Bayesians methods in system identification: equivalences, differences, and misunderstandings

Johan Schoukens and Carl Edward Rasmussen

ERNSI 2017 Workshop on System Identification

Lyon, September 24-27, 2017

Outline

Introduction: personal story and goal of presentation

What is prior knowledge

Setting the ideas: Impulse Response example

Nature of the modeling problem

Two approaches: Cross validation vs Bayesian inference

Birds Eye diagram

Priors: a robust concept

Conclusions

Prior Knowledge

Data-driven model relies not only on data:

- additional bits can take many forms: knowledge in the field, users beliefs, experience, assumptions, etc

Example

$$y_0 = f(u) = \sum_{i}^{\infty} \theta_i \phi_i(u), \qquad C(\theta),$$

where $C(\theta)$ is some form of constraint; we will call it the prior.

Example: Impulse Response

Number of data points (4096) and excitation (almost) white noise, noise level

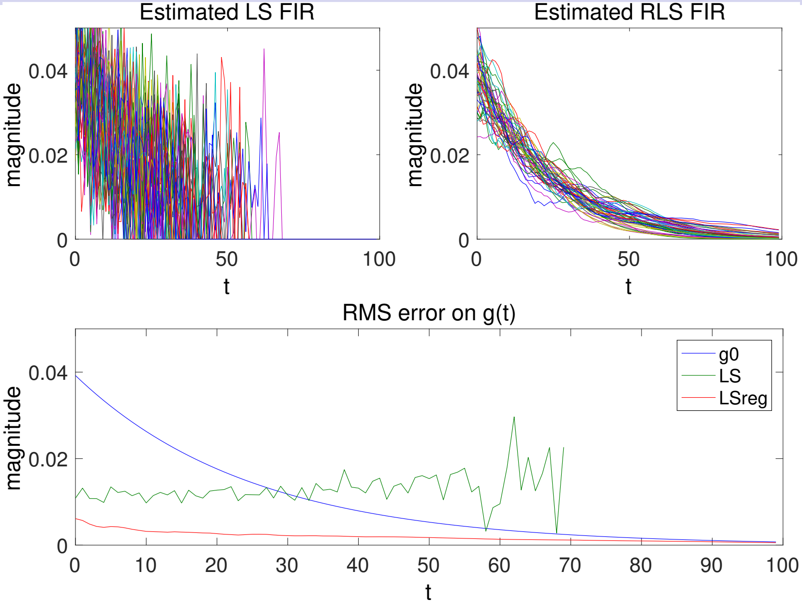
figures: $g(t)$ estimated by LS or by LS + reg

prior encodes: exponential decaying envelope, and smooth weights

second plot: RMS error as function of weigh delay (RMS values)

Akaike turns off coefficients above about $t = 40$.

Example: Impulse Response continued



The nature of the modeling problem

Prior information is often qualitative:

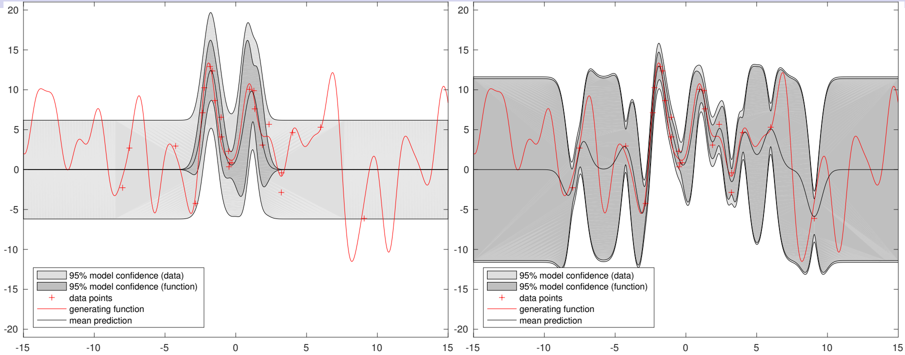
- stable
- smooth
- stationarity (invariances)
- positivity, monotonicity

The quantification (strength, or scale) of this information is inferred from the data through hyperparameters.

