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

Carl Edward Rasmussen and Johan Schoukens

Workshop on non-linear system identification Benchmarks

Brussels, April 24-26, 2017

Outline

Motivation: provide *insight* comparing *classical* with *Bayesian* method in system identification.

Four central questions:

- the goal of the modelling effort:
  input-output inference $\neq$ physical interpretation of the model
- marginalization versus optimization:
  what is the real difference between approaches?
- the role and interpretation of prior information:
  what is in a name?
- the impact of prior on model complexity control:
  equivalences/differences with classical AIC/MDL/etc.

Motivation

Many common tasks in SYSID are amenable to both *classical* and *Bayesian* treatments, yet puzzlingly, the dominant framework is mostly classical.

Many properties are common between the two frameworks, but differing terminology and misconceptions flourish.

In the following we will attempt to highlight similarities and differences, both in theory and practise.

The Goal of modeling

A spectrum of modeling tasks exists: from single (or multiple) physical parameters to entire (multivariate) functional relationships (of various degrees of mathematical abstraction. SYSID contains them all, black box, grey box, etc, e.g. step response functions.

Possible goals:

- *good enough* model, to elucidate particular properties
- parameter estimation vs prediction (parametric vs non-parametric)
- predictive models, one-step ahead or simulation
- at the complex end of the spectrum, it may be important to quantify the likely quality/reliability of the model (e.g. error-bars)

The goal of modelling; some typical differences

Classical	Bayesian
Provide a single simplified model, which is good enough (for some purpose).	Provide an explanation (generative model) of how it can be that the data is the way it is (irrespective of purpose).
Often the focus is on parameter estimation; the structure is either dictated by physics or assumed (approximated) known. The parameters themselves may be of interest.	Often flexible non-parametric models are used to avoid strong assumptions. Predictions from the model are of prime importance, usually (nuissance-) parameters themselves less so.
A single model is required.	A distribution over models (which reflects predictive error-bars) is required.

Preliminaries: experimental test bed

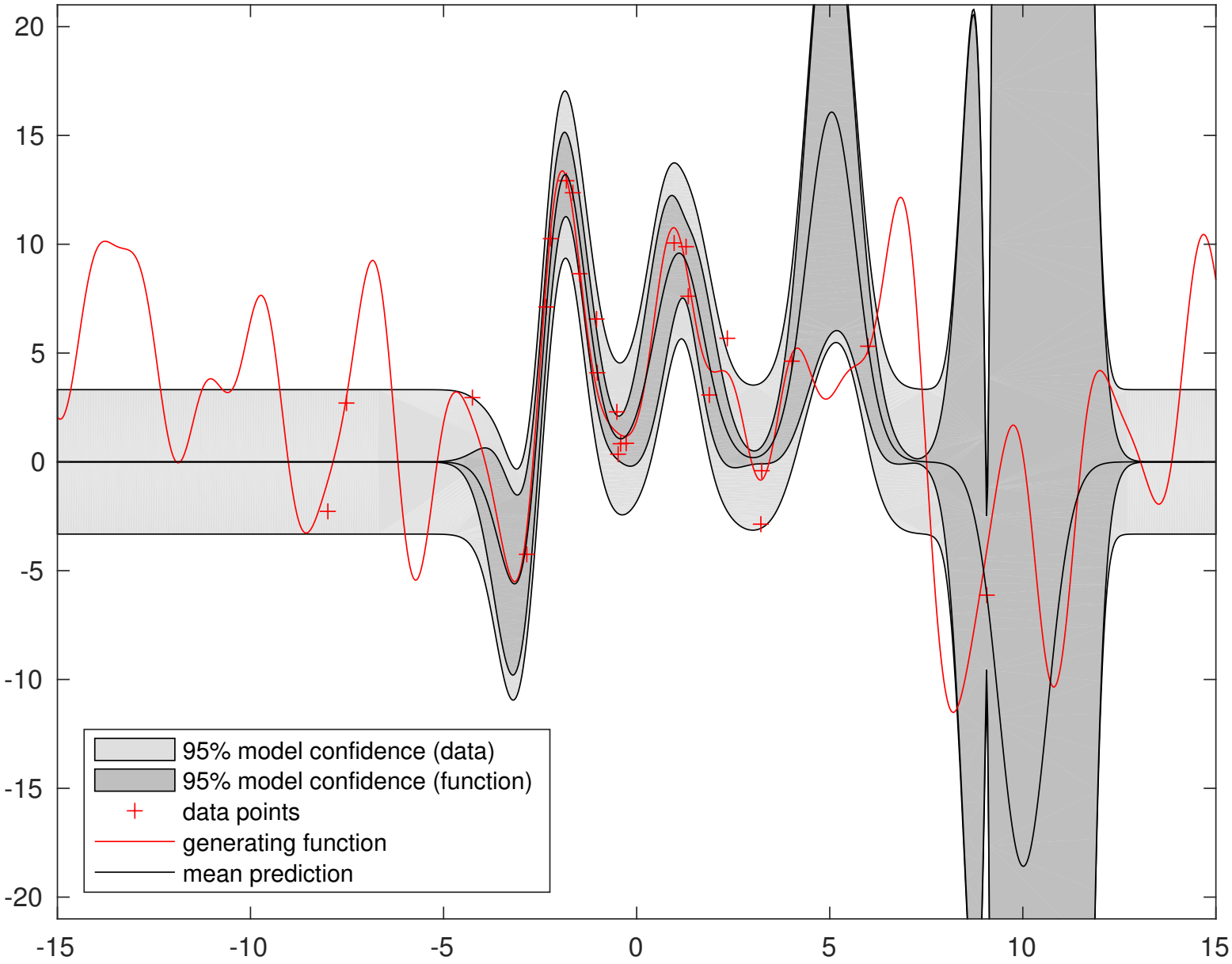
Important to anchor theoretical discussion in experimental results.

