An Introduction to Data Based Modeling

Rik Pintelon

Vrije Universiteit Brussel, dept. ELEC

April 2, 2021
Goal

Construction of a model from noisy data
Applications

Non-exhaustive list:
1. prediction
2. physical interpretation
3. control
4. computer aided design
5. fault detection
6. monitoring
7. classification
8. denoising of medical images
Developed in different scientific disciplines:

1. *regression analysis* in statistics
2. *system identification* in systems and control
3. *machine learning* in artificial intelligence
Techniques–Algorithms

Developed in different scientific disciplines:

1. *regression analysis* in statistics
2. *system identification* in systems and control
3. *machine learning* in artificial intelligence

Similarities: data, model, cost function, minimization algorithm
Differences: terminology used
Three simple examples reveal three critical issues in data based modeling:

1. estimation of the resistor value from DC-measurements
   choice of the cost function

2. polynomial curve fitting
   selection of the model complexity

3. estimation of the slope of a straight line
   nonlinear optimization cost function
Estimation Resistor Value

Choice Cost Function

\[ i_0 = 1 \text{A} \quad R_0 = 1 \Omega \]
\[ u_0 = 1 \text{V} \]

\[ i(k) = i_0 + n_i(k) \quad \text{with} \quad n_i(k) \in N(0, \sigma_i^2) \]
\[ u(k) = u_0 + n_u(k) \quad \text{with} \quad n_u(k) \in N(0, \sigma_u^2) \]

with \( \sigma_i = 0.5 \text{A} \) and \( \sigma_u = 0.2 \text{V} \)
Estimation Resistor Value

Choice Cost Function

Application Ohm’s law \( R(k) = \frac{u(k)}{i(k)} \)
Estimation Resistor Value
Choice Cost Function

Solution 1: Simple Approach – red line

\[
\hat{R}_{SA}(N) = \arg \min_R \sum_{k=1}^{N} \left( \frac{u(k)}{i(k)} - R \right)^2
\]

\[
= \frac{1}{N} \sum_{k=1}^{N} \frac{u(k)}{i(k)}
\]
Estimation Resistor Value

Choice Cost Function

Solution 1: Least Squares – green line

\[ \hat{R}_{LS}(N) = \arg \min_R \sum_{k=1}^{N} (u(k) - R_i(k))^2 \]

\[
= \frac{1}{N} \sum_{k=1}^{N} u(k) i(k) \frac{1}{N} \sum_{k=1}^{N} i^2(k)
\]
Solution 3: Maximum Likelihood – black line

\[ \hat{R}_{ML}(N) = \arg \min_R \left[ \min_{i_p} \sum_{k=1}^{N} (u(k) - R_i)^2 + \sum_{k=1}^{N} (i(k) - i_p)^2 \right] \]

\[ = \frac{1}{N} \sum_{k=1}^{N} u(k) \]

\[ = \frac{1}{N} \sum_{k=1}^{N} i(k) \]
Estimation Resistor Value

Choice Cost Function

Solution 3: Maximum Likelihood – five different realizations
$N = 17$ noisy samples $y(k)$, $k = 1, 2, \ldots, N$, of the arctan function are generated over the interval $[-2, 4]$ 

$$y(k) = \arctan u(k) + n_y(k) \text{ with } u(k) = 6 \frac{k - 1}{N - 1} - 2$$

and where $n_y(k) \in \mathcal{N}(0, \sigma_y^2)$
Polynomial Curve Fitting
Selection Model Complexity

Polynomial least squares estimate of order \( n_\theta - 1 = 3, 8 \) and 14

\[
\hat{\theta} = \arg \min_\theta \sum_{k=1}^{N} (y(k) - \sum_{r=1}^{n_\theta} \theta[r] u^{(r-1)}(k))^2
\]

\[
\hat{y}(k) = \sum_{r=1}^{n_\theta} \hat{\theta}[r] u^{(r-1)}(k)
\]

is repeated for \( M = 400 \) independent realizations of the disturbing noise \( n_y(k) \) and three different values of \( \sigma_y \)
Polynomial Curve Fitting

Selection Model Complexity

Sample bias

\[ \mathbb{E}\{\hat{y}(k) - \arctan u(k)\} \approx \frac{1}{M} \sum_{m=1}^{M} \hat{y}^{[m]}(k) - \arctan u(k) \]

and sample root mean squared error (RMSE)

\[ \mathbb{E}\left\{ (\hat{y}(k) - \arctan u(k))^2 \right\} \approx \frac{1}{M} \sum_{m=1}^{M} (\hat{y}^{[m]}(k) - \arctan u(k))^2 \]
Polynomial Curve Fitting
Selection Model Complexity

\[ y = 0 \]
\[ y = 0.1 \]
\[ y = 0.5 \]

polynomial order \( n_\theta - 1 = 3, 8 \) and 14
Estimation Slope Straight Line
Nonlinear Optimization Cost Function

Straight line through the origin

\[ y_0(k) = \theta_0 u_0(k) \quad \text{with} \quad u_0(k) \in N(0, 0.5^2) \quad \text{for} \quad k = 1, 2, \ldots, N \]

and where \( \theta_0 = 1 \)
The slope parameter $\theta_0$ is estimated from $N = 100$ noisy measurements $u(k)$ and $y(k)$ of, respectively, $u_0(k)$ and $y_0(k)$:

$$u(k) = u_0(k) + n_u(k)$$
$$y(k) = y_0(k) + n_y(k)$$

where $n_u(k) \in N(0, \sigma_u^2(k))$ and $n_y(k) \in N(0, \sigma_y^2(k))$ and with

$$\sigma_u(k) = 0.25 \left(1 + \sin \left(2\pi \frac{k - 1}{N}\right) \right)$$
$$\sigma_y(k) = 0.1 \left(1 + \cos \left(2\pi \frac{k - 1}{N}\right) \right)$$
The Gaussian maximum likelihood estimate of the slope parameter minimizes the following cost function

$$\frac{1}{N} \sum_{k=1}^{N} \frac{(y(k) - \theta u(k))^2}{\sigma_y^2(k) + \theta^2 \sigma_u^2(k)}$$
Linear least squares approximation

\[ \frac{1}{N} \sum_{k=1}^{N} (y(k) - \theta u(k))^2 \]

of which the solution is \( \hat{\theta}_{LS} = 0.90 \), is close enough to \( \hat{\theta}_{ML} = 0.99 \).
Outline

• Tools for Analyzing Estimators
• Linear Least Squares
• Nonlinear Least Squares
• Maximum Likelihood Method
• Bayesian Approach
• Neural Networks
• Tuning the Model Complexity
Outline

• Tools for Analyzing Estimators
• Linear Least Squares
• Nonlinear Least Squares
• Maximum Likelihood Method
• Bayesian Approach
• Neural Networks
• Tuning the Model Complexity
Tools for Analyzing Estimators

1. Stochastic Convergence
2. Consistency – Asymptotically Unbiased
3. Law of Large Numbers
4. Central Limit Theorem
5. Cramér-Rao Lower Bound – Asymptotic Efficiency
6. Robustness
7. Example 1: Estimation of the Resistor Value
8. Example 2: Estimation of the Slope of a Straight Line
• Tools for Analyzing Estimators
  1. Stochastic Convergence
  2. Consistency – Asymptotically Unbiased
  3. Law of Large Numbers
  4. Central Limit Theorem
  5. Cramér-Rao Lower Bound – Asymptotic Efficiency
  6. Robustness
  7. Example 1: Estimation of the Resistor Value
  8. Example 2: Estimation of the Slope of a Straight Line
Tools for Analyzing Estimators
Four Definitions of Stochastic Convergence

1. Convergence with probability one

\[ \text{Prob}\left( \lim_{N \to \infty} \hat{\theta}(N) = \theta_0 \right) = 1 \Rightarrow \begin{cases} 
\lim_{N \to \infty} \hat{\theta}(N) = \theta_0 \text{ w.p. 1} \\
a. s. \lim_{N \to \infty} \hat{\theta}(N) = \theta_0 
\end{cases} \]

2. Convergence in probability

\[ \forall \epsilon > 0, \lim_{N \to \infty} \text{Prob}\left( |\hat{\theta}(N) - \theta_0| < \epsilon \right) = 1 \Rightarrow \begin{cases} 
\lim_{N \to \infty} \hat{\theta}(N) = \theta_0 \text{ in prob.} \\
\text{plim}_{N \to \infty} \hat{\theta}(N) = \theta_0 
\end{cases} \]
3. Convergence in mean squared sense

\[ \lim_{N \to \infty} \mathbb{E}\{(\hat{\theta}(N) - \theta_0)^2\} = 0 \Rightarrow \begin{cases} 
\lim_{N \to \infty} \hat{\theta}(N) = \theta_0 \text{ in m.s.} \\
l. i. m. \hat{\theta}(N) = \theta_0
\end{cases} \]

4. Convergence in law

\[ \lim_{N \to \infty} F_N(\theta) = F(\theta) \Rightarrow \begin{cases} 
\lim_{N \to \infty} \hat{\theta}(N) = \theta_0 \text{ in distr.} \\
\lim_{N \to \infty} \hat{\theta}(N) = \theta_0
\end{cases} \]
Interrelations between the stochastic limits
Outline

• Tools for Analyzing Estimators
  1. Stochastic Convergence
  2. Consistency – Asymptotically Unbiased
  3. Law of Large Numbers
  4. Central Limit Theorem
  5. Cramér-Rao Lower Bound – Asymptotic Efficiency
  6. Robustness
  7. Example 1: Estimation of the Resistor Value
  8. Example 2: Estimation of the Slope of a Straight Line
Tools for Analyzing Estimators

Consistency – Asymptotically Unbiased

Consistency

**strongly consistent:**  \( \lim_{N \to \infty} \hat{\theta}(N) = \theta_0 \)

**weakly consistent:**  \( \text{plim} \ \hat{\theta}(N) = \theta_0 \)

**consistent:**  \( \text{l. i. m.} \ \hat{\theta}(N) = \theta_0 \)

(Asymptotically) unbiased

\[
\mathbb{E}\{\hat{\theta}(N)\} = \theta_0 \\
\lim_{N \to \infty} \mathbb{E}\{\hat{\theta}(N)\} = \theta_0
\]
Tools for Analyzing Estimators

Consistency – Asymptotically Unbiased

\[ N \to \infty \]

\[ f_{\hat{\theta}}(\hat{\theta} - \theta_0) \]

\[ f_{\hat{\theta}}(\hat{\theta} - \theta_0) \]

\[ N \to \infty \]
Outline

- Tools for Analyzing Estimators
  1. Stochastic Convergence
  2. Consistency – Asymptotically Unbiased
  3. Law of Large Numbers
  4. Central Limit Theorem
  5. Cramér-Rao Lower Bound – Asymptotic Efficiency
  6. Robustness
  7. Example 1: Estimation of the Resistor Value
  8. Example 2: Estimation of the Slope of a Straight Line
Consider the mean value $S(N)$ of $N$ random variables $x(k)$, $k = 1, 2, \ldots, N$,

$$S(N) = \frac{1}{N} \sum_{k=1}^{N} x(k)$$

Laws of large numbers

- **Strong law of large numbers**: $\text{a.s. } \lim_{N \to \infty} (S(N) - \mathbb{E}\{S(N)\}) = 0$
- **Weak law of large numbers**: $\text{plim}_{N \to \infty} (S(N) - \mathbb{E}\{S(N)\}) = 0$
- **Law of large numbers**: $\text{l.i.m.}_{N \to \infty} (S(N) - \mathbb{E}\{S(N)\}) = 0$
Outline

- Tools for Analyzing Estimators
  1. Stochastic Convergence
  2. Consistency – Asymptotically Unbiased
  3. Law of Large Numbers
  4. Central Limit Theorem
  5. Cramér-Rao Lower Bound – Asymptotic Efficiency
  6. Robustness
  7. Example 1: Estimation of the Resistor Value
  8. Example 2: Estimation of the Slope of a Straight Line
Consider the mean value $S(N)$ of $N$ random variables $x(k)$, $k = 1, 2, \ldots, N$,

$$S(N) = \frac{1}{N} \sum_{k=1}^{N} x(k)$$

Central limit theorem

$$\lim_{N \to \infty} \frac{S(N) - \mathbb{E}\{S(N)\}}{\sqrt{\text{var}(S(N))}} \in \mathcal{N}(0, 1)$$
Tools for Analyzing Estimators

Central Limit Theorem: Example

\( x(k) \) uniformly distributed with zero mean and variance \( 25/3 \)

(a) \( N = 1 \), (b) \( N = 2 \), and (c) \( N = 3 \)
• Tools for Analyzing Estimators
  1. Stochastic Convergence
  2. Consistency – Asymptotically Unbiased
  3. Law of Large Numbers
  4. Central Limit Theorem
  5. Cramér-Rao Lower Bound – Asymptotic Efficiency
  6. Robustness
  7. Example 1: Estimation of the Resistor Value
  8. Example 2: Estimation of the Slope of a Straight Line
Given noisy measurements $z = z_0 + v \in \mathbb{R}^N$, with $z_0$ the true unknown value and $v$ the disturbing noise, and a model $z_0 = h(\theta_0)$. 
Given noisy measurements $z = z_0 + \nu \in \mathbb{R}^N$, with $z_0$ the true unknown value and $\nu$ the disturbing noise, and a model $z_0 = h(\theta_0)$.

Using the pdf $f_z(z) = f_\nu(z - z_0)$ of the data and the model $z_0 = h(\theta_0)$, construct the pdf of the data $z$, given the true model parameters $\theta_0$

$$f_{z|\theta_0}(z|\theta_0) = f_\nu(z - h(\theta_0))$$
Given noisy measurements $z = z_0 + v \in \mathbb{R}^N$, with $z_0$ the true unknown value and $v$ the disturbing noise, and a model $z_0 = h(\theta_0)$.

Using the pdf $f_z(z) = f_v(z - z_0)$ of the data and the model $z_0 = h(\theta_0)$, construct the pdf of the data $z$, given the true model parameters $\theta_0$

$$f_{z|\theta_0}(z|\theta_0) = f_v(z - h(\theta_0))$$

Assumptions

1. first and second order derivatives $f_{z|\theta_0}(z|\theta_0)$ w.r.t. $\theta_0$ exist

2. the boundaries of the domain of $f_{z|\theta_0}(z|\theta_0)$ w.r.t. $z$ are $\theta_0$-independent
Tools for Analyzing Estimators
Cramér-Rao Lower Bound – Asymptotic Efficiency

Within the class of *unbiased estimators*, the covariance of the estimate $\hat{\theta}(z)$ is bounded below by the Cramér-Rao lower bound:

$$\text{Cov}(\hat{\theta}(z)) \geq \text{Fi}^{-1}(\theta_0)$$

with $\text{Fi}(\theta_0)$ the Fisher information matrix:

$$\text{Fi}(\theta_0) = \mathbb{E} \left\{ \left( \frac{\partial \log f_z|\theta_0(z|\theta_0)}{\partial \theta_0} \right)^T \left( \frac{\partial \log f_z|\theta_0(z|\theta_0)}{\partial \theta_0} \right) \right\}$$

$$= -\mathbb{E} \left\{ \frac{\partial^2 \log f_z|\theta_0(z|\theta_0)}{\partial \theta_0^2} \right\}$$

[proof: see lecture notes]
Tools for Analyzing Estimators

Cramér-Rao Lower Bound – Asymptotic Efficiency

\[
\text{Cov}(\hat{\theta}(z)) \geq F_i^{-1}(\theta_0)
\]

Notes

• asymptotically \((N \to \infty)\) valid for \textit{consistent estimators}
• Estimator is \textit{efficient} if \(\text{Cov}(\hat{\theta}(z)) = F_i^{-1}(\theta_0)\)
• Estimator is \textit{asymptotically efficient} if the equality holds asymptotically for \(N \to \infty\)
• Tools for Analyzing Estimators
  1. Stochastic Convergence
  2. Consistency – Asymptotically Unbiased
  3. Law of Large Numbers
  4. Central Limit Theorem
  5. Cramér-Rao Lower Bound – Asymptotic Efficiency
  6. Robustness
  7. Example 1: Estimation of the Resistor Value
  8. Example 2: Estimation of the Slope of a Straight Line
Tools for Analyzing Estimators

Robustness

Estimators are constructed under some assumptions on the data.

In practice, these assumptions are not perfectly met.

How sensitive the asymptotic properties of an estimator are w.r.t. deviations from the basic assumptions made?

Robustness analysis: check whether basic tools still valid under the relaxed conditions.
• **Tools for Analyzing Estimators**
  1. Stochastic Convergence
  2. Consistency – Asymptotically Unbiased
  3. Law of Large Numbers
  4. Central Limit Theorem
  5. Cramér-Rao Lower Bound – Asymptotic Efficiency
  6. Robustness
  7. **Example 1: Estimation of the Resistor Value**
  8. **Example 2: Estimation of the Slope of a Straight Line**
Tools for Analyzing Estimators

Estimation Resistor Value: Consistency

Simple Approach, Least Squares, and Maximum Likelihood

\[ \hat{R}_{SA}(N) \xrightarrow{\text{w.p. } 1} \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\left\{ \frac{u(k)}{i(k)} \right\} = \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\{u(k)\} \mathbb{E}\left\{ \frac{1}{i(k)} \right\} = \infty \]

\[ \hat{R}_{LS}(N) \xrightarrow{\text{w.p. } 1} \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\{u(k)i(k)\} = \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\{i(k)^2\} = \frac{u_0 i_0}{i_0^2 + \sigma_i^2} = \frac{R_0}{1 + \sigma_i^2 / i_0^2} \]

\[ \hat{R}_{ML}(N) \xrightarrow{\text{w.p. } 1} \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\{u(k)\} = \frac{u_0}{i_0} = R_0 \]
Tools for Analyzing Estimators

Estimation Resistor Value: Consistency

Bias Compensated Least Squares

\[ \hat{R}_{\text{BCLS}}(N) = \frac{1}{N} \sum_{k=1}^{N} u(k)i(k) \]
\[ \frac{1}{N} \sum_{k=1}^{N} i^2(k) - \sigma_i^2 \]

\[ \hat{R}_{\text{BCLS}}(N) \xrightarrow{\text{w.p. 1}} \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\{u(k)i(k)\} \]
\[ \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\{i(k)^2\} - \sigma_i^2 \]

\[ = \frac{u_0i_0}{i_0^2} = R_0 \]
Tools for Analyzing Estimators

Estimation Resistor Value: Asymptotic Variance

Least Squares

\[
\frac{1}{N} \sum_{k=1}^{N} u(k)i(k) = u_0i_0 + w_1(N) = u_0i_0(1 + \frac{w_1(N)}{u_0i_0})
\]

\[
\frac{1}{N} \sum_{k=1}^{N} i^2(k) = i_0^2 + \sigma_i^2 + w_2(N) = (i_0^2 + \sigma_i^2)(1 + \frac{w_2(N)}{i_0^2 + \sigma_i^2})
\]

where \(w_1(N)\) and \(w_2(N)\) are zero mean random variables

\[
w_1(N) = \frac{1}{N} \sum_{k=1}^{N} n_u(k)i_0 + n_i(k)u_0 + n_u(k)n_i(k)
\]

\[
w_2(N) = \frac{1}{N} \sum_{k=1}^{N} 2i_0n_i(k) + n_i^2(k) - \sigma_i^2
\]
Tools for Analyzing Estimators

Estimation Resistor Value: Asymptotic Variance

Strong law of large numbers

\[
\begin{align*}
\frac{w_1(N)}{u_0 i_0} & \xrightarrow{\text{w.p. } 1} 0 \\
\frac{w_2(N)}{i_0^2 + \sigma_i^2} & \xrightarrow{\text{w.p. } 1} 0
\end{align*}
\]

First order Taylor series expansion, with probability one,

\[
\hat{R}_{LS}(N) = \frac{R_0}{1 + \sigma_i^2 / i_0^2} \left( 1 + \frac{w_1(N)}{u_0 i_0}\right)
\]

\[
\approx \frac{R_0}{1 + \sigma_i^2 / i_0^2} \left( 1 + \frac{w_1(N)}{u_0 i_0} - \frac{w_2(N)}{i_0^2 + \sigma_i^2}\right)
\]
Tools for Analyzing Estimators

Estimation Resistor Value: Asymptotic Variance

Convergence w.p. 1 to a limit random variable \( \delta_{\text{RLS}}(N) \)

\[
\hat{R}_{\text{LS}}(N) - \frac{R_0}{1 + \sigma_i^2/i_0^2} \xrightarrow{\text{w.p. 1}} \delta_{\text{RLS}}(N) = \frac{R_0}{1 + \sigma_i^2/i_0^2} \left( \frac{w_1(N)}{u_0i_0} - \frac{w_2(N)}{i_0^2 + \sigma_i^2} \right)
\]

Asymptotic variance \( \hat{R}_{\text{LS}}(N) = \text{variance} \delta_{\text{RLS}}(N) \)

\[
\text{"var}(\hat{R}_{\text{LS}}(N))" = \text{var}(\delta_{\text{RLS}}(N)) = \frac{R_0^2}{(1 + \sigma_i^2/i_0^2)^2} \text{var} \left( \frac{w_1(N)}{u_0i_0} - \frac{w_2(N)}{i_0^2 + \sigma_i^2} \right)
\]

\[
= \frac{R_0^2}{N(1 + \sigma_i^2/i_0^2)^2} \left( \frac{\sigma_u^2}{u_0^2} + \frac{\sigma_i^2}{i_0^2} + \frac{\sigma_u^2 \sigma_i^2}{u_0^2 i_0^2} - 2 \frac{\sigma_i^4/i_0^4}{(1 + \sigma_i^2/i_0^2)^2} \right)
\]
Tools for Analyzing Estimators

Estimation Resistor Value: Asymptotic Variance

**Maximum Likelihood**

\[
\frac{1}{N} \sum_{k=1}^{N} u(k) = u_0 + w_1(N) = u_0 \left(1 + \frac{w_1(N)}{u_0}\right)
\]

\[
\frac{1}{N} \sum_{k=1}^{N} i(k) = i_0 + w_2(N) = i_0 \left(1 + \frac{w_2(N)}{i_0}\right)
\]

where \( w_1(N) \) and \( w_2(N) \) are zero mean random variables

\[
\begin{align*}
    w_1(N) &= \frac{1}{N} \sum_{k=1}^{N} n_u(k) \\
    w_2(N) &= \frac{1}{N} \sum_{k=1}^{N} n_i(k)
\end{align*}
\]
Tools for Analyzing Estimators

