Approximate system identification: Misfit versus latency

Ivan Markovsky

University of Southampton
\[
\frac{dx(t)}{dt} = f(x(t), e(t)), \quad x(0) = x_0, \quad y(t) = g(x(t), e(t))
\]
Linear or nonlinear, deterministic or stochastic?

- From simple to complex:

  \[
  \text{linear deterministic} \rightarrow \text{linear stochastic} \rightarrow \text{nonlinear deterministic} \rightarrow \text{nonlinear stochastic}
  \]

- 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:

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

Interpretation: $e$ is an unobserved latent variable

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

Least squares approximation $\iff$ latency minimization

$$\minimize_{x} L((A, b), x)$$
Total least squares \(\leftrightarrow\) Misfit

Total least squares:

\[
\begin{align*}
\min_{\Delta A, \Delta b, x} & \| [\Delta A \quad \Delta b] \|_F^2 \\
\text{subject to} & \quad (A + \Delta A)x = b + \Delta b
\end{align*}
\]

Interpretation: \(\Delta A, \Delta B\) are data corrections

\[
M((A, b), x) := \min_{\Delta A, \Delta b} \| [\Delta A \quad \Delta b] \|_F^2 \quad \text{s.t.} \quad (A + \Delta A)x = b + \Delta b
\]

\[
= \frac{\|Ax - b\|_2^2}{1 + \|x\|_2^2}
\]

Total least squares approximation \(\iff\) misfit minimization

\[
\begin{align*}
\minimize_x M((A, b), x)
\end{align*}
\]
Geometric interpretation of latency

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

\[ Ax = b + e =: \hat{b} \iff \begin{bmatrix} A & \hat{b} \end{bmatrix} \begin{bmatrix} x \\ -1 \end{bmatrix} = 0 \]

\[ \iff \begin{bmatrix} a_i & \hat{b}_i \end{bmatrix} \begin{bmatrix} x \\ -1 \end{bmatrix} = 0, \text{ for } i = 1, \ldots, m \]

\( (a_i \text{ is the } i\text{th row of } A) \)

- \((a_i, \hat{b}_i), \text{ for all } i, \text{ lie on the subspace } \perp \text{ to } (x, -1)\)

- “data point” \((a_i, b_i) = (a_i, \hat{b}_i) + (0, e_i)\)

- The approximation error \((0, e_i)\) is the vertical distance from \((a_i, b_i)\) to the subspace

- \( L((A, b), x) = \sum_{i=1}^{m} e_i^2 \) sum of the squared vertical distances
Geometric interpretation of latency

\[ \mathbb{R}^{n+1} \rightarrow \mathbb{R}^n \]

\[ (a_i, b_i) \quad e_i \quad \ker([x^\top \ -1]) \]

\[ \left[ \begin{array}{c} x \\ -1 \end{array} \right] \]
Geometric interpretation of misfit

- \( M((A, b), x) := \min_{\Delta A, \Delta b} \| [\Delta A \Delta b] \|_F^2 \) s.t. \((A + \Delta A)x = b + \Delta b\)

\[
\begin{align*}
(A + \Delta A)x = b + \Delta b & \iff \begin{bmatrix} \hat{A} & \hat{b} \end{bmatrix} \begin{bmatrix} x \\ -1 \end{bmatrix} = 0 \\
\iff \begin{bmatrix} \hat{a}_i & \hat{b}_i \end{bmatrix} \begin{bmatrix} x \\ -1 \end{bmatrix} = 0, \text{ for } i = 1, \ldots, m
\end{align*}
\]

- \((\hat{a}_i, \hat{b}_i)\), for all \(i\), lie on the subspace \(\perp\) to \((x, -1)\)

- “data point” \((a_i, b_i) = (\hat{a}_i, \hat{b}_i) + (\Delta a_i, \Delta b_i)\)

- \((\Delta a_i, \Delta b_i)\) is the orth. distance from \((a_i, b_i)\) to the subspace

- \(M((A, b), x) = \sum_{i=1}^{m} \| [\Delta a_i \Delta b_i] \|_2^2\) sum of squared orth. distances
Geometric interpretation of misfit

\[ \mathbb{R}^{n+1} \]  

\[ \mathbb{R}^1 \]  

\[ \ker(\begin{bmatrix} x^T & -1 \end{bmatrix}) \]  

\[ (a_i, b_i) \]  

\[ (\hat{a}_i, \hat{b}_i) \]  

\[ \begin{bmatrix} x \\ -1 \end{bmatrix} \]
Latency approach — correct the model in order to match the data

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

effect fit $\iff$ misfit $=$ latency $=$ 0

Both approaches reduce the approximate modelling problem to exact modelling problems.
Regression model:

\[ Ax = b + \varepsilon, \quad \text{where } \varepsilon \sim N(0, \sigma^2 I) \]

Maximum likelihood estimator \(\leftrightarrow\) latency minimization
Errors-in-variables (EIV) regression model:

\[(A + \delta A)x = b + \delta b, \quad \text{where} \quad \text{vec}([\delta A \quad \delta b]) \sim N(0, \sigma^2 I)\]

Maximum likelihood estimator \(\leftrightarrow\) misfit minimization
System identification: $w_d \mapsto \hat{B} \in \mathcal{M}$

Notation

- $w_d = (u_d, y_d)$ — given data (e.g., a vector time series)
- $\hat{B}$ — to be found model for $w_d$ (e.g., an LTI system)
- $\mathcal{M}$ — model class (e.g., bounded complexity LTI systems)

System identification

- defines a mapping $w_d \mapsto B$
- 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 $\hat{B} \in M$

**Stochastic point of view**

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

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 $\mathcal{B}$ with latent variable $e$

  $$L(w_d, \mathcal{B}_{\text{ext}}) := \min_{e} \|e\|^2 \quad \text{subject to} \quad (e, w_d) \in \mathcal{B}_{\text{ext}}$$

- **Misfit**: project $w_d$ on $\mathcal{B}$

  $$M(w_d, \mathcal{B}) := \min_{\hat{w}} \|w_d - \hat{w}\|^2 \quad \text{subject to} \quad \hat{w} \in \mathcal{B}$$

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: $\tilde{w} = (\tilde{u}, \tilde{y})$ — measurement errors

ARMAX model: $e$ — process noise

Assumptions: $\tilde{w}, e$ — 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

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

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

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

Notes:

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

<table>
<thead>
<tr>
<th>static</th>
<th>⇔</th>
<th>dynamic</th>
<th>⇔</th>
<th>concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>LS</td>
<td>regression</td>
<td>⇔</td>
<td>PEM</td>
<td>ARMAX</td>
</tr>
<tr>
<td>TLS</td>
<td>EIV regression</td>
<td>⇔</td>
<td>GTLS</td>
<td>EIV model</td>
</tr>
</tbody>
</table>
trajectory generated by a linear deterministic system

\[ \frac{d}{dt} x(t) = Ax(t), \quad y(t) = Cx(t) \]

of order (dim. of \( x \)) = 16
Thank you