Section 10

Advanced Data Analysis

Section 10 Advanced Data Analysis

- ► **Introduction**
- Law of Propagation of Uncertainty (LPU)
- Fitting a Model to Measured Data
- Linear Least Square Method
- Non-linear Least Square Method
- Lab sessions

Definitions

- **Data**
  - "Related items of (chiefly numerical) information considered collectively, typically obtained by scientific work and used for reference, analysis, or calculation". Some common examples include numbers, characters, images and sounds.
- **Information**
  - "Knowledge communicated concerning some particular fact, subject, or event; that of which one is apprised or told; intelligence, news".
- **Knowledge**
  - "That which is known; the sum of what is known".
  - In Computing, Knowledge is "Information in the form of facts, assumptions, and inference rules which can be accessed by a computer program".

(Oxford English Dictionary)

Data Analysis

- Descriptive statistics
  - Numerical values: median, mean, standard deviation, skewness, kurtosis
  - Graphs: histograms, bar charts, line graphs, boxplots
- Inferential statistics
  - Significance testing, t-test, F-test, etc.
  - ANOVA
  - Estimation (point and interval)
  - Regression
  - ...

Random Variables

- **Distributions and probability density functions (PDFs):**
  - *Discrete:* binomial; Poisson…
  - *Continuous:* normal (Gaussian); uniform (rectangular); triangular; chi-squared; Student's t; F; arc-sin; exponential; Cauchy…
- **Moments:**
  - An ordinary moment of kth order is defined as: $m_k=E(X^k)$
  - A central moment of kth order is defined as: $\mu_k=E[(X-E(X))^k]$, where X is the measurement error; E(X) is the expectation or average value of X; k=1, 2, …

Most Commonly Used Moments

- The **first-order ordinary moment**, i.e. **arithmetic average**: $m_1= E(X)$
- The **second-order central moment** is a measure of the **spread** of the measurement errors about the average value: $\mu_2=\sigma^2=V[X]=E[(X-E(X))^2]$.
- The **3rd-order central moment**, μ_3 is related to the asymmetry of the distribution, usually characterised by **skewness,** γ_3, i.e. the ratio of the 3rdorder central moment to the third power of the standard deviation: $\gamma_3=\mu_3/\sigma^3$. For a normal distribution, $\gamma_3= \mu_3=0$.
- The **4th-order central moment**, μ_3, is related to the graduation of the distribution, often characterised by the **kurtosis,** γ_4, i.e. the ratio of the 4th order central moment to the 4th power of the standard deviation: $\gamma_4=\mu_4/\sigma^4$. For a normal distribution, $\gamma_4=3$.

Univariate Distributions

- A continuous univariate distribution is specified by a probability density function (pdf) $f(x) \geq 0$:

$$P(a < X \leq b) = \int_a^b f(x)dx, \qquad P(-\infty < X \leq +\infty) = \int_{-\infty}^{+\infty} f(x)dx = 1.$$

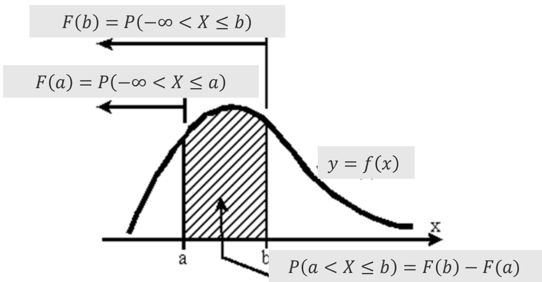
- Mean μ and variance σ^2:

$$\mu = E(X) = \int_{-\infty}^{+\infty} xf(x)dx,$$

$$\sigma^2 = V(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx.$$

- Properties:

$E(a + bX) = a + bE(X),$
$V(a + bX) = b^2V(X).$



Bivariate Distributions

- Bivariate density distribution $f(x, y) \geq 0$

- Covariance

$V(X,Y) = cov(X,Y) = E[(X - E(X))(Y - E(Y))] = \iint (x - \mu_X)(y - \mu_Y)f(x,y)dxdy$

where $\mu_X = E(X) = \iint xf(x,y)dxdy$, $\mu_Y = E(Y) = \iint yf(x,y)dxdy$.

- Measures the strength of linear dependence;
- X and Y are statistically independent if $f(x,y) = f(x)f(y)$;
- $V(X,Y) = 0 \; if \; X \; and \; Y$ are statistically independent.

- Correlation coefficient

$$\rho_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

$-1 \leq \rho_{X,Y} \leq 1.$

Multivariate Distributions

- Given random vectors $\boldsymbol{X} = (X_1, X_2, \dots, X_n)^T, \boldsymbol{Y} = (Y_1, Y_2, \dots Y_m)^T$,
- Covariance matrix

$$V(\boldsymbol{X}) = cov(\boldsymbol{X}, \boldsymbol{X}) = E[(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{X} - E(\boldsymbol{X}))^T]$$

– V(**X**) is the $n \times n$ matrix with $V_{jk} = V(X_j, X_k)$.

