High-dimensional regression:

How to pick the objective function in high-dimension

Derek Bean

UC Berkeley

March 11, 2013

Joint work with Noureddine El Karoui, Peter Bickel, Chingwhay Lim, and Bin Yu

Notation.

Standard linear model. Observe n pairs (X_i, Y_i):

$$Y_i = X_i^T \beta_0 + \epsilon_i.$$

- Errors $\epsilon_i \overset{iid}{\sim} f_\epsilon$
- $\dim(X_i) = dim(\beta_0) = p$

M-estimates.

$$\widehat{\beta}_\rho = \underset{\beta}{\text{argmin}} \sum_i \rho(Y_i - X_i^T \beta)$$

- ρ - "objective function", "loss function"

Classical theory: low-dimension.
Relles (1968); Huber (1973); Portnoy (1985)

Behavior of $\widehat{\beta} - \beta_0$:

$v \in \mathbb{R}^p$ set of p weights

- $v^T \widehat{\beta}$ unbiased for $v^T \beta_0$, asym. normal
  Variance:

$$\left[v^T \left(X^T X \right)^{-1} v \right] \times r^2(\rho, f_\epsilon)$$

- **Key:** p grows slowly with $n \Rightarrow p/n \approx 0$.

Given f_ϵ, compute $r^2 \Rightarrow$ possible to compare estimates

Best estimate: minimize r^2 over ρ.

Best objective function in low-dimension.

Given f_ϵ, $r^2(\rho, f_\epsilon)$ minimized by:

$$\rho_{opt} = -\log f_\epsilon$$

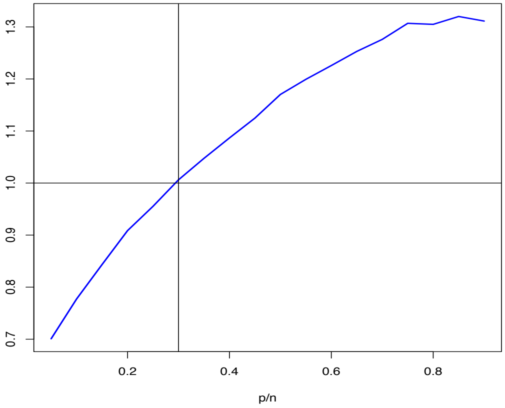
Well known: maximum likelihood estimate (MLE).

Example MLEs:

1. Normal errors: **Least squares** (LS): $\rho_{opt}(x) = x^2$.
2. Double exponential errors: **Least absolute deviations** (LAD): $\rho_{opt}(x) = |x|$. (Robust)

Surprising simulations!

D.E. errors: $\mathbb{E}\|\widehat{\beta}_{LAD} - \beta_0\|_2^2 / \mathbb{E}\|\widehat{\beta}_{LS} - \beta_0\|_2^2$



1000 samples, 1000 simulations.

M-estimates in high-dimension.
PNAS: El Karoui et. al. 2012, to appear

Assume:

- $p/n \to \kappa \in (0,1)$
- $X_i \stackrel{iid}{\sim} \mathcal{N}(0, I_p)$

Then: for set of p weights v, $\|v\|_2 = 1$:

1. $v^T\widehat{\beta}$ unbiased for $v^T\beta_0$, asym. normal.
2. Variance:

$$p^{-1} \times r^2(\rho, f_\epsilon; \kappa)$$

Can characterize r^2 (complicated!)

Best estimate: given f_ϵ AND κ, minimize r^2 across ρ

Optimal M-estimates in high-dimension

PNAS: Bean et. al. 2012, to appear

Key results: given error density f_ϵ,

1. For each dimension $p/n \approx \kappa$ there exists $r_{opt}(\kappa)$ such that $r(\rho, f_\epsilon; \kappa) \geq r_{opt}(\kappa)$ for all ρ
   - Can characterize $r_{opt}(\kappa)$
2. When f_ϵ is *log-concave*, r_{opt} is achieved by an "optimal loss function" ρ_{opt}.

Details of ρ_{opt}.

Let $f_{r,\epsilon} = \mathcal{N}(0, r^2) * f_\epsilon$.

