Model Selection for Gaussian Processes

Chris Williams

Institute for Adaptive and Neural Computation
School of Informatics, University of Edinburgh, UK

December 2006

Outline

- GP basics
- Model selection: covariance functions and parameterizations
- Criteria for model selection
- Marginal likelihood
- Cross-validation
- Examples
- Thanks to Carl Rasmussen (book co-author)

Features

- ✓ Bayesian learning
- ✓ Higher level regularization
- ✓ Predictive uncertainty
- ✓ Cross-validation
- ✓ Ensemble learning
- ✓ Search strategies (no more grid search)
- ✓ Feature selection
- ✓ More than 2 levels of inference

Gaussian Process Basics

- For a stochastic process $f(\mathbf{x})$, mean function is

$$\mu(\mathbf{x}) = E[f(\mathbf{x})].$$

  Assume $\mu(\mathbf{x}) \equiv 0 \ \forall \mathbf{x}$
- Covariance function

$$k(\mathbf{x}, \mathbf{x}') = E[f(\mathbf{x})f(\mathbf{x}')].$$

- Priors over function-space can be defined directly by choosing a covariance function, e.g.

$$E[f(\mathbf{x})f(\mathbf{x}')] = \exp(-w|\mathbf{x} - \mathbf{x}'|)$$

- Gaussian processes are stochastic processes defined by their mean and covariance functions.

Gaussian Process Prediction

- A Gaussian process places a prior over functions
- Observe data $\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^n$, obtain a posterior distribution

$$p(f|D) \propto p(f)p(D|f)$$

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

- For a Gaussian likelihood (regression), predictions can be made exactly via matrix computations
- For classification, we need approximations (or MCMC)

GP Regression

- Prediction at $\mathbf{x}_* \sim \mathcal{N}(\bar{f}(\mathbf{x}_*), \mathrm{var}(\mathbf{x}_*))$, with

$$\bar{f}(\mathbf{x}_*) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_*, \mathbf{x}_i)$$

where

$$\boldsymbol{\alpha} = (K + \sigma^2 I)^{-1} \mathbf{y}$$

and

$$\mathrm{var}(\mathbf{x}_*) = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}^T(\mathbf{x}_*)(K + \sigma^2 I)^{-1}\mathbf{k}(\mathbf{x}_*)$$

with $\mathbf{k}(\mathbf{x}_*) = (k(\mathbf{x}_*, \mathbf{x}_1), \ldots, \mathbf{k}(\mathbf{x}, \mathbf{x}_n))^T$

GP classification

- Response function $\pi(\mathbf{x}) = r(f(\mathbf{x}))$, e.g. logistic or probit
- For these choices log likelihood is concave $\Rightarrow$ unique maximum
- Use MAP or EP for inference (unconstrained optimization, c.f. SVMs)
- For GPR and GPC, some consistency results are available for non-degenerate kernels

Model Selection

- Covariance function often has some free parameters $\boldsymbol{\theta}$
- We can choose different families of covariance functions

$$k_{SE}(\mathbf{x}_p, \mathbf{x}_q) = \sigma_f^2 \exp\left(-\frac{1}{2}(\mathbf{x}_p - \mathbf{x}_q)^\top M(\mathbf{x}_p - \mathbf{x}_q)\right) + \sigma_n^2 \delta_{pq},$$

