Preface

Data Science and MATLAB

Data acquire their full value when they are properly analyzed and when the results are interpreted, communicated clearly, and distributed efficiently to end users. In this way data lead to making fully informed optimal decisions.

The field of *data science* is the key element in this chain. It is a modern discipline founded on principles, techniques and algorithms from statistics, computer science and mathematics. The objective is the extraction of information, knowledge, and value from data through a rigorous scientific approach.

Nowadays, *data science* has become a strategic investment priority for many companies, with the demand for *data scientist* positions exceeding that of all other disciplines by at least an order of magnitude. Rapid advances stemming from scientific research and sector investments are having an epoch-making impact which, according to some forecasts, will surpass that of the industrial revolution and the widespread diffusion of electric power in the last century. One example of these innovations is the strong emergence of Large Language Models (LLMs) and technologies such as GPT. Despite these extraordinary developments, it is essential that researchers maintain a solid grasp of the foundational elements of *data science* from which these advanced and revolutionary tools have arisen and been developed. Without these foundations, it will not be possible to understand the properties of these tools and, above all, to manage the results they produce when they are called upon to solve problems in mathematical-statistical contexts and in the analysis of complex data.

This book offers a broad treatment of the methods and techniques of *data science* algorithms, starting from the foundations of the discipline and arriving at robust dimension reduction. These methods can be used, for example, to predict the purchasing propensities of products and services for each individual customer and, more generally, to extract relevant information based on similarity metrics with other users. Customer profiling capabilities are now so advanced that, in some cases, roughly 70 likes on Facebook or Instagram from a user are sufficient to automatically and accurately infer highly sensitive information about their behaviour. For this reason, *data science* has also become the subject of ethical and philosophical reflection and legislative interventions to regulate its application, especially when the goal is to “learn” from data to equip machines with reasoning methods and adaptive behaviour, in the sense traditionally attributed to artificial intelligence.

Our book refers to MATLAB version 2025a and later versions, starting with the Copilot integration of ChatGPT within MATLAB for AI-assisted code generation and completions. R2025a is a revolution in MATLAB's interface and ergonomics. Included are the new web-based desktop, multi-tab figure management, performance improvements, artificial intelligence and advanced editing. These AI tools simplify and streamline the generation of code for statistical analysis and the creation of powerful graphics for the display of data and of the results of analyses.

This text introduces the discipline of *data science* from the ground up, following an innovative *coding* approach. Each concept is presented in detail and accompanied by practical application in MATLAB, which is one of the most widely used programming languages in this field, especially among physicists, engineers, and economists.

We chose MATLAB for this book because its classic development environment supports readable, open code (i.e., without resorting to complex practices for execution acceleration), ensures pedagogical clarity and excellent performance in numerical computation and vectorizable data structures (such as matrices). These properties yield high efficiency in simulations and 2D/3D graphical visualizations. We also chose MATLAB for its commitment to Open Science, an approach that promotes transparency and collaboration, key elements in the advancement of scientific research and dissemination of results.

The volume does not require specific prior skills, assuming only basic knowledge of descriptive and inferential statistics. A brief list of reference books on statistics is included at the end of the text under the heading "Suggestions for Further Reading". Our book also contains an introductory chapter section on MATLAB to make the material accessible to those without prior programming experience. Each method is presented with an exercise coded in MATLAB, together with its solution, to facilitate understanding of each step in the procedure. Applications to real-world problems are then illustrated to show the practical utility of the various methods described.

We believe that, thanks to these features, the volume is also useful to analysts in decision-making roles: business managers, marketing experts, financial operators, data managers and database administrators. They only require a basic knowledge of statistical indices and a willingness to apply data analysis techniques methodically in the support of business decisions.

There is often debate about the best language to accompany the study of *data science*. At the moment the preference leans toward R and Python, essentially because they are *open source* and therefore free. But see the penultimate bullet point in the list below.

We have said that transparency, pedagogical clarity, and code speed are the foundations of choosing MATLAB for this book. But there are also other points on which we invite the reader to reflect, especially after experimenting with the exercises in the book.

