Lecture 3: Approximate identification

Ivan Markovsky

ELEC doctoral school 2013

Vrije Universiteit Brussel

Outline

MPUM: complexity minimization

$$\begin{array}{ccc} \text{data} & & \text{model} \\ \mathscr{D} \subset \mathscr{U} & \xrightarrow{\text{exact identification}} & \mathscr{B}_{\text{mpum}}(\mathscr{D}) \end{array}$$

- $\mathscr{B}_{\text{mpum}}(\mathscr{D})$ most powerful unfalsified model of $\mathscr{D}$ in $\mathscr{L}$
- "most powerful" $\rightsquigarrow$ min. of complexity $c(\mathscr{B}) := (\mathtt{m}, \ell)$

  minimize over $\widehat{\mathscr{B}}$ $\quad c(\widehat{\mathscr{B}})$ $\quad$ subject to $\quad \mathscr{D} \subset \widehat{\mathscr{B}} \in \mathscr{M}$

  among all exact models, choose the least complicated
- user choice: $\mathscr{M}$ (LTI), no hyper parameters

- $\text{dist}(\mathscr{D}, \widehat{\mathscr{B}})$ — distance measure b/w $\mathscr{D}$ and $\mathscr{B}$
- $\mathscr{B}_{\text{mpum}}(w_{\text{d}})$ is a solution of

$$\text{minimize} \quad \text{over } \widehat{\mathscr{B}} \in \mathscr{M} \quad c(\widehat{\mathscr{B}}) \quad \text{s.t.} \quad \text{dist}(\mathscr{D}, \widehat{\mathscr{B}}) = 0$$

- the requirement that $\widehat{\mathscr{B}}$ is unfalsified is too restrictive

Approximate identification

$$\text{minimize} \quad \text{over } \widehat{\mathscr{B}} \in \mathscr{M} \quad \begin{bmatrix} c(\widehat{\mathscr{B}}) \\ \text{dist}(\mathscr{D}, \widehat{\mathscr{B}}) \end{bmatrix}$$

- biobjective optimization: complexity–accuracy trade-off
- user choices: $\mathscr{M}$ (LTI) and dist, no hyper parameters
- solution: set of Pareto optimal models
- selection of single model, requires a hyper parameter
  - upper bound e on the approximation error
  - upper bound r on the model complexity (LTI of bounded complexity)
  - trade-off parameter λ

Three possible scalarizations

- complexity minimization with error constraint

$$\min_{\widehat{\mathscr{B}} \in \mathscr{M}} \quad c(\widehat{\mathscr{B}}) \quad \text{subject to} \quad \text{dist}(\mathscr{D}, \widehat{\mathscr{B}}) \leq e$$

- error minimization with complexity constraint

$$\min_{\widehat{\mathscr{B}} \in \mathscr{M}} \quad \text{dist}(\mathscr{D}, \widehat{\mathscr{B}}) \quad \text{subject to} \quad c(\widehat{\mathscr{B}}) \leq r$$

- weighted sum of error and complexity minimization

$$\min_{\widehat{\mathscr{B}} \in \mathscr{M}} \quad \text{dist}(\mathscr{D}, \widehat{\mathscr{B}}) + \lambda c(\widehat{\mathscr{B}})$$

Approximation is needed when

1. the data generating system $\overline{\mathscr{B}}$ is not in $\mathscr{M}$
2. there are unobserved variables (disturbances)
3. the data is noisy due to measurement errors (ME)

Comments

- the importance of 1, 2, 3 depends on the application
- in many cases, 1 and 2 are the dominant sources
- as shown next, 1 and 3 are not essentially different

Deterministic vs stochastic approaches

- the error due to $\overline{\mathscr{B}} \notin \mathscr{M}$ is deterministic
- disturbances and measurement errors are often well modeled as stochastic processes
- stochastic estimation $\quad\leftrightarrow\quad$ deterministic approx.

$$\begin{array}{c} \text{noise} \\ \text{model} \end{array} \xleftrightarrow{\quad\text{maximum likelihood}\quad} \begin{array}{c} \text{cost} \\ \text{function} \end{array}$$

- also in control: LQG control $\leftrightarrow$ H_2 optimal control

- Ljung, page 74

  The noise model . . . is just an alibi for determining the predictor. . . . This also means that the difference between a "stochastic system" and a "deterministic" one is not fundamental.

