Supporting Information

Bean et al. 10.1073/pnas.1307845110

SI Text

Inferential Questions Regarding $v'\beta_0$. We mentioned the possibility of deriving a confidence interval for $v'\beta_0$ when v is a given deterministic vector. This is of interest, for instance, to get a confidence statement regarding one coordinate. Recall the stochastic representation

$$\hat{\beta} = \beta_0 + \Sigma^{-1/2} u r_\rho(p,n),$$

where u is uniform on the unit sphere of radius one, $r_\rho(p,n)$ is independent of u and is such that $r_\rho(p,n) = \|\hat{\beta}(\rho;0,\mathrm{Id}_p)\|$, and Σ is the covariance of the predictors. Therefore,

$$v'\hat{\beta} = v'\beta_0 + r_\rho(p,n) v'\Sigma^{-1/2} u.$$

Now $v'\Sigma^{-1/2}u$ is approximately $\mathcal{N}(0, v'\Sigma^{-1}v/p)$ as p tends to infinity. Similarly, $r_\rho(p,n)$ has a deterministic limit in our asymptotics, which we called $r_\rho(\kappa)$. So as p tends to infinity,

$$\frac{v'\hat{\beta} - v'\beta_0}{r_\rho(\kappa)\sqrt{v'\Sigma^{-1}v/p}} \Rightarrow \mathcal{N}(0,1).$$

Using the fact that our predictors are Gaussian, we know (1) that

$$\frac{v'\hat{\Sigma}^{-1}v}{v'\Sigma^{-1}v} \overset{\mathcal{L}}{=} \frac{n}{\chi^2_{n-p}}.$$

In the asymptotics we are considering, i.e., $p/n \to \kappa < 1$, while n and p tend to infinity, so

$$\frac{n}{\chi^2_{n-p}} = \frac{n}{n-p}\left[1 + \mathrm{O}_P\left(n^{-1/2}\right)\right].$$

So $\frac{n-p}{n} v'\hat{\Sigma}^{-1}v$ is a "$\sqrt{n}$-consistent" estimator of $v'\Sigma^{-1}v$ [in the sense that the ratio minus 1 is $\mathrm{O}_P(n^{-1/2})$].

An asymptotically 95% confidence interval for $v'\beta_0$ is therefore (in our asymptotics)

$$v'\hat{\beta} \pm 1.96 r_\rho(\kappa) p^{-1/2} \sqrt{\frac{n-p}{n} v'\hat{\Sigma}^{-1} v}.$$

When the distribution of the errors is known, which is needed to compute the optimal objective function, $r_\rho(\kappa)$ can be obtained by solving the system **S** in the main text. If the distribution of the errors is not known, at ρ given, leave-one-out methods can be used to yield an estimate of $r_\rho(\kappa)$ from the data. We do not discuss this issue further as it is not really relevant to the main theme of this paper.

The Case of Gaussian Errors. We give a full derivation of the following fact.

Fact. In the setting of independent identically distributed (i.i.d) Gaussian predictors, among all convex objectives, l_2 is optimal in regression when the errors are Gaussian.

Let us now justify this assertion.

In the Gaussian case, it is clear that $\phi_{r_{\mathrm{opt}}} \star f_\epsilon$ is a Gaussian density. We denote by p_2 the function such that $p_2(x) = x^2/2$. Hence, $(p_2 + r^2_{\mathrm{opt}} \log(\phi_{r_{\mathrm{opt}}} \star f_\epsilon))^*$ is a multiple of p_2 (up to an additive constant). It is easy to check that carrying out the algorithm, our proposal for ρ is

$$\rho_{\mathrm{opt}}(x) = \frac{x^2}{2}\left(\frac{p/n}{1-p/n}\right) - K.$$

Because this is, up to centering and scaling, $x^2/2$, we see that when the errors are Gaussian, l_2 objective is optimal (among all convex objectives) in any dimension.

