On robust regression with high-dimensional predictors

Noureddine El Karoui[a,1], Derek Bean[a], Peter J. Bickel[a,1], Chinghway Lim[b], and Bin Yu[a]

[a]Department of Statistics, University of California, Berkeley, CA 94720; and [b]Department of Statistics and Applied Probability, Faculty of Science, National University of Singapore, 119077

Contributed by Peter J. Bickel, April 25, 2013 (sent for review March 1, 2012)

We study regression *M*-estimates in the setting where *p*, the number of covariates, and *n*, the number of observations, are both large, but $p \leq n$. We find an exact stochastic representation for the distribution of $\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho(Y_i - X_i'\beta)$ at fixed *p* and *n* under various assumptions on the objective function ρ and our statistical model. A scalar random variable whose deterministic limit $r_\rho(\kappa)$ can be studied when $p/n \to \kappa > 0$ plays a central role in this representation. We discover a nonlinear system of two deterministic equations that characterizes $r_\rho(\kappa)$. Interestingly, the system shows that $r_\rho(\kappa)$ depends on ρ through proximal mappings of ρ as well as various aspects of the statistical model underlying our study. Several surprising results emerge. In particular, we show that, when p/n is large enough, least squares becomes preferable to least absolute deviations for double-exponential errors.

prox function | high-dimensional statistics | concentration of measure

In the "classical" period up to the 1980s, research on regression models focused on situations for which the number of covariates p was much smaller than n, the sample size. Least-squares regression (LSE) was the main fitting tool used, but its sensitivity to outliers came to the fore with the work of Tukey, Huber, Hampel, and others starting in the 1950s.

Given the model $Y_i = X_i'\beta_0 + \epsilon_i$ and M-estimation methods described in the Abstract, it follows from the discussion in ref. 1 (p. 170, for instance) that, if the design matrix X (an $n \times p$ matrix whose ith row is X_i) is nonsingular, under various regularity conditions on X, ρ, $\psi = \rho'$ and the [independent identically distributed (i.i.d)] errors $\{\epsilon_i\}_{i=1}^n$, $\hat{\beta}$ is asymptotically normal with mean β_0 and covariance matrix $C(\rho, \epsilon)(X'X)^{-1}$. Here, $C(\rho, \epsilon) = \mathbf{E}(\psi^2(\epsilon))/[\mathbf{E}(\psi'(\epsilon))]^2$ and ε has the same distribution as ϵ_i's.

It follows that, for p fixed, the relative efficiency of M-estimates such as least absolute deviations (LAD), to LSE, does not depend on the design matrix. Thus, LAD has the same advantage over LSE for heavy-tailed distributions as the median has over the mean.

In recent years, there has been great focus on the case where p and n are commensurate and large. Greatest attention has been paid to the "sparse" case where the number of nonzero coefficients is much smaller than n or p. This has been achieved by adding an ℓ_1 type of penalty to the quadratic objective function of LSE, in the case of the Least Absolute Shrinkage and Selection Operator (LASSO). Unfortunately, these types of methods result in biased estimates of the coefficients, and statistical inference, as opposed to prediction, becomes problematic.

Huber (2) was the first to investigate the regime of large p ($p \to \infty$ with n). His results were followed up by Portnoy (3) under weaker conditions [see also Bloomfield (4)]. Huber showed that the behavior found for fixed p persisted in regions such as $p^2/n \to 0$ and $p^3/n \to 0$. That is, estimates of coefficients and contrasts were asymptotically Gaussian and relative efficiencies of methods did not depend on the design matrix. His arguments were, in part, heuristic but well confirmed by simulation. He also pointed out a surprising feature of the regime, $p/n \to \kappa > 0$ for LSE; fitted values were not asymptotically Gaussian. He was unable to deal with this regime otherwise (see the discussion on p. 802 of ref. 2).

In this paper, we analyze what happens in robust regression when $p/n \to \kappa < 1$. We proceed in the manner of Huber (2), who developed heuristics that were highly plausible and buttressed by simulations. (While the paper was under review, we have managed to obtain rigorous proofs for many of our assertions. They will be presented elsewhere because they are very long and technical.) We give several results for covariates that are Gaussian or derived from Gaussian but present grounds that the behavior holds much more generally—the key being concentration of certain quadratic forms involving the vectors of covariates. We also investigate the sensitivity of our results to the geometry of the design matrix. [Further results with different designs can be found in our work (5).]

We find that (*i*) estimates of coordinates and contrasts that have coefficients independent of the observed covariates continue to be unbiased and asymptotically normal; and (*ii*) as in the fixed p case, this happens at scale $n^{-1/2}$, at least when the minimal and maximal eigenvalues of the covariance of the predictors stay bounded away from 0 and ∞, respectively.*

These findings are obtained by (*i*) using leave-one-out perturbation arguments both for the data units and predictors; (*ii*) exhibiting a pair of master equations from which the asymptotic mean square prediction error and the correct expressions for asymptotic variances can be recovered; and (*iii*) showing that these two quantities depend in a nonlinear way on p/n, the error distribution, the design matrix, and the form of the objective function, ρ.

It is worth noting that our findings go against the likelihood principle that the "ideal" objective function is the negative log-density of the error distribution. For example, we show that, when p/n is large enough, it becomes preferable to use least squares rather than LAD for double-exponential errors. We illustrate this point in Fig. 1.

Main Results contains a detailed presentation of our results. We give some examples and supporting simulations in *Examples*. We present our derivation in the last section.

Main Results

We consider the following robust regression problem: let $\hat{\beta}$ be

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho\left(Y_i - X_i'\beta\right). \quad [1]$$

Here, $X_i \in \mathbb{R}^p$, $Y_i = X_i'\beta_0 + \epsilon_i$, where $\beta_0 \in \mathbb{R}^p$, ϵ_i is a random (scalar) error independent of the vector $X_i \in \mathbb{R}^p$. ρ is a convex function. We assume that the pairs $\{\epsilon_i\}_{i=1}^n$ and $\{X_i\}_{i=1}^n$ are independent. Furthermore, we assume that X_i's are independent. Our aim is to characterize the distribution of $\hat{\beta}$. As we will discuss later, our approach is not limited to this "standard" robust regression setting: we can, for instance, shed light on similar questions in weighted regression.

