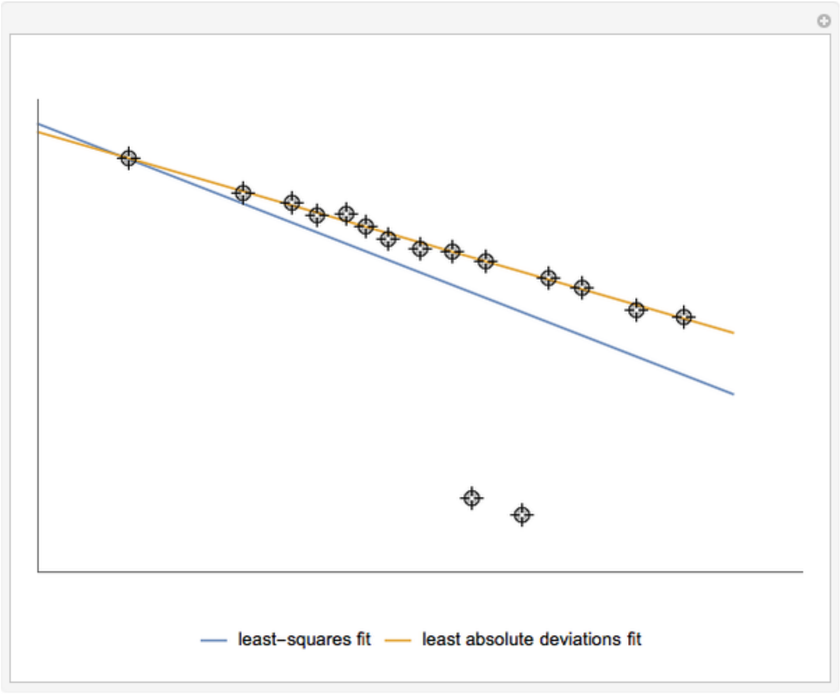
On Problems of Robust High-dimensional Statistics

Gilad Lerman
School of Mathematics
University of Minnesota

Robustness in Ancient Times of Science

- Example 1: Use of least absolute deviations in linear regression instead of least squares



$$y_i = \beta x_i + c + r_i$$

$$\min_{\beta,c} \sum |r_i|$$

instead of

$$\min_{\beta,c} \sum r_i^2$$

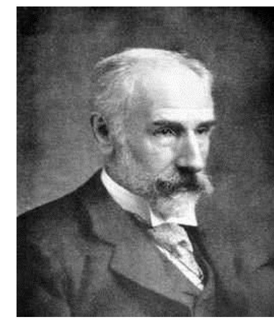
Robustness from Ancient Times of Science

- Example 1: Use of least absolute deviations in linear regression instead of least squares


Galileo Galilei

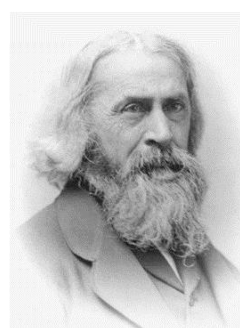

Boscovich

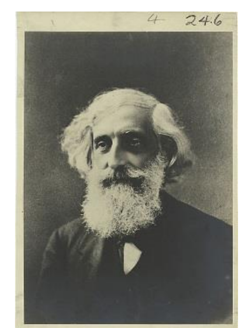

Laplace


Edgeworth

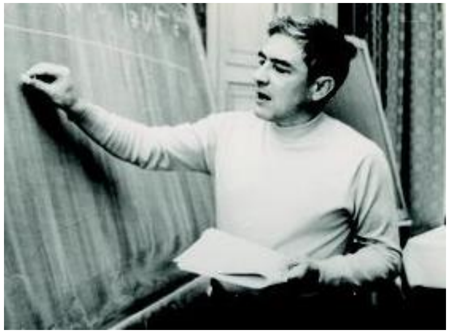
- Example 2: Direct rejection of outliers


Bessel

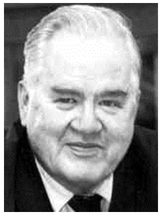

B. Peirce


W. Chauvenet

Theory of Robustness



- Huber: “Robustness signifies insensitivity to small deviations from the assumptions”
- Two influential works from 1960



  - J. W. Tukey, “A survey of sampling from contaminated distributions” (effect of deviation from model, initial analysis)

