Probabilistic Machine Learning

Probability, Bayesian Inference & Parsimonious Models

Mike Tipping

July 17, 2023

Department of
Computer Science

UNIVERSITY OF
BATH

Overview

- Some random thoughts on probability and uncertainty
- … and inference and decision-making
- Bayesian inference
  - a brief introduction
  - the “Seven Pillars of Bayesian Wisdom”
- Demo (sparse approximation of an image)
- Bayesian probabilistic ML for parsimonious models
- Real-world application examples
- Return to the image approximation demo
- Round-up and pointers to forthcoming attractions

“Nature Cannot Be Fooled” (Richard Feynman)

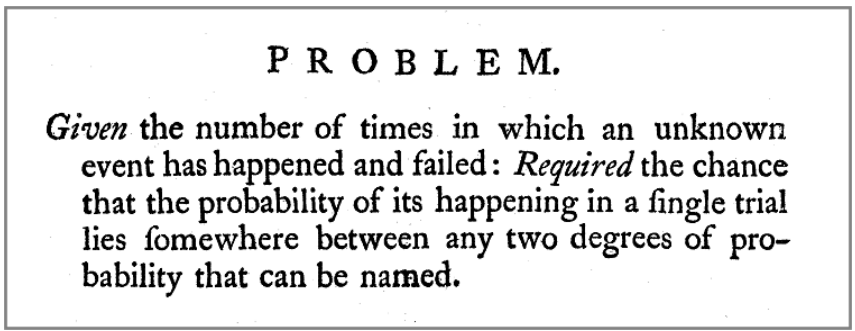
“A logarithmic plot suggested a straight line, so it was supposed that the erosion varied as the .58 power of the heat, the .58 being determined by a nearest fit. At any rate, adjusting some other numbers, it was determined that the model agreed with the erosion (to depth of one-third the radius of the ring).
There is nothing much so wrong with this as believing the answer!
Uncertainties appear everywhere …

When using a mathematical model, careful attention must be given to uncertainties in the model.”

Richard Feynman *— excerpt from Appendix F of the Report of the Presidential Commission on the Space Shuttle Challenger Accident (1986).*

Dawn of Probabilistic Inference

Thomas Bayes' original question, published in 1763:



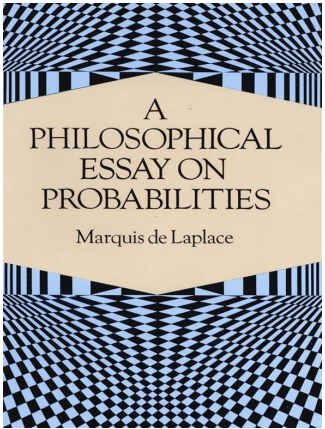
PROBLEM.

Given the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a ſingle trial lies ſomewhere between any two degrees of probability that can be named.

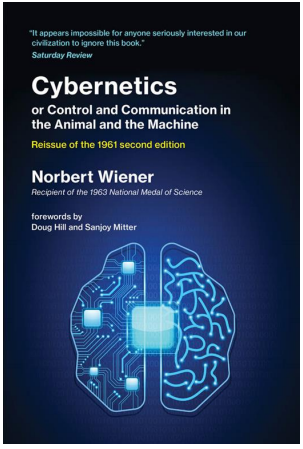
“Common Sense Reduced to Calculus”

“It is seen in this essay that the theory of probabilities is at bottom only common sense reduced to calculus; it makes us appreciate with exactitude that which exact minds feel by a sort of instinct without being able oft times to give a reason for it. … we shall see that there is no science more worthy of our meditations, and that **no more useful one could be incorporated in the system of public instruction.**”

Pierre-Simon Laplace *— from the closing paragraph of “A Philosophical Essay on Probabilities” (1814)*



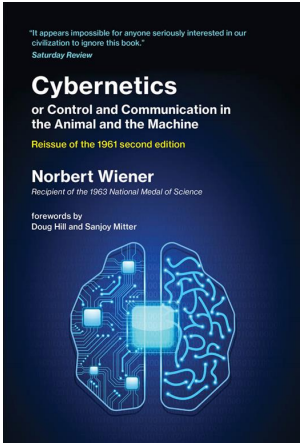
Probability in Machine Learning



> "Finally, as long as the automaton is running, its very rules of operation are susceptible to some change on the basis of the data which have passed through its receptors in the past, and this is not unlike the process of learning."

Norbert Wiener — *Cybernetics (1948)*

Probability in Machine Learning

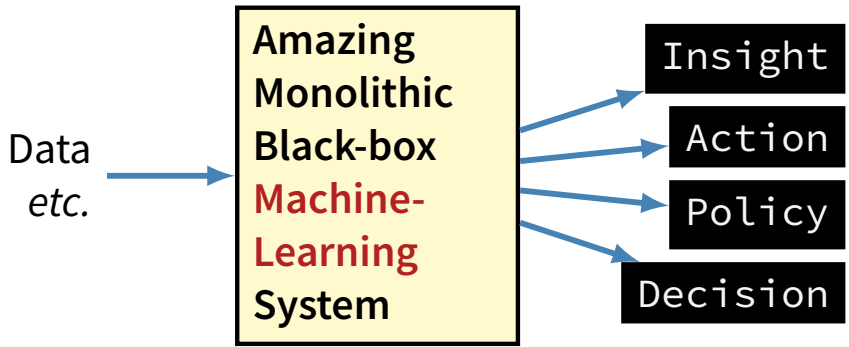


"I referred her to **Norbert Wiener, *Cybernetics*, 1948**. Of course, I wasn't referring to the robots depicted in illustrated papers, but to the lightning calculating machine, also known as the electronic brain … because **it has a greater power than the human brain to grasp information and assess its probability value**. Above all, however, the machine has no feelings, it feels no fear and no hope, which only disturb, it has no wishes with regard to the result, **it operates according to the pure logic of probability**."

Max Frisch — *excerpt from the novel "Homo Faber" (1957)*

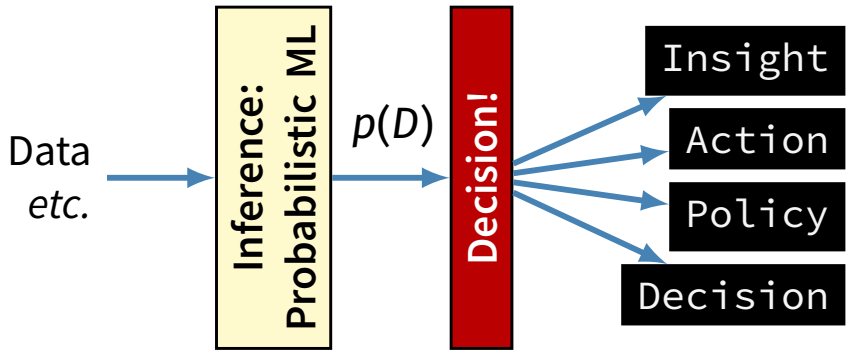
Separation of Inference and Decision

There could be trouble ahead …



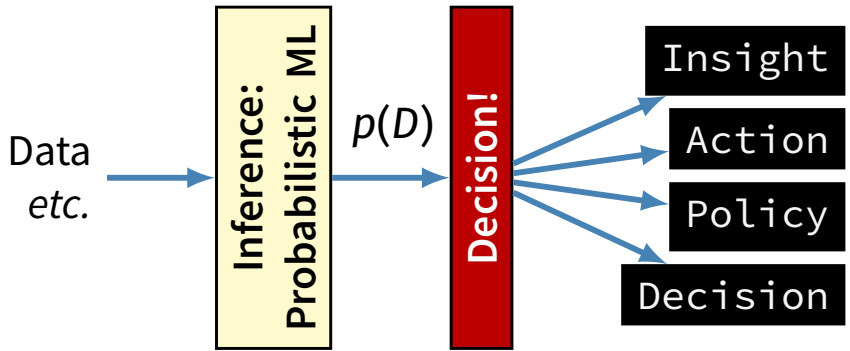
Separation of Inference and Decision

More appropriate ...



Separation of Inference and Decision

More appropriate … ?



"And, at least at the intuitive level, we have become rather good at this extended logic, and rather systematic. Before deciding what to do, our intuition organizes the preliminary reasoning into stages:
(I) Try to foresee all the possibilities that might arise;
(II) Judge how likely each is, based on everything you can see and all your past experience;
(III) In the light of this, judge what the probable consequences of various actions would be;
(IV) Now make your decision."

E. T. Jaynes — *from "Bayesian Methods: General Background" (1984)*

> "From the earliest times this process of plausible reasoning preceding decisions has been recognized. Herodotus, in about 500 BC, discusses the policy decisions of the Persian kings. **He notes that a decision was wise, even though it led to disastrous consequences, if the evidence at hand indicated it as the best one to make**; and that a decision was foolish, even though it led to the happiest possible consequences, if it was unreasonable to expect those consequences."
>
> ***E. T. Jaynes*** *— from "Bayesian Methods: General Background" (1984)*

> "I may be wrong. It is at best but a guess, and the world attaches wisdom to him that guesses right."

Horatio Nelson — *extract from a letter to Sir A. J. Ball (1804)*

"I tried to do the latter and I failed. But I don't admit that **my failure proved my view to be a wrong one, or that my success would have made it a right one**; though that's how we appraise such attempts nowadays — I mean, not by their essential soundness, but by their accidental outcomes."

Thomas Hardy *— from the novel "Jude the Obscure" (1895)*

Bayesian Probability — What and Why?

What is it?

- "Bayesian" implies that the rules of probability may be applied to model ***all sources of uncertainty***
- By contrast, in **orthodox** or **frequentist** statistics, only intrinsically **random variables** are treated probabilistically
- For example, probability can represent the **degree of belief** in, or **plausibility** of, a wide range unknowns, such as:
  - the outcome of a flip of a coin
  - whether it will rain tomorrow
  - the values a model parameter might take
  - the choice of model itself
  - the value of a measured physical constant (*e.g.* the mass of the moon)

The contrasting view of probability is more than a philosophical distinction!

Upside

The Bayesian framework offers a range of highly advantageous features when undertaking machine learning (see shortly)

Downside

But there is a major drawback: many required calculations are problematic to undertake and must be approximated (see later, under “voodoo”)

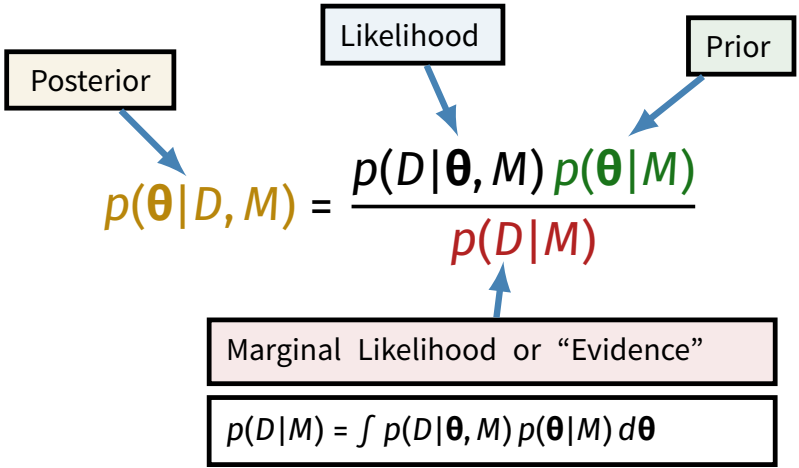
One Historical View on Bayesian Statistical Inference

"The subject is difficult. Some argue that this is a reason for not using it. But it is always harder to adhere to a strict moral code than to indulge in loose living … **Every statistician would be a Bayesian** if he took the trouble to read the literature thoroughly and was honest enough to admit that he might have been wrong."

