Robust PCA for High-Dimensional Data

Constantine Caramanis

The University of Texas at Austin
Department of Electrical and Computer Engineering
Wireless Networking and Communications Group

December 11, 2010

Two Problems

- PCA + Outliers
- Collaborative Filtering + Manipulators.

Theme: High-Dimensional Data

- High-dimensional data:
  #dimensionality $\approx$# observations.

Theme: High-Dimensional Data

- High-dimensional data:
  #dimensionality $\approx$# observations.
- DNA microarray, financial data, semantic indexing, images, etc
- Networks: user-behavior-aware network algorithms (Cognitive Networks)?

Theme: High-Dimensional Data

- High-dimensional data:
  #dimensionality $\approx$# observations.
- DNA microarray, financial data, semantic indexing, images, etc
- Networks: user-behavior-aware network algorithms (Cognitive Networks)?
- Traditional statistical tools do not work

Theme: High-Dimensional Data

- High-dimensional data:
  #dimensionality $\approx$# observations.
- DNA microarray, financial data, semantic indexing, images, etc
- Networks: user-behavior-aware network algorithms (Cognitive Networks)?
- Traditional statistical tools do not work for me

This Talk

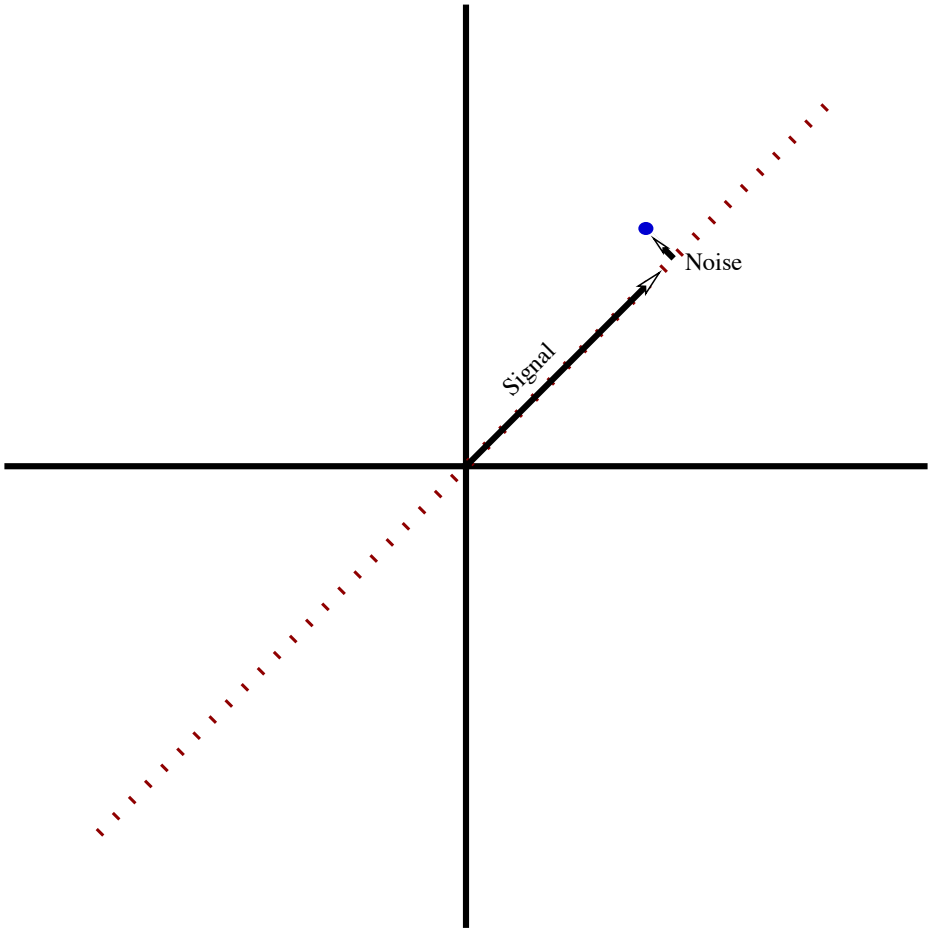
- Robust PCA, take 1
- Pitfalls in high dimensions
- Robust PCA, take 2
- Robust collaborative filtering

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

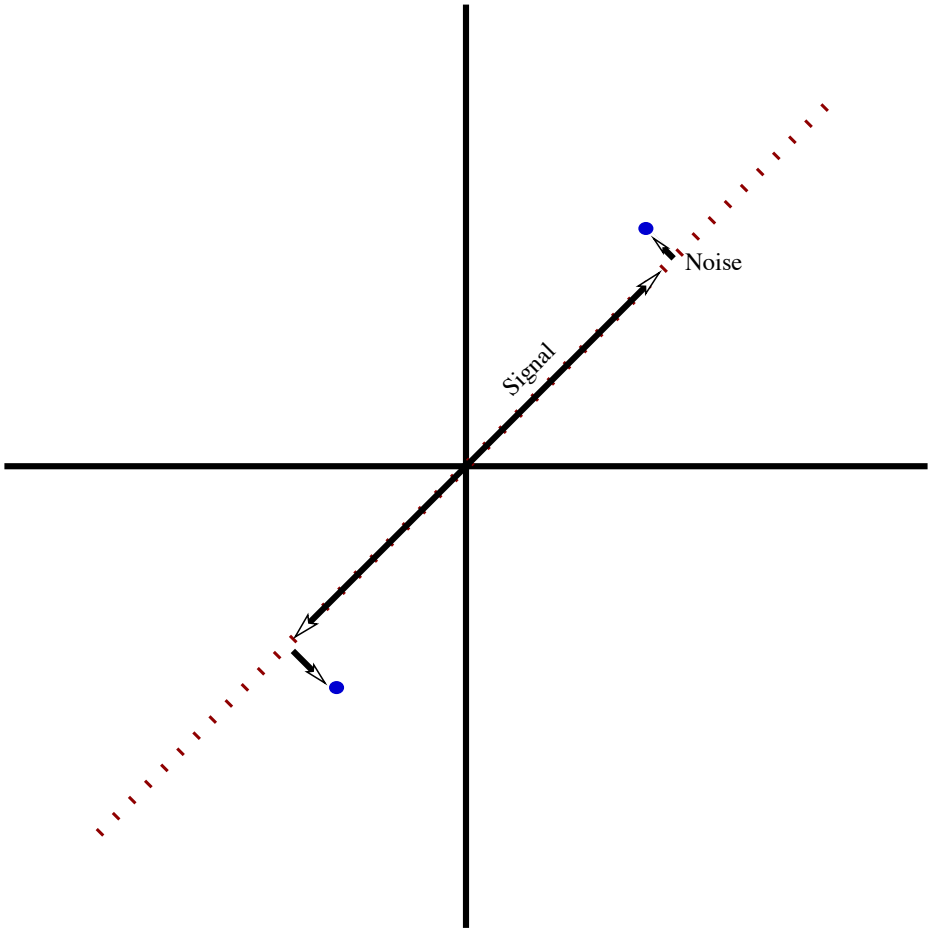
PCA - in Pictures

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



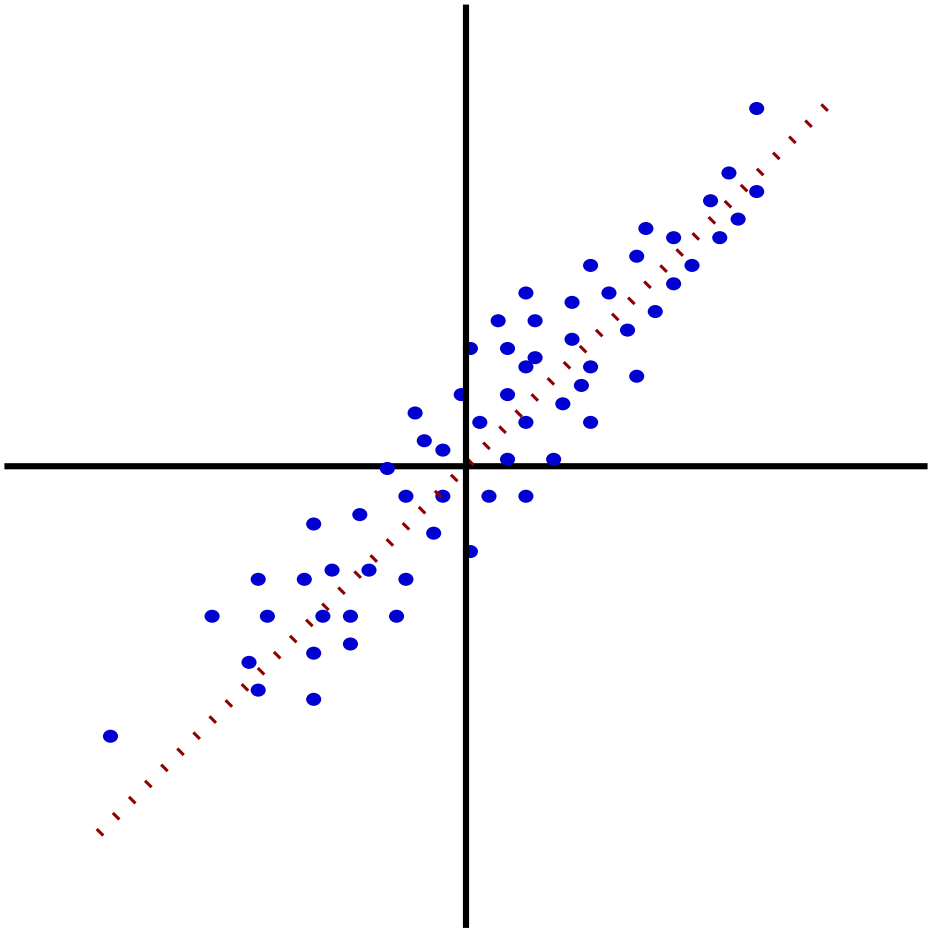
PCA - in Pictures

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



PCA - in Pictures

Observe points: $\mathbf{y} = A\mathbf{x} + \mathbf{v}$.
Goal: Find least-square-error subspace approximation.



PCA - in Math

- Least-square-error subspace approximation

  How: Singular value decomposition (SVD)

PCA - in Math

- Least-square-error subspace approximation
  How: Singular value decomposition (SVD)
- 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)

- 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.

