A Bayesian Perspective on Statistical Learning Theory

Daniel M. Roy (MIT)

9.520: Statistical Learning Theory, Spring 2009

Many figures and slides taken from:

Zoubin Ghahramani (Cambridge; CMU)

Carl Rasmussen (Cambridge)

Baback Moghaddam (JPL)

Charlie Frogner (MIT)

Overview

- Bayesian Basics
- Bayesian Philosophy
- Recap: Bayesian Linear Regression
- Recap: Bayesian Non-linear Regression (Gaussian Processes)
- How to choose the kernel?
  a.k.a. Model Selection
  a.k.a. Bayes Occam's Razor
- Nonparametric Philosophy: Occam's Hill v. Occam's Plateau
- Conclusion

Bayesian Inference

Posterior *Likelihood* x *Prior*

$$p(\theta \mid D, M) = \frac{p(D \mid \theta, M)\ p(\theta \mid M)}{p(D \mid M)}$$

Evidence

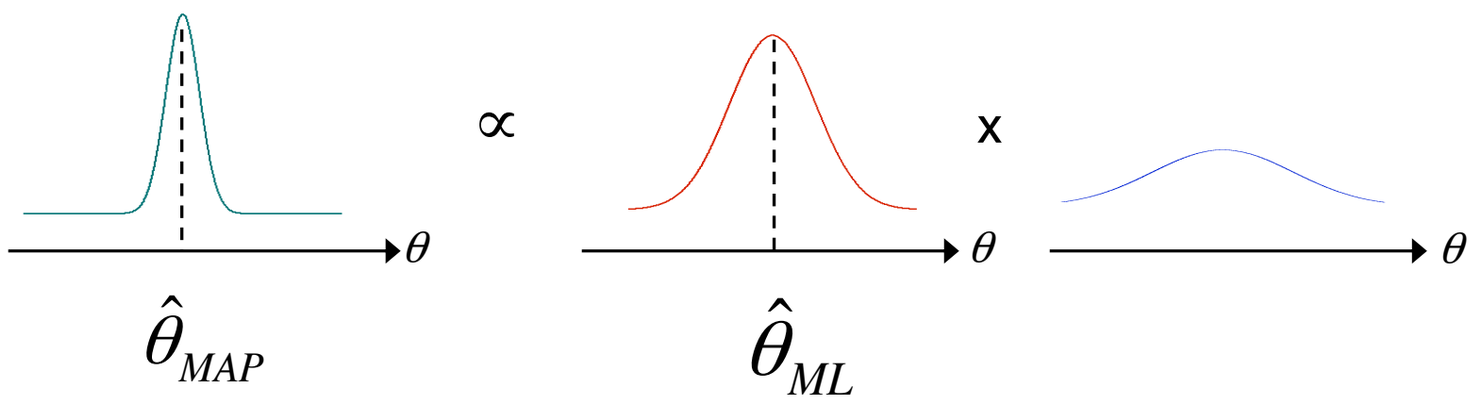
The *evidence* for our model $\boldsymbol{M}$ is also called "Marginal Likelihood'

$$p(D \mid M) = \int p(D \mid \theta, M)\ p(\theta \mid M)\, d\theta$$

Bayesian Nutshell

Posterior *Likelihood* x *Prior*

$$p(\theta \mid D, M) \propto p(D \mid \theta, M)\ p(\theta \mid M)$$



$\hat{\theta}_{MAP}$ $\hat{\theta}_{ML}$

Moghaddam

Probability = Degree of Belief

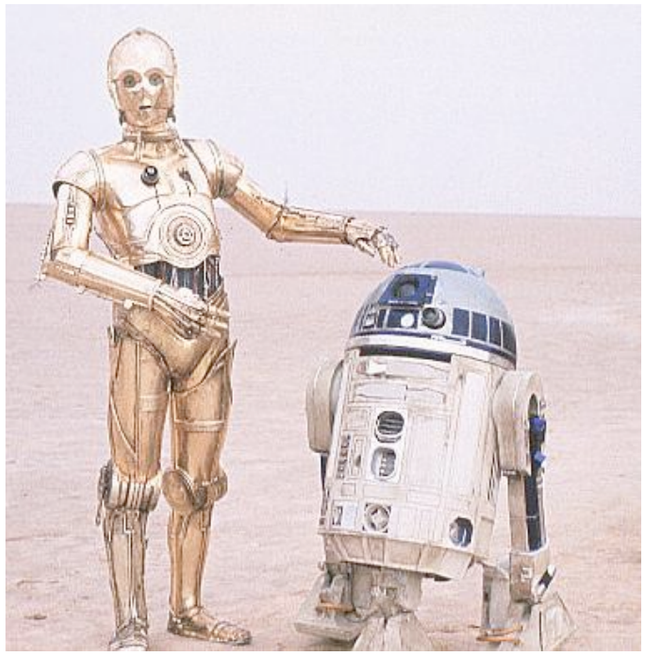
- Let C be the result of a coin toss, either heads (=h) or tails (=t), which is about to be revealed to you.

- **Frequentist Interpretation**
  P(C=h) = *p* "The long run frequency of heads is *p*"
  p is a *nonrandom* property of the coin/experiment

- **Bayesian Interpretation**
  P(C=h) = 1 "I'm absolutely certain the coin is heads"
  P(C=h) = 0 "I'm absolutely certain the coin is tails"
  P(C=h) = 1/2 "I'm completely uncertain"
  P(C=h) = *p* "The probability that the coin is heads is *p*."
  p is an uncertain quantity; i.e., we model it as *random* and put a distribution on it representing our uncertainty before learning C

Representing Beliefs (Artificial Intelligence)

Consider a robot. In order to behave intelligently the robot should be able to represent beliefs about propositions in the world:



"my charging station is at location (x,y,z)"

"my rangefinder is malfunctioning"

"that stormtrooper is hostile"

We want to represent the **strength** of these beliefs numerically in the brain of the robot, and we want to know what rules (calculus) we should use to manipulate those beliefs.

Zoubin Ghahramani

Representing Beliefs II

Let's use $b(x)$ to represent the stength of belief in (plausibility of) proposition x.

$0 \leq b(x) \leq 1$		
$b(x) = 0$	x	is definitely **not true**
$b(x) = 1$	x	is definitely **true**
$b(x\|y)$		strength of belief that x is true given that we know y is true

Cox Axioms (Desiderata):

- Strengths of belief (degrees of plausibility) are represented by real numbers
- Qualitative correspondence with common sense
- Consistency
  - If a conclusion can be reasoned in more than one way, then every way should lead to the same answer.
  - The robot always takes into account all relevant evidence.
  - Equivalent states of knowledge are represented by equivalent plausibility assignments.

Consequence: Belief functions (e.g. $b(x)$, $b(x|y)$, $b(x, y)$) must satisfy the rules of probability theory, including Bayes rule. (see Jaynes, *Probability Theory: The Logic of Science*)

The Dutch Book Theorem

Assume you are willing to accept bets with odds proportional to the stength of your beliefs. That is, $b(x) = 0.9$ implies that you will accept a bet:

$$\begin{cases} x & \text{is true} & \text{win} & \geq \$1 \\ x & \text{is false} & \text{lose} & \$9 \end{cases}$$

Then, unless your beliefs satisfy the rules of probability theory, including Bayes rule, there exists a set of simultaneous bets (called a "Dutch Book") which you are willing to accept, and for which **you are guaranteed to lose money, no matter what the outcome**.

The only way to guard against Dutch Books to to ensure that your beliefs are coherent: i.e. satisfy the rules of probability.

Zoubin Ghahramani

Where do ~~priors~~ models come from?

- Let $\{\mathbf{S_{c,t}}\}$ be stock prices for companies **c** and times **t**. We need a model of $P(S_{c,t})$.... where to begin?
- We should only believe the predictions from a model if we have faithfully encoded our knowledge into the probabilistic model.
- PAC-Bayes: Risk of using a model related to divergence between the distribution before and after receiving data.
    - Use Bayesian methods as a language to encode assumptions: Bayesian inference ensures that we never violate the assumptions that our distributions represent when updating our beliefs.

de Finetti's Theorem

- **Theorem**: Let C_1, C_2, ... be an infinite sequence of binary random variables. If the distribution of the sequence is invariant to permutations (i.e. if the sequence is ***exchangeable***), then

  there is a *random* variable **θ** with some distribution *F*
  such that conditioned on **θ**, the sequence is conditionally independent and identically distributed (i.i.d.).

  furthermore $F(\theta < t) = P\left(\lim_{n\to\infty} \sum_{i=1}^{n} C_i < t\right)$

- Bayesian justification for inventing a latent variable θ and assigning a (prior) distribution *Q*(θ).

$$P(\theta, C_1, C_2, \dots) = Q(\theta) \times P(C_1|\theta) \times P(C_2|\theta) \dots$$

Asymptotic Certainty

Assume that data set $\mathcal{D}_n$, consisting of n data points, was generated from some true θ^*, then under some regularity conditions, as long as $p(\theta^*) > 0$

$$\lim_{n\to\infty} p(\theta|\mathcal{D}_n) = \delta(\theta - \theta^*)$$

In the **unrealizable case**, where data was generated from some $p^*(x)$ which cannot be modelled by any θ, then the posterior will converge to

$$\lim_{n\to\infty} p(\theta|\mathcal{D}_n) = \delta(\theta - \hat{\theta})$$

where $\hat{\theta}$ minimizes $\mathrm{KL}(p^*(x), p(x|\theta))$:

$$\hat{\theta} = \operatorname*{argmin}_\theta \int p^*(x) \log \frac{p^*(x)}{p(x|\theta)}\, dx = \operatorname*{argmax}_\theta \int p^*(x) \log p(x|\theta)\, dx$$

Warning: careful with the regularity conditions, these are just sketches of the theoretical results

Asymptotic Consensus

Consider two Bayesians with *different priors*, $p_1(\theta)$ and $p_2(\theta)$,
who observe the *same data* $\mathcal{D}$.

