Robust PCA for High-Dimensional Data

Huan Xu, Constantine Caramanis and Shie Mannor

Talk by Shie Mannor, The Technion
Department of Electrical Engineering

June 2010

Thank you for staying for the graveyard session

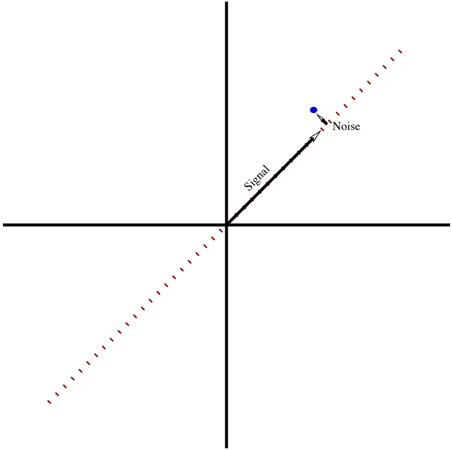
PCA - in Words

- Observe high-dimensional points
- Find least-square-error subspace approximation

- Many applications in feature-extraction and compression
  - data analysis
  - communication theory
  - pattern recognition
  - image processing

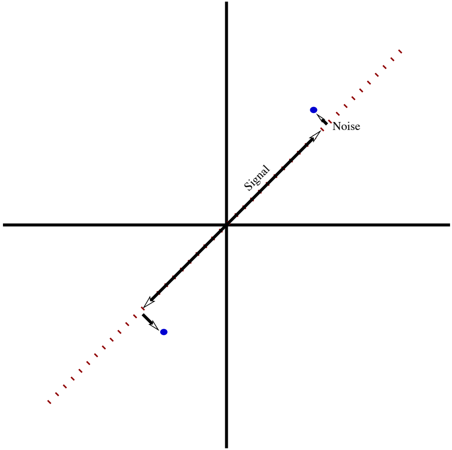
PCA - in Pictures

Observe points: $\mathbf{y} = A\mathbf{x} + \mathbf{v}$.



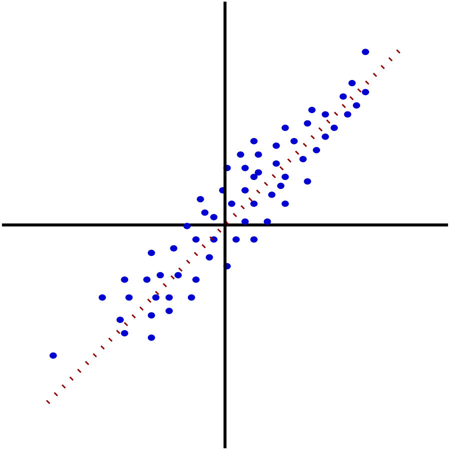
PCA - in Pictures

Observe points: $\mathbf{y} = A\mathbf{x} + \mathbf{v}$.



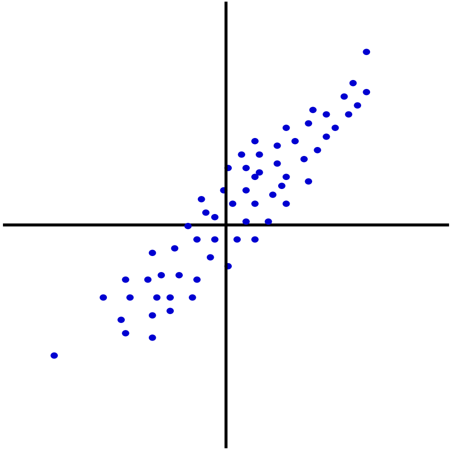
PCA - in Pictures

Observe points: $\mathbf{y} = A\mathbf{x} + \mathbf{v}$.



PCA - in Pictures

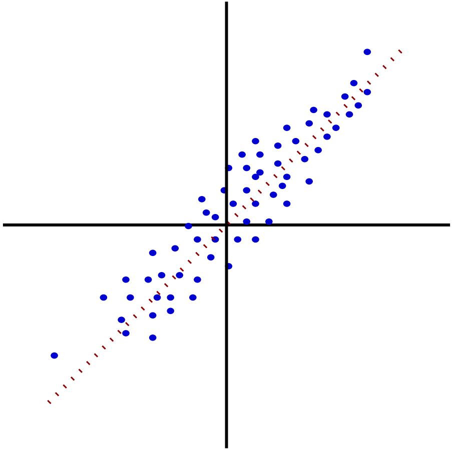
Observe points: $\mathbf{y} = A\mathbf{x} + \mathbf{v}$.



PCA - in Pictures

Observe points: $\mathbf{y} = A\mathbf{x} + \mathbf{v}$.

Goal: Find least-square-error subspace approximation.



PCA - in Math

- Least-square-error subspace approximation
- How: Singular value decomposition (SVD) performs eigenvector decomposition of the sample-covariance matrix

PCA - in Math

- Least-square-error subspace approximation
- How: Singular value decomposition (SVD) performs eigenvector decomposition of the sample-covariance matrix
- Magic of SVD: solving a non-convex problem
- Cannot replace quadratic objective here.

PCA - in Math

- Least-square-error subspace approximation
- How: Singular value decomposition (SVD) performs eigenvector decomposition of the sample-covariance matrix
- Magic of SVD: solving a non-convex problem
- Cannot replace quadratic objective here.
- Consequence: Sensitive to outliers
  - Even one outlier can make the output arbitrarily skewed;
  - What about a constant fraction of “outliers”?

This Talk: High Dimensions and Corruption

Two key differences to pictures shown

(A) High-dimensional regime: # observations $\leq$ dimensionality.

(B) A constant fraction of points arbitrarily corrupted.

Outline

1. **Motivation: PCA, High dimensions, corruption**
2. Where things get tricky: usual tools fail
3. HR-PCA: the algorithm
4. The Proof Ideas (and some details)
5. Conclusion

High-Dimensional Data

- What is high-dimensional data: $\#$dimensionality $\approx \#$ observations.
- Why high-dimensional data analysis:
  - Many practical examples: DNA microarray, financial data, semantic indexing, images, etc

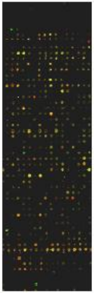


Figure: MicroArray: $24,401$ dim.

