Machine Intelligence 2

5.2 Maximum Likelihood & Estimation Theory

Prof. Dr. Klaus Obermayer

Fachgebiet Neuronale Informationsverarbeitung (NI)

SS 2018

Estimation theory

Estimator

An estimator $\hat{P}(X)$ is a function that maps from its sample space X (data) to a set of *sample estimates* W

An estimator ...

- is a function of a random variable
- is a random variable
- can be statistically characterized via its moments (mean, variance, ...)
  $\rightsquigarrow$ quality criteria: unbiasedness, efficiency

Probability distributions: an example

$$P\left(\left\{\underline{\mathbf{x}}^{(\alpha)}\right\};\underline{\mathbf{w}}^*\right)$$

set of observations: $\left\{\underline{\mathbf{x}}^{(\alpha)}\right\}, \alpha = 1, \ldots, p$ from true distribution

Goal: estimate "true" values $\underline{\mathbf{w}}^*$ from observed data

estimator $\widehat{\underline{\mathbf{w}}}$**:**

$$\widehat{\underline{\mathbf{w}}} = \widehat{\underline{\mathbf{w}}}\left(\left\{\underline{\mathbf{x}}^{(\alpha)}\right\}\right)$$

- procedure for the determination of $\underline{\mathbf{w}}^*$ given the observed data
- $\underline{\mathbf{w}}^*$ is a function of $\left(\left\{\underline{\mathbf{x}}^{(\alpha)}\right\}\right)$
- $\underline{\mathbf{x}}^{(\alpha)}$ are random variables $\rightarrow \widehat{\underline{\mathbf{w}}}$ is a **random variable!**

The Maximum Likelihood estimator

the likelihood function

$$\widehat{P}\left(\left\{\underline{\mathbf{x}}^{(\alpha)}\right\};\underline{\mathbf{w}}\right)$$

the $\log$**-likelihood function**

$$\ln\widehat{P}\left(\left\{\underline{\mathbf{x}}^{(\alpha)}\right\};\underline{\mathbf{w}}\right)=\sum_{\alpha=1}^{p}\ln\widehat{P}\left(\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}}\right)$$

the Maximum Likelihood estimator

$$\widehat{\underline{\mathbf{w}}}=\underset{\underline{\mathbf{w}}}{\operatorname{argmax}}\widehat{P}\left(\left\{\underline{\mathbf{x}}^{(\alpha)}\right\};\underline{\mathbf{w}}\right)$$

Quality criteria for estimators

What are good estimators?

bias: $$\underline{\mathbf{b}} = \underbrace{\left\langle \underline{\widehat{\mathbf{w}}} \right\rangle_{P(x^\alpha;w)}}_{\substack{\text{expectation} \\ \text{w.r.t the } \underline{\text{true}} \\ \text{distribution}}} - \underline{\mathbf{w}}^*$$

variance: $$\underline{\mathbf{\Sigma}} = \left\langle (\underline{\widehat{\mathbf{w}}} - < \underline{\hat{\mathbf{w}}} >)(\underline{\widehat{\mathbf{w}}} - < \underline{\hat{\mathbf{w}}} >)^T \right\rangle_{P(x^\alpha;w)}$$

Optimal estimators

no bias: $\underline{\mathbf{b}} \overset{!}{=} 0$ $\leftarrow$ only possible if true model within model class

minimal variance: $|\underline{\mathbf{\Sigma}}| \overset{!}{=} \min$

The sample mean

N observations $x^{(\alpha)}$

$$x^{(\alpha)} = A + \epsilon^{(\alpha)}$$

with $\epsilon^{(\alpha)} \sim N(0, \sigma^2)$

Examples for estimators for A:

$\hat{A} = \frac{1}{N} \sum x^{(\alpha)}$	unbiased
$\tilde{A} = \frac{1}{2N} \sum x^{(\alpha)}$	biased for $A \neq 0$
$\tilde{A} = k$	minimum variance but biased

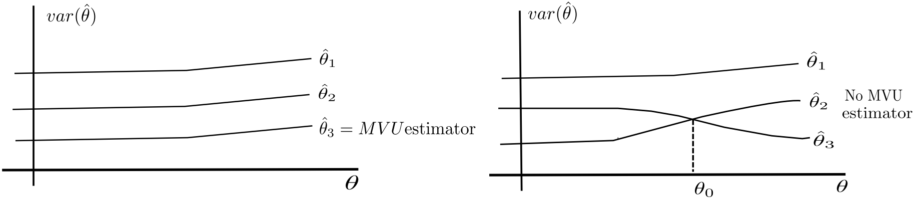
The minimum variance unbiased estimator

Optimal estimators

no bias: $\underline{\mathbf{b}} \stackrel{!}{=} 0$ $\leftarrow$ only possible if true model within model class

minimal variance: $|\underline{\mathbf{\Sigma}}| \stackrel{!}{=} \min$

MVU: criteria have to hold for ALL possible values of $\underline{\mathbf{w}}^*$!



