On high-dimensional robust regression

How to pick the loss in high-dimensional regression?

Noureddine El Karoui
joint with Bean, Bickel, Lim, Yu

Department of Statistics
UC, Berkeley

Stanford MMDS
July 10th, 2012
Consider linear regression model:

\[ Y_i = X_i' \beta_0 + \epsilon_i \quad i = 1, \ldots, n. \]

Here \( Y_i \in \mathbb{R}, \ X_i \in \mathbb{R}^p, \ \beta_0 \in \mathbb{R}^p \) and \( \epsilon_i \in \mathbb{R}. \)

- **Aim:** estimate \( \beta_0. \)
- **Setting:** \( X_i \)'s vectors of predictors. \( \epsilon_i \)'s noise.
- **Standard method:** (say \( p < n \)): pick \( \hat{\beta} \) as

\[ \hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho(Y_i - X_i' \beta), \quad \text{where} \ \rho \ \text{is a function}. \]

**Question:** how to pick \( \rho? \)
How to pick $\rho$?

Very classical question. Much work on this starting with Fisher in 30’s. Very nice work in the late 60’s, 70’s, 80’s on properties of these estimators. Contributors include: Relles, Huber (’72), Portnoy (’84-85), Mammen (’91), Yohai, Bickel, etc...

Short answer: in low dimension, if $f_\epsilon$ is density of $\epsilon$, $\epsilon$ i.i.d, pick

$$\rho = - \log f_\epsilon.$$

Remarkable fact: independent of design matrix, $X$. 
An example: double exponential errors

\( \epsilon_i \)'s double exponential, i.e \( f_{\epsilon}(x) = \exp(-|x|)/2 \).
According to classical results/intuition, \( l_1 \) should be optimal.

Figure: \( \frac{\mathbb{E}(\|\hat{\beta}_{L_1}\|^2)}{\mathbb{E}(\|\hat{\beta}_{OLS}\|^2)} \) and prediction, double exponential errors, 1000 simulations

\[ \frac{\mathbb{E}(\|\hat{\beta}_{L_1}\|^2)}{\mathbb{E}(\|\hat{\beta}_{OLS}\|^2)} \]
A proposition for $\rho$

Let $p_2(x) = x^2/2$. Suppose $\epsilon$ has log-concave density, $f_\epsilon$. For sake of argument, assume $f_\epsilon$ known. For reasons explained later, let us try

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

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

$\phi_r$: gaussian density with variance $r^2$.

$l_\epsilon(r)$: Fisher information of $\phi_r * f_\epsilon$

$g^*(x) = \sup_y (xy - g(y))$, Fenchel-Legendre dual of $g$
Comparison $\rho_{opt}$ to $\ell_1$, double exponential errors

Figure: $\mathbb{E} \left( \| \hat{\beta}_{opt} - \beta_0 \|^2 \right) / \mathbb{E} \left( \| \hat{\beta}_1 - \beta_0 \|^2 \right)$, double exponential errors.
Comparison $\rho_{opt}$ to $\ell_2$, double exponential errors

Figure: $E \left( \| \hat{\beta}_{opt} - \beta_0 \|^2 \right) / E \left( \| \hat{\beta}_{OLS} - \beta_0 \|^2 \right)$, double exponential errors.
Aim of talk

- Understand these pictures/phenomena
- Caveat: optimality now sensitive to design. Will get back to key properties of design
- Also: bootstrap appears problematic in this context
- Many interesting statistical phenomena at play

Why work under $p/n$ not close to 0?
Plan

1. Computation of risk of robust regression estimators
2. Inferential questions
3. Optimization with respect to loss function
4. Penalized case: risk computation and optimality in the $\ell_2$-penalized case
Suppose $p/n \to \kappa \in (0, 1)$. Temporarily, $X_i \overset{iid}{\sim} \mathcal{N}(0, \text{Id}_p)$.

**Proposition**

*Under regularity conditions on $\{\epsilon_i\}$ and $\rho$, $\|\hat{\beta} - \beta_0\|$ is asymptotically deterministic. Call $r_\rho(\kappa)$ its limit and $\hat{z}_\epsilon = \epsilon + r_\rho(\kappa)Z$, where $Z \sim \mathcal{N}(0, 1)$, independent of $\epsilon$. For a $c$ deterministic, we have*

\[
\begin{align*}
\mathbb{E} (\text{prox}_c(\rho)(\hat{z}_\epsilon)) &= 1 - \kappa, \\
\kappa r_\rho^2(\kappa) &= \mathbb{E} (\hat{z}_\epsilon - \text{prox}_c(\rho)(\hat{z}_\epsilon))^2).
\end{align*}
\]