Dennis V. Lindley *— comment on "Why Isn't Everyone a Bayesian?" (Bradley Efron) in The American Statistician, No. 40, 1986.*

Bayesian Inference in One Slide

If D is some observed data, and $\boldsymbol{\theta}$ encapsulates all the unknowns associated with a model M intended to describe that data, then **Bayesian inference** is:



Seven Pillars of Bayesian Wisdom

Generic

0. Fully Probabilistic Predictions

Priors & Posteriors

1. A Consistent Approach to Modelling All Uncertainty
2. Natural Adaptation to “Big” and “Small” Data
3. Intrinsic Handling of Streaming Data
4. Desired Properties are Incorporated in a Principled Way

Marginalisation

5. Predictions are Qualified and Informative
6. Irrelevant Variables are Factored Out
7. Implicit Implementation of Ockham’s Razor

Seven Pillars of Bayesian Wisdom

Generic

0. Fully Probabilistic Predictions

Priors & Posteriors

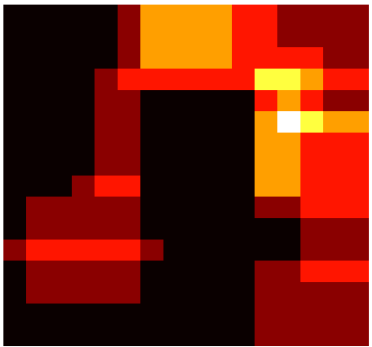
1. A Consistent Approach to Modelling All Uncertainty
2. Natural Adaptation to “Big” and “Small” Data
3. Intrinsic Handling of Streaming Data
4. Desired Properties are Incorporated in a Principled Way

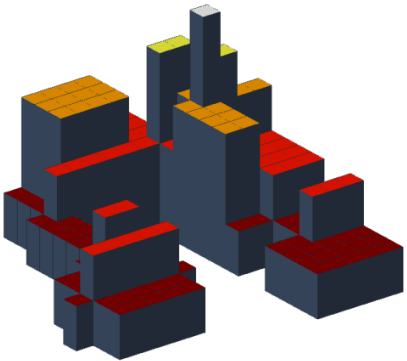
Marginalisation

5. Predictions are Qualified and Informative
6. Irrelevant Variables are Factored Out
7. Implicit Implementation of Ockham’s Razor

An Artificial Example: Efficient Image Decomposition (1)

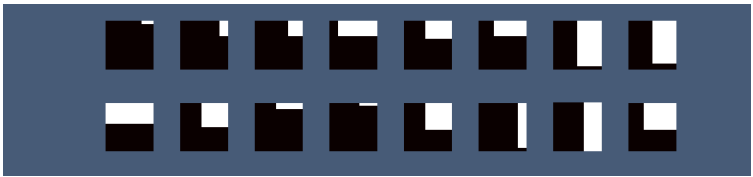
Consider modelling 16 × 16 images synthesised by noiselessly superposing randomly sized and located rectangular “blocks”:



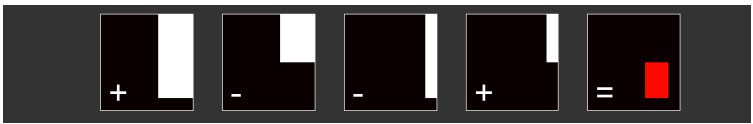


An Artificial Example: Efficient Image Decomposition (2)

- Model the image with a 256-element “integral” basis:



- Any “block” can be perfectly modelled by four appropriately-weighted integral functions:

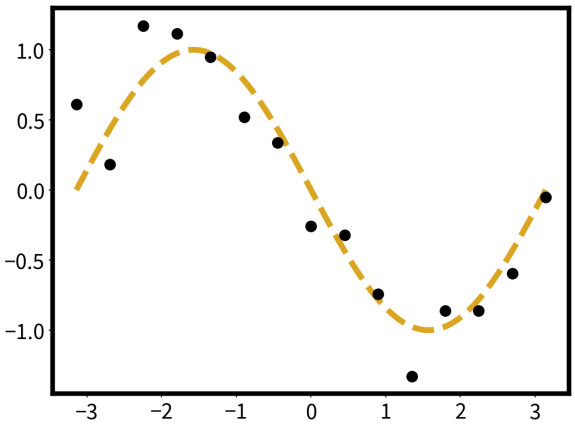


“Building block” (Image Approximation) Demo

- The demo compares an efficient sequential algorithm: **Order-Recursive Matching Pursuit** (ORMP) with an alternative Bayesian approach …

Bayesian Linear Regression

Example data set $\mathbf{t} = (t_1, \ldots, t_N)$ of $N = 15$ points generated from the function $f(x) = \sin(x)$ with added Gaussian noise of $\sigma = 0.25$



- Functional model is a linearly-weighted sum of M fixed basis functions:

$$\hat{f}(x; \mathbf{w}) = \sum_{m=1}^{M} w_m \phi_m(x)$$

- Basis functions are Gaussian (RBF), data-centred, so $M = 15$:

$$\phi_m(x) = \exp\left\{-(x - x_m)^2 / r^2\right\}$$

The Bayesian Linear Regression Framework

- **Likelihood model** (Gaussian, as usual):

$$p(\mathbf{t}|\mathbf{w}, \sigma^2) = (2\pi\sigma^2)^{-N/2} \prod_{n=1}^{N} \exp\left\{-\frac{\left[t_n - \hat{f}(x_n; \mathbf{w})\right]^2}{2\sigma^2}\right\}$$

- **Prior** (Gaussian, conjugate, incorporating **hyper-parameter** α):

$$p(\mathbf{w}|\alpha) = \prod_{m=1}^{M} (2\pi)^{-1/2} \alpha^{1/2} \exp\left\{-\frac{\alpha}{2} w_m^2\right\}$$

- **Hyper-priors** over α and σ^2:
  - uniform over a **logarithmic** scale, or,
  - Gamma(a, b), where a and b are fixed small values

Strictly Bayesian: the Ideal Framework

- We should proceed by computing the joint posterior over *all* unknowns, which from Bayes' rule, is:

$$p(\mathbf{w}, \alpha, \sigma^2 | \mathbf{t}) = \frac{p(\mathbf{t}, \mathbf{w}, \alpha, \sigma^2)}{p(\mathbf{t})} = \frac{p(\mathbf{t}|\mathbf{w}, \sigma^2)\, p(\mathbf{w}|\alpha)\, p(\alpha)\, p(\sigma^2)}{p(\mathbf{t})}$$

- The highlighted normalising factor is the "fully marginalised" likelihood:

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w}, \sigma^2)\, p(\mathbf{w}|\alpha)\, p(\alpha)\, p(\sigma^2)\, d\mathbf{w}\, d\alpha\, d\sigma^2$$

- It is ***almost always analytically intractable***!

Approximation & The Analytic Roadblock

Problem:

- We cannot compute $p(\mathbf{t})$ or $p(\mathbf{w}, \alpha, \sigma^2 | \mathbf{t})$ in closed form

Compromise "solution":

1. First perform any analytically computable integrations, then
2. **Approximate** remaining terms, perhaps by:
   - Stochastic techniques (*e.g.* Hamiltonian Monte Carlo sampling)
   - Variational techniques: *e.g.* $p(\mathbf{w}, \alpha, \sigma^2 | \mathbf{t}) \approx Q_{\mathbf{w}}(\mathbf{w})\, Q_{\alpha}(\alpha)\, Q_{\sigma^2}(\sigma^2)$
   - Type-II maximum likelihood
   - Laplace's method

Approximation is the "voodoo" of Bayesian machine learning

Approximation & The Analytic Roadblock

Problem:

- We cannot compute $p(\mathbf{t})$ or $p(\mathbf{w}, \alpha, \sigma^2 | \mathbf{t})$ in closed form

Compromise "solution":

1. First perform any analytically computable integrations, then
2. **Approximate** remaining terms, perhaps by:
   - Stochastic techniques (*e.g.* Hamiltonian Monte Carlo sampling)
   - Variational techniques: *e.g.* $p(\mathbf{w}, \alpha, \sigma^2 | \mathbf{t}) \approx Q_{\mathbf{w}}(\mathbf{w})\, Q_{\alpha}(\alpha)\, Q_{\sigma^2}(\sigma^2)$
   - Type-II maximum likelihood
   - Laplace's method

Approximation is the "voodoo" of Bayesian machine learning

Type-II Maximum Likelihood for Bayesian Regression

- The desired posterior, $p(\mathbf{w}, \alpha, \sigma^2 | \mathbf{t})$, can be written as:

$$p(\mathbf{w}, \alpha, \sigma^2 | \mathbf{t}) \equiv p(\mathbf{w} | \mathbf{t}, \alpha, \sigma^2)\, p(\alpha, \sigma^2 | \mathbf{t})$$

- First term is the weight posterior, derived from Bayes' rule:

$$p(\mathbf{w} | \mathbf{t}, \alpha, \sigma^2) = N(\mathbf{w} | \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

with

$$\boldsymbol{\Sigma} = \sigma^2 (\boldsymbol{\Phi}^\mathsf{T} \boldsymbol{\Phi} + \sigma^2 \alpha \mathbf{I})^{-1}$$

$$\boldsymbol{\mu} = \boldsymbol{\Sigma} \boldsymbol{\Phi}^\mathsf{T} \mathbf{t} / \sigma^2$$

Type-II Maximum Likelihood

- Second term $p(\alpha, \sigma^2 | \mathbf{t})$ is intractable: we will approximate it (very coarsely!) with a δ-function at its mode
- We find **most probable** values $\hat{\alpha}_{MP}$ and $\hat{\sigma}^2_{MP}$ which maximise:

$$p(\alpha, \sigma^2 | \mathbf{t}) = \frac{p(\mathbf{t} | \alpha, \sigma^2)\, p(\alpha)\, p(\sigma^2)}{p(\mathbf{t})}$$

- If we assume flat **uninformative priors** over $\log \alpha$ and $\log \sigma$, then we equivalently maximise $p(\mathbf{t} | \alpha, \sigma^2)$
- $p(\mathbf{t} | \alpha, \sigma^2)$ is the **marginal likelihood** (or "evidence")
- This procedure is known as **Type-II maximum likelihood**

The Marginal Likelihood (1)

- To find $\hat{\alpha}_{MP}$ and $\hat{\sigma}^2_{MP}$ we maximise:

$$p(\mathbf{t}|\alpha, \sigma^2) = \int p(\mathbf{t}|\mathbf{w}, \sigma^2)\, p(\mathbf{w}|\alpha)\, d\mathbf{w}$$
$$= (2\pi)^{-\frac{N}{2}} |\sigma^2\mathbf{I} + \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^{\mathsf{T}}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{t}^{\mathsf{T}}(\sigma^2\mathbf{I} + \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^{\mathsf{T}})^{-1}\mathbf{t}\right\}$$

