Terascale Analysis Center: Statistical Tools Group

20. January 2010 - DESY

Constrained Least Squares Methods

with

Correlated Data and Systematic Uncertainties

Volker Blobel – Universität Hamburg

. . . constrained least squares as a natural method
and a more general alternative to χ^2-function minimization

Keys during display: enter = next page; → = next page; ← = previous page; home = first page; end = last page (index, clickable); C-← = back; C-N = goto page; C-L = full screen (or back); C-+ = zoom in; C-− = zoom out; C-0 = fit in window; C-M = zoom to; C-F = find; C-P = print; C-Q = exit.

1. Alternative least squares methods for fitting/averaging

$\boldsymbol{x}_m$ = measured variables, with covariance matrix $\boldsymbol{V}_x$
$\boldsymbol{x}_u$ = unmeasured variables, "parameters"
$\boldsymbol{x} = (\boldsymbol{x}_m, \boldsymbol{x}_u)$ = measured *and* unmeasured variables $\quad$ $\boldsymbol{t}$ = independent coordinates

χ^2-function minimization

$$S(\boldsymbol{x}_u) = \sum_i \frac{((\boldsymbol{x}_m)_i - f(t_i, \boldsymbol{x}_u))^2}{\sigma_i^2} = \min$$
$$\rightarrow \boldsymbol{r}^{\mathrm{T}} \boldsymbol{V}_x^{-1} \boldsymbol{r} = \min$$

Residuals $\boldsymbol{r}$: χ^2-function to be minimized is sum of squares of residuals; problems, if residuals

- depend on > 1 measurement, and/or depend on > 1 error contribution, especially contributions changing the normalization.

Constrained Least Squares

$$S(\Delta\boldsymbol{x}_m) = \Delta\boldsymbol{x}_m^{\mathrm{T}} \boldsymbol{V}_x^{-1} \Delta\boldsymbol{x}_m = \min$$
$$h_j\left(\boldsymbol{x}_m + \Delta\boldsymbol{x}_m,\, \boldsymbol{x}_u + \Delta\boldsymbol{x}_u,\, \boldsymbol{t}\right) = 0 \quad j = 1,\, 2 \ldots m$$

Individuals corrections $\Delta\boldsymbol{x}_m$ for measured variables: expression to be minimized is sum of squares of corrections.

- Constraints $h_j(\boldsymbol{x}) = 0$ may be implicit expressions;
- bias reduced or avoided.

Both alternatives are equivalent, with identical results, for simple problems. In both alternatives the data may be correlated and the functions/constraints may be non-linear.

Comparison

χ^2-function minimization, e.g. using MINUIT

- User has to provide the function $S(\boldsymbol{x})$, which is "seen" by MINUIT. The user function includes all data, uncertainties, the physical and statistical model.
- MINUIT calculates by finite differences the first derivative of $S(\boldsymbol{x})$, and approximates, using the VM method, the full Hessian in $\geq n$ iterations for linear and non-linear problems.
- Variables are the parameters = unmeasured variables.

Constrained least squares, e.g. using APLCON 2.0

- User describes set of variables incl. covariance matrix, and individual model functions $h_j(\boldsymbol{x})$.
- APLCON calculates by finite differences the first derivative of all individual model functions $h_j(\boldsymbol{x})$, which allows to calculate the full Hessian during each iteration (Gauss-Newton matrix).
- Many variables: measured and unmeasured variables plus Lagrange multipliers.
- Principle used in HEP for > 50 years, mainly with kinematical constraints for particle reactions and decays; APLCON 1.0 in use for 33 years.

Constrained least squares fit program APLCON

$$\begin{array}{ll} \text{minimize} & \Delta\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}_x^{-1}\Delta\boldsymbol{x} \\ \text{subject to} & h_j\left(\boldsymbol{x},\boldsymbol{t}\right)=0 \quad j=1,2\ldots m \end{array}$$

Properties:

- Extreme form of constrained least squares, with separation into a quadratic expression, and a set of constraints $h_j(\boldsymbol{x})$ with all nonlinearities; solved using Lagrange multipliers;
- simple to use: derivatives calculated by numerical methods, no step definition necessary, no principle distinction between measured ($\mathbf{X}_m$) and unmeasured variables ($\mathbf{X}_u$); full initial and final covariance matrix, and pulls;
- Extension to **non-Gaussian** variables: selected variables can be treated e.g as Poisson- or log-normal-distributed;
- Extension to advanced analysis of uncertainties: profile likelihood
- APLCON is a method for difficult problems to follow accurately the assumed physical and statistical model of the measurement process, and to avoid a **bias** in the result;
- APLCON 1.0 and earlier test version 2.0 (Fortran) available from `www.desy.de/~blobel`

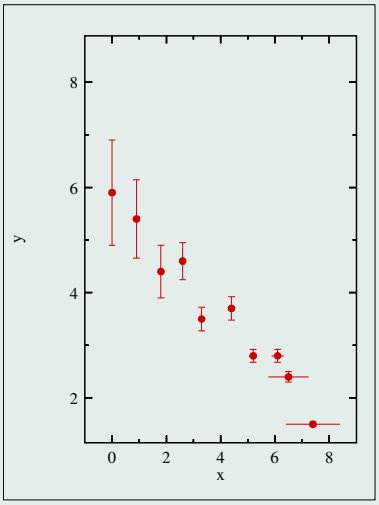
From publications: "In practice, the added technical complexity of a constrained fit with extra free parameters is not justified ..."

"The application of Lagrange multipliers is unnecessarily complicated and the linear approximation requires additional assumptions and iterations."

