Lecture 1: Classical vs behavioral paradigms for data modeling

Ivan Markovsky

ELEC doctoral school 2013

Vrije Universiteit Brussel

Outline

Course mechanics

- 4 sessions on Monday and Tuesday 13:00–17:00
- session $=$ 1h lecture $+$ 40' exercises $+$ 20' break

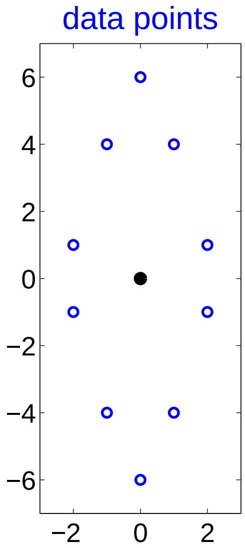
  "I hear, I forget; I see, I remember; I do, I understand."

- topics
  1. the behavioral paradigm for data modeling
  2. exact system identification
  3. approximate system identification

- book $+$ lecture slides

  homepages.vub.ac.be/∼imarkovs/doctoral-school.html

A line fitting example



- classic problem: fit points

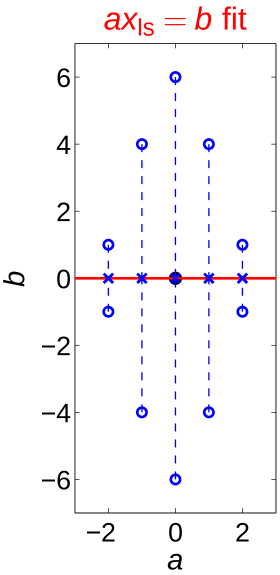
$$d_1 = \begin{bmatrix} 0 \\ 6 \end{bmatrix}, \; d_2 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}, \; \dots, \; d_{10} = \begin{bmatrix} -1 \\ 4 \end{bmatrix}$$

  by a line passing through the origin

- terminology and notation

data space	$\mathscr{U} = \mathbb{R}^2$
data	$\mathscr{D} = \{d_1, \dots, d_{10}\}$
model	$\mathscr{B} \subset \mathscr{U}$
model class	$\mathscr{M}$: lines through 0
fitting criterion	$\text{dist}(\mathscr{D}, \mathscr{B})$

A line fitting example (cont.)



- classic solution: define

$$d_i =: \begin{bmatrix} a_i \\ b_i \end{bmatrix}, \qquad \begin{aligned} A &:= \operatorname{col}(a_1, \dots, a_{10}) \\ B &:= \operatorname{col}(b_1, \dots, b_{10}) \end{aligned}$$

  and solve a least squares problem

$$Ax \approx B$$

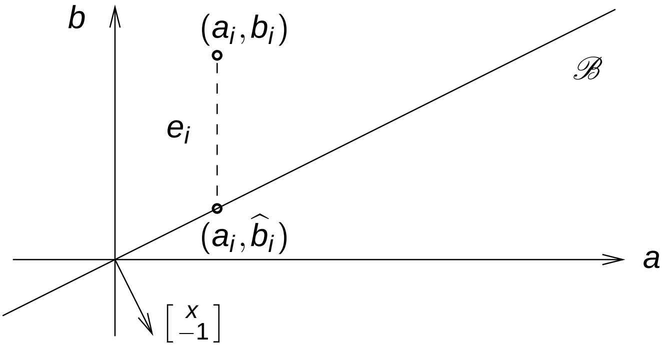
- the model is given by

$$\mathscr{B}_{\text{ls}} := \{\, (a,b) \in \mathbb{R}^2 \mid ax_{\text{ls}} = b \,\}$$

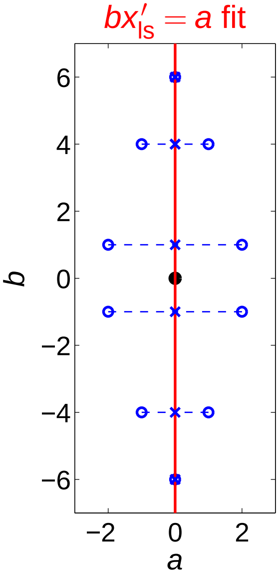
  representation of a line through 0

$\mathscr{B}_{\text{ls}}(x_{\text{ls}})$ minimizes the vertical distances from $\mathscr{D}$ to $\mathscr{B} \in \mathscr{M}$

$$\begin{aligned}\operatorname{dist}_{\text{ls}}(\mathscr{D},\mathscr{B}) &= \min_{\widehat{B}} \quad \|B-\widehat{B}\|_2 \quad \text{s.t.} \quad (a_i,\widehat{b}_i) \in \mathscr{B} \text{ for all } i \\ &= \|B - Ax_{\text{ls}}\|_2\end{aligned}$$



A line fitting example (cont.)



