Being Robust (in High Dimensions) Can Be Practical

Ilias Diakonikolas[1] Gautam Kamath[2] Daniel M. Kane[3] Jerry Li[2]
Ankur Moitra[4] Alistair Stewart[1]

[1]CS, USC

[2]EECS & CSAIL, MIT

[3]CSE & Math, UCSD

[4]Math & CSAIL, MIT

ICML, 2017

Presenter: Bargav Jayaraman

Outline

Outline

Introduction to Robust Estimation

- Given that samples come from a nice distribution, but adversary has corrupted a constant fraction of samples, goal is to robustly estimate the statistics - mean and covariance
- In one-dimension, robust alternatives to mean and covariance exist - *median* and *interquantile range*
- In high dimensions, there is a trade-off between robustness and computational efficiency
  - **Tukey median** - hard to compute; heuristics based computation does not scale with dimensions
  - **Minimum volume ellipsoid** - hard to compute; heuristics based computation scale poorly with dimensions

Outline

Related Work

- **Robust mean and covariance estimation** - [DKK+16] gave an algorithm for agnostically learning the parameters of a Gaussian $\mathcal{N}(\mu, \Sigma)$ that satisfy $d_{TV}(\mathcal{N}, \mathcal{N}') \leq O(\epsilon)$ where ϵ samples are corrupted, and the computational complexity of the algorithm is polynomial in dimensionality d and $1/\epsilon$. [LRV16] proposed unknown mean estimation where $d_{TV}(\mathcal{N}, \mathcal{N}') \leq O(\epsilon\sqrt{\log d})$.
- **Robust PCA** - [CLMW11] proposed robust PCA with semidefinite programming which can tolerate a constant fraction of corruptions. [XCS10] used semidefinite programming for robust PCA with outliers.

Outline

Contributions

1. Modification to the algorithm of [DKK+16] with the definition of *good sets* to estimate the mean with $O(d/\epsilon^2)$ samples and covariance with $O(d^2/\epsilon^2)$ samples.
2. Improvement to the number of corruptions that can be tolerated by empirically tuning the threshold for filtering of corrupt points.
3. Same bounds are shown to be valid even for weaker distributional assumptions of the underlying data.
4. Comparison of models via visual representation of genetic data that encodes the map of Europe.

Outline

Formal Framework

Notation: For a vector v, $\|v\|_2$ is the Euclidean norm, and for a matrix M, $\|M\|_2$ is the spectral norm and $\|M\|_F$ is the Frobenius norm. $X \in_u S$ means sample X is drawn from the empirical distribution defined by S.

Definition (ϵ-corruption)

Given $\epsilon > 0$ and a distribution family $\mathcal{D}$, the algorithm spcifies the number of samples m and the adversary generates m samples $X_1, X_2, \ldots, X_m$ from some $D \in \mathcal{D}$. It then draws $m' \sim Bin(\epsilon, m)$ from an appropriate distribution and replaces m' of the input samples with arbitrary points. The altered samples are given to the algorithm.

Goal of the algorithm is to return the parameters of $\hat{D}$ that are close to true distribution D. For mean, Euclidean distance is used and for covariance, Mahalanobis distance is used $\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I\|_F$.

Outline

Nearly Sample-Optimal Efficient Robust Learning

The overall filtering procedure for robust estimation is the following iterative procedure:

1. find some univariate test (via spectral methods) that is violated by the corrupted points
2. find some concrete tail bound violated by the corrupted set of points
3. thrown away all the points which violate this tail bound

Robust Mean Estimation of Sub-Gaussian

Theorem (1)

Let G be a sub-gaussian distribution on $\mathbb{R}^d$ with parameter $\nu = \Theta(1)$, mean μ^G, covariance matrix I, and $\epsilon > 0$. Let S be an ϵ-corrupted set of samples from G of size $\Omega((d/\epsilon^2)\mathrm{poly}\log(d/\epsilon))$. There exists an efficient algorithm that, on input S and $\epsilon > 0$, returns a mean vector $\hat{\mu}$ so that with probability at least 9/10 we have $\|\hat{\mu} - \mu^G\|_2 = O(\epsilon\sqrt{\log 1/\epsilon})$.

