The no free lunch principle in data modeling

Ivan Markovsky

VRIJE UNIVERSITEIT BRUSSEL

Improved performance is achieved by using more data or prior knowledge

"true system" generates data

prior knowledge: properties of the true system
(model class, noise distribution, . . .)

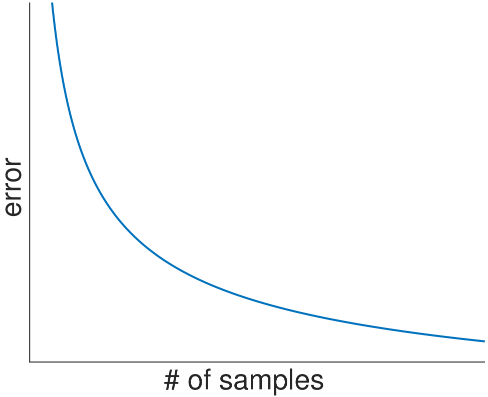
modeling: data $+$ prior knowledge $\rightsquigarrow$ model

objective:

- model $=$ true system
- use the model for filtering, control, . . .

Improved performance using more data $\rightsquigarrow$ consistent estimation

typical $1/\sqrt{\text{\# of samples}}$ estimation error decay rate



This talk is about improved performance using extra prior knowledge

System identification's view of prior knowledge

Linear algebra's view of prior knowledge

Example: ultrasound imaging

Next

System identification's view of prior knowledge

Linear algebra's view of prior knowledge

Example: ultrasound imaging

System identification aims to find "best" model in given model class

given:

- data $\mathscr{D}$
- model class $\mathscr{M}$
- distance measure $\text{dist}(\mathscr{D}, \mathscr{B})$

find: model $\widehat{\mathscr{B}}$, such that

$$\text{dist}(\mathscr{D}, \widehat{\mathscr{B}}) = \min_{\mathscr{B} \in \mathscr{M}} \text{dist}(\mathscr{D}, \mathscr{B})$$

The prior knowledge is the 1. model class, 2. distance measure

1. "true system" $\bar{\mathscr{B}}$ belongs to $\mathscr{M}$
2. dist($\mathscr{D}, \bar{\mathscr{B}}$) is "small" ($\leftrightarrow$ noise model)

Examples of prior knowledge

1. model class
   - input variables — not restricted
   - linear time-invariant (LTI), ...

2. distance measures
   - misfit (↔ measurement errors)
   - latency (↔ process noise)

The more general the model class, the weaker the prior knowledge

extreme cases

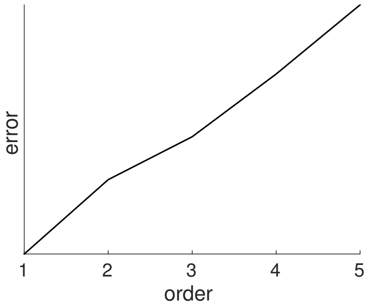
- all variables inputs $\rightsquigarrow$ trivial model (no restriction)
- all variables outputs $\rightsquigarrow$ autonomous model
- no "memory" (initial conditions) $\rightsquigarrow$ static model
- autonomous static model $\rightsquigarrow$ trivial model ($\mathscr{B} = \{ 0 \}$)

hyper parameters

- number of inputs
- number of initial conditions (order)
- model structure

The weaker the prior knowledge, the larger the estimation error

example: noise filtering



- true system $\bar{\mathscr{B}}$:

  autonomous

  LTI of order n

- measurement noise:

  $$y = \bar{y} + \widetilde{y}, \quad \bar{y} \in \bar{\mathscr{B}}$$
  $$\widetilde{y} \sim \text{Normal}(0, \sigma^2 I)$$

- estimation error:

  $$e = \|\bar{y} - \widehat{y}\|$$

Next

System identification's view of prior knowledge

Linear algebra's view of prior knowledge

Example: ultrasound imaging

Low-rank approximation: estimation with a rank constraint

given:

