0.1 Least-squares methods

Check out this source [1]. The general setting is that, given data $\{(x_j, y_j)\}_{j=1,\dots,n}$ we want to find a model function $y = f(x, \mu)$ that fits the pairs (x_j, y_j) in an optimal way by finding the paramere value μ that minimises the square of the residual $r = \sum_{j=1}^{n}(y_j - f(x_j, \mu))^2$.

0.2 Linear least-squares (LLSs)

Check out this source [2]. A least-squares problem is said to be linear if it is linear w.r.t. the parameter $\mu \in \mathbb{R}^p$ even though the relationship in the data between the independent x_j and dependent variables y_j is not linear. Linear least-squares problem are useful because the minimisation problem can be formulated as a linear set of equations and written compactly in matrix form.

0.2.1 Polynomial regression

Check out this source [3]. The model function is chosen as a univariate polynomial of degree $m-1$ so that the goal of this type of regression is to find the optimal set of m coefficients $\{c_0, \dots, c_{m-1}\}$ that minimises the residual when $f(x, \mu = (c_0, \dots, c_{m-1})) = \sum_{k=1}^{m} c_k x^{k-1}$. When we fit such model to the n pairs $\{(x_j, y_j)\}_{j=1,\dots,n}$ the problem can be written in matrix form as $\boldsymbol{A}\boldsymbol{c} - \boldsymbol{b} = \boldsymbol{r} \neq \boldsymbol{0}$ where

$$\mathbb{R}^{n\times m} \ni \underbrace{\boldsymbol{A} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{m-2} & x_1^{m-1} \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{m-2} & x_n^{m-1} \end{bmatrix}}_{\text{model matrix}}, \underbrace{\boldsymbol{c} = \begin{bmatrix} c_0 \\ \vdots \\ c_{m-1} \end{bmatrix}}_{\text{parameters}}, \underbrace{\boldsymbol{b} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}_{\text{data}}.$$

Each column $k = 1, \dots, m$ of the model matrix $\boldsymbol{A}$ might be thought as the values that a given basis function $\phi_k(x)$ takes in correspondance of the data i.e. $A_{jk} = \phi_k(x_j)$. The model function can thus be written in such basis as $f(x, (c_0, \dots, c_{m-1})) = \sum_{k=1}^{m-1} c_k \phi_k(x)$. The system is overdetermined due to the fact that $\boldsymbol{A}$ is tall and skinny ($n \gg m$) and therefore a solution $\boldsymbol{c}$ is guaranteed to exist if and only if $\boldsymbol{A}$ is full rank.

[1] https://en.wikipedia.org/wiki/Least_squares
[2] https://en.wikipedia.org/wiki/Linear_least_squares
[3] https://en.wikipedia.org/wiki/Polynomial_regression

