Data Fitting and Linear Least-Squares Problems

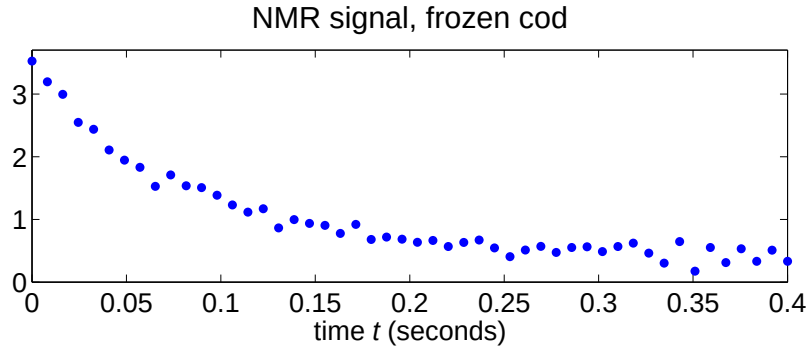
This lecture is based on the book

> P. C. Hansen, V. Pereyra and G. Scherer,
> *Least Squares Data Fitting with Applications*,
> Johns Hopkins University Press, to appear
> (the necessary chapters are available on CampusNet)

and we cover this material:

- Section 1.1: Motivation.
- Section 1.2: The data fitting problem.
- Section 1.4: The residuals and their properties.
- Section 2.1: The linear least squares problem.
- Section 2.2: The QR factorization.

Example: Parameter estimation



The measured NMR signal reflects the amount of different types of water environments in the meat.

The ideal time signal $\phi(t)$ from NMR is:

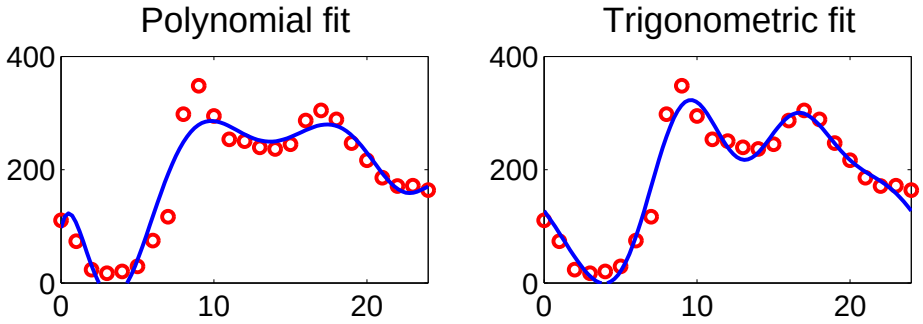
$$\phi(t) = x_1\, e^{-\lambda_1 t} + x_2\, e^{-\lambda_2 t} + x_3, \qquad \lambda_1, \lambda_2 > 0 \text{ areknown.}$$

Amplitudes x_1 and x_2 are proportional to the amount of water containing the two kinds of protons. The constant x_3 accounts for an undesired background (bias) in the measurements.

Goal: estimate the three unknown parameters and then compute the different kinds of water contents in the meat sample.

Example: Data approximation

Measurements of air pollution, in the form of the NO concentration, over a period of 24 hours, on H. C. Andersens Boulevard.



Fit a smooth curve to the measurements, so that we can compute the concentration at an arbitrary time between 0 and 24 hours.

Polynomial and periodic models:

$$\phi(t) = x_1\, t^p + x_2\, t^{p-1} + \cdots + x_p\, t + x_{p+1},$$

$$\phi(t) = x_1 + x_2\, \sin(\omega\, t) + x_3\, \cos(\omega\, t) + x_4\, \sin(2\omega\, t) + x_5\, \cos(2\omega\, t) + \cdots$$

The Underlying Problem: Data Fitting

Given: data (t_i, y_i) with measurement errors.

We want: to fit a model – a function $\phi(t)$ – to these data.

Requirement: $\phi(t)$ captures the "overall behavior" of the data without being too sensitive to the errors.

Data fitting is distinctly different from interpolation, where we seek a model that interpolates the given data, i.e., it satisfies $\phi(t_i) = y_i$ for all the data points.

In this data fitting approach there are *more data than unknown parameters*, which helps to decrease the uncertainty in the parameters of the model.

Formulation of the Data Fitting Problem

We assume that we are given m data points

$$(t_1, y_1),\ (t_2, y_2),\ \ldots,\ (t_m, y_m),$$

which can be described by the relation

$$y_i = \Gamma(t_i) + e_i, \qquad i = 1, 2, \ldots, m.$$

The function $\Gamma(t)$ describes the noise-free data; $e_1, e_2, \ldots, e_m$ are the data errors (we may have statistical information about them).

The data errors/noise represent measurement errors as well as random variations in the physical process that generates the data.

Without loss of generality we can assume that the abscissas t_i appear in non-decreasing order, i.e.,

$$t_1 \leq t_2 \leq \cdots \leq t_m.$$

The Linear Data Fitting Problem

We wish to compute an approximation to the data, given by the *fitting model* $M(\boldsymbol{x}, t)$.

The vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$ contains n parameters that characterize the model, to be determined from the given noisy data.

The linear data fitting problem:

$$\text{Linear fitting model:} \qquad M(\boldsymbol{x}, t) = \sum_{j=1}^{n} x_j \, f_j(t).$$

Here the functions $f_j(x)$ are chosen either because the reflect the underlying physical/chemical/... model (parameter estimation) or because they are "easy" to work with (data approximation).

The *order* of the fit n should be somewhat smaller than the number m of data points.

The Least Squares (LSQ) Fit

A standard technique for determining the parameters.