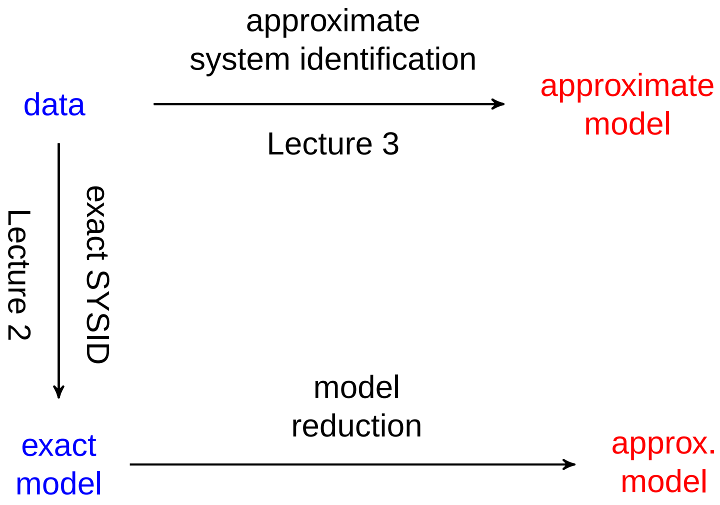
- Söderström and Stoica, pages 197, 198

  It should be stressed that it is a model assumption only that $e(t)$ is white noise. We can compute and apply the predictor even if this model assumption is not satisfied by the data. Thus the model assumption should be regarded as a tool to construct the predictor.

System identification as data compression

- the model is a concise representation of the data
- exact model $\leftrightarrow$ lossless compression (*e.g.*, `zip`)
- approximate model $\leftrightarrow$ lossy compression (*e.g.*, `mp3`)

Model reduction view of system identification



$\mathrm{dist}(\mathscr{D}, \mathscr{B})$ — misfit vs latency

uncert. source	deterministic	stochastic
1. $\bar{\mathscr{B}} \notin \mathscr{M}$ or 3. ME	misfit	EIV modeling
2. disturbances	latency	ARMAX model

Example in the static case

- misfit $\quad\leftrightarrow\quad$ total least squares

$$\min \left\| \begin{bmatrix} A - \widehat{A} & B - \widehat{B} \end{bmatrix} \right\|_{\mathrm{F}} \quad \text{s.t.} \quad \widehat{A}x = \widehat{B}$$

- latency $\quad\leftrightarrow\quad$ least squares

$$\min \|E\|_2 \quad \text{s.t.} \quad \begin{bmatrix} E & A \end{bmatrix} \begin{bmatrix} -1 \\ x \end{bmatrix} = B$$

Misfit

consider the case $\mathscr{D} = w_\text{d}$ (single trajectory)

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

orthogonal projection of w_d on $\mathscr{B}$

Misfit identification:

modify w_d as little as possible to obtain $\widehat{w}$, so that

$$\mathscr{B}_\text{mpum}(\widehat{w}) \in \mathscr{L}_{\mathtt{m},\ell}$$

then, the approximate model for w_d is

$$\widehat{\mathscr{B}}_\text{misfit} := \mathscr{B}_\text{mpum}(\widehat{w})$$

Latency

- augmented model class

$$(e, w) \in \mathscr{B}_{\text{ext}} \in \mathscr{L}_{\mathtt{m}+\mathtt{p},\ell}$$

e is a latent (unobserved) input $\qquad$ ($\leftrightarrow$ disturbance)

- given w_{d} and $\mathscr{B}_{\text{ext}} \in \mathscr{L}_{\mathtt{m}+\mathtt{p},\ell}$

$$\text{latency}(w_{\text{d}}, \mathscr{B}_{\text{ext}}) := \min_{e} \|e\| \quad \text{s.t.} \quad (e, w_{\text{d}}) \in \mathscr{B}_{\text{ext}}$$

- with, Π_w projector of (e, w) on w

$$\mathscr{B} = \Pi_w \mathscr{B}_{\text{ext}} \qquad \text{is the model for } w$$

- $\mathscr{C} \subset \mathscr{L}_{\mathtt{m+p},\ell}$ — models with bounded $e \mapsto y$ gain

Latency identification

augment w_d by, as small as possible e, so that

$$\mathscr{B}_\mathrm{mpum}\big((e, w_\mathrm{d})\big) \in (\mathscr{L}_{\mathtt{m+p},\ell} \cap \mathscr{C})$$

the approximate model for w_d is

$$\widehat{\mathscr{B}}_\mathrm{latency} := \Pi_w \mathscr{B}_\mathrm{mpum}\big((e, w_\mathrm{d})\big)$$

($\Pi_e \mathscr{B}_\mathrm{ext}$ is the disturbance model)

Computation of the misfit

$$\text{misfit}(w_\text{d}, \mathscr{B}) := \min_{\widehat{w} \in \mathscr{B}} \|w_\text{d} - \widehat{w}\|$$

