G. Least squares methods

V. Blobel
Universität Hamburg

April 2001

1. The least squares principle
2. Linear least squares
3. Properties of least squares estimates
4. Weighted least squares
5. Orthogonal polynomials
6. Large number of parameters
7. Nonlinear least squares
8. Least squares with constraints

2. Linear least squares

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

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

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 $\boldsymbol{a}$.

Least Squares solution

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

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

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).

Properties of solution

General scheme

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

Conditions:

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

Properties:

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

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

- 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 y (measured) vs. abscissa x

$$y_i = a_0 + a_1 \cdot x_i$$

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

$$\boldsymbol{A} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \boldsymbol{a} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}$$

Weight matrix is diagonal (*independent* measurements):

$$\boldsymbol{A}^T \boldsymbol{W} \boldsymbol{A} = \begin{pmatrix} \sum w_i & \sum w_i x_i \\ \sum w_i x_i & \sum w_i x_i^2 \end{pmatrix} \qquad \boldsymbol{A}^T \boldsymbol{W} \boldsymbol{y} = \begin{pmatrix} \sum w_i y_i \\ \sum w_i x_i y_i \end{pmatrix}$$

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

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

Normal equations for uncorrelated data

Usually the direct measurements, which are input to a least squares problem, are *uncorrelated*, i.e. the covariance matrix $\boldsymbol{V}(\boldsymbol{y})$ and the weight matrix $\boldsymbol{W}$ are *diagonal*. This property, which is assumed here, simplifies the computation of the matrix products

$$\boldsymbol{C} = \boldsymbol{A}^T \boldsymbol{W} \boldsymbol{A} \qquad \text{and} \qquad \boldsymbol{b} = \boldsymbol{A}^T \boldsymbol{W} \boldsymbol{y}$$

which are necessary for the solution

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

Normal equations for uncorrelated data contnd.

The diagonal elements of the weight matrix $\boldsymbol{W}$ are denoted by w_i, with $w_i = 1/\sigma_i^2$. Each data value y_i with its weight w_i makes an independent contribution to the final matrix products. Calling the i-th row $\boldsymbol{A}_i$, with

$$i\text{-th row of } \boldsymbol{A} \qquad \boldsymbol{A}_i = (d_1, d_2, \ldots, d_p)$$

the contributions of this row to $\boldsymbol{C}$ and $\boldsymbol{b}$ can be written as the $p \times p$-matrix $w_i \boldsymbol{A}_i^T \cdot \boldsymbol{A}_i$ and the p-vector $w_i \boldsymbol{A}_i^T \cdot y_i$. The contributions of a single row are:

$$\begin{array}{c|cccc} & d_1 & d_2 & \ldots & d_p \\ \hline d_1 & w_i d_1^2 & w_i d_1 d_2 & \ldots & w_i d_1 d_p \\ d_2 & & w_i d_2^2 & \ldots & w_i d_2 d_p \\ \ldots & & & \ldots & \ldots \\ d_p & & & & w_i d_p^2 \end{array} \qquad \begin{array}{c|c} & y_i \\ \hline d_1 & w_i d_1 z_i \\ d_2 & w_i d_2 z_i \\ \ldots & \ldots \\ d_p & w_i d_p z_i \end{array} ,$$

where the symmetric elements in the lower half are not shown.

Robust least square fit

Least square fit on observations y_i, yielding fitted values $\hat{y}_i$, and residuals $r_i = y_i - \hat{y}_i$. Estimate for the standard error of y_i (or of the r_i) is s_i.

- ”Clean” the data by pulling outliers towards their fitted values: winsorize the observations y_i and replace them by pseudo-observations y_i^*:

$$\begin{aligned} y_i^* &= y_i \,, & &\text{if} \quad |r_i| \leq c\, s_i \,, \\ &= \hat{y}_i - c\, s_i \,, & &\text{if} \quad r_i < -c\, s_i \,, \\ &= \hat{y}_i + c\, s_i \,, & &\text{if} \quad r_i > +c\, s_i \,. \end{aligned}$$

  The factor c regulates the amount of robustness, a goid choice is $c = 1.5$.

- Refit iteratively: the pseudo-observations y_i^* are used to calculate new parameters and new fitted values $\hat{y}_i$.

5. Orthogonal polynomials

If data interpolation required, but no parametrization known: Fit of normal general p-th polynomial to the data

$$f(x_i) \cong y_i \qquad \text{with} \quad f(x) = \sum_{j=0}^{p} a_j \, x^j$$

Generalization of straight-line fit to general p-th order polynomial straightforward: matrix $\boldsymbol{A}^T \boldsymbol{W} \boldsymbol{A}$ contains sum of all powers of x_i up to x_i^{2p}.

Disadvantage:

- fit numerically unstable
- determination of optimal order p difficult
- value of coefficients a_j depend on highest exponent p

Better: fit of orthogonal polynomial.

Straight line fit

Parametrization

$$f(x) = a_0 + a_1 x \quad \text{replaced by} \quad f(x) = a_0 \cdot p_0(x) + a_1 \cdot p_1(x)$$

The orthogonal polynomials $p_0(x)$ and $p_1(x)$ are defined by

$$\begin{aligned} p_0(x) &= b_{00} \\ p_1(x) &= b_{10} + b_{11} x = b_{11} \left(x - \langle x \rangle \right) \quad . \end{aligned}$$

with coefficients b_{00} and b_{11}

$$b_{00} = \left(\sum_{i=1}^{n} w_i \right)^{-1/2} \qquad b_{11} = \left(\sum_{i=1}^{n} w_i \left(x_i - \langle x \rangle \right)^2 \right)^{-1/2}$$

with the coefficient $b_{10} = -\langle x \rangle \, b_{11}$.