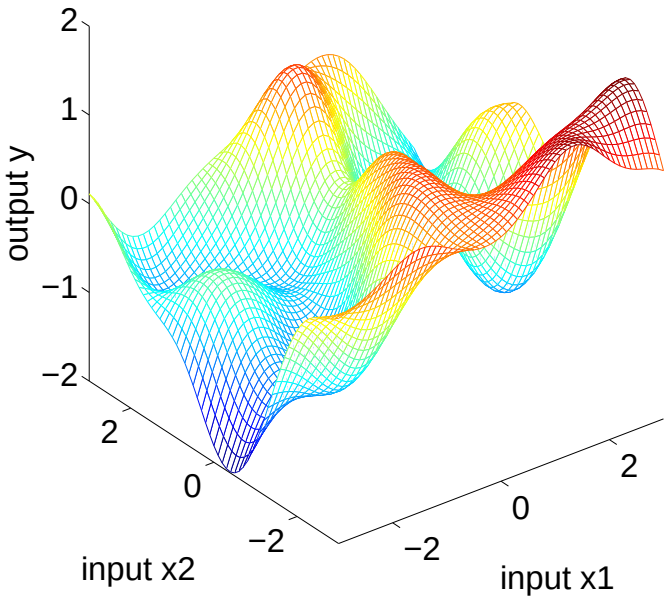
$$k_{\mathrm{NN}}(\mathbf{x}_p, \mathbf{x}_q) = \sigma_f^2 \sin^{-1}\left(\frac{2\tilde{\mathbf{x}}_p^\top M \tilde{\mathbf{x}}_q}{\sqrt{(1 + 2\tilde{\mathbf{x}}_p^\top M \tilde{\mathbf{x}}_p)(1 + 2\tilde{\mathbf{x}}_q'^\top M \tilde{\mathbf{x}}_q)}}\right) + \sigma_n^2 \delta_{pq},$$

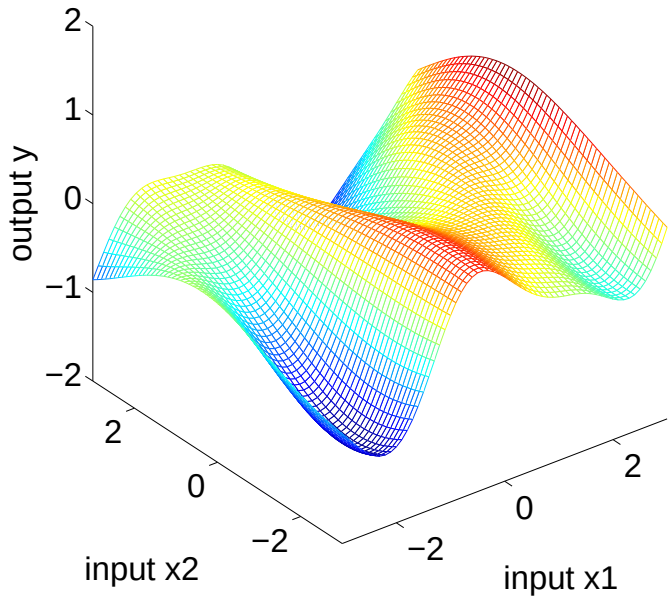
where $\tilde{\mathbf{x}} = (1, x_1, \ldots, x_D)^\top$

Automatic Relevance Determination

$$k_{SE}(\mathbf{x}_p, \mathbf{x}_q) = \sigma_f^2 \exp\left(-\frac{1}{2}(\mathbf{x}_p - \mathbf{x}_q)^\top M(\mathbf{x}_p - \mathbf{x}_q)\right) + \sigma_n^2 \delta_{pq}$$

- Isotropic $M = \ell^{-2} I$
- ARD: $M = \mathrm{diag}(\ell_1^{-2}, \ell_2^{-2}, \ldots, \ell_D^{-2})$





Further modelling flexibility

- We can *combine* covariance functions to make things more general
- Example, functional ANOVA, e.g.

$$k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{D} k_i(x_i, x_i') + \sum_{i=2}^{D} \sum_{j=1}^{i-1} k_{ij}(x_i, x_j; x_i', x_j')$$

