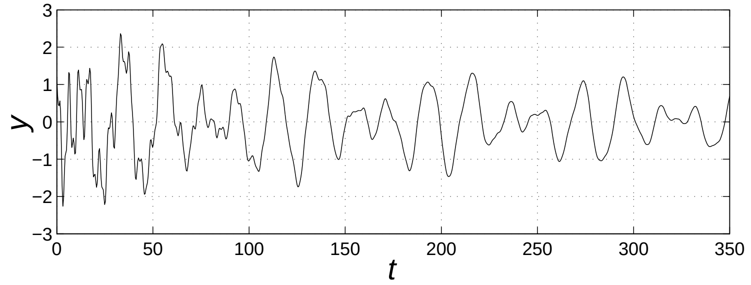
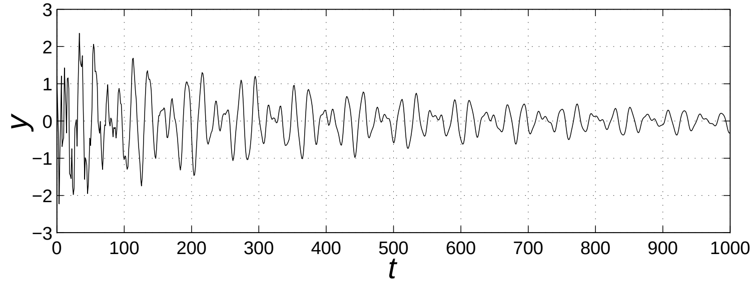
Approximate system identification: Misfit versus latency

Ivan Markovsky

University of Southampton

$\frac{d}{dt}x(t) = f\big(x(t), e(t)\big),\ x(0) = x_0,\ y(t) = g\big(x(t), e(t)\big)$





Linear or nonlinear, deterministic or stochastic?

- From simple to complex:

  linear deterministic → linear stochastic → nonlinear deterministic → nonlinear stochastic

- Exact linear system identification computationally involves solution of a linear system of equations (*i.e.*, easy).

- Maximum likelihood estimation of a linear stochastic system is a nonconvex optimization problem (*i.e.*, difficult).

  Evaluating the likelihood is least norm problem (*i.e.*, easy).

- For nonlinear stochastic systems, both the parameter optimization and the likelihood evaluating are difficult.

In this talk . . .

- linear systems
- initially deterministic
- eventually stochastic
- deterministic approximation vs stochastic estimation — two sides of the same coin

Least squares $\leftrightarrow$ Latency

Consider a linear static model $Ax \approx b$

A, b are given measurements, x is a model parameter

Least squares approximation:

$$\text{minimize}_{e,x} \|e\|_2^2 \quad \text{subject to} \quad Ax = b + e$$

Interpretation: e is unobserved latent variable

$$L\big((A,b),x\big) := \Big(\min_e \|e\|_2^2 \text{ s.t. } Ax = b + e\Big) = \|Ax - b\|_2^2$$

Least squares approximation $\iff$ latency minimization

$$\text{minimize}_x\, L\big((A,b),x\big)$$

Total least squares $\leftrightarrow$ Misfit

Total least squares:

$$\text{minimize}_{\Delta A,\Delta b,x} \big\| \begin{bmatrix} \Delta A & \Delta b \end{bmatrix} \big\|_{\mathrm{F}}^2 \text{ subject to } (A + \Delta A)x = b + \Delta b$$

Interpretation: ΔA, ΔB are data corrections

$$\begin{aligned} M\big((A,b),x\big) &:= \min_{\Delta A,\Delta b} \big\| \begin{bmatrix} \Delta A & \Delta b \end{bmatrix} \big\|_{\mathrm{F}}^2 \text{ s.t. } (A + \Delta A)x = b + \Delta b \\ &= \frac{\|Ax - b\|_2^2}{1 + \|x\|_2^2} \end{aligned}$$