A Lower Bound on $r^2_{\mathrm{opt}}(\kappa)$. Let us call ξ the function such that $\xi(r) = r^2 I(rZ + \epsilon)$. Since ξ is the information of $Z + \epsilon/r$, Stam's inequality (2) gives

$$\frac{1}{\xi(r)} \geq 1 + \frac{1}{r^2 I_\epsilon},$$

where I_ϵ is the information of ε. Therefore, if r is such that $\xi(r) = 1 - \eta$, we see that

$$r^2 \geq \left(\frac{1}{\eta} - 1\right)\frac{1}{I_\epsilon}.$$

So we see that

$$r^2_{\mathrm{opt}}(\kappa) \geq \frac{\kappa}{1-\kappa}\frac{1}{I_\epsilon}.$$

In particular, it tends to ∞ as p/n tends to 1.

Using suboptimality of least squares, we also see that $r^2_{\mathrm{opt}}(\kappa) \leq \frac{\kappa}{1-\kappa}\sigma^2_\epsilon$.

About $\xi(r)$ When $r \to \infty$. Recall that ξ is such that $\xi(r) = I(Z + \epsilon/r)$, where $Z \sim \mathcal{N}(0,1)$ and ε is independent of Z and has a log-concave density. In particular, ε has a variance (see ref. 3, p. 332).

It is well known, that for any random variable Y with a variance, $I(Y) \geq 1/\mathrm{var}(Y)$. So $\xi(r) \geq \frac{1}{1+\sigma^2_\epsilon/r^2}$. However, using Stam's inequality (2), we have $I(rZ + \epsilon) \leq 1$.

So we see that as $r \to \infty$, $\xi(r) \to 1$.

Simple computations also result in the fact that $\xi(r) = 1 - \frac{\sigma^2_\epsilon}{r^2} + \mathrm{o}(1/r^2)$ as $r \to \infty$ (see ref. 4 for more details). Using this fact, one can show that

$$\frac{r^2_{\mathrm{opt}}(\kappa)}{r^2_{\ell_2}(\kappa)} \to 1$$

as κ tends to 1.

More Details on $\|\hat{\boldsymbol{\beta}}_{\mathrm{opt}} - \boldsymbol{\beta}_0\|^2 / \|\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}_0\|^2$. Our simulations were performed in the case where $\beta_0 = 0$ and $\Sigma = \mathrm{Id}_p$, with double-exponential errors. We chose $n = 500$ and did 1,000 independent simulations.

Table S1 shows the 2.5 and 97.5 percentiles for $\|\hat{\beta}_{\mathrm{opt}}\|^2 / \|\hat{\beta}_{ols}\|^2$ over our 1,000 experiments.

We note that approximating $\mathbf{E}(r_\rho(p,n))$ by $r_\rho(\kappa)$ for $\kappa = p/n$ works very well even in this moderately sized setting, but larger problems are needed for $\|\hat{\beta}\|^2$ to become almost deterministic. We found that across values of p/n, the ratio $\sqrt{\mathrm{var}(r_\rho(p,n))}/\mathbf{E}(r_\rho(p,n))$ was about 10% in our simulations, for all of the estimators we looked at.

For the sake of completeness, we also present a brief comparison between the empirical behavior of $\hat{\beta}_{\text{opt}}$ and that of $\hat{\beta}_{\ell_1}$. The results are in Fig. S1. The maximal relative error in comparing empirical to theoretical values is 1%, achieved for $p/n = 0.5$.

Inf-Convolution and Conjugation. Recall that $p_2(x) = x^2/2$. We have

$$f \star_{\inf} p_2(x) = \inf_y \left[\frac{(x-y)^2}{2} + f(y) \right]$$
$$= \frac{x^2}{2} + \inf_y \left(-xy + \frac{y^2}{2} + f(y) \right)$$
$$= x^2/2 - \sup_y \left(xy - \frac{y^2}{2} - f(y) \right)$$
$$= \frac{x^2}{2} - (p_2 + f)^*(x).$$

It follows that

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

Plots of ρ_{opt}. Fig. S2 compares ρ_{opt} to other loss functions of potential interest, when $p/n = 0.2$. Fig. S3 does the same when $p/n = 0.5$. Fig. S4 plots ψ_{opt} when $p/n = 0.5$. In the plots, all of the objective functions are normalized to take value 0 at 0 and 1 at 1.