- we can also solve

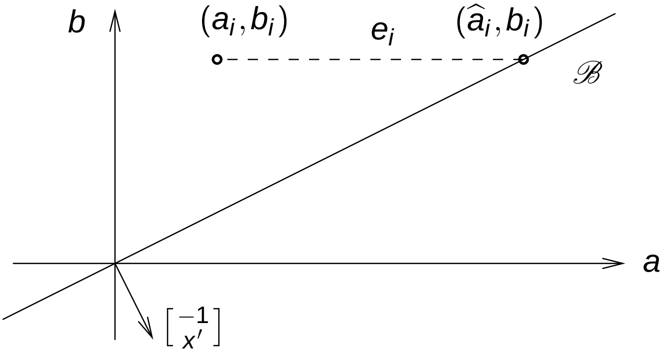
$$Bx' \approx A$$

- and obtain another fitting line

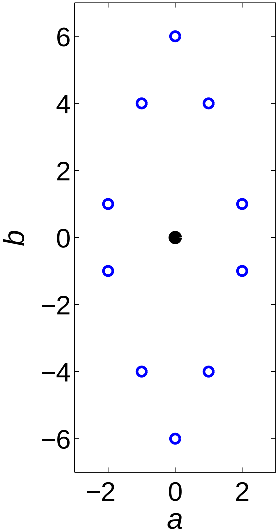
$$\mathscr{B}'_{\mathrm{ls}} := \{\, (a, b) \in \mathbb{R}^2 \mid bx'_{\mathrm{ls}} = a \,\}$$

$\mathscr{B}'_{\text{ls}}(x'_{\text{ls}})$ minimizes the horizontal distances from $\mathscr{D}$ to $\mathscr{B} \in \mathscr{M}$

$$\begin{aligned} \text{dist}'_{\text{ls}}(\mathscr{D},\mathscr{B}) &= \min_{\widehat{A}} \quad \|A-\widehat{A}\|_2 \quad \text{s.t.} \quad (\widehat{a}_i, b_i) \in \mathscr{B} \text{ for all } i \\ &= \|A - Bx'_{\text{ls}}\|_2 \end{aligned}$$



A line fitting example (cont.)



- total least squares problem:

$$\min_{x,\widehat{A},\widehat{B}} \quad \sqrt{\|A-\widehat{A}\|_2^2 + \|B-\widehat{B}\|_2^2}$$

$$\text{subject to} \quad \widehat{A}x = \widehat{B}$$

- x_{tls} does not exist ($x_{\text{tls}} = \infty$)
- however, $x'_{\text{tls}} = 0$ exists

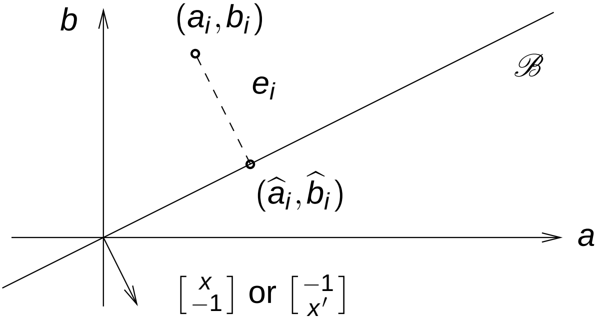
$$\min_{x',\widehat{A},\widehat{B}} \quad \sqrt{\|A-\widehat{A}\|_2^2 + \|B-\widehat{B}\|_2^2}$$

$$\text{subject to} \quad \widehat{B}x' = \widehat{A}$$

$\mathscr{B}_{\text{tls}}(x_{\text{tls}})$ minimizes orthogonal distances from $\mathscr{D}$ to $\mathscr{B} \in \mathscr{M}$

$$\text{dist}_{\text{tls}}(\mathscr{D}, \mathscr{B}) = \min_{\widehat{\mathscr{D}}} \quad \sqrt{\textstyle\sum_{i=1}^{N} \|d_i - \widehat{d}_i\|_2^2} \quad \text{s.t.} \quad \widehat{\mathscr{D}} \subset \mathscr{B}$$

$$= \frac{\|Ax_{\text{tls}} - B\|_2^2}{1 + \|x_{\text{tls}}\|_2^2}$$



Conclusions

- least squares is representation dependent
- total least squares is representation invariant
- total least squares may have no solution

What are the issues?