Estimation Resistor Value: Asymptotic Variance

Strong law of large numbers

\[
\frac{w_1(N)}{u_0} \xrightarrow{\text{w.p.} 1} 0
\]

\[
\frac{w_2(N)}{i_0} \xrightarrow{\text{w.p.} 1} 0
\]

First order Taylor series expansion, with probability one,

\[
\hat{R}_{ML}(N) = R_0 \left( 1 + \frac{w_1(N)}{u_0} \right) \left( 1 + \frac{w_2(N)}{i_0} \right)
\]

\[
\approx R_0 \left( 1 + \frac{w_1(N)}{u_0} - \frac{w_2(N)}{i_0} \right)
\]
Tools for Analyzing Estimators

Estimation Resistor Value: Asymptotic Variance

Convergence w.p. 1 to a limit random variable $\delta_{R_{ML}}(N)$

$$\hat{R}_{ML}(N) - R_0 \xrightarrow{w.p. 1} \delta_{R_{ML}}(N) = R_0\left(\frac{w_1(N)}{u_0} - \frac{w_2(N)}{i_0}\right)$$

Asymptotic variance $\hat{R}_{ML}(N) = \text{variance} \delta_{R_{ML}}(N)$

$$\text{"var}(\hat{R}_{ML}(N))" = \text{var}(\delta_{R_{ML}}(N)) = R_0^2 \text{var}\left(\frac{w_1(N)}{u_0} - \frac{w_2(N)}{i_0}\right) = \frac{R_0^2}{N}\left(\frac{\sigma_u^2}{u_0^2} + \frac{\sigma_i^2}{i_0^2}\right)$$
Tools for Analyzing Estimators

Estimation Resistor Value: Asymptotic Variance

Evaluation asymptotic variances for $N = 100$, $R_0 = 1 \Omega$, $i_0 = 1$ A, $u_0 = 1$ V and $\sigma = \sigma_i/i_0 = \sigma_u/u_0$

maximum likelihood, least squares, bias compensated least squares
Tools for Analyzing Estimators

Estimation Resistor Value: Asymptotic Normality

Apply the central limit theorem to the sums $w_1(N)$ and $w_2(N)$

\[
\delta_{R_{LS}}(N) = \frac{R_0}{1 + \sigma_i^2/i_0^2} \left( \frac{w_1(N)}{u_0i_0} - \frac{w_2(N)}{i_0^2 + \sigma_i^2} \right)
\]

\[
\delta_{R_{ML}}(N) = R_0 \left( \frac{w_1(N)}{u_0} - \frac{w_2(N)}{i_0} \right)
\]

\[
\delta_{R_{BCLS}}(N) = R_0 \left( \frac{w_1(N)}{u_0i_0} - \frac{w_2(N)}{i_0^2} \right)
\]

$\Rightarrow \hat{R}_{LS}(N), \hat{R}_{ML}(N)$ and $\hat{R}_{BCLS}(N)$ are asymptotically normally distributed

Reason: $\hat{R}(N)$ converges w.p. 1 to $\delta_R(N)$ and, hence, also in law
Tools for Analyzing Estimators

Estimation Resistor Value: Asymptotic Efficiency

Calculation Fisher information matrix

\[ z = \begin{bmatrix} u(1) & u(2) & \ldots & u(N) & i(1) & i(2) & \ldots & i(N) \end{bmatrix}^T \]

\[
f_{z|R_0,i_0}(z|R_0, i_0) = \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi \sigma_u}} e^{-\frac{(u(k)-R_0i_0)^2}{2\sigma_u^2}} \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi \sigma_i}} e^{-\frac{(i(k)-i_0)^2}{2\sigma_i^2}} \\
= \frac{1}{(2\pi \sigma_u \sigma_i)^N} e^{-\sum_{k=1}^{N} \frac{(u(k)-R_0i_0)^2}{2\sigma_u^2}} - \sum_{k=1}^{N} \frac{(i(k)-i_0)^2}{2\sigma_i^2} + \text{Cnst}
\]

\[- \log f_{z|R_0,i_0}(z|R_0, i_0) = \sum_{k=1}^{N} \frac{(u(k) - R_0i_0)^2}{2\sigma_u^2} + \frac{(i(k) - i_0)^2}{2\sigma_i^2} + \text{Cnst} \]
Tools for Analyzing Estimators

Estimation Resistor Value: Asymptotic Efficiency

\[
- \frac{\partial^2 \log f_{z|\theta_0}(z|\theta_0)}{\partial \theta_0^2} = \begin{bmatrix}
\sum_{k=1}^{N} \frac{i_0^2}{\sigma_u^2} & - \sum_{k=1}^{N} \frac{u(k) - 2R_0i_0}{\sigma_u^2} \\
- \sum_{k=1}^{N} \frac{u(k) - 2R_0i_0}{\sigma_u^2} & \sum_{k=1}^{N} \frac{R_0^2}{\sigma_u^2} + \frac{1}{\sigma_i^2}
\end{bmatrix}
\]

\[
Fi(\theta_0) = \mathbb{E} \left\{ - \frac{\partial^2 \log f_{z|\theta_0}(z|\theta_0)}{\partial \theta_0^2} \right\} = \begin{bmatrix}
N \frac{i_0^2}{\sigma_u^2} & N \frac{R_0i_0}{\sigma_u^2} \\
N \frac{R_0i_0}{\sigma_i^2} & N \frac{R_0^2 \sigma_i^2 + \sigma_u^2}{\sigma_u^2 \sigma_i^2}
\end{bmatrix}
\]

\[
Fi^{-1}(\theta_0) = \frac{1}{N} \begin{bmatrix}
\frac{R_0^2 \sigma_i^2 + \sigma_u^2}{i_0^2} & \frac{-R_0 \sigma_i^2}{i_0} \\
\frac{-R_0 \sigma_i^2}{i_0} & \frac{\sigma_i^2}{\sigma_i^2}
\end{bmatrix}
\]

\[
\Rightarrow \quad \text{var}(\hat{R}(N)) \geq \frac{R_0^2}{N} \left( \frac{\sigma_i^2}{i_0^2} + \frac{\sigma_u^2}{u_0^2} \right)
\]
Cramér-Rao lower bound

$$\text{var}(\hat{R}(N)) \geq \frac{R_0^2}{N} \left( \frac{\sigma_i^2}{i_0^2} + \frac{\sigma_u^2}{u_0^2} \right)$$

Asymptotic variance ML estimate

"$$\text{var}(\hat{R}_{ML}(N)) = \text{var}(\delta_{R_{ML}}(N)) = \frac{R_0^2}{N} \left( \frac{\sigma_u^2}{u_0^2} + \frac{\sigma_i^2}{i_0^2} \right)$$"

$$\Rightarrow$$ the ML estimate is asymptotically efficient
Tools for Analyzing Estimators

Estimation Resistor Value: Robustness

Assumptions Maximum Likelihood

1. noise on current measurement, $n_i(k)$, iid Gaussian
2. noise on voltage measurement, $n_u(k)$, iid Gaussian
3. $n_i(k)$ and $n_u(k)$ are mutually independent
Tools for Analyzing Estimators

Estimation Resistor Value: Robustness

Violation assumptions

1. Non-Gaussian iid $n_i(k)$ and $n_u(k)$:
   - consistent, asymptotically normally distributed, asymptotic variance expression remains valid
   - no longer asymptotically efficient

2. Non-Gaussian correlated $n_i(k)$ and $n_u(k)$:
   - consistent, asymptotically normally distributed
   - variance expression no longer valid and no longer asymptotically efficient
Outline

Tools for Analyzing Estimators

1. Stochastic Convergence
2. Consistency – Asymptotically Unbiased
3. Law of Large Numbers
4. Central Limit Theorem
5. Cramér-Rao Lower Bound – Asymptotic Efficiency
6. Robustness
7. Example 1: Estimation of the Resistor Value
8. Example 2: Estimation of the Slope of a Straight Line
Tools for Analyzing Estimators

Estimation Slope

ML cost function

\[ V(\theta, z) = \sum_{k=1}^{N} \frac{(y(k) - \theta u(k))^2}{\sigma_y^2(k) + \theta^2 \sigma_u^2(k)} \]

Estimate

\[ \hat{\theta}(z) = \arg \min_{\theta} V(\theta, z) \]

Scaling of the cost function

\[ V_N(\theta, z) = \frac{1}{N} V(\theta, z) \]
Tools for Analyzing Estimators

Estimation Slope: Consistency

Step 1: Expected value cost function is minimal in $\theta_0$

$$V_N(\theta) = \mathbb{E}\{ V_N(\theta, z) \} = \frac{1}{N} \sum_{k=1}^{N} \frac{(y_0(k) - \theta u_0(k))^2}{\sigma_y^2(k) + \theta^2 \sigma_u^2(k)} + 1$$
Tools for Analyzing Estimators

Estimation Slope: Consistency

Step 1: Expected value cost function is minimal in $\theta_0$

$$V_N(\theta) = \mathbb{E}\{V_N(\theta, z)\} = \frac{1}{N} \sum_{k=1}^{N} \frac{(y_0(k) - \theta u_0(k))^2}{\sigma_y^2(k) + \theta^2 \sigma_u^2(k)} + 1$$

Step 2: $V_N(\theta, z)$ convergences w.p. 1 to its expected value

$$\text{a.s.} \lim_{N \to \infty} (V_N(\theta, z) - V_N(\theta)) = 0$$
Tools for Analyzing Estimators

Estimation Slope: Consistency

Step 1: Expected value cost function is minimal in $\theta_0$

$$V_N(\theta) = \mathbb{E}\{V_N(\theta, z)\} = \frac{1}{N} \sum_{k=1}^{N} \frac{(y_0(k) - \theta u_0(k))^2}{\sigma_y^2(k) + \theta^2 \sigma_u^2(k)} + 1$$

Step 2: $V_N(\theta, z)$ convergences w.p. 1 to its expected value

$$\text{a.s.} \lim_{N \to \infty} (V_N(\theta, z) - V_N(\theta)) = 0$$

Step 3: The convergence is uniform in $\theta$ and the cost function has continuous derivative w.r.t. $\theta$

$$\text{a.s.} \lim_{N \to \infty} \hat{\theta}(z) = \theta_0$$
First order Taylor series with remainder

\[ V'_N(\hat{\theta}(z), z) = V'_N(\theta_0, z) + V''_N(\theta_1, z)(\hat{\theta}(z) - \theta_0) \]

where \( \theta_1 = t\hat{\theta}(z) + (1 - t)\theta_0 \) with \( t \in [0, 1] \)
First order Taylor series with remainder

\[ V'_N(\hat{\theta}(z), z) = V'_N(\theta_0, z) + V''_N(\theta_1, z)(\hat{\theta}(z) - \theta_0) \]

where \( \theta_1 = t\hat{\theta}(z) + (1 - t)\theta_0 \) with \( t \in [0, 1] \)

Since

1. By definition of \( \hat{\theta}(z) \): \( V'_N(\hat{\theta}(z), z) = 0 \)
2. Convergence \( \theta_1 \): \( \operatorname{a.s.} \lim_{N \to \infty} \theta_1 = t\theta_0 + (1 - t)\theta_0 = \theta_0 \)
3. Convergence second order derivative cost: \( \operatorname{a.s.} \lim_{N \to \infty} V''_N(\theta, z) = V''_N(\theta) \)
First order Taylor series with remainder

\[ V'_N(\hat{\theta}(z), z) = V'_N(\theta_0, z) + V''_N(\theta_1, z)(\hat{\theta}(z) - \theta_0) \]

where \( \theta_1 = t \hat{\theta}(z) + (1 - t) \theta_0 \) with \( t \in [0, 1] \)

Since

1. By definition of \( \hat{\theta}(z) \): \( V'_N(\hat{\theta}(z), z) = 0 \)
2. Convergence \( \theta_1 \): a.s. \( \lim_{N \to \infty} \theta_1 = t \theta_0 + (1 - t) \theta_0 = \theta_0 \)
3. Convergence second order derivative cost:
   a.s. \( \lim_{N \to \infty} V''_N(\theta, z) = V''_N(\theta) \)

we get

\[ \hat{\theta}(z) - \theta_0 \xrightarrow{w.p. 1} \delta_\theta(z) = -V^{-1}_N(\theta_0)V'_N(\theta_0, z) \]
Since $\mathbb{E}\{\delta_\theta(z)\} = 0$, we have that

$$\text{"var}(\hat{\theta}(z))" = \text{Cov}(\delta_\theta(z)) = V_N'' - 2(\theta_0)\mathbb{E}\{(V'_N(\theta_0, z))^2\}$$
Cramér-Rao lower bound
(see Chapter Maximum Likelihood Method)

\[
\text{var}(\hat{\theta}(z)) \geq \mathcal{F}_i^{-1}(\theta_0) \quad \text{with} \quad \mathcal{F}_i(\theta_0) = V_N''(\theta_0)
\]

\(\Rightarrow\) the ML estimate is not asymptotically efficient
Tools for Analyzing Estimators

Estimation Slope: Asymptotic Normality

Apply the central limit theorem to $V'_{N}(\theta_0, z)$

$$\delta_{\theta}(z) = -V''_{N}^{-1}(\theta_0)V'_{N}(\theta_0, z)$$

$\Rightarrow \hat{\theta}(z)$ is asymptotically normally distributed

Reason: $\hat{\theta}(z)$ converges w.p. 1 to $\delta_{\theta}(z)$ and, hence, also in law
Outline

• Tools for Analyzing Estimators
• **Linear Least Squares**
• Nonlinear Least Squares
• Maximum Likelihood Method
• Bayesian Approach
• Neural Networks
• Tuning the Model Complexity
• **Linear Least Squares**
  1. Linear Least Squares Estimator
  2. Stochastic Properties
  3. Numerical Stable Calculation
  4. Weighted Linear Least Squares
  5. Noisy Regression Matrix
  6. Bias Compensated Least Squares
  7. (Generalized) Total Least Squares
  8. Instrumental Variables
  9. Regularized Least Squares
  10. Outliers
Linear Least Squares

Linear Least Squares Estimator

Noisy observations \( y \in \mathbb{R}^{N \times 1} \) and a model \( y_0 = H\theta_0 \)

\[
y = H\theta_0 + \nu
\]

where the regression matrix \( H \in \mathbb{R}^{N \times n_\theta} \) is independent of the disturbing noise \( \nu \in \mathbb{R}^{N \times 1} \), with \( \mathbb{E}\{\nu\} = 0 \) and \( \text{Cov}(\nu) = C_\nu \).
Noisy observations $y \in \mathbb{R}^{N \times 1}$ and a model $y_0 = H\theta_0$

$$y = H\theta_0 + \nu$$

where the regression matrix $H \in \mathbb{R}^{N \times n_\theta}$ is independent of the disturbing noise $\nu \in \mathbb{R}^{N \times 1}$, with $\mathbb{E}\{\nu\} = 0$ and $\text{Cov}(\nu) = C_\nu$.

Linear least squares cost function

$$V_{LS}(\theta, y) = \frac{1}{2} (y - H\theta)^T (y - H\theta)$$
Linear Least Squares

Noisy observations $y \in \mathbb{R}^{N \times 1}$ and a model $y_0 = H\theta_0$

$$y = H\theta_0 + \nu$$

where the regression matrix $H \in \mathbb{R}^{N \times n_\theta}$ is independent of the disturbing noise $\nu \in \mathbb{R}^{N \times 1}$, with $\mathbb{E}\{\nu\} = 0$ and $\text{Cov}(\nu) = C_\nu$.

Linear least squares cost function

$$V_{\text{LS}}(\theta, y) = \frac{1}{2} (y - H\theta)^T (y - H\theta)$$

Minimization w.r.t. $\theta$, gives

$$\hat{\theta}_{\text{LS}}(y) = (H^T H)^{-1} H^T y$$

$$\hat{y}_{\text{LS}} = H(H^T H)^{-1} H^T y$$
• Linear Least Squares
  1. Linear Least Squares Estimator
  2. Stochastic Properties
  3. Numerical Stable Calculation
  4. Weighted Linear Least Squares
  5. Noisy Regression Matrix
  6. Bias Compensated Least Squares
  7. (Generalized) Total Least Squares
  8. Instrumental Variables
  9. Regularized Least Squares
  10. Outliers
Linear Least Squares

Stochastic Properties: Unbiased

Taking the expected value of the estimate, gives

$$\mathbb{E}\{\hat{\theta}_{LS}(y)\} = \theta_0 + (H^T H)^{-1} H^T \mathbb{E}\{v\} = \theta_0$$

for every $N \geq n_\theta$
Linear Least Squares
Stochastic Properties: Covariance

Calculating the covariance of the estimate, gives

$$\text{Cov}(\hat{\theta}_{LS}(y)) = (H^T H)^{-1} H^T C_v H (H^T H)^{-1}$$

for every $N \geq n_\theta$
For Gaussian distributed $v$, the pdf of $y$, given $\theta_0$, equals

$$f_{y|\theta_0}(y|\theta_0) = \frac{1}{\sqrt{(2\pi)^N\det C_v}}e^{-\frac{1}{2}(y-H\theta_0)^T C_v^{-1}(y-H\theta_0)}$$

with corresponding Fisher information matrix

$$Fi(\theta_0) = \mathbb{E} \left\{ -\frac{\partial^2 \log f_{z|\theta_0}(z|\theta_0)}{\partial \theta_0^2} \right\} = H^T C_v^{-1} H$$

$\Rightarrow \hat{\theta}_{LS}(y)$ is inefficient
Residual least squares estimate

\[ \delta\theta(v) = \hat{\theta}_{LS}(y) - \theta_0 = (H^T H)^{-1} H^T v \]
Linear Least Squares

Stochastic Properties: Consistency

Residual least squares estimate

\[ \delta_\theta(v) = \hat{\theta}_{LS}(y) - \theta_0 = (H^T H)^{-1} H^T v \]

If

1. \( \lim_{N \to \infty} \frac{1}{N} H^T H \) is of full rank
2. The correlation of the noise decreases sufficiently fast to zero [mixing condition of order \( P \geq 2 \)]
3. For each column of \( H/N \), the sum of the absolute values of the elements is finite
Linear Least Squares
Stochastic Properties: Consistency

Residual least squares estimate

\[ \delta_\theta(v) = \hat{\theta}_{LS}(y) - \theta_0 = (H^T H)^{-1} H^T v \]

If

1. \[ \lim_{N \to \infty} \frac{1}{N} H^T H \text{ is of full rank} \]
2. The correlation of the noise decreases sufficiently fast to zero [mixing condition of order \( P \geq 2 \)]
3. For each column of \( H/N \), the sum of the absolute values of the elements is finite

then \( \hat{\theta}_{LS}(y) \) is consistent

\[ \lim_{N \to \infty} \delta_\theta(v) = 0 \]
If \( \nu \) is mixing of order \( P = \infty \), then

\[
\sqrt{N} \delta_\theta(\nu) = \sqrt{N}(H^T H)^{-1} H^T \nu
\]

is asymptotically \((N \to \infty)\) normally distributed [central limit theorem] \(\Rightarrow\) also \( \hat{\theta}_{LS}(y) = \theta_0 + \delta_\theta(\nu) \)
Outline

- **Linear Least Squares**
  1. Linear Least Squares Estimator
  2. Stochastic Properties
  3. Numerical Stable Calculation
  4. Weighted Linear Least Squares
  5. Noisy Regression Matrix
  6. Bias Compensated Least Squares
  7. (Generalized) Total Least Squares
  8. Instrumental Variables
  9. Regularized Least Squares
  10. Outliers
Linear Least Squares
Numerical Stable Calculation

Singular value decomposition of the regression matrix $H \in \mathbb{R}^{N \times n_\theta}$

\[ H = U \Sigma V^T \quad \Rightarrow \quad H^T H = V \Sigma^2 V^T \]

with $U^T U = I_{n_\theta}$, $V^{-1} = V^T$ and $\Sigma = \text{diag}(\sigma_1(H), \ldots, \sigma_{n_\theta}(H))$

Condition number $\kappa(H)$

\[ \kappa(H) = \frac{\max_k \sigma_k(H)}{\min_k \sigma_k(H)} \quad \Rightarrow \quad \kappa(H^T H) = (\kappa(H))^2 \]
Rule of thumb for solving $Ax = b$ in a numerical reliable way

$$\log_{10} \kappa(A) \ll \text{number of significant digits}$$

$\Rightarrow$ never form the matrix product $H^T H$ explicitly
Linear Least Squares
Numerical Stable Calculation

Singular Value Decomposition \( H = U\Sigma V^T \)

\[ \hat{\theta}_{\text{LS}}(y) = V\Sigma^{-1}U^Ty \]
\[ \hat{y}_{\text{LS}} = UU^Ty \]

QR-factorization \( H = QR \)

\[ \hat{\theta}_{\text{LS}}(y) = R^{-1}Q^Ty \]
\[ \hat{y}_{\text{LS}} = QQ^Ty \]
Parameters mostly have different physical units $\Rightarrow$ scaling necessary to avoid mixing of physical units and to reduce the condition number

\[
y = HS^{-1} \underbrace{S\theta_0}_{\psi_0} + \nu \quad \text{with} \quad S_{[k,k]} = \sqrt{\sum_{l=1}^{N} H_{[l,k]}^2}, \quad k = 1, 2, \ldots, n_\theta
\]

$\Rightarrow$ apply SVD or QR on $HS^{-1}$ and recover $\hat{\theta}_{LS}(y)$ from $\hat{\psi}_{LS}(y)$ as

\[
\hat{\theta}_{LS}(y) = S^{-1}\hat{\psi}_{LS}(y)
\]
Approximation of \( y = \arctan(u) \) over \([0, 5]\), where \( u(k) = 5(k - 1)/(N - 1) \) with \( k = 1, 2, \ldots, N \) and \( N = 200 \), by a polynomial of order \( n_\theta - 1 \) in powers of \( u \) or powers of \( 2u/5 - 1 \)

\[
y(u, \theta) = \sum_{r=1}^{n_\theta} \theta[r] u^{r-1} \quad \Rightarrow \quad H[k,r] = u^{r-1}(k)
\]

\[
y(u, \theta) = \sum_{r=1}^{n_\theta} \theta[r] \left(\frac{2}{5}u - 1\right)^{r-1} \quad \Rightarrow \quad H[k,r] = \left(\frac{2}{5}u(k) - 1\right)^{r-1}
\]
Calculation least squares estimates for polynomial order for \( n_\theta - 1 = 30 \) in three different ways