- This is a ***zero-mean Gaussian*** distribution over the single N-dimensional dataset vector $\mathbf{t}$ with covariance matrix:

$$\sigma^2\mathbf{I} + \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^{\mathsf{T}}$$

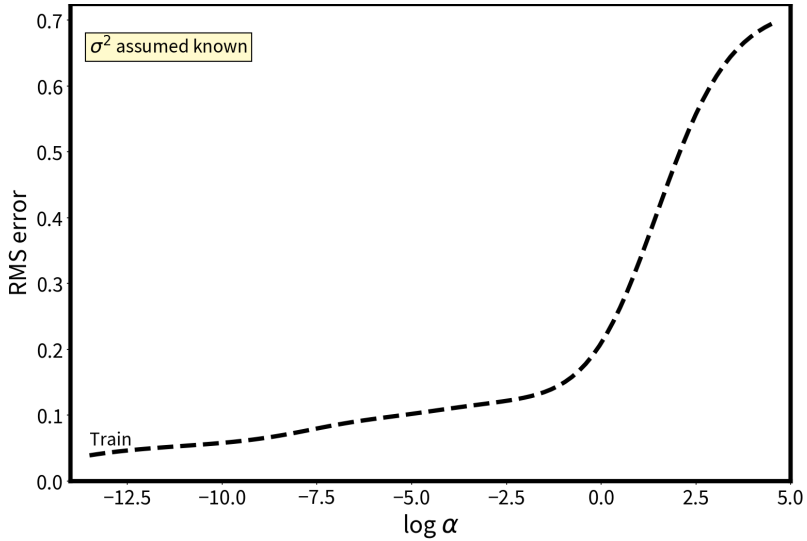
- It gives us a probability for the ***entire data set***, conditional on α and σ^2

The Marginal Likelihood (2)

- Computing the marginal likelihood is straightforward
  - it's just a Gaussian p.d.f., albeit potentially high-dimensional
- Given that, we can evaluate $p(\mathbf{t}|\alpha, \sigma^2)$ over a range of values of α and σ^2
- Then make our predictions using the **most probable** values $\hat{\alpha}_{MP}$ and $\hat{\sigma}^2_{MP}$
- This is not "strictly morally" Bayesian, but can often be useful in practice

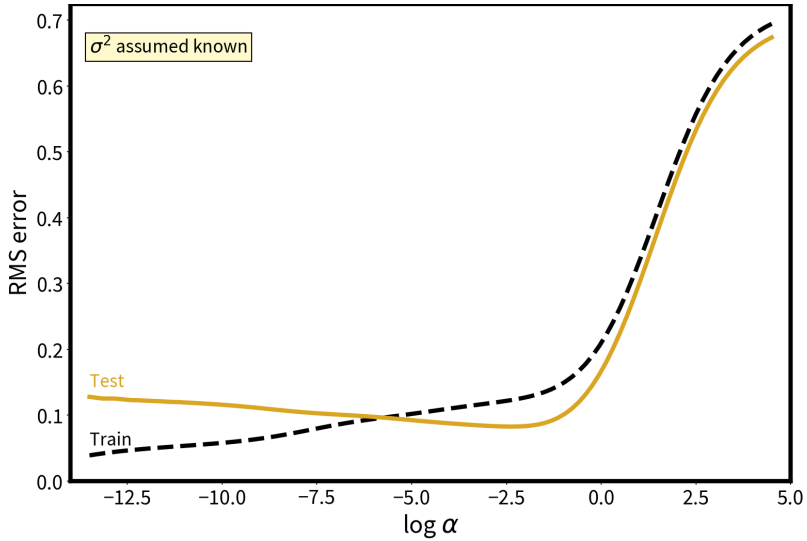
Marginal Likelihood to Estimate Single Hyper-parameter $\hat{\alpha}_{MP}$

Penalised least-squares estimation



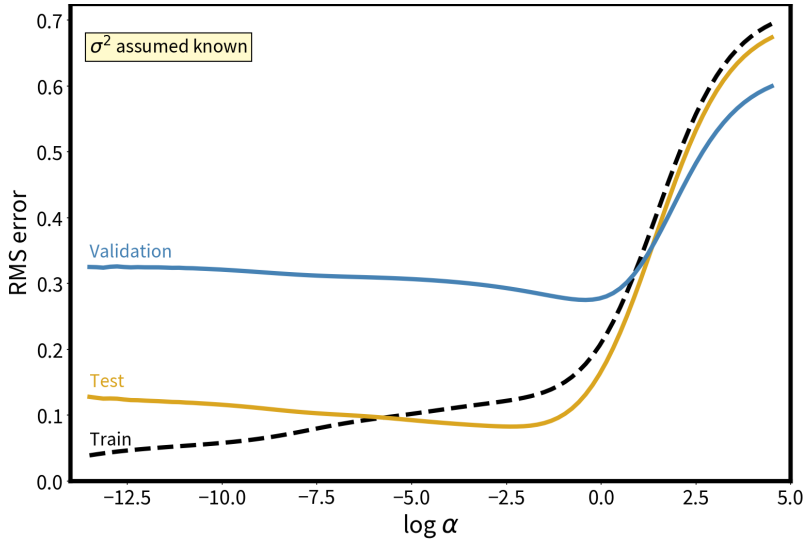
Marginal Likelihood to Estimate Single Hyper-parameter $\hat{\alpha}_{MP}$

Penalised least-squares estimation



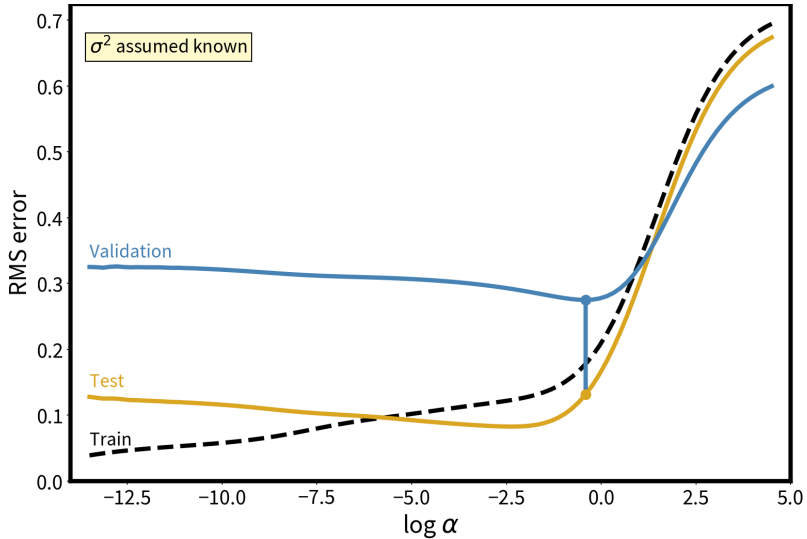
Marginal Likelihood to Estimate Single Hyper-parameter $\hat{\alpha}_{MP}$

Penalised least-squares estimation



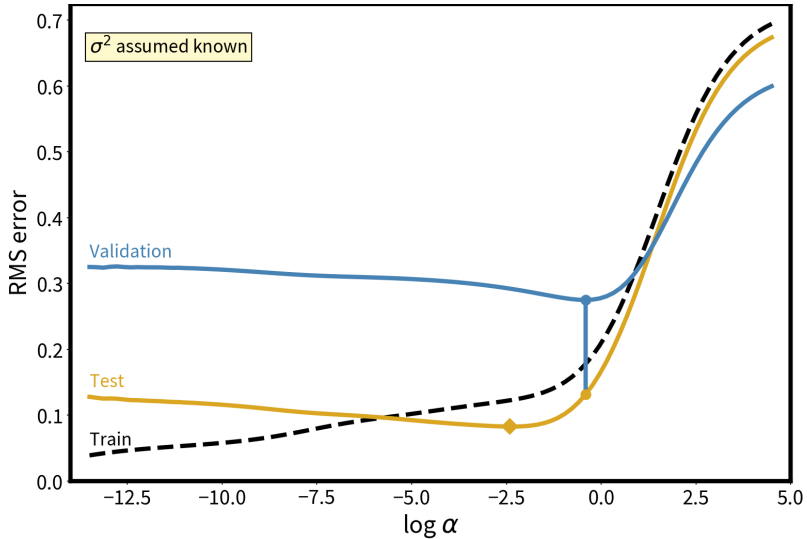
Marginal Likelihood to Estimate Single Hyper-parameter $\hat{\alpha}_{MP}$

Penalised least-squares estimation



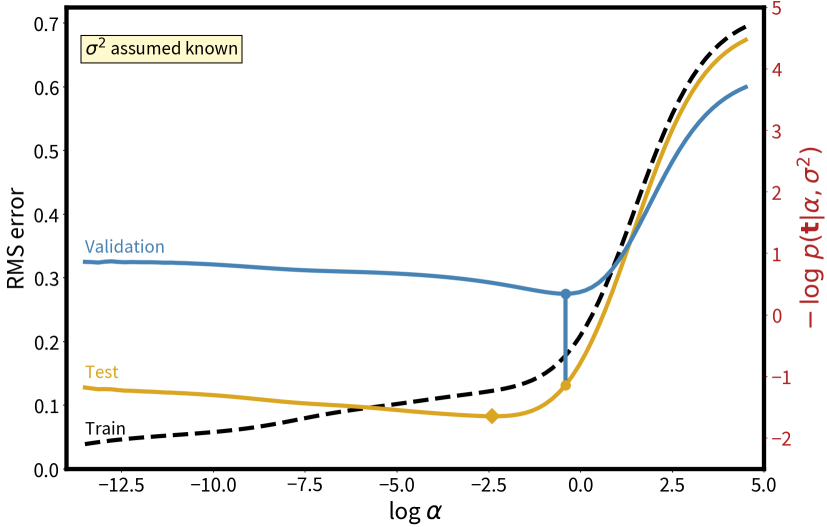
Marginal Likelihood to Estimate Single Hyper-parameter $\hat{\alpha}_{MP}$

Penalised least-squares estimation



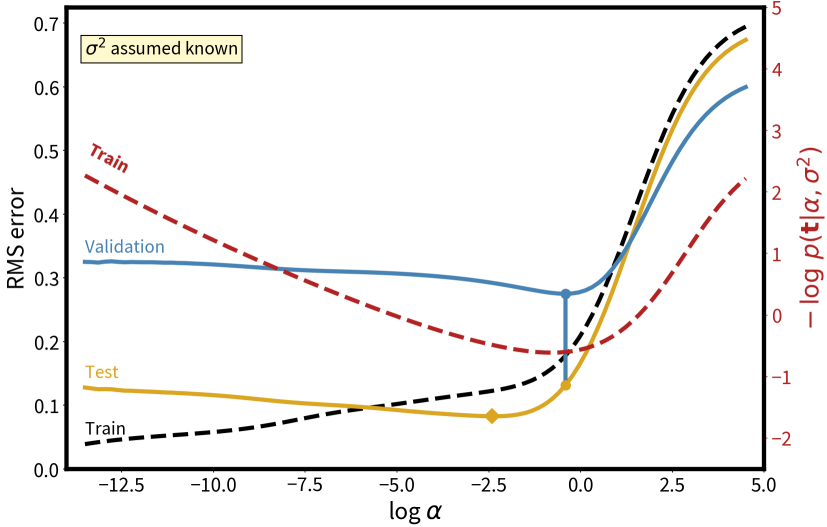
Marginal Likelihood to Estimate Single Hyper-parameter $\hat{\alpha}_{MP}$