We introduce the residual r_i associated with the data points as

$$r_i = y_i - M(\boldsymbol{x}, t_i), \qquad i = 1, 2, \ldots, m \ .$$

Note that each residual is a function of the parameter vector $\boldsymbol{x}$, i.e., $r_i = r_i(\boldsymbol{x})$. A *least squares fit* is a choice of the parameter vector $\boldsymbol{x}$ that minimizes the sum-of-squares of the residuals:

$$\text{LSQ fit:} \quad \min_{\boldsymbol{x}} \sum_{i=1}^{m} r_i(\boldsymbol{x})^2 = \min_{\boldsymbol{x}} \sum_{i=1}^{m} \big(y_i - M(\boldsymbol{x}, t_i)\big)^2 .$$

Sum-of-absolute-values makes the problem harder to solve:

$$\min_{\boldsymbol{x}} \sum_{i=1}^{m} |r_i(\boldsymbol{x})| = \min_{\boldsymbol{x}} \sum_{i=1}^{m} \big|y_i - M(\boldsymbol{x}, t_i)\big|$$

Summary: the Linear LSQ Data Fitting Problem

Given: data (t_i, y_i), $i = 1, \ldots, m$ with measurement errors.

Our data model: there is an (unknown) function $\Gamma(t)$ such that

$$y_i = \Gamma(t_i) + e_i, \qquad i = 1, 2, \ldots, m \ .$$

The data errors: e_i are are unknown, but we may have some statistical information about them.

Our linear fitting model:

$$M(\boldsymbol{x}, t) = \sum_{j=1}^{n} x_j \, f_j(t), \qquad f_j(t) = \text{given functions.}$$

The linear least squares fit:

$$\min_{\boldsymbol{x}} \sum_{i=1}^{m} r_i(\boldsymbol{x})^2 = \min_{\boldsymbol{x}} \sum_{i=1}^{m} \big(y_i - M(\boldsymbol{x}, t_i)\big)^2 \ ,$$

The Underlying Idea

Recall our underlying data model:

$$y_i = \Gamma(t_i) + e_i, \qquad i = 1, 2, \ldots, m$$

and consider the residuals for $i = 1, \ldots, m$:

$$\begin{aligned} r_i &= y_i - M(\boldsymbol{x}, t_i) \\ &= \big(y_i - \Gamma(t_i)\big) + \big(\Gamma(t_i) - M(\boldsymbol{x}, t_i)\big) \\ &= e_i + \big(\Gamma(t_i) - M(\boldsymbol{x}, t_i)\big) \ . \end{aligned}$$

- The *data error* e_i is from the measurements.
- The *approximation error* $\Gamma(t_i) - M(\boldsymbol{x}, t_i)$ is due to the discrepancy between the pure-data function and the fitting model.

A good fitting model $M(\boldsymbol{x}, t)$ is one for which the approximation errors are of the same size as the data errors.

Matrix-Vector Notation

Define the matrix $A \in \mathbb{R}^{m \times n}$ and the vectors $\boldsymbol{y}, \boldsymbol{r} \in \mathbb{R}^m$ as follows,

$$A = \begin{pmatrix} f_1(t_1) & f_2(t_1) & \cdots & f_n(t_1) \\ f_1(t_2) & f_2(t_2) & \cdots & f_n(t_2) \\ \vdots & \vdots & & \vdots \\ f_1(t_m) & f_2(t_m) & \cdots & f_n(t_m) \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}, \quad \boldsymbol{r} = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{pmatrix},$$

i.e., $\boldsymbol{y}$ is the vector of observations, $\boldsymbol{r}$ is the vector of residuals, and the matrix A is constructed such that the jth column is the jth model basis function sampled at the abscissas $t_1, t_2, \ldots, t_m$. Then

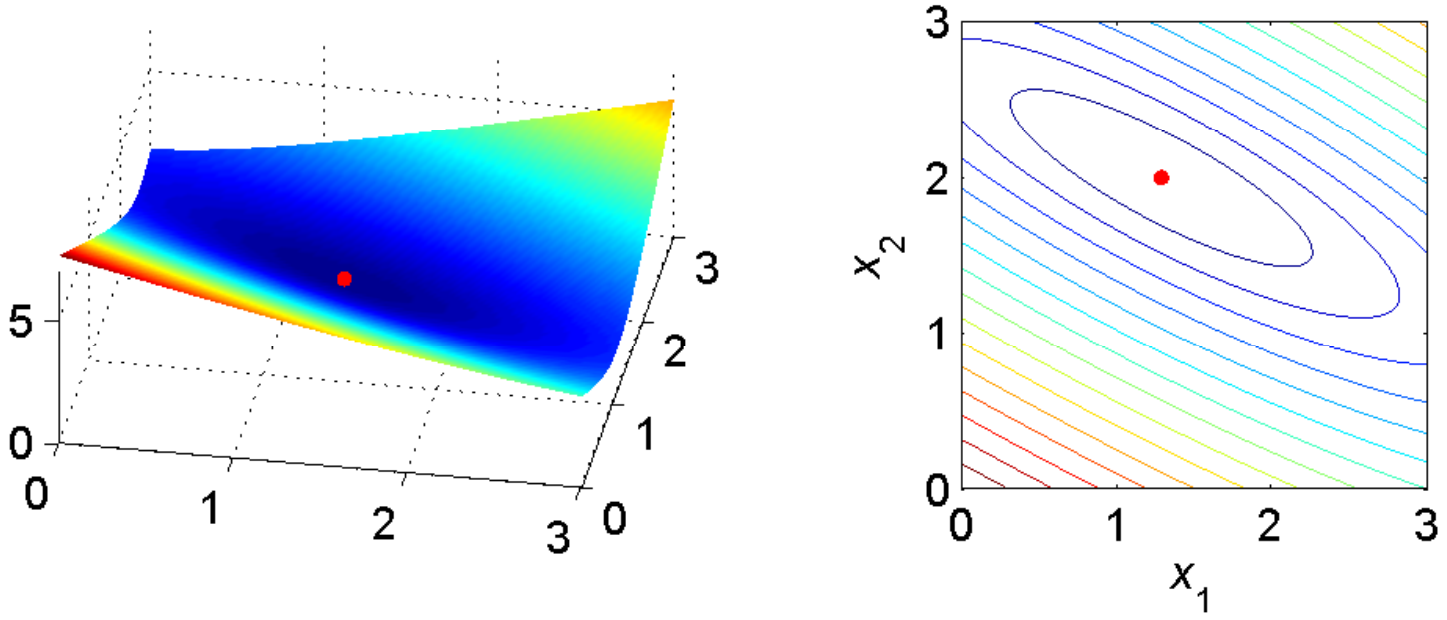
$$\boldsymbol{r} = \boldsymbol{y} - A\,\boldsymbol{x} \qquad \text{and} \qquad \rho(\boldsymbol{x}) = \sum_{i=1}^{m} r_i(\boldsymbol{x})^2 = \|\boldsymbol{r}\|_2^2 = \|\boldsymbol{y} - A\,\boldsymbol{x}\|_2^2.$$

