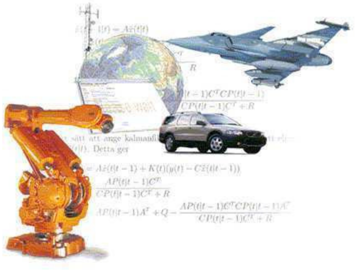
Using Multiple Kernel-based Regularization for Linear System Identification

What *are* the Structure Issues in System Identification?



Lennart Ljung
with coworkers; see last slide

Reglerteknik, ISY, Linköpings Universitet

Outline

Two parts:

- Prologue: Confessions of a Conventional System Identification Die-hard
- Some Technical Results on Choice of Regularization Kernels

System Identification

A Typical Problem

Given Observed Input-Output Data: Find the Impulse Response (IR) of the System that Generated the Data

Basic Approach

Find a suitable Model Structure, Estimate its parameters, and compute the IR of the resulting model

Techniques

Estimate the parameters by ML techniques/PEM (prediction error methods). Find the model structure by AIC, BIC or Cross Validation

Status of the "Standard Framework"

- The model structure is large enough (to contain a correct system description): The ML/PEM estimated model is (asymptotically) the best possible one. Has smallest possible variance (Cramér- Rao)
- The model structure is not large enough: The ML/PEM estimate converges to the best possible approximation of the system (for the experiment conditions in question). Smallest possible "asymptotic bias"
- The mean square error (MSE) of the estimate is MSE=Bias^2+Variance
- The choice of "size" of the models structure governs the Bias/Variance Trade Off.

What are the Structure Issues? - Part I

Structure = Model Structure

$\mathcal{M}(\theta)$ e.g.

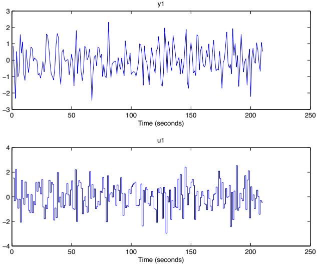
$$\begin{aligned} x(t+1) &= A(\theta)x(t) + B(\theta)u(t) + w(t) \\ y(t) &= C(\theta)x(t) + e(t) \end{aligned}$$

Find the parameterization!

Today: No particular internal structure, just need to determine the order $n = \dim x$. Also, no noise model ($w \equiv 0$) ("Output error models.")

A Simple Experiment

Look at data from a randomly generated system (selected, but typical)



Estimate models of different orders $k = 1, \ldots, 30$ by PEM/ML

```
m(k)= pem(data,k,’dist’,’no’);
```

Now we have 30 models, which one to pick?

Hypothesis Tests: Compare Loss Functions (Criteria)

Loss function (neg log likelihood):

$V = 1/N \sum_{t=1}^{N} |y(t) - \hat{y}(t|t-1)|^2$

Model Order	Log V
1	-0.41
2	-2.08
4	-2.40
6	-2.57
9	-2.76
11	-2.80
17	-2.88
19	-2.88
22	-2.96
29	-3.22

Hypothesis Tests: Compare Fits

Fit$= (1 - \sqrt{\frac{V}{1/N \sum |y|^2}}) * 100)$: The percentage of the output variation, reproduced by the model.

Model Order	Log V	Fit
1	-0.41	7.04
2	-2.08	61.28
4	-2.40	65.52
6	-2.57	68.28
9	-2.76	71.19
11	-2.80	71.68
17	-2.88	72.87
19	-2.88	72.91
22	-2.96	74.00
29	-3.22	77.25

Hypothesis Tests: Compare Fits for Validation Data

CVFit=Compute the model's fit on independent validation data.