High-Dimensional Data

- What is high-dimensional data: $\#$dimensionality $\approx \#$ observations.
- Why high-dimensional data analysis:
  - Many practical examples: DNA microarray, financial data, semantic indexing, images, etc
  - Networks: user-behavior-aware network algorithms (Cognitive Networks)?

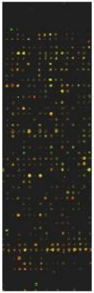


Figure: MicroArray: $24,401$ dim.

High-Dimensional Data

- What is high-dimensional data: $\#$dimensionality $\approx \#$ observations.
- Why high-dimensional data analysis:
  - Many practical examples: DNA microarray, financial data, semantic indexing, images, etc
  - Networks: user-behavior-aware network algorithms (Cognitive Networks)?
  - The kernel trick generates high-dimensional data

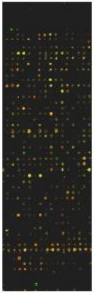


Figure: MicroArray: $24,401$ dim.

High-Dimensional Data

- What is high-dimensional data: $\#$dimensionality $\approx \#$ observations.
- Why high-dimensional data analysis:
  - Many practical examples: DNA microarray, financial data, semantic indexing, images, etc
  - Networks: user-behavior-aware network algorithms (Cognitive Networks)?
  - The kernel trick generates high-dimensional data
  - Traditional statistical tools do not work

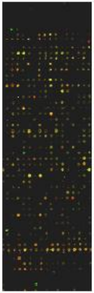


Figure: MicroArray: $24,401$ dim.

Corrupted Data

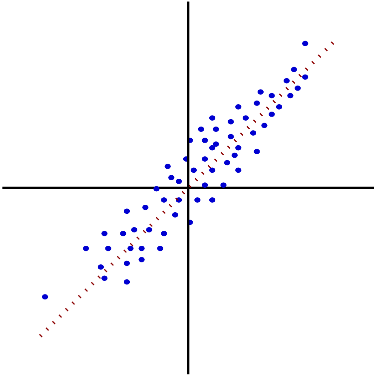


Figure: No Outliers

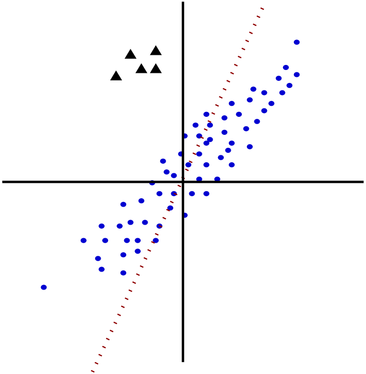


Figure: With Outliers

Corrupted Data

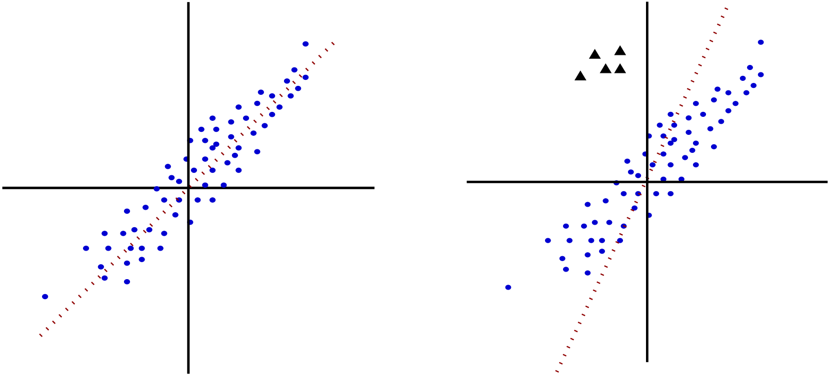


Figure: No Outliers　　Figure: With Outliers

- Some observations about the corrupted points:
  - They have a large magnitude.
  - They have a large (Mahalanobis) distance.
  - They increase the volume of the smallest containing ellipsoid.

Corrupted Data

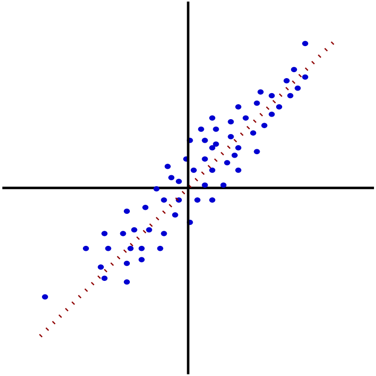


Figure: No Outliers

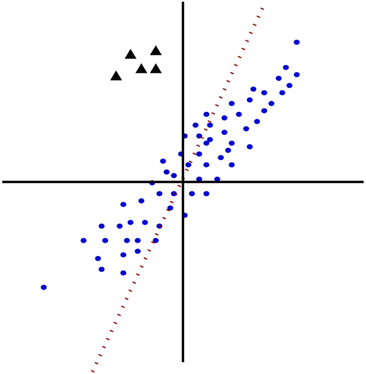


Figure: With Outliers

- Some observations about the corrupted points:
  - ~~They have a large magnitude~~.
  - ~~They have a large (Mahalanobis) distance~~.
  - ~~They increase the volume of the smallest containing ellipsoid~~.

Our Goal: Robust PCA

- Want robustness to arbitrarily corrupted data.
- One measure: Breakdown point
- Instead: bounded error measure between true PCs and output PCs.
- Bound will depend on:
  - Fraction of outliers.
  - Tails of true distribution.

Problem Setup

- "Authentic Samples" $\mathbf{z}_1, \cdots, \mathbf{z}_t \in \mathbb{R}^m$: $\mathbf{z}_i = A\mathbf{x}_i + \mathbf{n}_i$,

Problem Setup

- "Authentic Samples" $\mathbf{z}_1, \cdots, \mathbf{z}_t \in \mathbb{R}^m$: $\mathbf{z}_i = A\mathbf{x}_i + \mathbf{n}_i$,
  - $\mathbf{x}_i \in \mathbb{R}^d$. $\mathbf{x}_i \sim \mu$,
  - $\mathbf{n}_i \in \mathbb{R}^m$. $\mathbf{n}_i \sim \mathcal{N}(\mathbf{0}, I_m)$,
  - $A \in \mathbb{R}^{d \times m}$ and μ unknown. μ mean zero, covariance I.