The following lemma is easily shown by using the rotational invariance of the Gaussian distribution (*SI Text*).

Author contributions: N.E.K., P.J.B., and B.Y. designed research; N.E.K., D.B., P.J.B., C.L., and B.Y. performed research; and N.E.K. and P.J.B. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence may be addressed. E-mail: nkaroui@stat.berkeley.edu or bickel@stat.berkeley.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1307842110/-/DCSupplemental.

*And the vector defining the contrasts has norm bounded away from 0 and ∞.

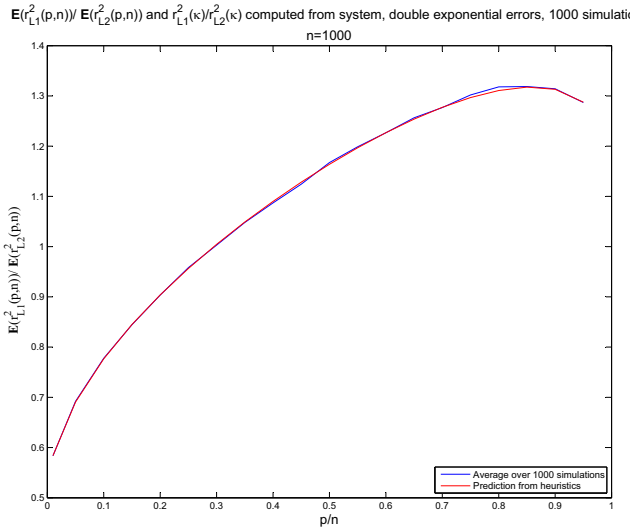

Fig. 1. Prediction [i.e., $r^2_{\ell_1}(\kappa)/r^2_{\ell_2}(\kappa)$] vs. realized value of $\mathbf{E}(r^2_{\ell_1}(p,n))/\mathbf{E}(r^2_{\ell_2}(p,n))$, double-exponential errors. Surprisingly, according to this measure, it becomes preferable to use ordinary least squares rather than l_1-regression when the errors are double exponential and κ is sufficiently large.

Lemma 1. *Suppose that $X_i = \lambda_i \mathcal{X}_i$, where $\mathcal{X}_i$'s are n i.i.d $\mathcal{N}(0,\Sigma)$, with Σ of rank p, and $\{\lambda_i\}_{i=1}^n$ are (nonzero) scalars, independent of $\{X_i\}_{i=1}^n$. Let $\hat{\beta}(\rho;\beta_0,\Sigma)$ be the solution† of* Eq. **1**. *When $n>p$, we have the stochastic representation*:

$$\hat{\beta}(\rho;\beta_0,\Sigma) \overset{\mathcal{L}}{=} \beta_0 + \left|\left|\hat{\beta}(\rho;\beta_0,\mathrm{Id}_p) - \beta_0\right|\right| \Sigma^{-1/2} u,$$

where u is uniform on the sphere of radius 1 in $\mathbb{R}^p$ and is independent of $||\hat{\beta}(\rho;0,\mathrm{Id}_p)-\beta_0||$. Furthermore, $\hat{\beta}(\rho;\beta_0,\mathrm{Id}_p)-\beta_0 \overset{\mathcal{L}}{=} \hat{\beta}(\rho;0,\mathrm{Id}_p)$.

In light of this result, it is clear that we just need to understand the distribution of $\hat{\beta}(\rho;0,\mathrm{Id}_p)$ to understand that of $\hat{\beta}(\rho;\beta_0,\Sigma)$.

Result 1. *Suppose that ρ is a nonlinear convex function. Let $r_\rho(p,n) = ||\hat{\beta}(\rho;0,\mathrm{Id}_p)||$. Assume that $X_i = \lambda_i \mathcal{X}_i$, where $\mathcal{X}_i$ are i.i.d $\mathcal{N}(0,\mathrm{Id}_p)$ and $\{\lambda_i\}_{i=1}^n$ are nonzero scalars, independent of $\{\mathcal{X}_i\}_{i=1}^n$. Assume also that $Y_i = \epsilon_i$ (i.e., $\beta_0 = 0$) and $\{\epsilon_i\}_{i=1}^n$ are independent of $\{X_i\}_{i=1}^n$.*

Then, under regularity conditions on $\{\epsilon_i\}_{i=1}^n$, $\{\lambda_i\}_{i=1}^n$ and ρ, $r_\rho(p,n)$ has a deterministic limit in probability as p and n tend to infinity while $p/n \to \kappa < 1$. We call this limit $r_\rho(\kappa)$.

Let us call $\hat{z}_\epsilon(i) = \epsilon_i + \lambda_i r_\rho(\kappa) Z_i$, where $Z_i \sim \mathcal{N}(0,1)$ are i.i.d and independent of $\{\epsilon\}_{i=1}^n$ and $\{\lambda_i\}_{i=1}^n$. We can determine $r_\rho(\kappa)$ through solving the following:

$$\begin{cases} \lim_{n\to\infty} \sum_{i=1}^n \dfrac{\mathbf{E}\left(\left[prox_{c\lambda_i^2}(\rho)\right]'\left(\hat{z}_\epsilon(i)\right)\right)}{n} = 1-\kappa, \\ \lim_{n\to\infty} \sum_{i=1}^n \dfrac{\mathbf{E}\left(\lambda_i^{-2}\left[\hat{z}_\epsilon(i) - prox_{c\lambda_i^2}(\rho)\left(\hat{z}_\epsilon(i)\right)\right]^2\right)}{n} = \kappa r_\rho^2(\kappa), \end{cases} \quad \textbf{[S1]}$$

where c is a positive deterministic constant to be determined from the above system. (The expectations above are taken with respect to the joint distribution of $\{\epsilon_i\}_{i=1}^n$, $\{\lambda_i\}_{i=1}^n$ and $\{Z_i\}_{i=1}^n$.)