MVUs do not always exist

The minimum variance unbiased estimator

given just observed sample conditionally independent observations with the 2 pdfs

$$x[0] \sim \mathcal{N}(\theta, 1) \qquad x[1] \sim \begin{cases} \mathcal{N}(\theta, 1) & \text{if } \theta \geq 0 \\ \mathcal{N}(\theta, 2) & \text{if } \theta < 0 \end{cases}$$

two estimators

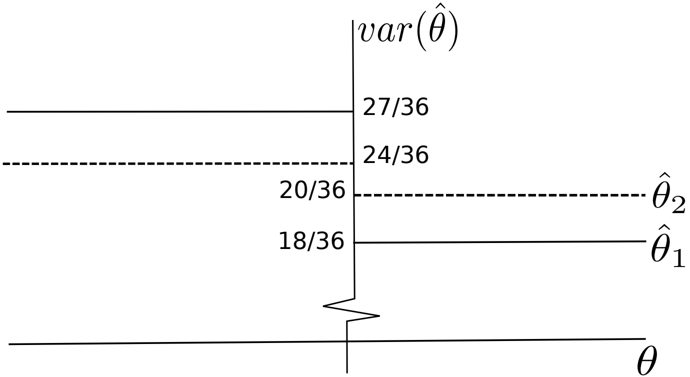
$$\hat{\theta}_1 = \frac{1}{2}(x[0] + x[1]) \qquad \text{and} \qquad \hat{\theta}_2 = \frac{2}{3}x[0] + \frac{1}{3}x[1]$$

variances:

$$var(\hat{\theta}_1) = \frac{1}{4}(var(x[0]) + var(x[1])) \qquad \begin{cases} \frac{18}{36} & \text{if } \theta \geq 0 \\ \frac{27}{36} & \text{if } \theta < 0 \end{cases}$$

$$var(\hat{\theta}_2) = \frac{4}{9}var(x[0]) + \frac{1}{9}var(x[1]) \qquad \begin{cases} \frac{20}{36} & \text{if } \theta \geq 0 \\ \frac{24}{36} & \text{if } \theta < 0 \end{cases}$$

Example for the non-existence of MVUs (Kay, 1993)



MVU vs. minimal mean squared error

$$MSE(\underline{\hat{\mathbf{w}}}) = E[(\underline{\hat{\mathbf{w}}} - \underline{\mathbf{w}^*})^2]$$

This however does not yield a realizable estimator because

$$\begin{aligned}
MSE(\hat{w}) &= E\left\{[(\hat{w} - E(\hat{w})) + (E(\hat{w}) - w^*)]^2\right\} \\
&= var(\hat{w}) + [E(\hat{w}) - w^*]^2 \\
&= variance + bias^2
\end{aligned}$$

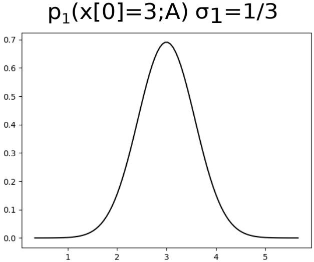
MSE trades bias against variance.

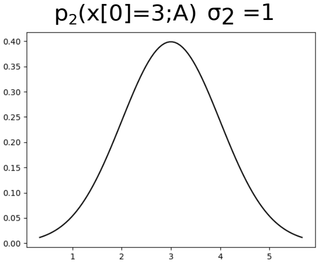
Cramer-Rao bound for unbiased estimators

The stronger a PDF depends on its parameters, the more accurate will their estimates be.

N observations $x^{(\alpha)}$ with $\epsilon^{(\alpha)} \sim N(0, \sigma^2)$

$$x^{(\alpha)} = A + \epsilon^{(\alpha)}, \qquad \hat{A} = \frac{1}{N} \sum x^{(\alpha)}$$





Accuracy can be measured by the 'sharpness' of the likelihood function ($\rightsquigarrow$ 2nd derivative of the neg. log likelihood).

Cramer-Rao bound for unbiased estimators

Fisher information matrix (Hessian matrix):

$$H_{ij} = -\left\langle \frac{\partial^2 \ln P}{\partial \mathrm{w}_i \partial \mathrm{w}_j} \right\rangle_{P(\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})} \Bigg|_{\underline{\mathbf{w}}}$$

For all unbiased estimators the following holds (Cramer-Rao Bound):

$$\underline{\mathbf{\Sigma}} - \left(\underline{\mathbf{H}}^{-1}\right) \text{ is a positive semidefinite matrix}$$

it follows:

$$var(\hat{w}_i) \geq [H^{-1}]_{ii} \text{ for all } i$$

Variance of an estimator > 1/ Fisher Information

This is a <u>universal</u> lower bound on the variance of estimators. The bound is tight.