1. Direct calculation of \( (H^T H)^{-1} H^T y \)
2. Via SVD of the regression matrix \( H \)
3. Via SVD of the scaled regression matrix \( HS^{-1} \)
Linear Least Squares
Numerical Stable Calculation: Example

Importance SVD, choice basis functions, and scaling

\[ y = \arctan(u) \]
## Linear Least Squares

### Numerical Stable Calculation: Example

Condition numbers for powers of $u$

<table>
<thead>
<tr>
<th></th>
<th>$LS_1$</th>
<th>$SVD_1$</th>
<th>$SVD_{sc,1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$3.5 \times 10^{32}$</td>
<td>$1.9 \times 10^{16}$</td>
<td>$1.3 \times 10^{16}$</td>
</tr>
</tbody>
</table>

Condition numbers for powers of $2u/5 - 1$

<table>
<thead>
<tr>
<th></th>
<th>$LS_2$</th>
<th>$SVD_2$</th>
<th>$SVD_{sc,2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$1.2 \times 10^{22}$</td>
<td>$1.1 \times 10^{11}$</td>
<td>$5.8 \times 10^{10}$</td>
</tr>
</tbody>
</table>
Outline

• **Linear Least Squares**
  1. Linear Least Squares Estimator
  2. Stochastic Properties
  3. Numerical Stable Calculation
  4. **Weighted Linear Least Squares**
  5. Noisy Regression Matrix
  6. Bias Compensated Least Squares
  7. (Generalized) Total Least Squares
  8. Instrumental Variables
  9. Regularized Least Squares
  10. Outliers
A symmetric positive definite weighting $W > 0$, $W \in \mathbb{R}^{N \times N}$, can be added to the least squares cost function

$$V_{WLS}(\theta, y) = \frac{1}{2} (y - H\theta)^T W (y - H\theta)$$
Linear Least Squares

Weighted Linear Least Squares

A symmetric positive definite weighting $W > 0$, $W \in \mathbb{R}^{N \times N}$, can be added to the least squares cost function

$$V_{WLS}(\theta, y) = \frac{1}{2} (y - H\theta)^T W (y - H\theta)$$

Minimization w.r.t. $\theta$, gives

$$\hat{\theta}_{WLS}(y) = (H^T WH)^{-1} H^T Wy$$
Linear Least Squares

Weighted Linear Least Squares

A symmetric positive definite weighting \( W > 0, \ W \in \mathbb{R}^{N \times N} \), can be added to the least squares cost function

\[
V_{WLS}(\theta, y) = \frac{1}{2} (y - H\theta)^T W (y - H\theta)
\]

Minimization w.r.t. \( \theta \), gives

\[
\hat{\theta}_{WLS}(y) = (H^T W H)^{-1} H^T W y
\]

with mean and covariance

\[
\mathbb{E}\{\hat{\theta}_{WLS}(y)\} = \theta_0 \\
\text{Cov}(\hat{\theta}_{WLS}(y)) = (H^T W H)^{-1} (H^T W C_v W H) (H^T W H)^{-1}
\]
Unbiased for any $W > 0$ and efficient if and only if $W = \lambda C_v^{-1}$, with $\lambda > 0$

$$\text{Cov}(\hat{\theta}_{WLS}(y)) = (H^T C_v^{-1} H)^{-1}$$

[compare with the Cramér-Rao lower bound]
Linear Least Squares
Weighted Linear Least Squares

Unbiased for any $W > 0$ and efficient if and only if $W = \lambda C_v^{-1}$, with $\lambda > 0$

$$\text{Cov}(\hat{\theta}_{WLS}(y)) = (H^T C_v^{-1} H)^{-1}$$

[compare with the Cramér-Rao lower bound]

Numerical stable implementation via SVD or QR of

1. $W^{1/2} H$
2. $(W^{1/2} H) S^{-1}$

with $W^{1/2}$ a square root of $W$
Outline

• **Linear Least Squares**
  1. Linear Least Squares Estimator
  2. Stochastic Properties
  3. Numerical Stable Calculation
  4. Weighted Linear Least Squares
  5. **Noisy Regression Matrix**
  6. Bias Compensated Least Squares
  7. (Generalized) Total Least Squares
  8. Instrumental Variables
  9. Regularized Least Squares
  10. Outliers
The basic assumption that $H$ is known exactly and independent of the disturbing noise $\nu$ is not always fulfilled.
Linear Least Squares

Noisy Regression Matrix

Two cases can be distinguished

- Case 1: A noisy observation of $H$ of the true unknown regression matrix $H_0$ is available

  \[ y = H_0 \theta_0 + \nu \]
  \[ H = H_0 + \Delta H \text{ with } \mathbb{E}\{\Delta H\} = 0 \]
  \[ \mathbb{E}\{\Delta H^T \nu\} = C_{H\nu} \]
  \[ \mathbb{E}\{\Delta H^T \Delta H\} = C_H \neq 0 \]

- Case 2: $H$ is known exactly, but is correlated with $\nu$

  \[ y = H\theta_0 + \nu \]
  \[ C_{H\nu} = \mathbb{E}\{H^T \nu\} \neq 0 \]
Linear Least Squares
Noisy Regression Matrix: Case 1

\[ \hat{\theta}_{LS}(y) = (H^T H)^{-1}(H^T H_0)\theta_0 + (H^T H)^{-1}(H^T v) \]
is inconsistent because

\[
\underset{N \to \infty}{\text{plim}} \hat{\theta}_{LS}(y) = \underset{N \to \infty}{\lim} (H_0^T H_0 + C_H)^{-1}((H_0^T H_0)\theta_0 + C_{Hv}) \\
\neq \theta_0
\]
Linear Least Squares

Noisy Regression Matrix: Example

Estimation resistor value from DC-measurements

\[ y = [u(1), u(2), \ldots, u(N)]^T \]
\[ H = [i(1), i(2), \ldots, i(N)]^T \]
\[ \theta = R \]
\[ \nu = [n_u(1), n_u(2), \ldots, n_u(N)]^T \]

Since, \( C_{H\nu} = 0 \) and \( C_H = N\sigma_i^2 \), we get

\[
\text{plim}_{N \to \infty} \hat{\theta}_{LS}(y) = \lim_{N \to \infty} \frac{Ni_0^2}{Ni_0^2 + N\sigma_i^2} R_0 = \frac{R_0}{1 + \sigma_i^2 / i_0^2}
\]
The LS estimate

\[ \hat{\theta}_{LS}(y) = \theta_0 + (H^T H)^{-1} (H^T v) \]

is inconsistent because

\[ \lim_{N \to \infty} \hat{\theta}_{LS}(y) = \theta_0 + \lim_{N \to \infty} (E\{H^T H\})^{-1} C_{Hv} \neq \theta_0 \]
Outline

• **Linear Least Squares**
  1. Linear Least Squares Estimator
  2. Stochastic Properties
  3. Numerical Stable Calculation
  4. Weighted Linear Least Squares
  5. Noisy Regression Matrix
  6. **Bias Compensated Least Squares**
  7. (Generalized) Total Least Squares
  8. Instrumental Variables
  9. Regularized Least Squares
  10. Outliers
Linear Least Squares
Bias Compensated Least Squares

Assuming that the second order moments are known, a consistent estimate is obtained as

Case 1: \[ \hat{\theta}_{BCLS}(y) = (H^T H - C_H)^{-1}(H^T y - C_{Hy}) \]

Case 2: \[ \hat{\theta}_{BCLS}(y) = (H^T H)^{-1}(H^T y - C_{Hy}) \]
Linear Least Squares
Bias Compensated Least Squares

Assuming that the second order moments are known, a consistent estimate is obtained as

Case 1: \[ \hat{\theta}_{BCLS}(y) = (H^T H - C_H)^{-1}(H^T y - C_{Hy}) \]

Case 2: \[ \hat{\theta}_{BCLS}(y) = (H^T H)^{-1}(H^T y - C_{Hy}) \]

The bias compensated least squares is weakly consistent

\[ \lim_{N \to \infty} \hat{\theta}_{BCLS}(y) = \theta_0 \]
Linear Least Squares

Bias Compensated Least Squares: Example

Estimation resistor value from DC-measurements

\[ \hat{R}_{\text{BCLS}}(N) = \frac{\sum_{k=1}^{N} i(k)u(k)}{\sum_{k=1}^{N} i^2(k) - N\sigma_i^2} = \frac{1}{N} \sum_{k=1}^{N} \frac{u(k)i(k)}{i^2(k) - \sigma_i^2} \]
Outline

• **Linear Least Squares**
  1. Linear Least Squares Estimator
  2. Stochastic Properties
  3. Numerical Stable Calculation
  4. Weighted Linear Least Squares
  5. Noisy Regression Matrix
  6. Bias Compensated Least Squares
  7. *(Generalized)* Total Least Squares
  8. Instrumental Variables
  9. Regularized Least Squares
  10. Outliers
The total least squares (TLS) handles Case 1 and the equation $y \approx H\theta_0$ is rewritten as

\[
\begin{pmatrix}
H & -y \\
\hat{H} & \hat{\theta}_0
\end{pmatrix}
\begin{pmatrix}
\theta_0 \\
1
\end{pmatrix} \approx 0 \implies \hat{H}\hat{\theta}_0 \approx 0
\]
Linear Least Squares

Total Least Squares

The total least squares (TLS) handles Case 1 and the equation $y \approx H\theta_0$ is rewritten as

$$\begin{bmatrix} H & -y \end{bmatrix} \begin{bmatrix} \theta_0 \\ 1 \end{bmatrix} \approx 0 \Rightarrow \tilde{H}\tilde{\theta}_0 \approx 0$$

Since

$$\mathbb{E}\{\tilde{H}\} = \tilde{H}_0 = \begin{bmatrix} H_0 & -H_0\theta_0 \end{bmatrix} \Rightarrow \tilde{H}_0\tilde{\theta}_0 = 0$$

it is natural to take as estimate the right singular vector corresponding to the smallest singular value of $\tilde{H}$

$$\tilde{H} = \tilde{U}\tilde{\Sigma}\tilde{V}^T \Rightarrow \hat{\theta} = \tilde{V}_{[\cdot, n_\theta+1]} \Rightarrow \hat{\theta}_{\text{TLS}}(y) = \hat{\theta}_{[1:n_\theta]} / \hat{\theta}_{[n_\theta+1]}$$
If the column covariance \( C_{\tilde{H}} = \mathbb{E}\{\Delta \tilde{H}^T \Delta \tilde{H}\} \) of the augmented regression matrix \( \tilde{H} \) satisfies

\[
C_{\tilde{H}} = \lambda I_{n_\theta + 1}
\]

then the total least squares (TLS) estimate is weakly consistent

\[
\lim_{N \to \infty} \hat{\theta}_{TLS}(y) = \theta_0
\]

[proof: see generalized total least squares]
Linear Least Squares

Generalized Total Least squares

If $C_{\tilde{H}} \neq \lambda I_{n_\theta + 1}$, then

1. Calculate a (symmetric) square root $C_{\tilde{H}}^{1/2}$ of $C_{\tilde{H}}$, for example,

$$C_{\tilde{H}} = V_C \Sigma_C V_C^T \Rightarrow C_{\tilde{H}}^{1/2} = V_C \Sigma_C^{1/2} V_C^T$$

2. Rewrite $\tilde{H} \tilde{\theta} \approx 0$ as

$$\tilde{H} C_{\tilde{H}}^{-1/2} C_{\tilde{H}}^{1/2} \tilde{\theta} \approx 0 \Rightarrow \tilde{H} C_{\tilde{H}}^{-1/2} \psi \approx 0$$

3. Calculate the SVD of $\tilde{H} C_{\tilde{H}}^{-1/2}$

$$\tilde{H} C_{\tilde{H}}^{-1/2} = \tilde{U} \tilde{\Sigma} \tilde{V}^T \Rightarrow \hat{\psi} = \tilde{V}[:, n_\theta + 1]$$

$$\Rightarrow \hat{\theta} = C_{\tilde{H}}^{-1/2} \hat{\psi}$$

$$\Rightarrow \hat{\theta}_{\text{GTLS}}(y) = \hat{\theta}[1:n_\theta]/\hat{\theta}[n_\theta + 1]$$
Linear Least Squares
Generalized Total Least squares

The GTLS estimate is weakly consistent $\lim_{N \to \infty} \hat{\theta}_{\text{GTLS}}(y) = \theta_0$

Proof:

- The column covariance of $\tilde{H}_S = \tilde{H}C_{\tilde{H}}^{-1/2}$ equals $I_{n_\theta + 1}$
- Using the SVD $\tilde{H}_{S,0} = \tilde{U}_0\tilde{\Sigma}_0\tilde{V}_0^T$, we find

\[
\mathbb{E}\{\tilde{H}_S^T\tilde{H}_S\} = \tilde{H}_{S,0}^T\tilde{H}_{S,0} + I_{n_\theta + 1} \\
= \tilde{V}_0\tilde{\Sigma}^2_0\tilde{V}_0^T + I_{n_\theta + 1} \\
= \tilde{V}_0(\tilde{\Sigma}^2_0 + I_{n_\theta + 1})\tilde{V}_0^T
\]

- $\Rightarrow$ right singular vectors of $\tilde{H}_S$ convergence weakly to the right singular vectors of $\tilde{H}_{S,0} = \tilde{H}_0C_{\tilde{H}}^{-1/2}$
Linear Least Squares
Generalized Total Least squares

If $\tilde{C}_{\tilde{H}}$ is not of full rank, then the GTLS estimate is calculated via the GSVD of the matrix pair $\tilde{H}, C_{\tilde{H}}^{1/2}$

$$\tilde{H} = U_1 \Sigma_1 X^{-1}$$
$$C_{\tilde{H}}^{1/2} = U_2 \Sigma_2 X^{-1}$$

with $U_1^T U_1 = I_{n_\theta + 1}$, $U_2^T U_2 = I_{n_\theta + 1}$, and $\Sigma_1, \Sigma_2$ diagonal matrices

$$\Rightarrow \hat{\psi} = X[:,k] \Rightarrow \hat{\theta}_{\text{GTLS}}(y) = \hat{\psi}[1:n_\theta] / \hat{\psi}[n_\theta + 1]$$

where $k$ is the index corresponding to the smallest generalized singular value

$$k = \arg \min_l \Sigma_1[l,l] / \Sigma_2[l,l]$$
Linear Least Squares

Generalized Total Least squares: Example

Estimation resistor value from DC-measurements

\[
\tilde{H} = \begin{bmatrix} i & -u \end{bmatrix} \quad \text{and} \quad C_{\tilde{H}}^{1/2} = \sqrt{N} \begin{bmatrix} \sigma_i & 0 \\ 0 & \sigma_u \end{bmatrix} \Rightarrow \hat{\theta}_{\text{GTLS}}(y)
\]
Outline

- **Linear Least Squares**
  1. Linear Least Squares Estimator
  2. Stochastic Properties
  3. Numerical Stable Calculation
  4. Weighted Linear Least Squares
  5. Noisy Regression Matrix
  6. Bias Compensated Least Squares
  7. (Generalized) Total Least Squares
  8. Instrumental Variables
  9. Regularized Least Squares
  10. Outliers
Linear Least Squares
Instrumental Variables

The asymptotic bias of the least squares solution originates from $C_H \neq 0$ (Case 1) and/or $C_{Hv} \neq 0$ (Cases 1 and 2)

Assume that we can construct a matrix $G$ of the same size as $H$ that is uncorrelated with the disturbing noise

$$
\mathbb{E}\{G^T v\} = 0
$$

$$
\mathbb{E}\{G^T \Delta H\} = 0
$$

and correlated with the noiseless part of $H$

$$
\lim_{N \to \infty} \frac{1}{N} \mathbb{E}\{G^T H\} \text{ is of full rank}
$$
Linear Least Squares
Instrumental Variables

Replacing $H^T$ in the LS solution by $G^T$ defines the instrumental variables (IV) estimator

$$\hat{\theta}_{IV}(y) = (G^T H)^{-1} G^T y$$

Hence,

Case 1:  $\hat{\theta}_{IV}(y) = (G^T H)^{-1}(G^T H_0)\theta_0 + (G^T H)^{-1}(G^T v)$

Case 2:  $\hat{\theta}_{IV}(y) = \theta_0 + (G^T H)^{-1}(G^T v)$

is weakly consistent

$$\lim_{N \to \infty} \hat{\theta}_{IV}(y) = \theta_0$$
Estimation resistor value from DC-measurements

*First attempt:* Use the first \( N - 1 \) current and voltage samples as data for \( y \) and \( H \), and the last \( N - 1 \) current samples as instrumental variables for constructing \( G \)

\[
y = [u(1), u(2), \ldots, u(N - 1)]^T
\]
\[
H = [i(1), i(2), \ldots, i(N - 1)]^T
\]
\[
G = [i(2), i(3), \ldots, i(N)]^T
\]

\[
\hat{R}_{IV}(N) = \frac{\sum_{k=1}^{N-1} u(k)i(k + 1)}{\sum_{k=1}^{N-1} i(k)i(k + 1)} = \frac{1}{N-1} \frac{\sum_{k=1}^{N-1} u(k)i(k + 1)}{\sum_{k=1}^{N-1} i(k)i(k + 1)}
\]
Second attempt: Split the current and voltage data records in two equal parts, giving

\[ y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \text{ and } H = \begin{bmatrix} H_1 \\ H_2 \end{bmatrix} \]

\[ y_1 = [u(1), u(2), \ldots, u(N/2)]^T \]

\[ y_2 = [u(N/2 + 1), u(N/2 + 2), \ldots, u(N)]^T \]

\[ H_1 = [i(1), i(2), \ldots, i(N/2)]^T \]

\[ H_2 = [i(N/2 + 1), i(N/2 + 2), \ldots, i(N)]^T \]

Use the current and voltage signals of the second part as instrumental variables for the first part and vice versa

\[ G = \begin{bmatrix} H_2 \\ H_1 \end{bmatrix} \]
\[ \hat{R}_{IV}(N) = \frac{\sum_{k=1}^{N/2} u(k)i(k + N/2) + u(k + N/2)i(k)}{2 \sum_{k=1}^{N/2} i(k)i(k + N/2)} \]
## Linear Least Squares

### BCLS, GTLS, IV: Summary

<table>
<thead>
<tr>
<th></th>
<th>BCLS</th>
<th>GTLS</th>
<th>IV&lt;sup&gt;1&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>consistent</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>formation matrix product</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>knowledge second order moments</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>

<sup>1</sup> choice instrumental variables
Outline

• Linear Least Squares
  1 Linear Least Squares Estimator
  2 Stochastic Properties
  3 Numerical Stable Calculation
  4 Weighted Linear Least Squares
  5 Noisy Regression Matrix
  6 Bias Compensated Least Squares
  7 (Generalized) Total Least Squares
  8 Instrumental Variables
  9 Regularized Least Squares
  10 Outliers
Increasing the model complexity decreases the bias but increases the variance ⇒ add a penalty term for model complexity

\[ V_{RLS}(\theta, y) = \frac{1}{2} (y - H\theta)^T (y - H\theta) + \frac{1}{2} \gamma \theta^T P^{-1} \theta \]

with \( P \) a symmetric positive (semi-)definite matrix – called the regularization matrix – and \( \gamma \geq 0 \)
Linear Least Squares
Regularized Least Squares

Increasing the model complexity decreases the bias but increases the variance ⇒ add a penalty term for model complexity

\[
V_{\text{RLS}}(\theta, y) = \frac{1}{2} (y - H\theta)^T (y - H\theta) + \frac{1}{2} \gamma \theta^T P^{-1} \theta
\]

with \( P \) a symmetric positive (semi-)definite matrix – called the regularization matrix – and \( \gamma \geq 0 \)

Minimization w.r.t. \( \theta \), gives the regularized linear least squares (RLS) estimate

\[
\hat{\theta}_{\text{RLS}}(y) = (PH^T H + \gamma I_{n_\theta})^{-1} PH^T y
\]

Note that \( P \) might be singular
Linear Least Squares
Regularized Least Squares

Basic idea regularization: Introduce a (small) bias giving a decrease in variance such that the resulting mean squared error is smaller.
Linear Least Squares
Regularized Least Squares

Basic idea regularization: Introduce a (small) bias giving a decrease in variance such that the resulting mean squared error is smaller

If $C_v = \sigma_v^2 I_N$, then the choice $\gamma = \sigma_v^2$ and $P = \theta_0\theta_0^T$ minimizes the mean squared error of $\hat{\theta}_{RLS}(y)$.
Linear Least Squares
Regularized Least Squares

Basic idea regularization: Introduce a (small) bias giving a decrease in variance such that the resulting mean squared error is smaller.

If \( C_v = \sigma_v^2 I_N \), then the choice \( \gamma = \sigma_v^2 \) and \( P = \theta_0\theta_0^T \) minimizes the mean squared error of \( \hat{\theta}_{\text{RLS}}(y) \).