We provide a, somewhat anecdotal, 1 dimensional illustration. Supposed to illustrate properties of common practise, not highly specialised approach.

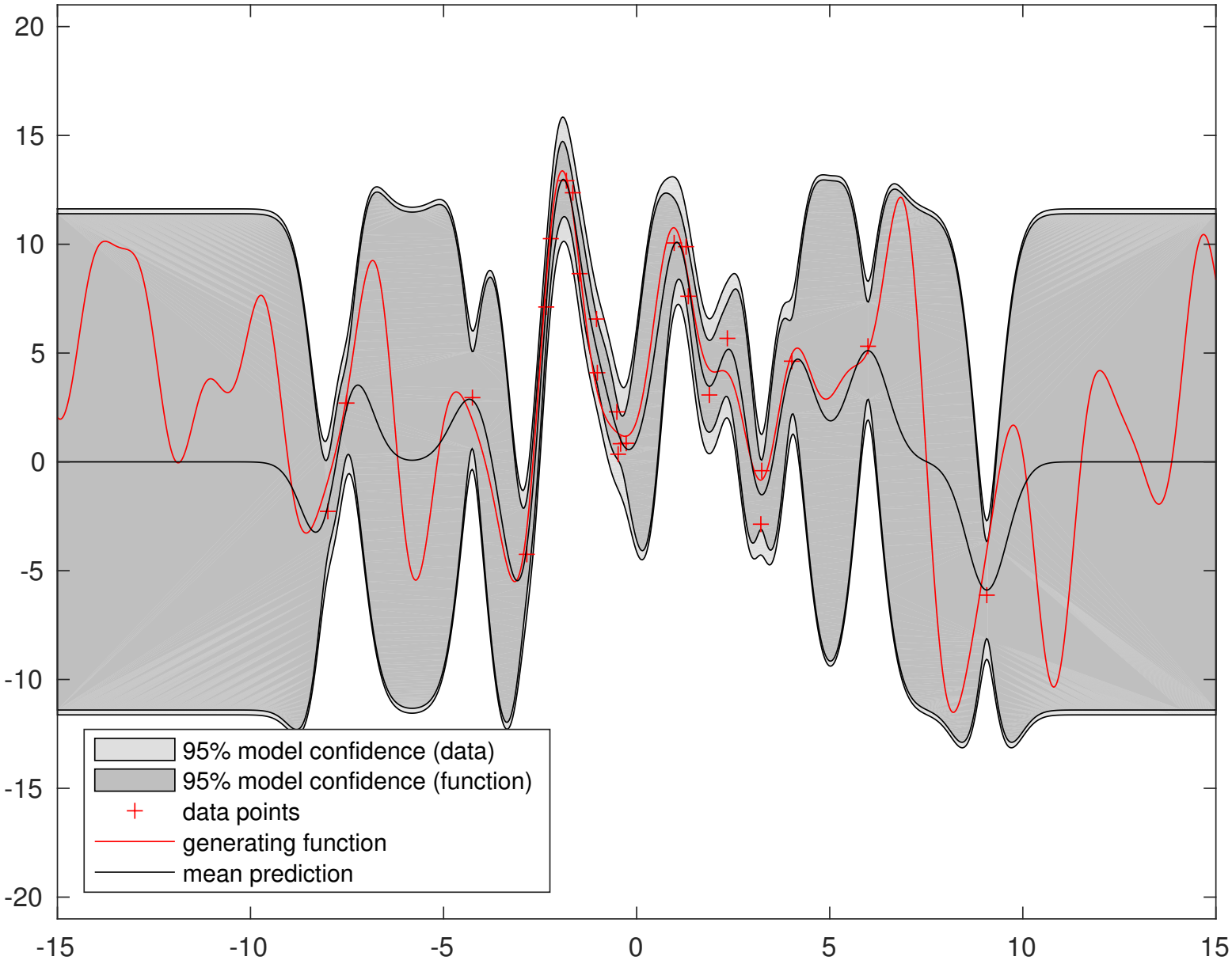
Classical: least squares: Gaussian basis functions with common width and individual means and amplitudes. Bayesian Information Criterion (BIC) to chose model order, Maximum Likelihood (least squares) to estimate parameters. 100 random restarts for each model order. Computation time, order 1 minute.

Bayes: Infinite number of Gaussian basis functions with common width. Training specifies 3 hyperparameters: prior on amplitudes, Gaussian widths, and noise level. One 3 dimensional optimisation of marginal likelihood, no restarts. Computation time, order 1 second.

