RWTH AACHEN UNIVERSITY

4. Dimensionality Reduction

Represent data in a low-dimensional space in an "optimal" way. Dim. 1,2,3 allow for visualization.

4.1 Principal Component Analysis (PCA)

Loose as little information as possible.

Given data $x_1, \ldots, x_n \in \mathbb{R}^p$.

Ex. MNIST data dim. $28 \cdot 28 = 784$.
$n = 70\,000$
grayscale encoding by a number $\in [0,1]$

a) Find a k-dim. subspace such that the projections of $x_1, \ldots, x_n$ thereon represent the original data on its best.

b) Preserve as much variance as possible in the projected data.

a) and b) are equivalent.

RWTH AACHEN UNIVERSITY

$x_1, \dots, x_n$ independently sampled from some distribution.

Sample mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Sample covariance matrix: $S_n = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^T$

$\bar{x}$: MLE, unbiased estimator of EX

S_n : nearly MLE, unbiased estimator of $Cov(X)$.

4.1.1. Find the best projection

Considers the following optimization problem

$$\min_{a\in\mathbb{R}^p} \sum_{i=1}^{n} \| x_i - a - Q(x_i - a)\|_F^2$$

Q orth. projections on a k-dim subspace

$$\min_{a,Q} \sum_{i=1}^{n} \|(x_i-a) - Q(x_i-a)\|^2$$

$$= \min_{a,Q} \sum_{i=1}^{n} \|(I-Q)(x_i-a)\|^2$$

$$= \min_{a,R} \sum_{i=1}^{n} \| R(x_i-a)\|^2, \quad R = I-Q \text{ orth. proj. as well.}$$

Thinking the Future
Zukunft denken

$$= \min_{a,R} \sum_{i=1}^{n} (x_i-a)^T R^T R (x_i-a)$$

$$= \min_{a,R} \sum_{i=1}^{n} \mathrm{tr.}\left((x_i-a)^T R (x_i-a)\right)$$

$$= \min_{a,R} \sum_{i=1}^{n} \mathrm{tr}\left(R (x_i-a)(x_i-a)^T\right)$$

$$\geq \min_{R} \sum_{i=1}^{n} \mathrm{tr}\left(R (x_i-\bar{x})(x_i-\bar{x})^T\right) \quad \text{(see MLE for } N(\mu,\Sigma)\text{)}$$

equality if $a = \bar{x}$

$$= \min_{R} \mathrm{tr}\left[R\left(\sum_{i=1}^{n} (x_i-\bar{x})(x_i-\bar{x})^T\right)\right]$$

$$= \min_{R} \mathrm{tr}\left(R\,(n-1)\,S_n\right)$$

$$= \min_{Q} (n-1)\,\mathrm{tr}\left(S_n (I-Q)\right)$$

It remains to solve

$$\max_{Q} \mathrm{tr}(S_n Q), \quad Q \text{ orth. proj.}, \; Q = \sum_{i=1}^{k} q_i q_i^T,$$

q_i orth., $Q = \tilde{Q}\tilde{Q}^T$,

$\tilde{Q} = (q_1, \ldots, q_k)$

$$= \max_{\tilde{Q}^T\tilde{Q} = I_k} \mathrm{tr}\left(\tilde{Q}^T S_n \tilde{Q}\right) = \sum_{i=1}^{k} \lambda_i(S_n) \quad \text{(Ky Fan Th. 2.4)}$$

where $\lambda_1(S_n) \geq \lambda_2(S_n) \geq \cdots \geq \lambda_n(S_n)$ are the eigenvalues of S_n in decreasing order.

RWTH AACHEN UNIVERSITY

The max is attained if $q_1, \dots, q_k$ are the orthogonal eigenvectors corresponding to $\lambda_1(S_n), \dots, \lambda_k(S_n)$.

4.1.2 Preserve most variance

Seek the k-dim. projection that preserves the most variance.

$$\max_{Q} \sum_{i=1}^{n} \| Q x_i - \frac{1}{n} \sum_{\ell=1}^{n} Q x_\ell \|_F^2$$

$$Q = \hat{Q} \hat{Q}^T, \quad \hat{Q}^T \hat{Q} = I_k$$

orth. proj.

$$= \max_{Q} \sum_{i=1}^{n} \| Q \{ x_i - Q \bar{x} \} \|^2$$

$$= \max_{Q} \sum_{i=1}^{n} \| Q (x_i - \bar{x}) \|^2$$

$$= \max_{Q} \sum_{i=1}^{n} \mathrm{tr}\left[(x_i - \bar{x})^T Q (x_i - \bar{x}) \right]$$

$$= \max_{Q} \mathrm{tr}\left(Q \sum_{i=1}^{n} (x_i - \bar{x}) (x_i - \bar{x})^T \right)$$

$$= \max_{Q} (n-1) \, \mathrm{tr}(Q S_n)$$

with the same solutions as above.

