High-Dimensional Statistics: A Non-Asymptotic Viewpoint

Martin J. Wainwright

Cambridge University Press 2019, xvii + 552 pages, £57.99, Hardback

ISBN 978-1-108-49802-9

Readership: Statistics/Machine Learning graduate students and researchers.

This is an excellent book. It provides a lucid, accessible and in-depth treatment of non-asymptotic high-dimensional statistical theory, which is critical as the underpinning of modern statistics and machine learning. It succeeds brilliantly in providing a self-contained overview of high-dimensional statistics, suitable for use in formal courses or for self-study by graduate-level students or researchers. The treatment is outstandingly clear and engaging, and the production is first-rate. It will quickly become essential reading and the key reference text in the field.

Conventional, classical statistics, as developed in the early 1900s, is founded on an asymptotic regime in which the dimension p of the parameter in the statistical model remains fixed as the sample size n grows to infinity. Standard laws of large numbers and the central limit theorem then furnish a general suite of inferential techniques, typically based on the asymptotic consistency, normality and efficiency of the maximum likelihood estimator. Such inferential techniques, which served as the bedrock of statistical analysis for decades, have been extended, from the 1980s onwards, by the development of refined, likelihood-based and bootstrap methods of distributional approximation: some of the key elements of this substantial theory of 'higher-order asymptotics' are described in the brief review article Young (2009). Typical focus in classical theory concerns a parametric model $F(y;\theta)$ indexed by a $p-$dimensional parameter $\theta = (\psi, \lambda)$, where ψ is a scalar interest parameter and λ is a $(p-1)-$dimensional nuisance parameter. Inference on ψ is based on a random sample Y of size n from $F(y;\theta)$ and, for instance, it is required to construct a confidence interval $I_\alpha(Y)$ for ψ, of nominal coverage $1-\alpha$. Bootstrap or analytic approximation is made for the sampling distribution of a 'pivot' $T(Y,\theta)$, such as the signed square root of the likelihood ratio statistic, or some modification thereof. This estimated sampling distribution is then used to construct an accurate confidence set $I_\alpha(Y)$, with the property, valid assuming only correctness of the model $F(y;\theta)$, that

$$Pr_\theta\{\psi \in I_\alpha(Y)\} = 1 - \alpha + O(n^{-r}),$$

for quantifiable r, typically $r = 1$ or $r = 3/2$. Though the error term $O(n^{-r})$ will typically depend on unknown quantities, so such a result is an asymptotic statement, the operational interpretation is immediate: as $n \to \infty$ the confidence set yields exactly the nominal coverage $1-\alpha$ under repeated sampling. In cases where λ is of low dimension, it is a rule of thumb that if $r = 3/2$ the confidence set will give essentially exact coverage for modest sample size, say $n = 10, 20$, though there are no guarantees from the theory of the magnitude of error for any finite n.

But, the data sets which arise in many areas of modern science and engineering generally have parameter dimension p of the same order, and often exceeding, sample size n, and for such problems classical statistical theory may fail to provide useful estimation or prediction, or simply break down completely. While investigations have established that accurate inference may often be obtained from higher-order asymptotics methodology in circumstances where the parameter dimension p is relatively large compared to available sample size n (see evidence contained in Barndorff-Nielsen and Cox, 1994, for instance), there has been relatively little systematic direct theoretical examination of what Wainwright terms 'high-dimensional asymptotics', where the pair (n,p) are taken to infinity simultaneously, in such a way that some scaling function of (n,p), and possibly other problem parameters, remains fixed or converges to some finite limit. Instead, focus in modern statistical theory has emphasized non-asymptotic results

in high-dimensional problems. In this theory, the pair (n, p) as well as other problem parameters are viewed as fixed, and high-probability statements, say about the error of a parameter estimator, are made as a function of them. As its title suggests, non-asymptotic, theoretical results of this type are the focus of this book. Chapter 1 gives a beautiful overview, illustrating through key examples involving linear discriminant analysis, covariance estimation and nonparametric regression, what can go wrong with classical statistics in high dimensions, and motivating persuasively the non-asymptotic viewpoint. The kind of non-asymptotic theory developed in the book, founded on obtaining bounds on the tails of a random quantity, or concentration inequalities which provide bounds on how a random variable deviates from some value, such as its mean, has its main value in being able to be used to predict some aspects of high-dimensional asymptotic phenomena, such as limiting forms, as (n, p) grow, of error probabilities in the linear discriminant problem. Further, the scaling functions that emerge in a non-asymptotic analysis can suggest the appropriate high-dimensional asymptotic analysis to perform in order to reveal relevant limiting distributional behavior. The non-asymptotic analysis is typically not used, or immediately available, to provide precise statements about behavior, say, of an estimator for a given, finite (n, p).