- The Correlation matrix is a scaled version of the covariance matrix with $C_{jk} = \frac{V_{jk}}{\sqrt{V_{jj}V_{kk}}}$.
- Cross-covariance matrix

$$V(\boldsymbol{X}, \boldsymbol{Y}) = cov(\boldsymbol{X}, \boldsymbol{Y}) = E[(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T]$$

Section 10 Advanced Data Analysis

Introduction

► **Law of Propagation of Uncertainty (LPU)**

Fitting a Model to Measured Data

Linear Least Square Method

Non-linear Least Square Method

Lab sessions

Law of Propagation of Uncertainty (LPU)

- Given $X=(X_1, X_2, \ldots, X_n)$ with mean $E(X)=\mathbf{x}=(x_1, x_2, \ldots, x_n)$ and variance matrix V
- If Y is a linear combination of X, i.e.:

  $Y=c_1X_1+c_2X_2+\ldots+c_nX_n=\mathbf{c}^TX$

  where $\mathbf{c}=(c_1, c_2, \ldots, c_n)$ are known constants

  then:

  $y=E(Y)=c_1E(X_1)+c_2E(X_2)+\ldots+c_nE(X_n)=c_1x_1+c_2x_2+\ldots+c_nx_n=\mathbf{c}^T\mathbf{x}$

  and

  $u^2(y)=V(Y)=\boldsymbol{c}^TV\boldsymbol{c}$
- If $V(X_i, X_j)=0$, $i\neq j$, then

$$u^2(y)=V(Y)=c_1^2V(X_1)+c_2^2V(X_2)+\cdots+c_n^2V(X_n)=c_1^2u_1^2+c_2^2u_2^2+\cdots+c_n^2u_n^2$$

Examples

- Sum of two independent variables

Assume $E(X_1)=x_1, V(X_1)=u_1^2, E(X_2)=x_2, V(X_2)=u_2^2$,

X_1 and X_2 are independent ➔ $V(X_1, X_2)=0$,

and

$$V_x=V(\boldsymbol{X})=\begin{bmatrix}u_1^2 & 0\\ 0 & u_2^2\end{bmatrix}, Y=\boldsymbol{c}^TX=X_1+X_2=[1\ 1]\begin{bmatrix}X_1\\ X_2\end{bmatrix}.$$

Then

$$E(Y)=\boldsymbol{c}^TE(\boldsymbol{X})=[1\ 1]\begin{bmatrix}x_1\\ x_2\end{bmatrix}=x_1+x_2,$$

and

$$u^2(y)=V(Y)=\boldsymbol{c}^TV_x\boldsymbol{c}=[1\ 1]\begin{bmatrix}u_1^2 & 0\\ 0 & u_2^2\end{bmatrix}\begin{bmatrix}1\\ 1\end{bmatrix}=u_1^2+u_2^2.$$

- Difference of two independent variables

$$Y=X_1-X_2, E(Y)=x_1-x_2, V(Y)=u_1^2+u_2^2.$$

LPU, Multivariate Case

- Given **X**=(X_1, X_2, …, X_n)^T with mean E(**X**)=**x** and variance matrix V
- If **Y** =(Y_1, Y_2, …, Y_m)^T=C**X**,

then

y=E(Y)=CE(X)=C**x**

and

$$u^2(y) = V(Y) = CVC^T$$

Example

- Measurements with a common systematic effect

Given $x_1 = a + \delta + \epsilon_1, x_2 = a + \delta + \epsilon_2$, and

$$E(\delta) = E(\epsilon_1) = E(\epsilon_2) = 0, V(\delta) = \beta^2, V(\epsilon_1) = V(\epsilon_2) = \sigma^2$$

- Let $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} a + \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \delta \end{bmatrix}$
- Assume the effects are independent , so their variance matrix is

$$V = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \beta^2 \end{bmatrix}$$

Applying LPU,

$$E(X) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} a,$$

$$V_X = CVC^T = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \beta^2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} = = \begin{bmatrix} \sigma^2 + \beta^2 & 0 & 0 \\ 0 & \sigma^2 + \beta^2 & 0 \\ 0 & 0 & \beta^2 \end{bmatrix}.$$

LPU, Nonlinear Case

- Assume Y=f(**X**), E(**X**)=**x**, V(**X**)=V_x
- Find the sensitivity coefficient (partial derivative):

  $c_j = \frac{\partial f}{\partial X_j}(\mathbf{x})$

  Then

  $\mathrm{E(Y)} \approx f(\mathbf{x})$,

  $u^2(y) = V(Y) \approx \boldsymbol{c}^T V_x \boldsymbol{c}^T$

- For multivariate case, if **Y** =f(**X**) and $C_{ij} = \frac{\partial f_i}{\partial X_j}$, then

  E(**Y**)≈f(**x**)

  and

  $V_y = V(\boldsymbol{Y}) \approx C V_x C^T$

Exercise 1: Comparison of two gauge blocks

- Suppose two gauge blocks of the same material and nominal length L_0 are measured. Calculate the 2×2 variance matrix V_L associated with their lengths $L = (L_1, L_2)^T$ for data:

  z_1 = 100.000 090 mm, z_2 = 100.000 050 mm,

  $u(z_1) = u(z_2) = u(z)$ = 0.000 050 mm,

  t_1 = 20.5 °C, t_2 = 20.3 °C,

  $u(t_1) = u(t_2) = u(t)$ = 0.1 °C,

  $c = 10 \times 10^{-6}$ m K^{-1}

  $u(c) = 1 \times 10^{-6}$

- Use V_L to calculate the uncertainties associated with $L_1 \pm L_2$.

Exercise 2.1: Mass calculations

- The mass of a cylindrical artefact is given by

$$M = \pi \rho h r^2,$$

where
ρ is the density at 20 °C,
r is the radius,
h is the height.

