Stochastic Signals and Systems

Contents

3 Parameter Estimation

The observations $(x_1,\ldots,x_n)^T = \mathbf{x}$ are realizations of the random variables $(X_1,\ldots,X_n)^T = \mathbf{X}$ with density $f_{\mathbf{X}}(\mathbf{x})$, which is element of a known set $\{f_{\mathbf{X}}(\mathbf{x}\,|\,\boldsymbol{\theta})\colon\ \boldsymbol{\theta}\in\Omega\}$ and whose parameter vector $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_p)^T$ is unknown.

Problem:

For given observations $\mathbf{x}$ we are looking for an estimate $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1,\ldots,\hat{\theta}_p)^T$ of $\boldsymbol{\theta}$ that depends on $\mathbf{x}$, i.e. $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{x})$.

In parameter estimation problems one distinguishes whether the parameter vector is a perfectly unknown quantity or whether the prior density of the parameter vector is supposed to be known.

In the latter case one interprets the parameter vector as random vector, whose posterior density is used for the subsequent statistical inference. This approach leads to Bayes estimators.

3.1 Estimating Function and Estimator

For determining θ_i, we denote the mapping

$$(x_1,\ldots,x_n) \mapsto \hat{\theta}_i(x_1,\ldots,x_n), \quad i = 1,\ldots,p$$

a estimating function for θ_i and its function value for a given set of observations an estimate. Since $x_1,\ldots,x_n$ can be considered as random values the estimates are also random and therefore only approximate the true values θ_i.

We examine the accuracy properties of the estimates on the basis of the corresponding random variables.

Therefore, we define the random variable

$$\hat{\Theta}_i = \hat{\theta}_i(X_1,\ldots,X_n), \quad i = 1,\ldots,p,$$

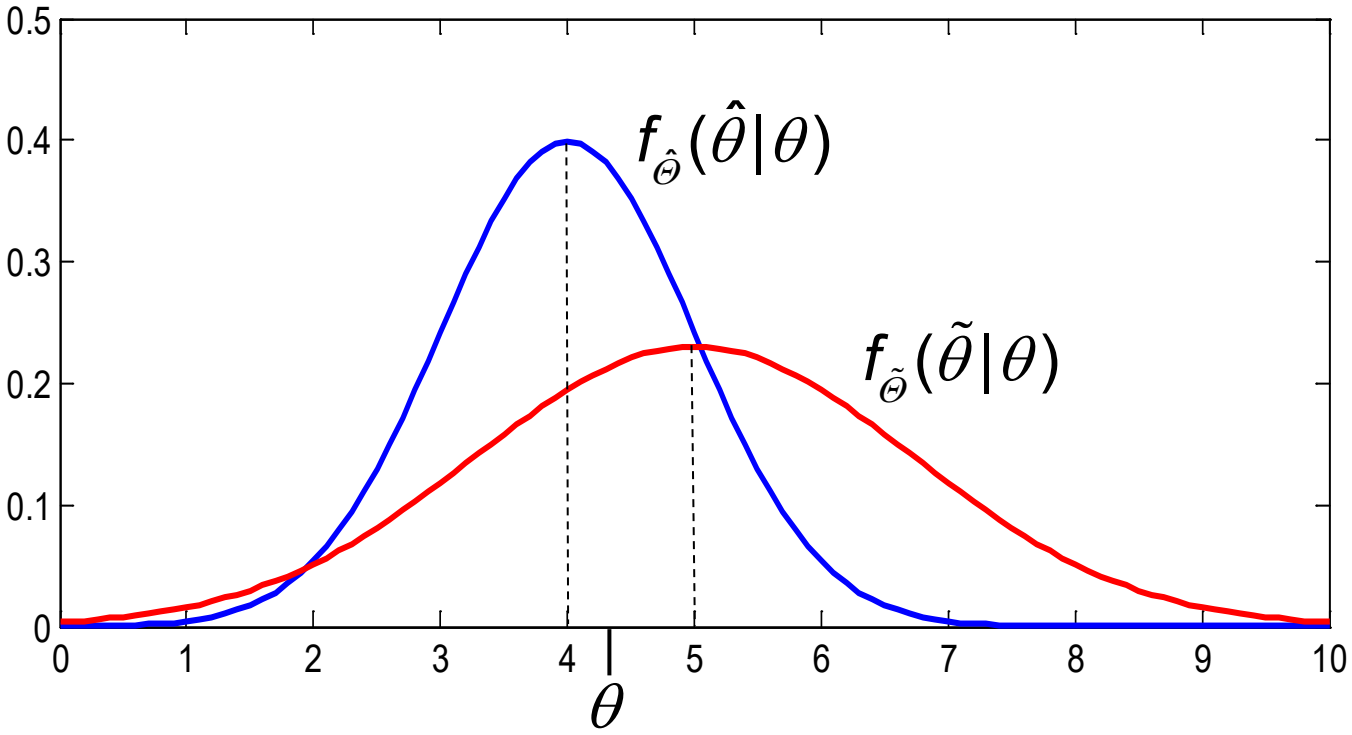
and denote it as estimator of θ_i.

To characterize the properties of the estimator

$$\hat{\boldsymbol{\Theta}} = (\hat{\Theta}_1,\ldots,\hat{\Theta}_p)^T = \hat{\boldsymbol{\theta}}(X_1,\ldots,X_n)$$

completely, the density function $f_{\hat{\boldsymbol{\Theta}}}(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta})$ that generally depends on $\boldsymbol{\theta}$ has to be determined, e.g. using the methods described in Chapter 1.7.

For simplicity let $p = 1$. Exemplarily, the densities of two alternative estimators $\hat{\Theta}$ and $\tilde{\Theta}$ for θ are depicted below.



The determination of the density function can be rather complicated. Therefore, one must be often satisfied with the determination of the moments of an estimator.

Bias:

The bias or systematic error is define by

$$\mathbf{b}(\hat{\boldsymbol{\Theta}}) = \left(b(\hat{\Theta}_1),\ldots,b(\hat{\Theta}_p)\right)^T = \left(\mathrm{E}(\hat{\Theta}_1)-\theta_1,\ldots,\mathrm{E}(\hat{\Theta}_p)-\theta_p\right)^T$$

$$= \left(\mathrm{E}(\hat{\Theta}_1),\ldots,\mathrm{E}(\hat{\Theta}_p)\right)^T - \left(\theta_1,\ldots,\theta_p\right)^T = \mathrm{E}(\hat{\boldsymbol{\Theta}})-\boldsymbol{\theta},$$

where

$$\mathrm{E}(\hat{\Theta}_i) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\hat{\theta}_i(x_1,\ldots,x_n)f_{X_1,\ldots,X_n}(x_1,\ldots,x_n\,|\,\boldsymbol{\theta})dx_1\cdots dx_n.$$

If $\mathbf{b}(\hat{\boldsymbol{\Theta}}) = \mathbf{0}$ the estimator $\hat{\boldsymbol{\Theta}}$ is said to be unbiased.

Covariance:

The covariance matrix is define by

$$\begin{aligned}\mathrm{Cov}(\hat{\boldsymbol{\Theta}}) &= \left[\mathrm{Cov}(\hat{\Theta}_i, \hat{\Theta}_j)\right]_{i,j=1,\ldots,p} \\ &= \left[\mathrm{E}\left(\left(\hat{\Theta}_i - \mathrm{E}(\hat{\Theta}_i)\right)\left(\hat{\Theta}_j - \mathrm{E}(\hat{\Theta}_j)\right)\right)\right]_{i,j=1,\ldots,p} \\ &= \mathrm{E}\left(\left(\hat{\boldsymbol{\Theta}} - \mathrm{E}(\hat{\boldsymbol{\Theta}})\right)\left(\hat{\boldsymbol{\Theta}} - \mathrm{E}(\hat{\boldsymbol{\Theta}})\right)^T\right),\end{aligned}$$

where the diagonal elements represent the variances

$$\begin{aligned}\mathrm{Var}(\hat{\Theta}_i) &= \mathrm{Cov}(\hat{\Theta}_i, \hat{\Theta}_i) \\ &= \mathrm{E}(\hat{\Theta}_i - \mathrm{E}\hat{\Theta}_i)^2 = \mathrm{E}(\hat{\Theta}_i^2) - (\mathrm{E}\hat{\Theta}_i)^2\end{aligned}$$

for $i = 1, \ldots, p$.

Mean Square Error:

The matrix-valued mean square error is define by

$$\begin{aligned}\mathrm{MSE}(\hat{\boldsymbol{\Theta}}) &= \left[\mathrm{E}\left((\hat{\Theta}_i-\theta_i)(\hat{\Theta}_j-\theta_j)\right)\right]_{i,j=1,\ldots,p}\\ &= \mathrm{E}\left((\hat{\boldsymbol{\Theta}}-\boldsymbol{\theta})(\hat{\boldsymbol{\Theta}}-\boldsymbol{\theta})^T\right)\\ &= \mathrm{Cov}(\hat{\boldsymbol{\Theta}})+\mathbf{b}(\hat{\boldsymbol{\Theta}})\mathbf{b}(\hat{\boldsymbol{\Theta}})^T,\end{aligned}$$

where the diagonal elements represent the mean square errors of the individual components $\hat{\Theta}_i$ given by

$$\mathrm{MSE}(\hat{\Theta}_i)=\mathrm{E}\left((\hat{\Theta}_i-\theta_i)^2\right)=\mathrm{Var}(\hat{\Theta}_i)+b(\hat{\Theta}_i)^2$$

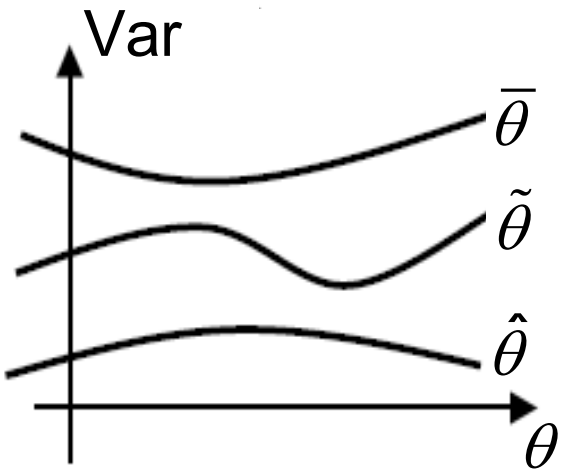
for $i = 1,\ldots,p$.

Minimum Variance Unbiased (MVU) Estimator:

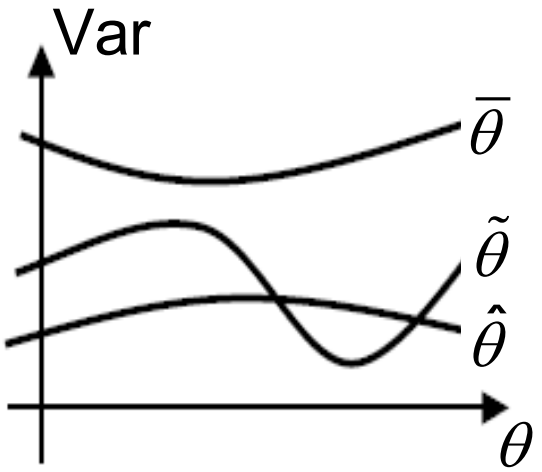
An unbiased estimator $\hat{\boldsymbol{\Theta}}$ is MVU if for any unbiased estimator $\tilde{\boldsymbol{\Theta}}$ the following inequality holds.

$$\mathbf{a}^T \mathrm{Cov}(\hat{\boldsymbol{\Theta}})\mathbf{a} \leq \mathbf{a}^T \mathrm{Cov}(\tilde{\boldsymbol{\Theta}})\mathbf{a} \quad \forall \mathbf{a} \in \mathbb{R}^p, \ \forall \boldsymbol{\theta} \in \Omega$$

MVU exists



no MVU exists



Note:

To emphasize that the variance is smallest for all $\boldsymbol{\theta} \in \Omega$ an MVU estimator is sometimes called uniformly minimum variance unbiased UMVU estimator.

Minimum Mean Square Error (MMSE) Estimator:

An estimator $\hat{\boldsymbol{\Theta}}$ provides an MMSE if for any estimator $\tilde{\boldsymbol{\Theta}}$ the following inequality holds.

$$\mathbf{a}^T \mathrm{MSE}(\hat{\boldsymbol{\Theta}})\mathbf{a} \leq \mathbf{a}^T \mathrm{MSE}(\tilde{\boldsymbol{\Theta}})\mathbf{a} \quad \forall \mathbf{a} \in \mathbb{R}^p, \ \forall \boldsymbol{\theta} \in \Omega$$

Linear Estimator:

An estimator $\hat{\boldsymbol{\Theta}}$ is called linear if it can be expressed as a linear function of the observations, i.e.

$$\hat{\boldsymbol{\Theta}} = \mathbf{A}\mathbf{X} \quad \text{with} \quad \mathbf{X} = (X_1, \ldots, X_n) \ \text{ and } \ \mathbf{A} = \left(a_{i,j}\right)_{i=1,\ldots,p;\, j=1,\ldots,n}$$

Exercise 3.1-1:
MVU and MMSE variance estimator

Consistency:

Often one is interested in the behavior of an estimator if the number of observations grows.