We have used different normalizations from the ones discussed in the main text to make visual comparisons easier. Therefore, some of the analytic comparisons made in the main text do not apply to the figures, because these comparisons are sensitive to the choice of centering and scaling.

The Question of Intercept. The normality assumption, and the invariance properties it entails, greatly simplify our arguments for obtaining inferential results. We show here how they also allow us to handle the issue of lack of intercept in the model described in the main text. The main conclusion of the brief discussion that follows is that we can take care of this issue by recentering predictors and responses before doing the regression.

Let us assume that X_i's are i.i.d $\mathcal{N}(\mu, \Sigma)$. μ and Σ depend on p, but the coming argument is almost entirely finite dimensional, so let us not mention p to make notations lighter. Let us assume that $Y_i = \epsilon_i + X_i'\beta_0$, where ϵ_i does not necessarily have mean 0. ϵ_i's are assumed to be independent of X_i's.

Call $\overline{Y}$ the sample mean of Y and $\hat{\mu}_X$ the sample mean of X_i's. Let us consider

$$\hat{\beta} = \text{argmin}_\beta \sum_{i=1}^n \rho\left(\left[Y_i - \overline{Y} \right] - (X_i - \hat{\mu}_X)'\beta \right) \; .$$

Of course, $Y_i - \overline{Y} = \epsilon_i - \overline{\epsilon} + (X_i - \hat{\mu}_X)'\beta_0$. Hence,

$$\left[Y_i - \overline{Y} \right] - (X_i - \hat{\mu}_X)'\beta = \epsilon_i - \overline{\epsilon} + (X_i - \hat{\mu}_X)'(\beta - \beta_0) \; .$$

If $Z_i = \Sigma^{-1/2}(X_i - \mu)$, $Z_i - \hat{\mu}_Z = \Sigma^{-1/2}(X_i - \hat{\mu}_X)$. Call $X - \overline{X}$ the $n \times p$ matrix whose ith row is $(X_i - \hat{\mu}_X)'$. Note that it is of course equal to $(Z - \overline{Z})\Sigma^{1/2}$, where Z_i is $\mathcal{N}(0, \text{Id}_p)$ and is the ith row of the $n \times p$ matrix Z.

Clearly, if 1_n is an $n \times 1$ vector with all entries equal to 1,

$$X - \overline{X} = \left(\text{Id}_n - \frac{1_n 1_n'}{n} \right) X = \left(\text{Id}_n - \frac{1_n 1_n'}{n} \right) Z \Sigma^{1/2}$$
$$= (Z - \overline{Z})\Sigma^{1/2}.$$

So, if e_i is the ith canonical basis vector in $\mathbb{R}^n$,

$$\hat{\beta} = \text{argmin}_\beta \sum_{i=1}^n \rho\left(\left[\epsilon_i - \overline{\epsilon} \right] - e_i'\left(\text{Id}_n - \frac{1_n 1_n'}{n} \right) Z \Sigma^{1/2}(\beta - \beta_0) \right).$$

We use $\stackrel{\mathcal{L}}{=}$ to denote equality in law. Let $w(\beta) = \Sigma^{1/2}(\beta - \beta_0)$. We note that this is a 1–1 reparametrization. Call $\hat{\beta}(\rho; 0, \text{Id}_p)$ the solution of our M-estimation problem when $\beta_0 = 0$ and $\Sigma = \text{Id}_p$. Note that $w(\hat{\beta}) \stackrel{\mathcal{L}}{=} \hat{\beta}(\rho; 0, \text{Id}_p)$, because $n \geq p + 1$, and therefore $\left(\text{Id}_n - \frac{1_n 1_n'}{n} \right) Z$ is of rank p with probability 1.

