Introduction to Bayesian Linear Regression & Model Selection

Prof. Nicholas Zabaras
Center for Informatics and Computational Science
https://cics.nd.edu/
University of Notre Dame
Notre Dame, Indiana, USA

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

September 18, 2018

Contents

- Empirical Bayes Example (Continuing from last lecture)
- Bayesian Computing and Machine Learning, Motivation to Bayesian inference via a regression example, Over fitting, Effect of Data Size, Model Selection, Over fitting and MLE, Regularization and Model Complexity
- Bayesian Inference and Prediction, Frequentist Vs Bayesian Paradigm, Bias in MLE (Gaussian Example), A Probabilistic View of Regression, MAP Estimate and Regularized Least Squares, Posterior Distribution, Predictive Distribution
- Model Selection and Cross Validation, AIC Information Criterion, Bayesian Model Selection, Bayesian Occam's Razor, Marginal Likelihood, Evidence Approximation, Examples
- Laplace Approximation, Bayesian Information Criterion, Akaike Information Criterion, Effect of the Prior, Empirical Bayes, Bayes Factors and Jeffreys Scale of Evidence, Examples, Jeffreys-Lindley Paradox
  - Chris Bishops' PRML book, Chapters 1 and 2
  - Kevin Murphy's, Machine Learning: A probabilistic perspective, Chapter 5
  - C P Robert, The Bayesian Choice: From Decision-Theoretic Motivations to Compulational Implementation, Springer-Verlag, NY, 2001 (online resource)
  - A. Gelman, JB Carlin, HS Stern and DB Rubin, Bayesian Data Analysis, Chapman and Hall CRC Press, 2nd Edition, 2003.
  - M Marin and C P Robert, The Bayesian Core, Spring Verlag, 2007 (online resource)
  - Bayesian Statistics for Engineering, Online Course at Georgia Tech, B. Vidakovic.

Empirical Bayes

- *Empirical Bayes violates the principle that the prior should be chosen independently of the data.*

- We can just view it as a cheap approximation to inference in a hierarchical Bayesian model, just as we viewed MAP estimation as an approximation to inference in the one level model $\boldsymbol{\theta} \rightarrow \mathcal{D}$.

- We can construct a hierarchy in which the more integrals one performs, the "more Bayesian" one becomes:

Method	Definition
Maximum likelihood	$\hat{\theta} = \text{argmax}_{\boldsymbol{\theta}}\, p(\mathcal{D}\|\theta)$
MAP estimation	$\hat{\theta} = \text{argmax}_{\boldsymbol{\theta}}\, p(\mathcal{D}\|\theta)p(\theta\|\eta)$
ML-II (Empirical Bayes)	$\hat{\eta} = \text{argmax}_{\boldsymbol{\eta}} \int p(\mathcal{D}\|\theta)p(\theta\|\eta)d\theta = \text{argmax}_{\boldsymbol{\eta}}\, p(\mathcal{D}\|\eta)$
MAP-II	$\hat{\eta} = \text{argmax}_{\boldsymbol{\eta}} \int p(\mathcal{D}\|\theta)p(\theta\|\eta)p(\eta)d\theta = \text{argmax}_{\boldsymbol{\eta}}\, p(\mathcal{D}\|\eta)p(\eta)$
Full Bayes	$p(\theta, \eta\|\mathcal{D}) \propto p(\mathcal{D}\|\theta)p(\theta\|\eta)p(\eta)$

Empirical Bayes: Gaussian-Gaussian Model

- We now consider an example where the data is real-valued. We use *a Gaussian likelihood and a Gaussian prior.*

- Suppose we have data from multiple related groups, e.g. x_{ij} is the test score for student i in school $j, j = 1:D, i = 1:N_j$. *We want to estimate the mean score for each school, θ_j.*

- *Since N_j may be small for some schools, we regularize the problem by using a hierarchical Bayesian model, where θ_j comes from a common prior, $\mathcal{N}(\mu, \tau^2)$.*

- The joint distribution has the following form:

$$p\left(\boldsymbol{\theta}, \mathcal{D} \mid \boldsymbol{\eta}, \sigma^2\right) = \prod_{j=1}^{D}\left(\prod_{i=1}^{N_j} \mathcal{N}\left(x_{ij} \mid \theta_j, \sigma^2\right)\mathcal{N}\left(\theta_j \mid \mu, \tau^2\right)\right), \boldsymbol{\eta} = (\mu, \tau)$$

- We assume for simplicity that σ^2 is known.

Empirical Bayes: Gaussian-Gaussian Model

- We rewrite the joint distribution exploiting the fact that N_j *Gaussian measurements with values* x_{ij} *and variance* σ^2 *are equivalent to one measurement* $\bar{x}_j = \frac{1}{N_j} \sum_{i=1:N_j} x_{ij}$ *with variance* $\sigma_j^2 = \sigma^2 / N_j$

- This yields the following unnormalized posterior

$$p(\theta, \mathcal{D} | \hat{\eta}, \sigma^2) = \prod_{j=1}^{D} \mathcal{N}\left(\theta_j | \hat{\mu}, \hat{\tau}^2\right) N\left(\bar{x}_j | \theta_j, \sigma_j^2\right)$$

where:

$$\hat{\mu} = \frac{1}{D} \sum_{j=1}^{D} \bar{x}_j \equiv \bar{x}\,, \; \sigma^2 + \hat{\tau}^2 = \frac{1}{D} \sum_{j=1}^{D} \left(\bar{x}_j - \bar{x}\right)^2$$

- From this, closing the square on θ_j, it follows that the posteriors are:

$$p\left(\theta_j | D, \hat{\mu}, \hat{\tau}^2\right) = \mathcal{N}\left(\theta_j | \hat{B}_j \hat{\mu} + \left(1 - \hat{B}_j\right) \bar{x}_j, \left(1 - \hat{B}_j\right) \sigma_j^2\right),$$

$$\hat{\mu} = \frac{1}{D} \sum_{j=1}^{D} \bar{x}_j \equiv \bar{x}\,, \; \sigma^2 + \hat{\tau}^2 = \frac{1}{D} \sum_{j=1}^{D} \left(\bar{x}_j - \bar{x}\right)^2 \quad , \hat{B}_j = \frac{\sigma_j^2}{\sigma_j^2 + \hat{\tau}^2}$$

Empirical Bayes: Gaussian-Gaussian Model

- Note that for constant σ_j^2 we can compute the evidence:

$$\int \mathcal{N}(\theta_j|\mu,\tau^2)\mathcal{N}(\bar{x}_j|\theta_j,\sigma^2)\,d\theta_j = \mathcal{N}(\bar{x}_j|\mu,\sigma^2+\tau^2) \Rightarrow$$

$$p(\mathcal{D}|\mu,\tau^2,\sigma^2) = \prod_{j=1}^{D} \mathcal{N}(\bar{x}_j|\mu,\sigma^2+\tau^2)$$

- We can now derive the previously shown estimates using MLE:

$$\hat{\mu} = \frac{1}{D}\sum_{j=1}^{D} \bar{x}_j \equiv \bar{x}\,,\ \sigma^2 + \hat{\tau}^2 = \frac{1}{D}\sum_{j=1}^{D}(\bar{x}_j - \bar{x})^2$$

- For non-constant σ_j^2, you need to use Expectation-Maximization to derive the empirical Bayes (EB) estimate.

James Stein Estimator

$$p(\theta_j|D,\hat{\mu},\bar{\tau}^2) = N(\theta_j|\hat{B}_j\hat{\mu} + (1-\hat{B}_j)\bar{x}_j, (1-\hat{B}_j)\sigma_j^2),$$

$$\hat{\mu} = \frac{1}{D}\sum_{j=1}^{D}\bar{x}_j \equiv \bar{x}\,,\ \sigma_j^2 + \hat{\tau}^2 = \frac{1}{D}\sum_{j=1}^{D}(\bar{x}_j - \bar{x})^2 \quad ,\ \hat{B}_j = \frac{\sigma_j^2}{\sigma_j^2 + \hat{\tau}^2}$$

- The quantity $0 \le \hat{B}_j \le 1$ controls the degree of *shrinkage towards the overall mean, $\hat{\mu}$*. If the data is reliable for group j, then σ_j^2 will be small relative to $\hat{\tau}^2$; hence $\hat{B}_j$ will be small, and we will put more weight on $\bar{x}_j$ when we estimate θ_j. However, groups with small N_j will get regularized (shrunk towards the overall mean $\hat{\mu}$) more heavily.

- For σ_j constant across j, the posterior mean becomes (James Stein estimator):

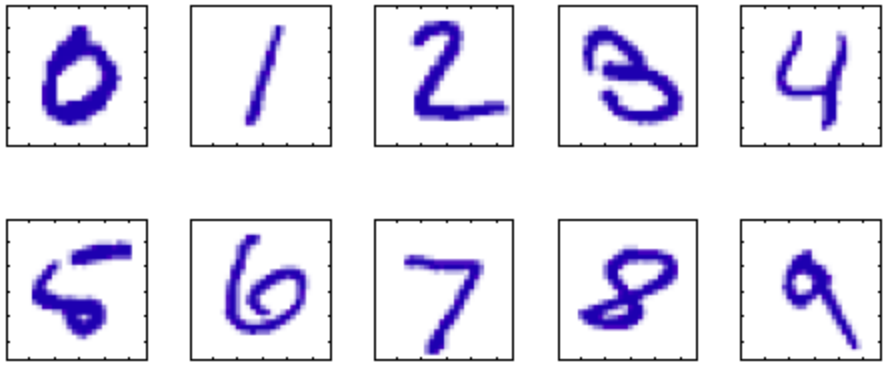
$$\hat{\theta}_j = \hat{B}\bar{x} + (1-\hat{B})\bar{x}_j = \bar{x} + (1-\hat{B})(\bar{x}_j - \bar{x}),\ \hat{B} = \frac{\sigma^2}{\sigma^2 + \hat{\tau}^2}$$

Typical Problems in Machine Learning

- Pattern Recognition: automatically *classifying the data* into different categories and use of these to take actions.
  - Example: handwritten recognition.

    Input: a vector $\boldsymbol{x}$ of pixel values.

    Output: A digit from 0 to 9.



Typical Problems in Machine Learning

- In the digits example, a large set of input vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, or *a training set* is used to tune the parameters of an adaptive model.

- The category of an input vector is expressed using a target vector t (identifying the corresponding digit).

- The result of a machine learning algorithm: $y(\boldsymbol{x})$ where the output y is encoded as the target vectors for *any* input $\boldsymbol{x}$.

Terminology

- Training or learning phase: determine $y(\boldsymbol{x})$ on the basis of the training data.
- Test set, generalization
- Feature extraction
  - Data pre-processing (rotation, scaling, etc.)
  - Using lower dimensional representation of the input and test data
- Supervised learning (input & target vectors in the training data)
  - Classification (discrete categories) or regression (continuous variables)

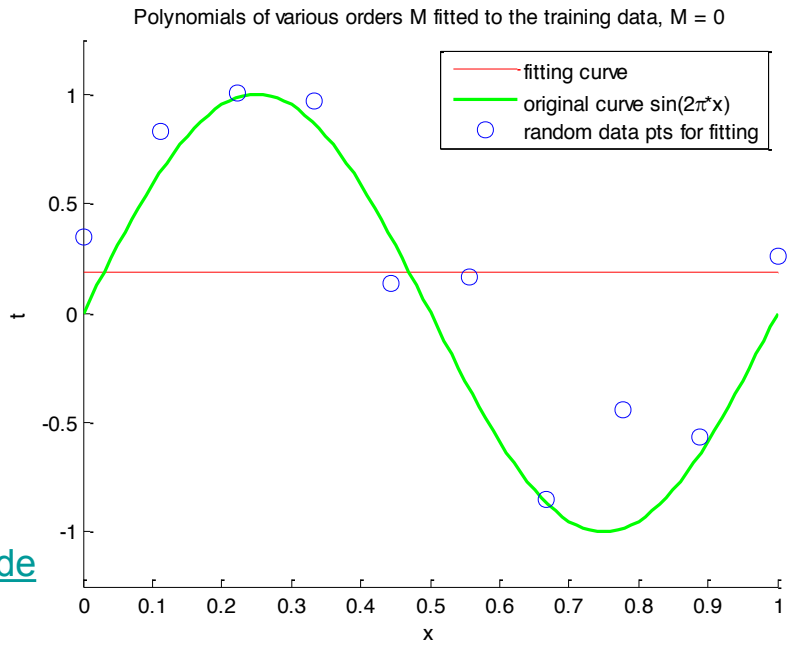
Terminology

- Unsupervised learning (no target vectors in the training data) also called clustering, or density estimation.
- Reinforcement learning (Richard S. Sutton and Andrew G. Barto)
  - credit assignment (rewards attributed to different moves at the end of a game)
  - exploration (of new actions)
  - exploitation (of high reward actions).

Motivation Example: Polynomial Curve Fitting

- Problem definition: implicitly trying to discover the underlying (generating) function $\sin(2\pi x)$ in a set of data.
- Some data points are known: $\boldsymbol{x} = (x_1, \ldots, x_N)^T$ as well as the corresponding target values $\boldsymbol{t} = (t_1, \ldots, t_N)^T$
- We fit the data using a polynomial function of the form

$$y(x, \boldsymbol{w}) = w_0 + w_1 x + \ldots + w_M x^M = \sum_{i=0}^{M} w_i x^i$$



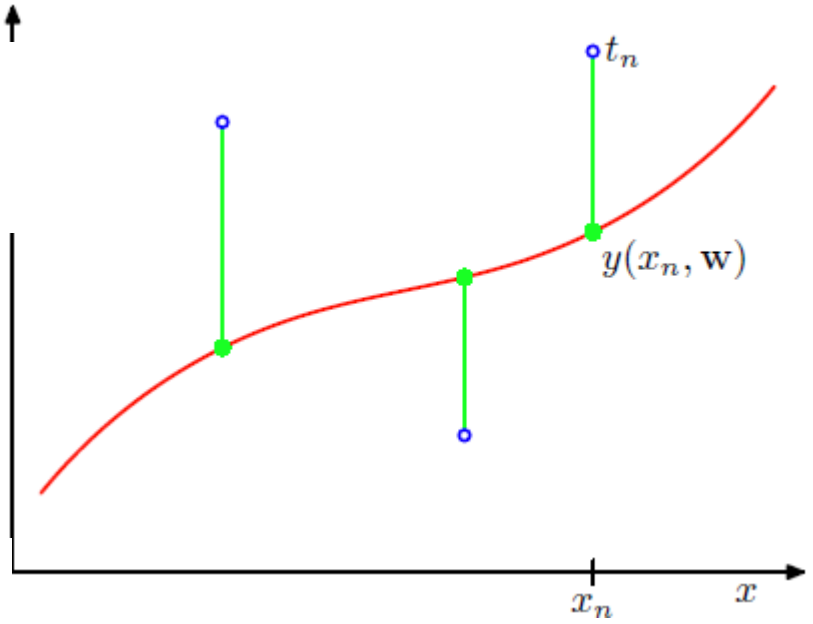
MatLab Code

Motivation Example: Polynomial Curve Fitting

- The values of the coefficients will be determined by fitting the polynomial $y(x, \boldsymbol{w}) = \sum_{j=0}^{M} w_j x^j$ to the training data.