- a representation is nonunique
- its choice should be independent of a fitting criterion
- a representation should cover all models in $\mathscr{M}$

Inputs and outputs

- the model considered is linear static with var. (a, b)
- two different representations:

$$\{(a,b) \in \mathbb{R}^2 \mid ax = b\} \qquad (*)$$

$$\{(a,b) \in \mathbb{R}^2 \mid bx' = a\} \qquad (**)$$

- $(*)$ and $(**)$ define $\mathscr{B}$ by functions
  - $a \mapsto b$ in $(*)$
  - $b \mapsto a$ in $(**)$
- input/output representations
  - in $(*)$, a is input, b is output   (a causes b)
  - in $(**)$, b is input, a is output   (b causes a)

Input/output representation

- separately $(*)$, $(**)$ don't represent all models in $\mathscr{M}$
- I/O representation: any $\mathscr{B} \in \mathscr{M}$ is representable as

$$\mathscr{B} = \mathscr{B}(x, \Pi) = \{\, \Pi \left[\begin{smallmatrix} a \\ b \end{smallmatrix}\right] \mid ax = b \,\}$$

  for some $x \in \mathbb{R}$ and a permutation matrix Π

- link to system of linear equations

$$\mathscr{D} \subset \mathscr{B}(x, \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]) \quad \iff \quad Ax = B$$

$$\mathscr{D} \subset \mathscr{B}(x, \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]) \quad \iff \quad Bx' = A$$

  where

$$A := \operatorname{col}(a_1, \ldots, a_{10}) \quad \text{and} \quad B := \operatorname{col}(b_1, \ldots, b_{10})$$

Kernel representation

- any $\mathscr{B} \in \mathscr{M}$ can be represented as

$$\mathscr{B} = \ker(R) := \{\, d \in \mathbb{R}^2 \mid Rd = R_1 a + R_2 b = 0 \,\}$$

for some nonzero vector $R \in \mathbb{R}^{1\times 2}$

- $Rd = 0$ defines a relation between a and b

- $\mathscr{D} \subset \ker(R)$ implies that

$$R \underbrace{\begin{bmatrix} d_1 & \cdots & d_N \end{bmatrix}}_{D} = 0$$

Image representation

- any $\mathscr{B} \in \mathscr{M}$ can be represented as

$$\mathscr{B} = \operatorname{image}(P) := \{\, d = P\ell \mid \ell \in \mathbb{R} \,\}$$

for some vector $P \in \mathbb{R}^{2\times 1}$

- $\mathscr{D} \subset \operatorname{image}(P)$ implies that

$$\begin{bmatrix} d_1 & \cdots & d_N \end{bmatrix} = PL$$

for some $L \in \mathbb{R}^{1\times N}$

Important observation

- common feature of the representations considered

$$\left.\begin{array}{llll} \exists\, x \in \mathbb{R} & Ax = B & \Longrightarrow \\ \exists\, x' \in \mathbb{R} & Bx' = A & \Longrightarrow \\ \exists\, x \in \mathbb{R}, \Pi \text{ permut.} & \begin{bmatrix} x & -1 \end{bmatrix} \Pi D = 0 & \Longleftrightarrow \\ \exists\, R \in \mathbb{R}^{1\times 2}, R \neq 0 & RD = 0 & \Longleftrightarrow \\ \exists\, P \in \mathbb{R}^{2\times 1}, L \in \mathbb{R}^{1\times N} & D = PL & \Longleftrightarrow \end{array}\right\} \operatorname{rank}(D) = 1$$

- representation free characterization of the exact data

$$\mathscr{D} \subset \mathscr{B} \text{ and}$$
$$\mathscr{B} \text{ is a line through } 0$$
$$\Updownarrow$$
$$\operatorname{rank}(D) = 1$$

Low-rank approximation

- representation free formulation
- exact modeling problem:

$$\exists \text{ exact model for } \mathscr{D} \iff D \text{ is rank deficient}$$