- data $\mathscr{D}$
- mapping $\mathscr{S} : \mathscr{D} \mapsto D \in \mathbb{R}^{m \times n}$ and $r \leq \min(m, n)$
- matrix norm $\|\cdot\|$

find: approximation $\widehat{\mathscr{D}}$ of $\mathscr{D}$ as a solution of

$$\text{minimize} \quad \text{over } \widehat{\mathscr{D}} \quad \left\| \mathscr{S}(\mathscr{D}) - \mathscr{S}(\widehat{\mathscr{D}}) \right\|$$
$$\text{subject to} \quad \operatorname{rank}\left(\mathscr{S}(\widehat{\mathscr{D}})\right) \leq r$$

The prior knowledge is the
1. rank constraint, 2. matrix norm

1. "true data" $\bar{\mathscr{D}}$ is such that rank $\big(\mathscr{S}(\bar{\mathscr{D}})\big) \leq r$

2. $\big\|\mathscr{S}(\bar{\mathscr{D}}) - \mathscr{S}(\widehat{\mathscr{D}})\big\|$ is "small" ($\leftrightarrow$ noise on $\bar{\mathscr{D}}$)

Example: Hankel matrix $\leftrightarrow$ LTI model class

$\mathscr{D} = \big(y(1), \dots, y(T)\big)$ — time series

Hankel matrix

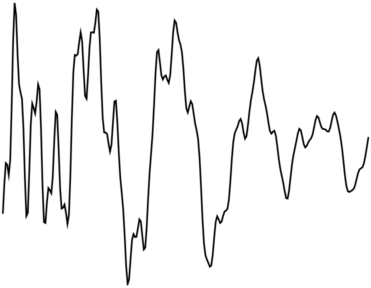
$$\mathscr{S}(\mathscr{D}) = \begin{bmatrix} y(1) & y(2) & \cdots & y(T-L+1) \\ y(2) & y(3) & \cdots & y(T-L+2) \\ y(3) & y(4) & \cdots & y(T-L+3) \\ \vdots & \vdots & & \vdots \\ y(L) & y(L+1) & \cdots & y(T) \end{bmatrix}$$

rank constraint r $\leftrightarrow$ model complexity $\leftrightarrow$ order n

Low-rank prior $\leftrightarrow$ sparsity prior

low-rank matrix $\leftrightarrow$ sparsity of the singular values

example:



- "time domain" dense
- "frequency domain" sparse (sum of 6 damped sines)
- low-rank property:

$$\operatorname{rank}\left(\mathscr{S}(y)\right) = 12$$

for $12 \leq L \leq T - 13$

Response of *n*-th order autonomous LTI system is constrained/structured/sparse

belongs to *n*-dimensional subspace

is linear combination of *n* signals

is parameterized by *n* parameters

Optimal filtering is projection on a model

problem: optimal filtering with given model

- given: 1. noisy data $y = \bar{y} + \widetilde{y}$
  2. model $\bar{\mathscr{B}}$, such that $\bar{y} \in \bar{\mathscr{B}}$ (prior knowledge)
- find: an estimate $\widehat{y}$ of $\bar{y}$

solution: project y on $\bar{\mathscr{B}}$ (ℓ_2-optimal approximation)

efficient recursive implementation for LTI systems

$\rightsquigarrow$ Kalman filter

What if the model $\bar{\mathscr{B}}$ is unknown?

use "higher-order" prior: $\bar{\mathscr{B}} \in \mathscr{M}$, with $\mathscr{M}$ given

classical definition of *n*-sparse signal

- y has n nonzero values
  (we don't know which ones)
- basis: unit vectors

n-th order autonomous LTI system's response

- y is sum of n complex exponentials
  (their frequencies and dampings are unknown)
- basis: damped complex exponentials

The low-order LTI prior makes ill-posed problems well-posed

noise filtering

- given: $y = \bar{y} + \widetilde{y}$, $\widetilde{y} \sim \text{Normal}(0, \sigma^2 I)$, and $\mathscr{M}$
- find: an estimate $\widehat{y}$ of $\bar{y} \in \bar{\mathscr{B}} \in \mathscr{M}$