Example: Gaussian basis functions, Least Squares, BIC



Example: Gaussian basis functions, Bayes



Quantitative comparison

test set mean squared error:
Least squares: 14.0
Bayes: 6.4

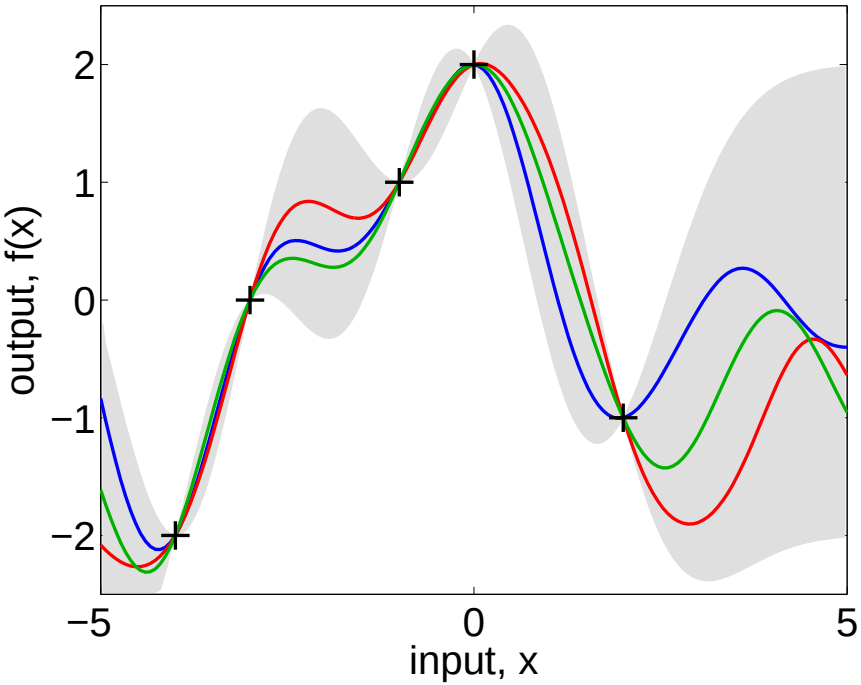
test set log probability:
Least squares: -3.1
Bayes: -2.1

This means that on average Bayes assigns $\exp(-2.1 + 3.1) = 2.7$ times higher density to the test points.

One difference rule them all

Both frameworks start identically by defining a restricted model or model class.

Fitting or *estimation* or *inference* is concerned with assigning structural and numerical values to elements in the model.



Both frameworks face the identical issue, that the data doesn't uniquely determine all elements.

The central difference between the approaches is that the classical approach attempts to find a best model (in some sense), whereas the Bayesian approach averages over possible models.

All the differences between the two approaches are caused by this choice.

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.

The role of the prior

Consider a model

$$f(\theta, x) = \sum_{j=1}^{J} \alpha_j \Phi_{\mu_j}(x) \tag{1}$$

- Flexibility grows with J
- J can become arbitrarily large
- A mathematical rule is needed to restrict the flexibility, for example

$$\sum_{j=1}^{J} \alpha_j^2 \leqslant c \tag{2}$$

- Use additional insight in the problem to tune the mathematical rule
- Prior distribution $P(\alpha)$: a structured and systematic representation of the mathematical rule
- the 'Prior' reflects what user is willing to assume about the model
- Equation (1) + prior $P(\alpha)$ defines the model class
- Prior is a user belief that adds information to the modelling problem that is not present in the data

Goal regularization: tune the bias/variance balance s.t. MSE is minimized

- Basic idea: pull parameters to zero

Simple scalar example

- True parameter $\theta_0 = 1$, unbiased estimate $\hat{\theta}$ with variance $\sigma^2 = 1$
- Scaled estimator

$$\tilde{\theta} = \lambda\hat{\theta} \tag{3}$$

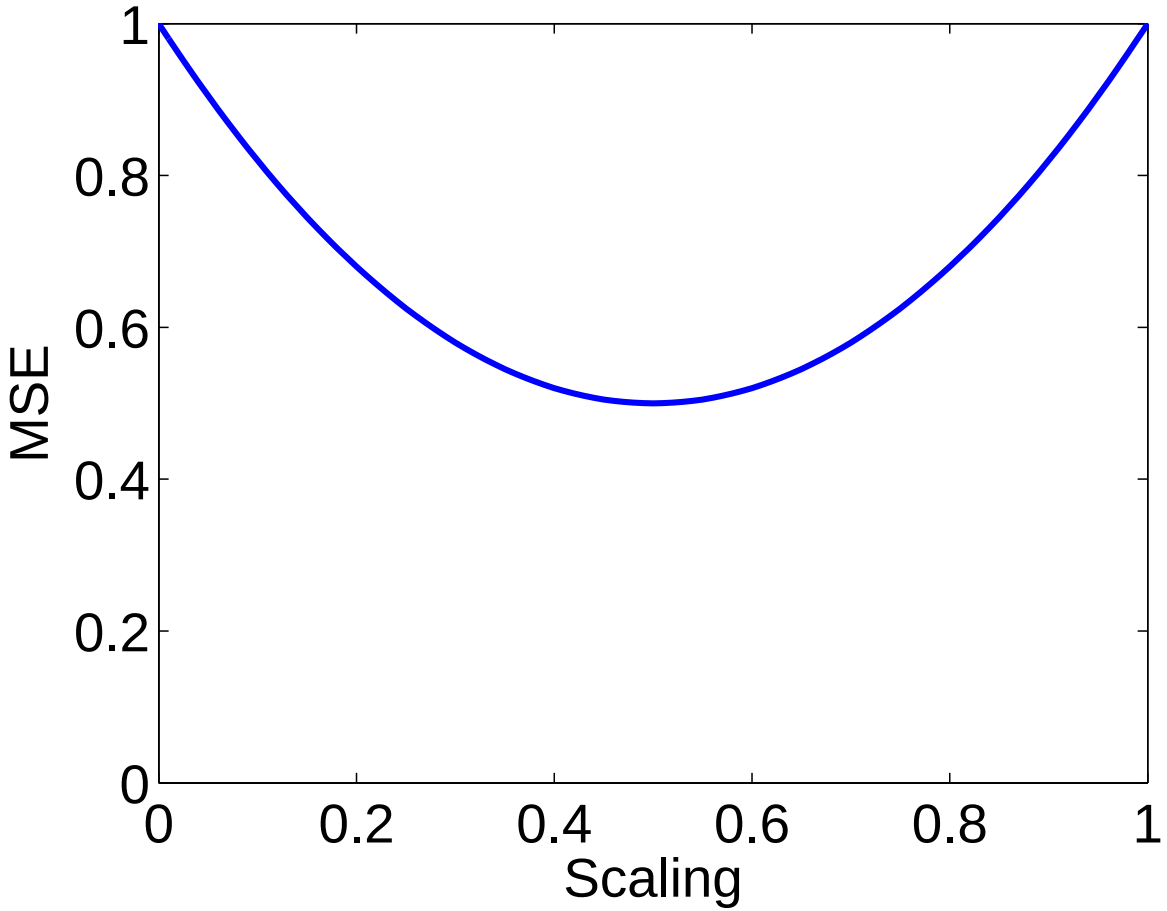
- Bias $\tilde{\theta}$ is $b = (1 - \lambda)$, variance $\tilde{\theta}$ is $\sigma^2_{\tilde{\theta}} = \lambda^2$
- MSE

