15.094J: Robust Modeling, Optimization, Computation

Lecture 22: Robust Statistics

Outline

1. Robust Regression
2. Support Vector Machines
3. The impact of Robustness
4. Conclusions

Robust Regression

- Given data (y_i, x_i), $y_i \in \mathbb{R}$, $x_i \in \mathbb{R}^m$, $i = 1, 2, \ldots, n$.
- Robust L_p Regression optimization problem:

$$\min_{\boldsymbol{\beta}, \beta_0} \max_{\boldsymbol{\Delta} X \in \mathcal{N}} \|y - (X + \boldsymbol{\Delta} X)\boldsymbol{\beta} - \beta_0 \mathbf{1}\|_p. \tag{1}$$

- $y = (y_1, \ldots, y_n)'$, $X = (x_1, \ldots, x_n)'$, and $\mathbf{1} = (1, \ldots, 1)'$
- $\mathcal{N}$ is the uncertainty set for $\boldsymbol{\Delta} X$.

Matrix Norms and Uncertainty Sets

- Norm $\| \bullet \|_{q,p}$ for an $n \times m$ matrix A:

$$\|A\|_{q,p} \equiv \sup_{x \in \mathbb{R}^m, x \neq \mathbf{0}} \frac{\|Ax\|_p}{\|x\|_q}, \; q, p \geq 1.$$

- Note that for any $x \in \mathbb{R}^m$ with $\|x\|_q = 1$, we have that $\|A\|_{q,p} = \|Ax\|_p$.
- $\mathcal{N}_1 = \{\mathbf{\Delta} X \in \mathbb{R}^{n \times m} \mid \|\mathbf{\Delta} X\|_{q,p} \leq \rho\}$.

Matrix Norms and Uncertainty Sets

- The p-Frobenius norm $\| \bullet \|_{p-F}$ of an $n \times m$ matrix A :

$$\|A\|_{p-F} \equiv \left(\sum_{i=1}^{n} \sum_{j=1}^{m} |A_{i,j}|^p \right)^{1/p}.$$

- For $p = 2$, we obtain the usual Frobenius norm.
- The dual norm of $\| \bullet \|_p$ is $\| \bullet \|_q : \frac{1}{p} + \frac{1}{q} = 1$.
- Thus, dual norm of $\| \bullet \|_p$ is $\| \bullet \|_{d(p)}$ with

$$d(p) = \frac{p}{p-1}, p \geq 1.$$

  Note $d(1) = \infty$ and $d(\infty) = 1$.
- $\mathcal{N}_2 = \{\mathbf{\Delta} X \in \mathbb{R}^{n \times m} \mid \|\mathbf{\Delta} X\|_{p-F} \leq \rho\}$.

Equivalence of Robustification and Reguralization

- Under uncertainty set $\mathcal{N}_1$, Problem (1) is equivalent to problem

$$\min_{\boldsymbol{\beta},\beta_0} \parallel y - X\boldsymbol{\beta} - \beta_0 \mathbf{1} \parallel_p + \rho \parallel \boldsymbol{\beta} \parallel_q .$$

- Under uncertainty set $\mathcal{N}_2$, Problem (1) is equivalent to problem

$$\min_{\boldsymbol{\beta},\beta_0} \parallel y - X\boldsymbol{\beta} - \beta_0 \mathbf{1} \parallel_p + \rho \parallel \boldsymbol{\beta} \parallel_{d(p)},$$

