PRACTICAL SCIENTIFIC ANALYSIS OF BIG DATA

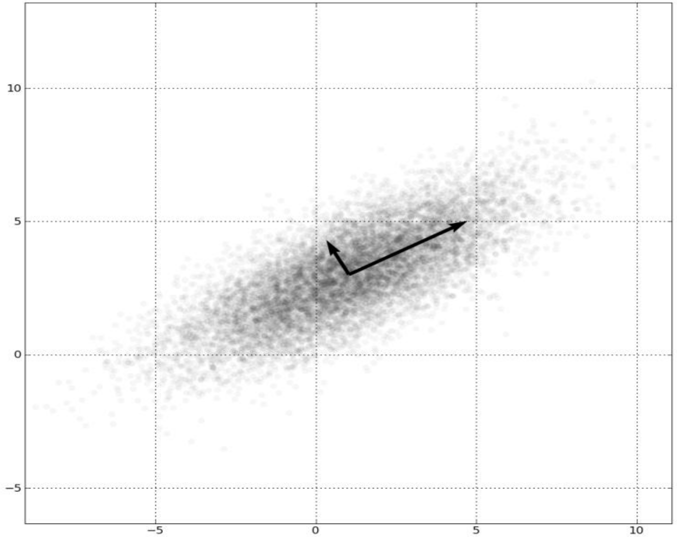
ROBUST STREAMING PCA

4/24/2013

Tamás Budavári / The Johns Hopkins University

Principal Component Analysis

- Principal directions
  - Directions of largest variations
  - Eigenproblem of covariances
  - Singular Value Decomposition
- Problems
  - Needs lots of memory
  - Only need largest ones
  - Very sensitive to outliers



Streams of Data

- Mean

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\mu_n = \frac{n-1}{n}\mu_{n-1} + \frac{1}{n}x_n$$

$$\mu = \gamma\,\mu_{\text{prev}} + (1-\gamma)x$$

Streams of Data

- Mean

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\mu_n = \frac{n-1}{n}\mu_{n-1} + \frac{1}{n}x_n$$

$$\mu = \gamma\,\mu_{\text{prev}} + (1-\gamma)x$$

- Covariance

$$C = \gamma\, C_{\text{prev}} + (1-\gamma)\, y\, y^{\mathrm{T}}$$

$$y = x - \mu_{\text{prev}}$$

Iterative evaluation!

Streaming PCA

- Initialization
  - Eigensystem of a small, random subset
  - Truncate at *p* largest eigenvalues

$$C \approx E_p \Lambda_p E_p^{\mathrm{T}}$$

- Incremental updates
  - Mean and the low-rank *A* matrix
  - SVD of *A* yields new eigensystem

$$C \approx \gamma E_p \Lambda_p E_p^{\mathrm{T}} + (1-\gamma) y y^{\mathrm{T}}$$

$$\approx A A^{\mathrm{T}}$$

- Randomized algorithm!

Robust Statistics

In a nutshell

Location

- M-estimates of the location

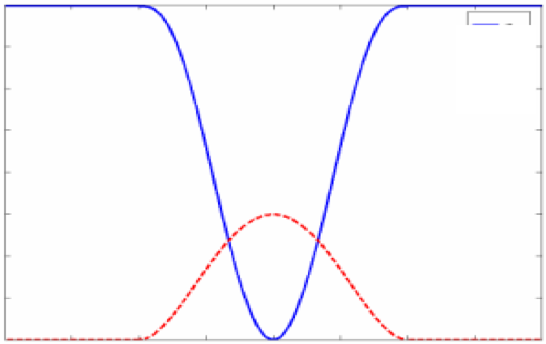
$$L(x_1, \ldots, x_n; \mu) = \prod_{i=1}^{n} f_0(x_i - \mu)$$

$$\widehat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \rho\,(x_i - \mu) \qquad \text{with} \qquad \rho = -\log f_0$$

$$\sum_{i=1}^{n} \rho'(x_i - \widehat{\mu}) = 0$$

  - E.g., mean if x^2, median if $|x|$

- Intuitive
  - Weights: $\sum_{i=1}^{n} W(x_i - \widehat{\mu})(x_i - \widehat{\mu}) = 0$



Dispersion

- M-estimates of the scale

$$\frac{1}{\sigma} f_0\left(\frac{x}{\sigma}\right)$$

$$\widehat{\sigma} = \arg\max_{\sigma} \frac{1}{\sigma^n} \prod_{i=1}^{n} f_0\left(\frac{x_i}{\sigma}\right)$$

$$\frac{1}{n} \sum_{i=1}^{n} \rho\left(\frac{x_i}{\widehat{\sigma}}\right) = 1$$

$$\boxed{\begin{aligned} \rho(t) &= t\psi(t) \\ \psi &= -f_0'/f_0 \end{aligned}}$$

