BOOK REVIEWS

BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 57, Number 3, July 2020, Pages 509–514
https://doi.org/10.1090/bull/1692
Article electronically published on January 28, 2020

High-dimensional statistics: A non-asymptotic viewpoint, by Martin J. Wainwright, Cambridge Series in Statistical and Probabilistic Mathematics, Vol. 48, Cambridge University Press, Cambridge, 2019, xvii+552 pp., ISBN 978-1-108-49802-9, US $79.99

1. Introduction

The book under review [5] provides a masterful exposition of various mathematical tools that are becoming increasingly common in the analysis of contemporary statistical problems. In addition to providing a rigorous and comprehensive overview of these tools, the author delves into the details of many illustrative examples to provide a convincing case for the general usefulness of the methods that are introduced. The book is clearly written from a theoretician's perspective, with only a smattering of concrete applications from computer science and engineering. However, the material can be appreciated by both theoretical and applied readers who are interested in gaining a deeper understanding of the inner workings of the statistical estimation procedures underlying many modern algorithms in machine learning and data science.

1.1. **High-dimensional statistics.** Theoretical statistics is, at its core, the study of convergence. As more data are acquired from a model, it becomes possible to perform increasingly accurate inference about the underlying data-generating mechanism. Much classical statistics focused on rigorously proving this intuition in an asymptotic sense, i.e., characterizing the limiting behavior of various statistical estimators as the number of data points tends to infinity [3]. However, a great deal of attention has been devoted in recent years to understanding the behavior of estimators when the amount of data available is not extremely large.

Of course, the size of a data set is relative; although the aforementioned results were proved under the assumption that the number of data points tends to infinity, such an assumption is clearly false in settings where the same statistical methods are regularly applied. Students in an introductory statistics class are often taught the "rule" that for a problem, such as hypothesis testing for a (univariate) population mean, the Central Limit Theorem and its associated normal approximation can be applied when the number of samples exceeds a threshold such as $n \geq 30$. Of course, this rule stems from the provable mathematical fact that for data distributions

2010 *Mathematics Subject Classification.* Primary 62-00; Secondary 62B10, 62F10, 62G05, 62H12.

satisfying appropriately mild regularity conditions, the empirical distribution of the samples is reasonably close to that of a normal distribution when $n \geq 30$. Even without delving into details, one can appreciate the fact that such a calculation depends on the fact that the data generating distribution (and therefore also the acquired samples) lies in one dimension. Indeed, if the data were 30-dimensional and the goal was to infer the 30-dimensional mean vector of the underlying distribution, it is hardly believable that only 30 samples would be sufficient—even if, as follows from classical theory in multivariate statistics [1], a version of the Central Limit Theorem indeed exists for the sample mean of 30-dimensional vectors as $n \to \infty$.

The example above describes a "high-dimensional" setting, for which various theoretical challenges arise in analyzing statistical estimators that are absent in low-dimensional settings. Indeed, sometimes the same estimation procedures must be altered in order to preserve their validity, e.g., constructing wider confidence intervals than would be appropriate in classical settings due to magnified noise. In other cases, high-dimensional considerations necessitate new classes of estimators that must consequently be studied rigorously. Again, we emphasize that the "high" dimensionality of a statistical problem is also a relative assessment, since the conclusion that the number of dimensions (which we henceforth denote by p) is large is simply due to the fact that the number of data samples is not substantially larger. Indeed, many contemporary statistical problems, which are motivated by real-world scientific problems, either possess the complicating property that the dimensionality of the data is relatively high (e.g., astronomical surveys of the night sky or genomic data), or the number of samples is relatively low (e.g., medical records about a rare disease)—or both. Concisely, we refer to such high-dimensional problems as settings where $p \gg n$, to distinguish them from the "p fixed, $n \to \infty$" setting of classical statistics.

1.2. **Nonasymptotic bounds.** Returning to the topic of convergence, any study of the $p \gg n$ setting must account for the size of p when considering the behavior of a statistical test or estimator as the sample size grows. Indeed, if $n \to \infty$, then $p \to \infty$, as well; however, classical multivariate statistical analysis only applies when p is a fixed quantity and $n \to \infty$. Philosophically, it is also unclear what would be meant by "sending both n and p to infinity" in this case: One can easily grasp the concept of characterizing the accuracy of a statistical inference procedure as more samples are acquired, but the dimensionality p is usually understood to be intrinsic to the problem and should not be "sent to infinity" as the sample size increases.