Assume both Bayesians agree on the set of possible and impossible values of θ:

$$\{\theta : p_1(\theta) > 0\} = \{\theta : p_2(\theta) > 0\}$$

Then, in the limit of $n \to \infty$, the posteriors, $p_1(\theta|\mathcal{D}_n)$ and $p_2(\theta|\mathcal{D}_n)$ will converge
(in uniform distance between distibutions $\rho(P_1, P_2) = \sup_E |P_1(E) - P_2(E)|$)

Probability Supports (Knowledge) Engineering

- **Universal Primitives**
  Independent uniformly distributed [0,1]-random variables sufficient for any computable distribution

- **Means of combination**
  Build complex models from simple pieces $P(X) \cdot P(Y|X) = P(X,Y)$

- **Means of abstraction**
  Create new primitives $P(Z) = \int \cdots \int P(A, B, \ldots, Y, Z)\, dA\, dB \cdots dY$

- Possible to devise **inference algorithms** that respect this structure as well (see e.g., MIT-Church [GMRBT2008])

Bayesian Linear Regression

- Prior knowledge: data *Y* are noisy measurements of a linear function $f_S(x) = x^T\theta$ at points $X=(x_1,...,x_n)^T$
- We are uncertain about the function, so as Bayesian we put a prior on the space of linear functions. We do so indirectly by placing a prior on θ

Bayesian Linear Regression

- Prior knowledge: data *Y* are noisy measurements of a linear function $f_S(x) = x^T\theta$ at points $X=(x_1,\ldots,x_n)^T$

- We are uncertain about the function, so as Bayesian we put a prior on the space of linear functions. We do so indirectly by placing a prior on θ

Model:

$$Y|X,\theta \sim \mathcal{N}\left(X\theta, \sigma_\varepsilon^2 I\right), \qquad \theta \sim \mathcal{N}(0, I)$$

Posterior:

$$\theta|X, Y \sim \mathcal{N}\left(\mu_{\theta|X,Y}, \Sigma_{\theta|X,Y}\right)$$

where

$$\hat{\theta}_{MAP}(Y|X) = \hat{\theta}_{BLS}(Y|X) = \begin{aligned} \mu_{\theta|X,Y} &= X^T(XX^T + \sigma_\varepsilon^2 I)^{-1} Y \\ \Sigma_{\theta|X,Y} &= I - X^T(XX^T + \sigma_\varepsilon^2 I)^{-1} X \end{aligned}$$

Charlie Frogner

Bayesian Linear Regression

- Prior knowledge: data *Y* are noisy measurements of a linear function $f_S(x) = x^T\theta$ at points $X=(x_1,\ldots,x_n)^T$

- We are uncertain about the function, so as Bayesian we put a prior on the space of linear functions. We do so indirectly by placing a prior on θ

Model:

$$Y|X,\theta \sim \mathcal{N}\left(X\theta, \sigma_\varepsilon^2 I\right), \qquad \theta \sim \mathcal{N}(0, I)$$

Posterior:

$$\theta|X, Y \sim \mathcal{N}\left(\mu_{\theta|X,Y}, \Sigma_{\theta|X,Y}\right)$$

where

$$\hat{\theta}_{RLS}(Y|X) = \hat{\theta}_{MAP}(Y|X) = \hat{\theta}_{BLS}(Y|X) = \mu_{\theta|X,Y} = X^T(XX^T + \sigma_\varepsilon^2 I)^{-1} Y$$
$$\Sigma_{\theta|X,Y} = I - X^T(XX^T + \sigma_\varepsilon^2 I)^{-1} X$$

Charlie Frogner

Bayesian Non-Linear Regression

- Prior knowledge: data *Y* are noisy measurements of a non-linear function $f_S(x) = \phi(x)^T\theta$ at points $X=(x_1,...,x_n)^T$

- We are uncertain about the function, so as Bayesian we put a prior on the space of functions with this basis. Again, we do so indirectly by placing a prior on the coefficients θ

Model:

$$Y|X,\theta \sim \mathcal{N}\left(\phi(X)\theta, \sigma_\varepsilon^2 I\right), \qquad \theta \sim \mathcal{N}(0, I)$$

Then:

$$\hat{\theta}_{MAP}(Y|X) = \phi(X)^T(K(X,X) + \sigma_\varepsilon^2 I)^{-1} Y$$

Estimated function?

$$\begin{aligned}\hat{f}_{MAP}(x) &= \phi(x)\hat{\theta}_{MAP}(Y|X) \\ &= \phi(x)\phi(X)^T(K(X,X) + \sigma_\varepsilon^2 I)^{-1} Y \\ &= K(x,X)(K(X,X) + \frac{\lambda}{2} I)^{-1} Y \\ &= \hat{f}_{RLS}(x)\end{aligned}$$

Charlie Frogner

Bayesian Non-Linear Regression (cont.)

- Prior knowledge: data *Y* are noisy measurements of a non-linear function $f_S(x) = \phi(x)^T\theta$ at points $X=(x_1,...,x_n)^T$
- We are uncertain about the function, so as Bayesian we put a prior on the space of functions with this basis. Again, we do so indirectly by placing a prior on the coefficients θ

Posterior:

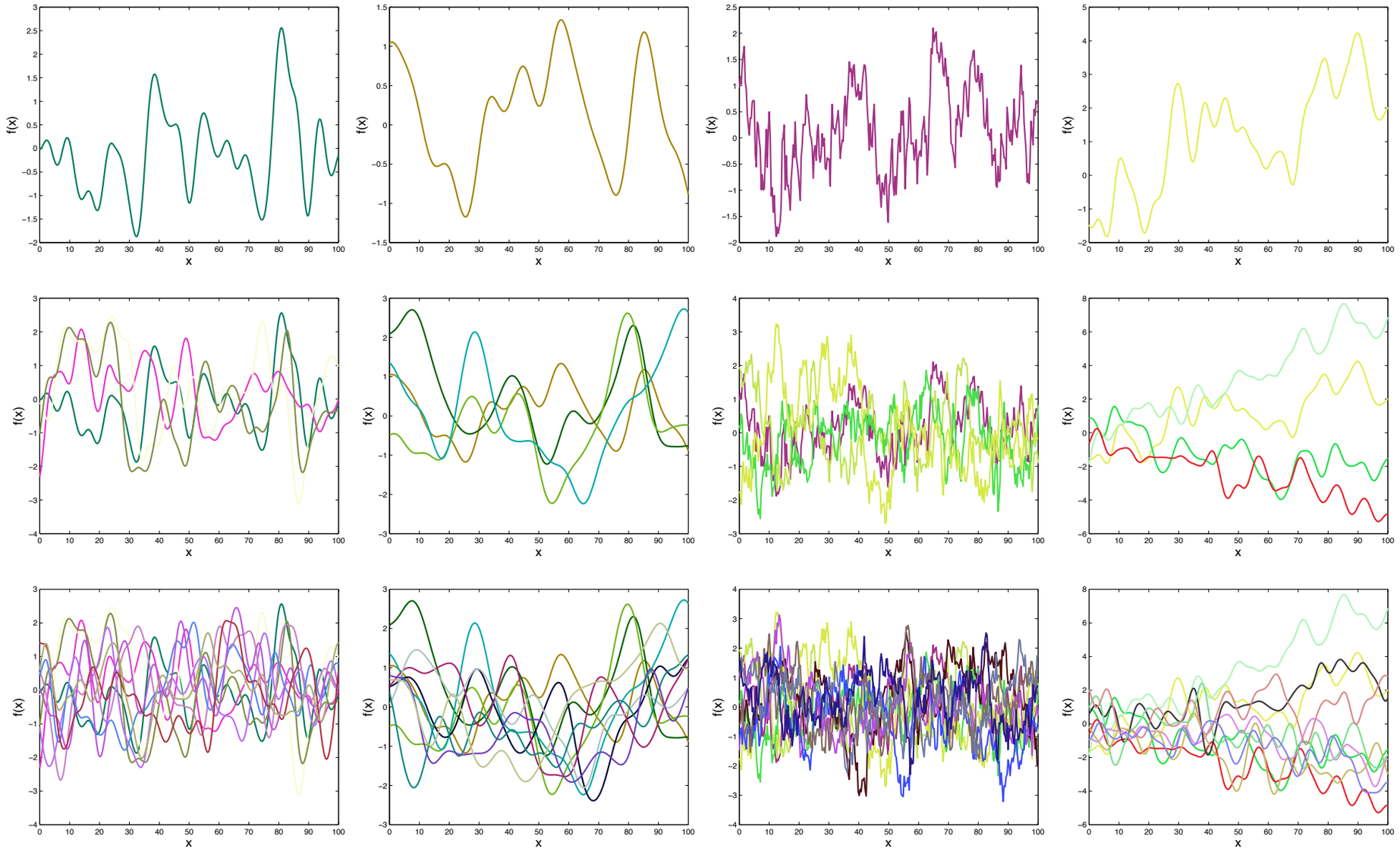
$$Y^*|X, Y \sim \mathcal{N}\left(\mu_{Y*|X,Y}, \Sigma_{Y^*|X,Y}\right)$$