Bayesian (Type-II) estimation



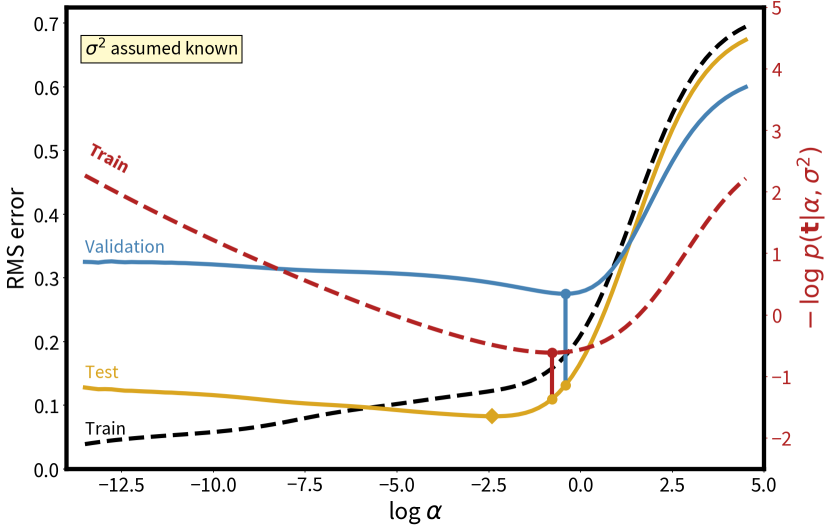
Marginal Likelihood to Estimate Single Hyper-parameter $\hat{\alpha}_{MP}$

Bayesian (Type-II) estimation



Marginal Likelihood to Estimate Single Hyper-parameter $\hat{\alpha}_{MP}$

Bayesian (Type-II) estimation

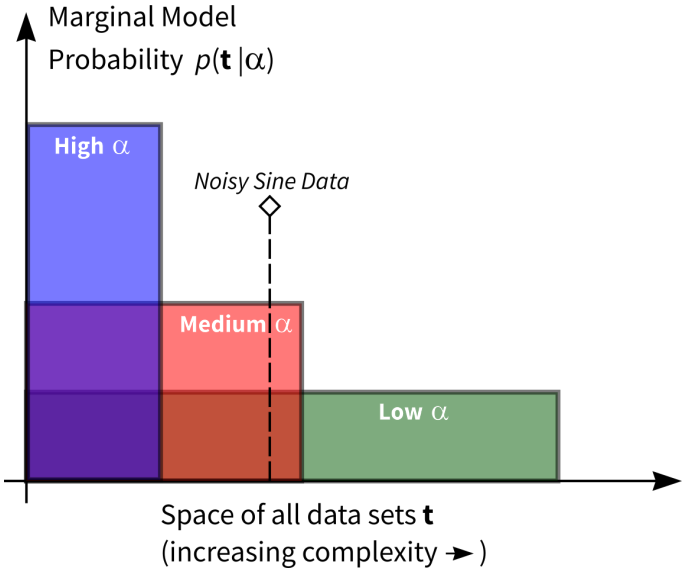


Ockham's Razor

> "Pluralitas non est ponenda sine neccesitate."
>
> *William of Ockham — 14th Century*

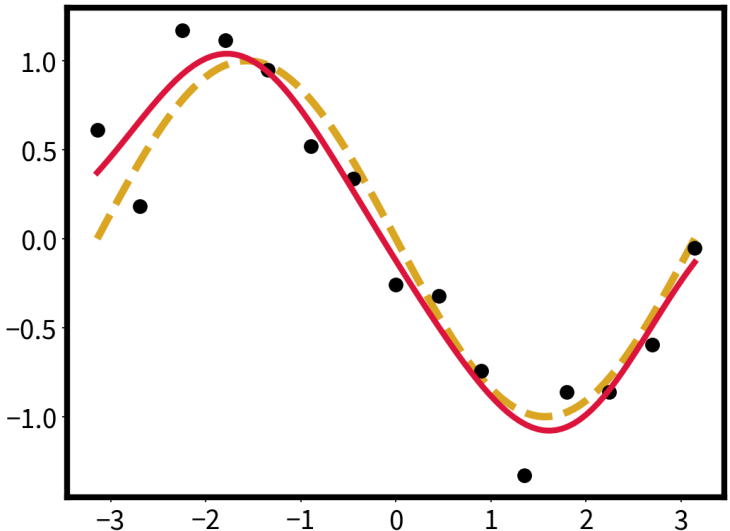
- Literally: "entities should not be multiplied unnecessarily"
- In the context of machine learning: ***models should be no more complex than is sufficient to explain the data***
- The Bayesian procedure is effectively implementing "Ockham's Razor" by assigning lower probability *both* to models that are too simple *and* too complex — how?

Preference for the “Just Right” Model



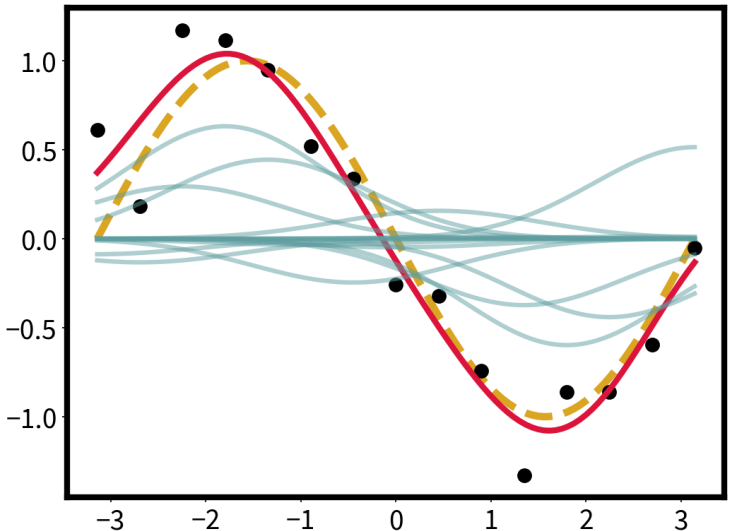
Type-II Bayesian Model (Based on $\hat{\alpha}_{MP}$ and $\hat{\sigma}^2_{MP}$)

Posterior mean predictor (15 data points)



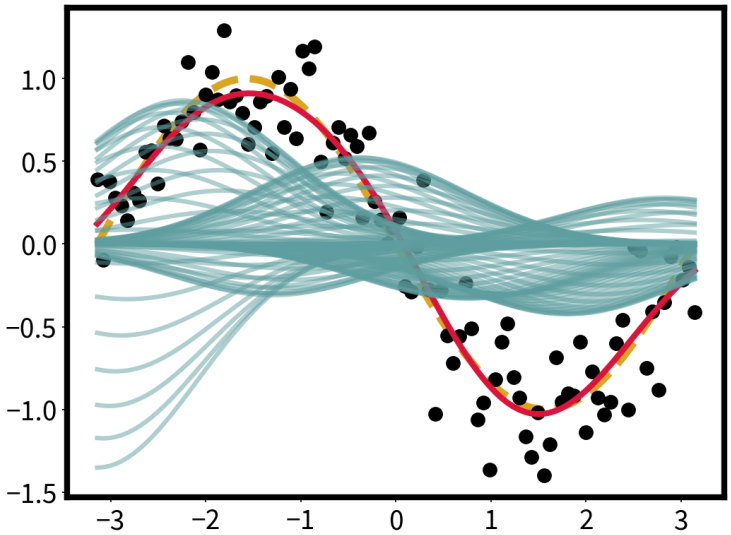
Type-II Bayesian Model (Based on $\hat{\alpha}_{MP}$ and $\hat{\sigma}^2_{MP}$)

Contribution of individual basis functions



Type-II Bayesian Model (Based on $\hat{\alpha}_{MP}$ and $\hat{\sigma}^2_{MP}$)

Contribution of individual basis functions — 100 data points!



Inferring Parsimonious Models

- It is often advantageous to fit a linear model:

$$\hat{f}(x; \mathbf{w}) = \sum_{m=1}^{M} w_m \phi_m(x)$$

  such that ***many w_m are set to zero*** (while retaining accuracy!)
- A model with few non-zero parameters is generally referred to as a **parsimonious**, or **sparse**, model
- Sparse (non-Bayesian) models have been historically popular:
  - The *Support Vector Machine* (SVM)
    — sparsity via L_2 regularisation ($\|\mathbf{w}\|_2$) and geometric constraints
  - The *Least Absolute Shrinkage and Selection Operator* (LASSO)
    — via L_1 regularisation ($\lambda\|\mathbf{w}\|_1$) and constrained optimisation
  - *Compressive Sensing* methods — via L_1 regularisation (mainly)

The Bayesian Approach To Parsimonious Inference

Express a preference for sparse model solutions via an appropriate choice of prior

- There are a number of possible candidates:
  - The *Laplace* distribution — analogous to L_1 regularisation

$$p(w|\alpha) \propto \exp\{-\alpha\|w\|\}$$

  - The *Spike and Slab* distribution — a mixture of two separate (*e.g.* Gaussian) distributions, one broad and one very narrow

$$p(w|\alpha_0, \alpha_1, \gamma) \propto \gamma . \exp\left\{-\alpha_0 |w|^2\right\} + (1 - \gamma) . \exp\left\{-\alpha_1 |w|^2\right\}$$

  - The *Logit-Normal Continuous Analogue of the Spike-and-Slab* (LN-CASS)
  - The *Student-t* distribution (indirectly)

$$p(w|a, b) \propto \left(b + \frac{w^2}{2}\right)^{-(a+\frac{1}{2})}$$

The Bayesian Approach To Parsimonious Inference

Express a preference for sparse model solutions via an appropriate choice of prior

- There are a number of possible candidates:
  - The *Laplace* distribution — analogous to L_1 regularisation

$$p(w|\alpha) \propto \exp\{-\alpha\|w\|\}$$

  - The *Spike and Slab* distribution — a mixture of two separate (*e.g.* Gaussian) distributions, one broad and one very narrow

$$p(w|\alpha_0, \alpha_1, \gamma) \propto \gamma.\exp\left\{-\alpha_0|w|^2\right\} + (1-\gamma).\exp\left\{-\alpha_1|w|^2\right\}$$

  - The *Logit-Normal Continuous Analogue of the Spike-and-Slab* (LN-CASS)
  - The *Student-t* distribution (indirectly)

$$p(w|a,b) \propto \left(b + \frac{w^2}{2}\right)^{-(a+\frac{1}{2})}$$

A Bayesian Prior for Sparse Models

- Previous regression prior (incorporating single hyper-parameter α):

$$p(\mathbf{w}|\alpha) = \prod_{m=1}^{M} (2\pi)^{-1/2} \alpha^{1/2} \exp\left\{-\frac{\alpha}{2} w_m^2\right\}$$

- **Sparse Prior** (M hyper-parameters $\alpha_1, \ldots, \alpha_M$):

$$p(\mathbf{w}|\alpha_1, \ldots, \alpha_M) = \prod_{m=1}^{M} (2\pi)^{-1/2} \alpha_m{}^{1/2} \exp\left\{-\frac{\alpha_m}{2} w_m^2\right\}$$

- **Posterior** over weights is: $p(\mathbf{w}|\mathbf{t}, \mathbf{A}, \sigma^2) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \sigma^2 \mathbf{A})^{-1} \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t}$$

$$\boldsymbol{\Sigma} = \sigma^2 (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \sigma^2 \mathbf{A})^{-1}$$

where $\mathbf{A} = \mathrm{diag}\left(\alpha_1, \ldots, \alpha_M\right)$, instead of $\mathbf{A} = \alpha \mathbf{I}$ previously

Why does this Prior Favour Sparse Solutions?

