- Computational Complexity of the conventional method

  Construct S_n : $\mathcal{O}(np^2)$
  Spectral decomp. : $\mathcal{O}(p^3)$

  $\Big\}$ both steps $\mathcal{O}(\max\{np^2, p^3\})$

  We can do better (assume $p<n$). Write

  $$X = (x_1, \dots, x_n), \quad S_n = \frac{1}{n-1} (X - \bar{x} 1_n^T)(X - \bar{x} 1_n^T)^T$$

  $$\text{SVD of } (X - \bar{x} 1_n^T) = \overset{p\times p}{U} \operatorname{diag}(\sigma_1, \dots, \sigma_p) \overset{p\times n}{V^T} \qquad (\blacktriangle)$$

  $$U = \mathcal{O}(p), \quad V^T V = I_p, \quad D = \operatorname{diag}(\sigma_1, \dots, \sigma_p)$$

  Then $S_n = \frac{1}{n-1} U D V^T V D U^T = \frac{1}{n-1} U D^2 U^T$

  Hence $U = (u_1, \dots, u_p)$ contains the eigenvalues of S_n.

- Computational complexity of $(\blacktriangle)$.

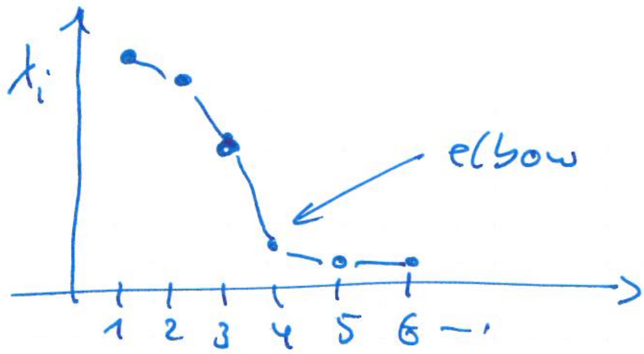
  SVD of $X - \bar{x} 1_n^T$: $\mathcal{O}(\min\{n^2 p, p^2 n\})$

  Only the top k eigenvectors : $\mathcal{O}(knp)$

- Finding the right k

Recall that $\sum_{i=1}^{k} \lambda_i(S_n)$, with $\lambda_1 \geq \dots \geq \lambda_p$ the eigenvalues of S_n, is the preserved variance in the projected points.

Choosing k by a <u>scree plot</u> in practice, depicting the ordered eigenvalues



Choose k as "elbow" − 1.

- Often PCA is applied before further processing, because these may be ineffective for high-dim. data.

4.1.4. The eigenvalue structure of S_n in high dim.

Assume:

$x_1, \dots, x_n \in \mathbb{R}^p$ independent samples of a Gaussian r.v. $\cancel{X \sim} N(0, \Sigma)$. Write $X = (x_1, \dots, x_n)$

Estimate Σ by $S_n = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T = \frac{1}{n} X X^T$.

It holds: $S_n \to \Sigma$ a.e. (p fixed) ($n \to \infty$)

Histogram and scree plot of eigenvalues of S_n for $n = 1000$, $p = 500$, S_n generated by $N(0, I_p)$

Th. 4.1. (Marchenko - Pastur, 1967)

Let $X_1, \dots, X_n \in \mathbb{R}^p$, i.i.d. r.v. with $E(X_i) = 0$ and $Cov(X_i) = \sigma^2 I_p$. $X = (X_1, \dots, X_n) \in \mathbb{R}^{p \times n}$,

$S_n = \frac{1}{n} X X^T \in \mathbb{R}^{p \times p}$, $\lambda_1, \dots, \lambda_p$ the eigenvalues of S_n.

Let $p, n \to \infty$ such that $\frac{p}{n} \to \gamma \in [0,1]$ ($n \to \infty$).

Then the sample distribution of $\lambda_1, \dots, \lambda_p$ (the histogram) converges a.s. to the density

$$f_\gamma(u) = \frac{1}{2\pi\sigma^2 u \gamma} \sqrt{(b-u)(u-a)}\,, \quad a \le u \le b$$

with $a = a(\gamma) = \sigma^2 (1 - \sqrt{\gamma})^2$, $b = b(\gamma) = \sigma^2 (1 + \sqrt{\gamma})^2$. $\lrcorner$

RWTH AACHEN UNIVERSITY

Remark: If $\gamma > 1$ there will be a mass point at zero.

Conclusion: even in the i.i.d. uncorrelated case there is a wide spectrum of eigenvalues.

Question: What happens, if there is a low dim. structure in the data? Is PCA useful.

4.1.5 Spike model

Model assumptions: $X_1, \dots, X_n \in \mathbb{R}^p$, i.i.d.

$Cov(X_i) = \Sigma = I_p + \beta v v^T$ for some $v \in \mathbb{R}^p$, $\|v\| = 1$, $\beta \geq 0$

Interpretation: $X_i = U_i + \sqrt{\beta}\, V_i v$,

$U_i \sim N(0, I_p)$ noise

$V_i \sim N(0, 1)$ signal, indep. of U_i

multiplied by a fixed $\sqrt{\beta}\, v \in \mathbb{R}^p$

Then $Cov(X_i) = Cov(U_i) + \beta\, Var(V_i)\, v v^T$

$= I_p + \beta v v^T$.

Thinking the Future
Zukunft denken

RWTH AACHEN UNIVERSITY

Th. 4.2. (BBP transition. Baik, Ben Arous, Péché [2005])

Assume $X_1, \dots, X_n \in \mathbb{R}^p$ r.v. $E(X_i) = 0$, $Cov(X_i) = I_p + \beta v v^T$,

$\beta \geq 0$, $v \in \mathbb{R}^p$, $\|v\| = 1$. $S_n = \frac{1}{n} X X^T$.

$n, p \to \infty$, $\frac{p}{n} \to \gamma$.

If $\beta \leq \sqrt{\gamma}$ then $\lambda_{max}(S_n) \to (1 + \sqrt{\gamma})^2$

and $|\langle v_{max}, v \rangle|^2 \to 0$

If $\beta > \sqrt{\gamma}$ then $\lambda_{max}(S_n) \to (1+\beta)(1 + \frac{\gamma}{\beta}) > (1 + \sqrt{\gamma})^2$

Ex.

and $|\langle v_{max}, v \rangle|^2 \to \dfrac{1 - \gamma/\beta^2}{1 - \gamma/\beta}$. $\lrcorner$

Thinking the Future
Zukunft denken