Bayesian Linear Regression: Model and Variable Selection, Equivalent Kernel Representation

Prof. Nicholas Zabaras
University of Notre Dame
Notre Dame, IN, USA

Email: nzabaras@gmail.com
URL: https://www.zabaras.com/

September 20, 2017

Contents

- Bayesian model comparison, Model Averaging and Model Selection, Model Complexity
- The evidence approximation for our regression example
- Another example of computing model evidence
- Limitations of fixed basis functions
- Laplace approximation, BIC criterion, Another Regression example and MatLab implementation of model selection
- Equivalent Kernel Representation
- Introduction to Variable Selection

Following closely Chris Bishops' PRML book, Chapter 3

Bayesian Model Comparison

- The Bayesian view of model comparison involves the use of probabilities to represent uncertainty in the choice of model.

- Suppose we wish to compare L models $\{\mathcal{M}_i\}$ where $i = 1, \ldots, L$. Here a model refers to a probability distribution over the observed data $\mathcal{D}$. Our uncertainty is expressed through a prior $p(\mathcal{M}_i)$. Given a training set $\mathcal{D}$, then the posterior distribution is:

$$p(\mathcal{M}_i \mid \mathcal{D}) \propto p(\mathcal{M}_i)\, \underbrace{p(\mathcal{D} \mid \mathcal{M}_i)}$$

Posterior Prior **Model evidence or marginal likelihood**

- We have already defined the ***Bayes Factor*** as the ratio of two model evidences

$$\frac{p(\mathcal{D} \mid \mathcal{M}_i)}{p(\mathcal{D} \mid \mathcal{M}_j)}$$

Model Averaging and Model Selection

- Once we know the posterior distribution over models, the predictive distribution is given, by

$$p(t \mid \mathbf{x}, \mathcal{D}) = \sum_{i=1}^{L} p(t \mid \mathbf{x}, \mathcal{M}_i, \mathcal{D})\, p(\mathcal{M}_i \mid \mathcal{D})$$

Test data

Training data

- This has the form of a mixture distribution in which the overall predictive distribution is obtained by averaging the predictive distributions $p(t \mid \mathbf{x}, \mathcal{M}_i, \mathcal{D})$ of individual models, weighted by the posterior probabilities $p(\mathcal{M}_i \mid \mathcal{D})$ of those models.

- A simple approximation to model averaging is to use the single most probable model alone to make predictions. This is known as model selection.

Model Evidence

- For a model governed by a set of parameters ***w***, the model evidence is given, from the sum and product rules of probability, by

$$p(\mathcal{D} \mid \mathcal{M}_i) = \int p\left(\mathcal{D} \mid \mathbf{w}_i, \mathcal{M}_i\right) p\left(\mathbf{w}_i \mid \mathcal{M}_i\right) d\mathbf{w}_i$$

- From a sampling perspective, the marginal likelihood can be viewed as **the probability of generating the data set $\mathcal{D}$ from a model whose parameters are sampled at random from the prior.**

- Also we can see the model evidence as normalizing factor:

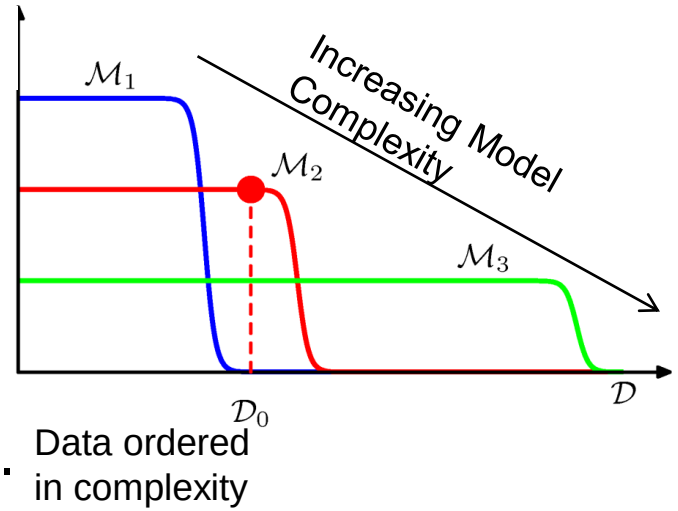
$$\mathbf{p}\left(\mathbf{w} \mid \mathcal{D}, \mathcal{M}_i\right) = \frac{p\left(\mathcal{D} \mid \mathbf{w}, \mathcal{M}_i\right) p\left(\mathbf{w} \mid \mathcal{M}_i\right)}{p(\mathcal{D} \mid \mathcal{M}_i)}$$

Occam's Razor and Model Selection

- Compare model classes $\mathcal{M}_i$ using their posterior probability given the data $\mathcal{D}$:

$$p(\mathcal{M}_i \mid \mathcal{D}) \propto p(\mathcal{M}_i)\, p(\mathcal{D} \mid \mathcal{M}_i) \qquad p(\mathcal{D} \mid \mathcal{M}_i) = \int_{\Theta_i} p(\mathcal{D} \mid \mathbf{w}_i, \mathcal{M}_i)\, p(\mathbf{w}_i \mid \mathcal{M}_i)\, d\mathbf{w}_i$$

- The marginal likelihood (Bayesian evidence) $p(\mathcal{D}|\mathcal{M}_i)$ is viewed as **the probability that randomly selected parameter values from the model class would generate the data set $\mathcal{D}$.**
  - Simple model classes are unlikely to generate $\mathcal{D}$.
  - Too complex model classes can generate many data sets so it is unlikely to generate the particular data set $\mathcal{D}$.
- Bayesian inference automatically implements Occam's Razor Principle: *Prefer simple than complex explanations*.

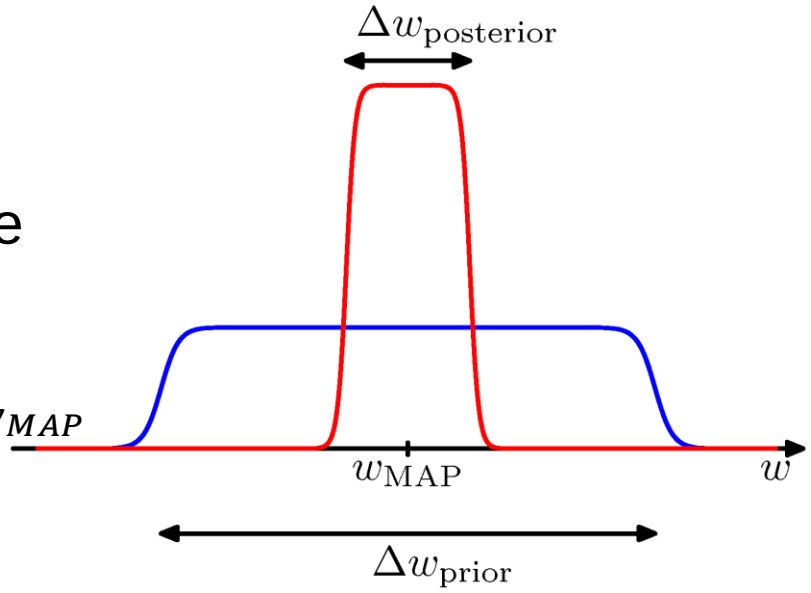


Bayesian Model Comparison

- For a given model (omit $|\mathcal{M}_i$) with a single parameter, w , consider the approximation (use Bayes rule)

$$p(\mathcal{D}) = \int p(\mathcal{D}|w)p(w)dw = \frac{p(\mathcal{D}|w)p(w)}{p(w|\mathcal{D})}\Big|_{w_{MAP}}$$

$$\simeq p(\mathcal{D}|w_{MAP})\frac{\Delta w_{posterior}}{\Delta w_{prior}}$$



where the posterior is assumed to be sharply peaked around w_{MAP} (we already have seen the more accurate Laplace approximation)

- Taking logs we obtain

$$lnp(\mathcal{D}) \simeq lnp(\mathcal{D}|w_{MAP}) + \underbrace{\ln\left(\frac{\Delta w_{posterior}}{\Delta w_{prior}}\right)}_{\text{Negative}}$$

Note: the evidence is not defined if the prior is improper.

Optimal Model Complexity

- For a model with *M* parameters, we can make a similar approximation for each parameter in turn. Assuming that all parameters have the same ratio of $\Delta w_{posterior} / \Delta w_{prior}$,

$$lnp(\mathcal{D}) \simeq lnp(\mathcal{D}|w_{MAP}) + \mathrm{M} \ln\left(\frac{\Delta w_{posterior}}{\Delta w_{prior}}\right)$$

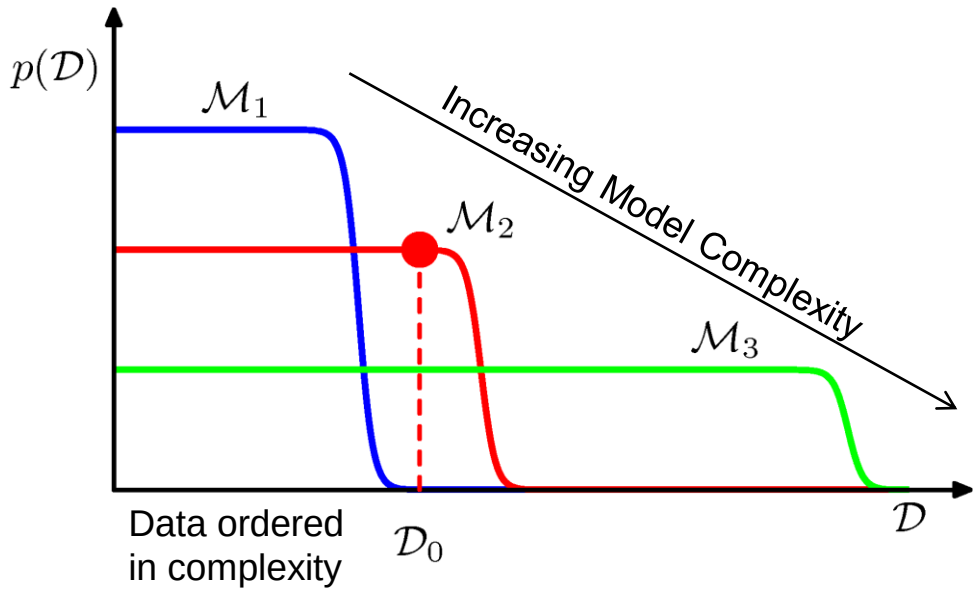
- The size of the complexity penalty increases linearly with *M*. As we increase the complexity of the model
  - the 1st term increases, because a more complex model is better able to fit the data,
  - whereas the 2nd term decreases due to the dependence on *M*.
- The optimal model complexity determined by maximum evidence will be given by a trade-off between these two terms.

Matching Data and Model Complexity

- The marginal likelihood favors models of intermediate complexity.

- Let us think of the regression model and consider the models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ represent linear, quadratic and cubic fitting.

- The data $\mathcal{D}$ are ordered in complexity – for a given model, we choose $\boldsymbol{w}$ from the prior $p(\boldsymbol{w})$, then sample the data from $p(\mathcal{D}|\boldsymbol{w})$.



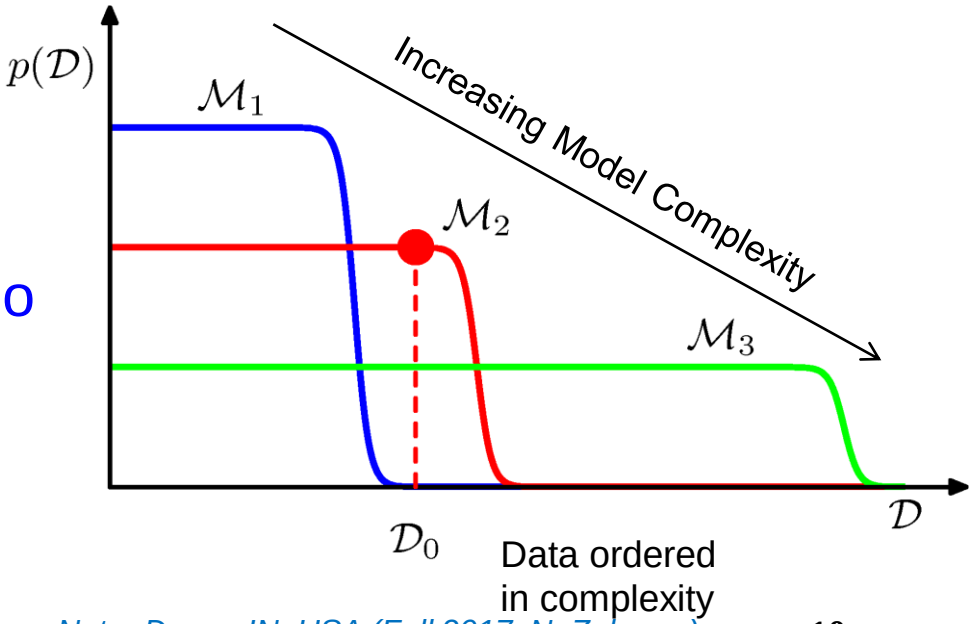
Matching Data and Model Complexity

- A 1st order polynomial has little variability, generates data that are similar, $p(\mathcal{D})$ is confined to a small region in the $\mathcal{D}$ axis.

- A 9th order polynomial generates a variety of different data, and so its $p(\mathcal{D})$ is spread over a large region in the $\mathcal{D}$ axis.

- Because $p(\mathcal{D}|\mathcal{M}_i)$ are normalized, a particular $\mathcal{D}_0$ can have the highest evidence for the model of intermediate complexity.

- The simpler model cannot fit the data well, whereas the more complex model spreads its predictive probability over too broad a range of data sets.



Matching Data and Model Complexity

- A Bayesian model comparison in an average (over the data $\mathcal{D}$) sense will favor the correct model.

