Errors-in-variables modelling

Ivan Markovsky

University of Southampton

Outline

Data modelling setup

Given:

- data generating system — the "true" system Σ_{true}
- from which data is observed — the data $\{ w^1, \dots, w^N \}$
- set of candidate models — the model class $\mathscr{M}$

Choose:

- a model in the model class that
- approximates "well" the true system

However, since the true system is unknown,

- the model is chosen to approximate "well" the data

User choices

- data preprocessing (centering, scaling, filtering, . . .)
- model class
- fitting criterion

often made by heuristic rules or by trail and error (experience)

User choices correspond to

- prior knowledge and/or
- assumptions

about the true system

System identification theory aims to

- justify particular fitting criteria (statistics)
- derive algorithms (optimization, numerical methods)

Classical paradigm

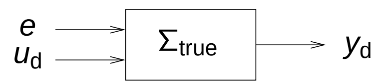
Assumptions:

- "true model in the model class" assumption
  the data generating system belongs to the model class
- input/output partitioning of the variables is a priori given
- the data-model mismatch is a stochastic process

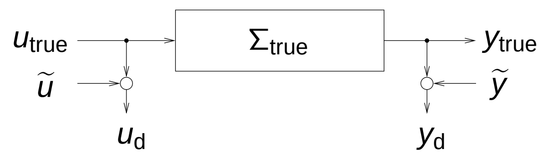
Two variations:

- treat observed inputs as exact — regression
  The uncertainty is attributed to unobserved latent inputs.
- treating all variables as noisy — errors-in-variables model
  The uncertainty is attributed to the measurement noise.

Latency model: e — process noise (unobserved)



EIV model: $\widetilde{w} = (\widetilde{u}, \widetilde{y})$ — measurement noise



Σ_{true} — data generating system, $w_d = (u_d, y_d)$ — observed data

Main results in the classical paradigm

Modelling methods that are

- consistent
- efficient and
- produce confidence bounds

under specified assumptions on the true model and the errors.

Are these assumptions reasonable in applications?

The true model in the model class assumption is often not realistic. The model-data mismatch is due to a combination of

1. wrong model class
2. process noise
3. measurement noise

In control and signal processing, 1 often dominates 2 and 3

- the model class consists of low-order LTI systems but the "true" system is high-order nonlinear time-varying
- the effect of the unobserved inputs is not strong
- the measurement devices are accurate

Model behavior

The behavior $\mathscr{B}$ of a model Σ is the set of all trajectories of Σ

$$\mathscr{B} = \{\, (u,y) \mid (u,y) \text{ is trajectory of } \Sigma \,\}$$

$\mathscr{B}$ completely specifies Σ and allow us to postpone the issue of choosing a model representation to a later stage of the analysis

$$w = (u,y) \in \mathscr{B} \quad \iff \quad y \text{ is an output of } \Sigma \text{ for an input } u$$

The inputs-outputs partition (u,y) of the trajectory w is not as essential as the classical setting implies.

The behavioral approach was introduced by Jan C. Willems

Deterministic approximation

Latency identification: given data w_{d} and model class $\mathscr{M}$

$$\text{minimize (over } \widehat{\mathscr{B}} \in \mathscr{M} \text{ and } e) \; \|e\| \text{ subject to } (e, w_{\mathrm{d}}) \in \widehat{\mathscr{B}}$$

EIV identification: given data w_{d} and model class $\mathscr{M}$

$$\text{minimize (over } \widehat{\mathscr{B}} \in \mathscr{M} \text{ and } \widehat{w}) \; \|w_{\mathrm{d}} - \widehat{w}\| \text{ subject to } \widehat{w} \in \widehat{\mathscr{B}}$$

". . . the noise model H in (3.1) is from this point of view just an alibi for determining the predictor. . . . This also means that the difference between a "stochastic system" (3.1) and a "deterministic" one (3.35) is not fundamental."