If \( C_v = \sigma_v^2 I_N \) and \( P = I_N \), then the choice \( \gamma = \sigma_v^2 \) is called the \( L_2 \) or Tikhonov regularization and it minimizes the mean squared error of \( \hat{\theta}_{\text{RLS}}(y) \).
Linear Least Squares

Regularized Least Squares: Example

Polynomial curve fitting of $y = \arctan(u)$ over $[-2, 4]$ of order $n_\theta - 1 = 14$

$$y(u, \theta) = \sum_{r=1}^{n_\theta} \theta_{[r]} (u - 1)^{r-1}$$

from $N = 17$ noisy samples $y(k), k = 1, 2, \ldots, N$

$$u(k) = 6(k - 1)/(N - 1) - 2 \Rightarrow H_{[k,r]} = (u(k) - 1)^{r-1}$$
Linear Least Squares

Regularized Least Squares: Example

Polynomial curve fitting of \( y = \arctan(u) \) over \([-2, 4]\) of order 
\( n_\theta - 1 = 14 \)

\[
y(u, \theta) = \sum_{r=1}^{n_\theta} \theta[r](u - 1)^{r-1}
\]

from \( N = 17 \) noisy samples \( y(k), k = 1, 2, \ldots, N \)

\[
u(k) = 6(k - 1)/(N - 1) - 2 \quad \Rightarrow \quad H[k,r] = (u(k) - 1)^{r-1}
\]

Three estimates for each of the 1000 Monte-Carlo runs

1. No regularization: \( \gamma = 0 \)
2. Tikhonov regularization: \( P = I_{n_\theta} \) and \( \gamma = \sigma_y^2 \)
3. Optimal regularization: \( P = \theta_0 \theta_0^T \) and \( \gamma = \sigma_y^2 \), where \( \theta_0 \) is obtained from the noiseless case
Linear Least Squares

Regularized Least Squares: Example

\[ \sigma_y = 0 \]

\[ \sigma_y = 0.5 \]

no, Tikhonov, and optimal regularization
Outline

- **Linear Least Squares**
  1. Linear Least Squares Estimator
  2. Stochastic Properties
  3. Numerical Stable Calculation
  4. Weighted Linear Least Squares
  5. Noisy Regression Matrix
  6. Bias Compensated Least Squares
  7. (Generalized) Total Least Squares
  8. Instrumental Variables
  9. Regularized Least Squares
  10. Outliers
Outliers are due to sensor malfunctioning and/or data transmission errors
Outliers are due to sensor malfunctioning and/or data transmission errors

Problem: the WLS cost functions overemphasize outliers resulting in an increased variability of the LS estimates
Outliers are due to sensor malfunctioning and/or data transmission errors

Problem: the WLS cost functions overemphasize outliers resulting in an increased variability of the LS estimates

Best practice: solve the hardware problem and to redo the experiments
Outliers are due to sensor malfunctioning and/or data transmission errors

Problem: the WLS cost functions overemphasize outliers resulting in an increased variability of the LS estimates

Best practice: solve the hardware problem and to redo the experiments

Sometimes not feasible because of time and cost, or because the experiments can not be redone, e.g. in case of historical data.
Linear Least Squares

Outliers

First solution: detect and discard the measurements with outliers from the data
Linear Least Squares

Outliers

First solution: detect and discard the measurements with outliers from the data

Second solution: reduce the sensitivity of the cost function to outliers, for example, using Huber’s robust norm
Polynomial approximation arctan function over $[-2, 4]$, see regularized least squares, with $n_\theta - 1 = 7$, $N = 100$ and $\sigma_y = 0.2$
Polynomial approximation of the arctan function over $[-2, 4]$, see regularized least squares, with $n\theta - 1 = 7$, $N = 100$ and $\sigma_y = 0.2$.

Two data sets
1. Noisy data without outliers
2. Noisy data where on 30% randomly chosen samples an outlier of amplitude 5 and random sign is added
Linear Least Squares
Outliers: Example

Polynomial approximation of the arctan function over $[-2, 4]$, see regularized least squares, with $n_\theta - 1 = 7$, $N = 100$ and $\sigma_y = 0.2$

Two data sets

1. Noisy data without outliers
2. Noisy data where on 30% randomly chosen samples an outlier of amplitude 5 and random sign is added

Three estimates on 400 Monte-Carlo runs

1. Linear least squares on the full data set ($N$ samples)
2. Huber’s estimate on the full data set ($N$ samples)
3. Linear least squares on the data set where samples with outliers are discarded ($0.7N$ samples)
Linear Least Squares

Outliers: Example

Data without outliers

![Graphs showing data comparison between Linear Least Squares (LLS) and Huber methods](image-url)

**RMS error (dB)**
Linear Least Squares

Outliers: Example

Data with outliers

LLS on reduced data set
Outline

• Tools for Analyzing Estimators
• Linear Least Squares
• Nonlinear Least Squares
• Maximum Likelihood Method
• Bayesian Approach
• Neural Networks
• Tuning the Model Complexity
• Nonlinear Least Squares
  1. The Nonlinear Least Squares Estimator
  2. Stochastic Properties
  3. Separable Least Squares
  4. Minimization Cost Function
  5. Generation Starting Values
Outline

- Nonlinear Least Squares
  1. The Nonlinear Least Squares Estimator
  2. Stochastic Properties
  3. Separable Least Squares
  4. Minimization Cost Function
  5. Generation Starting Values
Nonlinear Least Squares

The Nonlinear Least Squares Estimator

Noisy observations $y \in \mathbb{R}^{N \times 1}$ and a model $y_0 = H(\theta_0)$

$$y = H(\theta_0) + \nu$$

where $H(\theta_0) \in \mathbb{R}^{N \times 1}$ is independent of the disturbing noise $\nu \in \mathbb{R}^{N \times 1}$, with $\mathbb{E}\{\nu\} = 0$ and $\text{Cov}(\nu) = C_\nu$. 
Nonlinear Least Squares

The Nonlinear Least Squares Estimator

Noisy observations $y \in \mathbb{R}^{N \times 1}$ and a model $y_0 = H(\theta_0)$

$$y = H(\theta_0) + \nu$$

where $H(\theta_0) \in \mathbb{R}^{N \times 1}$ is independent of the disturbing noise $\nu \in \mathbb{R}^{N \times 1}$, with $\mathbb{E}\{\nu\} = 0$ and $\text{Cov}(\nu) = C_v$.

Nonlinear least squares cost function

$$V_{\text{NLS}}(\theta, y) = \frac{1}{2} (y - H(\theta))^T (y - H(\theta))$$
Nonlinear Least Squares

The Nonlinear Least Squares Estimator

Noisy observations \( y \in \mathbb{R}^{N \times 1} \) and a model \( y_0 = H(\theta_0) \)

\[
y = H(\theta_0) + v
\]

where \( H(\theta_0) \in \mathbb{R}^{N \times 1} \) is independent of the disturbing noise \( v \in \mathbb{R}^{N \times 1} \), with \( \mathbb{E}\{v\} = 0 \) and \( \text{Cov}(v) = C_v \).

Nonlinear least squares cost function

\[
V_{\text{NLS}}(\theta, y) = \frac{1}{2} (y - H(\theta))^T (y - H(\theta))
\]

Minimization w.r.t. \( \theta \) gives the nonlinear least squares estimate

\[
\hat{\theta}_{\text{NLS}}(y) = \arg \min_\theta V_{\text{NLS}}(\theta, y)
\]
Nonlinear Least Squares
The Nonlinear Least Squares Estimator

Following the same lines of the linear least squares, a symmetric positive definite weighting matrix $W \in \mathbb{R}^{N \times N}$ can be added to the cost function

$$V_{WNLS}(\theta, y) = \frac{1}{2} (y - H(\theta))^T W (y - H(\theta))$$

$$\hat{\theta}_{WNLS}(y) = \arg \min_{\theta} V_{WNLS}(\theta, y)$$
Nonlinear Least Squares
The Nonlinear Least Squares Estimator

Following the same lines of the linear least squares, a symmetric positive definite weighting matrix \( W \in \mathbb{R}^{N \times N} \) can be added to the cost function

\[
V_{WNLS}(\theta, y) = \frac{1}{2} (y - H(\theta))^T W (y - H(\theta))
\]

\[
\hat{\theta}_{WNLS}(y) = \arg \min_{\theta} V_{WNLS}(\theta, y)
\]

Goal: choose \( W \) such that

\[
"\text{Cov}(\hat{\theta}_{WNLS}(y))" < "\text{Cov}(\hat{\theta}_{NLS}(y))"
\]
Outline

• **Nonlinear Least Squares**
  1. The Nonlinear Least Squares Estimator
  2. Stochastic Properties
  3. Separable Least Squares
  4. Minimization Cost Function
  5. Generation Starting Values
Nonlinear Least Squares
Stochastic Properties

Technical conditions in a closed and bounded neighborhood of $\theta_0$

1. $H(\theta)$ is a continuous function of $\theta$ with continuous first and second order derivatives
2. Adding data should add information about $\theta$
   \[
   \lim_{N \to \infty} \frac{1}{N} H'(\theta)^T W H'(\theta)
   \]
   is of full rank

   where $H'(\theta)$ is the derivative of $H(\theta)$ w.r.t. $\theta$

3. For each column of the weighting matrix $W$ and the scaled derivative $H'(\theta)/N$, the sum of the absolute values is finite
4. The correlation of the noise over the measurements decreases sufficiently fast to zero (mixing condition of order $P \geq 2$)
Technical condition no. 2 is not fulfilled for the problem

\[ y[k] = Ae^{-\frac{(k-1)T_s}{\tau}} \cos(\omega(k - 1)T_s + \phi) + v[k] \]

where \( \tau > 0 \) and \( k = 1, 2, \ldots, N \)

The reason for this is that the signal part decays exponentially to zero as \( N \to \infty \)

\[ \Rightarrow \text{gathering more data (increasing } N) \text{ will not give additional information about the signal parameters.} \]
Nonlinear Least Squares
Stochastic Properties: Consistency

Step 1: The expected value of the WNLS cost function

\[ V_{WNLS}(\theta) = \frac{1}{2} (H(\theta_0) - H(\theta))^T W (H(\theta_0) - H(\theta)) + \frac{1}{2} \text{trace}(WC_v) \]

is minimal in the true model parameters \( \theta_0 \)
Nonlinear Least Squares
Stochastic Properties: Consistency

Step 1: The expected value of the WNLS cost function

\[ V_{WNLS}(\theta) = \frac{1}{2} (H(\theta_0) - H(\theta))^T W (H(\theta_0) - H(\theta)) + \frac{1}{2} \text{trace}(WC_v) \]

is minimal in the true model parameters \( \theta_0 \)

Step 2: Technical conditions 1, 3 and 4, with \( P = 4 \)

\[ \lim_{N \to \infty} \left( V_{WNLS}(\theta, y) - V_{WNLS}(\theta) \right) / N = 0 \]
Nonlinear Least Squares

Stochastic Properties: Consistency

Step 1: The expected value of the WNLS cost function

\[ V_{WNLS}(\theta) = \frac{1}{2} (H(\theta_0) - H(\theta))^T W (H(\theta_0) - H(\theta)) + \frac{1}{2} \text{trace}(WC_v) \]

is minimal in the true model parameters \( \theta_0 \)

Step 2: Technical conditions 1, 3 and 4, with \( P = 4 \)

\[ \text{l.i.m.}(V_{WNLS}(\theta, y) - V_{WNLS}(\theta)) / N = 0 \]

Step 3: Technical condition 2 + Steps 1 and 2, uniform in \( \theta \),

\[ \text{plim} \frac{\hat{\theta}_{WNLS}(y)}{N \to \infty} = \theta_0 \]
Nonlinear Least Squares

Stochastic Properties: Asymptotic Covariance

Follow the same lines of the estimation of the slope of a straight line (see Slide 65 and further)

\[ \hat{\theta}_{\text{WNLS}}(y) - \theta_0 \xrightarrow{\text{in prob.}} \delta_{\theta,\text{WNLS}}(v) = -(H_0'WH_0')^{-1}H_0'TWv \]

\[ \Rightarrow \text{the asymptotic covariance} \]

\[ "\text{Cov}(\hat{\theta}_{\text{WNLS}}(y))" = \text{Cov}(\delta_{\theta,\text{WNLS}}(v)) = (H_0'WH_0')^{-1}H_0'TWC_vWH_0'(H_0'TWH_0')^{-1} \]
Nonlinear Least Squares
Stochastic Properties: Asymptotic Efficiency

For Gaussian distributed $v$, the pdf of $y$, given $\theta_0$, equals

$$f_{y|\theta_0}(y|\theta_0) = \frac{1}{\sqrt{(2\pi)^N \text{det} C_v}} e^{-\frac{1}{2} (y - H(\theta_0))^T C^{-1}_v (y - H(\theta_0))}$$

with corresponding Fisher information matrix

$$\text{Fi}(\theta_0) = \mathbb{E} \left\{ - \frac{\partial^2 \log f_{z|\theta_0}(z|\theta_0)}{\partial \theta_0^2} \right\} = H_0' C_v^{-1} H_0'$$

$\Rightarrow \hat{\theta}_{WNLS}(y)$ is asymptotically efficient if and only if $W = \lambda C_v^{-1}$
If technical conditions 1–4 are fulfilled with $P = \infty$, then

$$\delta_{\theta, \text{WNLS}}(\nu) = -(H_0' W H_0')^{-1} H_0' W \nu$$

is asymptotically normally distributed (central limit theorem)
If technical conditions 1–4 are fulfilled with $P = \infty$, then

$$
\delta_{\theta, \text{WNLS}}(\nu) = - (H_0' W H_0')^{-1} H_0' W \nu
$$

is asymptotically normally distributed (central limit theorem).

Hence, $\hat{\theta}_{\text{WNLS}}(y)$ is also asymptotically normally distributed because convergence in probability implies convergence in law.
Outline

- Nonlinear Least Squares
  1. The Nonlinear Least Squares Estimator
  2. Stochastic Properties
  3. Separable Least Squares
  4. Minimization Cost Function
  5. Generation Starting Values
Some nonlinear models are linear in a subset of the model parameters

\[ y = H(\theta)C + \nu \]

with \( C \in \mathbb{R}^{n_c \times 1} \), \( H(\theta) \in \mathbb{R}^{N \times n_c} \), \( y \in \mathbb{R}^{N \times 1} \), and \( \nu \in \mathbb{R}^{N \times 1} \) zero mean disturbing noise with covariance \( C_\nu \).
Some nonlinear models are linear in a subset of the model parameters

\[ y = H(\theta)C + \nu \]

with \( C \in \mathbb{R}^{n_c \times 1} \), \( H(\theta) \in \mathbb{R}^{N \times n_c} \), \( y \in \mathbb{R}^{N \times 1} \), and \( \nu \in \mathbb{R}^{N \times 1} \) zero mean disturbing noise with covariance \( C_\nu \)

The corresponding separable nonlinear least squares (SNLS) cost function equals

\[ V_{\text{SNLS}}(\theta, C, y) = \frac{1}{2} (y - H(\theta)C)^T (y - H(\theta)C) \]
Nonlinear Least Squares

Separable Least Squares: Variable Projection Method

Step 1: Minimization $V_{SNLS}(\theta, C, y)$ w.r.t. $C$

$$C = \left( H^T(\theta)H(\theta) \right)^{-1}H^T(\theta)y$$
Nonlinear Least Squares

Separable Least Squares: Variable Projection Method

Step 1: Minimization \( V_{\text{SNLS}}(\theta, C, y) \) w.r.t. \( C \)

\[ C = (H^T(\theta)H(\theta))^{-1}H^T(\theta)y \]

Step 2: Substitution of \( C(\theta) \) in the cost function

\[ V_{\text{SNLS}}(\theta, y) = \frac{1}{2}(P(\theta)y)^T(P(\theta)y) \]

\[ P(\theta) = I_N - H(\theta)(H^T(\theta)H(\theta))^{-1}H^T(\theta) \]

with \( P(\theta) \) a symmetric idempotent projection matrix \( (P^2 = P) \)
Nonlinear Least Squares

Separable Least Squares: Variable Projection Method

Step 1: Minimization \( V_{\text{SNLS}}(\theta, C, y) \) w.r.t. \( C \)

\[
C = (H^T(\theta)H(\theta))^{-1}H^T(\theta)y
\]

Step 2: Substitution of \( C(\theta) \) in the cost function

\[
V_{\text{SNLS}}(\theta, y) = \frac{1}{2}(P(\theta)y)^T(P(\theta)y)
\]

\[
P(\theta) = I_N - H(\theta)(H^T(\theta)H(\theta))^{-1}H^T(\theta)
\]

with \( P(\theta) \) a symmetric idempotent projection matrix \( (P^2 = P) \)

Step 3: Minimization \( V_{\text{SNLS}}(\theta, y) \) w.r.t. \( \theta \) + back substitution to obtain \( C \)
Nonlinear Least Squares

Separable Least Squares: Variable Projection Method

Numerical stable implementation via SVD of $H(\theta)$

$$H(\theta) = U(\theta)\Sigma(\theta)V^T(\theta) \quad \text{and} \quad U^\perp(\theta)$$

$$\Rightarrow P(\theta) = U^\perp(\theta)U^\perp^T(\theta) \quad \text{and}$$

$$V_{SNLS}(\theta, y) = \frac{1}{2}(U^\perp^T(\theta)y)^T(U^\perp^T(\theta)y)$$

[proof: $I_N = [U \ U^\perp][U \ U^\perp]^T \quad \text{and} \quad U^\perp^TU^\perp = I_{N-n_c}$]
Nonlinear Least Squares
Separable Least Squares: Example

Estimation of the sum of harmonically related sinewaves

\[ y[k] = \sum_{l=0}^{L} A_{l,0} \cos\left(l \omega_0 (k - 1) T_s + \phi_{l,0}\right) + v[k] \]
Nonlinear Least Squares

Separable Least Squares: Example

Estimation of the sum of harmonically related sinewaves

\[ y[k] = \sum_{l=0}^{L} A_{l,0} \cos(l\omega_0(k-1)T_s + \phi_{l,0}) + v[k] \]

Using \( \cos(x + y) = \cos(x)\cos(y) - \sin(x)\sin(y) \), we get

\[ H_{[k,l+1]}(\theta) = \begin{cases} 
\cos(l\omega(k-1)T_s) & l = 0, 1, \ldots, L \\
\sin(l\omega(k-1)T_s) & l = L + 1, L + 2, \ldots, 2L 
\end{cases} \]

\[ C_{[l+1]} = \begin{cases} 
A_l \cos \phi_l & l = 0, 1, \ldots, L \\
-A_l \sin \phi_l & l = L + 1, L + 2, \ldots, 2L 
\end{cases} \]
Nonlinear Least Squares

Separable Least Squares: Example

Estimation of the sum of harmonically related sinewaves

\[ y[k] = \sum_{l=0}^{L} A_{l,0} \cos(l\omega_0(k-1)T_s + \phi_{l,0}) + v[k] \]

Using \( \cos(x + y) = \cos(x) \cos(y) - \sin(x) \sin(y) \), we get

\[ H_{[k,l+1]}(\theta) = \begin{cases} 
\cos(l\omega(k-1)T_s) & l = 0, 1, \ldots, L \\
\sin(l\omega(k-1)T_s) & l = L + 1, L + 2, \ldots, 2L 
\end{cases} \]

\[ C_{[l+1]} = \begin{cases} 
A_l \cos \phi_l & l = 0, 1, \ldots, L \\
-A_l \sin \phi_l & l = L + 1, L + 2, \ldots, 2L 
\end{cases} \]

\[ \Rightarrow \text{Reduction of } 2L + 2 \text{ parameters to 1} \]
• **Nonlinear Least Squares**
  1. The Nonlinear Least Squares Estimator
  2. Stochastic Properties
  3. Separable Least Squares
  4. Minimization Cost Function
  5. Generation Starting Values
Nonlinear Least Squares
Minimization Cost Function

Cost functions of the form

\[ V_{WNLS}(\theta, y) = \frac{1}{2} e^T(\theta, y)e(\theta, y) \]
\[ e(\theta, y) = W^{1/2}(y - H(\theta)) \]

where \( W^{1/2} \) is a square root of the positive definite matrix \( W \).
Cost functions of the form

\[ V_{WNLS}(\theta, y) = \frac{1}{2} e^T(\theta, y)e(\theta, y) \]

\[ e(\theta, y) = W^{1/2}(y - H(\theta)) \]

where \( W^{1/2} \) is a square root of the positive definite matrix \( W \)

Minimizing \( V_{WNLS}(\theta, y) \) is equivalent to solving the following set of nonlinear equations

\[ V_{WNLS}'(\theta, y) = 0 \Rightarrow e'(\theta, y)e(\theta, y) = 0 \]
Nonlinear Least Squares
Minimization Cost Function

Cost functions of the form

\[ V_{WNLS}(\theta, y) = \frac{1}{2} e^T(\theta, y)e(\theta, y) \]
\[ e(\theta, y) = W^{1/2}(y - H(\theta)) \]

where \( W^{1/2} \) is a square root of the positive definite matrix \( W \)

Minimizing \( V_{WNLS}(\theta, y) \) is equivalent to solving the following set of nonlinear equations

\[ V_{WNLS}'^T(\theta, y) = 0 \Rightarrow e'^T(\theta, y)e(\theta, y) = 0 \]

This is done via an iterative root finding algorithm
Nonlinear Least Squares
Minimization Cost Function: Newton-Raphson

The $p + 1$-th iteration step of the Newton-Raphson root finding algorithm for $F(\theta) = 0$ is of the form

$$F'(\theta^p) \Delta \theta^{p+1} = -F(\theta^p) \quad \text{with} \quad \Delta \theta^{p+1} = \theta^{p+1} - \theta^p$$
Nonlinear Least Squares

Minimization Cost Function: Newton-Raphson

The $p + 1$-th iteration step of the Newton-Raphson root finding algorithm for $F(\theta) = 0$ is of the form

$$F'(\theta[p]) \Delta \theta[p+1] = -F(\theta[p]) \quad \text{with} \quad \Delta \theta[p+1] = \theta[p+1] - \theta[p]$$

Applying this procedure to $V_{WNLS}^T(\theta, y) = 0$, we find

$$(J^T(\theta[p])J(\theta[p]) + D(\theta[p], y)) \Delta \theta[p+1] = -J^T(\theta[p])e(\theta[p], y)$$

$$J(\theta) = -W^{1/2}H'(\theta)$$

$$D_{[m,n]}(\theta, y) = -\sum_{k,l=1}^{N} W_{[l,k]}^{1/2} \frac{\partial^2 H[k](\theta)}{\partial \theta[m] \partial \theta[n]} e[l](\theta, y)$$

with $J(\theta)$ the Jacobian matrix
Nonlinear Least Squares

Minimization Cost Function: Gauss-Newton

Discard the second order derivative $D(\theta, y)$ in Newton-Raphson

$$J^T(\theta[p]) J(\theta[p]) \Delta\theta[p+1] = -J^T(\theta[p]) e(\theta[p], y)$$
Nonlinear Least Squares

Minimization Cost Function: Gauss-Newton

Discard the second order derivative \( D(\theta, y) \) in Newton-Raphson

\[
J^T(\theta^{[p]}) J(\theta^{[p]}) \Delta \theta^{[p+1]} = -J^T(\theta^{[p]}) e(\theta^{[p]}, y)
\]

Numerical stable implementation via QR or SVD of the Jacobian matrix \( J(\theta) = U(\theta) \Sigma(\theta) V^T(\theta) \)

\[
\Delta \theta^{[p+1]} = -V(\theta^{[p]}) \Sigma^{-1}(\theta^{[p]}) U^T(\theta^{[p]}) e(\theta^{[p]}, y)
\]
Nonlinear Least Squares
Minimization Cost Function: Gauss-Newton

Discard the second order derivative \( D(\theta, y) \) in Newton-Raphson

\[
J^T(\theta^{[p]}) J(\theta^{[p]}) \Delta \theta^{[p+1]} = -J^T(\theta^{[p]}) e(\theta^{[p]}, y)
\]

Numerical stable implementation via QR or SVD of the Jacobian matrix \( J(\theta) = U(\theta) \Sigma(\theta) V^T(\theta) \)

\[
\Delta \theta^{[p+1]} = -V(\theta^{[p]}) \Sigma^{-1}(\theta^{[p]}) U^T(\theta^{[p]}) e(\theta^{[p]}, y)
\]

Good practice: following the same lines of linear least squares, scale each column of \( J(\theta) \) by its 2-norm
Nonlinear Least Squares

Minimization Cost Function: Levenberg-Marquardt

Increase the convergence region of the Gauss-Newton method

\[
(J^T(\theta[p])J(\theta[p]) + \lambda^2 I_{n_\theta}) \Delta \theta[p+1] = -J^T(\theta[p])e(\theta[p], y)
\]
Nonlinear Least Squares
Minimization Cost Function: Levenberg-Marquardt

Increase the convergence region of the Gauss-Newton method

\[
(J^T(\theta[p])J(\theta[p]) + \lambda^2 I_{n_\theta}) \Delta \theta^{[p+1]} = -J^T(\theta[p])e(\theta[p], y)
\]

Numerical stable implementation via the SVD \( J(\theta) \)

\[
\Delta \theta^{[p+1]} = -V(\theta[p])\Lambda(\theta[p])U^T(\theta[p])e(\theta[p], y)
\]

\[
\Lambda(\theta) = \text{diag}(\frac{\sigma_1}{\sigma_1^2 + \lambda^2}, \ldots, \frac{\sigma_{n_\theta}}{\sigma_{n_\theta}^2 + \lambda^2}) \text{ with } \sigma_k = \sigma_k(J(\theta))
\]
Nonlinear Least Squares

Minimization Cost Function: Levenberg-Marquardt

Initialization and updating \( \lambda \):

1. GN steps \((\lambda = 0)\) till the cost increases.
2. Discard \( \theta[^{p+1}] \), and initialize \( \lambda \) as

\[
\lambda = 0.01 \max_{1 \leq k \leq n_\theta} (\sigma_k(J(\theta[^p])))
\]

3. Execute the following steps:
   (a) Calculate the LM update
   (b) If the cost decreases, then \( \lambda \) is decreased

\[
\lambda \rightarrow 0.4\lambda
\]

\( p \rightarrow p + 1 \), and go to step (a).