  - F. J. Anscombe, “Rejection of outliers” (insurance vs. significance, tradeoff with performance, computational cost ignored)

- Bickel (1975): Emphasis on computation

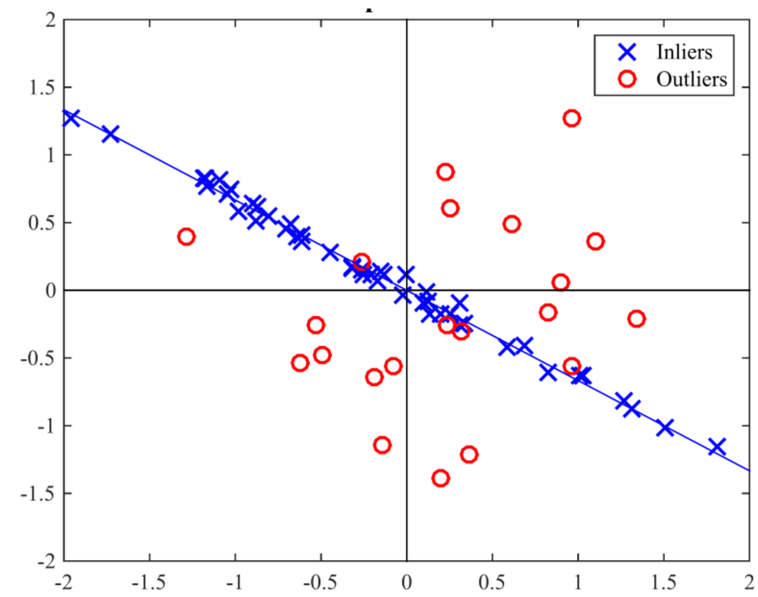
Outline

- Robust subspace recovery (RSR): Review and Insights
- New results on adversarial robustness in RSR
- Related problems and all about that base…

The Robust Subspace Recovery (RSR) Problem

- **Input:** Dataset $X = X_{\text{in}} \cup X_{\text{out}} \subset \mathbb{R}^D$
  X_{in} (inliers) lie near L_* a d-dim. subs. $\subset \mathbb{R}^D$
  X_{out} (outliers) from a different distribution

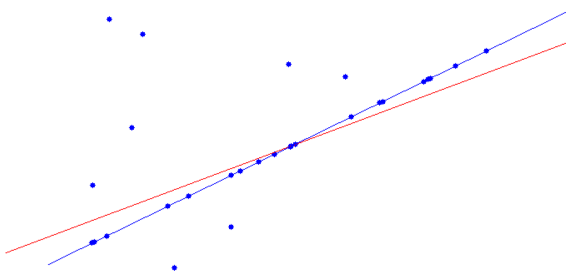
- **Desired Output:** L_*

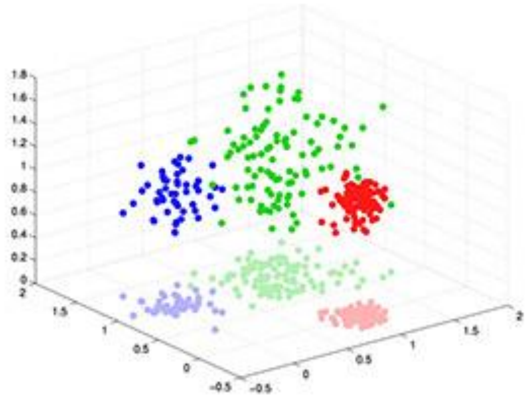


Review: L & Maunu (2018), Proceedings of the IEEE

Why should we care about RSR?

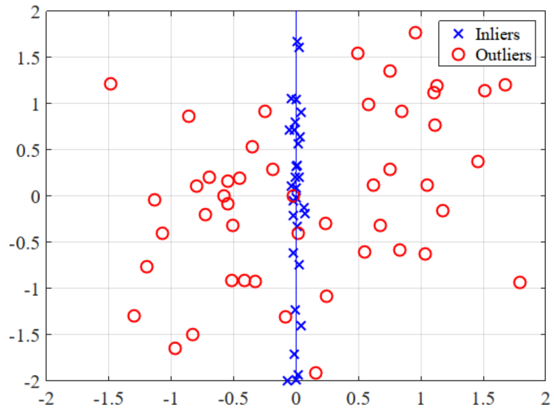
- We should care about PCA
- PCA subs. L minimizes $E_2(X,L)=\sum_{x\in X}\mathrm{dist}^2(x,L)$
- PCA – Basic preprocessing tool
- PCA is not robust to outliers
- Goal: develop an alternative to PCA, which is robust to outliers with nice properties



