Optimal M-estimation in high-dimensional regression

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

Department of Statistics, University of California, Berkeley, CA 94720

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

We consider, in the modern setting of high-dimensional statistics, the classic problem of optimizing the objective function in regression using M-estimates when the error distribution is assumed to be known. We propose an algorithm to compute this optimal objective function that takes into account the dimensionality of the problem. Although optimality is achieved under assumptions on the design matrix that will not always be satisfied, our analysis reveals generally interesting families of dimension-dependent objective functions.

robust regression | prox function

In this article, we study a classical statistical problem: how to pick optimally (among regression M-estimates) the objective to be minimized in a parametric regression when we know the error distribution. Our study is carried out in the modern setting where the number of predictors is proportional to the number of observations.

The classical answer to the problem we posed at the beginning, maximum likelihood, was given by Fisher (1) in the specific case of multinomial models and then at succeeding levels of generality by Cramér (2), Hájek (3), and, above all, Le Cam (4). For instance, for p fixed or $p/n \to 0$ fast enough, least squares is optimal for Gaussian errors, whereas least absolute deviations (LAD) is optimal for double-exponential errors. We shall show that this is no longer true in the regime we consider with the answer depending, in general, on the limit of the ratio p/n as well as the form of the error distribution. Our analysis in this paper is carried out in the setting of Gaussian predictors, although as we explain below, this assumption should be relaxable to a situation in which the distribution of the predictors satisfy certain concentration properties for quadratic forms.

We carry out our analysis in a regime that has been essentially unexplored, namely $0 \ll p/n < 1$, where p is the number of predictor variables and n is the number of independent observations. Because in most fields of application, situations in which p as well as n are large have become paramount, there has been a huge amount of literature on the case where $p/n \gg 0$ but the number of "relevant" predictors is small. In this case, the objective function, quadratic (least squares) or otherwise (ℓ_1 for LAD) has been modified to include a penalty (usually ℓ_1) on the regression coefficients, which forces sparsity (5). The price paid for this modification is that estimates of individual coefficients are seriously biased and statistical inference, as opposed to prediction, often becomes problematic.

In ref. 6, we showed[†] that this price need not be paid if p/n stays bounded away from 1, even if the regression does not have a sparse representation. We review the main theoretical results from this previous paper in *Result 1* below. Some of our key findings were as follows: (*i*) Surprisingly, when $0 \ll p/n < 1 - \epsilon$, it is no longer true that LAD is necessarily better than least squares for heavy-tailed errors. This behavior is unlike that in the classical regime p bounded or $p/n \to 0$ fast enough studied, for instance, in ref. 7. (*ii*) Linear combinations of regression coefficients are unbiased and still asymptotically Gaussian at rate[‡] $1/\sqrt{n}$.

This article contains three main parts: *Background and Main Results* contains needed background and a description of our findings. In *Computing the Optimal Objective*, we give two examples of interest to statisticians: the case of Gaussian errors and the case of double-exponential errors. We present our derivations in the last section.

Background and Main Results

We consider a problem in which we observe n independent, identically distributed pairs (X_i, Y_i), where X_i is a p-dimensional vector of predictors, and Y_i is a scalar response. We call the problem high-dimensional when the ratio p/n is not close to 0. In effect, we are considering an asymptotic setting where $\lim\ \inf\ p/n$ is not 0. We also limit ourselves to the case in which $\lim\ \sup\ p/n < 1$. As far as we know, all of the very large body of work developed in robust regression (following ref. 7) is concerned with situations in which p/n tends to 0, as n tends to infinity.

Let us briefly recall the details of the robust regression problem. We consider the estimator

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

where ρ is a function from $\mathbb{R}$ to $\mathbb{R}$, which we will assume throughout is convex.[§] Furthermore, we consider a linear regression model

$$Y_i = \epsilon_i + X_i'\beta_0,$$

where $\beta_0 (\in \mathbb{R}^p)$ is unknown and $\{\epsilon_i\}_{i=1}^n$ are random errors, independent of X_i's. Naturally, our aim is to estimate β_0 from our observations $\{(X_i, Y_i)\}_{i=1}^n$ and the question is therefore, which ρ we should choose. We can separate this into the following two questions: (*i*) What choice of ρ minimizes the asymptotic error for estimating an individual regression coefficient (or a given linear form in β_0)? (*ii*) What choice of ρ minimizes the asymptotic prediction error for a new observation (X_{new}, Y_{new}) given the training data? The answers to (*i*) and (*ii*) turn out to be the same in the high-dimensional and Gaussian setting we are considering, just as in the low-dimensional case, but the extension is surprising.

Some Recent High-Dimensional Results. In a recent paper (6), we found heuristically the following.

Result 1. [*See El Karoui et al.* (6).] *Suppose X_i are independent identically distributed* (*i.i.d*) *$\mathcal{N}(0, \Sigma)$, with Σ positive definite. Suppose $Y_i = \epsilon_i + X_i'\beta_0$, ϵ_i's are i.i.d, independent of X_i, $\beta_0 \in \mathbb{R}^p$ is deterministic*, *and $n \geq p$. Call*

$$\hat{\beta}(\rho; \beta_0, \Sigma) = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho(Y_i - X_i'\beta).$$

Then we have the stochastic representation

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

The authors declare no conflict of interest.

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

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1307845110/-/DCSupplemental.

[†] Heuristically.

[‡] Under conditions depending on the model and the linear combination; see details below.