Problem Setup

- "Authentic Samples" $\mathbf{z}_1, \cdots, \mathbf{z}_t \in \mathbb{R}^m$: $\mathbf{z}_i = A\mathbf{x}_i + \mathbf{n}_i$,
  - $\mathbf{x}_i \in \mathbb{R}^d$. $\mathbf{x}_i \sim \mu$,
  - $\mathbf{n}_i \in \mathbb{R}^m$. $\mathbf{n}_i \sim \mathcal{N}(\mathbf{0}, I_m)$,
  - $A \in \mathbb{R}^{d\times m}$ and μ unknown. μ mean zero, covariance I.

- The "Outliers" $\mathbf{o}_1, \cdots, \mathbf{o}_{n-t} \in \mathbb{R}^m$: generated arbitrarily.

Problem Setup

- "Authentic Samples" $\mathbf{z}_1, \cdots, \mathbf{z}_t \in \mathbb{R}^m$: $\mathbf{z}_i = A\mathbf{x}_i + \mathbf{n}_i$,
  - $\mathbf{x}_i \in \mathbb{R}^d$. $\mathbf{x}_i \sim \mu$,
  - $\mathbf{n}_i \in \mathbb{R}^m$. $\mathbf{n}_i \sim \mathcal{N}(\mathbf{0}, I_m)$,
  - $A \in \mathbb{R}^{d \times m}$ and μ unknown. μ mean zero, covariance I.

- The "Outliers" $\mathbf{o}_1, \cdots, \mathbf{o}_{n-t} \in \mathbb{R}^m$: generated arbitrarily.

- Observe: $\mathcal{Y} \triangleq \{\mathbf{y}_1 \cdots, \mathbf{y}_n\} = \{\mathbf{z}_1, \cdots, \mathbf{z}_t\} \bigcup \{\mathbf{o}_1, \cdots, \mathbf{o}_{n-t}\}$.

Problem Setup

- "Authentic Samples" $\mathbf{z}_1, \cdots, \mathbf{z}_t \in \mathbb{R}^m$: $\mathbf{z}_i = A\mathbf{x}_i + \mathbf{n}_i$,
  - $\mathbf{x}_i \in \mathbb{R}^d$. $\mathbf{x}_i \sim \mu$,
  - $\mathbf{n}_i \in \mathbb{R}^m$. $\mathbf{n}_i \sim \mathcal{N}(\mathbf{0}, I_m)$,
  - $A \in \mathbb{R}^{d \times m}$ and μ unknown. μ mean zero, covariance I.

- The "Outliers" $\mathbf{o}_1, \cdots, \mathbf{o}_{n-t} \in \mathbb{R}^m$: generated arbitrarily.

- Observe: $\mathcal{Y} \triangleq \{\mathbf{y}_1 \cdots, \mathbf{y}_n\} = \{\mathbf{z}_1, \cdots, \mathbf{z}_t\} \bigcup \{\mathbf{o}_1, \cdots, \mathbf{o}_{n-t}\}$.

- Regime of interest:
  - $n \approx m >> d$
  - $\sigma = ||A^\top A|| >> 1$ (scales slowly).
- Objective: Retrieve A

Outline

Features of the High Dimensional regime

- Noise Explosion in High Dimensions: noise magnitude scales faster than the signal noise;

- SNR goes to zero
  - If $\mathbf{n} \sim N(0, I_m)$, then $\mathbb{E}||\mathbf{n}||_2 = \sqrt{m}$, with very sharp concentration.
  - Meanwhile: $\mathbf{E}||Ax||_2 \leq \sigma\sqrt{d}$.

- Consequences:
  - Magnitude of true samples may be much bigger than outlier magnitude.

  - The direction of each sample will be approximately orthogonal to the direction of the signal;

Features of the High Dimensional regime: Pictures

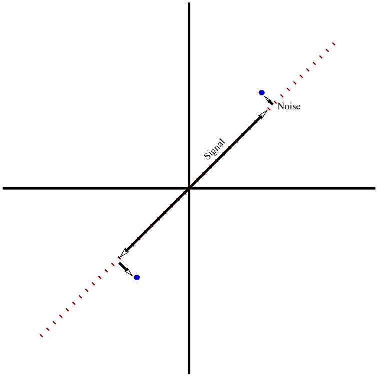


Figure: Recall low-dimensional regime

Features of the High Dimensional regime: Pictures

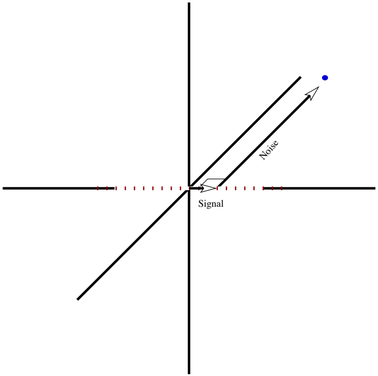


Figure: High dimensions are different: Noise $>>$ Signal

Features of the High Dimensional regime: Pictures

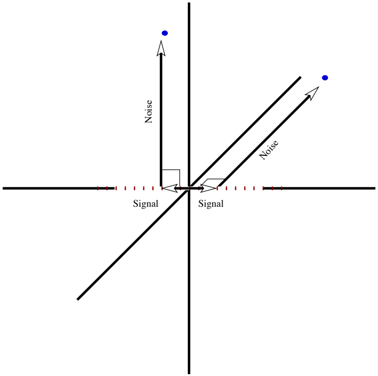


Figure: High dimensions are different: Noise $>>$ Signal

Features of the High Dimensional regime: Pictures

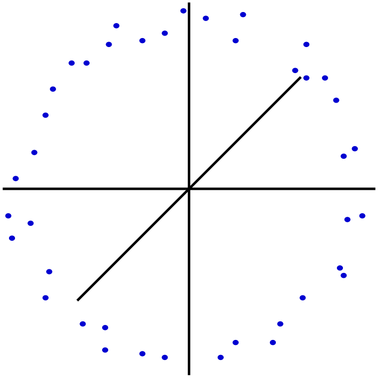


Figure: Every point equidistant from origin and from other points!

