Nonconvex Regularized Robust Regression: Properties and Algorithm

Xiaou Pan*, Qiang Sun* and Wen-Xin Zhou*

*University of California, San Diego and #University of Toronto

Nonconvex Regularized Robust Regression: Properties and Algorithm

High-Dimensional Regression: \{(y, x_i)\}_{i=1}^n are i.i.d. and obey the
linear model \(y = \beta^T x + \varepsilon\), where the dimension of \(d\) is
and \(\beta^*\) is sparse, \(\varepsilon\) is the regression error satisfying \(E[|\varepsilon|] = 0\).

Goal: Find the active set \(S = \{j: \beta_j^* \neq 0\}\) and estimate \(\beta^*_s\).

Regularized Regression: The Lasso (Tibshirani, 1996) estimator has two
downside: (i) Least squares methods are sensitive to the tails of
the error distributions; (ii) Convex penalties introduce estimation bias.

Huber Loss (Huber, 1964):
\[
\ell(x) = \begin{cases} \frac{x^2}{2} & \text{if } |x| \leq \tau, \\ \tau |x| - \frac{\tau^2}{2} & \text{if } |x| > \tau, \end{cases}
\]
where \(\tau\) is the robustification parameter.

Nonconvex Regularized Robust M-estimator:
\[
\hat{\theta} = \arg \min_{\theta(s, \alpha, \beta, \beta^*)} \left\{ L_r(\theta) + \sum p_i(\beta_i) \right\},
\]
where \(L_r(\theta) = (1/n) \sum \ell_r(y_i - \theta^T x_i), \) and \(p_i\) is a concave penalty function (eg. SCAD (Fan and Li, 2001)) and MCP (Zhang, 2010).

Conditions on Data (Condition 1): \(x\) is sub-Gaussian, \(\Sigma = \mathbb{E}[xx^T]\) is positive definite with \(\Sigma^2 = \text{maxlogdet}(\Sigma), \mathbb{E}[x^2] = \Sigma^2\) almost surely.

Conditions on Penalty (Condition 2): \(p_i(\beta) = \beta^T \lambda i(\beta)/\lambda_i\) with \(i(\beta) = i(-\beta)\) and \(p(0) = 0\); (i) \(i(\beta)\) is nondecreasing on \([0, \infty)\); (ii) \(i(\beta)\) is differentiable a.e., \(\lim_{\beta \to 0} i'(\beta) = 1\); (iv) \(i(\beta)\) is nonincreasing on \([0, \infty)\).

Conditions on Error (Condition 3): \(\varepsilon\) and \(x\) are independent. \(E(\varepsilon_i(x - \alpha))\) is a unique minimizer \(\alpha\) and \(\varepsilon([-\tau, \tau]) \leq \tau/2 > 0\).

The Role of \(\tau\): Consider \(\theta^* = \arg \min \mathbb{E} L_r(\theta)\) and the true parameter \(\theta^*\), under Condition 3, we have \(\beta^*_n = \beta^*_n + \alpha\) and \(\beta^*_n = \beta^*_n\).

Two-Stage Procedure: Tightening After Contraction

Sequential convex optimization problems:
\[
\begin{aligned}
& \min_{\theta(s, \alpha, \beta, \beta^*)} \left\{ L_r(\theta) + \sum p_i(\beta_i) \right\}, \\
& \text{(P_1)}
\end{aligned}
\]
It has two stages: contraction \((f = 1)\) and tightening \((f \geq 2)\).