$$e^2 = (1 - \lambda)^2 + \lambda^2 \tag{4}$$

- Implementation: modify costfunction

$$V_{regulized}(\theta) = V(\theta) + \lambda\theta^2 \tag{5}$$

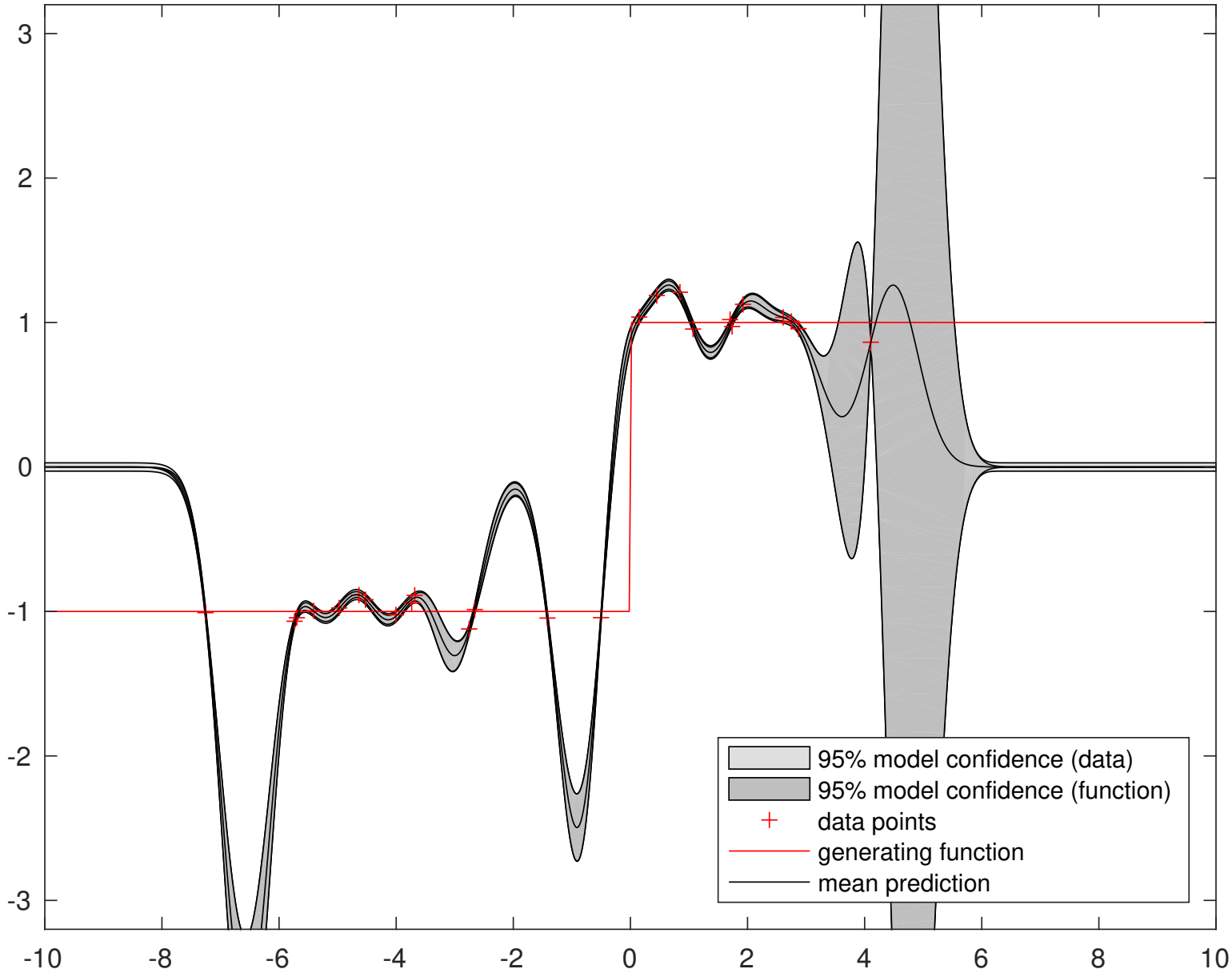
- λ is hyper parameter that needs to be tuned



