Proc. Int. Cong. of Math. – 2018
Rio de Janeiro, Vol. 3 (2845–2864)

RANDOM MATRICES AND HIGH-DIMENSIONAL STATISTICS: BEYOND COVARIANCE MATRICES

Noureddine El Karoui

Abstract

The last twenty-or-so years have seen spectacular progress in our understanding of the fine spectral properties of large-dimensional random matrices. These results have also shown light on the behavior of various statistical estimators used in multivariate statistics. In this short note, we will describe new strands of results, which show that intuition and techniques built on the theory of random matrices and concentration of measure ideas shed new light and bring to the fore new ideas about an arguably even more important set of statistical tools, namely M-estimators and certain bootstrap methods. All the results are obtained in the large n, large p setting, where both the number of observations and the number of predictors go to infinity.

1 Introduction

Random matrices have a very long history in multivariate statistics, going as far back as Wishart [1928]. Traditionally, they have been associated with problems arising from techniques such as Principal Components Analysis (PCA) Pearson [1901], Hotelling [1933], Anderson [1963], and Jolliffe [2002] or covariance matrix estimation where there is a natural focus on estimating spectral properties of large data matrices. We start by setting up precisely the problem and reviewing some of those important results before moving on to new statistical developments.

1.1 Setup. In most of this short review, we will be concerned with data stored in a matrix X, with n rows and p columns. n denotes the number of observations of p dimensional vectors available to the data analyst. The i-th row of X is denoted X_i' and $X_i \in \mathbb{R}^p$ is referred to as the i-th vector of covariates. p, the dimension of X_i, is the number of measurements per observation. If one works with financial data for instance Laloux, Cizeau,

The author gratefully acknowledges the support of grant NSF DMS-1510172. He would also like to thank Peter Bickel and Elizabeth Purdom for numerous discussions on these and related topics over the years.
MSC2010: primary 62F12; secondary 60F99, 62F40.

Bouchaud, and M. Potters [1999], p may be the number of assets in one's portfolio, n the number of days where those assets are monitored and $X_{i,j}$ may be the daily return of asset j on day i.

Traditional asymptotics. Traditionally, statistical theory has been concerned with studying the properties of estimators, i.e. functions of the data matrix X (and possibly other random variables), as $n \to \infty$ while p stayed fixed Anderson [1984] and Huber [1972] or was growing slowly with n Portnoy [1984] and Mammen [1989]. While mathematically and statistically interesting at the time, these sorts of problems are now really well-understood and their asymptotic analysis essentially amounts to doing probabilistic perturbation analysis (see more generally van der Vaart [1998]).

Modern developments. However, in the last two decades, technological advances in data collection have made it possible to work with datasets where both n and p are large: in genomics, p may be of order tens of thousands or millions and hundreds of observations Ramaswamy et al. [2001], data collected from internet companies may have millions of predictors *Criteo* [n.d.] and billions of observations, whereas financial data collected daily on a few hundreds of companies would yield after a year a dataset with hundreds of observations and hundreds of predictors Laloux, Cizeau, Bouchaud, and M. Potters [1999].

The case for "large p, large n". It is therefore now natural to study the so called "large n, large p" setting Johnstone [2001, 2007] where p and n grow to infinity but $p/n \to \kappa \in (0, \infty)$. On a more mathematical note, the ratio p/n can be somewhat informally seen as one measure of statistical difficulty of the problem. Fixing it amounts to doing asymptotics while the difficulty of the statistical problem stays constant and hence should (or at least could) yield asymptotic approximations of better quality than their traditional "fixed p, large n" counterparts. This is what we will see in some of the results described below. Furthermore, in the "fixed p, large n" settings, many asymptotic optimality results are meaningful only when it comes to relative errors, however absolute errors are typically infinitesimal and as such may not matter very much to applied statisticians and data analysts. By contrast, we will see that in the "large p, large n" setting, analyses predict substantial absolute differences between methods and as such may inform practitioners in the decision of what methods to use.

1.2 Modern random matrices. A key tool in multivariate statistics is the so-called sample covariance matrix, usually denoted, for an $n \times p$ data matrix X,

$$\widehat{\Sigma} = \frac{1}{n-1}(X - \bar{X})'(X - \bar{X}) .$$

Here $\bar{X} = 1_n \widehat{\mu}'$, where $\widehat{\mu} \in \mathbb{R}^p$ is the sample mean of the columns, i.e $\widehat{\mu} = X'1_n/n$. (We use $'$ to denote transposition throughout the paper; 1_n denotes the vector whose entries are all 1 in n dimension.). The $p \times p$ matrix $\widehat{\Sigma}$ therefore simply contains the empirical covariances between the various observed covariates.

This matrix is of course at the heart of much of multivariate statistics as it is the fundamental building block of principal components analysis (PCA) - probably the most widely used dimensionality reduction technique and the template for numerous modern variations - variants such as canonical correlation analysis Anderson [1984], and also plays a key role in the analysis of many supervised learning techniques.

To make things concrete, let us return to PCA. In that technique, practically speaking, the observations $\{X_i\}_{i=1}^n$ are projected onto the eigenvectors of $\widehat{\Sigma}$ to perform dimensionality reduction and allow for visualization; see Hastie, R. Tibshirani, and Friedman [2009] for a concrete introduction. A recurring question is how many dimensions should be used for this projection Cattell [1966]. This in turn revolves around estimation of eigenvalues questions.

Classical bulk results. To get a sense of the utility of large n, large p asymptotics in this context, we can return to a classical result Marčenko and L. A. Pastur [1967], which of course was later extended Wachter [1978], Silverstein [1995], Götze and Tikhomirov [2004], Pajor and L. Pastur [2009], and El Karoui [2009] and says the following :

Theorem 1.1 (Marchenko-Pastur)**.** *Suppose X_i's are independent and identically distributed (i.i.d) random variables with mean 0 and covariance identity, i.e.* $\operatorname{cov}(X_i) = \mathbf{E}\left((X_i - \mathbf{E}(X_i))(X_i - \mathbf{E}(X_i))'\right) = \mathrm{Id}_p$ *and mild concentration properties (see above references for details). Suppose further that $p/n \to \kappa \in (0,1)$. Then the empirical distribution of the eigenvalues of $\widehat{\Sigma}$ is asymptotically non-random and converges weakly almost surely to F_κ, a distribution whose density can be written*

$$f_\kappa(x) = \frac{\sqrt{(b_\kappa - x)(x - a_\kappa)}}{2\pi x \kappa} 1_{a_\kappa \le x \le b_\kappa} , \tag{1}$$

where $b_\kappa = (1 + \sqrt{\kappa})^2$ and $a_\kappa = (1 - \sqrt{\kappa})^2$.