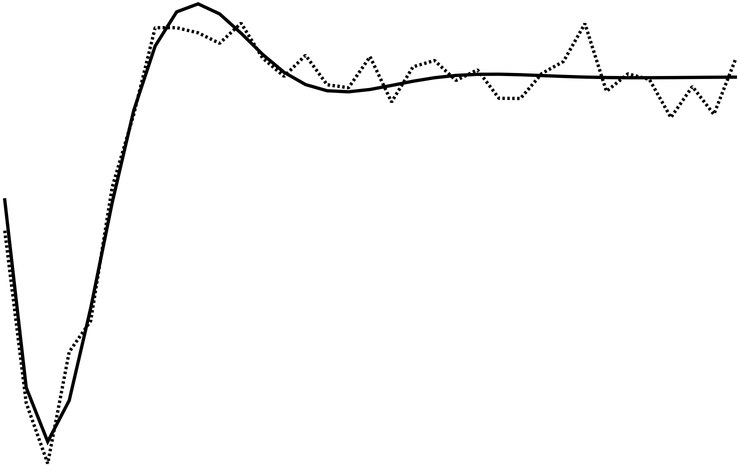
forecasting

- given: "past" samples $\big(y(-t), \ldots, y(0)\big)$ and $\mathscr{M}$
- find: "future" samples $\big(y(1), \ldots, y(t)\big)$

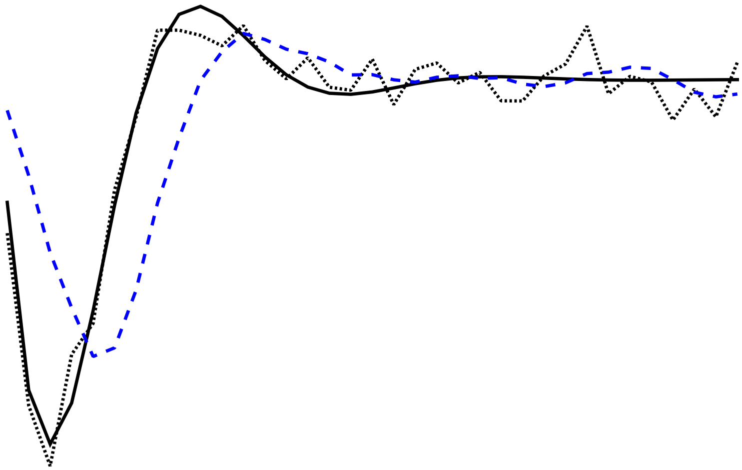
missing data estimation

- given: samples $y(t)$, $t \in \mathscr{T}_{\text{given}}$ and $\mathscr{M}$
- find: missing samples $y(t)$, $t \in \overline{\mathscr{T}_{\text{given}}}$

