Appendix P
Problems

P.1 (Least squares data fitting). Verify that the least squares fits, shown in Figure 1.1 on page 4, minimize the sums of squares of horizontal and vertical distances. The data points are:

$$d_1 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}, \quad d_2 = \begin{bmatrix} -1 \\ 4 \end{bmatrix}, \quad d_3 = \begin{bmatrix} 0 \\ 6 \end{bmatrix}, \quad d_4 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}, \quad d_5 = \begin{bmatrix} 2 \\ 1 \end{bmatrix},$$
$$d_6 = \begin{bmatrix} 2 \\ -1 \end{bmatrix}, \quad d_7 = \begin{bmatrix} 1 \\ -4 \end{bmatrix}, \quad d_8 = \begin{bmatrix} 0 \\ -6 \end{bmatrix}, \quad d_9 = \begin{bmatrix} -1 \\ -4 \end{bmatrix}, \quad d_{10} = \begin{bmatrix} -2 \\ -1 \end{bmatrix}.$$

P.2 (Distance from a data point to a linear model). The 2-norm distance from a point $d \in \mathbb{R}^{\mathtt{q}}$ to a linear static model $\mathscr{B} \subset \mathbb{R}^{\mathtt{q}}$ is defined as

$$\operatorname{dist}(d, \mathscr{B}) := \min_{\widehat{d} \in \mathscr{B}} \|d - \widehat{d}\|_2, \tag{dist}$$

i.e., $\operatorname{dist}(d, \mathscr{B})$ is the shortest distance from d to a point $\widehat{d}$ in $\mathscr{B}$. A vector $\widehat{d}^*$ that achieves the minimum of (dist) is a point in $\mathscr{B}$ that is closest to d.

Next we consider the special case when $\mathscr{B}$ is a linear static model.

1. Let
$$\mathscr{B} = \operatorname{image}(a) = \{\, \alpha a \mid \alpha \in \mathbb{R} \,\}.$$
Explain how to find $\operatorname{dist}\big(d, \operatorname{image}(a)\big)$. Find
$$\operatorname{dist}\big(\operatorname{col}(1,0), \operatorname{image}\big(\operatorname{col}(1,1)\big)\big).$$
Note that the best approximation $\widehat{d}^*$ of d in $\operatorname{image}(a)$ is the orthogonal projection of d onto $\operatorname{image}(a)$.
2. Let $\mathscr{B} = \operatorname{image}(P)$, where P is a given full column rank matrix. Explain how to find $\operatorname{dist}(d, \mathscr{B})$.
3. Let $\mathscr{B} = \ker(R)$, where R is a given full row rank matrix. Explain how to find $\operatorname{dist}(d, \mathscr{B})$.
4. Prove that in the linear static case, a solution $\widehat{d}^*$ of (dist) is always unique?

5. Prove that in the linear static case, the approximation error $\Delta d^* := d - \widehat{d}^*$ is orthogonal to $\mathscr{B}$. Is the converse true, *i.e.*, is it true that if for some $\widehat{d}$, $d - \widehat{d}$ is orthogonal to $\mathscr{B}$, then $\widehat{d} = \widehat{d}^*$?

P.3 (Distance from a data point to an affine model). Consider again the distance $\operatorname{dist}(d, \mathscr{B})$ defined in (dist). In this problem, $\mathscr{B}$ is an affine static model, *i.e.*,

$$\mathscr{B} = \mathscr{B}' + a,$$

where $\mathscr{B}'$ is a linear static model and a is a fixed vector.

1. Explain how to reduce the problem of computing the distance from a point to an affine static model to an equivalent problem of computing the distance from a point to a linear static model (Problem P.2).
2. Find
$$\operatorname{dist}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \ker(\begin{bmatrix} 1 & 1 \end{bmatrix}) + \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right).$$

P.4 (Geometric interpretation of the total least squares problem). Show that the total least squares problem

$$\begin{aligned} &\text{minimize} \quad \text{over } x \in \mathbb{R},\ \widehat{a} \in \mathbb{R}^N, \text{ and } \widehat{b} \in \mathbb{R}^N \quad \sum_{j=1}^{N} \left\| d_j - \begin{bmatrix} \widehat{a}_j \\ \widehat{b}_j \end{bmatrix} \right\|_2^2 \\ &\text{subject to} \quad \widehat{a}_j x = \widehat{b}_j, \quad \text{for } j = 1, \ldots, N \end{aligned} \tag{tls}$$

minimizes the sum of the squared orthogonal distances from the data points $d_1, \ldots, d_N$ to the fitting line

$$\mathscr{B} = \{\, \operatorname{col}(a,b) \mid xa = b \,\}$$

over all lines passing through the origin, except for the vertical line.

P.5 (Unconstrained problem, equivalent to the total least squares problem). A total least squares approximate solution x_{tls} of the linear system of equations $Ax \approx b$ is a solution to the following optimization problem

$$\text{minimize} \quad \text{over } x, \widehat{A}, \text{ and } \widehat{b} \quad \left\| \begin{bmatrix} A & b \end{bmatrix} - \begin{bmatrix} \widehat{A} & \widehat{b} \end{bmatrix} \right\|_{\text{F}}^2 \quad \text{subject to} \quad \widehat{A}x = \widehat{b}. \tag{TLS}$$

Show that (TLS) is equivalent to the unconstrained optimization problem

$$\text{minimize } f_{\text{tls}}(x), \quad \text{where} \quad f_{\text{tls}}(x) := \frac{\|Ax - b\|_2^2}{\|x\|_2^2 + 1}. \tag{TLS'}$$

Give an interpretation of the function f_{tls}.

P.6 (Lack of total least squares solution). Using the formulation (TLS'), derived in Problem P.5, show that the total least squares line fitting problem (tls) has no solution for the data in Problem P.1.

P.7 (Geometric interpretation of rank-1 approximation). Show that the rank-1 approximation problems

$$\begin{array}{ll} \text{minimize} & \text{over } R \in \mathbb{R}^{1\times 2},\ R \neq 0, \text{ and } \widehat{D} \in \mathbb{R}^{2\times N} \quad \|D-\widehat{D}\|_{\mathrm{F}}^2 \\ \text{subject to} & R\widehat{D} = 0. \end{array} \tag{lra$_R$}$$

and

$$\begin{array}{ll} \text{minimize} & \text{over } P \in \mathbb{R}^{2\times 1} \text{ and } L \in \mathbb{R}^{1\times N} \quad \|D-\widehat{D}\|_{\mathrm{F}}^2 \\ \text{subject to} & \widehat{D} = PL. \end{array} \tag{lra$_P$}$$

minimize the sum of the squared orthogonal distances from the data points $d_1,\ldots,d_N$ to the fitting line $\mathscr{B} = \ker(P) = \operatorname{image}(P)$ over all lines passing through the origin. Compare and contrast with the similar statement in Problem P.4.