The data fitting problem in linear algebra notation: $\boxed{\min_{\boldsymbol{x}} \rho(\boldsymbol{x}).}$

Contour Plots

NMR problem: fixing $x_3 = 0.3$ gives a problem $\min_{\boldsymbol{x}} \rho(\boldsymbol{x})$ with two unknowns x_1 and x_2.

Left – $\rho(\boldsymbol{x})$ versus $\boldsymbol{x}$. Right – level curves $\{\boldsymbol{x} \in \mathbb{R}^2 \mid \rho(\boldsymbol{x}) = c\}$.



Contours are ellipsoids, and cond(A) = eccentricity of the ellipsoids.

The Example Again

We return to the NMR data fitting problem. For this problem there are $m = 50$ data points and $n = 3$ model basis functions:

$$f_1(t) = e^{-\lambda_1 t}, \qquad f_2(t) = e^{-\lambda_2 t}, \qquad f_3(t) = 1.$$

The coefficient matrix:

```
A =
  1.0000e+000  1.0000e+000  1.0000e+000
  8.0219e-001  9.3678e-001  1.0000e+000
  6.4351e-001  8.7756e-001  1.0000e+000
  5.1622e-001  8.2208e-001  1.0000e+000
  4.1411e-001  7.7011e-001  1.0000e+000
  3.3219e-001  7.2142e-001  1.0000e+000
```

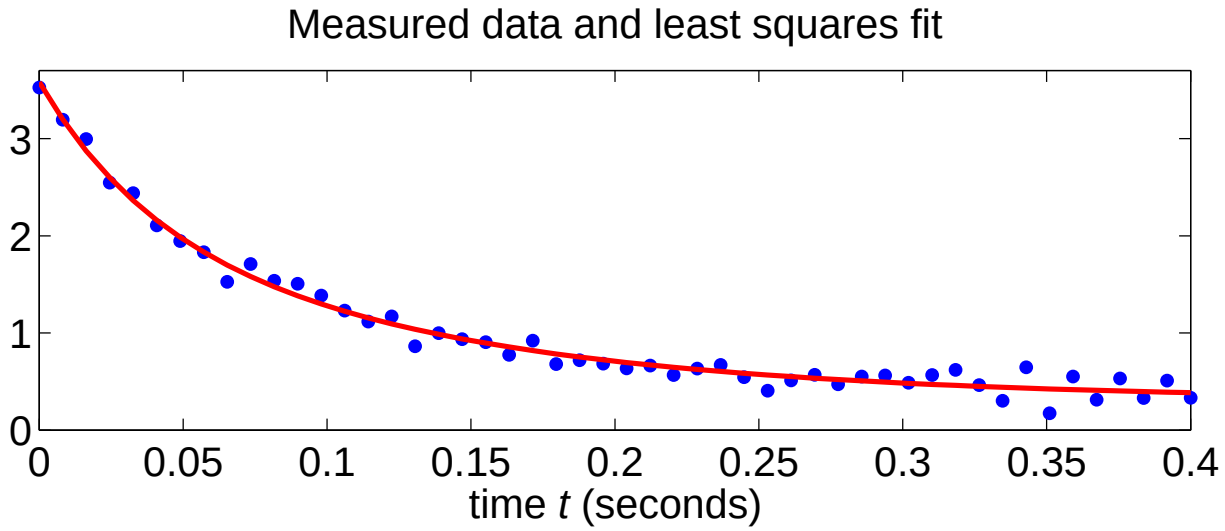
etc etc etc

The Example Again, Continued

The LSQ solution to the least squares problem is $\boldsymbol{x}^* = (x_1^*, x_2^*, x_3^*)^T$ with elements

$$x_1^* = 1.303, \qquad x_2^* = 1.973, \qquad x_3^* = 0.305.$$

The exact parameters used to generate the data are 1.27, 2.04, 0.3 .



Introduction of Weights

Some statistical assumptions about the noise:

$$\mathcal{E}(e_i) = 0, \qquad \mathcal{E}(e_i^2) = \varsigma_i^2, \qquad i = 1, 2, \ldots, m,$$

where ς_i is the standard deviation of e_i.

The *maximum likelihood* principle in statistics tells us that we should minimize the weighted residuals, with weights equal to the reciprocals of the standard deviations:

$$\min_{\boldsymbol{x}} \sum_{i=1}^{m} \left(\frac{r_i(\boldsymbol{x})}{\varsigma_i} \right)^2 = \min_{\boldsymbol{x}} \sum_{i=1}^{m} \left(\frac{y_i - M(\boldsymbol{x}, t_i)}{\varsigma_i} \right)^2 .$$

Identical standard deviations $\varsigma_i = \varsigma$ we have:

$$\min_{\boldsymbol{x}} \varsigma^{-2} \sum_{i=1}^{m} \left(y_i - M(\boldsymbol{x}, t_i) \right)^2$$

whose solution is independent of ς.