Total least squares approximation $\iff$ misfit minimization

$$\text{minimize}_x\, M\big((A,b),x\big)$$

Geometric interpretation of latency

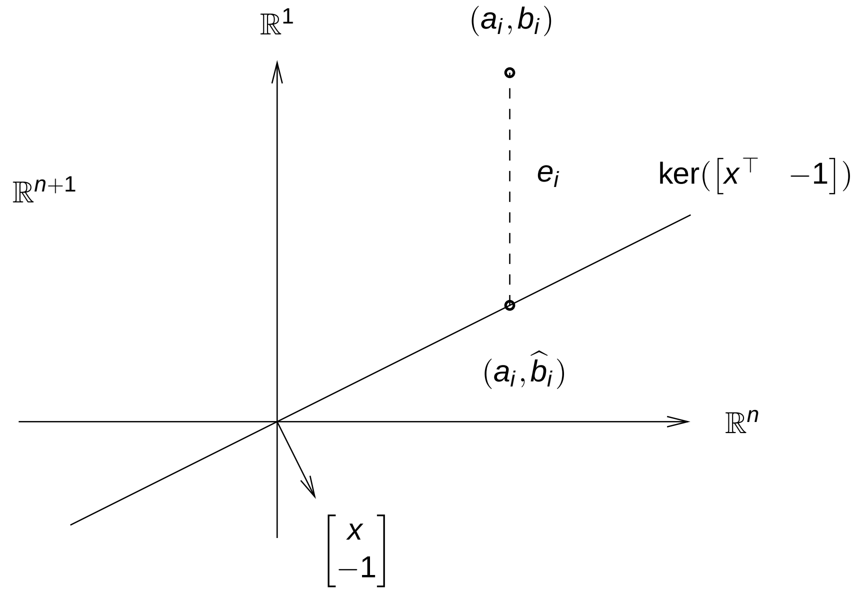
- $L\big((A,b),x\big) = \|Ax - b\|_2^2 =: \|e\|_2^2$

$$\begin{aligned} Ax = b + e =: \widehat{b} \quad &\iff \quad \begin{bmatrix} A & \widehat{b} \end{bmatrix} \begin{bmatrix} x \\ -1 \end{bmatrix} = 0 \\ &\iff \quad \begin{bmatrix} a_i & \widehat{b}_i \end{bmatrix} \begin{bmatrix} x \\ -1 \end{bmatrix} = 0, \text{ for } i = 1, \dots, m \end{aligned}$$

(a_i is the ith row of A)

- $(a_i, \widehat{b}_i)$, for all i, lie on the subspace $\perp$ to $(x, -1)$
- "data point" $(a_i, b_i) = (a_i, \widehat{b}_i) + (0, e_i)$
- The approximation error $(0, e_i)$ is the vertical distance from (a_i, b_i) to the subspace
- $L\big((A,b),x\big) = \sum_{i=1}^m e_i^2$ sum of the squared vertical distances

Geometric interpretation of latency



Geometric interpretation of misfit

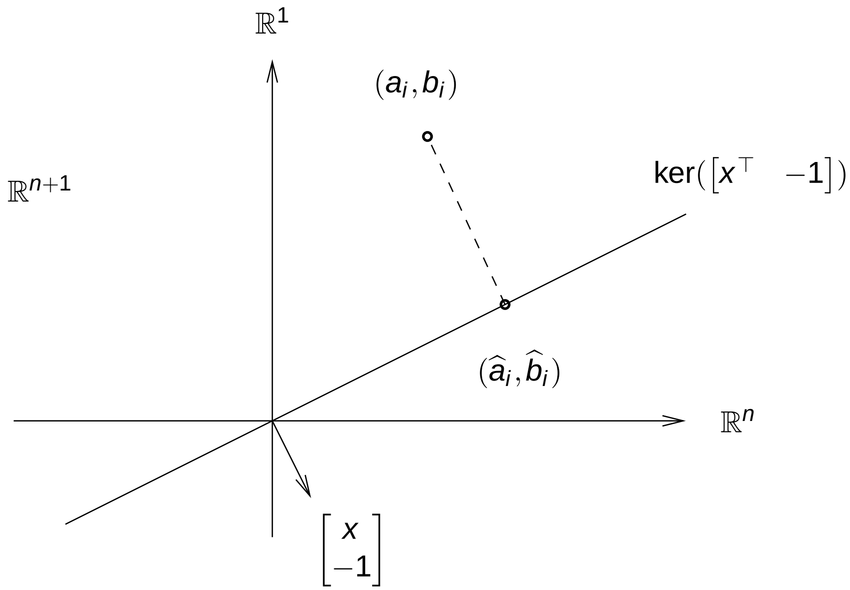
- $M\big((A,b),x\big) := \min\limits_{\Delta A,\Delta b} \left\| \begin{bmatrix} \Delta A & \Delta b \end{bmatrix} \right\|_{\mathrm{F}}^2$ s.t. $(A+\Delta A)x = b+\Delta b$