Corrupted Data

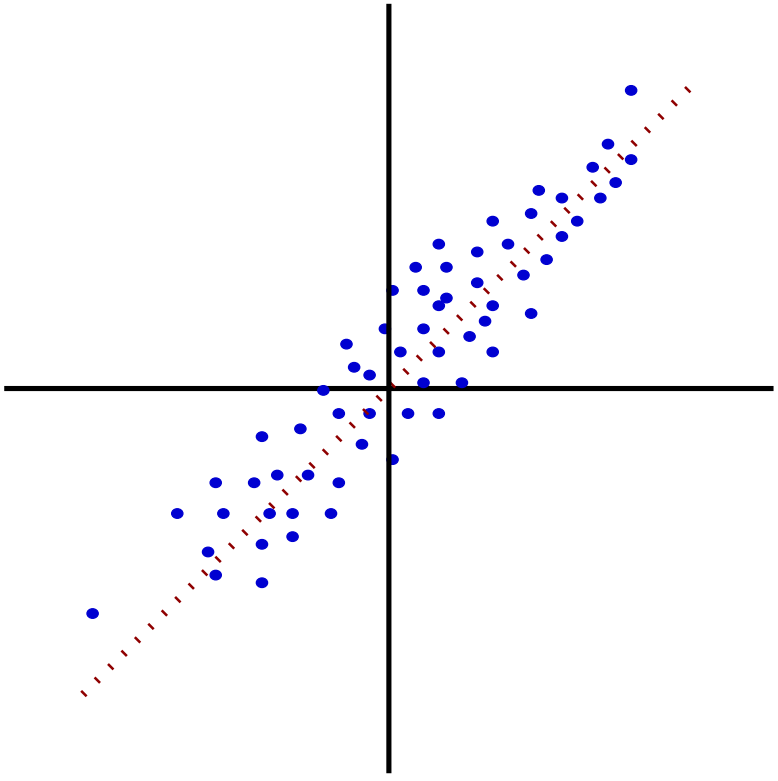


Figure: No Outliers

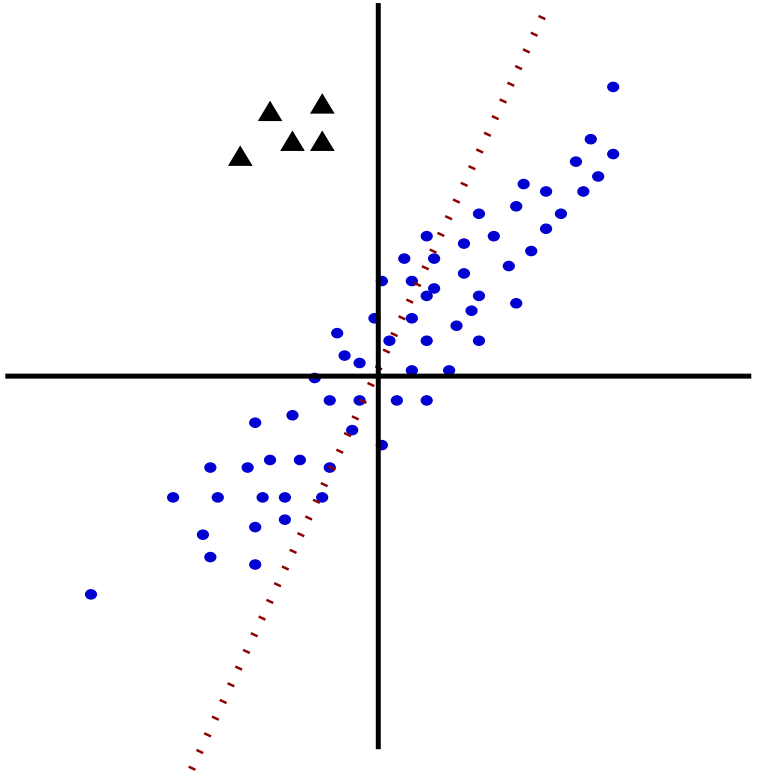


Figure: With Outliers

Corrupted Data

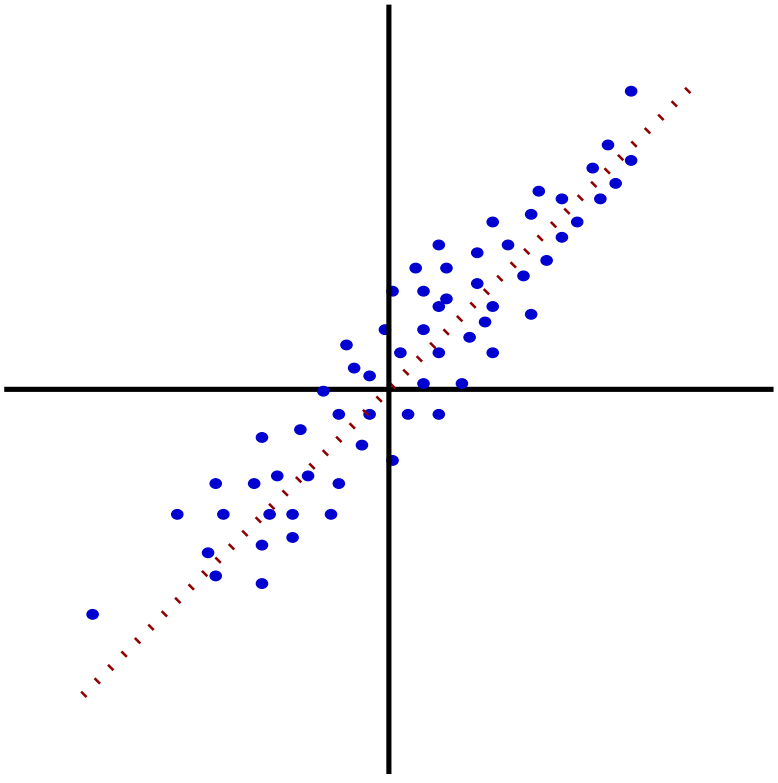


Figure: No Outliers

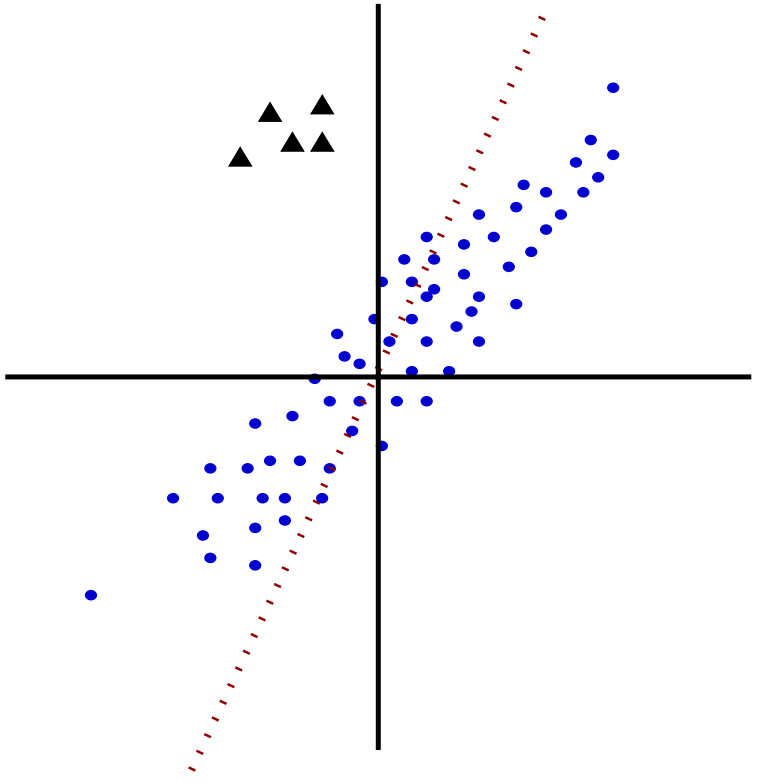


Figure: With Outliers

- 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

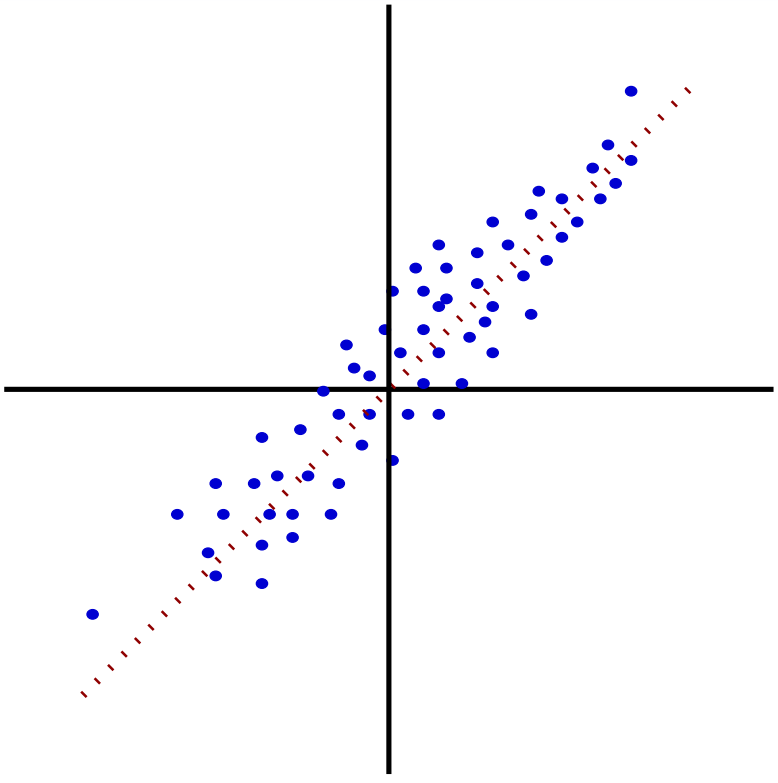


Figure: No Outliers

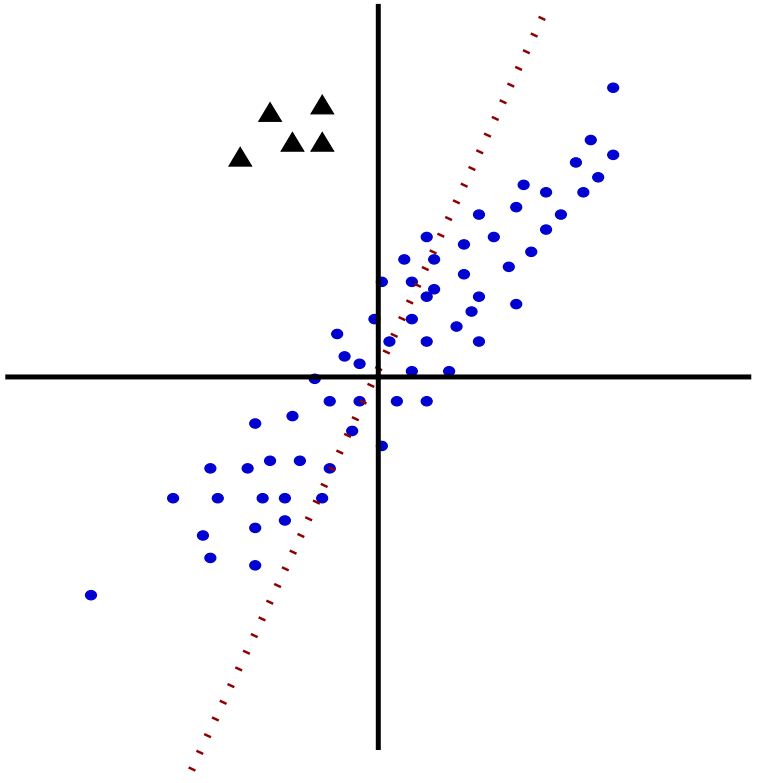


Figure: With Outliers

- 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~~.