  - E.g., rms if x^2

- Intuitive

$$\widehat{\sigma}^2 = \frac{1}{n\delta} \sum_{i=1}^{n} W\left(\frac{x_i}{\widehat{\sigma}}\right) x_i^2 \qquad \text{with} \qquad W(x) = \begin{cases} \rho(x)/x^2 & \text{if} \quad x \neq 0 \\ \rho''(0) & \text{if} \quad x = 0 \end{cases}$$

Robust PCA

- PCA minimizes σ_{RMS} of the residuals $r = y - Py$
  - Quadratic formula: Σr^2 extremely sensitive to outliers
- We optimize a robust M-scale σ^2 (Maronna 2005)
  - Implicitly given by

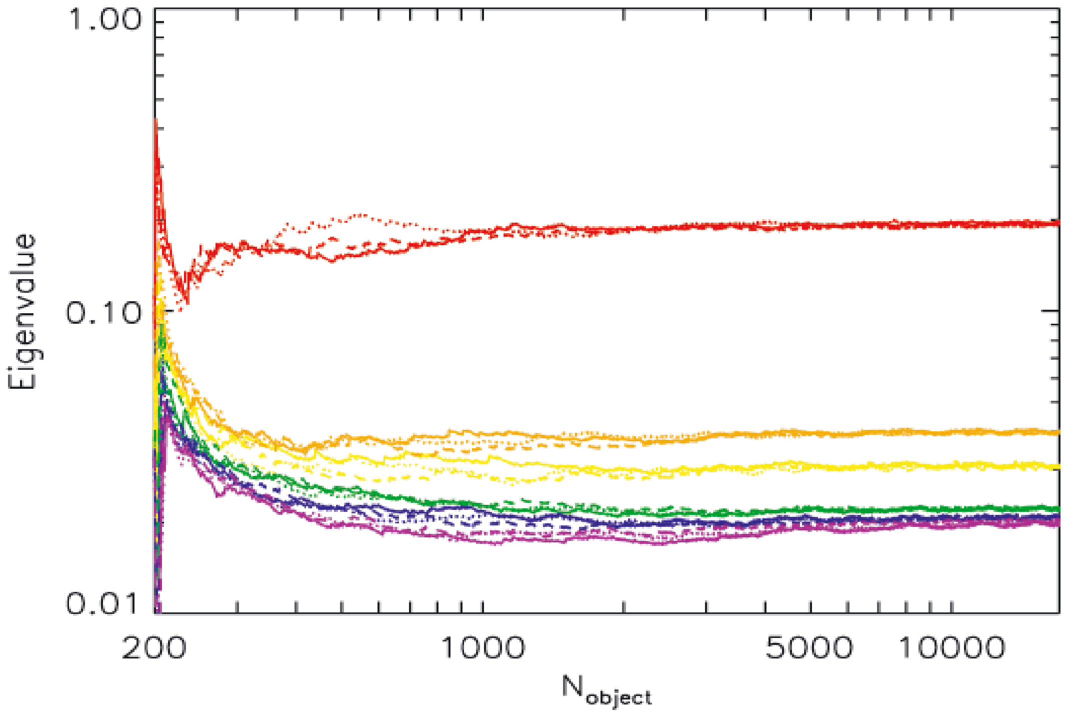
$$\frac{1}{N}\sum_{n=1}^{N}\rho\left(\frac{r_n^2}{\sigma^2}\right)=\delta$$

$$\boldsymbol{\mu}=\left(\sum w_n \boldsymbol{x}_n\right)\Big/\left(\sum w_n\right)$$

$$\boldsymbol{C}=\sigma^2\left[\sum w_n(\boldsymbol{x}_n-\boldsymbol{\mu})(\boldsymbol{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\right]\Big/\left(\sum w_n r_n^2\right)$$