This result already illustrates the great difference between modern (i.e. large n, large p) asymptotics and the classical setting where $p = o(n)$. In this latter case, the empirical

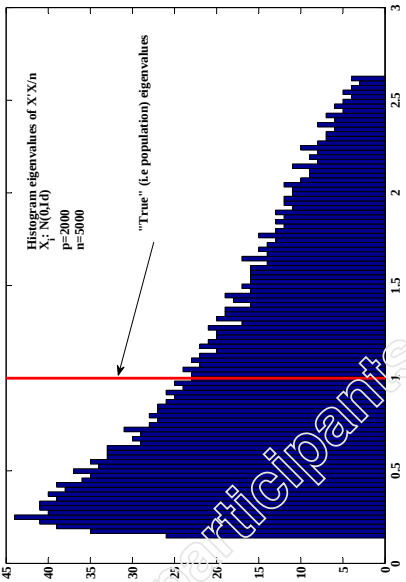

Figure 1: Illustration of Marchenko-Pastur law and high-dimensional estimation problem; n=500, p=200; $X_i \sim \mathcal{N}(0, \mathrm{Id}_p)$, i.i.d

distribution of eigenvalues goes, under the assumption of the previous theorem, to a point mass at 1; informally speaking all eigenvalues are consistently (loosely speaking correctly) estimated. The above theorem clearly shows that it is not the case in the "large n, large p" setting.

We can also illustrate the problem with a simple picture, comparing the histogram of observed eigenvalues of $\widehat{\Sigma}$ with the population eigenvalues, i.e. those of $\mathrm{cov}\,(X_i) = \Sigma$. See Figure 1, p. 2848.

This picture clearly illustrates the issue that the new paradigm of high-dimensional statistics creates: even though elementary concentration bounds show that entry-per-entry, i.e. in ℓ_∞ norm, estimation of Σ by e.g. $\widehat{\Sigma}$ is near trivial in the setup we consider, estimation of the spectrum of Σ may not be trivial. We refer the interested reader to El Karoui [2008] and Bickel and Levina [2008] (and Chaudhuri, Drton, and Richardson [2007] in the low-dimensional setting) for early work taking advantage of structure in the covariance matrix to improve estimation and to the recent Bun, Bouchaud, and Potters [2017] for a survey of applied random matrix theoretic work related to the questions we just discussed.

Right edge results. In the context of PCA, it is natural to ask questions about the largest eigenvalues of sample covariance matrices, as they could be used in a sequential testing fashion to determine how many components to keep in PCA.

A seminal result in this area in statistics is due to Johnstone who showed, building up on Tracy and Widom [1994b,a, 1996], the following remarkable result in Johnstone [2001].

Theorem 1.2 (Johnstone). *Suppose X_i's are i.i.d $\mathfrak{N}(0, \mathrm{Id}_p)$ and denote by l_1 the largest eigenvalue of $(n-1)\widehat{\Sigma}$. Then as p and n tend to infinity, while $p/n \to \kappa \in (0, \infty)$, we have*

$$\frac{l_1(\widehat{\Sigma}) - \mu_{n-2,p}}{\sigma_{n-2,p}} \Longrightarrow TW_1 \,, \tag{2}$$

with

$$\mu_{n,p} = (\sqrt{n} + \sqrt{p})^2 \quad \text{and } \sigma_{n,p} = \left(\sqrt{p} + \sqrt{n}\right)\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{p}}\right)^{1/3} .$$

Here TW_1 is the Tracy-Widom distribution appearing in the study of the Gaussian Orthogonal Ensemble Mehta [1991] and Deift [1999] and $\Longrightarrow$ denotes weak convergence.

In short, the largest eigenvalue of a sample covariance matrix computed from Gaussian data with identity covariance has fluctuations of size $n^{-2/3}$ around the edge of the Marchenko-Pastur distribution and the law of these fluctuations is asymptotically Tracy-Widom. Despite the fact that a great deal had been analytically known by statisticians about these questions James [1964], Constantine [1963], and Muirhead [1982] for a number of years, both the scale and the nature of the fluctuations discovered by Johnstone in his breakthrough paper came as a great surprise to the statistics community.

Johnstone's work is also connected to Forrester [1993] and Johansson [2000]. Later work extended Johnstone's result in many directions: to cite a few, see Soshnikov [2002] for results concerning the first k eigenvalues, for any fixed k, and relaxed distributional assumptions, Karoui [2003] for the case p/n tends to 0 or infinity at any rate, Baik, Ben Arous, and Péché [2005] for the discovery of very important phase transitions under low rank perturbation of $\Sigma = \mathrm{Id}_p$, El Karoui [2007] for the first result on general Σ and Lee and Schnelli [2016] for recent and powerful extensions of great potential in statistics.

This line of research continues with deep and insightful papers Bloemendal, Knowles, Yau, and Yin [2016] and has also benefited from progress in proving universality results - see for instance Erdős and Yau [2012] and Tao and Vu [2012].

One's enthusiasm for the broad applicability of such results in practice may nonetheless have been tempered by connections made with concentration of measure techniques Ledoux [2001] and Boucheron, Lugosi, and Massart [2013] for instance in Karoui and Koesters [2011]. Those results implied that most of the results above were intimately

linked to effectively geometric (and not probabilistic) assumptions made about the data and that when these easy-to-check-on-the-data assumptions were violated, the results mentioned above did not hold.

Other directions. The problems discussed above are of course very linear in nature. As such they have a broad reach beyond linear dimensionality reduction (see below and Karoui and Koesters [2011] for an example of a dimension-adaptive improvement of linear classification methods). Naturally, the reach of random matrix methods has extended beyond the strictly linear setup. For instance, the beautiful paper Koltchinskii and Giné [2000] studied the spectrum of so-called kernel random matrices, i.e. matrices with entries $K(i,j) = K(X_i, X_j)$ in the classical setting where p grows slowly with n. These results are important for understanding kernel methods in Statistics, which generalize standard methods to higher-dimensional spaces where the inner product between the de-facto observations is not the standard inner product anymore Wahba [1990] and Schölkopf and Smola [2002]. These matrices have been well understood in the high-dimensional case for quite a few years now El Karoui [2010] and Do and Vu [2013]. Random matrix results also have had interesting applications in randomized linear algebra and numerical optimization, and have been useful in speeding up various algorithms or allowing them to scale to very large data sizes - see for instance Achlioptas and McSherry [2007] and Drineas, Kannan, and Mahoney [2006] and follow-up results. These results typically use mathematically fairly coarse but very nice and broadly applicable bounds Tropp [2012] to prove the reliability of the algorithms under study, a function of the fact that they have to hold in a pretty general setting to be useful to practitioners.