The prox abbreviation refers to the proximal mapping, which is standard in convex optimization (see ref. 6). One of its definition is

$$\mathrm{prox}_c(\rho)(x) = \mathrm{argmin}_{y\in\mathbb{R}}\left(\rho(y) + \tfrac{(x-y)^2}{2c}\right).$$

Corollary 1 (Important Special Case). *When for all i, $\lambda_i^2 = 1$, and ϵ_i's are i.i.d, the same conclusions hold but the system characterizing $r_\rho(\kappa)$ becomes the following: if $\hat{z}_\epsilon = \epsilon + r_\rho(\kappa)Z$, where ε has the same distribution as ϵ_i and is independent of $Z \sim \mathcal{N}(0,1)$,*

$$\begin{cases} \mathbf{E}\left(\left[prox_c(\rho)\right]'\left(\hat{z}_\epsilon\right)\right) = 1-\kappa, \\ \mathbf{E}\left(\left[\hat{z}_\epsilon - prox_c(\rho)\left(\hat{z}_\epsilon\right)\right]^2\right) = \kappa r_\rho^2(\kappa). \end{cases} \quad \textbf{[S2]}$$

The asymptotic normality [in the finite dimensional (fidi) convergence sense] and unbiasedness of $\hat{\beta}$ is a consequence of Result 1 and Lemma 1. Note the complicated interaction of ρ, the distribution of ε, the distribution of the X_i's and κ in determining $r_\rho(\kappa)$. In the p fixed ($\kappa = 0$) case, $X_i \sim \mathcal{N}(0,\Sigma)$, the contribution of the design is just Σ^{-1}, which determines the correlation matrix of $\hat{\beta}$. In general, the correlation structure is the same but the variances also depend on the design.

As our arguments will show, we expect that the results concerning $r_\rho(\kappa)$ detailed in Result 1 will hold when the assumptions of normality on $\{\mathcal{X}_i\}_{i=1}^n$ are replaced by assumptions on concentration of quadratic forms in $\mathcal{X}_i$. Results on fidi convergence of $\hat{\beta}$ also appear likely to hold under these weakened restrictions.

The difference between the systems of equations characterizing $r_\rho(\kappa)$ in Result 1 and Corollary 1 highlights the importance of the geometry of the predictors, X_i, in the results. As a matter of fact, if we consider the case where $\Sigma = \mathrm{Id}_p$ and λ_i's are i.i.d with $\mathbf{E}(\lambda_i^2) = 1$, in both situations the X_i's have covariance $\mathrm{Id_p}$ and are nearly orthogonal to one another; however, in the setting of Result 1, $||X_i||/\sqrt{p}$ is close to $|\lambda_i|$ with high probability—hence variable with i—whereas in the setting of Corollary 1, the X_i's all have almost the same norm and hence are near a sphere. The importance of the geometry of the vectors of predictors in this situation is hence a generalization of similar phenomena that were highlighted in ref. 7 for instance. Further examples of a different nature detailed in ref. 5 (p. 27) illustrate the fundamental importance of our implicit geometric assumptions on the design matrix.

Our analysis also extends to the case where ρ is replaced by ρ_i, where for instance $\rho_i = w_i\rho$ (the weighted regression case) as long as $\{w_i\}_{i=1}^n$ is independent of $\{X_i\}_{i=1}^n$. (One simply needs to replace ρ by ρ_i in the system **S1** above and take expectation with respect to these quantities, too.) We refer the interested reader to ref. 5 (p. 26).

Examples

We illustrate the quality of our results on a few numerical examples, showing the importance of both the objective function and the distribution of the errors in the behavior of $r_\rho(\kappa)$. For simplicity, we focus only on the case where $\lambda_i^2 = 1$ for all i, i.e., the case of Gaussian predictors (an example with λ_i random is in *SI Text*). We also assume that ϵ_i's are i.i.d.

Least Squares. In this case, $\rho(x) = x^2/2$ and $\psi(x) = x$. Hence, $\mathrm{prox}_c(\rho) = \frac{1}{1+c}x$. Elementary computations then show that $c = \kappa/(1-\kappa)$. We also find that $r^2_{\ell_2}(\kappa) = \kappa/(1-\kappa)\sigma_\epsilon^2$, where σ_ϵ^2 is the variance of ε. Naturally, in the case of least squares, one can use results concerning Wishart distribution (8) as well as the explicit form of $\hat{\beta}$ to verify mathematically that this expression is correct. We also note that, in this case, the distribution of ε does not matter, only its variance.

Median Regression (LAD). This case, where ρ takes values $\rho(x) = |x|$, is substantially more interesting and reveals the importance of the interaction between objective function and error distribution. Clearly, we first have to compute the prox of the function ρ. It

†We write $\hat{\beta}(\rho;\beta_0,\Sigma)$ instead of $\hat{\beta}(\rho;\beta_0,\Sigma;\{\epsilon_i\}_{i=1}^n,\{\lambda_i\}_{i=1}^n)$ for simplicity.

is well known and not difficult to show that this prox is the soft-thresholding function. More formally, using the notation $x_+ = \max(x, 0)$, we have, for any $t > 0$, $\text{prox}_t(\rho)(y) = \text{sign}(y)(|y| - t)_+$. In this subsection, we use the notation r_{ℓ_1} instead of r_ρ.

Case of Gaussian errors. Let $s^2 = r^2_{\ell_1}(\kappa) + \sigma^2_\epsilon$. When ϵ_i's are i.i.d $\mathcal{N}(0, \sigma^2_\epsilon)$, $\hat{z}_\epsilon \sim \mathcal{N}(0, s^2)$. The first equation of our system **S2** therefore becomes $P(|Z| > c/s) = 1 - \kappa$, where $Z \sim \mathcal{N}(0, 1)$. Hence, $c/s = \Phi^{-1}((1+\kappa)/2)$, where Φ^{-1} is the quantile function for the standard normal distribution.

