Least Squares – Facts and Fiction

Properties of Least Squares Estimates

V. Blobel

II. Inst. für Experimentalphysik, Hamburg

1999-11-15

I. Introduction: Correlated Errors

II. Properties of Least Squares Estimates

III. Example: An Alignment Problem

IV. Millepede – a Fit Program for 1000 + x Parameters

V. Conclusions

1. Introduction: Correlated Errors

"Errors" are now called "deviations".

The terms "statistical errors" and "systematic errors" are out.

Distinguished are

- deviations, evaluated by statistical methods, and
- deviations, evaluated by other methods.

A deviation, caused by systematic effects, can be evaluated using statistical methods.

V. Blobel
II. Inst. für Experimentalphysik, Hamburg

Correlated Errors

Problem of fitting correlated data discussed by G. D'Agostini in

- talk "Use and Misuse of the Covariance Matrix in Fitting" (this seminar, 99-11-1), and
- paper "On the Use of the Covariance Matrix to Fit Correlated Data" (DESY 93-175) and `http://www-zeus.roma1.infn.it/~agostini`

Systematic uncertainties can produce (in the context of the Least Squares Method) *results that "contradict expectations."*

Fit a theoretical curve through experimental data by minimizing the χ^2

$$\chi^2 = \boldsymbol{\Delta}^T \boldsymbol{V}^{-1} \boldsymbol{\Delta} \tag{1}$$

where $\boldsymbol{V}$ is the error matrix and $\boldsymbol{\Delta}$ is the vector of the differences between the theoretical and the experimental values (standard method of Least Squares).

Numerical Example with discrepancy

Numerical example, discussed in the paper:

" … the result of two measurements, $8.0 \pm 2\%$ and $8.5 \pm 2\%$, have a 10 % common normalization error. 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 function. Minimizing χ^2 as defined in (1), with $\boldsymbol{V}$ estimated empirically by the data, one obtains a value of 7.87 ± 0.81, which is at least surprising[a], since the most probable result is outside the interval determined by the two measured values" ("Peelles pertinent Puzzle").

Origin of apparent discrepancy?

(D'A Paper) *"shows that the effect is a direct consequence of the hypothesis used to estimate the empirical covariance matrix, namely the linearization on which the usual error propagation rely."*

(D'A Talk) *Deep reason: linearization implicit in the usual error propagation.*

[a] The weigthed mean ignoring common error is $8.235 \pm 1.4\%$.

Properties of the Least Squares Method

Another interpretation: the covariance matrix used in the computation was *incorrect*.

The Least Squares Method has, under certain conditions, several optimal properties:

If the measured data are *unbiased*, and if the covariance matrix is *finite* and *known*, then it can be shown, that the fitted values are *unbiased* estimates of the parameters for a *linear* model; *no* specific distribution of the measured data is required. The fitted parameter value(s) are the *most accurate* among all possible linear estimates (Gauß-Markoff theorem).

If and only if the data follow the normal distribution:

the value of the quadratic expression will follow a χ^2-distribution with a certain number of degrees of freedom[a].

The result can be biased (systematically shifted), if the conditions of the method are no fulfilled.

[a] It is somewhat misleading to call the expression (1) itself a χ^2.

V. Blobel
II. Inst. für Experimentalphysik, Hamburg

Is χ^2 a principle?

G.D'Agostinis statements:

⇒ Here the trap is the $\underline{\chi^2 \text{ method}}$!

⇒ Why do we use χ^2?

"a principle"

⇒ why should we believe principles? By definition, they are based on nothing, i.e. they work until they work.

⇒ In fact, there is no general solution to this problem, if we hold this principle.

⇒ Change principle?