- The prior over weights appears to be Gaussian, but …
- … It is a **hierarchical** prior, parameterised by α_m, and we must marginalise to see its true form

$$p(w_m) = \int p(w_m|\alpha_m)\, p(\alpha_m)\, d\alpha_m$$

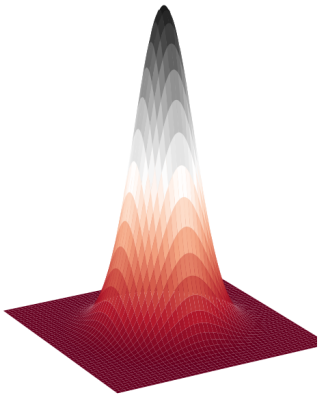
- An appropriate **hyper-prior** for α_m (a **scale parameter**) is a Gamma distribution (log-uniform is a special case)
- For $p(\alpha_m) = \text{Gamma}(\alpha_m|a, b)$, the marginal $p(w_m)$ is a *Student-t*

$$p(w_m) = \frac{b^a \Gamma(a + \frac{1}{2})}{(2\pi)^{\frac{1}{2}} \Gamma(a)} (b + w_m^2/2)^{-(a+\frac{1}{2})}$$

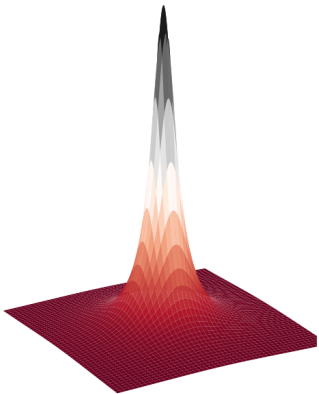
- If $a = b = 0$ (log-uniform case), then $p(w_m) \propto 1/|w_m|$

Priors Compared

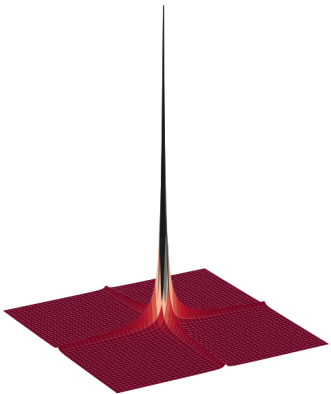
Gaussian



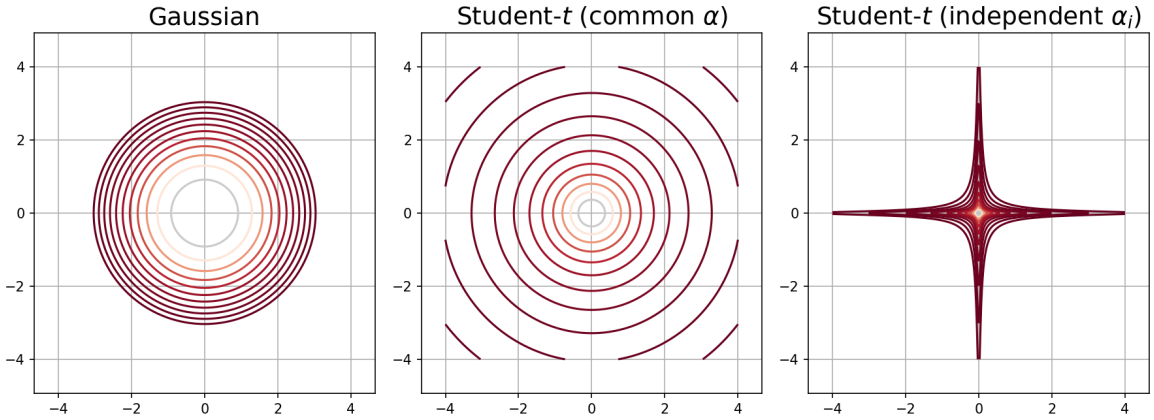
Student-t (common α)



Student-t (independent α_i)



Priors Compared



Type-II Maximum Likelihood for the Sparse Bayesian Model

- The marginal likelihood is given by:

$$p(\mathbf{t}|\mathbf{A}, \sigma^2) = \int p(\mathbf{t}|\mathbf{w}, \sigma^2)\, p(\mathbf{w}|\mathbf{A})\, d\mathbf{w}$$
$$= (2\pi)^{-\frac{N}{2}} |\mathbf{C}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{t}^{\mathsf{T}}\mathbf{C}^{-1}\mathbf{t}\right\}$$

- A ***Gaussian process*** with covariance matrix:

$$\mathbf{C} = \sigma^2\mathbf{I} + \mathbf{\Phi}\mathbf{A}^{-1}\mathbf{\Phi}^{\mathsf{T}} = \sigma^2\mathbf{I} + \sum_{m=1}^{M} \alpha_m^{-1}\boldsymbol{\phi}_m\boldsymbol{\phi}_m^{\mathsf{T}}$$

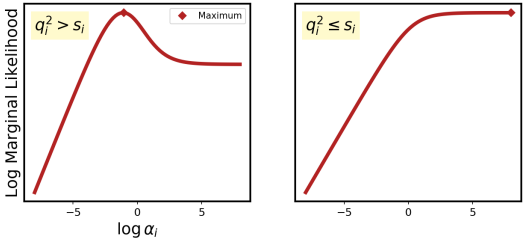
- It gives us a probability for the ***entire data set***, conditional on M values of $\alpha_1, \ldots, \alpha_M$ along with σ^2

Type-II Maximum Likelihood for the Sparse Bayesian Model

- How do we estimate all $\alpha_1, \ldots, \alpha_M$? (And σ^2?)
- It should be clear that the earlier ***empirical approach is impractical***!
- Possible algorithmic (iterative) approaches for maximising $\log p(\mathbf{t}|\mathbf{A}, \sigma^2)$:
  - off-the-shelf non-linear optimisation (gradient-based)
  - the expectation-maximisation (EM) algorithm
  - variational approximations $Q_\alpha(\alpha_m)$
- These work, but are slow to converge
- Alternative: exploit the specific properties of $\log p(\mathbf{t}|\mathbf{A}, \sigma^2)$ to derive an efficient ***co-ordinate (gradient) ascent*** algorithm

Properties of the Marginal Likelihood

- What does the function $\log p(\mathbf{t}|\alpha_1, \ldots, \alpha_M)$ look like?
- As a function of a *single* hyper-parameter α_i:



- Factors q_i and s_i are a function of all other α_m, but not α_i:

$$\text{“Quality factor”:} \quad q_i = \boldsymbol{\phi}_i^\mathsf{T} \mathbf{C}_{-i}^{-1} \mathbf{t}$$

$$\text{“Sparsity factor”:} \quad s_i = \boldsymbol{\phi}_i^\mathsf{T} \mathbf{C}_{-i}^{-1} \boldsymbol{\phi}_i$$

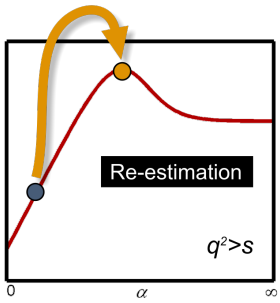
$$\text{with:} \quad \mathbf{C}_{-i} = \sigma^2 \mathbf{I} + \sum_{m \neq i} \alpha_m^{-1} \boldsymbol{\phi}_m \boldsymbol{\phi}_m^\mathsf{T}$$

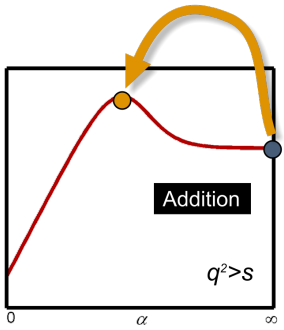
A Sparse Sequential Learning Algorithm Sketch

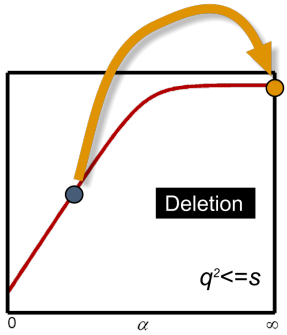
1. Initialise all $\alpha_m = \infty$: start with an "empty" model
2. Compute all q_m and s_m factors
3. Select a candidate basis function $\phi_i(\mathbf{x})$ from the set of all M
4. Examine quality and sparsity factors, q_i and s_i:
   - If $q_i^2 > s_i$ and $\alpha_i < \infty$: **re-estimate** $\alpha_i = s_i^2/(q_i^2 - s_i)$
   - If $q_i^2 > s_i$ and $\alpha_i = \infty$: **add** ϕ_i to the model with $\alpha_i = s_i^2/(q_i^2 - s_i)$
   - If $q_i^2 \leq s_i$ and $\alpha_i < \infty$: **delete** ϕ_i from the model (set $\alpha_i = \infty$)
5. If not converged, return to Step 2

