Nonlinear System Identification: A Palette from Off-white to Pit-black

Lennart Ljung
Automatic Control, ISY, Linköpings Universitet

This Presentation ...

... aims at

- a display of the essence of the problem of non-linear identification
- a color-coded overview of typical parametric approaches

A Common Frame

The world of nonlinear models is very diverse. A common framework: Discrete time observations of inputs and outputs:

$$Z^t = [u(1), u(2), ..., u(t), y(1), y(2), ..., y(t)]$$

A model is a parameterized predictor of the next output $y(t)$ made at time $t - 1$:

$$\hat{y}(t|t-1, \theta) = \hat{y}(t|\theta) = h(Z^{t-1}, \theta)$$

The parameters can be estimated using the prediction error method:

$$\hat{\theta} = \arg \min_\theta \sum_t ||y(t) - h(Z^{t-1}, \theta)||^2$$

(could be Maximum Likelihood)
What’s Special with Nonlinear Models?

The calculation of the true predictor is a nonlinear function of \( Z \). What makes the nonlinear problem much more difficult and rich than the linear problem?

Two major reasons:

- The richness of the model surface
- Propagation of noise signals to the output not immediate

Propagation of Noise Signals

- In linear systems that are cascaded we can always propagate the noise signals to the output: \( y = Gu + Hc \), where \( H \) picks up the coloring obtained by propagating the noise through a linear system.
- For nonlinear systems, this is generally not possible.
- Example: A linear system + noise, \( z = Gu + w \) is followed by a static nonlinearity \( f(z) \). At the output we have

\[
\begin{align*}
y(t) &= f(Gu + w) = f(Gu) + \tilde{w} \\
\tilde{w} &= f(Gu + w) - f(Gu)
\end{align*}
\]

Here, \( \tilde{w} \) is not really a “noise”: It is clearly contaminated with the input \( u \) which will create bias-problems when minimizing the output error. Indicates that the calculation of the true predictor could be challenging.

The Model Surface

Let us take \( Z^t = [u(t-1), u(t-2)] \) and a scalar output \( y(t) \). A model is then a surface in the space spanned by \([y(t), u(t-1), u(t-2)]\) and the estimation task is to estimate this surface.

Linear:

\[
\hat{y}(t) = a_1u(t-1) + a_2u(t-2)
\]

Nonlinear:

\[
\hat{y}(t) = h(u(t-1), u(t-2))
\]

The observations \( Z^t \) are points in this space.

The Palette of Nonlinear Models

- White: Known model
- Off-white: Careful Physical Modeling w or w/o noise models
- Smoke-grey: Semi-physical modeling (Could be used more!)
- Steel-grey: Local Linear Models
- Slate-grey: Block-oriented Models.
- Black: Flexible structures – universal approximators
- Pit-black: Non-Parametric Smoothing
Perform physical modeling (e.g. in MODELICA) and denote unknown physical parameters by $\theta$. Collect the model equations as

$$\dot{x}(t) = f(x(t), u(t), \theta)$$

$$y(t) = h(x(t), u(t), \theta)$$

(or in DAE, Differential Algebraic Equations, form.) For each parameter $\theta$ this defines a simulated output $\hat{y}(t|\theta)$ which is the parameterized function from sampled data

$$\hat{y}(t|\theta) = h(Z|\theta^{-1}, \theta) \quad (Z|\theta^{-1} = u|\theta^{-1})$$

in somewhat implicit form. To be a correct predictor this really assumes white measurement noise. Then the estimate is the $\theta$ that minimizes the output error fit $\sum ||y(t) - \hat{y}(t|\theta)||^2$

**Example: Missile**

10 inputs, 5 outputs, 16 unknown parameters.

The Equations

function [dx, y] = missile(t, x, p, u);

MISSILE A non-linear missile system.

Output equation. $y = \{x(1); \ldots; x(2); \ldots; x(3); \ldots$;

-p(18)*u(4)*(p(1)*x(5)+p(2)*u(3))/p(22); ...

-p(18)*u(4)*(p(3)*x(4)+p(4)*u(2))/p(22); ...

State equations. $dx =$

1/p(19)*(p(17)*p(18)*(p(5)*x(5)+0.5*p(6)*p(17)*x(1)/u(5)+ ... %

Angular velocity around x-axis.

p(7)*u(1)*u(4)-(p(21)-p(20))*x(2)^2*x(3)) + ... p(23)*(u(6)-x(1)); ... 1/p(20)*(p(17)*p(18)*(p(8)*x(4)+0.5*p(9)*p(17)*x(2)/u(5)+ ... ...

p(1)-p(25) unknown parameters u, y : measured inputs and outputs
Off-white Models with Noise Models

The (output error, off-white) approach is conceptually simple, but could be very demanding in practice.

