Lecture 4 Model Fitting

Mathematical Modeling

Prof. Dr. Jingzhi Li

Department of Mathematics,
Southern University of Science and Technology

2025 Spring

1. Preliminaries
2. About confidence intervals
3. Back to visual model fitting
4. Analytic Model Fitting

Data and Modeling: Three Situations in Data Modeling

Given a data set we may take three approaches:

1. **Fit an already selected model type to the data**
   - The model type is already fixed
   - For example: a linear, or a quadratic model; mass-action model
2. **Choose the most appropriate model from several alternative models that have been fitted to the data**
   - Decide whether the best-fitting exponential model is better than the best-fitting polynomial model
3. **Make predictions based solely on the data**
   - No hypothesis regarding the type of model
   - Predict intermediate and/or future behavior based just on the data set

Model Fitting vs. Interpolation

Model fitting

- The modeler has a **hypothesis** regarding the mathematical form of the model.
- It is just a matter of finding the numercial parameters that make the chosen model explain (fit) the experimental data set.
- Some deviations from the data are going to be willingly accepted.
- **Emphasis on the model.**

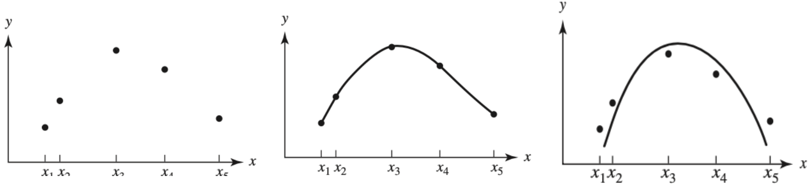
Interpolation

- **No a-priori hypothesis** regarding the model form.
- The modeler is strongly guided by the data.
- Aims to capture the data trend to predict in-between (or sometimes outside) the given points.
- **Emphasis on the data.**

Fitting and interpolating at the same time

- A fitted model may need to be replaced with an interpolating curve.
- The interpolating curve may have better mathematical properties for model analysis.
- Sometimes called **model approximation**.

- We are given a set of data
- Two approaches
  - Model fitting: we look for a **quadratic polynomial** to explain the data
  - Interpolation: look for a curve going exactly through the data points



Sources of error in modeling

Formulation errors

- Result from errors in the **model** formulation
- Significant variables were ignored
- Interrelationships between variables were ignored or simplified
- Relating the data to the model in the wrong way

Truncation errors

- Come from the math techniques used in building the model
- For example, an infinite series expansion may be truncated to a polynomial

Round-off errors

- Numerical errors coming from representing real numbers with finite precision

Measurement errors

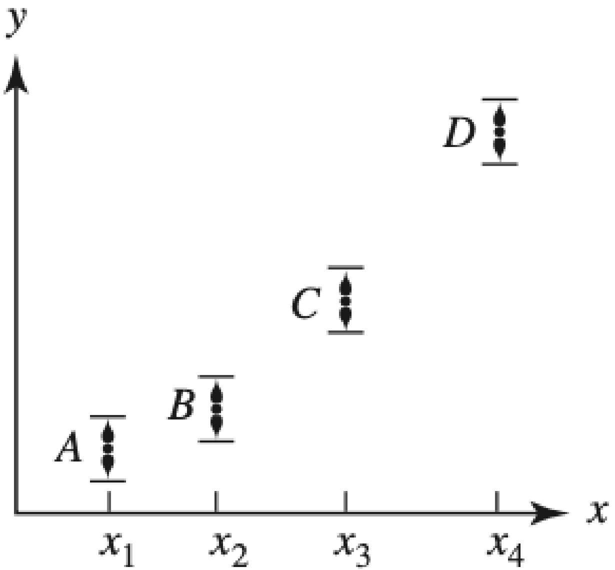
- Imprecision in the collection of data
- Physical limitations of the instruments
- Human errors

Content of this lecture

- **Model fitting**
  - Visual model fitting
  - (Confidence intervals)
  - Analytic model fitting
  - Choosing the best model

Visual model fitting

- The modeler has a hypothesis regarding the type of model to be fitted
- The numerical value of parameters needs to be fixed so that the model explains the available data
  - The range of the parameters is known
- Data set
  - The size of the data set is a trade-off between the cost of obtaining the data and the accuracy to be obtained for the model
  - **Minimum** at least as many data points as the number of parameters to fix
  - Spacing of the data points important
    - More data points on those intervals where the model should be fitted particularly well,
    - or where the maximum use of the model is expected,
    - or where abrupt changes in the model behavior are expected
  - Use the data together with its **confidence intervals**


y
A
B
C
D
x
x_1
x_2
x_3
x_4

Confidence intervals

Setup: a population and a parameter taking different values for different members of the population.