An estimator is said to be

- strongly consistent if $\hat{\boldsymbol{\Theta}}_n$ converges with probability 1 towards $\boldsymbol{\theta}$, i.e.
  $$\hat{\boldsymbol{\Theta}}_n = \hat{\boldsymbol{\theta}}(X_1,\ldots,X_n) \xrightarrow[n\to\infty]{a.s.} \boldsymbol{\theta}$$
- mean square consistent if $\hat{\boldsymbol{\Theta}}_n$ converges in mean square sense towards $\boldsymbol{\theta}$, i.e.
  $$\hat{\boldsymbol{\Theta}}_n = \hat{\boldsymbol{\theta}}(X_1,\ldots,X_n) \xrightarrow[n\to\infty]{m.s.} \boldsymbol{\theta}$$
- consistent if $\hat{\boldsymbol{\Theta}}_n$ converges in probability towards $\boldsymbol{\theta}$, i.e.
  $$\hat{\boldsymbol{\Theta}}_n = \hat{\boldsymbol{\theta}}(X_1,\ldots,X_n) \xrightarrow[n\to\infty]{P} \boldsymbol{\theta}$$

Asymptotic Normality:

In certain cases one can show that an estimator is asymptotically normally distributed such that

$$\lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\Theta}}_n - \boldsymbol{\theta}) \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}).$$

Consequently, for large n the density function of $\hat{\boldsymbol{\Theta}}_n$ can be approximated by

$$f_{\hat{\boldsymbol{\Theta}}_n}(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta}) = \frac{n^{p/2}}{(2\pi)^{p/2}\sqrt{\det(\boldsymbol{\Sigma})}} \cdot \exp\left\{-\frac{n}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\right\}.$$

3.2 Sufficient Statistic, Exponential Families

A function $\mathbf{T} = \mathbf{t}(\mathbf{X})$ that is only depending on the observation model $\mathbf{X}$ is called a statistic.

Let $\mathbf{X} = (X_1,\ldots,X_n)^T$ be a model of a binary sequence with probability p and $1-p$ of observing a one and a zero respectively. Furthermore, we suppose that the $X_1,\ldots,X_n$ are independent.

Thus, our model can be described by the probabilities

$$P(\mathbf{X} = \mathbf{x} \mid p) = \prod_{k=1}^{n} p^{x_k} (1-p)^{1-x_k}$$

$$= p^{\sum_{k=1}^{n} x_k} (1-p)^{n-\sum_{k=1}^{n} x_k} = p^{t(\mathbf{x})} (1-p)^{n-t(\mathbf{x})}$$

Now the following question arises. Do we collect all information in terms of inference about *p* by recording only

$$t(\mathbf{x}) = \sum_{k=1}^{n} x_k$$

The statistical understanding about the collection of all information is quantified in the following definition.

Definition:

A statistic $\mathbf{T} = \mathbf{t}(\mathbf{X})$ is called sufficient for the parameter $\boldsymbol{\theta}$ if the conditional distribution of $\mathbf{X}$ given $\mathbf{T} = \mathbf{t}(\mathbf{x})$ is independent of $\boldsymbol{\theta}$ for all $\mathbf{t}$, i.e.

$$F_{\mathbf{X}}\left(\mathbf{x} \mid \mathbf{T} = \mathbf{t}(\mathbf{x}); \boldsymbol{\theta}\right) = F_{\mathbf{X}}\left(\mathbf{x} \mid \mathbf{T} = \mathbf{t}(\mathbf{x})\right).$$

Hence, $\mathbf{T}$ contains "all information" about $\boldsymbol{\theta}$ included in $\mathbf{x}$.

Exercise 3.2-1:

Sufficiency of $t(\mathbf{x}) = \sum_{k=1}^{n} x_k$

Because the conditional distribution has to be determined a direct evaluation of sufficiency is usually difficult.

Fortunately, the following theorem exist whose conditions can be verified easily.

Theorem: *(factorization theorem for densities)*

A necessary and sufficient condition for a statistic $\mathbf{T} = \mathbf{t}(\mathbf{X})$ to be sufficient is that there exist non-negative functions $g(\mathbf{t} \mid \boldsymbol{\theta})$ and $h(\mathbf{x})$ such that $f_{\mathbf{x}}(\mathbf{x} \mid \boldsymbol{\theta})$ satisfies

$$f_{\mathbf{x}}(\mathbf{x} \mid \theta) = g\left(\mathbf{t}(\mathbf{x}) \mid \boldsymbol{\theta}\right) \cdot h(\mathbf{x}).$$

Exercise 3.2-2:
Proof of the Theorem

Exercise 3.2-3:
Sufficient statistic for mean and variance

Minimal Sufficient Statistic

A sufficient statistic **T** is said to be minimal if of all sufficient statistics it provides the greatest possible reduction of data, i.e. if for any sufficient statistic **T'** there exists a function **s** such that $\mathbf{T} = \mathbf{s}(\mathbf{T}')$.

Complete Sufficient Statistic

A sufficient statistic **T** is said to be complete (or unique) if the condition

$$\mathrm{E}_{\boldsymbol{\theta}}\left(\mathbf{f}(\mathbf{T})\right) = \mathbf{0} \quad \forall \boldsymbol{\theta} \in \Omega$$

implies $\mathbf{f}(\mathbf{T}) = \mathbf{0}$ with probability 1 for all $\boldsymbol{\theta}$.

Note: Completeness ensures minimality

Exercise 3.2-4:
Minimal and complete sufficient statistics

Uniqueness of unbiased estimating function

Completeness implies that there is just one estimating function of the sufficient statistic that provides an unbiased estimator of $\boldsymbol{\theta}$.

Let $\mathbf{T}$ be a complete sufficient statistic and $\mathbf{f}_1$ and $\mathbf{f}_2$ be two functions such that

$$\mathrm{E}_\theta\left(\mathbf{f}_1(\mathbf{T})\right) = \mathrm{E}_\theta\left(\mathbf{f}_2(\mathbf{T})\right) = \boldsymbol{\theta} \quad \forall \boldsymbol{\theta} \in \Omega.$$

Then

$$\mathrm{E}_\theta\left(\mathbf{f}_1(\mathbf{T}) - \mathbf{f}_2(\mathbf{T})\right) = \mathrm{E}_\theta\left(\mathbf{f}(\mathbf{T})\right) = \mathbf{0} \quad \forall \boldsymbol{\theta} \in \Omega$$

and due to the completeness of $\mathbf{T}$

$$\mathbf{f}(\mathbf{T}) = \mathbf{0} \Rightarrow \mathbf{f}_1(\mathbf{T}) = \mathbf{f}_2(\mathbf{T}) \quad \text{with probability 1} \quad \forall \boldsymbol{\theta} \in \Omega$$

Theorem: (Rao-Blackwell)

Let $\tilde{\boldsymbol{\Theta}}$ be an unbiased estimator of $\boldsymbol{\theta}$ and $\mathbf{T} = \mathbf{t}(\mathbf{X})$ be a sufficient statistic for $\boldsymbol{\theta}$. Then the estimator defined by

$$\hat{\boldsymbol{\Theta}} = \mathrm{E}(\tilde{\boldsymbol{\Theta}} \,|\, \mathbf{T})$$

is unbiased and improves on $\tilde{\boldsymbol{\Theta}}$ as follows.

$$\mathbf{a}^T \mathrm{Cov}(\hat{\boldsymbol{\Theta}})\mathbf{a} \leq \mathbf{a}^T \mathrm{Cov}(\tilde{\boldsymbol{\Theta}})\mathbf{a} \quad \forall \mathbf{a} \in \mathbb{R}^p$$

Theorem: (Lehmann-Scheffe)

If in addition to the assumptions employed for the Rao-Blackwell theorem the sufficient statistic is complete, then $\hat{\boldsymbol{\Theta}} = \mathrm{E}(\tilde{\boldsymbol{\Theta}} \,|\, \mathbf{T})$ is unique MVU estimator of $\boldsymbol{\theta}$.

Exercise 3.2-5:
Proof of the Theorems

Exponential Families

A family $\{F_{\mathbf{x}}(\mathbf{x}\,|\,\boldsymbol{\theta})\}$ of distributions is forming a *k*-dimensional exponential family if the distributions $F_{\mathbf{x}}(\mathbf{x}\,|\,\boldsymbol{\theta})$ have densities of the form

$$f_{\mathbf{x}}(\mathbf{x}\,|\,\boldsymbol{\theta}) = h(\mathbf{x}) \cdot \exp\left(\sum_{i=1}^{k} \xi_i(\boldsymbol{\theta}) t_i(\mathbf{x}) - B(\boldsymbol{\theta}) \right).$$

Frequently, it is more convenient to use the ξ_i as the parameters and write the density in the canonical form

$$f_{\mathbf{x}}(\mathbf{x}\,|\,\boldsymbol{\xi}) = h(\mathbf{x}) \cdot \exp\left(\sum_{i=1}^{k} \xi_i\, t_i(\mathbf{x}) - A(\boldsymbol{\xi}) \right).$$

Applying the factorization theorem for densities one can easily observe that

$$\mathbf{T} = \left(T_1,\ldots,T_k\right)^T = \left(t_1(\mathbf{X}),\ldots,t_k(\mathbf{X})\right)^T$$

constitutes a sufficient statistic for the exponential family.

Note:

The parameter space $\Xi \subset \mathbb{R}^k$ of the natural parameter vector $\boldsymbol{\xi} = (\xi_1,\ldots,\xi_k)^T$ is convex.

If the exponential family is of full rank, i.e. the parameter space Ξ contains a k-dimensional rectangle, then $\mathbf{T} = (T_1, \ldots, T_k)^T$ is complete.

For exponential families one can claim, that the integral

$$\int_{\mathbb{R}^n} f_{\mathbf{X}}(\mathbf{x}\,|\,\boldsymbol{\xi})\,dx = \int_{\mathbb{R}^n} h(\mathbf{x}) \cdot \exp\left(\sum_{i=1}^{k} \xi_i t_i(\mathbf{x}) - A(\boldsymbol{\xi}) \right) dx = 1$$

has derivatives of all orders with respect to the ξ_i which can be obtained by differentiating under the integral sign.

Exploiting the properties of the integral we can find

$$\mathrm{E}\left(T_i\right) = \mathrm{E}\left(t_i(\mathbf{X})\right) = \frac{\partial}{\partial \xi_i} A(\boldsymbol{\xi})$$

and

$$\mathrm{Cov}\left(T_i, T_j\right) = \mathrm{Cov}\left(t_i(\mathbf{X}), t_j(\mathbf{X})\right) = \frac{\partial^2}{\partial \xi_i \partial \xi_j} A(\boldsymbol{\xi})$$

Exercise 3.2-6:

Rayleigh distribution, mean and variance of the natural sufficient statistic

3.3 Linear Least Squares Estimation

Consider the linear model

$$\mathbf{X} = \mathbf{H}\boldsymbol{\theta} + \mathbf{Z} \quad \text{with} \quad X_i = h_{i1}\theta_1 + \ldots + h_{ip}\theta_p + Z_i, \;\; i = 1, \ldots, n,$$

where $\mathbf{X}$ and $\mathbf{Z}$ are $n \times 1$ vectors modeling the measurements and the measurement noise, respectively. Furthermore, $\mathbf{H}$ denotes a known $n \times p$ matrix and $\boldsymbol{\theta}$ the $p \times 1$ parameter vector that has to be estimated.

The measurement noise is supposed to be statistically characterized by $\mathrm{E}(\mathbf{Z}) = \mathbf{0}$ and $\mathrm{Cov}(\mathbf{Z}) = \mathrm{E}(\mathbf{Z}\mathbf{Z}^T) = \sigma_Z^2 \mathbf{I}$.

Hence, the measurement model $\mathbf{X}$ possess the mean vector $\mathrm{E}(\mathbf{X}) = \mathbf{H}\boldsymbol{\theta}$ and covariance matrix $\mathrm{Cov}(\mathbf{X}) = \sigma_Z^2 \mathbf{I}$.

Exercise 3.3-1:

Model examples,

- *polynomial curve fitting,*
- *amplitude and phase estimation of sinusoids*
- *FIR filter identification*

Now, the least squares criterion can be expressed by

$$q(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(x_i - (h_{i1}\theta_1 + \ldots + h_{ip}\theta_p) \right)^2$$

$$= (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T (\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) = \mathbf{x}^T\mathbf{x} - 2\boldsymbol{\theta}^T\mathbf{H}^T\mathbf{x} + \boldsymbol{\theta}^T\mathbf{H}^T\mathbf{H}\boldsymbol{\theta}.$$

Differentiating with respect to $\boldsymbol{\theta}$ by using the identities

$$\nabla_{\boldsymbol{\theta}}\, \boldsymbol{\theta}^T\mathbf{A}\,\mathbf{a} = \mathbf{A}\,\mathbf{a}, \quad \nabla_{\boldsymbol{\theta}}\, \boldsymbol{\theta}^T\mathbf{A}\,\boldsymbol{\theta} = (\mathbf{A} + \mathbf{A}^T)\boldsymbol{\theta}$$

results after equating to zero in the following so-called normal equation system

$$\mathbf{H}^T\mathbf{H}\hat{\boldsymbol{\theta}} = \mathbf{H}^T\mathbf{x}.$$

A solution of the normal equation system is given by

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T\mathbf{H})^{-}\mathbf{H}^T\mathbf{x},$$

where $(\mathbf{H}^T\mathbf{H})^-$ denotes a generalized inverse.