P.8 (Quadratically constrained problem, equivalent to rank-1 approximation). Show that (lra$_P$) is equivalent to the quadratically constrained optimization problem

$$\text{minimize } f_{\mathrm{lra}}(P) \quad \text{subject to} \quad P^\top P = 1, \tag{lra$'_P$}$$

where

$$f_{\mathrm{lra}}(P) = \operatorname{trace}\big(D^\top (I - PP^\top) D\big).$$

Explain how to find all solutions of (lra$_P$) from a solution of (lra$'_P$). Assuming that a solution to (lra$'_P$) exists, is it unique?

P.9 (Line fitting by rank-1 approximation). Plot the cost function $f_{\mathrm{lra}}(P)$ for the data in Problem P.1 over all P such that $P^\top P = 1$. Find from the graph of f_{lra} the minimum points. Using the link between (lra$'_P$) and (lra$_P$), established in Problem P.7, interpret the minimum points of f_{lra} in terms of the line fitting problem for the data in Problem P.1. Compare and contrast with the total least squares approach, used in Problem P.6.

P.10 (Analytic solution of a rank-1 approximation problem). Show that for the data in Problem P.1,

$$f_{\mathrm{lra}}(P) = P^\top \begin{bmatrix} 140 & 0 \\ 0 & 20 \end{bmatrix} P.$$

Using geometric or analytic arguments, conclude that the minimum of f_{lra} for a P on the unit circle is 20 and is achieved for

$$P^{*,1} = \operatorname{col}(0,1) \qquad \text{and} \qquad P^{*,2} = \operatorname{col}(0,-1).$$

Compare the results with those obtained in Problem P.9.

P.11 (Analytic solution of two-variate rank-1 approximation problem). Find an analytic solution of the Frobenius norm rank-1 approximation of a $2\times N$ matrix.

P.12 (Analytic solution of scalar total least squares). Find an analytic expression for the total least squares solution of the system $ax \approx b$, where $a, b \in \mathbb{R}^m$.

P.13 (Alternating projections algorithm for low-rank approximation). In this problem, we consider a numerical method for rank-r approximation:

$$\begin{array}{ll} \text{minimize} & \text{over } \widehat{D} \quad \|D-\widehat{D}\|_{\mathrm{F}}^2 \\ \text{subject to} & \operatorname{rank}(\widehat{D}) \leq \mathtt{m}. \end{array} \tag{LRA}$$

The alternating projections algorithm, outlined next, is based on an image representation $\widehat{D} = PL$, where $P \in \mathbb{R}^{\mathtt{q}\times\mathtt{m}}$ and $L \in \mathbb{R}^{\mathtt{m}\times N}$, of the rank constraint.

Algorithm 8 Alternating projections algorithm for low rank approximation

Input: A matrix $D \in \mathbb{R}^{\mathtt{q}\times N}$, with $\mathtt{q} \leq N$, an initial approximation $\widehat{D}^{(0)} = P^{(0)}L^{(0)}$, $P^{(0)} \in \mathbb{R}^{\mathtt{q}\times\mathtt{m}}$, $L^{(0)} \in \mathbb{R}^{\mathtt{m}\times N}$, with $\mathtt{m} \leq \mathtt{q}$, and a convergence tolerance $\varepsilon > 0$.

1: Set $k := 0$.
2: **repeat**
3: $\quad k := k+1$.
4: $\quad$ Solve: $P^{(k+1)} := \arg\min_P \|D - PL^{(k)}\|_{\mathrm{F}}^2$
5: $\quad$ Solve: $L^{(k+1)} := \arg\min_L \|D - P^{(k+1)}L\|_{\mathrm{F}}^2$
6: $\quad \widehat{D}^{(k+1)} := P^{(k+1)}L^{(k+1)}$
7: **until** $\|\widehat{D}^{(k)} - \widehat{D}^{(k+1)}\|_{\mathrm{F}} < \varepsilon$

Output: Output the matrix $\widehat{D}^{(k+1)}$.

1. Implement the algorithm and test it on random data matrices D of different dimensions with different rank specifications and initial approximations. Plot the approximation errors

$$e_k := \|D - \widehat{D}^{(k)}\|_{\mathrm{F}}^2, \qquad \text{for } k = 0,1,\ldots$$

as a function of the iteration step k and comment on the results.

* 2. Give a proof or a counter example for the conjecture that the sequence of approximation errors $e := (e_0, e_1, \ldots)$ is well defined, independent of the data and the initial approximation.

* 3. Assuming that e is well defined. Give a proof or a counter example for the conjecture that e converges monotonically to a limit point e_∞.

* 4. Assuming that e_∞ exists, give proofs or counter examples for the conjectures that e_∞ is a local minimum of (LRA) and e_∞ is a global minimum of (LRA).

P.14 (Two-sided weighted low rank approximation). Prove Theorem 2.29 on page 65.

P.15 (Most poweful unfalsified model for autonomous models). Given a trajectory

$$y = \big(y(1), y(2), \ldots, y(T)\big)$$

of an autonomous linear time-invariant system $\mathscr{B}$ of order $\mathtt{n}$, find a state space representation $\mathscr{B}_{\mathrm{i/s/o}}(A,C)$ of $\mathscr{B}$. Modify your procedure, so that it does not require prior knowledge of the system order $\mathtt{n}$ but only an upper bound $\mathtt{n}_{\max}$ for it.

P.16 (Algorithm for exact system identification). Develop an algorithm for exact system identification that computes a kernel representation of the model, *i.e.*, implement the mapping

$$w_{\mathrm{d}} \mapsto R(z), \qquad \text{where } \widehat{\mathscr{B}} := \ker\big(R(z)\big) \text{ is the identified model.}$$

You can assume that the system is single input single output and its order is known.

* **P.17 (When is $\mathscr{B}_{\mathrm{mpum}}(w_{\mathrm{d}})$ equal to the data generating system?).** Choose a (random) linear time-invariant system $\mathscr{B}_0$ (the "true data generating system") and a trajectory $w_{\mathrm{d}} = (u_{\mathrm{d}}, y_{\mathrm{d}})$ of $\mathscr{B}_0$. The aim is to recover the data generating system $\mathscr{B}_0$ back from the data w_{d}. Conjecture that this can be done by computing the most powerful unfalsified model $\mathscr{B}_{\mathrm{mpum}}(w_{\mathrm{d}})$. Verify whether and when in simulation $\mathscr{B}_{\mathrm{mpum}}(w_{\mathrm{d}})$ coincides with $\mathscr{B}_0$. Find counter examples when the conjecture is not true and based on this experience revise the conjecture. Find sufficient conditions for $\mathscr{B}_{\mathrm{mpum}}(w_{\mathrm{d}}) = \mathscr{B}_0$.