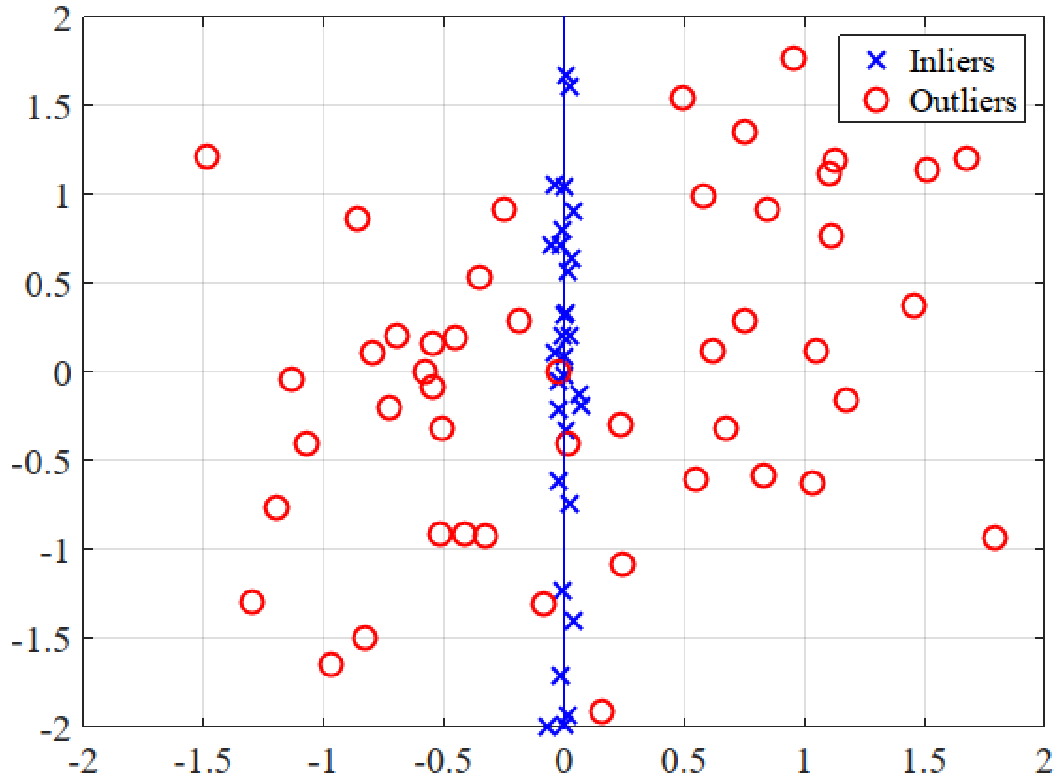


General Approaches to RSR

- Exhaustive subspace search (brute force)
- Rejection of Outliers (filtering)
- Energy minimization
  - Least absolute deviation – min. $E_1(X,L) = \sum_{x \in X} \mathrm{dist}(x,L)$
  - L_1-PCA
  - Projection pursuit
  - Robust covariance (Maronna, Tyler…)