where

$$\mu_{Y^*|X,Y} = \mu_{Y^*} + K(X^*, X)(K(X, X + \sigma^2 I)^{-1}(Y - \mu_Y)$$
$$\Sigma_{Y*|X,Y} = K(X^*, X^*) - K(X^*, X)(K(X, X) + \sigma^2 I)^{-1}K(X, X^*)$$

Charlie Frogner

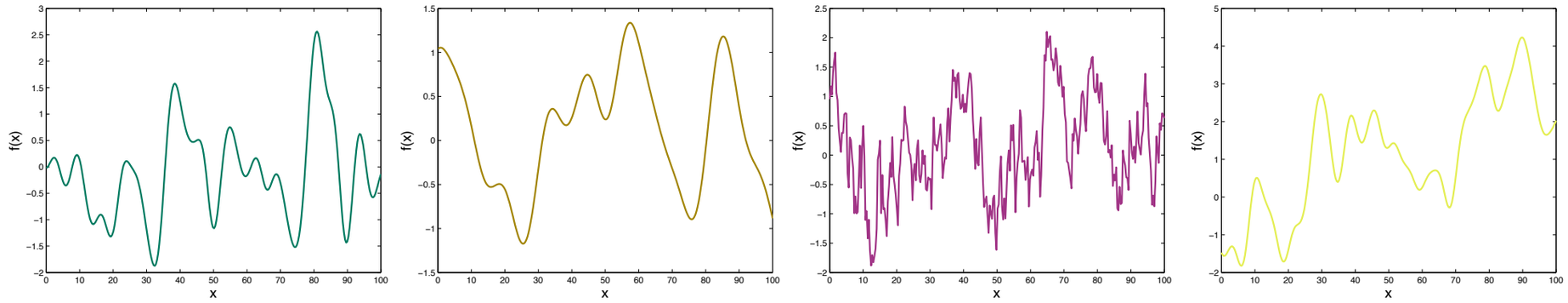
Samples from Gaussian processes with different $K(x, x')$



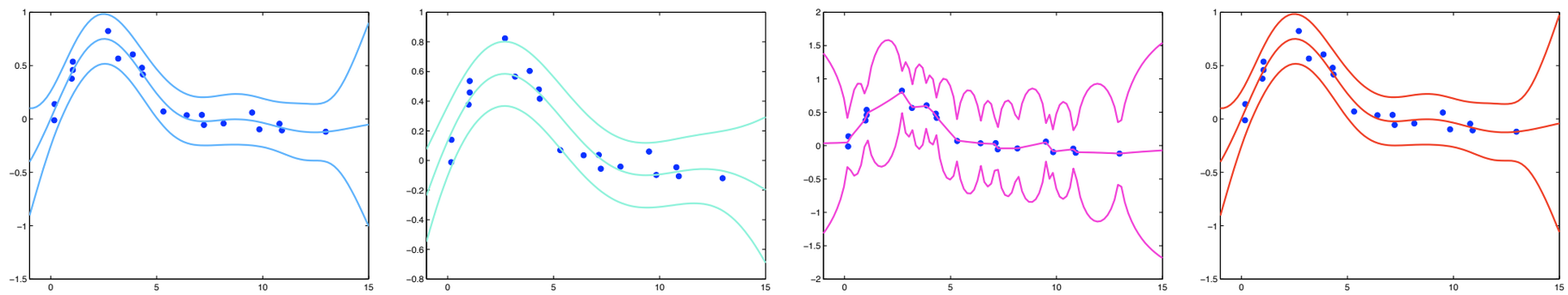
Zoubin Ghahramani

Prediction using GPs with different $K(x, x')$

A sample from the prior for each covariance function:



Corresponding predictions, mean with two standard deviations:



Zoubin Ghahramani

Squared Exponential (RBF) Kernel

$$K(x, x') = {\sigma_0}^2 \exp\left[-\frac{1}{2}\left(\frac{x - x'}{\lambda}\right)^2\right]$$

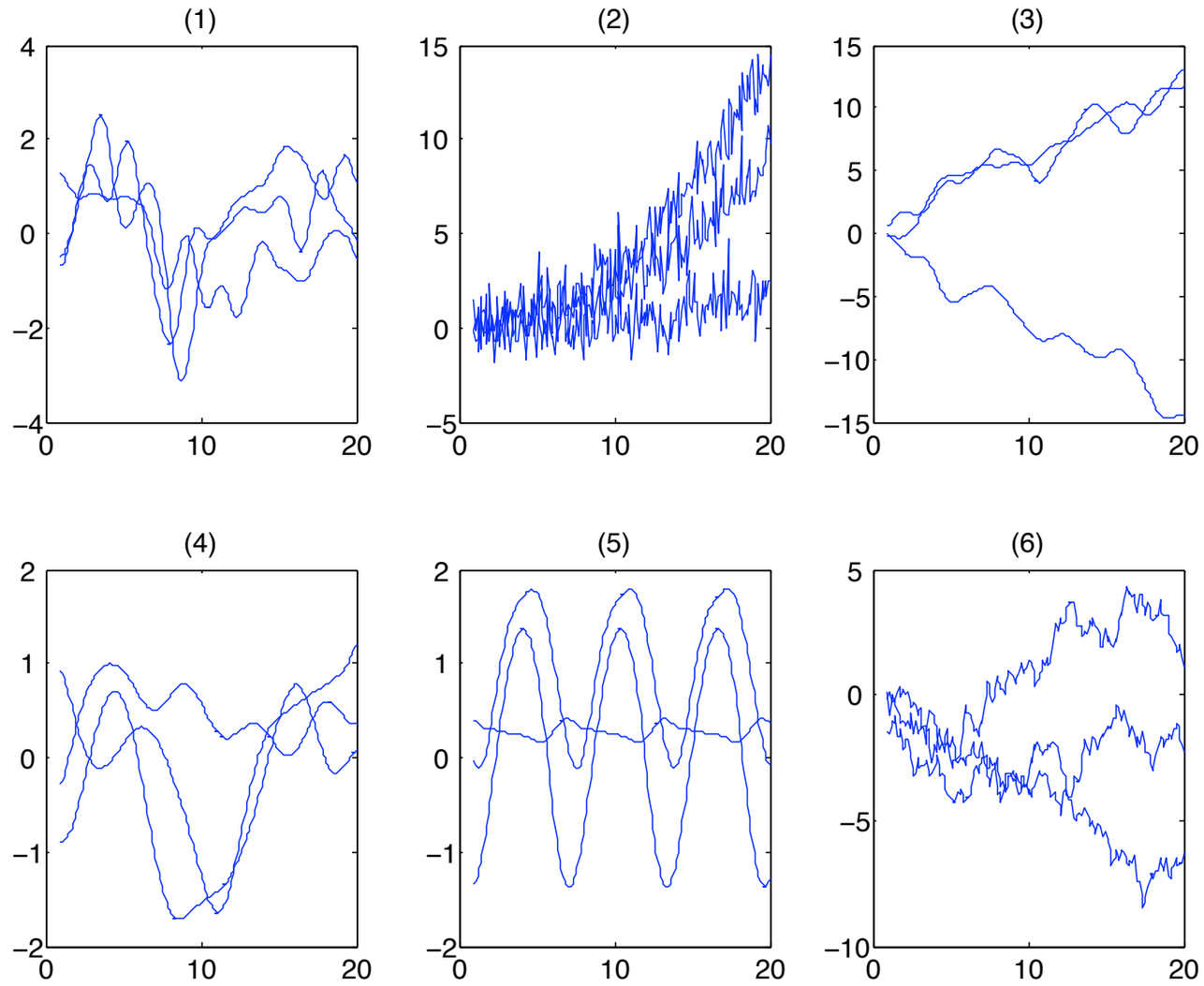
- Intuition: function variables close in input space are highly correlated, whilst those far away are uncorrelated
- λ, σ_0 — hyperparameters. λ: lengthscale, σ_0: amplitude
- Stationary: $K(x, x') = K(x - x')$ — invariant to translations
- Very smooth sample functions — infinitely differentiable

by *Carl Rasmussen*

Nonstationary Covariances

- Linear covariance: $K(x, x') = \sigma_0^2 + xx'$
- Brownian motion (Wiener process): $K(x, x') = \min(x, x')$
- Periodic covariance: $K(x, x') = \exp\left(-\frac{2\sin^2\left(\frac{x-x'}{2}\right)}{\lambda^2}\right)$

by *Carl Rasmussen*



(a) 1; $K(t,t') = \exp\big(-\frac{(t-t')^2}{2}\big)$

(b) 3; $K(t,t') = \exp\big(-\frac{(t-t')^2}{2}\big) + 0.8\,t\,t'$, second term adds linear trend