- Let $\mathcal{M}_1$ be the correct model and $\mathcal{M}_2$ another model. We can show that the evidence for model $\mathcal{M}_1$ is higher. Using the definition and properties of the Kullback-Leibler distance:

$$KL\left(p(\mathcal{D}\,|\,\mathcal{M}_1)\,\|\,p(\mathcal{D}\,|\,\mathcal{M}_2)\right)=\int \underbrace{p(\mathcal{D}\,|\,\mathcal{M}_1)}_{\substack{\text{Averaged with the}\\ \text{exact probability}}} \ln \underbrace{\frac{p(\mathcal{D}\,|\,\mathcal{M}_1)}{p(\mathcal{D}\,|\,\mathcal{M}_2)}}_{\text{Bayes factor}} d\mathcal{D} \geq 0$$

- This analysis assumes that the true distribution from which the data are generated is contained in our class of models.

The Evidence Approximation

- The fully Bayesian predictive distribution for our regression model is given by

The hyperparameters α and β are now random variables

$$p(t \mid \boldsymbol{t}) = \int\int\int \underbrace{p(t \mid \boldsymbol{w}, \beta)}_{\mathcal{N}\left(t \mid y(x,\boldsymbol{w}), \beta^{-1}\right)} \underbrace{p(\boldsymbol{w} \mid \boldsymbol{t}, \alpha, \beta)}_{\substack{\mathcal{N}(\boldsymbol{w} \mid \boldsymbol{m}_N, \boldsymbol{S}_N) \\ \boldsymbol{m}_N = \beta \boldsymbol{S}_N \boldsymbol{\Phi}^T \boldsymbol{t} \\ \boldsymbol{S}_N^{-1} = \alpha \boldsymbol{I} + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi}}} p(\alpha, \beta \mid \boldsymbol{t}) \, d\boldsymbol{w} d\alpha d\beta$$

Dependence on x and **x** not shown to simplify the notation

but this integral is intractable. Approximate with

$$p(t|\boldsymbol{t}) \simeq p(t|\boldsymbol{t}, \hat{\alpha}, \hat{\beta}) = \int p(t|\boldsymbol{w}, \hat{\beta}) p(\boldsymbol{w}|\boldsymbol{t}, \hat{\alpha}, \hat{\beta}) d\boldsymbol{w}$$

where $(\hat{\alpha}, \hat{\beta})$ is the mode of $p(\alpha, \beta \mid \boldsymbol{t})$, which is assumed to be sharply peaked.

a.k.a. empirical Bayes, type II or generalized maximum likelihood, or evidence approximation.

The Evidence Approximation

- From Bayes' theorem, the posterior distribution for α and β is given by

$$p(\alpha,\beta\,|\,\mathbf{t}) \propto p(\mathbf{t}\,|\,\alpha,\beta)\,p(\alpha,\beta)$$

- If the prior is relatively flat, then in the evidence framework the values of $(\hat{\alpha},\hat{\beta})$ are obtained by maximizing the marginal likelihood function $p(\mathbf{t}|\alpha, \beta)$.

- The marginal likelihood function $p(\mathbf{t}|\alpha, \beta)$ is obtained by integrating over the parameters $\mathbf{w}$, so that

$$p(\alpha,\beta\,|\,\mathbf{t}) \propto p(\mathbf{t}\,|\,\alpha,\beta) = \int \underbrace{p(\mathbf{t}\,|\,\mathbf{w},\beta)}_{\frac{1}{(2\pi)^{N/2}}\beta^{N/2}e^{-\beta E_D}} \underbrace{p(\mathbf{w}\,|\,\alpha)}_{\mathcal{N}\left(\mathbf{w},\alpha^{-1}\mathbf{I}_{M\times M}\right)} d\mathbf{w}$$

Evidence/ Marginal Likelihood

- One can evaluate this integral using the completion of the square procedure typical of Gaussian marginalizations.

The Evidence Approximation

- We can write the evidence function in the form

$$p(\boldsymbol{t}\,|\,\alpha,\beta)=\left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int \exp\{-E(\boldsymbol{w})\}\,d\boldsymbol{w}$$

where M is the dimensionality of $\boldsymbol{w}$, and we have defined

$$E(\boldsymbol{w})=\beta E_D(\boldsymbol{w})+\alpha E_W(\boldsymbol{w})=\frac{\beta}{2}\left\|\boldsymbol{t}-\boldsymbol{\Phi w}\right\|^2+\frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$E(\boldsymbol{w})=E(\boldsymbol{m}_N)+\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_N)^T\boldsymbol{A}(\boldsymbol{w}-\boldsymbol{m}_N)$$

- We have introduced here:

$$\boldsymbol{A}=\alpha\boldsymbol{I}+\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi},\quad E(\boldsymbol{m}_N)=\frac{\beta}{2}\left\|\boldsymbol{t}-\boldsymbol{\Phi m}_N\right\|^2+\frac{\alpha}{2}\boldsymbol{m}_N^T\boldsymbol{m}_N,\ \boldsymbol{m}_N=\beta\boldsymbol{A}^{-1}\boldsymbol{\Phi}^T\boldsymbol{t}$$

- Note that the Hessian matrix $\boldsymbol{A}$ corresponds to the matrix of 2nd derivatives of the error function:

$$\boldsymbol{A}=\nabla\nabla E(\boldsymbol{w})$$

The Evidence Approximation

- The integral over $\boldsymbol{w}$ can now be evaluated simply by appealing to the standard result for the normalization coefficient of a multivariate Gaussian, giving

$$\int \exp\{-E(\boldsymbol{w})\}\, d\boldsymbol{w} = \exp\{-E(\boldsymbol{m}_N)\} \int \exp\left\{-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_N)^T \boldsymbol{A}(\boldsymbol{w}-\boldsymbol{m}_N)\right\} d\boldsymbol{w}$$

$$= \exp\{-E(\boldsymbol{m}_N)\}(2\pi)^{M/2} |\boldsymbol{A}|^{-1/2}$$

- We can then write the log of the marginal likelihood in the form

$$p(\boldsymbol{t}|\alpha,\beta) = \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2} \int \exp\{-E(\boldsymbol{w})\}\, d\boldsymbol{w} = \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2} e^{-E(\boldsymbol{m}_N)}(2\pi)^{M/2}|\boldsymbol{A}|^{-1/2} \Rightarrow$$

$$\ln p(\boldsymbol{t}|\alpha,\beta) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - E(\boldsymbol{m}_N) - \frac{1}{2}\ln|\boldsymbol{A}| - \frac{N}{2}\ln(2\pi)$$

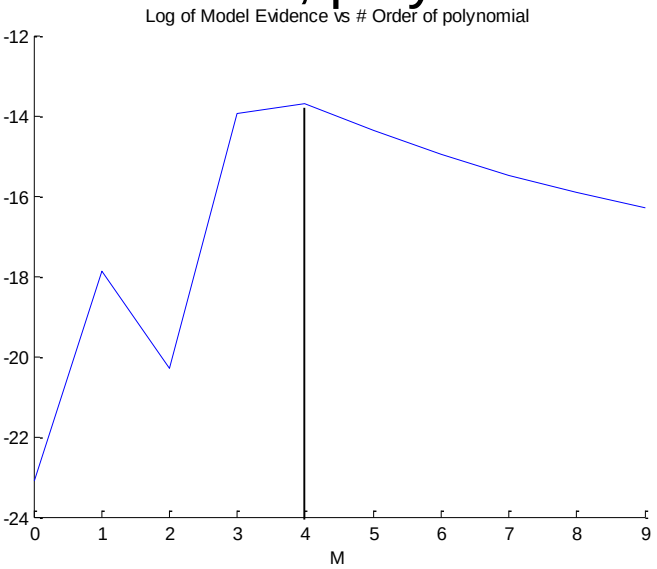
M : number of parameters in the model

N : number of data

The Evidence Approximation

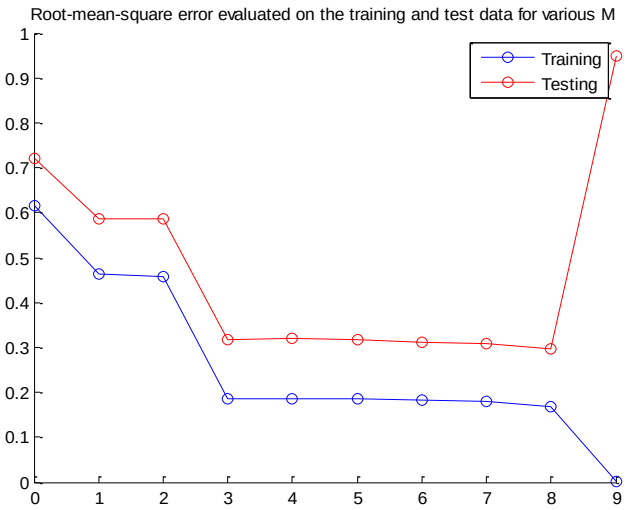
- Example: sinusoidal data, polynomial regression, $\alpha = 5 \times 10^{-3}$, $\beta = 11.1$

$p(\boldsymbol{t} \mid \alpha, \beta)$



MatLab Code and data

- From the plot of the model evidence for given α and β, we see that the evidence favors the model with $M = 5$ (4th degree polynomial)
- Looking at the non-Bayesian approach, one cannot distinguish the performance of polynomials of order 3….8



Maximizing the Evidence Function

- Let us first consider the maximization of $p(\boldsymbol{t}|\alpha, \beta)$ with respect to α. This can be done by first defining the following eigenvector equation

$$\left(\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\boldsymbol{u}_i = \lambda_i\boldsymbol{u}_i$$

- Thus, $\boldsymbol{A} = \alpha\boldsymbol{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}$ has eigenvalues $\alpha+\lambda_i$.

- Now consider the derivative of the term involving $ln|\boldsymbol{A}|$ with respect to α. We have

$$\ln p(\boldsymbol{t}\,|\,\alpha,\beta) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - E(\boldsymbol{m}_N) - \frac{1}{2}\ln|\,\boldsymbol{A}\,| - \frac{N}{2}\ln(2\pi)$$

$$\frac{d}{d\alpha}ln|\,\boldsymbol{A}\,| = \frac{d}{d\alpha}ln\prod_i(\lambda_i+\alpha) = \frac{d}{d\alpha}\sum_i ln(\lambda_i+\alpha) = \sum_i\frac{1}{\lambda_i+\alpha}$$

Maximizing the Evidence Function

$$\frac{d}{d\alpha}\ln|\boldsymbol{A}| = \frac{d}{d\alpha}\ln\prod_i(\lambda_i+\alpha) = \frac{d}{d\alpha}\sum_i\ln(\lambda_i+\alpha) = \sum_i\frac{1}{\lambda_i+\alpha}$$

- Thus the stationary points of

$$\ln p(\boldsymbol{t}\,|\,\alpha,\beta) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - E(\boldsymbol{m}_N) - \frac{1}{2}\ln|\boldsymbol{A}| - \frac{N}{2}\ln(2\pi)$$

$$E(\boldsymbol{m}_N) = \frac{\beta}{2}\|\boldsymbol{t}-\boldsymbol{\Phi}\boldsymbol{m}_N\|^2 + \frac{\alpha}{2}\boldsymbol{m}_N^T\boldsymbol{m}_N,\ \ \boldsymbol{m}_N = \beta\boldsymbol{A}^{-1}\boldsymbol{\Phi}^T\boldsymbol{t},\ \ \boldsymbol{A} = \alpha\boldsymbol{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}$$

with respect to α *satisfy* $$0 = \frac{M}{2\alpha} - \frac{1}{2}\boldsymbol{m}_N^T\boldsymbol{m}_N - \frac{1}{2}\sum_i\frac{1}{\lambda_i+\alpha}$$

- Multiplying through by 2α and rearranging, we obtain

$$\alpha\boldsymbol{m}_N^T\boldsymbol{m}_N = M - \alpha\sum_i\frac{1}{\lambda_i+\alpha} = \sum_i\left(1-\frac{\alpha}{\lambda_i+\alpha}\right) = \sum_i\frac{\lambda_i}{\lambda_i+\alpha} \equiv \gamma$$

$$\alpha = \frac{\gamma}{\boldsymbol{m}_N^T\boldsymbol{m}_N}$$

Implicit solution for α

1. Choose α
2. Calculate $\boldsymbol{m}_N, \gamma$
3. Re-estimate α

Maximizing the Evidence Function

Implicit Solution for Computing α

1. Choose α

2. Calculate $\boldsymbol{m}_N, \gamma$:

$$\boldsymbol{m}_N = \beta \boldsymbol{A}^{-1} \boldsymbol{\Phi}^T \boldsymbol{t}, \quad \boldsymbol{A} = a\boldsymbol{I} + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi}, \quad \left(\beta \boldsymbol{\Phi}^T \boldsymbol{\Phi}\right) \boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i$$

$$\gamma = \sum_i \frac{\lambda_i}{\lambda_i + \alpha}$$

3. Re-estimate α

$$\alpha = \frac{\gamma}{\boldsymbol{m}_N^T \boldsymbol{m}_N}$$

Maximizing the Evidence Function

- We can similarly maximize the log marginal likelihood with respect to β.

$$\ln p(\boldsymbol{t} \mid \alpha, \beta) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - E(\boldsymbol{m}_N) - \frac{1}{2}\ln|\boldsymbol{A}| - \frac{N}{2}\ln(2\pi)$$

- To do this, we note that the eigenvalues λ_i defined by

$$\left(\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\boldsymbol{u}_i = \lambda_i\boldsymbol{u}_i$$

are proportional to β, and hence $d\lambda_i / d\beta = \lambda_i / \beta$ giving

$$\frac{d}{d\beta}\ln|\boldsymbol{A}| = \frac{d}{d\beta}\ln\prod_i(\lambda_i + \alpha) = \frac{1}{\beta}\sum_i\frac{\lambda_i}{\lambda_i + \alpha} = \frac{\gamma}{\beta}$$