P.18 (Algorithms for approximate system identification).

1. Download the file `flutter.dat` from a Database for System Identification (De Moor, 1999).
2. Apply the function developed in Problem P.16 on the flutter data.
3. Compute the misfit between the flutter data and the model obtained in step 1.
4. *Misfit minimization* Partition the flutter data set into identification, *e.g.*, first 60%, and validation, *e.g.*, remaining 40%, parts. Compute a locally optimal model with lag $\mathtt{l} = 3$ for the identification part of the data. Validate the identified model by computing the misfit on the validation part of the data.

P.19 (Computing approximate common divisor with `slra`). Given polynomials p and q of degree n or less and an integer $\mathtt{d} < \mathtt{n}$, use `slra` to solve the Sylvester structured low rank approximation problem

$$\begin{aligned} &\text{minimize} && \text{over } \widehat{p}, \widehat{q} \in \mathbb{R}^{n+1} \quad \big\| [p \;\, q] - [\widehat{p} \;\, \widehat{q}] \big\|_{\mathrm{F}} \\ &\text{subject to} && \operatorname{rank}\big(\mathscr{R}_{\mathtt{d}}(\widehat{p}, \widehat{q})\big) \leq 2n - 2\mathtt{d} + 1 \end{aligned}$$

in order to compute an approximate common divisor c of p and q with degree at least $\mathtt{d}$. Verify the answer with the alternative approach developed in Section 3.2.

P.20 (Matrix centering). Prove Proposition 5.5.

P.21 (Mean computation as an optimal modeling). Prove Proposition 5.6.

P.22 (Nonnegative low rank approximation). Implement and test the algorithm for nonnegative low rank approximation (Algorithm 7 on page 176).

P.23 ((Luenberger, 1979, Page 53)). A thermometer reading 21°C, which has been inside a house for a long time, is taken outside. After one minute the thermometer reads 15°C; after two minutes it reads 11°C. What is the outside temperature? (According to Newton's law of cooling, an object of higher temperature than its environment cools at a rate that is proportional to the difference in temperature.)

P.24. Solve first Problem P.23. Consider the system of equations

$$\big[\mathbf{1}_{T-\mathtt{n}} \otimes G \quad \mathscr{H}_{T-\mathtt{n}}(\Delta y)\big] \begin{bmatrix} \bar{u} \\ \ell \end{bmatrix} = \operatorname{col}\Big(y\big((\mathtt{n}+1)t_s\big), \cdots, y\big(Tt_s\big)\Big), \qquad \text{(SYS DD)}$$

(the data driven-driven algorithm for input estimation on page 210) in the case of a first order single input single output system and three data points. Show that the solution of the system (SYS DD) coincides with the one obtained in Problem P.23.

P.25. Consider the system of equations (SYS DD) in the case of a first order single input single output system and N data points. Derive an explicit formula for the least squares approximate solution of (SYS DD). Propose a recursive algorithm that updates the current solution when new data point is obtained.

P.26. Solve first Problem P.25. Implement the solution obtained in Problem P.25 and validate it against the function `stepid_dd`.

References

De Moor B (1999) DaISy: database for the identification of systems. `www.esat.kuleuven.be/sista/daisy/`

Luenberger DG (1979) Introduction to Dynamical Systems: Theory, Models and Applications. John Wiley

Appendix S
Solutions

P.1 (Least squares data fitting). Minimization of the vertical distances (lse) for the data in the example is

$$\underbrace{\mathrm{col}(-2,-1,0,1,2,2,1,0,-1,-2)}_{\mathbf{a}}\, x = \underbrace{\mathrm{col}(1,4,6,4,1,-1,-4,-6,-4,-1)}_{\mathbf{b}}.$$

The least squares approximate solution is given by

$$x_{\mathrm{ls}} = (\mathbf{a}^\top \mathbf{a})^{-1}\mathbf{a}^\top \mathbf{b} = \frac{-2-4+0+4+2-2-4+0+4+2}{\mathbf{a}^\top \mathbf{a}} = 0,$$

so that the corresponding fitting line is

$$\mathscr{B}_{\mathrm{ls}} = \{\, d = \mathrm{col}(a,0) \mid a \in \mathbb{R}\,\}$$

the horizontal line passing through the origin.

Minimization of the horizontal distances (lse′), is $\mathbf{a} = \mathbf{b}x'$, with the $\mathbf{a}$ and $\mathbf{b}$ defined above. The least squares approximate solution in this case is

$$x'_{\mathrm{ls}} = (\mathbf{b}^\top \mathbf{b})^{-1}\mathbf{b}^\top \mathbf{a} = 0,$$

so that the corresponding fitting line is

$$\mathscr{B}'_{\mathrm{ls}} = \{\, d = \mathrm{col}(0,b) \mid b \in \mathbb{R}\,\}$$

the vertical line passing through the origin.

P.2 (Distance from a data point to a linear model).

1. Using the image representation $\mathrm{image}(P)$ of the model $\mathscr{B}$, the distance computation problem (tls”) is equivalent to the standard least squares problem

$$\mathrm{dist}(d,\mathscr{B}) := \min \|d - \widehat{d}\|_2 \quad \text{subject to} \quad \widehat{d} = P\ell.$$

Therefore, assuming that P is full column rank, the best approximation is

$$\widehat{d}^* = P(P^\top P)^{-1}P^\top d =: \Pi_P d \qquad (d^*)$$

and the distance of d to $\mathscr{B}$ is

$$\mathrm{dist}(d,\mathscr{B}) = \|d - \widehat{d}^*\|_2 = \sqrt{d^\top (I - \Pi_P)d}. \qquad (\mathrm{dist}_P)$$

The assumption that “P is full column rank” can be done without loss of generality because there are aways full column rank P’s such that $\mathrm{image}(P) = \mathscr{B}$ (choose any basis for $\mathscr{B}$).

2. Using the kernel representation $\ker(R)$ of the model $\mathscr{B}$, the distance computation problem (tls”) is equivalent to the problem

$$\mathrm{dist}(d,\mathscr{B}) := \min_{\widehat{d}} \|d - \widehat{d}\|_2 \quad \text{subject to} \quad R\widehat{d} = 0.$$

As written, this problem is not a standard least squares problem, however, with the change of variables $\Delta d := d - \widehat{d}$ it can be rewritten as an equivalent ordinary least norm problem

$$\mathrm{dist}(d,\mathscr{B}) := \min_{\widehat{d}} \|\Delta d\|_2 \quad \text{subject to} \quad R\Delta d = Rd.$$

Therefore, assuming that R is full row rank,

$$\Delta d^* = R^\top (RR^\top)^{-1}Rd = \Pi_{R^\top} d$$

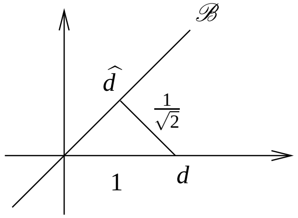
and

$$\mathrm{dist}(d,\mathscr{B}) = \|\Delta d^*\|_2 = \sqrt{d^\top \Pi_{R^\top} d}. \qquad (\mathrm{dist}_R)$$

Again, the assumption that R is full row rank is done without loss of generality because there are full row rank matrices R, such that $\ker(R) = \mathscr{B}$.

