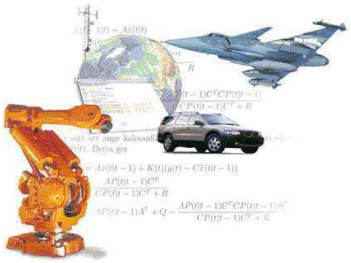
Full Bayesian Identification of Linear Dynamic Systems Using Stable Kernels



Lennart Ljung

Reglerteknik, ISY, Linköpings Universitet
Joint work with
Gianluigi Pillonetto. Padova University

Abstract

- A new technique to identifying linear black-box dynamic systems
- Outperforms existing methods
- Further development of kernel methods using “Empirical Bayes”
- We will work essentially with Linear Regressions.

Lennart Ljung
Paul van der Hof Symposium: Full Bayes Eindhoven April 19 2024
AUTOMATIC CONTROL
REGLERTEKNIK
LINKÖPINGS UNIVERSITET

Linear Regression for a General Linear Model

A general linear, single output model

$$y(t) = G(q,\theta)u(t) + H(q,\theta)e(t)$$

with nu inputs has an output predictor that is linear in past outputs and inputs:

$$\hat{y}(t|t-1) = \sum_{j=1}^{nu+1} \sum_{k=1}^{m_j} \tilde{h}_j(k)\tilde{u}_j(t-k)$$

where and $\tilde{h}$ are formed from G and H and $\tilde{u}_k = u_k, k = 1, ..., nu$ & $\tilde{u}_{nu+1}(k) = y(k)$.

Lennart Ljung
Paul van der Hof Symposium: Full Bayes Eindhoven April 19 2024
AUTOMATIC CONTROL
REGLERTEKNIK
LINKÖPINGS UNIVERSITET

A Simple Special Case: FIR model

Simple special case: A SISO FIR (finite impulse response) model :

$$y(t) = B(q)u(t) + e(t)$$
$$y(t) = b_1u(t-1) + \ldots + b_nu(t-n) + e(t) = \phi(t)\theta + e(t)$$
$$Y = \Phi^T\theta + E$$

Y and Φ^T collect the data for $t = 1 \ldots N$ in column vectors and θ is the impulse response of the system. (The general case with several inputs and an Autoregressive terms makes θ a “segmented long” matrix, with each segment being the impulse response from each of the inputs and from the past outputs.)

Lennart Ljung
Paul van der Hof Symposium: Full Bayes Eindhoven April 19 2024
AUTOMATIC CONTROL
REGLERTEKNIK
LINKÖPINGS UNIVERSITET

Linear Regression

Model

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

Idea #1, LS:

$$\min_\theta \|Y - \Phi^T\theta\|^2 \quad \rightarrow \quad \hat{\theta} = [\Phi\Phi^T]^{-1}\Phi^T Y$$

Idea #2, LS with regularization:

$$\min_\theta \|Y - \Phi^T\theta\|^2 + \theta^T D\theta \quad \rightarrow \quad \hat{\theta} = [\Phi\Phi^T + D]^{-1}\Phi^T Y$$

How to choose the regularization matrix D?

Bayesian setting: If a prior distribution $\theta \in N(0, P_0)$ is given and $E \in N(0, \sigma^2 I_N)$ then simple calculations show that the natural choice of D is

$$D = P_0^{-1}/\sigma^2$$

That makes

$$\hat{\theta} = [\Phi\Phi^T + D]^{-1}\Phi^T Y = \left(\Phi^T\Phi/\sigma^2 + P_0^{-1}\right)^{-1}\frac{\Phi^T}{\sigma^2}Y$$

which is the posterior mean $E\theta|Y$.

But P_0 and σ^2 are not known: Hyperparameters

The hyper parameters $\eta = \{\sigma^2, \alpha, \lambda\}$

- the noise variance σ^2
- the parameter covariance matrix $P_0(\alpha, \lambda)$
- $P_0(k, j) = \alpha \cdot \lambda^{\max(k,j)}$ (Size and Decay)
- Recall θ is the impulse response from the inputs (and past outputs) to the prediction.
- So the diagonal of P_0, i.e. the elements $\alpha\lambda^k$ give the decay of the impulse response. The correlation between two impulse response coefficients $\tilde{h}_j$ and $\tilde{h}_k$ decays like $\lambda^{|j-k|}$. It is convenient to measure both decay and correlation by the same entity λ , but it is easy to use different numbers for these (The "TC" and "DC" kernels, resp.)
- in the multi-input case there is one covariance matrix P_0 for each input (segment of the θ matrix)

The ML Estimate of the Hyperparameters

Actually, the ML estimate of the hyperparameter η is easily found.