- Writing MATLAB code is particularly natural and simple compared to R and Python, making it suitable for newcomers to the field. The highly readable and well-structured code makes maintenance and update easy.
- Its numerous libraries have been developed by experts in various application areas, ranging from aerospace to automotive engineering, from time series analysis to text mining algorithms.
- The development environment offers good integration with other programming languages. In addition to running Python or R functions from MATLAB, with the Coder toolbox, you can automatically translate MATLAB code to C/C++ to increase execution speed and to facilitate *porting* to embedded processors used in industrial production (e.g. ARM-based processors such as the Raspberry Pi).
- MATLAB provides a simple and intuitive environment for creating advanced graphical interfaces and inserting controls such as scroll bars, sliders, and drop-down menus, making it a valuable teaching and communication tool.
- It is also well known that MATLAB excels in data visualization, offering advanced tools for creating complex charts and interactive data manipulation. This makes verifying and interpreting results more immediate and intuitive.
- MATLAB has built-in version control routines with GitHub integration for collaborative software development, offering sophisticated web services for code review and sharing. So MATLAB does not require deep knowledge of software versioning services.
- Currently the on-line version of MATLAB is free for everyone, up to 20 hours per month. Therefore all the code presented here can be utilized without license.
- It is also worth noting that the code provided by MATLAB offers the guarantees required for use in critical sectors such as embedded systems. In other words, a data scientist programming in MATLAB need

not worry about unknowingly invoking “software coming with ABSOLUTELY NO WARRANTY”. In other words, MATLAB offers a robust official support system, including detailed documentation and dedicated forums. This can be particularly helpful not only for beginners but also for those facing complex issues that require competent technical support.

To this list we add a personal element, since we are the creators of a software project called FSDA, *Flexible Statistics and Data Analysis*, which extends MATLAB with hundreds of functions ranging from robust multidimensional data analysis to tools and graphical interfaces that simplify and automate advanced *data science* techniques. The toolbox installs with a simple click on Add-Ons|Get Add-Ons in the MATLAB HOME tab. We are convinced that a data scientist’s work can greatly benefit from using FSDA, unless one pretends that data collected in real applications are as we imagine them in a Platonic world: error-free, anomaly-free, and perfectly matching our assumed model.

FSDA includes many functions for robust data analysis and most are outside the scope of this already surprisingly long book. But we should all always be aware that information manifests itself in various kinds of relationships, sometimes unexpected, which may be hidden by contamination in the data. Understanding and knowledge of the underlying patterns and structures leads to new hypotheses and models that characterize, though generally simplify, the data-generation process. We hope that this book is a step in this direction.

Structure of the book

The structure of the book is roughly divided into two parts. The first covers the basic concepts and tools for using MATLAB in data science. The second consists of specialized chapters providing a deeper and technical analysis of more advanced methods in the field. The following chapters are part of the first logical section, dedicated to the fundamentals of data science with MATLAB.

- In the first we present the MATLAB environment from scratch. We illustrate the different MATLAB data types (character, string, array, table, struct, and dictionary), techniques for handling and processing real data, and basic programming concepts (constructs `if else end` and loops `for` and `while`).

- In the chapter “Linear Algebra” we review basic linear algebra concepts, which are a necessary prerequisite for data science.

- In the chapter “Exploratory Data Analysis and Pivot Tables” we recap the statistical indices for univariate analysis, with attention to their implementation. We illustrate frequency distributions, confidence-interval construction, and graphical presentation of results for the full sample or subgroups, with particular focus on pivot table construction.

- The chapter “Importing Real-Time Data from the Web” shows how to load into MATLAB data from major global providers and economic-financial market software. Data from GitHub (or any other version-control system), leading financial databases (e.g. Bloomberg), economic databases (e.g. FRED and LSEG), social media (e.g. 𝕏), or statistical institutes (e.g. the Italian Statistical Institute, ISTAT) can be loaded into MATLAB with a single line of code. We also discuss importing large datasets and advanced options for forcing variables into specific formats. Since financial and economic data often come as time series, this chapter also introduces `timetable` and tools for date handling and changing series periodicity.

- The chapter “Random Variables: Density, Distributions and Parameter Estimation” presents the main theoretical distributions that could have generated the observed empirical distribution. For each distribution we show how to compute its density function, cumulative distribution function, quantiles, and random numbers generation from it in MATLAB.

- The chapter “Preliminary Data Processing” is devoted to data preprocessing, a vital phase of analysis because it removes inconsistencies and errors in the collected information that could bias results. This “data cleaning” must always precede statistical processing and subsequent model building. Attention in this text is given to robust statistical algorithms that withstand outliers and measurement errors.

