Data Fitting

A. Godunov

1. Data modeling
2. Least-square fitting
3. Linear models
4. Non-linear models
5. Software

updated 5 April 2022

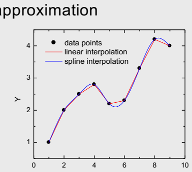
1

Part 1:
Data modeling

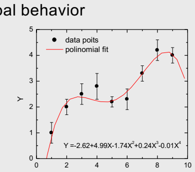
2

Data modeling vs. interpolation

Interpolation = local approximation



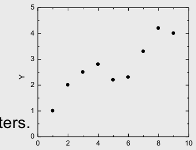
Data modeling = global behavior



3

3

Data and models



Given a set of observations, one often wants to condense and summarize the data by fitting it to a "model" that depends on adjustable parameters.

Sometimes the model is simply a convenient class of functions, such as polynomials or Gaussians, and the fit supplies the appropriate coefficients.

Other times, the model's parameters come from some underlying theory that the data are supposed to satisfy.

The basic approach is to find a set of parameters that minimize the difference between the data and the model.

4

4

Real data

There are important issues that go beyond the mere finding of best-fit parameters.

- Data are generally not exact. They are subject to *measurement errors* (called *noise* in the context of signal-processing).
- Thus, typical data never exactly fit the model that is being used, even when that model is correct.
- We need the means to assess whether or not the model is appropriate, that is, we need to test the goodness-of-fit against some useful statistical standard.
- We usually also need to know the accuracy with which parameters are determined by the data set. In other words, we need to know the likely errors of the best-fit parameters.

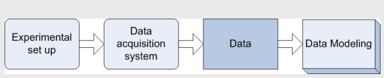
5

5

Steps and objectives

Objectives:

- Condense and summarize the data
- Using data in applications
- Getting deeper insight

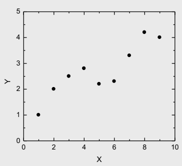


6

6

Major steps in data modeling

1. Getting data from normally observation (experiment)
   Data are generally not exact - measurement errors, noise
2. Selecting a model
   a) General: a function with adjustable parameters
      $g(x; a_1, a_2, \dots a_n)$
   b) Specific: reflecting the nature of data
3. Fitting procedure



7

7

Fitting procedure should provide

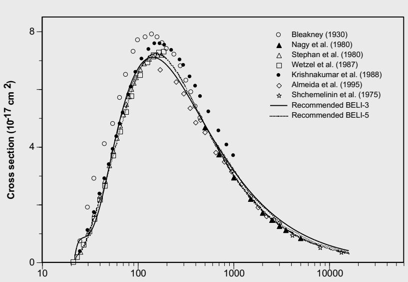
- Parameters a_j in $g(x; a_1, a_2, \dots a_n)$
- Error estimates on the parameters
- Statistical measure of goodness-of-fit

8

8

Example 1:

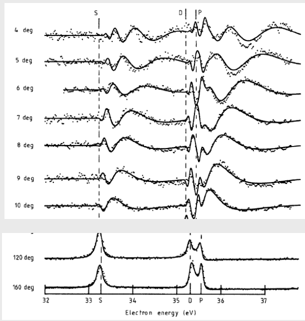
Godunov et al, Physica Scripta 59, 277 (1999)
On analytical fit for electron impact ionization cross sections



9

9

Example 2: data fit for $p + He \rightarrow p + He^{+} + e^{-}$



Experiment: Bordenave-Montesquieu et al (1996)

Fitting equation derived from theory Godunov et al 1995 (5 parameters)

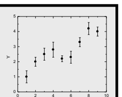
10

Part 2:

Least-square fitting

11

Least-square fitting



"Books have been written and careers have been spent discussing what is meant by a *good fit* to experimental data"*.

Assume that we have y_N data points from observations where $y(x)$. The observable data have the experimental uncertainty $y_i \pm \sigma_i,\ i = 1, 2, \dots N$

For simplicity we assume that all the errors σ_i occur in the dependent variable y_i (generally both x_i and y_i have errors).

Our goal is to determine how well a mathematical function $y = g(x)$ (also called a *model*) can describe y_i data.

Additionally, if the theory contains some parameters

$$g(x) \equiv g(x; a_1, a_2, \dots a_M) = g(x; \{a_M\})$$

our goal can be viewed as determining the best values for these parameters.

