3.4. Maximum Likelihood Estimation

$x_1, \dots, x_n$ sample, independent observations from pdf $f(x; \vartheta)$, ϑ parameter vector

$$L(x; \vartheta) = \prod_{i=1}^{n} f(x_i; \vartheta)$$

is called <u>likelihood function</u>.

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

is called <u>log-likelihood fct.</u>

Given $x_1, \dots, x_n$, consider L and ℓ as a fct. of ϑ.

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

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

$\hat{\vartheta}$ is called <u>maximum likelihood estimator</u>. (MLE)

<u>Th. 3.6.</u> $X \sim N_p(\mu, \Sigma)$, $x_1, \dots, x_n$ i.i.d. samples from X.
The MLE of μ and Σ are

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

RWTH AACHEN UNIVERSITY

Proof. Density of $\mathcal{N}_p(\mu, \Sigma)$

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

$$\ell(x_1, \dots, x_n; \mu, \Sigma) = \sum_{i=1}^{n} \left[\log \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} - \frac{1}{2}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)\right]$$

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

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

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

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

~~Steiner's rule:~~

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

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

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

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

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

~~max~~

Thinking the Future
Zukunft denken

RWTH AACHEN UNIVERSITY

$\max_{\Lambda} \ell^*(\mu^*, \Lambda)$

needs

$$\frac{\partial}{\partial \Lambda} \log|\Lambda| = \Lambda^{-1}$$

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

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

Solution $\Lambda^{-1} = S_n = \Sigma$ ∎

4. ~~Principal~~ Dimensionality Reduction

Represent data in a low-dim. space in an "optimal" way.

4.1. Principal component analysis (PCA)

Loose as little information as possible.

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

a) Find a k-dim subspace s.t. the projections of $x_1, .., x_n$ thereon represent the original data on its best.

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

a) and b) are equivalent! → later

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 EX

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

4.1.1. Find the best projection

Optimization problem:

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

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

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

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

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

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

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

$$\geq \min_{R} \operatorname{tr} \Big(R \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})(x_i - x)^T}_{} \Big)$$

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

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

It remains to solve

$$\max_{Q} \operatorname{tr} (S_n Q) \quad , \quad Q \text{ orth. proj.}$$

$$Q = \sum_{i=1}^{k} q_i q_i^T \, , \; q_i \text{ orth}$$

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

$$= \max_{\tilde{Q}^T \tilde{Q} = I_k} \operatorname{tr} (\tilde{Q}^T S_n \tilde{Q}) = \sum_{i=1}^{k} \lambda_i (S_n)$$

(Ky Fan, Th. 2.4)

where $\lambda_1(S_n) \geq \ldots \geq \lambda_p(S_n)$ are the eigenvalues of S_n.
The max attained if $q_1, \ldots, q_k$ are the orth.normal eigenvectors corresp. to ~~$\lambda_1(S_n)$~~ $\lambda_1(S_n), \ldots, \lambda_k(S_n)$.

4.1.2 Preserve most variance

Seek k-dim. projection preserving most variance.

$$\max_Q \sum_{i=1}^n \| Qx_i - \frac{1}{n}\sum_{\ell=1}^n Qx_\ell \|_F^2 , \quad Q = \tilde{Q}\tilde{Q}^T, \ \tilde{Q}^T\tilde{Q} = I_k$$

Q orth. proj.

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

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

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

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

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

with the same solution as above.

4.1.3 How to carry out PCA

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

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

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

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

spectral decomposition

Thinking the Future
Zukunft denken

RWTH AACHEN UNIVERSITY

- $v_1, \dots, v_k$ are called <u>principal eigenvector</u> to the principle eigenvalues $\lambda_1 \geq \dots \geq \lambda_k$.

  Projected points $\hat{x}_i = \begin{pmatrix} v_1^T \\ \vdots \\ v_k^T \end{pmatrix} x_i$

  (k-dim) (k × p) (p-dim)

Thinking the Future
Zukunft denken