2 Beyond covariance matrices: M-estimators

The previous section reviewed results in random matrix theory that could be useful for tasks in exploratory data analysis and generally unsupervised learning. However, much of statistics is concerned with the situation where one observes a scalar response, generically denoted $Y_i \in \mathbb{R}$, associated with the vector of predictors $X_i \in \mathbb{R}^p$. The simplest model of relationship between the two is the linear model where

(linear-model) $$\forall i\ , 1 \leq i \leq n, \quad Y_i = X_i'\beta_0 + \epsilon_i\ .$$

Here the data $\{Y_i, X_i\}_{i=1}^n$ are observed. The parameter of interest $\beta_0 \in \mathbb{R}^p$ is unobserved and so are the errors $\epsilon_i \in \mathbb{R}$. Typically, and in this short review, $\{\epsilon_i\}_{i=1}^n$ are assumed to be i.i.d from a certain distribution. The question the statistician faces is to estimate β_0. This is often done by solving an optimization problem, i.e. using a so-called M-estimator: for

a loss function ℓ chosen by the user, β_0 is estimated through

$$\widehat{\beta}_\ell = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \ell(Y_i; X_i'\beta) \ .$$

In the context of the linear model described above, one often uses the less general formulation

$$\widehat{\beta}_\rho = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho(Y_i - X_i'\beta) \ . \tag{3}$$

These estimators and the related family of generalized linear models McCullagh and Nelder [1989] are of fundamental importance in both theoretical and applied statistics and statistical learning, in academia and industry Chapelle, Manavoglu, and Rosales [2014] and Wood, Goude, and Shaw [2015].

2.1 Classical results: large n, small p. As such these estimators have received a great amount of attention Relles [1968] and Huber [1973, 1981]. In the classical case, i.e. p fixed and $n \to \infty$, Huber [1973] showed, under mild conditions, that $\widehat{\beta}_\rho$ is asymptotically normally distributed with mean 0 and covariance, if ϵ is a random variable with the same distribution as ϵ_i's mentioned in Equation (linear-model),

$$\text{cov}\left(\widehat{\beta}_\rho\right) = (X'X)^{-1} \frac{\mathbf{E}\left(\psi^2(\epsilon)\right)}{[\mathbf{E}\left(\psi'(\epsilon)\right)]^2} \ , \text{ where } \psi = \rho' \ .$$

This result is striking for at least two reasons : 1) the impact of the design matrix X, is decoupled from that of the error distribution ϵ; 2) finding the optimal estimator in this class is fairly simple as one just needs to find the function ψ that minimizes $\frac{\mathbf{E}(\psi^2(\epsilon))}{[\mathbf{E}(\psi'(\epsilon))]^2}$. In fact, Huber carried out this program and showed that in low-dimension, when ϵ has a density f_ϵ, the optimal loss function is

$$\rho_{\text{opt}} = -\log f_\epsilon \ .$$

In other words, the maximum likelihood estimator Fisher [1922] and Lehmann and Casella [1998] is optimal in this context, when one seeks to minimize the variability of the estimator.

Important work in the 70's, 80's and 90's extended some of these results to various situations where p was allowed to grow with n but $p = o(n)$ - see for instance Portnoy [1984, 1985, 1986, 1987], Mammen [1989], and Yohai [1974]. See also see Dümbgen,

Samworth, and Schuhmacher [2011] for more recent results in the classical dimensional framework and very interesting connections with the field of shape restricted estimation Groeneboom and Jongbloed [2014].

2.2 Modern high-dimensional results: large n, large p. It is natural to ask similar questions to those raised above in the modern context of large n, large p asymptotics, as in fact was done as far back as Huber [1973].

Before we proceed, let us say that much effort was devoted in the last two decades in statistics and statistical learning to understanding the properties of the estimators of the form

$$\widehat{\beta}_\lambda = \text{argmin}_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i - X_i'\beta) + \lambda P(\beta) \ ,$$

where P is a penalty function, for instance $P(\beta) = \|\beta\|_2^2$ or $P(\beta) = \|\beta\|_1$. However, works in this line of investigation put rather stringent conditions on β, such as dramatic sparsity (i.e. only a fixed number of coefficients of β_0 are allowed to not be equal to zero as $p \to \infty$), which essentially turns these problems into rather classical ones; their analysis depend essentially on well-understood methods, which nonetheless had to be adapted to these specific problems. See Bühlmann and van de Geer [2011] for a book-length survey of this line of work. Let us also note that in truly large case applications Chapelle, Manavoglu, and Rosales [2014], practitioners are note willing to make these stringent assumptions.

2.2.1 Behavior of the estimator. By contrast we make no such restrictions on β_0. We focus on the unpenalized case for ease of presentation. To get a sense of results in this context, let us recall the system obtained in N. El Karoui, D. Bean, Bickel, C. Lim, and B. Yu [2013]. Let us consider $\widehat{\beta}$ as in Equation (3). Suppose $p/n \to \kappa \in (0,1)$. For simplicity assume that are $X_i \overset{iid}{\sim} (0, \text{Id}_p)$, with i.i.d entries and certain moment conditions - see Karoui [2013] and El Karoui [2018] for technical details - we have

Theorem 2.1. *Under regularity conditions on $\{\epsilon_i\}$ and ρ (convex), $\|\widehat{\beta}_\rho - \beta_0\|_2$ is asymptotically deterministic. Call $r_\rho(\kappa)$ its limit and let $\hat{z}_\epsilon$ be a random variable with $\hat{z}_\epsilon = \epsilon + r_\rho(\kappa)Z$, where $Z \sim \mathfrak{N}(0,1)$, independent of ϵ, where ϵ has the same distribution as ϵ_i's. For c deterministic, we have*

$$\begin{cases} \mathbf{E}\left([prox(c\rho)]'(\hat{z}_\epsilon)\right) &= 1-\kappa \ , \\ \kappa r_\rho^2(\kappa) &= \mathbf{E}\left([\hat{z}_\epsilon - prox(c\rho)(\hat{z}_\epsilon)]^2\right) \ . \end{cases} \tag{4}$$

where by definition (see Moreau [1965]) for a convex function $f\ \mathbb{R} \mapsto \mathbb{R}$,

$$prox(f)(x) = \text{argmin}_{y \in \mathbb{R}} \left(f(y) + \frac{1}{2}(x-y)^2 \right) .$$

We note that the system generalizes easily to much more general setups (involving penalization) - see El Karoui [2018]. In particular, the system (4) is quite sensitive to the Euclidean geometry of the predictors, X_i's. For instance, if we had $X_i = \lambda_i Z_i$ where $Z_i \sim \mathfrak{N}(0, \text{Id}_p)$ and λ_i is an independent scalar "well-behaved" random variable with $\mathbf{E}\left(\lambda_i^2\right) = 1$, a similar type of result would hold, but it would depend on the distribution of λ_i and not only its second moment. In particular, $r_\rho(\kappa)$ would change, despite the fact that in both models, $\text{cov}\,(X_i) = \text{Id}_p$. As such, one cannot hope for strong universality results in this context. See also Donoho and Montanari [2016] for another point of view on this system.

We also note that the previous result can be generalized to the case where $\text{cov}\,(X_i) = \Sigma$ by simple and classical rotational invariance arguments - see Eaton [2007] and N. El Karoui, D. Bean, Bickel, C. Lim, and B. Yu [2013]. In the case where X_i's are Gaussian, N. El Karoui, D. Bean, Bickel, C. Lim, and B. Yu [2013] also uses those to characterize the distribution of $\widehat{\beta}_\rho - \beta_0$ in a non-asymptotic fashion.

