Least-Squares Method

Imagine we have a set of $n$ measurements, $y_i$, at points $x_i$.

- Each measurement has uncertainty $\sigma_i$.

We want to fit to some function $f(x; \alpha, \beta, \ldots, \epsilon_n)$ where the fit parameters are unknown.

Calculate the figure of merit $N$

$$S = \frac{1}{2} \sum_{i=1}^{n} \left[ \frac{y_i - f(x_i; \alpha, \beta, \ldots, \epsilon_n)}{\sigma_i} \right]^2$$

It would be a $\chi^2$ if errors Gaussian (not always the case).

Seek $\alpha_j$ for which $S$ is a minimum.
Analytically determining $a_j$ would require solving system of equations

$$\frac{ds}{da_j} = 0$$

* Not always an analytic solution
  - need to probe analytically for minimum computationally

Errors

Obtained from covariance matrix

$$\left( V_{ij}^{-1} \right)_{ij} = \frac{1}{2} \frac{\delta s}{\delta a_i \delta a_j} \text{ at minimum}$$

- change of $a_i$ near minimum
  - variance of $a_i$

$$V = \begin{bmatrix}
\sigma_1^2 & \text{cov}(1,2) & \text{cov}(1,3) \\
\text{cov}(2,1) & \sigma_2^2 & \text{cov}(2,3) \\
\text{cov}(3,1) & \text{cov}(3,2) & \sigma_3^2 \\
\end{bmatrix}$$
Straight Line Fit

Consider points fitted to a simple linear function \( y = ax + b \)

To get \( a \) and \( b \):

\[
\frac{\partial S}{\partial a} = 0 \quad \frac{\partial S}{\partial b} = 0
\]

Assigning

\[
A = \sum_i x_i^2 \quad B = \sum \frac{x_i}{\delta_i^2}
\]

\[
C = \sum \frac{x_i}{\delta_i^2} \quad D = \sum \frac{x_i y_i}{\delta_i^2}
\]

\[
E = \sum \frac{x_i y_i}{\delta_i^2}
\]

Give

\[
a = \frac{EB - CA}{DB - A^2}
\]

\[
b = \frac{DC - EA}{DB - A^2}
\]

Covariance matrix is yields

\[
\sigma_a^2 = B/(BD-A^2) \quad \sigma_b^2 = D/(BD-A^2)
\]

\[
\operatorname{cov}(a, b) = -A/(BD-A^2)
\]
Charged Particle Geometries

Pattern recognition

- individual hits in tracking detector

How do we go from this?

Do this?

It is impractical to try all combinations of hits to find the best associations for tracks but have advantage of considering full range of configurations.

Can also analyze hits by several calculable parameters

\[ \text{eg. } \frac{\Delta y}{\Delta x} \text{ for straight tracks} \]

\[ \Delta y / \Delta x \]
Read Method: Tracking

If we have a track with measured positions relative to that calculated

\[ \eta_i = y_i - \bar{y} = \bar{y} - \bar{e} \]

\[ = x_i \cdot a_1 + 1 \cdot a_2 \]

linear relation

\[ \bar{\eta} = X \cdot \bar{a} \]

To obtain \( \bar{a} \), we need the covariance matrix of \( \bar{y} \)

\[ C_y = \begin{pmatrix} 2 & 2 & 0 \\ 2 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \sigma_y^{-1} \]

diagonal because each measurement is independent.

Using least-squares, one minimizes

\[ \chi^2 = \bar{e}^T \sigma_y^{-1} \bar{e} \]

set

\[ \bar{a} = (X^T \sigma_y X)^{-1} X^T \sigma_y \bar{y} \quad \text{(values)} \]

\[ \sigma_a = (X^T \sigma_y X)^{-1} \quad \text{(cov)} \]
For large trackers with many hits
- may start at outer edge of detector
- consider 3 hit triplets
  - fit to a parabola
  - interpolate
  - insert
  - forming track segments this way
  - a series of segments is combined to form a final track
  - drop points with large 'residuals'
  - distance from track of good 
- refuse
Clustering

- a series of "hits" or "cells" in a 2D plane
  - after a 3rd coordinate involved

- a position measurement is determined
  for control of distribution

\[ \langle \eta \rangle = \frac{\sum \eta_i}{N} \]

- accuracy of location due to
  - granularity of detector
- characteristic size & fluctuations of
  "pseudomomentum" measurement
- "Energy" in calorimeters
A Simple Cone Algorithm