- Express the uncertainty associated with M in terms of the uncertainties associated with ρ, r and h.

Exercise 2.2: Taking into account temperature

- In Exercise 2.1, suppose r and h are measured at temperature t and that the coefficient of thermal expansion is c. Express the uncertainty associated with M in terms of the uncertainties associated with ρ, r, h, t and c.
- Determine the 2 × 2 variance matrix $V_{\mathbf{M}}$ for the masses M_1 and M_2 of two cylindrical artefacts made from the same material from the following data:

$r_1 = 25.005$ mm is the radius measured at temperature $t_1 = 20.5$ºC,
$h_1 = 63.510$ mm is the height measured at temperature t_1,
$r_2 = 25.020$ mm is the radius measured at temperature $t_2 = 20.8$ºC,
$h_2 = 63.515$ mm is the height measured at temperature t_2,
$\rho = 8000$ kg m^{-3},
$c = 10 \times 10^{-6}$ m K^{-1},
$u(r_1) = u(r_2) = u(h_1) = u(h_2) = 0.001$ mm,
$u(r_1) = u(r_2) = 0.1$ ºC,
$u(\rho) = 0.05$ kg m^{-3},
u(c) = 1×10^{-6} K^{-1}.

- Use V_M to evaluate the uncertainties associated with M1 ± M2.

Section 10 Advanced Data Analysis

Introduction

Law of Propagation of Uncertainty (LPU)

► **Fitting a Model to Measured Data**

Linear Least Square Method

Non-linear Least Square Method

Lab sessions

Functional Models

$$y = \varphi(\mathbf{x}),$$
$$\mathbf{x} = (x_1, x_2, \dots, x_n)^T$$

where

y is the response (or dependent) variable.

x are covariates (or independent variables).

φ is the function describing how y varies with (depends on) **x**.

Typically, values of **x** are controlled or known accurately, and the response y is measured subject to measurement uncertainty.

Functional Model: example 1

- The length of a gauge block depends on temperature:

$$y = \varphi(T, L, c) = L(1 + c(T - T_0))$$

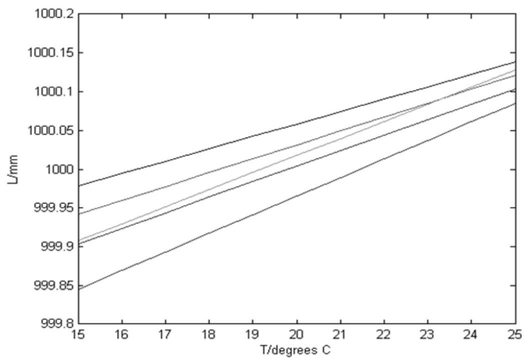
- In generic notation,

$$y = \varphi(x, \boldsymbol{a}) = a_1 + a_2(x - T_0))$$
$$\boldsymbol{a} = (a_1, a_2)^T, a_1 = L, a_2 = Lc.$$

- The parameters **a** are used to define the class (or *space*) of models that could describe the dependence of length on temperature.

Possible Behaviours

$$y = a_1 + a_2(x - T_0))$$

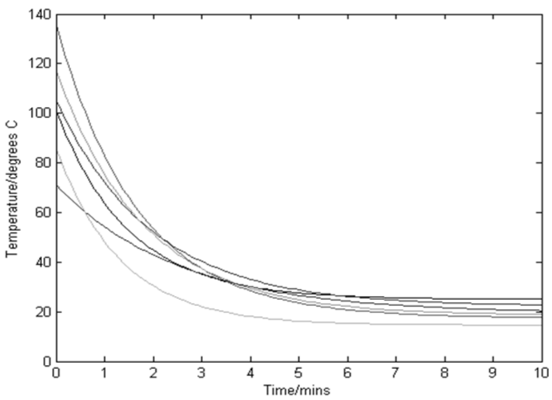


Functional Model: example 2

- Newton's law of cooling:
$$y = \varphi(t, T_0, T_b, k) = T_b + (T_0 - T_b)e^{-kt}$$
- Functional model involves three parameters T_b, T_0, k
- In generic notation,
$$y = \varphi(x, \boldsymbol{a}) = a_1 + (a_2 - a_1)e^{-a_3 x},$$
$$\boldsymbol{a} = (a_1, a_2, a_3)^T$$
- The parameters $\boldsymbol{a}$ are used to define the class (or *space*) of models that could describe the dependence of temperature on time.

Possible Behaviours

$$y = a_1 + (a_2 - a_1)e^{-a_3 x}$$



Mechanistic and Empirical Models

- Mechanistic models are based on the underlying physics governing the behaviour of the process, e.g.

$$T = T_b + (T_0 - T_b)e^{-kt}$$

  - They explain behaviour.
  - Parameters have a physical meaning.

- Empirical models are based on experience or data, e.g. Polynomials $\varphi(x, \boldsymbol{a}) = \sum_i \boldsymbol{a_i x^i}$ and Fourier series
  - They describe (summarise) behaviour.
  - Parameters may have no intrinsic meaning.

What is a Linear Model?