Finally, the behavior of the residuals $e_i = Y_i - X_i'\widehat{\beta}_\rho$ is very different in high-dimension from what it is in low-dimension; see N. El Karoui, D. Bean, Bickel, C. Lim, and B. Yu [ibid.] and follow-up papers for characterization. In particular, the residuals are not close in our framework to the "true errors", ϵ_i's, which is problematic as in many practical statistical methods - based on low-dimensional intuition - the residuals are used as proxies for those "true errors".

2.2.2 New loss functions. In light of the system (4), it is natural to ask which function ρ minimizes $r_\rho(\kappa)$, which is one measure of the inaccuracy of $\widehat{\beta}_\rho$ as an estimator of β_0. This question was investigated in Bean, Bickel, Karoui, and Yu [2013]. The following result is shown there.

Theorem 2.2. *Suppose that* ϵ *has a log-concave density, i.e.* $-\log f_\epsilon$ *is convex. Suppose* $r_\rho(\kappa)$ *is the solution of* (4)*. Then if* $p_2(x) = x^2/2$*, the optimal loss function that minimizes* $r_\rho(\kappa)$ *over convex* ρ *functions is*

$$\rho_{opt} = (p_2 + r_{opt}^2 \log \phi_{r_{opt}} \star f_\epsilon)^* - p_2 .$$

where $r_{opt} = \min\{r : r^2 I_\epsilon(r) = p/n\}$.

In the theorem above, ϕ_r is the density of a mean 0 Gaussian random variable with variance r^2, $\star$ denotes convolution, $I_\epsilon(r)$ is the Fisher information Lehmann and Casella [1998] of $\phi_r \star f_\epsilon$ and $g^*(x) = \sup_{y\in\mathbb{R}}(xy - g(y))$, is the Fenchel-Legendre dual of g Hiriart-Urruty and Lemaréchal [2001].

The function ρ_{opt} can be shown to be convex under the hypotheses of the theorem. It depends of course on p/n, our proxy for the statistical difficulty of the problem. In other words, this function quantifies the intuitively compelling notion that the loss function we use in these M-estimation problems should be adapted to the statistical hardness of the problem. Interestingly, the function in question is not the maximum likelihood estimator, which is the usual method that is used to determine on statistical grounds the loss function that should be used for a particular problem. We present a (limited) numerical comparison of these new loss functions and the maximum likelihood estimator in Figure 2.

Finally, it should be noted that the impact of choosing a better loss function is not limited to reducing uncertainty about the estimator. It also improves the quality of predictions, as the standard measure of expected prediction error Hastie, R. Tibshirani, and Friedman [2009] is closely tied to the size of $\mathbf{E}\left(\|\widehat{\beta}_\rho - \beta_0\|_2^2\right)$ in the models we consider.

3 Bootstrap and resampling questions

Modern statistics is increasingly computational and as such many methods have been devised to try to assess sampling variability of estimators through the use of simulations and without relying on asymptotic analyses. In other words, there are numerical ways to try to get at results such as those obtained in Theorems 1.2 and 2.1 for instance.

The most prominent of such methods is the bootstrap, proposed by Efron in the breakthrough paper Efron [1979]. Numerous variants of the bootstrap have appeared since then, and the bootstrap created an entire field of research, both theoretical and applied. See for instance Bickel and Freedman [1981], Efron [1982], Davison and Hinkley [1997], Hall [1992], and Efron and R. J. Tibshirani [1993] for classic references.

It is therefore natural to ask how the bootstrap performs in the modern high-dimensional context. Before we present some results in this direction, let us give a very brief introduction to the non-parametric bootstrap.

3.1 Non-parametric bootstrap and plug-in principle. As a so-called resampling method, the bootstrap seeks to re-use the data to assess for instance the variability of an estimator. Concretely, suppose we have data $\{X_i\}_{i=1}^n \in \mathbb{R}^p$, assumed to be i.i.d. and we are interested in the fluctuation behavior of a statistic/function of the data $\widehat{\theta} = \theta(X_1, \ldots, X_n)$. For instance, $\widehat{\theta}$ could be the sample mean of the X_i's or the largest eigenvalue of the sample covariance matrix of the X_i's.

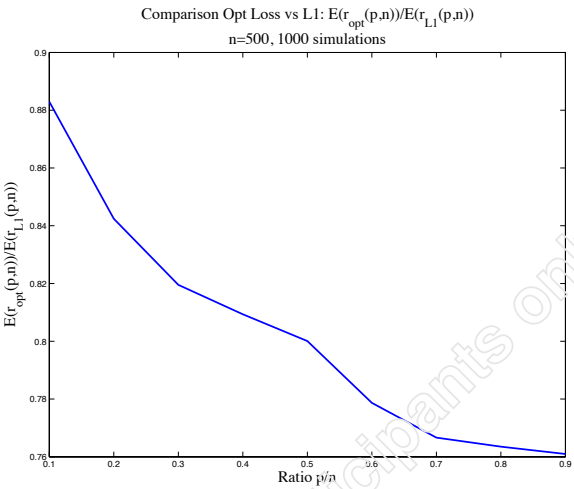

Figure 2: Numerical comparison of dimension-adaptive optimal loss and maximum likelihood loss: case where $f_\epsilon(x) = e^{-|x|}/2$, a.k.a. double exponential errors. We plot the ratio $\mathbf{E}\left(\|\widehat{\beta}_{opt} - \beta_0\|_2\right) / \mathbf{E}\left(\|\widehat{\beta}_{l_1} - \beta_0\|_2\right)$ as a function of p/n. The ratio is always less than 1 : ρ_{opt}, which varies with p/n and is used to compute $\widehat{\beta}_{opt}$, beats ℓ_1 loss, i.e. $\rho(x) = |x|$, the "optimal loss" in this context according to maximum likelihood theory. The curve is obtained by estimating the expectation through averaging over 1,000 independent simulations.

The non-parametric bootstrap uses the following algorithm :

- For $b = 1, \ldots, B$, repeat:
- Sample n times with replacement from $\{X_i\}_{i=1}^n$, to get bootstrapped dataset $D_b = \{X^*_{1,b}, \ldots, X^*_{n,b}\}$.
- Compute $\widehat{\theta}(X^*)_{n,b} = \theta(X^*_{1,b}, \ldots, X^*_{n,b})$.

Then the empirical distribution of $\{\widehat{\theta}(X^*)_{n,b}\}_{b=1}^B$ is used to assess the sampling variability of the original statistic $\widehat{\theta} = \theta(X_1, \ldots, X_n)$ for instance by computing the bootstrap estimate of variance (i.e. the empirical variance of $\{\widehat{\theta}(X^*)_{n,b}\}_{b=1}^B$ if the statistic is one-dimensional), or more sophisticated functions of the empirical distribution.