Because, for any $p \times p$ orthogonal matrix $\mathcal{O}$,

$$Z \stackrel{\mathcal{L}}{=} Z\mathcal{O},$$

we have, conditional on $\{\epsilon_i\}_{i=1}^n$, which we assume independent of X and therefore Z,

$$\hat{\beta}(\rho; 0, \text{Id}_p) \big| \{\epsilon_i\}_{i=1}^n \stackrel{\mathcal{L}}{=} \mathcal{O}'\hat{\beta}(\rho; 0, \text{Id}_p) \big| \{\epsilon_i\}_{i=1}^n.$$

Therefore, by a standard invariance argument,

$$\frac{\hat{\beta}(\rho; 0, \text{Id}_p)}{\|\hat{\beta}(\rho; 0, \text{Id}_p)\|_2} \Big| \{\epsilon_i\}_{i=1}^n \stackrel{\mathcal{L}}{=} u,$$

where u is uniform on the unit sphere in $\mathbb{R}^p$. Also, because the law of u does not depend on $\epsilon_i's$, we finally have

$$\hat{\beta} - \beta_0 \stackrel{\mathcal{L}}{=} \left\| \hat{\beta}(\rho; 0, \text{Id}_p) \right\|_2 \Sigma^{-1/2} u \;\; ,$$

and the two random variables in the product are independent.

If v is a fixed vector of norm 1, $\sqrt{p} v' u$ is nearly $\mathcal{N}(0, 1)$ in high dimension. (Of course, its exact distribution is known but these details are not needed here.)

So we conclude that, if W represents a $\mathcal{N}(0, 1)$ random variable,

$$\sqrt{p} \; \frac{v'(\hat{\beta} - \beta_0)}{\sqrt{v'\Sigma^{-1}v}} \Rightarrow \left\| \hat{\beta}(\rho; 0, \text{Id}_p) \right\|_2 W.$$

This argument shows that $\sqrt{p} \; \frac{v'(\hat{\beta} - \beta_0)}{\sqrt{v'\Sigma^{-1}v}}$ is asymptotically a scaled mixture of Gaussians. We expect the analysis of $\|\hat{\beta}(\rho; 0, \text{Id}_p)\|_2$ in this case to be similar the one we undertook in ref. 5, and we expect that it will again be asymptotically deterministic. That will give asymptotic normality of $\sqrt{p} \; \frac{v'(\hat{\beta} - \beta_0)}{\sqrt{v'\Sigma^{-1}v}}$. We have confirmed this fact in limited simulations.

We further note that when the errors have a symmetric distribution and the objective function ρ is symmetric, the system controlling $\|\hat{\beta}(\rho; 0, \text{Id}_p)\|_2$ in the case of recentered X_i's and Y_i's appears to be the same as the one in *Result 1*. Details concerning this situation will appear elsewhere as the derivation is long, technical, and tedious.

We have also investigated the situation in which we allow the intercept to be estimated directly in the M-estimation problem. Under further assumptions on $\mathbf{E}(X_i)$, we have obtained a characterization of this intercept and of $\|\hat{\beta} - \beta_0\|$ through a system of three nonlinear equations in three unknowns. This result will be presented elsewhere. The optimization of the objective function in that setting is naturally made harder by the presence of a third equation. We have not yet carried this task out.

Gaussian Design and Measure of Performance. In the case of Gaussian predictors, our stochastic representation gave

$$\hat{\beta}(\rho)-\beta_0 \stackrel{\mathcal{L}}{=} \|\hat{\beta}(\rho;0,\mathrm{Id}_p)\|_2 \Sigma^{-1/2} u,$$

where u is uniform on the unit sphere and the two random variables in the product are independent. Of course, ρ intervenes only in the distribution of $\|\hat{\beta}(\rho;0,\mathrm{Id}_p)\|_2$. So for any norm $\|\cdot\|_N$,

$$\mathbf{E}\left(\|\hat{\beta}(\rho)-\beta_0\|_N\right)=\mathbf{E}\left(\|\hat{\beta}(\rho;0,\mathrm{Id}_p)\|_2\right)\mathbf{E}\left(\|\Sigma^{-1/2}u\|_N\right).$$

Hence, the relative efficiency of the estimators obtained for different $\rho's$ should be the same regardless of the norm chosen to measure their accuracy, as the performance of the estimators in whatever norm is chosen effectively only depends on $\|\hat{\beta}(\rho;0,\mathrm{Id}_p)\|_2$.

In other words, the loss function we propose will lead to improvements of the regression estimators in any norm chosen by the user and not only in ℓ_2 norm.