(c) 4; $K(t,t') = \exp\big(-\frac{(t-t')^2}{8}\big)$

(d) 6; $K(t,t') = \min\{t,t'\}$ (Brownian motion, $\phi_t(x) = 1(x \in [0,t])$, hence $f = \int_0^t B(x)\,dx$, where $B(x)$ is white noise)

(e) 2; $K(t,t') = \frac{1}{1000}(t\,t')^2 + \delta(t,t')$, where $\delta(t,t') = 1$ if $t = t'$, 0 otherwise.

(f) $\emptyset$; $K(t,t') = \frac{1}{1000}(t\,t')^2 + 4\,\delta(t,t')$, same as 2 (quadratic trend) but much noisier

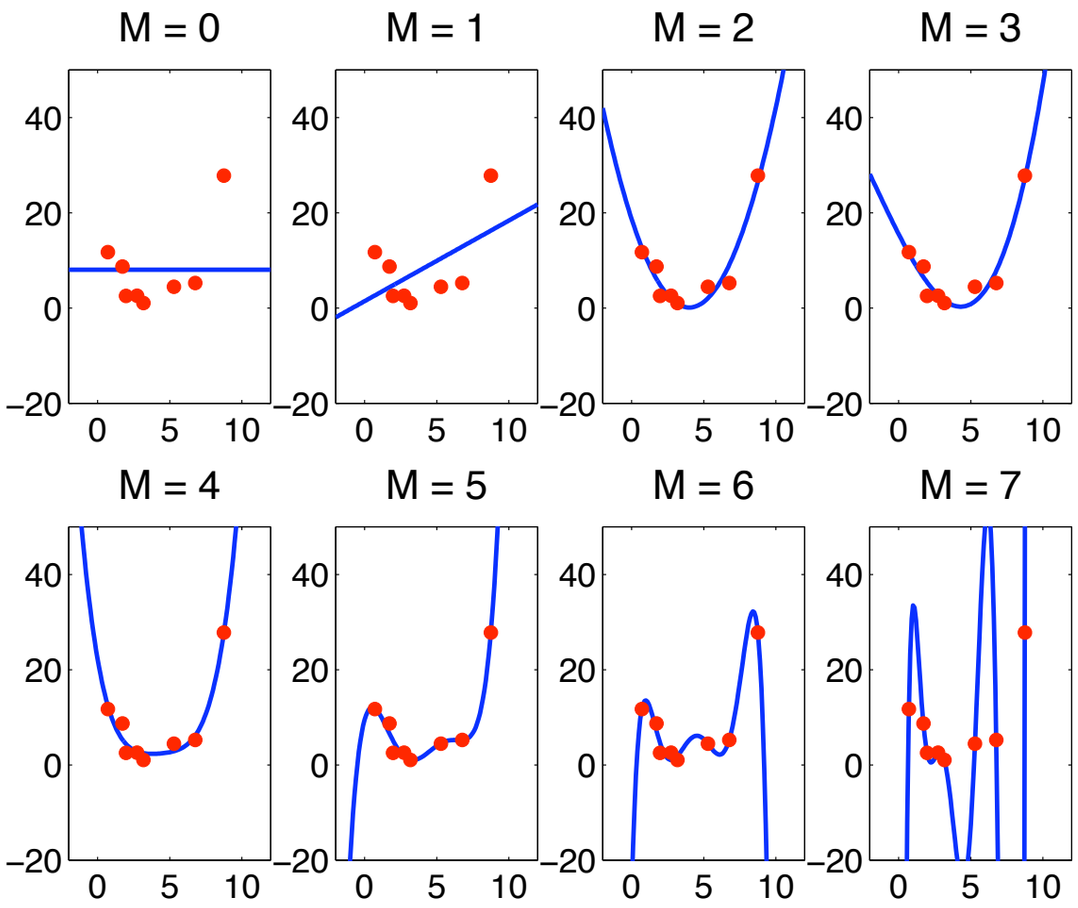
(g) 5; $K(t,t') = \sum_{n=0}^{3} e^{-n^2}\cos\big(n(t-t')\big)$, first several Fourier basis.

Which kernel?

- The previous model expressed no uncertainty in the basis $\phi(x)$ or equivalently no uncertainty in the kernel.
- Concrete example: what degree polynomial should we fit to the data?

$$K(x_i, x_j) = (x_i^T x_j + 1)^m$$

Model structure and overfitting:
A simple example: polynomial regression



Bayesian Occam's Razor and Model Comparison

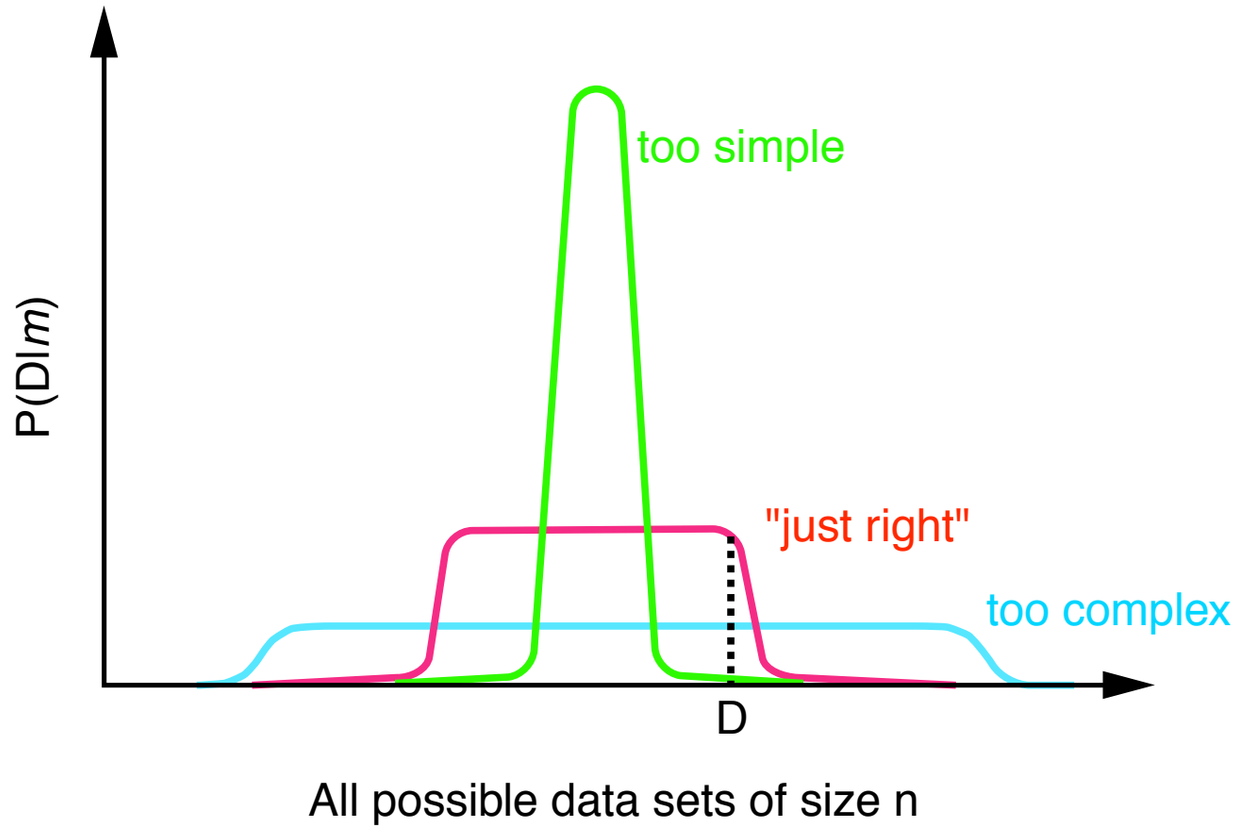
Compare model classes, e.g. m and m', using posterior probabilities given $\mathcal{D}$:

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)\, p(m)}{p(\mathcal{D})}, \qquad p(\mathcal{D}|m) = \int p(\mathcal{D}|\boldsymbol{\theta}, m)\, p(\boldsymbol{\theta}|m)\, d\boldsymbol{\theta}$$

Interpretation of the Marginal Likelihood ("evidence"): The probability that *randomly selected* parameters from the prior would generate $\mathcal{D}$.