Theoretical Bounds: Let \(K_{\alpha}(s, r, \theta)\) be a local curvature parameter over a local \(F\)-cone. Under Condition 2, with a proper \(\lambda\), we have
\[
\begin{align*}
\|\hat{\theta}_n - \theta^*\|_2 & \leq \frac{1 + \sqrt{2/\delta}}{K_{\alpha}(\delta^2 + 1, r, \theta^*)}, \\
\|\hat{\theta}_n - \theta^*\|_2 & \leq \frac{\sqrt{2}}{K_{\alpha}(s + 1/2, r, \theta^*)}.
\end{align*}
\]
\[
\begin{align*}
\|\hat{\theta}_n - \theta^*\|_2 & \leq \frac{\sqrt{2}}{K_{\alpha}(s + 1/2, r, \theta^*)} + \frac{\|\beta^*_n - \beta^*\|_2}{\varepsilon_n} \leq \frac{\sqrt{2}}{K_{\alpha}(s + 1/2, r, \theta^*)} + \frac{\|\beta^*_n - \beta^*\|_2}{\varepsilon_n}.
\end{align*}
\]

General Robust Losses

Conditions on Robust Loss: Suppose a general loss function \(\ell(\beta, \theta(x))\) is globally Lipschitz and locally quadratic, with \(\ell: \mathbb{R} \rightarrow [0, \infty)\) satisfying: (i) \(\ell(0) = 0\) and \(\ell'(0) = 1\); (ii) \(\ell''(x) = c_i\) and \(\ell''(x) \geq c\) for all \(x \leq c_i\); (iii) \(\ell''(x) - x \leq c_i x^2\).

Examples:

Examples of general robust loss functions and their derivatives up to order three.

Iterative Local Adaptive Majorize-Minimization (I-LAMM) Algorithm (Fan et al., 2018): At the k-th step with working parameter vector \(\theta^T(\theta^k)\), we use an isotropic quadratic function \(F(\theta, \phi, \theta^k)\)
\[
\begin{align*}
& = \mathcal{L}_r(\theta^T(\theta^k)) + L_r(\theta^T(\theta^k)) - \theta^T(\theta^k) + \frac{1}{2} \|\theta^T(\theta^k)\|_{2}^2
\end{align*}
\]
to locally majorize \(\mathcal{L}_r(\theta, \phi)\) such that \(F(\theta^k, \phi, \theta^k) \geq \mathcal{L}_r(\theta^k)\),
where \(\mathcal{L}_r(\theta, \phi)\) is a proper quadratic coefficient at the k-th update, and \(\theta^k\) is the solution to mini \(_{\theta} F(\theta, \phi, \theta^k, \theta^k) + \|\theta - \theta^k\|_{2}^2\).

1. Algorithm: \(\{\theta^k, \phi^k\}\) ← LAMM, \(\lambda^k \to \lambda^k(1)\), \(\phi^k \to \phi^k(1)\).
2. Input: \(\lambda^k, \phi^k(1), \phi^k(1)\).
3. Initialize: \(\phi^k(1) \leftarrow \max \{\phi^k(1), \gamma_n^k(1)\}\).
4. Repeat
5. \(\theta^k(n) \leftarrow T_{\lambda^k, \phi^k(n)}(\theta^k(n))\).
6. If \(F(\theta^k(n), \lambda(n)) < \mathcal{L}_r(\theta^k(n))\) then \(\phi^k(n) \leftarrow \gamma_n^k(\phi^k(n))\).
7. Until \(F(\theta^k(n), \lambda(n)) \geq \mathcal{L}_r(\theta^k(n))\)
8. Return \(\{\theta^k, \phi^k\}\)

Software and Empirical Study

Software: Available at https://github.com/XiaouPan/LAMM.

Curve: Consider Lasso, SCAD, Huber-SCAD, MCP, Huber-MCP under homoscedastic and heteroscedastic model, with errors generated from normal, Student’s t, lognormal and Pareto distributions.

Figure: Plots of ROC curves of the five methods under homoscedastic model with errors generated from four distributions: normal, Student’s t, lognormal and Pareto.

Figure: Plots of ROC curves of the five methods under heteroscedastic model with errors generated from four distributions: normal, Student’s t, lognormal and Pareto.

Figures show evident advantage of Huber-SCAD and Huber-MCP over least squares counterparts with a greater area under the curve (AUC).

UC San Diego Mathematics

xip024@ucsd.edu, qsun@stat.toronto.edu, wez243@ucsd.edu