1. Eaton ML (1983) *Multivariate Statistics: A Vector Space Approach*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics (Wiley, New York).
2. Stam AJ (1959) Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Inf Control* 2:101–112.
3. Karlin S (1968) *Total Positivity* (Stanford Univ Press, Stanford, CA), Vol I.
4. 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.
5. El Karoui N, Bean D, Bickel P, Lim C, Yu B (2012) On robust regression with high-dimensional predictors. *Proc Natl Acad Sci USA*, 10.1073/pnas.1307842110.

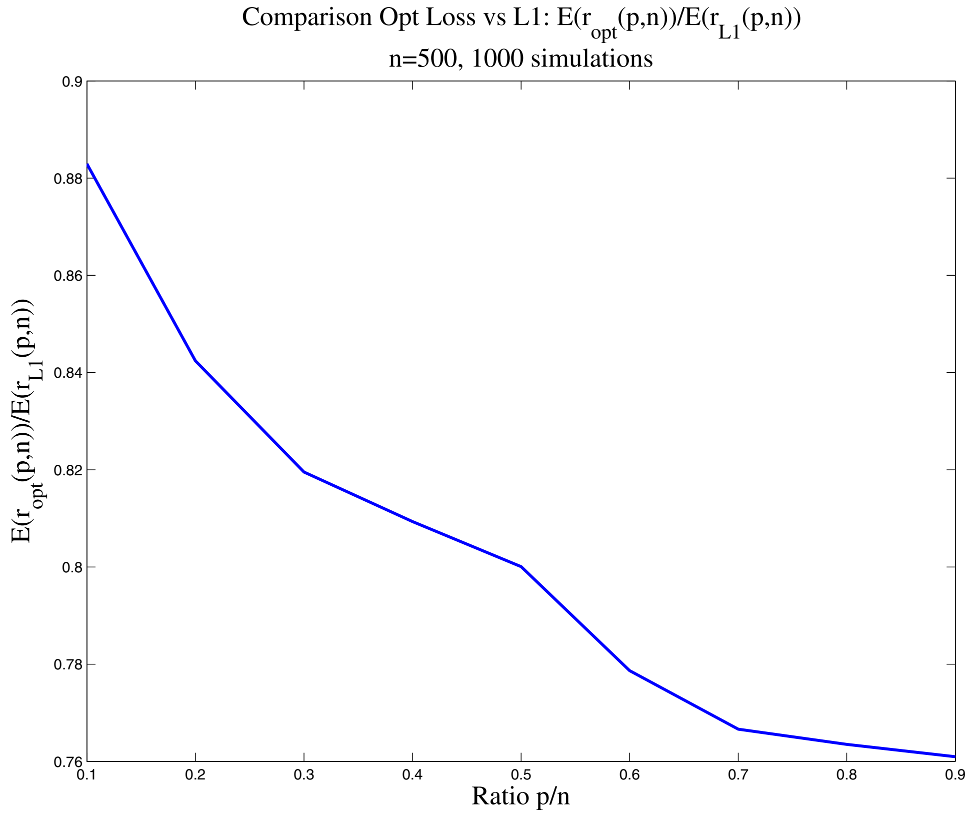

Fig. S1. The picture represents the ratio $\mathbf{E}(r_{\mathrm{opt}}(p,n))/\mathbf{E}(r_{\ell_1}(p,n))$ for different p's and $n=500$. The expectations are calculated numerically from 1,000 independent simulations.

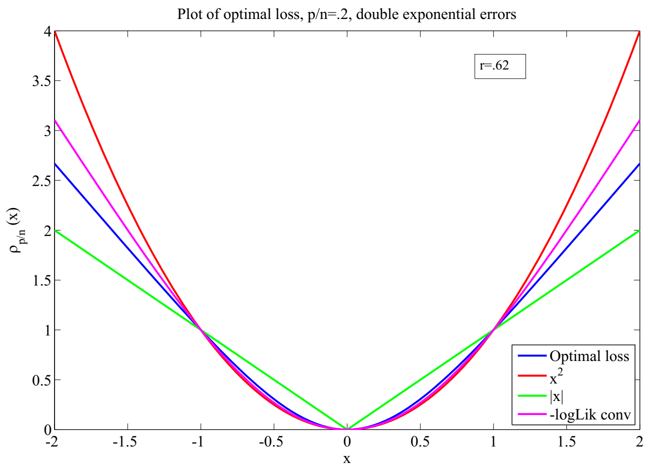

Fig. S2. $p/n = 0.2$: comparison of ρ_{opt} (optimal loss) to l_2, l_1, and $-\log f_{r_{\mathrm{opt}},\epsilon}$ (called −logLik conv in the legend of the plot). r_{opt} is the solution of $r^2 I_\epsilon(r) = p/n$; for $p/n = 0.2$, $r_{\mathrm{opt}} \simeq 0.62$.