Features of the High Dimensional regime: Pictures

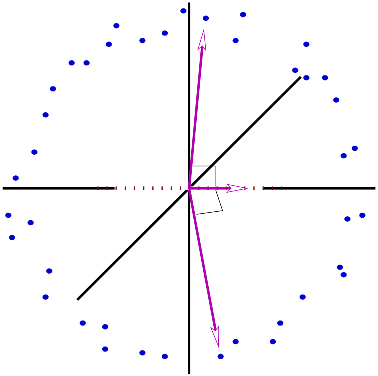


Figure: And every point perpendicular to signal space

Trouble in High Dimensions

- Some approaches that will not work:

- Leave-one-out (more generally, subsample, compare):

  - Either sample size very small: problem

  or

  - Have many corrupted points in each subsample: problem

Trouble in High Dimensions

- Some approaches that will not work:

- Leave-one-out (more generally, subsample, compare):

  - Either sample size very small: problem

    or

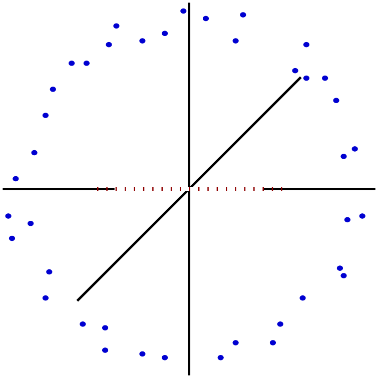
  - Have many corrupted points in each subsample: problem

- Standard Robust PCA: PCA on a robust estimation of the covariance
  - Consistency requires $\#(\text{observations}) \gg \#(\text{dimension})$
  - Not enough observations in high-dimensional case

Trouble in High Dimensions

- Some more approaches that will not work:

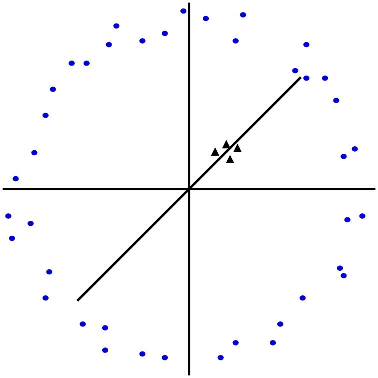
- Removing points with large magnitude



Trouble in High Dimensions

- Some more approaches that will not work:

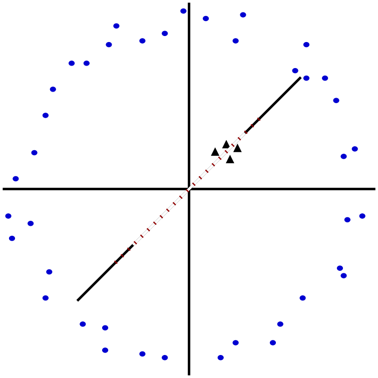
- Removing points with large magnitude



Trouble in High Dimensions

- Some more approaches that will not work:

- Removing points with large magnitude



Trouble in High Dimensions

- Some more approaches that will not work:

- Removing points with large magnitude

- Remove points with large Mahalanobis distance
  - Same example: All λn corrupted points: aligned, length $O(\sigma) << \sqrt{m}$.
  - Very large impact on PCA output.
  - But: Mahalanobis distance of outliers very small.

Trouble in High Dimensions

- Some more approaches that will not work:

- Removing points with large magnitude

- Remove points with large Mahalanobis distance
  - Same example: All λn corrupted points: aligned, length $O(\sigma) << \sqrt{m}$.
  - Very large impact on PCA output.
  - But: Mahalanobis distance of outliers very small.

- Remove points with large Stahel-Donoho distance

$$u_i \triangleq \sup_{\|\mathbf{w}\|=1} \frac{|\mathbf{w}^\top \mathbf{y}_i - \mathrm{med}_j(\mathbf{w}^\top \mathbf{y}_j)|}{\mathrm{med}_k|\mathbf{w}^\top \mathbf{y}_k - \mathrm{med}_j(\mathbf{w}^\top \mathbf{y}_j)|}.$$

  - Same example: impact large, but Stahel-Donoho outlyingness small.

Trouble in High Dimensions

- For these reasons: Some robust covariance estimators have breakdown point = $O(1/m)$, m = dimensions.
  - M-estimator,
  - Convex peeling, Ellipsoidal Peeling,
  - Classical outlier rejection
  - Iterative deletion, iterative trimming,
  - and others...
- These approaches cannot work in high-dimensional regime.

Trouble in High Dimensions

- Algorithmic Tractability

Trouble in High Dimensions

- Algorithmic Tractability
- Minimum volume ellipsoid; Minimum covariance determinant:

Trouble in High Dimensions

- Algorithmic Tractability

- Minimum volume ellipsoid; Minimum covariance determinant:
  - Ill-posed: many zero-volume ellipsoids containing data
  - Intractable: removing a fraction of points combinatorial.

Trouble in High Dimensions

- Algorithmic Tractability

- Minimum volume ellipsoid; Minimum covariance determinant:
  - Ill-posed: many zero-volume ellipsoids containing data
  - Intractable: removing a fraction of points combinatorial.
- Projection pursuit – maximize univariate estimator
  - Problems are non-convex: Intractable.

Trouble in High Dimensions

- Algorithmic Tractability

- Minimum volume ellipsoid; Minimum covariance determinant:
  - Ill-posed: many zero-volume ellipsoids containing data
  - Intractable: removing a fraction of points combinatorial.
- Projection pursuit – maximize univariate estimator
  - Problems are non-convex: Intractable.
  - Choosing subset of directions generated by points: authentic points $\perp$ to signal space, hence no good in high dimensions.

Outline

1. Motivation
2. Where things get tricky
3. **HR-PCA: the algorithm**
4. The Proof Ideas (and some details)
5. Conclusion

High-dimensional Robust PCA: Main Idea

- Get candidate directions from standard PCA (get **w**).
- Project, and use a robust variance estimator: variance of points nearer origin.
    - Outliers can be near origin. But: impact controlled.
- Random removal of “strange" points.