High-dimensional Robust PCA: What We Want

- Tractable (same complexity as standard PCA);
- Robust to outliers: Breakdown point
- More: performance guarantees;
- Asymptotically optimal: $t = o(n)$ perfect recovery.
- Easily kernelizable;

Problem Setup

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

Problem Setup

- "Authentic Samples" $\mathbf{z}_1, \cdots, \mathbf{z}_t \in \mathbb{R}^m$: $\mathbf{z}_i = A\mathbf{x}_i + \mathbf{n}_i$,
  - $\mathbf{x}_i \in \mathbb{R}^d$.
  - $\mathbf{n}_i \in \mathbb{R}^p$. $\mathbf{n}_i \sim \mathcal{N}(\mathbf{0}, I_p)$,
  - $A \in \mathbb{R}^{p \times d}$

Problem Setup

- "Authentic Samples" $\mathbf{z}_1, \cdots, \mathbf{z}_t \in \mathbb{R}^m$: $\mathbf{z}_i = A\mathbf{x}_i + \mathbf{n}_i$,
  - $\mathbf{x}_i \in \mathbb{R}^d$.
  - $\mathbf{n}_i \in \mathbb{R}^p$. $\mathbf{n}_i \sim \mathcal{N}(\mathbf{0}, I_p)$,
  - $A \in \mathbb{R}^{p \times d}$
- 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" $\mathbf{z}_1, \cdots, \mathbf{z}_t \in \mathbb{R}^m$: $\mathbf{z}_i = A\mathbf{x}_i + \mathbf{n}_i$,
  - $\mathbf{x}_i \in \mathbb{R}^d$.
  - $\mathbf{n}_i \in \mathbb{R}^p$. $\mathbf{n}_i \sim \mathcal{N}(\mathbf{0}, I_p)$,
  - $A \in \mathbb{R}^{p \times d}$
- 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 and scaling:
  - $n \approx p >> d$
  - $\sigma = ||A^\top A|| >> 1$ (scales slowly).

Problem Setup

- "Authentic Samples" $\mathbf{z}_1, \cdots, \mathbf{z}_t \in \mathbb{R}^m$: $\mathbf{z}_i = A\mathbf{x}_i + \mathbf{n}_i$,
  - $\mathbf{x}_i \in \mathbb{R}^d$.
  - $\mathbf{n}_i \in \mathbb{R}^p$. $\mathbf{n}_i \sim \mathcal{N}(\mathbf{0}, I_p)$,
  - $A \in \mathbb{R}^{p \times d}$
- 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 and scaling:
  - $n \approx p >> d$
  - $\sigma = ||A^\top A|| >> 1$ (scales slowly).
- Objective: Retrieve column space of A

Features of the High Dimensional regime

- Noise magnitude scales faster than the signal noise;
  - $\mathbf{n} \sim N(0, I_p)$: $\mathbb{E}||\mathbf{n}||_2 = \sqrt{p}$
  - $\mathbf{E}||Ax||_2 \leq \sigma\sqrt{d}$.
- 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