Model Order	Log V	Fit	CVFit
1	-0.41	7.04	-2.14
2	-2.08	61.28	57.40
4	-2.40	65.52	60.37
6	-2.57	68.28	61.29
9	-2.76	71.19	60.32
11	-2.80	71.68	61.43
17	-2.88	72.87	56.01
19	-2.88	72.91	58.07
22	-2.96	74.00	56.37
29	-3.22	77.25	-57.89

Hypothesis Tests: Compare AIC and BIC Criteria

AIC = log (Loss) + 2*dim(θ)/N

BIC = log (Loss) + log(N)*dim(θ)/N

N = number of observed data

Model Order	Log V	Fit	CVFit	AIC	BIC
1	-0.41	7.04	-2.14	6.01	4.50
2	-2.08	61.28	57.40	58.64	57.30
4	-2.40	65.52	60.37	63.52	59.85
6	-2.57	68.28	61.29	65.46	60.13
9	-2.76	71.19	60.32	67.26	59.40
11	-2.80	71.68	61.43	66.88	56.92
17	-2.88	72.87	56.01	65.40	48.04
19	-2.88	72.91	58.07	64.39	43.91
22	-2.96	74.00	56.37	64.34	39.67
29	-3.22	77.25	-57.89	65.25	30.49

Enter ZZZ: A New Method for Order Determination

H. Hjalmarsson gave me some new code: `mz = ZZZ(data)`.
His algorithm is not published yet. It is a way to find the simplest model that has a fit (sum of squared innovations) that is not falsified relative to a crude estimate of the innovations variance.

Model Order	Log V	Fit	CVFit	AIC	BIC	ZZZ
1	-0.41	7.04	-2.14	6.01	4.50	-
2	-2.08	61.28	57.40	58.64	57.30	-
4	-2.40	65.52	60.37	63.52	59.85	*
6	-2.57	68.28	61.29	65.46	60.13	-
9	-2.76	71.19	60.32	67.26	59.40	-
11	-2.80	71.68	61.43	66.88	56.92	-
17	-2.88	72.87	56.01	65.40	48.04	-
19	-2.88	72.91	58.07	64.39	43.91	-
22	-2.96	74.00	56.37	64.34	39.67	-
29	-3.22	77.25	-57.89	65.25	30.49	-

Where Are We Now?

We have computed 30 models of orders 1 to 30. We have four suggestion for which model to pick:

- Cross Validation: Order 11
- AIC Criterion: Order 9
- BIC Criterion: Order 6
- ZZZ Criterion: Order 4

Which choice is really best?

Enter the Oracle!

In this simulated case the true systems is known, and we can compute the actual fit between the true impulse response (from time 1 to 100) and responses of the 30 models:

Order	Log V	Fit	CVFit	AIC	BIC	ZZZ	Actual Fit
1	-0.41	7.04	-2.14	6.01	4.50	-	6.89
2	-2.08	61.28	57.40	58.64	57.30	-	77.01
4	-2.40	65.52	60.37	63.52	59.85	*	85.80
6	-2.57	68.28	61.29	65.46	60.13	-	83.18
9	-2.76	71.19	60.32	67.26	59.40	-	80.81
11	-2.80	71.68	61.43	66.88	56.92	-	79.57
17	-2.88	72.87	56.01	65.40	48.04	-	77.65
19	-2.88	72.91	58.07	64.39	43.91	-	79.66
22	-2.96	74.00	56.37	64.34	39.67	-	78.91
29	-3.22	77.25	-57.89	65.25	30.49	-	72.61

Lessons from This Test of the Traditional Approach

- Relatively straightforward (but somewhat time-consuming) to estimate all models.
- No definite rule to select the best model order.
- In this case Hjalmarsson's ZZZ order test gave the best advice (showing that there is much more to model order selection than the traditional tests)
- The fit 85.80% is the best fit among all the 30 models, showing that this is the best impulse response we can achieve within the traditional approach.

Enter XXX

Another friend of mine (Gianluigi Pillonetto) gave me an m-file to test:

`mx = xxx(data)`

It produces an FIR model `mx` of order 100. The fit of this model's impulse response to the true one is

87.51 %!!

Recall that the best possible fit among the traditional models was 85.80 %!

Well, `mx` is not a state space model of manageable order. But e.g. `m7=balred(mx,7)` is a 7th order state space model with a IR fit of 87.12 %. Note that the 7th order ML model had a fit of 77.56 %.

Some cracks in the foundation of the standard approach.

So what does `xxx` do?

XXX: Regularized FIR Models

From an (finite)impulse response model

$$y(t) = \sum_{k=1}^{n} g(k)u(t-k) + v(t);\ t = 1, \dots, N$$

a simple linear regression can be formed

$$Y = \Phi^T \theta + V$$

with θ being the vector of $g(k)$ and Φ constructed from the inputs $u(s)$.