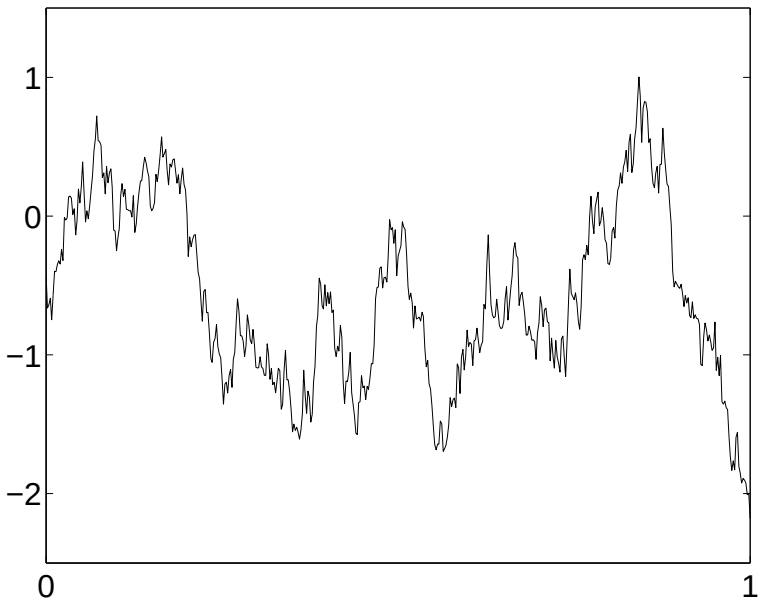
- Non-linear warping of the input space (Sampson and Guttorp, 1992)

The Baby and the Bathwater

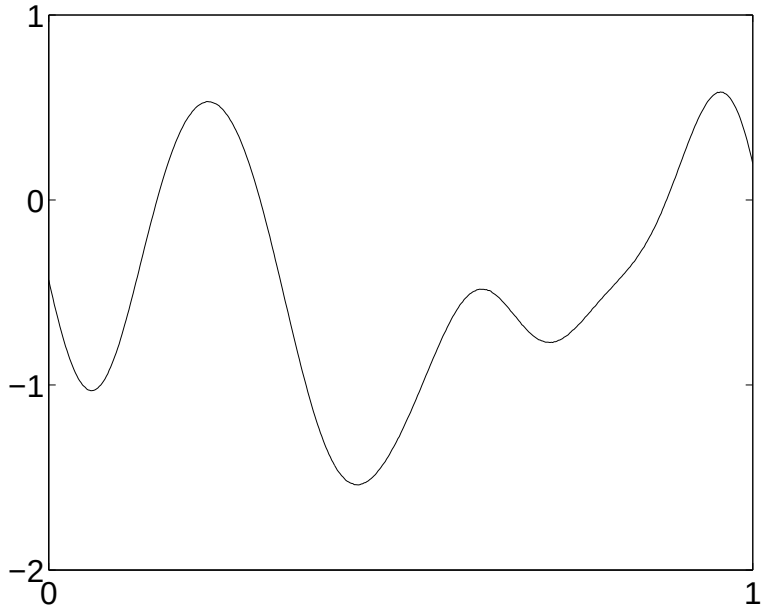
- MacKay (2003 ch 45): In moving from neural networks to kernel machines did we throw out the baby with the bathwater? i.e. the ability to learn hidden features/representations
- But consider $M = \Lambda\Lambda^\top + \mathrm{diag}(\boldsymbol{\ell})^{-2}$ for Λ being $D \times k$, for $k < D$
- The k columns of Λ can identify directions in the input space with specially high relevance (Vivarelli and Williams, 1999)