Robust Mean Estimation under Bounded Second Moment

Theorem (2)

Let P be a distribution on $\mathbb{R}^d$ with unknown mean vector μ^P and unknown covariance matrix $\Sigma_P \leq \sigma^2 I$. Let S be an ϵ-corrputed set of samples from P of size $\Theta((d/\epsilon)\log d)$. There exists an efficient algorithm that, on input S and $\epsilon > 0$, with probability 9/10 outputs $\|\hat{\mu} - \mu^P\|_2 = O(\sqrt{\epsilon}\sigma)$.

The main difference between this theorem and the previous one is the choice of filtering threshold. Instead of looking for a violation of a concentration inequality, here threshold is chosen at random.

Caution: This method may throw away even some uncorrupted points but it only rejects $O(\epsilon)$ samples with high probability.

Robust Covariance Estimation of Gaussian

Theorem (3)

Let $G \sim \mathcal{N}(0, \Sigma)$ be a Gaussian in d dimensions, and let $\epsilon > 0$. Let S be an ϵ-corrupted set of samples from G of size $\Omega((d^2/\epsilon^2)\mathrm{poly}\log(d/\epsilon))$. There exists an efficient algorithm that, given S and ϵ, returns the parameters of a Gaussian distribution $G' \sim \mathcal{N}(0, \Sigma)$ so that with probability at least 9/10, it holds $\|I - \Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2}\|_F = O(\epsilon \log(1/\epsilon))$.

Outline

Robust Mean Estimation

Algorithms which achieve Theorems 1 and 2 have common template, with three parameters:

- Thresh(ϵ) - threshold function to terminate if covariance has spectral norm bounded by Thresh(ϵ).
- Tail($T, d, \epsilon, \delta, \tau$) - univariate tail bound violated by only τ fraction of uncorrupted points, but more of corrupted points.
- $\delta(\epsilon, s)$ - slack function (required for technical reasons).

Robust Mean Estimation Algorithm Template

Algorithm 1 Filter-based algorithm template for robust mean estimation

Input: An ε-corrupted set of samples S, $\mathrm{Thres}(\varepsilon), \mathrm{Tail}(T, d, \varepsilon, \delta, \tau), \delta(\varepsilon, s)$
Compute the sample mean $\mu^{S'} = \mathbb{E}_{X \in_u S'}[X]$
Compute the sample covariance matrix Σ
Compute approximations for the largest absolute eigenvalue of Σ, $\lambda^* := \|\Sigma\|_2$, and the associated unit eigenvector v^*.
if $\|\Sigma\|_2 \leq \mathrm{Thres}(\varepsilon)$ **then**
    return $\mu^{S'}$.
Let $\delta = \delta(\varepsilon, \|\Sigma\|_2)$.
Find $T > 0$ such that

$$\Pr_{X \in_u S'} \left[|v^* \cdot (X - \mu^{S'})| > T + \delta \right] > \mathrm{Tail}(T, d, \varepsilon, \delta, \tau).$$

return $\{x \in S' : |v^* \cdot (x - \mu^{S'})| \leq T + \delta\}$.

Algorithm for Sub-Gaussian (Theorem 1)

Algorithm 2 Filter algorithm for a sub-gaussian with unknown mean and identity covariance

1: **procedure** FILTER-SUB-GAUSSIAN-UNKNOWN-MEAN(S', ε, τ)

input: A multiset S' such that there exists an (ε, τ)-good S with $\Delta(S, S') \leq 2\varepsilon$

output: Multiset S'' or mean vector $\widehat{\mu}$ satisfying Proposition A.7

2: Compute the sample mean $\mu^{S'} = \mathbb{E}_{X \in_u S'}[X]$ and the sample covariance matrix Σ , i.e., $\Sigma = (\Sigma_{i,j})_{1 \leq i,j \leq d}$ with $\Sigma_{i,j} = \mathbb{E}_{X \in_u S'}[(X_i - \mu_i^{S'})(X_j - \mu_j^{S'})]$.

3: Compute approximations for the largest absolute eigenvalue of $\Sigma - I$, $\lambda^* := \|\Sigma - I\|_2$, and the associated unit eigenvector v^*.

4: **if** $\|\Sigma - I\|_2 \leq O(\varepsilon \log(1/\varepsilon))$, **then return** $\mu^{S'}$.

