Optimal objective function in high-dimensional regression

Derek Bean [*], **Peter Bickel** [*] , **Noureddine El Karoui** [*] , **and Bin Yu** [*]

[*] University of California, Berkeley

Submitted to Proceedings of the National Academy of Sciences of the United States of America

We consider, for the first time in the modern setting of high-dimensional statistics, the classic problem of optimizing the objective function in regression. We propose an algorithm to compute this optimal objective function that takes into account the dimensionality of the problem.

robust regression | prox function | high-dimensional statistics

Abbreviations: EPE, expected prediction error; $\overset{\mathcal{L}}{=}$, equal in law; LAD, least absolute deviations

In this article we study a fundamental statistical problem: how to optimally pick the objective to be minimized in a parametric regression when we have information about the error distribution.

The classical answer to the problem we posed at the beginning, maximum likelihood, was given by Fisher (5) in the specific case of multinomial models and then at succeeding levels of generality by Cramér (3), Hájek (7) and above all Le Cam (11). For instance, for p fixed or $p/n \to 0$ fast enough, least squares is optimal for Gaussian errors while 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, though as we explain below, this assumption should be relaxable to a situation where the distribution of the predictors satisfy certain concentration properties for quadratic forms.

We carry out our analysis in a regime which 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. Since in most fields of application, situations where p as well as n is 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 ((1)). 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 (4), we showed [1] that this price need not be paid if p/n stays bounded away from 1. We review the main theoretical results from this previous paper in Result 1 below. From a practical standpoint, some of our key findings were:

1. 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 (8);
2. linear combinations of regression coefficients are unbiased and still asymptotically Gaussian at rate[2] $1/\sqrt{n}$.

This article contains three main parts: Section "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 $\liminf p/n$ is not 0. We also limit ourselves to the case where $\limsup p/n < 1$. As far as we know, all the very large body of work developed in robust regression (following (8)) 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

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

where ρ is a function from $\mathbb{R}$ to $\mathbb{R}$, which we will assume throughout is convex[3]. Furthermore, we consider a linear regression model

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

where β_0 ($\in \mathbb{R}^p$) is unknown and $\{\epsilon_i\}_{i=1}^n$ are random errors. Throughout, we will assume that $\{\epsilon_i\}_{i=1}^n$ are independent of X_i. 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 two questions. 1)What choice of ρ minimizes the asymptotic error for estimating an individual regression coefficient (or a given linear form in β_0)? 2) What choice of ρ minimizes the asymptotic prediction error for a new observation (X_{new}, Y_{new}) given the training data? The answers to 1) and 2) 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 (see (4)), we found heuristically the following.

Reserved for Publication Footnotes

[1] heuristically

[2] under conditions depending on the model and the linear combination; see details below.

[3] the properties of $\widehat{\beta}$ naturally depend on ρ - this dependence will be made clear later

www.pnas.org/cgi/doi/10.1073/pnas.0709640104

Result 1 (El Karoui et al.(4)). *Suppose X_i are i.i.d $\mathcal{N}(0,\Sigma)$, with Σ positive definite. Suppose $Y_i=\epsilon_i+X_i'\beta_0$, $\epsilon_i's$ are i.i.d, independent of X_i, $\beta_0\in\mathbb{R}^p$ is deterministic, and $n\geq p$. Call*

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

Then we have the stochastic representation

$$\begin{aligned}\widehat{\beta}(\rho;\beta_0,\Sigma)&\overset{\mathcal{L}}{=}\beta_0+\Sigma^{-1/2}\widehat{\beta}(\rho;0,\text{Id}_p)\;,\\ &\overset{\mathcal{L}}{=}\beta_0+\|\widehat{\beta}(\rho;0,\text{Id}_p)\|\Sigma^{-1/2}u\;,\end{aligned}$$

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

Let us call $r_\rho(p,n)=\|\widehat{\beta}(\rho;0,\text{Id}_p)\|$. As p and n tend to infinity, while $p\leq n$ 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

$$\left\{\begin{aligned}\mathbf{E}\left([prox_c(\rho)]'(\hat{z}_\epsilon)\right)&=1-\kappa\;,\\ \mathbf{E}\left([\hat{z}_\epsilon-prox_c(\rho)(\hat{z}_\epsilon)]^2\right)&=\kappa r_\rho^2(\kappa)\;,\end{aligned}\right. \tag*{[S]}$$

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 the section "Explanations" and the Appendix. 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 (4) 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 assumptions. System **[S]** is the basis of our analysis.