* R. Landau et. all Computational Physics, page 159

12

Least-square fitting (cont.)

We use the chi-square as a measure of how well a theoretical function g reproduces data (maximum likelihood estimation)

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - g(x_i; \{a_M\})}{\sigma_i}\right)^2$$

The definition χ^2 is such that smaller values of χ^2 are better fits, with $\chi^2 = 0$ occurring if the theoretical curve went through every data point.

Note that $1/\sigma_i^2$ factor means that measurements with larger errors contribute less to χ^2.

Least-squares fitting refers to adjusting the parameters in the theory until a minimum in χ^2 is found, that is, finding a curve that produces the least value for the summed squares of the deviations of the data from the function $g(x)$.

13

13

few notes

- Maximum likelihood estimation is entirely based on intuition
- It has no formal mathematical basis in and of itself
- It is based around normal distribution that is often wrong
  (Statistic is not a branch of mathematics)

There are three kinds of lies: lies, damned lies and statistics - Benjamin Disraeli (former British Prime Minister)

Statistics: The only science that enables different experts using the same figures to draw different conclusions – Evan Esar

14

14

Least-square fitting (cont.)

The M parameters $\{a_1, a_2, \dots a_M\}$ are found by solving the M equations:

$$\frac{\partial \chi^2}{\partial a_i} = 0$$

Attention!

For linear models

Example: $g(x) = a_0 + a_1 x + a_2 x^2$

a system of linear equations

For non-linear models

Example: $g(x) = (a_0 + a_1 x)e^{-a_2 x}$ (non-linear dependence on a_2)

a trial-and-error search through the M-dimensional parameter space. It can be a very challenging task!

Often a good guess is needed to find the best fit.

15

15

Part 3:

Linear models

16

A simple linear model

Consider a straight line

$$g(x) = a_0 + a_1 x$$

with two parameters.

Attention: a unique solution is not possible unless the number of data points is equal to or greater than the number of parameters.

$$\chi^2(a_0, a_1) = \sum_{i=1}^{N} \left(\frac{y_i - a_0 - a_1 x_i}{\sigma_i} \right)^2$$

After evaluating

$$\frac{\partial \chi^2(a_0, a_1)}{\partial a_0} = 0, \qquad \frac{\partial \chi^2(a_0, a_1)}{\partial a_1} = 0$$

and solving for a_0 and a_1 we have … (see the next slide)

17

17

Example: A simple linear model (cont.)

$$a_0 = \frac{S_{xx} S_y - S_x S_{xy}}{\Delta}, \qquad a_1 = \frac{S S_{xy} - S_x S_y}{\Delta}$$

$$S = \sum_{i=1}^{N} \frac{1}{\sigma_i^2}, \qquad S_x = \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}, \qquad S_y = \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2}$$

$$S_{xx} = \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2}, \qquad S_{xy} = \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2}, \qquad \Delta = S S_{xx} - S_x^2$$

Statistics also gives an expression for the *variance* or uncertainty in the deduced parameters:

$$\sigma_{a_0}^2 = \frac{S_{xx}}{\Delta}, \qquad \sigma_{a_1}^2 = \frac{S}{\Delta}$$

This is a measure of the uncertainties in the values of the fitted parameters arising from the uncertainties σ_i in the measured y_i values.

18

18

The correlation coefficient

A measure of the dependence of the parameters on each other is given by the correlation coefficient:

$$\rho(a_0, a_1) = \frac{cov\ (a_0, a_1)}{\sigma_{a_0}\sigma_{a_1}}, \qquad cov\ (a_0, a_1) = -\frac{S_x}{\Delta}$$

Here $cov\ (a_0, a_1)$ is the covariance of a_0 and a_1 and vanishes if a_0 and a_1 are independent.

The correlation coefficient $\rho(a_0, a_1)$ lies in the range $-1 \leq \rho \leq 1$, with a positive ρ indicating that the errors in a_0 and a_1 are likely to have the same sign, and a negative ρ indicating opposite signs.

19

19

Better for numerical calculations

The preceding analytic solutions for the parameters are of the form found in statistics books but are not optimal for numerical calculations because subtractive cancelation can make the answers unstable.