Maximizing the Evidence Function

$$\ln p(\boldsymbol{t}\,|\,\alpha,\beta)=\frac{M}{2}\ln\alpha+\frac{N}{2}\ln\beta-E(\boldsymbol{m}_N)-\frac{1}{2}\ln|\boldsymbol{A}|-\frac{N}{2}\ln(2\pi)$$

$$E(\boldsymbol{m}_N)=\frac{\beta}{2}\left\|\boldsymbol{t}-\boldsymbol{\Phi}\boldsymbol{m}_N\right\|^2+\frac{\alpha}{2}\boldsymbol{m}_N^T\boldsymbol{m}_N \qquad \frac{d}{d\beta}\ln|\boldsymbol{A}|=\frac{1}{\beta}\sum_i\frac{\lambda_i}{\lambda_i+\alpha}=\frac{\gamma}{\beta}$$

$$\boldsymbol{m}_N=\beta\boldsymbol{A}^{-1}\boldsymbol{\Phi}^T\boldsymbol{t},\quad \boldsymbol{A}=a\boldsymbol{I}+\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi},\quad \left(\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\boldsymbol{u}_i=\lambda_i\boldsymbol{u}_i,\quad \gamma=\sum_i\frac{\lambda_i}{\lambda_i+\alpha}$$

- Setting the derivative wrt β equal to zero, the stationary point of the marginal likelihood therefore satisfies

$$0=\frac{N}{2\beta}-\frac{1}{2}\sum_{n=1}^{N}\left\{t_n-\boldsymbol{m}_N^T\boldsymbol{\phi}(x_n)\right\}^2-\frac{\gamma}{2\beta}$$

- Rearranging we obtain

$$\frac{1}{\beta}=\frac{1}{N-\gamma}\sum_{n=1}^{N}\left\{t_n-\boldsymbol{m}_N^T\boldsymbol{\phi}(x_n)\right\}^2$$

Implicit solution for β

1. Choose β
2. Calculate $\boldsymbol{m}_N,\gamma$
3. Re-estimate β

Maximizing the Evidence Function

- It is interesting to note that in the evidence framework (using the optimal computed values of α and β), the following is true:

$$E(\boldsymbol{m}_N) = \frac{N}{2}$$

- This can be easily shown using the earliest derived results:

$$E(\boldsymbol{m}_N) = \frac{\beta}{2}\left\|\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{m}_N\right\|^2 + \frac{\alpha}{2}\boldsymbol{m}_N^T\boldsymbol{m}_N$$

with

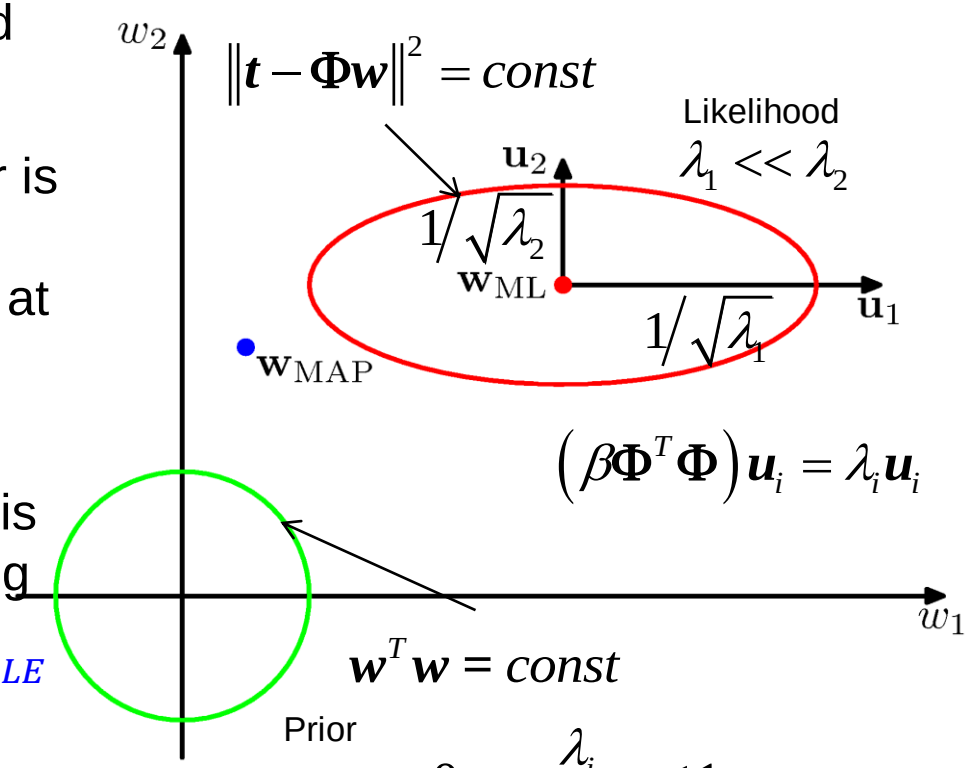
$$\alpha = \frac{\gamma}{\boldsymbol{m}_N^T\boldsymbol{m}_N}$$

and

$$\frac{1}{\beta} = \frac{1}{N-\gamma}\sum_{n=1}^{N}\left\{t_n - \boldsymbol{m}_N^T\boldsymbol{\phi}(x_n)\right\}^2 = \frac{1}{N-\gamma}\left\|\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{m}_N\right\|^2$$

Effective Number of Parameters

- Consider the contours of the likelihood & prior in which the axes in parameter space have been rotated to align with the eigenvectors $\mathbf{u}_i$.
- *For $\alpha = 0$,* the mode of the posterior is given by the MLE solution $\boldsymbol{w}_{\text{ML}}$, whereas for nonzero α the mode is at $\boldsymbol{w}_{MAP} = \boldsymbol{m}_N$.
- In the direction w_1 , λ_1 *is small* compared with α *and so* $\lambda_1/(\lambda_1 + \alpha)$ is close to zero, and the corresponding MAP value of w_1, $w_{1_{MAP}} = \frac{\lambda_1}{\lambda_1+\alpha} w_{1_{MLE}}$ *is also close to zero.*
- In the direction w_2, λ_2 is large compared with α and so $\lambda_2/(\lambda_2+\alpha)$ is close to unity, and the MAP value of w_2 is close to its MLE value.



$$0 < \frac{\lambda_i}{\lambda_i + \alpha} \leq 1,$$

$$0 \leq \gamma = \sum_i \frac{\lambda_i}{\lambda_i + \alpha} \leq M$$

Effective Number of Parameters

- In directions w_i, $\lambda_i << \alpha$, so $\lambda_i/(\lambda_i + \alpha)$ is close to zero, and the corresponding MAP value of w_i is also close to zero.

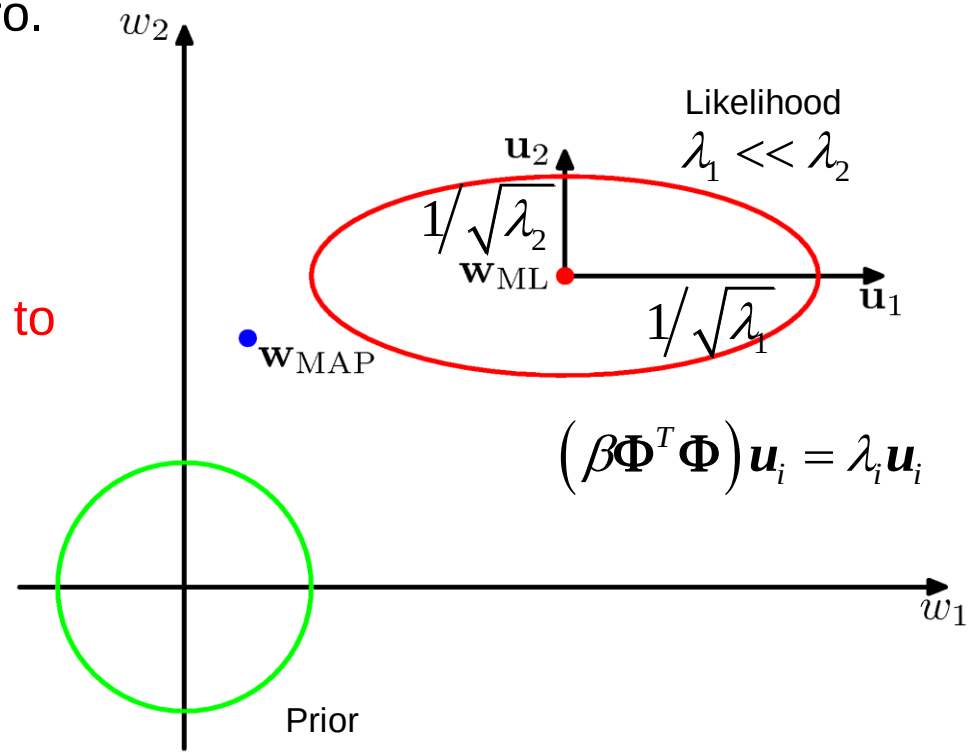
  These are directions in which the likelihood function is relatively insensitive to the parameter value and so the parameter has been set to a small value by the prior.

- The quantity γ

$$0 < \frac{\lambda_i}{\lambda_i + \alpha} \leq 1,$$

$$0 \leq \gamma = \sum_i \frac{\lambda_i}{\lambda_i + \alpha} \leq M$$

  therefore **measures the effective total number of well determined parameters.**



Effective Number of Parameters

- We can obtain some insight into the equation for β

$$\frac{1}{\beta} = \frac{1}{N-\gamma}\sum_{n=1}^{N}\left\{t_n - \mathbf{m}_N^T\boldsymbol{\phi}(x_n)\right\}^2$$

by comparing it with the MLE result derived earlier:

$$\frac{1}{\beta_{ML}} = \frac{1}{N}\sum_{n=1}^{N}\left\{t_n - \mathbf{m}_N^T\boldsymbol{\phi}(x_n)\right\}^2$$

- These formulas express the variance as an average of the squared differences between the targets and model predictions.

- They differ in that the # of data points
  - N in the MLE result is replaced by $N - \gamma$ in the Bayesian result.

Effective Number of Parameters

$$\frac{1}{\beta} = \frac{1}{N-\gamma}\sum_{n=1}^{N}\left\{t_n - \boldsymbol{m}_N^T\boldsymbol{\phi}(x_n)\right\}^2 \qquad \frac{1}{\beta_{ML}} = \frac{1}{N}\sum_{n=1}^{N}\left\{t_n - \boldsymbol{m}_N^T\boldsymbol{\phi}(x_n)\right\}^2$$

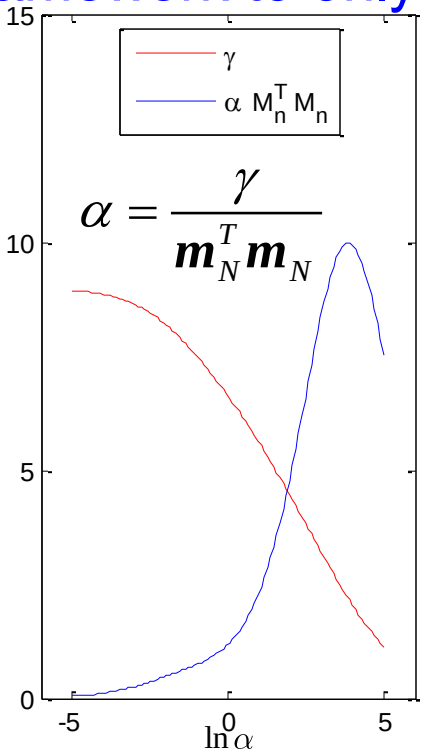
- The effective number of parameters determined by the data is γ.
- The remaining $M-\gamma$ parameters are set to small values by the prior.
- This is reflected in the Bayesian result for the variance that has a factor $N-\gamma$ in the denominator correcting for the bias of the MLE.
- These results are analogous to the estimation of the variance of a Gaussian:

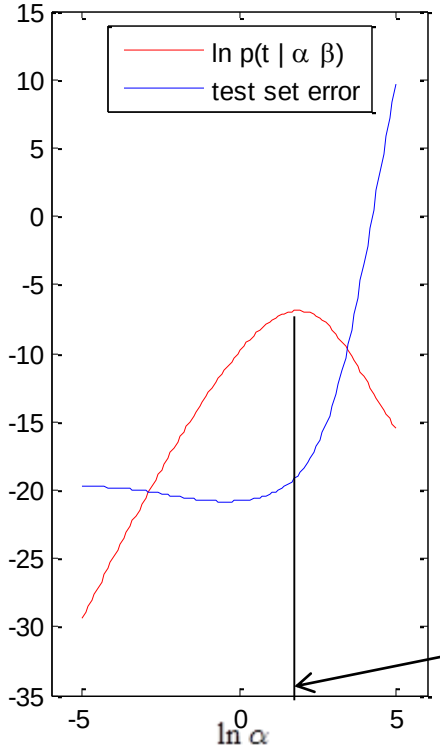
$$\sigma_{ML}^2 = \frac{1}{N}\sum_{n=1}^{N}\left\{x_n - \mu_{ML}\right\}^2 \quad vs. \quad \sigma_{MAP}^2 = \frac{1}{N-1}\sum_{n=1}^{N}\left\{x_n - \mu_{ML}\right\}^2$$

1 degree of freedom has been used to fit the mean and the MAP estimate for the variance accounts for that.

Effective Number of Parameters

- We illustrate the evidence framework for setting hyperparameters using the sinusoidal synthetic data, together with the 9 Gaussian basis functions. The total # of parameters is thus $M = 10$ including the bias.
- For simplicity, we set $\beta = 11.1$ (true value) and use the evidence framework to only determine α



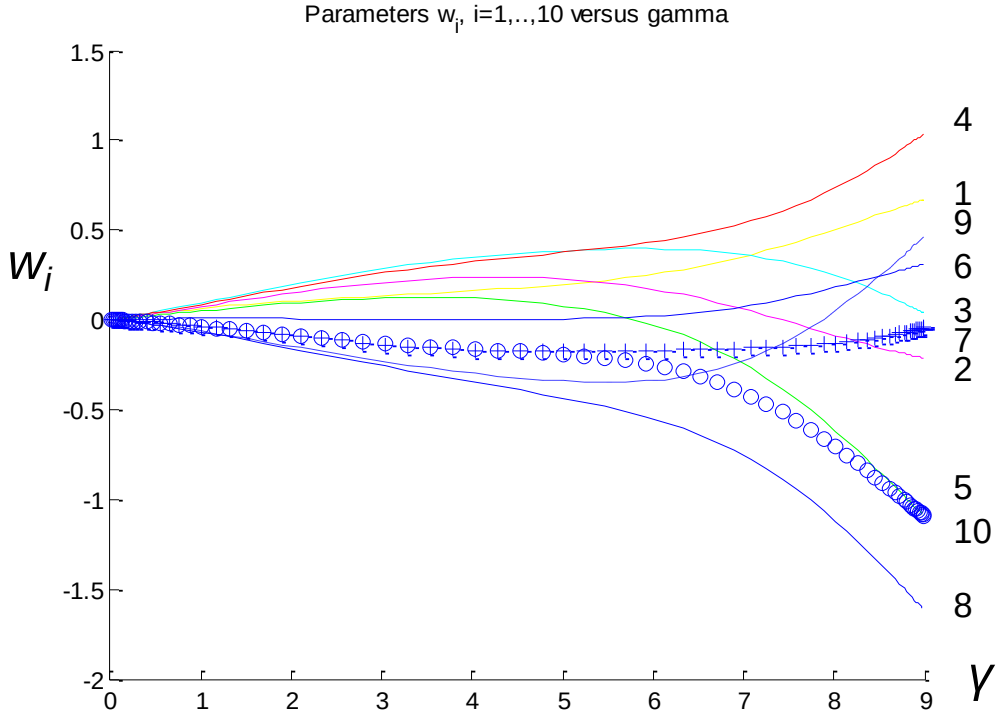


MatLab Code and data

Min generalization error

Effective Number of Parameters

- We can also see how α controls the magnitude of the parameters $\{w_i\}$, by plotting the individual parameters (posterior means) versus the effective number γ of parameters. We use the Gaussian basis model.