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

$$\text{var}\left(v'\widehat{\beta}(\rho;\beta_0,\Sigma)\right)=\frac{v'\Sigma^{-1}v}{p}\mathbf{E}\left(\|\widehat{\beta}(\rho;0,\text{Id}_p)\|^2\right)\;.$$

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

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

$$EPE=\mathbf{E}\left((Y_{new}-X_{new}'\widehat{\beta})^2\right)=\sigma_\epsilon^2+\mathbf{E}\left(\|\widehat{\beta}(\rho;0,\text{Id}_p)\|^2\right)\;.$$

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

Main result. We propose an algorithm to determine the asymptotically optimal objective function to use in robust regression. Just as in the classical case, it 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 (see (9; 13)) 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 $r\geq 0$. It is known that when ϵ has a density, $r^2I_\epsilon(r)$ is continuous in r (see (2), where it is explained that $I_\epsilon(\sqrt{r})$ is differentiable or see (6)).

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 $r_\rho\geq r_{opt}(\kappa)$, where $r_{opt}(\kappa)=\min\{r:r^2I_\epsilon(r)=\kappa\}$. Furthermore, $r_{opt}(\kappa)$ is the solution of System* **[S]** *when $\rho=\rho_{opt}$, and ρ_{opt} is the convex function*

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

(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 the Appendix.

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

1. Solve for r the equation
$$r^2I_\epsilon(r)=p/n\;. \tag*{[1]}$$
Define $r_{\text{opt}}=\min\{r:r^2I_\epsilon(r)=p/n\}$.
2. Use the objective function[4]
$$\rho_{\text{opt}}=\left(p_2+r_{\text{opt}}^2\log(\phi_{r_{\text{opt}}}\star f_\epsilon)\right)^*-p_2\;. \tag*{[2]}$$

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.

It should also be noted that at least numerically, computing ρ_{opt} is not very hard. Similarly solving Equation **[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 $\widehat{\beta}(\rho;0,\text{Id}_p)$. In the rest of the paper, we use the lighter notation

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

The dependence of $\widehat{\beta}_\rho$ 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.

[4] note that any $\lambda\rho_{\text{opt}}+\xi$, where λ and ξ are real-valued with $\lambda>0$, yields the same solution for $\widehat{\beta}$

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, $\left(p_2 + r_{\text{opt}}^2 \log(\phi_{r_{\text{opt}}} \star f_\epsilon)\right)^*$ 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 the SI.

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 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 (4), 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.

Though 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

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

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 $\widehat{\beta}_{\text{opt}}$ 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 $\widehat{\beta}_{\text{opt}}$. For this, we can use a well-known result in convex analysis, that says that for a convex function h (under regularity conditions) $(h^*)' = (h')^{-1}$ (see (14), Corollary 23.5.1).

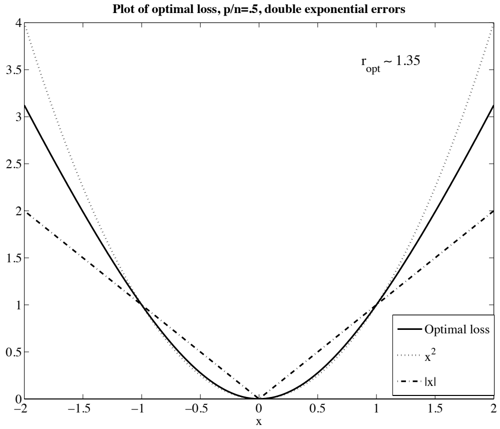

Fig. 1. $p/n = .5$: comparison of ρ_{opt} (optimal objective) to l_2 and l_1. r_{opt} is the solution of $r^2 I_\epsilon(r) = p/n$; for $p/n = .5$, $r_{opt} \simeq 1.35$

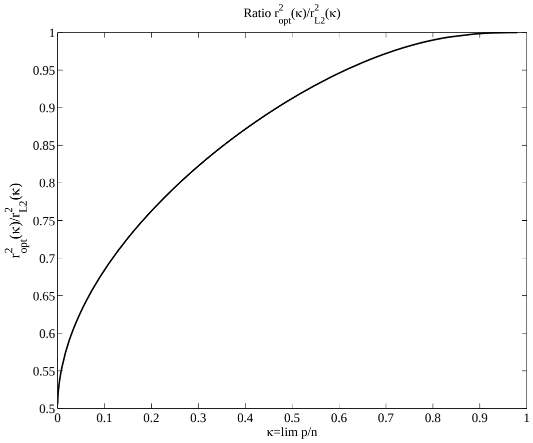

Fig. 2. Ratio $r_{\text{opt}}^2(\kappa)/r_{\ell_2}^2(\kappa)$ for double exponential errors: the ratio is always less than 1, showing the superiority of the objective we propose over ℓ_2.