A main shortcoming is the use of the output error criterion, which really assumes white measurement noise. Noise signals in nonlinear models cannot really be propagated to the output. If the size of the noise is non-trivial, more careful noise modeling should be done:

\[
\dot{x}(t) = f(x(t), u(t), w(t), \theta)
\]
\[
y(t) = h(x(t), u(t), \theta) + e(t)
\]

where \(w\) and \(e\) are white noises. To find correctly predicted outputs \(\hat{y}(t|Z^{t-1}, \theta) = E(y(t)|Z^{t-1}, \theta)\) is then the well-known “intractable” problem of nonlinear filtering. Often one has to resort to some simplistic observer.

Probabilistic Learning

Recently however, with the increasing computing power, new computing intensive simulation based methods have been developed for nonlinear filtering problem, and hence for applying the Maximum Likelihood method to non-linear state space models. Particle filtering, Markov Chain Monte Carlo, MCMC, Sequential Monte Carlo ....

Loosely, and briefly, these are based on simulation of the noisy state-space model, and evaluating the state probabilities, focusing on paths that give the measured output sequence.

It is a central current research area, Probabilistic Learning, to make these calculations as efficient as possible.

See e.g. Thomas B. Schön, Andreas Svensson, Lawrence Murray, and Fredrik Lindsten: Probabilistic learning of nonlinear dynamical systems using sequential Monte Carlo, ArXiv

The Palette of Nonlinear Models

- White: Known model
- Off-white: Careful Physical Modeling w or w/o noise models
- Smoke-grey: Semi-physical modeling (Could be used more!)
- Steel-grey: Local Linear Models
- Slate-grey: Block-oriented Models.
- Black: Flexible structures – universal approximators
- Pit-black: Non-Parametric Smoothing
Apply non-linear transformations to the measured data, so that the transformed data stand a better chance to describe the system in a linear relationship.

“Rules: Only high-school physics and max 10 minutes"

. . . Square the voltage! Sense morale: No excuse for not thinking over the basic physical facts!

Another example: . . .
Re-sample Data

\[ z = [y,u]; \; pf = \text{flow./level}; \]
\[ t = 1:\text{length}(z) \]
\[ \text{newt} = \text{interp1}([\text{cumsum}(pf),t],[pf(1):\text{sum}(pf)]'); \]
\[ \text{newz} = \text{interp1}([t,z], \text{newt}); \]

Semi-physical Model

\[
G(s) = \frac{0.8116}{1 + 110.28s} e^{-369.58s}
\]

The semi-physical model gives a sufficiently good description of the buffer, to allow proper time-marking of the pulp before and after.

The Palette of Nonlinear Models

- White: Known model
- Off-white: Careful Physical Modeling w or w/o noise models
- Smoke-grey: Semi-physical modeling (Could be used more!)
- Steel-grey: Local Linear Models
- Slate-grey: Block-oriented Models.
- Black: Flexible structures – universal approximators
- Pit-black: Non-Parametric Smoothing

Steel-Grey: Composite Local Models

Non-linear systems are often handled by linearization around a working point. The idea behind Composite Local (Local Linear) Models is to deal with the nonlinearities by selecting or averaging over relevant linearized models.

Example: Tank with inflow \( u \) and free outflow \( y \) and level \( h \): (Bernoulli’s) equations:

\[
\dot{h} = -\sqrt{h} + u; \quad y = \sqrt{h}
\]

Linearize around level \( h^* \) with corresponding flows \( u^* = y^* = \sqrt{h^*} \):

\[
\dot{h} = -\frac{1}{2\sqrt{h^*}}(h - h^*) + (u - u^*); \quad y = y^* + \frac{1}{2\sqrt{h^*}}(h - h^*)
\]
Tank Example, ctd

Sampled data model around level $h^*$ (Sampling time $T_s$):

$$y(t) = \gamma(h^*) + \alpha(h^*)y(t-T_s) + \beta(h^*)u(t-T_s) = \theta^T(h^*)\varphi(t)$$

An ARX-model with level-dependent parameters. Now compute linearized model for $d$ different levels, $h_1, h_2, \ldots, h_d$. Total model: select or average over these local models

$$\hat{y}(t) = \sum_{k=1}^{d} w_k(h(t), h_k)\theta^T(h_k)\varphi(t)$$

Choices of weights $w_k : \ldots$

Local Linear Models

Two levels (models) ($d=2$)  Five levels (models) ($d = 5$)

Composite Local Models: General Comments

Let the measured working point variable (tank level in example) be denoted by $\rho(t)$ (sometimes called regime variable or scheduling variable). If the regime variable is partitioned into $d$ values $\rho_k$, and model output according to value $\rho_k$ is $\hat{y}^{(k)}(t)$ the predicted output will be

$$\hat{y}(t) = \sum_{k=1}^{d} w_k(\rho(t), \rho_k)\hat{y}^{(k)}(t)$$

If the prediction $\hat{y}^{(k)}(t)$ corresponding to $\rho_k$ is linear in the parameters, $\hat{y}^{(k)}(t) = \varphi^T(t)\theta^{(k)}$, and the weights $w$ are fixed, the whole model will be a linear regression.