A Sparse Sequential Learning Algorithm Sketch

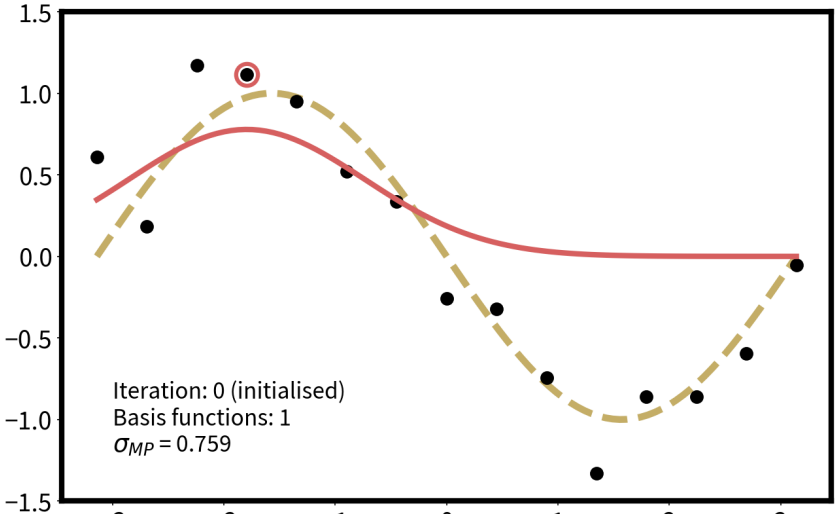
In pictures



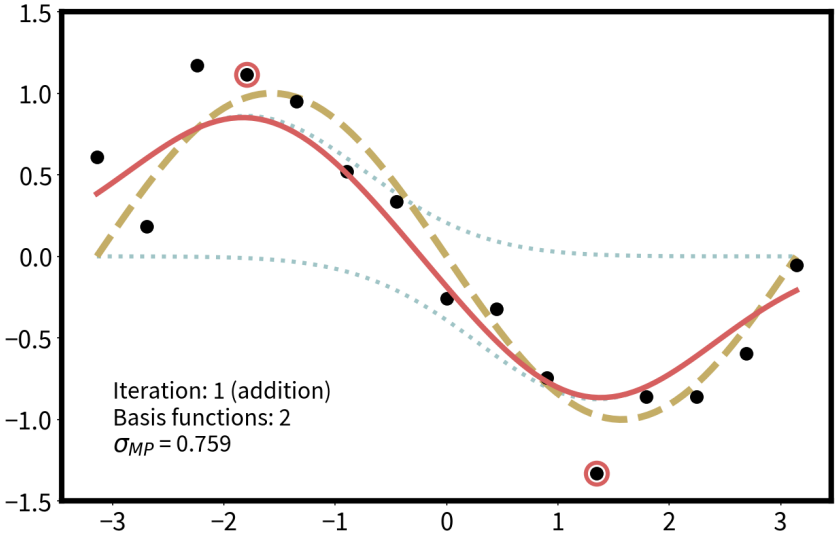




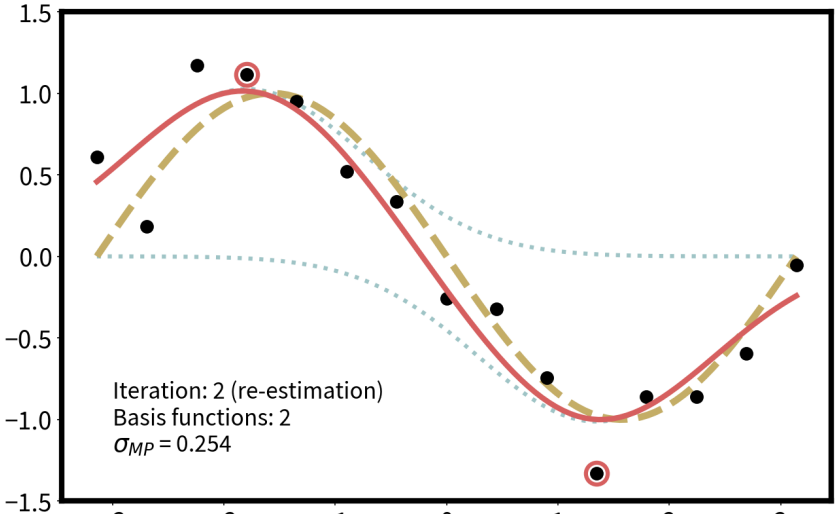
Example on Synthetic Data (N = 15)



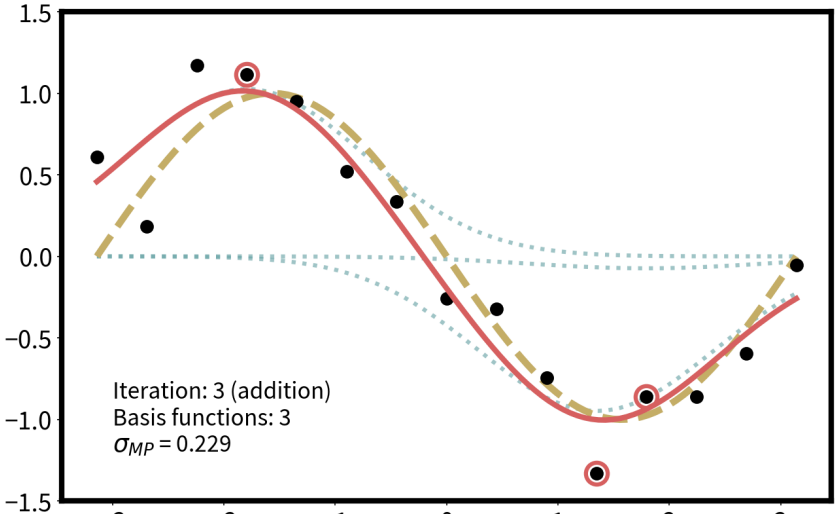
Example on Synthetic Data (N = 15)



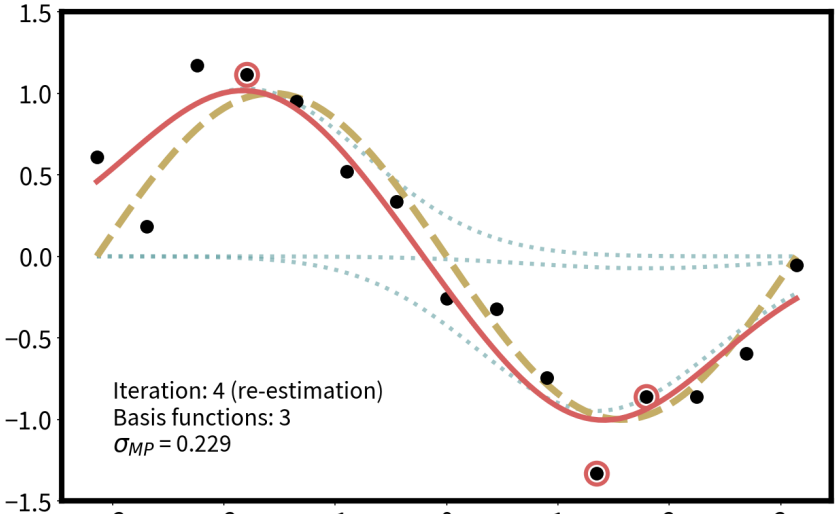
Example on Synthetic Data (N = 15)



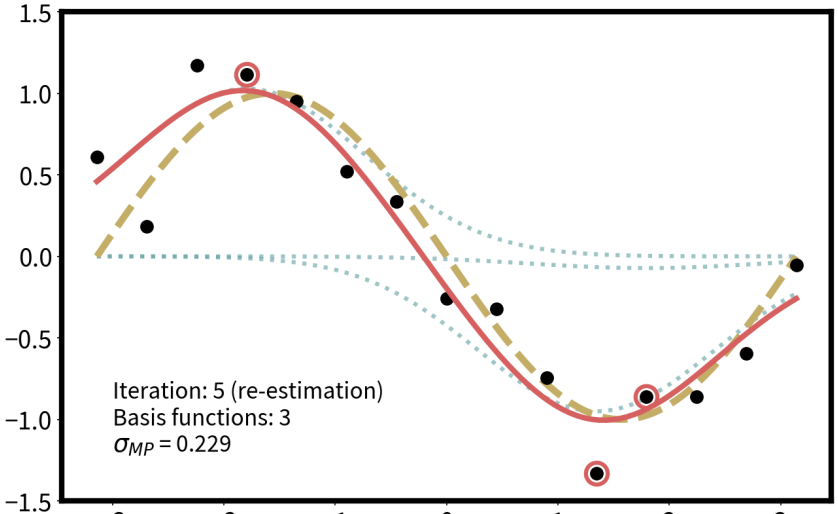
Example on Synthetic Data (N = 15)



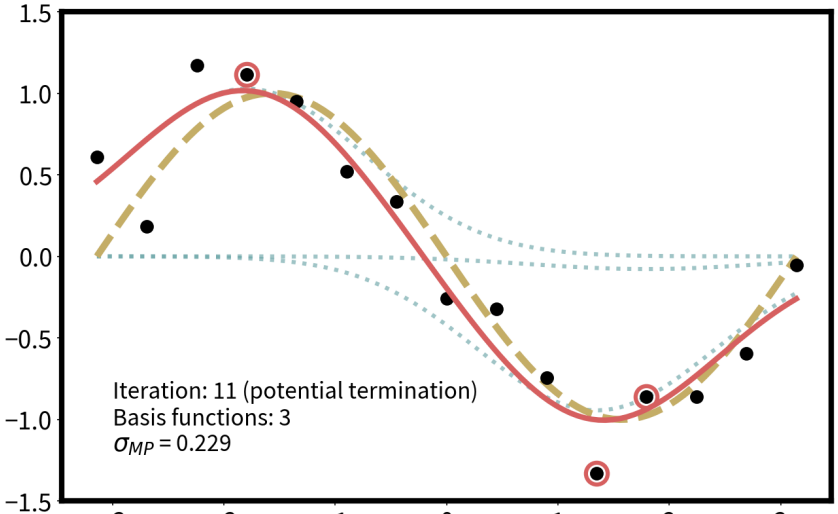
Example on Synthetic Data (N = 15)



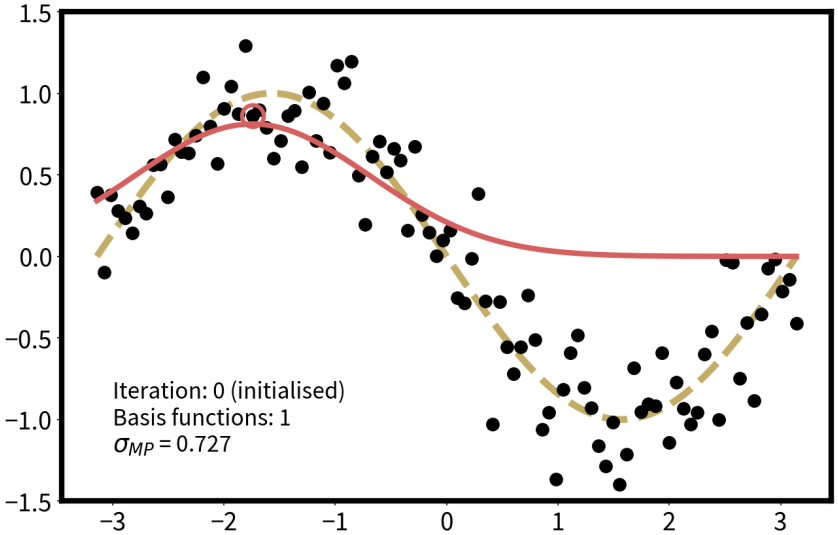
Example on Synthetic Data (N = 15)



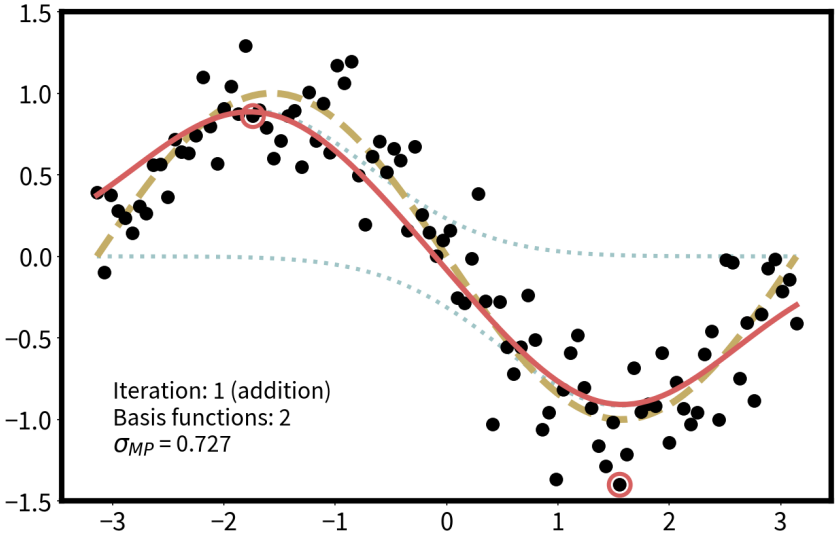
Example on Synthetic Data (N = 15)