⇒ Maximum likelihood

⇒ It can work better (it gives more often "resonable solution")

Correct treament

The model:

$$E[x_1] = \mu \qquad\qquad E[x_2] = \mu$$

(both single measurements are unbiased estimates of the same physical value)

$$\boldsymbol{V} = \begin{pmatrix} \sigma_1^2 + (\varepsilon\mu)^2 & (\varepsilon\mu)^2 \\ (\varepsilon\mu)^2 & \sigma_2^2 + (\varepsilon\mu)^2 \end{pmatrix}$$

(two contribution: a variance σ_1^2 and σ_2^2 for each single value, and a common normalization error, given as relative error).

Result from Least Squares:

$$8.235 \pm 1.4\% \pm 10\%$$

exactly identical to the expected result.

Approximate treatment

Replace $(\varepsilon\mu)$ by (εx_1) and (εx_2), respectively:

$$\boldsymbol{V} = \begin{pmatrix} \sigma_1^2 + (\varepsilon x_1)^2 & \varepsilon^2 x_1 x_2 \\ \varepsilon^2 x_1 x_2 & \sigma_2^2 + (\varepsilon x_2)^2 \end{pmatrix}$$

Result with "empirical" covariance matrix:

7.87 ± 0.81 discrepancy

Ansatz must be wrong, because normalization factor changed.

Geometrical explanation: angle φ of principal axis of error ellipse given by $\tan 2\varphi = 2V_{12}/(V_{22} - V_{11})$:

method	$\tan 2\varphi$	φ
correct treatment	250	44.8°
approximation	7.9	41.4°

V. Blobel
II. Inst. für Experimentalphysik,
Hamburg

The Poisson model

Is the model correct?

Data x_i are cross sections from event numbers n_i:

$$x_i = f \cdot n_i \qquad\qquad i = 1, 2$$

where the factor f may have relative uncertainty ε.

Optimal method: Maximum Likelihood based on Poisson statistic[a] with common mean value μ, with result

$$\hat{\mu} = \frac{1}{2}\,(n_1 + n_2)$$

(The sum of two Poisson distributed numbers follows the Poisson distribution).

Average cross section: $\hat{x} = 1/2\, f \cdot (n_1 + n_2)$.

[a] Example: the case $n_1 = 1$, $n_2 = 4$. The average based on the correct Poisson distribution is simply $1/2(n_1 + n_2) = 2.5$, while the weighted average from the two separate values $n_1 = 1 \pm 1$, $n_2 = 4 \pm 2$ would be 1.6.

V. Blobel

II. Inst. für Experimentalphysik, Hamburg

Lesson

Non-diagonal covariance matrices can produce strange results

($\Rightarrow$ Unfolding)

Extreme care necessary, if non-diagonal covariance matrix are defined "by hand"!

Maximum Likelihood requires full knowledge of distributions. If the distributions are Gaussian, then the Least Squares principle follows.

But Least Squares is a Principle of its own: it does not require knowledge of the distributions (just unbiased measurement of known variance is required).

V. Blobel

II. Inst. für Experimentalphysik, Hamburg

2. Properties of Least Squares Estimates

The model of Linear Least Squares: $\boldsymbol{y} = \boldsymbol{A}\,\boldsymbol{a}$

$$\begin{aligned} \boldsymbol{y} &= \text{vector of measured data} \\ \boldsymbol{A} &= \text{matrix (fixed)} \\ \boldsymbol{a} &= \text{vector of parameters} \\ \boldsymbol{r} &= \boldsymbol{y} - \boldsymbol{A}\boldsymbol{a} = \text{vector of residuals} \\ \boldsymbol{V}[\boldsymbol{y}] &= \text{covariance matrix of the data} \\ \boldsymbol{W} &= \boldsymbol{V}[\boldsymbol{y}]^{-1} \;\; \text{weight matrix} \end{aligned}$$

Least Squares Principle: minimize the expression

$$S(\boldsymbol{a}) = \boldsymbol{r}^T \boldsymbol{W} \boldsymbol{r} = (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{a})^T \; W \; (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{a})$$

with respect to $\boldsymbol{a}$.