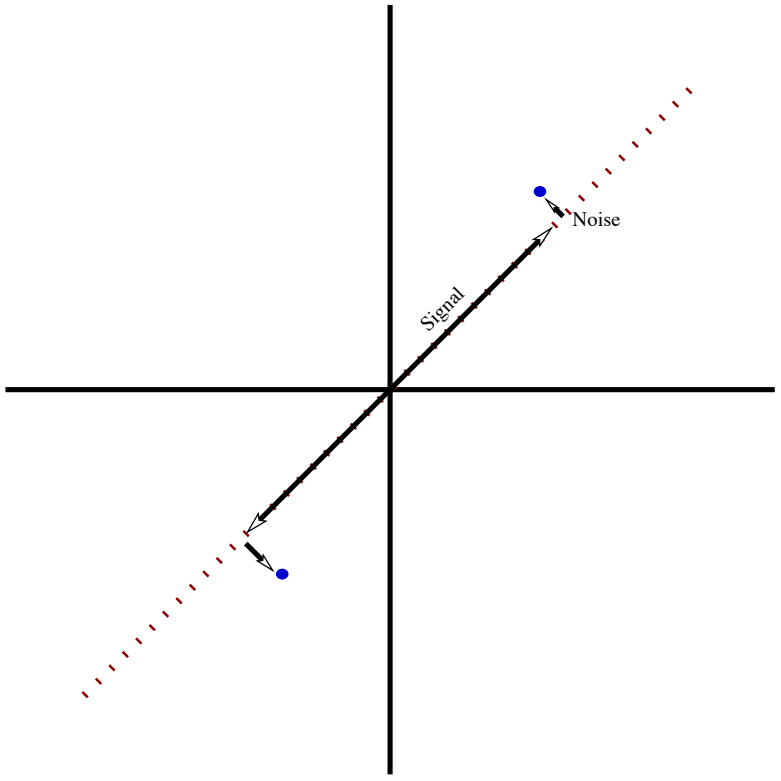


Figure: Recall low-dimensional regime

Features of the High Dimensional regime: Pictures

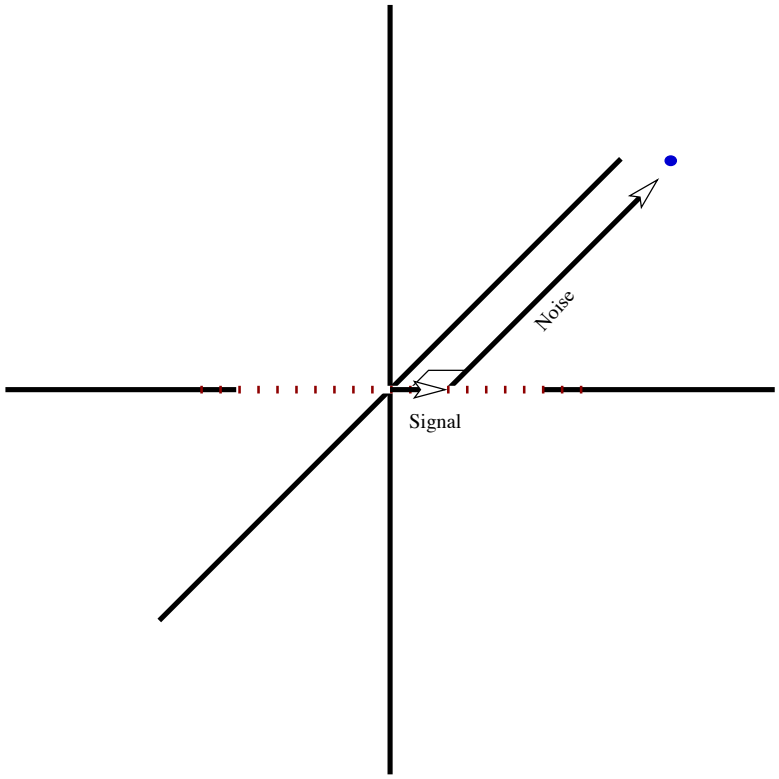


Figure: High dimensions are different: Noise $>>$ Signal

Features of the High Dimensional regime: Pictures

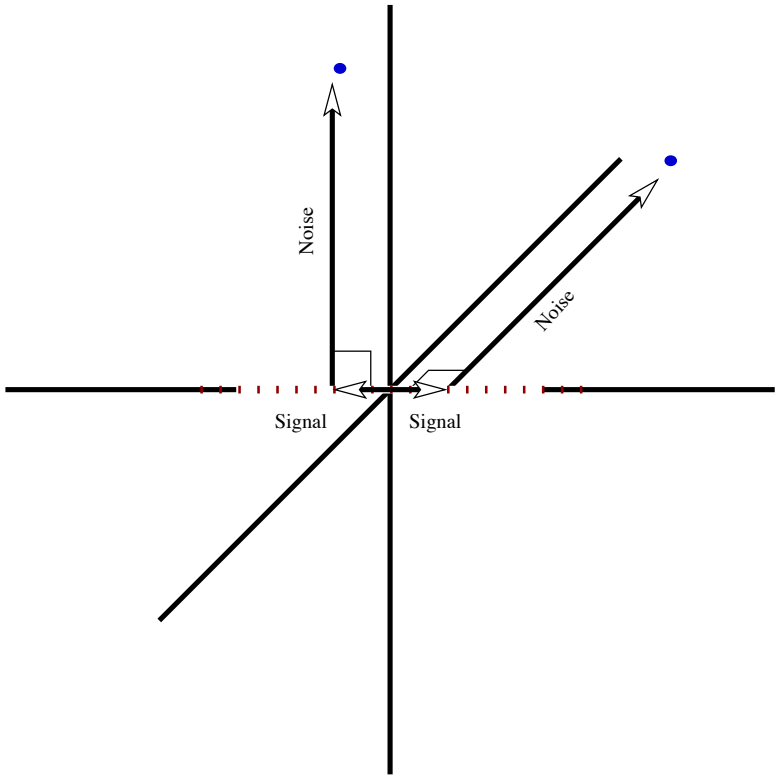


Figure: High dimensions are different: Noise $>>$ Signal

Features of the High Dimensional regime: Pictures

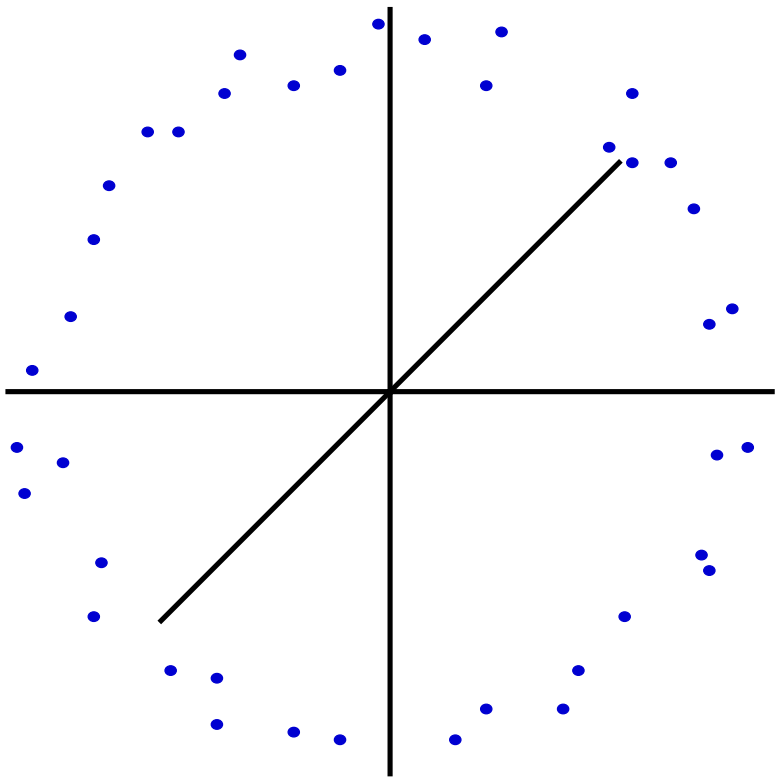


Figure: Every point equidistant from origin and from other points

Features of the High Dimensional regime: Pictures

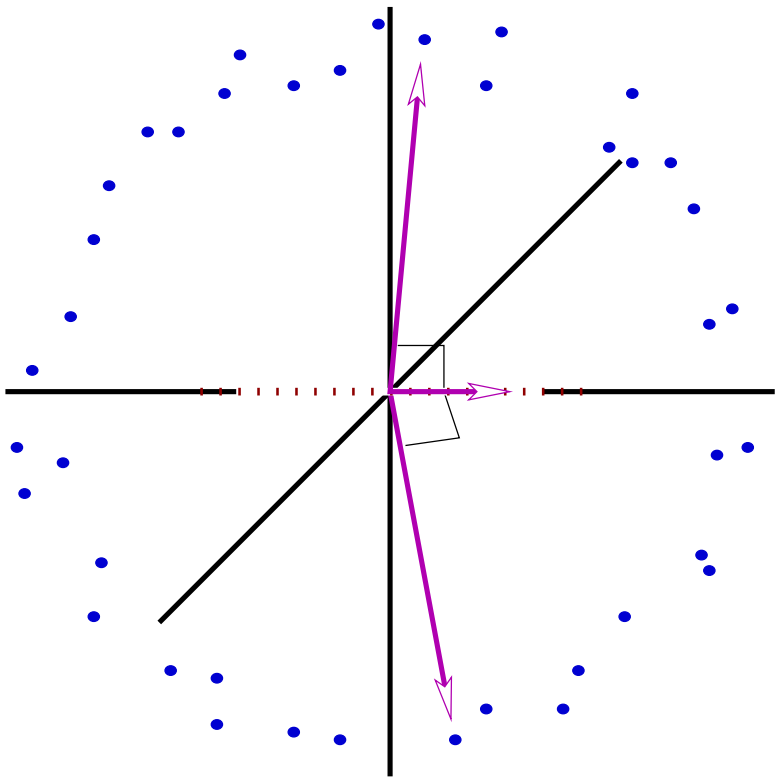


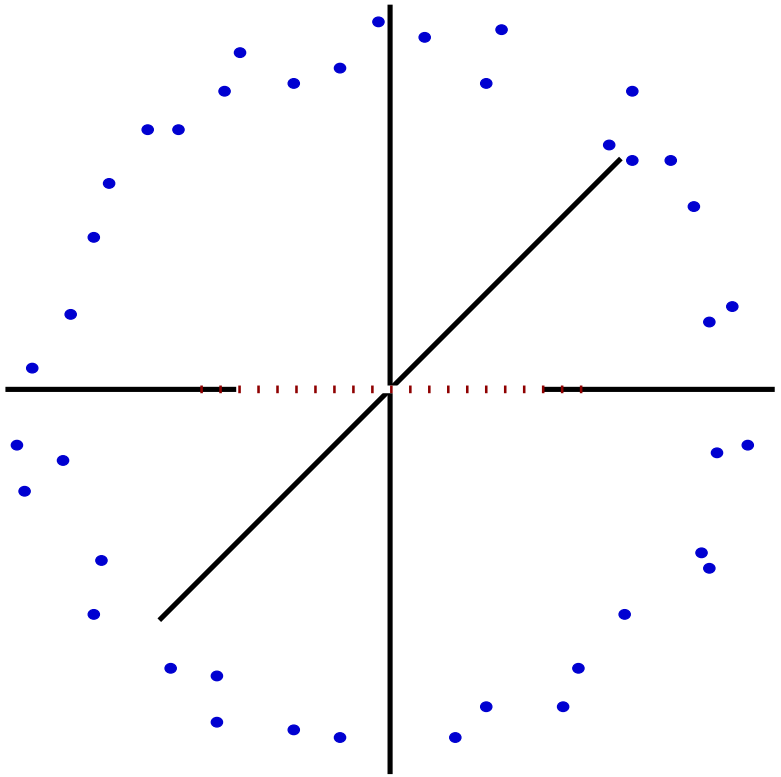
Figure: And every point perpendicular to signal space

Trouble in High Dimensions

- Some approaches that will not work:
- Standard Robust PCA: PCA on a robust estimation of the covariance
- Leave-one-out (more generally, subsample, compare):
- Remove points that have large magnitude (look strange)

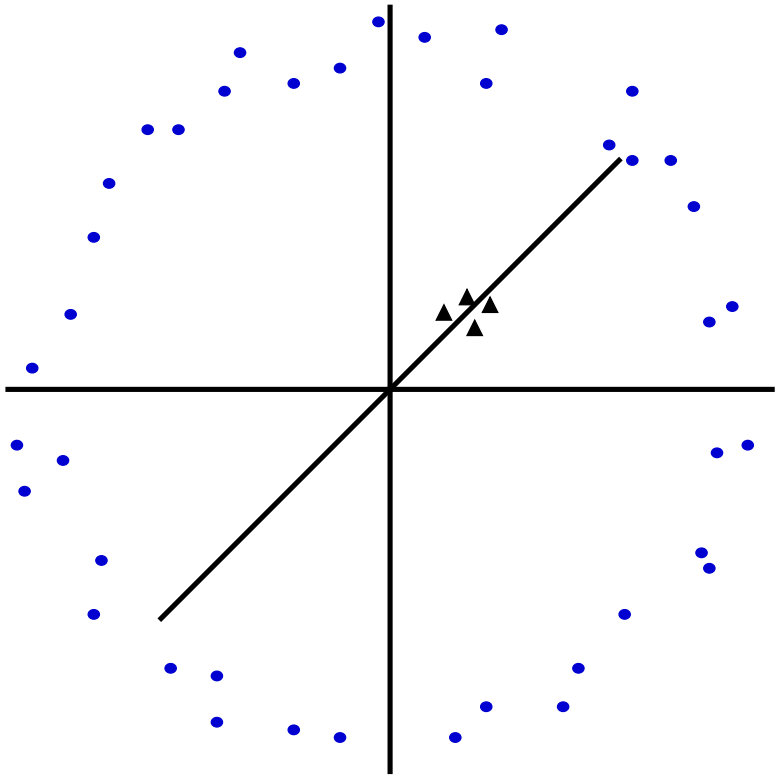
Trouble in High Dimensions

- Removing points with large magnitude



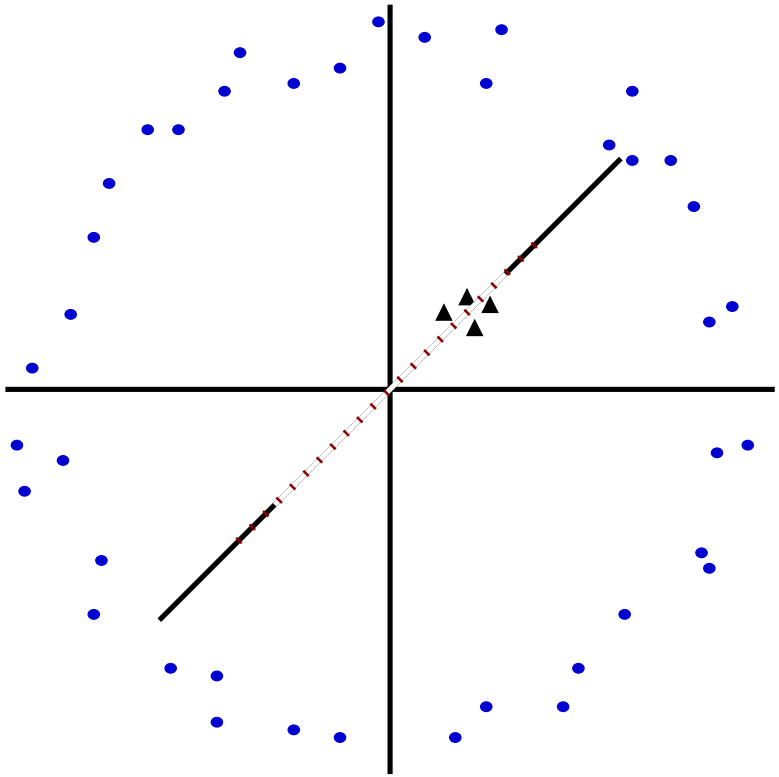
Trouble in High Dimensions

- Removing points with large magnitude



Trouble in High Dimensions

- Removing points with large magnitude



Trouble in High Dimensions

- For these reasons: Some robust covariance estimators have breakdown point = $O(1/p)$, p = 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
- Minimum volume ellipsoid: Ill-posed / Intractable
- Projection pursuit – maximize univariate estimator