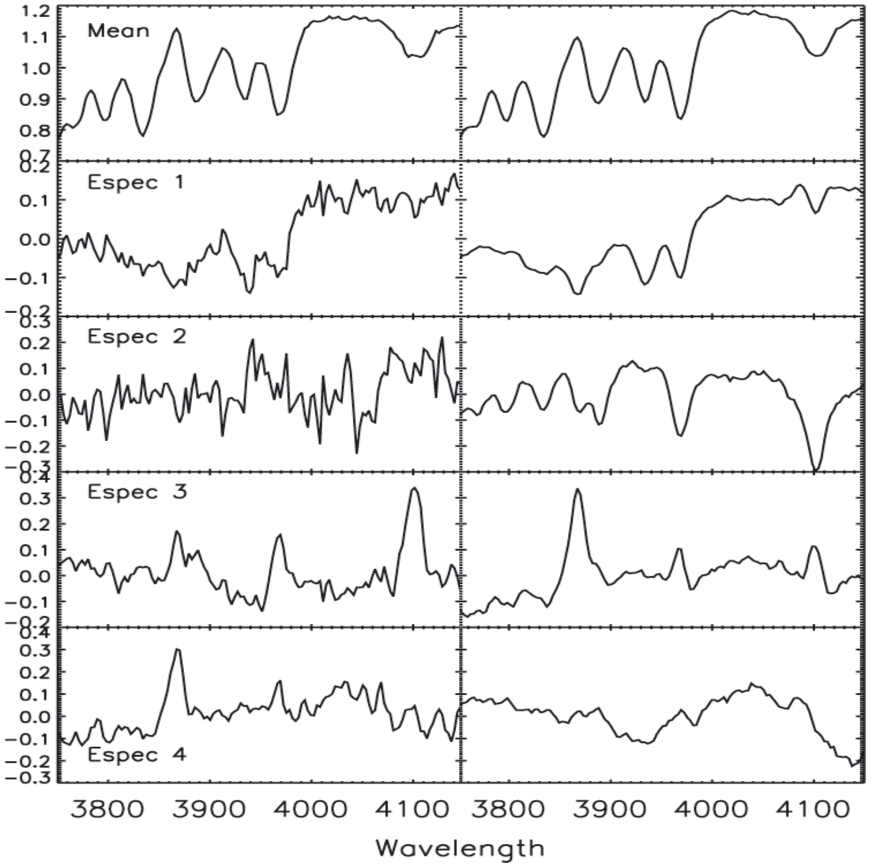
- Fits in with the iterative method!

Galaxy Spectra



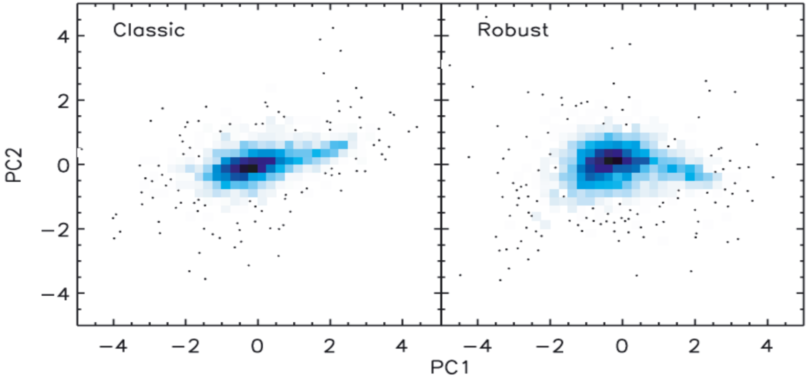
Galaxy Spectra

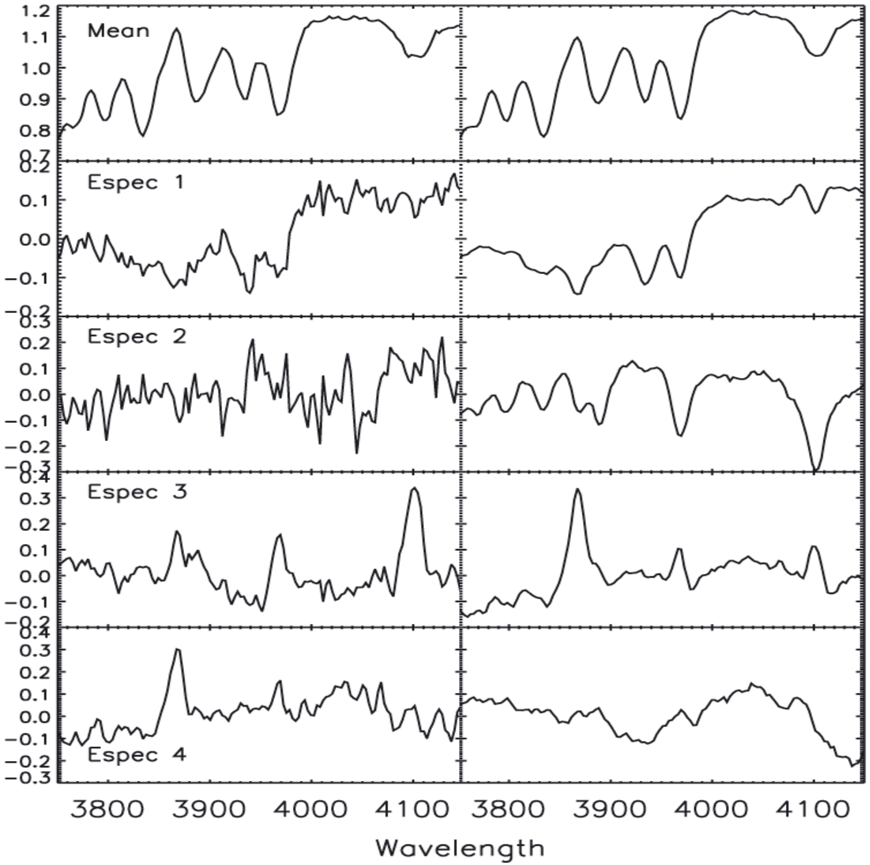
- High SNR eigenfunctions
  - Sign of robustness