Need to find the cluster of high energy hits cells in a calorimeter

- one method is to draw a cone in detector with vertex at interaction point

\[ \Delta R = \sqrt{\Delta \eta^2 + \Delta \phi^2} \]

- Look for cells with large energy deposition
- Draw cone \( \Delta R \) in radius around this cell
- Calculate a new \( \eta, \phi \) centroid

An iterative process is often needed
Quark jets:

Aftermath of parton fragmentation: jet
- a spray of particles arrayed roughly around direction of an initiating colored object: a quark or gluon
- this is not a perfect mapping between the kinematics of the "jet" and the parton

- at detector level
- final state particles from prompt decay of hadron fragmentation products
  - collision + shower in calorimeter
    - many overlapping showers

Several reconstruction schemes: two most common

ITERATIVE FIXED-CONE
JADE (energy or momentum space algorithm)
JADE Algorithm

We will iteratively consider the closest
of cells in momentum space
- as opposed to using $n_j$ of more locally

Use

$$Y_{ij} = \frac{2E_i E_j (1 - \cos \theta_{ij}) - M^2_{ij}}{E_{0+1s}^2}$$

Cells within a threshold on $y_{ij}$ are
combined:

yield a new "pseudojet" with its
own $E_j$ and direction

The specifics of many algorithms can
vary by experiment
Inverting (4.80), we find

\[ V = \frac{1}{A_{11}A_{22} - A_{12}^2} \begin{pmatrix} A_{22} & -A_{12} \\ -A_{12} & A_{11} \end{pmatrix} \tag{4.81} \]

so that

\[ \sigma^2(a) = \frac{A_{22}}{A_{11}A_{22} - A_{12}^2} = \frac{B}{BD - A^2}, \]

\[ \sigma^2(b) = \frac{A_{11}}{A_{11}A_{22} - A_{12}^2} = \frac{D}{BD - A^2}, \tag{4.82} \]

\[ \text{cov}(a, b) = -\frac{A_{12}}{A_{11}A_{22} - A_{12}^2} = -\frac{A}{BD - A^2}. \]

To complete the process, now, it is necessary to also have an idea of the \textit{quality of the fit}. Do the data, in fact, correspond to the function \( f(x) \) we have assumed? This can be tested by means of the chi-square. This is just the value of \( S \) at the minimum. Recalling Sect. 4.2.4, we saw that if the data correspond to the function and the deviations are Gaussian, \( S \) should be expected to follow a chi-square distribution with mean value equal to the degrees of freedom, \( v \). In the above problem, there are \( n \) independent data points from which \( m \) parameters are extracted. The degrees of freedom is thus \( v = n - m \). In the case of a linear fit, \( m = 2 \), so that \( v = n - 2 \). We thus expect \( S \) to be close to \( v = n - 2 \) if the fit is \textit{good}. A quick and easy test is to form the \textit{reduced chi-square}

\[ \frac{\chi^2}{v} = \frac{S}{v}, \tag{4.83} \]

which should be close to 1 for a \textit{good fit}.

A more rigorous test is to look at the probability of obtaining a \( \chi^2 \) value greater than \( S \), i.e., \( P(\chi^2 \geq S) \). This requires integrating the chi-square distribution or using cumulative distribution tables. In general, if \( P(\chi^2 \geq S) \) is greater than 5%, the fit can be accepted. Beyond this point, some questions must be asked.

An equally important point to consider is when \( S \) is very small. This implies that the points are \textit{not} fluctuating enough. Barring falsified data, the most likely cause is an overestimation of the errors on the data points. If the reader will recall, the error bars represent a 1\( \sigma \) deviation, so that about 1/3 of the data points should, in fact, be expected to fall outside the fit!

\textbf{Example 4.7} Find the best straight line through the following measured points

<table>
<thead>
<tr>
<th>( x )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>0.92</td>
<td>4.15</td>
<td>9.78</td>
<td>14.46</td>
<td>17.26</td>
<td>21.90</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>0.3</td>
<td>1.0</td>
<td>0.75</td>
<td>1.25</td>
<td>1.0</td>
<td>1.5</td>
</tr>
</tbody>
</table>

Applying (4.75) to (4.82), we find

\[ a = 4.227 \quad b = 0.878 \]

\[ \sigma(a) = 0.044 \quad \sigma(b) = 0.203 \quad \text{and} \]

\[ \text{cov}(a, b) = -0.0629. \]