- This can be done by minimizing an error function that measures the misfit between the function $y(x, \boldsymbol{w})$, for any given value of $\boldsymbol{w}$, and the training set data points.

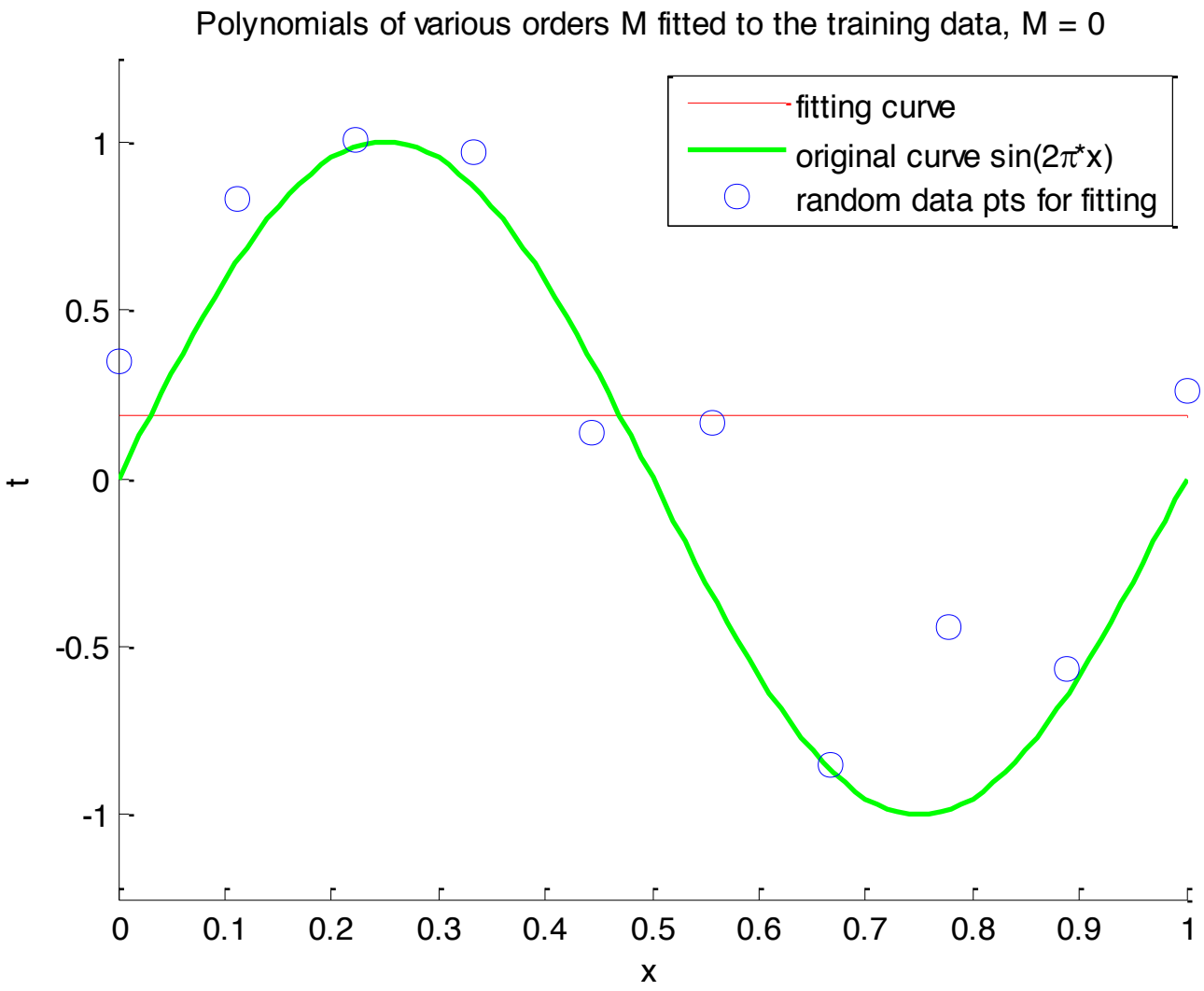
$$\min_{\boldsymbol{w}} E(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} \left(y(x_n, \boldsymbol{w}) - t_n \right)^2$$



- You can show that the minimizer is obtained by solving:

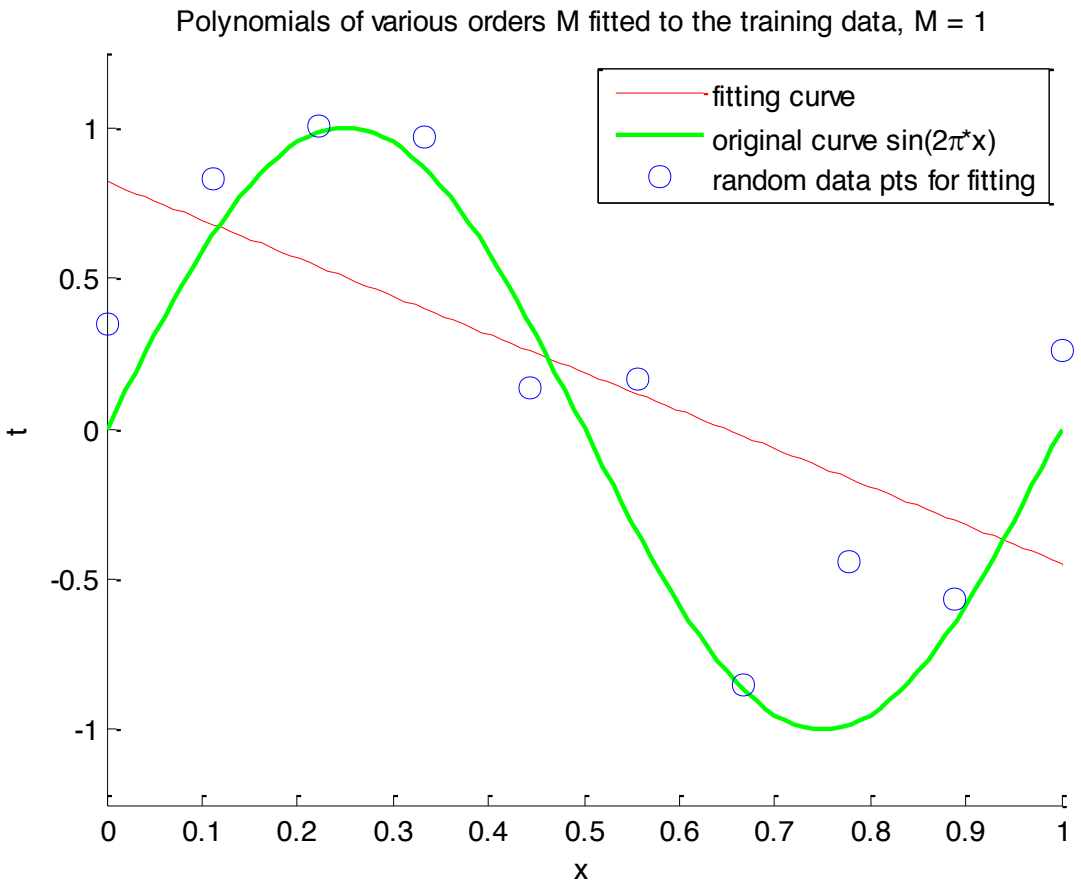
$$\sum_{j=0}^{M} A_{ij} w_j = T_i, \; A_{ij} = \sum_{n=1}^{N} x_n^{i+j}, \; T_i = \sum_{n=1}^{N} x_n^i t_n$$

Polynomial Curve Fitting: Overfitting



MatLab Code

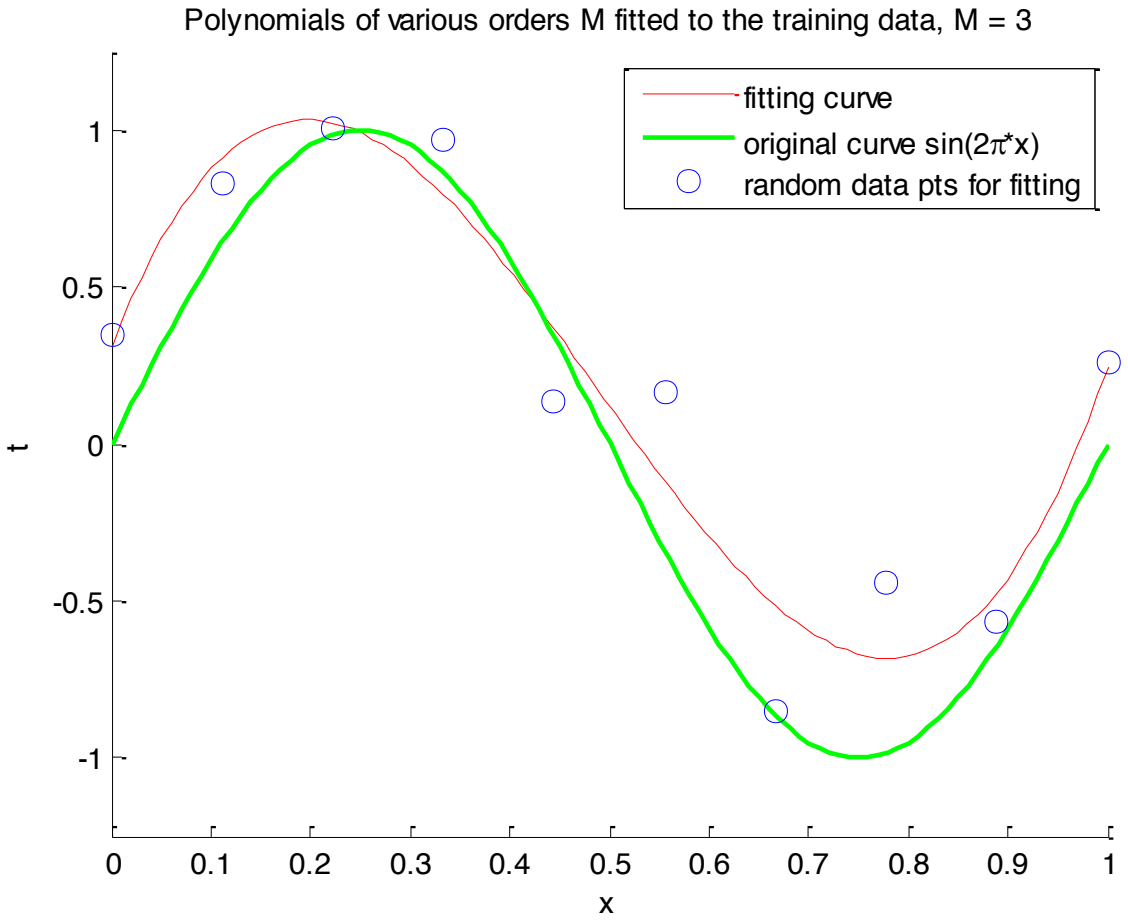
Polynomial Curve Fitting: Overfitting



MatLab Code

- 1st order (linear) polynomials give rather poor fit to the data and the $\sin(2\pi x)$.

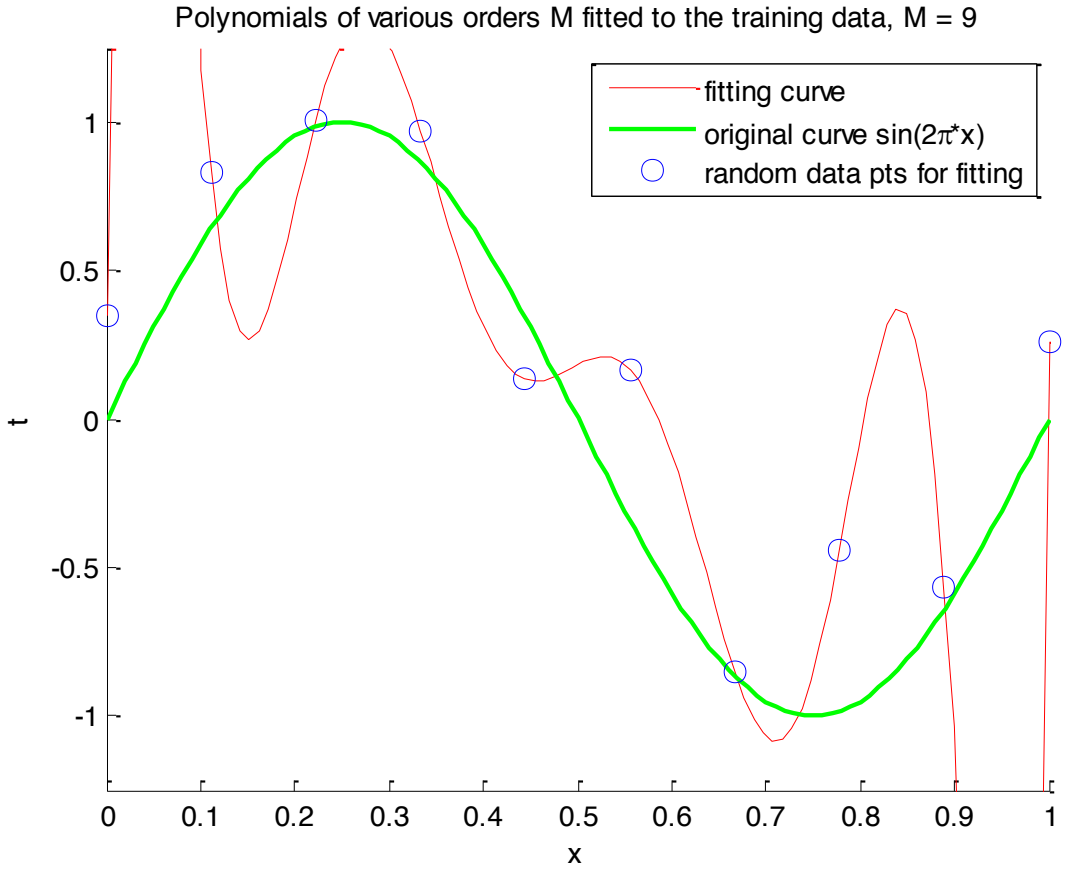
Polynomial Curve Fitting: Overfitting



MatLab Code

- The 3nd order polynomial seems to give the best fit to the function $\sin(2\pi x)$.