[§] The properties of $\hat{\beta}$ naturally depend on ρ—this dependence will be made clear later.

$$\hat{\beta}(\rho;\beta_0,\Sigma) \overset{\mathcal{L}}{=} \beta_0 + \Sigma^{-1/2}\hat{\beta}(\rho;0,\mathrm{Id}_p), \quad \overset{\mathcal{L}}{=} \beta_0 + \left\|\hat{\beta}(\rho;0,\mathrm{Id}_p)\right\|\Sigma^{-1/2}u,$$

where u is uniform on $\mathbb{S}_{p-1}$(the unit sphere in $\mathbb{R}^p$) and independent of $\|\hat{\beta}(\rho;0,\mathrm{Id}_p)\|$.

Let us call $r_\rho(p,n) = \|\hat{\beta}(\rho;0,\mathrm{Id}_p)\|$. As p and n tend to infinity, while $p \le n$ and $\lim_{n\to\infty} p/n = \kappa < 1$, $r_\rho(p,n) \to r_\rho(\kappa)$ in probability (under regularity conditions on ρ and ε), where $r_\rho(\kappa)$ is deterministic. Define $\hat{z}_\epsilon = \epsilon + r_\rho(\kappa)Z$, where $Z \sim \mathcal{N}(0,1)$ is independent of ε, and ε has the same distribution as ϵ_i. We can determine $r_\rho(\kappa)$ through

$$\begin{cases} \mathbf{E}\Big([prox_c(\rho)]'(\hat{z}_\epsilon)\Big) & = 1-\kappa, \\ \mathbf{E}\Big(\big[\hat{z}_\epsilon - prox_c(\rho)(\hat{z}_\epsilon)\big]^2\Big) & = \kappa r_\rho^2(\kappa), \end{cases} \qquad \textbf{[S]}$$

where c is a positive deterministic constant to be determined from the previous system.

The definition and details about the prox mapping are given in *Appendix: Reminders*. This formulation is important because it shows that what matters about an objective function in high dimension is not really the objective itself but rather its prox, in connection with the distribution of the errors. We also note that our analysis in ref. 6 highlights the fact that the result concerning $r_\rho(\kappa)$ should hold when normality of the predictors is replaced by a concentration of quadratic form assumption. System **S** is the basis of our analysis.

Consequences for estimation of β_0. If v is a given deterministic vector, we see that $v'\hat{\beta}(\rho;\beta_0,\Sigma)$ is unbiased for $v'\beta_0$ and

$$\mathrm{var}\big(v'\hat{\beta}(\rho;\beta_0,\Sigma)\big) = \frac{v'\Sigma^{-1}v}{p}\mathbf{E}\Big(||\hat{\beta}(\rho;0,\mathrm{Id}_p)||^2\Big).$$

In other words, in high dimension, the simple estimator $v'\hat{\beta}(\rho;\beta_0,\Sigma)$ is $\sqrt{p/(v'\Sigma^{-1}v)}$-consistent for $v'\beta_0$. We further note that $\sqrt{p}v'\hat{\beta}(\rho;\beta_0,\Sigma)$ is asymptotically normal, and its variance can be estimated, so inference about $v'\beta_0$ is easy. More details are in *SI Text*, including for instance explicit confidence intervals. Picking v to be the kth canonical basis vector, e_k, we also see that we can consistently estimate the kth coordinate of β_0, $\beta_0(k)$, at rate $\sqrt{p/(e_k'\Sigma^{-1}e_k)}$.

A similar analysis can be performed to obtain unbiased (and consistent) estimators of quadratic forms in β_0, i.e., quantities of the form $\beta_0'\Sigma_2\beta_0$, where Σ_2 is a given covariance matrix.

The situation in which the vector v is random (and for instance dependent on the design) is beyond the scope of this paper. We note that Huber (7) has early very partial results on this problem.

On expected prediction error. In the case in which (X_{new}, Y_{new}) follows the model above and is independent of $\{(X_i,Y_i)\}_{i=1}^n$, we immediately see that

$$EPE = \mathbf{E}\Big(\big(Y_{new} - X_{new}'\hat{\beta}\big)^2\Big) = \sigma_\epsilon^2 + \mathbf{E}\Big(||\hat{\beta}(\rho;0,\mathrm{Id}_p)||^2\Big).$$

Picking ρ to minimize the quantity $\mathbf{E}(||\hat{\beta}(\rho;0,\mathrm{Id}_p)||^2)$ (viewed as a function of ρ) will allow us to get the best estimators (in the class we are considering) for both expected prediction error (*EPE*) and, as it turns out, linear forms in β_0.

Potential Limitations of the Model. ***Assumptions concerning the predictors.*** The impact of changing the distributional assumptions of the predictors on *Result 1* is explored and explained in details in our papers (6) and especially ref. 8. We present in ref. 8 design matrices for which *Result 1* does not hold. For those designs, using the family of objective functions described in *Theorem 1* below will not yield optimal results—however, using dimension-dependent objective functions may still yield improvement over using standard, dimension-independent objective functions. We also note that parts of *Result 1* hold under less restrictive assumptions, but one key limitation is that our generalizations do not include categorical predictors.