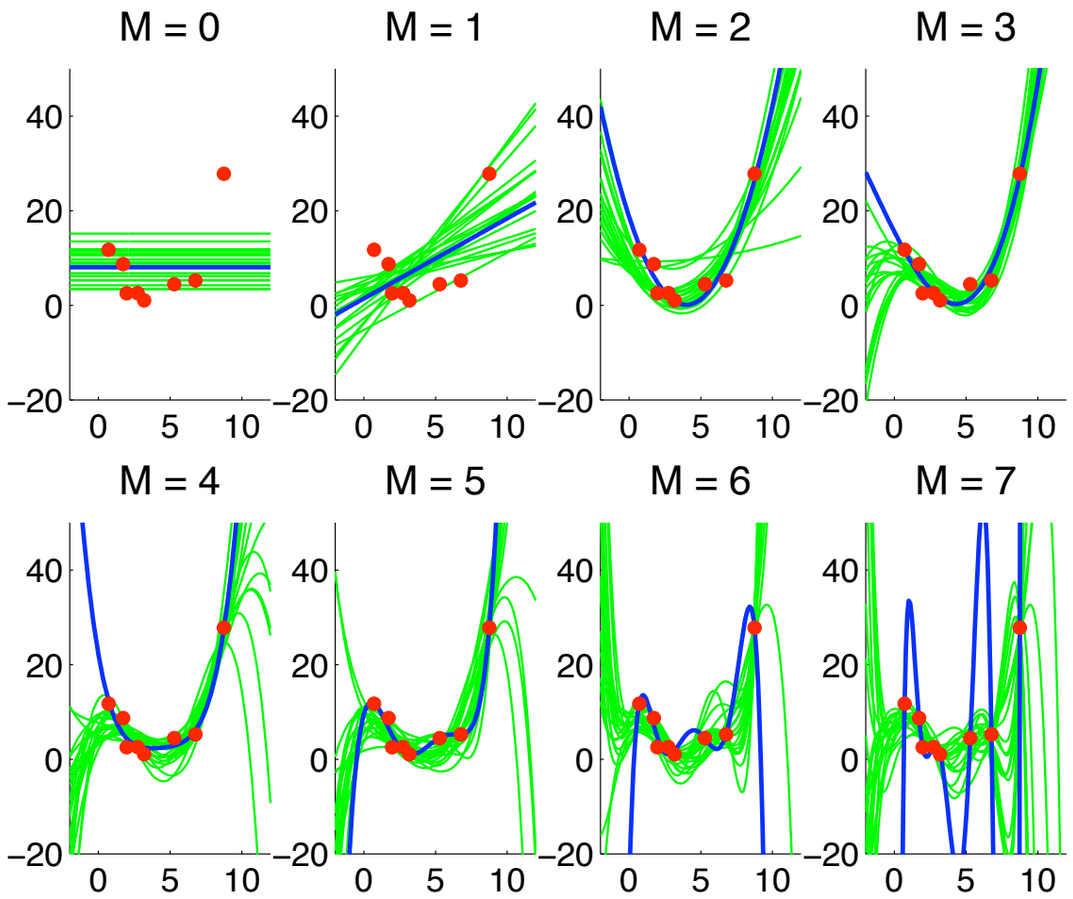
Model classes that are too simple are unlikely to generate the data set.

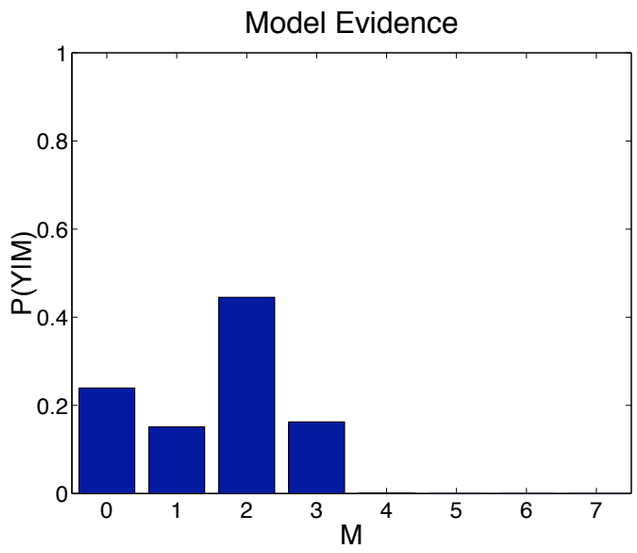
Model classes that are too complex can generate many possible data sets, so again, they are unlikely to generate that particular data set at random.



Zoubin Ghahramani

Bayesian Model Comparison: Occam's Razor at Work





$$\log p(\mathbf{y}|\mathbf{x}, M_i) = -\frac{1}{2}\mathbf{y}^\top K^{-1}\mathbf{y} - \frac{1}{2}\log|K| - \frac{n}{2}\log(2\pi)$$

Zoubin Ghahramani; Carl Rasmussen

Non-parametric Bayesian Models

- Bayesian methods are most powerful when your prior adequately captures your beliefs.

- Inflexible models (e.g. mixture of 5 Gaussians, 4th order polynomial) yield unreasonable inferences.

- Non-parametric models are a way of getting very flexible models.

- Many can be derived by starting with a finite parametric model and taking the limit as number of parameters $\rightarrow \infty$

- Non-parametric models can automatically infer an adequate model size/complexity from the data, without needing to explicitly do Bayesian model comparison.[2]

[2] Even if you believe there are infinitely many possible clusters, you can still infer how many clusters are *represented* in a finite set of n data points.

Nonparametric Bayesian Methods (Infinite Models)

We ought not to limit the complexity of our model a priori (e.g. number of hidden states, number of basis functions, number of mixture components, etc) since we don't believe that the real data was actually generated from a statistical model with a small number of parameters.

Therefore, regardless of how much training data we have, we should consider models with as many parameters as we can handle computationally.

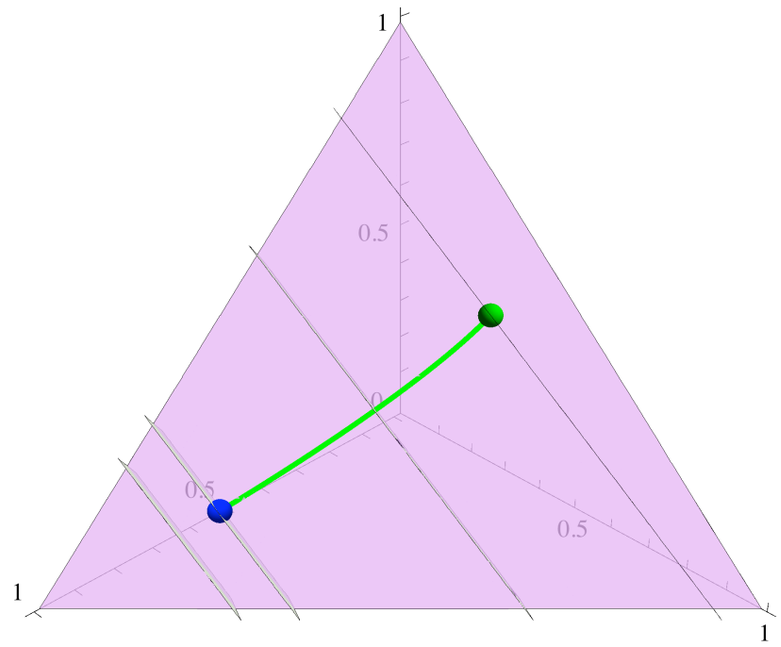
Here there is no model order selection task:

- No need to compare marginal likelihoods to select model order (which is often difficult).
- No need to use Occam's razor to limit the number of parameters in the model.

In fact, we may even want to consider doing inference in models with an infinite number of parameters...

Why Bayesian Nonparametrics?