L. Ljung, System identification: Theory for the user
Second edition, 1999, Page 74

Maximum likelihood estimation

The deterministic approximation problems yield maximum likelihood estimates assuming that

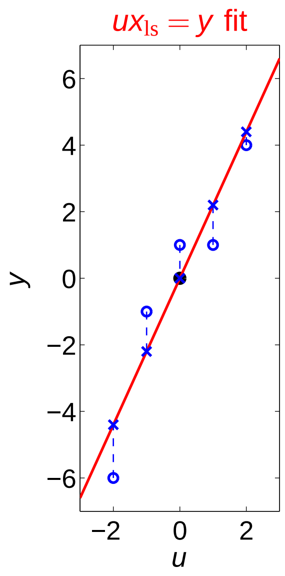
1. there is a true model $\mathscr{B}_{\text{true}}$ and
2. the model-data mismatch is a stochastic process
   - Latency model: there is e_{true}, such that $(w_{\mathrm{d}}, e_{\text{true}}) \in \mathscr{B}_{\text{true}}$
     e_{true} – realization of zero mean, white, Gaussian process
   - EIV model: there is $w_{\text{true}} \in \mathscr{B}_{\text{true}}$, such that $w_{\mathrm{d}} = w_{\text{true}} + \widetilde{w}$
     $\widetilde{w}$ – realization of zero mean, white, Gaussian process

Notes

- choosing representation of the model, makes the identification problems parameter optimization problems
- however, different representations lead to different parameter optimization problems
- latency and EIV are applications of a general principle:

  impose relevant prior knowledge by the selection of the model class and the data fitting criterion
- combined with the deterministic point of view, this principle leads to low-rank approximation problems

Static linear model: an example



Line fitting problem: Fit the points

$$w_{\text{d}}^1 = \begin{bmatrix} -2 \\ -6 \end{bmatrix}, \; w_{\text{d}}^2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \; \dots, \; w_{\text{d}}^5 = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$$

by a line passing through the origin.

Classic solution: Define $w_{\text{d}}^i =: \text{col}(u_{\text{d}}^i, y_{\text{d}}^i)$ and solve the least squares problem

$$xu_{\text{d}}^i = y_{\text{d}}^i, \quad \text{for } i = 1, \dots, 5$$

The model is the fitting line

$$\mathscr{B} := \{\, w = (u,y) \mid x_{\text{ls}} u = y \,\}$$

Static linear model

A static linear model $\mathscr{B}$ with q variables is a subspace of $\mathbb{R}^q$.

Complexity of $\mathscr{B}$ is defined to be the dimension of $\mathscr{B}$, which is equal to the number of inputs (free variables)

$\mathscr{L}_m$ — set of all static linear models with at most m inputs

Rank constraint on the data matrix

$$w_{\text{d}}^i \in \mathscr{B} \in \mathscr{L}_m, \; i = 1, \dots, N \quad \iff \quad \text{rank}\big(\begin{bmatrix} w_{\text{d}}^1 & \cdots & d_{\text{d}}^N \end{bmatrix}\big) \leq m$$

We will write $D \in \mathscr{B}$

$$D := \begin{bmatrix} w_{\text{d}}^1 & \cdots & d_{\text{d}}^N \end{bmatrix}$$

when each column of D is in $\mathscr{B}$.

Static linear EIV identification: given D and m

$$\text{minimize (over } \widehat{\mathscr{B}} \text{ and } \widehat{D}) \; \|D - \widehat{D}\| \text{ subject to } \widehat{D} \in \widehat{\mathscr{B}} \in \mathscr{L}_m$$

$$\Updownarrow$$

Low-rank approximation: given D and m

$$\text{minimize} \quad (\text{over } \widehat{D}) \quad \|D - \widehat{D}\|_{\text{F}} \quad \text{subject to} \quad \text{rank}(\widehat{D}) \leq m$$

- nonconvex, however, an analytic solution exists (SVD)
- also known as the principal component analysis

Modified low-rank approximation problems

- Weighted low-rank approximation

  $$\text{minimize} \quad \sum w_{ij}(D_{ij} - \widehat{D}_{ij})^2 \quad \text{subject to} \quad \text{rank}(\widehat{D}) \leq m$$

  Allows to treat missing data by setting weights w_{ij} to zero.