The question of intercept. The model and estimation problem described in *Result 1* do not include an intercept. Furthermore, both X_i and Y_i are assumed to have mean 0. However, if we recenter both the responses Y_i's and the predictors X_i's and fit a linear model (without intercept) on this recentered dataset, we have a solution to these problems that essentially fits into the framework of our analysis. The corresponding $\hat{\beta}(\rho;0,\mathrm{Id}_p)$ has the same stochastic representation properties as the ones described in *Result 1*, and hence inference about β_0 is possible in this case, too. More details and justifications are provided in *SI Text*. We note that, as presented in this paper, our results do not allow us to address inferential questions involving the intercept. However, in separate and ongoing work, we have obtained results that suggest that we will be able to answer some of these inferential questions as well as issues pertaining to the fitted values [see Huber (7), remarks 2 and 3, pp. 803–804, for early comments about the difficulty they pose in the high-dimensional setting].

The rest of the paper is focused on the model and statistical problem discussed in *Result 1*.

Main Result. We propose an algorithm to determine the asymptotically optimal objective function to use in robust regression. As explained in *SI Text*, under the assumptions of *Result 1*, this optimal objective function is the same regardless of the norm chosen to assess the performance of our regression estimator. Just as in the classical case, determining the optimal objective function requires knowledge of the distribution of the errors, which we call ε. We call the density of the errors f_ϵ and assume that f_ϵ is log-concave.

If ϕ_r is the normal density with variance r^2 and $f_{r,\epsilon} = \phi_r \star f_\epsilon$, where $\star$ is the usual convolution operation, $f_{r,\epsilon}$ is log-concave. As a matter of fact, it is well known (9, 10) that the convolution of two log-concave densities is log-concave.

We call $I_\epsilon(r) = \int (f_{r,\epsilon}')^2/f_{r,\epsilon}$ the information of $f_{r,\epsilon}$, which we assume exists for all $r \ge 0$. It is known that, when ε has a density, $r^2 I_\epsilon(r)$ is continuous in r (11), where it is explained that $I_\epsilon(\sqrt{r})$ is differentiable or see ref. 12.

Throughout the paper, we denote by p_2 the function taking value

$$p_2(x) = x^2/2.$$

Here is our theorem.

Theorem 1. *If r_ρ is a solution of system* **S**, *we have $r_\rho \ge r_{opt}(\kappa)$, where $r_{opt}(\kappa) = \min\{r : r^2 I_\epsilon(r) = \kappa\}$.*

Furthermore, $r_{opt}(\kappa)$ is the solution of system **S** *when $\rho = \rho_{opt}$, and ρ_{opt} is the convex function*

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

[*For a function g, g^* is its (Fenchel–Legendre) conjugate, i.e., $g^*(x) = \sup_y[xy - g(y)]$.*]

We give an alternative representation of ρ_{opt} in *Appendix: Reminders*.

We propose the following algorithm for computing the optimal objective function under the assumptions of the theorem.

1. Solve for r the equation

$$r^2 I_\epsilon(r) = p/n. \qquad [1]$$

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

2. Use the objective function[¶]

$$\rho_{\text{opt}} = \left(p_2 + r_{\text{opt}}^2 \log\left(\phi_{r_{\text{opt}}} \star f_\epsilon\right)\right)^* - p_2. \qquad [2]$$

The theorem and the algorithm raise a few questions: Is there a solution to the equation in step 1? Is the min well defined? Is the objective function in step 2 convex? We address all these questions in the course of the paper.

The significance of the algorithm lies in the fact that we are now able to incorporate dimensionality in our optimal choice of ρ. In other words, different objectives turn out to be optimal as the ratio of dimensions varies. Naturally, the optimality of the objective function is restricted to situations in which our assumptions on the design matrix are satisfied. However, more generally, our results provide practitioners with a "natural" family of objective functions that could be tried if one would like to use a dimension-dependent objective function in a high-dimensional regression problem. This family of functions is also qualitatively different from the ones that appear naturally in the small p setting, as our derivation (see below) shows.

Because our estimators are M-estimators, they are not immune to inadmissibility problems, as has been understood, in an even simpler context, since the appearance of James–Stein estimators. Picking ρ_{opt} instead of another ρ, although improving performance when our assumptions are satisfied, does not remove this potential problem.

It should also be noted that at least numerically, computing ρ_{opt} is not very hard. Similarly, solving Eq. **1** is not hard numerically. Hence, the algorithm is effective as soon as we have information about the distribution of ε.

As the reader will have noticed, a crucial role is played by $\hat{\beta}(\rho; 0, \text{Id}_p)$. In the rest of the paper, we use the lighter notation

$$\hat{\beta}_\rho \triangleq \hat{\beta}(\rho; 0, \text{Id}_p).$$

The dependence of $\hat{\beta}_\rho$ on p and n is left implicit in general, but will be brought back when there are any risks of confusion.

Next, we illustrate our algorithm in a few special cases.

Computing the Optimal Objective

The Case of Gaussian Errors.

Corollary 1. *In the setting of i.i.d Gaussian predictors, among all convex objective functions, l_2 is optimal in regression when the errors are Gaussian.*

In the case of Gaussian ε, it is clear that $\phi_{r_{\text{opt}}} \star f_\epsilon$ is a Gaussian density. Hence, $(p_2 + r_{\text{opt}}^2 \log(\phi_{r_{\text{opt}}} \star f_\epsilon))^*$ is a multiple of p_2 (up to centering) and so is ρ_{opt}. General arguments given later guarantee that this latter multiple is strictly positive. Therefore, ρ_{opt} is p_2, up to positive scaling and centering. Carrying out the computations detailed in the algorithm, we actually arrive at $\rho_{\text{opt}}(x) = \frac{x^2}{2}\left(\frac{p/n}{1-p/n}\right) - K$. Details are in *SI Text*.