(c) If the cost increases, then we keep the previous solution \( \theta[^p] \), increase \( \lambda \)

\[
\lambda \rightarrow 10\lambda
\]

and go to step (a).
If $\lambda^2 I_{n_\theta} \gg J^T(\theta) J(\theta)$, then Levenberg-Marquardt simplifies to the gradient descent method

$$\Delta \theta^{[p+1]} = -\frac{1}{\lambda^2} J^T(\theta^{[p]}) e(\theta^{[p]}, y)$$

where $\lambda$ can be initialized and adapted as in Levenberg-Marquardt
Nonlinear Least Squares

Minimization Cost Function: Summary NR, GN, LM and GD

<table>
<thead>
<tr>
<th></th>
<th>NR</th>
<th>GN</th>
<th>LM</th>
<th>GD</th>
</tr>
</thead>
<tbody>
<tr>
<td>conv. rate</td>
<td>quadratic</td>
<td>superlinear</td>
<td>$\geq$ linear</td>
<td>linear</td>
</tr>
<tr>
<td>matrix product</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>guaranteed conv.(^1)</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>

\(^1\) to a local minimum
Nonlinear Least Squares

Minimization Cost Function: Line Search

Basic idea: searching a minimum of the cost function along the direction of the parameter update $\Delta \theta^{[p+1]}$

$$\theta^{[p+1]} = \theta^{[p]} + \gamma \Delta \theta^{[p+1]}$$
Basic idea: searching a minimum of the cost function along the direction of the parameter update $\Delta \theta^{[p+1]}$

$$\theta^{[p+1]} = \theta^{[p]} + \gamma \Delta \theta^{[p+1]}$$

For example, construct a parabola $C(\gamma)$ through

$$C(0) = V_{\text{WNLS}}(\theta^{[p]}, y)$$
$$C(0.5) = V_{\text{WNLS}}(\theta^{[p]} + 0.5 \Delta \theta^{[p+1]}, y)$$
$$C(1) = V_{\text{WNLS}}(\theta^{[p+1]}, y)$$

This parabola is minimal at

$$\gamma_{\text{opt}} = -\frac{1}{2} \frac{4C(0.5) - C(1) - 3C(0)}{-4C(0.5) + 2C(1) + 2C(0)}$$
Nonlinear Least Squares
Minimization Cost Function: Line Search

Local parabolic (green) approximation cost function (black) using three cost function values \( \times \)
Nonlinear Least Squares

Minimization Cost Function: Stopping Criteria

Iterations are stopped if

\[
\frac{\|\theta[p+1] - \theta[p]\|_2}{\|\theta[p]\|_2} \leq \delta
\]

\[
\frac{|V_{\text{WNLS}}(\theta[p+1], y) - V_{\text{WNLS}}(\theta[p], y)|}{V_{\text{WNLS}}(\theta[p], y)} \leq \epsilon
\]

with $\delta$ and $\epsilon$ user defined values, or if a maximum number of iterations is reached.
The iterative algorithms require a starting value (initial estimate) of the model parameters.
Nonlinear Least Squares
Minimization Cost Function: Starting Values

The iterative algorithms require a starting value (initial estimate) of the model parameters

Three approaches

1. Prior knowledge
2. Brute force calculations
   - gridding
   - random initialization
   - random initialization + stochastic mechanism: simulated annealing, evolutionary and genetic algorithms
3. Convex approximation of the problem
Sinewave fitting from $N = 100$ noisy samples

$$y_{[k]} = A_0 + A_1 \cos(\omega_0(k - 1) T_s + \phi_1) + v_{[k]}$$

with $k = 1, 2, \ldots, N$, $f_0 = 0.75f_s/N$, $A_0 = -0.5$, $A_1 = 1$, and $\phi_1 = \pi/4$, and where $v(t)$ is normally distributed with zero mean and standard deviation $\sigma_v = 0.5$. 
Nonlinear Least Squares

Starting Values: Gridding

Noisy samples, true value

SLS cost, × DFT freq.
Nonlinear Least Squares

Starting Values: Convex Approximation

Estimation of the parameters of an exponentially decreasing function from noisy measurements

\[ y(t) = ae^{-t/\tau} + v(t) \]

with \( \tau > 0 \)
Nonlinear Least Squares
Starting Values: Convex Approximation

Estimation of the parameters of an exponentially decreasing function from noisy measurements

\[ y(t) = ae^{-t/\tau} + v(t) \]

with \( \tau > 0 \)

Using the variable projection method we get a nonlinear minimization problem in \( \tau \) \( \Rightarrow \) initial estimate for \( \tau \) needed
Nonlinear Least Squares
Starting Values: Convex Approximation

Estimation of the parameters of an exponentially decreasing function from noisy measurements

\[ y(t) = ae^{-t/\tau} + v(t) \]

with \( \tau > 0 \)

Using the variable projection method we get a nonlinear minimization problem in \( \tau \Rightarrow \) initial estimate for \( \tau \) needed

Taking the logarithm of the absolute value gives a LLS problem in \( \log |a| \) and \( 1/\tau \)

\[ \log |y(t)| \approx \log |a| - \frac{t}{\tau} \]
Nonlinear Least Squares
Starting Values: Convex Approximation

Example: $a = 1$, $\tau = 10 T_s$, $t = k T_s$, $k = 0, 1, \ldots, 99$, and $v(t)$ zero mean Gaussian noise with standard deviation $\sigma_v = 0.5$
Nonlinear Least Squares
Starting Values: Convex Approximation

Other examples:
① Estimation of the slope of a straight line
   (see Introduction)
② Rational approximation of functions
   (see Lecture Notes)
Nonlinear Least Squares
Starting Values: Convex Approximation

Other examples:

1. Estimation of the slope of a straight line
   (see Introduction)
2. Rational approximation of functions
   (see Lecture Notes)

Importance of convex optimization
(book of Stephen Boyd and Lieven Vandenberghe)
Outline

• Tools for Analyzing Estimators
• Linear Least Squares
• Nonlinear Least Squares
• Maximum Likelihood Method
• Bayesian Approach
• Neural Networks
• Tuning the Model Complexity
• Maximum Likelihood Method
  1. The Maximum Likelihood Estimator
  2. Stochastic Properties
  3. Example 1: Sample Mean, Sample Variance
  4. Example 2: Estimation Resistor value
  5. Example 3: Estimation Slope Straight Line
Outline

- Maximum Likelihood Method
  1. The Maximum Likelihood Estimator
  2. Stochastic Properties
  3. Example 1: Sample Mean, Sample Variance
  4. Example 2: Estimation Resistor value
  5. Example 3: Estimation Slope Straight Line
Maximum Likelihood Method

The Maximum Likelihood Estimator

Given noisy measurements $z \in \mathbb{R}^{N \times 1}$

$$z = z_0 + \nu$$

where $z_0$ is the true value and $\nu$ the noise, with known pdf $f_z(z) = f_\nu(z - z_0)$, and given and a model

$$z_0 = h(\theta_0)$$

with $\theta_0 \in \mathbb{R}^{n_\theta \times 1}$ the true model parameters
Maximum Likelihood Method

The Maximum Likelihood Estimator

Given noisy measurements \( z \in \mathbb{R}^{N \times 1} \)

\[
z = z_0 + \nu
\]

where \( z_0 \) is the true value and \( \nu \) the noise, with known pdf \( f_z(z) = f_\nu(z - z_0) \), and given and a model

\[
z_0 = h(\theta_0)
\]

with \( \theta_0 \in \mathbb{R}^{n_\theta \times 1} \) the true model parameters

Next, the pdf of \( z \), given \( \theta_0 \), is constructed as

\[
f_{z|\theta_0}(z|\theta_0) = f_\nu(z - h(\theta_0))
\]
Maximum Likelihood Method
The Maximum Likelihood Estimator

Given noisy measurements \( z \in \mathbb{R}^{N\times1} \)
\[
    z = z_0 + v
\]
where \( z_0 \) is the true value and \( v \) the noise, with known pdf
\[
    f_z(z) = f_v(z - z_0), \text{ and given and a model}
\]
\[
    z_0 = h(\theta_0)
\]
with \( \theta_0 \in \mathbb{R}^{n_\theta\times1} \) the true model parameters

Next, the pdf of \( z \), given \( \theta_0 \), is constructed as
\[
    f_{z|\theta_0}(z|\theta_0) = f_v(z - h(\theta_0))
\]

Finally, the maximum likelihood (ML) estimate is defined as
\[
    \hat{\theta}_{ML}(z) = \arg \max_{\theta} f_{z|\theta}(z|\theta)
\]
Maximum Likelihood Method

The Maximum Likelihood Estimator

Numerical stable implementation

\[ \hat{\theta}_{ML}(z) = \arg \min_{\theta} - \log f_{z|\theta}(z|\theta) \]
Maximum Likelihood Method
The Maximum Likelihood Estimator

Numerical stable implementation

\[ \hat{\theta}_{\text{ML}}(z) = \arg \min_\theta - \log f_{z|\theta}(z|\theta) \]

If the likelihood is differentiable, then \( \hat{\theta}_{\text{ML}}(z) \) is the solution of the following (non)linear set of equations

\[
\left( \frac{\partial - \log f_{z|\theta}(z|\theta)}{\partial \theta} \right)^T = 0
\]
Maximum Likelihood Method
The Maximum Likelihood Estimator

Example non-differentiable likelihood

\[ f_{z|\theta}(z|\theta) \]
Outline

• Maximum Likelihood Method
  1. The Maximum Likelihood Estimator
  2. Stochastic Properties
  3. Example 1: Sample Mean, Sample Variance
  4. Example 2: Estimation Resistor value
  5. Example 3: Estimation Slope Straight Line
Maximum Likelihood Method

Stochastic Properties

Technical conditions that should be fulfilled in a closed and bounded neighborhood of \( \theta_0 \):

1. The measurements are independently distributed
2. The likelihood function is continuous with continuous second order derivative w.r.t. \( \theta \)
3. The Hessian of the log-likelihood divided by \( N \) is of full rank
4. The boundaries of the likelihood are independent of \( \theta \)
5. The log-likelihood of one measurement has finite variance
6. The derivative of the log-likelihood of one measurement has finite moments of order \( 2 + \epsilon \)
7. \( n_\theta \) is independent of \( N \)
Maximum Likelihood Method

Stochastic Properties: Consistency

Step 1: The expected value $V_{ML}(\theta, z) = -\log f_{z|\theta}(z|\theta)$ is minimal in the true model parameters $\theta_0$
Step 1: The expected value \( V_{\text{ML}}(\theta, z) = -\log f_{z|\theta}(z|\theta) \) is minimal in the true model parameters \( \theta_0 \)

\[
V_{\text{ML}}(\theta) = \mathbb{E}\{-\log f_{z|\theta}(z|\theta)\} = -\int_{\mathcal{Z}} \log f_{z|\theta}(z|\theta)f_{z|\theta_0}(z|\theta_0)dz
\]
Maximum Likelihood Method

Stochastic Properties: Consistency

Step 1: The expected value $V_{ML}(\theta, z) = -\log f_{z|\theta}(z|\theta)$ is minimal in the true model parameters $\theta_0$

$$V_{ML}(\theta) = \mathbb{E}\{-\log f_{z|\theta}(z|\theta)\} = -\int_z \log f_{z|\theta}(z|\theta) f_{z|\theta_0}(z|\theta_0) dz$$

Derivative expected value cost function

$$V'_{ML}(\theta) = -\int_z \frac{f'_{z|\theta}(z|\theta)}{f_{z|\theta}(z|\theta)} f_{z|\theta_0}(z|\theta_0) dz$$
Maximum Likelihood Method

Stochastic Properties: Consistency

Step 1: The expected value $V_{ML}(\theta, z) = -\log f_{z|\theta}(z|\theta)$ is minimal in the true model parameters $\theta_0$

$$V_{ML}(\theta) = \mathbb{E}\{-\log f_{z|\theta}(z|\theta)\} = -\int_Z \log f_{z|\theta}(z|\theta)f_{z|\theta_0}(z|\theta_0)dz$$

Derivative expected value cost function

$$V'_{ML}(\theta) = -\int_Z \frac{f'_{z|\theta}(z|\theta)}{f_{z|\theta}(z|\theta)}f_{z|\theta_0}(z|\theta_0)dz$$

Evaluation in the true model parameters

$$V'_{ML}(\theta_0) = -\int_Z f'_{z|\theta_0}(z|\theta_0)dz = -\frac{\partial}{\partial \theta_0} \int_Z f_{z|\theta_0}(z|\theta_0)dz = -\frac{\partial 1}{\partial \theta_0} = 0$$
Maximum Likelihood Method

Stochastic Properties: Consistency

Step 2: Uniform almost sure convergence of the ML cost function
\[ V_{ML}(\theta, z) = -\log f_{z|\theta}(z|\theta) \]
to its expected value \[ V_{ML}(\theta) \]
Maximum Likelihood Method

Stochastic Properties: Consistency

Step 2: Uniform almost sure convergence of the ML cost function
\[ V_{\text{ML}}(\theta, z) = -\log f_{z|\theta}(z|\theta) \] to its expected value \( V_{\text{ML}}(\theta) \)

Independence of the measurements

\[ -\log f_{z|\theta}(z|\theta) = -\log \prod_{k=1}^{N} f_{z[k]|\theta}(z[k]|\theta) = -\sum_{k=1}^{N} \log f_{z[k]|\theta}(z[k]|\theta) \]
Maximum Likelihood Method

Stochastic Properties: Consistency

Step 2: Uniform almost sure convergence of the ML cost function
\[ V_{ML}(\theta, z) = - \log f_{z|\theta}(z|\theta) \] to its expected value \[ V_{ML}(\theta) \]

Independence of the measurements

\[ - \log f_{z|\theta}(z|\theta) = - \log \prod_{k=1}^{N} f_{z[k]|\theta}(z[k]|\theta) = - \sum_{k=1}^{N} \log f_{z[k]|\theta}(z[k]|\theta) \]

Strong law of large numbers for independent random variables with finite variance
Maximum Likelihood Method

Stochastic Properties: Consistency

Step 2: Uniform almost sure convergence of the ML cost function

\[ V_{ML}(\theta, z) = -\log f_{z|\theta}(z|\theta) \]

to its expected value \( V_{ML}(\theta) \)

Independence of the measurements

\[ -\log f_{z|\theta}(z|\theta) = -\log \prod_{k=1}^{N} f_{z[k]|\theta}(z[k]|\theta) = -\sum_{k=1}^{N} \log f_{z[k]|\theta}(z[k]|\theta) \]

Strong law of large numbers for independent random variables with finite variance

Step 3: Steps 1 and 2

\[ \text{a.s.} \lim_{N \to \infty} \hat{\theta}_{ML}(z) = \theta_0 \]
Maximum Likelihood Method

Stochastic Properties: Asymptotic Covariance and Efficiency

Follow the same lines of the estimation of the slope of a straight line (see Slide 65 and further)

\[
\hat{\theta}_{\text{ML}}(z) - \theta_0 \xrightarrow{\text{w.p. } 1} \delta_{\theta,\text{ML}}(z) = - V''_{\text{ML}}(\theta_0) V'_{\text{ML}}(\theta_0, z)
\]

⇒ asymptotic covariance

\[
\text{"Cov}(\hat{\theta}_{\text{ML}}(z))" = \text{Cov}(\delta_{\theta,\text{ML}}(z))
\]

\[
= V''_{\text{ML}}(\theta_0) \mathbb{E}\{ V'_{\text{ML}}(\theta_0, z) V'_{\text{ML}}(\theta_0, z) \} V''_{\text{ML}}^{-1}(\theta_0)
\]

\[
= F_i^{-1}(\theta_0)
\]

⇒ the ML estimate is asymptotically efficient
If technical conditions 1–6 are fulfilled, then

\[ \delta_{\theta, \text{ML}}(z) = -V''_{\text{ML}}^{-1}(\theta_0)V'_{\text{ML}}(\theta_0, z) \]

is asymptotically normally distributed (central limit theorem for independent random variables with finite moments of order $2 + \epsilon$)
Maximum Likelihood Method
Stochastic Properties: Asymptotic Normality

If technical conditions 1–6 are fulfilled, then

$$\delta_{\theta, \text{ML}}(z) = -V_{\text{ML}}''(\theta_0)^{-1}V_{\text{ML}}'(\theta_0, z)$$

is asymptotically normally distributed (central limit theorem for independent random variables with finite moments of order $2 + \epsilon$)

Hence, $\hat{\theta}_{\text{ML}}(z)$ is also asymptotically normally distributed because convergence w.p. 1 implies convergence in law
Maximum Likelihood Method

Stochastic Properties: Robustness

No general robustness statements possible – case specific results
Consider any function $\psi = g(\theta) \in \mathbb{R}^{n_\psi}$ with $n_\psi \leq n_\theta$.
Consider any function $\psi = g(\theta) \in \mathbb{R}^{n_\psi}$ with $n_\psi \leq n_\theta$

If $\hat{\theta}_{ML}$ is the maximum likelihood estimate of $\theta$, then the maximum likelihood estimate $\hat{\psi}_{ML}$ of $\psi$ is given by

$$\hat{\psi}_{ML} = g(\hat{\theta}_{ML})$$
Consider any function $\psi = g(\theta) \in \mathbb{R}^{n_\psi}$ with $n_\psi \leq n_\theta$

If $\hat{\theta}_{ML}$ is the maximum likelihood estimate of $\theta$, then the maximum likelihood estimate $\hat{\psi}_{ML}$ of $\psi$ is given by

$$\hat{\psi}_{ML} = g(\hat{\theta}_{ML})$$

This result follows from the equivalence

$$\max_\theta f_{z|\theta}(z|\theta) = \max_\psi \max_{\theta \text{ s.t. } g(\theta) = \psi} f_{z|\theta}(z|\theta)$$
• Maximum Likelihood Method
  1. The Maximum Likelihood Estimator
  2. Stochastic Properties
  3. Example 1: Sample Mean, Sample Variance
  4. Example 2: Estimation Resistor value
  5. Example 3: Estimation Slope Straight Line
Maximum Likelihood Method
Sample Mean, Sample Variance

Given $N$ independent samples

$$z[k], k = 1, 2, \ldots, N$$

from a normal distribution with unknown mean $\mu_0$ and variance $\sigma_0^2$
Maximum Likelihood Method  
Sample Mean, Sample Variance

Given $N$ independent samples

$$z[k], k = 1, 2, \ldots, N$$

from a normal distribution with unknown mean $\mu_0$ and variance $\sigma_0^2$

The likelihood function is

$$f_{z|\mu,\sigma^2}(z|\mu, \sigma^2) = \prod_{k=1}^{N} f_{z[k]|\mu,\sigma^2}(z[k]|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\sum_{k=1}^{N} \frac{(z[k]-\mu)^2}{2\sigma^2}}$$

with corresponding negative log likelihood

$$V_{ML}(\mu, \sigma^2, z) = \frac{N}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{k=1}^{N} (z[k] - \mu)^2$$
Maximum Likelihood Method
Sample Mean, Sample Variance