Objective & Performance Measurement

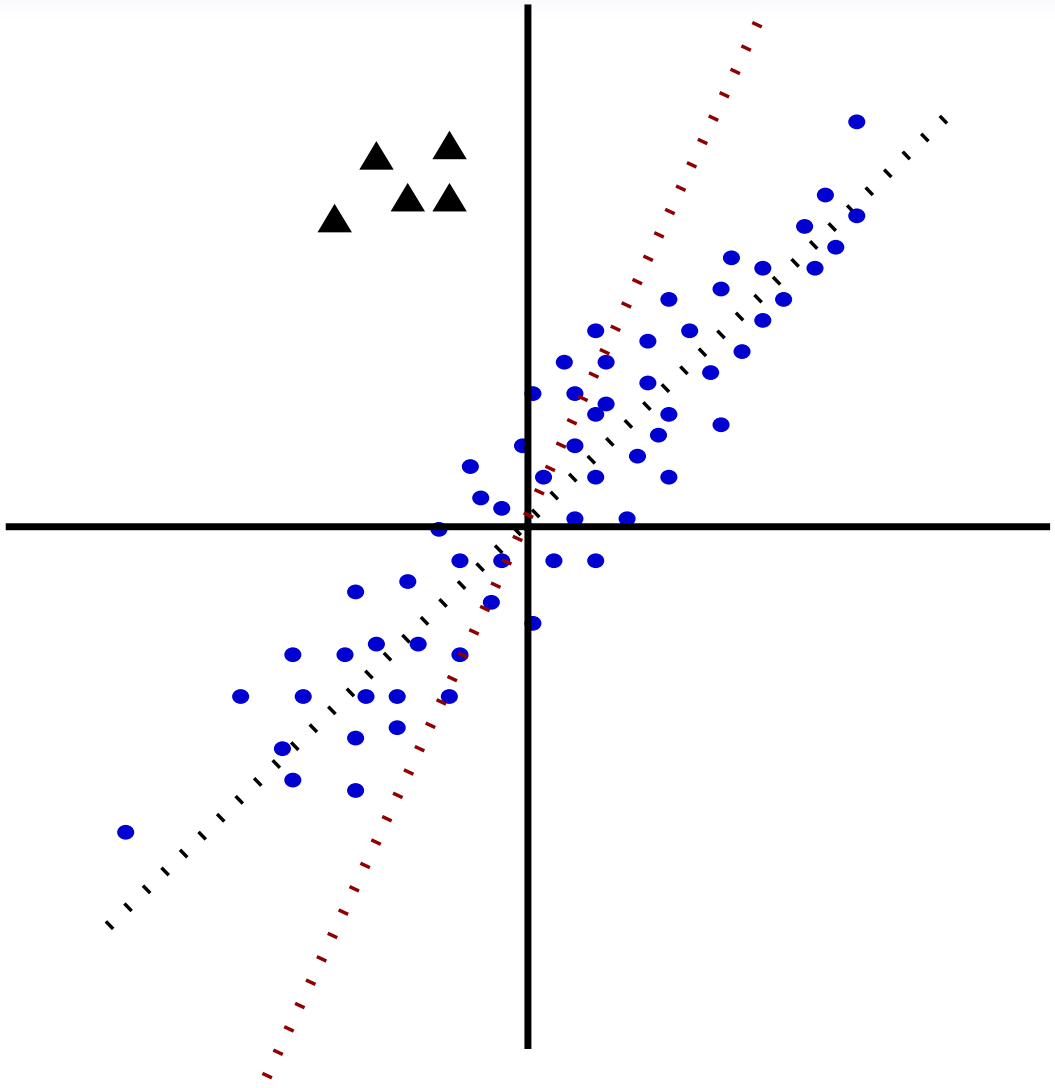


Figure: Angle makes sense for 1-dimensional recovery

Objective & Performance Measurement

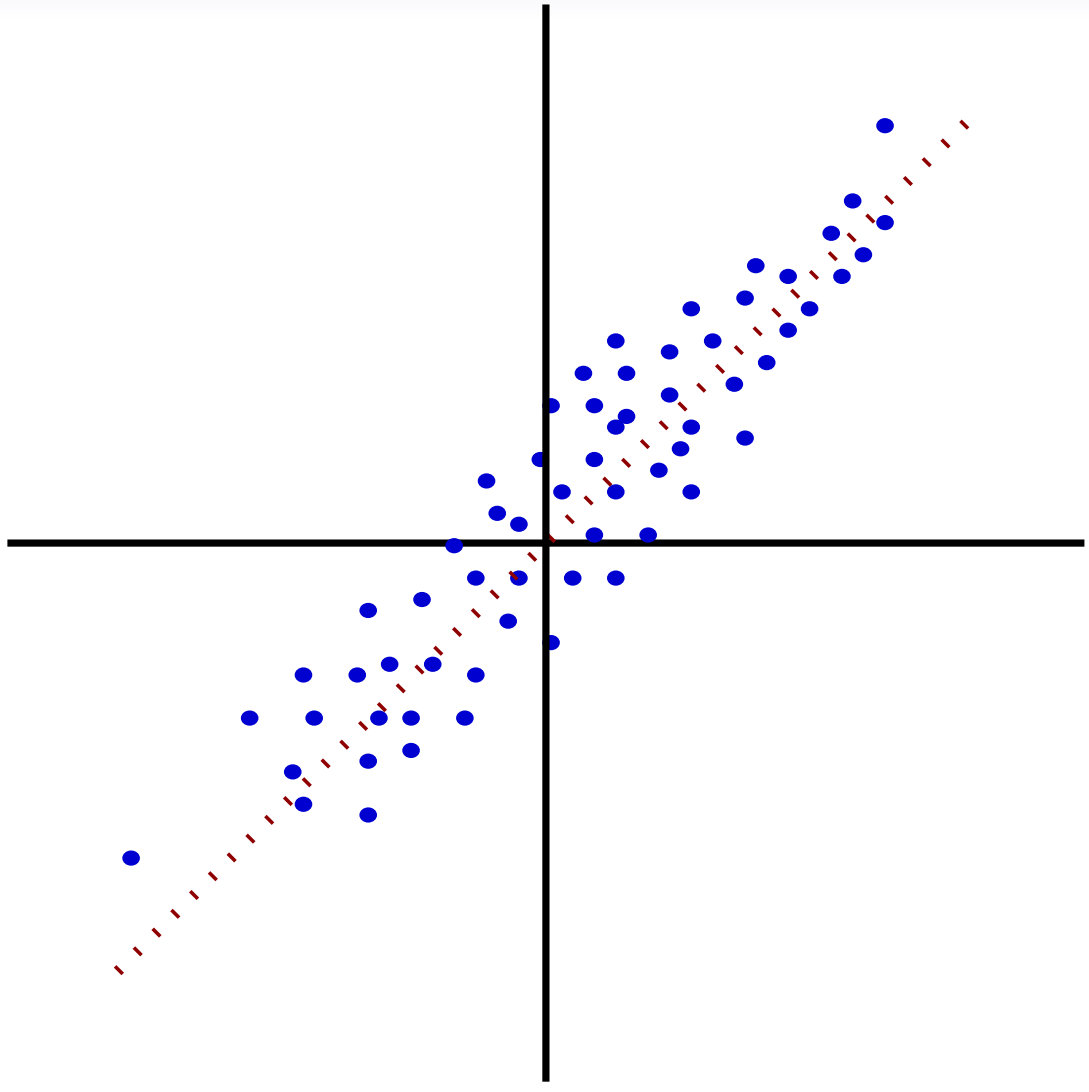


Figure: What we really want: How much variance is captured

Objective & Performance Measurement

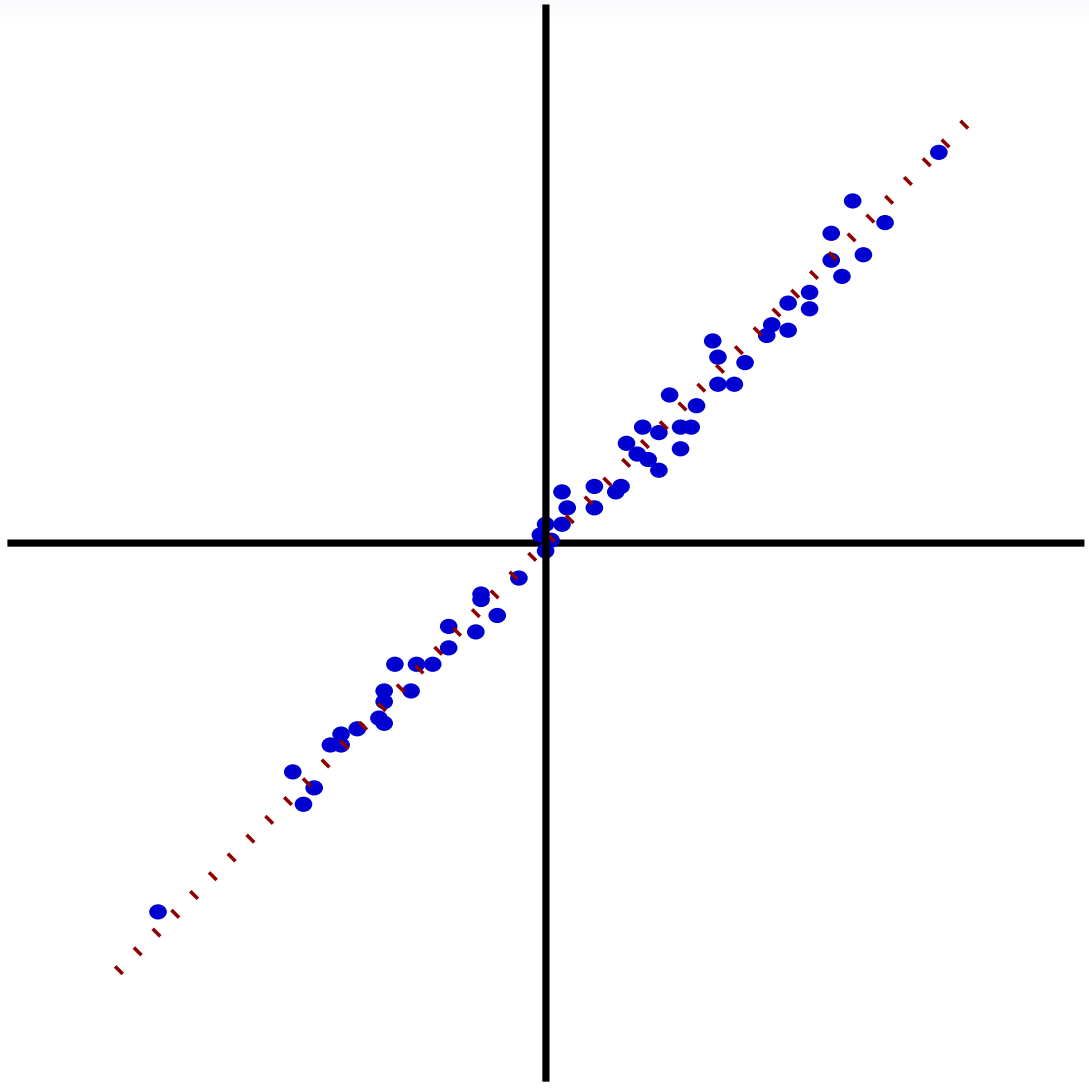


Figure: What we really want: How much variance is captured

Objective & Performance Measurement

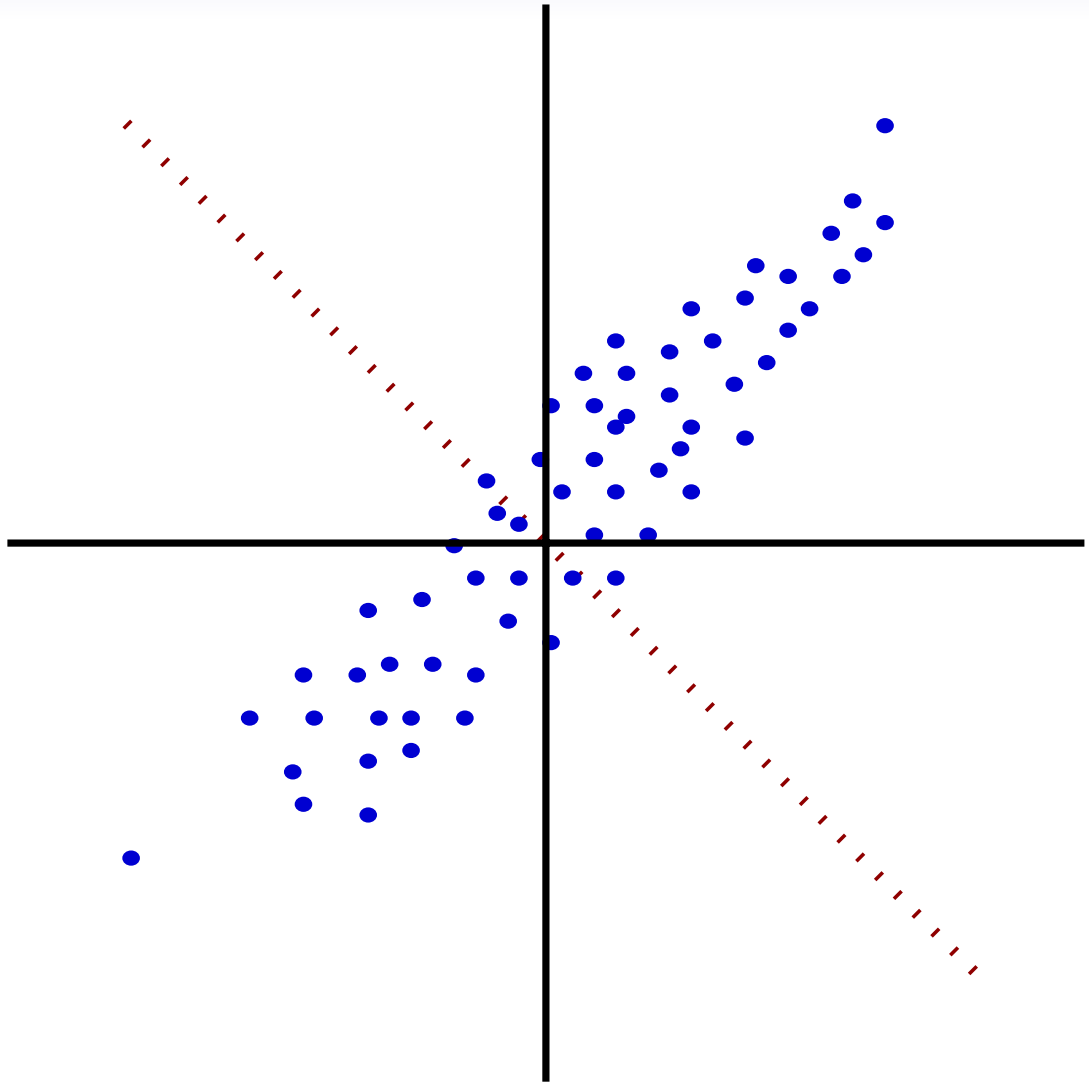