The Case of Double-Exponential Errors. We recall that in low dimension (e.g., p fixed, n goes to infinity), classic results show that the optimal objective is ℓ_1. As we will see, it is not at all of the case when p and n grow in such a way that p/n has a finite limit in $(0, 1)$. We recall that, in ref. 6, we observed that when p/n was greater than 0.3 or so, ℓ_2 actually performed better than ℓ_1 for double-exponential errors.

Although there is no analytic form for the optimal objective, it can be computed numerically. We discuss how and present a picture to get a better understanding of the solution of our problem.

The Optimal Objective. For $r > 0$, $r \in \mathbb{R}$, and Φ, the Gaussian cumulative distribution function, let us define

$$R_r(x) = r^2 \log\left(e^{\frac{\left(x-r^2\right)^2}{2r^2}} \Phi\left[\frac{x-r^2}{r}\right] + e^{\frac{\left(x+r^2\right)^2}{2r^2}} \Phi\left[-\frac{x+r^2}{r}\right]\right) + r^2 \log\left(\sqrt{\frac{\pi}{2}} r\right).$$

It is easy to verify that, when the errors are double exponential, $-r^2 \log(\phi_r \star f_\epsilon)(x) = x^2/2 - R_r(x)$. Hence, effectively the optimal objective is the function taking values

$$\rho_{\text{opt}}(x) = R^*_{r_{\text{opt}}}(x) - x^2/2.$$

It is of course important to be able to compute this function and the estimate $\hat{\beta}_{\text{opt}}$ based on it. We show below that R_r is a smooth convex function for all r. Hence, in the case we are considering, R'_r is increasing and therefore invertible. If we call $y^*(x) = (R'_{r_{\text{opt}}})^{-1}(x)$, we see that $\rho_{\text{opt}}(x) = xy^*(x) - R_{r_{\text{opt}}}(y^*(x)) - x^2/2$. We also need to be able to compute the derivative of ρ_{opt} (denoted ψ_{opt}) to implement a gradient descent algorithm to compute $\hat{\beta}_{\text{opt}}$. For this, we can use a well-known result in convex analysis, which says that for a convex function h (under regularity conditions) $(h^*)' = (h')^{-1}$ (see ref. 13, corollary 23.5.1).

We present a plot to get an intuitive feeling for how this function ρ_{opt} behaves (more can be found in *SI Text*). Fig. 1 compares ρ_{opt} to other objective functions of potential interest in the case of $p/n = 0.5$. All of the functions we compare are normalized so that they take value 0 at 0 and 1 at 1.

Comparison of Asymptotic Performance of ρ_{opt} Against Other Objective Functions. We compare r_{opt}^2 to the results we would get using other objective functions ρ in the case of double-exponential errors. Recall that our system **S** allows us to compute the asymptotic value of $\|\hat{\beta}_\rho\|^2$, r_ρ^2, as n and p go to infinity for any convex (and sufficiently regular) ρ.

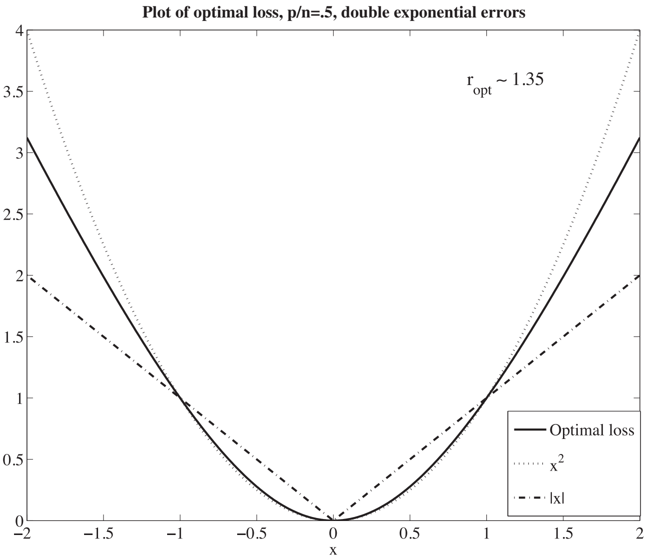

Fig. 1. $p/n = 0.5$: comparison of ρ_{opt} (optimal objective) to l_2 and l_1. r_{opt} is the solution of $r^2 I_\epsilon(r) = p/n$; for $p/n = 0.5$, $r_{opt} \simeq 1.35$.

[¶] Note that any $\lambda\rho_{\text{opt}} + \xi$, where λ and ξ are real valued with $\lambda > 0$, yields the same solution for $\hat{\beta}$.

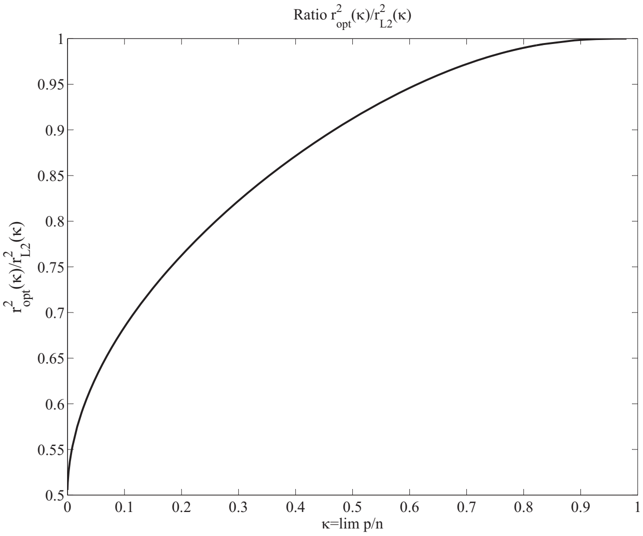

Fig. 2. Ratio $r^2_{\text{opt}}(\kappa)/r^2_{\ell_2}(\kappa)$ for double-exponential errors: the ratio is always less than 1, showing the superiority of the objective we propose over ℓ_2.

