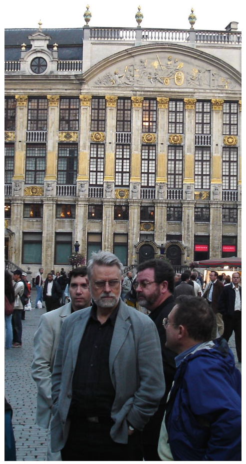
Hommage



Celebrating Anders's many great contributions to system science, including several on Estimation and System Identification.

Avoiding Local Maxima in ML Estimation

Joint work with Chengpu Yu and Michel Verhaegen.
(Automatica, to appear 2017)

Lennart Ljung

Reglerteknik, ISY, Linköpings Universitet

Outline

- The ML (or PEM) estimate of linear state-space models
- Major problem: Non-convex optimization $\Rightarrow$ Local minima (maxima)
- Essential for Gray-Box estimation
- A formulation via an initial subspace model estimate
- A DCP algorithm based on the initial model
- Numerical experiments

Identifying Linear State-Space Models

Model: $\dot{x} = A(\theta)x + B(\theta)u + w \qquad y = C(\theta)x + e$
Given input u and output y, find θ!
Maximum Likelihood (or Prediction Error) Method: Let $\hat{x}(t|\theta)$ be the Kalman filter estimate of $x(t)$ based on the model and the observations up to time $t-$. Then

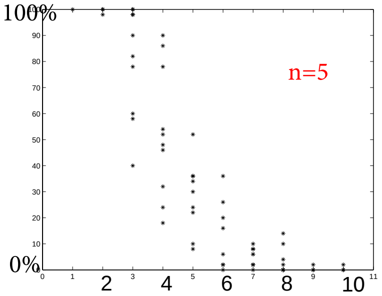
$$\hat{\theta} = \arg\min_{\theta} \sum_k \|y(t_k) - C(\theta)\hat{x}(t_k|\theta)\|^2$$

This estimate has many nice properties. (...) Any problems?

Minimize by iterative Gauss-Newton search. Main problem is that the sum is non-convex. May end up in non-global minima.
Is that a pressing problem?

Are local minima common?

[Parrilo and Ljung, 2003]
Pick an n-order system. Fix all parameters except k. Estimate these k parameters by minimizing the PEM criterion by stating in 100 different random values. Count how many time the minimization ends up in the true values. Each of the 10 stars is a specific system.
x-axis: k y-axis: success rate.



Lesson Learned

Very slim chances to reach the global minimum from random initial estimates for problems of realistic sizes!

Domains of attraction of the global minimum not very forgiving.

We need good initial estimates.

Well developed techniques for black-box type problems ...

Gray-box identification

Assume that that model is paramerized with parameters of physical importance/interest.

$$\dot{x} = A(\theta)x + B(\theta)u + w \qquad y = C(\theta)x + e$$

Not so restrictive to assume that the paramerization is affine:

$$A(\theta) = A_0 + \sum_{i=1}^{l} A_i\theta_i, \quad B(\theta) = B_0 + \sum_{i=1}^{l} B_i\theta_i,$$

$$C(\theta) = C_0 + \sum_{i=1}^{l} C_i\theta_i,$$

Example: DC-Servomotor model (time constant τ and gain β:

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ 0 & -1/\tau \end{bmatrix} x + \begin{bmatrix} 0 \\ \beta/\tau \end{bmatrix} u$$

is affine in $\theta_1 = -1/\tau$ and $\theta_2 = \beta/\tau$.

Start with a Total Model!

While estimating individual parameters in a state-space model, may be cumbersome, it is easy to estimate the full linear model by, e.g. subspace methods such at N4SID or MOESP. (For continuous time models do `d2c`.) So we can end up in a model A^*, B^*, C^* which is equivalent to the system with the true parameters $A(\theta_0), B(\theta_0), C(\theta_0)$. [But the state-space realization will be unknown.]
So, solve

$$C^*(sI - A^*)^{-1}B^* = C(\theta)(sI - A(\theta))^{-1}B(\theta) \text{ for } \theta!$$

Also a non-convex problem!