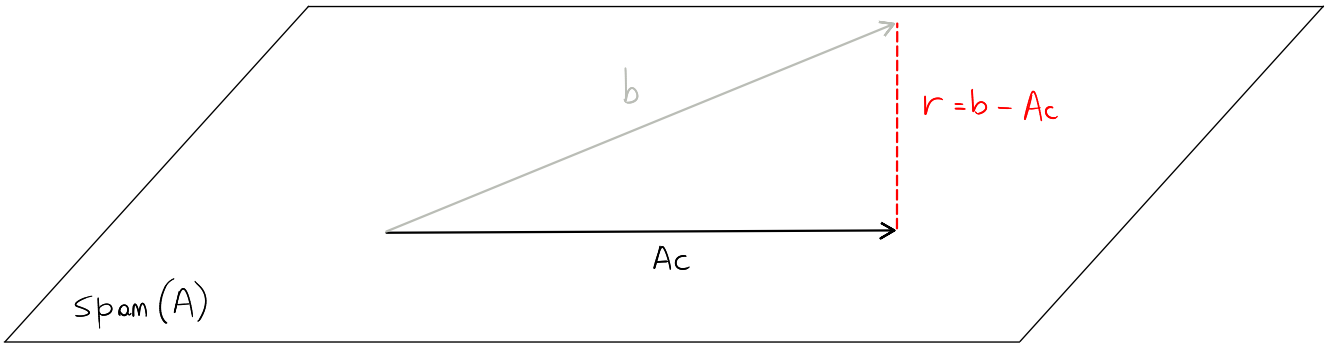

Figure 1: Visual relationship between the data $\boldsymbol{b} \in \mathbb{R}^n$, the least-squares solution $\boldsymbol{c} \in \mathbb{R}^m$ and its residual $\boldsymbol{r} \in \mathbb{R}^n$. The transformation $\boldsymbol{Ac}$ lives in span$(\boldsymbol{A})$ but $\boldsymbol{b}$ does not because $n \gg m$. To find the closest $\boldsymbol{x} \in$ span$(\boldsymbol{A})$ to $\boldsymbol{b}$ we look for the orthogonal projection of $\boldsymbol{b}$ on span$(\boldsymbol{A})$ and by minimising the norm of such distance give by $\boldsymbol{r} = \boldsymbol{b} - \boldsymbol{Ax}$.

Such solution is not unique and we are interested in finding the one that minimises the residual vector $\boldsymbol{r}$ in the L_2-norm sense. To do that we derive the analytical expression of such norm

$$\begin{aligned} ||\boldsymbol{r}||^2 &= \boldsymbol{r}^T\boldsymbol{r} = (\boldsymbol{b} - \boldsymbol{Ac})^T(\boldsymbol{b} - \boldsymbol{Ac}) = (\boldsymbol{b}^T - \boldsymbol{c}^T\boldsymbol{A}^T)(\boldsymbol{b} - \boldsymbol{Ac}) \\ &= \boldsymbol{b}^T\boldsymbol{b} - \boldsymbol{b}^T\boldsymbol{Ac} - \boldsymbol{c}^T\boldsymbol{A}^T\boldsymbol{b} + \boldsymbol{c}^T\boldsymbol{A}^T\boldsymbol{Ac}\,, \end{aligned} \tag{1}$$

We notice that since $\boldsymbol{c}^T\boldsymbol{A}^T\boldsymbol{b} = (\boldsymbol{Ac})^T\boldsymbol{b}$ is a number it will be equal to its transpose $\boldsymbol{b}^T\boldsymbol{Ac}$ which reduces (1) to

$$||\boldsymbol{r}||^2 = \boldsymbol{b}^T\boldsymbol{b} - 2\boldsymbol{b}^T\boldsymbol{Ac} + \boldsymbol{c}^T\underbrace{\boldsymbol{A}^T\boldsymbol{A}}_{\text{sym.}}\boldsymbol{c} = \boldsymbol{b}^T\boldsymbol{b} - 2\boldsymbol{b}^T\boldsymbol{Ac} + \boldsymbol{c}^T(\boldsymbol{A}^T\boldsymbol{A})^T\boldsymbol{c} =: r(\boldsymbol{c})\,, \tag{2}$$

where in the last equality we emphasize that the (squared) norm of the residual is a scalar quadratic function of the parameter vector $\boldsymbol{c}$. The quadratic dependence on $\boldsymbol{c}$ is encoded by the presence of the bilinear form [4]

$$\begin{aligned} a : V \times V &\to \mathbb{R} \\ \boldsymbol{x}, \boldsymbol{y} &\mapsto \boldsymbol{x}^T(\boldsymbol{A}^T\boldsymbol{A})^T\boldsymbol{y}\,, \end{aligned}$$

and its associated quadratic form [5]

$$\begin{aligned} q : V &\to \mathbb{R} \\ \boldsymbol{x} &\mapsto a(\boldsymbol{x}, \boldsymbol{x}) = \boldsymbol{x}^T(\boldsymbol{A}^T\boldsymbol{A})^T\boldsymbol{x}\,. \end{aligned}$$

The minimum of (2) is found by computing its gradient

$$\nabla r(\boldsymbol{c}) = 2\boldsymbol{c}^T\boldsymbol{A}^T\boldsymbol{A} - 2\boldsymbol{b}^T\boldsymbol{A}\,,$$

and imposing the equality to 0 yieling the so-called **normal equations**

$$\boldsymbol{c} = \underset{\boldsymbol{x}\in\mathbb{R}^m}{\operatorname{argmin}}\, r(\boldsymbol{x}) = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b} = \boldsymbol{A}^+\boldsymbol{b}\,, \tag{3}$$

where $(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T =: \boldsymbol{A}^+ \in \mathbb{R}^{m\times n}$ is the pseudoinverse of a non-square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times m}$ of rank m.