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

is a sum of two Gaussian variables, so Y is Gaussian with mean zero and covariance $P_Y(\eta) = \Phi^T P_0(\eta)\Phi + \sigma^2(\eta)I_N$. So, given η the pdf of Y is known, and the ML estimate of η will be

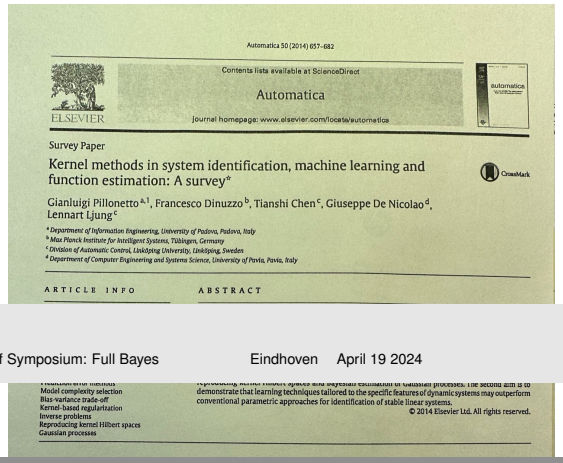
$$\hat{\eta}_{ML} = \arg\min_\eta Y^T P_Y(\eta)^{-1}Y + \log\det P_Y(\eta)$$

Empirical Bayes

Inserting the ML estimate $\hat{\eta}_{ML}$ of the hyperparameters in the regularization matrix

$$D = P_0^{-1}/\sigma^2$$

give the *Empirical Bayes* estimation method. This has proven to an effective method for identification of linear systems. Survey Paper



Automatica 50 (2014) 657–682

Contents lists available at ScienceDirect

Automatica

Journal homepage: www.elsevier.com/locate/automatica

Survey Paper

Kernel methods in system identification, machine learning and function estimation: A survey*

Gianluigi Pillonetto [a,1], Francesco Dinuzzo [b], Tianshi Chen [c], Giuseppe De Nicolao [d], Lennart Ljung [c]

[a] Department of Information Engineering, University of Padova, Padova, Italy
[b] Max Planck Institute for Intelligent Systems, Tübingen, Germany
[c] Division of Automatic Control, Linköping University, Linköping, Sweden
[d] Department of Computer Engineering and Systems Science, University of Pavia, Pavia, Italy

ARTICLE INFO

ABSTRACT

Model complexity selection
Bias-variance trade-off
Kernel-based regularization
Inverse problems
Reproducing kernel Hilbert spaces
Gaussian processes

demonstrate that learning techniques tailored to the specific features of dynamic systems may outperform conventional parametric approaches for identification of stable linear systems.

Beyond Empirical Bayes

We are looking for $\hat{\theta} = E(\theta|Y) = \int \theta \cdot p(\theta|Y)$ the mean of the posterior distribution. The posterior distribution of θ:

$$p(\theta|Y) = \int p(\theta, \eta|Y)d\eta = \int p(\theta|\eta, Y)p(\eta|Y)d\eta$$

If the posterior distribution $p(\eta|Y)$ is concentrated around the ML estimate $\hat{\eta}_{ML}$ we can use

$$p(\theta|Y) = \int p(\theta, \eta|Y)d\eta = \int p(\theta|\eta, Y)p(\eta|Y)d\eta \approx p(\theta|\hat{\eta}_{ML}, Y)$$

which is **Empirical Bayes (EB)**.