For example, Thompson (1992)* gives improved expressions that measure the data relative to their averages:

$$a_0 = y - a_1 x, \qquad a_1 = \frac{S_{xy}}{S_{xx}}, \qquad x = \frac{1}{N}\sum_{i=1}^{N} x_i\,, \qquad y = \frac{1}{N}\sum_{i=1}^{N} y_i\,,$$

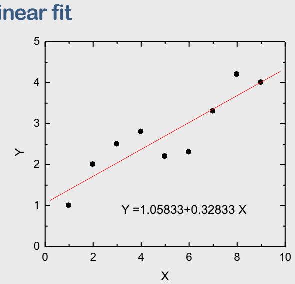
$$S_{xy} = \sum_{i=1}^{N} \frac{(x_i - x)(y_i - y)}{\sigma_i^2}, \qquad S_{xx} = \sum_{i=1}^{N} \frac{(x_i - x)^2}{\sigma_i^2},$$

* Thompson, W.J. (1992) *Computing for Scientists and Engineers*, John Wiley & Sons.

20

20

Example: linear fit



Parameter	Value	Error
a_0	1.05833	0.35504
a_1	0.32833	0.06309

21

21

Goodness-of-fit

The goodness-of-fit measures the agreement between data and the fitting model for a particular choice of the parameters

$$Q = \mathrm{gammaq}\left(\frac{N-2}{2}, \frac{\chi^2}{2}\right)$$

where gammaq is incomplete gamma functions

- if $Q > 0.1$ the goodness of fit is believable
- if $Q > 0.001$ the fit may be acceptable
- if $Q < 0.001$ change the model of fitting procedure

22

22

Issues to consider

- Errors in both coordinates
- Multidimensional fits

More can be found in Press et all "Numerical recipes" (multiple editions for Fortran, C++, Pascal, Java)

23

23

Part 4:
Non-linear models

24

Pro and cons non-linear fits

Pros:

- A fitting function can very well reflect the nature of data
- Lot of software available

Cons:

- Much more difficult to calculate. Trial-and-error approach.

284 *A. L. Godunov and P. B. Ivanov*

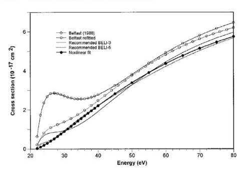


Fig. 12. The behaviour of a few fitting curves near the threshold for single ionisation of Ne by electron impact. The nonlinear fit has been obtained using the arc tangent formula (13) of Section 5, with the coefficients $A = 1117.10^{-16}\,\mathrm{eV\,cm^2}$, $B = 247.10^{-16}\,\mathrm{eV^2\,cm^2}$, $\beta = 0.24$.

25

25

Other methods

- Quick-and-dirty Monte-Carlo: The bootstrap method
- Genetic algorithm
- Simulated annealing
- and many more …

26

26

Part 5:

Software and libraries

27

A simple linear model

Program libraries:

- minpac
- lapack
- slatec
- sminpack
- napack
- …

Software

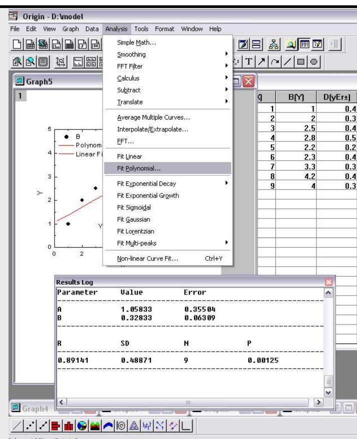
- Excel
- Origin
- MatLab
- Systat
- Statistica
- …