5: Let $\delta := 3\sqrt{\varepsilon \|\Sigma - I\|_2}$. Find $T > 0$ such that

$$\Pr_{X \in_u S'} \left[|v^* \cdot (X - \mu^{S'})| > T + \delta \right] > 8 \exp(-T^2/2\nu) + 8 \frac{\varepsilon}{T^2 \log\left(d \log(\frac{d}{\varepsilon\tau})\right)}.$$

6: **return** the multiset $S'' = \{x \in S' : |v^* \cdot (x - \mu^{S'})| \leq T + \delta\}$.

Algorithm with Bounded Second Moment (Theorem 2)

Algorithm 3 Filter under second moment assumptions

function FILTERUNDER2NDMOMENT(S)
  Compute μ^S, Σ^S, the mean and covariance matrix of S.
  Find the eigenvector v^* with highest eigenvalue λ^* of Σ^S.
  if $\lambda^* \leq 9$ **then**
    return μ^S
  else
    Draw Z from the distribution on $[0, 1]$ with probability density function $2x$.
    Let $T = Z \max\{|v^* \cdot x - \mu^S| : x \in S\}$.
    Return the set $S' = \{x \in S : |v^* \cdot (X - \mu^S)| < T\}$.

Robust Covariance Estimation (Theorem 3)

Algorithm 4 Filter algorithm for a Gaussian with unknown covariance matrix.

procedure FILTER-GAUSSIAN-UNKNOWN-COVARIANCE(S', ε, τ)

input: A multiset S' such that there exists an (ε, τ)-good set S with $\Delta(S, S') \leq 2\varepsilon$

output: Either a set S'' with $\Delta(S, S'') < \Delta(S, S')$ or the parameters of a Gaussian G' with $d_{TV}(G, G') = O(\varepsilon \log(1/\varepsilon))$.

Let $C > 0$ be a sufficiently large universal constant.

Let Σ' be the matrix $\mathbb{E}_{X \in_u S'}[XX^T]$ and let G' be the mean 0 Gaussian with covariance matrix Σ'.

if there is any $x \in S'$ so that $x^T(\Sigma')^{-1}x \geq Cd\log(|S'|/\tau)$ **then**

  return $S'' = S' - \{x : x^T(\Sigma')^{-1}x \geq Cd\log(|S'|/\tau)\}$.

Compute an approximate eigendecomposition of Σ' and use it to compute $\Sigma'^{-1/2}$

Let $x_{(1)}, \ldots, x_{(|S'|)}$ be the elements of S'.

For $i = 1, \ldots, |S'|$, let $y_{(i)} = \Sigma'^{-1/2}x_{(i)}$ and $z_{(i)} = y_{(i)}^{\otimes 2}$.

Let $T_{S'} = -I^\flat I^{\flat T} + (1/|S'|)\sum_{i=1}^{|S'|} z_{(i)}z_{(i)}^T$.

Approximate the top eigenvalue λ^* and corresponding unit eigenvector v^* of $T_{S'}$..

Let $p^*(x) = \frac{1}{\sqrt{2}}((\Sigma'^{-1/2}x)^T v^{*\sharp}(\Sigma'^{-1/2}x) - \mathrm{tr}(v^{*\sharp}))$

if $\lambda^* \leq (1 + C\varepsilon\log^2(1/\varepsilon))Q_{G'}(p^*)$ **then**

  return G'

Let μ be the median value of $p^*(X)$ over $X \in S'$.

Find a $T \geq C'$ so that

$$\Pr_{X \in_u S'}(|p^*(X) - \mu| \geq T + 4/3) \geq \mathrm{Tail}(T, d, \varepsilon, \tau)$$

return $S'' = \{X \in S' : |p^*(X) - \mu| < T\}$.

