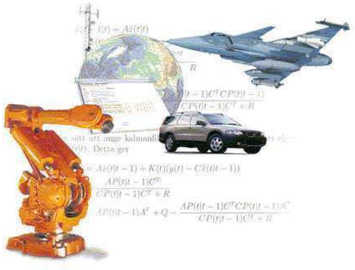
Parameter Estimation in a Moving Horizon Perspective

State and Parameter Estimation in Dynamical Systems



Lennart Ljung

Reglerteknik, ISY, Linköpings Universitet

OUTLINE

1. Problem Formulation
2. State Estimation with Sparse Process Disturbances in Linear Systems
3. Parameter and State Estimation in Unknown (Linear) Systems

State and Parameter Estimation in Dynamical Systems

Model

$$x(t+1) = f(x(t), u(t), w(t), \theta)$$
$$y(t) = h(x(t), u(t), \theta) + e(t)$$

Problem

Measure $y(t)$ and $u(t)$, $t = 1, \ldots, N$. Find $x(t)$ and θ

Linear Case

$$f(x(t), u(t), w(t), \theta) = A(\theta)x(t) + B(\theta)u(t) + w(t)$$
$$h(x(t), u(t), \theta) = C(\theta)x(t)$$

Known System Case

θ is a known vector

Maximum Likelihood

View $\Theta = [\theta, x(t), t = 1, \ldots, N]$ as unknown parameters. Assume $e(t) \in N(0, I)$. Then the negative log-likelihood function is

$$V(\theta, x(\cdot)) = \sum_{t=1}^{N} \|y(t) - h(x(t), u(t), \theta)\|^2$$

Too many parameters! $\Rightarrow$ Regularize!

Change of Parameterization

Do a (nonlinear) change of parameters:
View $\tilde{\Theta} = [\theta, x(1), w(1), \ldots, w(N-1)] = [\theta, w(\cdot)]$ as the new set of parameters,
$[x(k) = f(x(k-1), u(k-1), w(k-1), \theta) = x(k, \tilde{\Theta})]$
[The ML-estimate is unaffected by change of parameters!]
The negative log-likelihood function for $\tilde{\Theta}$ is

$$V(\tilde{\Theta}) = \sum_{t=1}^{N} \|y(t) - h(x(t, \tilde{\Theta}), u(t), \theta)\|^2$$

This to be minimized wrt $\tilde{\Theta} = [\theta, w(\cdot)]$.

Regularization

With regularization:

$$W(\tilde{\Theta}) = \sum_{t=1}^{N} \|y(t) - h(x(t,\tilde{\Theta}), u(t), \theta)\|^2 + \lambda R(\tilde{\Theta})$$

Choices of regularization:

$$R(\tilde{\Theta}) = \sum_{t=1}^{N} \|w(t)\|^2 \qquad \text{[Tichonov]}$$

or

$$R(\tilde{\Theta}) = \sum_{t=1}^{N} \|w(t)\| \qquad \text{[sum-of-norms]}$$

Classical Interpretation

Regularization curbs the flexibility of (large) model sets by pulling the parameters toward the origin.

- Tichonov: Regularization for Bias-Variance Trade-off
- Sum-of-norms: Regularization for Sparsity:
  Solutions with "many" $\|w(t)\| = 0$ are favored

Bayesian Interpretation

Suppose $w(t) \in N(0, I)$ and θ is a random vector with $\theta \in N(0, cI)$ (dim = d). Then the joint pdf of $\theta, y(\cdot), w(\cdot)$ is

$$-2 \log P(y(\cdot), w(\cdot), \theta) \sim \sum_{t=1}^{N} [\|y(t) - h(x(t, w(\cdot)), u(t), \theta)\|^2 + \|w(t)\|^2] + \|\theta\|^2/c + const$$

$$x(t, w(\cdot)) = f(x(t-1, w(\cdot)), u(t-1), w(t-1), \theta)$$

so the MAP estimate of $\tilde{\Theta}$ is

$$\hat{\tilde{\Theta}} = \arg\min W(\tilde{\Theta}) + \|\theta\|^2/c$$

which for $c \to \infty$ is the same as the Tichonov-regularized ML estimate of $\tilde{\Theta}$.