- A linear model is linear in the parameters $\boldsymbol{a}$.
- In practice, a linear model is usually defined as a linear combination of *basis functions*

$$\boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{a}) = a_1\varphi_1(\boldsymbol{x}) + a_2\varphi_2(\boldsymbol{x}) + \ldots + a_n\varphi_n(\boldsymbol{x})$$

- For a linear model, $\frac{\partial\varphi}{\partial a_i}$ is independent of $\boldsymbol{a}$
- Examples
  - polynomials

$$\varphi(x, \boldsymbol{a}) = a_0 + a_1x + a_2x^2 + \ldots + a_nx^n,$$

  - Fourier series

$$\varphi_0(x) = 1, \varphi_{2i-1} = \frac{cos2\pi ix}{L}, \varphi_{2i} = \frac{sin2\pi ix}{L}$$

Nonlinear Model

- For a nonlinear model, $\frac{\partial \varphi}{\partial a_i}$ is **not** independent of $\boldsymbol{a}$
- Examples: Newton's law of cooling

$$\varphi(t, T_0, T_b, k) = T_b + (T_0 - T_b)e^{-kt}$$

In this case, $\mathbf{a} = (\mathrm{T_b}, \mathrm{T_0}, k)^T$, $\frac{\partial \varphi}{\partial a_i}$ are:

$$\frac{\partial \varphi}{\partial T_b} = 1 - e^{-kt}, \frac{\partial \varphi}{\partial T_0} = e^{-kt}, \frac{\partial \varphi}{\partial k} = -t(T_0 - T_b)e^{-kt}$$

Fitting a Model to Measured Data

- Given data (x_i, y_i), i=1, 2,…,m and functional model $y = \varphi(\boldsymbol{x}, \boldsymbol{a})$
- Model predictions $\boldsymbol{\varphi}(\boldsymbol{a}) = [\varphi(x_1, \boldsymbol{a}), \dots, \varphi(x_m, \boldsymbol{a})]^T$
- $\mathbf{y}=(y_1, \dots, y_m)^T$ is a fixed point or vector in R^m
- $\boldsymbol{a} \mapsto \boldsymbol{\varphi}(\boldsymbol{a})$ is a surface in R^m
- Choose $\boldsymbol{a}$ such that $\varphi(\boldsymbol{a})$ is as close as possible to **y**
- $\min_{\boldsymbol{a}} E(\boldsymbol{a}) = \|\boldsymbol{y} - \boldsymbol{\varphi}(\boldsymbol{a})\|^2 = \sum_{i=1}^{m}[y_i - \varphi(x_i, a)]^2$

Geometrical Interpretation

- Data vector $\boldsymbol{y}=(y_1,\ldots,y_m)^{\mathrm{T}}$ defines a fixed point in $\mathcal{R}^m$
- Mapping $\boldsymbol{a}\mapsto\boldsymbol{\varphi}(\boldsymbol{a})$ defines an n-dimensional surface $S(\boldsymbol{a})$ in $\mathcal{R}^m$
- Best-fit parameter values $\widehat{\boldsymbol{a}}$ define the point $\boldsymbol{\phi}(\widehat{\boldsymbol{a}})$ on the surface $S(\boldsymbol{a})$ closest to $\boldsymbol{y}$
- Parameter estimation method is a method of associating with a data vector $\boldsymbol{y}$ a unique point on the surface $S(\boldsymbol{a})$

Section 10 Advanced Data Analysis

Introduction
Law of Propagation of Uncertainty (LPU)
Fitting a Model to Measured Data
► **Linear Least Square Method**
Non-linear Least Square Method
Lab sessions

Fitting with Linear Models Using Least Squares Method (1)

- $y_i = \varphi(x_i, \boldsymbol{a}) = a_1\varphi_1(x_i) + a_2\varphi_2(x_i)+\ldots+a_n\varphi_n(x_i)$
- $E(\boldsymbol{a}) = \|\boldsymbol{y} - \boldsymbol{\varphi}(\boldsymbol{a})\|^2 = \sum_{i=1}^{m}[y_i - \varphi(x_i, a)]^2$

$$= \sum_{i=1}^{m}[y_i - a_1\varphi_1(x_i) - a_2\varphi_2(x_i) - \ \ldots - a_n\varphi_n(x_i)]^2$$

- Let $\frac{\partial E}{\partial a_1} = \sum_{i=1}^{m} -2\varphi_1(x_i)[y_i - a_1\varphi_1(x_i) - a_2\varphi_2(x_i) - \ \ldots - a_n\varphi_n(x_i)] = 0,$

  we have:

  $$a_1\sum_{i=1}^{m}\varphi_1(x_i)\,\varphi_1(x_i) + \cdots + a_n\sum_{i=1}^{m}\varphi_1(x_i)\,\varphi_n(x_i) = \sum_{i=1}^{m}\varphi_1(x_i)y_i\,,$$

- Similarly,

  Let $\frac{\partial E}{\partial a_2} = \sum_{i=1}^{m} -2\varphi_2(x_i)[y_i - a_1\varphi_1(x_i) - a_2\varphi_2(x_i) - \ \ldots - a_n\varphi_n(x_i)] = 0,$

  we have:

  $$a_1\sum_{i=1}^{m}\varphi_2(x_i)\,\varphi_1(x_i) + \cdots + a_n\sum_{i=1}^{m}\varphi_2(x_i)\,\varphi_n(x_i) = \sum_{i=1}^{m}\varphi_2(x_i)y_i.$$

Fitting with Linear Models Using Least Squares Method (2)