3. Substituting $d = \left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$ and $P = \left[\begin{smallmatrix}1\\1\end{smallmatrix}\right]$ in (dist_P), we have

$$\begin{aligned}
\mathrm{dist}\left(\begin{bmatrix}1\\0\end{bmatrix}, \mathrm{image}\left(\begin{bmatrix}1\\1\end{bmatrix}\right)\right) &= \sqrt{\begin{bmatrix}1 & 0\end{bmatrix}\left(\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix} - \begin{bmatrix}1\\1\end{bmatrix}\left(\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}1\\1\end{bmatrix}\right)^{-1}\begin{bmatrix}1 & 1\end{bmatrix}\right)\begin{bmatrix}1\\0\end{bmatrix}} \\
&= \sqrt{\begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}1/2 & -1/2\\-1/2 & 1/2\end{bmatrix}\begin{bmatrix}1\\0\end{bmatrix}} = 1/\sqrt{2}
\end{aligned}$$



4. As shown in part 1, $\widehat{d}^*$ is unique (and can be computed by, *e.g.*, (d^*) and (dist$_R$)).
5. A vector Δd is orthogonal to the model $\mathscr{B}$ if and only if Δd is orthogonal to all vectors in $\mathscr{B}$. Using (d^*) and the basis P for $\mathscr{B}$, we have

$$\Delta d^{*\top} P = (d - \widehat{d}^*)^\top P = d^\top (I - \Pi_P) P = 0,$$

which shows that is Δd^* is orthogonal to $\mathscr{B}$.
The converse statement "$\Delta d = d - \widehat{d}$ being orthogonal to $\mathscr{B}$ implies that $\widehat{d}$ is the closest point in $\mathscr{B}$ to d" is also true. It completes the proof of what is known as the *orthogonality principle*—$\widehat{d}$ is an optimal approximation of a point d in a model $\mathscr{B}$ if and only if the approximation error $d - \widehat{d}$ is orthogonal to $\mathscr{B}$.

P.3 (Distance from a data point to an affine model).

- The problem of computing $\operatorname{dist}(d, \mathscr{B})$ reduces to an equivalent problem of computing the distance of a point to a subspace by the change of variables

$$d' := d - a.$$

We have

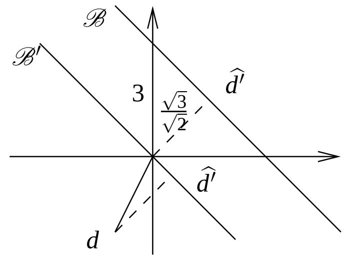
$$\operatorname{dist}(d, \mathscr{B}) = \min_{\widehat{d} \in \mathscr{B}} \|d - \widehat{d}\|_2 = \min_{\widehat{d}' \in \mathscr{B}'} \|d' - \widehat{d}'\|_2 = \operatorname{dist}(d', \mathscr{B}').$$

- Using the change of variables argument we have

$$\operatorname{dist}\left(\begin{bmatrix}0\\0\end{bmatrix}, \ker([1\ 1]) + \begin{bmatrix}1\\2\end{bmatrix}\right) = \operatorname{dist}\left(-\begin{bmatrix}1\\2\end{bmatrix}, \ker([1\ 1])\right).$$

Then using (dist$_R$) we have

$$\operatorname{dist}\left(-\begin{bmatrix}1\\2\end{bmatrix}, \ker([1\ 1])\right) = \sqrt{[1\ 2]\begin{bmatrix}1\\1\end{bmatrix}\left([1\ 1]\begin{bmatrix}1\\1\end{bmatrix}\right)^{-1}[1\ 1]\begin{bmatrix}1\\2\end{bmatrix}} = \sqrt{9/2}.$$



P.4 (Geometric interpretation of the total least squares problem). The constraint of (tls),

$$\widehat{a}_j x = \widehat{b}_j, \quad \text{for } j = 1, \dots, N$$

is equivalent to the constraint that

$$\widehat{d}_1 := (\widehat{a}_1, \widehat{b}_1), \quad \dots, \quad \widehat{d}_N := (\widehat{a}_N, \widehat{b}_N)$$

lie on the line

$$\mathscr{B}_{\text{i/o}}(x) := \{\, d = \operatorname{col}(a, b) \in \mathbb{R}^2 \mid ax = b \,\},$$

i.e., (tls) can be written as

$$\begin{array}{ll} \text{minimize} & \text{over } x \text{ and } \widehat{d}_1, \dots, \widehat{d}_N \quad \sum_{j=1}^N \|d_j - \widehat{d}_j\|_2^2 \\ \text{subject to} & \widehat{d}_j \in \mathscr{B}_{\text{i/o}}(x), \quad \text{for } j = 1, \dots, N. \end{array} \qquad \text{(tls')}$$

In turn, problem (tls') is equivalent to minimization of the function $f_{\text{tls}} : \mathbb{R} \to \mathbb{R}$ defined by

$$\begin{aligned} f_{\text{tls}}(x) := \min_{\widehat{d}_1, \dots, \widehat{d}_N} \sum_{j=1}^N \|d_j - \widehat{d}_j\|_2^2 \\ \text{subject to} \quad \widehat{d}_j \in \mathscr{B}_{\text{i/o}}(x), \quad \text{for } j = 1, \dots, N. \end{aligned} \qquad \text{(tls'')}$$

The minimization in (tls'') is separable in the variables $\widehat{d}_1, \dots, \widehat{d}_N$, *i.e.*, (tls'') decouples into N independent problems

$$f_{\text{tls},i}(x) = \min_{\widehat{d}_j} \|d_j - \widehat{d}_j\|_2^2 \quad \text{subject to} \quad \widehat{d}_j \in \mathscr{B}_{\text{i/o}}(x).$$

By the orthogonality principle, $f_{\text{tls},j}(x)$ is the squared orthogonal distance from d_j to the line $\mathscr{B}_{\text{i/o}}(x)$. Subsequently,

$$f_{\text{tls}}(x) = \sum_{j=1}^N f_{\text{tls},j}(x)$$

is the sum of squared orthogonal distances from the data points to the line $\mathscr{B}_{\text{i/o}}(x)$.