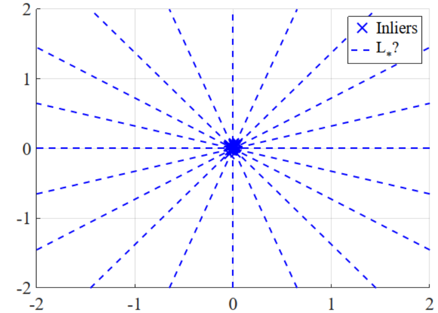


The RSR Formulation can be ill-defined

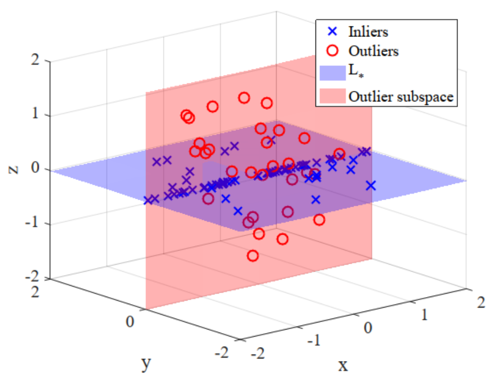


The RSR Formulation can be ill-defined

- **Example 1:** Only inliers at origin



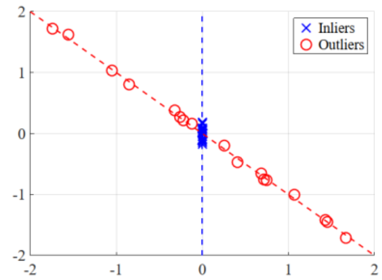
- **Example 2:** Inliers at low-dim. subs.



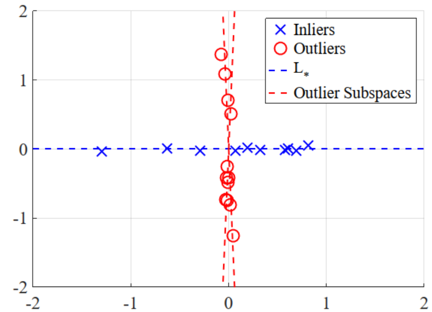
- **Principle I:** Inliers must permeate L_*