If we return to our original motivation of studying the $p \gg n$ scenario, we can remind ourselves that we are not actually interested in sending either p or n to infinity; rather, we simply wish to understand the validity of statistical inference procedures when p is large relative to n, possibly without imposing restrictions on the magnitude of n. Nonasymptotic bounds are specifically designed to address this problem: rather than providing statements about the limiting behavior of a statistical quantity as n and/or p tends to infinity, nonasymptotic bounds quantify the fluctuations of statistical quantities as a function of both n and p. Fixing p and sending n to infinity often then recovers the results of classical statistics; however, for the purpose of the contemporary problems mentioned earlier, we are primarily interested in settings where both p and n are of finite, comparable size. (Note that we measure "fluctuations" because the statistical quantity involved in an inference

procedure is nondeterministic, since it is computed using random draws from the data distribution. Thus, the nonasymptotic bounds will generally be guaranteed to hold with a certain high probability. Bounds on the expected value of functions of statistical estimators are also of interest, and they are similarly referred to as nonasymptotic bounds if they explicitly depend on n and p.)

2. Applications

To make the above discussion more concrete, we briefly describe two examples.

2.1. **Linear regression.** Suppose we have pairs of measurements $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$. We assume that the measurement pairs are independent and identically distributed (i.i.d.) and drawn from the linear model

$$y_i = x_i^T \beta^* + \epsilon_i,$$

where $\beta^* \in \mathbb{R}^p$ is the regression vector we wish to estimate. In other words, the x_i's, also known as the *predictors* or *covariates*, are drawn i.i.d. from some distribution, and the additive errors ϵ_i are drawn i.i.d. from a separate error distribution, giving rise to the y_i's.

The ordinary least squares (OLS) procedure seeks to minimize the sum of squared residuals

$$\widehat{\beta}_{OLS} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (y_i - x_i^T \beta)^2, \tag{2.1}$$

and in the regime where p is fixed and $n \to \infty$, classical statistics theory nicely characterizes the convergence of $\widehat{\beta}_{OLS}$ to β^*. On the other hand, if $p > n$, the solution to the minimization problem (2.1) will in general be nonunique, and the set of minimizers will be an unbounded, infinite set of vectors in $\mathbb{R}^p$.

One way to remedy this problem is to introduce a regularization term. Thus, instead of minimizing the expression (2.1), we solve

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^n (y_i - x_i^T \beta)^2 + \lambda \rho(\beta) \right\},$$

where $\rho : \mathbb{R}^p \to \mathbb{R}$ is an appropriately chosen function which encourages some sort of desirable structure in the solution and also makes the minimizer unique. The parameter $\lambda > 0$ controls the degree to which we enforce regularization vs. minimizing the original cost function. The Lasso [2] is perhaps the most famous example of regularization and corresponds to the choice $\rho(\beta) = \|\beta\|_1$. This choice of regularizer encourages the minimizer $\widehat{\beta}_{Lasso}$ to be "sparse," meaning it has a relatively small number of nonzero coordinates.

Turning to nonasymptotic bounds, it is desirable to quantify the accuracy of the regularization procedure when $p \gg n$. As mentioned earlier, the Lasso is used to promote sparsity in the solution. Thus, the bounds derived in high-dimensional statistics involve not only n and p, but also a third parameter $k = \big|\{j : \beta_j^* \neq 0\}\big|$, which counts the number of nonzero coordinates in the true parameter vector. The following is an example of a nonasymptotic bound in the literature (cf. Wainwright's book [5, Theorem 7.13 and Example 7.14]):

Theorem 2.1. *Suppose* $x_i \sim N(0, I_p)$ *and* $\epsilon_i \sim N(0, \sigma^2)$*, and suppose the regularization parameter is chosen such that* $\lambda = C_1\sigma\left(\sqrt{\frac{2\log p}{n}} + \delta\right)$*, for some* $\delta > 0$*. Then the Lasso solution satisfies the nonasymptotic error bound*

$$\|\widehat{\beta}_{Lasso} - \beta^*\|_2 \leq C_2\sqrt{k}\left(\sqrt{\frac{2\log p}{n}} + \delta\right),$$

with probability at least $1 - 2\exp(-n\delta^2/2)$*, where* $C_1, C_2 > 0$ *are universal constants which do not depend on* n, p, k*, or* σ*.*