Outline

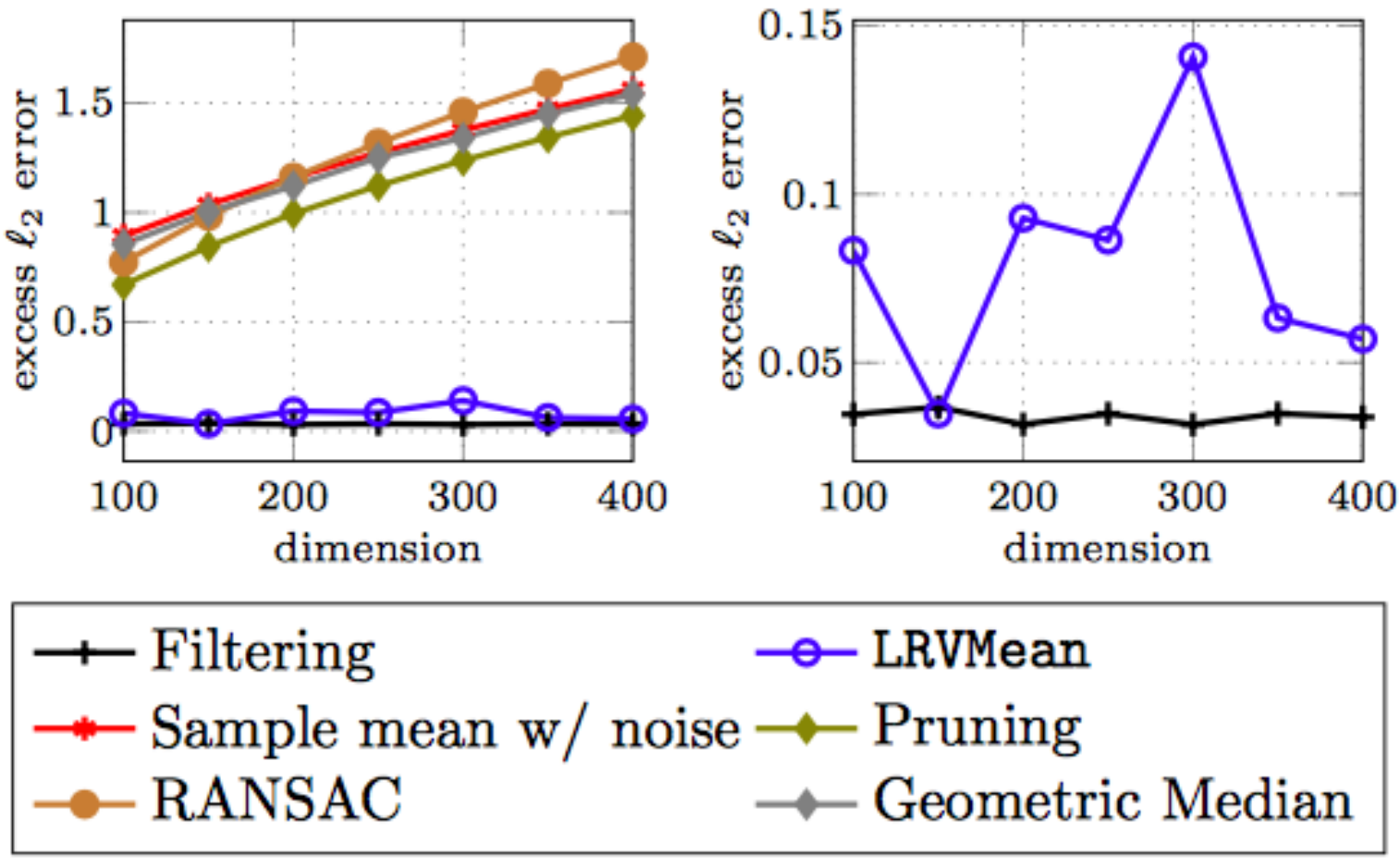
Experimental Settings

Experiments were performed over:

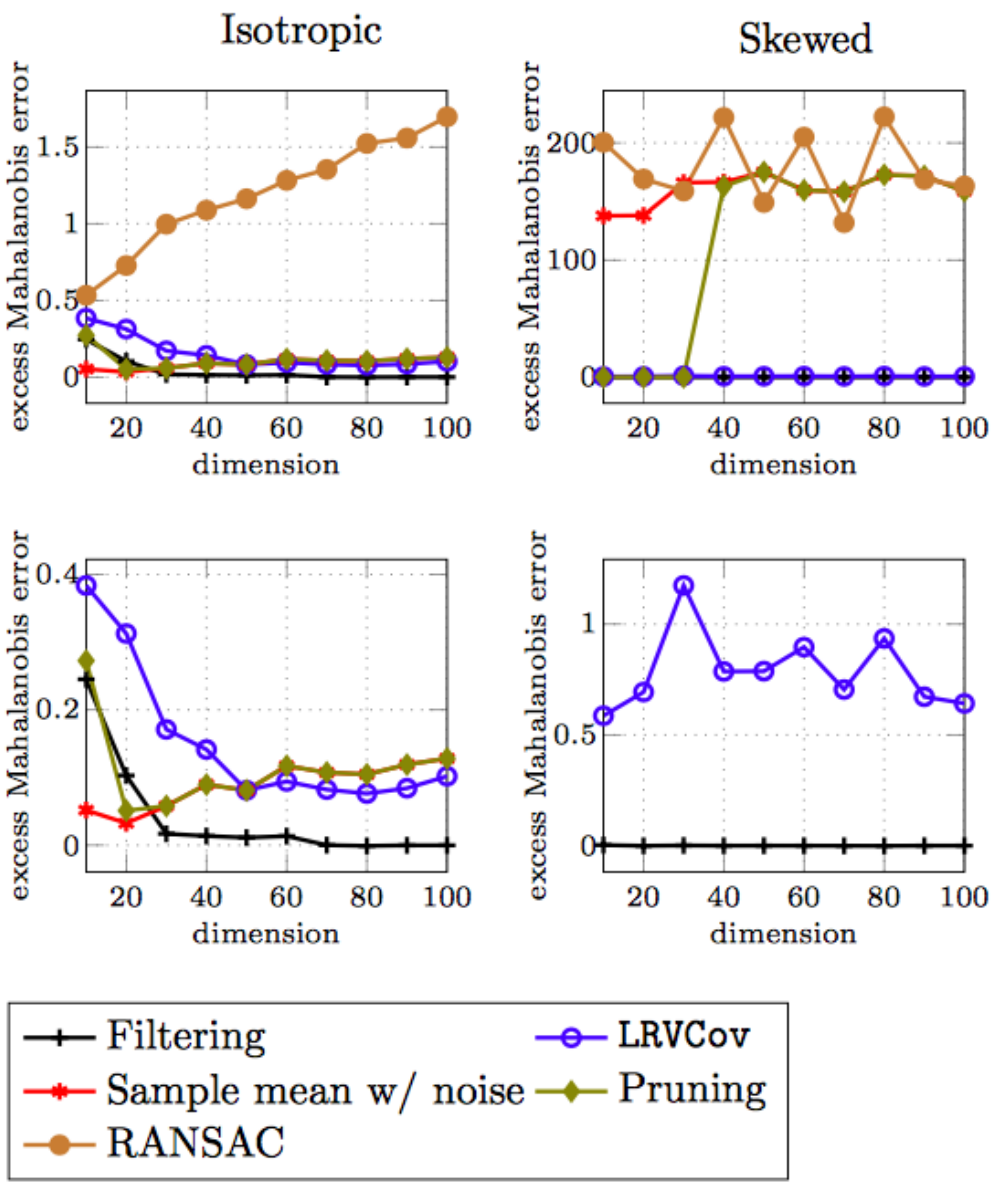
- **Synthetic dataset unknown mean** - $\epsilon = 0.1$, $d = [100, 150, \ldots, 400]$, $n = 10d/\epsilon^2$ samples are generated where $(1 - \epsilon)$ fraction come from $\mathcal{N}(\mu, I)$
- **Synthetic dataset unknown covariance** - $\epsilon = 0.1$, $d = [10, 20, \ldots, 100]$, $n = 0.5d/\epsilon^2$ samples are generated where $(1 - \epsilon)$ fraction come from $\mathcal{N}(0, \Sigma)$
- **Semi-synthetic dataset** - Genotype of thousands of individuals. PCA is used to project into two dimensions, which have a striking resemblance to the map of Europe.