1) If the rank$(\mathbf{H}) = p$, the number of unknown parameters, then $(\mathbf{H}^T\mathbf{H})^- = (\mathbf{H}^T\mathbf{H})^{-1}$ is the ordinary inverse.

2) Let rank$(\mathbf{H}) = M < p$, i.e. either $n < p$ or the columns of $\mathbf{H}$ are linearly dependent, then

$$(\mathbf{H}^T\mathbf{H})^- = (\mathbf{H}^T\mathbf{H})^+ = \sum_{m=1}^{M} \frac{1}{\lambda_m} \mathbf{u}_m \mathbf{u}_m^T$$

is the Moore-Penrose inverse, where $\lambda_1, \ldots, \lambda_M$ and $\mathbf{u}_1, \ldots, \mathbf{u}_M$ denote the non-zero eigenvalues and corresponding eigenvectors of $\mathbf{H}^T\mathbf{H}$, respectively.

Supplement:

A generalized inverse of $\mathbf{A}$ is defined by the property

$$\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}.$$

It is not unique. The Moore-Penrose inverse of $\mathbf{A}$ is defined by the properties

$$\mathbf{A}\mathbf{A}^{+}\mathbf{A} = \mathbf{A}, \qquad \mathbf{A}^{+}\mathbf{A}\mathbf{A}^{+} = \mathbf{A}^{+},$$
$$(\mathbf{A}\mathbf{A}^{+})^{T} = \mathbf{A}\mathbf{A}^{+} \qquad (\mathbf{A}^{+}\mathbf{A})^{T} = \mathbf{A}^{+}\mathbf{A}.$$

It is unique and provides the minimum length solution of the linear equation system $\mathbf{A}\mathbf{x} = \mathbf{b}$.

The Moore-Penrose inverse $\mathbf{A}^{+}$ can be determined by employing the singular value decomposition of $\mathbf{A}$.

Exercise 3.3-2:

Normal equation system and generalized inverse

Properties of the least squares estimator

The mean vector is given by

$$\mathrm{E}(\hat{\boldsymbol{\Theta}}) = (\mathbf{H}^T\mathbf{H})^{-}\mathbf{H}^T\mathrm{E}(\mathbf{X}) = (\mathbf{H}^T\mathbf{H})^{-}\mathbf{H}^T\mathbf{H}\boldsymbol{\theta}$$

$$= \begin{cases} \boldsymbol{\theta} & \mathrm{rank}(\mathbf{H}) = p \\ \mathbf{V}_1\mathbf{V}_1^T\,\boldsymbol{\theta} = (\mathbf{I} - \mathbf{V}_2\mathbf{V}_2^T)\boldsymbol{\theta} & \mathrm{rank}(\mathbf{H}) = M < p \end{cases},$$

where the matrices $\mathbf{V}_1$ and $\mathbf{V}_2$ can be derived from the singular value decomposition

$$\mathbf{H}^T\mathbf{H} = \mathbf{V}\,diag(\lambda_1,\ldots,\lambda_M,0,\ldots,0)\,\mathbf{V}^T$$

with

$$\mathbf{V} = \left(\mathbf{V}_1\,,\,\mathbf{V}_2\right), \quad \mathbf{V}_1 = (\mathbf{v}_1,\ldots,\mathbf{v}_M) \quad \mathbf{V}_2 = (\mathbf{v}_{M+1},\ldots,\mathbf{v}_p).$$

The covariance matrix of the least squares estimator

$$\mathbf{C}_{\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Theta}}} = \mathrm{Cov}(\hat{\boldsymbol{\Theta}}) = \mathrm{E}\left(\left(\hat{\boldsymbol{\Theta}} - \mathrm{E}(\hat{\boldsymbol{\Theta}})\right)\left(\hat{\boldsymbol{\Theta}} - \mathrm{E}(\hat{\boldsymbol{\Theta}})\right)^T \right)$$

results after exploiting

$$\hat{\boldsymbol{\Theta}} = (\mathbf{H}^T\mathbf{H})^{-}\mathbf{H}^T\mathbf{X} \quad \text{and} \quad \mathrm{E}(\hat{\boldsymbol{\Theta}}) = (\mathbf{H}^T\mathbf{H})^{-}\mathbf{H}^T\mathbf{H}\boldsymbol{\theta}$$

in the expression

$$\begin{aligned}
\mathbf{C}_{\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Theta}}} &= \mathrm{E}\left((\mathbf{H}^T\mathbf{H})^{-}\mathbf{H}^T(\mathbf{X} - \mathbf{H}\boldsymbol{\theta})(\mathbf{X} - \mathbf{H}\boldsymbol{\theta})^T\mathbf{H}(\mathbf{H}^T\mathbf{H})^{-} \right) \\
&= (\mathbf{H}^T\mathbf{H})^{-}\mathbf{H}^T\,\mathrm{E}\left((\mathbf{X} - \mathbf{H}\boldsymbol{\theta})(\mathbf{X} - \mathbf{H}\boldsymbol{\theta})^T \right)\mathbf{H}(\mathbf{H}^T\mathbf{H})^{-} \\
&= (\mathbf{H}^T\mathbf{H})^{-}\mathbf{H}^T\,(\sigma_Z^2\,\mathbf{I})\,\mathbf{H}(\mathbf{H}^T\mathbf{H})^{-} = \sigma_Z^2(\mathbf{H}^T\mathbf{H})^{-}.
\end{aligned}$$

Exercise 3.3-3:

LSE for p = 1 and sample mean

Theorem: (Gauß-Markov)

Given the model

$$\mathbf{X} = \mathbf{H\theta} + \mathbf{Z} \quad \text{with} \quad X_i = h_{i1}\theta_1 + \ldots + h_{ip}\theta_p + Z_i, \;\; i = 1, \ldots, n,$$

where

$$\mathrm{E}(\mathbf{X}) = \mathbf{H\theta}, \quad \mathbf{Cov}(\mathbf{X}) = \mathbf{Cov}(\mathbf{Z}) = \sigma_Z^2 \mathbf{I}$$

and

$$\mathrm{rank}(\mathbf{H}) = p.$$

Then the best linear unbiased estimator (BLUE) is the least squares estimator

$$\hat{\mathbf{\Theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{X}.$$

Exercise 3.3-4:
Proof of the Gauß-Markov Theorem

Consequently, the minimum of the sum of squares is

$$\begin{aligned} q(\hat{\boldsymbol{\theta}}) &= \left(\mathbf{x} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}\right)^T \left(\mathbf{x} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}\right) \\ &= (\mathbf{x} - \mathbf{P}\mathbf{x})^T(\mathbf{x} - \mathbf{P}\mathbf{x}) = \mathbf{x}^T(\mathbf{I} - \mathbf{P})^T(\mathbf{I} - \mathbf{P})\mathbf{x} \\ &= \mathbf{x}^T\mathbf{P}^{\perp T}\mathbf{P}^{\perp}\mathbf{x} = \mathbf{x}^T\mathbf{P}^{\perp}\mathbf{x}, \end{aligned}$$

where

$$\mathbf{P} = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T \quad \text{and} \quad \mathbf{P}^{\perp} = \mathbf{I} - \mathbf{P} = \mathbf{I} - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$$

are projection matrices, which project a vector $\mathbf{a} \in \mathbb{R}^n$ by $\mathbf{P}\mathbf{a}$ and $\mathbf{P}^{\perp}\mathbf{a}$ into $R(\mathbf{H})$ and $N(\mathbf{H}^T)$, respectively.

$R(\mathbf{H})$: range of $\mathbf{H}$, i.e. the space spanned by the columns of $\mathbf{H}$

$N(\mathbf{H}^T)$: nullspace of $\mathbf{H}^T$, i.e. the space that is orthogonal to $R(\mathbf{H})$

Thus, $\mathbf{H}^T\mathbf{P}^\perp=\mathbf{0}$, $\mathbf{P}^\perp\mathbf{H}=\mathbf{0}$ and $\mathbf{P}\mathbf{P}^\perp=\mathbf{P}^\perp\mathbf{P}=\mathbf{0}$. The projection matrices are also symmetric and idempotent, i.e.

$$\mathbf{P}=\mathbf{P}^T, \quad \mathbf{P}=\mathbf{P}\mathbf{P} \quad \text{and} \quad \mathbf{P}^\perp=\mathbf{P}^{\perp T}, \quad \mathbf{P}^\perp=\mathbf{P}^\perp\mathbf{P}^\perp.$$

Moreover, by employing the trace of a square matrix

$$\mathrm{tr}(\mathbf{A})=\sum_{i=1}^{n} a_{ii} \quad \text{with} \quad \mathbf{A}=\left(a_{ij}\right)_{i,j=1,\ldots,n}$$

together with its property

$$\mathrm{tr}(\mathbf{ABC})=\mathrm{tr}(\mathbf{BCA})=\mathrm{tr}(\mathbf{CAB}),$$

where $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ are $n\times k$, $k\times l$ and $l\times n$ matrices the minimum of the sum of squares can be expressed by

$$q(\hat{\boldsymbol{\theta}})=\mathbf{x}^T\mathbf{P}^\perp\mathbf{x}=\mathrm{tr}(\mathbf{P}^\perp\mathbf{x}\mathbf{x}^T).$$

Now we want to consider how the variance of the noise model σ_z^2 can be estimated. The expected value of the minimum of the sum of squares provides

$$\begin{aligned}
\mathrm{E}\left(q(\hat{\boldsymbol{\Theta}})\right) &= \mathrm{E}\left(\mathrm{tr}(\mathbf{P}^{\perp}\mathbf{X}\mathbf{X}^T)\right) = \mathrm{tr}\left(\mathbf{P}^{\perp}\mathrm{E}\left((\mathbf{H}\boldsymbol{\theta}+\mathbf{Z})(\mathbf{H}\boldsymbol{\theta}+\mathbf{Z})^T\right)\right) \\
&= \mathrm{tr}\left(\mathbf{P}^{\perp}\left(\mathbf{H}\boldsymbol{\theta}\boldsymbol{\theta}^T\mathbf{H}^T + \mathrm{E}(\mathbf{Z})\boldsymbol{\theta}^T\mathbf{H}^T + \mathbf{H}\boldsymbol{\theta}\,\mathrm{E}(\mathbf{Z}^T) + \mathrm{E}(\mathbf{Z}\mathbf{Z}^T)\right)\right) \\
&= \mathrm{tr}\left(\mathbf{P}^{\perp}(\mathbf{H}\boldsymbol{\theta}\boldsymbol{\theta}^T\mathbf{H}^T + \sigma_z^2\mathbf{I})\right) = \sigma_z^2\,\mathrm{tr}(\mathbf{P}^{\perp}) = (n-p)\sigma_z^2,
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{tr}(\mathbf{P}^{\perp}) &= \mathrm{tr}\left(\mathbf{I}_n - \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\right) \\
&= n - \mathrm{tr}\left((\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{H}\right) = n - \mathrm{tr}(\mathbf{I}_p) = n - p
\end{aligned}$$

has been exploited.

This result motivates

$$S^2 = q(\hat{\mathbf{\Theta}}) \big/ (n-p)$$

as an unbiased estimator for σ_Z^2.

Theorem:

If in addition to the Gauß-Markov theorem $\mathbf{Z}$ obeys a normal distribution, i.e. $\mathbf{Z} \sim \mathcal{N}_n(\mathbf{0}, \sigma_Z^2 \mathbf{I})$, the following holds.

a) $\hat{\mathbf{\Theta}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{X} \sim \mathcal{N}_p\left(\boldsymbol{\theta}, \sigma_Z^2 (\mathbf{H}^T\mathbf{H})^{-1}\right)$

b) $\hat{\mathbf{\Theta}}$ and S^2 are stochastically independent

c) $(n-p)\big/\sigma_Z^2 \cdot S^2 \sim \chi^2_{n-p}$

Exercise 3.3-5:
Proof of the Theorem

Exercise 3.3-6:
Amplitude and phase estimation of sinusoids

Exercise 3.3-7:
System identification (FIR filter)

In the following we generalize the linear model to

$$\mathbf{X} = \mathbf{H}\boldsymbol{\theta} + \mathbf{U} \quad \text{with} \quad X_i = h_{i1}\theta_1 + \ldots + h_{ip}\theta_p + U_i, \;\; i = 1, \ldots, n,$$

where the measurement noise possesses still the mean $\mathrm{E}(\mathbf{U}) = \mathbf{0}$ but now the covariance matrix $\mathrm{Cov}(\mathbf{U}) = \sigma_U^2 \mathbf{C}_{\mathbf{UU}}$.