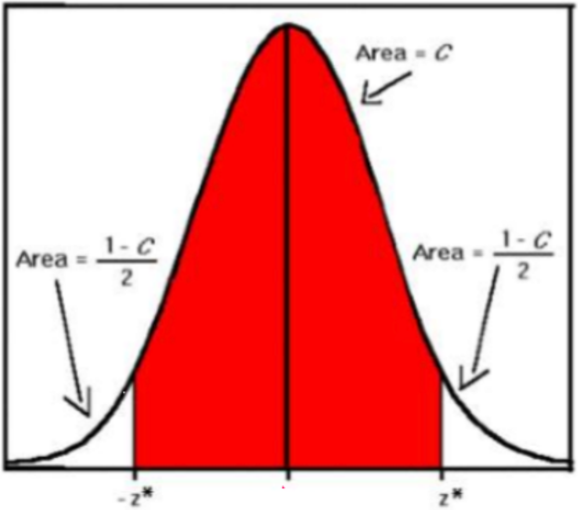
- Goal: We need to estimate the population mean value of that parameter only through a population sample
- Outcome: an interval where (the true value of) the population mean lies with large probability
- The interval is called confidence interval
- The probability is called confidence level

Example

- Measure the voting intentions for a certain party in an election
  - The answer may be presented as **40%**
  - It should be interpreted as an interval centered at 40% that is smaller or larger depending on the confidence level
  - The 90% confidence interval that could be calculated from the data might be for example 38% to 42%, while the 95% interval might be for example 36% to 44%

Confidence intervals (cont.): mathematical setup

- For a given sample, its mean is a stochastic variable
- If the measurements follow a normal distribution, the sample mean will have the distribution $N(\mu, \sigma/\sqrt{n})$
- From the calculated sample mean we need to report an interval where the real population mean μ lies with large probability
- The interval is related to the percentages of the area of the normal density curve
  - For example a 95% confidence interval should cover 95% of the area under the normal curve
- The value z^* of the point on the standard normal density curve $N(0, 1)$ such that the probability of observing a value greater than z^* is equal to p is known as the upper p-critical value of the standard normal distribution


Area = C
Area = $\frac{1-C}{2}$
Area = $\frac{1-C}{2}$
- z*
z*

Confidence intervals: unknown mean, known standard deviation

- Confidence intervals for unknown mean μ and known standard deviation σ
  - Given a sample of size n with mean m, a C-confidence interval for the population mean is:

$$(m - z^{*}\sigma/\sqrt{n}, \quad m + z^{*}\sigma/\sqrt{n})$$

    where z^{*} is the upper $(1 - C)/2$ critical value for the standard normal distribution
  - The error margin is $z^{*}\sigma/\sqrt{n}$

Note: the interval above is exact only for populations that are normally distributed. For other populations, the interval is approximately correct for large samples by the central limit theorem.

Confidence intervals: unknown mean, unknown standard deviation

- Confidence intervals for unknown mean μ and unknown standard deviation σ
  - The unknown standard deviation is replaced by the estimated standard deviation:

$$s^2 = \frac{1}{n-1} \sum (x_i - m)^2$$

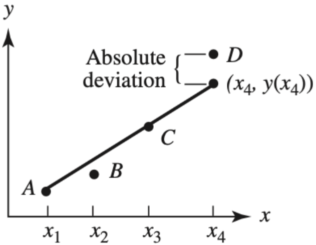
  - Given a sample of size n, its mean follows the t distribution $t(n-1)$ with mean μ and standard deviation $s/\sqrt{n}$
  - As the sample size n increases, the t distribution approaches the normal distribution
  - A C-confidence interval for the population mean is:

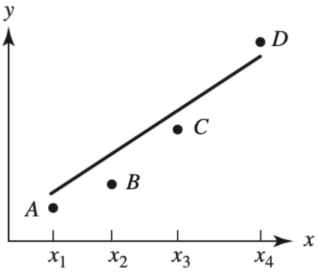
$$(m - t^* s/\sqrt{n}, \quad m + t^* s/\sqrt{n})$$

    where t^* is the upper $(1-C)/2$ critical value for the $t(n-1)$ distribution
  - The error margin is $t^* s/\sqrt{n}$

Visual model fitting

- Using the original data
  - Look at the deviations between the model prediction and the available data set
  - Aim to minimize the deviations
    - The largest deviation
    - Sum of all deviations
    - The sum of squares of deviations
    - ...
  - Example: fit a linear model $y = ax + b$ to the data on the previous slide
- **Note:** Although these methods of visually fitting the data may seem imprecise, they might be quite compatible with the accuracy of the modeling process
  - Grossness of the assumptions and the imprecision in the data collection may not warrantee a more sophisticated analysis


y
Absolute deviation
D
$(x_4, y(x_4))$
C
B
A
x
x_1 x_2 x_3 x_4


y
D
C
B
A
x
x_1 x_2 x_3 x_4

Visual model fitting

- Transforming the data
  - Much easier to fit visually to linear data
  - Problem: How about if the data is not linear?
  - Solution: transform the data!
- Examples:
  - Fit a model $y = Ce^x$ is the same as fitting a model $\ln(y) = \ln(C) + x$
    - Replace the (x, y) data with the (x, ln(y)) data (log-transform) and do a linear fit
  - Fit a model $y = Cx^a$ is the same as fitting a model $\ln(y) = \ln(C) + a\ln(x)$
    - Replace the (x, y) data with the (ln(x), ln(y)) data (log-log-transform) and do a linear fit

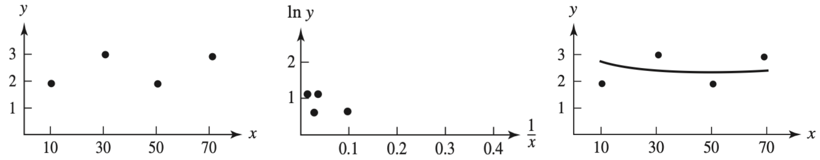
Data transformations

- When doing data transformations the concept of distance is also transformed
    - Fitting a transformed model to minimize the deviations may not yield a final model with minimum deviation to the original data
    - Always verify the final model against the original data!

Data transformations

Example

- Fit a model $y = Ce^{1/x}$
- Log-transform of the data: $\ln(y) = \ln(C) + \frac{1}{x}$
  - Data gets squeezed together
  - Absolute deviations appear small
  - Not all good fits on the log-scale will be good fits on the original scale



Data transformations

- Data may get squeezed
  - Modeler may be tricked in selecting a poor model based on the transformed data
  - Very important to remember when comparing alternative models
- Always compare using the original data
- Note: computer-based environments may use (hidden) implicit data transformations
  - Check how the indicators of model fit are computed

Analytic model fitting

- Methods for judging the fitness of a model
  - Chebyshev approximation criterion
  - Sum of absolute deviations
  - Least-squares criterion
  - ...

First criterion for goodness of a fit: Chebyshev approximation

- **Chebyshev criterion:**
  - Given a data set (x_i, y_i), $1 \le i \le m$, and a model $y = f(k, x)$, with k the vector of parameters to be fit, select those parameter values which minimize:

$$\max |y_i - f(x_i)|, \ 1 \le i \le m$$

- In other words: minimize the largest absolute deviation

Example: Chebyshev approximation

- We are given a segment AC and three measurements: of AC itself, of AB and of BC, where B is a point on the segment AC
  - **Example:** AC $= 19$, AB $= 13$, BC $= 7$
- Problem: find the values for the length of each segment that give the best fit using the **Chebyshev criterion**
- Notation: let x_1, x_2, $x_1 + x_2$ be the lengths of the segments AB, BC, and AC, respectively
- The deviations are:

$$x_1 - 13 = r_1, \quad x_2 - 7 = r_2, \quad x_1 + x_2 - 19 = r_3$$

Chebyshev approximation (cont.)