- general purpose solvers $\implies O(T^3)$ flops
- time-invariance of $\mathscr{B}$, implies Toeplitz structure
- $\ell < T$, implies banded structure
- structure exploiting misfit computation methods
  - structured matrix computations
  - Riccati recursions (Kalman smoother)
- have complexity $O(T)$

Approximate identification problems

General problem formulation

$$\text{minimize} \quad \text{over } \widehat{\mathscr{B}} \quad \text{dist}(\mathscr{D}, \widehat{\mathscr{B}}) \quad \text{subject to} \quad \widehat{\mathscr{B}} \in \mathscr{M}$$

Special cases

- Misfit:
$$\begin{aligned} &\text{minimize} && \text{over } \widehat{\mathscr{B}}, \widehat{w} \quad \|w_\text{d} - \widehat{w}\| \\ &\text{subject to} && \widehat{w} \in \widehat{\mathscr{B}} \in \mathscr{L}_{\mathtt{m},\ell} \end{aligned}$$

- Latency:
$$\begin{aligned} &\text{minimize} && \text{over } \widehat{\mathscr{B}}_\text{ext}, e \quad \|e\| \\ &\text{subject to} && (e, w_\text{d}) \in \widehat{\mathscr{B}}_\text{ext} \in (\mathscr{L}_{\mathtt{m}+\mathtt{p},\ell} \cap \mathscr{C}) \end{aligned}$$

Comments

- misfit and latency reduce approx. to exact SYSID:
  - $\widehat{\mathscr{B}}_{\text{misfit}}$ is exact for modified data $\widehat{w}$
  - $\widehat{\mathscr{B}}_{\text{latency}}$ is exact in extended model class $\mathscr{L}_{\mathtt{m}+\mathtt{p},\ell}$
- misfit approach: modifies w_{d}, does not change $\mathscr{M}$
- latency approach: modifies $\mathscr{M}$, does not change w_{d}

Maximum likelihood estimation (EIV setup)

- data generating model

$$w_{\rm d} = \overline{w} + \widetilde{w}, \quad \text{where} \quad \overline{w} \in \overline{\mathscr{B}}_{\rm ext} \in \mathscr{M} \quad \text{and} \quad \widetilde{w} \sim \mathrm{N}(0, s^2 I)$$

- log-likelihood function

$$L(\widehat{\mathscr{B}}, \widehat{w}) = \begin{cases} \text{const} - \frac{1}{2s^2} \|w_{\rm d} - \widehat{w}\|_2^2 & \text{if } \widehat{w} \in \widehat{\mathscr{B}} \\ -\infty & \text{otherwise} \end{cases}$$

- likelihood evaluation $\iff$ misfit computation
- $\widehat{\mathscr{B}}$ — maximum likelihood estimator of $\overline{\mathscr{B}}$
- $\widehat{\mathscr{B}}$ — consistent estimator of $\overline{\mathscr{B}}$

Maximum likelihood estimation (ARMAX)

- data generating model

$$(e, w_\mathrm{d}) \in \overline{\mathscr{B}} \in \mathscr{M}, \quad \text{where} \quad e \sim \mathrm{N}(0, s^2 I)$$

- log-likelihood function

$$L(\widehat{\mathscr{B}}_\mathrm{ext}, e) = \begin{cases} \text{const} - \frac{1}{2s^2}\|e\|_2^2 & \text{if } (e, w_\mathrm{d}) \in \widehat{\mathscr{B}}_\mathrm{ext} \\ -\infty & \text{otherwise} \end{cases}$$

- likelihood evaluation $\iff$ latency computation
- $\widehat{\mathscr{B}}$ — maximum likelihood estimator of $\overline{\mathscr{B}}$
- $\widehat{\mathscr{B}}$ — consistent estimator of $\overline{\mathscr{B}}$

Comments

- double minimization problems
- inner minimization is Kalman filtering/smoothing
- outer minimization is a nonconvex problem
- solution methods are based on local optimization
- initial approx. is obtained from heuristic methods

Generalizations

- multiple time-series $\mathscr{D} = \{w^1, \dots, w^N\}$

$$M(\mathscr{D}, \mathscr{B}) := \min_{\{\widehat{w}^1,\dots,\widehat{w}^N\} \subset \mathscr{B}} \sqrt{\sum_{i=1}^N \|w^i - \widehat{w}^i\|_2^2}$$

- fixed initial conditions w_{ini}

$$M(w, \mathscr{B}) := \min_{w_{\text{ini}} \wedge \widehat{w} \in \mathscr{B}} \|w - \widehat{w}\|_2$$

- fixed variables $\mathscr{I} \subset \{1, \dots, q\}$

$$M(w, \mathscr{B}) := \min_{\widehat{w} \in \mathscr{B},\ \widehat{w}_{\mathscr{I}} = w_{\mathscr{I}}} \|w - \widehat{w}\|_2$$