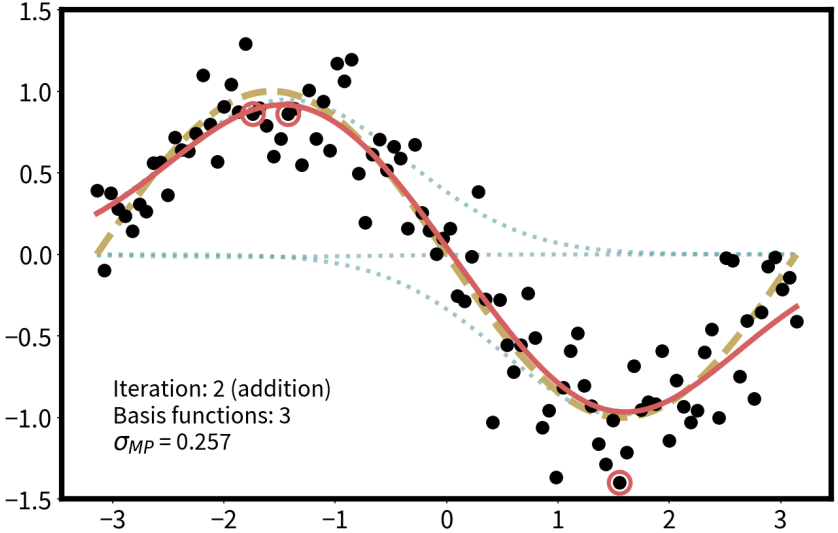
Example on Synthetic Data (N = 100)



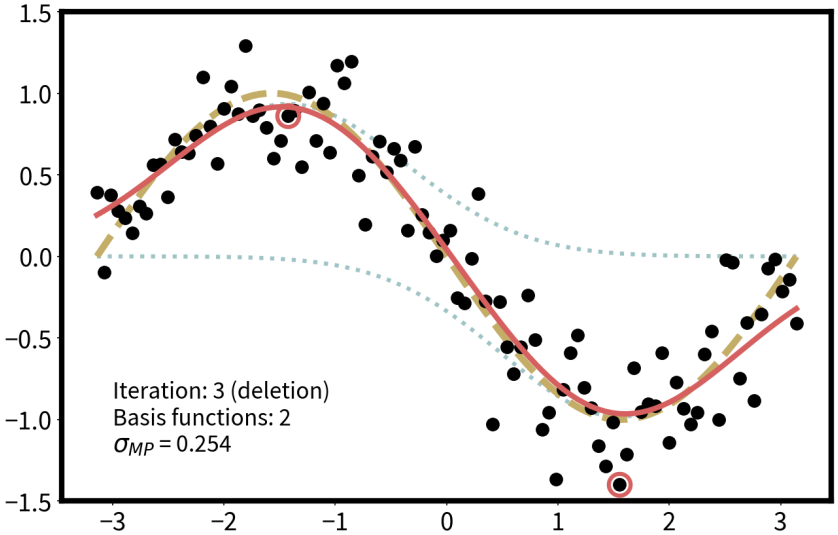
Example on Synthetic Data (N = 100)



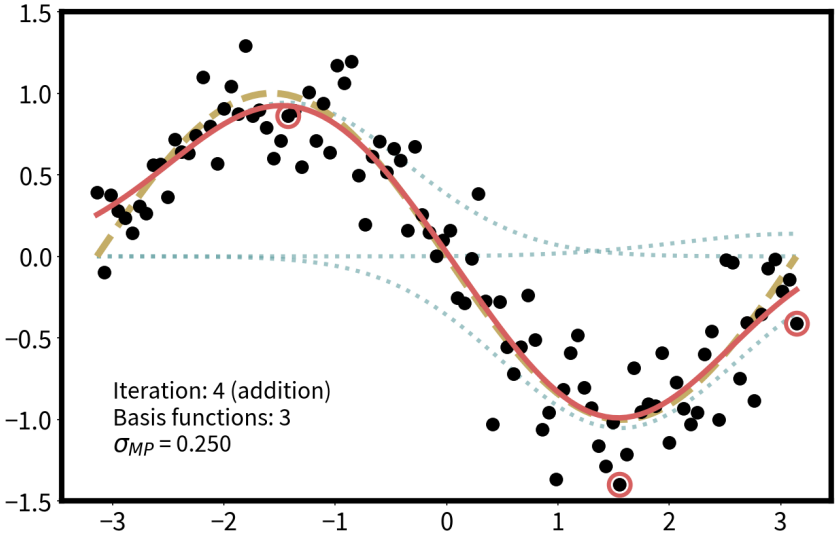
Example on Synthetic Data (N = 100)



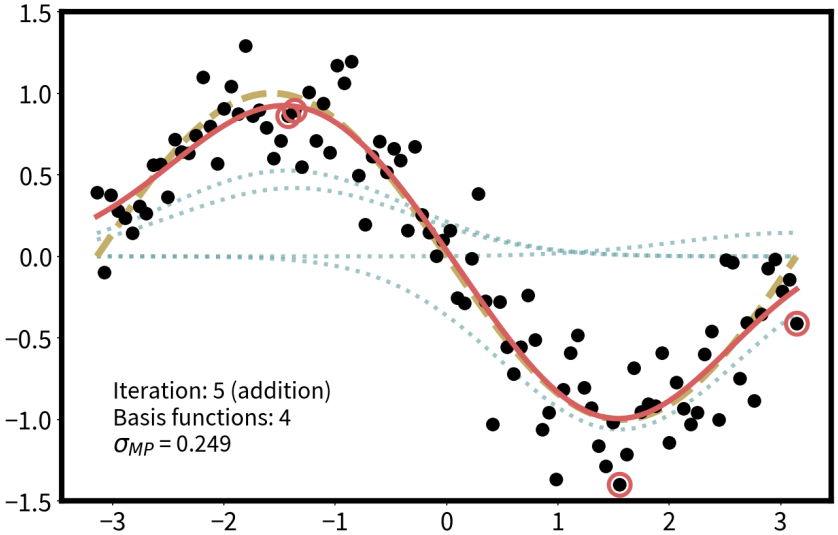
Example on Synthetic Data (N = 100)



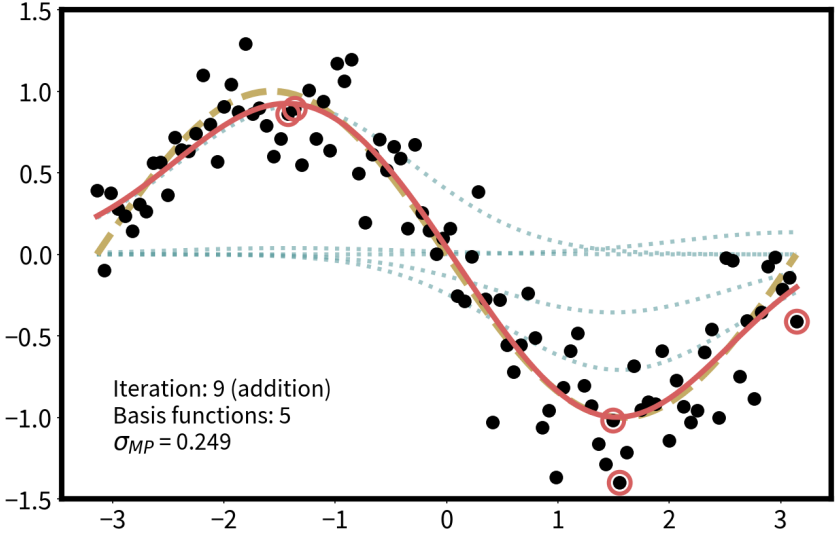
Example on Synthetic Data (N = 100)



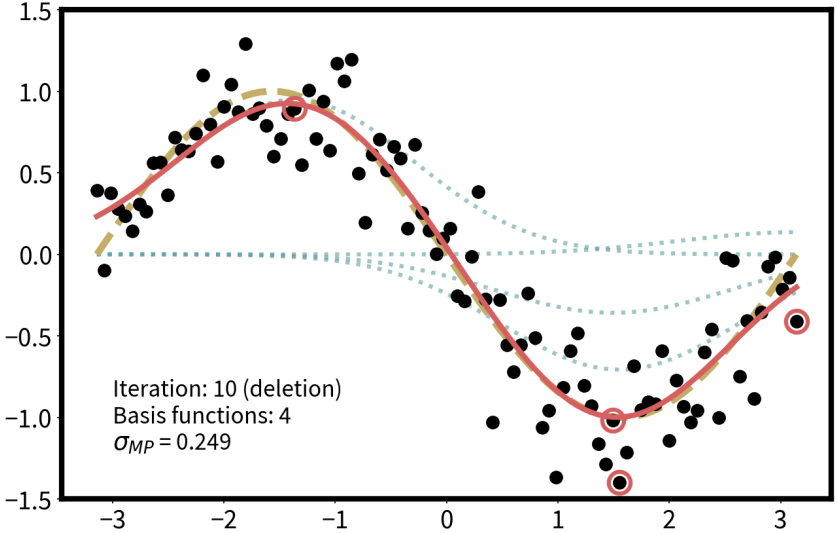
Example on Synthetic Data (N = 100)



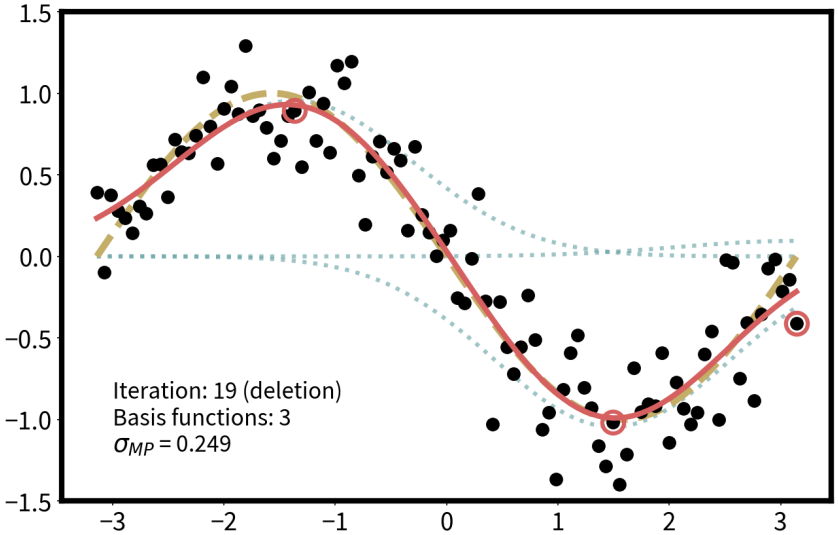
Example on Synthetic Data (N = 100)