Comparison of ρ_{opt} to ℓ_2. We compare r^2_{opt} to $r^2_{\ell_2}$ in Fig. 2. Interestingly, ρ_{opt} yields a $\hat{\beta}_\rho$ that is twice as efficient as $\hat{\beta}_{\ell_2}$ as p/n goes to 0. From classical results in robust regression (p bounded), we know that this is optimal because ℓ_1 objective is optimal in that setting, and also yields estimators that are twice as efficient as $\hat{\beta}_{\ell_2}$.

Comparison of ρ_{opt} to ℓ_1. We compare r^2_{opt} to $r^2_{\ell_1}$ in Fig. 3. Naturally, the ratio goes to 1 when p/n goes to 0, because ℓ_1, as we just mentioned, is known to be the optimal objective function for p/n tending to 0.

Simulations. We investigate the empirical behavior of estimators computed under our proposed objective function. We call those estimators $\hat{\beta}_{\text{opt}}$. Table 1 shows $\mathbf{E}(r^2_{\text{opt}}(p,n))/\mathbf{E}(r^2_{\ell_2}(p,n))$ over 1,000 simulations when $n=500$ for different ratios of dimensions and compares the empirical results to $r^2_{\text{opt}}(\kappa)/r^2_{\ell_2}(\kappa)$, the theoretical values. We used $\beta_0=0$, $\Sigma=\text{Id}_p$ and double-exponential errors in our simulations.

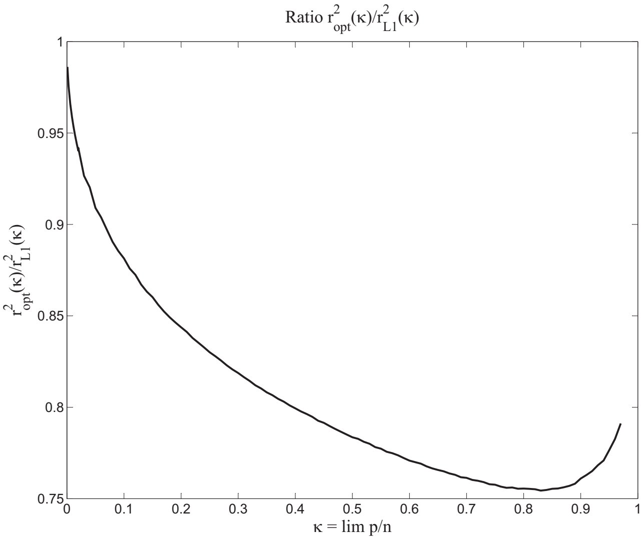

Fig. 3. Ratio $r^2_{\text{opt}}(\kappa)/r^2_{\ell_1}(\kappa)$: the ratio is always less than 1, showing the superiority of the objective we propose over ℓ_1. Naturally, the ratio goes to 1 at 0, because we know that ℓ_1 is the optimal objective when $p/n\to 0$ for double-exponential errors.

In *SI Text*, we also provide statistics concerning $||\hat{\beta}_{\text{opt}}-\beta_0||^2/||\hat{\beta}_{\ell_2}-\beta_0||^2$ computed over 1,000 simulations. We note that our predictions concerning $||\hat{\beta}_{\text{opt}}-\beta_0||^2$ work very well in expectation when p and n are a few 100's, even though in these dimensions, $||\hat{\beta}_{\text{opt}}-\beta_0||$ is not yet close to being deterministic (see *SI Text* for details—these remarks also apply to $||\hat{\beta}_\rho||^2$ for more general ρ).

Derivations

We prove *Theorem 1* assuming the validity of *Result 1*.

Phrasing the Problem as a Convex Feasibility Problem. Let us call $r_\rho(\kappa)=\lim_{n\to\infty}||\hat{\beta}(\rho;0,\text{Id}_p)||$, where $p/n\to\kappa<1$. We now assume throughout that $p/n\to\kappa$ and call $r_\rho(\kappa)$ simply r_ρ for notational simplicity. We recall that for $c>0$ $\text{prox}_c(\rho)=\text{prox}_1(c\rho)$ (*Appendix: Reminders*). From now on, we call prox_1 just prox. If r_ρ is feasible for our problem, there is a ρ that realizes it and the system $\mathbf{S}$ is therefore, with $\hat{z}_\epsilon=r_\rho Z+\epsilon$,

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

Now it is clear that if we replace ρ by $\lambda\rho$, $\lambda>0$, we do not change $\hat{\beta}_\rho$. In particular, if we call $\rho_0=c\rho$, where c is the real appearing in the system above, we have, if r_ρ is feasible: there exists ρ_0 such that

$$\begin{cases} \mathbf{E}\Big([\text{prox}(\rho_0)]'(\hat{z}_\epsilon)\Big) & =1-\kappa,\\ \mathbf{E}\Big(\left[\hat{z}_\epsilon-\text{prox}(\rho_0)(\hat{z}_\epsilon)\right]^2\Big) & =\kappa r_\rho^2.\end{cases}$$

