Quantitative Finance

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rquf20

High-Dimensional Probability: An Introduction with Applications in Data Science

Omiros Papaspiliopoulos

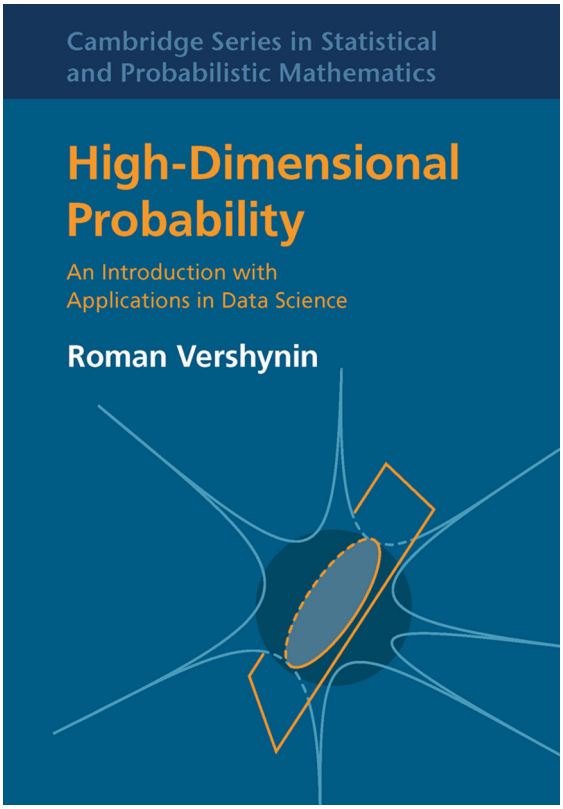
To cite this article: Omiros Papaspiliopoulos (2020) High-Dimensional Probability: An Introduction with Applications in Data Science, Quantitative Finance, 20:10, 1591-1594, DOI: 10.1080/14697688.2020.1813475

To link to this article: https://doi.org/10.1080/14697688.2020.1813475

Quantitative Finance, 2020
Vol. 20, No. 10, 1591–1594, https://doi.org/10.1080/14697688.2020.1813475

Routledge
Taylor & Francis Group

Book review



High-Dimensional Probability: An Introduction with Applications in Data Science,† by Roman Vershynin, Cambridge University Press (2018). ISBN 9781108231596. Online.

1. End quote

> Let us summarize our findings. A random projection of a set T in R^n onto an m-dimensional subspace approximately preserves the geometry of T if $m \gtrsim d(T)$. For smaller m, the projected set PT becomes approximately a round ball of diameter $\sim w_s(T)$ and its size does not shrink with m.

These are the last three sentences in Vershynin's excellent textbook (Section 11.3).

The fact that you can follow and learn something important from the last lines of a 300-page monograph on modern probability says a lot about this book and its priorities. The extract says quite a bit about the contents of this book, where concentration inequalities, random matrices, notions of sizes of sets and the probabilistic method are big recurring themes.

2. Concentration

Talagrand's well known quote (Talagrand 1996) goes like this: '*A random variable that depends in a smooth way on many independent random variables (but not too much on any of them) is essentially constant*'.

Concentration inequalities quantify how constant such variables are, specifically they bound their deviation from a central deterministic value, typically their mean or median. Strong laws of large numbers and central limit theorems give some clues on the concentration of averages of independent random variables around the population mean. Say, for iid X_i with mean μ and variance v,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i/n - \mu\right| > \epsilon\right) \lesssim \frac{2}{\sqrt{2\pi}} \mathrm{e}^{-n\epsilon^2/(2v)},$$

† An earlier and shorter version of this review originally appeared in the Bachelier Finance Society's January Newsletter.

but this inequality holds asymptotically in n. Contrast this with Talagrand's concentration inequality (Theorem 5.2.16) according to which for bounded iid X_i and f convex and Lipschitz

$$\|f(X_1,\dots,X_n) - \mathbb{E}[f(X_1,\dots,X_n)]\|_{\psi_2} \le C\|f\|_{Lip},$$

where the norm on the left is the so-called sub-Gaussian (Section 2.5.2) and that on the right the Lipschitz - and C a constant independent of n and ϵ. Applied to the function that returns the sample average we obtain now for a constant c,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i/n - \mu\right| > \epsilon\right) \le 2\,\mathrm{e}^{-cn\epsilon^2}.$$

Notice that now the result is non-asymptotic and entails the same dependence on n and ϵ as that implied by the CLT. In fact, there are way more direct concentration inequalities for averages (e.g. Hoeffding inequalities for this setting, Section 2.2), but it is remarkable that such a general result is also so precise.

3. High-dimensional probability: random structures, questions and answers

'High-dimensional probability is an area of probability theory that studies random objects in R^n, where the dimension n can be very large. The book places particular emphasis on random vectors, random matrices, and random projections'. (Preface)

The following three Data Science applications treated in the book provide a good motivation on random data structures, the type of questions high-dimensional probability is trying to answer and the role concentration inequalities (and the other building blocks developed in the book) play in this game.