MatLab Code and data

- For the simlation, α is varied $0 \leq \alpha \leq \infty$ causing γ to vary in the range $0 \leq \gamma \leq M$.

Case of N>>M

- For $N >> M$, all of the parameters are well determined by the data because $\Phi^T\Phi$ involves an implicit sum over data points, and so the eigenvalues λ_i increase with the size of the data set.

- In this case, $\gamma = M$, and the re-estimation equations for α and β become

$$\alpha = \frac{\gamma}{\mathbf{m}_N^T \mathbf{m}_N} = \frac{M}{2E_W(\mathbf{m}_N)} = \frac{M}{\mathbf{m}_N^T \mathbf{m}_N} \quad (\gamma = M)$$

$$\frac{1}{\beta} = \frac{1}{N}\sum_{n=1}^{N}\left\{t_n - \mathbf{m}_N^T \boldsymbol{\phi}(x_n)\right\}^2 \quad (N >> M)$$

- These results are useful as they do not require computing the eigenspectrum of the Hessian.

Another Example of Model Evidence

- We have seen in an earlier lecture that the conjugate prior for a Gaussian with unknown mean and unknown precision is a normal-gamma distribution.

- We can apply the same for the case of our likelihood

$$p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n \mid \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n), \beta^{-1}\right)$$

for which the conjugate prior for $\boldsymbol{w}$ and β is:

$$p(\boldsymbol{w}, \beta) = \mathcal{N}\left(\boldsymbol{w} \mid \boldsymbol{m}_0, \beta^{-1}\boldsymbol{S}_0\right)\mathcal{G}am\left(\beta \mid a_0, b_0\right)$$

- It can be shown that the corresponding posterior for this takes the form:

$$p(\boldsymbol{w}, \beta \mid \boldsymbol{t}) = \mathcal{N}\left(\boldsymbol{w} \mid \boldsymbol{m}_N, \beta^{-1}\boldsymbol{S}_N\right)\mathcal{G}am\left(\beta \mid a_N, b_N\right)$$

Posterior Distribution

- The posterior takes the form:

$$p(\boldsymbol{w},\beta\,|\,\boldsymbol{t}) \propto \beta^{N/2}\exp\left\{-\frac{1}{2}\boldsymbol{w}^T\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w}-\beta\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{t}-\frac{1}{2}\beta\sum_{n=1}^{N}t_n^2\right\}$$

$$\beta^{M/2}\exp\left\{-\frac{1}{2}\left(\boldsymbol{w}-\boldsymbol{m}_0\right)^T\beta\boldsymbol{S}_0^{-1}\left(\boldsymbol{w}-\boldsymbol{m}_0\right)\right\}\beta^{a_0-1}e^{-b_0\beta}$$

$$\propto \beta^{N/2}e^{-\beta(\boldsymbol{w}-\boldsymbol{m}_N)^T\boldsymbol{S}_N^{-1}(\boldsymbol{w}-\boldsymbol{m}_N)}e^{\frac{1}{2}\boldsymbol{m}_N^T\beta\boldsymbol{S}_N^{-1}\boldsymbol{m}_N}e^{-\frac{1}{2}\beta\sum_{n=1}^{N}t_n^2}\beta^{M/2}e^{-\frac{1}{2}\boldsymbol{m}_0^T\beta\boldsymbol{S}_0^{-1}\boldsymbol{m}_0}\beta^{a_0-1}e^{-b_0\beta}$$

$$\propto \beta^{M/2}e^{-\beta(\boldsymbol{w}-\boldsymbol{m}_N)^T\boldsymbol{S}_N^{-1}(\boldsymbol{w}-\boldsymbol{m}_N)}\frac{1}{\Gamma(a_N)}b_N^{a_N}\beta^{a_N-1}e^{-b_N\beta}$$

$$\propto \mathcal{N}\left(\boldsymbol{w}\,|\,\boldsymbol{m}_N,\beta^{-1}\boldsymbol{S}_N\right)\mathcal{G}am\left(\beta\,|\,a_N,b_N\right)$$

where:

$$\boldsymbol{m}_N=\boldsymbol{S}_N\left[\boldsymbol{S}_0^{-1}\boldsymbol{m}_0+\boldsymbol{\Phi}^T\boldsymbol{t}\right],\ \boldsymbol{S}_N^{-1}=\left[\boldsymbol{S}_0^{-1}+\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right]$$

$$a_N=a_0+\frac{N}{2},\ b_N=b_0+\frac{1}{2}\left(\boldsymbol{m}_0^{-1}\boldsymbol{S}_0^{-1}\boldsymbol{m}_0-\boldsymbol{m}_N^{-1}\boldsymbol{S}_N^{-1}\boldsymbol{m}_N+\sum_{n=1}^{N}t_n^2\right)$$

Model Evidence

- The model evidence for our example is given as:

$$p(\boldsymbol{t}) = \iint p(\boldsymbol{t} \mid \boldsymbol{w}, \beta)\, p(\boldsymbol{w} \mid \beta)\, d\boldsymbol{w} p(\beta) d\beta$$

$$= \iint \left(\frac{\beta}{2\pi}\right)^{N/2} \exp\left\{-\frac{\beta}{2}(\boldsymbol{t}-\boldsymbol{\Phi w})^T(\boldsymbol{t}-\boldsymbol{\Phi w})\right\}$$

$$\left(\frac{\beta}{2\pi}\right)^{M/2} |\boldsymbol{S}_0|^{-1/2} \exp\left\{-\frac{\beta}{2}(\boldsymbol{w}-\boldsymbol{m}_0)^T \boldsymbol{S}_0^{-1}(\boldsymbol{t}-\boldsymbol{m}_0)\right\} d\boldsymbol{w}$$

$$\Gamma(a_0)^{-1} b_0^{a_0} \beta^{a_0-1} \mathrm{e}^{-b_0\beta}\, d\beta =$$

$$= \frac{b_0^{a_0}}{\left((2\pi)^{M+N}|\boldsymbol{S}_0|\right)^{1/2}} \iint \exp\left\{-\frac{\beta}{2}(\boldsymbol{t}-\boldsymbol{\Phi w})^T(\boldsymbol{t}-\boldsymbol{\Phi w})\right\}$$

$$\exp\left\{-\frac{\beta}{2}(\boldsymbol{w}-\boldsymbol{m}_0)^T \boldsymbol{S}_0^{-1}(\boldsymbol{t}-\boldsymbol{m}_0)\right\} d\boldsymbol{w}$$

$$\Gamma(a_0)^{-1} \beta^{a_0-1} \beta^{N/2} \beta^{M/2} \mathrm{e}^{-b_0\beta}\, d\beta$$

Model Evidence

- Using some of our earlier results in deriving the posterior:

$$p(\boldsymbol{t}) = \frac{b_0^{a_0}}{\left((2\pi)^{M+N}\,|\boldsymbol{S}_0|\right)^{1/2}} \iint \exp\left\{-\frac{\beta}{2}(\boldsymbol{t}-\boldsymbol{m}_N)^T\,\boldsymbol{S}_N^{-1}(\boldsymbol{t}-\boldsymbol{m}_N)\right\} d\boldsymbol{w}$$

$$\exp\left\{-\frac{\beta}{2}\left(\boldsymbol{t}^T\boldsymbol{t} + \boldsymbol{m}_0^T\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 - \boldsymbol{m}_N^T\boldsymbol{S}_N^{-1}\boldsymbol{m}_N\right)\right\}$$

$$\Gamma(a_0)^{-1}\,\beta^{a_N-1}\beta^{M/2}\exp(-b_0\beta)\,d\beta$$

- Performing the integration in $\boldsymbol{w}$ and using the normalization factor for the Gamma distribution:

$$p(\boldsymbol{t}) = \frac{b_0^{a_0}}{\left((2\pi)^{M+N}\,|\boldsymbol{S}_0|\right)^{1/2}} (2\pi)^{M/2}\,|\boldsymbol{S}_N|^{1/2}\,\Gamma(a_0)^{-1} \underbrace{\int \beta^{a_N-1}\exp(-b_N\beta)\,d\beta}_{\Gamma(a_N)/b_N^{a_N}}$$

$$= \frac{1}{(2\pi)^{N/2}} \frac{|\boldsymbol{S}_N|^{1/2}}{|\boldsymbol{S}_0|^{1/2}} \frac{b_0^{a_0}}{b_N^{a_N}} \frac{\Gamma(a_N)}{\Gamma(a_0)}$$

Limitations of Fixed Basis Functions

- Up to now the basis functions $\phi_j(\boldsymbol{x})$ are fixed before the training data set is observed.
- As a consequence, the number of basis functions grows exponentially with the dimensionality D of the input space.
- There are two properties of real data sets that we can exploit to alleviate the curse of dimensionality:
  - the data vectors $\{\boldsymbol{x}_n\}$ typically lie close to a nonlinear manifold whose intrinsic dimensionality is smaller than that of the input space as a result of strong correlations between the input variables.
  - If we are using localized basis functions, we can arrange that they are scattered in input space only in regions containing data (radial basis functions, support vector and relevance vector machines).

Adaptive Basis Functions

- Neural network models using adaptive basis functions having sigmoidal nonlinearities, can adapt the parameters so that the regions of input space over which the basis functions vary correspond to the data manifold.

- The target variables may have significant dependence on only a small number of possible directions within the data manifold.

- Neural networks can exploit this property by choosing the directions in input space to which the basis functions respond.

Laplace Approximation

- As we have seen earlier, the Laplace approximation allows a Gaussian approximation of the parameter posterior about the maximum a posteriori (MAP) parameter estimate.

- Consider a data set $\mathcal{D}$ and M models $\mathcal{M}_i$, *i=1,..,M* with corresponding parameters $\boldsymbol{\theta}_i$, i=1,…M. We compare models using the posteriors:

$$p(\mathcal{M} \mid \mathcal{D}) \propto p(\mathcal{M})\, p(\mathcal{D} \mid \mathcal{M})$$

- For large sets of data $\mathcal{D}$ (relative to the model parameters), the parameter posterior is approximately Gaussian around the MAP estimate $\boldsymbol{\theta}_m^{MAP}$ (can also use 2nd order Taylor expansion of the log-posterior):

$$p(\boldsymbol{\theta}_m|\mathcal{D}, M_m) \simeq (2\pi)^{-d/2}|\boldsymbol{A}|^{1/2}\exp\left(-\frac{1}{2}(\boldsymbol{\theta}_m - \boldsymbol{\theta}_m^{MAP})^T \boldsymbol{A}(\boldsymbol{\theta}_m - \boldsymbol{\theta}_m^{MAP})\right),$$

$$A_{ij} = -\frac{\partial^2 \log P(\boldsymbol{\theta}_m|\mathcal{D}, M_m)}{\partial \boldsymbol{\theta}_{mi} \partial \boldsymbol{\theta}_{mj}}|_{\boldsymbol{\theta}_m^{MAP}}$$

Laplace Approximation

- We can write the model evidence as

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

- Using the Laplace approximation for the posterior of the parameters and evaluating the equation above at $\boldsymbol{\theta}_m^{MAP}$:

$$\log p(D|M_m) \simeq \log p(\theta_m^{MAP}, \mathcal{D}|M_m) - \log p(\theta_m^{MAP}|\mathcal{D}, M_m)$$
$$\simeq \log p(\mathcal{D}|\theta_m^{MAP}, M_m) + \log p(\theta_m^{MAP}|M_m) + \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|A|$$

- This Laplace approximation is used often for model comparison.

- Other approximations are also very useful:

  - Bayesian Information Criterion (BIC) (on the limit of N→∝)
  - MCMC (Sampling approach)
  - Variational Methods

Bayesian Information Criterion

- We start with the Laplace approximation on the limit of large data sets N→∝,

$$\log p(\mathcal{D}|M_m) \simeq \log p(\mathcal{D}|\boldsymbol{\theta}_m^{MAP}, M_m) + \log p(\boldsymbol{\theta}_m^{MAP}|M_m) + \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{A}|$$

- A N grows, $\boldsymbol{A}$ grows as $N\boldsymbol{A}_0$ for some fixed matrix $\boldsymbol{A}_0$, thus

$$\log|\boldsymbol{A}| \to \log|N\boldsymbol{A}_0| = \log\left(N^d\,|\boldsymbol{A}_0|\right) = d\log N + \log\left(|\boldsymbol{A}_0|\right) \xrightarrow{N\to\infty} d\log N$$

- Then the Laplace approximation is simplified as:

$$\log p(\mathcal{D}|M_m) \simeq \log p(\mathcal{D}|\boldsymbol{\theta}_m^{MAP}, M_m) - \frac{d}{2}\log N \quad (limit\ N \to \infty)$$

- Note interesting properties of (the easy to compute) BIC:
  - No dependence on the prior
  - One can use the MLE rather than the MAP estimate of $\boldsymbol{\theta}_m$
  - If not all parameters are well determined from the data, *d*=number of effective parameters.