We now turn our attention to the second equation in the system **S2**. We have $[y - \text{prox}_t(\rho)(y)]^2 = y^2 1_{y \le t} + t^2 1_{y \ge t}$. Using the fact that $c/s = \Phi^{-1}((1+\kappa)/2)$, computations show that the second equation in the system **S2** becomes the following:

$$\kappa r^2_{\ell_1}(\kappa) = s^2 h(\kappa) + c^2(1-\kappa),$$
$$= s^2 \left[h(\kappa) + (1-\kappa) \left(\Phi^{-1} \left[\frac{1}{2}(1+\kappa) \right] \right)^2 \right],$$

where h is the function such that for $t \in [0, 1]$,

$$h(t) = t - \sqrt{\frac{2}{\pi}} \Phi^{-1}([1+t]/2) \exp\left(- \left[\Phi^{-1}([1+t]/2) \right]^2 / 2 \right).$$

Finally, calling ζ the function such that for $t \in [0, 1]$, if φ denotes the standard normal density,

$$\zeta(t) = 2\Phi^{-1}(t) \left(\varphi\left[\Phi^{-1}(t) \right] - \Phi^{-1}(t)(1-t) \right),$$

further manipulations show that we can solve for s as a function of κ and therefore for $r_{\ell_1}(\kappa)$. Our final expression is that, when the ϵ_i's are i.i.d $\mathcal{N}(0, \sigma^2_\epsilon)$,

$$\boxed{r^2_{\ell_1}(\kappa) = \frac{\kappa - \zeta([1+\kappa]/2)}{\zeta([1+\kappa]/2)} \sigma^2_\epsilon}.$$

Fig. 2 compares this expression for $r^2_{\ell_1}(\kappa)$ to $\mathbf{E}(r^2_{\ell_1}(p, n))$ obtained by simulations. The comparison is done by computing relative errors. (Fig. S1 compares the actual values, which are also of interest.)

Case of errors with symmetric distribution. We call $f_{r,\epsilon}$ the density of $\hat{z}_\epsilon$ and drop the dependence of $r_\rho(\kappa)$ on ρ and κ from our notations for simplicity. The first equation of system **S2** still reads $P(|\hat{z}_\epsilon| > c) = 1 - \kappa$. Let us call $F_{r,\epsilon}$ the cumulative distribution function (cdf) of $\hat{z}_\epsilon$ and $\overline{F}_{r,\epsilon} = 1 - F_{r,\epsilon}$. Let us denote by $\overline{F}^{-1}_{r,\epsilon}$ the functional inverse of $\overline{F}_{r,\epsilon}$.

Integration by parts, symmetry of $f_{r,\epsilon}$, as well as the above characterization of c finally lead to the implicit characterization of $r_{\ell_1}(\kappa)$ (denoted simply by r for short in the next equation):

$$(1-\kappa) r^2 = 4 \int_{\overline{F}^{-1}_{r,\epsilon}((1-\kappa)/2)}^{\infty} x \overline{F}_{r,\epsilon}(x) dx - \sigma^2_\epsilon. \qquad [2]$$

We note in passing that $r^2 + \sigma^2_\epsilon = 4 \int_0^\infty x \overline{F}_{r,\epsilon}(x) dx$; therefore, the previous equation can be rewritten $\kappa r^2 = 4 \int_0^{\overline{F}^{-1}_{r,\epsilon}((1-\kappa)/2)} x \overline{F}_{r,\epsilon}(x) dx$, a convenient equation to work with numerically when κ is small.

Case of double-exponential errors. We now present a comparison of simulation results to numerical solutions of system **S2** when the errors are double exponential. It should be noted that, in this case, the cdf $F_{r,\epsilon}$ takes values

$$F_{r,\epsilon}(t) = \Phi\left[\frac{t}{r} \right] + \frac{e^{r^2/2}}{2} \left(e^t \Phi\left[-\frac{t + r^2}{r} \right] - e^{-t} \Phi\left[\frac{t - r^2}{r} \right] \right).$$

It is also clear in this case that $\sigma^2_\epsilon = 2$. We used all this information to solve Eq. **2** for r, by doing a dichotomous search. Fig. 3 illustrates our results by showing the relative errors between $\mathbf{E}(r^2_{\ell_1}(p, n))$ (computed from simulations) and numerical solutions of system **S2** with appropriate parameters. (Fig. S2 compares the actual values, which are also of interest.)

Other Objective Functions. We have carried out similar computations and validations of results for other objective functions, including the Huber objective functions, the objective functions appearing in quantile regression, as well as ℓ_3 and $\ell_{1.5}$ objective functions—the latter two more for their analytical tractability than for their statistical interest. We refer the reader to ref. 5 for details.

Further Remarks. The characterizations of $r_\rho(\kappa)$ allows us to compare the performance of various regression methods for various error distributions. One mathematical and statistical consequence is that we can optimize over ρ to minimize $r_\rho(\kappa)$ when the distribution of the errors is given and log-concave and we are in the setup of Gaussian predictors. We have done this in the companion paper (9).

Quite independently, we can investigate the performance of say median regression vs. least squares for a range of values of κ. In the case of double-exponential errors, it is well known (see, e.g., ref. 1) that median regression is twice as efficient as least squares when κ is close to 0. As our simulations and computations illustrate, this is not the case when κ is not close to zero. Indeed, when $\kappa > 0.3$ or so, $r_{\ell_2}(\kappa) < r_{\ell_1}(\kappa)$ for double-exponential errors. This should serve as caution against using "natural" maximum-likelihood methods in high dimension since they turn out to be suboptimal even in apparently favorable situations.

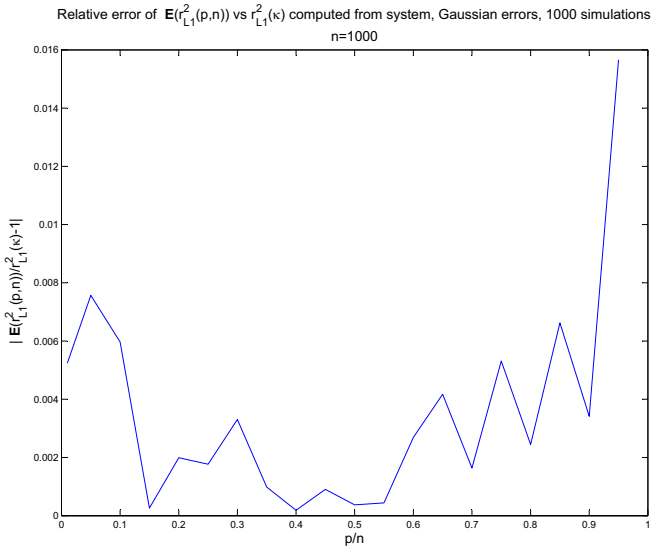

Fig. 2. Relative errors: $\left| \frac{\mathbf{E}(r^2_{\ell_1}(p,n))}{r^2_{\ell_1}(\kappa)} - 1 \right|$, Gaussian errors, 1,000 simulations.