Example: CRB for a scalar parameter w

The property of "positive semidefinite":

$$\sigma_{\mathrm{w}}^2 - \left\{ - \left\langle \frac{d^2 \ln P}{d\mathrm{w}^2} \right\rangle_{P(\underline{\mathbf{x}}^{(\alpha)};\mathrm{w})} \Bigg|_{\underline{\mathbf{w}}^*} \right\}^{-1} \geq 0$$

$$\sigma_{\mathrm{w}}^2 > - \frac{1}{\left\langle \frac{d^2 \ln P}{d\mathrm{w}^2} \right\rangle_{P(\underline{\mathbf{x}}^{(\alpha)};\mathrm{w})} \Big|_{\underline{\mathbf{w}}^*}}$$

Comment

Fisher information: precision of the estimator / interesting measure for evaluating data representations

Good estimators

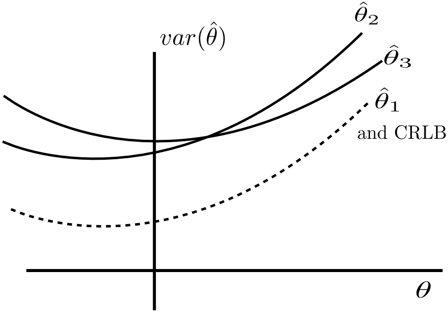
efficient estimator:

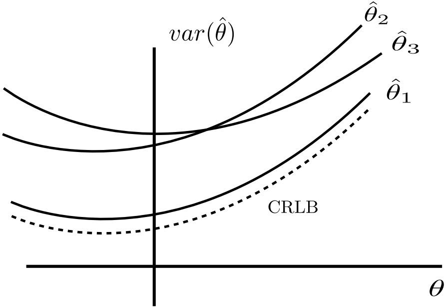
$$\underline{\mathbf{b}} = \underline{\mathbf{0}} \text{ and } \underline{\boldsymbol{\Sigma}} = \underline{\mathbf{H}}^{-1} \qquad \leftarrow \text{variance assumes lower bound}$$

unbiased minimum variance estimator:

$$\underline{\mathbf{b}} = \underline{\mathbf{0}} \text{ and } \left| \underline{\boldsymbol{\Sigma}} - \underline{\mathbf{H}}^{-1} \right| \stackrel{!}{=} \min_{\text{all estimators}}$$

Illustration: Cramer-Rao bound





Asymptotic optimality

An estimator is said to be asymptotically unbiased if for $p \to \infty$ (limit of infinite sample size):

$$E(\hat{\mathrm{w}}) \to \mathrm{w}^*$$

An estimator is said to be asymptotically efficient if for $p \to \infty$:

$$var(\hat{\mathrm{w}}) \to \text{Cramer Rao lower bound}$$

An estimator is said to be consistent if it converges to the true value for $p \to \infty$ and is asymptotically unbiased.

Results for the Maximum Likelihood estimator

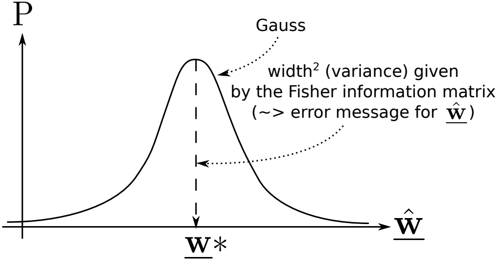
$P\left(\left\{\underline{\mathbf{x}}^{(\alpha)}\right\};\underline{\mathbf{w}}\right)$ normalized and two times differentiable

$H_{ij} = -\left\langle \frac{\partial^2 \ln P}{\partial \mathrm{w}_i \partial \mathrm{w}_j} \right\rangle_{P(\underline{\mathbf{x}}^{(\alpha)};\mathrm{w})} \Bigg|_{\underline{\mathbf{w}}^*}$ Fisher information matrix

The Maximum Likelihood estimator is consistent and asympotically unbiased and efficient.

$$\widehat{\underline{\mathbf{w}}} \sim \mathcal{N}\left(\underline{\mathbf{w}}^*, \underline{\mathbf{H}}^{-1}_{(\underline{\mathbf{w}}^*)}\right)$$

asymptotically Gaussian distributed



Summary

- An estimator is a random variable.
- $\Rightarrow$ It can only be analyzed statistically (e.g. mean, variance, shape of distribution).
- biased & unbiased estimators
- minimum variance unbiased estimator (MVU) has smallest variance for **all values** of the true parameter

MVUs and the Cramer-Rao bound

- minimum variance unbiased estimators do not always exist
- Cramer Rao Bound provides a universal bound but may not be realizable

Outlook

Inclusion of prior knowledge

- MLEs: no prior knowledge regarding 'reasonable' parameter values
- Maximum a Posteriori estimates (MAP) incorporate such knowledge via Bayes Theorem ($\rightsquigarrow$ regularisation)

$$p(\underline{\mathbf{w}}|\underline{\mathbf{x}}) \propto p(\underline{\mathbf{x}}|\underline{\mathbf{w}})p(\underline{\mathbf{w}})$$

- Beyond point estimates: Bayesian statistics. A complete (probabilistic) treatment should exploit the degrees of belief in a given model (set of parameters)