2. x-y-data with uncertainties in both coordinates

The subject is discussed by Press et. al. (Numerical Recipes) with the remarks:

> *"If experimental data are subject to measurement error not only in the y_i's, but also in the x_i's, then the task of fitting a straight-line model*
>
> $$y(x) = a + bx$$
>
> *is considerably harder ... Be aware that the literature on the seemingly straightforward subject of this section is generally confusing and sometimes plain wrong."*



(data in figure from C.A. Cantrell)

What is the uncertainty of residual $r_i = y_i - (a + b\,x_i)$?

C.A. Cantrell [*Atmos. Chem. Phys.*, **8**, 5477-5487, 2008] lists > 30 publications for methods (including methods giving wrong results), only for straight-line fits, almost all for uncorrelated data only.

$\rightsquigarrow$ try "Deming regression" or "error-in-variables-model (EIV)" or "total least squares (TLS)" in Google.

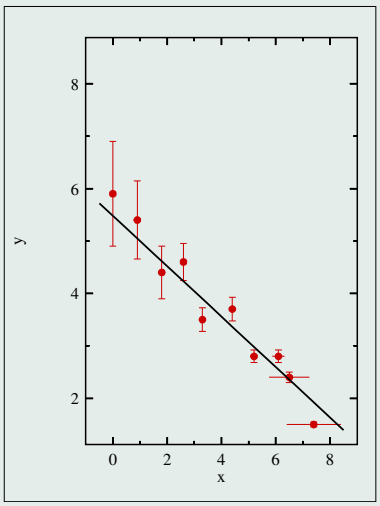
Example 1: uncertainties in both coordinates

Code:

X := ... *(variable array)*
$\mathbf{V}_x$:= ... *(matrix array)*
`aplcon(2*N+2,N)`
do
{
  for $j = 1$ to N
  {
    `h(j)` $= a + b \cdot x_j - y_j$
  }
  `aploop(X,VX,h,irep)`
} **while** (`irep` < 0);
result in **X** *and* $\mathbf{V}_x$

variable array

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_m \\ \hline \mathbf{X}_u \end{pmatrix} = \begin{pmatrix} x_1 \\ y_1 \\ x_2 \\ y_2 \\ \vdots \\ x_N \\ y_N \\ \hline a \\ b \end{pmatrix}$$



Note: order of measured and unmeasured variable irrelevant – distinguished by **zero** elements in input covariance matrix $\mathbf{V}_x$.

If measurement of slope b exists before: add variance of b to V, with no change in the program code

Example 2: Straight line and correlated data

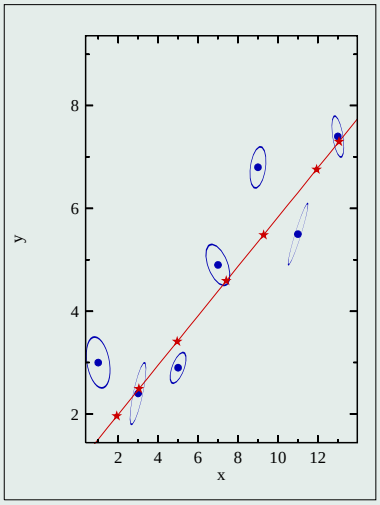
Now: correlation between x and y in data $\neq 0$, and fit of straight line required

Code:

X := ... *(variable array)*
$\mathbf{V}_x$:= ... *(matrix array)*
`aplcon(2*N+2,N)`
do
{
  for $j = 1$ to N
  {
    `h(j)` $= a + b \cdot x_j - y_j$
  }
  `aploop(X,VX,h,irep)`
} **while** (`irep` < 0);
result in **X** *and* $\mathbf{V}_x$

variable array

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_m \\ \hline \mathbf{X}_u \end{pmatrix} = \begin{pmatrix} x_1 \\ y_1 \\ x_2 \\ y_2 \\ \vdots \\ x_N \\ y_N \\ \hline a \\ b \end{pmatrix}$$



add off-diagonal elements to $\boldsymbol{V}_x$
no change of code

red star is fitted xy-value

Example 3: Parabola and correlated data

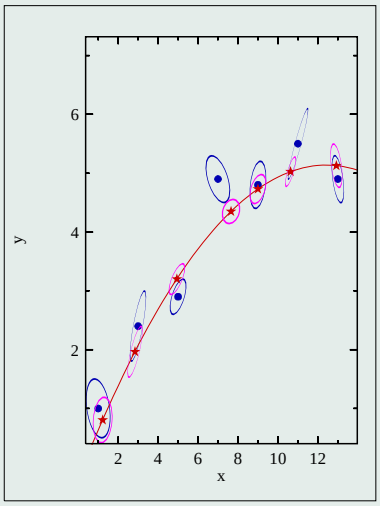
Correlation between x and y in data $\neq 0$, and fit of parabola required

Code:

X := ... *(variable array)*
$\mathbf{V}_x$:= ... *(matrix array)*
`aplcon(2*N+3,N)`
do
{
  for $j = 1$ to N
  {
    `h(j)` $= a + b \cdot x_j + c \cdot {x_j}^2 - y_j$
  }
  `aploop(X,VX,h,irep)`
} **while** (`irep` < 0);
result in **X** *and* $\mathbf{V}_x$

variable array

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_m \\ \hline \mathbf{X}_u \end{pmatrix} = \begin{pmatrix} x_1 \\ y_1 \\ x_2 \\ y_2 \\ \vdots \\ x_N \\ y_N \\ \hline a \\ b \\ c \end{pmatrix}$$



Only small change of code: include c in **X** and add $+ c \cdot {x_j}^2$ to `h(j)`

red star is fitted xy-value

3. Uncertainties of fit parameters

APLCON provides

full covariance matrix $\boldsymbol{V}_x$ for combined variables: fitted values of measured variables and of unmeasured variables ("parameters"), from the inverse of the Hessian (by the law of propagation of uncertainties);

pulls for all measured variables: should follow $N(0,1)$ distributions;

- Covariance matrix is accurate in simple cases: measured data Gaussian and constraints linear, or asymptotically in the limit of ∞ data;
- Matrix may be inaccurate (and non-Gaussian) for non-Gaussian data, constraints from non-linear models and low statistic $\rightsquigarrow$ statistically improved information is required on confidence intervals for important parameters.

confidence intervals on selected parameters by **profile analysis** (optional):
realized by repeated fits with one additional internal constraint;

contours for selected parameters pairs by **profile analysis** (optional):
realized by repeated fits with two additional internal constraints.