For any $x \in \mathbb{R}$, $\mathscr{B}_{\mathrm{i/o}}(x)$ is a line passing through the origin and any line passing through the origin, except for the vertical line, corresponds to a set $\mathscr{B}_{\mathrm{i/o}}(x)$, for some $x \in \mathbb{R}$. Therefore, the total least squares problem $\min_{x\in\mathbb{R}} f_{\mathrm{tls}}(x)$ minimizes the sum of squared orthogonal distances from the data points to a line, over all lines passing through the origin, except for the vertical line.

P.5 (Unconstrained problem, equivalent to the total least squares problem). The total least squares approximation problem (TLS) is $\min_x f_{\mathrm{tls}}(x)$, where

$$f_{\mathrm{tls}}(x) = \min_{\widehat{A},\widehat{b}} \left\| \begin{bmatrix} A & b \end{bmatrix} - \begin{bmatrix} \widehat{A} & \widehat{b} \end{bmatrix} \right\|_{\mathrm{F}}^2 \quad \text{subject to} \quad \widehat{A}x = \widehat{b} \qquad (f_{\mathrm{tls}})$$

or with the change of variables $\Delta A := A - \widehat{A}$ and $\Delta b := b - \widehat{b}$,

$$f_{\mathrm{tls}}(x) = \min_{\Delta A, \Delta b} \left\| \begin{bmatrix} \Delta A & \Delta b \end{bmatrix} \right\|_{\mathrm{F}}^2 \quad \text{subject to} \quad Ax - b = \Delta A x - \Delta b. \qquad (f'_{\mathrm{tls}})$$

Define

$$\Delta b := Ax - b, \qquad \Delta D := \begin{bmatrix} \Delta A & \Delta b \end{bmatrix}^\top, \quad \text{and} \quad r = \begin{bmatrix} x^\top & -1 \end{bmatrix}$$

in order to write (f'_{tls}) as a standard linear least norm problem

$$\min_{\Delta D} \|\Delta D\|_{\mathrm{F}}^2 \quad \text{subject to} \quad r\Delta D = \Delta b^\top.$$

The least norm solution for ΔD is

$$\Delta D^* = \frac{r^\top \Delta b}{rr^\top},$$

so that, we have

$$f_{\mathrm{tls}}(x) = \|\Delta D^*\|_{\mathrm{F}}^2 = \operatorname{trace}\left((\Delta D^*)^\top D^*\right) = \frac{\Delta b^\top \Delta b}{rr^\top} = \frac{\|Ax-b\|^2}{\|x\|^2+1}.$$

From Problem P.4 and the derivation of f_{tls}, we see that $f_{\mathrm{tls}}(x)$ is the sum of squared orthogonal distances from the data points to the model $\mathscr{B}_{\mathrm{i/o}}(x)$, defined by x.

P.6 (Lack of total least squares solution). The total least squares line fitting method, applied to the data in Problem P.1 leads to the overdetermined system of equations

$$\underbrace{\operatorname{col}(-2,-1,0,1,2,2,1,0,-1,-2)}_{\mathbf{a}}\, x = \underbrace{\operatorname{col}(1,4,6,4,1,-1,-4,-6,-4,-1)}_{\mathbf{b}}.$$

Therefore, using the (TLS') formulation, the problem is to minimize the function

$$f_{\mathrm{tls}}(x) = \frac{(\mathbf{a}x-\mathbf{b})^\top(\mathbf{a}x-\mathbf{b})}{x^2+1} = \cdots \begin{array}{c}\text{substituting } \mathbf{a} \text{ and } \mathbf{b} \text{ with}\\ \text{their numerical values}\end{array} \cdots = 20\frac{x^2+7}{x^2+1}.$$

The first derivative of f_{tls} is

$$\frac{\mathrm{d}}{\mathrm{d}x} f_{\mathrm{tls}}(x) = -\frac{240x}{(x^2+1)^2},$$

so that f_{tls} has a unique stationary point at $x = 0$. The second derivative of f_{tls} at $x = 0$ is negative, so that the stationary point is a maximum. This proves that the function f_{tls} has no minimum and therefore the total least squares problem has no solution.

Figure S.1 shows the plot of f_{tls} over the interval $[-6.3, 6.3]$. It can be verified that the infimum of f_{tls} is 20 and f_{tls} has asymptotes

$$f_{\mathrm{tls}}(x) \to 20 \qquad \text{for} \quad x \to \pm\infty,$$

i.e., the infimum is achieved asymptotically as x tends to infinity and to minus infinity.

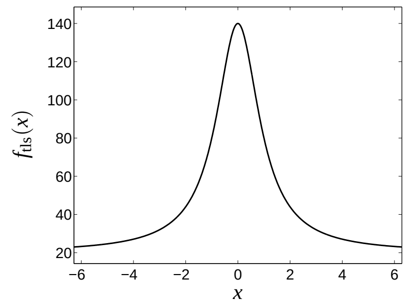


Fig. S.1 Cost function of the total least squares problem (TLS') in Problem P.6.

P.7 (Geometric interpretation of rank-1 approximation). In both problems (lra_R) and (lra_P) the cost function is

$$\|D - \widehat{D}\|_{\mathrm{F}}^2 = \sum_{j=1}^{N} \|d_j - \widehat{d}_j\|_2^2,$$

i.e., the sum of the squared distances from the data points d_j to their approximations $\widehat{d}_j$. The rank-1 constraint of

$$\widehat{D} = \begin{bmatrix} \widehat{d}_1 & \cdots & \widehat{d}_N \end{bmatrix},$$

however, is equivalent to the constraint that the approximations $\widehat{d}_j$ lie on a line $\mathscr{B}$ passing through the origin. In (lra_R), $\mathscr{B} = \ker(R)$ and, in (lra_P), $\mathscr{B} = \operatorname{image}(P)$. By the orthogonality principle, $\widehat{d}_j$ must be the orthogonal projection of d_j on $\mathscr{B}$, so that $\|d - \widehat{d}_j\|_2^2$ is the squared orthogonal distance from d_j to $\mathscr{B}$. Therefore, the rank-1 approximation problems (lra_R) and (lra_P) minimize the sum of the squared orthogonal distances from the data points to the fitting line $\mathscr{B} = \ker(P) = \text{image}(P)$ over all lines passing through the origin.

Comparing the geometric interpretations of the low rank approximation problems (lra_R) and (lra_P) and the total least squares problem (tls), we see that in both cases the same data fitting criterion is minimized, however, the minimization is over different sets of candidate solutions—in the low rank approximation problems *all* lines passing through the origin are considered, while in the total least squares problem all lines passing through the origin *except for* the vertical line are considered.