[4] https://en.wikipedia.org/wiki/Bilinear_form
[5] https://en.wikipedia.org/wiki/Quadratic_form

Statistical error analysis

Existing bounds in the literature Lol

New probabilistic upper bounds We formulate the polynomial least-squares problem derived above in a statistical sense so that we can perform rigorous error analysis. To start, we make explicit the dependence of the dependent variables y_j on some r.v.s which we assume to be standard distributed. The problem is thus formulated as follows: given a set of n nodes $\{x_1, \dots, x_n\}$ we record n observations of the form $y_j = f(x_j) + \xi_j$ where f is prescribing the deterministic trend in the data and it is subject to random perturbations modelled as r.v.s $\xi_j \sim \mathcal{N}(0, \sigma^2)$. The polynomial least-squares thus finds a set of $m < n$ coefficients $\{c_1, \dots, c_m\}$ by solving the normal equations (3). Let us denote $\boldsymbol{A}^+ =: \boldsymbol{H} \in \mathbb{R}^{m\times n}$, $\boldsymbol{b} =: \boldsymbol{y} \in \mathbb{R}^n$ and explicitly write the solution as $\boldsymbol{c} = \boldsymbol{H}\boldsymbol{y} \in \mathbb{R}^m$. Since we made the dependence of the observations of the dependent variables $\boldsymbol{y}$ on the normally distributed r.v.s $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)$ then we conclude that the coefficients are themselves r.v.s through a linear function $\boldsymbol{c} = g(\boldsymbol{y}) = \boldsymbol{H}\boldsymbol{y}$. Two questions arise: first is can we say something about the distribution of $\boldsymbol{c}$? And second is what are the errors associated to the solution $\boldsymbol{c}$? Given this two-fold analytical aspects we refer to this sub-paragraph as the statistical error analysis of a polynomial least-squares problem. Let us write the expression for the $j-$th coefficient as a solution of the least-squares problem

$$c_j = \sum_{k=1}^{n} H_{jk} y_k = \sum_{k=1}^{n} H_{jk} f(x_k) + \sum_{k=1}^{n} H_{jk}\xi_k = \overline{c_j} + \sum_{k=1}^{n} H_{jk}\xi_k \,. \tag{4}$$

Notice that the deterministic quantity $\overline{c_j} := \sum_{k=1}^{n} H_{jk} f(x_k)$ is nothing else then the solution of the least-squares problem (3) in the absence of noise (i.e. interpolating every observation $y_j = f(x_j)$) exactly so that the residual is $r(\boldsymbol{c}) = 0$. Furthermore $\overline{c_j}$ is the sample mean of the distribution of the c_j r.v., hence the overbar notation we adopted. It is reasonable to assume that $\mathbb{E}(c_j) \to \overline{c_j}$ in some asymptotic sense; in fact we notice that by defining the quantity

$$\hat{c}_j := c_j - \overline{c_j} = \sum_{k=1}^{n} H_{jk}\xi_k \,, \tag{5}$$

and assuming that $H_{jk} = \frac{1}{n}$, $k = 1, \dots, n$ then, by the Central Limit Theorem, we get

$$\sqrt{n}\hat{c}_j \xrightarrow{d} \mathcal{N}(0, \sigma^2) \,.$$

Since in general $H_{jk} \neq \frac{1}{n}$, we need to study the convergence of $\hat{c}_j$ with a different approach. Let us start by taking the absolute value (L_1-norm) of (5)

$$|\hat{c}_j| = |c_j - \overline{c_j}| = \left| \sum_{k=1}^{n} H_{jk}\xi_k \right| . \tag{6}$$