Confidence intervals

Confidence intervals $100(1-\alpha)\%$ on a parameter of a theory should have **coverage** (frequentists approach): interval includes true parameter at least $100(1-\alpha)\%$ of the time in repeated experiments.

Data model depends on k parameters $\boldsymbol{\pi}$ of interest, but also on additional nuisance parameters $\boldsymbol{\theta}$: full likelihood function $\mathcal{L}(\boldsymbol{\pi},\boldsymbol{\theta})$.

$$\textbf{profile likelihood } \lambda(\boldsymbol{\pi}_0) = \frac{\sup\{\mathcal{L}(\boldsymbol{\pi}_0,\boldsymbol{\theta})\}}{\sup\{\mathcal{L}(\boldsymbol{\pi},\boldsymbol{\theta})\}} = \frac{\text{suprenum over subspace with } \boldsymbol{\pi}=\boldsymbol{\pi}_0}{\text{suprenum over full parameter space}}$$

The profile likelihood λ is a function of $\boldsymbol{\pi}_0$ only, and $-2\log\lambda$ converges in distribution to a χ^2 random variable with $n_{\text{df}} = k$.

Coverage of confidence intervals computed by the profile likelihood is usually surprisingly good.[⋆]

APLCON provides two options for determination of confidence intervals:

$k=1$**: One-dimensional profile likelihood:** fit of $n-1$ parameters for many fixed values of a single parameter (like MINOS in MINUIT)

$k=2$**: Two-dimensional profile likelihood:** fit of $n-2$ parameters for many fixed points in the 2-parameter plane (like MNCONTOURS in MINUIT)

⋆ Wolfgang A. Rolke et al., Limits and confidence intervals in the presence of nuisance parameters, NIM A 551 (2005) 493 – 503

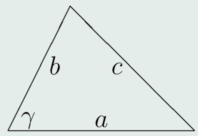
Example 4: Triangle parameters

Assume that three sides a, b and c of a triangle and one angle γ are measured. Three values are sufficient for a complete definition of a triangle. Thus the least squares method can be used to *improve* the measured values, ... and to determine the triangle area A.

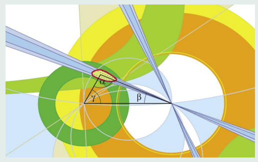
Code:

$\mathbf{X}$:= ... *(variable array)*
$\mathbf{V}_x$:= ... *(matrix array)*
`aplcon(5,2)` `aprofl(5,0)` `aprofl(5,2)`
do
{
  $p = (a + b + c)/2$! half the circumference
  $S = \sqrt{p(p-a)(p-b)(p-c)}$! area of triangle
  `h(1)` $= \tan(\gamma/2) - S/(p(p-c))$! angle constraint
  `h(2)` $= A - S$! area constraint
  `aploop(X,VX,h,irep)`
} **while** (`irep` < 0);

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_m \\ \hline \mathbf{X}_u \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \\ \gamma \\ \hline A \end{pmatrix}$$



variable	measured		fit result		pull
a	10	±0.05	10.01	±0.05	1.75
b	7	±0.2	7.06	±0.20	1.75
c	9	±0.2	8.72	±0.12	−1.75
γ	1	±0.02	1.019	±0.017	1.75
A			30.10	±0.87	



⇝ e.g. unitarity triangle, representing interactions between quarks

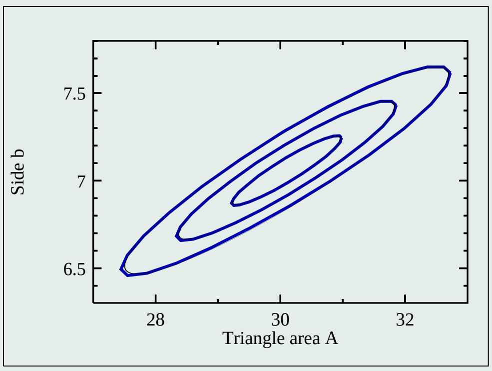
... more results from triangle fit

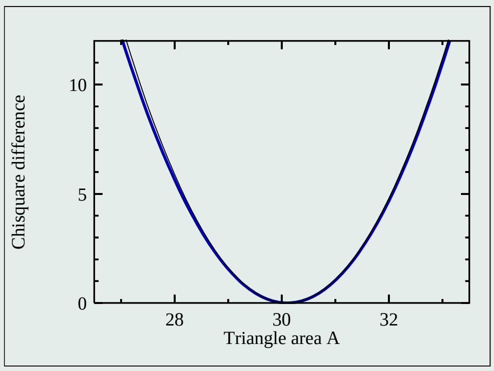
The parameter of interest is assumed to be the triangle area A. This parameter has no influence on the fit, but it is determined incl. the propagation of uncertainties because there is a constraint for A.

```
Matrix of correlation coefficients
   i     1   2   3   4   5
   1   100
   2    -2 100
   3    20  23 100
   4   -10 -11  90 100
   5    11  93  57  23 100
   i     1   2   3   4   5
(coefficients in %)
```

```
    Confidence intervals profile analysis for parameter 5: area
    Prob   sigmas ____________________________________  sigma units
0                 29.233       30.099       30.966
1   68.3 %  1.00  29.217        ...         30.979     -1.02   1.01
2   90.0 %  1.65  28.646        ...         31.544     -1.68   1.67
3   95.0 %  1.96  28.366        ...         31.819     -2.00   1.99
4   99.0 %  2.58  27.820        ...         32.357     -2.63   2.61
5   99.5 %  2.81  27.614        ...         32.559     -2.87   2.84
6   99.9 %  3.29  27.183        ...         32.979     -3.37   3.32
```