Overfitting



MatLab Code

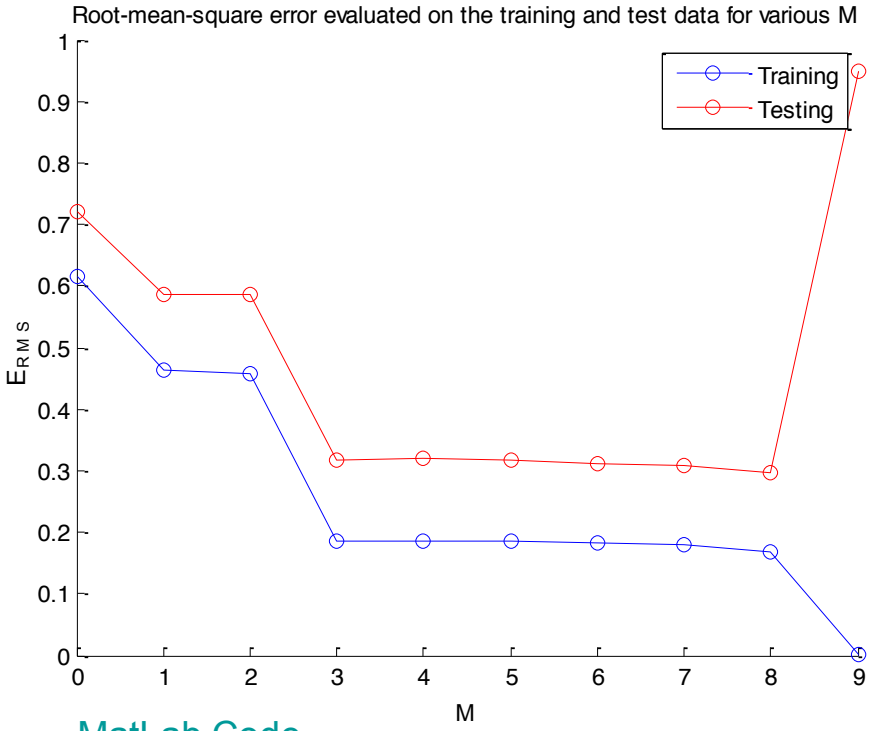
- For $M = 9$ we obtain a perfect fit to the training data. However, the fitted curve oscillates wildly and gives a very poor representation of $\sin(2\pi x)$. This is known as overfitting.

Training and Test Errors

- We often use the root-mean-square (RMS) error. *The division by N allows us to compare different data sizes. The square root makes E_{RMS} in units of t.*

$$E_{RMS} = \left(2E(\boldsymbol{w}^*)/N\right)^{1/2}$$

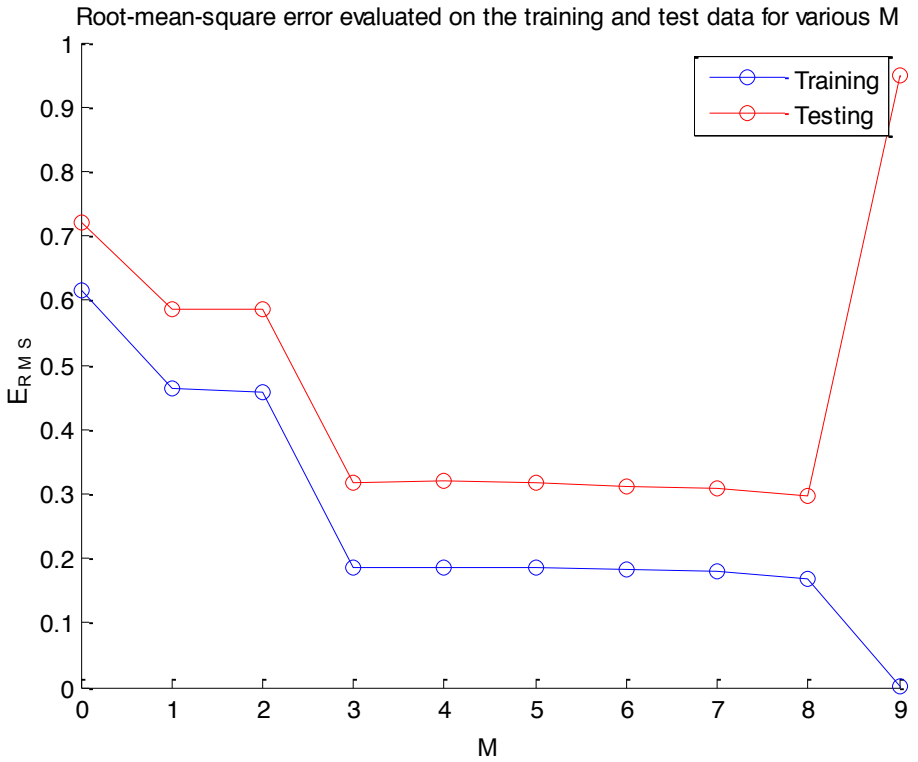
- The ***test set error*** is a measure of how well we are doing in predicting the values of t for new data observations of $\boldsymbol{x}$.



MatLab Code

Training and Test Errors

- Small values of M *give* relatively large values of the test set error. The corresponding polynomials are incapable of capturing the oscillations in $\sin(2\pi x)$.
- $3 < M \leq 8$ give small values for the test set error, and these also give reasonable representation of $\sin(2\pi x)$.



MatLab Code

Oscillation

- Obtain insight into the problem by examining the values of $\boldsymbol{w}$ obtained from polynomials of various order.
- As M increases, the magnitude of the coefficients typically gets larger.

1.0e+006 *

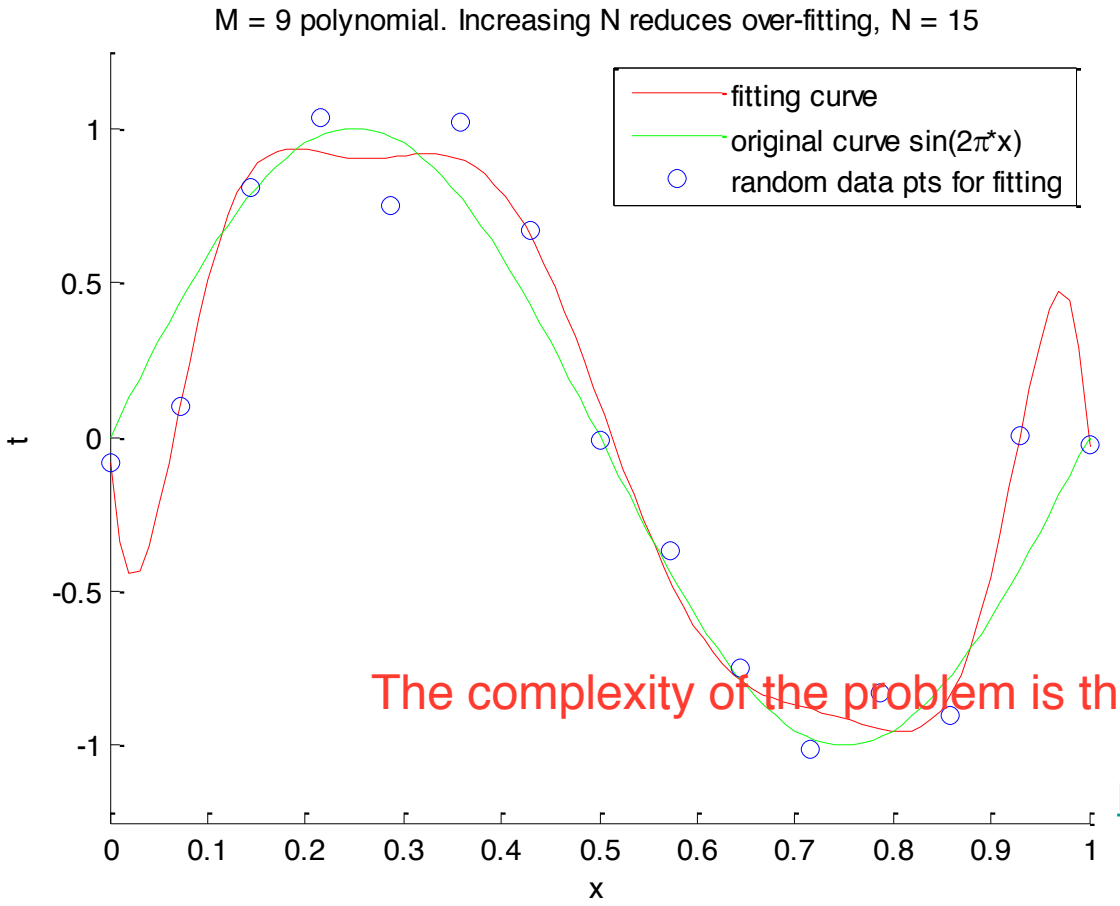
	$M=0$	$M=1$	$M=6$	$M=9$
w_0^*	0.0000	0.0000	0.0000	0.0000
w_1^*	0	-0.0000	0.0000	0.0002
w_2^*	0	0	-0.0000	-0.0053
w_3^*	0	0	0.0000	0.0486
w_4^*	0	0	0	-0.2316
w_5^*	0	0	0	0.6399
w_6^*	0	0	0	-1.0616
w_7^*	0	0	0	1.0422
w_8^*	0	0	0	-0.5576
w_9^*	0	0	0	0.1252

- For $M = 9$, the coefficients have become finely tuned to the data by developing large positive and negative values.
- The more flexible polynomials with larger values of M are becoming increasingly tuned to the random noise on the target values.

MatLab Code

Varying the Data Size

- It is also interesting to examine the behavior of a given model as the size of the data set is varied.

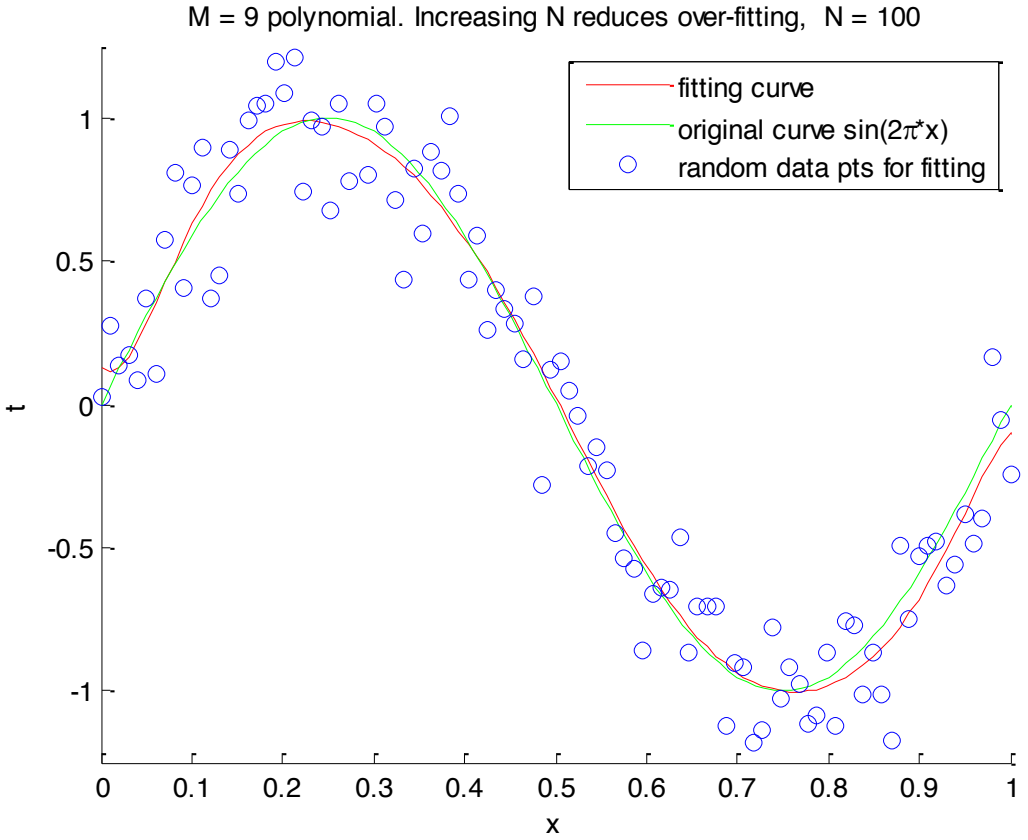


The complexity of the problem is the data size.

MatLab Code

Varying the Data Size

- For a given model complexity, the over-fitting problem becomes less severe as the size of the data set increases, i.e. the larger the data set, the more complex (in other words more flexible) the model that we can afford to fit to the data.



MatLab Code

Overfitting and Maximum Likelihood

- It is not reasonable having the number of parameters in a model (model complexity) limited according to the size of the available training set.

- It makes more sense to choose the complexity of the model according to the complexity of the problem being solved.

- The least squares approach to finding the model parameters is a specific case of maximum likelihood.

- The over-fitting problem is a general property of maximum likelihood.

Overfitting and Bayesian Approach

- By adopting a Bayesian approach, the over-fitting problem can be avoided.