V. Blobel

II. Inst. für Experimentalphysik, Hamburg

Least Squares solution

Derivatives of expression $S(\boldsymbol{a})$:

$$\begin{aligned} \frac{\partial S}{\partial \boldsymbol{a}} &= 2\left(-\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{y} + \left(\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}\right)\boldsymbol{a}\right) \\ \frac{\partial^2 S}{\partial \boldsymbol{a}^2} &= 2\left(\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}\right) \qquad = \text{constant} \end{aligned}$$

Solution (from $\partial S/\partial \boldsymbol{a} = 0$) is linear transf. of the data vector $\boldsymbol{y}$:

$$\hat{\boldsymbol{a}} = \left(\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}\right)^{-1} \boldsymbol{A}^T\boldsymbol{W}\boldsymbol{y} \qquad = \boldsymbol{B}\boldsymbol{y}$$

Covariance matrix of $\boldsymbol{a}$ by "error" propagation ($\boldsymbol{V}[\boldsymbol{y}] = \boldsymbol{W}^{-1}$):

$$\boldsymbol{V}[\hat{\boldsymbol{a}}] = \boldsymbol{B}\,\boldsymbol{V}[\boldsymbol{y}]\,\boldsymbol{B}^T = \left(\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}\right)^{-1}$$

(identical to inverse of second derivate of S).

V. Blobel
II. Inst. für Experimentalphysik, Hamburg

Properties of solution

General scheme

Starting from **Principles** properties of the solution are derived, which are valid under certain conditions.

Conditions:

- Data are unbiased: $E[\boldsymbol{y}] = \boldsymbol{A}\,\bar{\boldsymbol{a}}$ $\quad$ ($\bar{\boldsymbol{a}}$ = true parameter vector
- Covariance matrix $\boldsymbol{V}[\boldsymbol{y}]$ is known (correct) and finite

Properties:

- Estimated parameters are unbiased: $E[\hat{\boldsymbol{a}}] = \bar{\boldsymbol{a}}$

$$E[\hat{\boldsymbol{a}}] = \left(\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}\right)^{-1}\;\boldsymbol{A}^T\boldsymbol{W}\,E[\boldsymbol{y}] = \bar{\boldsymbol{a}}$$

- In the class of unbiased estimates $\boldsymbol{a}^*$, which are linear in the data, the Least Squares estimates $\hat{\boldsymbol{a}}$ have the smallest variance (Gauß-Markoff theorem)

Properties are not valid, if conditions violated.

Straight line fit

Example: track fit of z vs. $r\varphi$-length S

$$z_i = z_0 + (\cot\vartheta) \cdot S_i$$

Matrix $\boldsymbol{A}$ and parameter vector $\boldsymbol{a}$

$$\boldsymbol{A} = \begin{pmatrix} 1 & S_1 \\ 1 & S_2 \\ \vdots & \vdots \\ 1 & S_n \end{pmatrix} \qquad \boldsymbol{a} = \begin{pmatrix} z_0 \\ \cot\vartheta \end{pmatrix}$$

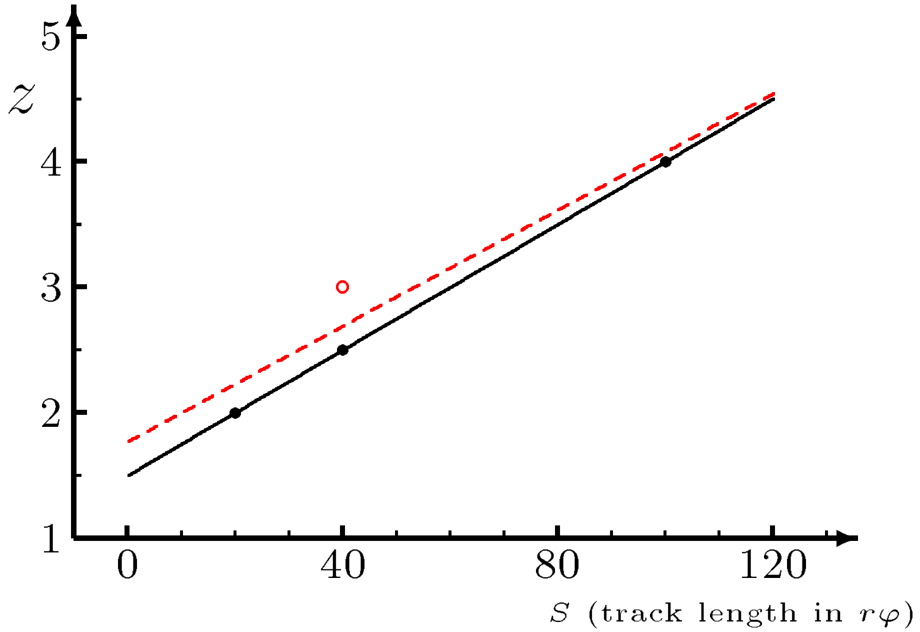
Weight matrix is diagonal (*independent* measurements):