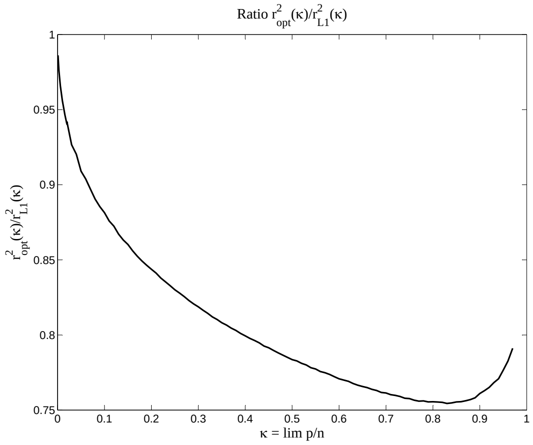

Fig. 3. Ratio $r_{\text{opt}}^2(\kappa)/r_{\ell_1}^2(\kappa)$: the ratio is always less than 1, showing the superiority of the objective we propose over ℓ_1. Naturally, the ratio goes to 1 at 0, since we know that ℓ_1 is the optimal objective when $p/n \to 0$ for double exponential errors.

We present a plot to get an intuitive feeling for how this function ρ_{opt} behaves (more can be found in the SI). Figure 1 compares ρ_{opt} to other objective functions of potential interest in the case of $p/n = .5$. All 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 $[\mathbf{S}]$ allows us to compute the asymp-

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

totic value of $\|\widehat{\beta}_\rho\|^2$, r_ρ^2, as n and p go to infinity for any convex (and sufficiently regular) ρ.

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

Comparison of $\boldsymbol{\rho_{\mathrm{opt}}}$ to $\boldsymbol{\ell_1}$ We compare r^2_{opt} to $r^2_{\ell_1}$ in Figure 3. Naturally, the ratio goes to 1 when p/n goes to 0, since 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 $\widehat{\beta}_{\mathrm{opt}}$. The table above shows $\mathbf{E}\left(r^2_{\mathrm{opt}}(p,n)\right)/\mathbf{E}\left(r^2_{\ell_2}(p,n)\right)$ over 1000 simulations when $n = 500$ for different ratios of dimensions and compares the empirical results to $r^2_{\mathrm{opt}}(\kappa)/r^2_{\ell_2}(\kappa)$, the theoretical values. We used $\beta_0 = 0$, $\Sigma = \mathrm{Id}_p$ and double exponential errors in our simulations.

In the SI, we also provide statistics concerning $\|\widehat{\beta}_{\mathrm{opt}} - \beta_0\|^2/\|\widehat{\beta}_{\ell_2} - \beta_0\|^2$ computed over 1000 simulations. We note that our predictions concerning $\|\widehat{\beta}_{\mathrm{opt}} - \beta_0\|^2$ work very well in expectation when p and n are a few 100's, even though in these dimensions, $\|\widehat{\beta}_{\mathrm{opt}} - \beta_0\|$ is not yet close to being deterministic (see SI for details - these remarks also apply to $\|\widehat{\beta}_\rho\|^2$ 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 $r_\rho(\kappa) = \lim_{n\to\infty}\|\widehat{\beta}(\rho; 0, \mathrm{Id}_p)\|$, 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 $c > 0$ $\mathrm{prox}_c(\rho) = \mathrm{prox}_1(c\rho)$ (see Appendix). 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 $\hat{z}_\epsilon = r_\rho Z + \epsilon$,

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

Now it is clear that if we replace ρ by $\lambda\rho$, $\lambda > 0$, we do not change $\widehat{\beta}_\rho$. 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

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

We can now rephrase this system using the fundamental equality (see (12) and the Appendix) $\mathrm{prox}(\rho)+\mathrm{prox}(\rho^*) = x$, where ρ^* is the (Fenchel-Legendre) conjugate of ρ. It becomes

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

Prox mappings are known to belong to subdifferentials of convex functions and to be contractive (see (12), p.292, Corollaire 10.c). Let us call $g = \mathrm{prox}(\rho_0^*)$ and recall that $f_{r,\epsilon}$ denotes the density of $\hat{z}_\epsilon = rZ + \epsilon$. Since g is contractive, $|g(x)/x| \leq 1$ as $|x| \to \infty$. Since $f_{r,\epsilon}$ is a log-concave density (as a convolution of two log-concave densities - see (9) and (13)) with support $\mathbb{R}$, it goes to zero at infinity exponentially fast (see (10), 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)

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

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

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

It follows that

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

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 $\mathrm{prox}(\rho_0^*)$) should saturate the Cauchy-Schwarz inequality above. Let $r_{\mathrm{opt}}(\kappa) = \min\{r : r^2 I_\epsilon(r) = \kappa\}$. A natural candidate is

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

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 (12). By Proposition 9.b p. 289 in (12), 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,