$$\underbrace{(A+\Delta A)}_{\widehat{A}}x = \underbrace{b+\Delta b}_{\widehat{b}} \iff \begin{bmatrix} \widehat{A} & \widehat{b} \end{bmatrix} \begin{bmatrix} x \\ -1 \end{bmatrix} = 0$$

$$\iff \begin{bmatrix} \widehat{a}_i & \widehat{b}_i \end{bmatrix} \begin{bmatrix} x \\ -1 \end{bmatrix} = 0, \text{ for } i = 1,\dots,m$$

- $(\widehat{a}_i, \widehat{b}_i)$, for all i, lie on the subspace $\perp$ to $(x,-1)$
- "data point" $(a_i,b_i) = (\widehat{a}_i,\widehat{b}_i) + (\Delta a_i, \Delta b_i)$
- $(\Delta a_i, \Delta b_i)$ is the orth. distance from (a_i,b_i) to the subspace
- $M\big((A,b),x\big) = \sum_{i=1}^{m} \left\| \begin{bmatrix} \Delta a_i \\ \Delta b_i \end{bmatrix} \right\|_2^2$ sum of squared orth. distances

Geometric interpretation of misfit



Notes

Latency approach — correct the model in order to match the data

Misfit approach — correct the data in order to match the model

$$\text{exact fit} \iff \text{misfit} = \text{latency} = 0$$

Both approaches reduce the approximate modelling problem to exact modelling problems.

Regression $\leftrightarrow$ Latency

Regression model:

$$Ax = b + \varepsilon, \qquad \text{where } \varepsilon \sim \mathrm{N}(0, \sigma^2 I)$$

Maximum likelihood estimator $\leftrightarrow$ latency minimization

Errors-in-variables regression $\leftrightarrow$ Misfit

Errors-in-variables (EIV) regression model:

$$(A+\delta A)x = b+\delta b, \qquad \text{where } \operatorname{vec}(\begin{bmatrix}\delta A & \delta b\end{bmatrix}) \sim \mathrm{N}(0,\sigma^2 I)$$

Maximum likelihood estimator $\leftrightarrow$ misfit minimization

System identification: $w_\mathrm{d} \mapsto \widehat{\mathscr{B}} \in \mathscr{M}$

Notation

- $w_\mathrm{d} = (u_\mathrm{d}, y_\mathrm{d})$ — given data (*e.g.*, a vector time series)
- $\widehat{\mathscr{B}}$ — to be found model for w_d (*e.g.*, an LTI system)
- $\mathscr{M}$ — model class (*e.g.*, bounded complexity LTI systems)

System identification

- defines a mapping $w_\mathrm{d} \mapsto \mathscr{B}$
- derives effective algorithms that realize the mapping, and
- develops efficient software that implements the algorithms

Stochastic estimation vs deterministic approximation

Deterministic point of view

- w_d can be generated by a nonlinear time-varying system
- The issue is how to best approximate w_d by $\widehat{\mathscr{B}} \in \mathscr{M}$

Stochastic point of view

- the data w_d is generated by an EIV or ARMAX model $\overline{\mathscr{B}}$
- The issue is how to best estimate $\overline{\mathscr{B}} \in \mathscr{M}$

An identification method can be given deterministic as well as stochastic interpretation.

Misfit vs latency

Two approaches to describe the model–data mismatch:

- Latency: augment $\mathscr{B}$ with latent variable e

$$L(w_\mathrm{d}, \mathscr{B}_\mathrm{ext}) := \min_e \|e\|^2 \quad \text{subject to} \quad (e, w_\mathrm{d}) \in \mathscr{B}_\mathrm{ext}$$

- Misfit: project w_d on $\mathscr{B}$

$$M(w_\mathrm{d}, \mathscr{B}) := \min_{\widehat{w}} \|w_\mathrm{d} - \widehat{w}\|^2 \quad \text{subject to} \quad \widehat{w} \in \mathscr{B}$$

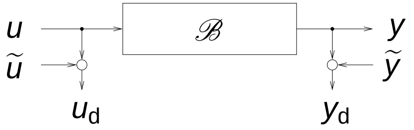
Computing misfit and latency are smoothing problems.