Prewhitening

If $\mathbf{C}_{\mathbf{UU}}$ is decomposed, e.g. by the Cholesky decomposition $\mathbf{C}_{\mathbf{UU}} = \mathbf{C}\mathbf{C}^T$, and

$$\mathbf{Y} = \mathbf{C}^{-1}\mathbf{X}, \quad \mathbf{K} = \mathbf{C}^{-1}\mathbf{H}, \quad \mathbf{W} = \mathbf{C}^{-1}\mathbf{U}$$

be introduced, the linear model can be reformulated to

$$\mathbf{Y} = \mathbf{C}^{-1}\mathbf{X} = \mathbf{C}^{-1}\mathbf{H}\boldsymbol{\theta} + \mathbf{C}^{-1}\mathbf{U} = \mathbf{K}\boldsymbol{\theta} + \mathbf{W}$$

with

$$\mathrm{E}(\mathbf{Y}) = \mathbf{K}\boldsymbol{\theta} = \mathbf{C}^{-1}\mathbf{H}\boldsymbol{\theta} \quad \text{and} \quad \mathrm{Cov}(\mathbf{Y}) = \mathrm{Cov}(\mathbf{W}) = \sigma_U^2 \mathbf{I}.$$

Hence, the weighted least squares criterion for the generalized linear model can be derived as follows.

$$q(\boldsymbol{\theta}) = (\mathbf{y} - \mathbf{K}\boldsymbol{\theta})^T (\mathbf{y} - \mathbf{K}\boldsymbol{\theta})$$

$$= \left(\mathbf{C}^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})\right)^T \left(\mathbf{C}^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})\right)$$

$$= (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T (\mathbf{C}^{-1})^T \mathbf{C}^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{C}_{\mathbf{UU}}^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}).$$

The normal equation system

$$\mathbf{K}^T\mathbf{K}\hat{\boldsymbol{\theta}} = \mathbf{K}^T\mathbf{y} \quad \text{resp.} \quad \mathbf{H}^T\mathbf{C}_{\mathbf{UU}}^{-1}\mathbf{H}\hat{\boldsymbol{\theta}} = \mathbf{H}^T\mathbf{C}_{\mathbf{UU}}^{-1}\mathbf{x}$$

and its solution

$$\hat{\boldsymbol{\theta}} = (\mathbf{K}^T\mathbf{K})^{-}\mathbf{K}^T\mathbf{y} = (\mathbf{H}^T\mathbf{C}_{\mathbf{UU}}^{-1}\mathbf{H})^{-}\mathbf{H}^T\mathbf{C}_{\mathbf{UU}}^{-1}\mathbf{x}$$

are obtained analogous to the white noise case.

Properties of the generalized least squares estimator

The mean vector is given by

$$\mathrm{E}(\hat{\boldsymbol{\Theta}}) = (\mathbf{H}^T \mathbf{C}_{\mathbf{UU}}^{-1} \mathbf{H})^{-} \mathbf{H}^T \mathbf{C}_{\mathbf{UU}}^{-1} \mathrm{E}(\mathbf{X}) = (\mathbf{H}^T \mathbf{C}_{\mathbf{UU}}^{-1} \mathbf{H})^{-} \mathbf{H}^T \mathbf{C}_{\mathbf{UU}}^{-1} \mathbf{H}\boldsymbol{\theta}$$

$$= \begin{cases} \boldsymbol{\theta} & \mathrm{rank}(\mathbf{H}) = p \\ \mathbf{V}_1 \mathbf{V}_1^T \boldsymbol{\theta} = (\mathbf{I} - \mathbf{V}_2 \mathbf{V}_2^T)\boldsymbol{\theta} & \mathrm{rank}(\mathbf{H}) = M < p \end{cases},$$

where the matrices $\mathbf{V}_1$ and $\mathbf{V}_2$ can be derived from the singular value decomposition

$$\mathbf{H}^T \mathbf{C}_{\mathbf{UU}}^{-1} \mathbf{H} = \mathbf{V}\, diag(\lambda_1, \ldots, \lambda_M, 0, \ldots, 0)\, \mathbf{V}^T$$

with

$$\mathbf{V} = \left(\mathbf{V}_1, \mathbf{V}_2\right), \quad \mathbf{V}_1 = (\mathbf{v}_1, \ldots, \mathbf{v}_M) \quad \mathbf{V}_2 = (\mathbf{v}_{M+1}, \ldots, \mathbf{v}_p).$$

The covariance matrix of the generalized least squares estimator

$$\mathbf{C}_{\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Theta}}} = \mathrm{Cov}(\hat{\boldsymbol{\Theta}}) = \mathrm{E}\left(\left(\hat{\boldsymbol{\Theta}} - \mathrm{E}(\hat{\boldsymbol{\Theta}})\right)\left(\hat{\boldsymbol{\Theta}} - \mathrm{E}(\hat{\boldsymbol{\Theta}})\right)^T \right)$$

results after exploiting

$$\hat{\boldsymbol{\Theta}} = (\mathbf{H}^T \mathbf{C}_{\mathbf{UU}}^{-1} \mathbf{H})^{-} \mathbf{H}^T \mathbf{C}_{\mathbf{UU}}^{-1} \mathbf{X} \quad \text{and} \quad \mathrm{E}(\hat{\boldsymbol{\Theta}}) = (\mathbf{H}^T \mathbf{C}_{\mathbf{UU}}^{-1} \mathbf{H})^{-} \mathbf{H}^T \mathbf{C}_{\mathbf{UU}}^{-1} \mathbf{H}\boldsymbol{\theta}$$

in the expression

$$\begin{aligned}
\mathbf{C}_{\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Theta}}} &= \mathrm{E}\left((\mathbf{H}^T \mathbf{C}_{\mathbf{UU}}^{-1}\mathbf{H})^{-} \mathbf{H}^T \mathbf{C}_{\mathbf{UU}}^{-1} (\mathbf{X} - \mathbf{H}\boldsymbol{\theta})(\mathbf{X} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{C}_{\mathbf{UU}}^{-1} \mathbf{H} (\mathbf{H}^T \mathbf{C}_{\mathbf{UU}}^{-1} \mathbf{H})^{-} \right) \\
&= (\mathbf{H}^T \mathbf{C}_{\mathbf{UU}}^{-1}\mathbf{H})^{-} \mathbf{H}^T \mathbf{C}_{\mathbf{UU}}^{-1} \mathrm{E}\left((\mathbf{X} - \mathbf{H}\boldsymbol{\theta})(\mathbf{X} - \mathbf{H}\boldsymbol{\theta})^T \right) \mathbf{C}_{\mathbf{UU}}^{-1} \mathbf{H} (\mathbf{H}^T \mathbf{C}_{\mathbf{UU}}^{-1} \mathbf{H})^{-} \\
&= (\mathbf{H}^T \mathbf{C}_{\mathbf{UU}}^{-1}\mathbf{H})^{-} \mathbf{H}^T \mathbf{C}_{\mathbf{UU}}^{-1} (\sigma_U^2 \mathbf{C}_{\mathbf{UU}}) \mathbf{C}_{\mathbf{UU}}^{-1} \mathbf{H} (\mathbf{H}^T \mathbf{C}_{\mathbf{UU}}^{-1} \mathbf{H})^{-} \\
&= \sigma_U^2 (\mathbf{H}^T \mathbf{C}_{\mathbf{UU}}^{-1} \mathbf{H})^{-}.
\end{aligned}$$

Theorem: (Gauß-Markov)

Given the model

$$\mathbf{X} = \mathbf{H}\boldsymbol{\theta} + \mathbf{U} \quad \text{with} \quad X_n = h_{i1}\theta_1 + \ldots + h_{ip}\theta_p + U_i, \;\; i = 1, \ldots, n,$$

where

$$\mathrm{E}(\mathbf{X}) = \mathbf{H}\boldsymbol{\theta}, \quad \mathbf{Cov}(\mathbf{X}) = \mathbf{Cov}(\mathbf{U}) = \sigma_U^2 \, \mathbf{C}_{\mathbf{UU}}$$

and

$$\mathrm{rank}(\mathbf{H}) = p.$$

Then the best linear unbiased estimator (BLUE) is the least squares estimator

$$\hat{\boldsymbol{\Theta}} = (\mathbf{H}^T \mathbf{C}_{\mathbf{UU}}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{C}_{\mathbf{UU}}^{-1} \mathbf{X}.$$

Hence, the minimum of the weighted sum of squares is

$$q(\hat{\boldsymbol{\theta}}) = \mathrm{tr}(\mathbf{P}^{\perp}\mathbf{y}\mathbf{y}^T) = \mathrm{tr}\left(\left(\mathbf{I}-\mathbf{K}(\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^T\right)\mathbf{y}\mathbf{y}^T\right)$$

$$= \mathrm{tr}\left(\left(\mathbf{I}-\mathbf{C}^{-1}\mathbf{H}\left(\mathbf{H}^T(\mathbf{C}^{-1})^T\mathbf{C}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T(\mathbf{C}^{-1})^T\right)\mathbf{C}^{-1}\mathbf{x}\mathbf{x}^T(\mathbf{C}^{-1})^T\right)$$

$$= \mathrm{tr}\left(\mathbf{C}_{UU}^{-1}\left(\mathbf{I}-\mathbf{H}\left(\mathbf{H}^T\mathbf{C}_{UU}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{C}_{UU}^{-1}\right)\mathbf{x}\mathbf{x}^T\right)$$

$$= \mathrm{tr}\left(\mathbf{C}_{UU}^{-1}\left(\mathbf{I}-\mathbf{P}_{\mathbf{C}_{UU}^{-1}}\right)\mathbf{x}\mathbf{x}^T\right) = \mathrm{tr}\left(\mathbf{C}_{UU}^{-1}\mathbf{P}_{\mathbf{C}_{UU}^{-1}}^{\perp}\mathbf{x}\mathbf{x}^T\right),$$

where

$$\mathbf{P}_{\mathbf{C}_{UU}^{-1}} = \mathbf{H}(\mathbf{H}^T\mathbf{C}_{UU}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}_{UU}^{-1} \quad \textbf{and} \quad \mathbf{P}_{\mathbf{C}_{UU}^{-1}}^{\perp} = \mathbf{I}-\mathbf{P}_{\mathbf{C}_{UU}^{-1}}$$

represent again projection matrices.

Again, we are interested in estimating the variance σ_Z^2 of the noise. The expected value of the minimum of the weighted sum of squares results in

$$\begin{aligned}
\mathrm{E}\left(q(\hat{\boldsymbol{\Theta}})\right) &= \mathrm{E}\left(\mathrm{tr}(\mathbf{C}_{\mathbf{UU}}^{-1}\mathbf{P}_{\mathbf{C}_{\mathbf{UU}}^{-1}}^{\perp}\mathbf{X}\mathbf{X}^T)\right) = \mathrm{tr}\left(\mathbf{C}_{\mathbf{UU}}^{-1}\mathbf{P}_{\mathbf{C}_{\mathbf{UU}}^{-1}}^{\perp}\mathrm{E}\left((\mathbf{H}\boldsymbol{\theta}+\mathbf{U})(\mathbf{H}\boldsymbol{\theta}+\mathbf{U})^T\right)\right) \\
&= \mathrm{tr}\left(\mathbf{C}_{\mathbf{UU}}^{-1}\mathbf{P}_{\mathbf{C}_{\mathbf{UU}}^{-1}}^{\perp}\left(\mathbf{H}\boldsymbol{\theta}\boldsymbol{\theta}^T\mathbf{H}^T+\mathrm{E}(\mathbf{U})\boldsymbol{\theta}^T\mathbf{H}^T+\mathbf{H}\boldsymbol{\theta}\mathrm{E}(\mathbf{U}^T)+\mathrm{E}(\mathbf{U}\mathbf{U}^T)\right)\right) \\
&= \mathrm{tr}\left(\mathbf{C}_{\mathbf{UU}}^{-1}\mathbf{P}_{\mathbf{C}_{\mathbf{UU}}^{-1}}^{\perp}(\mathbf{H}\boldsymbol{\theta}\boldsymbol{\theta}^T\mathbf{H}^T+\sigma_U^2\mathbf{C}_{\mathbf{UU}})\right) = \sigma_U^2\,\mathrm{tr}(\mathbf{P}_{\mathbf{C}_{\mathbf{UU}}^{-1}}^{\perp}) = (n-p)\sigma_U^2,
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{tr}(\mathbf{P}_{\mathbf{C}_{\mathbf{UU}}^{-1}}^{\perp}) &= \mathrm{tr}\left(\mathbf{I}_n - \mathbf{H}(\mathbf{H}^T\mathbf{C}_{\mathbf{UU}}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}_{\mathbf{UU}}^{-1}\right) \\
&= n - \mathrm{tr}\left((\mathbf{H}^T\mathbf{C}_{\mathbf{UU}}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}_{\mathbf{UU}}^{-1}\mathbf{H}\right) = n - \mathrm{tr}(\mathbf{I}_p) = n-p
\end{aligned}$$

has been utilized.

This result motivates

$$S^2 = q(\hat{\boldsymbol{\Theta}}) / (n-p)$$

as an unbiased estimator for σ_Z^2.