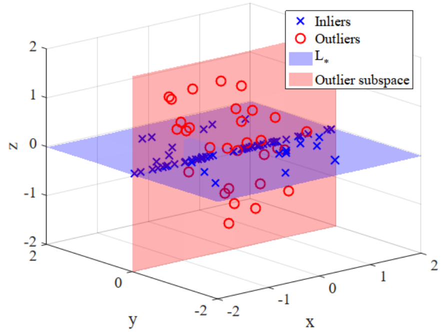
The RSR Formulation can be ill-defined

- **Example 1:** “Aligned” outliers



- **Example 2:** Other “aligned” outliers

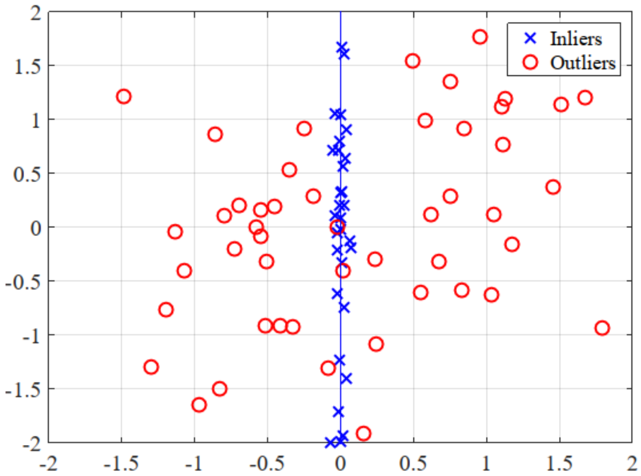




- **Principle II:** Restriction of out. alignment

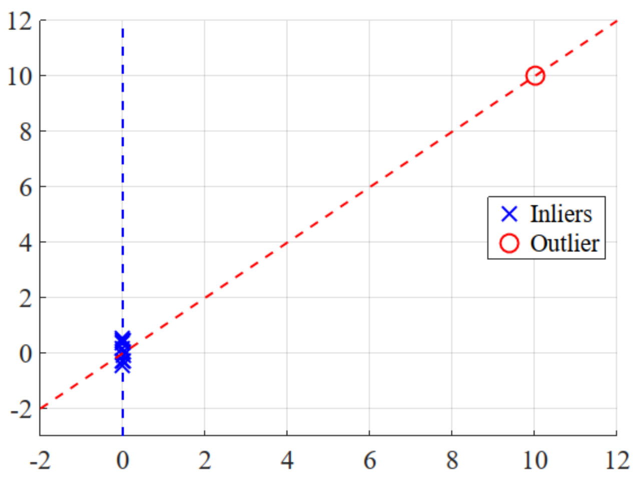
More Clarification for RSR

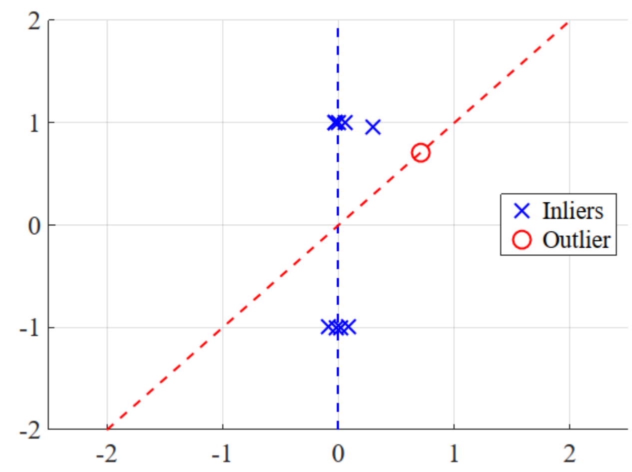
- Simplifying assumption: L_* is linear
- Nonconvex setting: Set of all d-dim. subs. in $\mathbb{R}^D$ (Grassmannian $G(D,d)$) is nonconvex



More Clarification for RSR

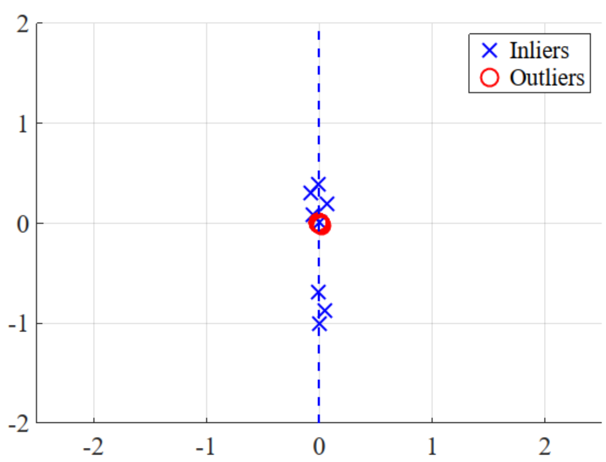
- Scale might be important
- A scale-invariant method does not weigh the magnitude of a point
- Scaling data points to the sphere makes any method scale invariant

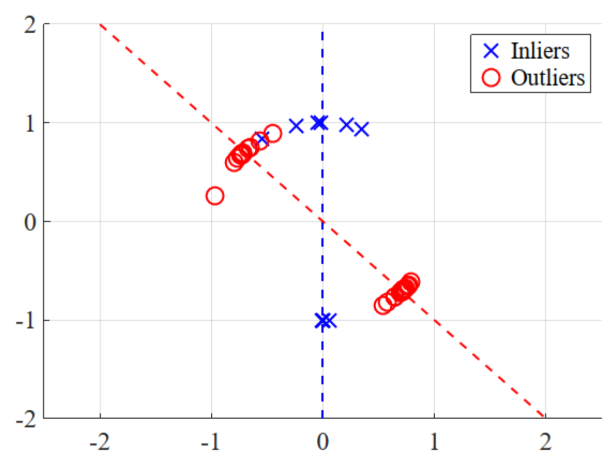




More Clarification for RSR

- Scale might be important
- A scale-invariant method does not weigh the magnitude of a point
- Scaling data points to the sphere makes any method scale invariant





Theoretical Settings for RSR

1. Adversarial outliers, permeated inliers, lower bound on SNR (signal to noise ratio)
   - SNR= fraction of inliers/outliers
   - Formulated for scale-invariant algorithms
2. Statistical model with low bound on SNR
3. Inliers & outliers in general position (GP) (Hardt & Moitra, Zhang, Arias-Castro & Wang)
   - GP: Any set of D points are linearly independent unless it includes at least d+1 points from L_*
   - If SNR<$d/(D$-$d)$, the problem is SSE-hard
   - If SNR>$d/(D$-$d)$, exact recovery by D/RF, TME
   - Restricted model, methods and noise analysis

Adversarial Outliers: Some Previous Works

- OP (Xu et al., 2012):
  $\text{SNR} \geq 121/9 \cdot d \cdot \mu$, μ-incoherence parameter

  Convex method, complexity: $O(TND^2)$, $T = O\left(\varepsilon^{-0.5}\right)$
  Perturbation to noise (121→1024, large error)
  Remark: unless otherwise stated N= O(D)
- TORP (Cherapanamjeri et al., 2017):
  $\text{SNR} \geq 128 \cdot d \cdot \mu^2$