3.1. Networks

Consider a so-called Erdös–Rényi random graph, a simple probabilistic model for real-world networks, according to which an edge exists between any two nodes with probability p. It is elementary that the expected degree of a node in this model with n nodes is $(n-1)p$. What is less obvious, and can be established using the Chernoff's concentration inequality (related to the Hoeffding discussed earlier), is that such graphs are almost regular provided that p is not too small; precisely, when $(n-1)p \ge C\log n$, with high probability all vertices have degrees $(n-1)p(1\pm 0.1)$ (Proposition 2.4.1).

3.2. Unsupervised learning

Spectral clustering is a method for the unsupervised classification of data into clusters. It assumes that n observations are represented as a graph and a symmetric adjacency martrix A contains a measure of affinity for each pair of observations. (For example, there might be d-dimensional data x_i, and $A_{ij} = k(x_i, x_j)$ for some positive definite kernel). Suppose for simplicity that A contains 0/1 values only. (For example, a K-nearest neighbour criterion is applied to the kernel values $k(x_i, x_j)$ to turn them into 0/1, see e.g. Section 14.5.3 of Hastie *et al.* 2001). Then, spectral clustering does the following; for simplicity suppose we wish to classify the data into two clusters; it obtains the eigenvector corresponding to second largest eigenvalue of A, and partitions the data according to the sign of the corresponding element of the eigenvector.† What can we say about the performance of this method in recovering clusters? Consider a data generating truth that is a two-community extension of the Erdös-Rényi random graph we saw earlier: the nodes of the graph are split in two communities, an edge between two nodes in the same community exists with probability p and between two in different communities with probability $q \ll p$; this is the stochastic block model (Section 4.5.1). Then, spectral clustering applied to the adjency matrix A computed on such data, with probability at least $1 - 4e^{-n}$ will identify the two communities correctly up to m misclassified vertices, where m depends on p, q but not n - and is proportional to $1/(p-q)^2$ (Theorem 4.5.6). The mechanics behind this result are as follows. (i) Basic linear algebra shows that the algorithm applied on $\mathbb{E}[A]$ (i.e. the matrix with elements p/q, not 0/1) yields the right clusters. (ii) We then need to understand the *concentration* of the *random matrix A* around its mean; this is achieved using bounds on the norm of sub-Gaussian random matrices (Section 4.4). (iii) Finally, we need to relate the spectra of the stochastic and deterministic matrices described before, which is typical of perturbation theory, e.g. the Davis–Kahan theorem (Theorem 4.5.5).

3.3. Supervised learning

We wish to build a predictive model for response y_i by taking a linear combination of p-dimensional predictors x_i, for $i = 1,\dots,n$. The Least Absolute Shrinkage and Selection Operator does so by solving the following *penalized likelihood* optimization problem:

$$\min \|Y - X\beta\|_2 \quad \text{s.t. } \|\beta\|_1 \le R,$$

where Y, X are vector and matrix containing the y_i's and x_i's respectively, $\|\cdot\|_1$ is the L1 norm, and R is a complexity penalty (Section 10.6). How good is the predictive performance of lasso? Here is a result on the *root mean squared error*, which is implied a by a number of results found in Section 10.6 of Vershynin's book. Suppose that the data generating truth is $Y = X\beta^* + w$, the x_i's are independent, isotropic, sub-Gaussian with $K = \max_i \|x_i\|_{\psi_2}$ (recall the sub-Gaussian norm we discussed earlier), and $\|\cdot\|_0$ denote the L_0 (pseudo)-norm, i.e. the number of non-zero entries. Then, provided $n \ge CK^4\|\beta^*\|_0 \log p$, and $R = \|\beta^*\|_1$, with probability at least $1 - 2\,\mathrm{e}^{-\|\beta^*\|\log p}$,

$$\frac{\|X(\beta-\beta^*)\|_2}{\sqrt{n}} \le CK\frac{\|w\|_2}{\sqrt{n}}\sqrt{\frac{\|\beta^*\|_0 \log p}{n}}.$$

† This is how spectral clustering is described in Chapter 4 of the book, although typically the method is applied a little differently.

This amazing result showcases the potential of lasso for predictive modelling in high-dimensional problems ($p \gg n$). Its proof involves the matrix deviation inequality (Chapter 9) and Talagrand's comparison inequality for random processes with sub-Gaussian entries (Chapter 8) - which in turn involves the spherical width $w_s(T)$ first discussed in the beginning of this review.