Outline

Synthetic Data - Mean Estimation



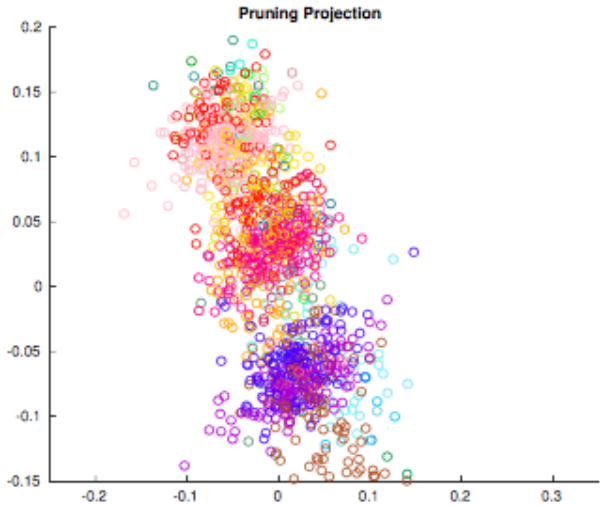
Synthetic Data - Covariance Estimation



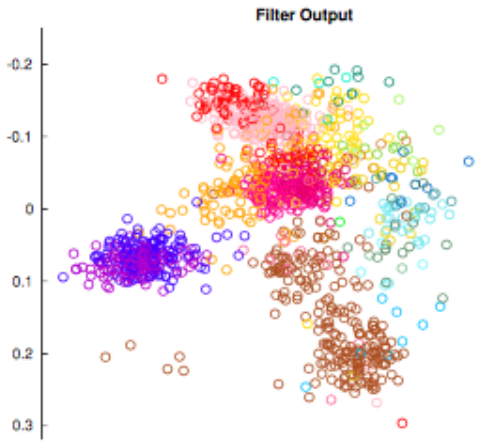
Outline

Semi-Synthetic Data

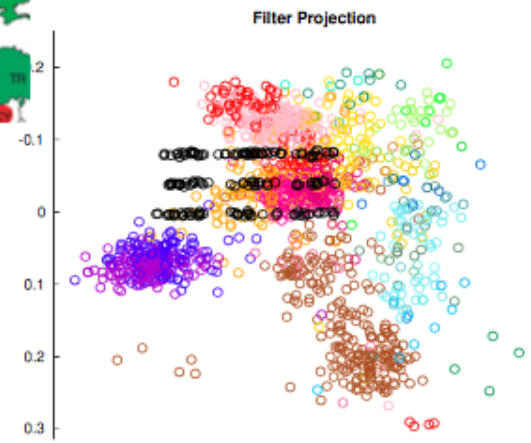
The data projected onto the top two directions of the original data set without noise



The data projected onto the top two directions of the noisy data set after pruning

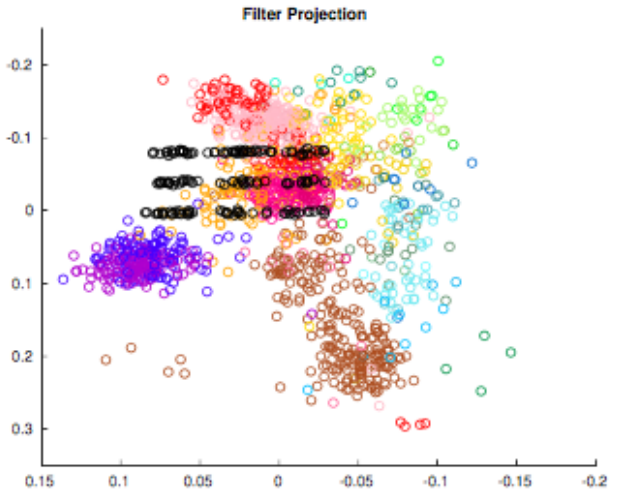


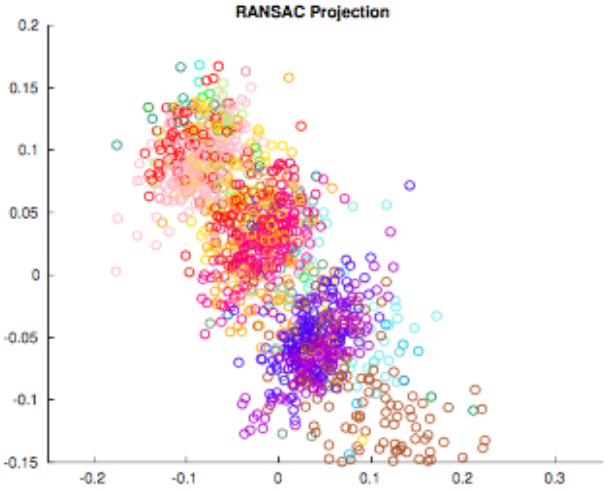
The filtered set of points projected onto the top two directions returned by the filter