High-dimensional Robust PCA: Main Idea

- Get candidate directions from standard PCA (get **w**).
- Project, and use a robust variance estimator: variance of points nearer origin.
    - Outliers can be near origin. But: impact controlled.
- Random removal of “strange" points.

- Desired properties of an algorithm:
    - Tractable (same complexity as standard PCA);
    - Robust to outliers: performance guarantees;
    - Asymptotically optimal: $t = o(n)$ perfect recovery.
    - Easily kernelizable;

Problem Setup

- "Authentic Samples" $\mathbf{z}_1, \cdots, \mathbf{z}_t \in \mathbb{R}^m$: $\mathbf{z}_i = A\mathbf{x}_i + \mathbf{n}_i$,
    - $\mathbf{x}_i \in \mathbb{R}^d$. $\mathbf{x}_i \sim \mu$,
    - $\mathbf{n}_i \in \mathbb{R}^m$. $\mathbf{n}_i \sim \mathcal{N}(\mathbf{0}, I_m)$,
    - $A \in \mathbb{R}^{d \times m}$ and μ unknown. μ mean zero, covariance I.

- The "Outliers" $\mathbf{o}_1, \cdots, \mathbf{o}_{n-t} \in \mathbb{R}^m$: generated arbitrarily.

- Observe: $\mathcal{Y} \triangleq \{\mathbf{y}_1 \cdots, \mathbf{y}_n\} = \{\mathbf{z}_1, \cdots, \mathbf{z}_t\} \bigcup \{\mathbf{o}_1, \cdots, \mathbf{o}_{n-t}\}$.
- Assumptions:
    - n, m scale to infinity together;
    - $\sigma = ||A^\top A||$ "big" (scales to infinity slowly);
    - μ: spherically symmetric; abs continuous; exponential tails.

Objective & Performance Measurement

- For output PCs $\mathbf{w}_1, \cdots, \mathbf{w}_d$, "Expressed Variance" w.r.t. $\mathbf{w}_1^{\text{true}}, \cdots, \mathbf{w}_d^{\text{true}}$

$$E_V(\mathbf{w}_1, \cdots, \mathbf{w}_d) \triangleq \frac{\sum_{i=1}^d \mathbf{w}_i^\top AA^\top \mathbf{w}_i}{\sum_{i=1}^d (\mathbf{w}_i^{\text{true}})^\top AA^\top \mathbf{w}_i^{\text{true}}} \leq 1.$$

- $E_V = 1$ if the subspace spanned by true PCs is recovered.

- For $d = 1$, $E_V(\mathbf{w}_1) = \cos^2(\angle \mathbf{w}_1, \mathbf{w}_1^{\text{true}})$.

A Robust Variance Estimator

- **Robust Variance Estimator**: $\overline{V}_{\hat{t}}(\mathbf{w}) \triangleq \frac{1}{n}\sum_{i=1}^{\hat{t}} |\mathbf{w}^\top \mathbf{y}|^2_{(i)}$.

- Order statistics: $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, then $\alpha_{(1)} \leq \alpha_{(2)} \leq \cdots \leq \alpha_{(n)}$.

- Idea: If outliers small, their impact is controlled.

The HR-PCA Algorithm

(1) Perform PCA on empirical covariance.

(2) If robust variance estimate in PC directions highest yet, record it, and PCs.

(3) Randomly remove a point in proportion to its variance along PCs.

(4) Repeat until “enough" points removed.

(5) Output the last PCs recorded.

The HR-PCA Algorithm

(1) Perform PCA on empirical covariance: $\{\mathbf{w}_1, \ldots, \mathbf{w}_d\}$.

(2) Compute $\mathfrak{b} = \mathrm{RVE}(\{\mathbf{w}_1, \ldots, \mathbf{w}_d\})$. If $\mathfrak{b} > \mathfrak{b}^*$,
- Update $\mathfrak{b}^* = \mathfrak{b}$
- Update $\{\mathbf{w}_1^*, \ldots, \mathbf{w}_d^*\} = \{\mathbf{w}_1, \ldots, \mathbf{w}_d\}$.

(3) Randomly remove a point in proportion to its variance along PCs.

(4) Repeat until all points removed.

(5) Output the last PCs recorded: $\{\mathbf{w}_1^*, \ldots, \mathbf{w}_d^*\}$.

The HR-PCA Algorithm: Pitfalls

- Things that can go wrong:

The HR-PCA Algorithm: Pitfalls

- Things that can go wrong:

∗ Remove authentic points

∗ May not ultimately report "best outcome."

∗ Corrupted points may contribute to ultimately reported PCs.

The HR-PCA Algorithm: Pitfalls

- Things that can go wrong:

  ∗ Remove authentic points

  ∗ May not ultimately report “best outcome.”

  ∗ Corrupted points may contribute to ultimately reported PCs.

- But: we show the error due to all such factors is controlled.

The Guarantees: Finite Sample + Asymptotic

- Results will depend on:
  - Fraction of outliers: λ.
  - Tails of μ.
- Define: $\mathcal{V} : [0,1] \to [0,1]$

$$\mathcal{V}(\alpha) = \int_{-c_\alpha}^{c_\alpha} x^2 \overline{\mu}(dx).$$

The Guarantees: Finite Sample + Asymptotic

Theorem: The following holds in probability (n, m, σ scale):

$$\text{E.V.(output)} \geq \max_{\kappa} \left[\frac{\mathcal{V}\left(1 - \frac{\lambda^*(1+\kappa)}{(1-\lambda^*)\kappa}\right)}{(1+\kappa)} \right] \times \left[\frac{\mathcal{V}\left(\frac{\hat{t}}{t} - \frac{\lambda^*}{1-\lambda^*}\right)}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)} \right].$$