A naive upper bound is derived for (6) by applying the triangle inequality and realising that since $\xi_k \sim \mathcal{N}(0, \sigma^2)$ then $P(|\xi_k| \leq 3\sigma) = 1$ and hence

$$|\hat{c}_j| = \left| \sum_{k=1}^{n} H_{jk}\xi_k \right| \leq \sum_{k=1}^{n} |H_{jk}||\xi_k| \overset{a.a.}{\leq} 3\sigma \sum_{k=1}^{n} |H_{jk}| \,. \tag{7}$$

From (7) we can deduce the upper estimate

$$|\hat{c}_j| \leq 3\sigma ||H||_\infty \leq 3\sigma n ||H||_{\max} \,, \tag{8}$$

where in we used the matrix norms $||H||_\infty = \max_{j=1,\dots,m} \sum_{k=1}^{n} |H_{jk}|$ and $||H||_{\max} = \max_{j,k} |H_{jk}|$. Notice that the less tight upper estimates in (8) provides an explicit dependence on n which is linear, meaning that it will monotonically increase as the number of observations increases. This is clearly not ideal however, as evidenced by Figure 2, we know that such increase is bounded from above. Nevertheless, we wish to derive a better upper estimate that decreses monotonically with n and to do that we must somehow exploit the structure of the pseudoinverse $\boldsymbol{H}$. To do this let us observe first that $|\hat{c}_j| = |\sum_{k=1}^{n} H_{jk}\xi_k|$ is the inner product between the $j-$th row vector $\boldsymbol{H}_j \in \mathbb{R}^n$ and the vector of i.i.d. normally-distributed r.v.s $\boldsymbol{\xi}$. Therefore using Cauchy-Schwarz inequality we can write

$$|\hat{c}_j| = \left| \sum_{k=1}^{n} H_{jk}\xi_k \right| = |\boldsymbol{H}_j \cdot \boldsymbol{\xi}| \leq ||\boldsymbol{H}_j||_2 ||\boldsymbol{\xi}||_2 \leq 3\sigma\sqrt{n}||\boldsymbol{H}_j||_2 \,, \tag{9}$$

where $||\cdot||_2$ denotes the L_2 vector norm. For any rectangular matrix $\boldsymbol{M} \in \mathbb{R}^{n\times m}$ with $m < n$ this norm induces the so-called *spectral norm* which corresponds exactly with the square root of the largest eigenvalue of $\boldsymbol{M}^T\boldsymbol{M}$. Now let $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ be the singular value decomposition of $\boldsymbol{M}$, where $\mathbb{R}^{n\times m} \ni \boldsymbol{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_m)$ then the spectral norm coincides with its largest singular value i.e. $||\boldsymbol{M}||_2 = \sqrt{\max_\lambda(\boldsymbol{M}^T\boldsymbol{M})} = \sigma_1$. Finally denoting $\boldsymbol{M}^+$ as the pseudoinverse of $\boldsymbol{M}$ then its singular value decomposition is well known $\boldsymbol{M}^+ = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^T$ where $\boldsymbol{\Sigma} = \text{diag}(\sigma_m^{-1}, \dots, \sigma_1^{-1})$ so that $||\boldsymbol{M}^+|| = \frac{1}{\sigma_m}$, i.e. the spectral norm of the pseudoinverse of $\boldsymbol{M}$ is the reciprocal of the smallest singular value of $\boldsymbol{M}$.