Understanding the prior

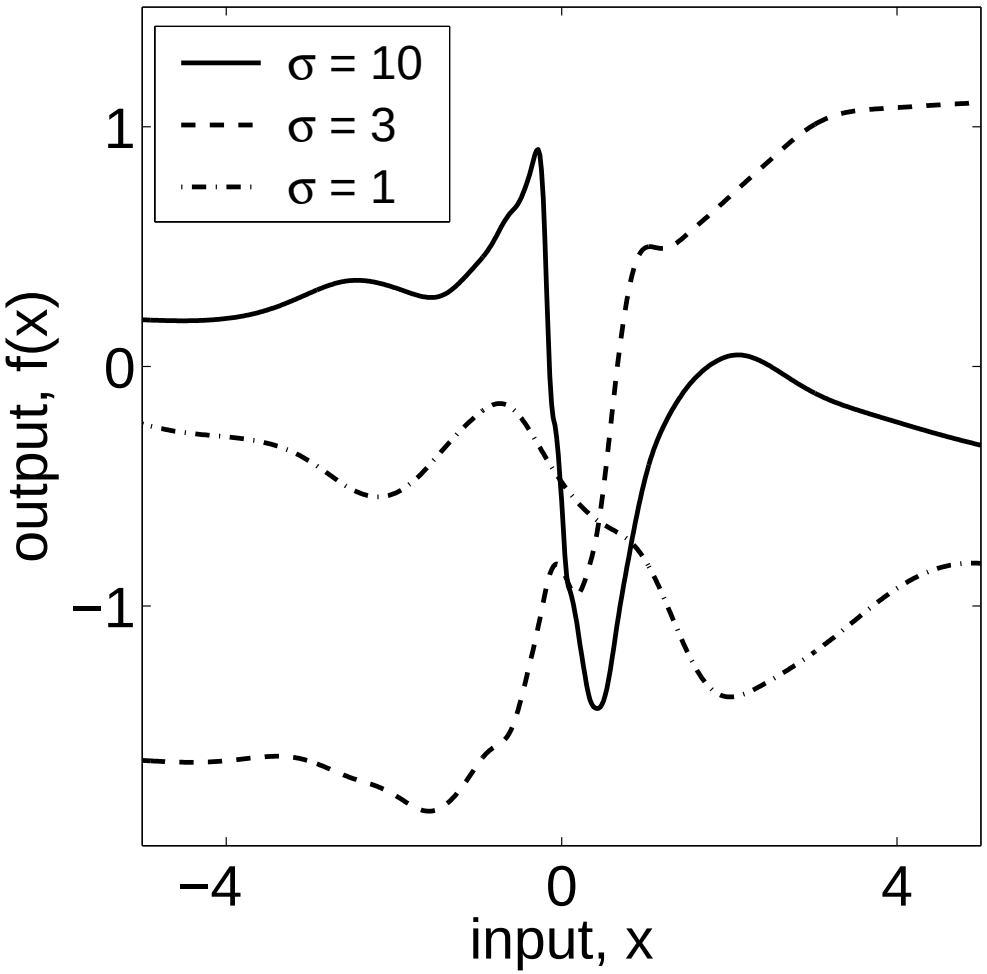
- We can analyze and draw samples from the prior



$k(x - x') = \exp -|x - x'|/\ell$
with $\ell = 0.1$



$k(x - x') = \exp -|x - x'|^2/2\ell^2$



Samples from the neural network covariance function

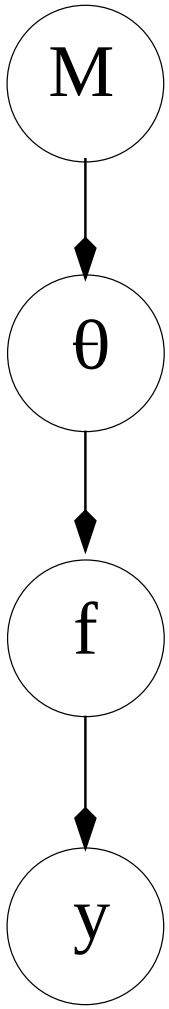
Criteria for Model Selection

- Bayesian marginal likelihood ("evidence") $p(\mathcal{D}|\text{model})$
- Estimate the generalization error (e.g. cross-validation)
- Bound the generalization error (e.g. PAC-Bayes)