More About Weights

Consider the expected value of the weighted sum-of-squares:

$$\begin{aligned}
\mathcal{E}\left(\sum_{i=1}^{m}\left(\frac{r_i(\boldsymbol{x})}{\varsigma_i}\right)^2\right) &= \sum_{i=1}^{m}\mathcal{E}\left(\frac{r_i(\boldsymbol{x})^2}{\varsigma_i^2}\right) \\
&= \sum_{i=1}^{m}\mathcal{E}\left(\frac{e_i^2}{\varsigma_i^2}\right) + \sum_{i=1}^{m}\mathcal{E}\left(\frac{(\Gamma(t_i) - M(\boldsymbol{x},t_i))^2}{\varsigma_i^2}\right) \\
&= m + \sum_{i=1}^{m}\frac{\mathcal{E}\big((\Gamma(t_i) - M(\boldsymbol{x},t_i))^2\big)}{\varsigma_i^2},
\end{aligned}$$

where we used that $\mathcal{E}(e_i) = 0$ and $\mathcal{E}(e_i^2) = \varsigma_i^2$.

Intuitive result:
we can allow the expected value of the approximation errors to be larger for those data (t_i, y_i) that have larger standard deviations (i.e., larger errors).

Weights in the Matrix-Vector Formulation

We introduce the diagonal matrix

$$W = \mathrm{diag}(w_1, \ldots, w_m), \qquad w_i = \varsigma_i^{-1}, \quad i = 1, 2, \ldots, m.$$

Then the weighted LSQ problem is $\boxed{\min_{\boldsymbol{x}} \rho_W(\boldsymbol{x})}$ with

$$\rho_W(\boldsymbol{x}) = \sum_{i=1}^{m} \left(\frac{r_i(\boldsymbol{x})}{\varsigma_i} \right)^2 = \|W\,(\boldsymbol{y} - A\,\boldsymbol{x})\|_2^2.$$

Same computational problem as before, with $\boldsymbol{y} \to W\,\boldsymbol{y}$, $A \to W\,A$.

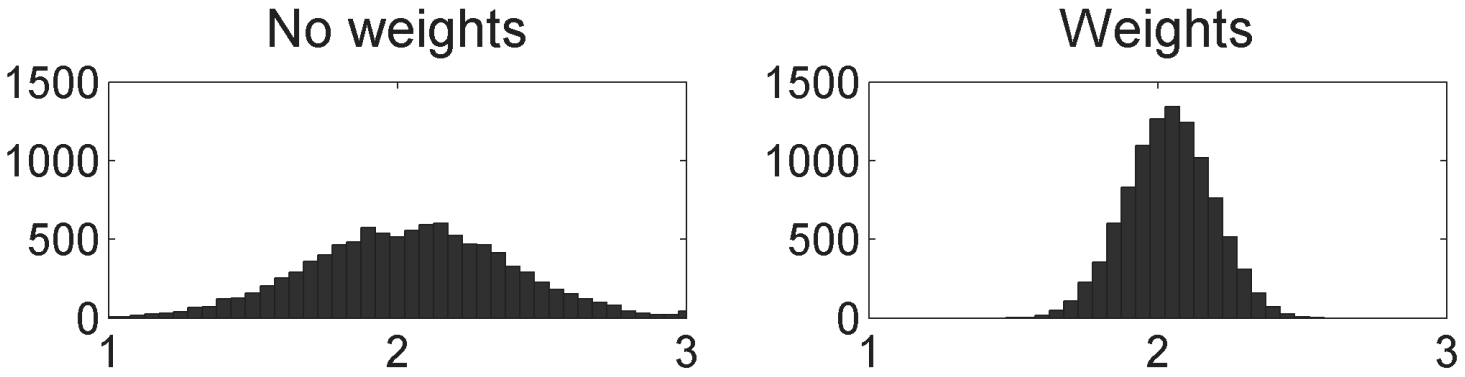
Example with Weights

The NMR data again, but we add larger Gaussian noise to the first 10 data points, with standard deviation 0.5:

- $\varsigma_i = 0.5$, $i = 1, 2, \ldots, 10$ (first 10 data with larger errors)
- $\varsigma_i = 0.1$, $i = 11, 12, \ldots, 50$ (remaining data with smaller errors).

The weights: $w_i = \varsigma_i^{-1}$ are $2, 2, \ldots, 2, 10, 10, \ldots, 10$.

We solve the problem with and without weights for $10,000$ instances of the noise, and consider x_2 (the exact value is 2.04).