Properties on matrix norms

- Definition:

$$[f(x,p)]_j = \begin{cases} sign(x_j)\left(\frac{|x_j|}{\| x \|_p}\right)^{p-1}, & x \neq \mathbf{0}, \\ 0, & x = \mathbf{0}, \end{cases} \quad j = 1, 2, \ldots, m,$$

where $sign(x) = \begin{cases} 1, & x \geq 0, \\ -1, & x < 0. \end{cases}$

- Proposition 1: (a) $[f(x,p)]'x = \|x\|_p$, (b) $\|f(x,p)\|_{d(p)} = 1$.
- Proposition 2: $\|A\|_{d(p),p} \leq \|A\|_{p-F}$.
- Proposition 3: For $u_1 \in \mathbb{R}^n$, $u_2 \in \mathbb{R}^m$, $p, q \geq 1$, $\|u_1 u_2'\|_{q,p} = \|u_1\|_p \|u_2\|_{d(q)}$.
- Proposition 4: For $u_1 \in \mathbb{R}^n$, $u_2 \in \mathbb{R}^m$, $p \geq 1$, $\|u_1 u_2'\|_{p-F} = \|u_1\|_p \|u_2\|_p$.

Proof of Equivalence of Reguralization and Robustification

$$\|y - (X + \mathbf{\Delta}X)\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p \ = \ \|y - X\boldsymbol{\beta} - \beta_0\mathbf{1} - \mathbf{\Delta}X\boldsymbol{\beta}\|_p$$

$$\leq \ \|y - X\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p \ + \ \|\mathbf{\Delta}X\boldsymbol{\beta}\|_p.$$

We obtain the bound

$$\|\mathbf{\Delta}X\boldsymbol{\beta}\|_p \ \leq \ \|\mathbf{\Delta}X\|_{q,p} \ \|\boldsymbol{\beta}\|_q,$$

Thus, for $\|\mathbf{\Delta}X\|_{q,p} \leq \rho$, $\|\mathbf{\Delta}X\boldsymbol{\beta}\|_p \ \leq \ \rho\|\boldsymbol{\beta}\|_q$, and for any $\mathbf{\Delta}X \in \mathcal{N}_1$,

$$\|y - (X + \mathbf{\Delta}X)\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p \ \leq \ \|y - X\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p \ + \ \rho\|\boldsymbol{\beta}\|_q.$$

Proof Continued

- Define

$$\boldsymbol{\Delta} X^o = \begin{cases} -\rho \dfrac{y - X\boldsymbol{\beta} - \beta_0 \mathbf{1}}{\|y - X\boldsymbol{\beta} - \beta_0 \mathbf{1}\|_p} [f(\boldsymbol{\beta}, q)]', & y - X\boldsymbol{\beta} - \beta_0 \mathbf{1} \neq \mathbf{0}, \\ -\rho u [f(\boldsymbol{\beta}, q)]', & y - X\boldsymbol{\beta} - \beta_0 \mathbf{1} = \mathbf{0}, \end{cases}$$

- where $f(x, p) \in \mathbb{R}^m$, $x \in \mathbb{R}^m$, $p \geq 1$, $u \in \mathbb{R}^n$, with $\|u\|_p = 1$.
- For $y - X\boldsymbol{\beta} - \beta_0 \mathbf{1} \neq \mathbf{0}$:

$$\|y - (X + \boldsymbol{\Delta} X^o)\boldsymbol{\beta} - \beta_0 \mathbf{1}\|_p = \|y - X\boldsymbol{\beta} - \beta_0 \mathbf{1} - \boldsymbol{\Delta} X^o \boldsymbol{\beta}\|_p$$

$$= \left\| y - X\boldsymbol{\beta} - \beta_0 \mathbf{1} + \rho \frac{y - X\boldsymbol{\beta} - \beta_0 \mathbf{1}}{\|y - X\boldsymbol{\beta} - \beta_0 \mathbf{1}\|_p} [f(\boldsymbol{\beta}, q)]' \boldsymbol{\beta} \right\|_p$$

Proof Continued

$$= \left\| (y - X\boldsymbol{\beta} - \beta_0\mathbf{1}) \left(1 + \frac{\rho\|\boldsymbol{\beta}\|_q}{\|y - X\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p} \right) \right\|_p \quad ([f(\boldsymbol{\beta}, q)]'\boldsymbol{\beta} = \|\boldsymbol{\beta}\|_q)$$

$$= \|y - X\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p + \rho\|\boldsymbol{\beta}\|_q.$$

Note that when $y - X\boldsymbol{\beta} - \beta_0\mathbf{1} = \mathbf{0}$, $\|y - (X + \boldsymbol{\Delta}X^o)\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p = \|y - X\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p + \rho\|\boldsymbol{\beta}\|_q$ as well.