- The solution is defined by the normal equations

$$\begin{bmatrix} \sum_i \varphi_1(x_i)\varphi_1(x_i) & \sum_i \varphi_1(x_i)\varphi_2(x_i) & \ldots & \sum_i \varphi_1(x_i)\varphi_n(x_i) \\ \sum_i \varphi_2(x_i)\varphi_1(x_i) & \sum_i \varphi_2(x_i)\varphi_2(x_i) & \ldots & \sum_i \varphi_2(x_i)\varphi_n(x_i) \\ \vdots & \vdots & \vdots & \vdots \\ \sum_i \varphi_n(x_i)\varphi_1(x_i) & \sum_i \varphi_n(x_i)\varphi_2(x_i) & \ldots & \sum_i \varphi_n(x_i)\varphi_n(x_i) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \sum_i \varphi_1(x_i)y_i \\ \sum_i \varphi_1(x_i)y_i \\ \vdots \\ \sum_i \varphi_1(x_i)y_i \end{bmatrix}$$

- Factor form

$$\sum_i \begin{bmatrix} \varphi_1(x_i) \\ \varphi_2(x_i) \\ \vdots \\ \varphi_n(x_i) \end{bmatrix} \begin{bmatrix} \varphi_1(x_i) \\ \varphi_2(x_i) \\ \vdots \\ \varphi_n(x_i) \end{bmatrix}^T \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \sum_i y_i \begin{bmatrix} \varphi_1(x_i) \\ \varphi_2(x_i) \\ \vdots \\ \varphi_n(x_i) \end{bmatrix}$$

Linear Least Squares Method Using Matrix-vector Form (1)

- Standard linear model:
  $y_i = \varphi(x_i, \boldsymbol{a}) + \epsilon_i = a_1\varphi_1(x_i) + a_2\varphi_2(x_i)+\ldots+a_n\varphi_n(x_i) + \epsilon_i = \boldsymbol{c}_i^T\boldsymbol{a} + \epsilon_i$, i=1, 2, …, m

  $E(\epsilon_i) = 0, V(\epsilon_i) = \sigma^2$
- In matrix-vector form
  $$\boldsymbol{y} = \boldsymbol{C}\boldsymbol{a} + \boldsymbol{\epsilon}, \qquad E(\boldsymbol{\epsilon}) = \boldsymbol{0}, V(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$$
  where
  $\boldsymbol{c}_i^T = [\varphi_1(x_i), \varphi_2(x_i), \ldots, \varphi_n(x_i)]$,
  $$C = \begin{bmatrix} \boldsymbol{c}_1^T \\ \boldsymbol{c}_2^T \\ \vdots \\ \boldsymbol{c}_m^T \end{bmatrix} = \begin{bmatrix} \varphi_1(x_1) & \varphi_2(x_1) & \ldots & \varphi_n(x_1) \\ \varphi_1(x_2) & \varphi_2(x_2) & \ldots & \varphi_n(x_2) \\ \vdots & \vdots & \vdots & \vdots \\ \varphi_1(x_m) & \varphi_2(x_m) & \ldots & \varphi_n(x_m) \end{bmatrix}, \boldsymbol{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{bmatrix}$$
- $E(\boldsymbol{y}) = E(C\boldsymbol{a} + \boldsymbol{\epsilon}) = C\boldsymbol{a} + E(\boldsymbol{\epsilon}) = C\boldsymbol{a}$,
- $V(\boldsymbol{y}) = \sigma^2\boldsymbol{I}$

Linear Least Squares Method Using Matrix-vector Form (2)

- Least-squares fit, the solution $\hat{\boldsymbol{a}}$ minimizes
  $$E(\boldsymbol{a}) = \|\boldsymbol{y} - C\boldsymbol{a}\|^2 = (\boldsymbol{y} - C\boldsymbol{a})^{\boldsymbol{T}}(\boldsymbol{y} - C\boldsymbol{a})$$
- $\frac{dE}{d\boldsymbol{a}} = 0$, i.e.
  $$2(\boldsymbol{y} - C\boldsymbol{a})^{\boldsymbol{T}}(-C) = \boldsymbol{0}$$
  $$C^T(\boldsymbol{y} - C\boldsymbol{a}) = \boldsymbol{0}$$
  $$C^T C\boldsymbol{a} = C^T\boldsymbol{y}$$
  $$\hat{\boldsymbol{a}} = (C^T C)^{-1}C^T\boldsymbol{y} = C^+\boldsymbol{y}$$
- Matrix $C^+ = (C^T C)^{-1}C^T$ is the pseudo-inverse of C
  $$C^+C = (C^T C)^{-1}C^T C = I$$

Geometrical Interpretation

- The columns $\mathbf{c}_j$ (j=1,…n) and $\mathbf{y}$ are vectors or points in R^m.
- The linear combinations $C\boldsymbol{a} = \sum_{j=1}^{n} a_j \boldsymbol{c_j} = a_1\boldsymbol{c_1} + \cdots + a_n\boldsymbol{c_n}$ of the vectors $\mathbf{c}_j$ defines points in the n-dimensional linear subspace $\mathcal{C}$ defined by these column vectors.
- The linear least square solution defines the point $\widehat{\boldsymbol{y}} = C\boldsymbol{a}$ on linear subspace $\mathcal{C}$ closest to y.
- The vector $\boldsymbol{y} - \widehat{\boldsymbol{y}}$ from $\boldsymbol{y}$ to $C\boldsymbol{a}$ must be orthogonal to the plane and perpendicular to every $\mathbf{c}_j$: $\boldsymbol{c}_j^T(\boldsymbol{y} - C\boldsymbol{a}) = 0$, j=1…, n.
- In matrix notation,

$$C^T(\boldsymbol{y} - C\boldsymbol{a}) = 0$$

or

$$C^T\boldsymbol{y} = C^TC\boldsymbol{a}$$

- C is known as the observation matrix.