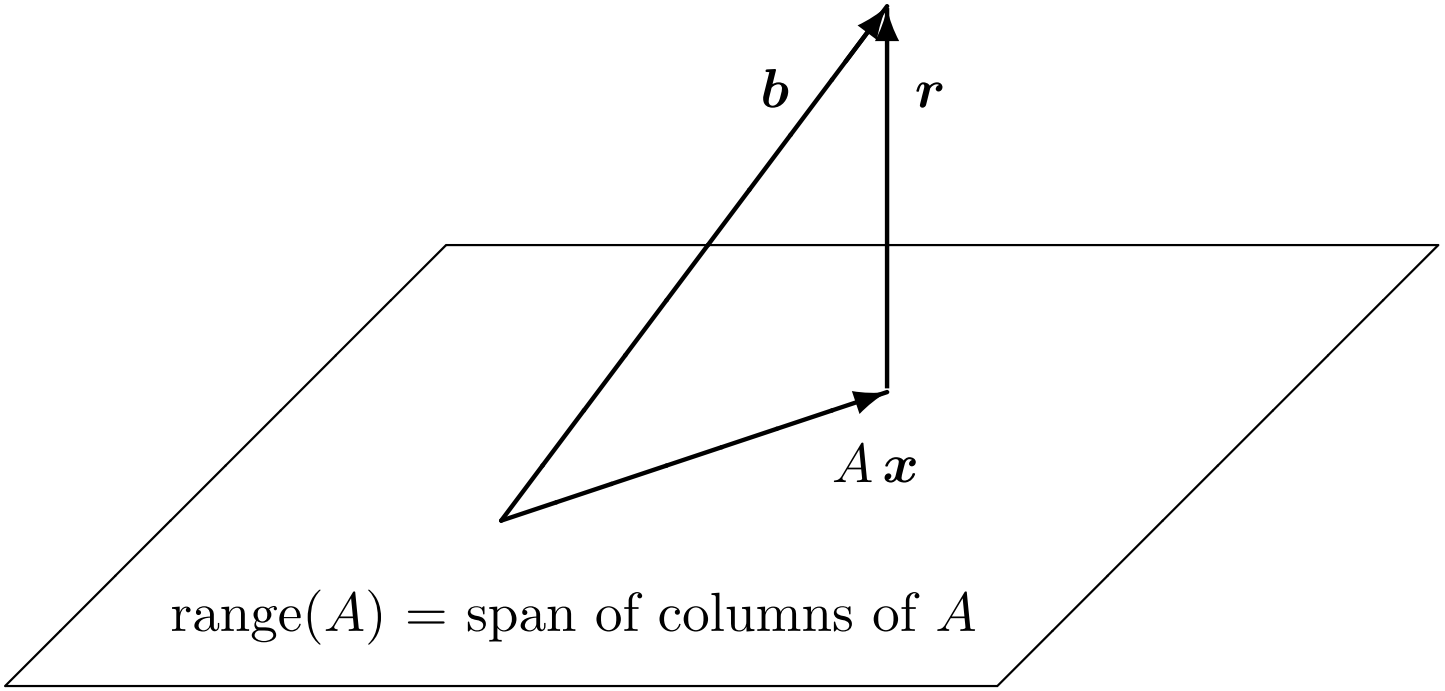


Geometric Characterization of the LSQ Solution

Our notation in Chapter 2 (no weights, and $\boldsymbol{y} \to \boldsymbol{b}$):

$$\boldsymbol{x}^* = \operatorname{argmin}_{\boldsymbol{x}} \|\boldsymbol{r}\|_2, \qquad \boldsymbol{r} = \boldsymbol{b} - A\,\boldsymbol{x}.$$

Geometric interpretation – smallest residual when $\boldsymbol{r} \perp \text{range}(A)$:



Derivation of Closed-Form Solution

The two components of $\boldsymbol{b} = A\,\boldsymbol{x} + \boldsymbol{r}$ must be orthogonal:

$$(A\,\boldsymbol{x})^T \boldsymbol{r} = 0 \quad \Leftrightarrow \quad \boldsymbol{x}^T A^T (\boldsymbol{b} - A\,\boldsymbol{x}) = 0 \quad \Leftrightarrow$$

$$\boldsymbol{x}^T A^T \boldsymbol{b} - \boldsymbol{x}^T A^T A\,\boldsymbol{x} = 0 \quad \Leftrightarrow \quad \boldsymbol{x}^T (A^T \boldsymbol{b} - A^T A\,\boldsymbol{x}) = 0$$

which leads to the *normal equations:*

$$A^T A\,\boldsymbol{x} = A^T \boldsymbol{b} \quad \Leftrightarrow \quad \boldsymbol{x}^* = (A^T A)^{-1} A^T \boldsymbol{b}.$$

Computational aspect – sensitivity to rounding errors controlled by

$$\mathrm{cond}(A^T A) = \mathrm{cond}(A)^2.$$

A side remark: the matrix $A^\dagger = (A^T A)^{-1} A^T$ in the above expression for $\boldsymbol{x}^*$ is called the *pseudoinverse* of A.

Avoiding the Normal Equations

QR factorization of $A \in \mathbb{R}^{m\times n}$ with $m > n$:

$$A = Q \begin{pmatrix} R_1 \\ 0 \end{pmatrix} \qquad \text{with} \qquad Q \in \mathbb{R}^{m\times m}, \quad R_1 \in \mathbb{R}^{n\times n},$$

where R_1 is upper triangular and Q is orthogonal, i.e.,

$$Q^T Q = I_m, \qquad Q\,Q^T = I_m, \qquad \|Q\,\boldsymbol{v}\|_2 = \|\boldsymbol{v}\|_2 \text{ for any } \boldsymbol{v}.$$

Splitting: $Q = (\ Q_1,\ Q_2\ ) \quad \Rightarrow \quad A = Q_1 R_1$, where $Q_1 \in \mathbb{R}^{m\times n}$.

`[Q,R] = qr(A)` $\rightarrow$ full QR factorization, `Q` $= Q$ and `R` $= \begin{pmatrix} R_1 \\ 0 \end{pmatrix}$.

`[Q,R] = qr(A,0)` $\rightarrow$ economy-size version with `Q` $= Q_1$ & `R` $= R_1$.