There are efficient algorithms in the state space (Kalman filter) and polynomial (Cholesky factorization of Toeplitz matrix) settings.

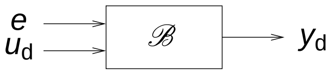
Statistical interpretation of misfit and latency

misfit	$\leftrightarrow$	errors-in-variables (EIV) model
latency	$\leftrightarrow$	ARMAX model

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



ARMAX model: e — process noise



Assumptions: $\widetilde{w}$, e — zero mean, stationary, white, ergodic, Gaussian, processes, $e \perp u$

Identification problems

Latency minimization (PEM): given $w_d \in (\mathbb{R}^w)^T$ and $n \in \mathbb{N}$, find

$$\widehat{\mathscr{B}}^*_{\text{ext}} := \arg\min_{\widehat{\mathscr{B}}_{\text{ext}}, e} \|e\|^2 \text{ s.t. } (e, \widehat{w}) \in \widehat{\mathscr{B}}_{\text{ext}} \text{ and order}(\widehat{\mathscr{B}}) \leq n$$

Misfit minimization (GTLS): given $w_d \in (\mathbb{R}^w)^T$ and $n \in \mathbb{N}$, find

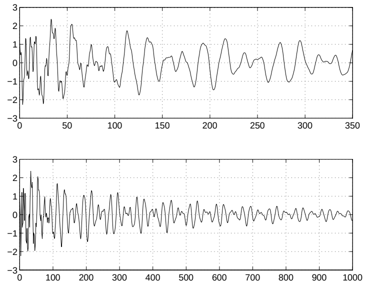
$$\widehat{\mathscr{B}}^* := \arg\min_{\widehat{\mathscr{B}}, \widehat{w}} \|w_d - \widehat{w}\|^2 \text{ s.t. } \widehat{w} \in \widehat{\mathscr{B}} \text{ and order}(\widehat{\mathscr{B}}) \leq n$$

Notes:

- nonconvex optimization problems
- solution methods based on local optimization methods
- initial approximation obtained from subspace methods

Conclusions

static		$\leftrightarrow$	dynamic		$\leftrightarrow$	concept
LS	regression	$\leftrightarrow$	PEM	ARMAX	$\leftrightarrow$	latency
TLS	EIV regression	$\leftrightarrow$	GTLS	EIV model	$\leftrightarrow$	misfit

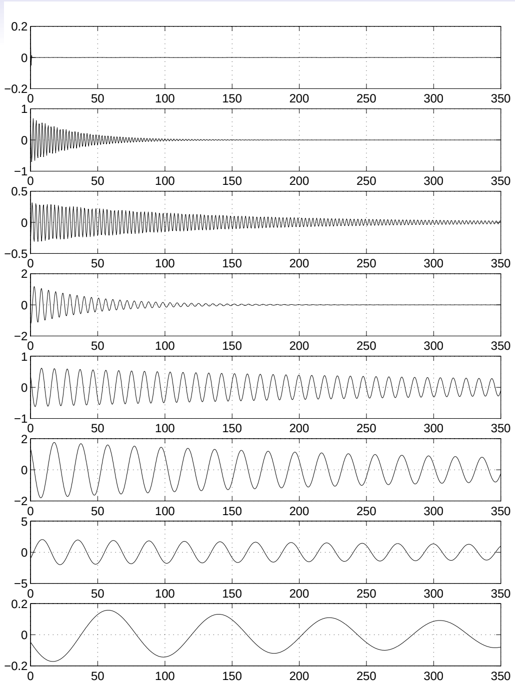


trajectory generated by a

linear deterministic system

$\frac{d}{dt}x(t) = Ax(t)$, $y(t) = Cx(t)$

of order (dim. of x) $= 16$


0.2
0
−0.2
0 50 100 150 200 250 300 350
1
0
−1
0.5
0
−0.5
2
0
−2
1
0
−1
2
0
−2
5
0
−5
0.2
0
−0.2

Thank you