Statistics Associated with the Least Squares Linear Model Fit

- $\widehat{\boldsymbol{a}} = C^+\boldsymbol{y},\ with\ C^+ = (C^TC)^{-1}C^T, E(\boldsymbol{y}) = C\boldsymbol{a}, V(\boldsymbol{y}) = \sigma^2\boldsymbol{I}$
- Applying LPU,

$$E(\widehat{\boldsymbol{a}}) = C^+E(\boldsymbol{y}) = C^+C\boldsymbol{a} = \boldsymbol{a}$$
$$\mathrm{V}(\widehat{\boldsymbol{a}}) = C^+V(\boldsymbol{y})(C^+)^T = (C^TC)^{-1}C^T\sigma^2\boldsymbol{I}C(C^TC)^{-1} = \sigma^2(C^TC)^{-1}$$

- If $\boldsymbol{y}$ is a random draw from a distribution with expectation $\boldsymbol{Ca}$ and variance $\sigma^2\boldsymbol{I}$, then $\widehat{\boldsymbol{a}}$ is a random draw from a distribution with expectation $\boldsymbol{a}$ and variance $\sigma^2(\boldsymbol{C}^T\boldsymbol{C})^{-1}$.
- If $m \gg n$, then the central limit theorem implies that

$$\widehat{\boldsymbol{a}} \in N(\boldsymbol{a}, \sigma^2(C^TC)^{-1})$$

Statistics Associated with Other Calculations

- Model predications at x_i

$$\hat{\boldsymbol{y}} = C\hat{\boldsymbol{a}},$$
$$V(\hat{\boldsymbol{y}}) = CV(\hat{\boldsymbol{a}})C^T = \sigma^2 C(C^T C)^{-1}C^T$$

- Vector of residuals

$$\boldsymbol{r} = \boldsymbol{y} - C\hat{\boldsymbol{a}} = (I - CC^+)\boldsymbol{y}$$
$$V_r = (I - CC^+)V_y(I - CC^+)^T = \sigma^2(I - C(C^T C)^{-1}C^T)$$

- Model prediction at a new point x

$$\hat{y} = \boldsymbol{c}^T\hat{\boldsymbol{a}}, \boldsymbol{c}^T = [\varphi_1(x), \varphi_2(x), \ldots, \varphi_n(x)]$$

$$\mathrm{V}(\hat{\boldsymbol{y}}) = \boldsymbol{c}^T V(\hat{\boldsymbol{a}})\boldsymbol{c} = \sigma^2\boldsymbol{c}^T(C^T C)^{-1}\boldsymbol{c}$$

Posterior Estimate of $\boldsymbol{\sigma}$

- If

$$\boldsymbol{y} = C\boldsymbol{a} + \boldsymbol{\epsilon}, C: m \times n, \boldsymbol{\epsilon} \in N(0, \sigma^2 I)$$
$$\hat{\boldsymbol{a}} = (C^T C)^{-1}C^T\boldsymbol{y}, \qquad \boldsymbol{r} = \boldsymbol{y} - C\hat{\boldsymbol{a}}$$

Then

$$\sum_i r_i^2 = \boldsymbol{r}^T\boldsymbol{r} \in \chi^2_{m-n}\,,\ E(\boldsymbol{r}^T\boldsymbol{r}) = m - n$$

- Posterior estimate of σ, V($\hat{\boldsymbol{a}}$):

$$\hat{\boldsymbol{\sigma}}^2 = \frac{\boldsymbol{r}^T\boldsymbol{r}}{m-n},$$
$$V(\hat{\boldsymbol{a}}) = \hat{\sigma}^2(C^T C)^{-1}$$

Exercise

Modify Matlab scripts r_line_fit_A.m to fit a quadratic $y = a + bx + cx^2$
(or cubic) to data.

What is the formula for

$$\sum_i u^2(r_i)$$

the sum of squares of the uncertainties associated with the residuals, in terms of σ, m and n?
Can you think/prove why this formula applies?

Coefficient of Determination, R^2

- R^2 measures how well the regression model approximates the real data points. An R^2 of 1 indicates that the regression line fits the data perfectly.
- $R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$
  $where\ TSS(total\ SS) = \sum_i (y_i - \bar{y})^2\, , ESS\ (Explained\ SS) = \sum_i (\hat{y}_i - \bar{y})^2\, ,$
  $RSS\ (Residual\ SS) = \sum_i (y_i - \hat{y}_i)^2 = \boldsymbol{r}^T\boldsymbol{r}\, , TSS = ESS + RSS.$
- Adjusted R^2 :
  $$\bar{R}^2 = 1 - \frac{\frac{RSS}{m-n}}{\frac{TSS}{m-1}} = 1 - (1 - R^2)\frac{m-1}{m-n}$$
- $\bar{R}^2$ can reduce the influence of m on the value of R^2.

