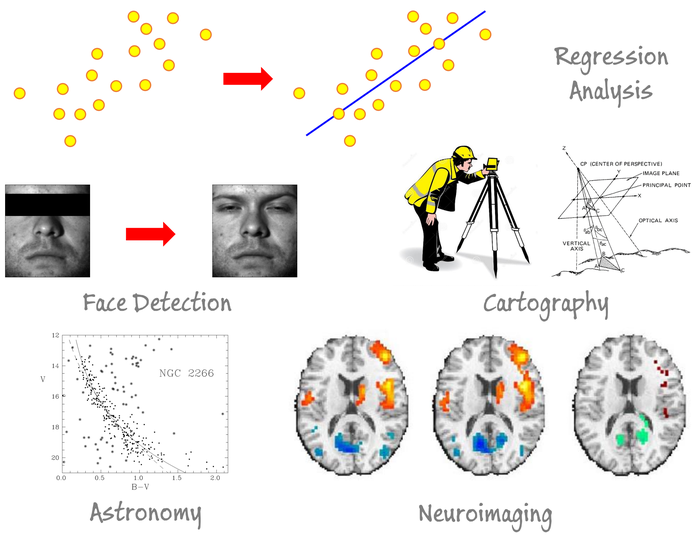
Robust Regression

Kush Bhatia, Prateek Jain and Purushottam Kar
Microsoft Research, Bengaluru, INDIA

Full Paper: http://tinyurl.com/robustreg

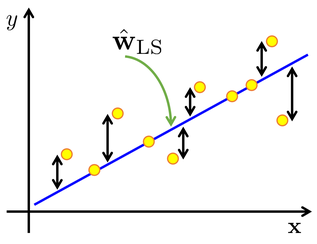
Goal: Perform statistical analysis of large-scale data in the presence of unbounded adversarial corruptions

Statistical Analysis with Corrupted Data



Ordinary Least Squares (OLS) Regression

Discovers a curve that "best fits" the entire data



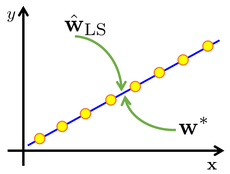
Data: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^p$

Response: $y_1, y_2, \ldots, y_n \in \mathbb{R}$

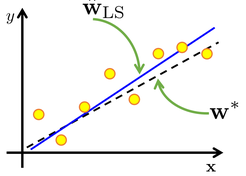
$$\hat{\mathbf{w}}_{\mathrm{LS}} = \arg\min_{\mathbf{w}} \sum_{i=1}^{n} (y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle)^2$$

Noise Model: $y_i = \langle \mathbf{w}^*, \mathbf{x}_i \rangle + b_i$

Favorable cases

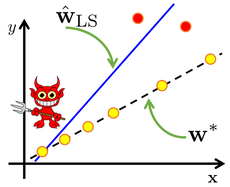


Exact recovery in noiseless case



Faithful recovery with white noise

Failure cases



Catastrophic failure under adversity

Problem Formulation: Recover the original curve in the face of a bounded number of adversarial corruptions

$$\arg\min_{\mathbf{w}, S} \sum_{i \in S} (y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle)^2, \qquad |\overline{S}| \leq \alpha \cdot n$$

Some existing approaches

Brute Force (Infeasible)

Try out all possible sets S and estimate $\mathbf{w}$ using each set

Random Selection (RANSAC) (Inefficient)

Try random sets S, estimate $\mathbf{w}$ using each and choose best

L_1 Relaxations (Expensive)

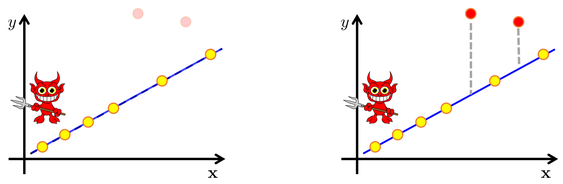
$$\hat{\mathbf{w}}_{L_1} = \arg\min_{\mathbf{w}} \sum_{i=1}^{n} |y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle|$$

KEY IDEA

Two unknowns: clean set of points S^*, original curve $\mathbf{w}^*$

Observation 1: given S^*, finding original curve $\mathbf{w}^*$ easy

Observation 2: given $\mathbf{w}^*$, finding clean points S^* easy



Proposal: can we alternate between estimating S^* and $\mathbf{w}^*$?