Derivation

We now turn our attention to the derivation of the system of equations **S1** presented in *Result 1*.

Our approach hinges on a "double leave-one-out approach," the use of concentration properties of certain quadratic forms and the Sherman–Woodbury–Morrison formula of linear algebra.

We focus on the case $\beta_0 = 0$ and $\Sigma = \text{Id}_p$. *Lemma 1* guarantees that we can do so without loss of generality. Note that, in this case, $Y_i = \epsilon_i$. We call $\hat{\beta}(\rho; 0, \text{Id}_p)$ simply $\hat{\beta}$ from now on. We also assume that ρ has two derivatives.

We call the residuals $R_i = \epsilon_i - X_i' \hat{\beta}$ and use the notation $X_{(i)} = \{X_j\}_{j \ne i}$. Recall that $\psi = \rho'$. We note that under our assumptions $\hat{\beta}$ satisfies the gradient equation:

STATISTICS

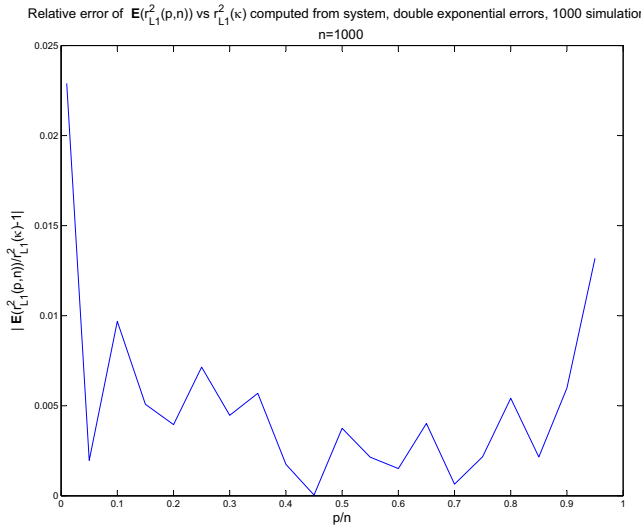

Fig. 3. Relative errors: $\left|\frac{\mathbf{E}(r^2_{t_1}(p,n))}{r^2_{t_1}(\kappa)}-1\right|$, double-exponential errors, 1,000 simulations.

$$\sum X_i\psi\left(\epsilon_i - X_i'\hat{\beta}\right) = 0. \qquad [3]$$

In the derivations that follow, we will use repeatedly the fact that if $\mathcal{X}_i$ are i.i.d $\mathcal{N}(0, \mathrm{Id}_p)$ and A_p is a sequence of deterministic symmetric matrices, under mild conditions on the growth of $\mathrm{trace}(A_p^k)$ with $k \in \mathbb{N}$, we have as n and p grow

$$\sup_{i=1,\ldots,n}\left|\frac{\mathcal{X}_i'A_p\mathcal{X}_i}{p} - \frac{\mathrm{trace}(A_p)}{p}\right| = o_P(1).$$

Many methods can be used to show this concentration result. A particularly simple one is to compute the second and fourth cumulants of $\mathcal{X}_i'A_p\mathcal{X}_i$. It shows that the result holds as soon as $\mathrm{trace}(A_p^4)p^{-4} + (\mathrm{trace}(A_p^2)\,p^{-2})^2 = o(1/n)$, a mild condition. This concentration result is easily extended to the case where A_p is random but independent of $\mathcal{X}_i$. The previous result also extends easily to $X_i = \lambda_i\mathcal{X}_i$, under mild conditions on λ_i's, to yield the following:

$$\sup_{i=1,\ldots,n}\left|\frac{X_i'A_pX_i}{p} - \lambda_i^2\frac{\mathrm{trace}(A_p)}{p}\right| = o_P(1). \qquad [4]$$

Leaving Out One Observation. Let us call $\hat{\beta}_{(i)}$ the usual leave-one-out estimator [i.e., the estimator we get by not using (X_i, Y_i) in our regression problem]. It solves

$$\sum_{j\neq i} X_j\psi\left(\epsilon_j - X_j'\hat{\beta}_{(i)}\right) = 0. \qquad [5]$$

Note that, when $\{X_i\}_{i=1}^n$ are independent, $\hat{\beta}_{(i)}$ is independent of X_i. For all j, $1 \le j \le n$, we call $\tilde{r}_{j,(i)}$

$$\tilde{r}_{j,(i)} = \epsilon_j - X_j'\hat{\beta}_{(i)}. \qquad [6]$$

When $j \neq i$, these are the residuals from this leave-one-out situation. For $j = i$, $\tilde{r}_{i,(i)}$ is the prediction error for observation i.

Intuitively, it is clear that under regularity conditions on ρ and ϵ_i's, when X_i's are i.i.d, for $i \neq j$, $R_j \simeq \tilde{r}_{j,(i)}$ (this means statistically that leave-one-out makes sense). However, it is easy to convince oneself (by looking, e.g., at the least-squares situation) that $\tilde{r}_{i,(i)}$ is very different from R_i in high dimension. The expansion we will get below will indeed confirm this fact in a more general setting than least squares.