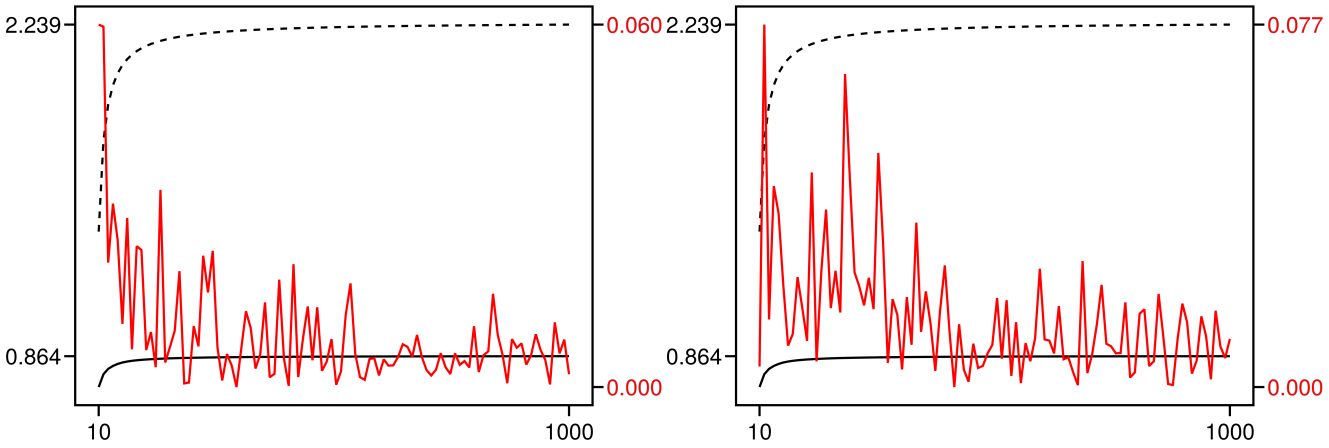

Figure 2: Comparison between the convergence with n of the exact numerical errors (red solid line) and the upper estimates $3\sigma||H||_\infty$ (black solid lines) and $3\sigma n||H||_{\max}$ (black dashed lines) in (8) for a quadratic ($m = 3$) least-square regression observations with standard deviation $\sigma = 0.1$. Linear coefficient c_1 on the left and the quadratic coefficient c_2 on the right.

Equipped with these facts we can find an upper estimate from (9)

$$|\hat{c}_j| \leq 3\sigma\sqrt{n}||\boldsymbol{H}_j||_2 \leq 3\sigma\sqrt{n}||\boldsymbol{H}||_2 = \frac{3\sigma\sqrt{n}}{\sigma_m} \,, \tag{10}$$

where σ_m denotes the smallest singular value of $\boldsymbol{A}$. We can now introduce the Frobenius matrix norm $||\boldsymbol{M}||_F = \sqrt{\sum_{j=1}^{m} \sigma_j^2}$ and state without proof that $||\boldsymbol{M}||_2 \leq ||\boldsymbol{M}||_F$. It follows trivially that $||\boldsymbol{M}||_F \leq \sigma_1\sqrt{m}$ and thus $||\boldsymbol{M}^+||_F \leq \frac{\sqrt{m}}{\sigma_m}$ which we use in (10) to get

$$|\hat{c}_j| \leq \frac{3\sigma\sqrt{n}}{\sigma_m} \leq \frac{3\sigma\sqrt{nm}}{\sigma_m} \,. \tag{11}$$

The dependence on n in (11) is still monotonically increasing albeit at a slower rate and we also made it explicit the dependence on the number of polynomial coefficients m.