The ML estimates \( \hat{\mu}_{ML}(z) \) and \( \hat{\sigma}^{2}_{ML}(z) \) are the solutions of

\[
\frac{\partial V_{ML}(\mu, \sigma^{2}, z)}{\partial \mu} = 0 \\
\frac{\partial V_{ML}(\mu, \sigma^{2}, z)}{\partial \sigma^{2}} = 0
\]
Maximum Likelihood Method
Sample Mean, Sample Variance

The ML estimates $\hat{\mu}_{ML}(z)$ and $\hat{\sigma}^2_{ML}(z)$ are the solutions of

$$\frac{\partial V_{ML}(\mu, \sigma^2, z)}{\partial \mu} = 0$$
$$\frac{\partial V_{ML}(\mu, \sigma^2, z)}{\partial \sigma^2} = 0$$

We find,

$$\hat{\mu}_{ML}(z) = \frac{1}{N} \sum_{k=1}^{N} z[k]$$
$$\hat{\sigma}^2_{ML}(z) = \frac{1}{N} \sum_{k=1}^{N} (z[k] - \hat{\mu}_{ML}(z))^2$$
Maximum Likelihood Method
Sample Mean, Sample Variance: Stochastic Properties

All technical conditions are met
Maximum Likelihood Method
Sample Mean, Sample Variance: Stochastic Properties

All technical conditions are met

⇒ strongly consistent, asymptotic covariance

\[
\text{Cov}(\hat{\theta}_\text{ML}(z)) = \mathcal{V}_\text{ML}^{-1}(\mu_0, \sigma_0^2) = \begin{bmatrix}
\frac{\sigma_0^2}{N} & 0 \\
0 & \frac{2\sigma_0^4}{N}
\end{bmatrix}
\]

asymptotically normally distributed, and asymptotically efficient
Maximum Likelihood Method
Sample Mean, Sample Variance: Stochastic Properties

Robustness

- **Consistency**: still valid for non-Gaussian iid measurements with finite 4th order moments, or non-Gaussian correlated measurements that are mixing of order 4

- **Asymptotic normality**: still valid for non-Gaussian iid measurements with finite 4th order moments, or non-Gaussian correlated measurements that are mixing of order infinity

- **Asymptotic covariance**: for non-Gaussian iid measurements the asymptotic variance of the sample mean remains valid

- **Asymptotic efficiency**: no longer valid for non-Gaussian measurements
Maximum Likelihood Method

Sample Mean, Sample Variance: Stochastic Properties

Expected value ML estimates

\[ \mathbb{E}\{\hat{\mu}_{\text{ML}}(z)\} = \mu_0 \]
\[ \mathbb{E}\{\hat{\sigma}^2_{\text{ML}}(z)\} = \frac{N - 1}{N} \sigma_0^2 \]

Removal finite sample bias

\[ \hat{\sigma}^2(z) = \frac{1}{N - 1} \sum_{k=1}^{N} (z[k] - \hat{\mu}_{\text{ML}}(z))^2 \]
• Maximum Likelihood Method
  1. The Maximum Likelihood Estimator
  2. Stochastic Properties
  3. Example 1: Sample Mean, Sample Variance
  4. Example 2: Estimation Resistor value
  5. Example 3: Estimation Slope Straight Line
Estimation resistor value from independent normally distributed DC current and voltage measurements \(i(k), u(k), k = 1, 2, \ldots, N\)

The corresponding negative log-likelihood equals

\[
V_{\text{ML}}(R, i, z) = \frac{1}{2\sigma^2_u} \sum_{k=1}^{N} (u(k) - Ri)^2 + \frac{1}{2\sigma^2_i} \sum_{k=1}^{N} (i(k) - i)^2 + \text{Cnst}
\]

[see slide 55]
The ML estimates $\hat{R}_{ML}(z)$ and $\hat{i}_{ML}(z)$ are the solutions of

\[
\frac{\partial V_{ML}(R, i, z)}{\partial R} = 0 \quad \frac{\partial V_{ML}(R, i, z)}{\partial i} = 0
\]
Maximum Likelihood Method
Estimation Resistor Value

The ML estimates \( \hat{R}_{ML}(z) \) and \( \hat{i}_{ML}(z) \) are the solutions of

\[
\frac{\partial V_{ML}(R, i, z)}{\partial R} = 0
\]
\[
\frac{\partial V_{ML}(R, i, z)}{\partial i} = 0
\]

We find,

\[
\hat{R}_{ML}(z) = \frac{1}{N} \sum_{k=1}^{N} u(k)
\]
\[
\hat{i}_{ML}(z) = \frac{1}{N} \sum_{k=1}^{N} i(k)
\]
Maximum Likelihood Method

Estimation Resistor Value: Stochastic Properties

All technical conditions are met
Maximum Likelihood Method

Estimation Resistor Value: Stochastic Properties

All technical conditions are met

⇒ strongly consistent, asymptotic covariance

\[ \text{Cov}(\hat{\theta}_{\text{ML}}(z)) = V_{\text{ML}}^{-1}(R_0, i_0) = \frac{1}{N} \begin{bmatrix} \frac{R_0^2 \sigma_i^2 + \sigma_u^2}{i_0^2} & -\frac{R_0 \sigma_i^2}{i_0} \\ -\frac{R_0 \sigma_i^2}{i_0} & \sigma_i^2 \end{bmatrix} \]

asymptotically normally distributed, and asymptotically efficient
Maximum Likelihood Method
Estimation Resistor Value: Stochastic Properties

All technical conditions are met
⇒ strongly consistent, asymptotic covariance

\[
\text{Cov}(\hat{\theta}_{\text{ML}}(z)) = \frac{1}{N} \left[ \begin{array}{ccc}
\frac{R_0^2 \sigma_i^2 + \sigma_u^2}{i_0^2} & - \frac{R_0 \sigma_i^2}{i_0} \\
- \frac{R_0 \sigma_i^2}{i_0} & \sigma_i^2
\end{array} \right]
\]

asymptotically normally distributed, and asymptotically efficient

Robustness: see Slide 59
Outline

• Maximum Likelihood Method
  1 The Maximum Likelihood Estimator
  2 Stochastic Properties
  3 Example 1: Sample Mean, Sample Variance
  4 Example 2: Estimation Resistor value
  5 Example 3: Estimation Slope Straight Line
Maximum Likelihood Method
Estimation Slope Straight Line

Given independent normally distributed measurements of abscissa $u(k)$ and ordinate $y(k)$, $k = 1, 2, \ldots, N$ [see Slide 18]

The $2N + 1$ true model parameters, the abscissas $u_0(k)$, the ordinates $y_0(k)$, $k = 1, 2, \ldots, N$, and the slope $\theta_0$ are related as

$$y_0(k) = \theta_0 u_0(k)$$
Maximum Likelihood Method
Estimation Slope Straight Line

Given independent normally distributed measurements of abscissa \( u(k) \) and ordinate \( y(k) \), \( k = 1, 2, \ldots, N \) [see Slide 18]

The \( 2N + 1 \) true model parameters, the abscissas \( u_0(k) \), the ordinates \( y_0(k) \), \( k = 1, 2, \ldots, N \), and the slope \( \theta_0 \) are related as

\[
y_0(k) = \theta_0 u_0(k)
\]

\( \Rightarrow \) the number of free model parameters equals the total number of model parameters minus the number of constraints giving

\[
2N + 1 - N = N + 1
\]

free model parameters s.t. technical condition no. 7 is not fulfilled
Maximum Likelihood Method
Estimation Slope Straight Line

Independent and normally distributed measurements

\[ f_{z|\theta,u_p}(z|\theta, u_p) = \prod_{k=1}^{N} \frac{1}{2\pi\sigma_y(k)\sigma_u(k)} e^{-\frac{(y(k) - \theta u_p(k))^2}{2\sigma_y^2(k)}} - \frac{(u(k) - u_p(k))^2}{2\sigma_u^2(k)} \]

with \( z \in \mathbb{R}^{2N \times 1} \) the noisy abscissa and ordinate measurements, and \( u_p \in \mathbb{R}^{N \times 1} \) the unknown abscissa values.
Maximum Likelihood Method

Estimation Slope Straight Line

Independent and normally distributed measurements

\[
f_{z|\theta,u_p}(z|\theta, u_p) = \prod_{k=1}^{N} \frac{1}{2\pi \sigma_y(k) \sigma_u(k)} e^{-\frac{(y(k) - \theta u_p(k))^2}{2\sigma_y^2(k)} - \frac{(u(k) - u_p(k))^2}{2\sigma_u^2(k)}}
\]

with \( z \in \mathbb{R}^{2N \times 1} \) the noisy abscissa and ordinate measurements, and \( u_p \in \mathbb{R}^{N \times 1} \) the unknown abscissa values

The corresponding negative log likelihood equals

\[
V_{ML}(\theta, u_p, z) = \sum_{k=1}^{N} \frac{(y(k) - \theta u_p(k))^2}{2\sigma_y^2(k)} + \sum_{k=1}^{N} \frac{(u(k) - u_p(k))^2}{2\sigma_u^2(k)} + \text{Cnst}
\]

\[
= \frac{248}{392}
\]
Maximum Likelihood Method
Estimation Slope Straight Line

The abscissa parameters $u_p$ can analytically be eliminated as

$$\frac{\partial V_{ML}(\theta, u_p, z)}{\partial u_p(k)} = 0 \implies u_p(k) = \frac{\theta \sigma_u^2(k)y(k) + \sigma_y^2(k)u(k)}{\theta^2 \sigma_u^2(k) + \sigma_y^2(k)}$$
Maximum Likelihood Method

Estimation Slope Straight Line

The abscissa parameters $u_p$ can analytically be eliminated as

$$
\frac{\partial V_{ML}(\theta, u_p, z)}{\partial u_p(k)} = 0 \quad \Rightarrow \quad u_p(k) = \frac{\theta \sigma_u^2(k)y(k) + \sigma_y^2(k)u(k)}{\theta^2 \sigma_u^2(k) + \sigma_y^2(k)}
$$

Back substitution in $V_{ML}(\theta, u_p, z)$, gives

$$
V_{ML}(\theta, z) = \frac{1}{2} \sum_{k=1}^{N} \frac{(y(k) - \theta u(k))^2}{\sigma_y^2(k) + \theta^2 \sigma_u^2(k)}
$$
Maximum Likelihood Method

Estimation Slope Straight Line

The abscissa parameters $u_p$ can analytically be eliminated as

$$\frac{\partial V_{ML}(\theta, u_p, z)}{\partial u_p(k)} = 0 \quad \Rightarrow \quad u_p(k) = \frac{\theta \sigma_u^2(k)y(k) + \sigma_y^2(k)u(k)}{\theta^2\sigma_u^2(k) + \sigma_y^2(k)}$$

Back substitution in $V_{ML}(\theta, u_p, z)$, gives

$$V_{ML}(\theta, z) = \frac{1}{2} \sum_{k=1}^{N} \frac{(y(k) - \theta u(k))^2}{\sigma_y^2(k) + \theta^2\sigma_u^2(k)}$$

$\Rightarrow$ the ML estimate of the slope requires a nonlinear optimization

$$\hat{\theta}_{ML}(z) = \arg \min_{\theta} V_{ML}(\theta, z)$$
Maximum Likelihood Method

Estimation Slope Straight Line: Consistency

Step 1: The expected value of the cost function $V_{ML}(\theta)$ is minimal in the true value $\theta_0$ of the slope

$$V_{ML}(\theta) = E\left\{V_{ML}(\theta, z)\right\} = \frac{1}{N^2} \sum_{k=1}^{N} \left( y_0(k) - \theta u_0(k) \right)^2 \sigma^2 y(k) + \theta^2 \sigma^2 u(k)$$
Maximum Likelihood Method
Estimation Slope Straight Line: Consistency

Step 1: The expected value of the cost function $V_{ML}(\theta)$ is minimal in the true value $\theta_0$ of the slope

$$V_{ML}(\theta) = \mathbb{E}\{ V_{ML}(\theta, z) \} = \frac{N}{2} + \frac{1}{2} \sum_{k=1}^{N} \frac{(y_0(k) - \theta u_0(k))^2}{\sigma_y^2(k) + \theta^2 \sigma_u^2(k)}$$
Maximum Likelihood Method
Estimation Slope Straight Line: Consistency

Step 1: The expected value of the cost function $V_{ML}(\theta)$ is minimal in the true value $\theta_0$ of the slope

$$V_{ML}(\theta) = \mathbb{E}\{V_{ML}(\theta, z)\} = \frac{N}{2} + \frac{1}{2} \sum_{k=1}^{N} \frac{(y_0(k) - \theta u_0(k))^2}{\sigma_y^2(k) + \theta^2 \sigma_u^2(k)}$$

Step 2: Uniform strong convergence of $V_{ML}(\theta, z)$ to $V_{ML}(\theta)$
Maximum Likelihood Method

Estimation Slope Straight Line: Consistency

Step 1: The expected value of the cost function $V_{ML}(\theta)$ is minimal in the true value $\theta_0$ of the slope

$$V_{ML}(\theta) = \mathbb{E}\{V_{ML}(\theta, z)\} = \frac{N}{2} + \frac{1}{2} \sum_{k=1}^{N} \frac{(y_0(k) - \theta u_0(k))^2}{\sigma_y^2(k) + \theta^2 \sigma_u^2(k)}$$

Step 2: Uniform strong convergence of $V_{ML}(\theta, z)$ to $V_{ML}(\theta)$

Application of the strong law of large numbers for independent random variables with finite variance
Maximum Likelihood Method
Estimation Slope Straight Line: Consistency

Step 1: The expected value of the cost function $V_{ML}(\theta)$ is minimal in the true value $\theta_0$ of the slope

$$V_{ML}(\theta) = \mathbb{E}\{V_{ML}(\theta, z)\} = \frac{N}{2} + \frac{1}{2} \sum_{k=1}^{N} \frac{(y_0(k) - \theta u_0(k))^2}{\sigma_y^2(k) + \theta^2 \sigma_u^2(k)}$$

Step 2: Uniform strong convergence of $V_{ML}(\theta, z)$ to $V_{ML}(\theta)$
Application of the strong law of large numbers for independent random variables with finite variance

Step 3: Steps 1 and 2 $\Rightarrow$ strongly consistent

$$\text{a. s. lim}_{N \to \infty} \hat{\theta}_{ML}(z) = \theta_0$$
Maximum Likelihood Method

Estimation Slope Straight Line: Consistency

However, ML estimates of the abscissa values

\[ \hat{u}_{ML}(k) = \frac{\hat{\theta}_{ML}(z)\sigma_u^2(k)y(k) + \sigma_y^2(k)u(k)}{\hat{\theta}_{ML}^2(z)\sigma_u^2(k) + \sigma_y^2(k)} \]

are not consistent!
Maximum Likelihood Method

Estimation Slope Straight Line: Consistency

However, ML estimates of the abscissa values

\[ \hat{u}_{ML}(k) = \frac{\hat{\theta}_{ML}(z)\sigma^2_u(k)y(k) + \sigma^2_y(k)u(k)}{\hat{\theta}^2_{ML}(z)\sigma^2_u(k) + \sigma^2_y(k)} \]

are not consistent!

Reason: new measurements give no information about previous abscissa values
The asymptotic variance has been derived on Slide 65 and further

\[
\text{var}(\hat{\theta}_{\text{ML}}(z)) = \text{Cov}(\delta_{\theta_{\text{ML}}}(z)) = V_{\text{ML}}''(\theta_0) \mathbb{E}\{(V'_{\text{ML}}(\theta_0, z))^2\}
\]
Maximum Likelihood Method

Estimation Slope Straight Line: Asymptotic Inefficiency

The Fisher information matrix corresponding to the likelihood function is given by

$$\mathbf{F}_i(\psi_0) = -\frac{\partial^2 \log \mathbb{E}\{f_{z|\theta_0, u_0}(z|\theta_0, u_0)\}}{\partial \psi_0^2}$$

with $$\psi = [\theta, u_p^T]^T$$
Maximum Likelihood Method

Estimation Slope Straight Line: Asymptotic Inefficiency

The Fisher information matrix corresponding to the likelihood function is given by

\[ F_i(\psi_0) = -\frac{\partial^2 \log \mathbb{E}\{f_{z|\theta_0, u_0}(z|\theta_0, u_0)\}}{\partial \psi_0^2} \]

with \( \psi = [\theta, u_p^T]^T \)

Elaborating \( F_i^{-1}(\psi_0) \) using the block matrix inverse lemma, gives

\[ F_i^{-1}(\psi_0) = \begin{bmatrix} F_i^{-1}(\theta_0) & \cdots \\ \vdots & \ddots \end{bmatrix} \]

where \( F_i(\theta_0) = V''_{ML}(\theta_0) \) is the Fisher information of the slope parameter [proof: see lecture notes]
ML estimate of the slope parameter

- **Consistency**: still valid for non-Gaussian iid measurements with finite 4th order moments, or non-Gaussian correlated measurements that are mixing of order 4

- **Asymptotic normality**: still valid for non-Gaussian iid measurements with finite \((4 + \epsilon)\)th order moments, or non-Gaussian correlated measurements that are mixing of order infinity

- **Asymptotic variance**: general expression remains unchanged, but its value will dependent on the pdf of the noise and the correlation
Outline

• Tools for Analyzing Estimators
• Linear Least Squares
• Nonlinear Least Squares
• Maximum Likelihood Method
• Bayesian Approach
• Neural Networks
• Tuning the Model Complexity
Outline

- Bayesian Approach
  1. The Bayesian Estimator
  2. Stochastic Properties
  3. Example: Bayesian Estimation Slope Straight Line
  4. Example: Bayesian Linear Regression
  5. Gaussian Process Modeling
  6. Example: Gaussian Process Linear Regression
  7. Example: Gaussian Process FIR Estimation
• Bayesian Approach

1. The Bayesian Estimator
2. Stochastic Properties
3. Example: Bayesian Estimation Slope Straight Line
4. Example: Bayesian Linear Regression
5. Gaussian Process Modeling
6. Example: Gaussian Process Linear Regression
7. Example: Gaussian Process FIR Estimation
Given

1. noisy data \( z \in \mathbb{R}^{N \times 1} \) with known pdf \( f_z(z) \)
2. prior knowledge about (some of) the model parameters \( \theta \in \mathbb{R}^{n_{\theta}} \) to be estimated under the form of a probability density function (pdf) \( f_\theta(\theta) \)
Bayesian Approach
The Bayesian Estimator

Given

1. noisy data $z \in \mathbb{R}^{N \times 1}$ with known pdf $f_z(z)$
2. prior knowledge about (some of) the model parameters $\theta \in \mathbb{R}^{n_\theta}$ to be estimated under the form of a probability density function (pdf) $f_\theta(\theta)$

construct via Bayes’ rule the (posterior) conditional pdf $f_{\theta|z}(\theta|z)$ of the model parameters $\theta$, given the data $z$,

$$f_{\theta|z}(\theta|z) = \frac{f_{z|\theta}(z|\theta)f_\theta(\theta)}{f_z(z)} = \frac{f_{z,\theta}(z, \theta)}{f_z(z)}$$

where the likelihood $f_{z|\theta}(z|\theta)$ is derived from the pdf $f_z(z)$ and the model [see Slide 205 and further]
Bayesian Approach

The Bayesian Estimator

Different estimators \( \hat{\theta}(z) \) can be defined from \( f_{\theta|z}(\theta|z) \), for example,

- **posterior mean:** \( \hat{\theta}(z) = \mathbb{E}\{\theta|z\} = \int_{-\infty}^{+\infty} \theta f_{\theta|z}(\theta|z) d\theta \)

- **posterior median:** \( \int_{-\infty}^{\hat{\theta}(z)} f_{\theta|z}(\theta|z) d\theta = \frac{1}{2} \)

- **posterior mode:** \( \hat{\theta}(z) = \arg \max_{\theta} f_{\theta|z}(\theta|z) \)

For Gaussian measurements \( z \), and a Gaussian prior on \( \theta \), the three estimators coincide. The posterior mode is called the Bayes estimator, which – using Bayes rule – can be written as

\[
\hat{\theta}_{\text{Bayes}}(z) = \arg \min_{\theta} -\log f_{z,\theta}(z,\theta)
\]
Bayesian Approach

The Bayesian Estimator

Different estimators $\hat{\theta}(z)$ can be defined from $f_{\theta|z}(\theta|z)$, for example,

posterior mean: $\hat{\theta}(z) = \mathbb{E}\{\theta|z\} = \int_{-\infty}^{+\infty} \theta f_{\theta|z}(\theta|z) d\theta$

posterior median: $\int_{-\infty}^{\hat{\theta}(z)} f_{\theta|z}(\theta|z) d\theta = \frac{1}{2}$

posterior mode: $\hat{\theta}(z) = \arg \max_{\theta} f_{\theta|z}(\theta|z)$

For Gaussian measurements $z$, and a Gaussian prior on $\theta$, the three estimators coincide
Bayesian Approach

The Bayesian Estimator

Different estimators $\hat{\theta}(z)$ can be defined from $f_{\theta|z}(\theta|z)$, for example,

posterior mean: $\hat{\theta}(z) = \mathbb{E}\{\theta|z\} = \int_{-\infty}^{+\infty} \theta f_{\theta|z}(\theta|z) d\theta$

posterior median: $\int_{-\infty}^{\hat{\theta}(z)} f_{\theta|z}(\theta|z) d\theta = \frac{1}{2}$

posterior mode: $\hat{\theta}(z) = \arg \max_{\theta} f_{\theta|z}(\theta|z)$

For Gaussian measurements $z$, and a Gaussian prior on $\theta$, the three estimators coincide.

The posterior mode is called the Bayes estimator, which – using Bayes rule – can be written as

$\hat{\theta}_{\text{Bayes}}(z) = \arg \min_{\theta} \log f_{z,\theta}(z, \theta)$
• Bayesian Approach
  1  The Bayesian Estimator
  2  Stochastic Properties
  3  Example: Bayesian Estimation Slope Straight Line
  4  Example: Bayesian Linear Regression
  5  Gaussian Process Modeling
  6  Example: Gaussian Process Linear Regression
  7  Example: Gaussian Process FIR Estimation
The Bayes estimator combines prior knowledge of the model parameters with information from data.
The Bayes estimator combines prior knowledge of the model parameters with information from data.

As the amount of data increases, the data information becomes more important than the prior knowledge.
The Bayes estimator combines prior knowledge of the model parameters with information from data.

As the amount of data increases, the data information becomes more important than the prior knowledge.