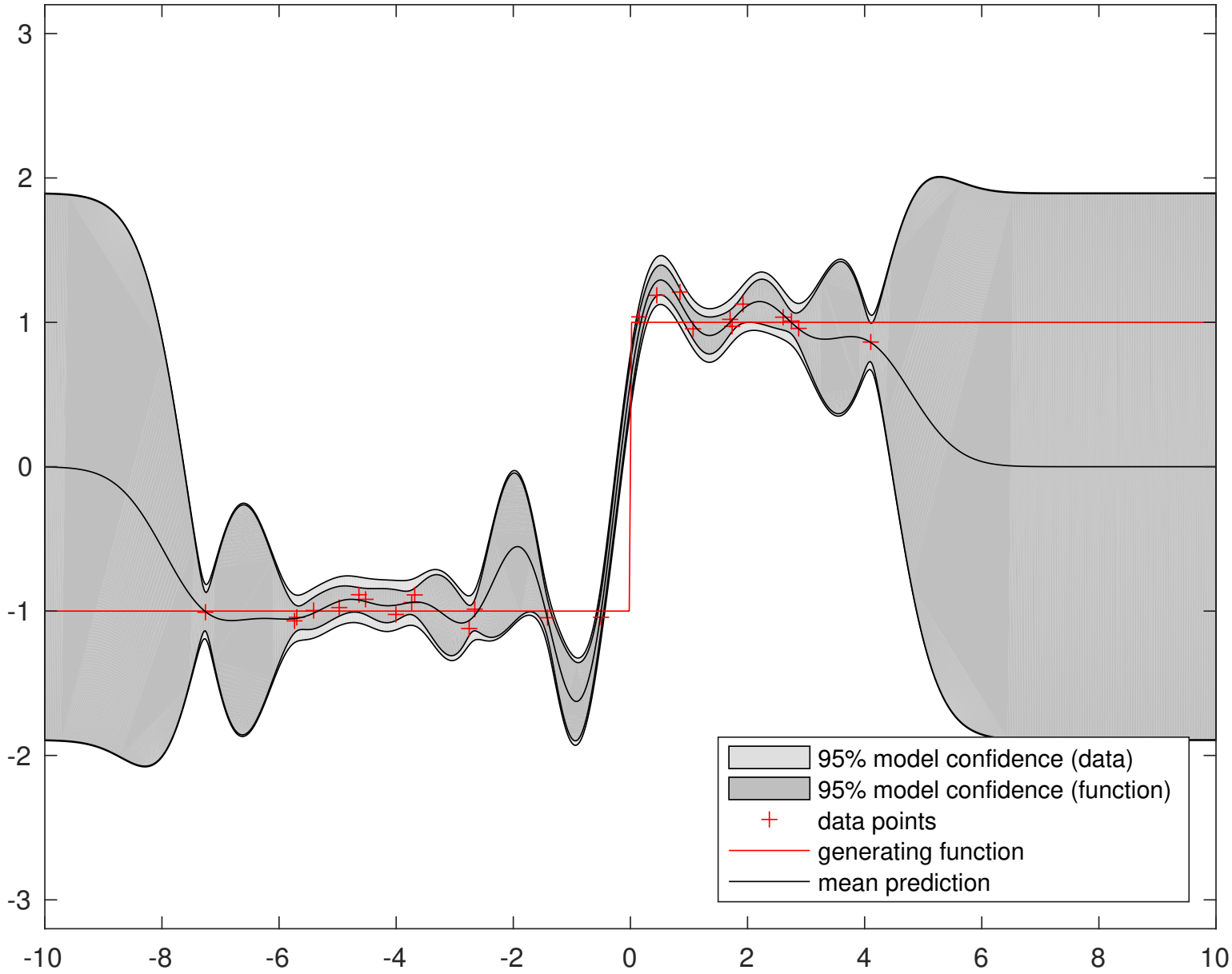
- Example Gaussian Prior: $\theta \sim N(0, R^{\{-1\}})$