This is the so-called plug-in principle: one considers that the bootstrap data-generating process mimics the "true" (i.e. sampling from the population) data-generating process and

proceeds with bootstrap data as one would do with data sampled from the population. As such the bootstrap offers the promise of uncertainty assessment for arbitrarily complicated statistics without much need for mathematical understanding.

One natural question is of course to know when the bootstrap works (and what it means for the bootstrap to work). The first such results appeared in the pioneering Bickel and Freedman [1981]; nowadays, a common way to look at this problem is by looking at θ as a function over probability distributions - $\widehat{\theta}$ being θ applied to the empirical distribution of the data - and requiring θ to be sufficiently smooth in an appropriate sense van der Vaart [1998].

3.2 Bootstrapping regression M-estimates. Because of the lack of closed formulae to characterize the behavior of estimators such as $\widehat{\beta}_\rho$ defined in Equation (3), the bootstrap became early on an appealing tool to use for this task Shorack [1982] and questions related to the ones we raise in the high-dimensional setting were addressed in setting where $p/n \to 0$ in Wu [1986] and Mammen [1989, 1993].

In Karoui and Purdom [n.d.], various results concerning the bootstrap in high-dimension regression are presented. Bootstrapping as described above the observations $\{(Y_i, X_i)\}_{i=1}^n$ is called the pairs bootstrap in this setting. Elementary algebra shows that the pairs bootstrap amounts to fitting weighted regression models, i.e for bootstrap weights $\{w_i^*\}$,

$$\widehat{\beta}^*_{\rho,w} = \text{argmin}_{\beta\in\mathbb{R}^p} \sum_{i=1}^n w_i^* \rho(Y_i - X_i'\beta) \ .$$

For instance, it is shown that (for precise technical details see Karoui and Purdom [ibid.]):

Theorem 3.1. *Suppose weights $(w_i)_{i=1}^n$ are i.i.d., $\mathbf{E}(w_i) = 1$, have sufficiently many moments and are bounded away from 0. Let $X_i \overset{iid}{\backsim} \mathcal{N}(0, \mathrm{Id}_p)$ and let v be a (sequence of) deterministic unit vector.*

Suppose $\widehat{\beta}$ is obtained by solving a least-squares problem, i.e $\rho(x) = x^2/2$ and that the linear model holds. Let us call $\mathrm{var}(\epsilon_i) = \sigma_\epsilon^2$ and corresponding bootstrapped estimates $\widehat{\beta}^*_w$.

If $\lim p/n = \kappa < 1$ then asymptotically as $n \to \infty$

$$p\mathbf{E}\left(\mathrm{var}\left(v'\widehat{\beta}^*_w\right)\right) \to \sigma_\epsilon^2 \left[\kappa \frac{1}{1-\kappa-\mathbf{E}\left(\frac{1}{(1+cw_i)^2}\right)} - \frac{1}{1-\kappa}\right] ,$$

where c is the unique solution of

$$\mathbf{E}\left(\frac{1}{1+cw_i}\right) = 1-\kappa\ .$$

We note that in the previous context, it is not complicated to show that

$$p\mathrm{var}\left(v'\widehat{\beta}\right) \to \sigma_\epsilon^2 \frac{\kappa}{1-\kappa}\ .$$

Therefore the type of bootstraps described above fails at the very simple task of estimating the variance $v'\widehat{\beta}$, even for least squares. Figure 3 on p. 2866 gives a graphical illustration of the problem, showing that the bootstrap overestimates the variance of our estimator.

Karoui and Purdom [ibid.] contains many other results concerning other types of bootstraps and other resampling techniques, such as the jackknife. In general, the results show that even when classical bootstrap theory would suggest that the bootstrap should work (i.e. the statistics of interest are sufficiently "smooth"), it does not work in high-dimension, even when the statistician has very minimal requirements about what it means to work. Problematically, various bootstraps can fail in many ways, yielding confidence intervals with either too much or not enough coverage for instance. See Karoui and Purdom [ibid.] for details and relations to relevant literature as well as Bickel and Freedman [1983] for an early take on closely related questions, with however different requirements concerning bootstrap performance and analysis of a different kind of bootstraps.

3.3 Bootstrap and eigenvalues. It is also natural to wonder whether the bootstrap would be able to "automatically discover" results such as Theorem 1.2 and adapt to phase transitions such as the one discovered in Baik, Ben Arous, and Péché [2005]. Analysis of the bootstrap for eigenvalues in low-dimension goes as far back as Beran and Srivastava [1985] and Eaton and Tyler [1991]. In Karoui and Purdom [2016], questions of that type are investigated in high-dimension through a mix of theory and simulations, for various statistics related to eigenvalues of random matrices. Many mathematical questions remain open; however the results are generally negative, in that typically bootstrap confidence intervals do not have the right coverage probabilities. The only positive results about the bootstrap in that context are situations where the population covariance Σ has very isolated eigenvalues, and the problem is hence effectively low-dimensional and therefore of limited mathematical interest.

As such the bootstrap appears as of this writing to be a genuinely perturbation analytic technique and hence to be poorly suited to the kind of problems discussed in this short review.

4 Conclusions

We have a presented a small overview of recent results in theoretical statistics focused on the high-dimensional case, where the number of measurements per observations grows with the number of observations.

Mathematical analysis in this setup reveals the breakdown of basic consistency results. Furthermore, classical optimality results (based essentially on the likelihood principle) do not hold, yielding results and methods that upended many practitioners' intuition.

Interestingly, the analyses summarized above led the way to the proposal of new loss functions outside of "standard" families and adapted to the statistical difficulty of the problem, as measured by p/n.

Finally, standard data-driven methods of uncertainty assessment such as the bootstrap seem to completely break down in this setup, where they are most needed by practitioners given the complexity of the problems.

As such the large n, large p setting is much more than just a technical hurdle for theoreticians but seems to call for a serious rethinking of tools used by statisticians, whether they be involved in theory, methodology or applications.

Much mathematically stimulating work remains to be done to be able to develop improved methods (both for estimation and uncertainty assessment) and improve our understanding of statistics in this still novel and challenging framework.

References

Dimitris Achlioptas and Frank McSherry (2007). "Fast computation of low-rank matrix approximations". *J. ACM* 54.2, Art. 9, 19. MR: 2295993 (cit. on p. 2850).

T. W. Anderson (1963). "Asymptotic theory for principal component analysis". *Ann. Math. Statist.* 34, pp. 122–148. MR: 0145620 (cit. on p. 2845).

– (1984). *An introduction to multivariate statistical analysis*. Second. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, pp. xviii+675. MR: 771294 (cit. on pp. 2846, 2847).

Jinho Baik, Gérard Ben Arous, and Sandrine Péché (2005). "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices". *Ann. Probab.* 33.5, pp. 1643–1697. MR: 2165575 (cit. on pp. 2849, 2857).

D. Bean, P. J. Bickel, N. El Karoui, and B. Yu (2013). "Optimal m-estimation in high-dimensional regression". *Proceedings of the National Academy of Sciences* 110 (36), pp. 14563–14568 (cit. on p. 2853).