TORRENT

1. Start with any arbitrary curve $\mathbf{w}^0$ and set timer $t \leftarrow 0$
2. Repeat until convergence
   - i. Create *active set* S_t using points "close" to $\mathbf{w}^t$
   - ii. Create updated* curve $\mathbf{w}^{t+1}$ using active set S_t
   - iii. Increment time counter $t \leftarrow t + 1$
3. Return final curve

*4 update variants FC, GD, HYB for low-dimensional and HD for high-dimensional problems

Thresholding Operator-based Robust RegrEssioN meThod

Theoretical Guarantees

Theorem: TORRENT can resist even an all powerful adversary that corrupts after viewing data and the curve (if < 1/65th fraction of data is corrupted)

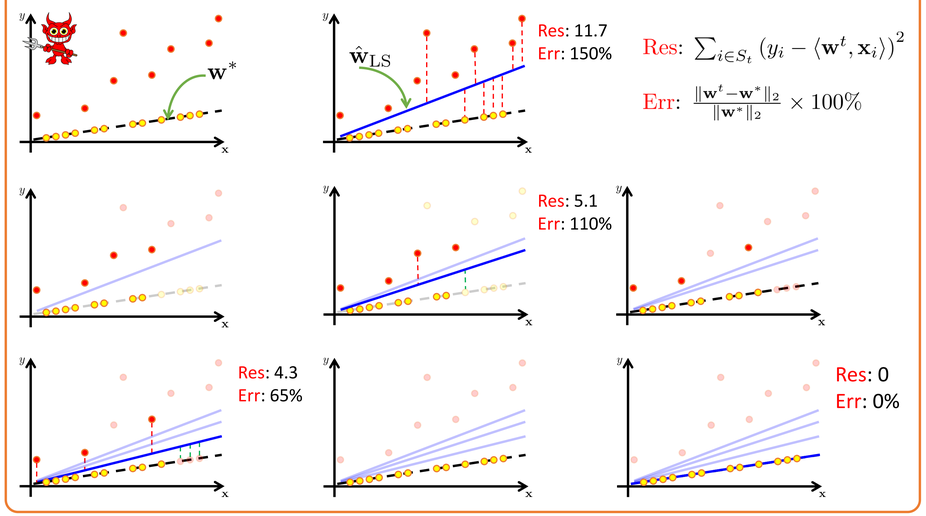
Theorem: TORRENT converges to an ϵ-optimal solution i.e.

$$\|\mathbf{w}^t - \mathbf{w}^*\|_2 \leq \epsilon$$

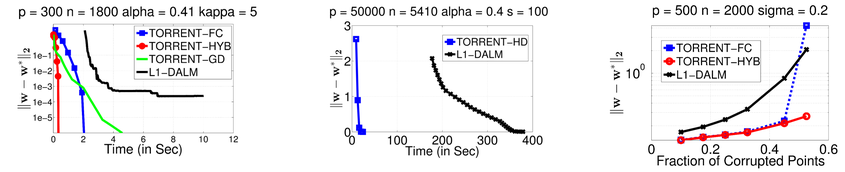
after no more than $\log \frac{1}{\epsilon}$ iterations.

In contrast, RANSAC offers no guarantees, and L_1 relaxation based methods give very poor guarantees in the face of all-powerful adversaries

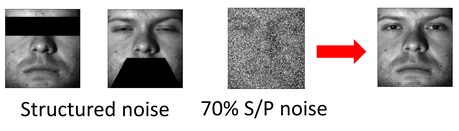
TORRENT in action



Real Data Experiments

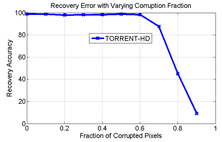


On regression analysis tasks, TORRENT is up to 20x faster than leading methods on low, as well as high dimensional data and can tolerate up to 40% corruption!



Structured noise

70% S/P noise



On face recognition tasks, TORRENT is able to recover the correct identity of the person in the presence of as much as 70% corruption!