Constructing the posterior distribution by MCMC

To improve on the EB estimate we should construct the posterior distribution $p(\eta|Y)$ more seriously by treating also the hyperparameters as random variables and

- Defining priors for the hyperparameters η.
- Define a chain of randomly perturbed values of η.
- For given values of the hyperparameters the estimate of θ is well defined by regularized LS.
- Monitor the statistics of the corresponding sequence of θ -values (MCMC: Markov Chain Monte Carlo - process)

Hyperparameter *Priors* for σ^2, α, γ

- σ^2, α, are positive real numbers with *Jeffrey's prior*
  - "noninformative prior", its posteriors sample with the inverse gamma distribution $I_g(a, b)$ (with a, b given from observations of data)
  - $I_g(a, b)$ pdf $p(x) \sim x^{-a-1}e^{b/x}$
- λ positive scalar less than 1.

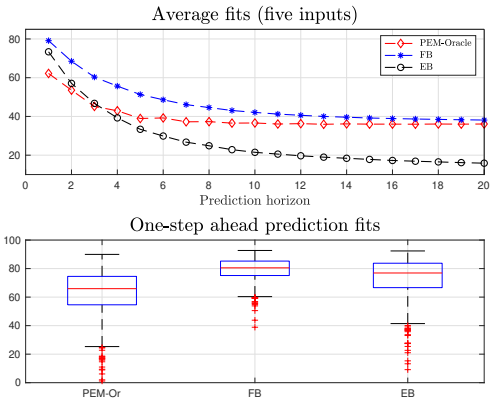
The MCMC Steps: Hyperparameters $\sigma^2, \lambda, \alpha$

1. Initialize: θ^0 is the LS estimate, σ_0^2= the loss function for the LS estimate. Set e.g. $\lambda = 0.95, \alpha = 1$. Set $\Delta = 0.02$
2. For $i = 1, 2, ..., M$ ($M \sim 10000$) repeat
   - resample the value for σ_i^2 as the loss for the current model $\sigma_i^2 = \|Y - \Phi\theta^{i-1}\|^2/(N-1)$. $[\sigma_i^2 \sim I_g(N/2, \|Y - \Phi\theta^{i-1}\|^2/2)]$
   - similarily, resample α as a value drawn from an inverse gamma distribution with parameters given by the current estimate of θ^{i-1}
   - make a random change of size Δ in λ. Accept the change with a probability that depends on the data fit with new and old λ values
   - Compute the posterior distribution of θ for the new hyperparameters and μ^i be mean of this distribution, while the new estimate θ^i be drawn from this distribution.
3. The resulting estimate will be $\hat{\theta} = \frac{1}{M}\sum_{i=1}^{M} \mu^i$
4. The sequence $\{\theta^i\}$ is the sampled posterior density $p(\theta|Y)$

Example: Simulated Models

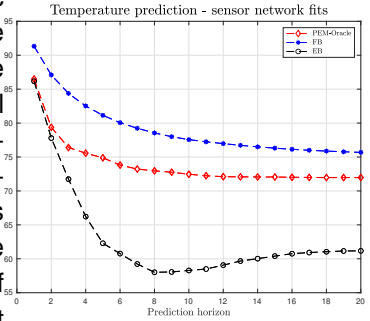
Generate randomly a system with 5 inputs. Estimate models with three different methods and compute and compare m-step predictions (1000 systems generated)

1. PEM-Oracle (red): The standard, "classical" method. Try ARMAX models of different orders. Select the order that works best (with an *oracle* – ideal) $\sim 10sec$
2. EB: (black) Empirical Bayes $< 1sec$.
3. FB:(blue) Full Bayes (using MCMC),$\sim 10sec$



Real Experiment: Temperature Sensor Network

A building is covered by a network of 24 thermodynamic sensors(temperature, relative humidity, and similar). One sensor is selected to model the temparature from the measurements in the 23 other sensors. An ARMAX structure is set up with 23 inputs and one output, and the parameters of this are estimated in different ways.



Conclusions

The "Full Bayesian Approach" has a clear edge over "Empirical Bayes" and "classical methods" and could be seen as the preferred method to estimate linear black box models. No need to deal with order/stucture determination of ARX, ARMAX, Box-Jenkins models.

Full story in

G. Pillonetto and L. Ljung: Full Bayesian identification of linear dynamic systems using stable kernels. Proc National Academy of Sciences, PNAS, Vol 120 (18), 2023, e2218197120

Take it Easy — Keep on Working!

Lennart Ljung

Paul van der Hof Symposium: Full Bayes Eindhoven April 19 2024

AUTOMATIC CONTROL
REGLERTEKNIK
LINKÖPINGS UNIVERSITET