- There is *no difficulty in employing models for which the number of parameters greatly exceeds the number of data points.*

- ***In a Bayesian model, "the effective number" of parameters adapts automatically to the size of the data set.***

Regularization Technique

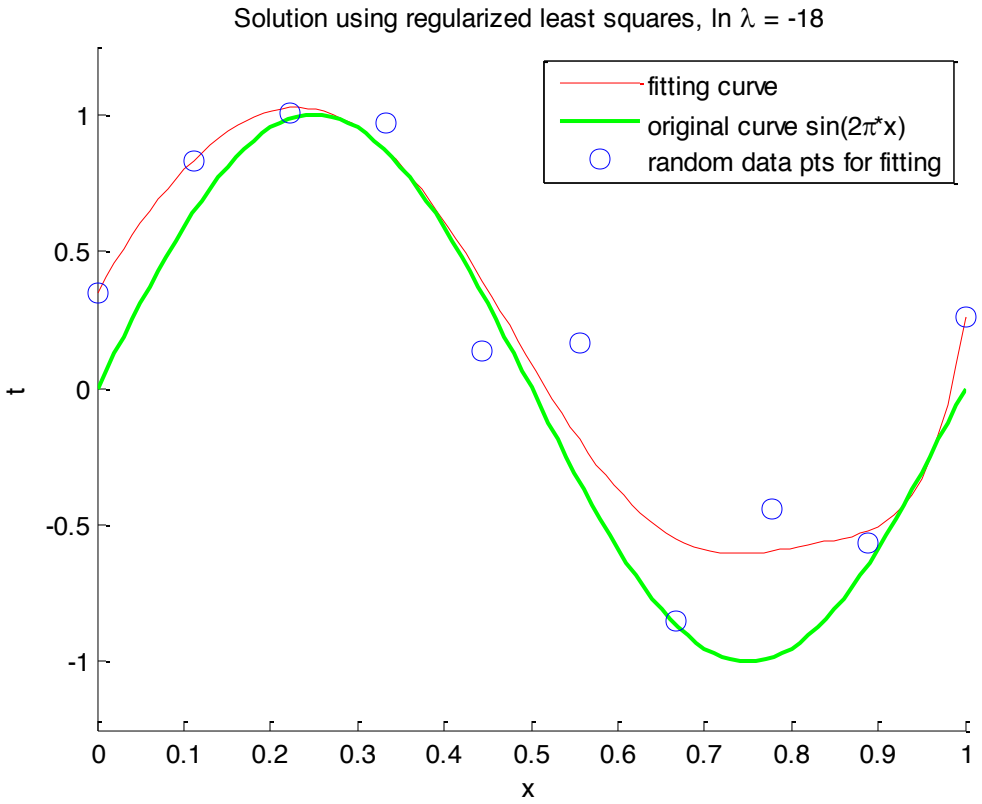
- To control the over-fitting we use *regularization, e.g.* adding "a penalty term" to the error function in order to discourage the coefficients from reaching large values.
- The simplest such penalty term takes the form of a sum of squares of all of the coefficients, leading to a modified error function of the form

$$\overline{E}(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(y(x_n, \boldsymbol{w}) - t_n\right)^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|^2$$

- The regularization parameter λ governs the relative importance of the regularization term compared with the sum-of-squares error term.
- The minimizer is similar to that given earlier but with

$$\sum_{j=0}^{M} A_{ij} w_j = T_i,\ A_{ij} = \sum_{n=1}^{N} x_n^{i+j} + \lambda\delta_{ij}, T_i = \sum_{n=1}^{N} x_n^i t_n$$

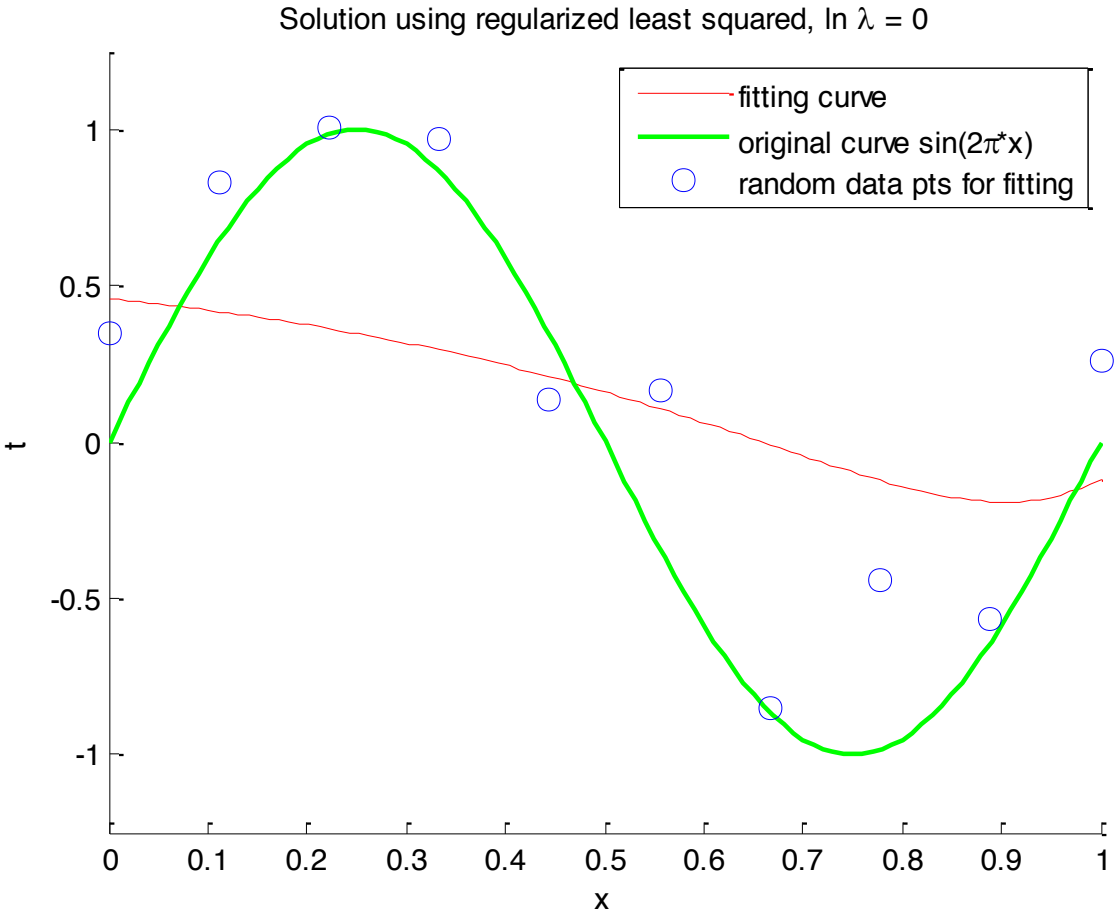
Regularization Controls Model Complexity



MatLab Code

- We now fit the polynomial of order $M = 9$ to the same data set as before but now using the regularized error function.
- For $lnl = -18$, the over-fitting has been suppressed and we obtain a much closer representation of $\sin(2\pi x)$.

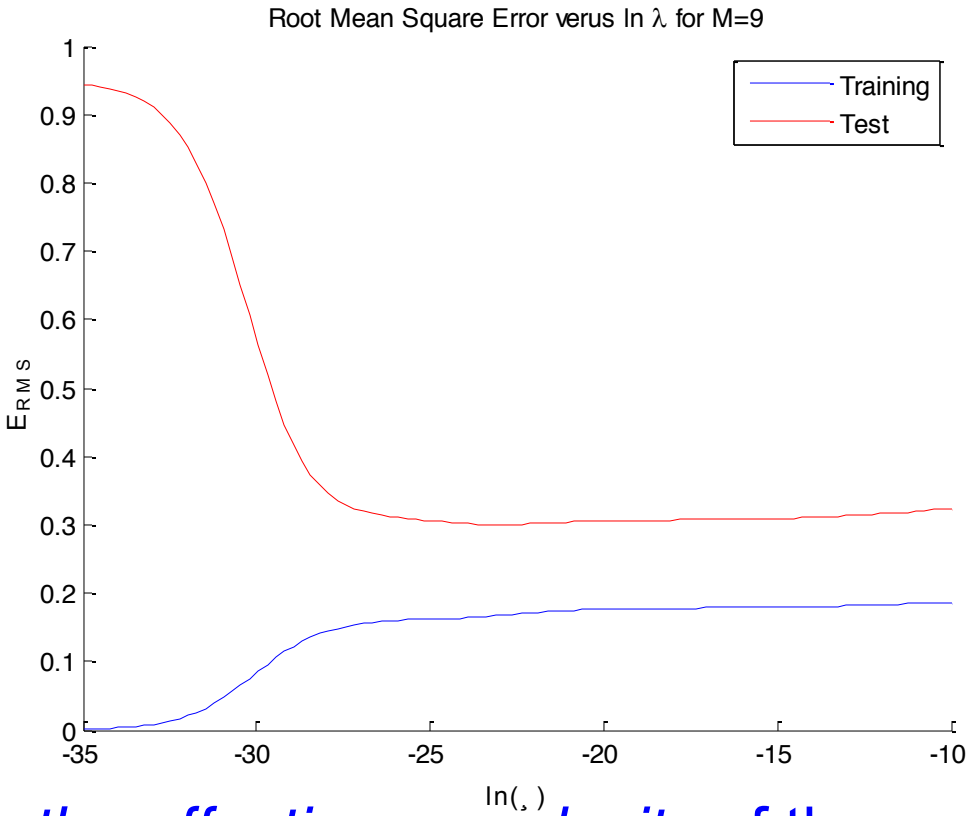
Regularization Controls Model Complexity



MatLab Code

- If, however, we use too large a value for λ then we again obtain a poor fit, as shown for $\ln\lambda = 0$.

Regularization Controls Model Complexity



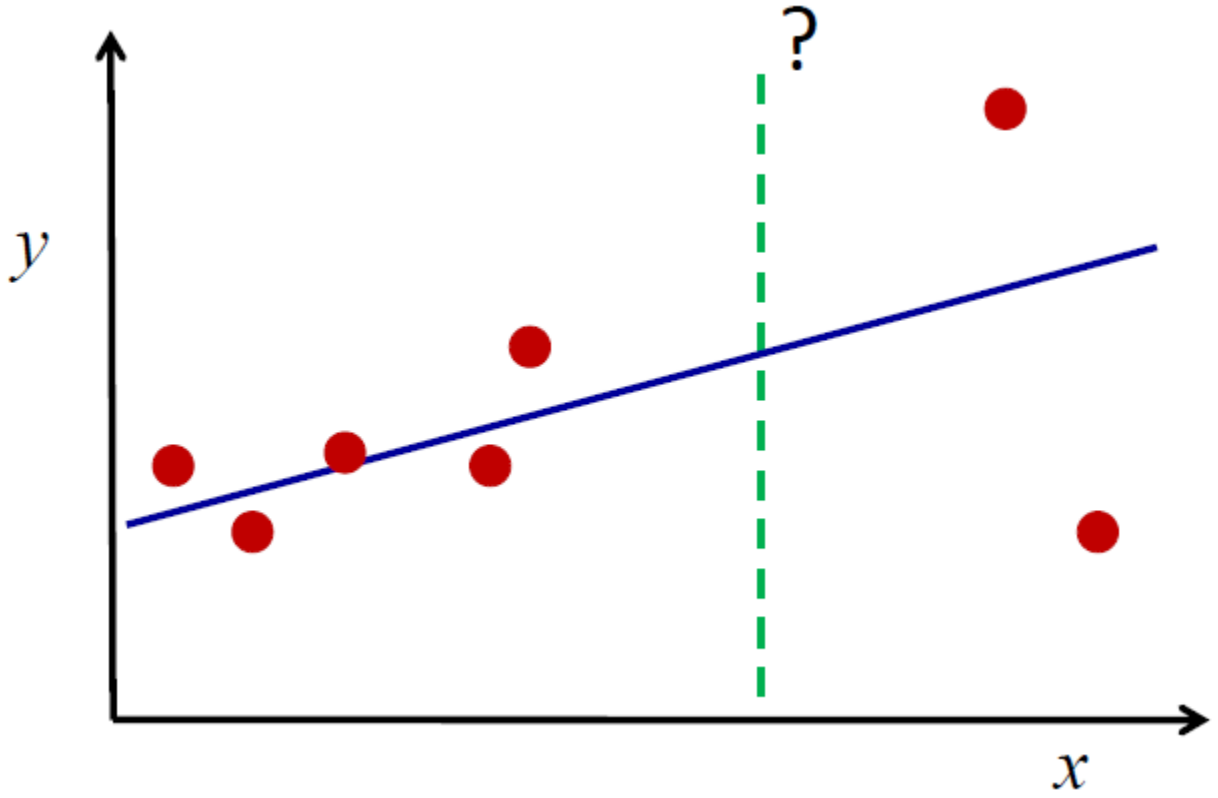
MatLab Code

- λ *controls the effective complexity* of the model and hence determines the degree of over-fitting.

- We will soon re-examine this problem with a Bayesian approach that avoids the over-fitting problem.

Prior Knowledge is Essential

- We cannot do everything simply based on data – prior knowledge is essential to inference and prediction.



Bayesian Probabilities

- For example in the regression problem with the observed data $\mathcal{D} = \{t_1, \ldots, t_N\}$, we can obtain the conditional probability $p(\boldsymbol{w}|\mathcal{D})$ by Bayes' theorem

$$p(\boldsymbol{w} \mid \boldsymbol{\mathcal{D}}) = \frac{p(\boldsymbol{\mathcal{D}} \mid \boldsymbol{w})\, p(\boldsymbol{w})}{p(\boldsymbol{\mathcal{D}})}, \quad p(\boldsymbol{\mathcal{D}}) = \int p(\boldsymbol{\mathcal{D}} \mid \boldsymbol{w})\, p(\boldsymbol{w}) d\boldsymbol{w}$$

- The quantity $p(\mathcal{D}|\boldsymbol{w})$ on the right-hand side of Bayes' theorem is evaluated for the observed data set $\mathcal{D}$ and can be viewed as a function of the parameter vector $\boldsymbol{w}$ (*likelihood function*).

- Given this definition of likelihood, we can state Bayes' theorem in words

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

Frequentist Versus Bayesian Paradigms

- The likelihood $p(\mathcal{D}|\boldsymbol{w})$ is essential in both Bayesian and frequentist approaches but it is used in different roles.
- In a frequentist approach
  - $\boldsymbol{w}$ is a fixed parameter computed by an estimator (e.g. maximum likelihood estimator).
  - Error bars on this point estimate are computed by considering the distribution of all possible data sets $\mathcal{D}$ (e.g. variability of predictions between different bootstrap data sets)
- In Bayesian approach
  - there is only one set of data $\mathcal{D}$ and
  - the uncertainty in $\boldsymbol{w}$ is introduced with appropriate prior and computing posterior probabilities over $\boldsymbol{w}$.