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

It is clear that $r^2\log\int_{-\infty}^{\infty}\mathrm{e}^{(xy/r^2)}\mathrm{e}^{-y^2/(2r^2)} f_\epsilon(y)dy$ 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 (12). We denote the inf-convolution operation by $\star_{\text{inf}}$. More details about $\star_{\text{inf}}$ are given in the SI and Appendix. If γ is a (proper, closed) convex function, and $\zeta = p_2 \star_{\text{inf}} \gamma$, we have (see (12), p.286)

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

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 (see SI) 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,

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

Convexity of $\boldsymbol{\rho_{\text{opt}}}$ We still need to make sure that the function ρ_{opt} we have obtained is convex. We once again appeal to (12), Proposition 9.b. Since $H_{r_{\text{opt}},\epsilon}$ is less convex than p_2, $p_2 - H_{r_{\text{opt}},\epsilon}$ is convex. However, since $H_{r_{\text{opt}},\epsilon}$ 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 Equation [**3**], which says that for any feasible r_ρ, when $p/n \to \kappa$, $\kappa \leq r^2_\rho(\kappa) I_\epsilon(r_\rho(\kappa))$. Our theorem requires solving the equation $r^2 I_\epsilon(r) = \kappa$. Let us study the properties of the solutions of this equation.

Let us call ξ the function such that $\xi(r) = r^2 I_\epsilon(r)$. We note that $\xi(r)$ is the information of $Z + \epsilon/r$, where $Z \sim \mathcal{N}(0,1)$ and independent of ϵ. Hence $\xi(r) \to 0$ as $r \to 0$ and $\xi(r) \to 1$ as $r \to \infty$. This is easily established using the information inequality $I(X+Y) \leq I(X)$ when X and Y are independent (I is the Fisher information; see e.g (15)). As a matter of fact, $\xi(r) = r^2 I(rZ+\epsilon) \leq r^2 I(\epsilon) \to 0$ as $r \to 0$. On the other hand, $\xi(r) = I(Z + \epsilon/r) \leq I(Z) = 1$. Finally, as $r \to \infty$, it is clear that $\xi(r) \to I(Z) = 1$ (see SI for details). Using the fact that ξ is continuous (see e.g (2)), 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 $r_{\text{opt}}(\kappa) = \min\{r : r^2 I_\epsilon(r) = \kappa\}$. Denote $r_1 = \inf\{r : r^2 I_\epsilon(r) = \kappa\}$. We need to show two facts to guarantee optimality of r_{opt}: 1) the inf is really a min. 2) $r \geq r_{\text{opt}}(\kappa)$, for all feasible r's (i.e r's such that $r^2 I_\epsilon(r) \geq \kappa$).

1) follows easily from the continuity of ξ and lower bounds on $\xi(r)$ detailed in the SI.

We now show that for all feasible r's, $r \geq r_{\text{opt}}(\kappa)$. Suppose it is not the case. Then, there exists r_2, which is asymptotically feasible and $r_2 < r_{\text{opt}}(\kappa)$. Since r_2 is asymptotically feasible, $\xi(r_2) \geq \kappa$. 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, since $\xi(r_2) > \kappa$, there exists $r_3 \in (0, r_2)$ such that $\xi(r_3) = \kappa$. But $r_3 < r_2 < r_{\text{opt}}(\kappa)$, 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 (14), p.34). If f and g are two functions,

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

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, $(f^*)^* = f$ ((14), 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 the SI),

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

The prox function. The prox function seems to have been introduced in convex analysis by Moreau (see (12), (14), pp.339-340). 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,

$$\begin{aligned}
\text{prox}_1(f)(x) &= \text{prox}(f)(x) = \text{argmin}_y \frac{(x-y)^2}{2} + f(y) \ , \\
\text{prox}_c(f)(x) &= \text{prox}(cf)(x) = \text{argmin}_y \frac{(x-y)^2}{2c} + f(y) \ , \\
\text{prox}(f)(x) &= (\text{Id} + \partial f)^{-1}(x) \ .
\end{aligned}$$

In the last equation, ∂f is in general a subdifferential of f. Though this could be a multi-valued 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

$$\text{prox}(f)(x) + \text{prox}(f^*)(x) = x \ .$$

An alternative representation for ψ_{opt}

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

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

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

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

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 $\widetilde{\psi}_{\text{opt}}$ will be essentially $-f'_\epsilon/f_\epsilon$, the objective derived from maximum likelihood theory.

ACKNOWLEDGMENTS. Bean gratefully acknowledges support from NSF grant DMS-0636667 (VIGRE). Bickel gratefully acknowledges support from NSF grant DMS-0907362. El Karoui gratefully acknowledges support from an Alfred P. Sloan Research Fellowship and NSF grant DMS-0847647 (CAREER). Yu gratefully acknowledges support from NSF Grants SES-0835531 (CDI), DMS-1107000 and CCF-0939370.

References

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