- In the chapter “Correlation and Cograduation” we illustrate methods for analyzing and testing relationships between quantitative variables.

The topics covered in the subsequent, more advanced, chapters are as follows:

1. Univariate, bivariate, and multidimensional graphical representations, which provide a first impression of the patterns present in the data.

2. Association, which studies relationships between nominal or ordinal qualitative variables.

3. Distances and similarity indices, which highlight differences and analogies among the statistical units examined and form the necessary basis for behavioural segmentation or, more generally, data profiling, as well as for understanding advanced dimensionality-reduction techniques.

4. Principal component analysis, to reduce problem dimensionality with respect to variables, focusing attention on the most important aspects. Unlike other *data science* texts that treat these complex topics as "black boxes", with commands to invoke certain procedures and produce output from input, and unlike classic statistics texts that focus predominantly on mathematical aspects without practical follow-through, our text unpacks every instruction, line by line, needed to replicate each intermediate step. We believe this study method is indispensable for fully mastering the details of the statistical technique. For those who already know the details of various techniques, we provide GUIs that allow interactive engagement with the automatically produced output.

At the end of each chapter, a series of summary exercises are provided to deepen understanding of the concepts and explore additional aspects. Solutions to the exercises, along with extra material, are available online at the publisher's site: `https://biblioteca.giappichelli.it/studenti/`, accessible with the credentials associated with the book. The online pages dedicated to the book are organized according to the chapter index and are interactive, allowing readers to download supplementary material and answer questions to test their comprehension.

Every successful project is accompanied by certain assumptions and a dream. The assumption behind this book is that the ingredients for doing *data science* boil down to minimal mathematical tools and some programming: all one needs is a curious mind and the willingness to invest effort. The dream is to provide curious people with the right tools to enjoy the subject, illustrate the basics of statistical programming, and launch a *data scientist* career without too much struggle.

Acknowledgments

This book is the English version of the third edition of an Italian book. Its transformation into this edition was made possible through our close collaboration with MathWorks, the company that produces MATLAB. Our annual visits to their headquarters in Natick (Boston, US), attending the MATLAB Research Summit and MATLAB Advisory Board events, have been instrumental in fostering a rich exchange of ideas. These interactions allowed us to

engage closely with both the management and the developers at MathWorks. Through these collaborative experiences, we became aware of the broader potential and relevance of the content we had developed for the Italian audience. This realization inspired us to revise and update our work, at the same time translating the book into English. The intention is to reach an international audience. It is our hope that this English edition will continue to advance understanding and innovation within the global data science community.

We particularly want to express our heartfelt gratitude to the staff and management at The MathWorks Corporation for their unwavering support and invaluable insights, which have greatly enriched this process. We particularly thank Rob Purser, Vijay Iyer, Jos Martin, Andy Campbell and Fred Smith for significantly boosting our group's and the FSDA project's effectiveness and productivity with their contributions and those of their teams. A fuller list of those we thank is available in the README.md file of the GitHub repo of our toolbox FSDA: https://github.com/UniprJRC/FSDA.

Within MathWorks Italy, we would like to mention Paolo Panarese, Stefano Olivieri, Francesca Perino and Giovanna Galliano for recognizing our passion for statistical programming and encouraging and supporting advanced MATLAB use. We also thank Paola Vallauri, Fabrizia Grande, and Alessio Conte, who played important roles in facilitating MATLAB's educational adoption at the University of Parma. We cannot forget that it is thanks to all of them at the Italian branch of MathWorks that our group could be introduced to the specialists and expert developers at the company's headquarters in Natick.

Our final thanks go to the University of Parma, the Joint Research Center of the European Commission and the London School of Economics for granting the MATLAB licences, and to the students who reported typos in the preliminary Italian versions of the book. This English edition incorporates the valuable feedback we received. Special thanks to Federico Baio, Angela Borrello, Giacomo Boschi, Georgiana Flotta, Marika Palme and Eleonora Sula. We intend to continue this form of interaction.

Marco Riani
Anthony C. Atkinson
Fabrizio Laurini
Domenico Perrotta
Aldo Corbellini
Luigi Grossi
Gianluca Morelli
Francesca Torti