Important connections to active research areas

- LPV (Linear Parameter-Varying) Models
- Hybrid Models ($\approx w(\cdot, \cdot)$ is estimated too.)
\[ x(t+1) = A(\rho)x(t) + B(\rho)u(t) \]
\[ y(t) = C(\rho)x(t) \]

is a linear model for each fixed \( \rho \).

If \( \rho \in \Omega = \{ \rho_1, \ldots, \rho_d \} \) it is a set of local linear models. If \( \rho = \rho(t) \) is time varying, we have a Linear Parameter Varying model. A basic difficulty is to find a common state basis from input/output observations and to manage the time variable in \( \rho(t) \) in relation to the observations.

### Slate Grey: Block-oriented Models

**Building Blocks:**

- **Linear Dynamic System**
  \[ G(s) \]

- **Nonlinear static function**
  \[ f(u) \]
Other Combinations

Active Research Field:

With the linear blocks parameterized as a linear dynamic system and
the static blocks parameterized as a function (“curve”), this gives a
parameterization of the output as

$$\hat{y}(t|\theta) = h(Z^{t-1}, \theta)$$

and the general approach of model fitting can be applied.
However, in this contexts many algorithmic variants have been
suggested.

Example: Hydraulic Crane Data

These are data from a forest harvest machine:

Input: Hydraulic Pressure.
Output: Tip Position

Linear Model

Black: Measured Output
Blue: Model Simulated Output

Hammerstein Model of the Hydraulic Crane

The Hammerstein Model gives a good fit. The extra flexibility
offered by the input nonlinearity is quite useful,
(even though no direct physical explanation is obvious.)
Noise Effects in Hammerstein-Wiener?

There is frequently reason to assume that some noise enters before the output nonlinearity $g$.

What happens if we propagate that noise to the output and apply an Output Error criterion to the above 2 input 2 output system?

#### Maximum Likelihood (EM) for HW Models

It is clear that more effort must be paid to the noise structure. We turn to the Maximum Likelihood method for the HW model structure. It is a complication that the ML criterion cannot easily be formed. But if the unmeasured noise $\nu$ were known it is easy to compute the ML criterion. So, treat it as “incomplete data” $X$ and apply the EM algorithm, which iterates between estimating $X$ and estimating the model for this $X$.


#### Results for the ML-EM Method for HW Models

Blue curve: Plots for the true system
Red curves: Median and standard deviations for estimated systems over 80 Monte Carlo runs
Number of observed data: 2000

Number of observed data: 2000

Red curves: Median and standard deviations for estimated systems over 80 Monte Carlo runs
Number of observed data: 2000

Blue curve: Plots for the true system
Red curves: Median and standard deviations for estimated systems over 80 Monte Carlo runs
Number of observed data: 2000
It is natural to think of Taylor expansions: $\kappa(z) = z^k$. If $na = 0$ (NLFIR), this becomes the classical Volterra series expansion. But note that if $\dim z = r$, then $z^k$ has $r^k$ components!

A more common choice is to form all the basis functions $\kappa_k$ from one mother function $\kappa$ and scale and position the argument differently:

$$
\kappa(z) = \kappa(\beta_k(\varphi - \gamma_k))
$$

$$
\hat{y}(t|\theta) = \sum_{k=1}^{d} a_k \kappa(\beta_k(\varphi - \gamma_k)), \quad \theta = \{a_k, \beta_k, \gamma_k\}
$$

Intuitive picture: Think of a scalar $\varphi$ and let $\kappa(z)$ be a unit pulse for $0 \leq z \leq 1$. Then $\kappa(\beta(\varphi - \gamma))$ is a pulse of width $1/\beta$ starting in $\varphi = \gamma$. The sum above is then a piecewise constant function, capable of approximation "any" function arbitrary well for large enough $d$. ⇒ ANN, LS-SVM etc (Sjöberg et al, Automatica 1995)
The Palette of Nonlinear Models

- White: Known model
- Off-white: Careful Physical Modeling w or w/o noise models
- Smoke-grey: Semi-physical modeling (Could be used more!)
- Steel-grey: Local Linear Models
- Slate-grey: Block-oriented Models.
- Black: Flexible structures – universal approximators
- Pit-black: Non-Parametric Smoothing

Conclusions

- Confusingly many approaches!
- A user-oriented roadmap would be excellent!

Pit-black Models: Non-Parametric Smoothing Methods

Form the model surface \( h(\varphi(t)) \) by smoothing over the observation points in the space!

- Even Blacker!
- Huge literature – Mostly in the statistical community and now also in machine learning
- Important to find lower dimensional manifolds (– counterpart of PCA in linear modelling). Concepts like Manifold Learning and Local Linear Embedding become central.