Noise filtering, $\bar{\mathscr{B}}$ autonomous LTI 2nd order

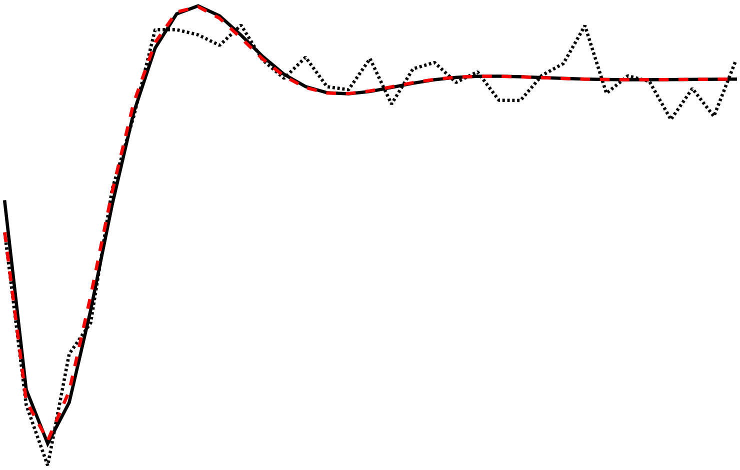


Heuristic: smooth the data by low-pass filter

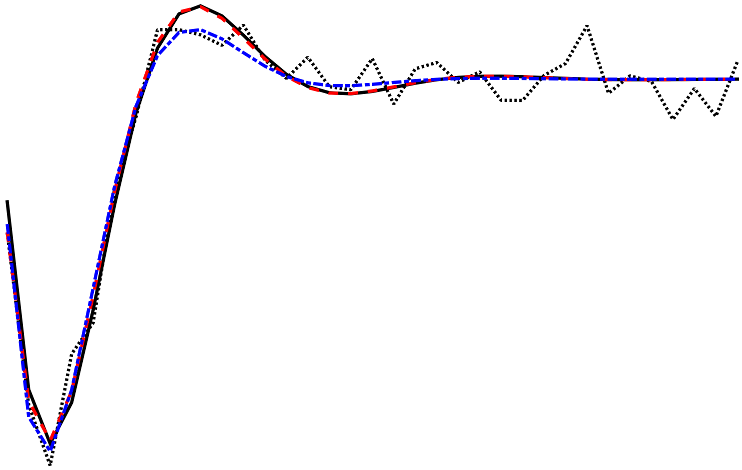


Optimal (Kalman) filtering requires a model

The best (but unrealistic) option is to use $\bar{\mathscr{B}}$



Optimal filtering using identified model $\widehat{\mathscr{B}}$, with the 2nd order LTI model class prior



Next

System identification's view of prior knowledge

Linear algebra's view of prior knowledge

Example: ultrasound imaging

High-resolution ultrasound imaging requires data compression

sensor array (64 antennas)

high sampling rate (40MHz)

generates 2.5 GB / second

Compression techniques based on skipping samples require missing data estimation

a priori known property of the data:

joint sparsity in a known basis

the signals are band limited by the sensor

the sensor's bandwidth is a priori known

Joint-sparsity $\rightsquigarrow$ low-rank

D — $T \times N$ data matrix (T samples, N channels)

$d_j = c_{j1} \exp_{\omega_1} + \cdots + c_{jr} \exp_{\omega_r}$ — reduced Fourier basis

$D = FC$, where F is $T \times r$ and C is $r \times N$, therefore

$$\text{rank}(D) \leq r$$

The low-rank property allows compression down to rN samples

$\omega_1, \dots \omega_r$ a priori known $\implies$ F is known

moreover, $\frac{1}{N}F$ is orthonormal

compression: transmit the rN coefficients

$$C = \frac{1}{N} F^T D$$

Extra prior: C is "close" to rank deficiency

quantify the distance to rank deficiency by

$$\|C\|_* = \text{sum of the singular values} \qquad \text{(nuclear norm)}$$

sampling operator $S(\cdot)$ — select $r'N < rN$ samples

extra compression using the extra prior

$$\text{minimize} \quad \text{over } C \quad \|C\|_* + \alpha \|S(X) - S(FC)\|$$

This work is in collaboration with UZ Leuven and VUB ETRO

Miaomiao Zhang (formerly UZL)

Jan D'hooge (UZL)

Colas Schretter (ETRO)

Incomplete prior by tuning hyper-parameters

order selection (rank estimation)

- Akaike information criterion
- minimum description length
- ...

Bayesian methods with parameterized prior

these methods use "hyper-prior knowledge"

hence "no free lunch"

The prior knowledge, used in data modeling, is often implicit, although it's crucial

"classical" prior:
1. model class
2. noise distribution

low-rank approximation problem

connection to sparse estimation

Outlook

other types of prior

- nonnegativity
- ...

related work

- regularization techniques
- Bayesian methods
- ...

how to come up with prior knowledge?

- parameters tuning (hyper-prior)
- ...