- Finite-dimensional models are low-dimensional manifolds in probability space



- We want distributions with (nearly) full support

Nonparametric models as limits

- Look at putting priors directly on infinite dimensional RKHS
- Start with D component Fourier model

$$y(x) = a_0 + \sum_{d=1}^{D} a_d \sin(dx) + b_d \cos(dx),$$

$$\mathbf{w} = \{a_0, a_1, b_1, \ldots, a_D, b_D\}. \qquad p(\mathbf{w}|S, \mathbf{c}) \propto \exp\Big(-\frac{S}{2}\big[c_0 a_0^2 + \sum_{d=1}^{D} c_d(a_d^2 + b_d^2)\big]\Big),$$

$$K(x, x') = \Big[\sum_{d=0}^{D} \cos\big(d(x - x')\big)/c_d\Big]/S.$$

Rasmussen&Ghahramani NIPS*2001

Still finite, but something amiss

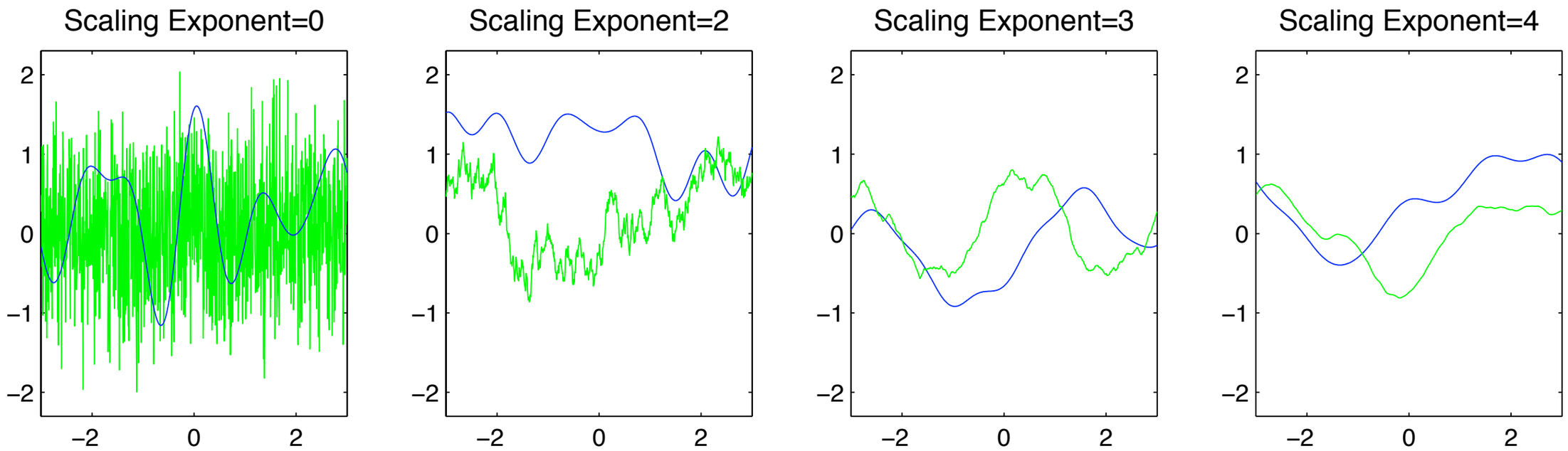


Figure 3: Functions drawn at random from the Fourier model with order $D = 6$ (dark) and $D = 500$ (light) for four different scalings; limiting behaviour from left to right: discontinuous, Brownian, borderline smooth, smooth.

Parameter scales fixed; model size increasing

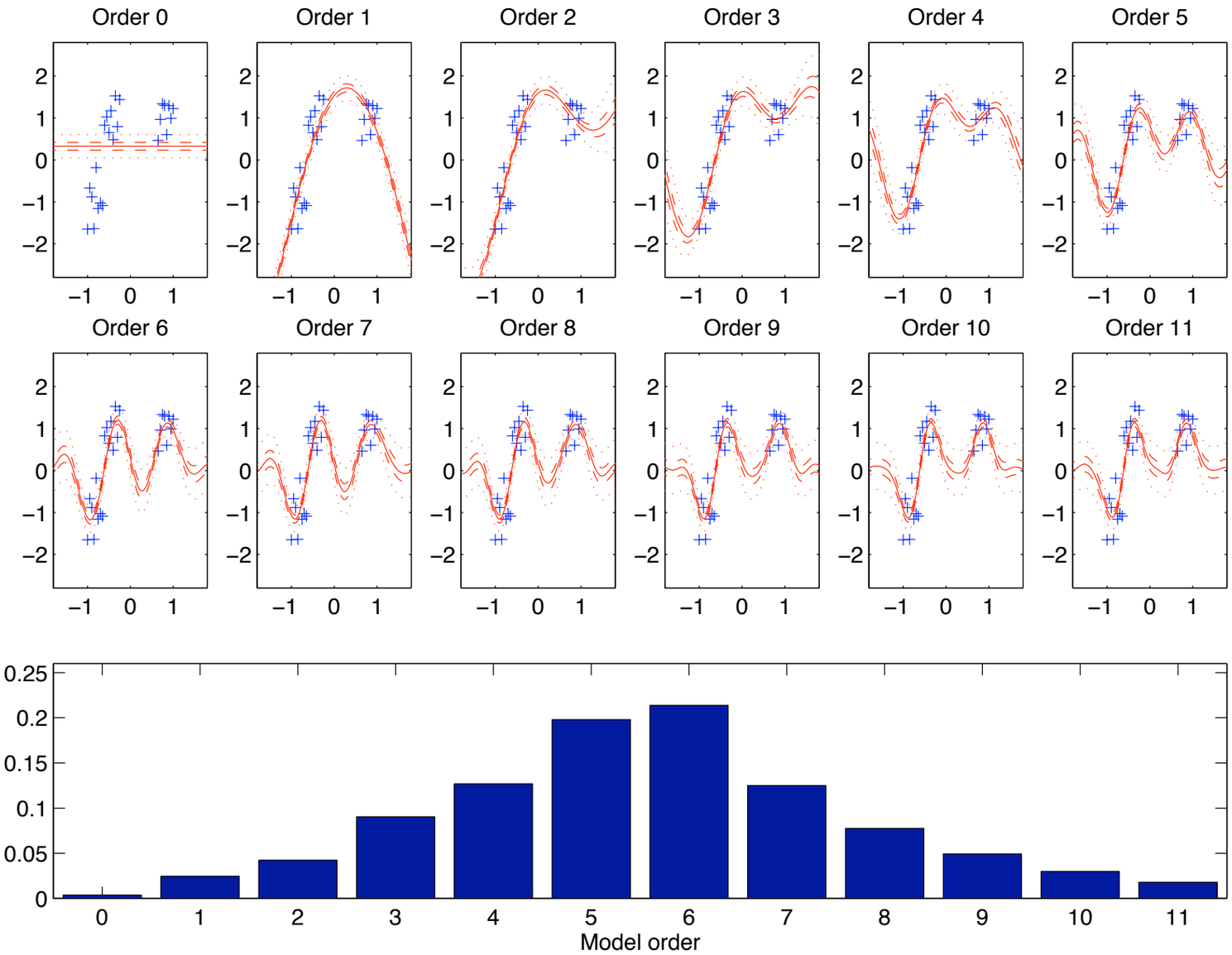


Figure 2: Top: 12 different model orders for the "unscaled" model: $c_d \propto 1$. The mean predictions are shown with a full line, the dashed and dotted lines limit the 50% and 95% central mass of the predictive distribution (which is student-t). Bottom: posterior probability of the models, normalised over the 12 models. The probabilities of the models exhibit an Occam's Hill, discouraging models that are either "too small" or "too big".