Significance Testing of Single Parameter

- Hypothesis:

$$H_0: a_i = 0 \quad (i = 1, 2, \dots, n)$$
$$H_1: a_i \neq 0 \quad (i = 1, 2, \dots, n)$$

- Calculate t statistic:

$$t(\hat{a}_i) = \frac{\hat{a}_i - a_i}{S(\hat{a}_i)} = \frac{\hat{a}_i}{S(\hat{a}_i)} \sim t(m-n)$$

where $S(\hat{a}_i) = \sqrt{v_{ii}}$, c_{ii} is the ith element on diagonal of $V(\hat{\boldsymbol{a}})$.

- Choose a suitable significance level, e.g. α=0.05 (Confidence level=1-α=0.95=95%)

- If $|t(\hat{a}_i)| > t_{\frac{\alpha}{2}}$(m-n), reject H_0. a_i has significance influence y. Otherwise, accept H_0, a_i should be excluded in the model.

Prediction Interval

- Model prediction at a new point x

$$\hat{y} = \boldsymbol{c}^T\boldsymbol{a},\ \boldsymbol{c}^T = [\varphi_1(x), \varphi_2(x), \dots, \varphi_n(x)]$$

$$\mathrm{V}(\hat{y}) = \boldsymbol{c}^{\boldsymbol{T}} V(\hat{\boldsymbol{a}})\boldsymbol{c} = \sigma^2 \boldsymbol{c}^T (C^T C)^{-1} \boldsymbol{c}$$

- Since $y = \boldsymbol{c}^T\boldsymbol{a} + \boldsymbol{\epsilon} = \hat{y} + \boldsymbol{\epsilon}$

$$\mathrm{V}(y) = V(\hat{y}) + \sigma^2 = \sigma^2(1 + \boldsymbol{c}^T (C^T C)^{-1} \boldsymbol{c})$$

- $\frac{y-\hat{y}}{S(\boldsymbol{y})} = \frac{y-\hat{y}}{\hat{\boldsymbol{\sigma}}\sqrt{(1+\boldsymbol{c}^T(C^TC)^{-1}\boldsymbol{c})}} \sim t_{\frac{\alpha}{2}}$(m−n)

- The prediction interval (confidence level (1-α)%) is thus

$$[\hat{y} - S(y)\ t_{\frac{\alpha}{2}}(\mathrm{m-n}), \hat{y} + S(y)\ t_{\frac{\alpha}{2}}(\mathrm{m-n})]$$

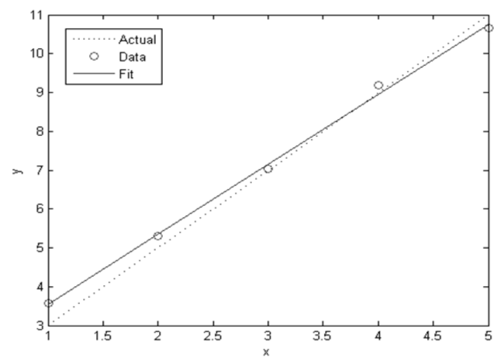
Example: Straightline Fit

$$y_i = a + bx_i + e_i,$$
$$e_i \in N(0, \sigma^2), i = 1,2, \dots, m.$$

C is a $m \times 2$ observation matrix:

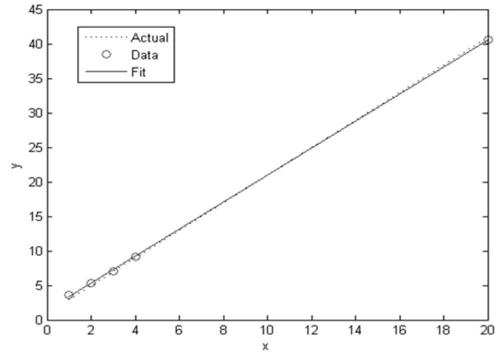
$$C(i, 1{:}2) = [1\ x_i]$$

Data Set A



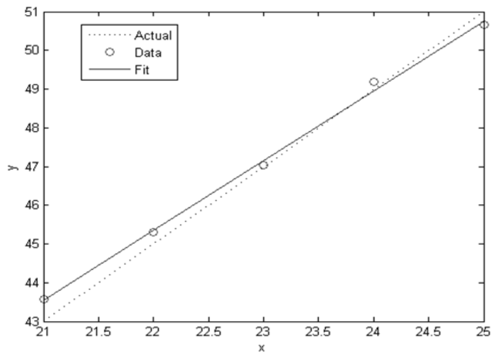
ua	Ca	
0.5244	1.0000	-0.9045
0.1581	-0.9045	1.0000

Data Set B



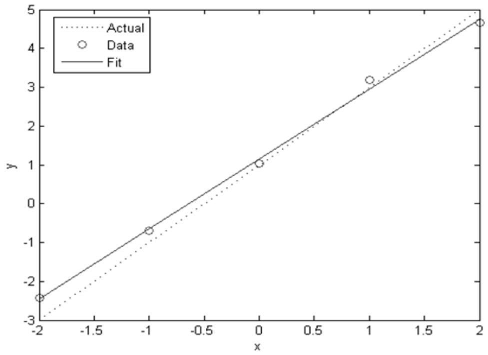
ua	Ca	
0.2933	1.0000	-0.6470
0.0316	-0.6470	1.0000

Data Set C



ua	Ca	
3.6435	1.0000	-0.9981
0.1581	-0.9981	1.0000

Data Set D



ua	Ca	
0.2236	1.0000	0
0.1581	0	1.0000

Section 10 Advanced Data Analysis

- Introduction
- Law of Propagation of Uncertainty (LPU)
- Fitting a Model to Measured Data
- Linear Least Square Method
- ► **Non-linear Least Square Method**
- Lab sessions