Rudolf Beran and Muni S. Srivastava (1985). "Bootstrap tests and confidence regions for functions of a covariance matrix". *Ann. Statist.* 13.1, pp. 95–115. MR: 773155 (cit. on p. 2857).

P. J. Bickel and D. A. Freedman (1983). "Bootstrapping regression models with many parameters". In: *A Festschrift for Erich L. Lehmann*. Wadsworth Statist./Probab. Ser. Wadsworth, Belmont, Calif., pp. 28–48. MR: 689736 (cit. on p. 2857).

Peter J. Bickel and David A. Freedman (1981). "Some asymptotic theory for the bootstrap". *Ann. Statist.* 9.6, pp. 1196–1217. MR: 630103 (cit. on pp. 2854, 2856).

Peter J. Bickel and Elizaveta Levina (2008). "Covariance regularization by thresholding". *Ann. Statist.* 36.6, pp. 2577–2604. MR: 2485008 (cit. on p. 2848).

Alex Bloemendal, Antti Knowles, Horng-Tzer Yau, and Jun Yin (2016). "On the principal components of sample covariance matrices". *Probab. Theory Related Fields* 164.1-2, pp. 459–552. MR: 3449395 (cit. on p. 2849).

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart (2013). *Concentration inequalities*. A nonasymptotic theory of independence, With a foreword by Michel Ledoux. Oxford University Press, Oxford, pp. x+481. MR: 3185193 (cit. on p. 2849).

Peter Bühlmann and Sara van de Geer (2011). *Statistics for high-dimensional data*. Springer Series in Statistics. Methods, theory and applications. Springer, Heidelberg, pp. xviii+556. MR: 2807761 (cit. on p. 2852).

Joël Bun, Jean-Philippe Bouchaud, and Marc Potters (2017). "Cleaning large correlation matrices: tools from random matrix theory". *Phys. Rep.* 666, pp. 1–109. MR: 3590056 (cit. on p. 2848).

R. Cattell (1966). "The scree test for the number of factors". *Multivariate Behav. Res.* 1, pp. 245–276 (cit. on p. 2847).

O. Chapelle, E. Manavoglu, and R. Rosales (Dec. 2014). "Simple and scalable response prediction for display advertising". *ACM Trans. Intell. Syst. Technol.* 5 (4), 61:1–61:34 (cit. on pp. 2851, 2852).

Sanjay Chaudhuri, Mathias Drton, and Thomas S. Richardson (2007). "Estimation of a covariance matrix with zeros". *Biometrika* 94.1, pp. 199–216. MR: 2307904 (cit. on p. 2848).

A. G. Constantine (1963). "Some non-central distribution problems in multivariate analysis". *Ann. Math. Statist.* 34, pp. 1270–1285. MR: 0181056 (cit. on p. 2849).

Criteo (n.d.). Criteo public datasets (cit. on p. 2846).

A. C. Davison and D. V. Hinkley (1997). *Bootstrap methods and their application*. Vol. 1. Cambridge Series in Statistical and Probabilistic Mathematics. With 1 IBM-PC floppy disk (3.5 inch; HD). Cambridge University Press, Cambridge, pp. x+582. MR: 1478673 (cit. on p. 2854).

P. A. Deift (1999). *Orthogonal polynomials and random matrices: a Riemann-Hilbert approach*. Vol. 3. Courant Lecture Notes in Mathematics. New York University, Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence, RI, pp. viii+273. MR: 1677884 (cit. on p. 2849).

Yen Do and Van Vu (2013). "The spectrum of random kernel matrices: universality results for rough and varying kernels". *Random Matrices Theory Appl.* 2.3, pp. 1350005, 29. MR: 3109422 (cit. on p. 2850).

David Donoho and Andrea Montanari (2016). "High dimensional robust M-estimation: asymptotic variance via approximate message passing". *Probab. Theory Related Fields* 166.3-4, pp. 935–969. MR: 3568043 (cit. on p. 2853).

Petros Drineas, Ravi Kannan, and Michael W. Mahoney (2006). "Fast Monte Carlo algorithms for matrices. II. Computing a low-rank approximation to a matrix". *SIAM J. Comput.* 36.1, pp. 158–183. MR: 2231644 (cit. on p. 2850).

Lutz Dümbgen, Richard Samworth, and Dominic Schuhmacher (2011). "Approximation by log-concave distributions, with applications to regression". *Ann. Statist.* 39.2, pp. 702–730. MR: 2816336 (cit. on p. 2851).

Morris L. Eaton (2007). *Multivariate statistics*. Vol. 53. Institute of Mathematical Statistics Lecture Notes—Monograph Series. A vector space approach, Reprint of the 1983 original [MR0716321]. Institute of Mathematical Statistics, Beachwood, OH, pp. viii+512. MR: 2431769 (cit. on p. 2853).

Morris L. Eaton and David E. Tyler (1991). "On Wielandt's inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix". *Ann. Statist.* 19.1, pp. 260–271. MR: 1091849 (cit. on p. 2857).

B. Efron (1979). "Bootstrap methods: another look at the jackknife". *Ann. Statist.* 7.1, pp. 1–26. MR: 515681 (cit. on p. 2854).

Bradley Efron (1982). *The jackknife, the bootstrap and other resampling plans*. Vol. 38. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa., pp. vi+92. MR: 659849 (cit. on p. 2854).

Bradley Efron and Robert J. Tibshirani (1993). *An introduction to the bootstrap*. Vol. 57. Monographs on Statistics and Applied Probability. Chapman and Hall, New York, pp. xvi+436. MR: 1270903 (cit. on p. 2854).

Noureddine El Karoui (2007). "Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices". *Ann. Probab.* 35.2, pp. 663–714. MR: 2308592 (cit. on p. 2849).

– (2008). "Operator norm consistent estimation of large-dimensional sparse covariance matrices". *Ann. Statist.* 36.6, pp. 2717–2756. MR: 2485011 (cit. on p. 2848).

– (2009). "Concentration of measure and spectra of random matrices: applications to correlation matrices, elliptical distributions and beyond". *Ann. Appl. Probab.* 19.6, pp. 2362–2405. MR: 2588248 (cit. on p. 2847).

– (2010). "The spectrum of kernel random matrices". *Ann. Statist.* 38.1, pp. 1–50. MR: 2589315 (cit. on p. 2850).

– (2018). "On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators". *Probab. Theory Related Fields* 170.1-2, pp. 95–175. MR: 3748322 (cit. on pp. 2852, 2853).

László Erdős and Horng-Tzer Yau (2012). "Universality of local spectral statistics of random matrices". *Bull. Amer. Math. Soc. (N.S.)* 49.3, pp. 377–414. MR: 2917064 (cit. on p. 2849).

R. A. Fisher (1922). "On the mathematical foundations of theoretical statistics". *Philosophical Transactions of the Royal Society, A* 222, pp. 309–368 (cit. on p. 2851).