4. Why high-dimensional probability and not just probability

Given that the discussed Data Science examples can be considered regardless of dimension and that the concentration inequalities on performance are non-asymptotic, maybe it is worth summarizing in what ways the high-dimensions are relevant in the previous examples - and in what ways concentration inequalities interweave with high-dimensions to lead to modern asymptotics. The result about the degree distribution in networks holds for any network size n, provided p is not too small. For smaller values of p one can use the concentration inequalities discussed in the example to carry out asymptotics in n to answer similar questions. The probability that the stated spectral clustering performance holds is increasing with the training data size n, and the relative misclassification error gets smaller with n, therefore here we really have a high-dimensional probability result. Additionally, the inequality obtained allows for refined analysis as either n or $p-q$ vary in different ways. This type of analysis is a major feat of the result we obtained for the prediction problem, which allows high-dimensional regimes in both n and p to be simultaneously considered. There is yet another, more subtle way in which high-dimensions and the inequalities involved in these results interact; there are constants that depend on the structure of the problem and the inequalities used in the derivations (but not on n and p); with increasing dimensions the constants are less critical for the obtained result to be practically relevant.

5. The book

A good textbook is as much about learning as about learning something specific. Vershynin's high-dimensional probability is a good textbook. When developing a topic, it starts from the simplest idea, it examines its weaknesses and builds up to a better idea; this is superbly done when bounding the tail probabilities of binomial distributions in Chapter 2. It always prioritizes high-level narrative to technical details; the reader never loses sight of the main theme, arguments are kept to their essense, side results are given as exercises and important special cases are given priority over the most general statements. Intuition is at least as important as the techniques; this is usually the hardest to communicate in a book, compared for example to a classroom presentation, but it comes across beautifully in this book, as for example in the proof of the 'decoupling' theorem in Chapter 6. Finally, it shows sympathy to the reader, when sympathy is due (and much appreciated!): *'The definition of the VC dimension may take some time to fully comprehend. We work out a few examples to illustrate this notion'*. (Chapter 8).

One fundamental theme in the book is that of concentration inequalities. Another broad theme is that of geometric and combinatorial notions of 'size' of a set. Chapter 4 introduces and studies the notions of ϵ-nets, covering and packing numbers. A multiscale version of ϵ-nets, known as chaining, is developed in Chapter 8, where also the VC dimension is introduced, explained and related to packing. Chapter 7 introduces the notions of Gaussian and spherical width of a set, the concept of the stable dimension of a set (and its surprising and sharp contrast from that of the familiar linear-algebraic notion of dimension) as well as that of stable rank of a matrix; interestingly, the notion of stable dimension is introduced at this level of generality for the first time in this book.

Random matrices and random projections is another recurring topic. Not in the sense of the Tracy-Widom distribution and related ideas, which are not at all covered in this textbook. Neither in terms of randomized linear algebra, which is also not discussed at all but I think it would fit nicely given the existing structure. Rather, about norms of random matrices (Chapter 4) -bounds on eigenvalues and eigenvectors are obtained by means of basic perturbation theory- concentration inequalities for sums of random matrices (Chapter 5) and quadratic forms (Chapter 6), dimension reduction by random projections and the well-known Johnson-Lindenstrauss lemma (Chapters 5-7-9-11, a development that concludes with the sentences that open this review).

6. Other books

It might be worth to put this book in perspective relative to two (out of a good number of) recent publications with related themes. Boucheron *et al.* (2013) is an exhaustive treatment of concentration inequalities and the probabilistic method. Wainwright (2019), which is the next title in the same CUP series, covers both high dimensional probability and its applications to statistics and machine learning.

7. Buy it!

I could characterize Vershynin's book as I did for Talagrand's concentration inequality: it is remarkable that something so general is also so precise. In that respect, content-wise, it is a real addition relative to anything else I have seen published.

I think everyone who works in modern stochastics should be familiar with this book. Parts of the book could and probably should be incorporated in any modern intermediate probability course. Interestingly, measure-theoretic probability is never really required in this textbook. This potentially opens the possibility of designing and redesigning graduate probability courses along the narrative in this textbook for graduate programs in Data Science. As a reference book, I very much value that I can jump into any part in the book

and relatively easily follow the presentation, even when not recalling all previous developments.

But books, even on mathematics, or maybe even more so those in mathematics, should be enjoyable. And this one is unbeatable in this respect!

References

Boucheron, S., Lugosi, G. and Massart, P., *Concentration Inequalities: A Nonasymptotic Theory of Independence*, 2013 (OUP: Oxford).

Hastie, T., Tibshirani, R. and Friedman, J., *The Elements of Statistical Learning*, Springer Series in Statistics, 2001 (Springer New York Inc.: New York, NY).

Talagrand, M., A new look at independence. *Ann. Probab.*, 1996, **24**(1), 1–34.

Wainwright, M.J., *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge Series in Statistical and Probabilistic Mathematics, 2019 (Cambridge University Press: Cambridge).

Omiros Papaspiliopoulos
ICREA and Universitat Pompeu Fabra

Omiros Papaspiliopoulos is an ICREA Research Professor, based at Universitat Pompeu Fabra, and Director of the Data Science Center at Barcelona Graduate School of Economics. His research intersects Statistics, Machine Learning and Applied Mathematics and is primarily on computational statistics and algorithms. He currently serves as co-editor of *Biometrika* and has served as an Associate Editor for *Biometrika*, *Journal of the Royal Statistical Society series B, SIAM Journal of Uncertainty Quantification*. In 2010 he received the Royal Statistical Society's Guy Medal.