Taking the difference between Eqs. **3** and **5**, we get, after using Taylor expansions for $j \neq i$ (and truncating the expansion at first order),

$$X_i\psi\left(\epsilon_i - X_i'\hat{\beta}\right) + \sum_{j\neq i}\psi'\left(\tilde{r}_{j,(i)}\right)X_jX_j'\left(\hat{\beta}_{(i)} - \hat{\beta}\right) \simeq 0.$$

We call $S_i = \sum_{j\neq i}\psi'(\tilde{r}_{j,(i)})X_jX_j'$. This suggests that

$$\hat{\beta} - \hat{\beta}_{(i)} \simeq S_i^{-1}X_i\psi\left(\epsilon_i - X_i'\hat{\beta}\right). \qquad [7]$$

Note that S_i is independent of X_i. Hence, multiplying the previous expression by X_i', we get, using the approximation given in Eq. **4** (which amounts to assuming that S_i^{-1} is "nice enough"),

$$R_i - \tilde{r}_{i,(i)} \simeq -\lambda_i^2\mathrm{trace}\left(S_i^{-1}\right)\psi(R_i).$$

Experience in random matrix theory as well as the form of the matrix S_i suggest that $\mathrm{trace}(S_i^{-1})$ should have a deterministic limit (again under conditions[‡] on ρ, λ_i's and ϵ_i's). Then, by symmetry between the observations, all $\mathrm{trace}(S_i^{-1})$ are approximately the same, i.e., when p and n are large, $\mathrm{trace}(S_i^{-1}) \simeq c$. Hence,

$$\boxed{R_i - \tilde{r}_{i,(i)} \simeq -\lambda_i^2 c\psi(R_i)}. \qquad [8]$$

Note that because X_i and $\hat{\beta}_{(i)}$ are independent when X_i's are independent and independent of $\{\epsilon_i\}_{i=1}^n$ and $\{\lambda_i\}_{i=1}^n$, much can be said about the distribution of $\tilde{r}_{i,(i)}$. However, at this point in the derivation, it is not clear what the value of c should be.

Leaving Out One Predictor. Let us consider what happens when we leave the pth predictor out. Because we are assuming that X_i is $\mathcal{N}(0, \mathrm{Id}_p)$ and $\beta_0 = 0$, all of the predictors play a symmetric role, so we pick the pth to simplify notations. There is nothing particular about it and the same analysis can be done with any other predictors.

Let us call $\hat{\gamma}(\in \mathbb{R}^{p-1})$ the corresponding optimal regression vector for the loss function ρ. We use the notations and partitions

$$X_i = \begin{bmatrix} V_i \\ X_i(p) \end{bmatrix}, \hat{\beta} = \begin{bmatrix} \hat{\beta}_{S_p} \\ \hat{\beta}_p \end{bmatrix}.$$

We have $V_i \in \mathbb{R}^{p-1}$. Naturally, $\hat{\gamma}$ satisfies

$$\sum_{i=1}^n V_i\psi\left(\epsilon_i - V_i'\hat{\gamma}\right) = 0.$$

We call

$$r_{i,[p]} = \epsilon_i - V_i'\hat{\gamma},$$

i.e., the residuals based on $p-1$ predictors. Note that $\{r_{i,[p]}\}_{i=1}^n$ is independent of $\{X_i(p)\}_{i=1}^n$ under our assumptions [because V_i is independent of $X_i(p)$ and the X_i's are i.i.d].

It is intuitively clear that[§] $R_i \simeq r_{i,[p]}$, for all i, because adding a predictor will not help us much in estimating $\beta_0 = 0$. Hence the residuals should not be much affected by the addition of one

[‡]To help with intuition, note that in the least-squares case, $S_i = \sum_{j\neq i}X_jX_j'$, a sample covariance matrix multiplied by $n-1$.

[§]Under regularity conditions on ρ, ϵ_i's and λ_i's.

predictor. Taking the difference of the equations defining $\hat{\beta}$ (Eq. **3**) and $\hat{\gamma}$, we get the following:

$$\sum_i X_i \psi\left(\epsilon_i - X_i'\hat{\beta}\right) - \begin{bmatrix} V_i \\ 0 \end{bmatrix} \psi\left(\epsilon_i - V_i'\hat{\gamma}\right) = 0.$$

This p-dimensional equation separates into a scalar and a vector equation, namely,

$$\sum_i X_i(p)\psi\left(\epsilon_i - X_i'\hat{\beta}\right) = 0,$$

$$\sum_i V_i\left[\psi(R_i) - \psi\left(r_{i,[p]}\right)\right] = 0_{p-1}.$$

Using a first-order Taylor expansion of $\psi(R_i)$ around $\psi(r_{i,[p]})$ and noting that $R_i - r_{i,[p]} = V_i'(\hat{\gamma} - \hat{\beta}_{S_p}) - X_i(p)\hat{\beta}_p$, we can transform the first equation above into

$$\sum_i X_i(p)\left[\psi\left(r_{i,[p]}\right) + \psi'\left(r_{i,[p]}\right)\left(V_i'\left(\hat{\gamma} - \hat{\beta}_{S_p}\right) - X_i(p)\hat{\beta}_p\right)\right] \simeq 0.$$

This gives the near identity

$$\hat{\beta}_p \simeq \frac{\sum X_i(p)\left[\psi\left(r_{i,[p]}\right) + \psi'\left(r_{i,[p]}\right)V_i'\left(\hat{\gamma} - \hat{\beta}_{S_p}\right)\right]}{\sum X_i^2(p)\psi'\left(r_{i,[p]}\right)}.$$

Working similarly on the equations involving V_i, we get

$$\sum_i \psi'\left(r_{i,[p]}\right)V_i\left[R_i - r_{i,[p]}\right] \simeq 0.$$

Because $R_i - r_{i,[p]} = -\hat{\beta}_p X_i(p) + V_i'(\hat{\gamma} - \hat{\beta}_{S_p})$, the previous equation reads

$$\left[\sum_i \psi'\left(r_{i,[p]}\right)V_i V_i'\right]\left(\hat{\gamma} - \hat{\beta}_{S_p}\right) - \hat{\beta}_p \sum_i \psi'\left(r_{i,[p]}\right)V_i X_i(p) \simeq 0.$$

Calling

$$\mathbb{S}_p = \sum_i \psi'\left(r_{i,[p]}\right)V_i V_i', \quad \text{and} \quad u_p = \sum_i \psi'\left(r_{i,[p]}\right)V_i X_i(p),$$

we see that $(\hat{\gamma} - \hat{\beta}_{S_p}) \simeq \hat{\beta}_p \mathbb{S}_p^{-1} u_p$. Using this approximation in the previous equation for $\hat{\beta}_p$, we have finally

$$\boxed{\hat{\beta}_p \simeq \frac{\sum X_i(p)\psi\left(r_{i,[p]}\right)}{\sum X_i^2(p)\psi'\left(r_{i,[p]}\right) - u_p'\mathbb{S}_p^{-1}u_p}}. \qquad [9]$$

Approximation of This Denominator. Let us write in matrix form $u_p'\mathbb{S}_p^{-1}u_p = X(p)'AX(p)$, where $A = D^{1/2}P_V D^{1/2}$, $P_V = D^{1/2}V(V'DV)^{-1}V'D^{1/2}$ and D is a diagonal matrix with $D(i,i) = \psi'(r_{i,[p]})$. Note that P_V is a projection matrix of rank $p-1$ in general.