Therefore, the Bayes estimator converges for $N \to \infty$ to the maximum likelihood estimator

$$\hat{\theta}_{\text{Bayes}}(z) \xrightarrow{\text{in stoch. sense}} \hat{\theta}_{\text{ML}}(z)$$

and, hence, inherits all its asymptotic properties.
• **Bayesian Approach**
  1. The Bayesian Estimator
  2. Stochastic Properties
  3. Example: Bayesian Estimation Slope Straight Line
  4. Example: Bayesian Linear Regression
  5. Gaussian Process Modeling
  6. Example: Gaussian Process Linear Regression
  7. Example: Gaussian Process FIR Estimation
Bayesian Approach
Bayesian Estimation Slope Straight Line

Consider the estimation the slope of a straight line from normally distributed abscissa and ordinate measurements [see Slide 18]
Consider the estimation the slope of a straight line from normally distributed abscissa and ordinate measurements [see Slide 18]

Assume that the prior distribution of the slope parameter is Gaussian with mean value $\mu_\theta$ and standard deviation $\sigma_\theta$

$$f_\theta(\theta) = \frac{1}{\sqrt{2\pi\sigma_\theta^2}} e^{-\frac{(\theta - \mu_\theta)^2}{2\sigma_\theta^2}}$$
Consider the estimation the slope of a straight line from normally distributed abscissa and ordinate measurements [see Slide 18]

Assume that the prior distribution of the slope parameter is Gaussian with mean value $\mu_\theta$ and standard deviation $\sigma_\theta$

$$f_\theta(\theta) = \frac{1}{\sqrt{2\pi\sigma^2_\theta}} e^{-\frac{(\theta-\mu_\theta)^2}{2\sigma^2_\theta}}$$

No prior knowledge concerning the abscissa values
Bayesian Approach

Bayesian Estimation Slope Straight Line

The corresponding Bayes cost function equals

$$V_{\text{Bayes}}(\theta, u_p, z) = -\log f_{z|\theta,u_p}(z|\theta, u_p) - \log f_{\theta}(\theta)$$

$$= \sum_{k=1}^{N} \left( \frac{(y(k) - \theta u_p(k))^2}{2\sigma_y^2(k)} \right) + \sum_{k=1}^{N} \left( \frac{(u(k) - u_p(k))^2}{2\sigma_u^2(k)} \right)$$

$$+ \left( \frac{(\theta - \mu_\theta)^2}{2\sigma_\theta^2} \right) + C_1$$
Bayesian Approach
Bayesian Estimation Slope Straight Line

The corresponding Bayes cost function equals

\[ V_{\text{Bayes}}(\theta, u_p, z) = -\log f_{z|\theta, u_p}(z|\theta, u_p) - \log f_{\theta}(\theta) \]

\[ = \sum_{k=1}^{N} \frac{(y(k) - \theta u_p(k))^2}{2\sigma^2_y(k)} + \sum_{k=1}^{N} \frac{(u(k) - u_p(k))^2}{2\sigma^2_u(k)} + \frac{(\theta - \mu_{\theta})^2}{2\sigma^2_{\theta}} + C_1 \]

Analytic elimination \( u_p \), gives

\[ V_{\text{Bayes}}(\theta, z) = \frac{1}{2} \sum_{k=1}^{N} \frac{(y(k) - \theta u(k))^2}{\sigma^2_y(k) + \theta^2 \sigma^2_u(k)} + \frac{(\theta - \mu_{\theta})^2}{2\sigma^2_{\theta}} \]
Correct prior mean $\mu_\theta = 1$

$\sigma_\theta = 0.01$, $\sigma_\theta = 0.1$ and $\sigma_\theta = \infty$, 
Wrong prior mean $\mu_\theta = 1.1$

\[\sigma_\theta = 0.01, \sigma_\theta = 0.1 \text{ and } \sigma_\theta = \infty,\]
Bayesian Approach
Bayesian Estimation Slope Straight Line

If $\sigma_u(k) = 0$ and $\sigma_y(k) = \sigma_y$ for $k = 1, 2, \ldots, N$, then the Bayes cost function can be minimized analytically

$$
\hat{\theta}_{\text{Bayes}}(z) = \frac{\sigma_\theta^2 \sum_{k=1}^{N} y(k)u(k) + \sigma_y^2 \mu_\theta}{\sigma_\theta^2 \sum_{k=1}^{N} u^2(k) + \sigma_y^2}
$$
Bayesian Approach

Bayesian Estimation Slope Straight Line

If $\sigma_u(k) = 0$ and $\sigma_y(k) = \sigma_y$ for $k = 1, 2, \ldots, N$, then the Bayes cost function can be minimized analytically

$$\hat{\theta}_{\text{Bayes}}(z) = \frac{\sigma^2 \sum_{k=1}^{N} y(k)u(k) + \sigma^2 \mu_{\theta}}{\sigma^2 \sum_{k=1}^{N} u^2(k) + \sigma_y^2}$$

Three special cases

$$\lim_{\sigma_y \to \infty} \hat{\theta}_{\text{Bayes}}(z) = \lim_{\sigma_{\theta} \to 0} \hat{\theta}_{\text{Bayes}}(z) = \mu_{\theta}$$

$$\lim_{\sigma_{\theta} \to \infty} \hat{\theta}_{\text{Bayes}}(z) = \frac{\sum_{k=1}^{N} y(k)u(k)}{\sum_{k=1}^{N} u^2(k)}$$

$$\hat{\theta}_{\text{Bayes}}(z) \bigg|_{N \gg 1} \approx \frac{\sum_{k=1}^{N} y(k)u(k)}{\sum_{k=1}^{N} u^2(k)}$$
Outline

- Bayesian Approach
  1. The Bayesian Estimator
  2. Stochastic Properties
  3. Example: Bayesian Estimation Slope Straight Line
  4. Example: Bayesian Linear Regression
  5. Gaussian Process Modeling
  6. Example: Gaussian Process Linear Regression
  7. Example: Gaussian Process FIR Estimation
Consider the linear regression problem

\[ y = H\theta + \nu \]

where \( H \in \mathbb{R}^{N \times n_\theta} \) is noise free, and where \( \nu \in \mathbb{R}^N \) is normally distributed with zero mean and covariance \( \sigma^2 I_N \)

\[
f_{\nu}(\nu) = \frac{1}{\sqrt{(2\pi)^N N\sigma^2}} e^{-\frac{\nu^T \nu}{2\sigma^2}}
\]
Bayesian Approach
Bayesian Linear Regression

Consider the linear regression problem

\[ y = H\theta + v \]

where \( H \in \mathbb{R}^{N \times n_\theta} \) is noise free, and where \( v \in \mathbb{R}^N \) is normally distributed with zero mean and covariance \( \sigma^2 I_N \)

\[ f_v(v) = \frac{1}{\sqrt{(2\pi)^N N\sigma^2}} e^{-\frac{v^T v}{2\sigma^2}} \]

A Gaussian prior with zero mean and covariance \( C_\theta \) is imposed on the model parameters \( \theta \)

\[ f_\theta(\theta) = \frac{1}{\sqrt{2\pi} \det C_\theta} e^{-\frac{1}{2} \theta^T C_\theta^{-1} \theta} \]
The data $y$ and the model parameters $\theta$ are jointly Gaussian distributed with mean and covariance

$$\mathbb{E}\left\{\begin{bmatrix} \theta \\ y \end{bmatrix}\right\} = 0 \quad \text{Cov}\left(\begin{bmatrix} \theta \\ y \end{bmatrix}\right) = \begin{bmatrix} C_\theta & C_\theta H^T \\ HC_\theta & HC_\theta H^T + \sigma^2 I_N \end{bmatrix}$$
Bayesian Approach
Bayesian Linear Regression

The data $y$ and the model parameters $\theta$ are jointly Gaussian distributed with mean and covariance

$$E\left\{ \begin{bmatrix} \theta \\ y \end{bmatrix} \right\} = 0 \quad \text{Cov} \left( \begin{bmatrix} \theta \\ y \end{bmatrix} \right) = \begin{bmatrix} C_\theta & C_\theta H^T \\ HC_\theta & HC_\theta H^T + \sigma^2 I_N \end{bmatrix}$$

Two possible approaches

1. Via the joint pdf
   $$\hat{\theta}_{\text{Bayes}}(y) = \arg \min_\theta - \log f_{z,\theta}(y, \theta)$$
   using the block matrix inversion formula

2. Via the posterior pdf
   $$\hat{\theta}(y) = \arg \max_\theta f_{\theta|z}(\theta|y)$$
   where $f_{\theta|y}(\theta|y)$ is derived from $f_{y,\theta}(y, \theta)$ via a property of jointly distributed Gaussian random variables
Bayesian Approach
Bayesian Linear Regression

Let $x_1$ and $x_2$ be jointly Gaussian distributed with mean and covariance

$$\mathbb{E}\left\{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right\} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

$$\text{Cov} \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix}$$

If $\Sigma_{22}$ is regular, then, $x_1$ given $x_2$, denoted as $x_1|x_2$, is normally distributed with mean and covariance

$$\mathbb{E}\{x_1|x_2\} = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \mu_2)$$

$$\text{Cov}(x_1|x_2) = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{12}^T$$

[proof: see lecture notes]
\[
\mathbb{E}\{\theta | y\} = C_\theta H^T (HC_\theta H^T + \sigma^2 I_N)^{-1} y \\
= (C_\theta H^T H + \sigma^2 I_{n_\theta})^{-1} C_\theta H^T y
\]

\[
\text{Cov}(\theta | y) = C_\theta - C_\theta H^T (HC_\theta H^T + \sigma^2 I_N)^{-1} HC_\theta \\
= C_\theta - (C_\theta H^T H + \sigma^2 I_{n_\theta})^{-1} C_\theta H^T HC_\theta
\]
Bayesian Approach
Bayesian Linear Regression

\[ \mathbb{E}\{\theta|y\} = C_\theta H^T (HC_\theta H^T + \sigma^2 I_N)^{-1} y \]
\[ = (C_\theta H^T H + \sigma^2 I_{n_\theta})^{-1} C_\theta H^T y \]
\[ \text{Cov}(\theta|y) = C_\theta - C_\theta H^T (HC_\theta H^T + \sigma^2 I_N)^{-1} HC_\theta \]
\[ = C_\theta - (C_\theta H^T H + \sigma^2 I_{n_\theta})^{-1} C_\theta H^T HC_\theta \]

\[ \Rightarrow \text{Bayes estimate model parameters} \]

\[ \hat{\theta}_{\text{Bayes}}(y) = \mathbb{E}\{\theta|y\} = (C_\theta H^T H + \sigma^2 I_{n_\theta})^{-1} C_\theta H^T y \]
\[ \text{Cov}\left(\hat{\theta}_{\text{Bayes}}(y)\right) = (C_\theta H^T H + \sigma^2 I_{n_\theta})^{-1} C_\theta H^T \sigma^2 I_N \]
\[ HC_\theta (H^T HC_\theta + \sigma^2 I_{n_\theta})^{-1} \]

equals the regularized least squares solution for the choice
\[ P = C_\theta^{-1} \text{ and } \gamma = \sigma^2 \] [see Slide 126]
• Bayesian Approach
  1. The Bayesian Estimator
  2. Stochastic Properties
  3. Example: Bayesian Estimation Slope Straight Line
  4. Example: Bayesian Linear Regression
  5. Gaussian Process Modeling
  6. Example: Gaussian Process Linear Regression
  7. Example: Gaussian Process FIR Estimation
Gaussian process (GP) modeling is a special case of the Bayesian framework where the data $z$ and the prior on the model parameters $\theta$ is Gaussian.
Gaussian process (GP) modeling is a special case of the Bayesian framework where the data $z$ and the prior on the model parameters $\theta$ is Gaussian

$\Rightarrow$ the posterior mean, the posterior median and the posterior mode are all equal
Gaussian process (GP) modeling is a special case of the Bayesian framework where the data $z$ and the prior on the model parameters $\theta$ is Gaussian

$\Rightarrow$ the posterior mean, the posterior median and the posterior mode are all equal

Contrary to the classical Bayesian framework, Gaussian process modeling can handle partial prior knowledge via parametrization of the covariance (and the mean) of the model parameters $\theta$ with some hyper-parameters $\psi$
Bayesian Approach

Gaussian Process Modeling: Estimation Procedure

Step 1: The posterior pdf of $\theta$, $f_{\theta|z}(\theta|z)$, is conditioned on the hyper-parameters $\psi$ describing the prior covariance (and mean) of $\theta$.

$$f_{z,\theta|\psi}(z, \theta|\psi)$$
Bayesian Approach

Gaussian Process Modeling: Estimation Procedure

Step 1: The posterior pdf of $\theta$, $f_{\theta|z}(\theta|z)$, is conditioned on the hyper-parameters $\psi$ describing the prior covariance (and mean) of $\theta$

$$f_{z,\theta|\psi}(z, \theta|\psi)$$

Step 2: The marginal likelihood function $f_{z|\psi}(z|\psi)$ is constructed by marginalizing out the model parameters $\theta$ from the joint pdf

$$f_{z|\psi}(z|\psi) = \int_{\Theta} f_{z,\theta|\psi}(z, \theta|\psi) d\theta$$

where $\Theta$ is the domain of $\theta$
Bayesian Approach

Gaussian Process Modeling: Estimation Procedure

Step 1: The posterior pdf of $\theta$, $f_{\theta|z}(\theta|z)$, is conditioned on the hyper-parameters $\psi$ describing the prior covariance (and mean) of $\theta$

$$f_{z,\theta|\psi}(z, \theta|\psi)$$

Step 2: The marginal likelihood function $f_{z|\psi}(z|\psi)$ is constructed by marginalizing out the model parameters $\theta$ from the joint pdf

$$f_{z|\psi}(z|\psi) = \int_{\Theta} f_{z,\theta|\psi}(z, \theta|\psi) d\theta$$

where $\Theta$ is the domain of $\theta$

Step 3: $\psi$ is tuned by minimizing the negative log marginal likelihood

$$\hat{\psi} = \arg\min_{\psi} - \log f_{z|\psi}(z|\psi)$$
Step 4: the posterior mean is calculated for \( \psi = \hat{\psi} \), giving

\[
\hat{\theta}_{GP}(z) = \mathbb{E}\{\theta|z, \hat{\psi}\}
\]
Bayesian Approach

Gaussian Process Modeling: Estimation Procedure

Step 4: the posterior mean is calculated for $\psi = \hat{\psi}$, giving

$$\hat{\theta}_{GP}(z) = \mathbb{E}\{\theta|z, \hat{\psi}\}$$

Notes:

1. Since (part of) the prior parameter distribution is estimated from the data (Step 3), $\hat{\theta}_{GP}(z)$ is called the *empirical Bayes estimate*.

2. For Gaussian distributions, the marginalization (Step 2) and the posterior mean (Step 4) calculations are very simple.
Outline

• Bayesian Approach
  1 The Bayesian Estimator
  2 Stochastic Properties
  3 Example: Bayesian Estimation Slope Straight Line
  4 Example: Bayesian Linear Regression
  5 Gaussian Process Modeling
  6 Example: Gaussian Process Linear Regression
  7 Example: Gaussian Process FIR Estimation
Bayesian Approach

Gaussian Process Linear Regression

Consider again the Bayesian linear regression problem on Slide 286, and assume that the noise variance \( \sigma^2 \) is unknown and that the prior parameter covariance \( C_\theta \) is parametrized as

\[
C_\theta(\alpha)
\]

with \( \alpha \in \mathbb{R}^{n\alpha \times 1} \)
Consider again the Bayesian linear regression problem on Slide 286, and assume that the noise variance $\sigma^2$ is unknown and that the prior parameter covariance $C_\theta$ is parametrized as

$$C_\theta(\alpha)$$

with $\alpha \in \mathbb{R}^{n_\alpha \times 1}$

The hyper-parameters $\psi$ of the Gaussian process modeling problem are then

$$\psi = \begin{bmatrix} \sigma^2 \\ \alpha \end{bmatrix}$$
Bayesian Approach

Gaussian Process Linear Regression

Using \( y = H\theta + \nu \), the mean and covariance of the data \( y \), given the hyper-parameters \( \psi \), are readily found

\[
\mathbb{E}\{y|\psi\} = 0
\]

\[
\text{Cov}\{y|\psi\} = HC_\theta(\alpha)H^T + \sigma^2 I_N
\]
Bayesian Approach

Gaussian Process Linear Regression

Using $y = H\theta + \nu$, the mean and covariance of the data $y$, given the hyper-parameters $\psi$, are readily found

$$E\{y|\psi\} = 0$$

$$\text{Cov}\{y|\psi\} = HC_\theta(\alpha)H^T + \sigma^2 I_N$$

Hence, $f_{y|\psi}(y|\psi)$ equals

$$-2 \log f_{y|\psi}(y|\psi) = y^T \left( HC_\theta(\alpha)H^T + \sigma^2 I_N \right)^{-1} y$$

$$+ \log \det \left( HC_\theta(\alpha)H^T + \sigma^2 I_N \right) + N \log(2\pi)$$

[this circumvents the marginalization in Step 2]
Bayesian Approach
Gaussian Process Linear Regression

Minimizing

\[-2 \log f_{y|\psi}(y|\psi) = y^T \left( HC_\theta(\alpha)H^T + \sigma^2 I_N \right)^{-1} y\]

\[+ \log \det \left( HC_\theta(\alpha)H^T + \sigma^2 I_N \right) + N \log(2\pi)\]

w.r.t. $\psi$ provides a trade-off between

1. data-fit: first term
2. model complexity: second term
Bayesian Approach
Gaussian Process Linear Regression

Minimizing

\[-2 \log f_{y|\psi}(y|\psi) = y^T \left( HC_\theta(\alpha)H^T + \sigma^2 I_N \right)^{-1} y \]
\[+ \log \det \left( HC_\theta(\alpha)H^T + \sigma^2 I_N \right) + N \log(2\pi) \]

w.r.t. $\psi$ provides a trade-off between

1. data-fit: first term
2. model complexity: second term

Finally, the minimizing argument $\hat{\psi}$ is used to calculate the posterior mean

\[\hat{\theta}_{\text{GP}}(y) = (C_\theta(\hat{\alpha})H^T H + \hat{\sigma}^2 I_{n_\theta})^{-1} C_\theta(\hat{\alpha})H^T y\]
Outline

• Bayesian Approach
  1. The Bayesian Estimator
  2. Stochastic Properties
  3. Example: Bayesian Estimation Slope Straight Line
  4. Example: Bayesian Linear Regression
  5. Gaussian Process Modeling
  6. Example: Gaussian Process Linear Regression
  7. Example: Gaussian Process FIR Estimation
Bayesian Approach
Gaussian Process FIR Estimation

Consider the estimation of the impulse response coefficients $g(t)$ of a discrete-time system

$$y(t) = \sum_{n=0}^{\infty} g(n)u(t - n) + v(t)$$

where $v(t)$ is zero mean white Gaussian noise with variance $\sigma^2$
Bayesian Approach
Gaussian Process FIR Estimation

Consider the estimation of the impulse response coefficients \( g(t) \) of a discrete-time system

\[
y(t) = \sum_{n=0}^{\infty} g(n)u(t - n) + v(t)
\]

where \( v(t) \) is zero mean white Gaussian noise with variance \( \sigma^2 \)

The convolution product is approximated by a finite impulse response (FIR) model

\[
\sum_{r=0}^{R} g(r)u(t - r)
\]
The estimation of $g(r)$, $r = 0, 1, \ldots, R$, from $N$ known input $u(t)$ and $N$ noisy output $y(t)$ samples, $t = 0, 1, \ldots, N - 1$ leads to a linear least squares problem $y = H\theta + ν$, with

$$y = \begin{bmatrix} y(R) & y(R + 1) & \ldots & y(N - 1) \end{bmatrix}^T$$

$$\theta = \begin{bmatrix} g(0) & g(1) & \ldots & g(R) \end{bmatrix}^T$$

$$H_{[k,l]} = u(R + k - l) \quad \text{with} \quad \begin{cases} k = 1, 2, \ldots, N - R \\ l = 1, 2, \ldots, R + 1 \end{cases}$$

$\Rightarrow N \geq 2R + 1$ samples are needed for estimating an FIR model of order $R$. 
Bayesian Approach
Gaussian Process FIR Estimation

The impulse response coefficients $\theta$ are now regarded as a particular realization of a Gaussian process with zero mean value and a parametrized covariance matrix $C_\theta(\alpha)$.
Bayesian Approach
Gaussian Process FIR Estimation

The impulse response coefficients $\theta$ are now regarded as a particular realization of a Gaussian process with zero mean value and a parametrized covariance matrix $C_\theta(\alpha)$

For example, the diagonal and correlated covariance (kernel)

$$C_\theta[k,l] = \begin{cases} c\rho|k-l|\lambda^{(k+l)/2} & k, l = 1, 2, \ldots \\ 0 & k, l = 0, -1, \ldots \end{cases}$$

Role hyper-parameters

1. Stability: $|\lambda| < 1$
2. Smoothness: $|\rho| \leq 1$
3. Prior knowledge versus data: $c$
Bayesian Approach

Gaussian Process FIR Estimation

Second order discrete-time system

\[ G_0(z^{-1}) = \frac{0.58761z^{-1} + 0.53813z^{-2}}{1 - 0.65202z^{-1} + 0.77777z^{-2}} \]

excited by zero mean Gaussian noise with standard deviation one
Bayesian Approach

Gaussian Process FIR Estimation

Second order discrete-time system

\[ G_0(z^{-1}) = \frac{0.58761z^{-1} + 0.53813z^{-2}}{1 - 0.65202z^{-1} + 0.77777z^{-2}} \]

excited by zero mean Gaussian noise with standard deviation one

Output disturbed by zero mean Gaussian noise with \( \sigma_v = 0.1 \)
Bayesian Approach
Gaussian Process FIR Estimation

Second order discrete-time system

\[ G_0(z^{-1}) = \frac{0.58761z^{-1} + 0.53813z^{-2}}{1 - 0.65202z^{-1} + 0.77777z^{-2}} \]

excited by zero mean Gaussian noise with standard deviation one

Output disturbed by zero mean Gaussian noise with \( \sigma_v = 0.1 \)