Outline

2

State Estimation with Sparse Process Disturbances in Linear Systems

3

Parameter and State Estimation in Unknown (Linear) Systems

Linear System with Sparse Process Disturbances

$$x(t+1) = Ax(t) + Bu(t) + w(t)$$
$$y(t) = Cx(t) + e(t).$$

Here, e is white measurement noise and w is a process disturbance. In many applications, w is mostly zero, and strikes, $w(t*) \neq 0$, only occasionally. Examples of applications:

- Control: Load disturbance
- Tracking: Sudden maneuvers
- FDI: Additive system faults
- Recursive Identification (x=parameters): model segmentation

AUTOMATIC CONTROL
REGLERTEKNIK
LINKÖPINGS UNIVERSITET

Approaches

- Find the jump times t ($w(t) \neq 0$) and/or the smoothed state estimates $\hat{x}_s(t|N),\ t = 1, \ldots, N.$

Common methods:

- Say $w(t^*) \neq 0$. View t^* and $w(t^*)$ as unknown parameters and estimate them. (Willsky-Jones GLR)
- Set the process noise variance to a small number and use Kalman Smoothing to estimate x (and $w(t)$)
- Branch the KF at each time instant: jump/no jump. Prune/merge trajectories (IMM).
- It is a non-linear filtering problem (linear but not Gaussian noise), so try particle filtering

All methods require some design variables that reflect the trade-off between measurement noise sensitivity and jump alertness.

More on Willsky-Jones GLR

For one jump at time t^*, estimate t^* and $w(t^*)$ as parameters.

$$x(t+1) = Ax(t) + Bu(t) + w(t); \quad y(t) = Cx(t) + e(t).$$

- If t^* is known it is a simple LS problem to estimate $w(t^*)$. $x(t)$ is a linear function of $w(t^*)$:

$$\min_{w(t^*)} \sum \|y(t) - Cx(t)\|^2$$

- Using the variance of the estimate, the significance of the jump size can be decided in a χ^2 test.
- The time of the most significant jump is the t^* that minimizes

$$\min_{t^*} \min_{w(t^*)} \sum \|y(t) - Cx(t)\|^2$$

Willsky-Jones as a Constrained Optimization Problem

- Can be written as

$$\min_{w(k),k=1,\ldots,N-1} \sum_{t=1}^{N} \|y(t) - Cx(t)\|^2$$

$$\text{s.t. } \|W\|_0 = 1;\ W = \left[\|w(1)\|_2, \ldots, \|w(N-1)\|_2\right]$$

$$\text{such that} \quad x(t+1) = Ax(t) + Bu(t) + w(t);\ x(1) = 0.$$

- k jumps: ...
- Adjustable number of jumps:

$$\min_{w(k),k=1,\ldots,N-1} \sum_{t=1}^{N} \|y(t) - Cx(t)\|^2 + \lambda \|W\|_0$$

Do the ℓ_1 Trick! ($\ell_0 \to \ell_1$)

This problem is computationally forbidding, so relax the ℓ_0 norm:

$$\min_{w(k),k=1,\dots,N-1} \sum_{t=1}^{N} \left\|y(t) - Cx(t)\right\|^2 + \lambda \|W\|_1$$

$$= \min_{w(k),k=1,\dots,N-1} \sum_{t=1}^{N} \left\|y(t) - Cx(t)\right\|^2 + \lambda \sum_{t=1}^{N} \|w(t)\|_2$$

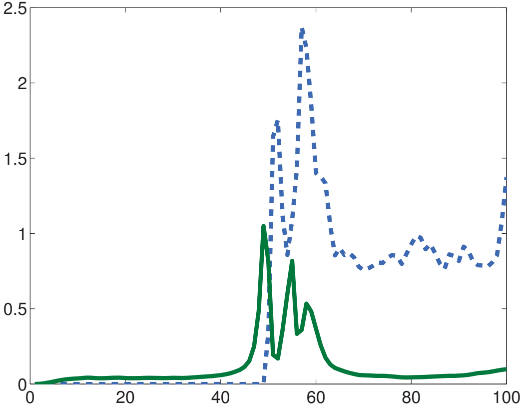
[StateSON] This is our Moving Horizon State estimation problem with SON-regularization.