Let us call ξ_n the denominator of $\hat{\beta}_p$ divided by n. We have

$$\xi_n = \frac{1}{n}X(p)'(D - A)X(p).$$

Let us call $\mathbb{S}_p(i) = \mathbb{S}_p - \psi'(r_{i,[p]})V_i V_i'$. Using the Sherman–Morrison–Woodbury formula (see ref. 10, p. 19, and *SI Text*), we see that $P_V(i,i) = 1 - \frac{1}{1+\psi'(r_{i,[p]})V_i'[\mathbb{S}_p(i)]^{-1}V_i}$. We notice that $\mathbb{S}_p(i)^{-1}$ can be approximated by a matrix M_i^{-1}, which is independent of V_i (by using our leave-one-predictor-out observations) for which $V_i'M_i^{-1}V_i/\lambda_i^2 = \mathrm{trace}(M_i^{-1}) + o_P(1)$ by Eq. **4** (these two approximations naturally require some regularity conditions on ρ, etc... so that M_i^{-1} is "nice enough"). Hence,

$$P_V(i,i) = 1 - \frac{1}{1 + \lambda_i^2\psi'\left(r_{i,[p]}\right)\mathrm{trace}\left(\left[\mathbb{S}_p(i)\right]^{-1}\right)} + o_P(1).$$

Therefore, using the approximations $r_{i,[p]} \simeq R_i$ and $\mathrm{trace}([\mathbb{S}_p(i)]^{-1}) \simeq c$ (because $\mathrm{trace}([\mathbb{S}_p(i)]^{-1}) \simeq \mathrm{trace}([\mathbb{S}_p]^{-1})$ using Sherman–Morrison–Woodbury), we also have

$$1 - P_V(i,i) = \frac{1}{1 + \psi'\left(r_{i,[p]}\right)V_i'\left[\mathbb{S}_p(i)\right]^{-1}V_i} \simeq \frac{1}{1 + \lambda_i^2 c\psi'\left(r_{i,[p]}\right)}.$$

Because P_V is a rank $(p-1)$ projection matrix, we have $\mathrm{trace}(P_V) = p - 1 = \sum_i P_V(i,i)$, and therefore

$$\frac{1}{n}\sum_{i=1}^n \frac{1}{1 + \lambda_i^2 c\psi'\left(r_{i,[p]}\right)} = 1 - \frac{p}{n} + o_P(1).$$

Using concentration properties of $X(p)$ conditional on $\{\lambda_i\}_{i=1}^n$, we have

$$\begin{aligned}\xi_n &= \frac{1}{n}\mathrm{trace}(D_\lambda(D - A)D_\lambda) + o_P(1) \\ &= \frac{1}{n}\sum_{i=1}^n \lambda_i^2\psi'\left(r_{i,[p]}\right)(1 - P_V(i,i)) + o_P(1).\end{aligned}$$

Replacing $1 - P_V(i,i)$ by its approximate value, we get

$$\begin{aligned}\xi_n &\simeq \frac{1}{n}\sum_{i=1}^n \frac{\lambda_i^2\psi'\left(r_{i,[p]}\right)}{1 + c\lambda_i^2\psi'\left(r_{i,[p]}\right)}, \\ &= \frac{1}{c}\left(1 - \frac{1}{n}\sum_{i=1}^n \frac{1}{1 + c\lambda_i^2\psi'\left(r_{i,[p]}\right)}\right) \simeq \frac{1}{c}\frac{p}{n}.\end{aligned}$$

So finally,

$$\boxed{\hat{\beta}_p \simeq \frac{\sum X_i(p)\psi\left(r_{i,[p]}\right)/n}{\frac{p}{n}\Big/c} \simeq c\frac{1}{p}\sum_{i=1}^n \lambda_i\psi\left(r_{i,[p]}\right)\mathcal{X}_i(p)}. \qquad [10]$$

Using again $\psi(r_{i,[p]}) \simeq \psi(R_i)$, we see that

$$\mathbf{E}\left(||\hat{\beta}||^2\right) \simeq \frac{n}{p}\left[\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left(c^2\lambda_i^2\psi^2(R_i)\right)\right], \qquad [11]$$

assuming that we can take expectations in all these approximations.

From Approximations to Functional System. Our approximations concerning the residuals and $\hat{\beta}_p$ shed considerable light on them. Our focus is now on $||\hat{\beta}||$.

From Eq. **8**, we got the approximation

$$\tilde{r}_{i,(i)} \simeq R_i + c\lambda_i^2\psi(R_i).$$

Recall that, for a convex, proper, and closed function ρ, whose subdifferential we call ψ, and $t > 0$, $\mathrm{prox}_t(\rho) = (\mathrm{Id} + t\psi)^{-1}$. It is an important fact that the prox is indeed a function and not a multivalued mapping, even when ρ is not differentiable everywhere. We therefore get the approximation

$$R_i \simeq \mathrm{prox}_{c\lambda_i^2}(\rho)\left(\tilde{r}_{i,(i)}\right).$$

Recalling Eq. **6** and using the independence of $\hat{\beta}_{(i)}$ and X_i, we

have $\tilde{r}_{i,(i)} \overset{\mathcal{L}}{=} \epsilon_i + |\lambda_i| ||\hat{\beta}_{(i)}|| Z_i$, where Z_i is $\mathcal{N}(0,1)$ and independent of ϵ_i, λ_i and $||\hat{\beta}_{(i)}||$ ($\overset{\mathcal{L}}{=}$ denotes "equal in law").