- missing data: $w_j^i(t) = \texttt{NaN} \implies w_j^i(t)$ is missing

Rank deficient Hankel matrices

$$\mathscr{H}_L(w) := \begin{bmatrix} w(1) & w(2) & w(3) & \cdots & w(T-L+1) \\ w(2) & w(3) & w(4) & \cdots & w(T-L+2) \\ w(3) & w(4) & w(5) & \cdots & w(T-L+3) \\ \vdots & \vdots & \vdots & & \vdots \\ w(L) & w(L+1) & w(L+2) & \cdots & w(T) \end{bmatrix}$$

- single time series

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

- multiple time-series $\quad \rightsquigarrow \quad$ mosaic-Hankel matrix
- complexity minimization $\quad \leftrightarrow \quad$ rank minimization

Variable projection

- using kernel representation

$$\operatorname{rank}\big(\mathscr{H}_{\ell+1}(w)\big) \leq r \quad \iff \quad R\mathscr{H}_{\ell+1}(w) = 0$$

where $R \in \mathbb{R}^{\mathrm{p}\times q(\ell+1)}$ is full row rank (f.r.r.)

- the approximate identification problem

$$\text{minimize} \quad \text{over } \widehat{\mathscr{B}} \in \mathscr{M} \quad \operatorname{dist}(\mathscr{D}, \widehat{\mathscr{B}})$$

becomes

$$\begin{array}{ll} \text{minimize} & \text{over } \widehat{w} \text{ and f.r.r. } R \quad \|w_{\mathrm{d}} - \widehat{w}\| \\ \text{subject to} & R\mathscr{H}_{\ell+1}(\widehat{w}) = 0 \end{array}$$

- with $\|\cdot\| = \|\cdot\|_2$, the minimization over $\widehat{w}$

$$f(R) := \min_{\widehat{w}} \ \|w - \widehat{w}\| \quad \text{subject to} \quad R\mathscr{H}_{\ell+1}(\widehat{w}) = 0$$

  is a least-norm problem with analytic solution

$$M(R) = \operatorname{vec}^\top(w)\Gamma^{-1}(R)\operatorname{vec}(w)$$

  where Γ is a positive definite banded Toeplitz matrix

- the identification problem is then

$$\text{minimize} \quad \text{over } R \quad M(R) \quad \text{subject to} \quad R \text{ is f.r.r.}$$

- nonconvex optimization problem on a manifold

SLRA software package

- efficient evaluation of $M(R)$ exploiting the structure
- different strategies for enforcing "R to be f.r.r."
  - $RR^\top = I_\mathrm{p} \quad \rightsquigarrow \quad$ quadratic constraint
  - $R\Pi = \begin{bmatrix} X & I_\mathrm{p} \end{bmatrix}$, Π is a permutation, X is a free var.
- different local optimization methods
  - Gauss-Newton
  - Levenberg-Marquardt
  - trust region methods
- software implementation

Software

- mosaic-Hankel low-rank approximation

  homepages.vub.ac.be/∼imarkovs/slra/software.html

- `[sysh,info,wh] = ident(w, m, ell, opt)`
  - `sysh` — I/S/O representation of the identified model
  - `opt.sys0` — I/S/O repr. of initial approximation
  - `opt.wini` — initial conditions
  - `opt.exct` — exact variables
  - `info.Rh` — parameter R of kernel repr.
  - `info.M` — misfit

- `[M, wh, xini] = misfit(w, sysh, opt)`

Summary

- exact SYSID — complexity minimization
- approx. SYSID — complexity–accuracy trade-off

uncert. source	deterministic	stochastic
1. $\bar{\mathscr{B}} \notin \mathscr{M}$ or 3. ME	misfit	EIV modeling
2. disturbances	latency	ARMAX model

- double minimization $\rightsquigarrow$ variable projection
- approximate SYSID $\leftrightarrow$ mosaic-Hankel LRA

Exercise 1: Misfit computation

- given data w_d and an LTI system $\mathscr{B}$, represented by
  - image $\big(P(\sigma)\big)$
  - $\mathscr{B}(A,B,C,D)$
- explain how to compute misfit$(w_\mathrm{d},\mathscr{B})$ in 2-norm
- *i.e.*, find the orthogonal projection of w_d on $\mathscr{B}$
- HW: misfit computation using ker $\big(R(\sigma)\big)$

Exercise 2: Latency computation

- given data w_d and an LTI system $\mathscr{B} = \ker\big(R(\sigma)\big)$
- explain how to compute latency$(w_\mathrm{d}, \mathscr{B})$ in 2-norm
- HW: latency computation using $\mathscr{B}(A, B, C, D)$