$$V_{regulized}(\theta) = V(\theta) + \lambda \theta^t R \theta \tag{6}$$

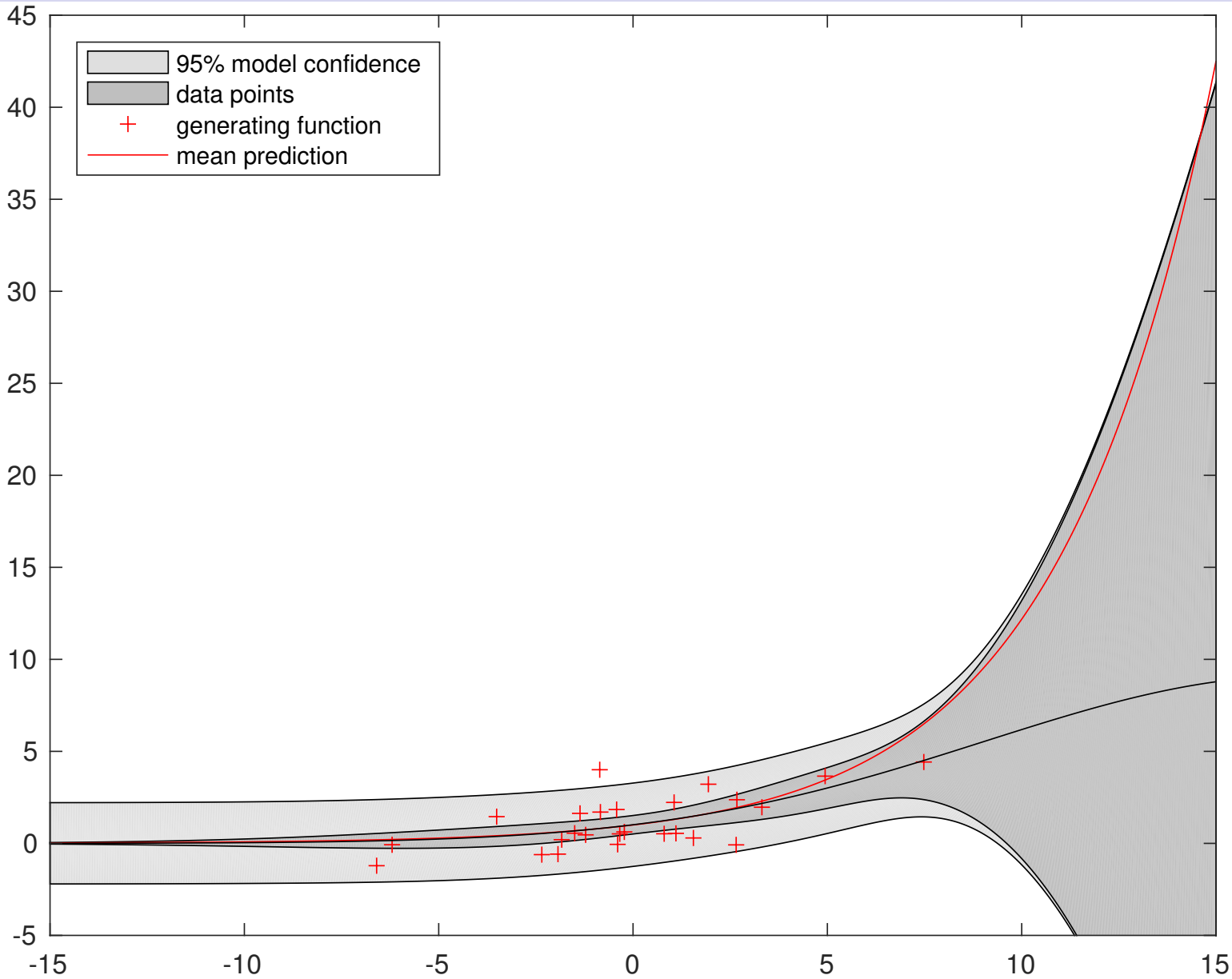
Example: Gaussian basis functions, Least Squares, BIC



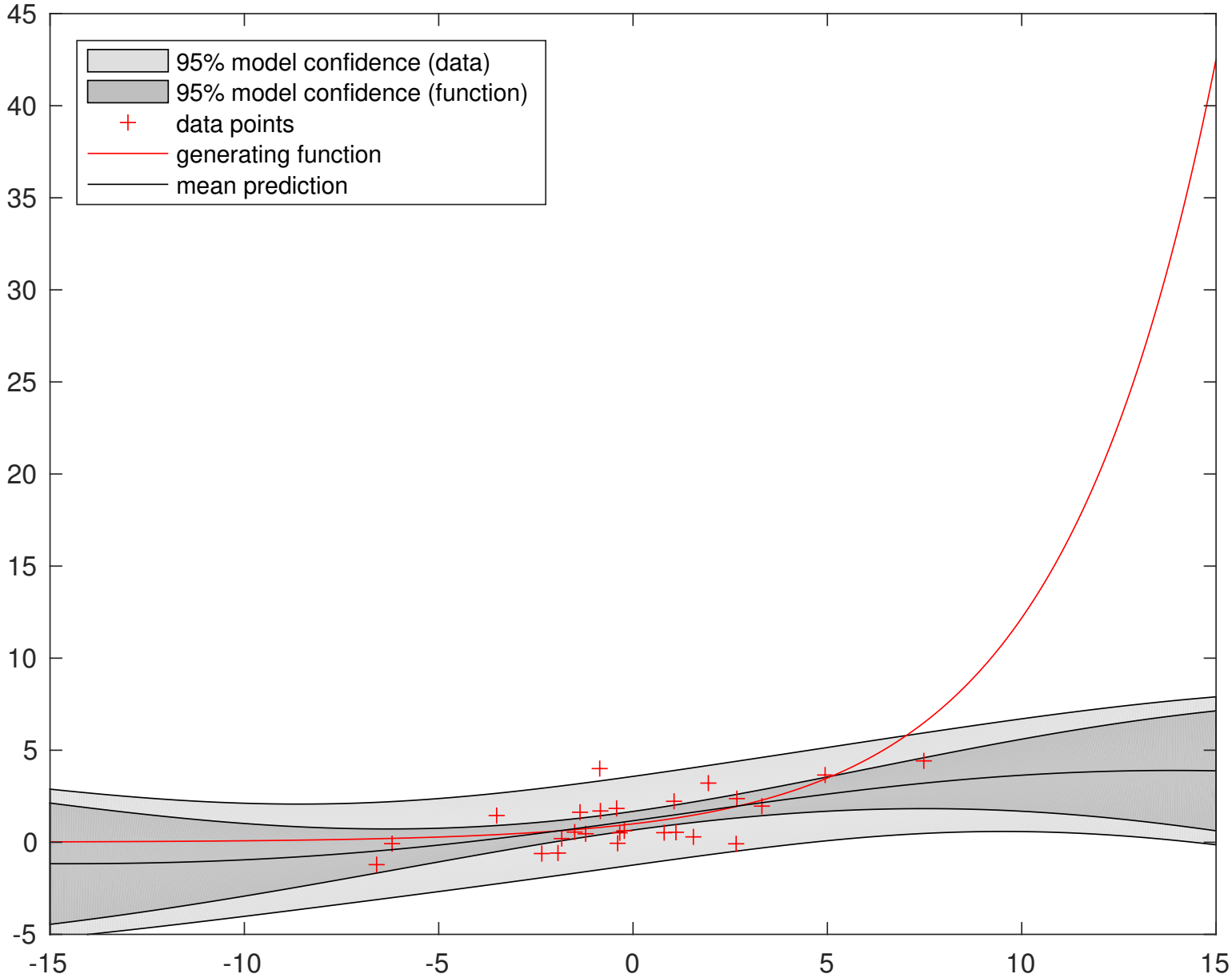
Example: Gaussian basis functions, Bayes