28

28

Example: Origin

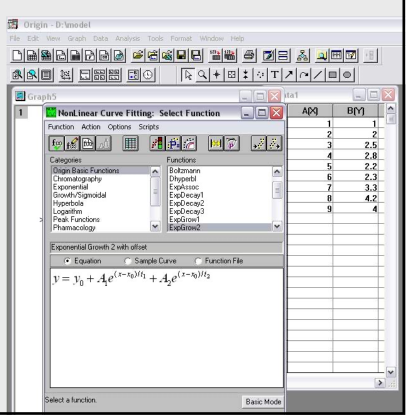
Polynomial fit



29

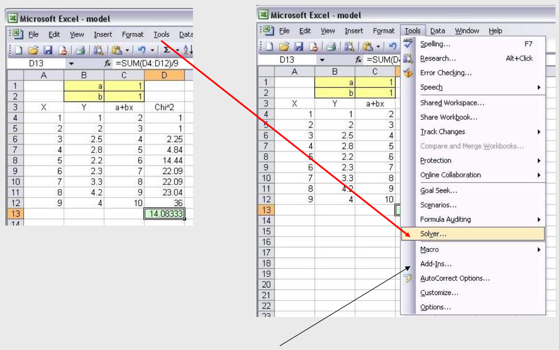
Example: Origin

Non-linear fit



30

Example: Excel

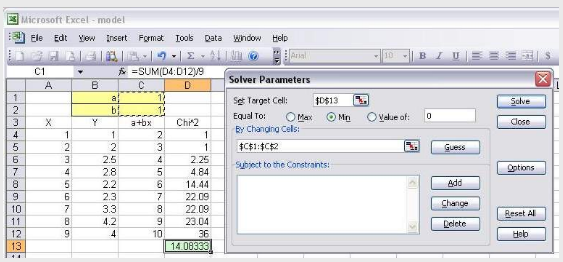


if you can not see “Solver” – check “Add-ins…”

31

31

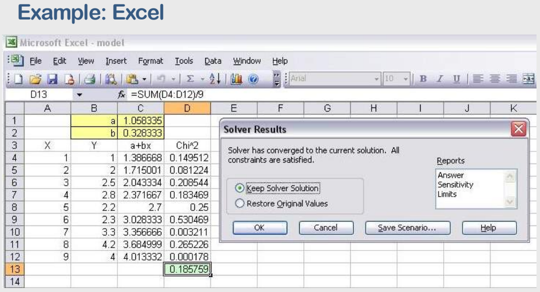
Example: Excel



32

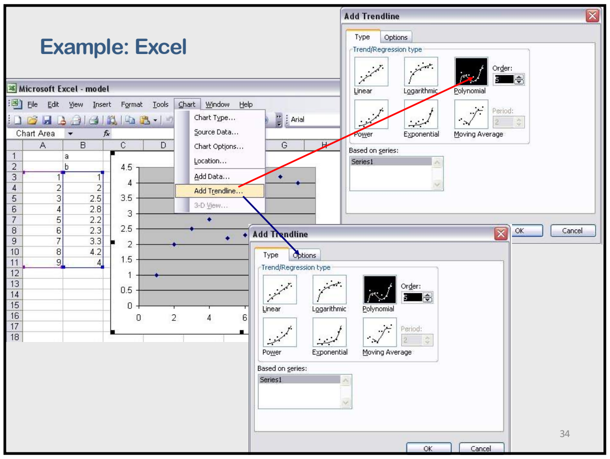
32

Example: Excel

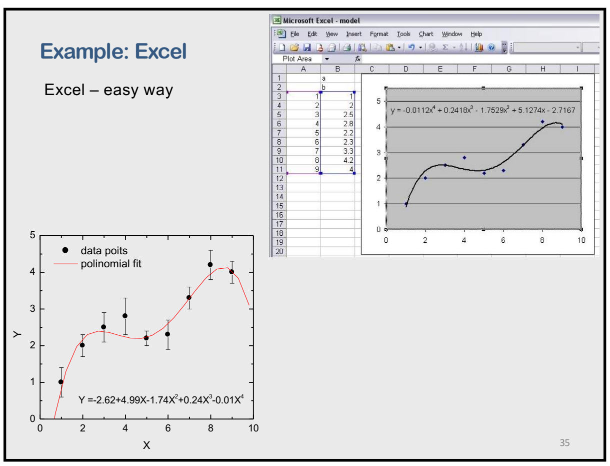


33

33


Example: Excel
Add Trendline
Type
Options
Trend/Regression type
Linear
Logarithmic
Polynomial
Order:
Power
Exponential
Moving Average
Period:
Based on series:
Series1
OK
Cancel
Microsoft Excel - model
Chart Type...
Source Data...
Chart Options...
Location...
Add Data...
Add Trendline...
3-D View...
34

34


Example: Excel
Excel – easy way
Microsoft Excel - model
y = -0.0112x^4 + 0.2418x^3 - 1.7529x^2 + 5.1274x - 2.7167
data poits
polinomial fit
Y =-2.62+4.99X-1.74X^2+0.24X^3-0.01X^4
X
Y
35

35