Choice of λ : ....

Ohlsson, Gustafsson, Ljung, Boyd: Smoothed state estimates under abrupt changes using sum-of-norms regularization. *Automatica* 48(4):595-605, April 2012

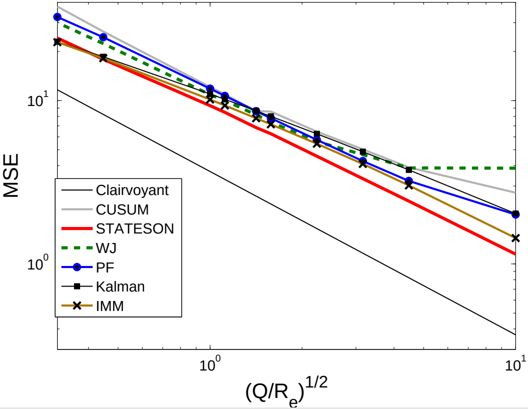
How Does it Work?

DC motor with impulse disturbances at $t = 49, 55$
State RMSE over 500 realizations as a function of time $t : 0 \to 100$.
Dashed blue: Willsky-Jones, Solid green: StateSON



Varying SNRs

Same system. Jump probability $\mu = 0.015$. Varying SNR: Q = jump size, R_e = measurement noise variance. For each SNR, RMSE averages over 500 MC runs. Many different approaches.



Conclusions Sparse State Estimation

- Solving the (moving horizon) state estimation problem with Sum-of-Norm (ℓ_1) regularization is a good way to handle sparse process noise.
- Performance is at least as good as for more complicated (hypothesis-testing) routines

State and Parameter Estimation

New problem: No longer assume that the parameter vector is known.

How to estimate also the parameter θ in the system description?

State and Parameter Estimation

Recall:

$$\tilde{\Theta} = [\theta, x(1), w(1), \ldots, w(N-1)] = [\theta, w(\cdot)]$$
$$x(k) = f(x(k-1), u(k-1), w(k-1), \theta) = x(k, \tilde{\Theta})$$
$$\min_{\tilde{\Theta}} \sum_{t=1}^{N} [\|y(t) - h(x(t, \tilde{\Theta}), u(t), \theta)\|^2 + \|w(t)\|^2]$$

This **is** (a) ML/MAP joint estimate of the states and the parameter vector.
View it as minimization over

$$\Theta = [\theta, x(1), x(2), \ldots, x(N)] = [\theta, x(\cdot)]$$
$$x(k) = f(x(k-1), u(k-1), w(k-1, x(\cdot)), \theta)$$

Joint State and Parameter Estimate

$$\min_{\Theta} V(\theta, x(\cdot))$$

$$V(\theta, x(\cdot)) = \sum_{t=1}^{N} \|y(t) - h(x(t), u(t), \theta)\|^2 + \|w(t, x)\|^2$$

$$'' \sim P(Y|\theta, X)''$$

$$[\hat{\theta}^J, x^s(t, \hat{\theta}^J)] = \hat{\Theta} = \arg\min_{\Theta} V(\theta, x(\cdot))$$

$$x^s(t, \theta^*) = \text{The smoothed states for given parameter } \theta^*$$

The states are *nuisance parameters* in the estimation of θ.

Possible Parameter Estimates

$\hat{\theta}^J$ as above; joint estimate with states

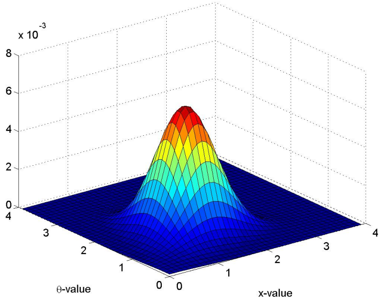
$$\hat{\theta}^{ML} = \arg\max P(Y|\theta) \quad \sim \arg\max \int P(Y|\theta, X)P(X|\theta)dX$$