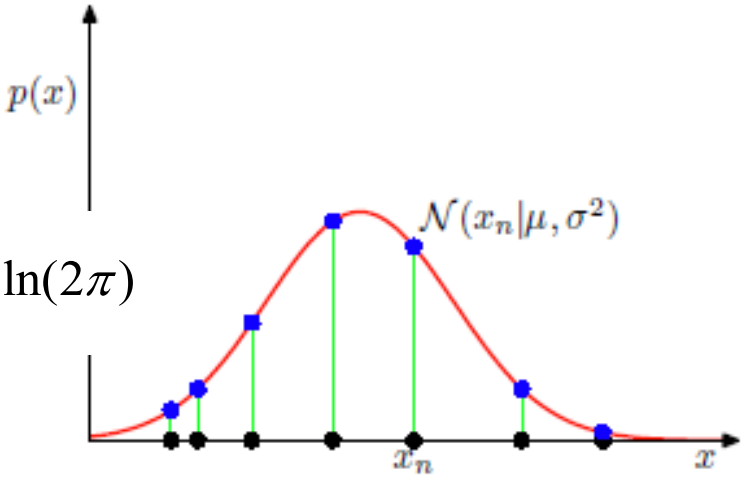
Gaussian Distribution

- Consider the Gaussian distribution

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\left(2\pi\sigma^2\right)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

- The likelihood function for the Gaussian distribution is

$$p(\mathcal{D} \mid \mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n \mid \mu, \sigma^2)$$



- The log likelihood takes the form*

$$\ln p(\mathcal{D} \mid \mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

- Maximum likelihood solution

$$\mu_{ML} = \frac{\sum_{n=1}^{N} x_n}{N}, \quad \sigma^2_{ML} = \frac{\sum_{n=1}^{N}(x_n - \mu_{ML})^2}{N}$$

* We work often with log-likelihood to avoid underflow (taking products of small probabilities) and for simplifying the algebra.

MLE for a Gaussian Distribution

$$\mu_{ML} = \underbrace{\frac{1}{N}\sum_{i=1}^{N} x_i}_{\text{Sample mean}}, \quad \sigma^2_{ML} = \underbrace{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{ML})^2}_{\text{Sample variance wrt ML mean (not the exact mean)}}$$

- The MLE approach underestimates the variance (bias) – this is in the root of the over-fitting problem e.g. in polynomial curve fitting.
- The maximum likelihood solutions μ_{ML}, σ^2_{ML} are functions of the data set values $x_1, \ldots, x_N$. Consider the expectations of these quantities with respect to the data set values, which come from a Gaussian.
- Using the point estimates above you can show that :

$$\mathbb{E}\left[\mu_{ML}\right] = \mu, \quad \mathbb{E}\left[\sigma^2_{ML}\right] = \frac{N-1}{N}\sigma^2$$

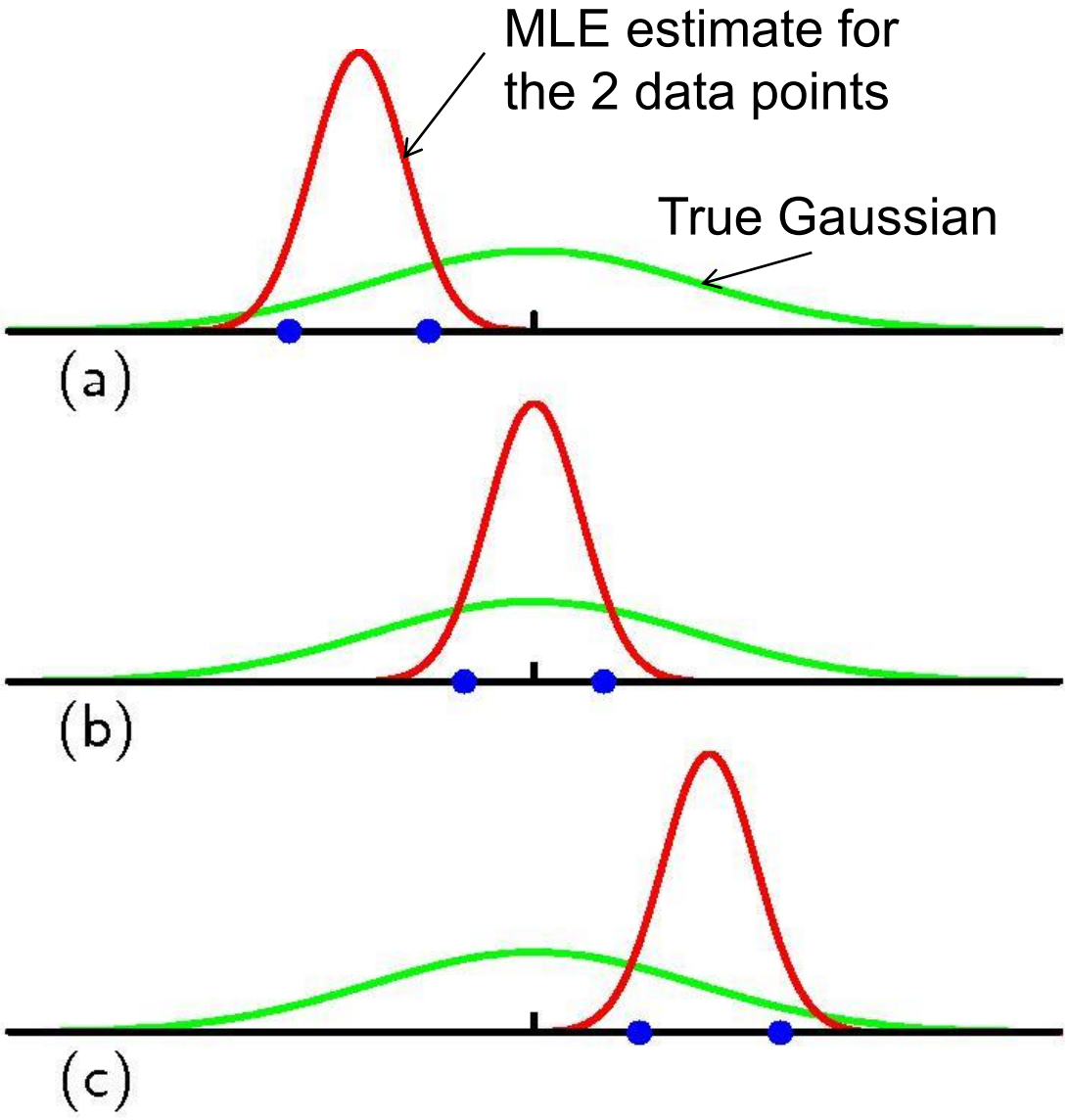
In this derivation you need to use :

$$\mathbb{E}\left[x_i x_j\right] = \mu^2 \quad \text{for } i \neq j$$

$$\mathbb{E}\left[x_i^2\right] = \sigma^2 + \mu^2$$

MLE: Underestimating the Variance

- In the schematic from Bishop's PRML, we consider 3 cases each with 2 data points extracted from the true Gaussian.
- The mean of the 3 distributions predicted via MLE (i.e. averaged over the data) is correct.
- However, ***the variance is underestimated*** since it is a variance with respect to the sample mean and NOT the true mean.

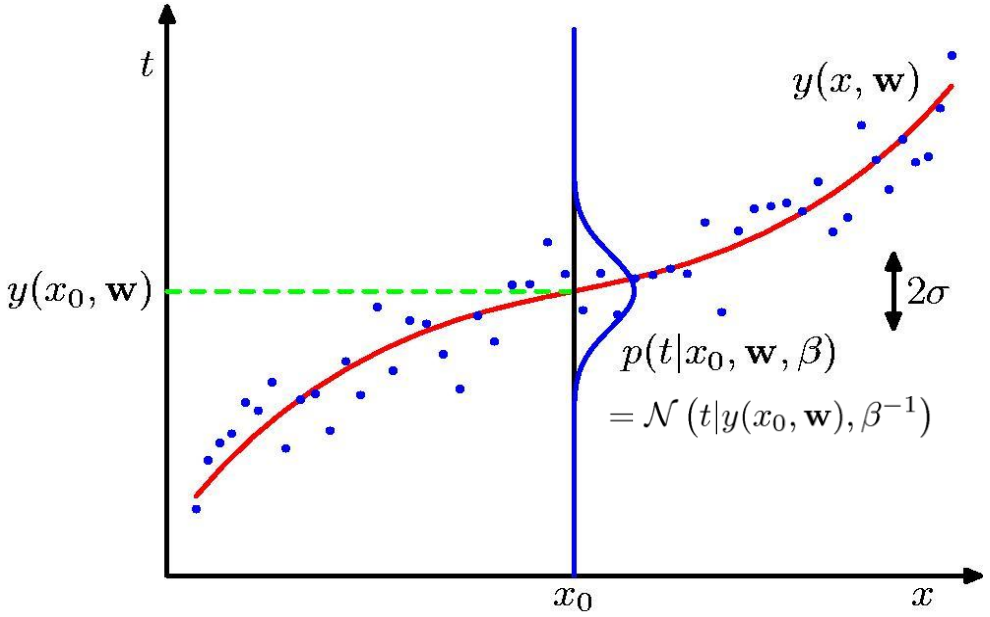


Curve Fitting Revisited

- Consider a set of training data comprising N input $\boldsymbol{x} = (x_1, \ldots, x_N)^T$ & the corresponding target values $\boldsymbol{t} = (t_1, \ldots, t_N)^T$

- We assume that, given the value of x, the corresponding value of t has a Gaussian distribution with a mean equal to the value $y(x, \boldsymbol{w})$ and precision (inverse of the variance) β

$$p(t \mid x, \boldsymbol{w}, \beta) = \mathcal{N}(t \mid y(x, \boldsymbol{w}), \beta^{-1})$$



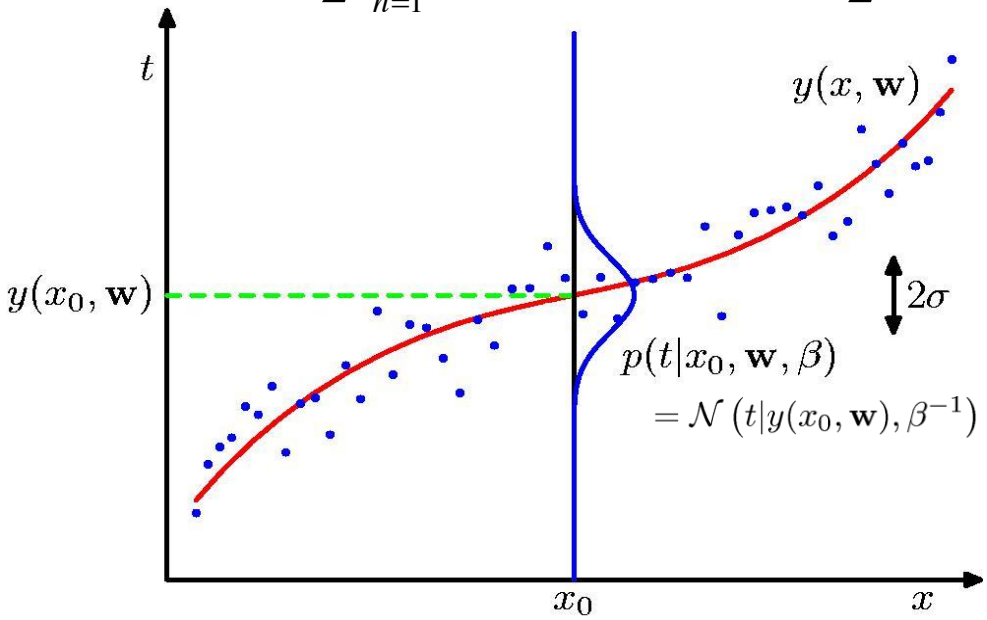
Curve Fitting Revisited

- The likelihood function is

$$p(\boldsymbol{t} \mid \boldsymbol{x}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid y(x_n, \boldsymbol{w}), \beta^{-1})$$

- From this, the log-likelihood takes the form:

$$\ln p(\boldsymbol{t} \mid \boldsymbol{x}, \boldsymbol{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)$$



A Frequentist Approach

- Consider first the MLE estimate for $\boldsymbol{w}$. Note that maximizing the log-likelihood to obtain $\boldsymbol{w}_{ML}$ is the same as minimizing the sum of squares error function (residual sum of squares, $RSS(\boldsymbol{w})$):

$$\max_{\boldsymbol{w}} \ln p(\boldsymbol{t} \mid \boldsymbol{x}, \boldsymbol{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\left\{y(x_n, \boldsymbol{w}) - t_n\right\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) \Leftrightarrow$$

$$\min_{\boldsymbol{w}} \frac{1}{2}\sum_{n=1}^{N}\left\{y(x_n, \boldsymbol{w}) - t_n\right\}^2$$

- We can also determine the MLE estimate of β:

$$\frac{1}{\beta_{ML}} = \frac{1}{N}\sum_{n=1}^{N}\left\{y(x_n, \boldsymbol{w}_{ML}) - t_n\right\}^2$$

- We can now make (a frequentist, plug-in approximation) prediction as follows: $p(t \mid x, \boldsymbol{w}_{ML}, \beta_{ML}) = \mathcal{N}(t \mid y(x, \boldsymbol{w}_{ML}), \beta_{ML}^{-1})$

Bayesian Curve Fitting Approach

- Now let us take a step towards a Bayesian approach and introduce a prior distribution over the coefficients $\boldsymbol{w}$. Consider a Gaussian distribution

$$p(\boldsymbol{w}\,|\,\alpha) = \mathcal{N}(\boldsymbol{w}\,|\,0, \alpha^{-1} I) = \left(\frac{\alpha}{2\pi}\right)^{M/2} \exp\left\{-\frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}\right\}$$

where α is the precision of the distribution, and M is the total number of elements in the vector $\boldsymbol{w}$ for an $(M-1)^{th}$ order polynomial.*

- Using Bayes' theorem:

$$p(\boldsymbol{w}\,|\,\boldsymbol{x},\boldsymbol{t},\alpha,\beta) \propto p(\boldsymbol{t}\,|\,\boldsymbol{x},\boldsymbol{w},\beta)\,p(\boldsymbol{w}\,|\,\alpha)$$

* One should not penalize the bias term w_0 as it does not contribute to overfitting.