We now argue that $||\hat{\beta}||$ is asymptotically deterministic. Using the relationship between $\hat{\beta}$ and $\hat{\beta}_{(i)}$ in Eq. **7** and taking squared norms, we see that

$$\left\|\hat{\beta}\right\|^2 \simeq \left\|\hat{\beta}_{(i)}\right\|^2 + 2\hat{\beta}'_{(i)} S_i^{-1} X_i \psi(R_i) + X_i' S_i^{-2} X_i \psi^2(R_i).$$

Assuming that the smallest eigenvalue of S_i/n remains bounded, which is automatically satisfied with high probability for strongly convex functions ρ, we see that $||\hat{\beta}||^2 - ||\hat{\beta}_{(i)}||^2$ is $\mathrm{O}_P(1/n)$, provided $||\hat{\beta}_{(i)}||$ remains bounded and ψ and ψ' do not grow too fast at infinity. Applying the Efron–Stein inequality, we see that $\mathrm{var}(||\hat{\beta}||^2) = \mathrm{O}(1/n)$ if we take squared expectations in our approximations. It follows that $||\hat{\beta}||^2$ is asymptotically deterministic.

These arguments suggest that, as p and n become large, $\tilde{r}_{i,(i)} \overset{\mathcal{L}}{=} \epsilon_i + |\lambda_i| r_\rho(\kappa) Z_i + \mathrm{o}_P(1)$, where $Z_i \sim \mathcal{N}(0,1)$, independent of λ_i and ϵ_i, and $r_\rho(\kappa)$ is deterministic. We also note that Z_i are i.i.d, because $\mathcal{X}_i$ are.

Because $c\lambda_i^2 \psi(R_i) \simeq \tilde{r}_{i,(i)} - R_i \simeq \tilde{r}_{i,(i)} - \mathrm{prox}_{c\lambda_i^2}(\rho)(\tilde{r}_{i,(i)})$, we see that Eq. **11** now becomes asymptotically

$$\kappa r_\rho^2(\kappa) = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left(\lambda_i^{-2}\left[\tilde{r}_{i,(i)} - \mathrm{prox}_{c\lambda_i^2}(\rho)\left(\tilde{r}_{i,(i)}\right)\right]^2\right),$$

where the expectations are over the joint distribution of λ_i's, ϵ_i's, and Z_i's. (We note that our arguments do not depend on independence of λ_i's or ϵ_i's, although both families of random variables need to be independent of $\{\mathcal{X}_i\}_{i=1}^n$.) This is the second equation of system **S1**.

We now recall that using the fact that the matrix P_V above was a projection matrix, we had argued that asymptotically

$$\frac{1}{n}\sum_{i=1}^n \frac{1}{1 + c\lambda_i^2 \psi(R_i)} = 1 - \kappa + \mathrm{o}_P(1).$$

We observe that $(\mathrm{prox}_{c\lambda_i^2}(\rho))'(\tilde{r}_{i,(i)}) = \frac{1}{1 + c\lambda_i^2 \psi(\mathrm{prox}_{c\lambda_i^2}[\tilde{r}_{i,(i)}])}$ and therefore $\frac{1}{1+c\lambda_i^2\psi(R_i)} \simeq (\mathrm{prox}_{c\lambda_i^2}(\rho))'(\tilde{r}_{i,(i)})$. This allows us to conclude that under regularity conditions,

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left(\left(\mathrm{prox}_{c\lambda_i^2}(\rho)\right)'\left(\tilde{r}_{i,(i)}\right)\right) = 1 - \kappa.$$

This is the first equation of our system **S1**.

A Note on Nondifferentiable ρ's. One of the appeals of our systems **S1** and **S2** is that they yield expressions even in the case of nondifferentiable ρ, because the prox is well defined. However, we derived the systems assuming smoothness of ρ. To go around this hurdle, one can approximate ρ by a family ρ_η of smooth convex functions such that $\rho_\eta \to \rho$ as $\eta \to 0$ in an appropriate sense. Intuitively, it is quite clear that $r_{\rho_\eta}(\kappa)$ should tend to $r_\rho(\kappa)$ as η tends to 0 under appropriate regularity conditions on ϵ_i's and λ_i's. We then just need to take limits in our systems to justify them for nondifferentiable ρ's.

ACKNOWLEDGMENTS. D.B. gratefully acknowledges support from National Science Foundation (NSF) Grant DMS-0636667 (Vertical Integration of Research and Education in the Mathematical Sciences); P.J.B. gratefully acknowledges support from NSF Grant DMS-0907362; N.E.K. gratefully acknowledges support from an Alfred P. Sloan Research Fellowship and NSF Grant DMS-0847647 (Faculty Early Career Development); and B.Y. gratefully acknowledges support from NSF Grants SES-0835531 (Cyber-Enabled Discovery and Innovation), DMS-1107000, and CCF-0939370.

1. Huber PJ, Ronchetti EM (2009) *Robust Statistics*. Wiley Series in Probability and Statistics (Wiley, Hoboken, NJ), 2nd Ed.
2. Huber PJ (1973) Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann Stat* 1:799–821.
3. Portnoy S (1984) Asymptotic behavior of *M*-estimators of *p* regression parameters when p^2/n is large. I. Consistency. *Ann Stat* 12(4):1298–1309.
4. Bloomfield P (1974) On the distribution of the residuals from a fitted linear model (Department of Statistics, Princeton Univ, Princeton, NJ), Technical Report 56, Series 2.
5. El Karoui N, Bean D, Bickel P, Lim C, Yu B (2012) On robust regression with high-dimensional predictors (Department of Statistics, Univ of California, Berkeley, CA), Technical Report 811.
6. Moreau J-J (1965) Proximité et dualité dans un espace hilbertien. *Bull Soc Math France* 93:273–299. French.
7. El Karoui N (2010) High-dimensionality effects in the Markowitz problem and other quadratic programs with linear constraints: Risk underestimation. *Ann Stat* 38: 3487–3566.
8. Eaton ML (1983) *Multivariate Statistics: A Vector Space Approach* (Wiley, New York); reprinted (2007) Institute of Mathematical Statistics Lecture Notes—Monograph Series (Institute of Mathematical Statistics, Beachwood, OH), Vol 53.
9. Bean D, Bickel PJ, El Karoui N, Yu B (2013) Optimal M-estimation in high-dimensional regression. *Proc Natl Acad Sci USA* 110:14563–14568.
10. Horn RA, Johnson CR (1994) *Topics in Matrix Analysis* (Cambridge Univ Press, Cambridge, UK), corrected reprint of the 1991 original.