Proof Continued

- From Propositions 1, 3, we have that if $y - X\boldsymbol{\beta} - \beta_0\mathbf{1} \neq \mathbf{0}$,

$$\|\boldsymbol{\Delta}X^o\|_{q,p} = \rho \left\| \frac{y - X\boldsymbol{\beta} - \beta_0\mathbf{1}}{\|y - X\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p} \right\|_p \|f(\boldsymbol{\beta}, q)\|_{d(q)} = \rho,$$

- and if $y - X\boldsymbol{\beta} - \beta_0\mathbf{1} = \mathbf{0}$,

$$\|\boldsymbol{\Delta}X^o\|_{q,p} = \rho\|u\|_p\|f(\boldsymbol{\beta}, q)\|_{d(q)} = \rho,$$

  and thus, $\boldsymbol{\Delta}X^o \in \mathcal{N}_1$.

- Hence

$$\max_{\boldsymbol{\Delta}X \in \mathcal{N}_1} \|y - (X + \boldsymbol{\Delta}X)\boldsymbol{\beta} - \beta_0\mathbf{1}\|_p \;=\; \|y - X + \boldsymbol{\beta} - \beta_0\mathbf{1}\|_p + \rho\|\boldsymbol{\beta}\|_q$$

Support Vector Machines

- Given categorical data (y_i, x_i), $y_i \in \{1, -1\}$, $x_i \in \mathbb{R}^m$, $i \in \{1, 2, \dots, n\}$, we define the separation error $S(\boldsymbol{\beta}, \beta_0, y, X)$ of the hyperplane classifier $\boldsymbol{\beta}' x + \beta_0 = 0$, $x \in \mathbb{R}^m$, in space $\mathbb{R}^m$ by

$$S(\boldsymbol{\beta}, \beta_0, y, X) = \sum_{i=1}^{n} \max(0, 1 - y_i(\boldsymbol{\beta}' x_i + \beta_0)), \tag{2}$$

- The hyperplane which minimizes the separation error is the solution of the optimization problem

$$\min_{\boldsymbol{\beta}, \beta_0} S(\boldsymbol{\beta}, \beta_0, y, X), \tag{3}$$

which can be expressed as the linear optimization problem

$$\begin{array}{ll} \min\limits_{\boldsymbol{\beta}, \beta_0, \boldsymbol{\xi}} & \sum\limits_{i=1}^{n} \xi_i \\ \text{s.t.} & y_i(\boldsymbol{\beta}' x_i + \beta_0) \geq 1 - \xi_i, i \in \{1, 2, \dots, n\} \\ & \xi_i \geq 0, i \in \{1, 2, \dots, n\}. \end{array}$$

Robustification

- Consider uncertainty set

$$\mathcal{N}_3 = \left\{ \boldsymbol{\Delta} X \in \mathbb{R}^{n \times m} \mid \sum_{i=1}^{n} \|\boldsymbol{\Delta} x_i\|_p \leq \rho \right\}. \tag{4}$$

- The robust version of Problem (3):

$$\min_{\boldsymbol{\beta}, \beta_0} \max_{\boldsymbol{\Delta} X \in \mathcal{N}_3} S(\boldsymbol{\beta}, \beta_0, y, X + \boldsymbol{\Delta} X). \tag{5}$$