P.8 (Quadratically constrained problem, equivalent to rank-1 approximation). Consider the rank-1 approximation problem (lra_P) and observe that for a fixed parameter $P \in \mathbb{R}^{2\times 1}$, problem ($\text{lra}_P$) becomes a least squares problem in the parameter $L \in \mathbb{R}^{1\times N}$

$$\text{minimize} \quad \text{over } L \quad \|D - PL\|_{\text{F}}^2$$

Assuming that P is full column rank (*i.e.*, $P \neq 0$), the solution is unique and is given by

$$L^* = (P^\top P)^{-1} P^\top D.$$

Then the minimum $f_{\text{lra}}(P) = \|D - PL^*\|_{\text{F}}^2$ is given by

$$f_{\text{lra}}(P) = \text{trace}\Big(D^\top \big(I - P(P^\top P)^{-1} P^\top\big) D\Big).$$

The function f_{lra}, however, depends only on the direction of P, *i.e.*,

$$f_{\text{lra}}(P) = f_{\text{lra}}(\alpha P), \quad \text{for all } \alpha \neq 0.$$

Therefore, without loss of generality we can assume that $\|P\|_2 = 1$. This argument and the derivation of f_{lra} show that problem (lra'_P) is equivalent to problem (lra_P). All solutions of (lra_P) are obtained from a solution P'^* of (lra'_P) by multiplication with a nonzero scalar and vice verse a solution P^* of (lra_P) is reduced to a solution of (lra'_P) by normalization $P^*/\|P^*\|$. A solution to (lra'_P), however, is still not unique because if P'^* is a solution so is $-P'^*$.

P.9 (Line fitting by rank-1 approximation). The set of vectors $P \in \mathbb{R}^2$, such that $P^\top P = 1$, is parametrized by

$$P(\theta) = \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix},$$

where $\theta \in [0, 2\pi)$. The plot of $f_{\text{lra}}\big(P(\theta)\big)$ over θ is shown in Figure S.2. The global minimum points are

$$\theta^{*,1} = \pi/2 \qquad \text{and} \qquad \theta^{*,2} = 3\pi/2$$

(indicated with dots on the figure) and the global minimum is

$$f_{\text{lra}}\big(P(\theta^{*,1})\big) = f_{\text{lra}}\big(P(\theta^{*,2})\big) = 20.$$

The minimum points

$$\theta^{*,1} = \pi/2 \qquad \text{and} \qquad \theta^{*,2} = 3\pi/2$$

correspond to optimal parameters

$$P^{*,1} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad \text{and} \qquad P^{*,2} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \tag{$*$}$$

which in the context of the line fitting problem correspond to the vertical line passing through the origin. The link between the low rank approximation (lra_P) and total least squares (tls) problems allow us to compare their respective cost functions f_{lra} and f_{tls}. In particular, we see that f_{lra} achieves the infimum of f_{tls}.

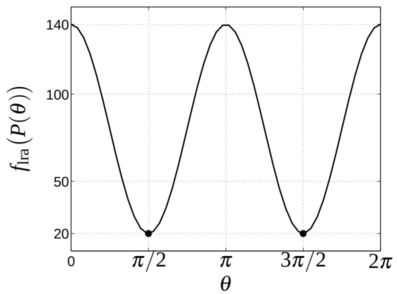


Fig. S.2 Cost function of the rank-1 approximation problem (lra'_P) in Problem P.9.

P.10 (Analytic solution of a rank-1 approximation problem).

a) MATLAB code for the alternating least squares algorithm:

250a ⟨*Alternating least squares algorithm for low rank approximation* 250a⟩≡

```
function [dh, e] = lra_als(d, p, l, tol, maxiter)
dh = p*l;
e(1) = norm(d - dh, 'fro') ^ 2;
for i = 2:maxiter
    p = d * l' / (l * l'); l = (p' * p) \ p' * d;
    dh_old = dh; dh = p * l; e(i) = norm(d - dh, 'fro') ^ 2;
    if norm(dh_old - dh, 'fro') ^ 2 < tol, break, end
end
```

Defines:
`lra_als`, used in chunk 250b.

A typical error convergence plot (for a 10×10 matrix with rank specification $r = 5$) is shown on Figure S.3.

250b ⟨*test* `lra_als` 250b⟩≡

```
q = 10; N = 10; r = 5;
```

```
d = rand(q, N); p = rand(q, r); l = rand(r, N);
[dh, e] = lra_als(d, p, l, 0.0, 25); plot(e)
print_fig('als-conv')
```

Uses `lra_als` 250a.

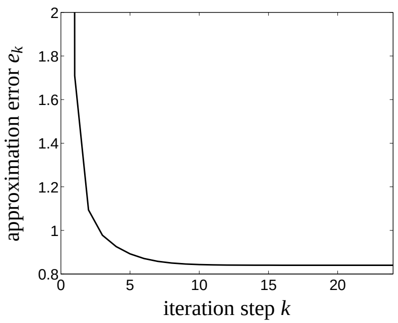


Fig. S.3 Error convergence plot for the alternating least squares algorithm in Problem P.13 (random 10×10 matrix, rank specification $r = 5$, and random initial approximation).

The convergence is monotonic. The approximation error drops significantly in the first few iteration steps and after that decreases slowly.

b) The sequence e is well defined when $P^{(k)}$ and $L^{(k)}$ are full rank for all $k = 0, 1, \ldots$, however, it may not be well defined when $P^{(k)}$ or $L^{(k)}$ become rank deficient at certain iteration step k. Indeed, rank efficiently of $P^{(k)}$ or $L^{(k)}$ implies that a solution for, respectively, $L^{(k)}$ or $P^{(k+1)}$ is not unique. Then, depending on the choice of the solution different values of e_{k+1} may be obtained.
For example, the data Problem P.1 with the initial approximation $L^{(0)} = \mathbf{1}_{10}^{\top}$, results in $P^{(1)} = 0$, which implies that $L^{(1)}$ is arbitrary. Choosing $L^{(1)} = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}$, leads to the error sequence

$$e_1 = 160,\ e_2 = 20,\ e_3 = 20,\ \ldots$$

while $L^{(1)} = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}$, leads to the error sequence

$$e_1 = 160,\ e_2 = 116,\ e_3 = 29,\ e_4 = 20.1996,\ e_5 = 20.0041,\ e_6 = 20.0001,\ \ldots$$

c) See, the proof of Theorem 5.17 on page 236.
d) See, the proof of Theorem 5.17 on page 236.