The Guarantees: Finite Sample + Asymptotic

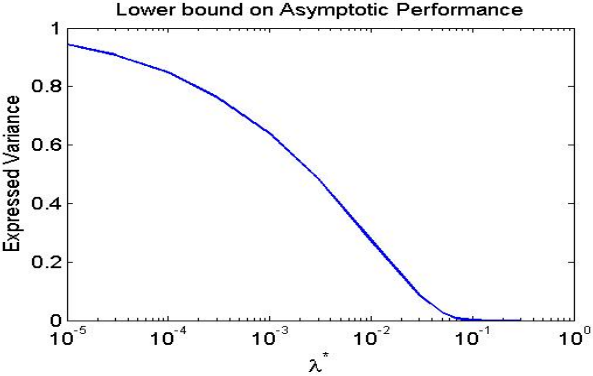
Theorem: The following holds in probability (n, m, σ scale):

$$\text{E.V.(output)} \geq \max_{\kappa} \left[\frac{\mathcal{V}\left(1 - \frac{\lambda^*(1+\kappa)}{(1-\lambda^*)\kappa}\right)}{(1+\kappa)} \right] \times \left[\frac{\mathcal{V}\left(\frac{\hat{t}}{t} - \frac{\lambda^*}{1-\lambda^*}\right)}{\mathcal{V}\left(\frac{\hat{t}}{t}\right)} \right].$$

- The Bound:
  - Term 1: May not remove all outliers, and some authentic points may be removed.
  - Term 2: May have small outliers that alter PC directions.
- If $t = o(n)$, RHS = 1: optimal recovery.
- Breakdown point: $1/2$.

Asymptotic Performance Guarantee

E.V. is lower bounded by



If the *proportion* of outliers goes to zero: the Expressed Variance equals 1.

Proof Idea

(1) "Blessing of dimensionality": empirical covariance estimates good, even for high-dimensional regime;

(2) Random removal: have a "good" solution, or outlier is removed with large probability;

(3) Therefore: at some early iteration, algorithm finds a "good" solution.

(4) Output of algorithm has higher robust variance estimate than the "good" solution. We show output must then also be (almost as) "good."

Proof Idea - Step 1

With high probability:

(1.a) Largest eigenvalue of the empirical noise covariance matrix is bounded:

$$\sup_{\mathbf{w}\in\mathcal{S}_m} \frac{1}{n}\sum_{i=1}^{t}(\mathbf{w}^\top \mathbf{n}_i)^2 \leq c.$$

(1.b) Largest eigenvalue of the signals in original space converges to 1:

$$\sup_{\mathbf{w}\in\mathcal{S}_d} |\frac{1}{t}\sum_{i=1}^{t}(\mathbf{w}^\top \mathbf{x}_i)^2 - 1| \leq \epsilon.$$

Proof Idea - Step 1

(1.c) RVE is a valid variance estimator for the $d-$dimensional signals $\mathbf{x}$:

$$\sup_{\mathbf{w}\in\mathcal{S}_d} \left|\frac{1}{t}\sum_{i=1}^{\hat{t}} |\mathbf{w}^\top \mathbf{x}|^2_{(i)} - \mathcal{V}\left(\frac{\hat{t}}{t}\right)\right| \le \epsilon.$$

(1.d) RVE is a valid estimator of the variance of the authentic samples, $\mathbf{z} = A\mathbf{x} + \mathbf{n}$: uniformly over all $\mathbf{w} \in \mathcal{S}_m$,

$$(1-\epsilon)\|\mathbf{w}^\top A\|^2 \mathcal{V}\left(\frac{t'}{t}\right) - c\|\mathbf{w}^\top A\| \le \frac{1}{t}\sum_{i=1}^{t'} |\mathbf{w}^\top \mathbf{z}|^2_{(i)} \le$$

$$(1+\epsilon)\|\mathbf{w}^\top A\|^2 \mathcal{V}\left(\frac{t'}{t}\right) + c\|\mathbf{w}^\top A\|.$$

Proof - Step 1.a - details

(1.a) Largest eigenvalue of the variance of noise matrix is bounded:

$$\sup_{\mathbf{w} \in \mathcal{S}_m} \frac{1}{n} \sum_{i=1}^{t} (\mathbf{w}^\top \mathbf{n}_i)^2 \leq c.$$

- Two keys: "blessing of dimensionality" and uniform laws of large numbers.

Proof - Step 1.a - details

(1.a) Largest eigenvalue of the variance of noise matrix is bounded:

$$\sup_{\mathbf{w} \in \mathcal{S}_m} \frac{1}{n} \sum_{i=1}^{t} (\mathbf{w}^\top \mathbf{n}_i)^2 \leq c.$$

- Two keys: "blessing of dimensionality" and uniform laws of large numbers.
- Step 1 (a): Need basic Lemma:
- **Lemma**: For Γ a $m \times t$ matrix ($m \leq t$), $\Gamma_{ij} \sim \mathcal{N}(0, 1)$, i.i.d.:

$$\Pr\left(\sigma_{\max}(\Gamma) > \sqrt{m} + \sqrt{t} + \sqrt{t}\epsilon\right) \leq \exp(-t\epsilon^2/2).$$

Proof - Step 1.a - details

(1.a) Largest eigenvalue of the variance of noise matrix is bounded:

$$\sup_{\mathbf{w} \in \mathcal{S}_m} \frac{1}{n} \sum_{i=1}^{t} (\mathbf{w}^\top \mathbf{n}_i)^2 \leq c.$$

- Two keys: “blessing of dimensionality” and uniform laws of large numbers.
- Step 1 (a): Need basic Lemma:
- **Lemma**: For Γ a $m \times t$ matrix ($m \leq t$), $\Gamma_{ij} \sim \mathcal{N}(0, 1)$, i.i.d.:

$$\Pr\left(\sigma_{\max}(\Gamma) > \sqrt{m} + \sqrt{t} + \sqrt{t}\epsilon\right) \leq \exp(-t\epsilon^2/2).$$

- **Observation**:

$$\sup_{\mathbf{w} \in \mathcal{S}_m} \frac{1}{t} \sum_{i=1}^{t} (\mathbf{w}^\top \mathbf{n}_i)^2 = \lambda_{\max}(\Gamma\Gamma^\top)/t = \sigma^2_{\max}(\Gamma)/t.$$

Proof - Step 1.a - An Aside

- Where do these results come from:

- Basic idea: *dimension-free* concentration of measure
- **Theorem**: Let F be L-Lipschitz w.r.t. Euclidean norm, $X \sim N(0, I)$ standard Gaussian measure. M_F the mean of $F(X)$. Then

$$\mathbb{P}(F(X) \geq M_F + \xi) \leq e^{-\xi^2/2L^2}.$$

- Basic observation: $\sigma_{\max}(\cdot) : \mathbb{R}^{n_1 \times n_2} \longrightarrow \mathbb{R}$ is 1-Lipschitz.