Figure: What we really want: How much variance is captured

Objective & Performance Measurement

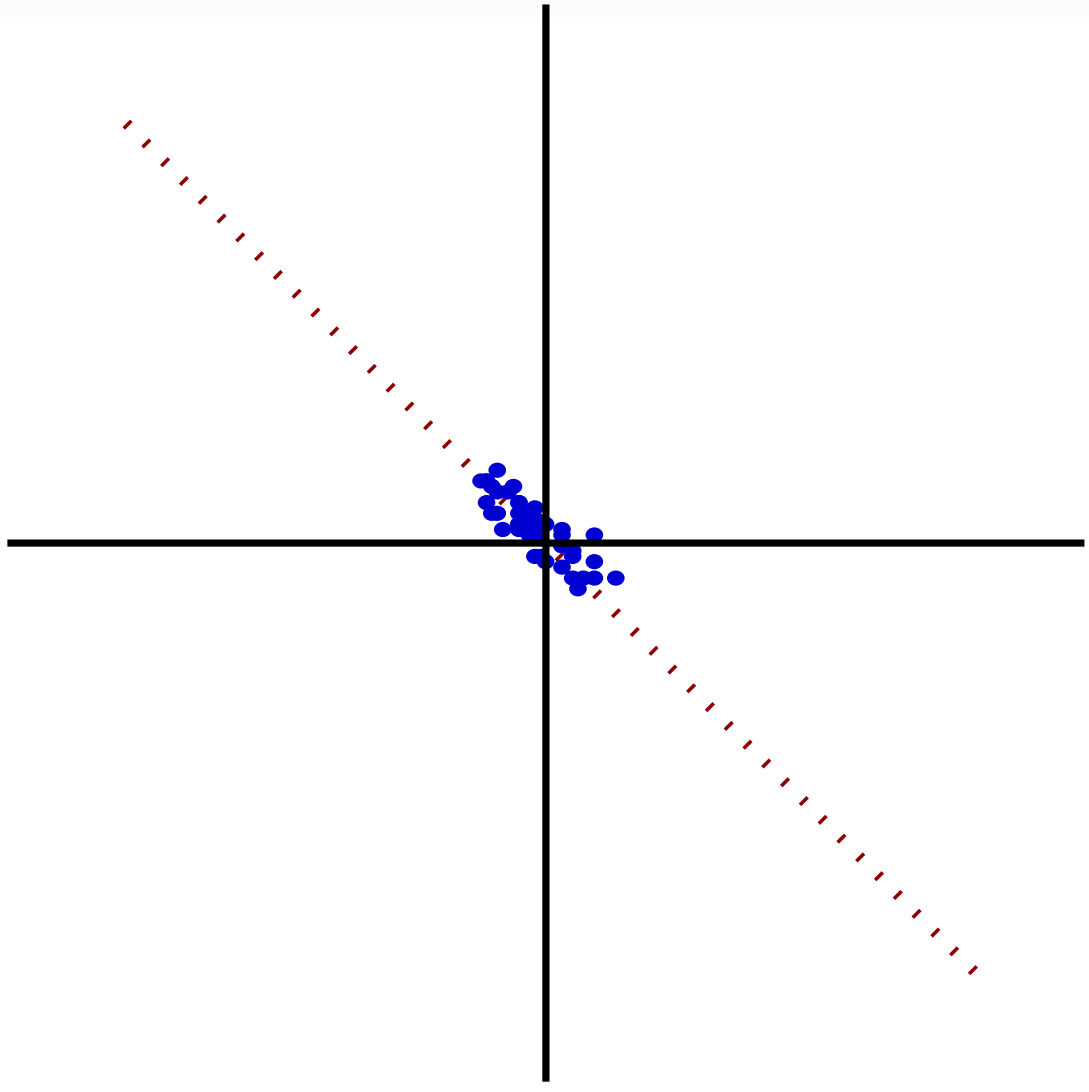


Figure: What we really want: How much variance is captured

A Robust Variance Estimator

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

A Robust Variance Estimator - Pictures

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

A Robust Variance Estimator - Pictures

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

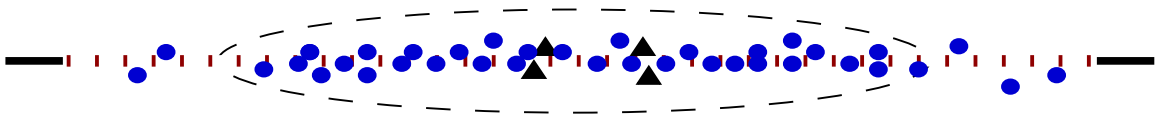