  Nonconvex, complexity O(NDd)
  Small perturbation to noise (large error)
  Nice stability with Gaussian noise (128→1024)
  Requires knowledge of fraction of outliers
- Related algorithms for a different problem:
  - RLG (Diakonikolas et al., 2018)
  - RR (Steinhardt et al., 2018)

Adversarial Outliers (Joint Work with T. Maunu)

Adversarial Outliers: Motivating Questions

- What is the lowest SNR with well-defined setting?
- Is there a hardness result for lowest SNR?
- What is the lowest SNR obtained by a reasonable-time algorithm for exact RSR?
- Any other competitive and flexible algorithm for adversarial outliers?

Best SNR for Well-defined Formulation

- Initial setting: $X_{\text{in}} \subset L_*,\ X_{\text{out}} = X \setminus L_* \subset \mathbb{R}^D$
- Recall: Problem is ill-defined for general X_{in}
- Xu, Caramanis, Sanghavi (2012) explain O(d) SNR with an example, but the inliers in this example are not permeated

Best SNR for Well-defined Formulation

- Initial setting: $X_{\text{in}} \subset L_*$, $X_{\text{out}} = X \setminus L_* \subset \mathbb{R}^D$
- Recall: Problem is ill-defined for general X_{in}
- For general position X_{in} (d points in L_* span L_*),

  well-defined setting if $N_{\text{in}} > (N-d+1)/2$,

  that is,

$$\text{SNR} > \frac{N-d+1}{N+d-1} = 1 - o(1)$$

- If also X_{out} is in GP $(\max_{L \in G(D,d) \setminus L_*} \#(X \cap L) = d)$,

  then the problem is well-defined if $N_{\text{in}} > d$

  (SNR = $d/(N\text{-}d)$=o(1))

Hardness Result

- Recall Hardt & Moitra (2013):
  If SNR< $d/(D\text{-}d)$, the problem is SSE-hard
- Thus too small SNR in the case of GP inliers & outliers results in SSE-hard formulation
- It is also relevant for GP inliers and adversarial outliers when $D\text{-}d$=O(1) and $d/(D\text{-}d)=O(d)$=O(D)

Best SNR for an Algorithm

Best SNR for RANSAC

- RANSAC-type algorithm for RSR

Input: dataset $\mathcal{X}$, subspace dimension d, tolerance τ, consensus number m, max iterations n
Output: $L_* \in G(D, d)$

$k = 0,\ i = 1$
while $i \leq n$ **do**
  $\mathcal{Y} \leftarrow$ random subset of $\mathcal{X}$ that spans a d-subspace
  $L = \mathrm{Sp}(\mathcal{Y})$
  $c = \#(\{\boldsymbol{x} \in \mathcal{X} : \angle(\boldsymbol{x}, L) \leq \tau\})$
  if $c > k$ **then**
    $L_* = L$
    $k = c$
  end
  if $k > m$ **then**
    return L_*
  end
  $i \leftarrow i + 1$
end
return L_*

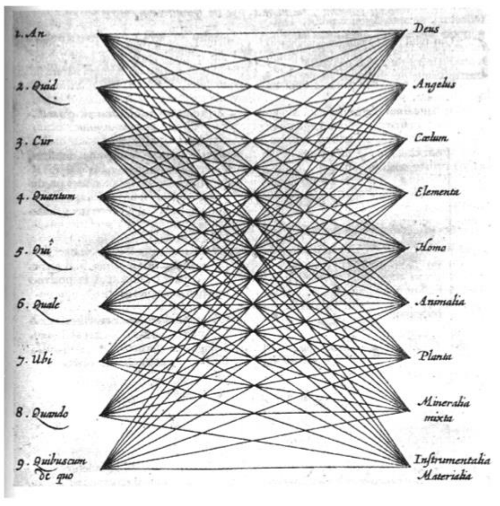
Best SNR for RANSAC

- RANSAC-type algorithm for RSR
- For permeated inliers, SNR>1, $n \gg 1, \tau \ll 1$ and $m \geq N/2$, L_* is recovered w.h.p.
- If also SNR $\geq$ cd, L_* recovered w.h.p. when n=O(1) and the complexity of the algorithm is $O(NDd)$

