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

References: N.R. book, chapter 15; C&H, chp. 7

1. Introduction:

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

(a) Why curve-fitting?

- To condense and summarize data, so we needn't carry a "lookup" table of data.
- To interpolate & extrapolate data to wider ranges of parameters. (Risks! Need to know the underlying physics reasonably 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 accuracy of fitting and the likely error of the fitting parameters.
§ 2. Least-square fitting

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

Model eqn: \(y(x) = y(x, a_1, a_2, \ldots, a_M)\)

(a) Least-square fit:

Criterion is to minimize the least square sum:
\[
L = \sum_{i=1}^{N} \left[ y_i - y(x_i, a_1, a_2, \ldots, a_M) \right]^2
\]

with respect to the parameters \(a_1, \ldots, a_M\).

\((N-M: \text{degree of freedom})\)

Taking \(\partial L / \partial a_j = 0, \ j = 1, \ldots, M:\)

\[
\sum_{i=1}^{N} \left[ y_i - y(x_i, a_1, a_2, \ldots, a_M) \right] \frac{\partial y}{\partial a_j} x_i = 0
\]

\(M\) equations for \(M\) unknowns: \(a_1, \ldots, a_M\)

[Nonlinear equations: Newton's method?]

(b) Why this criterion (square sum)?

It turns out that the least-square fit generates a "maximum likelihood estimator" of the chosen \(a_j\)'s maximizes the probability or likelihood that the model equation, \((x_i, y_i)\)

Suppose the underlying physics indeed gives \(y = y(x, a_1, \ldots, a_M)\), 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 \( \sigma \);

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

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

Why is the normal distribution so (Bell-curve) 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 \( N \) gets large.

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

Now the probability of \((x_i, y_i)\) being \( \delta y \) of the mean \( y(x_i) \) is:

\[ P_i(\delta y) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(\delta y)^2}{2\sigma^2}} \]

and the probability of realizing the \( N \) independent events is:

\[ P \propto \prod_{i=1}^{N} \left\{ \left. \left[ y_i - y(x_i) \right] \right| \frac{1}{2\sigma^2} \right\} \]
\[ \ln P = N \ln y + C - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left[ \frac{y_i - y(x_i)}{\sigma_i} \right]^2 \]

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 deviations \( \sigma_i \) at different \( x_i \) values? From the above derivation, we should minimize

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

This is known as Chi-square fitting.

* This downplays data points subject to larger scatter in the \( \chi^2 \) criterion.
* How would we know \( \sigma_i \)? Maybe from the nature of the measurements.
* What if we don't know \( \sigma_i \)? 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 \( \sigma \), standard deviation?
  It indicates the amount of scatter in data.

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

\[ \sigma^2 = \bar{y}^2 - \bar{y}^2 \]

For normal distribution

\[ P(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left( -\frac{(y - \mu)^2}{2\sigma^2} \right) \]

\[ \mathcal{N}(\mu, \sigma^2) \]

\[ \sigma^2 \Rightarrow \text{variance} \]

\[ \sigma = \sqrt{\text{variance}} \]

\[ \mu \Rightarrow \text{mean} \]

\[ \bar{y} \Rightarrow \text{average of } y_i \]

\[ \mathcal{N}(\mu, \sigma^2) \Rightarrow 68\% \text{ chance} \]