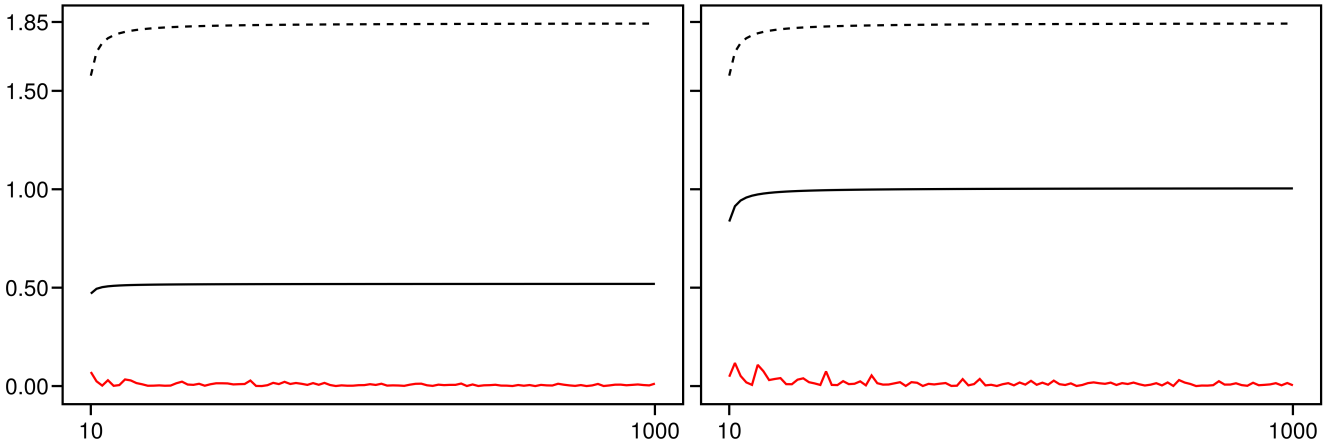

Figure 3: Comparison between the convergence with n of the exact numerical errors (red solid lines) and the upper estimates (9) (black solid lines) and (11) (black dash lines) for the same fit of Figure 2.

A new approach is required to derive the desired asymptotic decay of the error with the number of samples n. So far in fact we treated the RHS of the estimate as a statistical quantity due to the presence of the normally-distributed r.v.s ξ_j but have left the error itself on the LHS $|\hat{c}_j| = |c_j - \overline{c_j}|$ as a deterministic quantity. Therefore let us modify (6) by taking its expected value

$$\mathbb{E}(\hat{c}_j) = \mathbb{E}\left(\left|\sum_{k=1}^{n} H_{jk}\xi_k\right|\right). \tag{12}$$

In general, if x is a r.v. and g a reasonable function of x then it holds that $\mathbb{E}(g(x)) \neq g(\mathbb{E}(x))$. However, as we mentioned at the beginning of the sub-paragraph, if g is linear, then, by linearity of the expectation operator, the inequality becomes an identity $\mathbb{E}(g(x)) = g(\mathbb{E}(x))$. In addition let $\{x_1, \ldots, x_n\}$ be i.i.d. with $x_j \sim \mathcal{N}(0, \sigma^2)$, then it holds

$$y = g(\{x_1, \ldots, x_n\}) = \sum_{j=1}^{n} \alpha_j x_j \sim \mathcal{N}(0, \sigma^2 \sum_{j=1}^{n} \alpha_j^2 =: \overline{\sigma^2}). \tag{13}$$

Being $y \sim \mathcal{N}(0, \overline{\sigma^2})$ we know that $|y|$ has a folded normal distribution whose expectation value is

$$\mathbb{E}(|y|) = \sqrt{\frac{2}{\pi}}\overline{\sigma},$$

which by (13) it becomes

$$\mathbb{E}\left(\left|\sum_{j=1}^{n} \alpha_j x_j\right|\right) = \sigma\sqrt{\frac{2}{\pi}\sum_{j=1}^{n}\alpha_j^2}. \tag{14}$$

We can now put (12) into (14) to get

$$\mathbb{E}(|\hat{c}_j|) = \sigma\sqrt{\frac{2}{\pi}\sum_{k=1}^{n} H_{jk}^2} \leq \sigma\sqrt{\frac{2}{\pi}}||\boldsymbol{H}||_2 = \sigma\sqrt{\frac{2}{\pi}\max_{\lambda}(\boldsymbol{H}^T\boldsymbol{H})} = \sigma\sqrt{\frac{2}{\pi}\max_{\lambda}(\boldsymbol{H}\boldsymbol{H}^T)},$$

which for convenience we square to get

$$\mathbb{E}^2(|\hat{c}_j|) \leq \frac{2\sigma^2}{\pi}\lambda_{\max}(\boldsymbol{H}\boldsymbol{H}^T), \tag{15}$$