Another Proposal for Adversarial Outliers

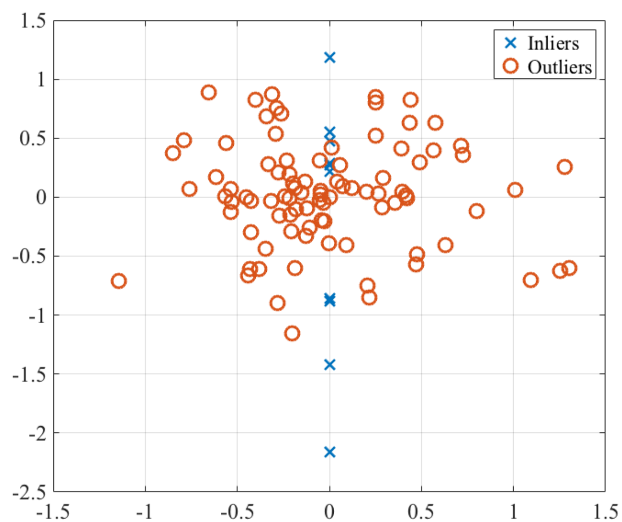
- Review of GGD (geodesic gradient descent) and of the well-tempered landscape (WTL) of
  $$E_1(X,L) = \sum_{x \in X} \mathrm{dist}(x,L)$$
  (Maunu, Zhang, L)

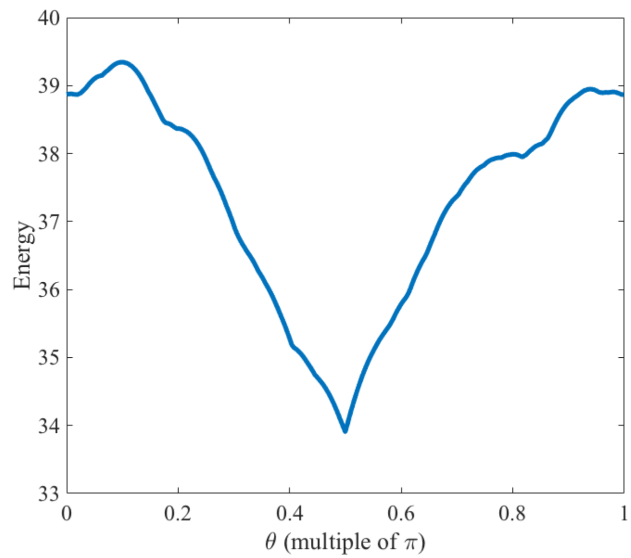
THE WELL-TEMPERED EXPOSITION



Review of WTL & GGD

- Under a generic "stability" condition, there exists a neighborhood of L_*, where L_* is the only minimizer of $E_1(X,L)=\sum_{x\in X}\operatorname{dist}(x,L)$ and all other points have a direction of strictly decreasing cost





Review of WTL & GGD

- Under a generic "stability" condition, there exists a neighborhood of L_*, where L_* is the only minimizer and all other points have a direction of strictly decreasing cost
- GGD initialized in this neighborhood converges to L_* sufficiently fast
- Under a similar condition, PCA initializes in this neighborhood

Spherical GGD with Adversarial Outliers

- SGGD: spherize data + GGD
- Spherize: $x_i \to \tilde{x}_i := x_i / \|x_i\|, \; X \to \tilde{X}$
- Spherical condition number of inliers

$$\kappa_d(X_{\text{in}}) = \frac{\lambda_1\left(\tilde{X}_{\text{in}}\tilde{X}_{\text{in}}^T\right)}{\lambda_d\left(\tilde{X}_{\text{in}}\tilde{X}_{\text{in}}^T\right)}$$

- Stability condition for SGGD initialized in $B(L_*,\gamma)$:

$$\text{SNR} \geq \frac{d \cdot \kappa_d(\tilde{X}_{\text{in}})}{\cos(\gamma)}$$

Spherical GGD with Adversarial Outliers

- Condition for SPCA in $B(L_*, \gamma)$:

$$\mathrm{SNR} \geq \frac{\sqrt{2} \cdot d \cdot \kappa_d(\tilde{X}_{\mathrm{in}})}{\sin(\gamma)}$$