RWTH AACHEN UNIVERSITY

4.1.3. How carry out PCA

Given $x_1, \dots, x_n \in \mathbb{R}^p$, fix $k \ll p$

Compute $S_n = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$

$$S_n = V \Lambda V^T, \quad \Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_p) \quad (*)$$

$$\lambda_1 \geq \dots \geq \lambda_p, \quad V = (v_1, \dots, v_p) \in \mathcal{O}(p)$$

spectral decomposition.

- $v_1, \dots, v_k$ are called the k principal eigenvectors to the principle eigenvalues $\lambda_1, \dots, \lambda_k$.

- Projected points

$$\hat{x}_i = \begin{pmatrix} v_1^T \\ \vdots \\ v_k^T \end{pmatrix} x_i, \quad i = 1, \dots, n$$

$$\hat{x}_i \in \mathbb{R}^k$$

- Computational complexity of the conventional method $(*)$

$$\left.\begin{array}{l} \text{Compute } S_n: \ \mathcal{O}(np^2) \\ \text{Spectral decomposition: } \mathcal{O}(p^3) \end{array}\right\} \text{both steps:}$$

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

Thinking the Future
Zukunft denken

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

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

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

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

Then
$$S_n = \frac{1}{n-1} UDV^TVDU^T = \frac{1}{n-1} UD^2U^T$$

Hence $U = [u_1, \dots, u_p]$ contains the eigenvectors of S_n.

- Computational complexity of $(\triangle)$

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

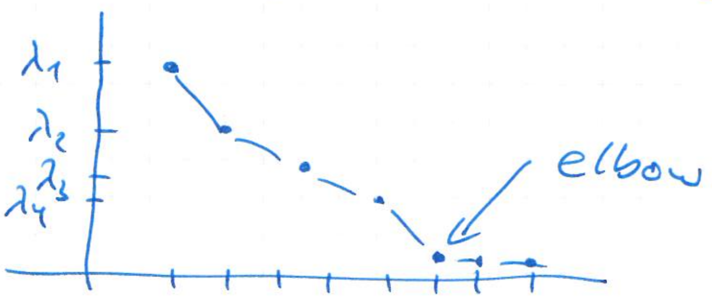
  Only the top k eigenvectors: $\mathcal{O}(knp)$

- Finding the right k

  Visualization: $k = 2, 3$

  In Applications: • low-dim. data corrupted by high-dim noise

  Apply a scree plot, plot the ordered eigenvalues



4.1.4 The eigenvalue structure of S_n in high dimensions

Assume:

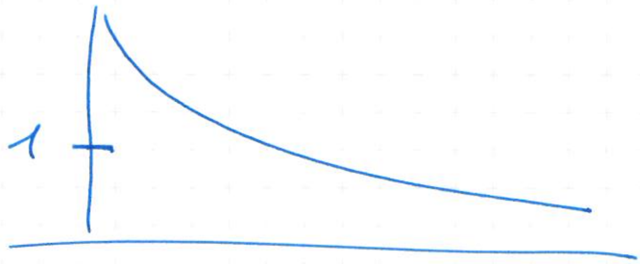
$x_1, \dots, x_n \in \mathbb{R}^p$ independent~~ly~~ samples from a Gaussian r.v. $X \sim N_p(0, \Sigma)$. Write $X = [x_1, \dots, x_n]$

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

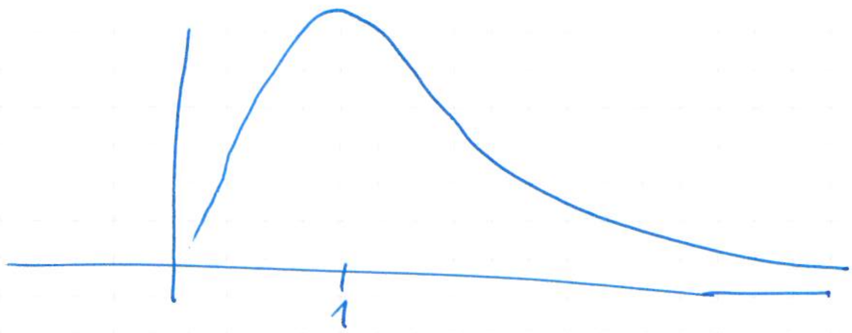
It holds: $S_n \to \Sigma$ a.e. (p fixed)

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

Scree plot:



Histogram:



Many eigenvalues > 1 and < 1.

Conclusion: data has a low dimensional structure ?