P. J. Forrester (1993). "The spectrum edge of random matrix ensembles". *Nuclear Phys. B* 402.3, pp. 709–728. MR: 1236195 (cit. on p. 2849).

Friedrich Götze and Alexander Tikhomirov (2004). "Rate of convergence in probability to the Marchenko-Pastur law". *Bernoulli* 10.3, pp. 503–548. MR: 2061442 (cit. on p. 2847).

Piet Groeneboom and Geurt Jongbloed (2014). *Nonparametric estimation under shape constraints*. Vol. 38. Cambridge Series in Statistical and Probabilistic Mathematics. Estimators, algorithms and asymptotics. Cambridge University Press, New York, pp. xi+416. MR: 3445293 (cit. on p. 2852).

Peter Hall (1992). *The bootstrap and Edgeworth expansion*. Springer Series in Statistics. Springer-Verlag, New York, pp. xiv+352. MR: 1145237 (cit. on p. 2854).

Trevor Hastie, Robert Tibshirani, and Jerome Friedman (2009). *The elements of statistical learning*. Second. Springer Series in Statistics. Data mining, inference, and prediction. Springer, New York, pp. xxii+745. MR: 2722294 (cit. on pp. 2847, 2854).

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal (2001). *Fundamentals of convex analysis*. Grundlehren Text Editions. Abridged version of ıt Convex analysis and minimization algorithms. I [Springer, Berlin, 1993; MR1261420 (95m:90001)] and ıt II [ibid.; MR1295240 (95m:90002)]. Springer-Verlag, Berlin, pp. x+259. MR: 1865628 (cit. on p. 2854).

H. Hotelling (1933). "Analysis of a complex of statistical variables into principal components". *Journal of Educational Psychology* 24, pp. 417–441 (cit. on p. 2845).

Peter J. Huber (1972). "The 1972 Wald lecture. Robust statistics: A review". *Ann. Math. Statist.* 43, pp. 1041–1067. MR: 0314180 (cit. on p. 2846).

– (1973). "Robust regression: asymptotics, conjectures and Monte Carlo". *Ann. Statist.* 1, pp. 799–821. MR: 0356373 (cit. on pp. 2851, 2852).

– (1981). *Robust statistics*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, pp. ix+308. MR: 606374 (cit. on p. 2851).

Alan T. James (1964). "Distributions of matrix variates and latent roots derived from normal samples". *Ann. Math. Statist.* 35, pp. 475–501. MR: 0181057 (cit. on p. 2849).

Kurt Johansson (2000). "Shape fluctuations and random matrices". *Comm. Math. Phys.* 209.2, pp. 437–476. MR: 1737991 (cit. on p. 2849).

Iain M. Johnstone (2001). "On the distribution of the largest eigenvalue in principal components analysis". *Ann. Statist.* 29.2, pp. 295–327. MR: 1863961 (cit. on pp. 2846, 2849).

– (2007). "High dimensional statistical inference and random matrices". In: *International Congress of Mathematicians. Vol. I*. Eur. Math. Soc., Zürich, pp. 307–333. MR: 2334195 (cit. on p. 2846).

I. T. Jolliffe (2002). *Principal component analysis*. Second. Springer Series in Statistics. Springer-Verlag, New York, pp. xxx+487. MR: 2036084 (cit. on p. 2845).

N. El Karoui and E. Purdom (n.d.). "Can we trust the bootstrap in high-dimension?" (). Technical Report 824, UC Berkeley, Department of Statistics, February 2015 (cit. on pp. 2856, 2857).

Noureddine El Karoui (Sept. 2003). "On the largest eigenvalue of Wishart matrices with identity covariance when n, p and p/n tend to infinity". arXiv: math/0309355 (cit. on p. 2849).

– (Nov. 2013). "Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators : rigorous results". arXiv: 1311.2445 (cit. on p. 2852).

Noureddine El Karoui and Holger Koesters (May 2011). "Geometric sensitivity of random matrix results: consequences for shrinkage estimators of covariance and related statistical methods". arXiv: 1105.1404 (cit. on pp. 2849, 2850).

Noureddine El Karoui and Elizabeth Purdom (Aug. 2016). "The bootstrap, covariance matrices and PCA in moderate and high-dimensions". arXiv: 1608.00948 (cit. on p. 2857).

Vladimir Koltchinskii and Evarist Giné (2000). "Random matrix approximation of spectra of integral operators". *Bernoulli* 6.1, pp. 113–167. MR: 1781185 (cit. on p. 2850).

L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters (1999). "Noise dressing of financial correlation matrices". *Phys. Rev. Lett.* 83 (7), pp. 1467–1470 (cit. on pp. 2845, 2846).

Michel Ledoux (2001). *The concentration of measure phenomenon*. Vol. 89. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, pp. x+181. MR: 1849347 (cit. on p. 2849).

Ji Oon Lee and Kevin Schnelli (2016). "Tracy–Widom distribution for the largest eigenvalue of real sample covariance matrices with general population". *Ann. Appl. Probab.* 26.6, pp. 3786–3839. arXiv: 1409.4979. MR: 3582818 (cit. on p. 2849).

E. L. Lehmann and George Casella (1998). *Theory of point estimation*. Second. Springer Texts in Statistics. Springer-Verlag, New York, pp. xxvi+589. MR: 1639875 (cit. on pp. 2851, 2854).

Enno Mammen (1989). "Asymptotics with increasing dimension for robust regression with applications to the bootstrap". *Ann. Statist.* 17.1, pp. 382–400. MR: 981457 (cit. on pp. 2846, 2851, 2856).

– (1993). "Bootstrap and wild bootstrap for high-dimensional linear models". *Ann. Statist.* 21.1, pp. 255–285. MR: 1212176 (cit. on p. 2856).

V. A. Marčenko and L. A. Pastur (1967). "Distribution of eigenvalues in certain sets of random matrices". *Mat. Sb. (N.S.)* 72 (114), pp. 507–536. MR: 0208649 (cit. on p. 2847).

P. McCullagh and J. A. Nelder (1989). *Generalized linear models*. Monographs on Statistics and Applied Probability. Second edition [of MR0727836]. Chapman & Hall, London, pp. xix+511. MR: 3223057 (cit. on p. 2851).

Madan Lal Mehta (1991). *Random matrices*. Second. Academic Press, Inc., Boston, MA, pp. xviii+562. MR: 1083764 (cit. on p. 2849).

Jean-Jacques Moreau (1965). "Proximité et dualité dans un espace hilbertien". *Bull. Soc. Math. France* 93, pp. 273–299. MR: 0201952 (cit. on p. 2853).

Robb J. Muirhead (1982). *Aspects of multivariate statistical theory*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, pp. xix+673. MR: 652932 (cit. on p. 2849).

N. El Karoui, D. Bean, P. J. Bickel, C. Lim, and B. Yu (2013). *On robust regression with high-dimensional predictors* (cit. on pp. 2852, 2853).