- Stability condition for SGGD (+SPCA):

$$\mathrm{SNR} \geq \sqrt{3} \cdot d \cdot \kappa_d(\tilde{X}_{\mathrm{in}})$$

- Condition for linear convergence:

$$\mathrm{SNR} \geq \left(2 + \frac{d-1}{N_{\mathrm{out}}}\right) \cdot \sqrt{3} \cdot d \cdot \kappa_d(\tilde{X}_{\mathrm{in}})$$

- For isotropic Gaussian distribution of inliers:

$$\kappa_d(\tilde{X}_{\mathrm{in}}) = O(1) \text{ unlike } \mu = O\left(\max(1, \log N / d)\right)$$

Table of Comparisons

SPCA [31]	SNR $> \dfrac{d\kappa_d(\widetilde{\boldsymbol{X}_{\text{in}}})}{\sin(\gamma)/\sqrt{2}}$ *Upsides:* *Fast γ-approximation.* *Downsides:* *No exact recovery (only γ-approximation).*
OP [47]	SNR $\geq \dfrac{121\mu d}{9}$ *Upsides:* *Convex algorithm.* *Downsides:* *Incoherence μ can be large, poor constants, no strong noise analysis, requires parameter tuning.*
TORP [6]	SNR $\geq 128\mu^2 d - 1$ *Upsides:* *Nice analysis for Gaussian noise.* *Downsides:* *Parameter μ can be large, poor constants, requires parameter tuning.*
RR ([42])	SNR ≥ 2 *Upsides:* *Fast approximation, best constant SNR breakdown,* *Downsides:* *Constant approximation, no exact recovery, returns a matrix of rank at most 15d, requires resilient inliers.*
RLG ([9])	SNR $\geq \dfrac{1-\epsilon}{\epsilon}$ *Upsides:* *Fast ϵ-approximation for a different problem.* *Downsides:* *For RSR, reduces to only Gaussian inliers and no exact recovery (ϵ-approximation).*
SGGD ([38] and this work)	SNR $> \sqrt{3}d\kappa_d(\widetilde{\boldsymbol{X}_{\text{in}}})$ *Upsides:* *Efficient, linear convergence with another condition, good constants, adapts to other statistical models of data.* *Downsides:* *No strong noise analysis.*
RANSAC ([15] and this work)	SNR $\geq cd$ *Upsides:* *Good constants.* *Downsides:* *Potentially sensitive and unstable to noise, requires parameter tuning.*

SGGD under Statistical Models

- (S)GGD adapts well to other statistical models with low SNR (Maunu, Zhang, L)
- In the case of the Needle-Haystck Model (Gaussian inliers & outliers) its lowest SNR adapts to different sample regimes.
- In particular, it can address any SNR when $N_{\text{out}} \geq C \cdot \max(D^3 \log^3(N), (d/\text{SNR})^6)$
- Complexity under this model: $\mathrm{O}(NDd)$

Open Directions in RSR

- Robustness to noise or under spiked model
- Large sample & high-dimensional limits for RSR
- Estimation of subspace dimension
- Clarification of tradeoffs
- Case for applications
- Beyond gradient descent?
- Specific issues: affine case, improved guarantees, missing values, virtue of dimension reduction, phase transitions

Relevance to Other Problems

- Robust covariance estimation
- Robust subspace/manifold clustering
- Robust synchronization
- Estimation of camera locations from corrupted pairwise directions
- Robust fundamental/essential camera estimation

Take-Home Message

- RSR is a basic problem that raises interesting questions
- Clean treatment of the case of adversary outliers
- The LAD RSR optimization problem is nonconvex.
  The non-convex SGGD is shown to be fast and flexible
- Ideas seem to extend to other problems

Thanks

- Highlighted Collaboration:
  Tyler Maunu (MIT),

- Relevant past and present collaborations:
  Arthur Szlam (Facebook), Ery Arias-Castro (UCSD),
  Teng Zhang (UCF), Yi Grace Wang (CSUDH),
  Mike McCoy (Cruise Automation), Joel Tropp (Caltech),
  John Goes (Lord Abbett), Boaz Nadler (Weizmann),
  Vahan Huroyan (UA), Yunpeng Shi (UMN),
  Wei-Kuo Chen (UMN), Madeline Handschy (UMN)
- Support by NSF, NGA, NASA
- Contact: lerman@umn.edu