistical analysis of masked data. *Journal of Official statistics*, 9(2):407, 1993.
- G. Livan, M. Novaes, and P. Vivo. *Introduction to random matrices: theory and practice*, volume 26. Springer, 2018.
- P. C. Mahalanobis. On the generalized distance in statistics. volume 2 (1), pages 49–55. National Institute of Science of India, 1936.
- B. B. Mandelbrot. Possible refinement of the lognormal hypothesis concerning the distribution of energy dissipation in intermittent turbulence. In *Statistical models and turbulence*, pages 333–351. Springer, 1972.
- B. B. Mandelbrot. Intermittent turbulence in self-similar cascades: divergence of high moments and dimension of the carrier. *Journal of fluid Mechanics*, 62(2):331–358, 1974.
- B. B. Mandelbrot. *Fractals: Form, Chance and Dimension*. W. H. Freeman and Company, San Francisco, 1977.
- B. B. Mandelbrot. *Fractals and scaling in finance: Discontinuity, concentration, risk. Selecta volume E*. Springer Science & Business Media, 2013.
- B. B. Mandelbrot and B. B. Mandelbrot. *The fractal geometry of nature*, volume 1. WH freeman New York, 1982.

- H. Markowitz. Portfolio selection. *The Journal of Finance*, 7:77–91, 1952.
- H. M. Markowitz. *Portfolio selection*. Yale university press, 1968.
- G. Marsaglia and J. Marsaglia. Evaluating the anderson-darling distribution. *Journal of statistical software*, 9(1):1–5, 2004.
- G. Marti, F. Nielsen, M. Biúkowski, and P. Donnat. A review of two decades of correlations, hierarchies, networks and clustering in financial markets. *Progress in Information Geometry*, pages 245–274, 2021.
- V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.
- G. P. Massara and T. Aste. Learning clique forests. *arXiv preprint arXiv:1905.02266*, 2019.
- G. P. Massara, T. Di Matteo, and T. Aste. Network filtering for big data: Triangulated maximally filtered graph. *Journal of complex Networks*, 5(2):161–178, 2017.
- F. J. Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(235):68–78, 1951.
- H. Matsuda. Physical nature of higher-order mutual information: Intrinsic correlations and frustration. *Physical review E*, 62(3):3096, 2000.
- M. M. Mayfield and D. B. Stouffer. Higher-order interactions capture unexplained complexity in diverse communities. *Nature ecology & evolution*, 1(3):1–7, 2017.
- C. D. McGillem and G. R. Cooper. *Continuous and discrete signal and system analysis*. Harcourt School, 1991.
- R. Metzler and J. Klafter. The random walk’s guide to anomalous diffusion: a fractional dynamics approach. *Physics reports*, 339(1):1–77, 2000.
- A. D. Mirlin, Y. V. Fyodorov, F.-M. Dittes, J. Quezada, and T. H. Seligman. Transition from localized to extended eigenstates in the ensemble of power-law random banded matrices. *Physical Review E*, 54(4):3221, 1996.
- R. H. Myers and R. H. Myers. *Classical and modern regression with applications*, volume 2. Duxbury press Belmont, CA, 1990.
- P. Nakkiran. More data can hurt for linear regression: Sample-wise double descent. *arXiv preprint arXiv:1912.07242*, 2019.
- A. Natekin and A. Knoll. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21, 2013.
- N. Nava, T. Di Matteo, and T. Aste. Anomalous volatility scaling in high frequency financial data. *Physica A: Statistical Mechanics and its Applications*, 447:434–445, 2016a.
- N. Nava, T. Di Matteo, and T. Aste. Time-dependent scaling patterns in high frequency financial data. *The European Physical Journal Special Topics*, 225(10):1997–2016, 2016b.
- J. Nešetřil, E. Milková, and H. Nešetřilová. Otakar boruvka on minimum spanning tree problem translation of both the 1926 papers, comments, history. *Discrete mathematics*, 233(1-3):3–36, 2001.
- M. Newman. *Networks*. Oxford university press, 2018.
- J. Neyman and E. S. Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.
- J. Nocedal and S. Wright. *Numerical optimization*. Springer Science & Business Media, 2006.
- I. Nourdin. *Selected aspects of fractional Brownian motion*, volume 4. Springer, 2012.
- E. Novikov. Intermittency and scale similarity in the structure of a turbulent plow. *Journal of Applied Mathematics and Mechanics*, 35(2):231–241, 1971.
- E. Olbrich, N. Bertschinger, N. Ay, and J. Jost. How should complexity scale with system size? *The European Physical Journal B*, 63(3):407–415, 2008.
- G. Parisi. Complex systems: a physicist’s viewpoint. *arXiv preprint cond-mat/0205297*, 2002.
- J. Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019.
- J. Pearl and D. Mackenzie. Ai can’t reason why. *Wall Street Journal*, 2018.