As an example of a non-asymptotic result, we take illustration from Wainwright (Example 7.14), concerning the classical linear model

$$Y = X\beta + \epsilon,$$

where the design matrix $X \in \mathbb{R}^{n\times p}$ is deterministic and the noise vector ϵ has independent elements, identically distributed as $N(0, \sigma^2)$. Assume that X satisfies some (checkable, since X is fixed) 'restricted eigenvalue' condition, and that it has $\max_{j=1,\ldots,p} \frac{\|X_j\|_2}{\sqrt{n}} \leq C$, where X_j is the jth column of X. Suppose further that the vector β is supported on a subset $S \subseteq \{1, 2, \ldots, p\}$ with $|S| = s$, so that β is sparse, with s non-zero elements. If we define the Lasso estimator $\hat{\beta}$ of β as the minimizer of

$$\frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1,$$

we have that, for constant K,

$$\|\hat{\beta} - \beta\|_2 \leq K\sqrt{s}\left\{\sqrt{\frac{2\log p}{n}} + \delta\right\},$$

with probability at least $1 - 2e^{-n\delta^2/2}$, for any $\delta > 0$, if we set $\lambda = 2C\sigma\left(\sqrt{\frac{2\log p}{n}} + \delta\right)$. Then we see that, provided K, which is defined in terms of C, σ^2 and the eigenvalue condition, stays fixed as (n, p) increase, the Lasso estimator $\hat{\beta}$ is consistent, as long as $\log p$ is dominated by n, if the size of the true β, as determined by s, remains fixed. Such a result predicts rather little, though, about the finite sample behaviour of the estimator $\hat{\beta}$: the bound on $\|\hat{\beta} - \beta\|_2$ depends, inter alia, on the unknown error variance σ^2 and the unknown true β, through its sparsity level s. As is the case for classical statistical theory developed for the fixed p regime, non-asymptotic results of this kind only offer precise guarantees, therefore, asymptotically. But, the beauty of the theory of high-dimensional statistics as described by Wainwright is precisely that non-asymptotic results can yield strong operational support to practical methods of data analysis, such as the Lasso described above. For instance, while the Lasso estimator is not universally optimal, it comes close to mimicing the properties of an oracle estimator (which knows the true state of nature) in many sparse settings, in estimation of the mean $E(Y)$ of Y when X is given. The Lasso predicts $E(Y)$ almost as well as an oracle which knows which of the elements of β are non-zero.

Chapter 1 of Wainwright's book gives also a very clear account of what enables statistics in the high-dimensional setting. What saves us is the reasonable expectation that high-dimensional data is actually endowed with some form of low-dimensional structure, typically some form of sparsity, which might

crudely be expressed as meaning that only s of the p parameters of the model are non-zero or non-negligible, where s is much smaller than p. Much of high-dimensional statistics therefore involves constructing models of intrinsically high-dimensional phenomena, but where the models incorporate some implicit form of low-dimensional structure, which can be successfully revealed from sample data. The introductory chapter is, in its own right, a tour de force, but sets the scene for a marvelous account of the mathematics and methodology of all the main elements of high-dimensional statistical theory. Beautifully signposted, the subsequent material of the book really divides into two types. Foundational material on tools and techniques, such as concentration inequalities, concentration of measure, uniform laws of large numbers, notions of covering and packing, reproducing kernel Hilbert spaces and techniques for obtaining minimax lower bounds is elegantly described. Of mathematical interest in its own right, this material is directed here to derive theory that is broadly applicable in high-dimensional statistics. Crucially for those interested in statistical practice, the book also provides a thorough account of the models and estimators used in data analysis. The text includes a series of chapters each focused on a particular class of statistical estimation problems, including covariance estimation, the sparse linear model, principal component analysis, estimators based on decomposable regularizers, estimation of low-rank matrices, graphical models and least squares estimation in a nonparametric setting: the principal aim is to shed light on the theoretical guarantees that are offered by widely used estimation techniques. Careful attention is paid throughout to computational considerations, such as the gains that are made by deriving from an initial otherwise NP-hard optimization problem a convex criterion that can be optimized efficiently, while ensuring that the resulting statistical procedure is almost as good as that initially considered. Ideas and formalities are illustrated throughout the text with carefully chosen examples, primarily of a theoretical, rather than data analytic, nature. A minor quibble of this reviewer is the lack of a glossary of notation: someone who is not immersed in the area might wonder what precisely $\asymp$ and $\precsim$ mean.

In summary, this is an authoritative, scholarly and highly useful summary of high-dimensional statistics. It is dense and not for the faint of heart, though the author offers excellent routemaps through the material. As a text, it stands natural comparison with Giroud (2015) and Bühlmann and van de Geer (2011), but to suggest any preference here would be invidious.

References

Barndorff-Nielsen, O.E. and Cox, D.R. (1994). *Inference and Asymptotics*. CRC Press.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Giroud, C. (2015). *Introduction to High-Dimensional Statistics*. CRC Press.

Young, G.A. (2009). Routes to higher-order accuracy in parametric inference. *Aust. N.Z. J. Stat.*, **51**, 115–126 .

G. Alastair Young
Department of Mathematics
Imperial College London
180 Queens Gate
London SW7 2AZ
U.K.
alastair.young@imperial.ac.uk