Another Example of Model Selection

- Let us consider a regression model with the following particulars (d=k+1 dimensional data):[a]

$$y \mid \boldsymbol{w}, \sigma^2, \boldsymbol{\Phi} \sim \mathcal{N}_n(\boldsymbol{\Phi w}, \sigma^2 \boldsymbol{I}_n)$$

$$\boldsymbol{w} \mid \sigma^2, \boldsymbol{\Phi} \sim \mathcal{N}_{k+1}\left(\boldsymbol{w}_0, \sigma^2 \boldsymbol{M}^{-1}\right), \boldsymbol{M}\ a\ (k+1)\times(k+1)\ pos.\,def.\,symm.\,matrix$$

$$\sigma^2 \mid \boldsymbol{\Phi} \sim \mathit{InvGamma}(a,b), a,b>0$$

$$\boldsymbol{M} = \boldsymbol{I}_{k+1}/c, c>0\ and\ \boldsymbol{w}_0 = 0_{k+1}$$

- Our data are in a matrix form (dimension $N \times (k+1)$):

$$\boldsymbol{\Phi} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ 1 & x_{31} & x_{32} & \dots & x_{3k} \\ . & . & . & . & \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$$

We only slightly change our notation here to conform with the MatLab program implementing this example (from Zoubin Ghahramani)

Example: Likelihood Calculation

- We will derive the model evidence analytically. At first the likelihood can be written as:

$$\ell(\boldsymbol{w},\sigma^2 \mid \boldsymbol{y},\boldsymbol{\Phi}) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{\Phi w})^T(\boldsymbol{y}-\boldsymbol{\Phi w})\right)$$

- With simple algebra, we can rewrite the likelihood introducing the MLE estimate of the parameters as follows:

$$\ell(\boldsymbol{w},\sigma^2 \mid \boldsymbol{y},\boldsymbol{\Phi}) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{\Phi w}_{ML})^T(\boldsymbol{y}-\boldsymbol{\Phi w}_{ML}) - \frac{1}{2\sigma^2}(\boldsymbol{w}-\boldsymbol{w}_{ML})^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}(\boldsymbol{w}-\boldsymbol{w}_{ML})\right) =$$

$$= \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}s^2 - \frac{1}{2\sigma^2}(\boldsymbol{w}-\boldsymbol{w}_{ML})^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}(\boldsymbol{w}-\boldsymbol{w}_{ML})\right)$$

where

$$\boldsymbol{w}_{ML} = \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\boldsymbol{y}$$

$$s^2 \triangleq (\boldsymbol{y}-\boldsymbol{\Phi w}_{ML})^T(\boldsymbol{y}-\boldsymbol{\Phi w}_{ML})$$

Computing Model Evidence

- Our posterior is then of the following form:

$$p\left(\boldsymbol{w},\sigma^2 \mid \boldsymbol{w}_{ML},\boldsymbol{s}^2,\boldsymbol{\Phi}\right)=\frac{1}{(2\pi)^{N/2}}\frac{b^a}{\Gamma(a)}\sigma^{-k-n-2a-3}\exp\left(-\frac{1}{2\sigma^2}\left\{s^2+2b+\boldsymbol{w}^T\left(\frac{1}{c}\boldsymbol{I}+\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\boldsymbol{w}-2\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w}_{ML}+\boldsymbol{w}_{ML}^T\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\boldsymbol{w}_{ML}\right\}\right)$$

- The evidence is now computed as (use first the normalization of the *Inv Gamma* distribution):

$$p(\boldsymbol{y}/\mathcal{M})=\int\int \ell(\boldsymbol{w},\sigma^2\mid \boldsymbol{y},\boldsymbol{\Phi})\; p\left(\boldsymbol{w},\sigma^2 \mid \boldsymbol{w}_{ML},\boldsymbol{s}^2,\boldsymbol{\Phi}\right)d\boldsymbol{w}d\sigma^2=$$

$$\frac{1}{(2\pi)^{N/2}}\frac{b^a}{\Gamma(a)}\frac{1}{(2\pi)^{d/2}}c^{d/2}\int\int\sigma^{-k-n-2a-3}\exp\left(-\frac{1}{2\sigma^2}\underbrace{\left\{s^2+2b+\boldsymbol{w}^T\left(\frac{1}{c}\boldsymbol{I}+\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\boldsymbol{w}-2\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w}_{ML}+\boldsymbol{w}_{ML}^T\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\boldsymbol{w}_{ML}\right\}}_{A}\right)d\sigma^2d\boldsymbol{w}=$$

$$\frac{1}{(2\pi)^{N/2}}\frac{b^a}{\Gamma(a)}\frac{1}{(2\pi)^{d/2}}c^{d/2}\int\int\sigma^{-k-n-2a-3}\exp\left(-\frac{1}{2\sigma^2}A\right)d\sigma^2d\boldsymbol{w}=\frac{1}{(2\pi)^{N/2}}\frac{b^a}{\Gamma(a)}\frac{1}{(2\pi)^{d/2}}c^{d/2}\int\int\left(\sigma^2\right)^{(-k-n-2a-3)/2}\exp\left(-\frac{A/2}{\sigma^2}\right)d\sigma^2d\boldsymbol{w}=$$

$$\frac{1}{(2\pi)^{N/2}}\frac{b^a}{\Gamma(a)}\frac{1}{(2\pi)^{d/2}}c^{d/2}\int\int\left(\sigma^2\right)^{\underbrace{-(k+n+2a+1)/2-1}_{\alpha}}\exp\left(-\underbrace{(A/2)}_{\beta}\frac{1}{\sigma^2}\right)d\sigma^2d\boldsymbol{w}=\frac{1}{(2\pi)^{N/2}}\frac{b^a}{\Gamma(a)}\frac{1}{(2\pi)^{d/2}}c^{d/2}\int\frac{\Gamma\left((k+n+2a+1)/2\right)}{(A/2)^{(k+n+2a+1)/2}}d\boldsymbol{w}=$$

$$\frac{1}{(2\pi)^{N/2}}\frac{b^a}{\Gamma(a)}\frac{1}{(2\pi)^{d/2}}c^{d/2}\Gamma\left((k+n+2a+1)/2\right)2^{(k+n+2a+1)/2}\int\left[s^2+2b+\boldsymbol{w}^T\left(\frac{1}{c}\boldsymbol{I}+\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\boldsymbol{w}-2\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w}_{ML}+\boldsymbol{w}_{ML}^T\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\boldsymbol{w}_{ML}\right]^{-(k+n+2a+1)/2}d\boldsymbol{w}$$

- Useful formulas for the *Inv Gamma*: $p(\theta)=\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{-(\alpha+1)}e^{-\beta/\theta},\quad \theta>0$

Computing Model Evidence

- The evidence can be further simplified (we will use now the normalization ot the multivariate ***Student–t*** distribution)

$$p(\boldsymbol{y}/\mathcal{M}) = \frac{1}{(2\pi)^{N/2}} \frac{b^a}{\Gamma(a)} \frac{1}{(2\pi)^{d/2}} c^{d/2} \Gamma\left(\frac{d+n}{2}+a\right) 2^{\frac{d+n}{2}+a}$$

$$\boldsymbol{\mu} = \left(\frac{1}{c}\boldsymbol{I} + \boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1} \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\boldsymbol{w}_{ML}$$

$$\int \left[\underbrace{s^2 + 2b + \boldsymbol{w}_{ML}^T\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\boldsymbol{w}_{ML} - \boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}}_{g} + (\boldsymbol{w}-\boldsymbol{\mu})^T \left[\underbrace{\left(\frac{1}{c}\boldsymbol{I} + \boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}}_{\Sigma} \right]^{-1} (\boldsymbol{w}-\boldsymbol{\mu}) \right]^{-(d+n)/2-a} d\boldsymbol{w}$$

$$\upsilon = n + 2a$$

$$\frac{1}{(2\pi)^{N/2}} \frac{b^a}{\Gamma(a)} \frac{1}{(2\pi)^{d/2}} c^{d/2} \Gamma\left(\frac{d+n}{2}+a\right) 2^{\frac{d+n}{2}+a} g^{-(d+n)/2-a} \int \left[1 + \frac{1}{g}\left\{(\boldsymbol{w}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{w}-\boldsymbol{\mu})\right\}\right]^{-(d+n+2a)/2} d\boldsymbol{w} =$$

$$\frac{1}{(2\pi)^{N/2}} \frac{b^a}{\Gamma(a)} \frac{1}{(2\pi)^{d/2}} c^{d/2} \Gamma\left(\frac{d+n}{2}+a\right) 2^{\frac{d+n}{2}+a} g^{-(d+n)/2-a} \int \left[1 + \frac{1}{\upsilon}\left\{(\boldsymbol{w}-\boldsymbol{\mu})^T\left(\frac{g}{n+2a}\Sigma\right)^{-1}(\boldsymbol{w}-\boldsymbol{\mu})\right\}\right]^{-(d+\upsilon)/2} d\boldsymbol{w}$$

Useful formulas for the ***Student-t***

$$p(\theta) = \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)\nu^{d/2}\pi^{d/2}} |\Sigma|^{-1/2} \times \left(1 + \frac{1}{\nu}(\theta-\mu)^T\Sigma^{-1}(\theta-\mu)\right)^{-(\nu+d)/2}$$

Computing Model Evidence

- Performing the integration in $\boldsymbol{\beta}$ in the last slide:

$$p(\boldsymbol{y}/\mathcal{M}) = \frac{1}{(2\pi)^{N/2}} \frac{b^a}{\Gamma(a)} \frac{1}{(2\pi)^{d/2}} c^{d/2} 2^{\frac{d+n}{2}+a} g^{-(d+n)/2-a} \Gamma\left(\frac{n}{2}+a\right)(n+2a)^{d/2} \pi^{d/2} \left(\frac{g}{n+2a}\right)^{d/2} \left|\left(\frac{1}{c}\boldsymbol{I}+\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\right|^{-1/2} =$$

$$\frac{1}{(2\pi)^{N/2}} \frac{b^a}{\Gamma(a)} c^{d/2} \Gamma\left(\frac{n}{2}+a\right)\left(\frac{1}{2}s^2+b+\frac{1}{2}\boldsymbol{w}_{ML}^T\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\boldsymbol{w}_{ML}-\frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}\right)^{-n/2-a} \left|\left(\frac{1}{c}\boldsymbol{I}+\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\right|^{-1/2} =$$

$$\frac{1}{(2\pi)^{N/2}} \frac{b^a}{\Gamma(a)} c^{d/2} \Gamma\left(\frac{n}{2}+a\right)\left(\underbrace{\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{w}_{ML})^T(\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{w}_{ML})+b+\frac{1}{2}\boldsymbol{w}_{ML}^T\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\boldsymbol{w}_{ML}-\frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}}_{\mathcal{E}}\right)^{-n/2-a} \left|\left(\frac{1}{c}\boldsymbol{I}+\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\right|^{-1/2}$$

- Using some of the earlier definitions,

$$\boldsymbol{w}_{ML} = \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\boldsymbol{y} \quad s^2 \triangleq (\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{w}_{ML})^T(\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{w}_{ML}) \quad \boldsymbol{\mu} = \left(\frac{1}{c}\boldsymbol{I}+\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\boldsymbol{w}_{ML}$$

we can simplify $\mathcal{E}$ as:

$$\mathcal{E} = \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{w}_{ML})^T(\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{w}_{ML})+b+\frac{1}{2}\boldsymbol{w}_{ML}^T\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\boldsymbol{w}_{ML}-\frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu} = b+\frac{1}{2}\boldsymbol{y}^T\boldsymbol{y}+\frac{1}{2}\boldsymbol{w}_{ML}^T\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\boldsymbol{w}_{ML}-\boldsymbol{y}^T\boldsymbol{\Phi}\boldsymbol{w}_{ML}+\frac{1}{2}\boldsymbol{w}_{ML}^T\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\boldsymbol{w}_{ML}$$

$$-\frac{1}{2}\boldsymbol{w}_{ML}^T\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\left(\frac{1}{c}\boldsymbol{I}+\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\left(\frac{1}{c}\boldsymbol{I}+\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\left(\frac{1}{c}\boldsymbol{I}+\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\boldsymbol{w}_{ML} = b+\frac{1}{2}\boldsymbol{y}^T\boldsymbol{y}+\boldsymbol{y}^T\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\boldsymbol{y}$$

$$-\boldsymbol{y}^T\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\boldsymbol{y}-\frac{1}{2}\boldsymbol{w}_{ML}^T\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\left(\frac{1}{c}\boldsymbol{I}+\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\boldsymbol{y} = b+\frac{1}{2}\boldsymbol{y}^T\boldsymbol{y}-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{\Phi}\left(\frac{1}{c}\boldsymbol{I}+\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\boldsymbol{y}$$

Model Evidence

- The final evidence in analytical form is given as:

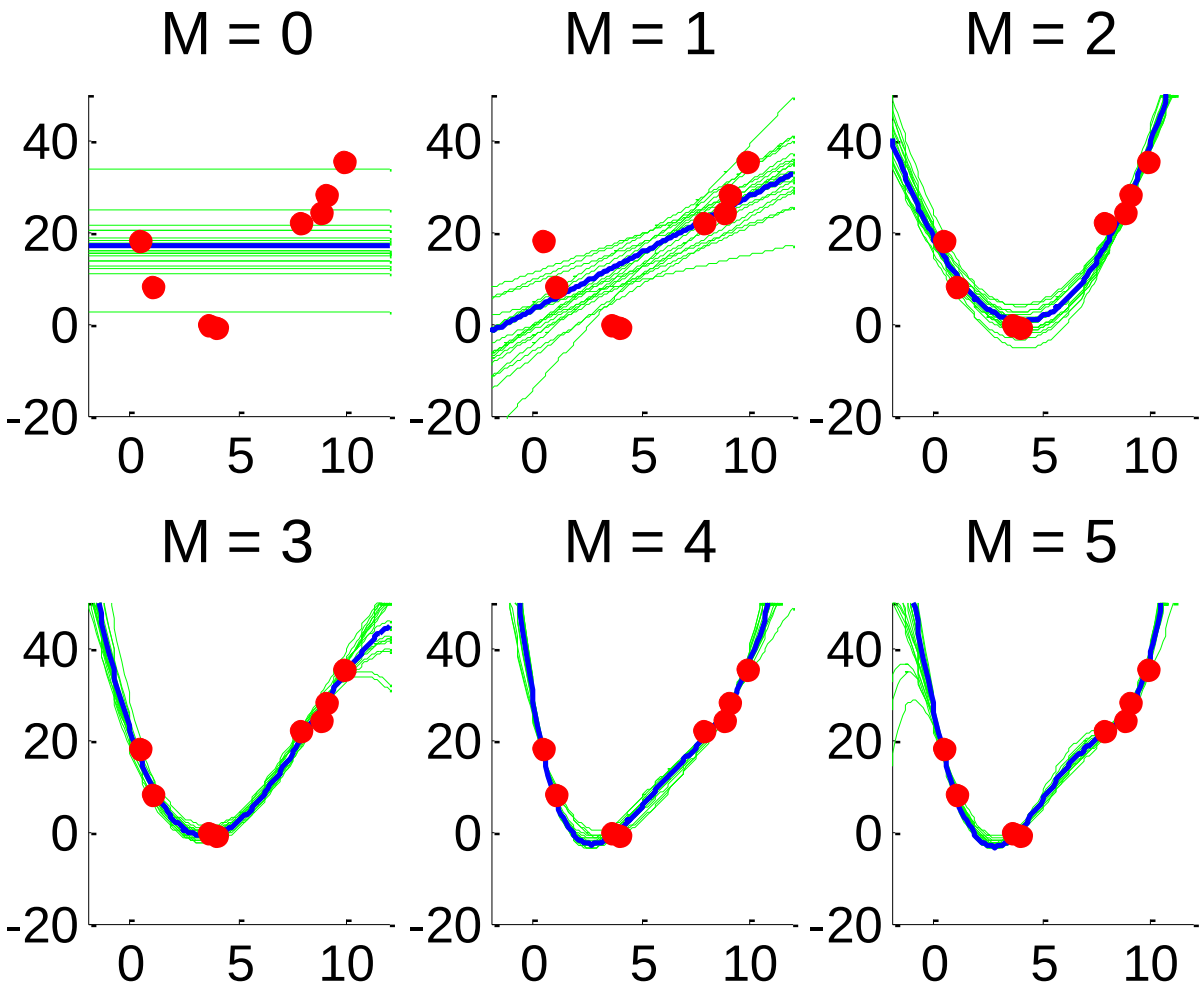
$$p(\boldsymbol{y}/\mathcal{M}) = \frac{1}{(2\pi)^{N/2}} \frac{b^a}{\Gamma(a)} c^{d/2} \Gamma\left(\frac{n}{2}+a\right) \left(b + \frac{1}{2}\boldsymbol{y}^T\boldsymbol{y} - \frac{1}{2}\boldsymbol{y}^T\boldsymbol{\Phi}\left(\frac{1}{c}\boldsymbol{I} + \boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\boldsymbol{y} \right)^{-n/2-a} \left| \left(\frac{1}{c}\boldsymbol{I} + \boldsymbol{\Phi}^T\boldsymbol{\Phi}\right) \right|^{-1/2}$$

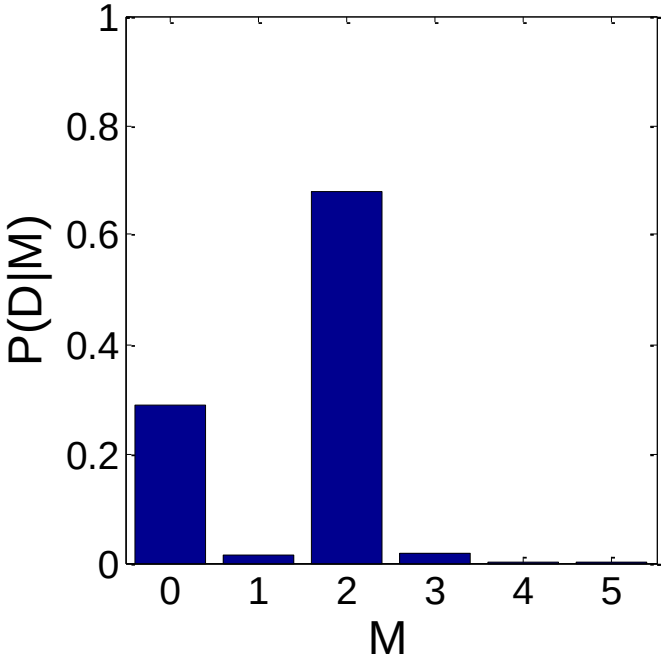
- Compare this with what is given in this MatLab Implementation.

- The model evidence and samples of different order (M) regression models are given below. The specific data of the problem can be found in the MatLab file.

- We are looking for the order of the polynomial that maximizes the evidence. Note that the MatLab implementation utilized random input/output data.

Bayesian Model Comparison





MatLab implementation of Bayesian Model Selection (from Zoubin Ghahramani)

Linear Models Of Regression: Equivalent Kernel Representation

Equivalent Kernel

- The predictive mean in our regression model can be written as:

$$y(\boldsymbol{x}, \boldsymbol{m}_N) = \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{m}_N = \beta \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{S}_N \boldsymbol{\Phi}^T \boldsymbol{t} = \sum_{n=1}^{N} \underbrace{\beta \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{S}_N \boldsymbol{\phi}(\boldsymbol{x}_n)}_{\text{Equivalent kernel } k(\boldsymbol{x}, \boldsymbol{x}_n)} t_n \Rightarrow$$

$$y(\boldsymbol{x}, \boldsymbol{m}_N) = \sum_{n=1}^{N} k(\boldsymbol{x}, \boldsymbol{x}_n) t_n$$

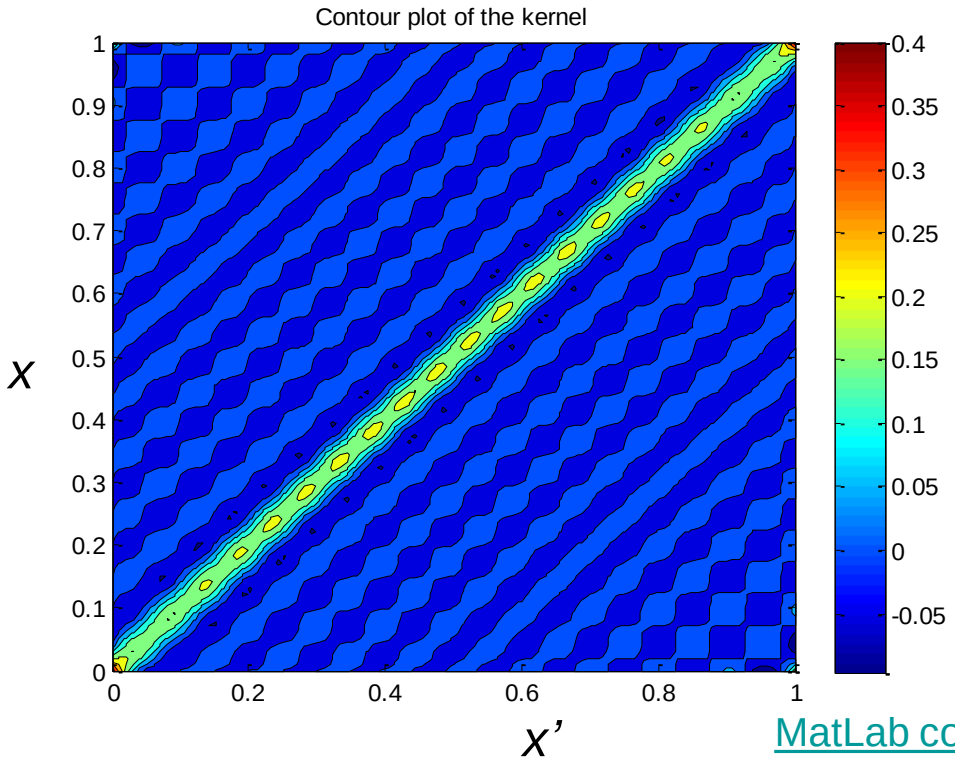
$$k(\boldsymbol{x}, \boldsymbol{x}_n) = \beta \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{S}_N \boldsymbol{\phi}(\boldsymbol{x}_n)$$

The kernel is shown here as a plot of *x vs. x'*.

20 samples x_n equally spaced in *[0,1]*

Gaussian kernels (s=0.05), $\alpha = 5 \times 10^{-3}$, $\beta = 11.1$.

$$\phi_j(x) = \exp\left(-\frac{(x - x_j)^2}{2s^2}\right)$$



MatLab code

Equivalent Kernel

- The predictive mean can be written as follows:

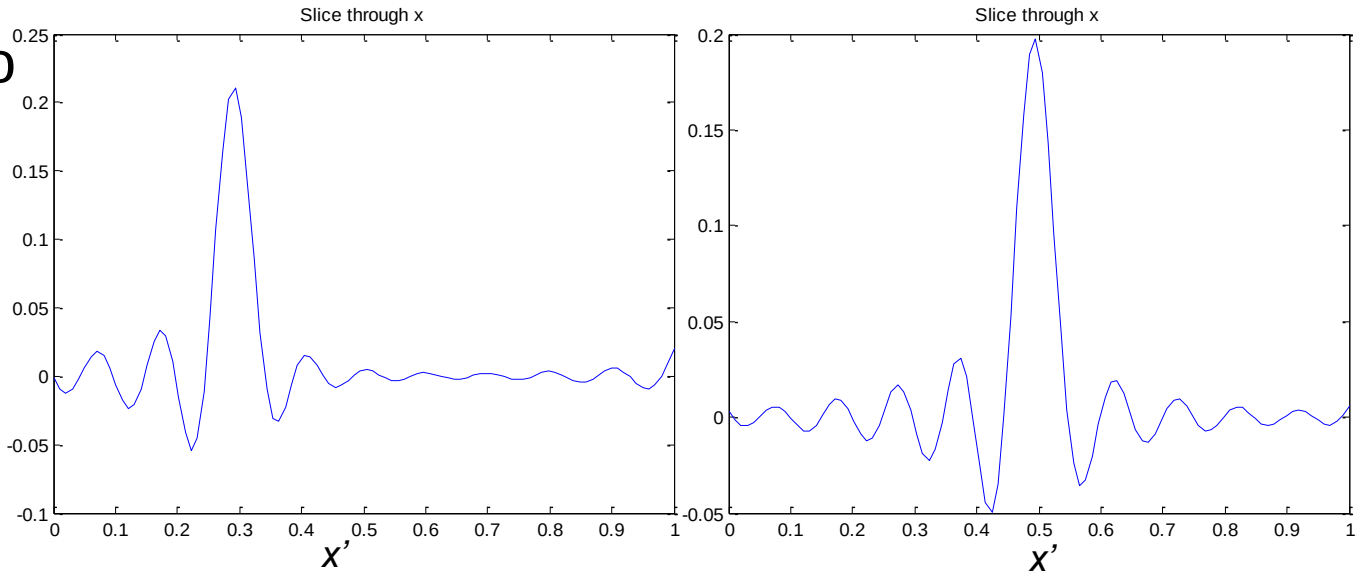
$$y(\boldsymbol{x}, \boldsymbol{m}_N) = \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{m}_N = \beta \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{S}_N \boldsymbol{\Phi}^T \boldsymbol{t} = \sum_{n=1}^{N} \underbrace{\beta \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{S}_N \boldsymbol{\phi}(\boldsymbol{x}_n)}_{\text{Equivalent kernel } k(\boldsymbol{x}, \boldsymbol{x}_n)} t_n \Rightarrow$$

$$y(\boldsymbol{x}, \boldsymbol{m}_N) = \sum_{n=1}^{N} k(\boldsymbol{x}, \boldsymbol{x}_n) t_n \qquad k(\boldsymbol{x}, \boldsymbol{x}_n) = \beta \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{S}_N \boldsymbol{\phi}(\boldsymbol{x}_n)$$

Two slices corresponding to *x=0.3, x=0.5.*

20 Gaussians (s=0.05)

Plots for 100 points *x* uniformly in (0,1)