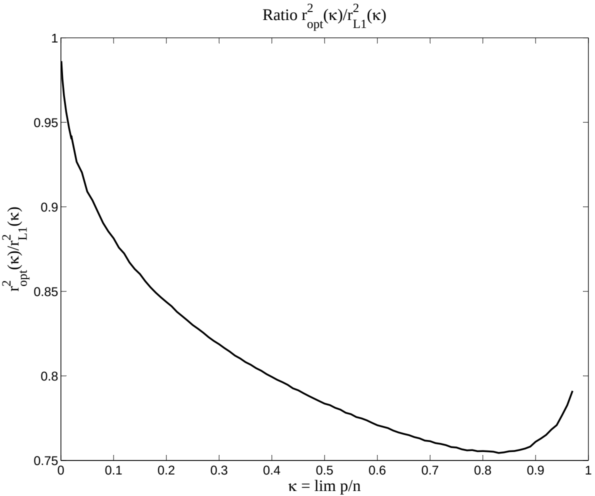
Write $r_{opt} = r_{opt}(\kappa)$. Optimal loss:

$$\rho_{opt}(x) = \left(P_2 + r_{opt}^2 \log f_{r_{opt},\epsilon}\right)^* (x) - P_2(x),$$

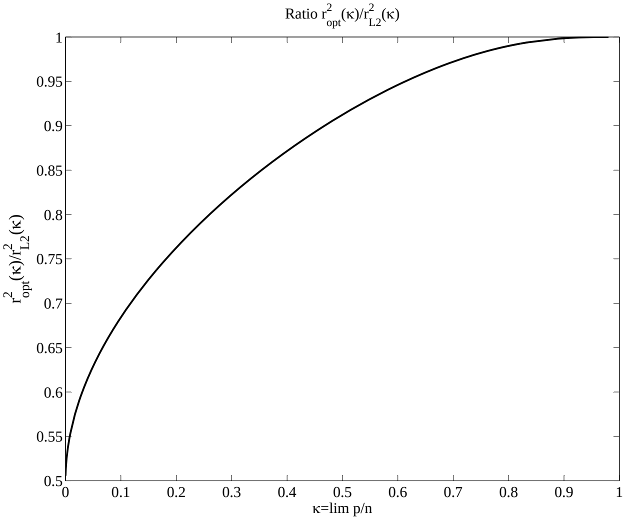
$\Rightarrow$ optimal objective **adaptive** to dimension!

- $P_2(x) = x^2/2$
- g^* is the *conjugate dual* of generic convex g.

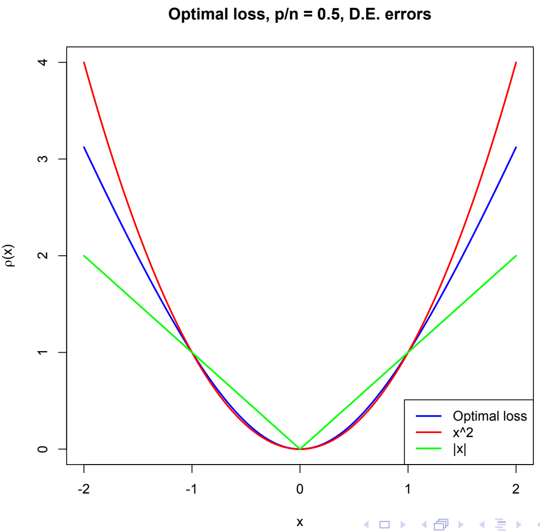
Optimal loss vs. LAD, D.E. errors



Optimal loss vs. LS, D.E. errors



Example: behavior of optimal loss function



Takeaway messages.

1. Low-dimensional intuition upended in high-dimensional setting

Takeaway messages.

1. Low-dimensional intuition upended in high-dimensional setting
2. **Can** get precise distributional behavior in high-dimensions
   - Random vs. fixed design...

Takeaway messages.

1. Low-dimensional intuition upended in high-dimensional setting
2. **Can** get precise distributional behavior in high-dimensions
   - Random vs. fixed design...
3. **Can** optimize the loss in high-dimensions
   - A new family of dimension-adaptive loss functions

Takeaway messages.

1. Low-dimensional intuition upended in high-dimensional setting
2. **Can** get precise distributional behavior in high-dimensions
   - Random vs. fixed design...
3. **Can** optimize the loss in high-dimensions
   - A new family of dimension-adaptive loss functions
4. (Not presented) **Extensions** to penalized estimates
   - E.g. LASSO, ridge-type estimates