MAP Estimate

- We can determine $\boldsymbol{w}$ (point estimate) by finding the most probable value of $\boldsymbol{w}$ given the data, i.e. maximizing the posterior.

- This technique is called *maximum posterior (MAP).*

- The maximum of the posterior is given by the minimum of

$$\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\boldsymbol{w})-t_n\}^2+\frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}$$

- Note that the MAP point estimate is equivalent to regularized sum of squares error function with regularization parameter

$$\lambda=\frac{\alpha}{\beta}=\frac{precision\ of\ \ prior}{precision\ in\ the\ data}$$

Posterior Distribution

$$p(\boldsymbol{w}\,|\,\alpha)=\left(\frac{\alpha}{2\pi}\right)^{M/2}\exp\left\{-\frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}\right\},\; p(\boldsymbol{t}\,|\,\boldsymbol{x},\boldsymbol{w},\beta)\propto\exp\left\{-\frac{\beta}{2}\sum_{n=1}^{N}\left(\boldsymbol{\phi}(x_n)^T\boldsymbol{w}-t_n\right)^2\right\}\Rightarrow$$

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

Quadratic in $\boldsymbol{w}$

- We now have the product of two Gaussians and the posterior is easily computed as:

Completing the square $-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu})^T\boldsymbol{S}_N^{-1}(\boldsymbol{w}-\boldsymbol{\mu})$

$$p(\boldsymbol{w}\,|\,\boldsymbol{x},\boldsymbol{t},\alpha,\beta)\propto$$

$$\exp\left(-\frac{1}{2}\boldsymbol{w}^T\alpha\boldsymbol{I}_{M\times M}\boldsymbol{w}-\frac{1}{2}\boldsymbol{w}^T\sum_{n=1}^{N}\beta\boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^T\boldsymbol{w}+\beta\boldsymbol{w}^T\sum_{n=1}^{N}t_n\boldsymbol{\phi}(x_n)\right)$$

$$\propto\mathcal{N}\left(\boldsymbol{w}\,|\,\beta\boldsymbol{S}_N\sum_{n=1}^{N}t_n\boldsymbol{\phi}(x_n),\boldsymbol{S}_N\right),\;\boldsymbol{S}_N^{-1}=\alpha\boldsymbol{I}+\sum_{n=1}^{N}\beta\boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^T$$

Notation: $\boldsymbol{\phi}(x_n)=\begin{Bmatrix}1\\x_n\\x_n^2\\:\\x_n^{M-1}\end{Bmatrix},\;\boldsymbol{\phi}(x)^T=\left\{1\quad x\quad x^2\quad ..\quad x^{M-1}\right\},\;\boldsymbol{I}=unit\ matrix\ M\times M$

Polynomial order: $M-1$

Predictive Distribution

- In a full Bayesian treatment, we want to compute the predictive distribution, i.e. given the training data $\boldsymbol{x}$ and $\boldsymbol{t}$ and a new test point x, we want the distribution:

$$p(t \mid x, \boldsymbol{x}, \boldsymbol{t}) = \int p(t \mid x, \boldsymbol{w})\, p(\boldsymbol{w} \mid \boldsymbol{x}, \boldsymbol{t}) d\boldsymbol{w}, \quad where$$

$$p(t \mid x, \boldsymbol{w}, \beta) = \mathcal{N}(t \mid y(x, \boldsymbol{w}), \beta^{-1}) = \mathcal{N}(t \mid \boldsymbol{\phi}(x)^T \boldsymbol{w}, \beta^{-1}) \; and$$

$$p(\boldsymbol{w} \mid \boldsymbol{x}, \boldsymbol{t}, \alpha, \beta) \propto \mathcal{N}\left(\boldsymbol{w} \mid \beta \boldsymbol{S}_N \sum_{n=1}^{N} t_n \boldsymbol{\phi}(x_n), \boldsymbol{S}_N \right), \; \boldsymbol{S}_N^{-1} = \alpha \boldsymbol{I} + \sum_{n=1}^{N} \beta \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x_n)^T$$

Gaussian Distribution

- To integrate $\boldsymbol{w}$ out as shown on the predictive distribution expression, we use a fundamental result for linear Gaussian models (see relevant lecture for proof).

Appendix: Linear Gaussian Models

$$p(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}, \Lambda^{-1}\right)$$

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{y} \mid A\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{L}^{-1}\right)$$

- For the above linear model, we have shown in an earlier lecture, the following very useful results about marginal and conditional Gaussian models.

$$p(\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{y} \mid A\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{L}^{-1} + A\Lambda^{-1}A^T\right)$$

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{x} \mid \left(\Lambda + A^T\boldsymbol{L}A\right)^{-1}\left(\Lambda\boldsymbol{\mu} + A^T\boldsymbol{L}(\boldsymbol{y} - \boldsymbol{b})\right), \left(\Lambda + A^T\boldsymbol{L}A\right)^{-1}\right)$$

Predictive Distribution

$$p(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}, \Lambda^{-1}\right)$$

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{L}^{-1}\right)$$

$$p(\boldsymbol{w} \mid \boldsymbol{x}, \boldsymbol{t}, \alpha, \beta) = \mathcal{N}\left(\boldsymbol{w} \mid \beta \boldsymbol{S}_N \sum_{n=1}^{N} t_n \boldsymbol{\phi}(x_n), \boldsymbol{S}_N\right)$$

$$p(t \mid x, \boldsymbol{w}, \beta) = \mathcal{N}(t \mid y(x, \boldsymbol{w}), \beta^{-1})$$

- Thus for our problem:

$$\boldsymbol{x} \leftarrow \boldsymbol{w}, \boldsymbol{\mu} = \beta \boldsymbol{S}_N \sum_{n=1}^{N} t_n \boldsymbol{\phi}(x_n), \Lambda^{-1} = \boldsymbol{S}_N$$

$$\boldsymbol{y} \leftarrow t, \boldsymbol{A} = \boldsymbol{\phi}(x)^T, \boldsymbol{b} = 0, \boldsymbol{L}^{-1} = \beta^{-1}$$

- The predictive distribution now takes the form:

$$p(\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{L}^{-1} + \boldsymbol{A}\Lambda^{-1}\boldsymbol{A}^T\right) \Rightarrow$$

$$p(t) = \mathcal{N}\left(t \mid \boldsymbol{\phi}(x)^T \beta \boldsymbol{S}_N \sum_{n=1}^{N} t_n \boldsymbol{\phi}(x_n), \beta^{-1} + \boldsymbol{\phi}(x)^T \boldsymbol{S}_N \boldsymbol{\phi}(x)\right)$$

Predictive Distribution

- In a full Bayesian treatment, we want to compute the predictive distribution, i.e. given the training data $\boldsymbol{x}$ and $\boldsymbol{t}$ and a new test point x, we want the distribution:

$$p(t \mid x, \boldsymbol{x}, \boldsymbol{t}) = \int p(t \mid x, \boldsymbol{w})\, p(\boldsymbol{w} \mid \boldsymbol{x}, \boldsymbol{t})\, d\boldsymbol{w}, \quad p(t \mid x, \boldsymbol{w}, \beta) = \mathcal{N}(t \mid y(x, \boldsymbol{w}), \beta^{-1}) \Rightarrow$$

$$p(t \mid x, \boldsymbol{x}, \boldsymbol{t}) = \mathcal{N}(t \mid m(x), s^2(x))$$

where the mean and variance were shown in the earlier slide to be:

$$m(x) = \beta \boldsymbol{\phi}(x)^T \boldsymbol{S}_N \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) t_n$$

$$s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^T \boldsymbol{S}_N \boldsymbol{\phi}(x)$$

$$\boldsymbol{S}_N^{-1} = \alpha \boldsymbol{I} + \beta \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x)^T$$

Bayesian Curve Fitting

- The notation used here is as follows:

$$p(t \mid x, \boldsymbol{x}, \boldsymbol{t}) = \mathcal{N}(t \mid m(x), s^2(x))$$

$$m(x) = \beta \boldsymbol{\phi}(x)^T \boldsymbol{S}_N \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) t_n$$

$$s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^T \boldsymbol{S}_N \boldsymbol{\phi}(x)$$

$$\boldsymbol{S}_N^{-1} = \alpha \boldsymbol{I} + \beta \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x_n)^T$$

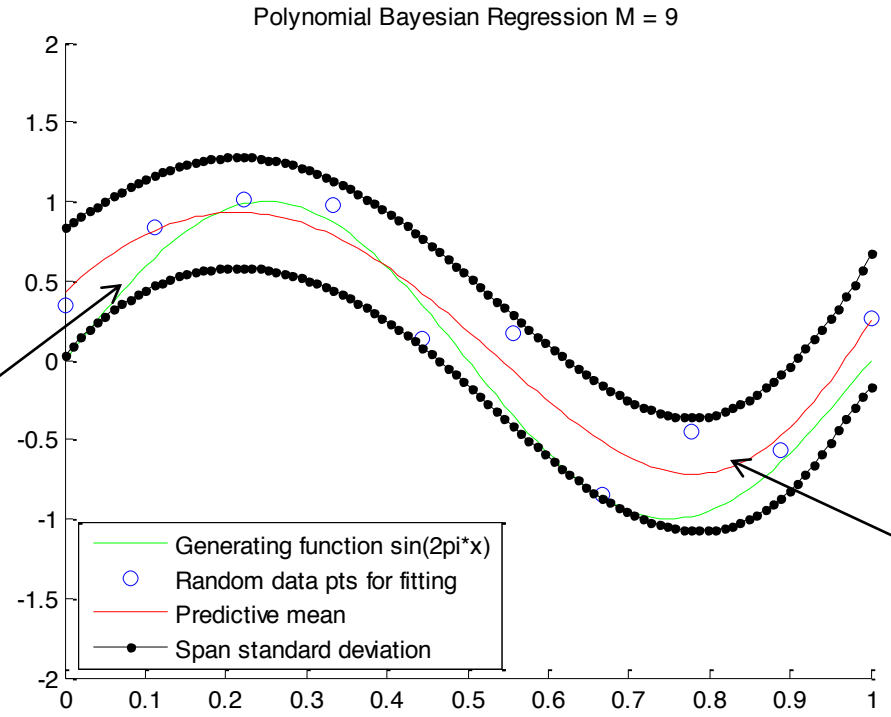
Predictive mean and variance are functions of x.

$$\boldsymbol{\phi}(x_n) = \begin{Bmatrix} 1 \\ x_n \\ x_n^2 \\ \vdots \\ x_n^{M-1} \end{Bmatrix}, \boldsymbol{\phi}(x)^T = \begin{Bmatrix} 1 & x & x^2 & .. & x^{M-1} \end{Bmatrix}, \boldsymbol{I} = unit\ matrix\ M \times M$$

Predictive Distribution

- The predictive distribution using an $M = 9$ polynomial, with fixed parameters $\alpha = 5 \times 10^{-3}$ *and* $\beta = 11.1$ (corresponding to a known noise variance).

- The red curve denotes the mean of the predictive distribution and the red region corresponds to $\pm$ standard deviation around the mean.



Generating function

MatLab Code

Mean of the predictive distribution

Model Selection

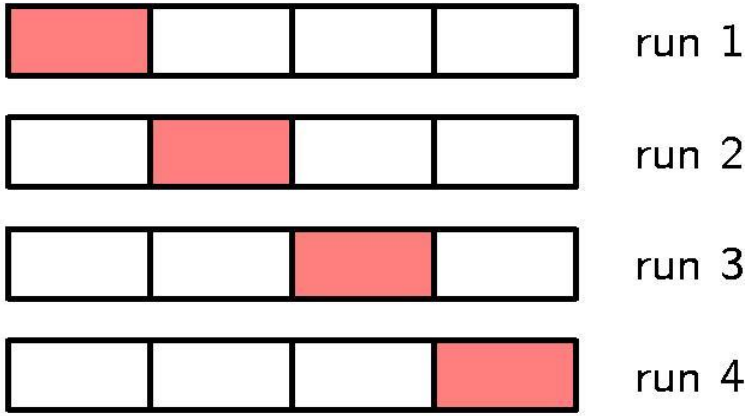
- A number of complexity parameters (polynomial order, regularization parameter, etc.) need to be selected to optimize performance/predictive capability. This is a model selection problem.

- In MLE, *the performance on the training set is not a good indicator of predictive performance due to the problem of over-fitting.*

- We often use some of the available data to train a range of models (or a given model with a range of values for its complexity parameters) and then to compare them on a validation set. We then select the one having the best predictive performance.

- Some over-fitting to the validation data can occur and a third test set on which the performance of the selected model is finally evaluated maybe needed.

Model Selection: Cross Validation



- The technique of S-fold cross-validation (here $S = 4$) involves taking the available data and partitioning it into S groups.

- $S - 1$ of the groups are used to train a set of models that are then evaluated on the remaining group. This procedure is repeated for all S possible choices for the held-out group and the performance scores from the S runs are then averaged.

Akaike Information Criterion

- The cross-validation cost increases by a factor of S.
- We should allow multiple hyperparameters and model types to be compared in a single training run.
- To correct for the bias of MLE, we use different information criteria (here $M = \#$ of parameters in the model), e.g.:

  Akaike Information Criterion (AIC): $\ln p(\mathcal{D} \mid w_{ML}) - M$

  We choose the model for which the AIC is largest.

- *AIC* does not account for uncertainty in model parameters. *It favor simple models.* We will visit Bayesian model selection criteria in follow up lectures.

Bayesian Model Selection

- In general, when faced with a set of models (i.e., families of parametric distributions) of different complexity, how should we choose the best one? This is called the model selection problem.

- Examples:
  - *a low order polynomial in linear regression underfitts while a high order polynomial overfitts*
  - *a small regularization parameter λ results in overfitting and too large λ in underfitting*.

- Can use CV to estimate the generalization error of all the candidate models, and then to pick the model that performs the best. This requires fitting each model K times, where K is the number of CV folds. More efficient approach is to *compute the posterior over models.*

$$p(m \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid m)\, p(m)}{\sum_{m' \in M} p(m', \mathcal{D})}$$

- From this, we can easily compute the MAP model $\overline{m} = \max_m p(m \mid \mathcal{D})$

Model Evidence

- If we use a uniform prior over models, $p(m) \sim 1$, this amounts to picking the model which maximizes the marginal likelihood:

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

- This quantity is called the evidence for model m.
- The details on how to perform this integral will be discussed with examples later on.
- An intuitive interpretation of model evidence is discussed next.