$$\boldsymbol{A}^T \boldsymbol{W} \boldsymbol{A} = \begin{pmatrix} \sum w_i & \sum w_i S_i \\ \sum w_i S_i & \sum w_i S_i^2 \end{pmatrix}$$

If one measured z_i-value is shifted (biased), then

- parameters biased, and
- χ^2-value very high.

Straight line fit



The full line is a straight line fit to three well aligned data points (black dots).

The dashed curve is the straight line fit, if the middle point is "badly aligned" (circle).

V. Blobel

II. Inst. für Experimentalphysik, Hamburg

3. An Alignment Problem

z-coordinate measured in detector with many (e.g. ten) parallel planes, with model

$$z_i = z_0 + (\cot\vartheta) \cdot S_i$$

Some or all planes may be misaligned (Δz), visible in bad χ^2 of fits.

Standard method in HEP based on residuals:

- make straight line fits,
- plot histograms of residuals for all planes,
- determine in all histograms mean residuum $\Rightarrow \Delta z$ correction

With corrected z-values: the χ^2 of fits is much smaller!

$\Rightarrow$ alignment called "successful"!

HEP *Alignment – properties*

Is this alignment really successful?

Although the χ^2 of fits is much smaller, the parameter values are almost unchanged and still biased!

... the original fit result was biased, because a wrong model (without shifts) has been used.

Further properties of method based on residuals:

Corrections Δz_j

- depend on chamber (in)efficiencies,
- are unstable, if "alignment" is done repeatedly.

V. Blobel

II. Inst. für Experimentalphysik, Hamburg

Improved Alignment

Improvement I: try to use correct model

$$z_i = z_0 + (\cot \vartheta) \cdot S_i + \Delta z_j$$

where Δz_j = z-correction of plane j.

Problem: 10 measured values z_i, but $2 + 10$ parameters to be fitted (no solution).

Improvement II: use correct model simultaneously for N tracks.

$10 \times N$ measured values, and $2 \times N + 10$ parameters $\Rightarrow$ fit possible, for example $N = 1000$: 10 000 measured values 2010 parameters

This model is still unstable – matrix $\left(\boldsymbol{A}^T \boldsymbol{W} \boldsymbol{A}\right)$ is singular with rank defect 2: overall shift plus rotation of detector not fixed!

Improvement III: add two constraints for shifts Δz_j (simple in simultaneous fit, difficult in residual method).

4. Millepede – a Fit Program for 1000 + x Parameters

Millepede – a general program for Linear Least Squares (LLS) with a large number of parameters, especially for alignment and calibration problems.

LLS with solution $\boldsymbol{a} = \boldsymbol{C}^{-1}\,\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{y}$ requires inversion of symmetric matrix

$$\boldsymbol{C} = \boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}$$

Execution time $\propto n^3$ for $n \times n$ matrix, memory $\propto n^2$.

$\Rightarrow$ Practical limit around $n = 1000$: time $\approx$ few minutes, 4 Mbytes for matrix $\boldsymbol{C}$ (double precision).