- approximate modeling problem:

$$\begin{array}{ll} \text{minimize} & \text{over } \widehat{\mathscr{D}} \quad \operatorname{dist}(\mathscr{D}, \widehat{\mathscr{D}}) \\ \text{subject to} & \exists \text{ exact model for } \widehat{\mathscr{D}} \end{array}$$

$$\Updownarrow$$

$$\begin{array}{ll} \text{minimize} & \text{over } \widehat{D} \quad \operatorname{dist}(D, \widehat{D}) \\ \text{subject to} & \widehat{D} \text{ is rank deficient} \end{array}$$

Generalizations

1. multivariable data fitting $\mathscr{U} = \mathbb{R}^q$
   - linear static model $\leftrightarrow$ subspace
   - model complexity $\leftrightarrow$ subspace dimension
   - rank(D) $\leftrightarrow$ upper bound on the model complexity
2. nonlinear static modeling
   - $\mathscr{D} \mapsto D$ — nonlinear function
   - nonlinearly structured low-rank approximation
3. linear time-invariant dynamical models
   - $\mathscr{D} \mapsto$ Hankel matrix D
   - Hankel structured low-rank approximation
4. nonlinear dynamic (2. + 3.)

Related frameworks

- behavioral approach in systems and control theory
  - representation free: model $=$ set (the behavior)
  - no a priori separation of inputs and outputs
- errors-in-variables modeling
  - all variables are perturbed by noise
  - maximum likelihood estimation $\leftrightarrow$ LRA
- principal component analysis
  - another statistical setting for LRA
- factor analysis
  - factors $\leftrightarrow$ latent variables in image repr.

Overview of applications

- systems and control
  - approximate realization / model reduction
  - system identification
- signal processing
  - approximate deconvolution
  - image deblurring
- machine learning
  - dimensionality reduction
  - recommender systems
- computer algebra
  - approximate common divisor

Structure $\mathscr{S}$ $\leftrightarrow$ Model class $\mathscr{M}$

unstructured	linear static
Hankel	scalar LTI
$q \times 1$ Hankel	q-variate LTI
q$\times N$ Hankel	N equal length traj.
mosaic Hankel	N general traj.
$\begin{bmatrix}\text{Hankel} & \text{unstructured}\end{bmatrix}$	finite impulse response
block-Hankel Hankel-block	2D linear shift-invariant

Problems with analytic solutions

- unstructured, unweighted $\qquad (\|\cdot\|_\mathrm{F} := \|\operatorname{vec}(\cdot)\|_2)$

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

- unstructured, with some fixed rows

- unstructured, with left/right weighting matrices

$$\begin{array}{ll} \text{minimize} & \text{over } \widehat{D} \quad \|W_\mathrm{l}(D - \widehat{D})W_\mathrm{r}\|_\mathrm{F} \\ \text{subject to} & \operatorname{rank}(\widehat{D}) \leq r \end{array}$$

- circulant structure, unweighted

Theorem (Eckart–Young–Mirsky)

Let $D = U\Sigma V^\top$ be the (thin) SVD of $D \in \mathbb{R}^{q\times N}$ and define

$$U =: \begin{array}{cc} & \\ \begin{array}{cc} r & q-r \end{array} & \\ \begin{bmatrix} U_1 & U_2 \end{bmatrix} & q \end{array}, \quad \Sigma =: \begin{array}{cc} \begin{array}{cc} r & q-r \end{array} & \\ \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} & \begin{array}{c} r \\ q-r \end{array} \end{array}, \quad V =: \begin{array}{cc} \begin{array}{cc} r & q-r \end{array} & \\ \begin{bmatrix} V_1 & V_2 \end{bmatrix} & N \end{array}$$

An optimal low-rank approximation (a solution of (LRA)) is

$$\widehat{D}^* = U_1\Sigma_1 V_1^\top, \qquad \widehat{\mathscr{B}}^* = \ker(U_2^\top) = \operatorname{image}(U_1).$$

It is unique if and only if $\sigma_r \neq \sigma_{r+1}$.

- “truncated SVD”

- $\widehat{\mathscr{B}}^*$ depends only on the left singular vectors

- in general
  - structures other than circular
  - norms other than 2-norm
  - weights other than “left/right” multiplication of $D - \widehat{D}$

  lead to hard non-convex optimization problems

- there are many (heuristic) solution methods

Overview of algorithms

- global solution methods
  - SDP relaxations of rational function min. problem
  - systems of polynomial equations (computer algebra)
  - branch-and-bound, simulating annealing, . . .
- local optimization methods
  - variable projections (Lecture 3)
  - alternating projections
  - variations (parameterization $+$ optimization method)
- convex relaxations / multistage methods
  - subspace methods (Lecture 2)
  - nuclear norm heuristic

Summary

- linear static model $=$ subspace
- model representations
  - input/output (a function, system $AX = B$)
  - kernel (implicit function, relation)
  - image (introduces latent variables)
- representation invariant problem formulation $\rightsquigarrow$ LRA
- different representations $\rightsquigarrow$ different solution methods

"... most linear resistors let us treat current as a function of voltage or voltage as a function of current, since R is neither zero nor infinite. But in the two limiting cases - the short circuit and the open circuit - that's not true. To fit these cases neatly in a unified framework, we shouldn't think of the relation between current and voltage as defining a function. It's just a relation!"