A. Pajor and L. Pastur (2009). "On the limiting empirical measure of eigenvalues of the sum of rank one matrices with log-concave distribution". *Studia Math.* 195.1, pp. 11–29. MR: 2539559 (cit. on p. 2847).

K. Pearson (1901). "On lines and planes of closest fit to systems of points in space". *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2, pp. 559–572 (cit. on p. 2845).

Stephen Portnoy (1984). "Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency". *Ann. Statist.* 12.4, pp. 1298–1309. MR: 760690 (cit. on pp. 2846, 2851).

– (1985). "Asymptotic behavior of M estimators of p regression parameters when p^2/n is large. II. Normal approximation". *Ann. Statist.* 13.4, pp. 1403–1417. MR: 811499 (cit. on p. 2851).

– (1986). "Asymptotic behavior of the empiric distribution of M-estimated residuals from a regression model with many parameters". *Ann. Statist.* 14.3, pp. 1152–1170. MR: 856812 (cit. on p. 2851).

– (1987). "A central limit theorem applicable to robust regression estimators". *J. Multivariate Anal.* 22.1, pp. 24–50. MR: 890880 (cit. on p. 2851).

S. Ramaswamy et al. (2001). "Multiclass cancer diagnosis using tumor gene expression signatures". 98, pp. 15149–15154 (cit. on p. 2846).

Daniel Arthur Relles (1968). *Robust Regression by Modified Least Squares*. Thesis (Ph.D.)–Yale University. ProQuest LLC, Ann Arbor, MI, p. 135. MR: 2617863 (cit. on p. 2851).

B. Schölkopf and A. J. Smola (2002). *Learning with kernels*. Cambridge, MA: The MIT Press (cit. on p. 2850).

Galen R. Shorack (1982). "Bootstrapping robust regression". *Comm. Statist. A—Theory Methods* 11.9, pp. 961–972. MR: 655465 (cit. on p. 2856).

Jack W. Silverstein (1995). "Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices". *J. Multivariate Anal.* 55.2, pp. 331–339. MR: 1370408 (cit. on p. 2847).

Alexander Soshnikov (2002). "A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices". *J. Statist. Phys.* 108.5-6. Dedicated to David Ruelle and Yasha Sinai on the occasion of their 65th birthdays, pp. 1033–1056. MR: 1933444 (cit. on p. 2849).

Terence Tao and Van Vu (2012). "Random covariance matrices: universality of local statistics of eigenvalues". *Ann. Probab.* 40.3, pp. 1285–1315. MR: 2962092 (cit. on p. 2849).

Craig A. Tracy and Harold Widom (1994a). "Fredholm determinants, differential equations and matrix models". *Comm. Math. Phys.* 163.1, pp. 33–72. MR: 1277933 (cit. on p. 2849).

– (1994b). "Level-spacing distributions and the Airy kernel". *Comm. Math. Phys.* 159.1, pp. 151–174. MR: 1257246 (cit. on p. 2849).

– (1996). "On orthogonal and symplectic matrix ensembles". *Comm. Math. Phys.* 177.3, pp. 727–754. MR: 1385083 (cit. on p. 2849).

Joel A. Tropp (2012). "User-friendly tail bounds for sums of random matrices". *Found. Comput. Math.* 12.4, pp. 389–434. MR: 2946459 (cit. on p. 2850).

A. W. van der Vaart (1998). *Asymptotic statistics*. Vol. 3. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, pp. xvi+443. MR: 1652247 (cit. on pp. 2846, 2856).

Kenneth W. Wachter (1978). "The strong limits of random matrix spectra for sample matrices of independent elements". *Ann. Probability* 6.1, pp. 1–18. MR: 0467894 (cit. on p. 2847).

Grace Wahba (1990). *Spline models for observational data*. Vol. 59. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, pp. xii+169. MR: 1045442 (cit. on p. 2850).

J. Wishart (1928). "The generalised product moment distribution in samples from a normal multivariate population". *Biometrika* 20 (A), pp. 32–52 (cit. on p. 2845).

Simon N. Wood, Yannig Goude, and Simon Shaw (2015). "Generalized additive models for large data sets". *J. R. Stat. Soc. Ser. C. Appl. Stat.* 64.1, pp. 139–155. MR: 3293922 (cit. on p. 2851).

C.-F. J. Wu (1986). "Jackknife, bootstrap and other resampling methods in regression analysis". *Ann. Statist.* 14.4. With discussion and a rejoinder by the author, pp. 1261–1350. MR: 868303 (cit. on p. 2856).

Víctor J. Yohai (1974). "Robust estimation in the linear model". *Ann. Statist.* 2. Collection of articles dedicated to Jerzy Neyman on his 80th birthday, pp. 562–567. MR: 0365875 (cit. on p. 2851).

Received 2017-12-07.

Noureddine El Karoui
Department of Statistics, University of California at Berkeley
and
Criteo Research
nkaroui@berkeley.edu
n.elkaroui@criteo.com

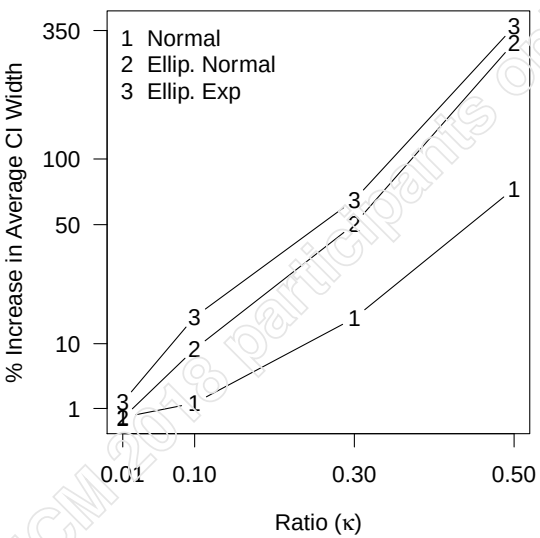

Figure 3. **Comparison of width of 95% confidence intervals of $e_1'\widehat{\beta}_\rho$ for L_2 loss:** $\rho(x) = x^2/2$; e_1 is the first canonical basis vector in $\mathbb{R}^p$; y-axis is the percent increase of the average confidence interval width based on simulation ($n = 500$), as compared to exact theoretical result for least squares; the percent increase is plotted against the ratio $\kappa = p/n$ (x-axis). Shown are three different choices in simulating the entries of the design matrix X: (1) Normal: $X_i \overset{iid}{\backsim} \mathfrak{N}(0, \mathrm{Id}_p)$ (2) Ellip. Normal: $X_i = \lambda_i Z_i$ with $\lambda_i \overset{iid}{\backsim} N(0,1)$ and independently $Z_i \overset{iid}{\backsim} \mathfrak{N}(0, \mathrm{Id}_p)$ and (3) Ellip. Exp : $X_i = \lambda_i Z_i$ with $\lambda_i \overset{iid}{\backsim} Exp(\sqrt{2})$. The errors ϵ_i's are i.i.d $\mathfrak{N}(0,1)$