Example: Gaussian basis functions, Least Squares, BIC



Example: Gaussian basis functions, Bayes



Quantitative comparison

Step function example: test set mean squared error:
Least squares: 0.51
Bayes: 0.20

test set log probability:
Least squares: −285
Bayes: −2.9

Exponential example: test set mean squared error:
Least squares: 1.3
Bayes: 1.7

test set log probability:
Least squares: −1.47
Bayes: −1.63

Robustness of prior

- The better the match between prior and reality, the better the model
- The choice of the prior is not critical
- Example: comparison LS and BIC and BI with prior

Conclusion: BI is robust with respect to the choice of the prior

more discussion about the prior

The prior is often criticised, for being arbitrary, subjective or not available in practice. In fact it just encodes assumptions, which of course is present in the classical approach too.

People often ask 'what is the *right* prior'? This is a misguided question. The right prior would just be a delta function on the actual generating process.

Another misguided idea is the 'non-informative' prior.

In practice the prior reflects what the modeller is willing to assume about the data. It will hopefully contain some models that are good explanations of the data, but also many other ones;

Typically, hierarchical priors are used, which can state eg that smoothness is assumed, but the exact amount ot smoothness is not specified (hyperparameter to be inferred).

Prior as a tool to control model complexity

Classical SI controls scans over models with different complexity, and selects the *best* one using either

- Validation set
- Extended cost function with model complexity term that predicts Validation cost: AIC, BIC, MDL

Regularized SI allows for the use of too complex models

- relies on the regularized cost function to reduce the impact of over-fitting
- a hyper parameter balances the data information and the regularization information

Bayesian Inference does not select a specific model

- averages over the whole set of models (marginalization)
- relies on the Prior to reduce impact of too complex models: these become very unlikely
- a hyper parameter balances the data information and the Prior information

The impact of the prior on complexity control

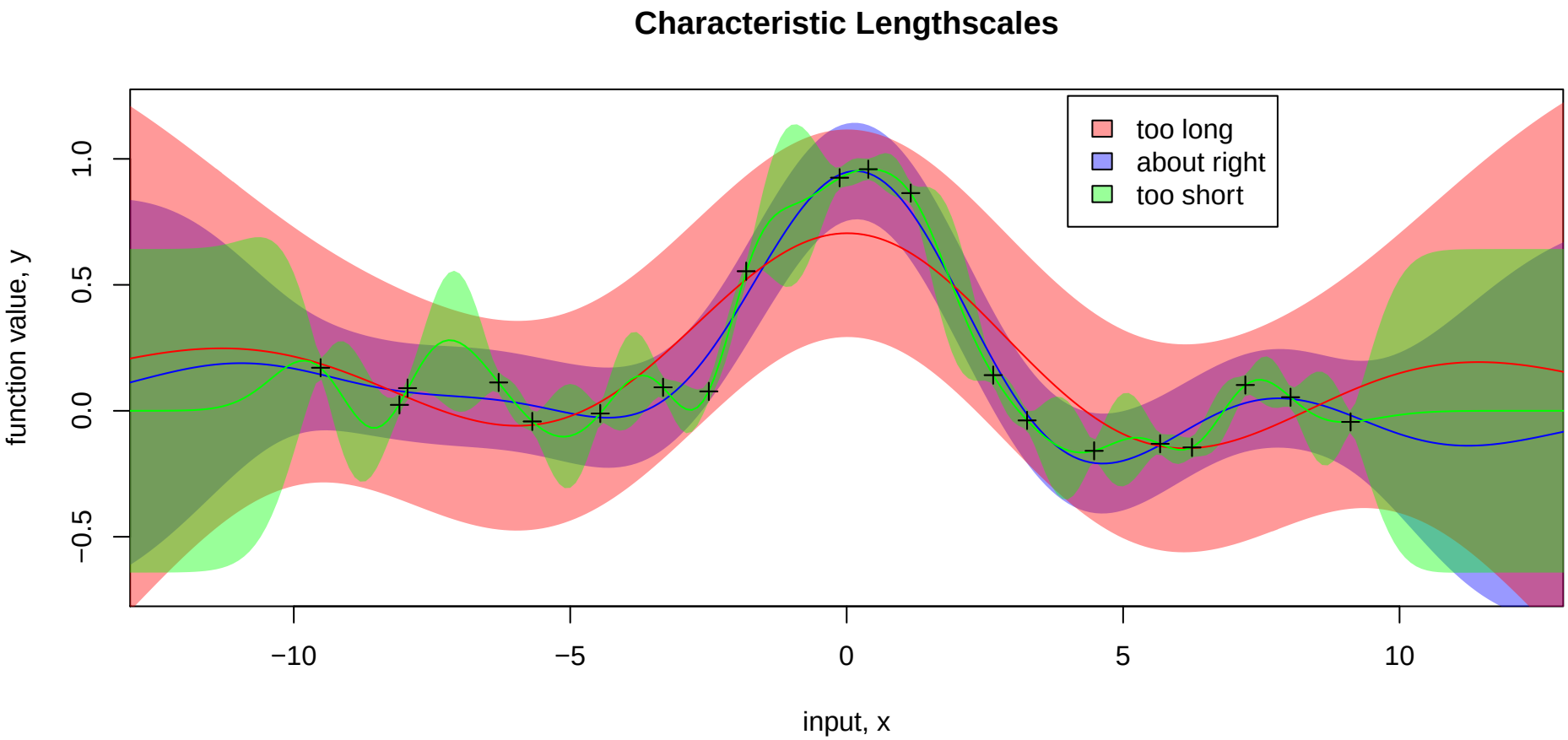


Illustration of Bayesian learning with infinite numbers of Gaussian basis functions. The predictive mean and error-bars are shown. In the red the Gaussian are wide, in blue intermediate, and in green narrow.