Bayesian Model Selection

- For GPR, we can compute the marginal likelihood $p(\mathbf{y}|X, \boldsymbol{\theta}, M)$ exactly (integrating out $\mathbf{f}$). For GPC it can be approximated using Laplace approx or EP
- Can also use MCMC to sample
- $p(M_i|\mathbf{y}, X) \propto p(\mathbf{y}|X, M_i)p(M_i)$ where

$$p(\mathbf{y}|X, M_i) = \int p(\mathbf{y}|X, \boldsymbol{\theta}_i, M_i)p(\boldsymbol{\theta}_i|M_i)\, d\boldsymbol{\theta}_i$$

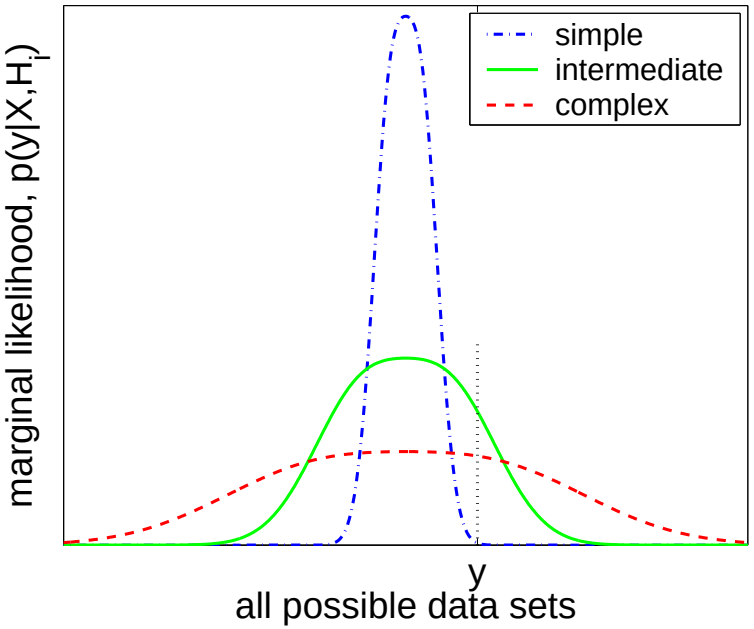


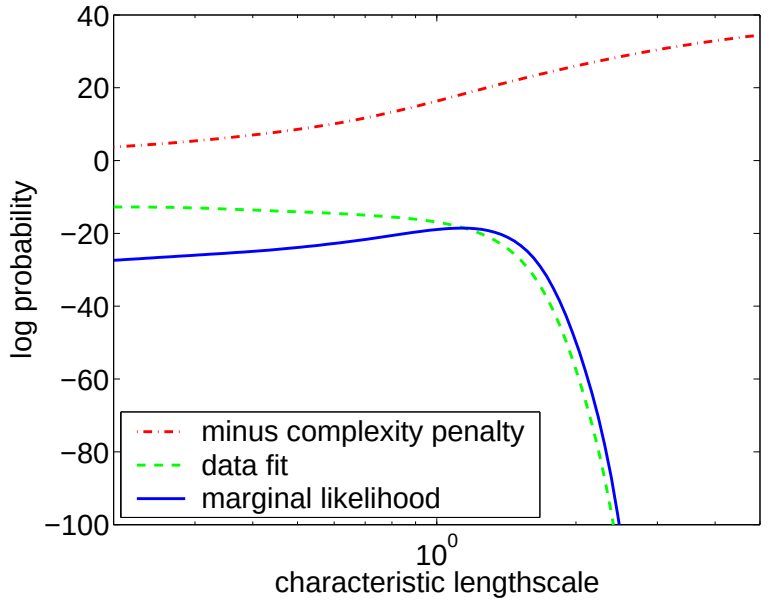
Type-II Maximum Likelihood

- Type-II maximum likelihood maximizes the marginal likelihood $p(\mathbf{y}|X, \boldsymbol{\theta}_i, M_i)$ rather than integrates
- $p(\mathbf{y}|X, \boldsymbol{\theta}_i, M_i)$ is differentiable wrt $\boldsymbol{\theta}$: no more grid search!