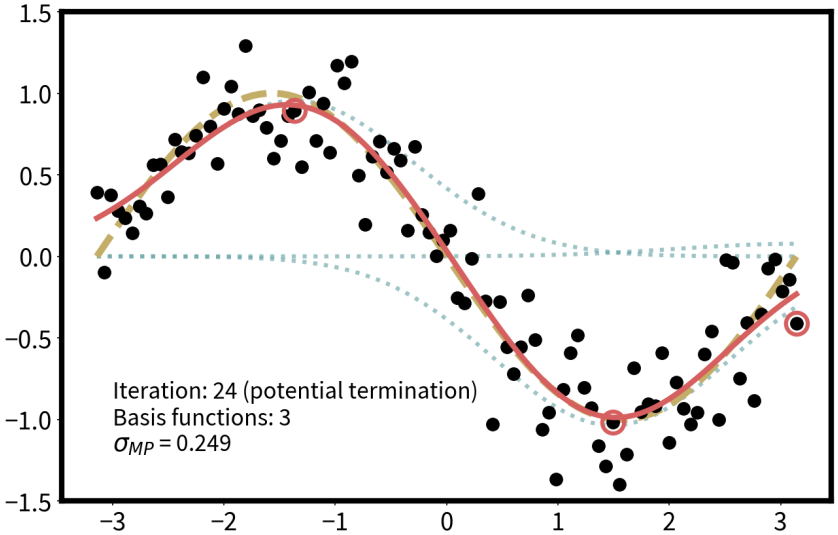
Example on Synthetic Data (N = 100)



Example on Synthetic Data (N = 100)



Example on Synthetic Data (N = 100)

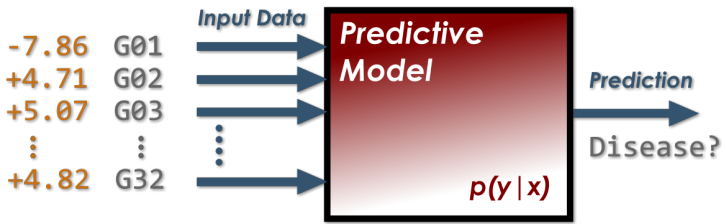


Sparse Modelling in Real-World Applications

Prediction of disease based on gene expression micro-array data

Logistic regression model: 100% accurate

➥ needs expression data from all 32 genes

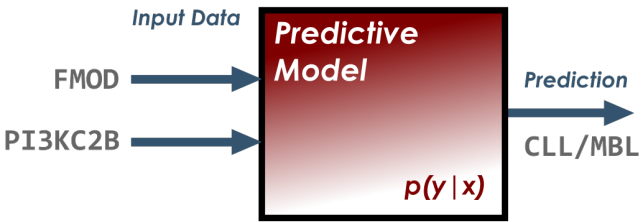


Sparse Modelling in Real-World Applications

Prediction of disease based on gene expression micro-array data

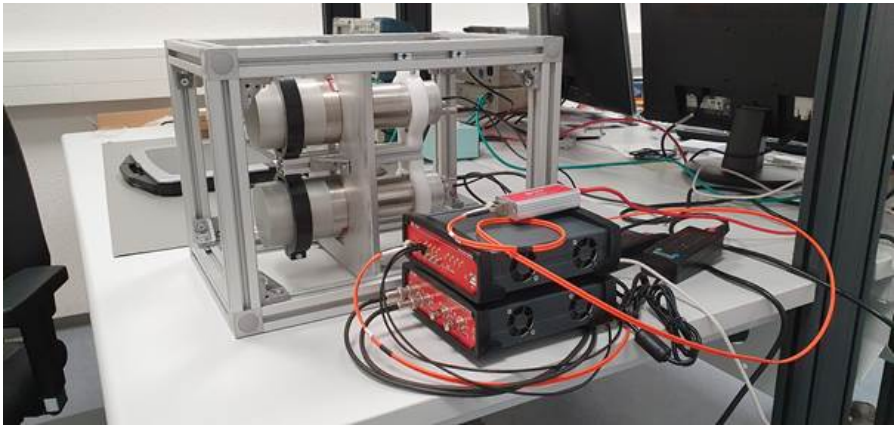
Sparse Bayesian model: 100% accurate

➥ needs knowledge of *only two genes*



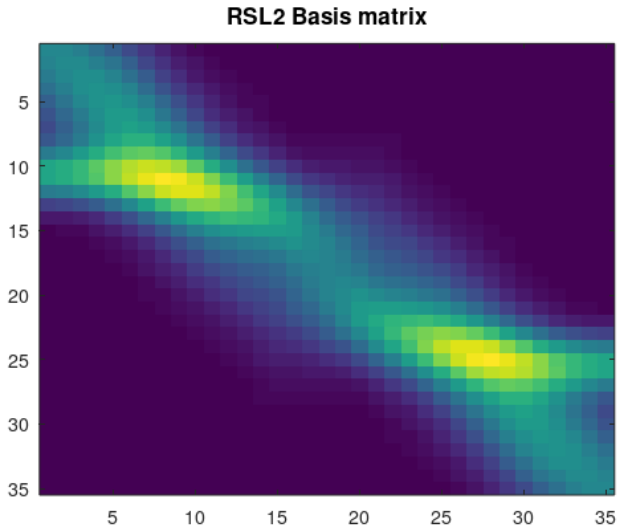
Correctly distinguishes 100% of CLL and MBL cases from normal polyclonal and mono/oligoclonal B lymphocytes.

A “Compton camera” setup, for the resolving of Gamma rays:



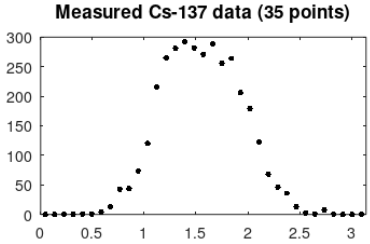
Sparse Modelling in Real-World Applications

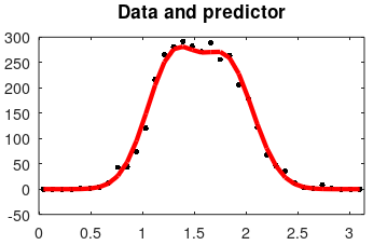
Localisation of radioactive sources using a Compton Camera

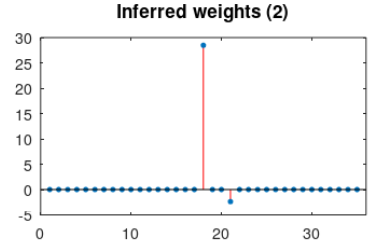


Sparse Modelling in Real-World Applications

Localisation of radioactive sources using a Compton Camera



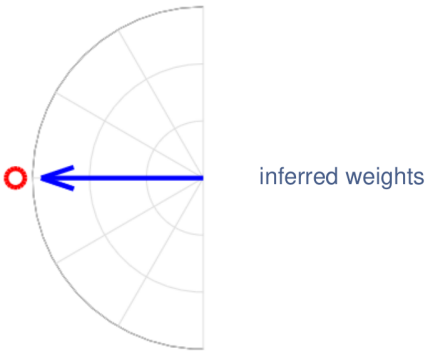




Sparse Modelling in Real-World Applications

Localisation of radioactive sources using a Compton Camera

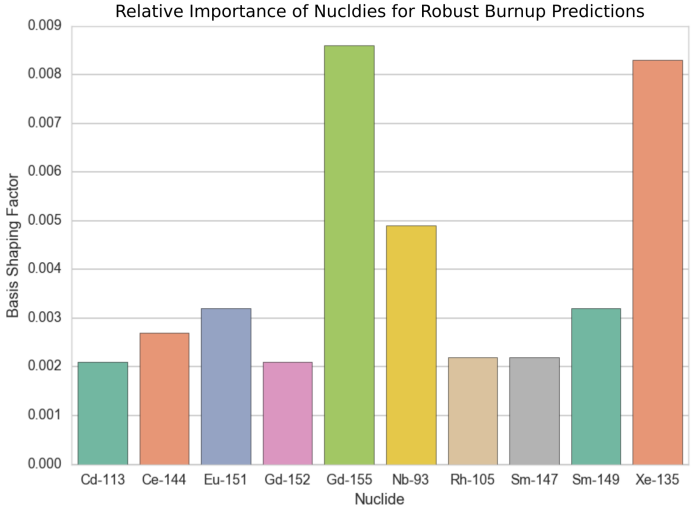
- This is an **inverse modelling problem**
- The sparse solution allows the gamma radiation source(s) to be localised by angle
- More details: “Machine learning techniques applied to Compton cameras” at *Applied Inverse Problems*, Göttingen, September 2023
- Data and illustrations courtesy of: *Hellma Materials GmbH* and *Hochschule Zittau/Görlitz*



RVM Output: blue arrows
True Location: red dot

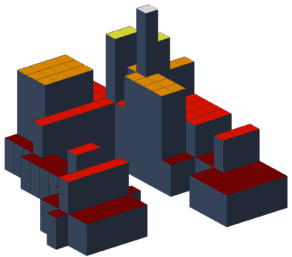
Sparse Modelling in Real-World Applications

Prediction of nuclear reactor core burn-up based on nuclide samples (Dayman et al.)



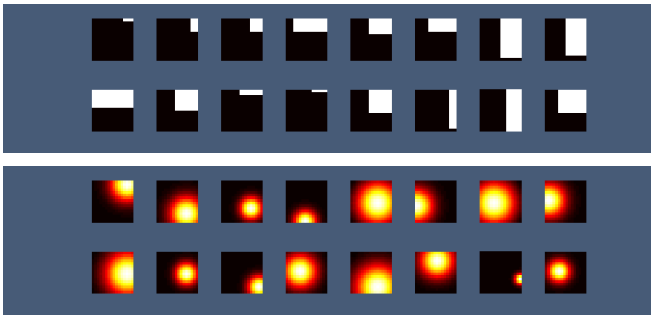
Could The Sparse Bayesian Algorithm Be Too "Greedy"?

- At every step, we updated the basis function $\phi_m(\mathbf{x})$ that most increased the objective function $\log p(\mathbf{t}|\alpha_1, \ldots, \alpha_M)$
- A number of (non-Bayesian) algorithms work in a very similar way — *e.g.* **orthogonal matching pursuit** (in *scikit-learn*)
- Such algorithms are termed **greedy**: they are efficient, but "early" additions can be significantly sub-optimal
- Return to the example …



Example: Building Blocks Revisited

- The previous demo was perhaps too easy?
- Let's make things tougher: include a basis of Gaussians centred on every pixel (256), each with four different widths:



- This makes 1280 basis functions, of which 80% are "confusers"
  - the basis is **over-complete**, and so …
  - the problem is **under-constrained**

Demo

- “Building block” demo (reloaded) …

High-Level Summary

- Appropriate integration of probability in real-world applications is crucial
  - to handle multiple sources of uncertainty
  - to enable crucial separation of inference and decision-making
- Bayesian probabilistic approaches can offer numerous advantages
  - we generally try to “adhere to a strict moral code” …
  - … but can obtain useful results with a little judicious “loose living”
  - the “art” is in the choice of approximations and how we apply them

And Finally, Coming Soon ...

And Finally, Coming Soon ...

And Finally, Coming Soon ...

Oppenheimer on “Style”

“The problem of doing justice to the implicit, the imponderable, and the unknown is of course not unique in politics. It is always with us in science, it is with us in the most trivial of personal affairs, and it is one of the great problems of writing and of all forms of art. The means by which it is solved is sometimes called *style*. It is style which complements affirmation with limitation and with humility; it is style which makes it possible to act effectively, but not absolutely; it is style which, in the domain of foreign policy, enables us to find a harmony between the pursuit of ends essential to us, and the regard for the views, the sensibilities, the aspirations of those to whom the problem may appear in another light; **it is style which is the deference that action pays to uncertainty**; it is above all style through which power defers to reason.”

J. Robert Oppenheimer *— from “The Open Mind” lecture collection (1955)*