Bayesian Occam's Razor

- One might think that using $p(\mathcal{D}|m)$ to select models would always favor the model with the most parameters.
- This is true if we use $p(\mathcal{D}|\hat{\theta}_m)$ to select models, where $\hat{\theta}_m$ is the MLE or MAP estimate of the parameters for model m - *models with more parameters will fit the data better, and hence achieve higher likelihood.*
- However, *if we integrate out the parameters, rather than maximizing them, we are automatically protected from overfitting*.
- Models with more parameters do not necessarily have higher marginal likelihood.
- This is called the Bayesian Occam's razor effect (MacKay 1995b; Murray and Ghahramani 2005)
- *Occams Razor Principle: one should pick the simplest model that adequately explains the data.*

Bayesian Occam's Razor

- The marginal likelihood can be rewritten as follows:

$$p(\mathcal{D}) = p(y_1)p(y_2 \mid y_1)p(y_3 \mid y_{1:2})\ldots p(y_N \mid y_{1:N-1})$$

where we have dropped the conditioning on m for brevity.

- This is similar to a leave-one-out cross-validation estimate of the likelihood, since we predict each future point given all the previous ones.

- If a model is too complex, it will overfit the early examples and will then predict the remaining ones poorly.

Bayesian Model Validation

- Suppose we have two models M_1 and M_2
- Each is associated with a set of parameters θ_1 and θ_2
- We consider priors $p_i(\theta_i | M_i)$, *likelihoods* $f_i(\boldsymbol{x} | \theta_i, M_i)$ and posteriors $p_i(\theta_i | \boldsymbol{x}, M_i)$

$$\pi_i(\theta_i \mid \boldsymbol{x}, M_i) = \frac{f_i(\boldsymbol{x} \mid \theta_i, M_i)\pi_i(\theta_i \mid M_i)}{\pi_i(\boldsymbol{x} \mid M_i)}$$

- We define as the *best* model the one that is more *probable to have generated* the data $\boldsymbol{x}$ that we observed.

Bayesian Model Validation

From data we can learn the parameters for each model and then the model itself

$$\boldsymbol{x} \Rightarrow \pi_i(\theta_i \mid \boldsymbol{x}, M_i) = \frac{f_i(\boldsymbol{x} \mid \theta_i, M_i)\pi_i(\theta_i \mid M_i)}{\pi_i(\boldsymbol{x} \mid M_i)} \Rightarrow \quad \pi_i(M_i \mid \boldsymbol{x}) = \frac{\pi_i(\boldsymbol{x} \mid M_i)\pi_i(M_i)}{\pi(\boldsymbol{x})}$$

Noting that

$$\pi_i(\boldsymbol{x} \mid M_i) = \int f_i(\boldsymbol{x} \mid \theta_i, M_i)\pi_i(\theta_i \mid M_i)d\theta_i$$

we can find the best model that represents the data by computing:

$$\frac{\pi(M_1 \mid \boldsymbol{x})}{\pi(M_2 \mid \boldsymbol{x})} = \frac{\pi(\boldsymbol{x} \mid M_1)\pi(M_1)}{\pi(\boldsymbol{x} \mid M_2)\pi(M_2)} = \underbrace{\frac{\int f_1(\boldsymbol{x} \mid \theta_1, M_1)\pi_1(\theta_1 \mid M_1)d\theta_1}{\int f_2(\boldsymbol{x} \mid \theta_2, M_2)\pi_2(\theta_2 \mid M_2)d\theta_2}}_{B_{10}^{\pi}\ \text{Ratio of Bayes' factors}} \underbrace{\frac{\pi(M_1)}{\pi(M_2)}}_{\text{Ratio of Priors}}$$

Bayesian Model Validation - Example

Consider the coin flipping example

Let θ the probability of getting heads

Consider two models:

M_1 Coin is Fair: $\theta|M_1 \sim \mathcal{B}(100,100)$

M_2 Coin is Unfair: $\theta|M_2 \sim \mathcal{B}(0.5,0.5)$

Data $\boldsymbol{x} = \{2H, 3T\}$

Bayes Factors $$\underbrace{\frac{\int f_1(\boldsymbol{x}\,|\,\theta, M_1)\pi_1(\theta\,|\,M_1)d\theta}{\int f_2(\boldsymbol{x}\,|\,\theta, M_2)\pi_2(\theta\,|\,M_2)d\theta}}_{\substack{\text{Ratio of} \\ \text{Bayes' factors}}} = \frac{\int \theta^2(1-\theta)^3\theta^{99}(1-\theta)^{99}\,/\,beta(100,100)d\theta}{\int \theta^2(1-\theta)^3\theta^{-0.5}(1-\theta)^{-0.5}\,/\,beta(0.5,0,5)d\theta} = \frac{0.031}{0.012}$$

Model Validation $$\frac{\pi(M_1\,|\,\boldsymbol{x})}{\pi(M_2\,|\,\boldsymbol{x})} = \underbrace{\frac{\int f_1(\boldsymbol{x}\,|\,\theta, M_1)\pi_1(\theta\,|\,M_1)d\theta}{\int f_2(\boldsymbol{x}\,|\,\theta, M_2)\pi_2(\theta\,|\,M_2)d\theta}}_{\substack{\text{Ratio of} \\ \text{Bayes' factors}}} \underbrace{\frac{\pi(M_1)}{\pi(M_2)}}_{\substack{\text{Ratio of} \\ \text{Priors}}} = 2.58\frac{\pi(M_1)}{\pi(M_2)}$$

Bayesian Model Validation - Example

Consider the coin flipping example

Let θ probability of getting heads

Two models:

M_1 Coin is Fair: $\theta|M_1 \sim \mathcal{B}(100,100)$

M_2 Coin is Unfair: $\theta|M_2 \sim \mathcal{B}(0.5,0.5)$

Data $\boldsymbol{x} = \{5H\}$

Bayes Factor

$$\underbrace{\frac{\int f_1(\boldsymbol{x}|\theta,M_1)\pi_1(\theta|M_1)d\theta}{\int f_2(\boldsymbol{x}|\theta,M_2)\pi_2(\theta|M_2)d\theta}}_{\substack{Ratio\ of \\ Bayes'\ factors}} = \frac{\int \theta^5\theta^{99}(1-\theta)^{99}/beta(100,100)d\theta}{\int \theta^5\theta^{-0.5}(1-\theta)^{-0.5}/beta(0.5,0,5)d\theta} = \frac{0.033}{0.25}$$

Model Validation

$$\frac{\pi(M_1|\boldsymbol{x})}{\pi(M_2|\boldsymbol{x})} = \underbrace{\frac{\int f_1(\boldsymbol{x}|\theta,M_1)\pi_1(\theta|M_1)d\theta}{\int f_2(\boldsymbol{x}|\theta,M_2)\pi_2(\theta|M_2)d\theta}}_{\substack{Ratio\ of \\ Bayes'\ factors}} \underbrace{\frac{\pi(M_1)}{\pi(M_2)}}_{\substack{Ratio\ of \\ Priors}} = 0.13\frac{\pi(M_1)}{\pi(M_2)}$$

Bayes' factors and posterior model PDFs should be used with caution when non-informative priors are applied.

Bayesian Occam's Razor

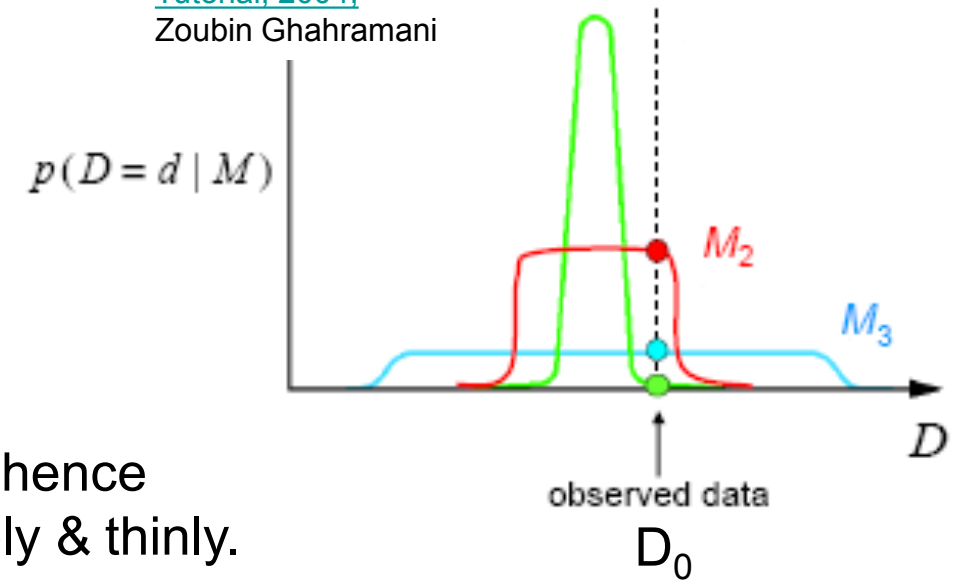
- To further understand the Bayesian Occam's razor effect is to note that probabilities must sum to one (sum over all possible data sets)

$$\sum_{\mathcal{D}'} p\left(\mathcal{D}' \mid m\right) = 1$$

Bayesian Methods for Machine Learning, ICML Tutorial, 2004, Zoubin Ghahramani

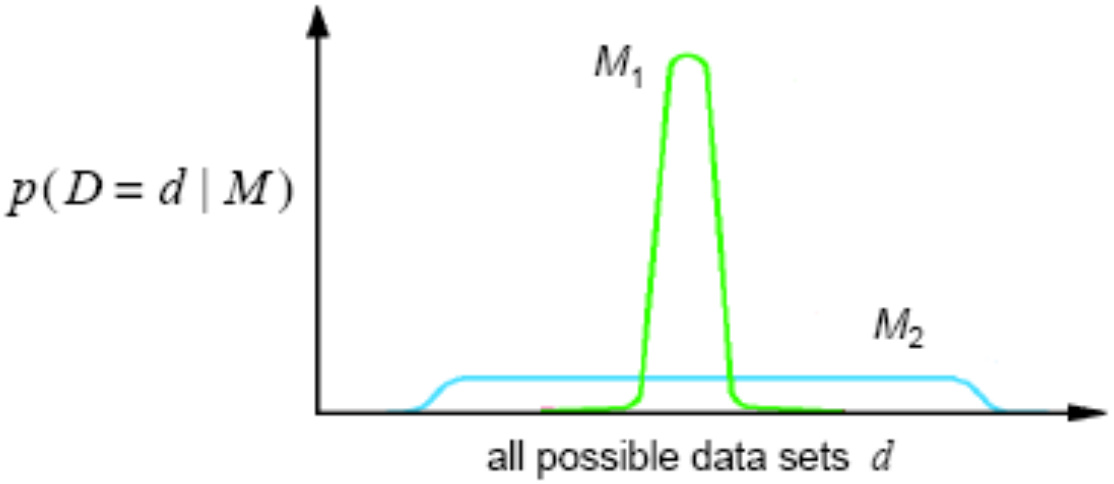
- Model 1 is too simple and assigns low probability to D_0.

- Model 3 also assigns D_0 relatively low probability, because it can predict many data sets, and hence it spreads its probability quite widely & thinly.



- Model 2 is "just right": it predicts the observed data with a reasonable degree of confidence, but does not predict too many other things. Hence model 2 is the most probable model.

Bayesian Occam's Razor



$$For\ any\ model\ M: \quad \sum_{all\ \boldsymbol{d} \in D} p(D = \boldsymbol{d} \mid M) = 1$$

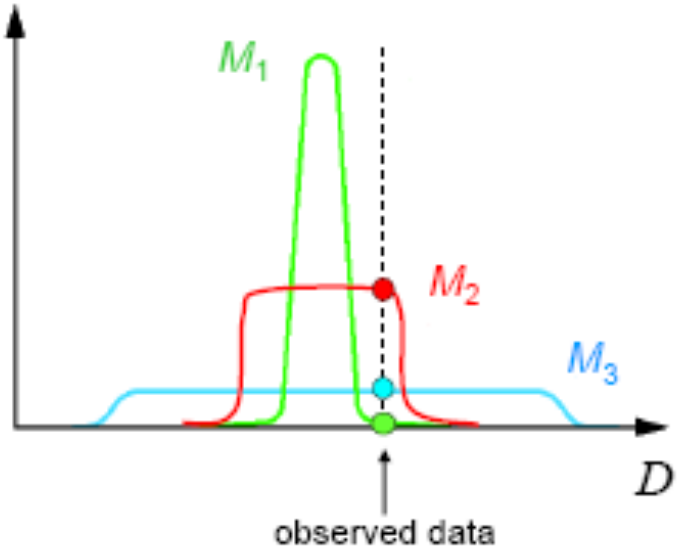
The law of *conservation of belief states that models that explain many* possible data sets must necessarily assign each of them a low probability

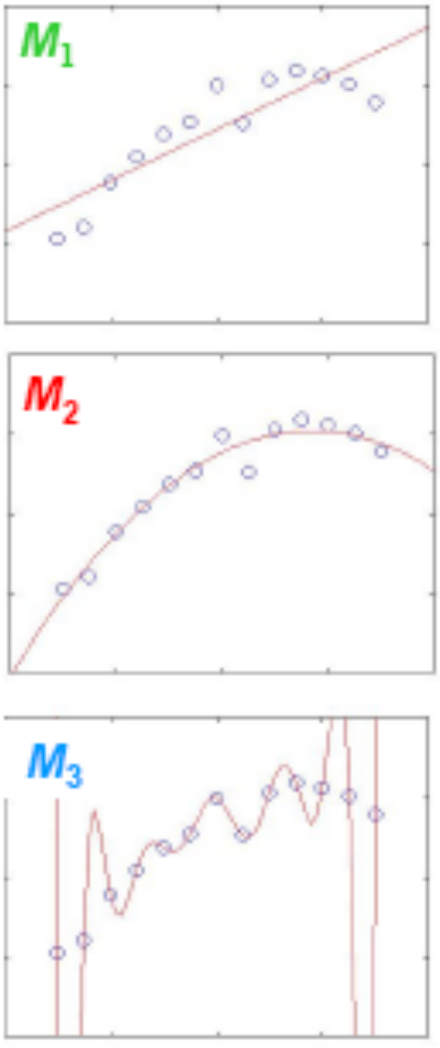
- A note on the evidence and Bayesian Occam's razor, I. Murray and Z. Ghahramani (2005), Gatsby Unit Technical Report GCNU-TR 2005-003
- Occam's Razor, C. Rasmussen and Z. Ghahramani, In T.K. Leen, T.G. Dietterich and V. Tresp (edts), Neural Information Procesing Systems 13, pp. 294-300, 2001, MIT Press