Contour $A - b$ from a **2-dim profile analysis**: Confidence interval from a **1-dim profile analysis**:





4. Averaging and systematic uncertainties

$$x_{\rm ave} = \sum_i w_i\, x_i \qquad \text{with} \quad \sum_i w_i = 1 \qquad (\rightsquigarrow x_{\rm ave} \text{ unbiased, if } x_i \text{ unbiased})$$

Definition of optimal weights w_i with minimal variance $\sigma^2_{\rm ave}$ from least squares requirement:

Uncorrelated data $x_i \pm \sigma_i$:

$$w_i = \left(\sum_i \frac{1}{\sigma_i^2}\right)^{-1} \cdot \frac{1}{\sigma_i^2} \qquad\qquad \sigma^2_{\rm ave} = \left(\sum_i \frac{1}{\sigma_i^2}\right)^{-1}$$

Correlated data x_i **with covariance matrix** $\boldsymbol{V}_x$**:** needs inverse $\boldsymbol{V}_x^{-1}$

$$w_i = \left(\sum_{j,k} \left(\boldsymbol{V}_x^{-1}\right)_{jk}\right)^{-1} \cdot \sum_j \left(\boldsymbol{V}_x^{-1}\right)_{ij} \qquad\qquad \sigma^2_{\rm ave} = \sum_{ij} w_i w_j \left(\boldsymbol{V}_x\right)_{ij}$$

Common additive uncertainty : $x_i \pm \sigma_i \pm \Delta$ (identical *systematic* error Δ)

- (PDG:) first average $x_i \pm \sigma_i$, then combine error with Δ^2, or
- (PDG:) apply factor $(1 + \Delta^2 \left(\sum_i 1/\sigma_i^2\right))^{1/2}$ to all errors, and treat as uncorrelated,or
- define covariance matrix with $(\boldsymbol{V}_x)_{ii} = \sigma_i^2 + \Delta^2 \qquad (\boldsymbol{V}_x)_{ij} = \Delta^2 \quad i \neq j$

  all three methods are equivalent.

Common multiplicative uncertainty: e.g. $(x_i \pm \sigma_i)\,(1 \pm \Delta)$ $\rightsquigarrow$ more complicated, discussed later

Average of two correlated data

Covariance matrix $\boldsymbol{V}_x$ and its inverse $\boldsymbol{V}_x^{-1}$ (weight matrix) depend on σ_1, σ_2 and ρ:

$$\boldsymbol{V}_x = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \qquad \boldsymbol{V}_x^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} 1/\sigma_1^2 & -\rho/(\sigma_1\sigma_2) \\ -\rho/(\sigma_1\sigma_2) & 1/\sigma_2^2 \end{pmatrix}$$

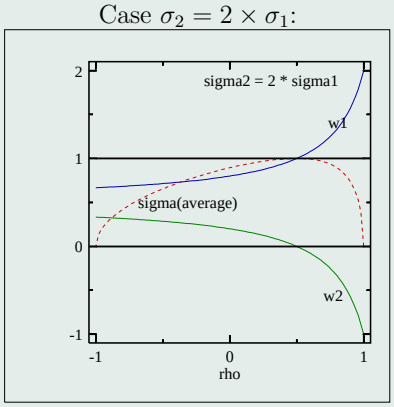
Average is $x_{\rm ave} = w_1 x_1 + w_2 x_2$, with

$$w_1 = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \qquad w_2 = \frac{\sigma_1^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

$$\sigma^2_{\rm ave} = \left(1-\rho^2\right) \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \qquad (\text{dashed red curve} \Longrightarrow)$$

- Weight $w_2 < 0$ for large correlation $\rho > +\frac{\sigma_1}{\sigma_2}$
  $\rightsquigarrow$ **not meaningful**
- $x_{\rm ave} \equiv x_1$ and $\sigma_{\rm ave} \equiv \sigma_1$, **no improvement** for $\rho = \frac{\sigma_1}{\sigma_2}$,
- expected difference^2 $E\left[(x_1 - x_2)^2\right] = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$
- smaller value of $\sigma_{\rm ave}$ for negative correlation.

Case $\sigma_2 = 2 \times \sigma_1$:



$\sigma_{\rm ave}/\sigma_1$ and weights vs ρ

Averaging of two values, assumed to be uncorrelated ($\rho = 0$): expected χ^2-value $= 1.0$, at $n_{\rm df} = 1$.
A small χ^2-value (< 1) can be caused by

- overestimated errors: ($\rightsquigarrow$ reduce $\sigma_{\rm ave}$?), or it can indicate an
- unknown positive correlation between the two values: $\rightsquigarrow$ increase $\sigma_{\rm ave}$– no gain in accuracy!

Example 5: Averaging correlated data

50 years old data on the isospin 1/2 and 3/2 scattering lengths in πp-scattering in the s-state:

Experiment (1): $a_1 = 0.170 \pm 0.0240$; $a_3 = -0.107 \pm 0.0197$; corr. coefficient $\rho = -39.1\%$.
Experiment (2): $a'_3 = -0.104 \pm 0.006$.