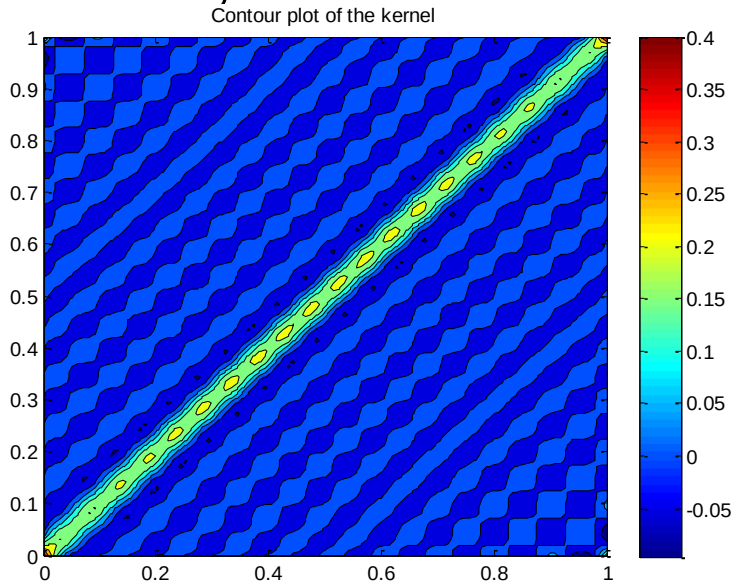


MatLab code

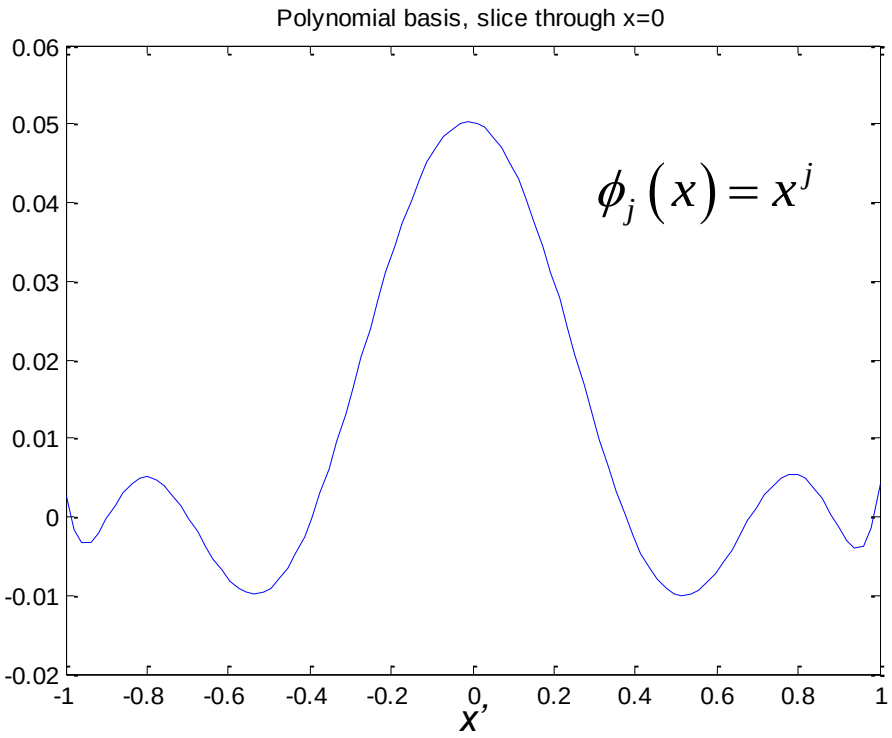
Equivalent Kernel

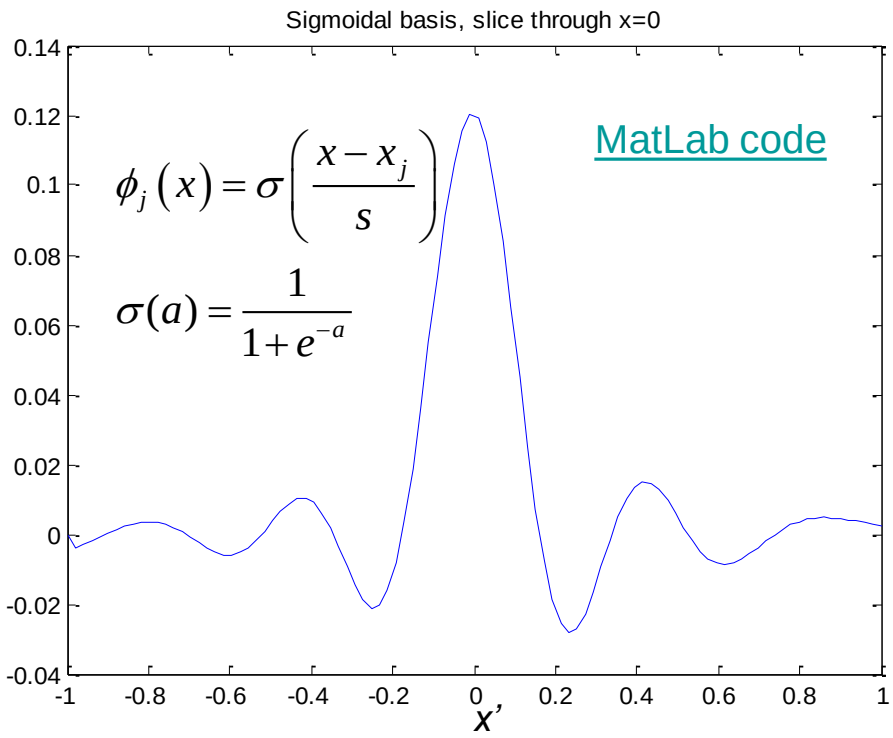
$$y(\boldsymbol{x}, \boldsymbol{m}_N) = \sum_{n=1}^{N} k(\boldsymbol{x}, \boldsymbol{x}_n) t_n \qquad k(\boldsymbol{x}, \boldsymbol{x}_n) = \beta \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{S}_N \boldsymbol{\phi}(\boldsymbol{x}_n)$$

- Note that we make predictions by taking linear combinations of the training set target values (*linear smoothers).*
- The equivalent kernel depends on the input values $\boldsymbol{x}_n$ since these appear in S_N.
- See from the figure that its localized around x, and so the mean of the predictive distribution $y(x,\boldsymbol{m}_N)$, is obtained by forming a weighted combination of t_n in which data points close to x are given higher weight than points (evidence) further removed from x.



Equivalent Kernel





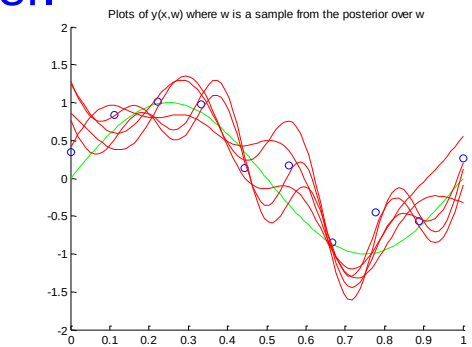
- Examples of equivalent kernels *k(x, x')* for *x = 0* plotted as a function of *x'*. 10 basis functions were used in each case. The sigmoidal basis is centered at 10 equally spaced x_n in [-1,1].
- These are localized functions of *x'* even though the corresponding basis functions are nonlocal.

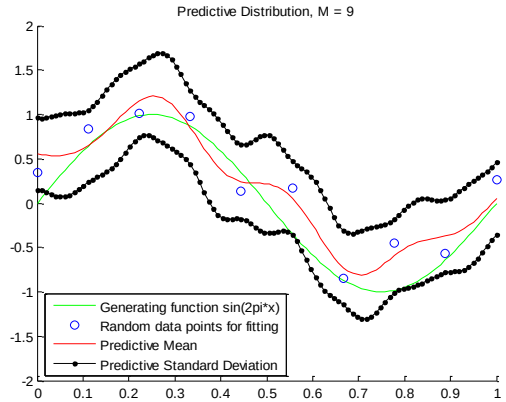
Equivalent Kernel

- Consider the covariance between $y(\boldsymbol{x})$ and $y(\boldsymbol{x}')$, which is given by (recall that $p(\boldsymbol{w} \mid \boldsymbol{x}, \boldsymbol{t}, \alpha, \beta) = \mathcal{N}\left(\boldsymbol{w} \mid \beta \boldsymbol{S}_N \sum_{n=1}^{N} t_n \boldsymbol{\phi}(x_n), \boldsymbol{S}_N\right)$):

$$\operatorname{cov}\left[y(\boldsymbol{x}), y(\boldsymbol{x}')\right] = \operatorname{cov}\left[\boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{w}, \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}')\right] = \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{S}_N \boldsymbol{\phi}(\boldsymbol{x}') = \beta^{-1} k(\boldsymbol{x}, \boldsymbol{x}')$$

- From the earlier discussion on the equivalent kernel, we see that the predictive mean at nearby points will be highly correlated, whereas for more distant points the correlation will be smaller.
- By drawing samples from the posterior distribution over $\boldsymbol{w}$, and plotting the corresponding model functions $y(x,\boldsymbol{w})$, we are visualizing the joint uncertainty in the posterior distribution between the y values at two (or more) x values, as governed by the equivalent kernel.
- This is contrary to the Figure here where we visualized pointwise uncertainty in the predictions.





Equivalent Kernel

- We can avoid the use of basis functions and define the kernel function directly, leading e.g. to *Gaussian Processes* (more on Kernel methods and GPs later on in this course).
- *Note that*

$$\sum_{n=1}^{N} k(\boldsymbol{x}, \boldsymbol{x}_n) = 1$$

for all values of **x.** However, the equivalent kernel

$$k(\boldsymbol{x}, \boldsymbol{x}_n) = \beta \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{S}_N \boldsymbol{\phi}(\boldsymbol{x}_n)$$

may be negative for some values of ***x.***

- Like all kernel functions, the equivalent kernel can be expressed as an inner product:

$$k(\boldsymbol{x}, \boldsymbol{z}) = \beta \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{S}_N \boldsymbol{\phi}(\boldsymbol{z}) \Rightarrow \quad k(\boldsymbol{x}, \boldsymbol{z}) = \psi(\boldsymbol{x})^T \psi(\boldsymbol{z})$$
$$\psi(\boldsymbol{x}) = \beta^{1/2} \boldsymbol{S}_N^{1/2} \boldsymbol{\phi}(\boldsymbol{x})$$

- This property can be used to "design" interesting kernels.

Introduction to Variable Selection

Variable Selection

- Let us return to our regression model with one dependent random variable y and a set of *k* $\{x_1, x_2, ..., x_k\}$ explanatory variables.
- Are all the x_i's needed in the regression?
- We assume that every q-subset $\{i_1, i_2, ..., i_q\}$, $0 \le q \le k$, of the explanatory variables,

$$\{1, x_{i_1}, x_{i_2}, ..., x_{i_q}\}$$

is a proper set of explanatory variables for the regression of y (as before, the intercept is included in all models).

- We have a total of 2^k models to select from!

Variable Selection

- Following earlier notation, we denote: $\boldsymbol{X} = \left[\mathbf{1}_n \; x_1 \; x_2 \ldots x_k\right]$ as the matrix that contains $\mathbf{1}_n$ and the *k* potential predictor variables.
- Each model M_γ is associated with binary indicator vector

$$\gamma \in \Gamma = \{0,1\}^k$$

where γ_i=1 means that the variable x_i is included in the model M_γ and γ_i=0 that it is not.

- The number of variables included in the model M_γ is:

$$q_\gamma = \mathbf{1}^T \boldsymbol{\gamma}$$

- The indices of the variables included in the model and not included in the model are denoted, respectively, as: $t_1(\gamma), t_0(\gamma)$

Variable Selection – Models in Competition

- For $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$ and $\boldsymbol{X}$, we define $\boldsymbol{\beta}_\gamma$ as the sub-vector

$$\boldsymbol{\beta}_\gamma = \left(\beta_0, (\beta_i)_{i \in t_1(\gamma)}\right)$$

- Let $\boldsymbol{X}_\gamma$ be the submatrix of $\boldsymbol{X}$ where only the column $\mathbf{1}_n$ and the columns in $t_1(\gamma)$ have been left.

- The model M_γ is then defined as:

$$\boldsymbol{y} \,|\, \gamma, \boldsymbol{\beta}_\gamma, \sigma^2, X \sim \mathcal{N}_n\left(\boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma, \sigma^2 \boldsymbol{I}_n\right)$$

where $\beta_\gamma \in \mathbb{R}^{q_\gamma + 1}, \sigma^2 \in \mathbb{R}_+^*$ are the unknown parameters.

- The σ^2 is common to all models and we use the same prior for all models.

Variable Selection – Models in Competition

- We have a high number 2^k of models in competition.
- We cannot specify a prior on every M_γ in a completely subjective and autonomous manner.
- We derive all priors from a single global prior **associated with the full model** that corresponds to $\gamma = (1, \ldots, 1)$.

Zellner's Informative Prior: Variable Selection

- For the full model that corresponds to $\gamma = (1, \ldots, 1)$, we use the Zellner's informative G-prior:

$$\boldsymbol{\beta}|\sigma^2, \boldsymbol{X} \sim \mathcal{N}_{k+1}\left(\tilde{\boldsymbol{\beta}}, c\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)$$
$$\sigma^2 \sim \pi(\sigma^2|\boldsymbol{X}) \propto \sigma^{-2}\ improper\ Jeffreys\ prior$$

- For each model M_γ, the prior distribution of β_γ conditional on σ^2 is fixed as:

$$\boldsymbol{\beta}_\gamma\ |\boldsymbol{\gamma}, \sigma^2 \sim \mathcal{N}_{q_\gamma+1}\left(\tilde{\boldsymbol{\beta}}_\gamma, c\sigma^2\left(\boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma\right)^{-1}\right)$$

where $\tilde{\boldsymbol{\beta}}_\gamma = \left(X_\gamma^T\boldsymbol{X}_\gamma\right)^{-1}\boldsymbol{X}_\gamma^T\tilde{\boldsymbol{\beta}}$ and same prior on σ^2.

Zellner's Informative Prior: Variable Selection – Prior

- The joint prior for model M_γ is the improper prior

$$\pi(\boldsymbol{\beta}_\gamma, \sigma^2|\boldsymbol{\gamma}) \propto (\sigma^2)^{-(q_\gamma+1)/2-1} \exp\left[-\frac{1}{2(c\sigma^2)}(\boldsymbol{\beta}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)^T(\boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma)(\boldsymbol{\beta}_\gamma - \tilde{\boldsymbol{\beta}}_\gamma)\right]$$

- Infinitely many ways of defining a prior on the model index γ:

  Our choice is a uniform prior $p(\gamma|\boldsymbol{X}) = 2^{-k}$.