John Baez

Instructions

- individual work (1–5 min.)
- discussion within groups ($\sim$ 4 people)
- Konstantin, Mariya, and Ivan are around to help
- give us a sign when you are ready
- class discussion
- there are problems for homework

Exercise 1: Constant approximation

- the approximate modeling problem:

$$\begin{array}{ll} \text{minimize} & \text{over } \widehat{D} \quad \|D-\widehat{D}\|_{\mathrm{F}} \\ \text{subject to} & \{\widehat{d}_1,\ldots,\widehat{d}_N\} \subset \text{ line through } 0 \end{array} \qquad (*)$$

  is equivalent to low-rank approximation

- the solution is given by the EYM theorem (SVD of D)

- solve $(*)$ with additional constraint

$$\widehat{d}_i = \widehat{d}, \quad \text{for some } \widehat{d} \in \mathbb{R}^q \qquad (**)$$

Exercise 2: Line fitting

- we considered the problem of fitting

$$\mathscr{D} = \{ d_1, \ldots, d_N \} \subset \mathbb{R}^2$$

to a line passing through the origin

- the solution is rank-1 approximation of the matrix

$$D = \begin{bmatrix} d_1 & \cdots & d_N \end{bmatrix}$$

- consider now the problem of fitting $\mathscr{D}$ to any line in $\mathbb{R}^2$

- is it also equivalent to low-rank approximation?

Exercise 3: Conic section fitting

- conic section model

$$\mathscr{B}(S,u,v) = \{\, d \in \mathbb{R}^2 \mid d^\top S d + u^\top d + v = 0 \,\}$$

  where $S = S^\top$, u, v are model parameters

- express the exact fitting condition $\mathscr{D} \subset \mathscr{B}(S,u,v)$ as a rank constraint on a matrix $D(\mathscr{D})$

- find a conic section fitting the points

$$d_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \quad d_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad d_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad d_4 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Exercise 4: Subspace clustering

- this problem is a special case of the Generalized PCA

- union of two lines model

$$\mathscr{B}(R^1, R^2) = \{\, d \in \mathbb{R}^2 \mid (R^1 d)(R^2 d) = 0 \,\}$$

  where $R^1, R^2 \in \mathbb{R}^{1\times 2}$, $R^1, R^2 \neq 0$ are model parameters

- express the exact fitting condition $\mathscr{D} \subset \mathscr{B}(R^1, R^2)$ as a rank constraint on a matrix $D(\mathscr{D})$

- find a union of two lines model fitting the points

$$d_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \quad d_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad d_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad d_4 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Exercise 5: LTI autonomous system fitting

- linear time-invariant autonomous system

$$\mathscr{B}|_T(R) = \{\, \big(w(1),\dots,w(T)\big) \mid \\ R_0 w(t) + R_1 w(t+1) + \dots + R_\ell w(t+\ell) = 0, \\ \text{for } t = 1,\dots,T-\ell \,\}$$

- express the exact fitting condition $w \in \mathscr{B}(R)$ as a rank constraint on a matrix constructed from w

- find the smallest ℓ, for which $\exists\ R \in \mathbb{R}^{1\times(\ell+1)}$, such that

$$w_{\rm d} := (1,2,4,7,13,24,44,81) \in \mathscr{B}_8(R)$$

Exercise 6: Polynomial common divisor

- the polynomials

$$p(z) = p_0 + p_1 z + \cdots + p_{\ell_p} z^{\ell_p}$$
$$q(z) = q_0 + q_1 z + \cdots + q_{\ell_q} z^{\ell_q}$$

  have a common divisor

$$c(z) = c_0 + c_1 z + \cdots + c_{\ell_c} z^{\ell_c}$$

  iff $p = ca$ and $q = cb$ for some polynomials a and b

- express the common divisor condition as a rank constraint on a matrix constructed from p, q