We can now rephrase this system using the fundamental equality (see ref. 14 and *Appendix: Reminders*) $\text{prox}(\rho)+\text{prox}(\rho^*)=x$, where ρ^* is the (Fenchel–Legendre) conjugate of ρ. It becomes

$$\begin{cases} \mathbf{E}\Big([\text{prox}(\rho_0^*)]'(\hat{z}_\epsilon)\Big) & =\kappa,\\ \mathbf{E}\Big(\left[\text{prox}(\rho_0^*)(\hat{z}_\epsilon)\right]^2\Big) & =\kappa r_\rho^2.\end{cases}$$

Prox mappings are known to belong to subdifferentials of convex functions and to be contractive (see ref. 14, p. 292, corollaire 10. c). Let us call $g=\text{prox}(\rho_0^*)$ and recall that $f_{r,\epsilon}$ denotes the density of $\hat{z}_\epsilon=rZ+\epsilon$. Because g is contractive, $|g(x)/x|\leq 1$ as $|x|\to\infty$. Because $f_{r,\epsilon}$ is a log-concave density (as a convolution of two log-concave densities—see refs. 9 and 10) with support $\mathbb{R}$, it goes to zero at infinity exponentially fast (see ref. 15, p. 332). We can therefore use integration by parts in the first equation to rewrite the previous system as (we now use r instead of r_ρ for simplicity)

$$\begin{cases} -\displaystyle\int g(x)f'_{r,\epsilon}(x)dx & =\kappa,\\ \displaystyle\int g^2(x)f_{r,\epsilon}(x)dx & =\kappa r^2.\end{cases}$$

Because $f_{r,\epsilon}(x)>0$, for all x, we can multiply and divide by $\sqrt{f_{r,\epsilon}}$ inside the integral of the first equation and use the Cauchy–Schwarz inequality to get

$$\begin{cases} \kappa & =-\displaystyle\int g(x)f'_{r,\epsilon}(x)dx\leq\sqrt{\int g^2 f_{r,\epsilon}}\sqrt{\int\frac{\left(f'_{r,\epsilon}\right)^2}{f_{r,\epsilon}}},\\ \kappa r^2 & =\displaystyle\int g^2(x)f_{r,\epsilon}(x)dx.\end{cases}$$

Table 1. $n = 500$: $\text{mean}(\|\hat{\boldsymbol{\beta}}_{\text{opt}} - \boldsymbol{\beta}_0\|^2)/\text{mean}(\|\hat{\boldsymbol{\beta}}_{\ell_2} - \boldsymbol{\beta}_0\|^2)$ over 1,000 independent simulations

p/*n*	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Observed mean ratio	0.6924	0.7732	0.8296	0.8862	0.9264	0.9614	0.9840	0.9959	0.9997
Predicted mean ratio	0.6842	0.7626	0.8224	0.8715	0.9124	0.9460	0.9721	0.9898	0.9986
|Relative error|, %	1.2	1.4	0.8	1.7	1.5	1.6	1.2	0.6	0.1

It follows that

$$\boxed{\lim_{n\to\infty}\frac{p}{n}=\kappa\le r_p^2(\kappa)I_\epsilon\left(r_\rho(\kappa)\right).} \qquad [3]$$

We now seek a ρ to achieve this lower bound on $r_\rho^2(\kappa)I_\epsilon(r_\rho(\kappa))$.

Achieving the Lower Bound. It is clear that a good g [which is $\text{prox}(\rho_0^*)$] should saturate the Cauchy–Schwarz inequality above. Let $r_{\text{opt}}(\kappa)=\min\{r : r^2I_\epsilon(r)=\kappa\}$. A natural candidate is

$$g_{\text{opt}} = -r_{\text{opt}}^2(\kappa)\frac{f'_{r_{\text{opt}}(\kappa),\epsilon}}{f_{r_{\text{opt}}(\kappa),\epsilon}} = -r_{\text{opt}}^2(\kappa)\left[\log f_{r_{\text{opt}}(\kappa),\epsilon}\right]'.$$

It is easy to see that for this function g_{opt}, the two equations of the system are satisfied (the way we have chosen r_{opt} is of course key here). However, we need to make sure that g_{opt} is a valid choice; in other words, it needs to be the prox of a certain (convex) function.

We can do so by using ref. 14. By proposition 9.b on p. 289 in ref. 14, it suffices to establish that, for all $r>0$,

$$H_{r,\epsilon}(x) = -r^2\log f_{r,\epsilon}(x)$$

is convex and less convex than p_2. That is, there exists a convex function γ such that $H_{r,\epsilon}=p_2-\gamma$.

When ε has a log-concave density, it is well known that $f_{r,\epsilon}$ is log-concave. $H_{r,\epsilon}$ is therefore convex.

Furthermore, for a constant K,

$$H_{r,\epsilon}(x)=\frac{x^2}{2}-r^2\log\int_{-\infty}^{\infty} e^{(xy/r^2)}e^{-y^2/(2r^2)}f_\epsilon(y)dy+K.$$

It is clear that $r^2\log\int_{-\infty}^{\infty}e^{(xy/r^2)}e^{-y^2/(2r^2)}f_\epsilon(y)dy$ is convex in x. Hence, $H_{r,\epsilon}$ is less convex than p_2. Thus, g_{opt} is a prox function and a valid choice for our problem.