Diagram of methodologies

Maximum likelihood	Regularization framework	Bayes
Cost: negative log likelihood + discrete penalty	Cost: negative log likelihood + lambda times regulariser	Cost: negative log likelihood + negative log prior
Classical: SI + AIC, BIC, CV	Regularized SI	
Procedure: double optimisation parameters (continuous) model (discrete)	Procedure: two levels: parameters + hyperparameters	
	optimize: CV or marginal likelihood for hypers	optimize: marginal likelihood for hypers
	optimisation for parameters (condition on hypers)	parameter posterior
Results: parameter point estimate parameter covariance (data driven)	Results: parameter point estimate parameter covariance (data + regularizer)	Results: parameter posterior predictive distribution (MCMC or approx)

Regression comparison

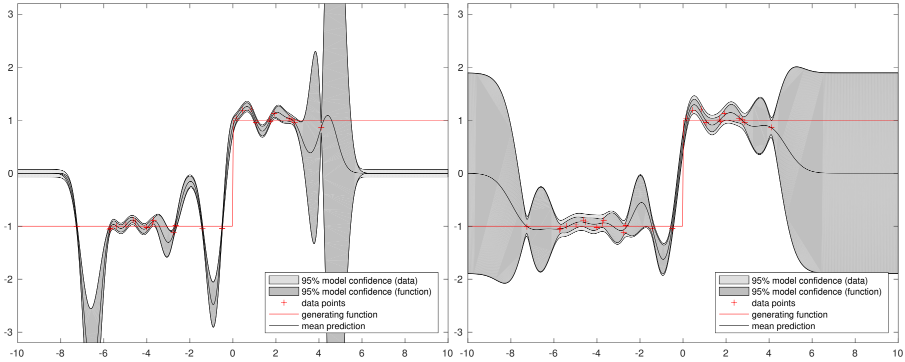


Comparing BIC with Gaussian process. Model uncertainty (95% confidence) in light grey, data uncertainty in dark grey.

BIC uses a small number of basis functions, leads to under-fitting and overconfidence.

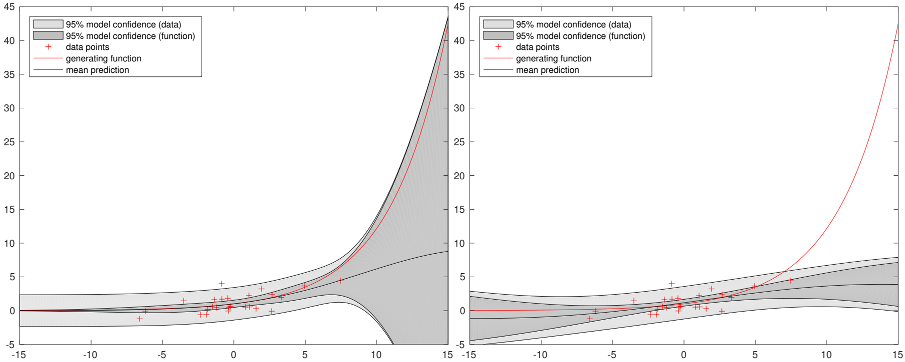
GP uses infinitely many basis functions.

Robustness of solutions to prior, step example



Comparing BIC with Gaussian process. Data from step function, prior for stationary function, not well matched.

Robustness of solutions to prior, exp example



Data from exp function, prior for stationary function.

BIC selects a single basis function, overconfident for neg arguments. GP extrapolates poorly.

Marginalize or optimise?

When optimising, overfitting becomes a problem. The optimiser will find solutions which agree well with the particular training set observed, but doesn't generalize well. This motivates regularisation and working with small models (Akaike, etc). Often external information (validation sets) are used to control complexity.

When marginalising, overfitting does not happen. Instead, in large models with vague priors the large uncertainties will remain; the predictive error-bars will be large. Internal measures (the marginal likelihood) will show the problem (no external information is required).

Unfortunately, whereas non-linear optimisation is hard, marginalisation is REALLY hard. Bayesian methods generally require 1) MCMC techniques for inference, or 2) specific model classes, such as Gaussian processes, or 3) analytic approximations (eg variational).

Marginalize or Optimize

Although the use of a regulariser and prior look very similar (sometime even identical expressions), in fact these a quite different: In optimisation only the properties of the regulariser around the optimum are important, but in Bayes the whole prior distribution is important. This fact is typically overlooked.

Marginalisation is mostly harder, and leads to a less convenient result, but may provide better uncertainty estimates.

The tricky thing may become understanding the prior distribution.

possible discussion points

priors can be useful for interpretation for generative models

prior specification is invariance to how much data will be available

Conclusions

Main message:

- data driven modeling uses more information than in the data
- Bayesian framework is systematic way of dealing with non-data information

so: max likelihood systematic treatment of noisy data, Bayes systematic treatment of noisy data and priors

The regularization framework may be interpreted as a Bayesian procedure in the mono-modal (Gaussian) case

Posterior interpretation of prior can help interpretation

Bayesian complexity control