where we adopted the notation $\max_\lambda(\boldsymbol{H}\boldsymbol{H}^T) =: \lambda_{\max}(\boldsymbol{H}\boldsymbol{H}^T)$. Notice that since $\boldsymbol{H} = \boldsymbol{A}^+ = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$ then

$$\boldsymbol{H}\boldsymbol{H}^T = \Big((\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\Big)\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\,,$$

which by eigendecomposition it thus gives us that

$$\sigma(\boldsymbol{H}\boldsymbol{H}^T) = \frac{1}{\sigma(\boldsymbol{A}^T\boldsymbol{A})}\,,$$

and in particular

$$\lambda_{\max}(\boldsymbol{H}\boldsymbol{H}^T) = \frac{1}{\lambda_{\min}(\boldsymbol{A}^T\boldsymbol{A})}\,. \tag{16}$$

Putting (16) into (15) makes the estimate to depend explicitly on the model matrix $\boldsymbol{A}$

$$\mathbb{E}^2(|\hat{c_j}|) \leq \frac{2\sigma^2}{\pi\lambda_{\min}(\boldsymbol{A}^T\boldsymbol{A})}\,, \tag{17}$$

which now enables us to exploit the structure of $\boldsymbol{A}$ being a Vandermonde matrix. To investigate that let us first write explicitly

$$\boldsymbol{A}^T\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 & 1 \\ x_1 & x_2 & x_3 & \cdots & x_{n-1} & x_n \\ x_1^2 & x_2^2 & x_3^2 & \cdots & x_{n-1}^2 & x_n^2 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ x_1^{m-1} & x_2^{m-1} & x_3^{m-1} & \cdots & x_{n-1}^{m-1} & x_n^{m-1} \end{bmatrix} \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{m-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{m-1} \\ 1 & x_3 & x_3^2 & \cdots & x_3^{m-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n-1} & x_{n-1}^2 & \cdots & x_{n-1}^{m-1} \\ 1 & x_n & x_n^2 & \cdots & x_n^{m-1} \end{bmatrix} = \left[\sum_{\ell=1}^{n} x_\ell^{j+k-2}\right]_{j,k=1}^{m}\,.$$

We are interested in the structure of $\boldsymbol{A}^T\boldsymbol{A}$ as a function of n therefore we write

$$\boldsymbol{A}^T\boldsymbol{A} = n\left[\frac{1}{n}\sum_{\ell=1}^{n} x_\ell^{j+k-2}\right]_{j,k=1}^{m}\,. \tag{18}$$

Putting (18) into (17) gives us

$$\mathbb{E}^2(|\hat{c_j}|) \leq \frac{2\sigma^2}{\pi n\lambda_{\min}(\frac{1}{n}\boldsymbol{A}^T\boldsymbol{A})}\,, \tag{19}$$

which is already promising since the dependence on n is now monotonically decreasing compared to the previous estimates. The next step is realising that each entry in (18) is nothing else but the normalised Riemann sum of the univariate monomial $f_{j,k} = x^{j+k-2}$.

Highlight

Let $[a,b] \subset \mathbb{R}$ and $f : [a,b] \to \mathbb{R}$ bounded on $[a,b]$. Then $\forall n \in \mathbb{Z}$ we let $\{a = x_0, \ldots, x_n = b\}$ to be an ordered uniform partition of $[a,b]$ s.t. $\Delta x_j = x_j - x_{j-1} = \frac{b-a}{n}$, $j = 1, \ldots, n$. The (right) Riemann sum is thus derived as

$$S_n(f) := \sum_{j=1}^{n} f(x_j)\Delta x_j = \frac{b-a}{n}\sum_{j=1}^{n} f(x_j)\,.$$

If we let $I_{[a,b]}(f) := \int_a^b f(x)dx$ and $\overline{I_{[a,b]}}(f) := \frac{1}{b-a}I(f)$ then in the $n \to \infty$ asymptotic limit we get

$$S_n(f) \to I_{[a,b]}(f) \;\Rightarrow\; \frac{1}{b-a}S_n(f) = \frac{1}{n}\sum_{j=1}^{n} f(x_j) \to \overline{I_{[a,b]}}(f)\,.$$