Figure: If corrupted points small, they can only have limited impact on variance

A Robust Variance Estimator - Pictures

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

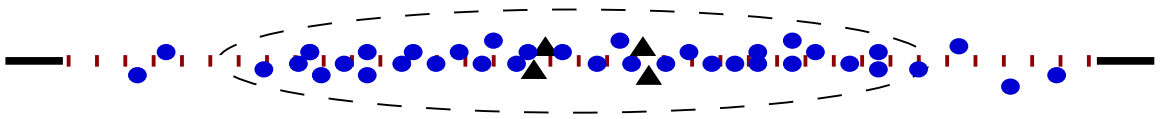


Figure: If corrupted points small, they can only have limited impact on variance

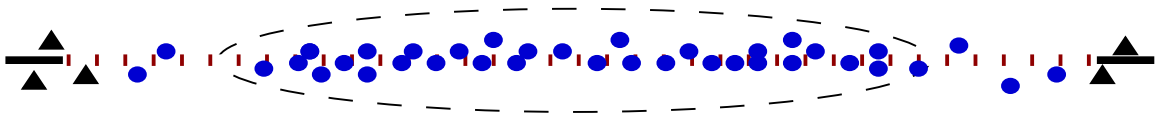
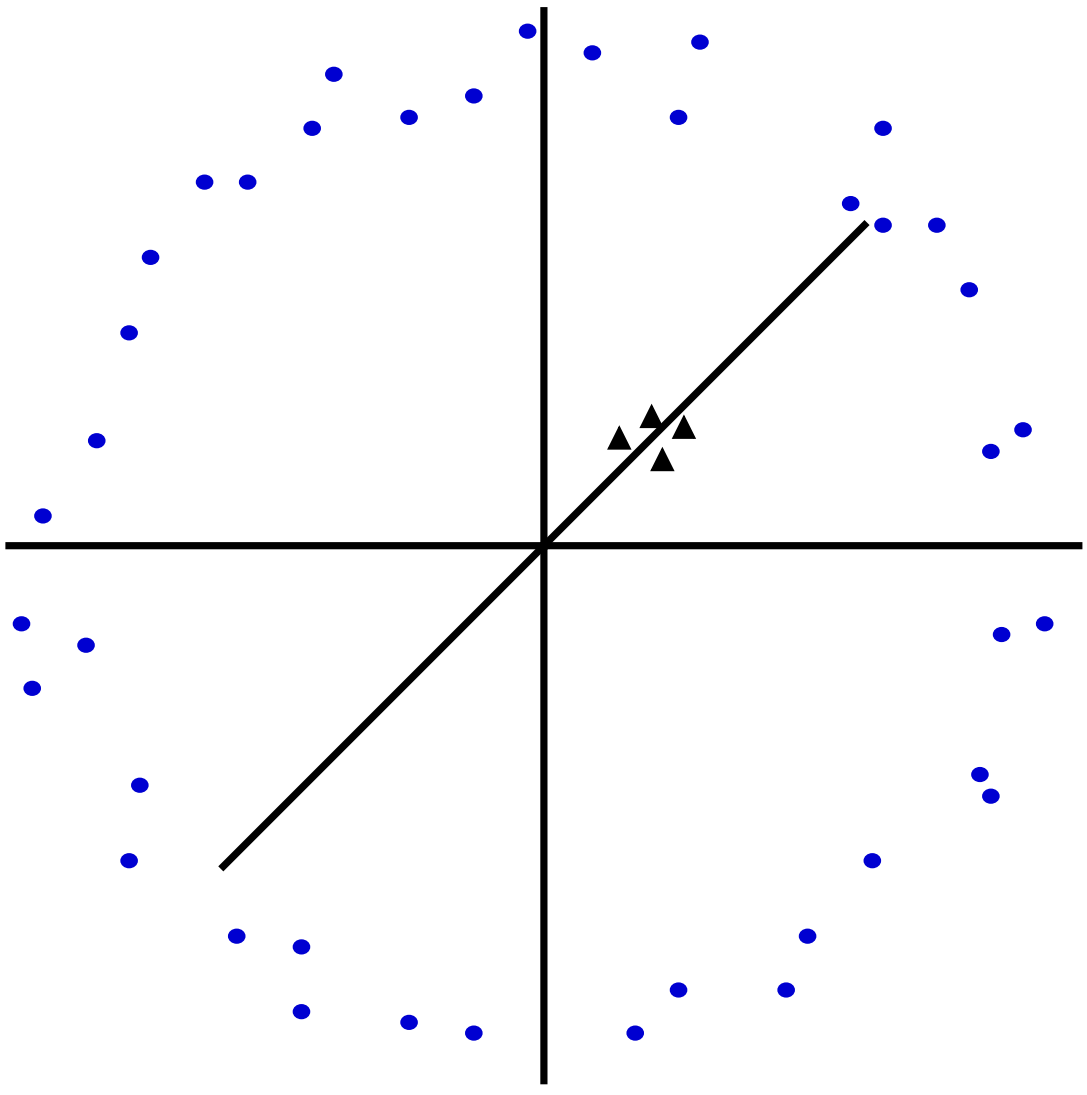
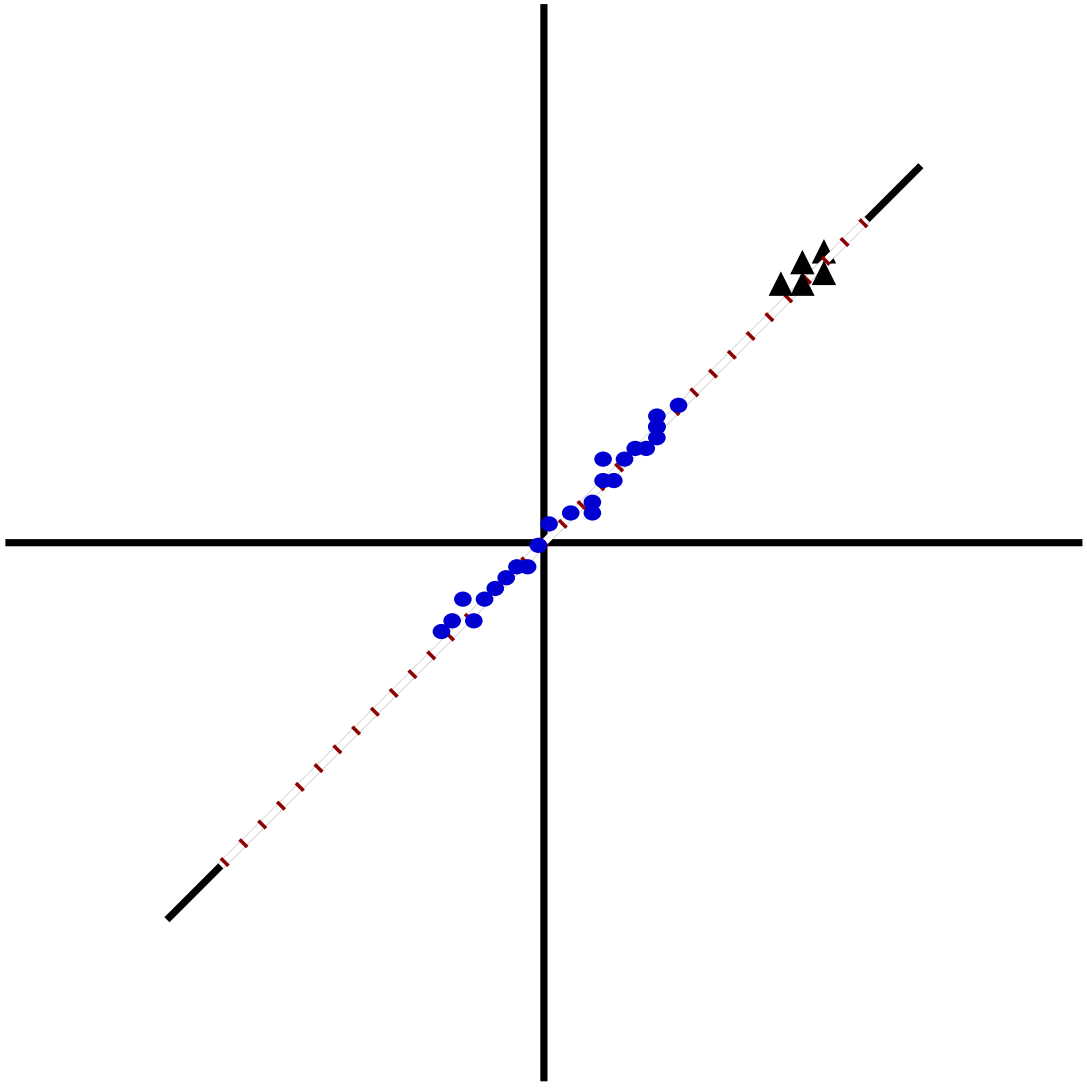


Figure: If corrupted points large, they are discarded and do not impact variance much