Input to the APLCON fit to average the two a_3-values and, at the same time, improve the correlated a_1:

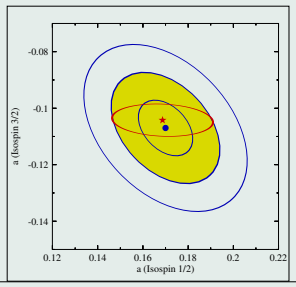
$$\boldsymbol{x}_m = \begin{pmatrix} a_1 \\ a_3 \\ a'_3 \end{pmatrix} = \begin{pmatrix} 0.170 \pm 0.0240 \\ -0.107 \pm 0.0197 \\ -0.104 \pm 0.0060 \end{pmatrix} \qquad \boldsymbol{V}_x = \begin{pmatrix} 0.580 & -0.185 & 0 \\ -0.185 & 0.388 & 0 \\ 0 & 0 & 0.036 \end{pmatrix} \times 10^{-3}$$

and after the code $h_1 = a_3 - a'_3$ the result by APLCON is

$$\boldsymbol{x} = \begin{pmatrix} a_1 \\ a_3 \\ a'_3 \end{pmatrix} = \begin{pmatrix} 0.169 \pm 0.0220 \\ -0.1043 \pm 0.0057 \\ -0.1043 \pm 0.0057 \end{pmatrix} \qquad \boldsymbol{V}_x = \left(\begin{array}{cc|c} 0.499 & -0.0157 & -0.0157 \\ -0.0157 & 0.0329 & 0.0329 \\ \hline -0.0157 & 0.0329 & 0.0329 \end{array}\right) \times 10^{-3} .$$

Plot of the two scattering length a_1 and a_3:

- The yellow ellipse is 1-σ contour of experiment (1).
- The star is the average with ellipse indicating the 1-σ contour of the average.



Note: $\chi^2/n_{\rm df} = 0.02123$ means p-value of 88.4 %.

Systematics - either additional variables ... ... or contribution to $\boldsymbol{V}_x$

Data with common additive systematic uncertainty

... as correlated data ...
$x_1 = x'_1 + a$
$x_2 = x'_2 + a$

$$a = \begin{array}{rl} x'_1 & \pm \sigma_1 \\ x'_2 & \pm \sigma_2 \\ 0 & \pm \Delta \end{array} \qquad \boldsymbol{V}\begin{pmatrix} x'_1 \\ x'_2 \\ a \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \Delta^2 \end{pmatrix}$$

Equivalent: non-diagonal covariance matrix by law of (linear) propagation of uncertainties:

$$\boldsymbol{V}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \boldsymbol{V}\begin{pmatrix} x'_1 \\ x'_2 \\ a \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 + \Delta^2 & \Delta^2 \\ \Delta^2 & \sigma_2^2 + \Delta^2 \end{pmatrix}$$

Data with common multiplicative systematic uncertainty

... as correlated data ...
$x_1 = x'_1 \times a$
$x_2 = x'_2 \times a$

$$a = \begin{array}{rl} x'_1 & \pm \sigma_1 \\ x'_2 & \pm \sigma_2 \\ 1 & \pm \Delta \end{array} \qquad \boldsymbol{V}\begin{pmatrix} x'_1 \\ x'_2 \\ a \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \Delta^2 \end{pmatrix}$$

Non-diagonal covariance matrix by law of propagation of uncertainties: non-linear transformation

$$\boldsymbol{V}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a & 0 & x_1 \\ 0 & a & x_2 \end{pmatrix} \boldsymbol{V}\begin{pmatrix} x'_1 \\ x'_2 \\ a \end{pmatrix} \begin{pmatrix} a & 0 \\ 0 & a \\ x_1 & x_2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 + x_1^2\Delta^2 & x_1x_2\Delta^2 \\ x_1x_2\Delta^2 & \sigma_2^2 + x_2^2\Delta^2 \end{pmatrix} \quad \text{with } a = 1$$

Elements of the transformation matrix are not constant; the two representations are **not equivalent**.

Averaging with normalization uncertainty "χ^2-function" minimization

In a publication (NIM A) the following measurement for two data points x_1, x_2 and a common normalization factor α with uncertainty ϵ is given:

$$x_1 = 8.0 \pm 2\% \qquad x_2 = 8.5 \pm 2\% \qquad \alpha = 1 \pm \epsilon \qquad \text{with} \quad \epsilon = 0.1$$

"Assuming that the two measurements refer to the same physical quantity, the best estimate of its true value can be obtained by fitting the points to a constant" (from the publication).

A simple straightforward average would be $x_{\text{ave}} = (x_1 + x_2)/2 = 8.25$, but ...

Publication: average x_{ave} by "χ^2-function minimization", the covariance matrix $\boldsymbol{V}$ is defined to include the normalization uncertainty:

$$\chi^2 = \boldsymbol{\Delta}^{\mathrm{T}} \boldsymbol{V}^{-1} \boldsymbol{\Delta} = \text{minimum} \qquad \text{with} \quad \boldsymbol{V} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} + \epsilon^2 \cdot \begin{pmatrix} x_1^2 & x_1 x_2 \\ x_1 x_2 & x_2^2 \end{pmatrix}$$

($\boldsymbol{\Delta}$ is "the vector of the differences" between x_i and average x_{ave}).

Resulting average is $x_{\text{ave}} = 7.87 \pm 0.81$, outside (!) the range of the two input values

... apparently wrong $\rightsquigarrow$ large bias with constructed non-diagonal covariance matrix. ⇒**more**

Note: weights $w_1 = +1.25$ and $w_2 = -0.25$ because $\sigma_1 < \sigma_2$;

Example 6: Normalization uncertainty

With two constraints the average x_{ave} is forced to agree with the two measurements, multiplied by the normalization factor α:

Code:

X := ... *(variable array)*
$\mathbf{V}_x$:= ... *(matrix array)*
`aplcon(4,2)`
do
{
`h(1)` = $\alpha \cdot x_1 - x_{\text{ave}}$
`h(2)` = $\alpha \cdot x_2 - x_{\text{ave}}$
`aploop(X,VX,h,irep)`
} **while** (`irep` < 0);

variable array

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_m \\ \hline \mathbf{X}_u \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \alpha \\ \hline x_{\text{ave}} \end{pmatrix}$$

variable	measured		fit result		pull
x_1	8.0	±2%	8.235	±0.116	2.14
x_2	8.5	±2%	8.235	±0.116	−2.14
α	1	±10%	1.000	±0.100	−2.14
x_{ave}			8.235	±0.832	

$\rightsquigarrow$ no problem with normalization uncertainty with constrained least squares.