Advantage of new ansatz

$$\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A} = \begin{pmatrix} \sum w_i p_0^2(x_i) & \sum w_i p_0(x_i)p_1(x_i) \\ \sum w_i p_0(x_i)p_1(x_i) & \sum w_i p_1^2(x_i) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{y} = \begin{pmatrix} \sum w_i p_0(x_i) y_i \\ \sum w_i p_1(x_i) y_i \end{pmatrix} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix}$$

i.e. covariance matrix is unit matrix and parameters $\boldsymbol{a}$ are calculated by sums (no matrix inversion).

General orthogonal polynomial polynomial:

$$f(x) = \sum_{j=1}^{p} a_j p_j(x)$$

Title

Construction of higher order polynomial $p_j(x)$ by recurrence relation:

$$\boxed{\gamma p_j(x) = (x_i - \alpha)\, p_{j-1}(x) - \beta p_{j-2}(x)}$$

from the two previous functions $p_{j-1}(x)$ and $p_{j-2}(x)$ with parameters α and β defined by

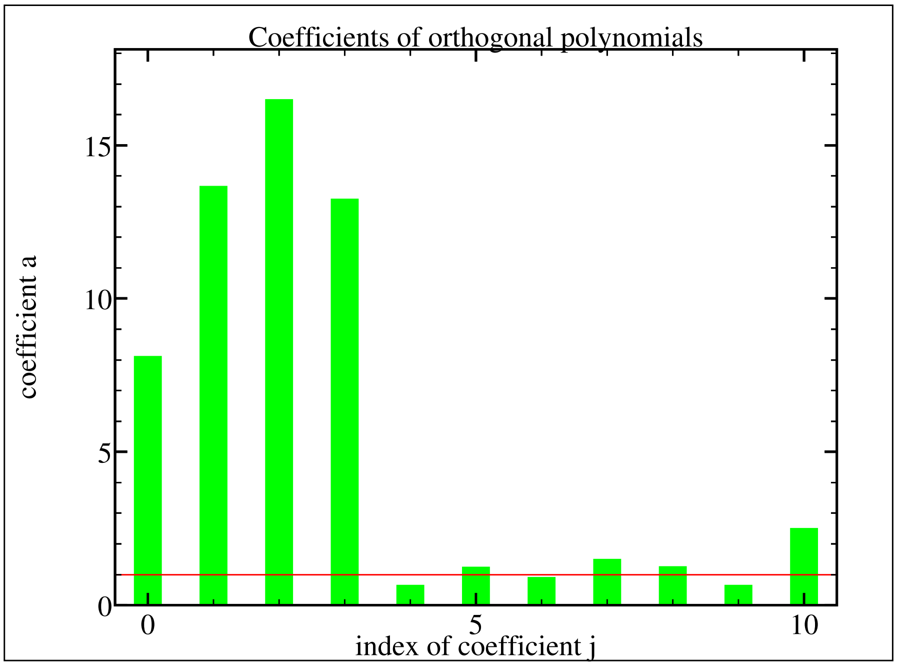
$$\alpha = \sum_{i=1}^{n} w_i x_i p_{j-1}^2(x_i) \qquad \beta = \sum_{i=1}^{n} w_i x_i p_{j-1}(x_i) p_{j-2}(x_i)$$

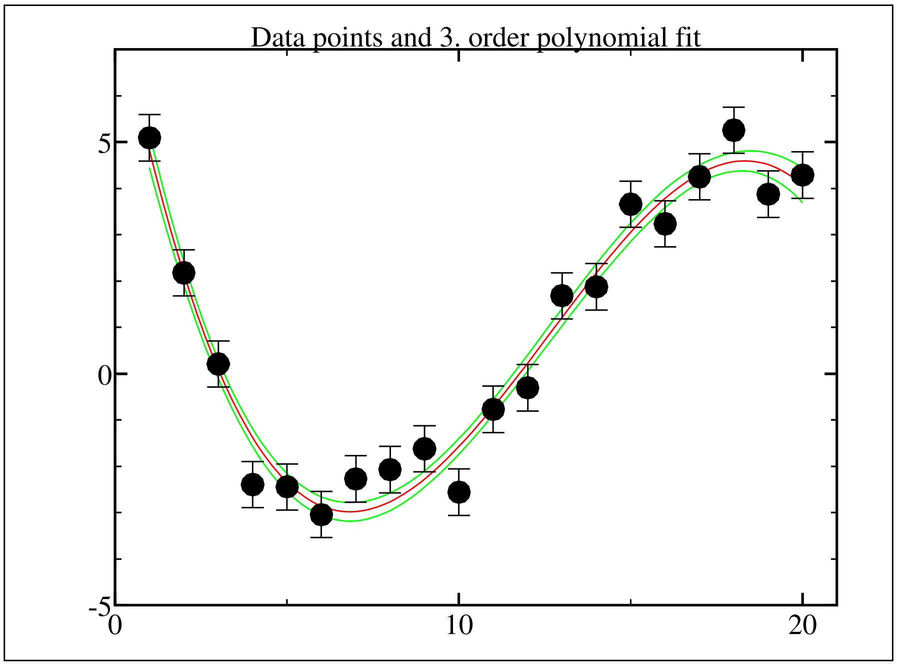
Normalization factor γ given by

$$\gamma^2 = \sum_{i=1}^{n} w_i \left[(x_i - \alpha)\, f_{j-1}\,(x_i) - \beta f_{j-2}\,(x_i)\right]^2$$

Parameters $\widehat{a}_j$ are determined from data by

$$\widehat{a}_j = \sum_{i=1}^{n} w_i p_j(x_i)\; y_i$$


Coefficients of orthogonal polynomials
coefficient a
index of coefficient j
0
5
10
15


Data points and 3. order polynomial fit
-5
0
5
10
20