If we take $\delta = \sqrt{\frac{2\log p}{n}}$ in Theorem 2.1, we see that for the corresponding choice of λ, we have a bound of the form

$$\|\widehat{\beta}_{Lasso} - \beta^*\|_2 \leq C_2'\sqrt{\frac{k\log p}{n}},$$

which holds with probability at least $1 - 2\exp(-\log p)$. Thus, the error incurred by the Lasso can still be small when the dimensionality p is substantially larger than n, provided the true sparsity level k is somewhat smaller than n (so that $k\log p \ll n$).

Intuitively, if we knew a priori that the unknown vector β^* was supported on a specific subset of k coordinates, we could simply perform an ordinary least squares fit of the data upon the relevant coordinates, which would be expected to succeed as long as n is large relative to k. In our setting, we have knowledge that β^* is supported on k coordinates, but we do not know *which* coordinates. The theorem above shows that by using the Lasso regularization technique, we can still perform accurate estimation when n is large relative to $k\log p$. Thus, although the sample size needs to inflate by a factor of $\log p$ in comparison to the case where we know the actual support set of β^*, the number of required samples is not nearly as large as the ambient dimension p.

2.2. **Covariance estimation.** As a second example, we turn to the problem of covariance matrix estimation. Suppose we have i.i.d. data $\{x_i\}_{i=1}^n$, where $x_i \in \mathbb{R}^p$. The goal is to estimate the covariance matrix $\Sigma = \mathrm{Cov}(x_i)$ of the data-generating distribution. Again, classical statistics theory studies the convergence of the sample covariance matrix $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$ to Σ as $n \to \infty$ (for simplicity, we assume that the data have been recentered to have zero mean). However, situations may arise when p is comparably large relative to n, so the convergence results that assume p is fixed and $n \to \infty$ no longer apply. It is therefore important to derive nonasymptotic bounds on the deviations between $\widehat{\Sigma}$ and Σ as a function of both n and p. Such deviation bounds would be useful, for instance, in analyzing the behavior of principal component analysis (PCA), where the goal is to reduce the dimensionality of data by projecting it onto the leading eigenvectors of Σ. Since only the estimate $\widehat{\Sigma}$ is computable from the data set, we wish to determine how the accuracy of the approximation will depend on the number of samples and/or the ambient dimensionality of the data.

An example of a nonasymptotic bound on the sample covariance matrix is the following (cf. Wainwright's book [5, Theorem 6.1 and Example 6.3]):

Theorem 2.2. *Suppose $x_i \sim N(0, \Sigma)$. For any $\delta > 0$, we have*

$$\|\widehat{\Sigma} - \Sigma\|_2 \leq \|\Sigma\|_2 \left(2\sqrt{\frac{p}{n}} + 2\delta + \left(\sqrt{\frac{p}{n}} + \delta\right)^2\right),$$

with probability at least $1 - 2\exp(-n\delta^2/2)$, where $\|\cdot\|_2$ denotes the matrix spectral norm.

Taking $\delta = \sqrt{\frac{p}{n}}$ in Theorem 2.2, we see that the relative error satisfies the bound

$$\frac{\|\widehat{\Sigma} - \Sigma\|_2}{\|\Sigma\|_2} \leq c\left(\sqrt{\frac{p}{n}} + \frac{p}{n}\right),$$

with probability at least $1 - \exp(p/2)$. In particular, the relative error can be controlled by the ratio $\frac{p}{n}$.

As in the case of linear regression discussed in the previous subsection, more sophisticated techniques may be employed to obtain tighter bounds under structural assumptions on the unknown matrix. For example, suppose Σ is known to have at most k nonzero entries per row. We can consider the thresholded matrix $T_\lambda(\widehat{\Sigma})$, which simply replaces each entry of $\widehat{\Sigma}$ by 0 if it lies in the interval $[-\lambda, \lambda]$. Then we have the following result (cf. Wainwright's book [5, Theorem 6.23 and Corollary 6.24]):

Theorem 2.3. *Suppose $x_i \sim N(0, \Sigma)$, where $\|\Sigma\|_2 \leq \sigma^2$ and Σ has at most k nonzero entries per row. If $n > \log p$, then for any $\delta > 0$, the thresholded matrix $T_\lambda(\widehat{\Sigma})$ with $\lambda = \sigma^2\left(8\sqrt{\frac{\log p}{n}} + \delta\right)$ satisfies*

$$\|T_\lambda(\widehat{\Sigma}) - \Sigma\|_2 \leq 2k\lambda,$$

with probability at least $1 - 8\exp\left(-n\min\{\delta, \delta^2\}/16\right)$.