XXX then estimates θ as the regularized Least Squares estimate

$$\hat{\theta}_N = \arg\min_{\theta} \|Y - \Phi^T\theta\|^2 + \theta^T D^{-1} \theta$$

for some carefully chosen *regularization matrix* D.

Structure Issues – Part II: How to Choose the Regularization Matrix D?

The focus of the *question of suitable structures for the identification problem* is then shifted from discrete model orders to *continuous tuning of D.*

The bias-variance trade-off has thus become a richer problem.

There are not many concrete analytical method for how to parameterize and tune the regularization matrix (which contains $\approx n^2/2$, $n \sim 100$ elements). The more technical part of this presentations will discuss one particular parametrization and tuning algorithm.

Choice of D: Classical Perspective

From a classical, frequentist point of view we can compute the MSE matrix of the impulse response vector: Let $EVV^T = I$, $R = \Phi\Phi^T$ and θ_0 be the true impulse response. Then

$$\begin{aligned} MSE(D) =& E(\hat{\theta}_N - \theta_0)(\hat{\theta}_N - \theta_0)^T = \\ & (R + D^{-1})^{-1}(R + D^{-1}\theta_0{\theta_0}^T D^{-T})(R + D^{-1})^{-1} \end{aligned}$$

This is minimized wrt D (also in matrix sense) by

$$D_{opt} = \theta_0\theta_0^T$$

What is the best average MSE over a set $\{\theta_0\}$ with $E\theta_0\theta_0^T = P$?

$$E\ MSE(D) = (R + D^{-1})^{-1}(R + D^{-1}PD^{-T})(R + D^{-1})^{-1}$$

Minimized by $D_{opt} = P$. Notice the link to Bayesian framework!

Parameterization of D

So, the matrix – or the Kernel –D should mimic typical behavior of the impulse responses, like exponential decay and smoothness. A common choice is TC (“Tuned/Correlated”) (what was used in XXX);

$$D^{TC}_{j,k}(\alpha) = C\min(\lambda^k, \lambda^j),\ \lambda < 1 \quad \alpha = [C, \lambda]$$

Related, common kernels are DC(Diagonal/Correlated) and SS (Stable Splines).

$$D^{DC}_{j,k}(\alpha) = C\lambda^{(j+k)/2}\rho^{|j-k|}, \quad \alpha = [C, \lambda, \rho]$$

$$D^{SS}_{j,k}(\alpha) = C\frac{\lambda^{2k}}{2}(\lambda^j - \frac{\lambda^k}{3}), k \geq j, \quad \alpha = [C, \lambda]$$

Tuning of the Parameters D

The kernel $D(\alpha)$ depends on the hyper-parameters α. They can be tuned by invoking a Bayesian interpretation:

$$\begin{aligned}
&Y = \Phi^T\theta + V \\
&V \in N(0, \sigma^2 I),\ \theta \in N(0, D(\alpha)),\ \Phi \text{ known} \\
&Y \in N(0, \Sigma(\alpha)),\ \Sigma(\alpha) = \Phi^T D(\alpha)\Phi + \sigma^2 I
\end{aligned}$$

ML estimate of α: ("Empirical Bayes")

$$\hat{\alpha} = \arg\min_{\alpha} Y^T \Sigma(\alpha)^{-1} Y + \log\det\Sigma(\alpha)$$

(Typically Non-Convex Problem)

Wish List for D: Three Properties

1. Should have a flexible structure so that diverse and complicated dynamics can be captured
2. Should make the non-convex hyper-parameter estimation problem (”the empirical Bayes estimate”) easy to solve
   - an efficient algorithm and implementation to tackle the marginal likelihood maximization problem
3. Should have the capability to tackle problems of finding sparse solutions arising in system identification
   - sparse dynamic network identification problem
   - segmentation of linear systems
   - change detection of linear systems

Suggested Solution: Multiple Kernels

The multiple kernel given by a conic combination of certain suitably chosen fixed kernels has these features.

$$D(\alpha) = \sum_{i=1}^{m} \alpha_i P_i, \quad \alpha = \left[\alpha_1, \cdots, \alpha_m\right]^T \tag{1}$$

where $P_i \succeq 0$ and $\alpha_i \geq 0$, $i = 1, \cdots, m$

- The fixed kernels P_i can be instances of any existing kernels, such as SS, TC and DC for selected values of their hyper-parameters
- The fixed kernels P_i can also be constructed as

  $$P_i = \hat{\theta}_i \hat{\theta}_i^T \tag{2}$$

  where $\hat{\theta}_i$ contains the impulse response coefficients of a preliminary model.