Theorem:

If in addition to the Gauß-Markov theorem $\mathbf{U}$ obeys a normal distribution, i.e. $\mathbf{U} \sim \mathcal{N}_n(\mathbf{0}, \sigma_U^2 \mathbf{C_{UU}})$, the following holds.

a) $\hat{\boldsymbol{\Theta}} = (\mathbf{H}^T \mathbf{C}_{\mathbf{UU}}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{C}_{\mathbf{UU}}^{-1} \mathbf{X} \sim \mathcal{N}_p \left(\boldsymbol{\theta}, \sigma_U^2 (\mathbf{H}^T \mathbf{C}_{\mathbf{UU}}^{-1} \mathbf{H})^{-1} \right)$

b) $\hat{\boldsymbol{\Theta}}$ and S^2 are stochastically independent

c) $(n-p) / \sigma_U^2 \cdot S^2 \sim \chi_{n-p}^2$

3.4 Confidence Intervals

Now, an unknown parameter θ is considered, where the density $f_{\mathbf{X}}(\mathbf{x}|\theta)$ of $\mathbf{X}$ and an estimator $\hat{\Theta} = \hat{\theta}(\mathbf{X})$ for θ possessing the density $f_{\hat{\Theta}}(\hat{\theta}\,|\,\theta)$ are given.

With the knowledge of $f_{\hat{\Theta}}(\hat{\theta}\,|\,\theta)$ and a given α, e.g. $\alpha = 0.05$, we can derive from

$$1-\alpha = P\left(\theta - a_1 < \hat{\Theta} \le \theta + a_2\right) = F_{\hat{\Theta}}\left(\theta + a_2\,|\,\theta\right) - F_{\hat{\Theta}}\left(\theta - a_1\,|\,\theta\right)$$

the probability equation

$$1-\alpha = P\left(\theta - a_1 < \hat{\Theta} \le \theta + a_2\right) = P\left(-a_1 < \hat{\Theta} - \theta \le a_2\right)$$

$$= P\left(-\hat{\Theta} - a_1 < -\theta \le a_2 - \hat{\Theta}\right) = P\left(\hat{\Theta} - a_2 \le \theta < \hat{\Theta} + a_1\right)$$

which states that random interval $[\hat{\Theta} - a_2, \hat{\Theta} + a_1)$ covers the parameter θ with probability $1-\alpha$.

For given observations **x** the set

$$C(\mathbf{x}) = \left\{ \theta : \ \hat{\theta}(\mathbf{x}) - a_2 \le \theta < \hat{\theta}(\mathbf{x}) + a_1 \right\}$$

is called

- confidence interval for θ with the confidence coefficient $1-\alpha$

or alternatively

- interval estimate to the confidence level $1-\alpha$.

In case of an unknown parameter vector $\boldsymbol{\theta}$ an interval estimate can be constructed if a function $g(\mathbf{x}\,|\,\boldsymbol{\theta})$ exists such that the corresponding random variable

$$Y = g(\mathbf{x}\,|\,\boldsymbol{\theta})$$

obeys a known distribution which is independent of $\boldsymbol{\theta}$.

Frequently one can find such a function, which depends on $\mathbf{x}$ only over the estimate $\hat{\boldsymbol{\theta}}(\mathbf{x})$, i.e.

$$g(\mathbf{x}\,|\,\boldsymbol{\theta}) = h(\hat{\boldsymbol{\theta}}\,|\,\boldsymbol{\theta}).$$

Examples are given in the subsequent exercises.

Exercise 3.4-1:

Signal + white noise, confidence interval for the

- *signal amplitude estimate (known noise variance)*
- *signal amplitude estimate (unknown noise variance)*
- *noise variance estimate*

Exercise 3.4-2:

Sinusoids + white noise, confidence interval for the amplitude estimates (unknown noise variance)

3.5 Cramer-Rao Lower Bound

In the following only unbiased estimators, i.e.

$$\mathrm{E}(\hat{\boldsymbol{\Theta}}) = \int_{\mathbb{R}^n} \hat{\boldsymbol{\theta}}(\mathbf{x}) f_{\mathbf{x}}(\mathbf{x} \mid \boldsymbol{\theta}) d\mathbf{x} = \boldsymbol{\theta},$$

are considered, whose covariance matrix

$$\mathbf{C}_{\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Theta}}} = \mathrm{Cov}(\hat{\boldsymbol{\Theta}}) = \mathrm{E}\left((\hat{\boldsymbol{\Theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\Theta}} - \boldsymbol{\theta})^T \right)$$

exist. The Cramer-Rao inequality provides now a lower bound to the covariance matrix of any unbiased estimator of $\boldsymbol{\theta}$. Due to the following regularity conditions certain expected values can be determined, which are of importance for deriving the Cramer-Rao inequality.

Regularity conditions:

a) The parameter set Ω is an interval in $\mathbb{R}^p$.

b) The gradient $\nabla_{\boldsymbol{\theta}} \ln\left(f_{\mathbf{x}}(\mathbf{x}\,|\,\boldsymbol{\theta})\right)$ exists for any $\mathbf{x}$ and $\boldsymbol{\theta}$.

c) The gradient of $\int_{\mathbb{R}^n} f_{\mathbf{x}}(\mathbf{x}\,|\,\boldsymbol{\theta})\,d\mathbf{x}$ with respect to $\boldsymbol{\theta}$ can be obtained by taking the gradient under the integral sign, i.e. $\nabla_{\boldsymbol{\theta}} \int_{\mathbb{R}^n} f_{\mathbf{x}}(\mathbf{x}\,|\,\boldsymbol{\theta})\,d\mathbf{x} = \int_{\mathbb{R}^n} \nabla_{\boldsymbol{\theta}}\, f_{\mathbf{x}}(\mathbf{x}\,|\,\boldsymbol{\theta})\,d\mathbf{x}$.

d) The gradient of $\mathrm{E}\left(\hat{\boldsymbol{\theta}}(\mathbf{X})\right)$ with respect to $\boldsymbol{\theta}$ can be obtained by taking the gradient under the integral sign, i.e. $\nabla_{\theta} \int_{\mathbb{R}^n} \hat{\boldsymbol{\theta}}(\mathbf{x}) f_{\mathbf{x}}(\mathbf{x}\,|\,\boldsymbol{\theta})\,d\mathbf{x} = \int_{\mathbb{R}^n} \hat{\boldsymbol{\theta}}(\mathbf{x}) \nabla_{\boldsymbol{\theta}}\, f_{\mathbf{x}}(\mathbf{x}\,|\,\boldsymbol{\theta})\,d\mathbf{x}$.

Corollary:

If the former regularity conditions hold, then

1) $\mathrm{E}\left(\nabla_{\boldsymbol{\theta}} \ln\left(f_{\mathbf{X}}(\mathbf{X}\,|\,\boldsymbol{\theta})\right)\right) = \mathbf{0}$

2) $\mathrm{E}\left(\left(\nabla_{\boldsymbol{\theta}} \ln\left(f_{\mathbf{X}}(\mathbf{X}\,|\,\boldsymbol{\theta})\right)\right)(\hat{\boldsymbol{\Theta}} - \boldsymbol{\theta})^{T}\right) = \mathbf{I}$

Proof of 1):

$$\mathrm{E}\left(\nabla_{\boldsymbol{\theta}} \ln\left(f_{\mathbf{X}}(\mathbf{X}\,|\,\boldsymbol{\theta})\right)\right) = \int_{\mathbb{R}^n} \nabla_{\boldsymbol{\theta}} \ln\left(f_{\mathbf{X}}(\mathbf{x}\,|\,\boldsymbol{\theta})\right) f_{\mathbf{X}}(\mathbf{x}\,|\,\boldsymbol{\theta})\, d\mathbf{x} =$$

$$= \int_{\mathbb{R}^n} \frac{\nabla_{\boldsymbol{\theta}} f_{\mathbf{X}}(\mathbf{x}\,|\,\boldsymbol{\theta})}{f_{\mathbf{X}}(\mathbf{x}\,|\,\boldsymbol{\theta})} f_{\mathbf{X}}(\mathbf{x}\,|\,\boldsymbol{\theta})\, d\mathbf{x} = \int_{\mathbb{R}^n} \nabla_{\boldsymbol{\theta}} f_{\mathbf{X}}(\mathbf{x}\,|\,\boldsymbol{\theta})\, d\mathbf{x}$$

$$= \nabla_{\boldsymbol{\theta}} \int_{\mathbb{R}^n} f_{\mathbf{X}}(\mathbf{x}\,|\,\boldsymbol{\theta})\, d\mathbf{x} = \nabla_{\boldsymbol{\theta}} 1 = \mathbf{0}$$

Proof of 2):

$$\mathrm{E}\left(\left(\nabla_{\boldsymbol{\theta}} \ln\left(f_{\mathbf{X}}(\mathbf{X}\,|\,\boldsymbol{\theta})\right)\right)\left(\hat{\boldsymbol{\Theta}}-\boldsymbol{\theta}\right)^{T}\right) =$$

$$= \int_{\mathbb{R}^n} \nabla_{\boldsymbol{\theta}} \ln\left(f_{\mathbf{X}}(\mathbf{x}\,|\,\boldsymbol{\theta})\right)\left(\hat{\boldsymbol{\theta}}(\mathbf{x})-\boldsymbol{\theta}\right)^{T} f_{\mathbf{X}}(\mathbf{x}\,|\,\boldsymbol{\theta})\,d\mathbf{x}$$

$$= \int_{\mathbb{R}^n} \frac{\nabla_{\boldsymbol{\theta}} f_{\mathbf{X}}(\mathbf{x}\,|\,\boldsymbol{\theta})}{f_{\mathbf{X}}(\mathbf{x}\,|\,\boldsymbol{\theta})}\left(\hat{\boldsymbol{\theta}}(\mathbf{x})-\boldsymbol{\theta}\right)^{T} f_{\mathbf{X}}(\mathbf{x}\,|\,\boldsymbol{\theta})\,d\mathbf{x}$$

$$= \int_{\mathbb{R}^n} \nabla_{\boldsymbol{\theta}} f_{\mathbf{X}}(\mathbf{x}\,|\,\boldsymbol{\theta})\left(\hat{\boldsymbol{\theta}}(\mathbf{x})-\boldsymbol{\theta}\right)^{T} d\mathbf{x}$$

$$= \nabla_{\boldsymbol{\theta}} \int_{\mathbb{R}^n} f_{\mathbf{X}}(\mathbf{x}\,|\,\boldsymbol{\theta})\,\hat{\boldsymbol{\theta}}(\mathbf{x})^{T} d\mathbf{x} - \nabla_{\boldsymbol{\theta}} \int_{\mathbb{R}^n} f_{\mathbf{X}}(\mathbf{x}\,|\,\boldsymbol{\theta})\,d\mathbf{x}\;\boldsymbol{\theta}^{T}$$

$$= \nabla_{\boldsymbol{\theta}}\,\boldsymbol{\theta}^{T} - \left(\nabla_{\boldsymbol{\theta}}\,1\right)\boldsymbol{\theta}^{T} = \mathbf{I}$$

Definition: *(Fisher information matrix)*

The $p \times p$ Fisher information matrix is defined by

$$\mathcal{I}(\boldsymbol{\theta}) = \mathrm{E}\left(\nabla_{\boldsymbol{\theta}} \ln\left(f_{\mathbf{X}}(\mathbf{X} \mid \boldsymbol{\theta}) \right) \nabla_{\boldsymbol{\theta}}^{T} \ln\left(f_{\mathbf{X}}(\mathbf{X} \mid \boldsymbol{\theta}) \right) \right) =$$

$$= \int_{\mathbb{R}^n} \nabla_{\boldsymbol{\theta}} \ln\left(f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta}) \right) \nabla_{\boldsymbol{\theta}}^{T} \ln\left(f_{\mathbf{X}}(\mathbf{X} \mid \boldsymbol{\theta}) \right) f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta}) d\mathbf{x}.$$

Hence, the entries of the Fisher information matrix are for $i, j = 1, \ldots, p$ given by

$$\mathcal{I}_{ij}(\boldsymbol{\theta}) = \mathrm{E}\left(\frac{\partial \ln\left(f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta}) \right)}{\partial \theta_i} \cdot \frac{\partial \ln\left(f_{\mathbf{X}}(\mathbf{X} \mid \boldsymbol{\theta}) \right)}{\partial \theta_j} \right)$$

$$= \int_{\mathbb{R}^n} \frac{\partial \ln\left(f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta}) \right)}{\partial \theta_i} \cdot \frac{\partial \ln\left(f_{\mathbf{X}}(\mathbf{X} \mid \boldsymbol{\theta}) \right)}{\partial \theta_j} f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta}) d\mathbf{x}.$$

Theorem: (Cramer-Rao inequality)

Suppose the former assumptions and regularity conditions hold and the Fisher information matrix $\mathcal{I}(\boldsymbol{\theta})$ is positive definite. Then the variance of any estimator $\hat{\Theta}_{\mathbf{a}} = \mathbf{a}^T \hat{\boldsymbol{\Theta}}$ is bounded downwards by

$$\mathrm{Var}(\hat{\Theta}_{\mathbf{a}}) = \mathbf{a}^T \mathbf{C}_{\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Theta}}}\, \mathbf{a} \geq \mathbf{a}^T \mathcal{I}(\boldsymbol{\theta})^{-1} \mathbf{a} \quad \forall \mathbf{a} \in \mathbb{R}^p.$$

If particularly $\mathbf{a} = \mathbf{e}_i$, i.e. the i-th Cartesian unit vector, the inequality provides the lower limit for the variance of the i-th component of $\hat{\boldsymbol{\Theta}}$.