5. Non-Gaussian variables

Least squares requires data with constant variance, independent of fit result.
What happens, if x_i are not normal distributed or do not have constant variance? $\rightsquigarrow$ **Bias**

Example: Average of data following (or proportional to) Poisson distribution

$x_1 = \ 9 \pm 3$
$x_2 = 16 \pm 4$

Weighted mean (LS) $x_{\text{ave}} = 11.52 \pm 2.40$
Using Poisson statistic (ML) $x_{\text{ave}} = 12.5 \pm 2.5$

APLCON can treat Poisson distributed measured variables using ML formalism, avoiding the bias, by `apoiss`(*index*).

Data in HEP are often given with uncertainty in %, i.e relative uncertainty.

This indicates the log-normal (instead of the normal) distribution with constant *relative* uncertainty.

Normalization factors will approximately follow the log-normal distribution, as a consequence of the Central Limit-Theorem: product of many factors with small uncertainty.

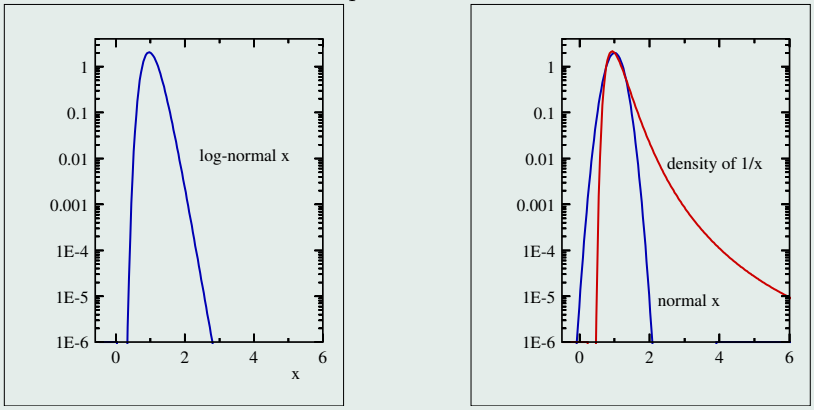
APLCON can treat log-normal distributed measured variables by a transformation, avoiding a potential bias.

Transformations of variables in APLCON

Log-normal distribution for e.g normalization factors: `aplogn`(*index*)

log-normal variable (with uncertainty $\propto$ value): external $\alpha \ \Rightarrow \ \exp[\ \alpha'\ ]$
with new internal variable $\alpha' \ \equiv \ \ln\alpha$

Example: $\alpha = 1 \pm 0.2$



Similar: square-root-transformation for variable with uncertainty $\propto \sqrt{\text{value}}$ by `apsqrt`(*index*).

6. Cross section measurement

Least squares popular in particle physics for cross section fits and averaging, using data from ≥ 1 experiment.
Cross sections x_i are measured via counted numbers $n_i = S \cdot x_i$ of events:

$$\text{cross section} \quad x_i = S^{-1} \cdot n_i, \quad i = 1, \ldots$$

where the sensitivity factor S is a product

$$S = A_1 \cdot A_2 \cdots A_a \cdot \int \mathcal{L}\, \mathrm{d}t \cdot \Delta x$$

of many factors (trigger, detection, reconstruction ... probabilities, luminosity, bin width).

number n_i: follows Poisson distribution,

sensitivity S: will follow a log-normal distribution (log of S normal distributed) – the inverse S^{-1} will follow a log-normal distribution too:

$$\text{cross section } x_i \; = S^{-1} \text{ (log-normal)} \times n_i \text{ (Poisson)} \qquad \text{– possible with APLCON}$$

cross section: in practice assumed to follow the normal distribution with x_i assumed to be independent (with diagonal covariance matrix); even resolution-corrected ("unfolded") cross sections usually assumed to be independent (!); in addition there is a common normalization factor $\alpha = 1$ with uncertainty ϵ.

Residual vs Constraint – in case of systematic uncertainties

Cross section data x_i rely on e.g. energy measurement by calorimeter with uncertainty in calibration. Method: repeat determination of x_i with calibration constant changed by ± 1 standard deviation, to obtain $x_i^{(-)}$ and $x_i^{(+)}$, and estimate systematic uncertainty v_i.

$$\text{Additive systematic uncertainty} \quad v_i = \frac{x_i^{(+)} - x_i^{(-)}}{2} \qquad i = 1, \ldots$$

Residual $r_i = \alpha \, (\, x_i + \sum_k \beta_k \, v_{ik} \,) - f_i$

(α: normalization $1 \pm \epsilon$; β_k: factor 0 ± 1; f_i: expectation)

Constraint:

$$\alpha \, (\, x_i + \sum_k \beta_k \, v_{ik} \,) - f_i = 0$$

Residuals r_i are influenced by uncertainties in ≥ 3 measured quantities: x_i, α and β_k $\rightsquigarrow$ standard deviation of residual unclear ... $\rightsquigarrow$ χ^2-function minimization impossible or difficult, with potential bias. (in practice complicated expressions for variance in denominator to avoid or reduce bias in result).

Constraints: individual corrections fitted for each variable: x_i, α, β_k ... with individual variance.

Cross section averaging – example with "χ^2-function" minimization

... from: Measurement of the Inclusive ep Scattering Cross Section at Low $\boldsymbol{Q}^2$ and $\boldsymbol{x}$ at HERA, H1 Collaboration, DESY 08-171 and arXiv :0904.0929.