- Nonnegative low-rank approximation

  $$\begin{aligned} \text{minimize} \quad \|D - \widehat{D}\| \quad \text{subject to} \quad & \text{rank}(\widehat{D}) \leq m \\ \text{and} \quad & \widehat{D}_{ij} \geq 0, \quad \text{for all } i,j \end{aligned}$$

- Structured low-rank approximation

  $$\begin{aligned} \text{minimize} \quad \|D - \widehat{D}\| \quad \text{subject to} \quad & \text{rank}(\widehat{D}) \leq m \\ & \text{and } \widehat{D} \text{ has the same structure as } D \end{aligned}$$

  Allows to treat dynamical models $\rightsquigarrow$ using Hankel structure.

Linear time-invariant (LTI) models

Consider the time series

$$\big(w(1),\dots,w(T)\big), \qquad w(t)\in\mathbb{R}^q$$

Two special cases:

- $q=1$ variable, $m=1$ input — scalar, autonomous model
- $q=2$ var., $m=1$ input — single input, single output model

The difference equation

$$r_0w(t)+r_1w(t+1)+\cdots+w(t+n)=0, \quad r_i\in\mathbb{R}^{1\times q} \qquad \text{(DE)}$$

defines an LTI model with $m=q-1$ inputs of order at most n.

$\mathscr{L}_{m,n}$ — LTI model CLASS $\leq m$ inputs and order $\leq n$

Note that $\mathscr{L}_{m,0}=\mathscr{L}_m$ — the class of static models.

Linear prediction problem

Future values of w are estimated as linear comb. of past values

$$w(t)=c_1w(t-1)+c_2w(t-2)+\cdots+c_nw(t-n) \qquad \text{(LP)}$$

c_i are the linear prediction coefficients

Given an observed signal w_d, how do we find the coefficients c_i?

There are many methods for doing this:

- Pisarenko, Prony, Kumaresan–Tufts methods
- subspace methods
- frequency domain methods
- maximum likelihood method $\equiv$ Hankel low-rank approx.

Sum-of-damped-exponentials model

Model the signal w as

$$w(t)=\sum_{i=1}^{n}a_ie^{d_it}e^{\mathrm{i}(\omega_it+\phi_i)} \qquad \text{(SDE)}$$

where a_i, d_i, ϕ_i, and ω_i are parameters of the model

a_i	—	amplitudes	d_i	—	dampings
ω_i	—	frequencies	ϕ_i	—	initial phases

For all $\{a_i,d_i,\omega_i,\phi_i\}$ there are c_i and $w(-n+1),\dots,w(0)$, s.t. the solution of (LP) coincides with (SDE) and vice verse.

the LP problem $\iff$ modeling by (SDE)

LTI models and Hankel structured matrices

$$r_0w(t)+r_1w(t+1)+\cdots+r_nw(t+n)=0, \qquad w(t)\in\mathbb{R}^{1\times q}$$

for $t=1,\dots,T-n$, is equivalent to the system of equations

$$\begin{bmatrix} r_0 & \cdots & r_n\end{bmatrix}\underbrace{\begin{bmatrix} w(1) & w(2) & w(3) & \cdots & w(T-n)\\ w(2) & w(3) & w(4) & \cdots & w(T-n+1)\\ w(3) & w(4) & w(5) & \cdots & w(T-n+1)\\ \vdots & \vdots & \vdots & & \vdots\\ w(n+1) & w(n+2) & w(n+3) & \cdots & w(T)\end{bmatrix}}_{\mathscr{H}_{n+1}(w)}=0$$

$$\iff \quad \operatorname{rank}\big(\mathscr{H}_{n+1}(w)\big)\leq m(n+1)+n, \quad m:=q-\operatorname{rowdim}(r)$$

- the subspace methods are based on the SVD of $\mathscr{H}_{n+1}(w)$ (unstructured low-rank approximation)
- the maximum-likelihood method preserves the structure

EIV identification

$$w \in \mathscr{B} \in \mathscr{L}_{m,n} \quad \Longleftrightarrow \quad \operatorname{rank}\big(\mathscr{H}_{n+1}(w)\big) \leq m(n+1)+n$$