Robustification leads to Support Vector Machines

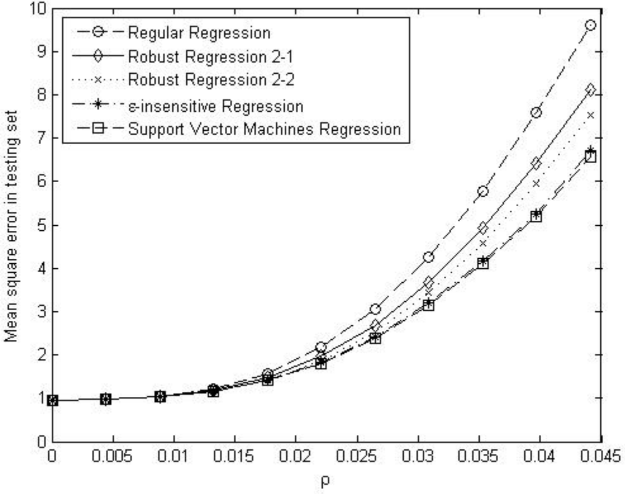
- Definition: The set of data (y_i, x_i), $i \in \{1, 2, \dots, n\}$, is separable if there exists a hyperplane $\boldsymbol{\beta}'x + \beta_0 = 0$ in $\mathbb{R}^m$, such that for any $i \in \{1, 2, \dots, n\}$, $y_i(\boldsymbol{\beta}'x_i + \beta_0) \geq 0$.
- Theorem: If the set of data (y_i, x_i), $i \in \{1, 2, \dots, n\}$ is not separable, Problem (5) is equivalent to problem

$$
\begin{aligned}
\min_{\boldsymbol{\beta}, \beta_0, \boldsymbol{\xi}} \quad & \sum_{i=1}^{n} \xi_i + \rho \parallel \boldsymbol{\beta} \parallel_{d(p)} \\
\text{s.t.} \quad & \xi_i \geq 1 - y_i(\boldsymbol{\beta}'x_i + \beta_0), i \in \{1, 2, \dots, n\} \\
& \xi_i \geq 0, i \in \{1, 2, \dots, n\},
\end{aligned}
$$

The impact of Robustness

- $x_i \sim N(\mathbf{1}, 5I_3),\ \ i = 1, \ldots, 200.$
- $y_i = \boldsymbol{\beta}' x_i + \beta_0 + r$, $\beta_0 = 1$, $\boldsymbol{\beta} = (1, -3, 1)'$, $r \sim N(0, 1)$.
- Training set (50%), testing set (50%).
- $\Delta x_i \sim \rho N(0, \mathbf{1})$.
- The procedure was repeated 30 times and the average performance of each estimate was recorded.

The impact of Robustness



The impact of Robustness

- As ρ increases, the difference in the out-of-sample performance between the robust and the respective classical estimates increases, with the robust estimates always yielding better results.
- The robust regression estimate with $p = 2$ and $q = 2$ shows better performance than the robust regression estimate with $p = 2$ and $q = 1$.

Performance on Real Data

Data set	n	m
Abalone	4177	9
Auto MPG	392	8
Comp Hard	209	7
Concrete	1030	8
Housing	506	13
Space shuttle	23	4
WPBC	46	32

- Training (50%), validation (25%), and testing (25%).
- For each ρ, prediction error on validation set was measured, and ρ with highest performance on validation was used in testing.
- Prediction errors were averaged over the 30 partitions.

Mean square error

Data set	Regular	Rob 2-1	Rob 2-2	Supp vector
Abalone	5.7430	5.6345	5.5369	5.0483*
Auto MPG	18.7928	18.6981	18.5829	12.5251*
Comp Hard	2026.0024	1965.7531	1925.1250*	2348.2914
Concrete	132.4700	131.0820	129.3135	127.0923*
Forest Fires	5525.9994	4994.8077*	5266.3994	5229.5167
Housing	39.8084	39.4257	39.0716	24.6867*
Space shuttle	0.5323	0.5225*	0.5265	0.5501
WPBC	4723.0723	4630.1946	4489.2032	4410.4599*

Conclusions

- Reguralization and Robustfication are Equivalent!
- Insights on the norms to use in regurilization.
- Support vector, method developed 3 decades ago, is one of the strongest performing classification methods.
- Robustness provides insights on the reasons.