Least squares solution

$$\begin{aligned}
\|\boldsymbol{b} - A\,\boldsymbol{x}\|_2^2 &= \left\| \boldsymbol{b} - Q \begin{pmatrix} R_1 \\ 0 \end{pmatrix} \boldsymbol{x} \right\|_2^2 = \left\| Q \left[Q^T \boldsymbol{b} - \begin{pmatrix} R_1 \\ 0 \end{pmatrix} \boldsymbol{x} \right] \right\|_2^2 \\
&= \left\| \begin{pmatrix} Q_1^T \boldsymbol{b} \\ Q_2^T \boldsymbol{b} \end{pmatrix} - \begin{pmatrix} R_1 \\ 0 \end{pmatrix} \boldsymbol{x} \right\|_2^2 = \left\| \begin{pmatrix} Q_1^T \boldsymbol{b} - R_1\,\boldsymbol{x} \\ Q_2^T \boldsymbol{b} \end{pmatrix} \right\|_2^2 \\
&= \|Q_1^T \boldsymbol{b} - R_1\,\boldsymbol{x}\|_2^2 + \|Q_2^T \boldsymbol{b}\|_2^2 .
\end{aligned}$$

Last term is independent of $\boldsymbol{x}$. First term is minimal when

$$R_1\,\boldsymbol{x}^* = Q_1^T \boldsymbol{b} \quad \Leftrightarrow \quad \boldsymbol{x}^* = R_1^{-1} Q_1^T \boldsymbol{b}\ .$$

That's easy! Multiply with Q_1^T followed by backsolve with R_1.

Sensitivity to rounding errors controlled by $\mathrm{cond}(R) = \mathrm{cond}(A)$.

MATLAB: `x = A\b;` uses the QR factorization.

NMR Example Again, Again

A and Q_1:

$$\begin{pmatrix} 1.00 & 1.00 & 1 \\ 0.80 & 0.94 & 1 \\ 0.64 & 0.88 & 1 \\ \vdots & \vdots & \vdots \\ 3.2 \cdot 10^{-5} & 4.6 \cdot 10^{-2} & 1 \\ 2.5 \cdot 10^{-5} & 4.4 \cdot 10^{-2} & 1 \\ 2.0 \cdot 10^{-5} & 4.1 \cdot 10^{-2} & 1 \end{pmatrix}, \quad \begin{pmatrix} 0.597 & -0.281 & 0.172 \\ 0.479 & -0.139 & 0.071 \\ 0.384 & -0.029 & -0.002 \\ \vdots & \vdots & \vdots \\ 1.89 \cdot 10^{-5} & 0.030 & 0.224 \\ 1.52 \cdot 10^{-5} & 0.028 & 0.226 \\ 1.22 \cdot 10^{-5} & 0.026 & 0.229 \end{pmatrix},$$

$$R_1 = \begin{pmatrix} 1.67 & 2.40 & 3.02 \\ 0 & 1.54 & 5.16 \\ 0 & 0 & 3.78 \end{pmatrix}, \qquad Q_1^T \boldsymbol{b} = \begin{pmatrix} 7.81 \\ 4.32 \\ 1.19 \end{pmatrix}.$$

Normal equations vs QR factorization

Normal equations

- Squared condition number: $\text{cond}(A^T A) = \text{cond}(A)^2$.
- OK for well conditioned A.
- Work = $mn^2 + (1/3)n^3$ flops.

QR factorization

- No squaring of condition number, $\text{cond}(R) = \text{cond}(A)$
- Can better handle ill conditioned A.
- Work = $2mn^2 - (2/3)n^3$ flops.

Normal eqs. always faster, but risky for ill-conditioned matrices.

QR factorization is always stable $\rightarrow$ better black-box method.

Residual Analysis (Section 1.4)

Must choose the fitting model $M(\boldsymbol{x}, t) = \sum_{j=1}^{n} x_j \, f_j(t)$ such that the data errors and the approximation errors are balanced.

Hence we must analyze the residuals $r_1, r_2, \ldots, r_m$:

1. The model captures the pure-data function "well enough" when the approximation errors are smaller than the data errors. Then the residuals are dominated by the data errors and some of the statistical properties of the errors carry over to the residuals.
2. If the fitting model does not capture the behavior of the pure-data function, then the residuals are dominated by the approximation errors. Then the residuals will tend to behave as a sampled signal and show strong local correlations.

Use the "simplest" model (i.e., the smallest n) that satisfies 1.

Residual Analysis Assumptions and Techniques

We make the following assumptions about the data errors e_i:

- They are random variables with mean zero and identical variance, i.e., $\mathcal{E}(e_i) = 0$ and $\mathcal{E}(e_i^2) = \varsigma^2$ for $i = 1, 2, \ldots, m$.
- They belong to a normal distribution, $e_i \sim \mathcal{N}(0, \varsigma^2)$.

We describe two tests with two different properties.

- Randomness test: check for randomness of the signs of r_i.
- Autocorrelation test: check if the residuals are uncorrelated.

Test for Random Signs

Can we consider the signs of the residuals to be random?

Run test from time series analysis.

Given a sequence of two symbols – in our case, "+" and "−" for positive and negative residuals r_i – a *run* is defined as a succession of identical symbols surrounded by different symbols.

The sequence "$+ + + - - - - - + + - - - - - + + +$" has:

- $m = 17$ elements,
- $n_+ = 8$ pluses,
- $n_- = 9$ minuses, and
- $u = 5$ runs: $+ + +$, $- - - - -$, $++$, $- - - - -$, and $+ + +$.

The Run Test

The distribution of runs u (not the residuals!) can be approximated by a normal distribution with mean μ_u and standard deviation ς_u given by

$$\mu_u = \frac{2\,n_+\,n_-}{m} + 1, \qquad \varsigma_u^2 = \frac{(\mu_u - 1)\,(\mu_u - 2)}{m-1}.$$

With a 5 % significance level we will accept the sign sequence as random if

$$z_\pm = \frac{|u - \mu_u|}{\varsigma_u} < 1.96\ .$$

In the above example with 5 runs we have $z_\pm = 2.25$ and the sequence of signs cannot be considered random.

Test for Correlation

Are short sequences of residuals $r_i, r_{i+1}, r_{i+2} \ldots$ are correlated?