Least Squares Approximation with Nonlinear Models

- Model: $y = \varphi(\boldsymbol{x}, \boldsymbol{a})$, e.g.
$$\varphi(t, T_0, T_b, k) = T_b + (T_0 - T_b)e^{-kt}$$
- Data: $(\boldsymbol{x_i}, y_i), i = 1,2, \dots, m$
- Least squares parameter estimates $\hat{\boldsymbol{a}}$ minimizes
$$E(\boldsymbol{a}) = \tfrac{1}{2}\sum_{i=1}^{m} f_i^2(\boldsymbol{a}) = \frac{1}{2}\boldsymbol{f}^T\boldsymbol{f},$$
$$f_i(\boldsymbol{a}) = y_i - \varphi(\boldsymbol{x_i}, \boldsymbol{a}), \boldsymbol{f} = (f_1, f_2, \dots, f_m)^T$$

Geometrical Interpretation

- $\boldsymbol{a} \mapsto \varphi(\boldsymbol{a})$ is a n-surface in R^m.
- We look for the point on the surface closest to y.
- At the solution $\varphi(\hat{\boldsymbol{a}})$, the vector $\mathbf{f} = \mathbf{y} - \varphi(\hat{\boldsymbol{a}})$ is orthogonal to the surface at $\hat{a}$.
- The tangent plane at $\hat{\boldsymbol{a}}$ is
$$\varphi(\boldsymbol{a}) \approx \varphi(\hat{\boldsymbol{a}}) + J^T(\boldsymbol{a} - \hat{\boldsymbol{a}}), \mathrm{J_{ij}} = \frac{\partial f_i}{\partial \boldsymbol{a_j}}(\hat{\boldsymbol{a}}),$$
- so $\hat{\boldsymbol{a}}$ must solve
$$J^T(a)f(a) = J^T(a)\big(y - \varphi(a)\big) = 0.$$
- Linear case: $C^T(\boldsymbol{y} - C\boldsymbol{a}) = 0$.

Nonlinear Least Squares Method

- The gradient of the objective function is

$$\boldsymbol{g} = \frac{\partial E}{\partial \boldsymbol{a}} = \frac{1}{2}\left(2\boldsymbol{f}^T\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{a}}\right) = \boldsymbol{f}^T J = J^{\boldsymbol{T}}\boldsymbol{f},$$

$$J_{ij} = \frac{\partial f_i}{\partial a_j}.$$

- The Hessian matrix:

$$H = \frac{\partial^2 E}{\partial \boldsymbol{a}^2} = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{a}} = J^T\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{a}} + \boldsymbol{f}^T\frac{\partial J}{\partial \boldsymbol{a}} = J^T J + G,$$

$$G_{jk} = \sum_{i=1}^{m} f_i \frac{\partial^2 f_i}{\partial a_j a_k}$$

- The necessary condition $\boldsymbol{g} = 0$, i.e.

$$\boldsymbol{f}^T J = \boldsymbol{0},$$

or $J^T\boldsymbol{f} = \boldsymbol{0}$.

Nonlinear Least Squares Method (Newton method)

- Newton method to find the zero of function $g(a) = 0$
- Given estimate a, linearise $g(a + \Delta a) \approx g(a) + g'(a)\Delta a$
- Given an initial a, we try to find Δa such that $g(a + \Delta a) = 0$, hence

$$g'(a)\Delta a = -g(a)$$

- Multivariate case:

$$\boldsymbol{g}(\boldsymbol{a} + \Delta\boldsymbol{a}) = \boldsymbol{g}(\boldsymbol{a}) + H\Delta\boldsymbol{a}$$

$$H\Delta\boldsymbol{a} = -\boldsymbol{g}$$

- Solve the above equation in LS sense to find $\mathbf{p} = \Delta\boldsymbol{a} = -H\backslash\boldsymbol{g}$
- Update $\hat{\mathbf{a}}_{\mathrm{k+1}} = \hat{\mathbf{a}}_{\mathrm{k}} + \mathbf{p}$.

Nonlinear Least Squares Method (Gauss-Newton method)

- Use Newton's algorithm
$$H\Delta\boldsymbol{a} = -\boldsymbol{g}$$
- Approximate the Hessian matrix (when f_i is small):
$$H = J^T J + G \approx J^T J$$
- Hence
$$J^T J\Delta\boldsymbol{a} = -\boldsymbol{g} = -J^T\boldsymbol{f}$$
- Normal equations for $J\Delta\boldsymbol{a} = -\boldsymbol{f}$
- Solve the above equation in LS sense to find $\mathbf{p} = \Delta\boldsymbol{a} = -J\backslash\boldsymbol{f}$ (using QR factorisation, $J = QR$)
- Update $\hat{\mathbf{a}}_{\mathrm{k}+1} = \hat{\mathbf{a}}_{\mathrm{k}} + \mathbf{p}$.
- Uncertainty $\mathrm{V}(\hat{\boldsymbol{a}}) \approx \sigma^2(J^T J)^{-1}$

Section 10 Advanced Data Analysis

Introduction

Law of Propagation of Uncertainty (LPU)

Fitting a Model to Measured Data

Linear Least Square Method

Non-linear Least Square Method

► **Lab sessions**