Taking $\delta = \sqrt{\frac{\log p}{n}}$ in Theorem 2.3, we see that deviations of $T_\lambda(\widehat{\Sigma})$ from Σ can thus be controlled by the ratio $\frac{k^2 \log p}{n}$, which may be considerably smaller than the quantity $\frac{p}{n}$ appearing in the bound of Theorem 2.2 (which holds more generally for nonsparse matrices).

3. The book

Wainwright divides his book chapters into two rough categories: "Tools and techniques" (Chapters 2–5, 12, 14–15) and "Models and estimators" (Chapters 6–11, 13). Thus, the first third of the book effectively sets the stage by developing key concentration results that will form the backbone of the derivations needed to analyze the estimators presented in later chapters. The latter settings include linear regression, matrix estimation, edge structure estimation for graphical models, and nonparametric regression. In addition to demonstrating how the key technical tools can be used to derive nonasymptotic bounds on the statistical error of various high-dimensional estimators, the author closes the book in Chapter 15 by presenting general methods for deriving minimax lower bounds for estimation, meaning a minimal amount of error which must be incurred by *any* statistical estimator, due to random fluctuations in the data.

In addition to well-organized and crystal clear exposition, the author provides an extensive list of exercises at the end of each chapter. The exercises range from

easy (filling in omitted details in the proofs) to quite difficult (pointing the reader toward derivations of key technical contributions in research papers). Although a casual reader can gain a fairly comprehensive understanding of the topic without working through the exercises, a more serious reader will appreciate the fact that the exercises will help them gain confidence in applying the key techniques to a wider range of applications. The book is thus an excellent primary text for a graduate-level course, and the author notes that drafts of the book have already been used successfully in statistics curricula at various universities prior to printing.

Another nice feature of the book is the concluding section on "Bibliographic details and background" included in each chapter. These sections contain historic notes on where the ideas in the chapter first appeared in literature, as well as citations for the proofs of the theorems presented, where appropriate. Many of these sections also mention reference texts or survey papers for further reading.

It is worth mentioning the book [4], published one volume earlier in the same Cambridge University Press series. The two books contain a fairly large amount of overlap in terms of technical tools, in that they both begin with self-contained introductions of the basic building blocks used to derive concentration inequalities in high-dimensional statistics. The main difference between the two books lies in the application areas used by the authors to illustrate the power of these tools, in part influenced by the authors' individual research interests. For instance, the book under review provides a more comprehensive overview of statistical estimation rates of vectors and matrices with various structural constraints, as well as minimax lower bounds; in contrast, the book [4] covers topics such as community detection in networks, graph cuts, and random projections, which are all of contemporary interest in data science. Thus, an interested reader could benefit from reading both books in conjunction, although some of the introductory sections and supporting exercises might be skipped to avoid redundancy.

References

[1] T. W. Anderson, *An introduction to multivariate statistical analysis*, 3rd ed., Wiley Series in Probability and Statistics, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2003. MR1990662

[2] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Roy. Statist. Soc. Ser. B **58** (1996), no. 1, 267–288. MR1379242

[3] A. W. van der Vaart, *Asymptotic statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, vol. 3, Cambridge University Press, Cambridge, 1998. MR1652247

[4] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*, Cambridge Series in Statistical and Probabilistic Mathematics, vol. 47, Cambridge University Press, Cambridge, 2018. With a foreword by Sara van de Geer. MR3837109

[5] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, Cambridge Series in Statistical and Probabilistic Mathematics, vol. 48, Cambridge University Press, Cambridge, 2019. MR3967104

Po-Ling Loh
Department of Statistics
University of Wisconsin-Madison
1300 University Avenue
Madison, Wisconsin 53706
Email address: ploh@stat.wisc.edu