$$\mathrm{Var}(\hat{\Theta}_i) = \mathbf{e}_i^T \mathbf{C}_{\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Theta}}}\, \mathbf{e}_i \geq \mathbf{e}_i^T \mathcal{I}(\boldsymbol{\theta})^{-1} \mathbf{e}_i = \mathcal{I}(\boldsymbol{\theta})^{-1}_{ii}$$

$$i = 1, \ldots, p$$

Proof:

First, the random vector

$$\mathbf{Y} = \hat{\boldsymbol{\Theta}} - \boldsymbol{\theta} - \mathcal{I}(\boldsymbol{\theta})^{-1} \nabla_{\boldsymbol{\theta}} \ln\left(f_{\mathbf{X}}(\mathbf{X} \mid \boldsymbol{\theta})\right)$$

is introduced and its second order moment matrix

$$\begin{aligned}
\mathrm{E}(\mathbf{Y}\mathbf{Y}^T) = \mathrm{E}\left((\hat{\boldsymbol{\Theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\Theta}} - \boldsymbol{\theta})^T\right) \\
- \mathrm{E}\left((\hat{\boldsymbol{\Theta}} - \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^T \ln\left(f_{\mathbf{X}}(\mathbf{X} \mid \boldsymbol{\theta})\right)\right) \mathcal{I}(\boldsymbol{\theta})^{-1} \\
- \mathcal{I}(\boldsymbol{\theta})^{-1} \mathrm{E}\left(\nabla_{\boldsymbol{\theta}} \ln\left(f_{\mathbf{X}}(\mathbf{X} \mid \boldsymbol{\theta})\right)(\hat{\boldsymbol{\Theta}} - \boldsymbol{\theta})^T\right) \\
+ \mathcal{I}(\boldsymbol{\theta})^{-1} \mathrm{E}\left(\nabla_{\boldsymbol{\theta}} \ln\left(f_{\mathbf{X}}(\mathbf{X} \mid \boldsymbol{\theta})\right) \nabla_{\boldsymbol{\theta}}^T \ln\left(f_{\mathbf{X}}(\mathbf{X} \mid \boldsymbol{\theta})\right)\right) \mathcal{I}(\boldsymbol{\theta})^{-1}.
\end{aligned}$$

is determined. After exploiting the former corollary the

second order moment matrix can be simplified to

$$\mathrm{E}(\mathbf{Y}\mathbf{Y}^T) = \mathrm{E}\left((\hat{\boldsymbol{\Theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\Theta}} - \boldsymbol{\theta})^T\right) - \mathcal{I}(\boldsymbol{\theta})^{-1}.$$

Since

$$\mathbf{a}^T\mathrm{E}(\mathbf{Y}\mathbf{Y}^T)\mathbf{a} = \mathrm{E}(\mathbf{a}^T\mathbf{Y}\mathbf{Y}^T\mathbf{a}) = \mathrm{E}\left((\mathbf{a}^T\mathbf{Y})^2\right) \geq 0 \quad \forall \mathbf{a} \in \mathbb{R}^p$$

the second order moment matrix is positive semidefinite. Thus, we can write

$$\mathbf{a}^T\mathrm{E}(\mathbf{Y}\mathbf{Y}^T)\mathbf{a} = \mathbf{a}^T\mathbf{C}_{\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Theta}}}\mathbf{a} - \mathbf{a}^T\mathcal{I}(\boldsymbol{\theta})^{-1}\mathbf{a} \geq 0 \quad \forall \mathbf{a} \in \mathbb{R}^p$$

and the assertion

$$\mathbf{a}^T\mathbf{C}_{\hat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Theta}}}\mathbf{a} \geq \mathbf{a}^T\mathcal{I}(\boldsymbol{\theta})^{-1}\mathbf{a} \quad \forall \mathbf{a} \in \mathbb{R}^p$$

of the theorem follows.

Under the following regularity condition an alternative and often more convenient expression for calculating the information matrix can be presented.

Regularity condition:

e) The Hessian matrix of

$$\int_{\mathbb{R}^n} f_{\mathbf{x}}(\mathbf{x}\,|\,\boldsymbol{\theta})\,d\mathbf{x}$$

with respect to $\boldsymbol{\theta}$ can be obtained by taking all the required second partial derivatives under the integral sign, i.e.

$$\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}}^{T} \int_{\mathbb{R}^n} f_{\mathbf{x}}(\mathbf{x}\,|\,\boldsymbol{\theta})\,d\mathbf{x} = \int_{\mathbb{R}^n} \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}}^{T} f_{\mathbf{x}}(\mathbf{x}\,|\,\boldsymbol{\theta})\,d\mathbf{x}.$$

Corollary:

If the former regularity condition holds, the Fisher information matrix can be alternatively determined by

$$\mathcal{I}(\boldsymbol{\theta}) = \mathrm{E}\left(\nabla_{\boldsymbol{\theta}} \ln\left(f_{\mathbf{X}}(\mathbf{X}\,|\,\boldsymbol{\theta})\right) \nabla_{\boldsymbol{\theta}}^{T} \ln\left(f_{\mathbf{X}}(\mathbf{X}\,|\,\boldsymbol{\theta})\right)\right)$$

$$= -\mathrm{E}\left(\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}}^{T} \ln\left(f_{\mathbf{X}}(\mathbf{X}\,|\,\boldsymbol{\theta})\right)\right).$$

Consequently, the elements of the Fisher information matrix can be expressed by

$$\mathcal{I}_{ij}(\boldsymbol{\theta}) = \mathrm{E}\left(\frac{\partial \ln\left(f_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta})\right)}{\partial \theta_i} \cdot \frac{\partial \ln\left(f_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta})\right)}{\partial \theta_j}\right) = -\mathrm{E}\left(\frac{\partial^2 \ln\left(f_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta})\right)}{\partial \theta_i \partial \theta_j}\right)$$

$$i, j = 1, \ldots, p.$$

Proof:

$$\mathrm{E}\left(\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}^{T}\ln\left(f_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta})\right)\right)=\mathrm{E}\left(\nabla_{\boldsymbol{\theta}}\frac{\nabla_{\boldsymbol{\theta}}^{T}f_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta})}{f_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta})}\right)=$$

$$=\mathrm{E}\left(\frac{\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}^{T}f_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta})\cdot f_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta})-\nabla_{\boldsymbol{\theta}}f_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta})\cdot\nabla_{\boldsymbol{\theta}}^{T}f_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta})}{f_{\mathbf{X}}^{2}(\mathbf{X}\,|\,\boldsymbol{\theta})}\right)$$

$$=\mathrm{E}\left(\frac{\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}^{T}f_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta})}{f_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta})}\right)-\mathrm{E}\left(\frac{\nabla_{\boldsymbol{\theta}}f_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta})\cdot\nabla_{\boldsymbol{\theta}}^{T}f_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta})}{f_{\mathbf{X}}^{2}(\mathbf{X}|\boldsymbol{\theta})}\right)$$

$$=\int_{\mathbb{R}^n}\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}^{T}f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})\,d\mathbf{x}-\mathrm{E}\left(\nabla_{\boldsymbol{\theta}}\ln\left(f_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta})\right)\nabla_{\boldsymbol{\theta}}^{T}\ln\left(f_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta})\right)\right)$$

$$=\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}^{T}\int_{\mathbb{R}^n}f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})\,d\mathbf{x}-\mathcal{I}(\boldsymbol{\theta})=\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}^{T}1-\mathcal{I}(\boldsymbol{\theta})=-\mathcal{I}(\boldsymbol{\theta})$$

Theorem:

Let $f_{\mathbf{x}}(\mathbf{x} \mid \boldsymbol{\xi})$ belong to the exponential family in canonical form and $\mathbf{T}$ denote the corresponding sufficient statistic.

1) The regularity conditions a) – e) are satisfied and the information matrix is positive definite

$$\mathcal{I}(\boldsymbol{\xi}) = \mathrm{E}\left(\nabla_{\boldsymbol{\xi}} \ln\left(f_{\mathbf{x}}(\mathbf{X} \mid \boldsymbol{\xi})\right) \nabla_{\boldsymbol{\xi}}^{T} \ln\left(f_{\mathbf{x}}(\mathbf{X} \mid \boldsymbol{\xi})\right)\right)$$

$$= -\mathrm{E}\left(\nabla_{\boldsymbol{\xi}} \nabla_{\boldsymbol{\xi}}^{T} \ln\left(f_{\mathbf{x}}(\mathbf{X} \mid \boldsymbol{\xi})\right)\right)$$

$$= \nabla_{\boldsymbol{\xi}} \nabla_{\boldsymbol{\xi}}^{T} A(\boldsymbol{\xi}) = \mathrm{Cov}(\mathbf{T}) = \mathbf{C}_{\mathbf{TT}}.$$

Thus, for unbiased estimators the Cramer-Rao inequality holds.

2) Let the vector valued function

$$\boldsymbol{\xi} = \mathbf{g}(\boldsymbol{\theta}),\ \boldsymbol{\xi} \in \mathbb{R}^k,\ \boldsymbol{\theta} \in \mathbb{R}^p \quad \text{with} \quad p \le k$$

be one-to-one and differentiable with

$$\mathbf{G}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\, \mathbf{g}(\boldsymbol{\theta})^T = \left(\frac{\partial g_j(\boldsymbol{\theta})}{\partial \theta_i} \right)_{i=1,\ldots,p;\, j=1,\ldots,k}.$$

Then the regularity conditions a) - e) are satisfied with respect to $\boldsymbol{\theta}$ and the information matrix can be expressed by

$$\mathcal{I}(\boldsymbol{\theta}) = -\mathrm{E}\left(\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}}^T \ln\left(f_{\mathbf{X}}(\mathbf{X} \mid \boldsymbol{\theta}) \right) \right) = \mathbf{G}(\boldsymbol{\theta})\, \mathcal{I}(\boldsymbol{\xi})\big|_{\boldsymbol{\xi}=\mathbf{g}(\boldsymbol{\theta})}\, \mathbf{G}(\boldsymbol{\theta})^T$$

$$= \mathbf{G}(\boldsymbol{\theta}) \left(\nabla_{\boldsymbol{\xi}} \nabla_{\boldsymbol{\xi}}^T A(\boldsymbol{\xi}) \right)\Big|_{\boldsymbol{\xi}=\mathbf{g}(\boldsymbol{\theta})}\, \mathbf{G}(\boldsymbol{\theta})^T.$$

Now, one asks oneself immediately, does there exist an unbiased estimator for which the information inequality takes the equality sign, i.e. whose covariance matrix is equal to the inverse Fisher information matrix.

An estimator for which the information inequality takes the equality sign, i.e. whose covariance matrix coincides with the inverse Fisher information matrix (Cramer-Rao lower bound), is called efficient.

Efficient estimators can be achieved only in special cases, since the random vector $\mathbf{Y}$, introduced for proving the information inequality, must be the zero-vector with probability 1.

Exercise 3.5-1:

Non/efficiency of linear least squares estimators when the noise is i.i.d. and normally distributed

The previous exercise showed, that the deviation of the covariance matrix from the inverse Fisher information matrix can be neglected for large sample sizes.

This motivates the definition of the limit

$$\boldsymbol{\Gamma}(\boldsymbol{\theta}) = \lim_{n\to\infty} \frac{1}{n} \mathcal{I}_n(\boldsymbol{\theta}),$$

where n indicates the growing sample size. If the matrix $\boldsymbol{\Gamma}(\boldsymbol{\theta})$ exists and is not singular, then unbiased and consistent estimators $\hat{\boldsymbol{\Theta}}_n$ are of interest that satisfy the limit

$$\lim_{n\to\infty} n\,\mathbf{C}_{\hat{\boldsymbol{\Theta}}_n\hat{\boldsymbol{\Theta}}_n} = \boldsymbol{\Gamma}(\boldsymbol{\theta})^{-1}.$$

Simultaneously, such estimators are then even consistent in the mean square sense.

Estimators used in practice are often neither unbiased nor mean square consistent. In these cases the limiting distributions of such estimators are considered. If the limiting distribution satisfies the property

$$\lim_{n\to\infty} \sqrt{n}\,(\hat{\boldsymbol{\Theta}}_n - \boldsymbol{\theta}) \sim \mathcal{N}_p\left(\mathbf{0}, \mathbf{D}(\boldsymbol{\theta})\right) \quad \text{and} \quad \mathbf{D}(\boldsymbol{\theta}) = \boldsymbol{\Gamma}(\boldsymbol{\theta})^{-1},$$

the estimator is said to be asymptotically efficient.

Furthermore, under certain regularity conditions (similar to the former) one can show that in the class of asymptotically normally distributed estimators the following lower bound can be proved.

$$\mathbf{a}^T \mathbf{D}(\boldsymbol{\theta})\,\mathbf{a} \geq \mathbf{a}^T \boldsymbol{\Gamma}(\boldsymbol{\theta})^{-1}\mathbf{a} \quad \mathbf{a} \in \mathbb{R}^p$$

Exercise 3.5-2:

Asymptotic efficiency of linear least squares estimators when the noise is i.i.d. and normally distributed

3.6 Maximum Likelihood Estimation

The maximum Likelihood method is one of the most important procedures for the construction of estimators.

For given observations $\mathbf{x}$ the procedure selects that $\boldsymbol{\theta}$ as maximum likelihood estimate $\hat{\boldsymbol{\theta}}(\mathbf{x})$, for which the density function $f_{\mathbf{x}}(\mathbf{x}\,|\,\boldsymbol{\theta})$ takes its maximum.