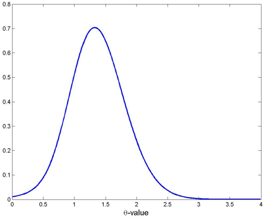
$$\hat{\theta}^{SEM} = \arg\min_{\theta} \sum_{t=1}^{N} \|y(t) - h(x^s(t,\theta), u(t), \theta)\|^2$$ "smoothing error est"

- $\hat{\theta}^J$ is conceptually simple to compute (in line with MPC) - could be a lot of numerical work, though.
- $\hat{\theta}^{SEM}$ sounds like a good idea: "Smoothing Error minimization should be better than Prediction Error minimization"
- $\hat{\theta}^{ML}$ has good credentials, but the ML criterion for nonlinear models involves solving the non-linear filtering problem.
- (The marginalization wrt x above is an extensive task.)

Marginalization: Picture

The integration will of course in general affect the maximum:





One more Estimation Method: EM

When the likelihood function is difficult to form, it may be advantageous to extend the problem with latent variables for a well defined likelihood function, and iterate between estimating these variables and the parameters.

This is the EM-algorithm, and in our case the states x can serve as these latent variables.

Take expectation of $V(\theta, x(.))$ under the assumption that x has been generated by the model with the parameter value α:

$$
\begin{aligned}
Q(\theta, \alpha) &= E[V(\theta, x(\cdot))|Y, \theta = \alpha)] \\
\theta^k &= \arg\min_\theta Q(\theta, \theta^{k-1}) \\
\hat{\theta}^{EM} &= \lim_{k \to \infty} \theta^k \qquad [\approx \theta^{ML}?]
\end{aligned}
$$

- How much work is required to form $Q(\theta, \alpha)$?

AUTOMATIC CONTROL
REGLERTEKNIK
LINKÖPINGS UNIVERSITET

Linear Models

How are these estimates related - and are they any good?

3

Parameter and State Estimation in Unknown Linear Systems

Linear Model: (Joint discussions with Thomas Schön and David Törnqvist)

$$x(t+1) = A(\theta)x(t) + B(\theta)u(t) + w(t)$$
$$y(t) = C(\theta)x(t) + e(t)$$
$$Ew(t)w^T(t) = Q(\theta) \quad Ee(t)e^T(t) = R(\theta)$$

Specialize to (without much loss of generality):

$$u(t) \equiv 0, \quad Q(\theta) = I, \quad R(\theta) = I$$

Notation

$$X^T = \begin{bmatrix} x(1)^T & x(2)^T & \cdots & x(N)^T \end{bmatrix}$$

W^T;E^T, and Y^Tanalogously

$$F_\theta = \begin{bmatrix} I & 0 & 0 & \cdots & 0 \\ A(\theta) & I & 0 \cdots & 0 & \\ A^2(\theta) & A(\theta) & I \cdots & 0 & \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A^{N-1}(\theta) & A^{N-2}(\theta) & A^{N-3}(\theta) & \cdots & 0 \end{bmatrix}$$

$$H_\theta = \begin{bmatrix} C(\theta) & 0 & \cdots & 0 \\ 0 & C(\theta) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & C(\theta) \end{bmatrix}$$

Matrix Formulation

$$X = F_\theta W, \quad Y = H_\theta X + E$$

W and E are Gaussian random vectors $N(0, I)$.

$$Y = H_\theta F_\theta W + E$$

$$Y \in N(0, R_\theta), \quad R_\theta = H_\theta F_\theta F_\theta^T H_\theta^T + I$$

$$-2 \log P(Y|\theta) = Y^T R_\theta^{-1} Y + \log \det R_\theta$$

$$-2 \log P(Y|\theta, X) = \|Y - H_\theta X\|^2$$

In a Bayesian setting with $\theta \in N(0, cI)$

$$P(Y, X, \theta) = P(Y|X, \theta) P(X|\theta) P(\theta)$$

$$V(Y, X, \theta) = -2 \log P(Y, X, \theta) = ||Y - H_\theta X\|^2 + \|F_\theta^{-1} X\|^2 + \|\theta\|^2 / c$$