This is a clear indication of trends.

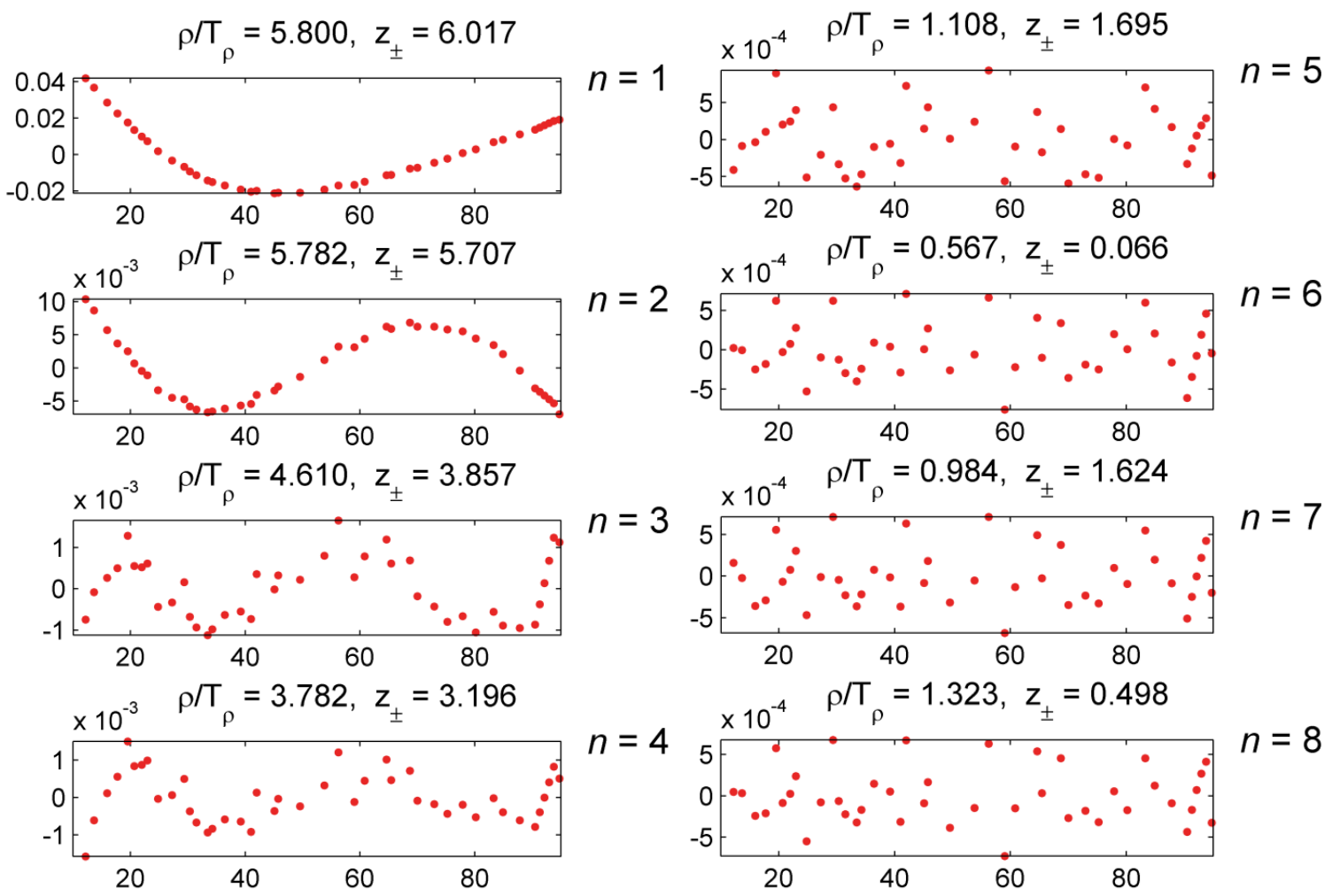
Define the *autocorrelation* ϱ of the residuals, as well as the trend threshold T_ϱ, as the quantities

$$\varrho = \sum_{i=1}^{m-1} r_i \, r_{i+1}, \qquad T_\varrho = \frac{1}{\sqrt{m-1}} \sum_{i=1}^{m} r_i^2.$$

Trends are likely to be present in the residuals if the absolute value of the autocorrelation exceeds the trend threshold, i.e., if $|\varrho| > T_\varrho$.

In some presentations, the mean of the residuals is subtracted before computing ϱ and T_ϱ.

Not necessary here, since we assume that the errors have zero mean.



Run test: for $n \geq 5$ we have $z_\pm < 1.96 \rightarrow$ random residuals.

Autocorrelation test: for $n = 6$ and 7 we have $|\varrho| < T_\varrho$.

Covariance Matrices

Recall that $b_i = \Gamma(t_i) + e_i$. Covariance matrix for $\boldsymbol{b}$:

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{b}) &= \mathcal{E}(\boldsymbol{b}\,\boldsymbol{b}^T) = \mathcal{E}\big((\boldsymbol{\Gamma} + \boldsymbol{e})\,(\boldsymbol{\Gamma} + \boldsymbol{e})\big) \\ &= \mathcal{E}\big(\boldsymbol{\Gamma}\,\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}\,\boldsymbol{e}^T + \boldsymbol{e}\,\boldsymbol{\Gamma}^T + \boldsymbol{e}\,\boldsymbol{e}^T\big) = \mathcal{E}(\boldsymbol{e}\,\boldsymbol{e}^T). \end{aligned}$$

Covariance matrix for $\boldsymbol{x}^*$:

$$\mathrm{Cov}(\boldsymbol{x}^*) = \mathrm{Cov}\big((A^T A)^{-1} A^T \boldsymbol{b}\big) = (A^T A)^{-1} A^T \mathrm{Cov}(\boldsymbol{b}) A\,(A^T A)^{-1}.$$

