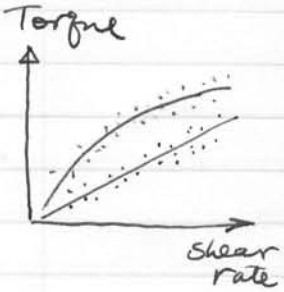
Chapter 2: Curve-fitting, or the Modeling of Data

References: N.R. book, chapter 15; C&M: Chpt. 7

§1. Introduction:



Imagine any type of experimental measurement — take rheometer to be exact, the data collected are scattered.

(a) Why curve-fitting?

- To condense and summerize data, so we needn't carry a "lookup" table of data.
- To interpolate & extropolate data to wider ranges of parameters. (risks! Need to know the underlying physics reasonebly well.)

(b) How to curve-fit?

(i) To come up with the "model": functional form with undetermined parameters.

* Convenience: polynomials, e.g., or other simple forms.
* Knowledge of underlying physics: theoretical guidance to model equation.

(ii) To determine the fitting parameters.

* Criterion for fitting: to minimize a suitably defined discrepancy or "merit function" $\Rightarrow$ best-fitting parameters.
* Advanced: to estimate acuracy of fitting and the likely error of the fitting parameters.

§2. Least-square fitting

Data set: (x_i, y_i), $i = 1, \cdots, N$

Model equ: $y(x) = y(x, \underbrace{a_1, a_2, \cdots, a_M}_{\text{parameters}})$

(a) Least-square fit:

Criterion is to minimize the ~~least~~ square sum:

$$\mathcal{L} = \sum_{i=1}^{N} [y_i - y(x_i, a_1, \cdots, a_M)]^2$$

with respect to the parameters $a_1, \cdots, a_M$.
($N-M$: degree of freedom).

Taking $\partial \mathcal{L} / \partial a_j = 0$, $j = 1, \cdots, M$:

$$\sum_{i=1}^{N} [y_i - y(x_i, a_1, \cdots, a_M)] \frac{\partial y}{\partial a_j}\Big|_{x_i} = 0$$

M-equations for M unknowns: $a_1, \cdots, a_M$.
[Nonlinear equations: Newton's method ?].

(b) Why this criterion (square-sum)?

It turns out that the least square fit generates a "maximum likelihood estimator". the chosen a_j's maximizes the probability or likelihood ~~that~~ the model equation, (of the actual data set (x_i, y_i)) for

Suppose the underlying physics indeed gives $y = y(x, a_1, \cdots, a_M)$, then when doing experiments, (x_i, y_i) has the greatest chance of being realized if a_j's are so chosen.

To see why this is the case, we need 2 assumptions:

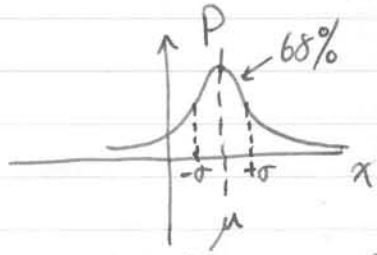
* y_i obeys the normal (or Gaussian) distribution around the exact value $y(x_i)$, with a standard deviation σ;

* All data points (x_i, y_i) are independent of each other and have the same σ.

$\Rightarrow$ Digression into normal distribution:

Probability distribution function

$$P(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



(Bell-curve)

Why is the normal distribution so important? "Central limit theorem": if you take the mean of many random variables, each having its own distribution, then the mean approaches Normal distribution when the # gets large.

Thus, if we take a particular measurement as the result of many random factors, it's reasonable to assume that the result (x_i, y_i) obeys a Normal distribution. $\Leftarrow$ End of digression.

Now the probability of (y_i) being within Δy of the mean $y(x_i)$ is:

$$P_i(\Delta y) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{[y_i - y(x_i)]^2}{2\sigma^2}} \Delta y.$$

and the probability of realizing the N independent events is:

$$P \propto \prod_{i=1}^{N} \left\{ \Delta y\, e^{-\frac{[y_i - y(x_i)]^2}{2\sigma^2}} \right\}$$

$$\ln P = N \ln \sigma_y + C - \frac{1}{2\sigma^2} \sum_{i=1}^{N} [y_i - y(x_i)]^2$$

$\Rightarrow$ To maximize probability P is equivalent to minimizing the square sum of all errors.

(c) Chi-square fitting.

What if the data have different standard deviation σ_i at different x_i values? From the above derivation, we should minimize

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i - y(x_i, a_1, a_2, \cdots, a_M)}{\sigma_i} \right]^2$$

This is known as Chi-square fitting.

* This downplays data points subject to larger scatter in the χ^2 criterion.
* How would we know σ_i? Maybe from the nature of the measurements.
* What if we don't know σ_i? Set them to constant (reverts to before), or maybe proportional to $|y_i|$ (depends on actual protocol & instrumentation of measurement.)
* What is the meaning of σ, standard deviation? It indicates the amount of scatter in data.

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \mu)^2, \quad \mu = \frac{1}{N} \sum_{i=1}^{N} y_i$$

$$\sigma^2 = \overline{y^2} - \overline{y}^2 \quad \left[\sigma^2 = \frac{1}{N} \sum (y_i^2 - 2 y_i \mu + \mu^2) = \frac{1}{N} \sum y_i^2 - \mu^2. \right]$$

For normal distribution

$$P(y) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}},$$