- Posterior distribution of γ is central to variable selection since it is proportional to marginal density of **y** on M_γ (or evidence of M_γ)

Zellner's Informative Prior: Variable Selection – Prior

- Posterior distribution of γ is proportional to the marginal density of **y** on M_γ (so it can also be used to compute Bayes factors)

$$\pi(\boldsymbol{\gamma}\,|\,\mathbf{y},\mathbf{X}) \propto f(\mathbf{y}\,|\,\boldsymbol{\gamma},\mathbf{X})\pi(\boldsymbol{\gamma}\,|\,\mathbf{X}) \propto f(\mathbf{y}\,|\,\boldsymbol{\gamma},\mathbf{X})$$
$$= \int\left(\int f(\mathbf{y}\,|\,\boldsymbol{\gamma},\boldsymbol{\beta},\sigma^2,\mathbf{X})\pi(\boldsymbol{\beta}\,|\,\boldsymbol{\gamma},\sigma^2,\mathbf{X})d\boldsymbol{\beta}\right)\pi(\sigma^2\,|\,\mathbf{X})d\sigma^2$$

where it can be shown

$$f(\mathbf{y}|\boldsymbol{\gamma},\sigma^2,\mathbf{X}) = \int f(\mathbf{y}|\boldsymbol{\gamma},\boldsymbol{\beta},\sigma^2)\pi(\boldsymbol{\beta}|\boldsymbol{\gamma},\sigma^2)d\boldsymbol{\beta} =$$
$$= (c+1)^{-(q_\gamma+1)/2}(2\pi)^{-n/2}(\sigma^2)^{-n/2} \times$$
$$\exp\left(-\frac{1}{2\sigma^2}\mathbf{y}^T\mathbf{y} + \frac{1}{2\sigma^2(c+1)}\left\{c\mathbf{y}^T\mathbf{X}_\gamma\left(\mathbf{X}_\gamma^T\mathbf{X}_\gamma\right)^{-1}\mathbf{X}_\gamma^T\mathbf{y} - \tilde{\boldsymbol{\beta}}_\gamma^T\mathbf{X}_\gamma^T\mathbf{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma + 2\mathbf{y}^T\mathbf{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma\right\}\right)$$

Zellners Informative Prior: Variable Selection - Prior

$$\pi(\gamma \mid y, X) \propto \int f\left(y \mid \gamma, \sigma^2, X\right)\pi\left(\sigma^2 \mid X\right)d\sigma^2 = \int f\left(y \mid \gamma, \sigma^2, X\right)\frac{1}{\sigma^2}d\sigma^2$$

- Posterior distribution of γ is then given as:

$$f(\gamma|\boldsymbol{y}, \boldsymbol{X}) \propto (c+1)^{-(q_\gamma+1)/2} \times$$
$$\left(\boldsymbol{y}^T\boldsymbol{y} - \frac{c}{(c+1)}\boldsymbol{y}^T\boldsymbol{X}_\gamma\left(\boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma\right)^{-1}\boldsymbol{X}_\gamma^T\boldsymbol{y} + \frac{1}{c+1}\tilde{\boldsymbol{\beta}}_\gamma^T\boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma - \frac{2}{c+1}\boldsymbol{y}^T\boldsymbol{X}_\gamma\tilde{\boldsymbol{\beta}}_\gamma\right)^{-n/2}$$

- This is of similar form to a familiar result for a fixed c:

$$f(\boldsymbol{y}|\boldsymbol{X}, c) = (c+1)^{-(k+1)/2}\pi^{-n/2}\Gamma(\frac{n}{2})\left[\boldsymbol{y}^T\boldsymbol{y} - \frac{c}{c+1}\boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} + \frac{1}{c+1}\tilde{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\tilde{\boldsymbol{\beta}} - \frac{2}{c+1}\boldsymbol{y}^T\boldsymbol{X}\tilde{\boldsymbol{\beta}}\right]^{-n/2}$$

Model Selection

- Most likely models ordered by decreasing posterior probabilities using Zellner's informative G-prior with *c=100*.

$t_1(\gamma)$ $\pi(\gamma \mid \boldsymbol{y}, \boldsymbol{X})$

t1_gamma	pi(gamma | y, X)
0 1 2 4 5	0.231543
0 1 2 4 5 9	0.037358
0 1 9	0.034435
0 1 2 4 5 10	0.032975
0 1 4 5	0.030606
0 1 2 9	0.025016
0 1 2 4 5 7	0.024144
0 1 2 4 5 8	0.023784
0 1 2 4 5 6	0.023735
0 1 2 3 4 5	0.023207
0 1 6 9	0.014587
0 1 2 3 9	0.014491
0 9	0.014281
0 1 2 6 9	0.013551
0 1 4 5 9	0.012761
0 1 3 9	0.011712
0 1 2 8	0.011477
0 1 8	0.009519
0 1 2 3 4 5 9	0.009036
0 1 2 4 5 6 9	0.009031

The data set corresponds to the caterpillar regression problem (details in Bayesian Core) with 10 explanatory variables.

Model Selection

- Model M_γ with the highest posterior probability is $t_1(\gamma) = (1, 2, 4, 5)$, which corresponds to the variables
  - altitude,
  - slope,
  - height of the tree sampled in the center of the area, and
  - diameter of the tree sampled in the center of the area.

Model Selection

- For the Zellner's non-informative prior with $\pi(c)=1/c$, we have ($\tilde{\boldsymbol{\beta}} = \mathbf{0}_{k+1}$) :

$$\pi\left(\boldsymbol{\gamma} \mid \boldsymbol{y}, \boldsymbol{X}\right) = \sum_{c=1}^{\infty} c^{-1}\left(c+1\right)^{-\left(q_{\gamma}+1\right)/2}\left[\boldsymbol{y}^{T}\boldsymbol{y} - \frac{c}{c+1}\boldsymbol{y}^{T}\boldsymbol{X}_{\gamma}\left(\boldsymbol{X}_{\gamma}^{T}\boldsymbol{X}_{\gamma}\right)^{-1}\boldsymbol{X}_{\gamma}^{T}\boldsymbol{y}\right]^{-n/2}$$

- Again this is of the same form as (see Bayesian Core)

$$f\left(\boldsymbol{y} \mid \boldsymbol{X}\right) = \sum_{c=1}^{\infty} f\left(\boldsymbol{y} \mid \boldsymbol{X}, c\right) c^{-1} \propto$$

$$\sum_{c=1}^{\infty} c^{-1}(c+1)^{-(k+1)/2}\left[\boldsymbol{y}^{T}\boldsymbol{y} - \frac{c}{c+1}\boldsymbol{y}^{T}\boldsymbol{X}\left(\boldsymbol{X}^{T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}\boldsymbol{y}\right]^{-n/2}$$

Model Selection

- Most likely models ordered by decreasing posterior probabilities using Zellner's non-informative G-prior.

$t_1(\gamma)$	$\pi(\gamma \mid \boldsymbol{y}, \boldsymbol{X})$
t1_gamma	pi(gamma | y, X)
0 1 2 4 5	0.092914
0 1 2 4 5 9	0.032553
0 1 2 4 5 10	0.029512
0 1 2 4 5 7	0.023114
0 1 2 4 5 8	0.022843
0 1 2 4 5 6	0.022807
0 1 2 3 4 5	0.022409
0 1 2 3 4 5 9	0.016733
0 1 2 4 5 6 9	0.016725
0 1 2 4 5 8 9	0.013726
0 1 4 5	0.011031
0 1 2 4 5 9 10	0.009933
0 1 2 3 9	0.009698
0 1 2 9	0.009316
0 1 2 4 5 7 9	0.009253
0 1 2 6 9	0.009189
0 1 4 5 9	0.008756
0 1 2 3 4 5 10	0.007933
0 1 2 4 5 8 10	0.007901
0 1 2 4 5 7 10	0.007896

Stochastic Search for the Most Likely Model

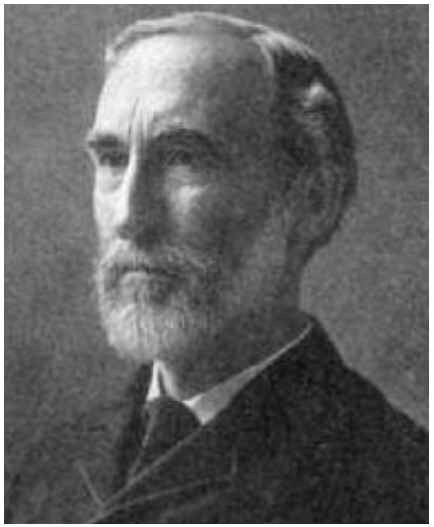
- When k is large, it becomes computationally intractable to compute the posterior probabilities of the 2^k models.

- Need of a tailored algorithm that samples from $\pi(\gamma|\mathbf{y},\mathbf{X})$ and selects the most likely models.

- Can be done by Gibbs sampling*, given the availability of the full conditional posterior probabilities of the γ_i's. If

$\gamma_{-i} = (\gamma_1, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots, \gamma_k)$ $(1 \le i \le k)$
then
$\pi(\gamma_i | \mathbf{y}, \gamma_{-i}, \mathbf{X}) \propto \pi(\gamma | \mathbf{y}, \mathbf{X})$

(to be evaluated in both $\gamma_i = 0$ and $\gamma_i = 1$)



* Gibbs and other sampling algorithms will be introduced and discussed in detail in forthcoming lectures.

Gibbs Sampling for Variable Selection

Initialization: Draw γ^0 from the uniform distribution on Γ

Iteration t: Given $\left(\gamma_1^{(t-1)},...,\gamma_k^{(t-1)}\right)$, generate

- $\gamma_1^{(t)}$ *according to* $\pi\left(\gamma_1 \mid \boldsymbol{y}, \gamma_2^{(t-1)},...,\gamma_k^{(t-1)}, \boldsymbol{X}\right)$
- $\gamma_2^{(t)}$ *according to* $\pi\left(\gamma_2 \mid \boldsymbol{y}, \gamma_1^{(t)}, \gamma_3^{(t-1)},...,\gamma_k^{(t-1)}, \boldsymbol{X}\right)$
- ..
- ..
- $\gamma_k^{(t)}$ *according to* $\pi\left(\gamma_k \mid \boldsymbol{y}, \gamma_1^{(t)}, \gamma_2^{(t)},...,\gamma_{k-1}^{(t)}, \boldsymbol{X}\right)$

Gibbs Sampling for Variable Selection

Question: How to sample γ_1^t according to $\pi\left(\gamma_1 \mid \boldsymbol{y}, \gamma_2^{(t-1)}, \ldots, \gamma_k^{(t-1)}, \boldsymbol{X}\right)$

1. The conditional distribution $\pi\left(\gamma_1 \mid \boldsymbol{y}, \gamma_2^{(t-1)}, \ldots, \gamma_k^{(t-1)}, \boldsymbol{X}\right)$ is proportional to $\pi\left(\gamma \mid \boldsymbol{y}, \boldsymbol{X}\right)$

2. Since γ_1^t only has two possible values which are 0 and 1, we get

$$p_0 \propto \pi\left(\gamma_1 = 0, \gamma_2^{(t-1)}, \ldots, \gamma_k^{(t-1)} \mid \boldsymbol{y}, \boldsymbol{X}\right)$$

$$p_1 \propto \pi\left(\gamma_1 = 1, \gamma_2^{(t-1)}, \ldots, \gamma_k^{(t-1)} \mid \boldsymbol{y}, \boldsymbol{X}\right)$$

So the probability that $\gamma_1^t = 1$ is p_1, then we can use Gibbs Sampling to approximate the distribution of $\{\gamma_i^t\}$

Gibbs Sampling: Posterior Probabilities

- After T ≫ 1 MCMC iterations, we approximate the posterior probabilities p(γ|**y**,**X**) by empirical averages

$$\hat{\pi}(\gamma|\boldsymbol{y},\boldsymbol{X}) = \frac{1}{T-T_0+1}\sum_{t=T_0}^{T} \mathbb{I}_{\gamma^{(t)}=\gamma}$$

The T_0 first values (burn in) in the MCMC chain are eliminated.

Model Choice Comparison: Gibbs Estimates

First level Informative G-prior model with ($\tilde{\boldsymbol{\beta}} = 0_{11}$, c=100) compared with the Gibbs estimates of the top ten posterior probabilities

$t_1(\gamma)$	$\pi(\gamma \mid \boldsymbol{y}, \boldsymbol{X})$	$\hat{\pi}(\gamma \mid \boldsymbol{y}, \boldsymbol{X})$
t1_gamma	pi(gamma | y, X)	pi_hat(gamma | y, X)
0 1 2 4 5	0.231543	0.239276
0 1 2 4 5 9	0.037358	0.034397
0 1 9	0.034435	0.032397
0 1 2 4 5 10	0.032975	0.030097
0 1 4 5	0.030606	0.029397
0 1 2 9	0.025016	0.025297
0 1 2 4 5 7	0.024144	0.022498
0 1 2 4 5 8	0.023784	0.024898
0 1 2 4 5 6	0.023735	0.023598
0 1 2 3 4 5	0.023207	0.022998

Model Choice Comparison: Gibbs Estimates

Non-informative G-prior variable model choice compared with the Gibbs estimates of the top ten posterior probabilities

$t_1(\gamma)$	$\pi(\gamma \mid \boldsymbol{y}, \boldsymbol{X})$	$\hat{\pi}(\gamma \mid \boldsymbol{y}, \boldsymbol{X})$
t1_gamma	pi(gamma | y, X)	pi_hat(gamma | y, X)
0 1 2 4 5	0.092914	0.093391
0 1 2 4 5 9	0.032553	0.033097
0 1 2 4 5 10	0.029512	0.032597
0 1 2 4 5 7	0.023114	0.025097
0 1 2 4 5 8	0.022843	0.023098
0 1 2 4 5 6	0.022807	0.022498
0 1 2 3 4 5	0.022409	0.021698
0 1 2 3 4 5 9	0.016733	0.015998
0 1 2 4 5 6 9	0.016725	0.014899
0 1 2 4 5 8 9	0.013726	0.013399

Gibbs Sampling: Probabilities of Inclusion

- An approximation of the probability to include the i-th variable:

$$\hat{P}^{\pi}(\gamma_i = 1 | \boldsymbol{y}, \boldsymbol{X}) = \frac{1}{T - T_0 + 1} \sum_{t=T_0}^{T} \mathbb{I}_{\gamma_i^{(t)}=1}$$

Probabilities of Inclusion Estimates

Informative ($\tilde{\boldsymbol{\beta}} = 0_{11}$, c=100) and non-informative G-prior variable inclusion estimates (based on the same Gibbs output as in the earlier two tables)

γ_i	$\hat{P}^{\pi}(\gamma_i = 1 \mid \boldsymbol{y}, \boldsymbol{X})$	$\hat{P}^{\pi}(\gamma_i = 1 \mid \boldsymbol{y}, \boldsymbol{X})$
gamma_i	P(gamma_i|y, X)	P(gamma_i|y, X)
gamma_1	0.8733	0.8806
gamma_2	0.7100	0.7789
gamma_3	0.1515	0.2958
gamma_4	0.6842	0.7422
gamma_5	0.6635	0.7234
gamma_6	0.1659	0.2992
gamma_7	0.1343	0.2812
gamma_8	0.1478	0.2740
gamma_9	0.3942	0.5015
gamma_10	0.1135	0.2556