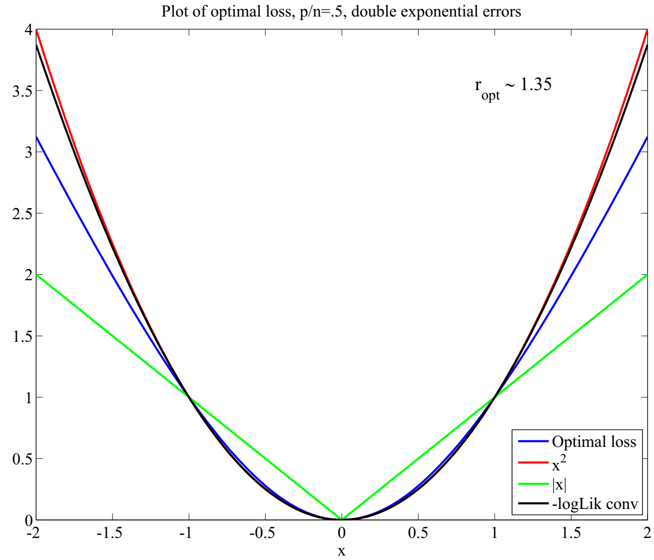

Fig. S3. $p/n = 0.5$: comparison of ρ_{opt} (optimal loss) to l_2, l_1, and $-\log f_{r_{\mathrm{opt}},\epsilon}$ (called −logLik conv in the legend of the plot). r_{opt} is the solution of $r^2 I_\epsilon(r) = p/n$; for $p/n = 0.5$, $r_{opt} \simeq 1.35$.

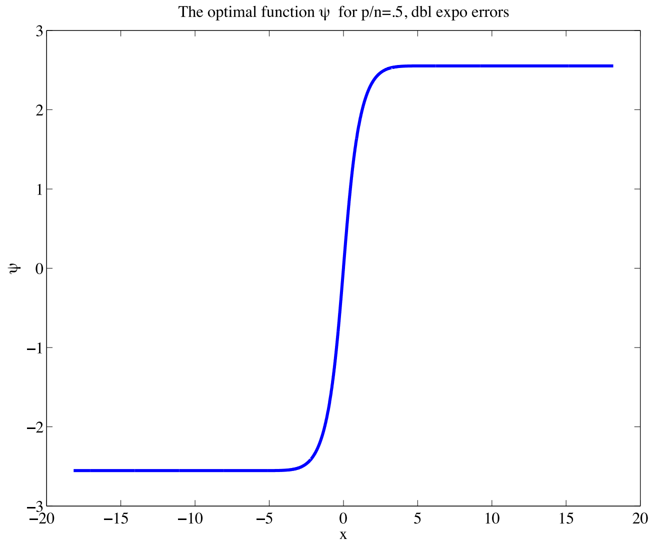

Fig. S4. $p/n=0.5$: representation of $\psi_{\mathrm{opt}}=\rho'_{\mathrm{opt}}$ for double-exponential errors. The normalization is the same as above. Numerically, $\lim_{x\to\infty}\psi(x)\simeq 2.55$.

Table S1. Case n = 500: statistics of the distribution of $\|\hat{\beta}_{\mathrm{opt}}\|^2/\|\hat{\beta}_{\ell_2}\|^2$ 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
2.5%-tile	0.4781	0.6061	0.7009	0.7725	0.8365	0.8925	0.9463	0.9824	0.9972
97.5%-tile	0.9675	0.9674	0.9792	0.9906	0.9974	1.0058	1.0077	1.0039	1.0013