Thus, the maximum likelihood estimate can be interpreted heuristically as that parameter vector that makes the occurrence of the data observed most likely.

Instead of determining the maximum of $f_{\mathbf{x}}(\mathbf{x}\,|\,\boldsymbol{\theta})$, one can due to the monotonicity of the logarithmic function alternatively maximize $\ln\left(f_{\mathbf{x}}(\mathbf{x}\,|\,\boldsymbol{\theta})\right)$.

Definition:

Suppose $\mathbf{X}$ possesses the density $f_{\mathbf{x}}(\mathbf{x}\,|\,\boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \Omega$ and the vector of observations $\mathbf{x} = (x_1, \ldots, x_n)^T$ is given. Then $f_{\mathbf{x}}(\mathbf{x}\,|\,\boldsymbol{\theta})$ and $\ln\left(f_{\mathbf{x}}(\mathbf{x}\,|\,\boldsymbol{\theta})\right)$ are called likelihood function and log-likelihood function respectively, and

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Omega}{\operatorname{argmax}}\, f_{\mathbf{x}}(\mathbf{x}\,|\,\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}\in\Omega}{\operatorname{argmax}} \ln\left(f_{\mathbf{x}}(\mathbf{x}\,|\,\boldsymbol{\theta})\right)$$

is called the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$.

If the gradient of $f_{\mathbf{x}}(\mathbf{x}\,|\,\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ exists and if $f_{\mathbf{x}}(\mathbf{x}\,|\,\boldsymbol{\theta})$ is positive for the given $\mathbf{x}$, one can try to find the MLE by solving the likelihood equation system

$$\nabla_{\theta} \ln\left(f_{\mathbf{x}}(\mathbf{x}\,|\,\boldsymbol{\theta})\right)\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\mathbf{x})} = \mathbf{0}.$$

Remarks:

- Generally the likelihood equation system is nonlinear.
- The likelihood equation system can possess several solutions.
- The solutions of the likelihood equation system do not necessarily correspond to relative maxima.
- If the absolute maximum of the likelihood function lies on the boundary of Ω it can not be described by a solution of the likelihood equation system.
- Consequently, one has generally to employ sophisticated numerical optimization techniques for determining maximum likelihood estimates.

Properties of Maximum Likelihood Estimation:

1) Let $\mathbf{g}: \mathbb{R}^q \to \mathbb{R}^p,\ p \le q$ be a continuous mapping of the parameter vector $\boldsymbol{\eta}$ and $\hat{\boldsymbol{\eta}}$ the MLE of $\boldsymbol{\eta}$. Then the MLE of $\boldsymbol{\theta} = \mathbf{g}(\boldsymbol{\eta})$ is given by $\hat{\boldsymbol{\theta}} = \mathbf{g}(\hat{\boldsymbol{\eta}})$.

2) Suppose $\mathbf{t}(\mathbf{x})$ is a sufficient transformation for $\boldsymbol{\theta}$, i.e.

$$f_{\mathbf{x}}(\mathbf{x} \mid \boldsymbol{\theta}) = g\left(\mathbf{t}(\mathbf{x}) \mid \boldsymbol{\theta}\right) \cdot h(\mathbf{x}).$$

Then the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ only depends over $\mathbf{t}(\mathbf{x})$ on $\mathbf{x}$.

3) If $f_{\mathbf{x}}(\mathbf{x} \mid \boldsymbol{\xi})$ belongs to the exponential family in canonical form the related likelihood equation system

$$\nabla_{\boldsymbol{\xi}} \ln\left(f_{\mathbf{x}}(\mathbf{x} \mid \boldsymbol{\xi})\right) = \mathbf{t}(\mathbf{x}) - \nabla_{\boldsymbol{\xi}} A(\boldsymbol{\xi}) = \mathbf{0}$$

possesses a unique solution.

Exercise 3.6-1:
Maximum likelihood estimation for linear models when the noise is i.i.d. and normally distributed

Exercise 3.6-2:
Sinusoids + white noise, maximum likelihood estimation of amplitude, phase and frequency when the noise is i.i.d. and normally distributed

Exercise 3.6-3:
Example for inconsistent maximum likelihood estimation

3.7 Bayesian Estimation

In the Bayesian theory of parameter estimation the unknown parameter vector $\boldsymbol{\theta}$ is treated itself as a realization of a random experiment that obeys its own so called prior distribution $f_{\Theta}(\boldsymbol{\theta})$.

The objective is to use $f_{\Theta}(\boldsymbol{\theta})$ together with the measurements $\mathbf{x}$ drawn from the distribution $f_{\mathbf{X}}(\mathbf{x}\,|\,\boldsymbol{\theta})$ to turn the prior distribution $f_{\Theta}(\boldsymbol{\theta})$ into a posterior distribution

$$f_{\Theta}(\boldsymbol{\theta}\,|\,\mathbf{x}) = \frac{f_{\mathbf{X}\Theta}(\mathbf{x},\boldsymbol{\theta})}{f_{\mathbf{X}}(\mathbf{x})} = \frac{f_{\mathbf{X}}(\mathbf{x}\,|\,\boldsymbol{\theta})f_{\Theta}(\boldsymbol{\theta})}{\int_{\mathbb{R}^p} f_{\mathbf{X}\Theta}(\mathbf{x},\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{f_{\mathbf{X}}(\mathbf{x}\,|\,\boldsymbol{\theta})f_{\Theta}(\boldsymbol{\theta})}{\int_{\mathbb{R}^p} f_{\mathbf{X}}(\mathbf{x}\,|\,\boldsymbol{\theta})f_{\Theta}(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

that is used for the statistical inference.

Loss function

The quality of the estimate $\hat{\boldsymbol{\theta}}(\mathbf{x})$ is measured by a real-valued loss function $L(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}(\mathbf{x}))$, e.g. by the Euclidean distance between $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}(\mathbf{x})$, i.e.

$$L(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}(\mathbf{x})) = \Big(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\mathbf{x})\Big)^T \Big(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\mathbf{x})\Big).$$

Risk or loss average

The risk is the loss $L(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}(\mathbf{x}))$ averaged over the distribution of the measurements for any fixed parameter vector $\boldsymbol{\theta} \in \Omega$, i.e.

$$R_L(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta}) = \mathrm{E}_{\mathbf{X}|\boldsymbol{\theta}}\Big(L(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}(\mathbf{X}))\Big) = \int_{\mathbb{R}^n} L(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}(\mathbf{x})) f_{\mathbf{x}}(\mathbf{x} \mid \boldsymbol{\theta})\, d\mathbf{x}.$$

Bayes risk

The Bayes risk is the risk $R_L(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\Theta})$ averaged over the prior distribution $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$, i.e.

$$R_L(\hat{\boldsymbol{\theta}}) = \mathrm{E}_{\boldsymbol{\Theta}}\left(R_L(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\Theta})\right) = \mathrm{E}_{\boldsymbol{\Theta}}\left(\mathrm{E}_{\mathbf{X}|\boldsymbol{\theta}}\left(L(\boldsymbol{\Theta}, \hat{\boldsymbol{\theta}}(\mathbf{X}))\right)\right)$$

$$= \int_{\mathbb{R}^p} \int_{\mathbb{R}^n} L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\mathbf{x})) f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) d\mathbf{x} d\boldsymbol{\theta}$$

$$= \int_{\mathbb{R}^{n+p}} L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\mathbf{x})) f_{\mathbf{X}\boldsymbol{\Theta}}(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x} d\boldsymbol{\theta} = \mathrm{E}_{\mathbf{X}\boldsymbol{\Theta}}\left(L(\boldsymbol{\Theta}, \hat{\boldsymbol{\theta}}(\mathbf{X}))\right)$$

The Bayes risk measures the average risk one incurs if the estimator $\hat{\boldsymbol{\Theta}} = \hat{\boldsymbol{\theta}}(\mathbf{X})$ is used for the random experiment at hand. Hence, $R_L(\hat{\boldsymbol{\theta}})$ is depending only on the estimating function $\hat{\boldsymbol{\theta}}(\mathbf{x})$.

Bayes estimator

Minimization of the Bayes risk $R_L(\hat{\boldsymbol{\theta}})$ over the class of all estimating functions $\hat{\boldsymbol{\theta}}(\mathbf{x})$ for which $R_L(\hat{\boldsymbol{\theta}})$ exists, provides the Bayes estimating function

$$\hat{\boldsymbol{\theta}}_L(\mathbf{x}) = \underset{\hat{\boldsymbol{\theta}}(\mathbf{x})}{\arg\min}\, R_L(\hat{\boldsymbol{\theta}}) = \underset{\hat{\boldsymbol{\theta}}(\mathbf{x})}{\arg\min}\, \mathrm{E}_{\mathbf{X},\boldsymbol{\Theta}}\left(L(\boldsymbol{\Theta},\hat{\boldsymbol{\theta}}(\mathbf{X}))\right)$$

and therefore the Bayes estimator $\hat{\boldsymbol{\Theta}}_L = \hat{\boldsymbol{\theta}}_L(\mathbf{X})$.

In the following Bayes estimators are considered in more detail for the square error, the absolute error and uniform error loss functions.

3.7.1 Minimum Mean Square Error Estimation

Non-Linear Minimum Mean Square Error Estimation

To estimate the random parameter vector $\boldsymbol{\Theta}$ by means of the realization of the random vector $\mathbf{X}$ in the minimum mean square error sense we have to find an estimating function $\hat{\boldsymbol{\theta}}(\mathbf{x})$ such that

$$R_{MSE}(\hat{\boldsymbol{\theta}}) = \mathrm{E}\left\{\left(\boldsymbol{\Theta}-\hat{\boldsymbol{\theta}}(\mathbf{X})\right)^T\left(\boldsymbol{\Theta}-\hat{\boldsymbol{\theta}}(\mathbf{X})\right)\right\}$$

$$= \int_{\mathbb{R}^{n+p}}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}(\mathbf{x})\right)^T\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}(\mathbf{x})\right) f_{\mathbf{X}\boldsymbol{\Theta}}(\mathbf{x},\boldsymbol{\theta})\, d\mathbf{x}\, d\boldsymbol{\theta}$$

is minimum. Over the class of all functions $\hat{\boldsymbol{\theta}}(\mathbf{x})$ for which the expected value exists, $R_{MSE}(\hat{\boldsymbol{\theta}})$ is minimized by

$$\hat{\boldsymbol{\theta}}_{MMSE}(\mathbf{x}) = \mathrm{E}\left(\boldsymbol{\theta} \mid \mathbf{X} = \mathbf{x}\right),$$

cf. Exercise 1.9-18.

Linear Minimum Mean Square Error Estimation

Now, we wish to estimate the random parameter vector $\boldsymbol{\Theta}$ by the linear estimating function

$$\hat{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$$

such that the mean square error

$$R_{LMSE}(\mathbf{A},\mathbf{b}) = \mathrm{E}\left(\left(\boldsymbol{\Theta} - (\mathbf{A}\mathbf{X} + \mathbf{b})\right)^T \left(\boldsymbol{\Theta} - (\mathbf{A}\mathbf{X} + \mathbf{b})\right)\right)$$

$$= \int_{\mathbb{R}^{n+p}} \left(\boldsymbol{\theta} - (\mathbf{A}\mathbf{x} + \mathbf{b})\right)^T \left(\boldsymbol{\theta} - (\mathbf{A}\mathbf{x} + \mathbf{b})\right) f_{\mathbf{X}\boldsymbol{\Theta}}(\mathbf{x},\boldsymbol{\theta})\, d\mathbf{x}\, d\boldsymbol{\theta}$$

is minimized by varying the parameter matrix $\mathbf{A}$ and vector $\mathbf{b}$. Thus, we have to solve the minimization problem

$$(\hat{\mathbf{A}},\hat{\mathbf{b}}) = \underset{\mathbf{A},\mathbf{b}}{\arg\min}\left(R_{LMSE}(\mathbf{A},\mathbf{b})\right).$$

The mean square error is minimum if

$$\frac{\partial}{\partial \mathbf{A}} R_{LMSE}(\mathbf{A},\mathbf{b}) = -2\mathrm{E}\left(\left(\boldsymbol{\Theta} - (\hat{\mathbf{A}}\mathbf{X} + \hat{\mathbf{b}})\right)\mathbf{X}^T\right) = \mathbf{0}$$

$$\nabla_{\mathbf{b}} R_{LMSE}(\mathbf{A},\mathbf{b}) = -2\mathrm{E}\left(\boldsymbol{\Theta} - (\hat{\mathbf{A}}\mathbf{X} + \hat{\mathbf{b}})\right) = \mathbf{0}$$

are satisfied. Applying the expectation operator we obtain

$$\mathbf{R}_{\Theta\mathbf{X}} - \hat{\mathbf{A}}\mathbf{R}_{\mathbf{X}\mathbf{X}} - \hat{\mathbf{b}}\boldsymbol{\mu}_{\mathbf{X}}^T = \mathbf{0} \quad \text{and} \quad \boldsymbol{\mu}_{\Theta} - \hat{\mathbf{A}}\boldsymbol{\mu}_{\mathbf{X}} - \hat{\mathbf{b}} = \mathbf{0},$$

where

$$\mathbf{R}_{\mathbf{X}\mathbf{X}} = \mathrm{E}(\mathbf{X}\mathbf{X}^T),\ \mathbf{R}_{\Theta\mathbf{X}} = \mathrm{E}(\boldsymbol{\Theta}\mathbf{X}^T),\ \boldsymbol{\mu}_{\mathbf{X}} = \mathrm{E}(\mathbf{X}) \text{ and } \boldsymbol{\mu}_{\Theta} = \mathrm{E}(\boldsymbol{\Theta}).$$