$$\frac{\partial}{\partial \theta_j} \log p(\mathbf{y}|X, \boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^\top K^{-1} \frac{\partial K}{\partial \theta_j} K^{-1}\mathbf{y} - \frac{1}{2}\mathrm{trace}\left(K^{-1}\frac{\partial K}{\partial \theta_j}\right)$$

- This was in Williams and Rasmussen (NIPS*95)
- Can also use MAP estimation or MCMC in $\boldsymbol{\theta}$-space, see e.g. Williams and Barber (1998)

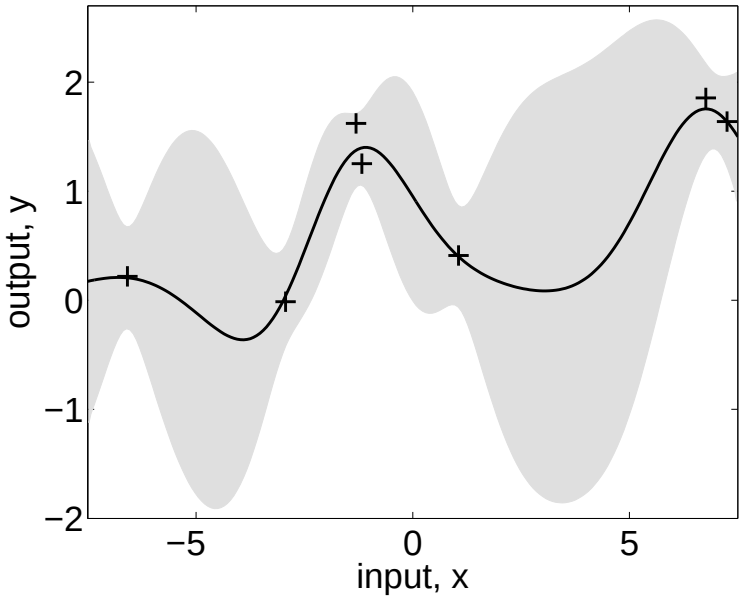


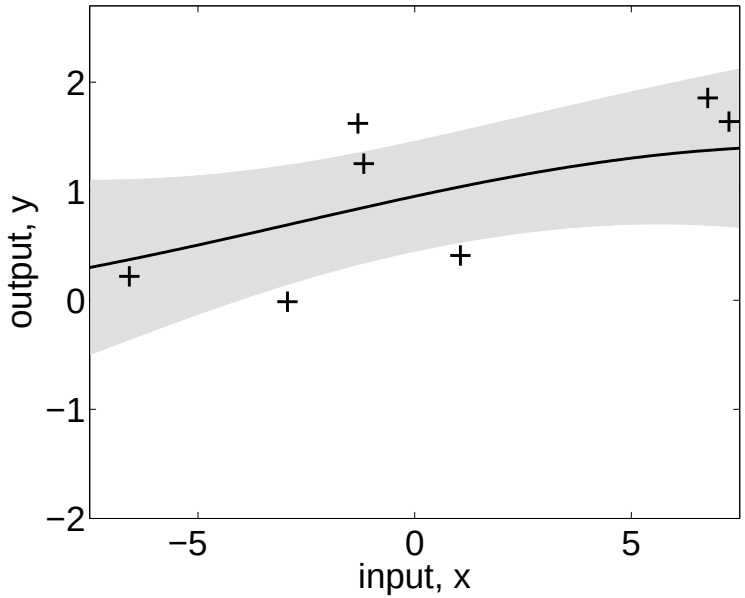


- Marginal likelihood automatically incorporates a trade-off between model fit and model complexity

$$\log p(\mathbf{y}|X, \boldsymbol{\theta}_i, M_i) = -\frac{1}{2}\mathbf{y}^T K_y^{-1}\mathbf{y} - \frac{1}{2}\log|K_y| - \frac{n}{2}\log 2\pi$$