Problem formulation: find the minimal r such that:

$$|r_1| \leq r, \quad |r_2| \leq r, \quad |r_3| \leq r$$

Equivalently:

$$-r \leq r_i \leq r \Rightarrow r - r_i \geq 0, \quad r + r_i \geq 0$$

In our example:

$$\begin{aligned} r - x_1 + 13 &\geq 0 \\ r + x_1 - 13 &\geq 0 \\ r - x_2 + 7 &\geq 0 \\ r + x_2 - 7 &\geq 0 \\ r - x_1 - x_2 + 19 &\geq 0 \\ r + x_1 + x_2 - 19 &\geq 0 \end{aligned}$$

Linear programming! Solved through the simplex method

Chebyshev approximation

Note: in general, the resulting optimization problem may not be linear!

- Example: $f(x) = \sin(kx)$
- For this reason, the criterion is not much used in practice

Second criterion for goodness of a fit: sum of absolute deviations

Criterion:

- Given a data set (x_i, y_i), $1 \leq i \leq m$, and a model $y = f(k, x)$, with k the vector of parameters to be fit, select those parameter values which minimize:

$$\sum_{1 \leq i \leq m} |y_i - f(x_i)|$$

- In other words: minimize the sum of absolute deviations
- General approach: differentiate the sum with respect to each parameter, solve the 0-equations to find the critical points
- Problem: because of the modules, the derivatives may not be continuous

Example on the previous slide:

- Find $x_1, x_2 \geq 0$ such that $|x_1 - 13| + |x_2 - 7| + |x_1 + x_2 - 19|$ is minimal

Third criterion for goodness of a fit: least-squares

Criterion:

- Given a data set (x_i, y_i), $1 \leq i \leq m$, and a model $y = f(k, x)$, with k the vector of parameters to be fit, select those parameter values which minimize:

$$\sum_{1 \leq i \leq m} |y_i - f(x_i)|^2$$

- In other words: minimize the sum of squares of absolute deviations

Third criterion for goodness of a fit: least-squares

- Widely used criterion because the resulting problem can be easily solved using calculus: if f is mathematically "well-behaved" function (say, analytical), then so is the sum of squares

Example on the previous slide:

- Find $x_1, x_2 \geq 0$ such that

$$(x_1 - 13)^2 + (x_2 - 7)^2 + (x_1 + x_2 - 19)^2$$

  is minimal
- **Solution:** differentiate (partial derivative) with respect to the two parameters:
  - $2(x_1 - 13) + 2(x_1 + x_2 - 19) = 0$
  - $2(x_2 - 7) + 2(x_1 + x_2 - 19) = 0$
  - Equivalently: $2x_1 + x_2 = 32$, $x_1 + 2x_2 = 26$
  - Solution: $x_1 = 12 + \frac{2}{3}$, $x_2 = 6 + \frac{2}{3}$, $x_1 + x_2 = 19 + \frac{1}{3}$

Example: fitting a straight line

- Fitting a straight line with the least squares criterion
  - look for a model $y = Ax + B$
  - data set: (x_i, y_i), $1 \leq i \leq m$
  - Denote the least-squares solution by $y = ax + b$

- the minimization of $\sum_{1\leq i\leq m}(y_i - ax_i - b)^2$
- Necessary conditions: the partial derivatives of the sum with respect to a and b are zero:

$$\begin{cases} \dfrac{\partial S}{\partial a}(a,b) = -2\displaystyle\sum_{i=1}^{m}(y_i - ax_i - b)x_i = 0 \\ \dfrac{\partial S}{\partial b}(a,b) = -2\displaystyle\sum_{i=1}^{m}(y_i - ax_i - b) = 0 \end{cases}$$

$$\begin{cases} a = \dfrac{m\sum_{i=1}^{m} x_i y_i - \sum_{i=1}^{m} x_i \sum_{i=1}^{m} y_i}{m\sum_{i=1}^{m} x_i^2 - \left(\sum_{i=1}^{m} x_i\right)^2} \\ \\ b = \dfrac{\sum_{i=1}^{m} x_i^2 \sum_{i=1}^{m} y_i - \sum_{i=1}^{m} x_i y_i \sum_{i=1}^{m} x_i}{m\sum_{i=1}^{m} x_i^2 - \left(\sum_{i=1}^{m} x_i\right)^2} \end{cases}$$