1. Capability to Better Capture Diverse Dynamics

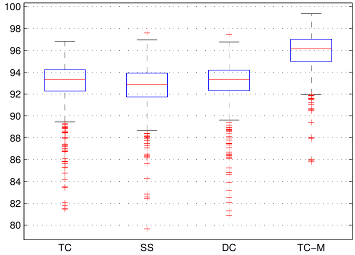
Consider second order systems in the form of

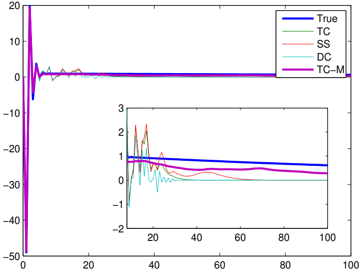
$$G_0(q) = \frac{z_1 q^{-1}}{1 - p_1 q^{-1}} + \frac{z_2 q^{-1}}{1 - p_2 q^{-1}} \tag{3}$$

where $z_1 = 1, z_2 = -50$ and $p_i, i = 1,2$ are generated as p_1 = `rand(1)/2+0.5` and p_2 = `sign(randn(1))*rand(1)/2`. Compare conventional kernels (TC, SS, DC) with a multiple kernel consisting of 20 fixed TC kernels for different vales of λ (TC-M).

Boxplots of Fits over 1000 Systems.

The Fit is as before the relative fit between the impulse responses of the true system and the model, in %. (100% is a perfect fit)





2. Efficient Hyper-parameter Estimation

Recall the Empirical Bayes kernel tuning:

$$[\hat{\alpha}, \hat{\sigma}^2] = \arg\min_{\sigma,\alpha \geq 0} H(\alpha, \sigma^2)$$

$$H(\alpha, \sigma^2) = Y^T \Sigma(\alpha, \sigma^2)^{-1} Y + \log |\Sigma(\alpha, \sigma^2)|$$

$$\Sigma(\alpha, \sigma^2) = \Phi^T D(\alpha) \Phi + \sigma^2; \quad D(\alpha) = \sum \alpha_i P_i$$

Note that for the multiple kernel approach, $D(\alpha)$ is linear in α, so

- $Y^T \Sigma(\alpha, \sigma^2)^{-1} Y$ is convex in $\alpha \geq 0$ and $\sigma^2 > 0$.
- $\log |\Sigma(\alpha, \sigma^2)|$ is concave in $\alpha \geq 0$ and $\sigma^2 > 0$.

So H is a difference of two convex functions, which means that the minimization is a difference of convex programming (DCP) problem Such problems can be solved efficiently as a sequence of convex optimization problems, for example by the Majorization Minimization (MM) method.

3. Sparse Solutions for Structure Detection

Unknown structural issues may be model order, existing or non-existing connecting links in networks, abrupt changes at some time instant and so on.

A Generous parameterization, with zero/non-zero parameters defining structures is thus a desired feature.

That is, an estimation routine that favors sparse solutions is a important asset.

It is easy to use many kernels in the multiple kernel approach, since the estimation problem is a DCP problem. Kernel terms can be introduced, that correspond to structural issues as above.

But, does the algorithm favor sparse solutions?