By definition, (Moreau '65), for convex function $f$,

\[
\text{prox}_c(f)(x) = \arg\min_y \left(f(y) + \frac{1}{2c}(x - y)^2\right).
\]

Much more can be said: elliptical models, heteroskedastic $\epsilon_i$'s, weighted robust regression, no need for normality of $X_i$ etc... Approach can handle penalized versions.
Call $R_i = Y_i - \hat{\beta}'X_i$, the $i$-th residual. In the asymptotic limit,

$$R_i \overset{\mathcal{L}}{=} \text{prox}_c(\rho)(\epsilon_i + r_\rho(\kappa)Z_i)$$

where $Z_i \sim \mathcal{N}(0, 1)$ independent of $\epsilon_i$. Somewhat complicated relationship between $\rho$, distribution of $\epsilon_i$ and distribution of $R_i$. Very different from classical setting of $p/n$ close to 0.
Suppose $X_i = \lambda_i X_i$, where $X_i$ is $\mathcal{N}(0, \text{Id}_p)$, $\lambda_i$ random variables independent of $X_i$.

$\|\hat{\beta} - \beta_0\|$ still asymptotically deterministic, limit denoted by $r_\rho(\kappa)$.

**Proposition**

Let us now 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

$$\begin{cases}
\lim_{n \to \infty} \sum_{i=1}^n \frac{\mathbb{E}\left(\left[\text{prox}_{c\lambda_i^2(\rho)}\right]'(\hat{z}_{\epsilon}(i))\right]}{n} = 1 - \kappa, \\
\lim_{n \to \infty} \sum_{i=1}^n \frac{\mathbb{E}\left(\lambda_i^{-2}[\hat{z}_{\epsilon}(i) - \text{prox}_{c\lambda_i^2(\rho)}(\hat{z}_{\epsilon}(i))]^2\right]}{n} = \kappa r_\rho^2(\kappa),
\end{cases}$$

(S1)

where $c$ again positive deterministic constant determined from above system.
How proof and heuristics work?

Key elements

- concentration of quadratic forms in $X_i$; consequence: geometry of dataset influences crucially result. No universality.
- leave-one-out ideas.
- martingale ideas
- connection with ideas in random matrix theory and convex analysis

Surprise, in particular in connection to $l_1$ regression: it is a random matrix problem!
When $X_i$ are i.i.d $\mathcal{N}(0, \Sigma)$, then

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

where $u$ unif of sphere of radius 1 in $\mathbb{R}^p$, independent of $\|\hat{\beta}(\rho; 0, \text{Id}_p)\|$. Consequences:

- easy to handle case $\beta_0 \neq 0$ and $\Sigma \neq \text{Id}_p$.
- elliptical setting works similarly
- can do inference on $v'\beta_0$, any $v$ given. (Surprise to experts.)
- not complicated to include intercept (several manners to deal with that)
- side note: bootstrap

Fact: quality of inference depends only on $\|\hat{\beta}(\rho; 0, \text{Id}_p)\|$; its limit $r_\rho(\kappa)$ characterized before.

Natural to optimize $r_\rho(\kappa)$ over $\rho$. 
Suppose wish to measure quality of estimator as

$$E \left( \| \hat{\beta} - \beta_0 \|_q \right), q \neq 2 \text{ for instance.}$$

Stochastic representation yields immediately

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

Hence optimizing $r_{\rho}(\kappa)$ yields asymptotically optimal performance in any $\ell_q$ norm, not only $\ell_2$. 

Noureddine El Karoui joint with Bean, Bickel, Lim, Yu

On high-dimensional robust regression
Recall system: if $\hat{z}_\epsilon = \epsilon + r_\rho(\kappa)Z$, with $Z \sim \mathcal{N}(0,1)$,

\[
\begin{align*}
\mathbb{E}\left(\text{prox}_{c}(\rho)(\hat{z}_\epsilon)\right) &= 1 - \kappa, \\
\kappa r_\rho^2(\kappa) &= \mathbb{E}\left([\hat{z}_\epsilon - \text{prox}_{c}(\rho)(\hat{z}_\epsilon)]^2\right) .
\end{align*}
\]

Possible to optimize $r_\rho(\kappa)$ over $\rho$!
Write problem as feasibility problem in $r$

Use Moreau’s fundamental prox-identity:

$$x = \prox_1(\rho)(x) + \prox_1(\rho^*)(x).$$

to rewrite system. Natural variable: $\prox_1(\rho^*)$

Cauchy-Schwarz yields lower bound on possible values of $r^2 l_\epsilon(r)$, where $l_\epsilon(r)$ is Fisher information of $\epsilon + rZ$

Come up with good $\prox_1(\rho^*)$ for which lower bound is achieved.