Example: fitting a power curve

- Fitting a power curve with the least squares criterion
  - Goal: fit a curve of the form $y = ax^n$ where n is fixed, to a given data set
  - Minimize the following sum:

$$S = \sum_{i=1}^{m}(y_i - f(x_i))^2 = \sum_{i=1}^{m}(y_i - ax_i^n)^2$$

  - The only parameter here is a
  - Necessary condition: the derivative $\frac{dS}{da}(a) = 0$

$$\frac{dS}{da}(a) = -2\sum_{i=1}^{m} x_i^n(y_i - ax_i^n) = 0$$

$$a = \frac{\sum_{i=1}^{m} x_i^n y_i}{\sum_{i=1}^{m} x_i^{2n}}$$

Relating the 3 criteria

Geometric intuition

- Chebyshev criterion: minimize the largest absolute deviation
  - more weight given to the worst point
- Minimize the sum of absolute deviations
  - tends to treat each data point equally and to average the deviations
- Least-squares
  - somewhat in-between

Relating the 3 criteria

Analytical relationship between Chebyshev and least-squares

- Let $f_1(x)$ be the solution given by the Chebyshev criterion
  - Let $c_i = |y_i - f_1(x_i)|$
  - Let $c_{\max}$ be the largest c_i: the largest absolute deviation
  - f_1 is calculated so that $c_{\max}$ is minimal
- Let $f_2(x)$ be the solution given by the least-squares criterion
  - Let $d_i = |y_i - f_2(x_i)|$
  - f_2 is calculated so that $d_1^2 + d_2^2 + \cdots + d_m^2$ is minimal
  - Let $d_{\max}$ be the largest d_i

- Optimality of the Chebyshev solution: $c_{\max} \leq d_{\max}$
- Optimality of the least-squares solution: $d_1^2 + \cdots + d_m^2 \leq c_1^2 + \cdots + c_m^2$
- $c_m^2 \leq mc_{\max}^2$
- Define $D = \sqrt{(d_1^2 + \cdots + d_m^2)/m}$
- **Conclusion:** $D \leq c_{\max} \leq d_{\max}$

Note: least-squares is more widely used; if the difference between D and $d_{\max}$ is however considerable, consider using Chebyshev instead

Choosing a best model

- For the same data set and the same type of model different results may be obtained depending on the type of fit
  - other methods for model fitting also exist, potentially giving different results for the same data set (nondeterministic methods)
- **Question:** which model to choose as being "best"?
  - if one were allowed to change the type of math model (e.g., look for a cubic, rather than a quadratic polynomial) even better fits may be found
  - The answer should be given depending on the purpose of the model, the required precision, accuracy of the data, etc.
    - a preset sum of squares may be enough to judge a model "good"
  - Careful about applying the numerical criteria blindly
  - Example: for the following 4 data sets, the model $y = x$ yields the same sum of squared deviations

Fit quality

- Various methods for defining a quantitative measure for the quality of a model fit
  - Here present just one, from Kuhnel et al, BMC Systems Biology (2008)
  - Only one data set at a time
  - Gives a measure of the average deviation of the model prediction from the experimental data, normalized by (the average of) the absolute values of the model prediction
  - This measure of fit quality does not discriminate against models aiming to explain experimental data with large absolute values

- Let exp be the experimental data; m the number of experimental points

$$\texttt{qual(exp)} = \sqrt{\frac{\texttt{sum\_of\_squared\_deviations}}{m \cdot \texttt{mean\_of\_predicted\_values}}} \cdot 100\%$$

- Rule of thumb (Kuhnel et al): lower than 20% value for qual(exp) can be considered as a good fit

Learning objectives

- Understand the concepts of **model fitting** and **data interpolation**
- Indicate several possible sources of errors in modeling
- Understand the concept of confidence interval
- Ability to formulate the following **3 fitting criteria** for a given model:
    - Chebyshev approximation
    - Sum of absolute deviation
    - Least squares