Philippe Rigollet
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue,
Cambridge, MA 02139-4307, USA
rigollet@math.mit.edu

Introduction to High-Dimensional Statistics

Christophe Giraud. Boca Raton, FL: Chapman & Hall/CRC, 2015. xv+252 pp. ISBN: 978-1-482-23794-8

The information era has witnessed an explosion in the collection of data that contain potentially useful information for a wide range of applications such as biology, sociology, pattern recognition, marketing and finance. As a result of this inflation, statisticians have developed a new set of tools that combine fundamental statistical concepts with profoundly new ideas. This set of problems and tools is broadly referred to as *high-dimensional statistics*.

The book *Introduction to High-Dimensional Statistics* by Christophe Giraud succeeds singularly at providing a structured introduction to this active field of research. It describes a statistical pipeline where statistical *principles* enable the development of new *methods*, which, in turn, require a new mathematical *analysis*.

Most of "classical statistics" focus on situations where the number p of parameters of a model is much smaller than the number n of observations. As a result, asymptotic results where $n \to \infty$ and p is fixed prevailed in the 20th century, relying on results such as the law of large number or the central limit theorem for example. In high dimensional statistics, it often the case that $p \gg n$ so that traditional asymptotic results no longer provide a theoretical explanation for the behavior of a statistical procedure. Several approaches can be found in the literature and this book follows a preponderant trend that consists in replacing sharp asymptotic statements such as "consistency" or "asymptotic normality" with *finite sample risk bounds*. Unfortunately, without further assumptions, such bounds are typically too large to be informative when $p \gg n$. In this context, the effectiveness of high-dimensional statistical methods has relied on structural assumptions. One such assumption that Giraud describes in great details is *sparsity*

The sparsity assumption essentially postulates that there exists a much smaller sub-model that can explain the data quite well. Finding this sub-model falls in the general framework of *model selection* and can be achieved using penalization. A systematic finite sample analysis of these techniques was developed in the nineties under a very general

setup devoid of any computational considerations. Chapter 2 revisits this general theory and instantiates it on several variations of sparsity assumption. A recent alternative to model selection, called *aggregation*, is explored in Chapter 3. It leads to statistical procedures that have optimality properties similar to model selection procedures but that are quite different in nature. Unfortunately, in the context of sparsity, both approaches require prohibitively high computation.

The statistical principles exploited in Chapter 2 and 3 do not represent a major break from the finite sample analysis of classical statistical procedures such as model selection in (low-dimensional) linear regression. A salient feature of high-dimensional statistics is the appearance of *computationally efficient* methods with provable guarantees as opposed to heuristics such as stepwise regression for example. The popular Lasso estimator employed in linear regression is one such example. Chapter 4 carries out detailed mathematical analysis of the prediction performance of this estimator. In particular, this analysis surveys the key ingredients of the proof in such a way that it is completely demystified. Moreover, the chapter explores several state-of-the-art algorithms to compute the Lasso estimator and points to R implementations. It is worth mentioning that this chapter focuses exclusively on prediction performance and leaves other important questions such as variable selection and parameter estimation as guided exercises.

The Lasso estimator, like many other estimators employed in high-dimensional statistics, requires tuning one or more parameters that significantly affect their performance. Procedures for choosing these turning parameters, including the celebrated "cross-validation" method, are presented mostly without proofs in Chapter 5, together with a thorough analysis of the square-root Lasso (a popular twist on the Lasso that can be tuned without knowing the variance of the noise).

More recently, high-dimensional statistics have witnessed the rise problems where the parameter of interest is no longer a vector but a matrix or a graph. In this context, surprisingly similar techniques apply not only under the sparsity assumption but also under the assumption that the matrix of interest has low rank, which turns out to be natural in this context. This is perhaps the first book to offer a clear and detailed treatment of matrix estimation in the context of *multivariate regression*. The treatment of *graphical models* is more superficial due to extra technicalities but the methods are clearly defined and motivated in view of the earlier chapters on sparse linear regression.

The rest of the book departs somewhat from the first chapters to describe other issues arising in high-dimensional statistics thus providing a bit of diversity to the reader interested in exploring other topics. On the one hand, Chapter 8 is devoted to *multiple*

testing with emphasis on the False Discovery Rate (FDR). The proofs in this chapter are much simpler than the first part of the book since they are quite different in nature and no connection with the results of the previous chapters is made. On the other hand, Chapter 9 covers supervised classification, a core topic in *statistical learning theory*. The treatment is fairly classical with upper bounds on the excess risk using fundamental tools from empirical process theory such as symmetrization and VC dimension. This chapter also provides a fairly detailed treatment of risk convexification, a technique that can be used to provide a unified framework to several classical machine learning methods such as boosting and support vector machines.

A striking aspect of this book is the omnipresence of computational considerations across chapters. The author carefully points to potential implementations, R packages and algorithmic details that have now become inherent to modern high-dimensional statistical research. Beyond a unified and cohesive treatment, Giraud also offers informative and fairly comprehensive bibliographical notes that point to the main results of the field as well as connected work. Once the subject has been mastered, the reader is invited to attempt at solving some of numerous exercises provided at the end of each chapter. These exercises provide detailed guidelines on how to derive key results from the recent literature and are one of the best features of this book. Put together, the exercises amount to the equivalent of another book and provide a lot of insight on how the core concepts encountered in the main text extend to other problems.

This book is not a global overview of all the aspects of high-dimensional statistics and focuses primarily on prediction performance, much in the spirit of statistical learning theory. However, it is arguably the most accessible overview yet published of the mathematical ideas and principles that one needs to master in order to enter the field of high-dimensional statistics. Indeed, several years have passed since the publication of *Statistics for High-Dimensional Data* by Peter Bühlmann and Sara van de Geer (2001), which is widely considered as the main reference on the theory of high-dimensional statistics. These years have allowed Giraud to distill the core elements of the literature and simplify some of the arguments so that *Introduction to High-Dimensional Statistics* can serve as a gentle introduction to the more advanced text. This feeling is reinforced by an engaging introduction (Chapter 1) that brings forward the key challenges associated to high dimensional data and a nice account of useful probabilistic inequalities in the appendix. It should be recommended to anyone interested in the main results of current research in high-dimensional statistics as well as anyone interested in acquiring the core mathematical skills to enter this area of research.

Philippe RIGOLLET
Massachusetts Institute of Technology

REFERENCES

Bühlmann, Peter and van de Geer, Sara (2011), *Statistics for High-Dimensional Data*, Springer Series in Statistics. Springer, Heidelberg.