Galaxy Spectra

- High SNR eigenfunctions
  - Sign of robustness
- It makes a difference



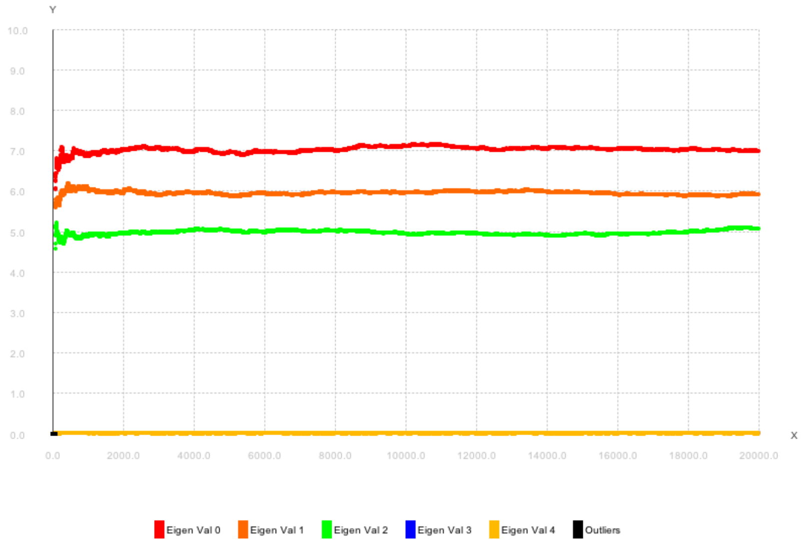


Synthetic Streams

- 3D Gaussian rotated into 50D
  - Stretches: 7, 6, 5
  - Total Var = 110

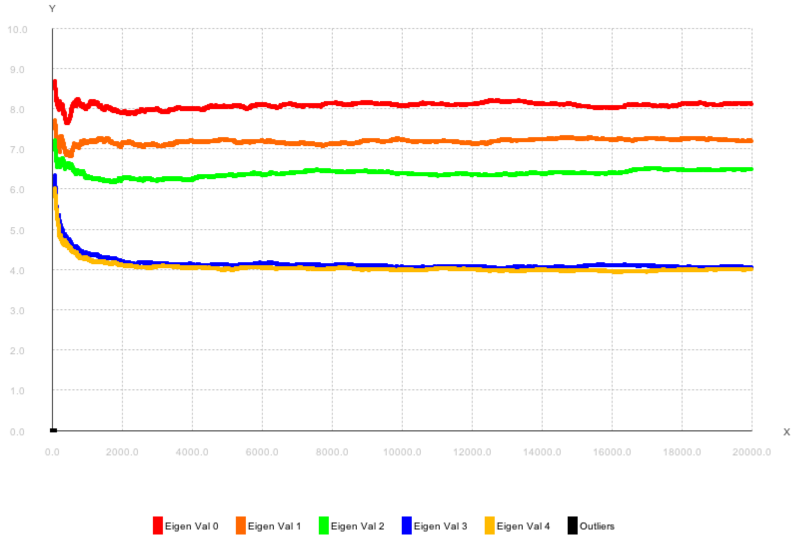
- Plotting square roots of the top 5 eigenvalues