White noise in the data: $\mathrm{Cov}(\boldsymbol{b}) = \varsigma^2\, I_m \Rightarrow$

$$\mathrm{Cov}(\boldsymbol{x}^*) = \varsigma^2\,(A^T A)^{-1}.$$

Recall that:

$$\begin{aligned} [\mathrm{Cov}(\boldsymbol{x}^*)]_{ij} &= \mathrm{Cov}(x_i^* x_j^*), \quad i \neq j \\ [\mathrm{Cov}(\boldsymbol{x}^*)]_{ii} &= \mathrm{st.dev}(x_i^*)^2 \end{aligned}$$

Estimation of Noise Standard Deviation

White noise in the data: $\mathrm{Cov}(\boldsymbol{b}) = \varsigma^2 I_m$. Can show that:

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{r}^*) &= \varsigma^2 Q_2 Q_2^T \\ \mathcal{E}(\|\boldsymbol{r}^*\|_2^2) &= \|Q_2^T \boldsymbol{\Gamma}\|_2^2 + \mathcal{E}(\|Q_2^T \boldsymbol{e}\|_2^2) \\ &= \|Q_2^T \boldsymbol{\Gamma}\|_2^2 + (m-n)\,\varsigma^2. \end{aligned}$$

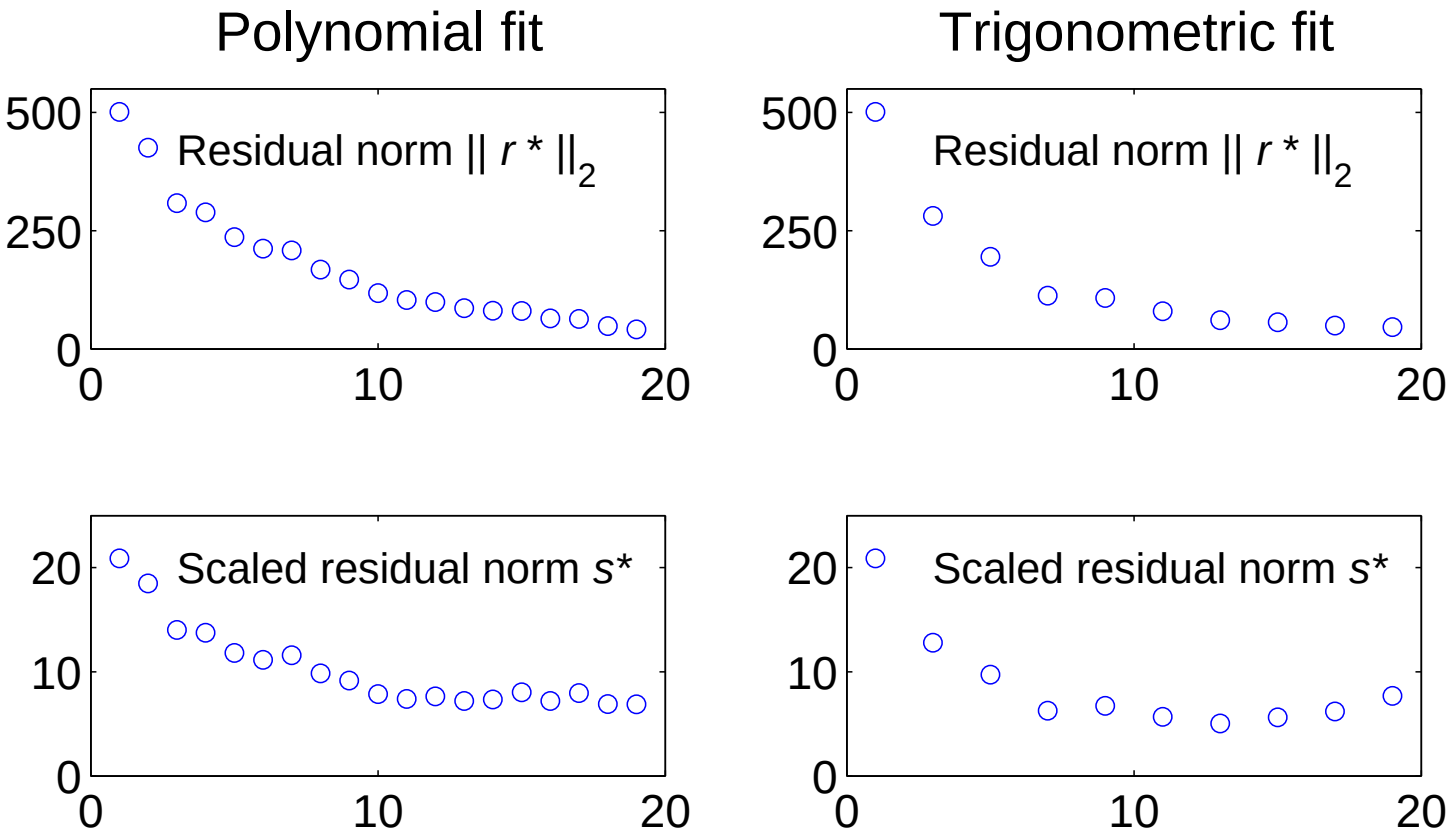
Hence, if the approximation errors are smaller than the data errors then $\|\boldsymbol{r}^*\|_2^2 \approx (m-n)\,\varsigma^2$ and the scaled residual norm

$$s^* = \frac{\|\boldsymbol{r}^*\|_2}{\sqrt{m-n}}$$

is an estimate of the standard deviation ς of the errors in the data.

Monitor s^* as a function of n to estimate ς.

Residual Norm and Scaled Residual Norm



The residual norm is monotonically decreasing while s^* has a plateau/minimum.