The averaging and the phenomenological analysis of the data is done using the χ^2 definition eq. (31)

$$\chi^2_{\text{exp}}\left(\boldsymbol{m},\boldsymbol{b}\right) = \sum_i \frac{\left[m^i - \sum_j \gamma^i_j m^i b_j - \mu^i\right]^2}{\delta^2_{i,\text{stat}}\mu^i\left(m^i - \sum_j \gamma^i_j m^i b_j\right) + \left(\delta_{i,\text{uncor}} m^i\right)^2} + \sum_j {b_j}^2 \,.$$

"*... Correlated and uncorrelated systematic errors are to a good approximation proportional to the central values (multiplicative errors), whereas the statistical errors scale with the square root of the expected number of events.*" The χ^2 definition should avoid "*a small bias to lower cross sections since the measurements with lower central values have smaller absolute uncertainties.*"

Measured value is μ^i with statistical and uncorrelated systematic uncertainties $\Delta_{i,\text{stat}}$ and $\Delta_{i,\text{uncor}}$. Relative correlated systematic, statistical and uncorreleated uncertainties are $\gamma^i_j = \Gamma^i_j/\mu^i$, $\delta_{i,\text{stat}} = \Delta_{i,\text{stat}}/\mu^i$ and $\delta_{i,\text{uncor}} = \Delta_{i,\text{uncor}}/\mu^i$. The underlying physical quantities are m_i (vector $\boldsymbol{m}$).

There are correlated systematic error sources of type j, with a central value α_j and uncertainty Δ_{α_j}, where $\partial\mu^i/\partial\alpha_j$ quantifies the sensitivity of the measurement μ^i to the systematic source j. Summation over j extends over all correlated systematic sources. The variables $b_j = (a_j - \alpha_j)/\Delta_{\alpha_j}$ and $\Gamma^i_j = (\partial\mu^i/\partial\alpha_j)\Delta_{\alpha_j}$ are introduced.

Data sets are consistent: $\chi^2/n_{\text{df}} = 19.5/39$ and $86.2/125$, corresponding to p-values of 99.62 % and 99.68 %.

Summary

Constrained least squares fit as alternative to "χ^2-function"?

Properties of constrained least squares fit program Aplcon:

- Aplcon avoids potential bias of fit result in complicated cases of "residuals" from > 1 measured quantities, and of systematic uncertainties, and background;
- allows to perform a constrained simultaneous fit, taking into account the non-Gaussian character of certain variables, and providing confidence intervals by profile analysis;
- equivalent to "χ^2-function" minimization in those cases, where "χ^2-function" minimization can be applied;
- Hessian matrix equivalent to Gauss-Newton matrix, no variable-metric iteration necessary – fewer iterations than MINUIT; ... but requires larger memory space and, due to numerical derivatives of many variables, slower than Minuit in "equivalent" cases;
- clear and general method to follow assumed model of measurement process, simple to use.

Solution of problem with constraints

Method of Lagrange multipliers, introducing $\lambda_j,\ j = 1,\ 2 \ldots m$:

Linearization of $h_j(\boldsymbol{x}_m, \boldsymbol{x}_u) = 0\ \ j = 1,\ 2 \ldots m:$ $\qquad \boldsymbol{A}_m \Delta\boldsymbol{x}_m + \boldsymbol{A}_u \Delta\boldsymbol{x}_u - \boldsymbol{c} = 0 \qquad (\boldsymbol{A})_{ji} = \dfrac{\partial h_j(\boldsymbol{x})}{\partial(\boldsymbol{x})_i}$

$$\text{Lagrange function} \quad \mathcal{L}(\boldsymbol{\Delta x}_m, \boldsymbol{\Delta x}_u, \boldsymbol{\lambda}) = \boldsymbol{\Delta x}_m^{\mathrm{T}} \boldsymbol{V}_x^{-1} \boldsymbol{\Delta x}_m + 2\boldsymbol{\lambda}^{\mathrm{T}} \left(\boldsymbol{A}_m \Delta\boldsymbol{x}_m + \boldsymbol{A}_u \Delta\boldsymbol{x}_u - \boldsymbol{c}\right)$$

Matrix equation to be solved for new corrections $\boldsymbol{\Delta x}_m$, $\boldsymbol{\Delta x}_u$, and $\boldsymbol{\lambda}$:

$$\begin{pmatrix} \boldsymbol{V}_m^{-1} & 0 & \boldsymbol{A}_m^{\mathrm{T}} \\ 0 & 0 & \boldsymbol{A}_u^{\mathrm{T}} \\ \boldsymbol{A}_m & \boldsymbol{A}_u & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\Delta x}_m \\ \boldsymbol{\Delta x}_u \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \boldsymbol{c} \end{pmatrix}$$

Matrix inversion: first part skipped ($\boldsymbol{V}_m^{-1}$), remaining part inverted, making use of symmetry of matrix:

$$\left(\begin{array}{ccc} \boldsymbol{V}_m^{-1} & \boldsymbol{0} & \boldsymbol{A}_m^{\mathrm{T}} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_u^{\mathrm{T}} \\ \boldsymbol{A}_m & \boldsymbol{A}_u & \boldsymbol{0} \end{array}\right) \Longrightarrow \left(\begin{array}{c|cc} -\boldsymbol{V}_m & \boldsymbol{0} & \boldsymbol{V}_m \boldsymbol{A}_m^{\mathrm{T}} \\ \hline \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_u^{\mathrm{T}} \\ \boldsymbol{A}_m \boldsymbol{V}_m & \boldsymbol{A}_u & -\boldsymbol{A}_m \boldsymbol{V}_m \boldsymbol{A}_m^{\mathrm{T}} \end{array}\right) \Longrightarrow \text{inverse} = \text{covar. matrix}$$

Bias in averaging — "χ^2-function" minimization

Determine average x_{ave} of data x_j, which have common normalization factor α with uncertainty ϵ. Best estimate for x_{ave} is $\overline{x} = \sum_j x_j/n$ with normalization factor α unchanged (no information on α from averaging).