- Two nice references: (a) Davidson and Szarek: Operators, Random Matrices & Banach Spaces; (b) Matousek: Lectures on Discrete Geometry.

Proof Idea

(1) “Blessing of dimensionality”: empirical covariance estimates good, even for high-dimensional regime;

(2) Random removal: have a “good” solution, or outlier is removed with large probability;

(3) Therefore: at some early iteration, algorithm finds a “good” solution.

(4) Output of algorithm has higher robust variance estimate than the “good” solution. We show output must then also be (almost as) “good.”

Proof Idea - Step 2

- Let $\mathcal{Z}(s)$, $\mathcal{O}(s)$ be remaining authentic/outlier points.
- Fix $\kappa > 0$ and call step s a "Good Event", $\mathcal{G}(s)$ if:

Proof Idea - Step 2

- Let $\mathcal{Z}(s)$, $\mathcal{O}(s)$ be remaining authentic/outlier points.
- Fix $\kappa > 0$ and call step s a "Good Event", $\mathcal{G}(s)$ if:

$$\sum_{j=1}^{d} \sum_{\mathbf{z}_i \in \mathcal{Z}(s-1)} (\mathbf{w}_j(s)^\top \mathbf{z}_i)^2 \geq \frac{1}{\kappa} \sum_{j=1}^{d} \sum_{\mathbf{o}_i \in \mathcal{O}(s-1)} (\mathbf{w}_j(s)^\top \mathbf{o}_i)^2 \}.$$

Proof Idea - Step 2

- Let $\mathcal{Z}(s)$, $\mathcal{O}(s)$ be remaining authentic/outlier points.
- Fix $\kappa > 0$ and call step s a "Good Event", $\mathcal{G}(s)$ if:

$$\underbrace{\sum_{j=1}^{d} \sum_{\mathbf{z}_i \in \mathcal{Z}(s-1)} (\mathbf{w}_j(s)^\top \mathbf{z}_i)^2}_{\text{variance of authentic pts}} \geq \frac{1}{\kappa} \underbrace{\sum_{j=1}^{d} \sum_{\mathbf{o}_i \in \mathcal{O}(s-1)} (\mathbf{w}_j(s)^\top \mathbf{o}_i)^2}_{\text{variance of corrupted pts}} .$$

Proof Idea - Step 2

- Let $\mathcal{Z}(s)$, $\mathcal{O}(s)$ be remaining authentic/outlier points.
- Fix $\kappa > 0$ and call step s a "Good Event", $\mathcal{G}(s)$ if:

$$\underbrace{\sum_{j=1}^{d} \sum_{\mathbf{z}_i \in \mathcal{Z}(s-1)} (\mathbf{w}_j(s)^\top \mathbf{z}_i)^2}_{\text{variance of authentic pts}} \geq \frac{1}{\kappa} \underbrace{\sum_{j=1}^{d} \sum_{\mathbf{o}_i \in \mathcal{O}(s-1)} (\mathbf{w}_j(s)^\top \mathbf{o}_i)^2}_{\text{variance of corrupted pts}}.$$

- This means: variance on the direction of found PCs is mostly due to the authentic samples.
- Hence: $\{\mathbf{w}_1, \ldots, \mathbf{w}_d\}$ must be close to true PCs.

Proof Idea - Step 2

- Let $\mathcal{Z}(s)$, $\mathcal{O}(s)$ be remaining authentic/outlier points.
- Fix $\kappa > 0$ and call step s a "Good Event", $\mathcal{G}(s)$ if:

$$\underbrace{\sum_{j=1}^{d} \sum_{\mathbf{z}_i \in \mathcal{Z}(s-1)} (\mathbf{w}_j(s)^\top \mathbf{z}_i)^2}_{\text{variance of authentic pts}} \geq \frac{1}{\kappa} \underbrace{\sum_{j=1}^{d} \sum_{\mathbf{o}_i \in \mathcal{O}(s-1)} (\mathbf{w}_j(s)^\top \mathbf{o}_i)^2}_{\text{variance of corrupted pts}}.$$

- This means: variance on the direction of found PCs is mostly due to the authentic samples.
- Hence: $\{\mathbf{w}_1, \ldots, \mathbf{w}_d\}$ must be close to true PCs.
- **Theorem**: If $\mathcal{G}^c(s)$ — step s is not good — then next point removed is an outlier with probability at least $\frac{\kappa}{1+\kappa}$.

Proof Idea

(1) "Blessing of dimensionality": empirical covariance estimates good, even for high-dimensional regime;

(2) Random removal: have a "good" solution, or outlier is removed with large probability;

(3) Therefore: at some early iteration, algorithm finds a "good" solution.

(4) Output of algorithm has higher robust variance estimate than the "good" solution. We show output must then also be (almost as) "good."

Proof Idea - Step 3

- **Theorem**: With high probability, we have a "good event" by time at most $s_0 > \lambda n[(1+\kappa)/\kappa]$.

Proof Idea - Step 3

- **Theorem**: With high probability, we have a "good event" by time at most $s_0 > \lambda n[(1+\kappa)/\kappa]$.

- Intuition: Suppose subsequent steps were independent.
  - Since, "expected number of corrupted points removed each step" is $\kappa/(1+\kappa)$.
  - After M steps, expected corrupted points removed is $M\frac{\kappa}{1+\kappa}$.
  - Therefore: All the outliers removed after $M = \lambda n \frac{1+\kappa}{\kappa}(1+\varepsilon)$ steps, with exponentially high probability.

Proof Idea - Step 3

- **Theorem**: With high probability, we have a "good event" by time at most $s_0 > \lambda n[(1+\kappa)/\kappa]$.

