Estimating Covariance Spectrum

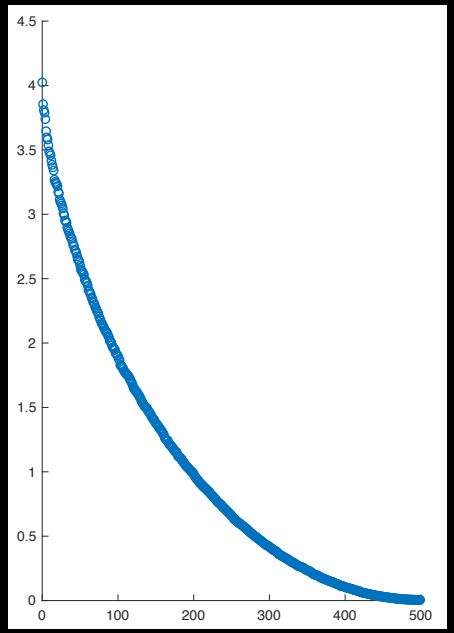
Weihao Kong
Joint work with Gregory Valiant

May 12, TOCA

A Case Study

Data matrix $\mathbf{Y}$: 500 iid samples with dimension 500.
Compute PCA.

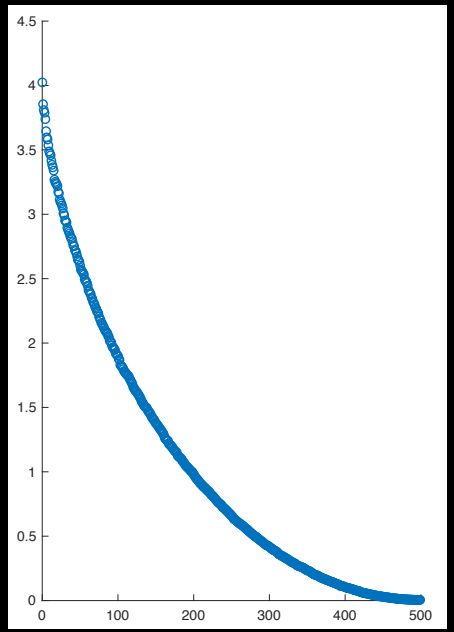
Plot singular values of sample covariance i.e. svd($\frac{1}{500}\mathbf{Y}^\top\mathbf{Y}$):



A Case Study

Data matrix $\mathbf{Y}$: 500 iid samples with dimension 500.
Compute PCA.

Plot singular values of sample covariance i.e. svd($\frac{1}{500}\mathbf{Y}^\top\mathbf{Y}$):



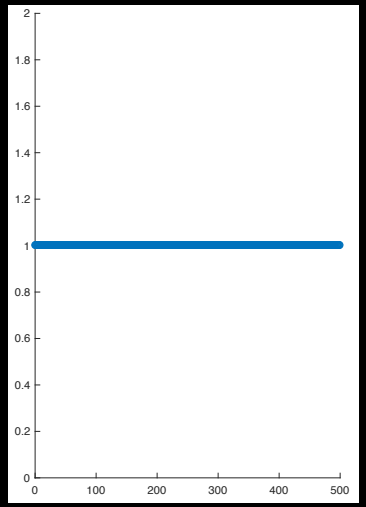
Sample eigenvalues misleading,
distribution has IDENTITY covariance!
$\boldsymbol{Y}_i \sim N(0, \boldsymbol{I}_d)$

Question: Is there anything REALLY meaningful in the data?

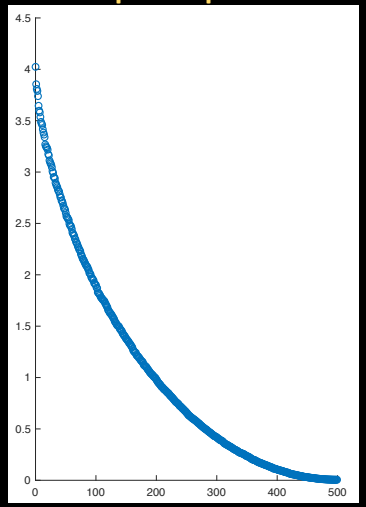
Answer: Estimating Covariance Spectrum

Observation

Population spectrum

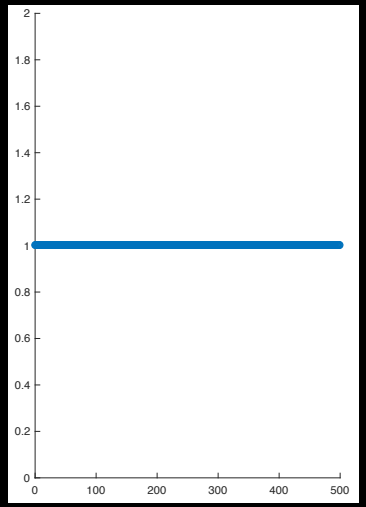


Sample spectrum

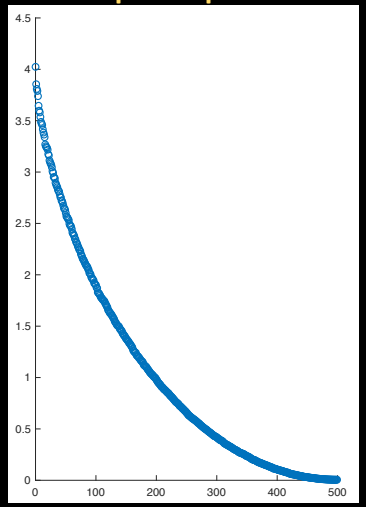


Answer: Estimating Covariance Spectrum

Population spectrum



Sample spectrum



Observation

Thm (informal): Given n samples from d dimensional distribution. Can estimate spectrum to (L1) error εd with sample size $n = O(d^{1-\varepsilon/C})$ w.h.p.

Further Questions

Question:	Answer
How much information do top principal components capture?	Accurately estimate the actual variance explained by top principal components.
What's the best covariance estimator?	Almost-optimal covariance estimator for variety of metrics (Frobenius, Schatten norm)

Both with the help of the knowledge of covariance spectrum