P.11 (Analytic solution of two-variate rank-1 approximation problem). A solution is given by the eigenvalue decomposition of the 2×2 matrix

$$S := DD^{\top} = \begin{bmatrix} s_1 & s_{12} \\ s_{21} & s_2 \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{N} d_{1j}^2 & \sum_{j=1}^{N} d_{1j} d_{2j} \\ \sum_{j=1}^{N} d_{1j} d_{2j} & \sum_{j=1}^{N} d_{2j}^2 \end{bmatrix}.$$

Let λ_1 and λ_2 be the eigenvalues of S. We have

$$\begin{aligned} \lambda_1 + \lambda_2 &= s_1 + s_2 \\ \lambda_1 \lambda_2 &= s_1 s_2 - s_{12}^2 \end{aligned} \quad \Longrightarrow \quad \lambda_2 = s_1 + s_2 - \lambda_1$$

Substituting the expression for λ_2 in the second equation, we have

$$\lambda_1^2 - (s_1 + s_2)\lambda_1 + (s_1 s_2 - s_{12}^2) = 0,$$

so that

$$\lambda_{1,2} = \frac{1}{2}\left(s_1 + s_2 \pm \sqrt{(s_1 - s_2)^2 + 4 s_{12}^2} \right).$$

Let $\lambda_{\min}$ be the smaller eigenvalue. (It corresponds to the minus sign.)
Next, we solve for an eigenvector v, corresponding to $\lambda_{\min}$:

$$(s - \lambda_{\min} I)v = 0$$
$$\Updownarrow$$
$$\begin{bmatrix} s_1 - s_2 + \sqrt{(s_1 - s_2)^2 + 4 s_{12}^2} & 2 s_{12} \\ 2 s_{12} & s_2 - s_1 + \sqrt{(s_1 - s_2)^2 + 4 s_{12}^2} \end{bmatrix} v = 0.$$

Provided, $s_{12} \neq 0$, *i.e.*, the rows of D are not perpendicular,

$$v = \alpha \begin{bmatrix} x \\ -1 \end{bmatrix}, \qquad \text{where} \quad x := \frac{s_2 - s_1 + \sqrt{(s_1 - s_2)^2 + 4 s_{12}^2}}{2 s_{12}}, \tag{$*$}$$

and α is an arbitrary nonzero constants.
In this case, parameters of kernel and image representations of the optimal model are

$$R = \begin{bmatrix} x & -1 \end{bmatrix}, \qquad \text{and} \qquad P = \begin{bmatrix} 1 \\ x \end{bmatrix}.$$

(We fixed $\alpha = 1$.) Finally, the optimal approximation $\widehat{D}$ of D is

$$\widehat{D} = P(P^{\top}P)^{-1}P^{\top}D = \frac{x}{1 + x^2} \begin{bmatrix} \frac{1}{x} d_{11} + d_{21} & \cdots & \frac{1}{x} d_{1N} + x d_{2N} \\ d_{11} + x d_{21} & \cdots & d_{1N} + x d_{2N} \end{bmatrix}.$$

Note that in the case $s_{12} \neq 0$, alternative formulas for the eigenvector v, corresponding to $\lambda_{\min}$ can be derived.

P.12 (Analytic solution of scalar total least squares). In the case when a is not perpendicular to b, the total least squares solution exists and is unique. In this case, it is given by $(*)$ (derived in Problem P.11). In the case when $a \perp b$, but $\|a\| > \|b\|$, the total least squares solution is $x = 0$. Otherwise, a total least squares solution does not exists.

P.13 (Alternating projections algorithm for low-rank approximation). We have

$$\begin{aligned} f_{\text{lra}}(P) &= \operatorname{trace}\left(D^\top (I - PP^\top) D\right) \\ &= \operatorname{trace}\left((I - PP^\top) DD^\top\right) \\ &= \cdots \text{ substituting the data and using } P^\top P = p_1^2 + p_2^2 = 1 \ \cdots \\ &= \operatorname{trace}\left(\begin{bmatrix} p_2^2 & -p_1 p_2 \\ -p_1 p_2 & p_1^2 \end{bmatrix} \begin{bmatrix} 20 & 0 \\ 0 & 140 \end{bmatrix} \right) \\ &= 140 p_1^2 + 20 p_2^2 = 140 \sin^2(\theta) + 20 \cos^2(\theta). \end{aligned}$$

From the analytic expression of f_{lra} it is easy to see that

$$20 \leq f_{\text{lra}}\big(P(\theta)\big) \leq 140$$

and the minimum is achieved for $(*)$, which is the result established in Problem P.9 by less rigorous methods.

P.14 (Two-sided weighted low rank approximation). Define

$$D_{\text{m}} := \sqrt{W_{\text{l}}} D \sqrt{W_{\text{r}}} \qquad \text{and} \qquad \widehat{D}_{\text{m}} := \sqrt{W_{\text{l}}} \widehat{D} \sqrt{W_{\text{r}}}.$$

Since W_{l} and W_{r} are nonsingular,

$$\operatorname{rank}(\widehat{D}) = \operatorname{rank}(\widehat{D}_{\text{m}}).$$

Then, from (WLRA2), we obtain the equivalent problem

$$\begin{aligned} &\text{minimize} && \text{over } \widehat{D}_{\text{m}} \quad \|D_{\text{m}} - \widehat{D}_{\text{m}}\|_{\text{F}} \\ &\text{subject to} && \operatorname{rank}(\widehat{D}_{\text{m}}) \leq \mathtt{m}, \end{aligned} \tag{WLRA2'}$$

which is an unweighted low rank approximation.

P.15 (Most poweful unfalsified model for autonomous models). Realization of $H : \mathbb{Z}_+ \to \mathbb{R}^{\mathtt{p} \times \mathtt{m}}$ is equivalent to exact modeling of the time series

$$w_{\text{d},1} = (u_{\text{d},1}, y_{\text{d},1}) := (\delta e_1, h_1), \quad \ldots \quad, w_{\text{d},\mathtt{m}} = (u_{\text{d},\mathtt{m}}, y_{\text{d},\mathtt{m}}) := (\delta e_{\mathtt{m}}, h_{\mathtt{m}}).$$

Consider the impulse response H of the system

$$\mathscr{B}_{\text{i/s/o}}\left(A, \begin{bmatrix} b_1 & \cdots & b_{\mathtt{m}} \end{bmatrix}, C, \bullet\right)$$

and the responses $y_1, \ldots, y_{\mathtt{m}}$ of the autonomous system $\mathscr{B}_{\text{i/s/o}}(A, C)$ due to the initial conditions $b_1, \ldots, b_{\mathtt{m}}$. It is easy to verify that

$$\sigma H = \begin{bmatrix} y_1 & \cdots & y_{\mathtt{m}} \end{bmatrix}.$$

Thus, with the obvious substitution

$$B = \begin{bmatrix} x_0^1 & \cdots & x_0^{\mathtt{m}} \end{bmatrix},$$

where $x_0^1, \ldots, x_0^{\mathtt{m}}$ are the initial conditions generating the responses $h_1, \ldots, h_{\mathtt{m}}$, realization algorithms can be used for exact identification of an autonomous system and vice verse; algorithms for identification of an autonomous systems can be used for realization.

P.16 (Algorithm for exact system identification). If the order n of the system is knowning, the problem reduces to the computation of a basis for the left kernel of the Hankel matrix $\mathscr{H}_{n+1}(w_{\text{d}})$.