Determining ρ_{opt} from g_{opt}. Let us now recall another result of ref. 14. We denote the inf-convolution operation by $\star_{\text{inf}}$. More details about $\star_{\text{inf}}$ are given in *SI Text* and *Appendix: Reminders*. If γ is a (proper, closed) convex function, and $\zeta=p_2\star_{\text{inf}}\gamma$, we have (see ref. 14, p. 286)

$$\nabla\zeta=\text{prox}\left(\gamma^*\right).$$

Recall that $g_{\text{opt}}=\text{prox}(\rho_{\text{opt}}^*)=\nabla H_{r_{\text{opt}},\epsilon}$. So up to constants that do not matter, we have

$$H_{r_{\text{opt}},\epsilon}=p_2\star_{\text{inf}}\rho_{\text{opt}}.$$

It is easy to see (*SI Text*) that, for any function f,

$$f\star_{\text{inf}}p_2=p_2-(f+p_2)^*.$$

So we have $H_{r_{\text{opt}},\epsilon}=p_2-(\rho_{\text{opt}}+p_2)^*$. Now, for a proper, closed, convex function γ, we know that $\gamma^{**}=\gamma$. Hence,

$$\boxed{\rho_{\text{opt}}=\left(p_2-H_{r_{\text{opt}},\epsilon}\right)^*-p_2.}$$

Convexity of ρ_{opt}. We still need to make sure that the function ρ_{opt} we have obtained is convex. We once again appeal to ref. 14, proposition 9.b. Because $H_{r_{\text{opt}},\epsilon}$ is less convex than p_2, $p_2-H_{r_{\text{opt}},\epsilon}$ is convex. However, because $H_{r_{\text{opt}},\epsilon}$ is convex, $p_2-H_{r_{\text{opt}},\epsilon}$ is less convex than p_2. Therefore, $(p_2-H_{r_{\text{opt}},\epsilon})^*$ is more convex than p_2, which implies that ρ_{opt} is convex.

Minimality of r_{opt}. The fundamental inequality we have obtained is Eq. **3**, which says that, for any feasible r_ρ, when $p/n\to\kappa$, $\kappa\le r_\rho^2(\kappa)I_\epsilon(r_\rho(\kappa))$. Our theorem requires solving the equation $r^2I_\epsilon(r)=\kappa$. Let us study the properties of the solutions of this equation.

Let us call ξ the function such that $\xi(r)=r^2I_\epsilon(r)$. We note that $\xi(r)$ is the information of $Z+\epsilon/r$, where $Z\sim\mathcal{N}(0,1)$ and independent of ε. Hence $\xi(r)\to0$ as $r\to0$ and $\xi(r)\to1$ as $r\to\infty$. This is easily established using the information inequality $I(X+Y)\le I(X)$ when X and Y are independent (I is the Fisher information; see, e.g., ref. 16). As a matter of fact, $\xi(r)=r^2I(rZ+\epsilon)\le r^2I(\epsilon)\to0$ as $r\to0$. However, $\xi(r)=I(Z+\epsilon/r)\le I(Z)=1$. Finally, as $r\to\infty$, it is clear that $\xi(r)\to I(Z)=1$ (see *SI Text* for details). Using the fact that ξ is continuous (see, e.g., ref. 11), we see that the equation $\xi(r)=\kappa$ has at least one solution for all $\kappa\in[0,1)$.

Let us recall that we defined our solution as $r_{\text{opt}}(\kappa)=\min\{r : r^2I_\epsilon(r)=\kappa\}$. Denote $r_1=\inf\{r : r^2I_\epsilon(r)=\kappa\}$. We need to show two facts to guarantee optimality of r_{opt}: (i) the inf is really a min; (ii) $r\ge r_{\text{opt}}(\kappa)$, for all feasible r's (i.e., r's such that $r^2I_\epsilon(r)\ge\kappa$). ($i$) follows easily from the continuity of ξ and lower bounds on $\xi(r)$ detailed in *SI Text*.

We now show that, for all feasible r's, $r\ge r_{\text{opt}}(\kappa)$. Suppose it is not the case. Then, there exists r_2, which is asymptotically feasible and $r_2<r_{\text{opt}}(\kappa)$. Because r_2 is asymptotically feasible, $\xi(r_2)\ge\kappa$. Clearly, $\xi(r_2)>\kappa$, for otherwise we would have $\xi(r_2)=\kappa$ with $r_2<r_{\text{opt}}(\kappa)$, which would violate the definition of $r_{\text{opt}}(\kappa)$. Now recall that $\xi(0)=0$. By continuity of ξ, because $\xi(r_2)>\kappa$, there exists $r_3\in(0,r_2)$ such that $\xi(r_3)=\kappa$. However, $r_3<r_2<r_{\text{opt}}(\kappa)$, which violates the definition of $r_{\text{opt}}(\kappa)$.

Appendix: Reminders

Convex Analysis Reminders. ***Inf-convolution and conjugation.*** Recall the definition of the inf-convolution (see, e.g., ref. 13, p. 34). If f and g are two functions,

$$f\star_{\text{inf}}g(x)=\inf_y[f(x-y)+g(y)].$$

Recall also that the (Fenchel–Legendre) conjugate of a function f is

$$f^*(x)=\sup_y[xy-f(y)].$$

A standard result says that, when f is closed, proper, and convex, $(f^*)^*=f$ (ref. 13, theorem 12.2).