EIV identification: given data w_d, # of inputs m, and order n

minimize (over $\widehat{\mathscr{B}} \in \mathscr{M}$ and $\widehat{w}$) $\|w_\mathrm{d} - \widehat{w}\|$ subject to $\widehat{w} \in \widehat{\mathscr{B}}$

$$\Updownarrow$$

Hankel structured low-rank approximation: given w_d and k

minimize over $\widehat{w}$ $\|w_\mathrm{d} - \widehat{w}\|$ subject to $\operatorname{rank}\big(\mathscr{H}_{n+1}(\widehat{w})\big) \leq k$

Structured low-rank approximation

No closed form solution is known for the general SLRA problem

$$\widehat{w}^* := \arg\min_{\widehat{w}} \|w_\mathrm{d} - \widehat{w}\| \quad \text{subject to} \quad \operatorname{rank}\big(\mathscr{S}(\widehat{w})\big) \leq n$$

NP-hard, consider solution methods based on local optimization

Representing the constraint in a kernel form, the problem is

$$\min_{r,\, rr^\top = 1} \Big(\min_{\widehat{w}} \|w_\mathrm{d} - \widehat{w}\| \quad \text{subject to} \quad r\mathscr{S}(\widehat{w}) = 0 \Big)$$

Double minimization with bilinear equality constraint.

There is a matrix $G(r)$, such that $r\mathscr{S}(\widehat{w}) = 0 \iff \widehat{w}G(r) = 0$.

Variable projection vs. alternating projections

Two ways to approach the double minimization:

- Variable projections (VARPRO):
  solve the inner minimization analytically

$$\min_{r,\, rr^\top = 1} r\mathscr{S}(w_\mathrm{d})\big(G^\top(r)G(r)\big)^{-1}\mathscr{S}^\top(w_\mathrm{d})r^\top$$

  $\rightsquigarrow$ a nonlinear least squares problem for r only.

- Alternating projections (AP):
  alternate between solving two least squares problems

VARPRO is globally convergent with a super linear conv. rate.

AP is globally convergent with a linear convergence rate.

Algorithmic details using the VARPRO approach

The structured low-rank approximation problem is equivalent to

$$\min_{r,\, rr^\top = 1} r\mathscr{S}(w_\mathrm{d})\big(G^\top(r)G(r)\big)^{-1}\mathscr{S}^\top(w_\mathrm{d})r^\top$$

To evaluate the cost function we need to solve for z

$$\big(G^\top(r)G(r)\big)z = \big(r\mathscr{S}(w_\mathrm{d})\big)$$

What special structure does $G^\top G$ have?

Banded Toeplitz for any $\mathscr{S} = \begin{bmatrix} \mathscr{S}_1 & \cdots & \mathscr{S}_q \end{bmatrix}$, where $\mathscr{S}_i$ is Toeplitz, Hankel, Toeplitz+Hankel, unstructured, or fixed.

Special case: sum-of-damped-exp. modeling

In the sum-of-damped-exp. modeling, the structure is

$$\mathscr{S}(w) = \mathscr{H}_{n+1}(w)$$

What matrix G satisfies

$$r\,\mathscr{H}_{n+1}(w) = 0 \quad \iff \quad wG(r) = 0$$

for all r and w? What is the structure of $G^\top G$?

Special case: sum-of-damped-exp. modeling

$$\begin{bmatrix} r_0 & r_1 & \cdots & r_n \end{bmatrix} \underbrace{\begin{bmatrix} w(1) & w(2) & \cdots & w(T-n) \\ w(2) & w(3) & \cdots & w(T-n+1) \\ \vdots & \vdots & & \vdots \\ w(n+1) & w(n+2) & \cdots & w(T) \end{bmatrix}}_{\mathscr{H}_{n+1}(w)}$$

$$= \begin{bmatrix} w_1 & w_2 & \cdots & w_T \end{bmatrix} \underbrace{\begin{bmatrix} r_0 & & & \\ r_1 & r_0 & & \\ \vdots & r_1 & \ddots & \\ r_n & \vdots & \ddots & r_0 \\ & r_n & & r_1 \\ & & \ddots & \vdots \\ & & & r_n \end{bmatrix}}_{G(r)}$$

Special case: sum-of-damped-exp. modeling

Therefore,

$$G^\top G = \begin{bmatrix} r_0 & r_1 & \cdots & r_n & & & \\ & r_0 & r_1 & \cdots & r_n & & \\ & & \ddots & \ddots & & \ddots & \\ & & & r_0 & r_1 & \cdots & r_n \end{bmatrix} \begin{bmatrix} r_0 & & & \\ r_1 & r_0 & & \\ \vdots & r_1 & \ddots & \\ r_n & \vdots & \ddots & r_0 \\ & r_n & & r_1 \\ & & \ddots & \vdots \\ & & & r_n \end{bmatrix}$$

(All missing elements are zeros.)