$$s^2 = \frac{1}{n}\sum_j (x_j - \overline{x})^2 = \frac{1}{n}\sum_j x_j^2 - (\overline{x})^2 \quad \rightsquigarrow \quad \frac{1}{n}\sum_j x_j^2 \approx (\overline{x})^2 + \sigma^2$$

χ^2-function $\qquad F(x_{\text{ave}}, \alpha) = \displaystyle\sum_j \left(\frac{\alpha \cdot x_j - x_{\text{ave}}}{\sigma}\right)^2 + \left(\frac{\alpha - 1}{\epsilon}\right)^2$

The minimum of $F(x_{\text{ave}}, \alpha)$ is determined from the two derivative conditions:

$$\frac{1}{2}\frac{\partial F}{\partial \alpha} = \sum_j \left(\frac{\alpha x_j - x_{\text{ave}}}{\sigma^2}\right) x_j + \left(\frac{\alpha - 1}{\epsilon^2}\right) = 0 \qquad \frac{1}{2}\frac{\partial F}{\partial x_{\text{ave}}} = -\sum_j \left(\frac{\alpha x_j - x_{\text{ave}}}{\sigma^2}\right) = 0$$

The second equation gives the estimate $x_{\text{ave}} = \alpha \cdot \overline{x}$. The estimate $\widehat{\alpha}$ for the normalization factor is obtained from the first equation; the result is biased:

$$\widehat{\alpha} = \frac{1}{1 + n\epsilon^2} \qquad x_{\text{ave}} = \frac{1}{1 + n\epsilon^2}\,\overline{x}\ .$$

The two χ^2-functions below give unbiased results (but incorrect parameter uncertainties from the Hessian):

$$F(x_{\text{ave}}, \alpha) = \sum_j \left(\frac{\alpha \cdot x_j - x_{\text{ave}}}{\alpha \cdot \sigma}\right)^2 + \left(\frac{\alpha - 1}{\epsilon}\right)^2 \qquad F(x_{\text{ave}}, \alpha) = \sum_j \left(\frac{x_j - \alpha \cdot x_{\text{ave}}}{\sigma}\right)^2 + \left(\frac{\alpha - 1}{\epsilon}\right)^2 .$$

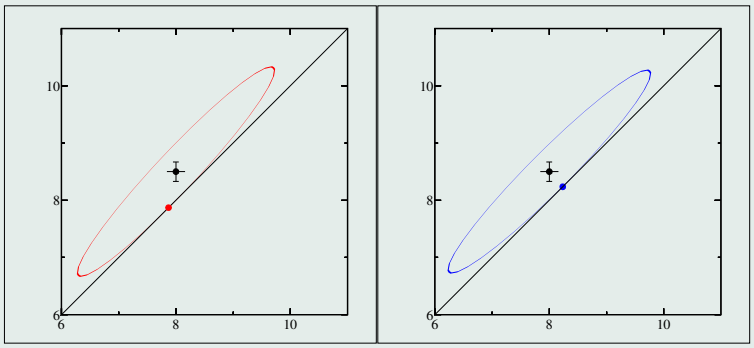
Covariance matrix plot

χ^2-function" minimization

Axis of covariance ellipse is slightly tilted (left) because input values x_1 and x_2 (and σ_1, σ_2) are not equal; this causes the "strange" value of the average.

$$\chi^2 = \boldsymbol{\Delta}^{\mathrm{T}} \boldsymbol{V}^{-1} \boldsymbol{\Delta} = \text{minimum} \qquad \text{with} \quad \boldsymbol{V} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} + \epsilon^2 \cdot \begin{pmatrix} x_1^2 & x_1 x_2 \\ x_1 x_2 & x_2^2 \end{pmatrix}$$

($\boldsymbol{\Delta}$ is 'the vector of the differences" between x_i and average x_{ave}).

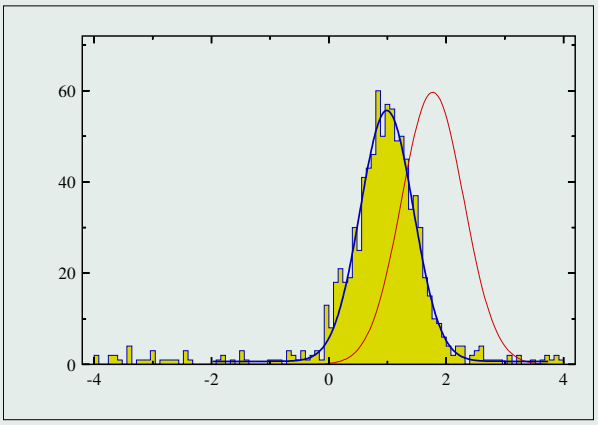


Axis of covariance ellipse is not tilted for $\sigma_1 = \sigma_2$ (right).

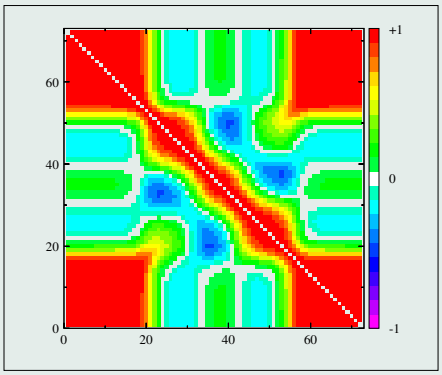
Example 7: fit of a histogram

Fitting of a curve to histogram data – not the natural application for constrained fitting ...

but it is possible:



	Fit result	
$\widehat{n}$	800.8	$\pm$ 29.6
$\widehat{\mu}$	0.993	$\pm$ 0.018
$\widehat{\sigma}$	0.480	$\pm$ 0.015
$\widehat{b}$	0.95	$\pm$ 0.17



Plot of correlation coefficients of fitted bin contents

Poisson distribution assumed for bin contents (a few bin contents are zero).

Contents