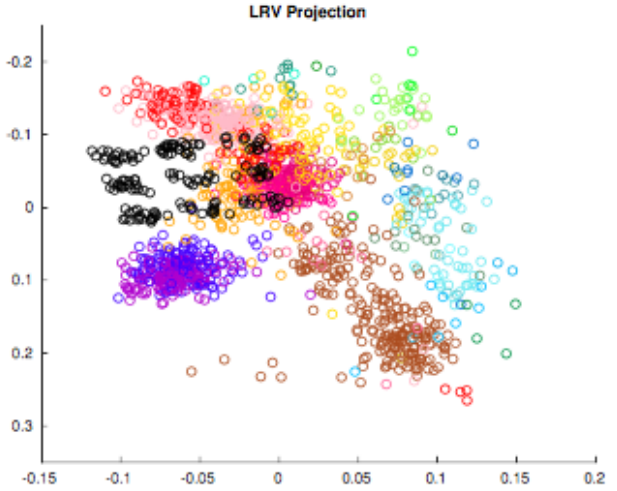


The data projected onto the top two directions returned by the filter

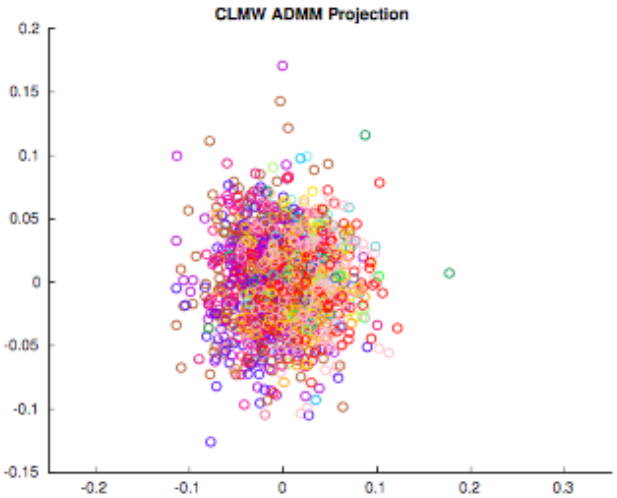
Semi-Synthetic Data

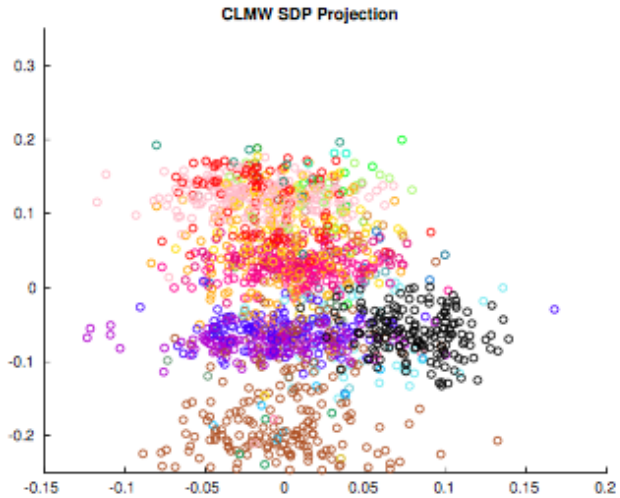


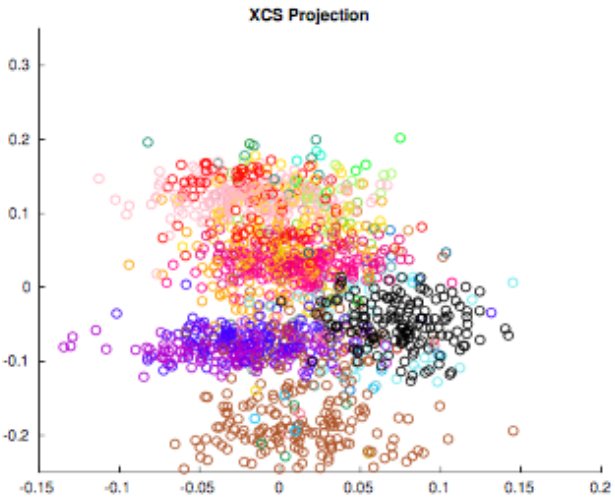




Semi-Synthetic Data







Conclusion

- Robust estimation of mean and covariance in high-dimensional data
- Experiments over synthetic data comparing the statistical accuracy of proposed method over existing approaches
- Experiments over Genotype dataset to visually represent the performance of various approaches in recovering the map of Europe
- Theoretical proof of correctness of the proposed approach

Roadmap of Proof of Theorem 1