3. Capability to Find Sparse Solutions

The kernel estimation problem is

$$\hat{\alpha} = \arg\min Y^T(\Phi^T[\sum_{i=1}^{p} \alpha_i P_i]\Phi + \sigma^2 I)^{-1} Y + \log|\Phi^T[\sum_{i=1}^{p} \alpha_i P_i]\Phi + \sigma^2 I|$$

Define $x_i = \alpha_i/\sigma^2,\ Q_i = \Phi^T P_i \Phi$ For a given σ^2, the estimation problem is equivalent to

$$\hat{x} = \arg\min_{x \geq 0} \quad Y^T(\sum_{i=1}^{p} x_i Q_i + I)^{-1} Y + \sigma^2 \log|\sum_{i=1}^{p} x_i Q_i + I|$$

Clearly, there exists σ^2_{max} such that $\hat{x} = 0$ for $\sigma^2 \geq \sigma^2_{max}$. The value of σ^2 will also control the sparsity of the minimizing x.

Same as the tuning of the regularization parameter in l_1-norm regularization techniques, e.g., LASSO. σ^2 can also be tuned by CV.

Back to Our Test System

Recall that we had fits to the true impulse response of
PEM + CV: 79.57 %
PEM + AIC: 80.81 %
PEM + BIC: 83.16 %
PEM + ZZZ: 85.80 %
Regularization by TC kernel: 87.51 %

Now, test it with Multiple kernels regularization: 90.27 %

Monte Carlo Tests over 2000 Systems

Methods:

- AIC, CV and ZZZ are Parametric methods (PEM/ML) with different order selections.
- TC, SS , DC are regularized FIR models with common kernels
- OE-M, DC-M, TCSS-M are multiple kernels containing 6, 54, and 29 fixed kernels

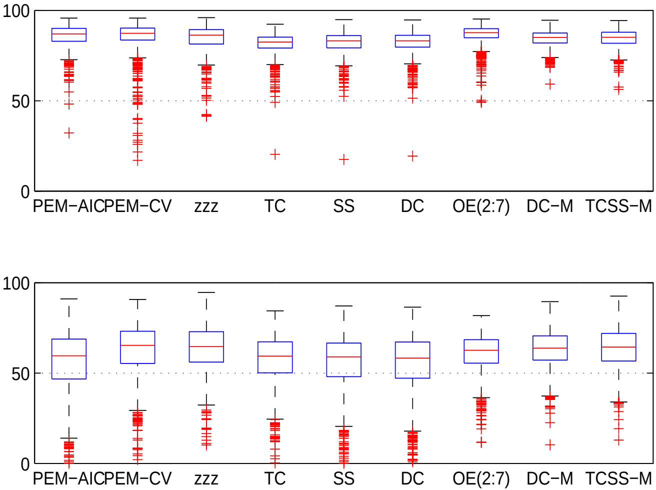
Data:

- Data: D1, D2 are 1000 systems with WGN input and SNR 10 and 1, resp.   210 data points

Legend: x|m: x average fit; m number of ”failures” (fit < 0).

AF|NO	PEM-AIC	PEM-CV	ZZZ	TC	SS	DC	OE(2:7)-M	DC-M	TCSS-M
D1	85.9|0	83.8|9	84.6|0	81.5|0	82.1|0	82.1|0	86.6|0	84.4|0	84.4|0
D2	56.5|7	62.2|13	63.3|2	55.9|25	56.1|6	54.3|24	61.1|0	63.2|0	63.7|0

Boxplots of Fits over 1000 + 1000 Random Systems



Conclusions

- Regularization in simple FIR models is a valuable alternative to conventional system identification techniques for estimation of unstructured linear systems
- The Regularization approach offers a greater variety of tuning instruments (kernels, regularization matrices) as an alternative to model orders for the bias-variance trade-off
- Regularization kernels that are formed as linear combinations of fixed, given kernels offer several advantages:
  - Potentially greater flexibility to handle diverse systems
  - Hyper-parameter tuning employing efficient convex programming techniques
  - Potential to handle sparsity in the estimation problems

References and Acknowledgments

- First part, “the confessions”, was based on and inspired by:
  T. Chen, H. Ohlsson and L. Ljung: On the estimation of transfer functions, regularization and Gaussian Processes – Revisited. *Automatica*, Aug 2012.
- Second part was based on:
  T. Chen, M.S. Andersen, L. Ljung, A. Chiuso, G. Pillonetto: System identification via sparse kernel-based regularization using sequential convex optimization techniques. *Submitted to the special issue of IEEE Trans. Autom. Control.*
- Funded by the ERC advanced grant LEARN

European Research Council
Established by the European Commission