We also need a simple remark about relation between inf-convolution and conjugation. Recall that $p_2(x)=x^2/2$. Then (we give details in *SI Text*),

$$\boxed{f\star_{\text{inf}}p_2=p_2-(f+p_2)^*.}$$

The prox function. The prox function seems to have been introduced in convex analysis by Moreau (refs. 13, pp. 339–340, and 14). The definition follows. We assume that f is a proper, closed, convex function. Then, when f: $\mathbb{R} \to \mathbb{R}$, and $c > 0$ is a scalar,

$$\operatorname{prox}_1(f)(x) = \operatorname{prox}(f)(x) = \operatorname{argmin}_y \frac{(x-y)^2}{2} + f(y),$$

$$\operatorname{prox}_c(f)(x) = \operatorname{prox}(cf)(x) = \operatorname{argmin}_y \frac{(x-y)^2}{2c} + f(y),$$

$$\operatorname{prox}(f)(x) = (\mathrm{Id} + \partial f)^{-1}(x).$$

In the last equation, ∂f is in general a subdifferential of f. Although this could be a multivalued mapping when f is not differentiable, the prox is indeed well defined as a (single-valued) function.

A fundamental result connecting prox mapping and conjugation is the equality

$$\operatorname{prox}(f)(x) + \operatorname{prox}\left(f^*\right)(x) = x.$$

An Alternative Representation for ψ_{opt}. We give an alternative representation for ψ_{opt}. Recall that we had $g_{\mathrm{opt}} = \operatorname{prox}(\rho^*_{\mathrm{opt}}) = -r^2_{\mathrm{opt}} f'_{r_{\mathrm{opt}},\epsilon}/f_{r_{\mathrm{opt}},\epsilon}$. Using $\operatorname{prox}(\rho_{\mathrm{opt}}) = \mathrm{Id} - \operatorname{prox}(\rho^*_{\mathrm{opt}})$, we see that $\operatorname{prox}(\rho_{\mathrm{opt}}) = \mathrm{Id} + r^2_{\mathrm{opt}} f'_{r_{\mathrm{opt}},\epsilon}/f_{r_{\mathrm{opt}},\epsilon}$. In the case in which ρ_{opt} is differentiable, this gives immediately

$$\psi_{\mathrm{opt}}\left(x + r^2_{\mathrm{opt}} \frac{f'_{r_{\mathrm{opt}},\epsilon}(x)}{f_{r_{\mathrm{opt}},\epsilon}(x)}\right) = -r^2_{\mathrm{opt}} \frac{f'_{r_{\mathrm{opt}},\epsilon}(x)}{f_{r_{\mathrm{opt}},\epsilon}(x)}.$$

Because ψ_{opt} is defined up to a positive scaling factor,

$$\tilde{\psi}_{\mathrm{opt}}\left(x + r^2_{\mathrm{opt}} \frac{f'_{r_{\mathrm{opt}},\epsilon}(x)}{f_{r_{\mathrm{opt}},\epsilon}(x)}\right) = -\frac{f'_{r_{\mathrm{opt}},\epsilon}(x)}{f_{r_{\mathrm{opt}},\epsilon}(x)}$$

is an equally valid choice.

Interestingly, for $\kappa = \lim p/n$ near 0, r_{opt} will be near zero too, and the previous equation shows that $\tilde{\psi}_{\mathrm{opt}}$ will be essentially $-f'_\epsilon/f_\epsilon$, corresponding to the objective derived from maximum-likelihood theory.

ACKNOWLEDGMENTS. We thank the referees for their interesting and constructive comments. We thank Prof. Steven Portnoy for several thought-provoking remarks and discussions. 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). B.Y. gratefully acknowledges support from NSF Grants SES-0835531 (Cyber-Enabled Discovery and Innovation), DMS-1107000, and CCF-0939370.

1. Fisher RA (1922) On the mathematical foundations of theoretical statistics. *Philos Trans R Soc Lond A* 222:309–368.
2. Cramér H (1946) *Mathematical Methods of Statistics*. Princeton Mathematical Series (Princeton Univ Press, Princeton), Vol 9.
3. Hájek J (1972) Local asymptotic minimax and admissibility in estimation. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (University of California, Berkeley, California, 1970/1971)* (Univ of California Press, Berkeley, CA), Vol I: Theory of Statistics, pp 175–194.
4. Le Cam L (1970) On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Ann Math Stat* 41:802–828.
5. Bühlmann P, van de Geer S (2011) *Statistics for High-Dimensional Data. Methods, Theory and Applications*. Springer Series in Statistics (Springer, Heidelberg).
6. El Karoui N, Bean D, Bickel P, Lim C, Yu B (2012) On robust regression with high-dimensional predictors. *Proc Natl Acad Sci USA* 110:14557–14562.
7. Huber PJ (1973) Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann Stat* 1:799–821.
8. 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.
9. Ibragimov IA (1956) On the composition of unimodal distributions. *Teor Veroyatnost i Primenen* 1:283–288.
10. Prékopa A (1973) On logarithmic concave measures and functions. *Acta Sci Math (Szeged)* 34:335–343.
11. Costa MHM (1985) A new entropy power inequality. *IEEE Trans Inf Theory* 31:751–760.
12. Guo D, Wu Y, Shamai S, Verdú S (2011) Estimation in Gaussian noise: Properties of the minimum mean-square error. *IEEE Trans Inf Theory* 57:2371–2385.
13. Rockafellar RT (1997) *Convex analysis*. Princeton Landmarks in Mathematics (Princeton Univ Press, Princeton), Reprint of the 1970 original, Princeton Paperbacks.
14. Moreau J-J (1965) Proximité et dualité dans un espace hilbertien. *Bull Soc Math France* 93:273–299. French.
15. Karlin S (1968) *Total positivity* (Stanford Univ Press, Stanford, CA), Vol I.
16. Stam AJ (1959) Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Inf Control* 2:101–112.