Resolving for $\hat{\mathbf{A}}$ and $\hat{\mathbf{b}}$ leads after some manipulations to

$$\hat{\mathbf{A}} = \left(\mathbf{R}_{\Theta\mathbf{x}} - \boldsymbol{\mu}_{\Theta}\boldsymbol{\mu}_{\mathbf{x}}^{T}\right)\left(\mathbf{R}_{\mathbf{xx}} - \boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\mu}_{\mathbf{x}}^{T}\right)^{-1}$$

$$\hat{\mathbf{b}} = \boldsymbol{\mu}_{\Theta} - \left(\mathbf{R}_{\Theta\mathbf{x}} - \boldsymbol{\mu}_{\Theta}\boldsymbol{\mu}_{\mathbf{x}}^{T}\right)\left(\mathbf{R}_{\mathbf{xx}} - \boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\mu}_{\mathbf{x}}^{T}\right)^{-1}\boldsymbol{\mu}_{\mathbf{x}}$$

and therefore to the estimating function

$$\hat{\boldsymbol{\theta}}_{LMMSE}(\mathbf{x}) = \hat{\mathbf{A}}\mathbf{x} + \hat{\mathbf{b}} = \boldsymbol{\mu}_{\Theta} + \left(\mathbf{R}_{\Theta\mathbf{x}} - \boldsymbol{\mu}_{\Theta}\boldsymbol{\mu}_{\mathbf{x}}^{T}\right)\left(\mathbf{R}_{\mathbf{xx}} - \boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\mu}_{\mathbf{x}}^{T}\right)^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}\right).$$

After exploiting the relationships between the covariance and the first and second order moments

$$\mathbf{C}_{\mathbf{xx}} = \mathbf{R}_{\mathbf{xx}} - \boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\mu}_{\mathbf{x}}^{T} \quad \text{and} \quad \mathbf{C}_{\Theta\mathbf{x}} = \mathbf{R}_{\Theta\mathbf{x}} - \boldsymbol{\mu}_{\Theta}\boldsymbol{\mu}_{\mathbf{x}}^{T}$$

the estimating function can also be expressed by

$$\hat{\boldsymbol{\theta}}_{LMMSE}(\mathbf{x}) = \boldsymbol{\mu}_{\Theta} + \mathbf{C}_{\Theta\mathbf{x}}\mathbf{C}_{\mathbf{xx}}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}\right).$$

Theorem:

Let $\mathbf{Y}$ be a random vector composed of the random vectors $\mathbf{X}$ and $\boldsymbol{\Theta}$ and obeying the distribution

$$\mathbf{Y} = \begin{pmatrix} \mathbf{X} \\ \boldsymbol{\Theta} \end{pmatrix} \sim \mathcal{N}_{n+p}\left(\begin{pmatrix} \boldsymbol{\mu}_{\mathbf{X}} \\ \boldsymbol{\mu}_{\boldsymbol{\Theta}} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{XX}} & \boldsymbol{\Sigma}_{\mathbf{X}\boldsymbol{\Theta}} \\ \boldsymbol{\Sigma}_{\boldsymbol{\Theta}\mathbf{X}} & \boldsymbol{\Sigma}_{\boldsymbol{\Theta}\boldsymbol{\Theta}} \end{pmatrix} \right).$$

Then $\mathbf{X}$ and $\boldsymbol{\Theta}$ possessing the marginal distributions

$$\mathbf{X} \sim \mathcal{N}_n\left(\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}_{\mathbf{XX}}\right), \quad \boldsymbol{\Theta} \sim \mathcal{N}_p\left(\boldsymbol{\mu}_{\boldsymbol{\Theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\Theta}\boldsymbol{\Theta}}\right)$$

and the conditional distribution

$$\boldsymbol{\Theta} \,|\, \mathbf{X} = \mathbf{x} \sim \mathcal{N}_p\left(\boldsymbol{\mu}_{\boldsymbol{\Theta}} + \boldsymbol{\Sigma}_{\boldsymbol{\Theta}\mathbf{X}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}}), \boldsymbol{\Sigma}_{\boldsymbol{\Theta}\boldsymbol{\Theta}} - \boldsymbol{\Sigma}_{\boldsymbol{\Theta}\mathbf{X}}\boldsymbol{\Sigma}_{\mathbf{XX}}^{-1}\boldsymbol{\Sigma}_{\mathbf{X}\boldsymbol{\Theta}}\right)$$

cf. exercise 1.10-7.

Theorem:

Given the linear model

$$\mathbf{X} = \mathbf{H\Theta} + \mathbf{Z},$$

where $\mathbf{H}$ is a known matrix, $\mathbf{\Theta}$ and $\mathbf{Z}$ are statistically independent random vectors with $\mathbf{\Theta} \sim \mathcal{N}_p(\boldsymbol{\mu}_\Theta, \boldsymbol{\Sigma}_{\Theta\Theta})$ and $\mathbf{Z} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{\Sigma}_{ZZ})$. Then

$$\mathbf{Y} = \begin{pmatrix} \mathbf{X} \\ \mathbf{\Theta} \end{pmatrix} \sim \mathcal{N}_{n+p}\left(\begin{pmatrix} \mathbf{H}\boldsymbol{\mu}_\Theta \\ \boldsymbol{\mu}_\Theta \end{pmatrix}, \begin{pmatrix} \mathbf{H}\boldsymbol{\Sigma}_{\Theta\Theta}\mathbf{H}^T + \boldsymbol{\Sigma}_{ZZ} & \mathbf{H}\boldsymbol{\Sigma}_{\Theta\Theta} \\ \boldsymbol{\Sigma}_{\Theta\Theta}\mathbf{H}^T & \boldsymbol{\Sigma}_{\Theta\Theta} \end{pmatrix} \right)$$

and the MMSE estimate can be expressed by

$$\hat{\boldsymbol{\theta}}_{MMSE}(\mathbf{x}) = \boldsymbol{\mu}_\Theta + \boldsymbol{\Sigma}_{\Theta\Theta}\mathbf{H}^T \left(\mathbf{H}\boldsymbol{\Sigma}_{\Theta\Theta}\mathbf{H}^T + \boldsymbol{\Sigma}_{ZZ} \right)^{-1} (\mathbf{x} - \mathbf{H}\boldsymbol{\mu}_\Theta).$$

Exercise 3.7-1:

Proof of the Theorem

3.7.2 Minimum Mean Absolute Error Estimation

The mean absolute error estimation is considered for the single parameter case only. To estimate the random parameter Θ by means of the realization of the random vector $\mathbf{X}$ in the minimum mean absolute error sense

$$R_{MAE}(\hat{\theta}) = \mathrm{E}\left(\left|\Theta - \hat{\theta}(\mathbf{X})\right|\right) = \int_{\mathbb{R}^{n+1}} |\theta - \hat{\theta}(\mathbf{x})|\, f_{\mathbf{X}\Theta}(\mathbf{x},\theta)\, d\mathbf{x}\, d\theta$$

has to be minimized. As shown in the subsequent exercise the minimization of $R_{MAE}(\hat{\theta})$ leads to

$$\int_{-\infty}^{\hat{\theta}_{MMAE}(\mathbf{x})} f_{\Theta}(\theta \mid \mathbf{x})\, d\theta = \int_{\hat{\theta}_{MMAE}(\mathbf{x})}^{\infty} f_{\Theta}(\theta \mid \mathbf{x})\, d\theta.$$

Hence, the $\hat{\theta}_{MMAE}(\mathbf{x})$ is the median of the posterior density function $f_{\Theta}(\theta \mid \mathbf{x})$.

Exercise 3.7-2:

Proof of median

3.7.3 Maximum A Posteriori Estimation

Another estimation technique that fits within the Bayesian framework is the maximum a posteriori estimation (MAP).

Single parameter case

To motivate the approach the uniform loss function

$$L(\theta,\hat{\theta}(\mathbf{x})) = \begin{cases} 0 & |\theta - \hat{\theta}(\mathbf{x})| \le \delta \\ 1 & |\theta - \hat{\theta}(\mathbf{x})| > \delta \end{cases} \quad \text{where} \quad \delta > 0$$

is introduced. The Bayes risk is then given by

$$R_L(\hat{\theta}) = \mathrm{E}\Big(L(\theta,\hat{\theta}(\mathbf{X}))\Big) = \int_{\mathbb{R}^{n+1}} L(\theta,\hat{\theta}(\mathbf{x}))\, f_{\mathbf{X}\Theta}(\mathbf{x},\theta)\, d\mathbf{x}\, d\theta$$

$$= \int_{\mathbb{R}^n} \left(\int_{-\infty}^{\infty} L(\theta,\hat{\theta}(\mathbf{x}))\, f_{\Theta}(\theta \,|\, \mathbf{x})\, d\theta \right) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}.$$

Since $f_{\mathbf{X}}(\mathbf{x})$ and the integral in the brackets are always positive it is sufficient to minimize

$$\int_{-\infty}^{\infty} L(\theta,\hat{\theta}(\mathbf{x}))f_{\Theta}(\theta|\mathbf{x})d\theta = \int_{-\infty}^{\hat{\theta}(\mathbf{x})-\delta} f_{\Theta}(\theta|\mathbf{x})d\theta + \int_{\hat{\theta}(\mathbf{x})+\delta}^{\infty} f_{\Theta}(\theta|\mathbf{x})d\theta$$

or equivalently to maximize

$$1-\left(\int_{-\infty}^{\hat{\theta}(\mathbf{x})-\delta} f_{\Theta}(\theta|\mathbf{x})d\theta + \int_{\hat{\theta}(\mathbf{x})+\delta}^{\infty} f_{\Theta}(\theta|\mathbf{x})d\theta\right) = \int_{\hat{\theta}(\mathbf{x})-\delta}^{\hat{\theta}(\mathbf{x})+\delta} f_{\Theta}(\theta|\mathbf{x})d\theta$$

for every **x.**

For δ arbitrary small $\hat{\theta}(\mathbf{x})$ determines the mode of $f_{\Theta}(\theta|\mathbf{x})$ and is termed maximum a posteriori (MAP) estimate, i.e.

$$\hat{\theta}_{MAP}(\mathbf{x}) = \arg\max_{\theta} f_{\Theta}(\theta|\mathbf{x}) = \arg\max_{\theta} f_{\mathbf{X}}(\mathbf{x}\,|\,\theta)f_{\Theta}(\theta)$$
$$= \arg\max_{\theta} \left(\ln\left(f_{\mathbf{X}}(\mathbf{x}\,|\,\theta)\right) + \ln\left(f_{\Theta}(\theta)\right)\right)$$

Multiple parameter case

Employing the marginal posterior density functions

$$f_{\Theta_i}(\theta_i \mid \mathbf{x}) = \int_{\mathbb{R}^{p-1}} f_{\boldsymbol{\Theta}}(\boldsymbol{\theta} \mid \mathbf{x}) d\theta_1 \ldots d\theta_{i-1} d\theta_{i+1} \ldots d\theta_p$$

the MAP estimates are given by

$$\hat{\theta}_{i,MAP}(\mathbf{x}) = \arg\max_{\theta_i} f_{\Theta_i}(\theta_i \mid \mathbf{x}) \quad i = 1, \ldots, p.$$

However, due to the required integration we might propose the alternative vector valued MAP estimate

$$\hat{\boldsymbol{\theta}}_{MAP}(\mathbf{x}) = \arg\max_{\boldsymbol{\theta}} f_{\boldsymbol{\Theta}}(\boldsymbol{\theta} \mid \mathbf{x}) = \arg\max_{\boldsymbol{\theta}} f_{\mathbf{X}}(\mathbf{x} \mid \boldsymbol{\theta}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}).$$

Remark:

Generally, both MAP estimates are not equivalent.

Exercise 3.7-3:

MMSE, MMAE and MAP estimation of the parameter of an exponential distribution, where the parameter obeys also an exponential distribution

Exercise 3.7-4:

Signal + noise, MMSE signal amplitude estimation

References to Chapter 3

[1] M. Fisz, *Probability Theory and Mathematical Statistics*, Krieger Publishing Company, 1980

[2] C.W. Helstrom, *Elements of Signal Detection and Estimation,* Prentice Hall, 1994

[3] S. Kay, *Fundamentals of Statistical Signal Processing, Vol. 1: Estimation Theory*, Prentice Hall, 1993

[4] E.L. Lehmann and G. Casella, *Theory of Point Estimation*, Springer, 2003

[5] H.V. Poor, *An Introduction to Signal Detection and Estimation*, Springer, 1994

[6] C.R. Rao, *Linear Statistical Inference and Its Application*, John Wiley, 1973

[7] L.L. Scherf, *Statistical Signal Processing*, Addison Wesley, 1990