The Estimates

Joint Criterion:

$$W(\theta, X) = \|Y - H_\theta X\|^2 + \|F_\theta^{-1} X\|^2 \qquad \text{"}(c \to \infty)\text{"}$$

Estimates:

State: $X^s(\theta) = F_\theta F_\theta^T H_\theta^T R_\theta^{-1} Y \qquad (Y - H_\theta X^s(\theta) = \ldots = R_\theta^{-1} Y)$

Joint: $\theta^J = \arg\min \|R_\theta^{-1} Y\|^2 + \|F_\theta^{-1} F_\theta F_\theta^T H_\theta^T R_\theta^{-1} Y\|^2$

$= \arg\min Y^T R_\theta^{-1} Y$

Smoothed: $\hat{\theta}^{SEM} = \arg\min \|R_\theta^{-1} Y\|^2 = \arg\min Y^T R_\theta^{-2} Y$

ML: $\hat{\theta}^{ML} = \arg\min Y^T R_\theta^{-1} Y + \log\det R_\theta$

EM: $Q(\theta, \alpha) = \ldots$

Expected Values of the Criteria

Let the true covariance matrix of Y be $R_0 = EYY^T$

$$\begin{aligned} \text{ML:} \quad & \text{trace} R_0 R_\theta^{-1} + \log\det R_\theta \\ \text{J:} \quad & \text{trace} R_0 R_\theta^{-1} \\ \text{SEM:} \quad & \text{trace} R_\theta^{-1} R_0 R_\theta^{-1} \end{aligned}$$

Note

$$\text{trace}\, BA^{-1} + \log\det A \geq \dim B + \log\det B \quad \forall A$$

$$\text{equality iff all eigenvalues of } BA^{-1} \equiv 1$$

So ML is consistent (but not the others!)

Numerical Illustration

The values that minimize the expected value of the criterion functions (= the limiting estimates as the number of observations tend to infinity) for the system

$$x(t+1) = ax(t) + w(t); \quad y(t) = x(t) + e(t); \quad a = 0.7$$

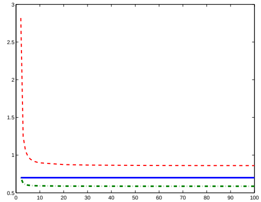


Figure: The minimizing values of the expected criterion functions as a function of N. Blue solid line: ML, Green dash-dotted line: SEM, Red dashed line: J.

EM-algorithm for this simple case

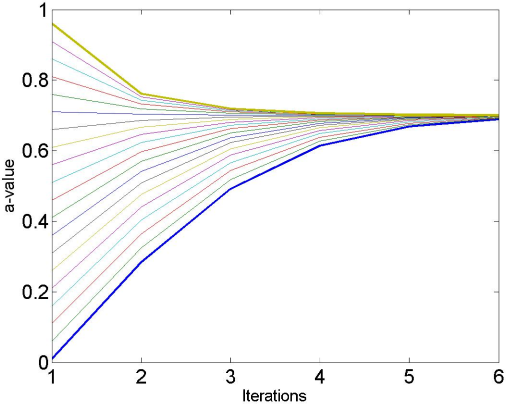


Figure: The estimates over the first six iterations of the EM-algorithm for different initial guesses

AUTOMATIC CONTROL
REGLERTEKNIK
LINKÖPINGS UNIVERSITET

Some Observations

- J $\sim W(\theta, X)$ ML $\sim$ " $\int W(\theta, X) dX$"
- J$\sim Y^T R_\theta^{-1} Y$ ML $\sim Y^T R_\theta^{-1} Y + \log\det R_\theta$
- J is not consistent,(but ML is, of course)
- J and ML are different maxima of $W(\theta, X)$
- The marginalization of $W(\theta, X)$ only leads to a data-independent (regularization) term $\log\det R_\theta$
- Is a similar result true also in the non-linear case?
- How would EM work in the non-linear case? (Schön, Wills, Ninness: Automatica 2011.)

Conclusions: State and Parameter Estimation

- Tempting to use MPC-thinking for model estimation using Moving Horizon Estimation - "Just" minimize over θ as well.
- This however leads to inconsistent estimates.
- Can it be saved by thoughtful regularization?