Rasmussen&Ghahramani NIPS*2001

Scaling the parameters with model size

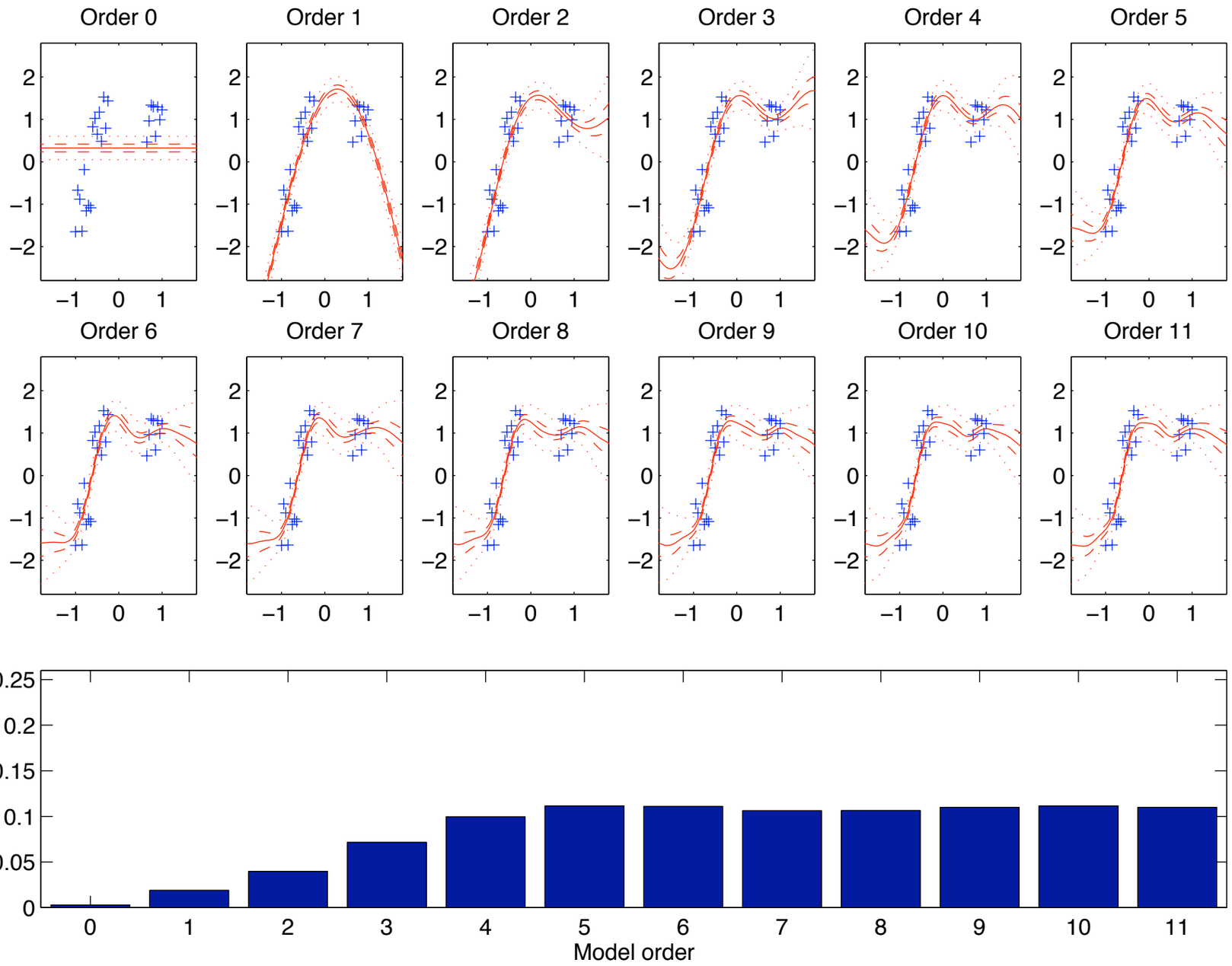
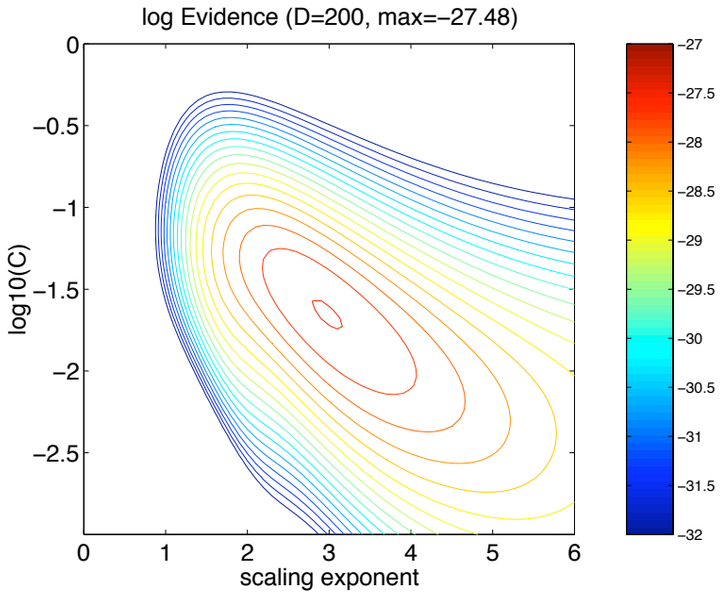


Figure 4: The same as figure 2, except that the scaling $c_d = d^3$ was used here, leading to a prior which converges to smooth functions as $D \to \infty$. There is no Occam's Razor; instead we see that as long as the model is complex enough, the evidence is flat. We also notice that the predictive density of the model is unchanged as long as D is sufficiently large.

Rasmussen&Ghahramani NIPS*2001

Scaling the parameters with model size

$$G(\Delta) = E[(f(x) - f(x+\Delta))^2],$$

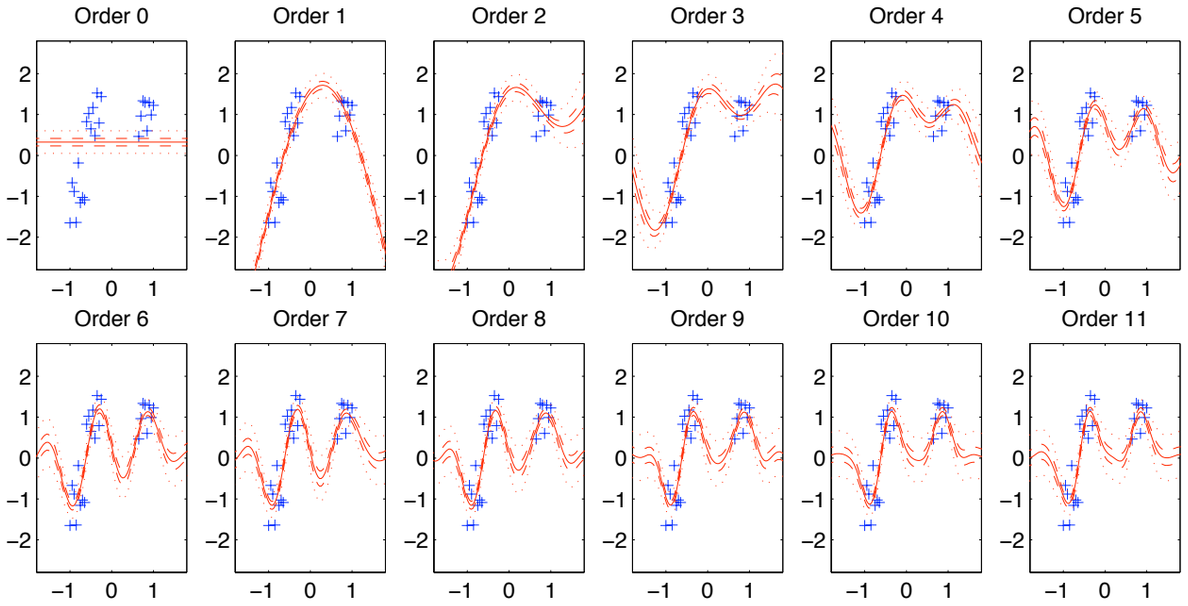


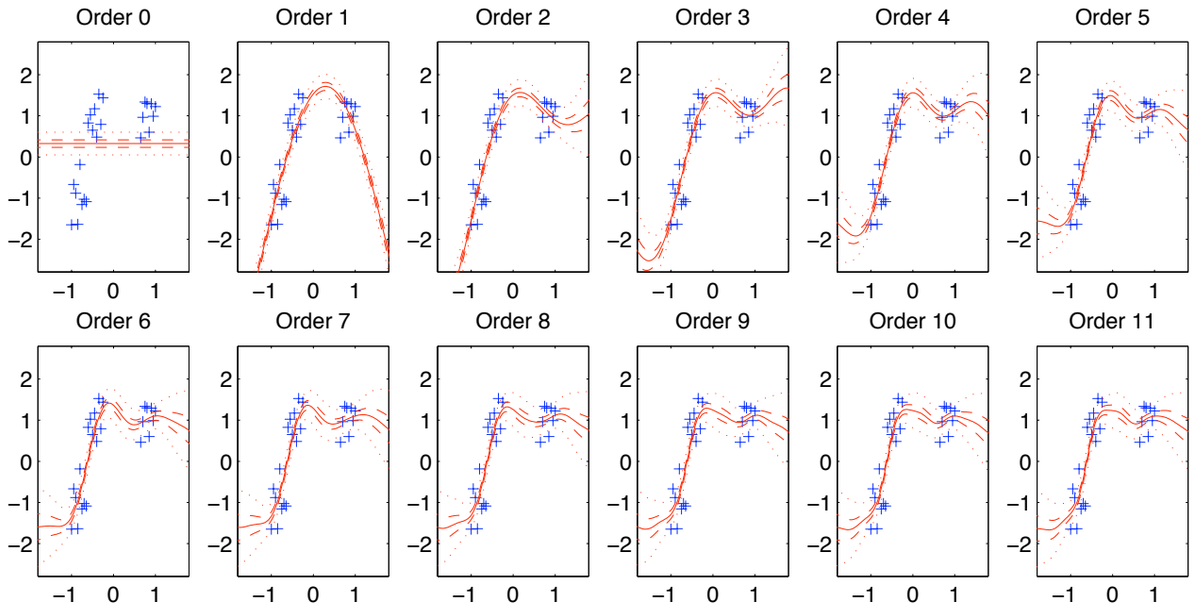
γ	$\lim_{\Delta\to 0} G(\Delta)$	properties
≤ 1	1	discontinuous
2	Δ	Brownian
3	$\Delta^2(1-\ln\Delta)$	borderline smooth
> 3	Δ^2	smooth