Streaming Classic PCA



Adding Noise

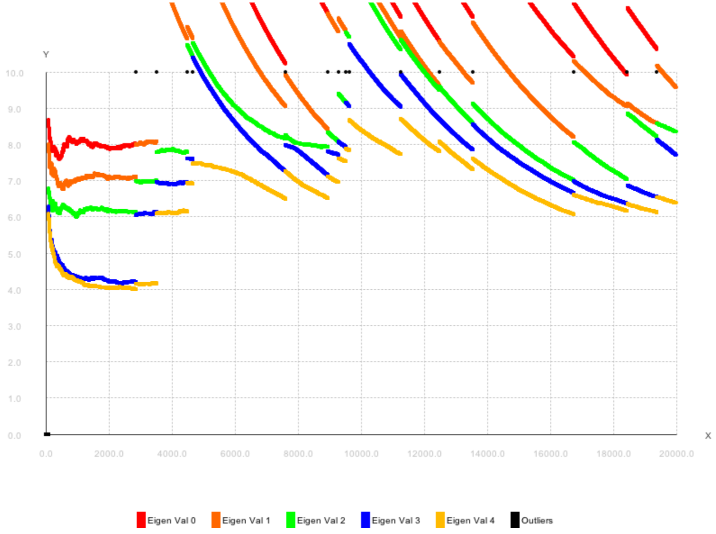
- Uncorrelated noise
  - Gaussian with $\sigma = 4$
  - Total Var = 800
- Same signal
  - With Var = 110
- Effective window size
  - $1-\alpha = 4/20k$



With Outliers

- Adding 0.1% outliers
  - σ = 100 in each bin
- Outliers take over the PCs
  - Instability, no convergence

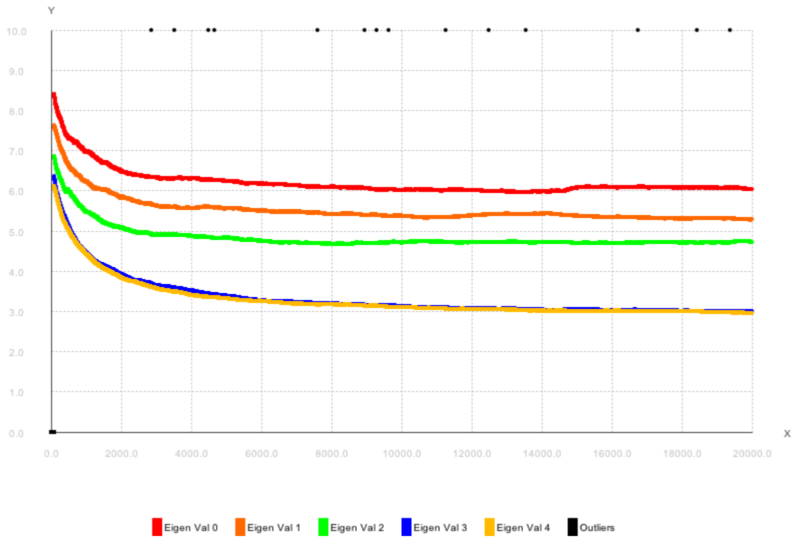
Streaming Classic PCA



Robust Algorithm

- Outliers under control
  - Marked on top
- Initialized with SVD
  - On a set of 100 vectors

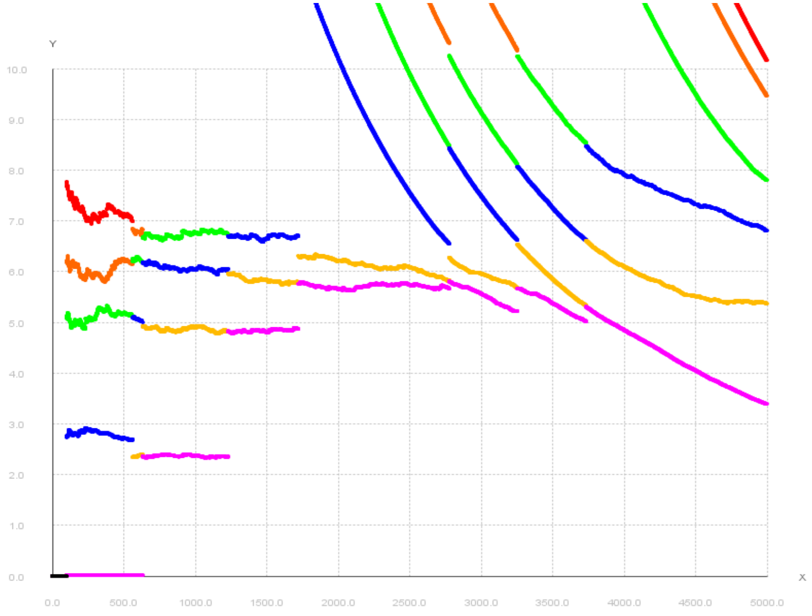
Streaming Robust PCA



Comparison

Tamás Budavári

- Classic



- Robust