Matrix often singular or close to singular $\Rightarrow$ special inversion routine for symmetric matrices, which determines invertible part of the matrix (remaining elements and corresponding parameters $\Rightarrow$ zero).

Code and manual on homepage: `http://www.desy.de/~blobel/`

V. Blobel

II. Inst. für Experimentalphysik, Hamburg

Alignment/Calibration Problems

Example: tracks in z-chambers with $z_i = z_0 + (\cot\vartheta)\cdot S_i + \Delta z_j$

Often large number (≈ 1000) of global parameters

plus

large number N of tracks ($\approx 10\,000$) with e.g. $2 \times N$ local parameters.

$\Rightarrow$ standard solution not possible, but:

Interest only in global parameters: huge matrix C can be separated into submatrices related to

- global parameters and
- local parameters,

and due to special structure, can be reduced to system of equations for global parameters only!

V. Blobel
II. Inst. für Experimentalphysik, Hamburg

System of normal equations

The simultaneous fit of global and local parameters:

$$
\left(\begin{array}{c|c|c|c}
\sum \boldsymbol{C}_i & \cdots & \boldsymbol{G}_i & \cdots \\
\hline
\vdots & \ddots & 0 & 0 \\
\hline
\boldsymbol{G}_i^T & 0 & \boldsymbol{\Gamma}_i & 0 \\
\hline
\vdots & 0 & 0 & \ddots
\end{array}\right)
\; . \;
\left(\begin{array}{c}
\boldsymbol{a} \\
\hline
\vdots \\
\hline
\boldsymbol{\alpha}_i \\
\hline
\vdots
\end{array}\right)
= \; . \;
\left(\begin{array}{c}
\sum \boldsymbol{b}_i \\
\hline
\vdots \\
\hline
\boldsymbol{\beta}_i \\
\hline
\vdots
\end{array}\right)
$$

Reduction of matrix size

The matrices $\boldsymbol{C}_i$, $\boldsymbol{\Gamma}_i$, $\boldsymbol{G}_i$ and the vectors $\boldsymbol{b}_i$ and $\boldsymbol{\beta}_i$ contain contributions from the i-th track. The only connection between the local parameters of different tracks is given via the sub-matrices $\boldsymbol{G}_i$ und $\boldsymbol{C}_i$.

Ignoring the global parameters one could solve the normal equations $\boldsymbol{\Gamma}_i \boldsymbol{\alpha}_i = \boldsymbol{\beta}_i$ for each track separately by $\boldsymbol{\alpha}_i = \boldsymbol{\Gamma}_i^{-1} \boldsymbol{\beta}_i$.

Huge matrix is reduced to n normal equations for global parameters

$$\left(\quad \boldsymbol{C}' \quad \right) \left(\boldsymbol{a} \right) = \left(\boldsymbol{b}' \right) ,$$

with a modified matrix $\boldsymbol{C}'$ and a modified vector $\boldsymbol{b}'$,

$$\boldsymbol{C}' = \sum_i \boldsymbol{C}_i - \sum_i \boldsymbol{G}_i \boldsymbol{\Gamma}_i^{-1} \boldsymbol{G}_i^T \qquad \boldsymbol{b}' = \sum_i \boldsymbol{b}_i - \sum_i \boldsymbol{G}_i \left(\boldsymbol{\Gamma}_i^{-1} \boldsymbol{\beta}_i \right) .$$

V. Blobel
II. Inst. für Experimentalphysik, Hamburg

System of normal equations

The solution

$$\boldsymbol{a} = \boldsymbol{C'}^{-1}\,\boldsymbol{b'}$$

with covariance matrix $\boldsymbol{C'}^{-1}$ is direct, no iterations are required.

Iterations required however by presence of outliers (e.g. bad tracks), with loose cuts initially and tight final cuts.