Therefore we can write an asymptotic equivalence for (19)

$$\frac{1}{n}\boldsymbol{A}^T\boldsymbol{A} = \left[\frac{1}{n}\sum_{\ell=1}^{n} x_\ell^{j+k-2}\right]_{j,k=1}^{m} \xrightarrow{n\to\infty} \left[\frac{1}{b-a}\int_a^b x^{j+k-2}\right]_{j,k=1}^{m} =: \boldsymbol{G}\,,$$
$$\mathbb{E}^2(|\hat{c_j}|) \leq \frac{2\sigma^2}{\pi n \lambda_{\min}\left(\frac{1}{n}\boldsymbol{A}^T\boldsymbol{A}\right)} \xrightarrow{n\to\infty} \frac{2\sigma^2}{\pi n \lambda_{\min}(\boldsymbol{G})}\,, \tag{20}$$

where $\boldsymbol{G}$ denotes the Gram matrix associated to the inner products on the set of functions $\{\ell_j(x)\}_{j=1,\dots,m}$ given by the m monomials $\{1, x, x^2, \dots, x^{m-1}\}$ up to degree $m-1$, i.e $\ell_j(x) = x^{j-1}$

$$G_{j,k} = \frac{1}{b-a}\int_a^b \ell_j^*(x)\ell_k(x)dx = \frac{1}{b-a}\int_a^b x^{(j-1)}x^{(k-1)}dx = \frac{1}{b-a}\int_a^b x^{j+k-2}dx\,.$$

Notice that the monomials $\{1, x, x^2, \dots, x^{m-1}\}$ are L_2 functions on a closed and finite interval $[a, b]$. The new estimate (20) not only shows the desired monotonic behaviour with n but it is also surprisingly tight as shown in Figure 4 below.

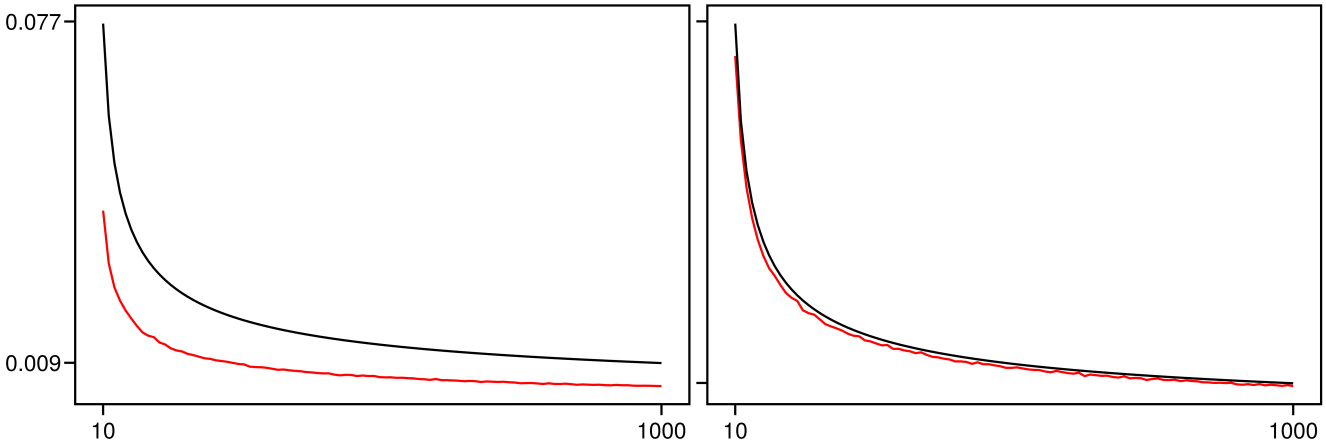

Figure 4: Comparison between the convergence with n of the expected error (red lines) averaged over an ensemble of $5 \cdot 10^3$ measurements and the square root of the upper estimate (20) (solid black lines) for the same fit of Figure 2.