Marginal Likelihood and Local Optima





- There can be multiple optima of the marginal likelihood
- These correspond to different interpretations of the data

Cross-validation for GPR

$$\log p(y_i|X, \mathbf{y}_{-i}, \boldsymbol{\theta}) = -\frac{1}{2}\log(2\pi\sigma_i^2) - \frac{(y_i - \mu_i)^2}{2\sigma_i^2}$$

$$L_{\mathrm{LOO}}(X, \mathbf{y}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(y_i|X, \mathbf{y}_{-i}, \boldsymbol{\theta})$$

- Leave-one-out predictions can be made efficiently (e.g. Wahba, 1990; Sundararajan and Keerthi, 2001)
- We can also compute derivatives $\partial L_{LOO}/\partial\theta_j$
- LOO for squared error ignores predictive variances, and does not determine overall scale of the covariance function
- For GPC LOO computations are trickier, but the cavity method (Opper and Winther, 2000) seems to be effective

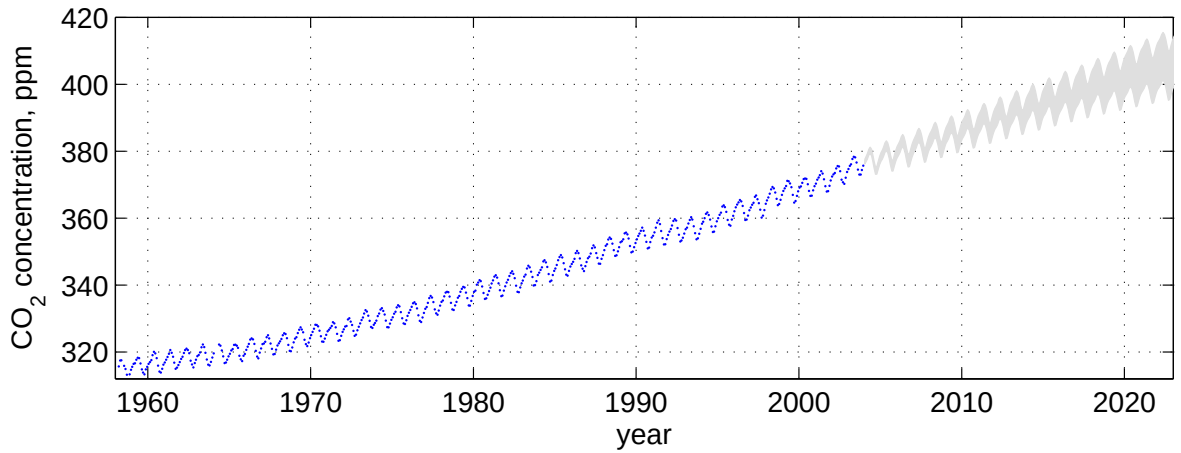
Comparing marginal likelihood and LOO-CV

$$L = \sum_{i=1}^{n} \log p(y_i | \{y_j,\ j < i\}, \boldsymbol{\theta})$$

$$L_{LOO} = \sum_{i=1}^{n} \log p(y_i | \{y_j,\ j \neq i\}, \boldsymbol{\theta})$$

- Marginal likelihood tells us the probability of the data given the assumptions of the model
- LOO-CV gives an estimate of the predictive log probability, whether or not the model assumptions are fulfilled
- CV procedures should be more robust against model mis-specification (e.g. Wahba, 1990)

