3.4. Maximum Likelihood Estimation

$x_1, \dots, x_n$ independent sample from pdf
$f(x, \vartheta)$, ϑ a parameter.

$$L(x;\vartheta) = \prod_{i=1}^{n} f(x_i, \vartheta) \qquad \underline{\text{likelihood function}}$$

$$\ell(x;\vartheta) = \log L(x;\vartheta) = \sum_{i=1}^{n} \log f(x_i, \vartheta)$$

$\underline{\text{log-likelihood fct.}}$

Find ϑ which fits the data best, i.e.,

$$\hat{\vartheta} = \arg\max_{\vartheta} \ell(x;\vartheta)$$

$\hat{\vartheta}$ is called ML estimator. (MLE)

<u>Th. 3.6.</u> $X \sim N_p(\mu, \Sigma)$, $x_1, \dots, x_n$ i.i.d. sample of X.
The MLEs of μ, Σ are:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x} \quad , \quad \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T = S_n \, .$$

<u>Proof.</u> Density of $N_p(\mu, \Sigma)$

$$f(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right), \; x \in \mathbb{R}^p$$

$$\ell(x_1, \dots, x_n; \mu, \Sigma)$$

$$= \sum_{i=1}^{n} \left[\log \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} - \frac{1}{2} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \right]$$

$$= \underbrace{n \log \frac{1}{(2\pi)^{p/2}}}_{\text{constant}} + \frac{n}{2} \log |\Sigma^{-1}| - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)$$

Leave the constant, set $\Lambda = \Sigma^{-1}$

$$\ell^*(\mu, \Sigma) = \frac{n}{2} \log |\Lambda| - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^T \Lambda (x_i - \mu)$$

$$= \frac{n}{2} \log |\Lambda| - \frac{1}{2} \sum_{i=1}^{n} tr\left(\Lambda (x_i - \mu)(x_i - \mu)^T\right)$$

$$= \frac{n}{2} \log |\Lambda| - \frac{1}{2} tr\left(\Lambda \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T\right)$$

Steiners rule:

$$\sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T$$

$$= \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T}_{n S_n} + (\bar{x} - \mu)(\bar{x} - \mu)^T$$

$$\geq n S_n \quad (\text{equality if } \mu = \bar{x})$$

$$\leq \frac{n}{2} \log |\Lambda| - \frac{n}{2} tr(\Lambda S_n) = \ell^*(\mu^*, \Lambda)$$

$$\max \, \ell^*(\mu^*, \Lambda)$$

Need $\frac{\partial}{\partial \Lambda} \log |\Lambda| = (\Lambda^{-1})^T$

$$\frac{\partial}{\partial \Lambda} \operatorname{tr}(\Lambda A) = A^T$$

$$\frac{\partial}{\partial \Lambda} \ell^*(\mu^*, \Lambda) = \frac{n}{2} \Lambda^{-1} - \frac{n}{2} S_n \overset{!}{=} 0_{p \times p}$$

$$\Leftrightarrow \quad \Sigma^* = S_n \qquad \boxed{}$$

4. Dimensionality Reduction

Represent high dim. 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 $x_1, \dots, x_n \in \mathbb{R}^p$.

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

b) Preserve as much variance as possible.

a) and b) are equivalent. $\rightarrow$ later

Thinking the Future
Zukunft denken

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

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

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

$\left(\begin{array}{l} \bar{x} : \text{ unbiased estimator of } E(X) \\ S_n : \text{ unbiased estimator of } Cov(X) \end{array} \right)$

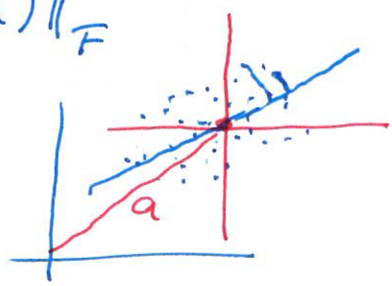
(Ex. MNIST data, $n = 500$, $p = 28 \cdot 28 = 784$)

4.1.1. Find the best projection.

Consider the opt. problem

$$\min_{\substack{a \in \mathbb{R}^p, \\ Q \text{ orth. proj. on a } k\text{-dim. subspace}}} \sum_{i=1}^{n} \| x_i - a - Q(x_i - a) \|_F^2$$

Q orth. proj. 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

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

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

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

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

$$\geq \min_{R} \operatorname{tr}\left(R \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T \right) \quad \left(\begin{array}{l} \text{see MLE for } N(\mu, \Sigma) \\ \text{equality if } a = \bar{x} \end{array} \right)$$

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

$$= \min_{Q} (n-1) \operatorname{tr} S_n (I - Q)$$

It remains to solve

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

$$Q = \tilde{Q} \tilde{Q}^T, \ \tilde{Q} = (q_1, \dots, q_k)$$

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

where $\lambda_1(S_n) \geq \dots \geq \lambda_k(S_n) \geq \dots \geq \lambda_p(S_n)$ are the eigenvalues of S_n in decreasing order.

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

4.1.2 Preserve most variance

Seek a hyperplane so that the proj. data has most variance.

$$\cancel{\max_{Q} \sum_{i=1}^{n} \| Q x_i - \frac{1}{n} \sum_{l=1}^{n} Q x_l \|^2}$$

$$\max_{Q} \sum_{i=1}^{n} \| Q x_i - \frac{1}{n} \sum_{l=1}^{n} Q x_l \|^2 \quad , \quad Q = \tilde{Q}\tilde{Q}^T, \quad \tilde{Q}^T\tilde{Q} = I_k$$

orth. proj.

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

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

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

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

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

with the same solution as above.

Thinking the Future
Zukunft denken

RWTH AACHEN UNIVERSITY

4.1.3 How to carry out PCA

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

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

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

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

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

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

Thinking the Future
Zukunft denken