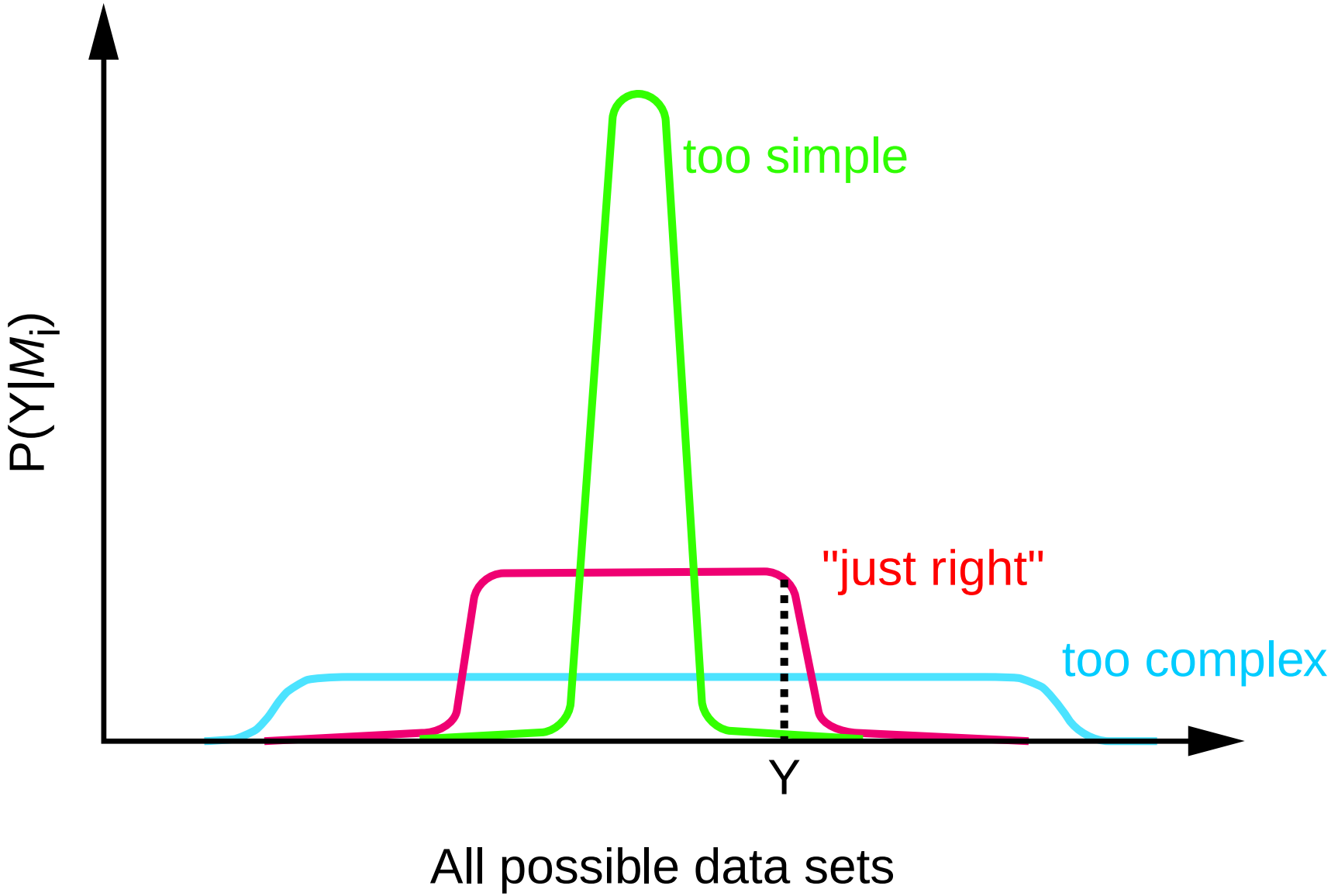
The red model is under-fitting. The inferred function is over-smoothed, and the model has turned up the noise level in order to be able to explain the observed data points. The marginal likelihood is poor, because of the poor data fit.

In green the basis functions are narrow. The mean function is close to every data point, but the predictive error-bars shoot up away from the data. The marginal likelihood for this model is poor because of the (automatic) Occam's razor.

The blue model represents are good balance, and is the one selected by maximising the marginal likelihood.

The hierarchical specification: expect smoothness, but not sure *how much* is extremely powerful.

Bayesian complexity control



Conclusions

Typical goals are somewhat overlapping and somewhat distinct.

Modelling philosophy is very different:

- classical: optimize to a *best fit*, use restriction on model order and regularization to avoid overfitting (requires a validation set).
- Bayes: marginalize (don't optimize). Don't restrict model order. Encode knowledge (weak or strong) in hierarchical prior. No need for validation set (all properties calculated on training set.

Even if the (log) prior and rgularizer have the same form, predictions are still different. There is no such thing are a right/wrong prior or regularizer.

Marginalization is typically computationally hard. GPs are an exception.