Go from optimal $\prox_1(\rho^*)$ to optimal $\rho$

Following this strategy, we get that, if $p_2(x) = x^2/2$, if $-\log f_\epsilon$ convex,

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

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

Noureddine El Karoui joint with Bean, Bickel, Lim, Yu

On high-dimensional robust regression
Further remarks on optimal loss

- For Gaussian errors, $\ell_2$ still optimal
- As $p/n \rightarrow 1$, performance of $\ell_2$ becomes optimal
- However, limit of optimal loss not $\ell_2$
- Also, $\rho_{opt}$ proposed above convex
Plot for $p/n = .5$, double exponential errors

**Figure:** Some “natural” objective functions
What about the case of penalized regression, i.e:

\[ \hat{\beta} = \text{argmin}_\beta \rho(Y_i - \beta'X_i) + \tau P(\beta). \]

Can handle that, too. At this point, need

- \( P(\beta) = \sum_{i=1}^{p} f_i(\beta_i), \)
- \( \text{cov}(X_i) = \text{Id}_p \)

Possible to characterize the limit \( \hat{\beta} - \beta_0. \)
See also work on Lasso of Donoho-Maleki-Montanari, Bayati-Montanari. Approach is different.
System for $\|\hat{\beta} - \beta_0\|$, elliptical setting

$\|\hat{\beta} - \beta_0\|$ asymptotically deterministic. Call $\hat{\beta}_{(i)}$ leave-one-out estimate of $\beta$ and $\tilde{r}_{i,(i)} = \epsilon_i - (\hat{\beta}_{(i)} - \beta_0)'X_i$. Below $\nu(\tau)$ and $c_\tau$ are unknowns. Call

$$Z_k \overset{\mathcal{L}}{=} \mathcal{N} \left( \beta_0(k), \frac{1}{n\nu(\tau)^2} \mathbb{E} \left( \frac{\left[ \text{prox}_{c_\tau \lambda_i^2(\rho)} \left( \tilde{r}_{i,(i)} \right) - \tilde{r}_{i,(i)} \right]^2}{\lambda_i^2} \right) \right).$$

We have asymptotically, when $p/n$ has finite limit,

$$\begin{cases} 
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left( \left[ \text{prox}_{c_\tau \lambda_i^2(\rho)} \right]' \left( \tilde{r}_{i,(i)} \right) \right) = 1 - \nu(\tau), \\
\lim_{n \to \infty} \frac{p}{n} \frac{1}{p} \sum_{k=1}^{p} \mathbb{E} \left( \left[ \text{prox}_{K_\tau} \left( f_k \right) \right]' \left( Z_k \right) \right) = \nu(\tau), \\
\forall 1 \leq k \leq p, \text{prox}_{K_\tau} \left( f_k \right) \left[ Z_k \right] \overset{\mathcal{L}}{=} \hat{\beta}_k.
\end{cases}$$

with $K_\tau = \frac{\tau c_\tau}{n\nu(\tau)}$

Last $p$ equations relate asymptotic value of $\|\hat{\beta} - \beta_0\|^2$ to $\nu(\tau)$ and $c_\tau$. Yields a system of 3 equations in three unknowns $\|\hat{\beta} - \beta_0\|$, $\nu(\tau)$ and $c_\tau$. 

Noureddine El Karoui joint with Bean, Bickel, Lim, Yu — On high-dimensional robust regression
Optimization when $P(\beta) = \|\beta\|^2/2$

Suppose want to minimize $\|\hat{\beta} - \beta_0\|_2$ for Tikhonov penalty. Then optimal loss is again in family found above. However, $r_{opt}$ changes. Now it is

$$r_{opt} = \min \left\{ r : r^2 = \frac{\|\beta_0\|^2}{1 + \frac{n}{p} I_{\epsilon}(r) \|\beta_0\|^2} \right\}.$$ 

(Also, $\tau_{opt}/n = p/n - r_{opt}^2 I_{\epsilon}(r_{opt})$)
Saw interplay between loss function and error distribution in high-dimensional robust regression

Optimal loss computable in high-dimension

It is not maximum-likelihood

Inference is possible in our context

Problems with the bootstrap (not touched in much details here)

Can do penalized regression

Optimal loss “canonical” as it is also optimal in the Tikhonov regularized context.