Alternative Formulation – Via Similarity Transformations

- Via similarity transformations (Xie and Ljung, 2002): Solve for θ and Q

  $$QA^* = A(\theta)Q, \quad QB^* = B(\theta), \quad C^* = C(\theta)Q \quad \text{or}$$

  $$\min_{\theta,Q} \|QA^* - A(\theta)Q\|_F^2 + \|QB^* - B(\theta)\|_F^2 + \|C^* - C(\theta)Q\|_F^2$$

  A bilinear minimization problem!
- Many local minima! Difficult to find the global one!
- Is it a fundamental problem to find the identifiable parameters in grey box?
- No: Criterion can be rewritten as a Sum-Of-Squares Problem (Parrilo-Ljung, 2003)
- No: "Any identifiable Model Structure can be rearranged as a linear regression" (Ljung-Glad, 1994)

Alternative Formulation – Via Hankel Matrix

Solve $C^*(sI - A^*)^{-1}B^* = C(\theta)(sI - A(\theta))^{-1}B(\theta)$ for θ!

- Via the Hankel Matrix of Impulse responses. (Yu et al)
  Let $Y = H_n^*$ be the Hankel Matrix of impulse response from A^*, B^*, C^* $\quad C^*(A^*)^k B^*$ (known matrix)

  We then have

  $$Y = H_n(\theta) = \begin{bmatrix} \cdots & C(\theta)A^k(\theta)B(\theta) & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}$$

Solve that equation for θ! Quite complex expressions...

More on the Hankel-based Equation

We have to solve

$$\min_{\theta} \|Y - H_n(\theta)\|_F^2$$

We shall approach this non-linear problem with a lifting technique: Introduce many more free variables, so that the minimization problem becomes quadratic and "take back" the freedom by introducing many linear constraints and a rank constraint.

Think: $H_n = OC$ (observability and controllability matrices)
$\tilde{O} = OA(\theta), \qquad \tilde{C} = A(\theta)C \qquad$ (shifted matrices)

New Formulation

$$\min_{\theta,X,C,O,\tilde{O},\tilde{C},\tilde{A}} \|Y - X\|_F^2 \qquad (''X = H_n(\theta) = OC'')$$

$s.t. \quad \text{rank } Z = n \quad$ (model order)

$$Z = \begin{bmatrix} X & O & \tilde{O} \\ C & I_n & A(\theta) \\ \tilde{C} & A(\theta) & \tilde{A} \end{bmatrix} = n$$

$$O(1:p,:) = C(\theta)$$
$$\tilde{O}(1:(n-1)p,:) = O(p+1:np,:)$$
$$C(:,1:m) = B(\theta)$$
$$\tilde{C}(:,1:(n-1)m = C(:,m+1:nm)$$

Note: Z linear in all the minimization variables

A DCP Algorithm

Let $f_n(Z)$ denote the sum of the n largest singular values of Z. Then

rank $Z = n \iff \|Z\|_* - f_n(Z) = 0$ $\quad \|\cdot\|_*$ is the nuclear norm

$\|Z\|_*$ and $f_n(Z)$ are convex in Z, so (the constrained problem with Lagrange multiplier λ)

$$\min_{\theta,X,C,O,\bar{O},\bar{C},\bar{A}} \|Y - X\|_F^2 + \lambda(\|Z\|_* - f_n(Z))$$

is a difference of convex programming (DCP) problem, which can be solved by a sequential convex relaxation method (involving linearization of the concave term $-f_n(Z)$).

Numerical Results

Repeat the experiments from slide 3: Collect data from a randomly generated 5th order system. Select k elements as unknown in an otherwise correct system model. Estimate these k elements with

1. Minimization of the PEM (ML) criterion from randomly chosen initial values
2. Initial values from the DCP method just described.

Over 100 experiments, how often are the parameters estimated correctly?

k	ML	DCP
6	22%	88%
10	2%	20%

Conclusions

- Important problem to find initial parameter estimates for ML optimization not to end up in non-global extremal points
- Important subproblem: Do this for affine parameterizations!
- Then an initial model, estimated by subspace techniques may be very useful. Do model matching between this estimate and the structured model.
- We have shown one method based on Hankel matrix matching and DCP.
- Clear improvements in avoiding local extrema!