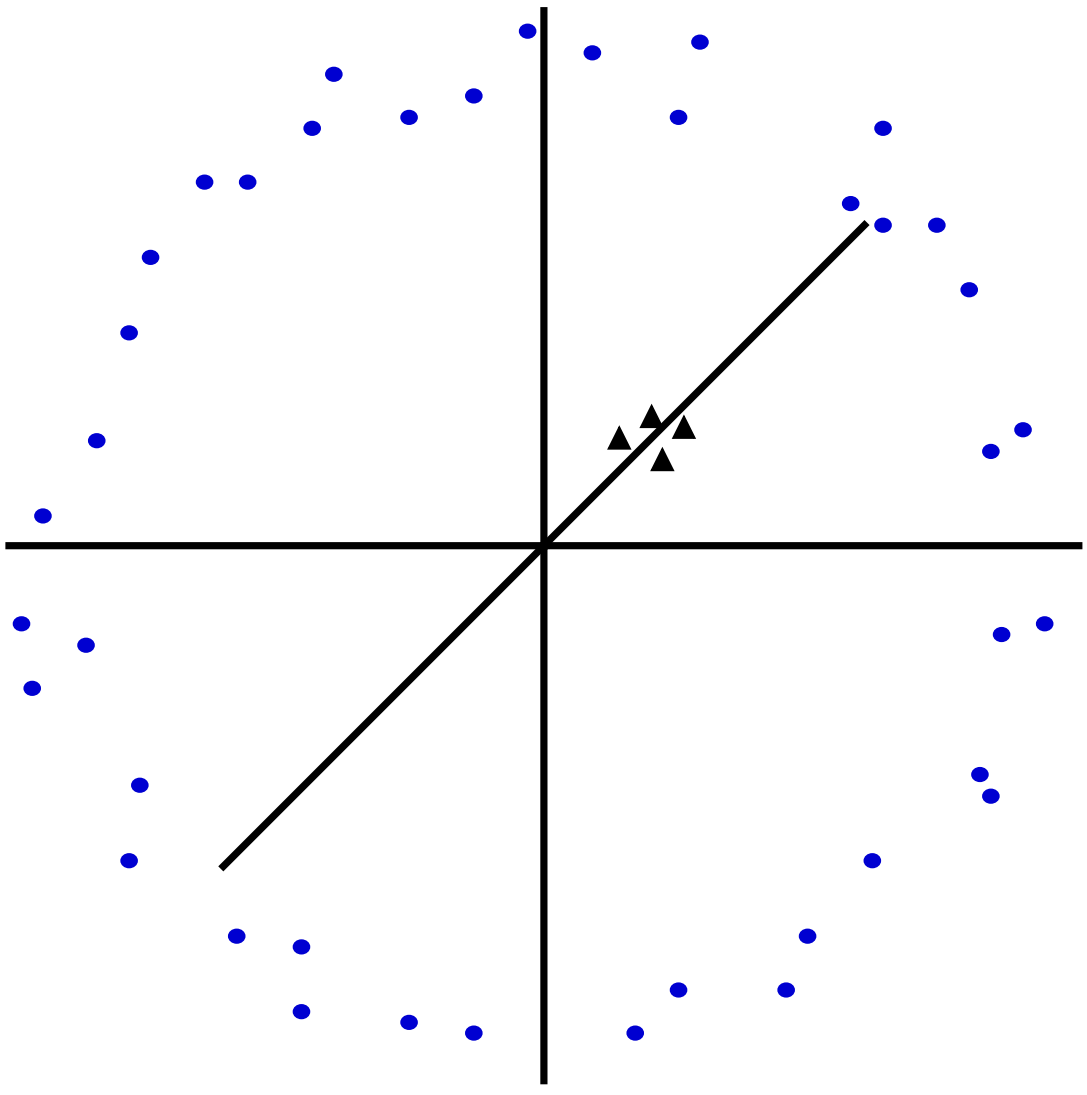
A Robust Variance Estimator - Pictures



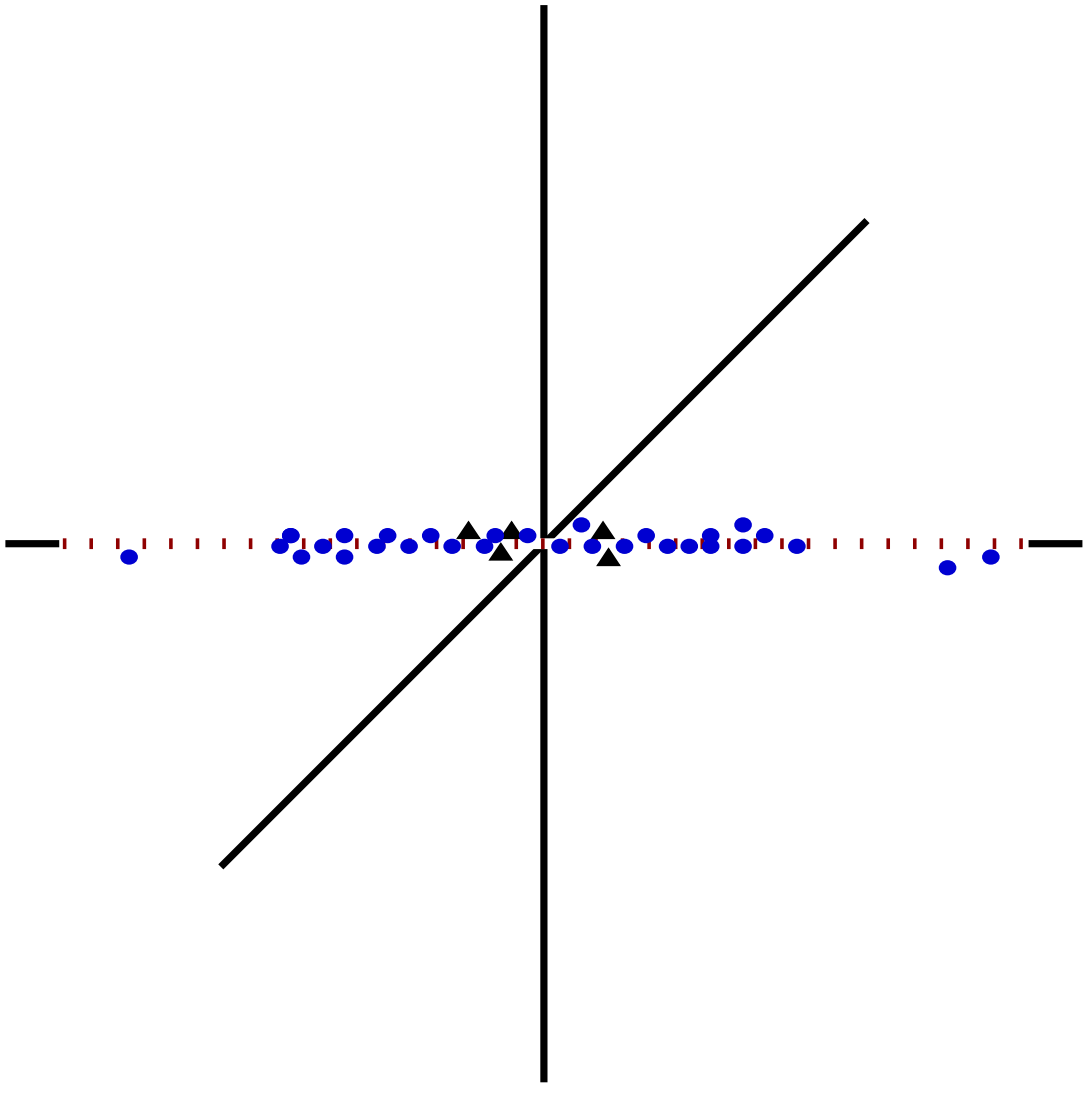
A Robust Variance Estimator - Pictures



A Robust Variance Estimator - Pictures



A Robust Variance Estimator - Pictures



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

Theorem: The algorithm presented:

- Breakdown point $1/2$.
- Perfect recovery for $o(n)$ corrupted points.
- Explicit lower bounds.

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."

A Different Approach: Convex Optimization

- Observation — without noise: $\mathbf{y} = A\mathbf{x}$

$$\begin{bmatrix} | & | & | & | & | \\ A\mathbf{x}_1 & \mathbf{o} & \cdots & \mathbf{o} & A\mathbf{x}_n \\ | & | & | & | & | \end{bmatrix}$$

- Exploit algebraic structure of low-rank matrices:

$$\underbrace{M}_{\text{observed matrix}} = \underbrace{L}_{\text{low rank matrix}} + \underbrace{C}_{\text{column-sparse matrix}}$$

A Different Approach: Convex Optimization

$$\underbrace{M}_{\text{observed matrix}} = \underbrace{L}_{\text{low rank matrix}} + \underbrace{C}_{\text{column-sparse matrix}}$$

- Can we recover L and C by solving?

$$\begin{aligned} \text{minimize} \quad & \|L\|_* + \lambda \|C\|_{1,2} \\ \text{subject to} \quad & M = L + C. \end{aligned}$$

A Different Approach: Convex Optimization

$$\underbrace{M}_{\text{observed matrix}} = \underbrace{L}_{\text{low rank matrix}} + \underbrace{C}_{\text{column-sparse matrix}}$$

- Can we recover L and C by solving?

$$\begin{aligned} \text{minimize} \quad & \|L\|_* + \lambda \|C\|_{1,2} \\ \text{subject to} \quad & M = L + C. \end{aligned}$$

- Problem: No.

A Different Approach: Convex Optimization

$$\underbrace{M}_{\text{observed matrix}} = \underbrace{L}_{\text{low rank matrix}} + \underbrace{C}_{\text{column-sparse matrix}}$$

- Can we recover L and C by solving?

$$\begin{aligned} \text{minimize} \quad & \|L\|_* + \lambda \|C\|_{1,2} \\ \text{subject to} \quad & M = L + C. \end{aligned}$$

- Problem: No.

- Solution: Don't need that.

A Different Approach: Convex Optimization

$$\underbrace{M}_{\text{observed matrix}} = \underbrace{L}_{\text{low rank matrix}} + \underbrace{C}_{\text{column-sparse matrix}}$$

- Can we recover L and C by solving?

$$\begin{aligned} \text{minimize} \quad & \|L\|_* + \lambda \|C\|_{1,2} \\ \text{subject to} \quad & M = L + C. \end{aligned}$$

- Problem: No.
- Solution: Don't need that.
- Oracle problem.

A Different Approach: Convex Optimization

$$\underbrace{M}_{\text{observed matrix}} = \underbrace{L}_{\text{low rank matrix}} + \underbrace{C}_{\text{column-sparse matrix}}$$

- Exact recovery and outlier identification.
- And:
- Methodology for problems where exact recovery not possible.

Robust Collaborative Filtering

- Collaborative filtering: matrix completion.
- Robust collaborative filtering: resistance to manipulators.

Robust Collaborative Filtering

- Collaborative filtering: matrix completion.
- Robust collaborative filtering: resistance to manipulators.
- Same as before, but only partial observations:

$$\underbrace{M}_{\text{observed matrix}} = \underbrace{L}_{\text{low rank matrix}} + \underbrace{C}_{\text{column-sparse matrix}}$$

Get to see:

$$X = P_\Omega(M), \qquad \Omega \subset \{1, 2, \ldots, n\} \times \{1, 2, \ldots, n\}.$$

Robust Collaborative Filtering

- Collaborative filtering: matrix completion.
- Robust collaborative filtering: resistance to manipulators.
- Same as before, but only partial observations:

$$\underbrace{M}_{\text{observed matrix}} = \underbrace{L}_{\text{low rank matrix}} + \underbrace{C}_{\text{column-sparse matrix}}$$

Get to see:

$$X = P_\Omega(M), \qquad \Omega \subset \{1, 2, \ldots, n\} \times \{1, 2, \ldots, n\}.$$

- Can recover L and C by solving:

$$\begin{aligned} \text{minimize} \quad & \|L\|_* + \lambda \|C\|_{1,2} \\ \text{subject to} \quad & P_\Omega(L + C) = X. \end{aligned}$$

- Key: oracle-based techniques of previous work.

Conclusion

- Robust PCA: high dimensions; many outliers
- Extensions: Very general noise model: sub-Gaussian, log-concave.
- Extensions: More general stochastic programming with corrupted sampled data?

Find out more:

- http://users.ece.utexas.edu/~cmcaram
- caramanis@mail.utexas.edu

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 $\mathbf{x}$:

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

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

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

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

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:

$$\text{Variance of Authentic Points} \geq \frac{1}{\kappa} \text{Variance of Corrupted Points}$$

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

(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:

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

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

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

Proof Idea - Step 1.b - details

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

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

- Key idea: matrix concentration of measure result
- For X a random vector in $\mathbb{R}^n$, with controlled fourth moment, then

$$\mathbb{P}\left(\| \frac{1}{N}\sum_{i=1}^{N} X_i X_i^\top - \mathbb{E}(X_i X_i^\top) \| \geq \epsilon \right),$$

decays exponentially.

Proof Idea - Step 1.c - details

(1.c) RVE is a valid variance estimator for the $d-$dimensional signals:

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

- Uniform Laws of Large Numbers, VC Dimension
  - Prove first a one-dimensional version
  - Choose an ε-net of $\mathcal{S}^d$ (recall d bounded). Use 1-dimensional result + Union Bound.
  - Relate ε-net to full version.

Proof Idea - Step 1.c - details

(1.c) RVE is a valid variance estimator for the $d-$dimensional signals:

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

- Uniform Laws of Large Numbers, VC Dimension
  - Prove first a one-dimensional version
  - Choose an ε-net of $\mathcal{S}^d$ (recall d bounded). Use 1-dimensional result + Union Bound.
  - Relate ε-net to full version.
- (Painful, but doable...)

Proof Idea - Step 1.d - details

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

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

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

Proof Idea - Step 1.d - details

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

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

- This is quite plausible... (if I don't say so myself)

Proof Idea - Step 1.d - details

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

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

- This is quite plausible... (if I don't say so myself)
- No noise: 1.c times expansion factor of $\|\mathbf{w}^\top A\|^2$.

Proof Idea - Step 1.d - details

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

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

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

- This is quite plausible... (if I don't say so myself)
- No noise: 1.c times expansion factor of $\|\mathbf{w}^\top A\|^2$.
- With noise: 1.a promises noise effect bounded by constant c.

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.