Example 1: Mauna Loa CO_2

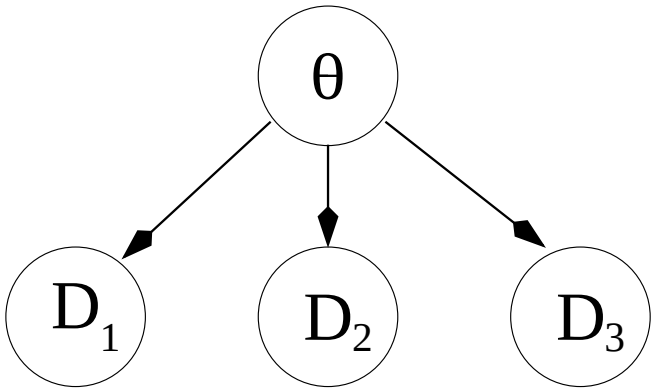


- Fit this data with sum of four covariance functions, modelling (i) smooth trend (ii) periodic component (with some decay) (iii) medium term irregularities (iv) noise
- Optimize and compare models using marginal likelihood

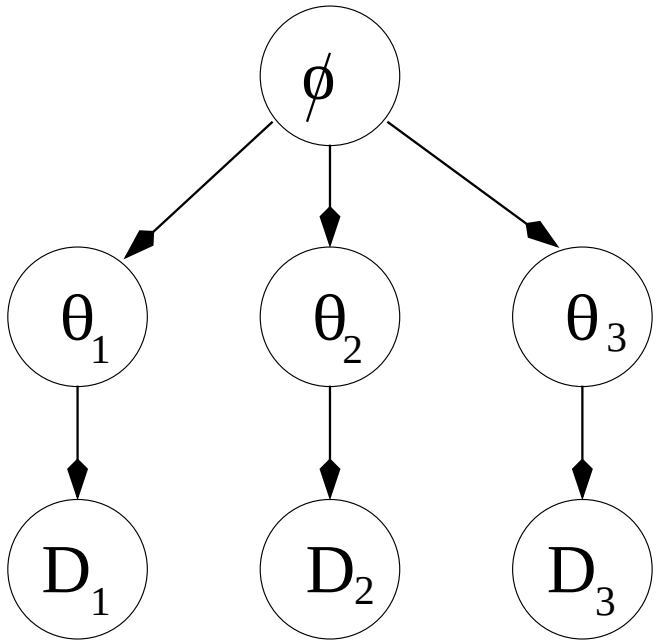
Example 2: Robot Arm Inverse Dynamics

- 44,484 training, 4,449 test examples, in 21-dimensions
- Map from 7 joint positions, velocities and accelerations of 7 joints to torque
- Use SE (Gaussian) covariance function with ARD $\Rightarrow$ 23 hyperparameters, optimizing marginal likelihood or L_{LOO}
- Similar accuracy for both SMSE and mean standardized log loss, but marginal likelihood optimzation is quicker

Multi-task Learning



Minka and Picard (1999)



Summary

- ✓ Bayesian learning
- ✓ Higher level regularization
- ✓ Predictive uncertainty
- ✓ Cross-validation
- ✓ Ensemble learning
- ✓ Search strategies (no more grid search)
- ✓ Feature selection
- ✓ More than 2 levels of inference
- Model selection is much more than just setting parameters in a covariance function

Relationship to alignment

$$A(K, \mathbf{y}) = \frac{\mathbf{y}^\top K \mathbf{y}}{n \|K\|_F}$$

where $\mathbf{y}$ has $+1$/-1 elements

$$\log A(K, \mathbf{y}) = \log \left(\mathbf{y}^\top K \mathbf{y} \right) - \log \mathrm{tr}(K)$$

$$\log q(\mathbf{y}|K) = -\frac{1}{2} \hat{\mathbf{f}}^\top K^{-1} \hat{\mathbf{f}} + \log p(\mathbf{y}|\hat{\mathbf{f}}) - \frac{1}{2} \log |B|,$$

where $B = I + W^{\frac{1}{2}} K W^{\frac{1}{2}}$, $\hat{\mathbf{f}}$ is the maximum of the posterior found by Newton's method, and W is the diagonal matrix $W = -\nabla\nabla \log p(\mathbf{y}|\hat{\mathbf{f}})$.