Applications:

- CST Silicon tracker aligment
- CJC Central Jet Chamber alignment (global parameters: 360 half-cell dependent drift velocity corrections, . . .)
- CIZ/COZ z chamber alignment

V. Blobel
II. Inst. für Experimentalphysik, Hamburg

Simulation

Track measurement by z-chamber:

Straight tracks through ten planes perpendicular to detector axis. 1000 tracks generated with ≈ 90 % chamber efficiency, with hits shifted by Δ_j , smeared by standard deviation σ.

Reconstruction of shifts by MILLEPEDE:

- First plane fixed, overall rotation constrained to be zero.
- 9 free parameters plus 1 constraint ($+2 \times 1000$ local parameters).

Result from MILLEPEDE:

- high accuracy ($\sigma = 20\mu$m) planes aligned with 1μm precision,
- low accuracy ($\sigma = 300\mu$m) planes aligned with 10μm precision.

Simulation parameters and results

i	x [cm]	σ [cm]	Δ [cm]	Δ_f MILLEPEDE	Δ_r fr. resid.
1	5	0.0020	0.0000	0.0000 ± 0.0000	-0.0274
2	9	0.0020	-0.0400	-0.0399 ± 0.0001	-0.0636
3	20	0.0300	0.1500	0.1495 ± 0.0009	0.1269
4	25	0.0300	0.0300	0.0293 ± 0.0009	0.0099
5	30	0.0300	-0.0750	-0.0755 ± 0.0012	-0.0921
6	35	0.0300	0.0450	0.0483 ± 0.0010	0.0346
7	40	0.0300	0.0350	0.0342 ± 0.0010	0.0216
8	45	0.0300	-0.0800	-0.0814 ± 0.0010	-0.0954
9	50	0.0300	0.0900	0.0902 ± 0.0011	0.0840
10	55	0.0300	-0.0500	-0.0494 ± 0.0011	-0.0513

Δ_f from MILLEPEDE fit

Δ_r from residual distributions – remaining bias by factor ≈ 100 larger, and unstable

Initial and final χ^2 distribution

Histogram of χ^2/n_{df} for local fits done automatically in MILLEPEDE

Initial χ^2 distribution (first iteration: corrections zero)

```
Total 1000 local fits, 631 rejected.
Histogram of Chi**2/ndf:  Total 1000 entries with 871 overflows
                                             X
                                           X X   X
                                           X XX  X
                                   XX   X  X XX  X X
                               X  XXXX  XX X XX XXXX
                           X   X  XXXXX XX X XX XXXX
                           XXX X XXXXXX XXXX XXXXXXX
             X   X  X    XXXXX XXXXXXXXXXXXX XXXXXXX
        X   XXX  X XX    XXXXXXXXXXXXXXXXXXXXXXXXXXX
  ----+----1----+----2----+----3----+----4----+----5
 0.0       2.0       4.0       6.0       8.0      10.0

        1   121  2 12    224331523566542653819735856
```

Final χ^2 distribution (last iteration: final corrections)

```
Total 1000 local fits, 2 rejected.
Histogram of Chi**2/ndf:  Total 1000 entries with 0 overflows
     XX
    XXXX
    XXXX
   XXXXX
   XXXXXX
   XXXXXX
   XXXXXXXX
   XXXXXXXXX
  XXXXXXXXXXX
  ----+----1----+----2----+----3----+----4----+----5
 0.0       2.0       4.0       6.0       8.0      10.0

   11111
  203553856321
  453929162903274211
```


V. Blobel
II. Inst. für Experimentalphysik, Hamburg

5. Conclusions

- The Principle of Least Squares is alive and well!
- Proofs exist for certain optimal properties of LLS results, but input has to satisfy certain conditions!
- "Handmade" covariance matrices may be dangerous!
- Efficient methods exist for huge number of parameters in typical aligment/calibration problems! High precision is reached, because simultaneous fit makes full use of all the correlations between data, local and global parameters.!

Least Squares Methods, invented 200 years ago (GAUSS, LAPLACE, DE MOIVRE ...) can be recommended also for the next Millenium.

V. Blobel
II. INST. FÜR EXPERIMENTALPHYSIK, HAMBURG