254a ⟨*wn2r* 254a⟩≡

```
function R = wn2r(w, n), R = null(blkhank(w, n + 1)')';
```

With unknown order, we can proceed iteratively by trying to find an exact model for increasing orders $n = 1, 2, \ldots$, till an exact model exists. This appraoch guarantees that the computed model is the most powerful unfalsified model for the data in the model class of LTI systems.

254b ⟨*w2r* 254b⟩≡

```
function R = w2r(w)
⟨reshape w and define q, T (never defined)⟩
nmax = floor((T - ttw) / (ttw + 1));
for n = 1:nmax
  R = wn2r(w, n); if ~isempty(R), break, end
end
```

P.18 (Algorithms for approximate system identification).

P.19 (Computing approximate common divisor with `slra`).

P.20 (Matrix centering).

$$\begin{aligned} \mathbf{E}\left(\mathbf{C}(D)\right) &= \mathbf{E}\left(D - \mathbf{E}(D)\mathbf{1}_N^\top\right) \\ &= \frac{1}{N}\left(D - \frac{1}{N} D \mathbf{1}_N \mathbf{1}_N^\top\right)\mathbf{1}_N \\ &= \frac{1}{N} D \mathbf{1}_N - \frac{1}{N^2} D \mathbf{1}_N \underbrace{\mathbf{1}_N^\top \mathbf{1}_N}_{N} = 0. \end{aligned}$$

P.21 (Mean computation as an optimal modeling). The optimization problem is a linear least squares problem and its solution is

$$\widehat{c} = D\mathbf{1}_N (\mathbf{1}_N^\top \mathbf{1}_N)^{-1} = \frac{1}{N} D \mathbf{1}_N = \mathbf{E}(D).$$

P.23 ((Luenberger, 1979, Page 53)). Let $y(t)$ be the reading of the thermometer at time t and let $\bar{u}$ be the environmental temperature. From Newton's law of cooling, we have that

$$\frac{\mathrm{d}}{\mathrm{d}t} y = a\left(\bar{u}s - y\right)$$

for some unknown constant $a \in \mathbb{R}$, $a > 0$, which describes the cooling process. Integrating the differential equation, we obtain an explicit formula for y in terms of the constant a, the environmental temperature $\bar{u}$, and the initial condition $y(0)$

$$y(t) = e^{-at}y(0) + (1 - e^{-at})\overline{u}, \qquad \text{for } t \geq 0 \tag{$*$}$$

The problem is to find $\overline{u}$ from ($*$) given that

$$y(0) = 21, \qquad y(1) = 15, \qquad \text{and} \qquad y(2) = 11.$$

Substituting the data in ($*$), we obtain a nonlinear system of two equations in the unknowns $\overline{u}$ and $f := e^{-a}$

$$\begin{cases} y(1) = fy(0) + (1-f)\overline{u} \\ y(2) = f^2y(0) + (1-f^2)\overline{u} \end{cases} \tag{$**$}$$

We may stop here and declare that the solution can be computed by a method for solving numerically a general nonlinear system of equations. (Such methods and software are available, see, *e.g.*, (Dennis and Schnabel, 1987).)

System ($**$), however, can be solved without using "nonlinear" methods. Define Δy to be the temperature increment from one measurement to the next, *i.e.*,

$$\Delta y(t) := y(t) - y(t-1), \qquad \text{for all } t.$$

The increments satisfy the homogeneous differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\Delta y(t) = a\Delta y(t),$$

so that

$$\Delta y(t+1) = e^{-a}\Delta y(t) \qquad \text{for } t = 0, 1, \ldots \tag{$***$}$$

From the given data we evaluate

$$\Delta y(0) = y(1) - y(0) = 15 - 21 = -6, \qquad \Delta y(1) = y(2) - y(1) = 11 - 15 = -4.$$

Substituting in ($***$), we find the constant

$$f = e^{-a} = 2/3.$$

With f known, the problem of solving ($**$) in $\overline{u}$ is linear, and the solution is found to be $\overline{u} = 3°\mathrm{C}$.

P.24 The system (SYS DD) is

$$\begin{bmatrix} g & \Delta y(2) \\ g & \Delta y(3) \end{bmatrix} \begin{bmatrix} \overline{u} \\ \ell \end{bmatrix} = \begin{bmatrix} y(2) \\ y(3) \end{bmatrix},$$

and has the unique solution

$$\begin{bmatrix} \overline{u} \\ \ell \end{bmatrix} = \frac{1}{g\big(y(1) - 2y(2) + y(3)\big)} \begin{bmatrix} y(1)y(3) - y^2(2) \\ g\big(y(3) - y(2)\big) \end{bmatrix}.$$

It can be shown that in this case

$$e^{-a} = f = \frac{\ell}{\ell - 1} \qquad \text{and} \qquad \ell = \frac{f}{f-1}.$$

P.25 The system (SYS DD) is

$$\begin{bmatrix} g & \Delta y(2) \\ \vdots & \vdots \\ g & \Delta y(T) \end{bmatrix} \begin{bmatrix} \overline{u} \\ \ell \end{bmatrix} = \begin{bmatrix} y(2) \\ \vdots \\ y(T) \end{bmatrix},$$

and the corresponding normal equations are

$$\begin{bmatrix} (T-1)g^2 & g\sum_{t=2}^T \Delta y(t) \\ g\sum_{t=2}^T \Delta y(t) & \sum_{t=2}^T \Delta y^2(t) \end{bmatrix} \begin{bmatrix} \overline{u} \\ \ell \end{bmatrix} = \begin{bmatrix} g\sum_{t=2}^T y(t) \\ \sum_{t=2}^T \Delta y(t)y(t) \end{bmatrix}.$$

The least squares approximation of $\overline{u}$ is

$$\widehat{u} = \frac{1}{g\Big((T-1)\sum_{t=2}^T \Delta y^2(t) - \big(\sum_{t=2}^T \Delta y(t)\big)^2\Big)} \Big(\sum_{t=2}^T \Delta y^2(t) \sum_{t=2}^T y(t) - \sum_{t=2}^T \Delta y(t) \big(\sum_{t=2}^T \Delta y^2(t) \sum_{t=2}^T \Delta y(t)y(t)\big)\Big)$$

A recursive algorithm for computing $\widehat{u}$ in real time requires only the four running sums

$$\sum_{\tau=2}^t \Delta y^2(\tau), \qquad \sum_{\tau=2}^t \Delta y(\tau), \qquad \sum_{\tau=2}^t y(\tau), \quad \text{and} \quad \sum_{\tau=2}^t \Delta y(\tau)y(\tau).$$

References

Dennis J, Schnabel R (1987) Numerical Methods for Unconstrained Optimization and Nonlinear Equations. Society for Industrial Mathematics

Luenberger DG (1979) Introduction to Dynamical Systems: Theory, Models and Applications. John Wiley