The LS and GP estimates of the FIR model of order \( R = 50 \) are calculated for two different values of \( N \):

\[ N - R = R + 1 \]
\[ N - R = R + 2 \]
Bayesian Approach
Gaussian Process FIR Estimation

Root mean squared error (RMSE) LS (blue lines) and GP (red lines) estimates over 400 Monte-Carlo runs
Outline

• Tools for Analyzing Estimators
• Linear Least Squares
• Nonlinear Least Squares
• Maximum Likelihood Method
• Bayesian Approach
• Neural Networks
• Tuning the Model Complexity
Outline

- Neural Networks
  1. Neural Network Model
  2. Universal Approximation Property
  3. Training/Learning Neural Networks
  4. Example: Function Approximation
• Neural Networks
  ① Neural Network Model
  ② Universal Approximation Property
  ③ Training/Learning Neural Networks
  ④ Example: Function Approximation
A neural network consists of three basic components:

(i) input and output signal nodes
(ii) artificial neurons
(iii) output neurons
Neural Networks

Neural Network Model: Signal Nodes

Input and output signal nodes

\[ u_n \rightarrow y_m \]
Neural Networks

Neural Network Model: Artificial Neurons

Artificial neuron

\[ z = \phi(b + \sum_{i=1}^{n} w_i x_i) \]
Neural Networks

Neural Network Model: **Artificial Neurons**

**Artificial neuron**

\[ z = \phi(b + \sum_{i=1}^{n} w_i x_i) \]

**Biological neuron**
Neural Networks

Neural Network Model: Artificial Neurons

Activation functions $\phi$ (non-exhaustive list)
Output neuron

\[ y = b + \sum_{i=1}^{n} w_i x_i \]

Note the linear activation function!
Three basic neural network topologies:

(i) feedforward
(ii) recurrent
(iii) convolutional
Neural Networks

Neural Network Model: topologies

Feedforward neural network

Shallow: one hidden layer    Deep: more than one hidden layer
Neural Networks

Neural Network Model: topologies

Output feedback neural network
Neural Networks

Neural Network Model: topologies

State feedback neural network
• **Neural Networks**
  1. Neural Network Model
  2. Universal Approximation Property
  3. Training/Learning Neural Networks
  4. Example: Function Approximation
Any function $y = f(u)$ with continuous derivative over $u \in [u_1, u_2]$ can be approximated arbitrary well over $u \in [u_1, u_2]$ by a feedforward neural network with

(i) Fixed depth, increasing width
(ii) Fixed width, increasing depth

[see lecture notes for the details]
Neural Networks
Universal Approximation Property

Any function $y = f(u)$ with continuous derivative over $u \in [u_1, u_2]$ can be approximated arbitrary well over $u \in [u_1, u_2]$ by a feedforward neural network with

(i) *Fixed depth, increasing width*

(ii) *Fixed width, increasing depth*

[see lecture notes for the details]

Note that this property requires that the activation functions of the output layer are linear
Outline

- Neural Networks
  1. Neural Network Model
  2. Universal Approximation Property
  3. Training/Learning Neural Networks
  4. Example: Function Approximation
Consider a feedforward neural network with $n_u$ inputs and $n_y$ outputs and assume that $N$ input $u(k)$ and output $y(k)$ samples, $k = 1, 2, \ldots, N$ are available.
Neural Networks

Training/Learning: Cost Function

Consider a feedforward neural network with $n_u$ inputs and $n_y$ outputs and assume that $N$ input $u(k)$ and output $y(k)$ samples, $k = 1, 2, \ldots, N$ are available.

The weights (and hyper-parameters) are obtained by minimizing

$$V(y, \theta) = \frac{1}{2} \sum_{k=1}^{N} E^T(k, \theta) E(k, \theta)$$

$$E(k, \theta) = y(k) - \Phi_L(W_L\Phi_{L-1}(W_{L-1} \ldots \Phi_1(W_1u(k))))$$

with

- $Z_l = \Phi_l(V_l) \in \mathbb{R}^{n_l \times 1}$: the activation functions $\phi_l$ of layer $l$
- $W_l \in \mathbb{R}^{n_l \times n_{l-1}}$: the weighting matrix of layer $l$
- $\theta = [\text{vec}(W_L)^T \text{vec}(W_{L-1})^T \ldots \text{vec}(W_1)^T]^T$
Neural Networks
Training/Learning: Glossary

Connection between the terminology in system identification, function regression and machine learning

• to minimize a cost function = to train, to learn
• to validate = to generalize
• model structure = network
• model parameters = weights (and hyper-parameters)
• estimation data = training data
• validation data = generalization data
• overmodeling, overfitting = overtraining, overlearning
• undermodeling = oversmoothing
• step size $1/\lambda^2$ in gradient descent = learning rate
Machine learning specific terms

- *batch size*: Amount of training data entries seen before updating the weights of the network via the stochastic gradient descent method.
Machine learning specific terms

- **batch size**: Amount of training data entries seen before updating the weights of the network via the stochastic gradient descent method
- **epochs**: Parameter defining the number of times the entire training (identification) data set is seen by the learning (minimization) algorithm
Machine learning specific terms

- **batch size**: Amount of training data entries seen before updating the weights of the network via the stochastic gradient descent method
- **epochs**: Parameter defining the number of times the entire training (identification) data set is seen by the learning (minimization) algorithm
- **validation (generalization) data set**: The validation data set is a subset of the data used only to assess the performance of a given (network) model
Machine learning specific terms

- **batch size**: Amount of training data entries seen before updating the weights of the network via the stochastic gradient descent method
- **epochs**: Parameter defining the number of times the entire training (identification) data set is seen by the learning (minimization) algorithm
- **validation (generalization) data set**: The validation data set is a subset of the data used only to assess the performance of a given (network) model
- **test data set**: Data set that is independent of the training (identification) and validation (generalization) data, used to assess the division of the data in training and validation sets
The training/learning (minimization) is performed with the gradient descent algorithm

\[
\Delta \theta[p+1] = - \frac{1}{\lambda^2} \frac{\partial V(y, \theta[p])}{\partial \theta[p]} 
\]

It requires

\[
\frac{\partial V(y, \theta)}{\partial W_l[n,m]} = \sum_{k=1}^{N} \left( \frac{\partial E(k, \theta)}{\partial W_l[n,m]} \right)^T E(k, \theta)
\]

with \( W_l[n,m] \) the weight of the \( m \)-th input of the \( n \)-th node in layer \( l \).
Applying the chain rule of the derivative, we find,

$$\frac{\partial E(k, \theta)}{\partial W_i[n,m]} = - \frac{\partial \Phi_L}{\partial V_L} W_L \frac{\partial \Phi_{L-1}}{\partial V_{L-1}} W_{L-1} \ldots \frac{\partial \Phi_{i+1}}{\partial V_{i+1}} W_{i+1} 1_n \phi'_i(V_i[n]) Z_{i-1}[m]$$

with

- $V_i = W_i Z_{i-1}$ and $\partial \Phi_i / \partial V_i$, $i = 1, 2, \ldots, L$, diagonal matrices
- $1_n$ a vector of zeros except at entry $n$ where it is equal to one
Neural Networks

Training/Learning: Back Propagation

Time-efficient calculation of the derivative of one term in the cost function $V(y, \theta)$

$$
\left( \frac{\partial E(k, \theta)}{\partial W_{l[n,m]}} \right)^T \nabla(k, \theta) = -\phi'_i(V_l[n])Z_{l-1[m]}1_n^T \delta_{l+1}(k)
$$

$$
\delta_{l+1}(k) = W_{l+1}^T \frac{\partial \Phi_{l+1}}{\partial V_{l+1}} \delta_{l+2}(k)
$$

$$
\vdots
$$

$$
\delta_{L-1}(k) = W_{L-1}^T \frac{\partial \Phi_{L-1}}{\partial V_{L-1}} \delta_{L}(k)
$$

$$
\delta_{L}(k) = W_L^T \frac{\partial \Phi_{L}}{\partial V_{L}} E(k, \theta)
$$
Neural Networks
Training/Learning: Stochastic Gradient

For large-scale problems (big data), the computational burden of the gradient descent method combined with the back propagation algorithm can still be too large.
For large-scale problems (big data), the computational burden of the gradient descent method combined with the back propagation algorithm can still be too large.

The stochastic gradient method trades computational burden for a slower convergence rate.
Neural Networks

Training/Learning: Stochastic Gradient

At every iteration, the stochastic gradient descent approximates the gradient on a subset of $M$ randomly selected of samples

$$\frac{\partial V(y, \theta)}{\partial W_{l[n,m]}} \approx \sum_{k=1}^{N} \nu_k \left( \frac{\partial E(k, \theta)}{\partial W_{l[n,m]}} \right)^T E(k, \theta)$$

For example, the $M$ randomly selected samples have the same weight $\nu_k = N/M$, and $\nu_k = 0$ for all other indices

$M$ is called the batch size
Neural Networks

Training/Learning: Stochastic Gradient

At every iteration, the stochastic gradient descent approximates the gradient on a subset of $M$ randomly selected of samples

$$\frac{\partial V(y, \theta)}{\partial W_{l[n,m]}} \approx \sum_{k=1}^{N} \nu_k \left( \frac{\partial E(k, \theta)}{\partial W_{l[n,m]}} \right)^T E(k, \theta)$$

For example, the $M$ randomly selected samples have the same weight $\nu_k = N/M$, and $\nu_k = 0$ for all other indices

$M$ is called the batch size

After $N/M$ iterations all data is – on average – seen by the training algorithm, and this is called an epoch

Accelerated versions exist (work of Yurii Nestorov)
Neural Networks

Training/Learning: Early Stopping

Very often, the number of parameters (weights) of a neural network is (much) larger than the amount of training data.
Neural Networks
Training/Learning: Early Stopping

Very often, the number of parameters (weights) of a neural network is (much) larger than the amount of training data.

Problem: This results in overfitting (overtraining) if the (stochastic) gradient descent minimization of the cost function is executed till a local minimum is reached.

Solution: Evaluate the neural network during training on an independent data set, and stop the minimization as soon as the validation cost increases during a user defined number of epochs.
Initialization of the weights is a critical issue, especially for deep neural networks, and has become a research topic on its own.
Outline

• Neural Networks
  1. Neural Network Model
  2. Universal Approximation Property
  3. Training/Learning Neural Networks
  4. Example: Function Approximation
Neural Networks

Function Approximation: *Universal Approximation Property*

The arctan function \( y(u) = \arctan(u) \) is evaluated at \( N = 400 \) values

\[
u(k) = 10 \frac{k - 1}{N - 1} \quad \text{with} \quad k = 1, 2, \ldots, N
\]

of which \( N/2 \) are used for training and \( N/2 \) for validation

- **Training**: \( u(2k + 1) \) with \( k = 0, 1, \ldots, \frac{N}{2} - 1 \)
- **Validation**: \( u(2k) \) with \( k = 1, 2, \ldots, \frac{N}{2} \)
Neural Networks

Function Approximation: Universal Approximation Property

Two feedforward neural networks with ReLU activation functions are trained on the noiseless data:

1. Two hidden layers with each $m$ nodes $\Rightarrow m^2 + 4m + 1$ weights
2. One hidden layer with $n$ nodes $\Rightarrow 3n + 1$ weights
Two feedforward neural networks with ReLU activation functions are trained on the noiseless data:

1. Two hidden layers with each $m$ nodes $\Rightarrow m^2 + 4m + 1$ weights
2. One hidden layer with $n$ nodes $\Rightarrow 3n + 1$ weights

For $m = 50$ and $n = 900$ both networks have the same number of weights: $2701 \gg N = 400$
Neural Networks
Function Approximation: Universal Approximation Property

Two feedforward neural networks with ReLU activation functions are trained on the noiseless data:

1. Two hidden layers with each $m$ nodes $\Rightarrow m^2 + 4m + 1$ weights
2. One hidden layer with $n$ nodes $\Rightarrow 3n + 1$ weights

For $m = 50$ and $n = 900$ both networks have the same number of weights: $2701 \gg N = 400$

Both networks are trained with the resilient back propagation algorithm of Matlab.
Neural Networks

Function Approximation: **Universal Approximation Property**

\[ y = \arctan(u) \]

- **y** vs **u**
- **Error** vs **u**
- **MSE (dB)** vs **epochs**

*shallow* and *deep* neural network
Neural Networks

Function Approximation: Universal Approximation Property

Root mean squared error over 400 Monte-Carlo runs

shallow and deep neural network
The arctan function $y_0(u) = \arctan(u)$ is evaluated at $N = 20$ values

$$u(k) = 10 \frac{k - 1}{N - 1} \quad \text{with} \quad k = 1, 2, \ldots, N$$
Neural Networks

Function Approximation: Early Stopping

The arctan function $y_0(u) = \arctan(u)$ is evaluated at $N = 20$ values

$$u(k) = 10 \frac{k - 1}{N - 1} \quad \text{with} \quad k = 1, 2, \ldots, N$$

Zero mean Gaussian noise $\nu(k)$ with standard deviation 0.25 is added to the function values

$$y(k) = y_0(k) + \nu(k)$$

for $k = 1, 2, \ldots, N$
The arctan function $y_0(u) = \arctan(u)$ is evaluated at $N = 20$ values

$$u(k) = 10 \frac{k - 1}{N - 1} \quad \text{with} \quad k = 1, 2, \ldots, N$$

Zero mean Gaussian noise $v(k)$ with standard deviation 0.25 is added to the function values

$$y(k) = y_0(k) + v(k)$$

for $k = 1, 2, \ldots, N$

Two independent data sets of $N = 20$ samples each are generated, one for training and one for validation.
These data sets are used for training a shallow (one hidden layer) feedforward neural network with $n = 900$ nodes.
These data sets are used for training a shallow (one hidden layer) feedforward neural network with $n = 900$ nodes.

The number of weights (network parameters) to be estimated (2701) is much larger than the amount of training samples (20).
These data sets are used for training a shallow (one hidden layer) feedforward neural network with \( n = 900 \) nodes.

The number of weights (network parameters) to be estimated (2701) is much larger than the amount of training samples (20).

The whole procedure is repeated for 400 independent random realizations of the disturbing noise and the initialization of the network weights.
Neural Networks

Function Approximation: Early Stopping

Estimates one run with and without early stopping
Root mean squared error over the 400 Monte-Carlo runs with and without early stopping

![Graph showing RMSE error mean over u with and without early stopping]

Error mean value × and its 95% uncertainty bound —— of the estimates with early stopping
Outline

- Tools for Analyzing Estimators
- Linear Least Squares
- Nonlinear Least Squares
- Maximum Likelihood Method
- Bayesian Approach
- Neural Networks
- Tuning the Model Complexity
• Tuning the Model Complexity
  1. Problem Statement
  2. Use of Validation Data
  3. Use of Penalty Terms
  4. Example: Polynomial Function Approximation
Outline

- Tuning the Model Complexity
  1. Problem Statement
  2. Use of Validation Data
  3. Use of Penalty Terms
  4. Example: Polynomial Function Approximation
Tuning the Model Complexity

Problem Statement

Observation:

- model too simple $\Rightarrow$ bias error dominant
- model too complex $\Rightarrow$ variance error dominant
Tuning the Model Complexity

Problem Statement

Observation:
• model too simple ⇒ bias error dominant
• model too complex ⇒ variance error dominant

Goal: select the model with minimal mean squared error
Observation:

- model too simple $\Rightarrow$ bias error dominant
- model too complex $\Rightarrow$ variance error dominant

Goal: select the model with minimal mean squared error

Two approaches:

1. Use of validation data
2. Use of penalty terms
Outline

• Tuning the Model Complexity
  1. Problem Statement
  2. Use of Validation Data
  3. Use of Penalty Terms
  4. Example: Polynomial Function Approximation
Basic idea:

• Estimate the model on an identification (training) data set
• Evaluate the model on an independent (uncorrelated) validation data set
• Select the model that minimizes the validation cost
Tuning the Model Complexity
Use of Validation Data

Ideal case: Two (large) independent (uncorrelated) data sets are available, one for identification (training) and one for validation (generalization)
Tuning the Model Complexity

Use of Validation Data

Ideal case: Two (large) independent (uncorrelated) data sets are available, one for identification (training) and one for validation (generalization)

Practice: The available data is split into an independent (uncorrelated) identification (training) and validation data set
Tuning the Model Complexity
Use of Validation Data

Three approaches:

1. **Split the N data samples:**
   - First 2/3 for identification and last 1/3 for validation
   - Used in system identification and time series analysis
Tuning the Model Complexity

Use of Validation Data

Three approaches:

1. *Split the N data samples*:
   - First 2/3 for identification and last 1/3 for validation
   - Used in system identification and time series analysis

2. *Leave-one-out-cross-validation (LOOCV)*:
   - All N data samples, except one, are used for identification, and this is repeated N times
   - The N estimates can be averaged
   - Appropriate for short data records
Tuning the Model Complexity

Use of Validation Data

Three approaches:

1. **Split the \( N \) data samples:**
   - First 2/3 for identification and last 1/3 for validation
   - Used in system identification and time series analysis

2. **Leave-one-out-cross-validation (LOOCV):**
   - All \( N \) data samples, except one, are used for identification, and this is repeated \( N \) times
   - The \( N \) estimates can be averaged
   - Appropriate for short data records

3. **\( p \)-fold cross-validation (typical 10-fold):**
   - The \( N \) samples are randomly partitioned in \( p \) independent sets of \( N/p \) samples, of which one is used for validation and the \( p - 1 \) others for identification, and this is repeated \( p \) times
   - The \( p \) estimates can be averaged
   - Appropriate for long data records
Tuning the Model Complexity

Use of Validation Data

Notes:

• In a final step the selected model complexity can be used to estimate the model parameters using all data (training + validation)

• In LOOCV and $p$-fold cross-validation each data sample is exactly used once for validation

• The selection of training and validation data sets not straightforward in case of correlated disturbing noise
• Tuning the Model Complexity
  1. Problem Statement
  2. Use of Validation Data
  3. Use of Penalty Terms
  4. Example: Polynomial Function Approximation
Tuning the Model Complexity

Use of penalty Terms

Basic idea:

- Add a penalty term for model complexity to the identification cost function
- Minimize the modified identification cost function over a discrete or continuous set of models
Tuning the Model Complexity

Use of penalty Terms: Discrete Tuning

A cost function of the form

$$V \left( \hat{\theta}(z), z \right) p(n_\theta, N)$$

is minimized over the considered discrete set of models, with

- $\hat{\theta}(z)$ is the minimizer of $V(\theta, z)$
- $n_\theta = \text{dim}(\theta)$
- $N$ the number of data samples
- penalty terms $p(n_\theta, N)$

FPE:

$$p(n_\theta, N) = \frac{N + n_\theta}{N - n_\theta}$$

AIC:

$$p(n_\theta, N) = e^{\frac{2n_\theta}{N}}$$

MDL, BIC:

$$p(n_\theta, N) = e^{n_\theta \frac{\log N}{N}}$$
Tuning the Model Complexity

Use of penalty Terms: Continuous Tuning

Generalized $L_2$-regularization

$$V(\theta, z) + \gamma \theta^T P^{-1} \theta$$

Classical $L_2$-regularization

$$V(\theta, z) + \gamma \|\theta\|_2^2$$

$L_1$ regularization

$$V(\theta, z) + \gamma \|\theta\|_1$$

also called LASSO (least absolute shrinkage and selection operator) or basis pursuit

Elastic net regularization

$$V(\theta, z) + \gamma \left( \rho \|\theta\|_1 + (1 - \rho) \|\theta\|_2^2 \right)$$
Tuning the Model Complexity

Use of penalty Terms: Continuous Tuning

\[ \theta_{[2]} \]

\[ \theta_{[1]} \]

\[ \theta_{[2]} \]

\[ \theta_{[1]} \]

\[ \theta_{[2]} \]

\[ \theta_{[1]} \]
Tuning the Model Complexity

Use of penalty Terms: Summary

Discrete tuning:

- Estimate as many models as defined by the discrete set
- Select the model that minimizes FPE, or AIC or MDL

Continuous tuning:

- Estimation of the model parameters and tuning of the model complexity is performed in one step
• Tuning the Model Complexity
  1. Problem Statement
  2. Use of Validation Data
  3. Use of Penalty Terms
  4. Example: Polynomial Function Approximation
Tuning the Model Complexity

Polynomial Function Approximation

Retake the polynomial approximation of the arctan function [see Slide 131] and calculate for each of the 1000 Monte-Carlo runs

1. The linear least squares estimate for the fixed polynomial order
   \[ n_\theta - 1 = 14 \]

2. The linear least squares estimate that minimizes the AIC criterion
   \[ n_\theta - 1 \in [0, 14] \]

3. The linear least squares estimate that minimizes the MDL criterion over \( n_\theta - 1 \in [0, 14] \)

4. The linear least squares estimate that minimizes the LOOCV cost over \( n_\theta - 1 \in [0, 14] \)

5. The \( L_2 \)-regularized estimate with \( \gamma = \sigma_y^2 \)

6. The generalized \( L_2 \)-regularized estimate with \( \gamma = \sigma_y^2 \) and
   \[ P = \theta_0 \theta_0^T, \] and where \( \theta_0 \) is obtained from a linear least squares fit on the noiseless data
Tuning the Model Complexity
Polynomial Function Approximation

Result for one realization of the 1000 Monte-Carlo runs

MDL, AIC, and LOOCV
Tuning the Model Complexity

Polynomial Function Approximation

Number of times a polynomial order $n_{\theta} - 1$ has been selected (zero for orders $n_{\theta} - 1 \geq 10$)

<table>
<thead>
<tr>
<th>$n_{\theta} - 1$</th>
<th>AIC</th>
<th>MDL</th>
<th>LOOCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>326</td>
<td>494</td>
<td>249</td>
</tr>
<tr>
<td>2</td>
<td>410</td>
<td>380</td>
<td>324</td>
</tr>
<tr>
<td>3</td>
<td>117</td>
<td>75</td>
<td>157</td>
</tr>
<tr>
<td>4</td>
<td>91</td>
<td>45</td>
<td>141</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>4</td>
<td>62</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>2</td>
<td>36</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>0</td>
<td>27</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Tuning the Model Complexity

Polynomial Function Approximation

\[ n_\theta - 1 = 14, \text{ Tikhonov, MDL, AIC, LOOCV, optimal } L_2 \]