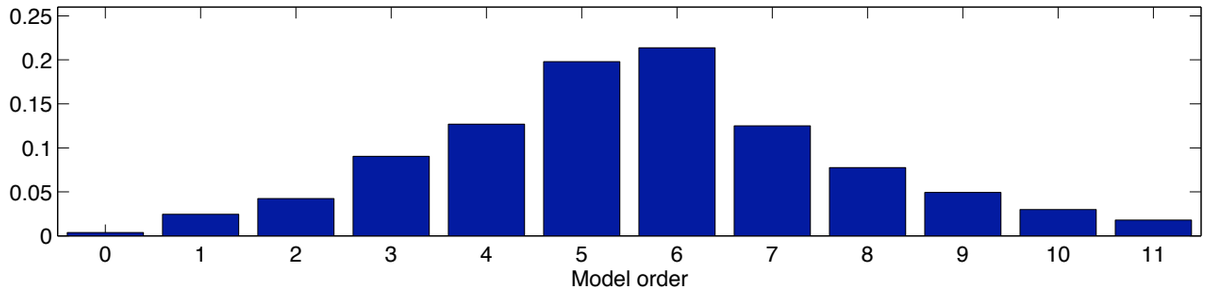
Figure 5: Left panel: the evidence as a function of the scaling exponent, γ and overall scale C, has a maximum at $\gamma = 3$. The table shows the characteristics of functions for different values of γ. Examples of these functions are shown in figure 3.

Rasmussen&Ghahramani NIPS*2001

Scaling the parameters with model size







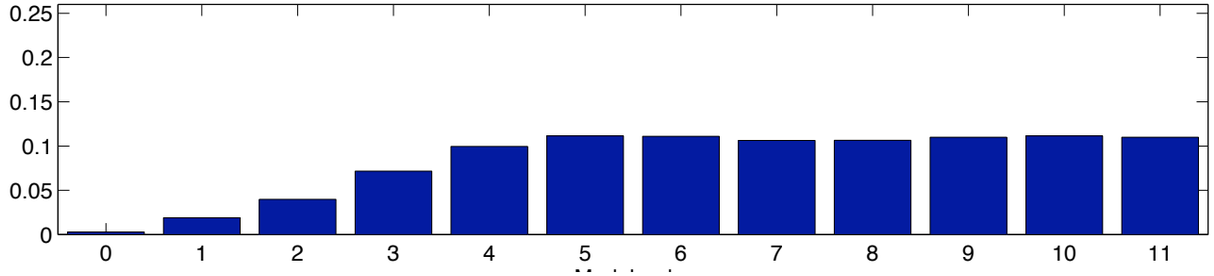


Figure 2: Top: 12 different model orders for the "unscaled" model: $c_d \propto 1$. The mean predictions are shown with a full line, the dashed and dotted lines limit the 50% and 95% central mass of the predictive distribution (which is student-t). Bottom: posterior probability of the models, normalised over the 12 models. The probabilities of the models exhibit an Occam's Hill, discouraging models that are either "too small" or "too big".

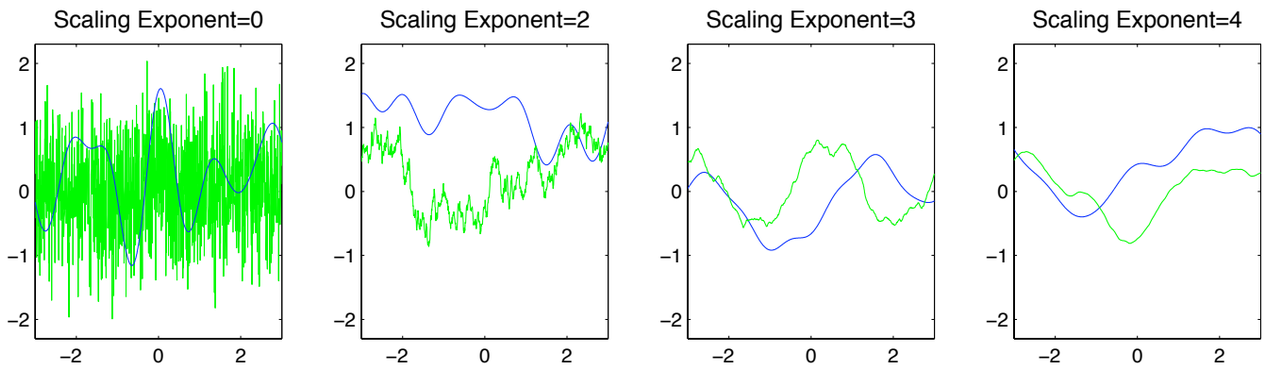
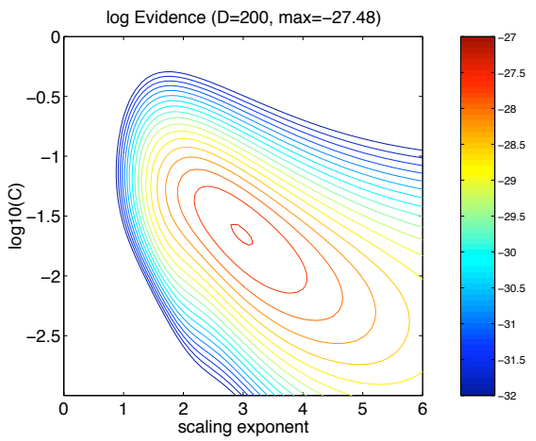


Figure 3: Functions drawn at random from the Fourier model with order $D = 6$ (dark) and $D = 500$ (light) for four different scalings; limiting behaviour from left to right: discontinuous, Brownian, borderline smooth, smooth.



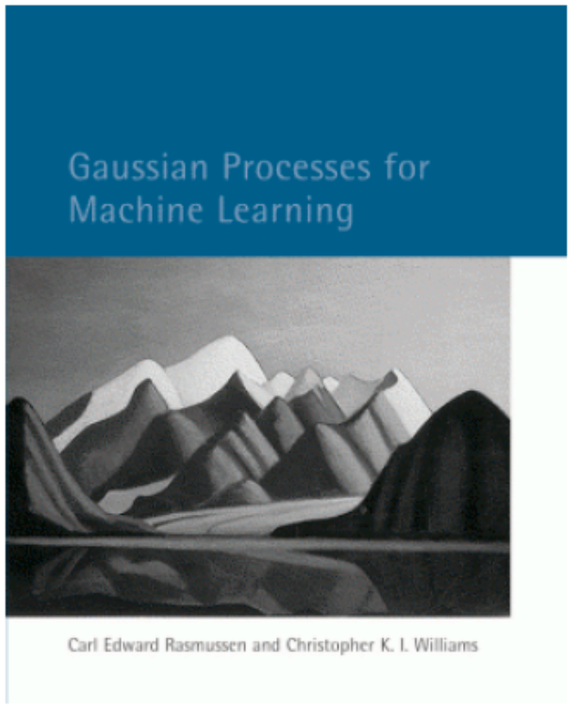
γ	$\lim_{\Delta \to 0} G(\Delta)$	properties
≤ 1	1	discontinuous
2	Δ	Brownian
3	$\Delta^2(1 - \ln \Delta)$	borderline smooth
> 3	Δ^2	smooth

Figure 5: Left panel: the evidence as a function of the scaling exponent, γ and overall scale C, has a maximum at $\gamma = 3$. The table shows the characteristics of functions for different values of γ. Examples of these functions are shown in figure 3.

Rasmussen&Ghahramani NIPS*2001

The GP Bible (for ML folk)

all the chapters
are available online!



The GP book: Rasmussen and Williams, 2006

Basic GP (Matlab) code available:
http://www.gaussianprocess.org/gpml/

Take-home

- Inference in Nonparametric models does not require model selection as there are an infinite number of parameters to fit
    - Nonparametric models essentially turn a structure learning problem (how many components) into a parameter estimation problem

- Gaussian Processes are a fully Bayesian alternative to RLS
  Provides error bars on predictions; marginal likelihood tractable
  Structure in the kernel induces structure in the output
  Kernel composition laws provide a rich space of models
  Cubic/training, linear/test performance

- A Bayesian machine learning approach explicitly models uncertainty by treating unknown variables as random

References

- Ghahramani slides taken from ICML 2004 tutorial on Bayesian Machine Learning and UAI 2005 tutorial on Nonparametric Bayesian Methods. Moghaddam slides taken from Lecture 17 of Machine Learning class at CalTech.

- [GMRBT2008] Goodman, Mansinghka, Roy, Bonawitz, Tenenbaum. Church: a language for generative models. *Uncertainty in Artificial Intelligence*, 2008.