- J. Pearl et al. Models, reasoning and inference. *Cambridge, UK: Cambridge University Press*, 19, 2000.
- K. Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.
- K. Pearson. X. contributions to the mathematical theory of evolution.—ii. skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London.(A.)*, (186): 343–414, 1895.
- K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.
- K. Pearson. The problem of the random walk. *Nature*, 72(1867):342–342, 1905.
- K. Popper. *The logic of scientific discovery*. Julius Springer, 1934.
- F. Pozzi, T. Di Matteo, and T. Aste. Exponential smoothing weighted correlations. *The European Physical Journal B*, 85(6):1–21, 2012.
- W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.
- R. C. Prim. Shortest connection networks and some generalizations. *The Bell System Technical Journal*, 36(6):1389–1401, 1957.
- P. F. Proccaci and T. Aste. Forecasting market states. *Quantitative Finance*, 19(9):1491–1498, 2019.
- A. Rényi. On the dimension and entropy of probability distributions. *Acta Mathematica Academiae Scientiarum Hungarica*, 10(1-2):193–215, 1959a.
- A. Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10(3-4):441–451, 1959b.
- A. Rényi et al. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1. Berkeley, California, USA, 1961.
- G. Ringel. *Map color theorem*, volume 209. Springer Science & Business Media, 1974.
- L. Rokach and O. Maimon. Clustering methods. In *Data mining and knowledge discovery handbook*, pages 321–352. Springer, 2005.
- F. E. Rosas, P. A. Mediano, M. Gastpar, and H. J. Jensen. Quantifying high-order interdependencies via multivariate extensions of the mutual information. *Physical Review E*, 100(3): 032305, 2019.
- S. M. Ross. *Introduction to probability models*. Academic press, 2014.
- S. M. Sanchez and T. W. Lucas. Exploring the world of agent-based simulations: Simple models, complex analyses. In *Proceedings of the Winter Simulation Conference*, volume 1, pages 116–126. IEEE, 2002.
- F. Santosa and W. W. Symes. Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1307–1330, 1986.
- T. Schreiber. Measuring information transfer. *Physical review letters*, 85(2):461, 2000.
- G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.
- C. E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.
- B. W. Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.
- M. Smoluchowski. The kinetic theory of brownian molecular motion and suspensions. *Ann. Phys*, 21:756–780, 1906.
- W.-M. Song, T. Di Matteo, and T. Aste. Building complex networks with platonic solids. *Physical Review E*, 85(4):046115, 2012a.
- W.-M. Song, T. Di Matteo, and T. Aste. Hierarchical information clustering by means of topologically embedded graphs. *PLoS one*, 7(3):e31929, 2012b.
- Y.-Y. Song and L. Ying. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130, 2015.
- Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

- H. A. Sturges. The choice of a class interval. *Journal of the american statistical association*, 21(153):65–66, 1926.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.
- A. N. Tikhonov, A. Goncharsky, V. Stepanov, and A. G. Yagola. *Numerical methods for the solution of ill-posed problems*, volume 328. Springer Science & Business Media, 1995.
- D. Tikk, L. T. Kóczy, and T. D. Gedeon. A survey on universal approximation and its limits in soft computing techniques. *International Journal of Approximate Reasoning*, 33(2):185–202, 2003.
- H. K. Ting. On the amount of information. *Theory of Probability & Its Applications*, 7(4):439–447, 1962.
- C. Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52(1):479–487, 1988.
- C. Tsallis. Introduction to nonextensive statistical mechanics: approaching a complex world. *Springer*, 1(1):2–1, 2009a.
- C. Tsallis. Nonadditive entropy and nonextensive statistical mechanics—an overview after 20 years. *Brazilian Journal of Physics*, 39:337–356, 2009b.
- M. Tumminello, T. Aste, T. Di Matteo, and R. N. Mantegna. A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences*, 102(30):10421–10426, 2005.
- M. Tumminello, T. Di Matteo, T. Aste, and R. N. Mantegna. Correlation based networks of equity returns sampled at different time horizons. *The European Physical Journal B*, 55(2):209–217, 2007.
- M. Tumminello, S. Micciche, F. Lillo, J. Piilo, and R. N. Mantegna. Statistically validated networks in bipartite complex systems. *PloS one*, 6(3):e17994, 2011.
- S. Umarov, C. Tsallis, and S. Steinberg. On aq-central limit theorem consistent with nonextensive statistical mechanics. *Milan journal of mathematics*, 76(1):307–328, 2008.
- S. Verdu and T. Weissman. The information lost in erasures. *IEEE Transactions on Information Theory*, 54(11):5030–5058, 2008.
- A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269, 1967.
- J. Vrbik. Small-sample corrections to kolmogorov-smirnov test statistic. *Pioneer Journal of Theoretical and Applied Statistics*, 15(1-2):15–23, 2018.
- Y. Wang and T. Aste. Homological neural networks. *in preparation*, 2022.
- L. Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5:501–532, 2018.
- N. Wiener. The theory of prediction. *Modern mathematics for engineers*, 1956.
- N. Wiener. *Cybernetics or Control and Communication in the Animal and the Machine*. MIT press, 2019.
- J. H. Wilkinson. Modern error analysis. *SIAM review*, 13(4):548–568, 1971.
- J. H. Wilkinson. *Rounding errors in algebraic processes*. Courier Corporation, 1994.
- S. S. Wilks. Certain generalizations in the analysis of variance. *Biometrika*, 24(3/4):471–494, 1934.
- P. L. Williams and R. D. Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.
- P. L. Williams and R. D. Beer. Generalized measures of information transfer. *arXiv preprint arXiv:1102.1507*, 2011.
- R. W. Yeung. A new outlook on shannon’s information measures. *IEEE transactions on information theory*, 37(3):466–474, 1991.
- N. Young. *An introduction to Hilbert space*. Cambridge university press, 1988.