- Intuition: Suppose subsequent steps were independent.
  - Since, "expected number of corrupted points removed each step" is $\kappa/(1+\kappa)$.
  - After M steps, expected corrupted points removed is $M\frac{\kappa}{1+\kappa}$.
  - Therefore: All the outliers removed after $M = \lambda n \frac{1+\kappa}{\kappa}(1+\varepsilon)$ steps, with exponentially high probability.
  - The Problem: not i.i.d.
  - The Fix: use martingales and Azuma-Hoeffding.

Proof Idea - Step 3 - details

- Let $T = \min\{s | \mathcal{G}(s) \text{ is true}\}$.

Proof Idea - Step 3 - details

- Let $T = \min\{s | \mathcal{G}(s) \text{ is true}\}$.
- Define the random variable (w.r.t. natural filtration $\mathcal{F}_s$):

$$X_s = \begin{cases} |\mathcal{O}(T-1)| + \frac{\kappa}{1+\kappa} \cdot (T-1), & \text{if } T \leq s; \\ |\mathcal{O}(s)| + \frac{\kappa}{1+\kappa} \cdot s, & \text{if } T > s. \end{cases}$$

Note: $X_0 = \lambda n$.

Proof Idea - Step 3 - details

- Let $T = \min\{s | \mathcal{G}(s) \text{ is true}\}$.
- Define the random variable (w.r.t. natural filtration $\mathcal{F}_s$):

$$X_s = \begin{cases} |\mathcal{O}(T-1)| + \frac{\kappa}{1+\kappa} \cdot (T-1), & \text{if } T \leq s; \\ |\mathcal{O}(s)| + \frac{\kappa}{1+\kappa} \cdot s, & \text{if } T > s. \end{cases}$$

  Note: $X_0 = \lambda n$.
- **Lemma**: $\{X_s, \mathcal{F}_s\}$ is a supermartingale.

Proof Idea - Step 3 - details

- Let $T = \min\{s | \mathcal{G}(s) \text{ is true}\}$.
- Define the random variable (w.r.t. natural filtration $\mathcal{F}_s$):

$$X_s = \begin{cases} |\mathcal{O}(T-1)| + \frac{\kappa}{1+\kappa} \cdot (T-1), & \text{if } T \leq s; \\ |\mathcal{O}(s)| + \frac{\kappa}{1+\kappa} \cdot s, & \text{if } T > s. \end{cases}$$

  Note: $X_0 = \lambda n$.
- **Lemma**: $\{X_s, \mathcal{F}_s\}$ is a supermartingale.

- Now we have: for $s_0 = \lambda n[(1+\kappa)/\kappa](1+\varepsilon)$

$$\mathbb{P}(T > s_0) \leq \mathbb{P}\left(X_{s_0} \geq \frac{\kappa s_0}{1+\kappa}\right) = \mathbb{P}(X_{s_0} \geq (1+\epsilon)\lambda n)$$

Proof Idea - Step 3 - details

- Let $T = \min\{s | \mathcal{G}(s) \text{ is true}\}$.
- Define the random variable (w.r.t. natural filtration $\mathcal{F}_s$):

$$X_s = \begin{cases} |\mathcal{O}(T-1)| + \frac{\kappa}{1+\kappa} \cdot (T-1), & \text{if } T \leq s; \\ |\mathcal{O}(s)| + \frac{\kappa}{1+\kappa} \cdot s, & \text{if } T > s. \end{cases}$$

  Note: $X_0 = \lambda n$.
- **Lemma**: $\{X_s, \mathcal{F}_s\}$ is a supermartingale.

- Now we have: for $s_0 = \lambda n[(1+\kappa)/\kappa](1+\varepsilon)$

$$\mathbb{P}(T > s_0) \leq \mathbb{P}\left(X_{s_0} \geq \frac{\kappa s_0}{1+\kappa}\right) = \mathbb{P}(X_{s_0} \geq (1+\epsilon)\lambda n)$$

- Azuma-Hoeffding completes the proof.

Proof Idea

(1) "Blessing of dimensionality": empirical covariance estimates good, even for high-dimensional regime;

(2) Random removal: have a "good" solution, or outlier is removed with large probability;

(3) Therefore: at some early iteration, algorithm finds a "good" solution.

(4) Output of algorithm has higher robust variance estimate than the "good" solution. We show output must then also be (almost as) "good."

Proof Idea - Step 4

- Putting it all together:

- An early iteration produces directions $\hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_d$ that have "most of" the variance.

- Bound quality on these directions:

$$E_V(\hat{\mathbf{w}}_1, \cdots, \hat{\mathbf{w}}_d) \triangleq \frac{\sum_{i=1}^d \hat{\mathbf{w}}_i^\top AA^\top \hat{\mathbf{w}}_i}{\sum_{i=1}^d (\mathbf{w}_i^{\text{true}})^\top AA^\top \mathbf{w}_i^{\text{true}}}.$$

- The final algorithm only produces directions $\mathbf{w}_1^*, \ldots, \mathbf{w}_d^*$ with biggest robust variance estimator.

- Bound quality on these directions:

$$E_V(\mathbf{w}_1^*, \cdots, \mathbf{w}_d^*) \triangleq \frac{\sum_{i=1}^d (\mathbf{w}_i^*)^\top AA^\top \mathbf{w}_i^*}{\sum_{i=1}^d \sum_{i=1}^d \hat{\mathbf{w}}_i^\top AA^\top \hat{\mathbf{w}}_i}.$$

Kernelization

- Using a kernel function $k(\cdot, \cdot)$ to represent a feature mapping $\Upsilon(\cdot)$
- PCA can be kernelized using Kernel PCA, with output in a form $\mathbf{v}_q = \sum_{i=1}^{n-s} \alpha_i(q)\Upsilon(\hat{\mathbf{y}}_i), \quad q = 1, \cdots, d$.
- HR-PCA Algorithm requires:
  - Computing PCA;
  - Computing Robust Variance Estimator;
- Both steps can be done.

Conclusion

- Methodology for handling dimensionality reduction when:
  1. $\#$(Observation) $\sim$ $\#$(Dimension)
  2. $\#$(*Outliers*) is “large"
- The key idea: verify projections statistics behave in a certain way, if not - probabilistic point removal
- Works well in simulations

On the todo list:

- Generalize to other identification problems with outliers: when a probabilistic model is available
- Extend to stochastic programming with corrupted sampled data
- Looking for an online algorithm.