Special case: sum-of-damped-exp. modeling

$$G^\top G = \begin{bmatrix} \sum_{i=0}^n r_i r_i & \sum_{i=1}^n r_i r_{i-1} & \cdots & r_n r_0 & & & \\ \sum_{i=1}^n r_{i-1} r_i & \ddots & & & \ddots & & \\ \vdots & & \ddots & & & \ddots & \\ r_0 r_n & & & \ddots & & & r_n r_0 \\ & \ddots & & & \ddots & & \vdots \\ & & \ddots & & & \ddots & \sum_{i=1}^n r_i r_{i-1} \\ & & & r_0 r_n & \cdots & \sum_{i=1}^n r_{i-1} r_i & \sum_{i=0}^n r_i r_i \end{bmatrix}$$

banded Toeplitz, bandwidth $2n+1$

Data fitting by a second order model

$$\mathscr{B}(A,b,c) := \{\, w \in \mathbb{R}^q \mid w^\top A w + b^\top w + c = 0 \,\}, \quad \text{with } A = A^\top$$

Consider first exact data:

$$w \in \mathscr{B}(A,b,c) \iff w^\top A w + b^\top w + c = 0$$

$$\iff \big\langle \underbrace{\operatorname{col}(w \otimes_s w, w, 1)}_{w_{\text{ext}}}, \underbrace{\operatorname{col}\big(\operatorname{vec}_s(A), b, c\big)}_{\theta} \big\rangle = 0$$

$$\{w_1,\dots,w_N\} \in \mathscr{B}(\theta) \iff \theta \in \operatorname{leftker} \underbrace{\begin{bmatrix} w_{\text{ext},1} & \cdots & w_{\text{ext},N} \end{bmatrix}}_{D_{\text{ext}}}, \quad \theta \neq 0$$

$$\iff \operatorname{rank}(D_{\text{ext}}) \leq q-1$$

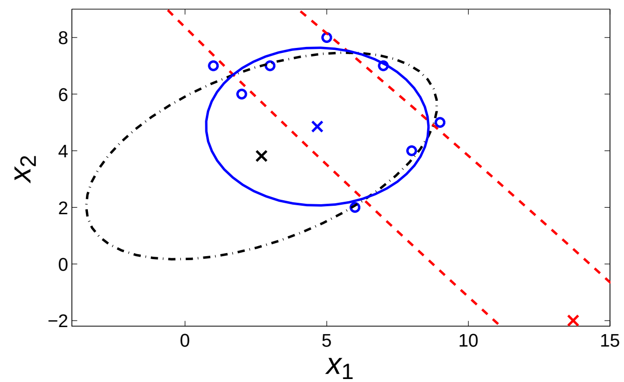
Therefore, for measured data $\rightsquigarrow$ LRA of D_{ext}.

Notes:

- Special case $\mathscr{B}$ an ellipsoid (for $A > 0$ and $4c < b^\top A^{-1} b$).
- Related to kernel PCA

Example: ellipsoid fitting

benchmark example of (Gander *et al.* 94), called “special data”



dashed — LRA solid — modified LRA

dashed-dotted — orthogonal regression (geometric fitting)

$\circ$ — data points $\times$ — centers

Conclusions

- User choices should impose prior knowledge for the true systems rather than convenient theoretical assumption.
- Stochastic estimation and deterministic approximation are two sides of the same coin.
- The behavioral setting leads to elegant and useful data modeling philosophy.
- Its algorithmic implementation is low-rank approximation.
- EIV static linear model identification — unstructured LRA.
  EIV dynamic LTI model identification — Hankel LRA.
- Algorithms based on VARPRO and alternating projections.
- Nonlinear model identification via data transformation.

Thank you