Bayesian Occam's Razor

Bayesian Methods for Machine Learning, ICML Tutorial, 2004, Zoubin Ghahramani

$p(D = d \mid M)$





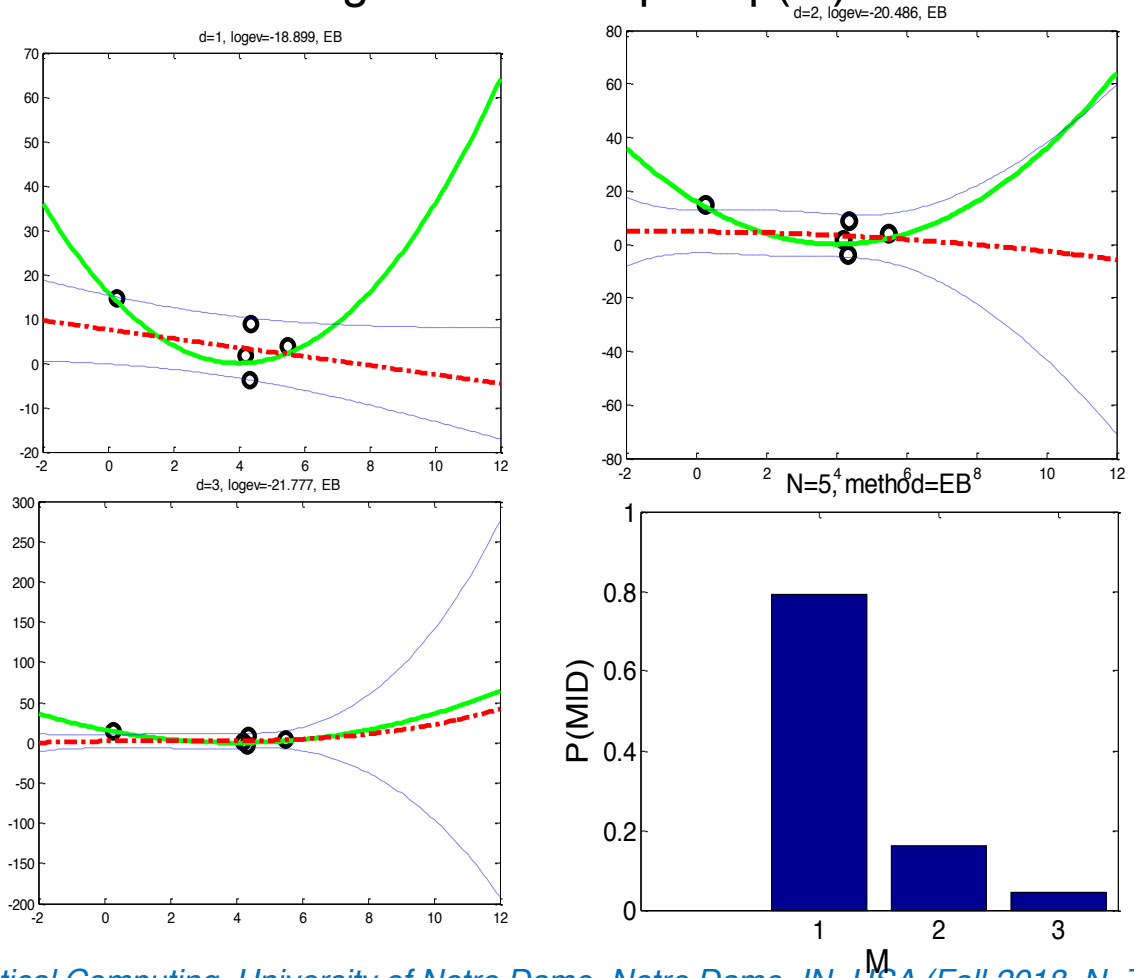
M_1 : the too simple model is unlikely to generate this data

M_3 : the too complex model explains poorly a lots of data sets and it is a little better but still unlikely to have generated our data

M_2 : the just right model has the highest marginal likelihood

Bayesian Occam's Razor

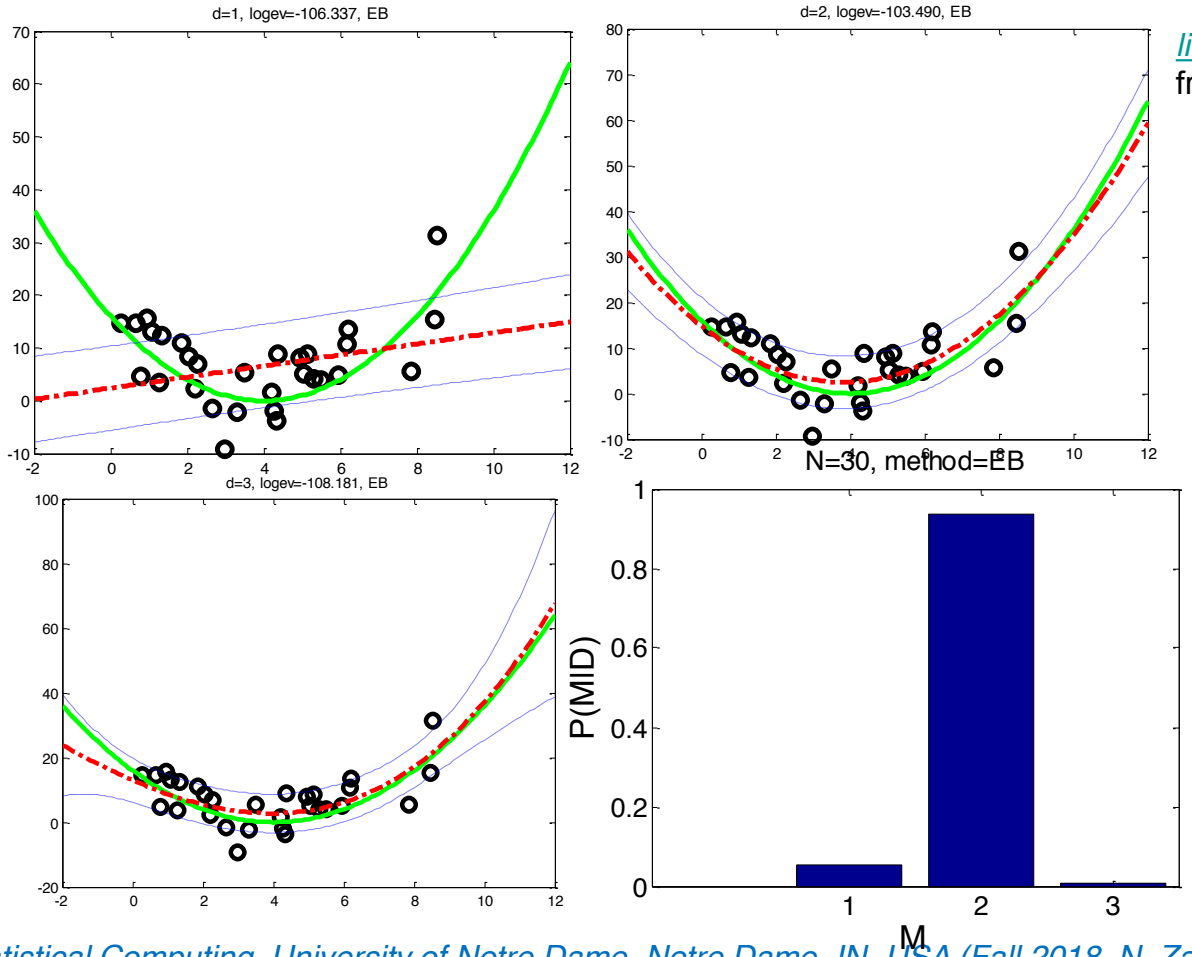
- Polynomials of degrees $1, 2, 3$ fit to $N = 5$ data points using empirical Bayes. Solid green curve is the true function, Dashed red curve is the prediction (dotted blue lines represent $\pm\sigma$ around the mean). The posterior over models $p(m|\mathcal{D})$ is also shown using a Gaussian prior p(m).



linregEbModelSelVsN from Kevin Murphys' PMTK

Bayesian Occam's Razor

- Polynomials of degrees $1, 2, 3$ fit to $N = 30$ data points using empirical Bayes. Solid green curve is the true function, Dashed red curve is the prediction (dotted blue lines represent $\pm\sigma$ around the mean). The posterior over models $p(m|\mathcal{D})$ is also shown using a Gaussian prior $p(m)$.



linregEbModelSelVsN
from Kevin Murphys' PMTK

Marginal Likelihood (Evidence)

- When discussing parameter inference for a fixed model, we often write

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

- We thus ignore the normalization constant $p(\mathcal{D}|m)$. This is valid since $p(\mathcal{D}|m)$ is constant wrt θ.

- However, when comparing models, we need to know how to compute the marginal likelihood, $p(\mathcal{D}|m)$.

- In general, this can be quite hard, since we have to integrate over all possible parameter values, but when we have a conjugate prior, it is easy to compute.

$$p(\mathcal{D} \mid m) = \int p(\mathcal{D} \mid \theta, m)\, p(\theta \mid m)\, d\theta$$

Marginal Likelihood (Evidence)

- Let $p(\theta) = q(\theta)/Z_0$ be our prior, where $q(\theta)$ is an unnormalized distribution, and Z_0 is the normalization constant of the prior.
- Let $p(\mathcal{D}|\theta) = q(\mathcal{D}|\theta)/Z_l$ be the likelihood, where Z_l contains any constant factors in the likelihood.
- Let $p(\theta|\mathcal{D}) = q(\theta|\mathcal{D})/Z_N$ be our posterior, where $q(\theta|\mathcal{D}) = q(\mathcal{D}|\theta)q(\theta)$ is the unnormalized posterior, and Z_N is the normalization constant of the posterior.

$$p(\theta|\mathcal{D}) = \frac{p(\theta)p(\mathcal{D}|\theta)}{p(\mathcal{D})} \Rightarrow \frac{q(\theta|\mathcal{D})}{Z_N} = \frac{q(\theta)q(\mathcal{D}|\theta)}{Z_0 Z_l p(\mathcal{D})} \Rightarrow p(\mathcal{D}) = \frac{Z_N}{Z_0 Z_l}$$

- So assuming the relevant normalization constants are tractable, we have an easy way to compute the marginal likelihood.
- Several examples are presented next.

Beta-Binomial Model

- Let us apply the above result to the Beta-binomial model. Since we know $p(\theta|D) = \mathcal{B}(\theta|a', b')$, where $a' = a + N_1$ and $b' = b + N_0$, we know the normalization constant of the posterior is $\mathcal{B}(a', b')$. Hence

$$p(\theta|\mathcal{D}) = \frac{p(\theta)p(\mathcal{D}|\theta)}{p(\mathcal{D})} = \frac{1}{p(\mathcal{D})}\left[\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}\right]\left[\binom{N}{N_1}\theta^{N_1}(1-\theta)^{N_0}\right]$$

$$\frac{1}{B(a+N_1, b+N_0)} = \frac{1}{p(\mathcal{D})}\binom{N}{N_1}\frac{1}{B(a,b)}$$

$$p(\mathcal{D}) = \binom{N}{N_1}\frac{B(a+N_1, b+N_0)}{B(a,b)}$$

- The marginal likelihood for the Beta-Bernoulli model is the same as above, but without the $\binom{N}{N_1}$ term.

Dirichlet-Multinoulli Model

- One can show that the marginal likelihood for the Dirichlet-multinoulli model is given by

$$p(\mathcal{D}) = \frac{B(N+\alpha)}{B(\alpha)}, B(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}$$

- Hence we can rewrite the above result in the following form, which is more often used

$$p(\mathcal{D}) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\Gamma\left(N + \sum_{k=1}^{K} \alpha_k\right)} \prod_{k=1}^{K} \frac{\Gamma(N_k + \alpha_k)}{\Gamma(\alpha_k)}$$

Gaussian-Gaussian-Wishart Model

- Consider the case of a $\mathcal{MVN}$ with a conjugate $\mathcal{NIW}$ prior. Let Z_0 be the normalizer for the prior, Z_N be normalizer for the posterior, and let $Z_l = (2p)^{ND/2}$ be the normalizer for the likelihood. Then it is easy to see that

$$p(\mathcal{D}) = \frac{Z_N}{Z_0 Z_l} = \frac{1}{\pi^{ND/2}} \frac{1}{2^{ND/2}} \frac{\left(\frac{2\pi}{\kappa_N}\right)^{D/2} |S_N|^{-\nu_N/2} 2^{(\nu_0+N)D/2} \Gamma_D(\nu_N/2)}{\left(\frac{2\pi}{\kappa_0}\right)^{D/2} |S_0|^{-\nu_N/2} 2^{\nu_0 D/2} \Gamma_D(\nu_0/2)}$$

$$= \frac{1}{\pi^{ND/2}} \left(\frac{\kappa_0}{\kappa_N}\right)^{D/2} \frac{|S_0|^{\nu_0/2}}{|S_N|^{\nu_N/2}} \frac{\Gamma_D(\nu_N/2)}{\Gamma_D(\nu_0/2)}$$

- This equation will prove useful later.

$$\mathcal{NIW}(\mu, \Sigma \mid \mu_0, \kappa_0, S_0, \nu_0) = \mathcal{N}\left(\mu \mid \mu_0, \frac{1}{\kappa_0}\Sigma\right) \mathcal{InvWis}(\Sigma \mid S_0, \nu_0) =$$

$$= \frac{1}{Z_{NIW}} |\Sigma|^{-1/2} exp\left(-\frac{\kappa_0}{2}(\mu-\mu_0)^T \Sigma^{-1}(\mu-\mu_0)\right) |\Sigma|^{-(\nu_0+D+1)/2} \exp\left(-\frac{1}{2} Tr\left(\Sigma^{-1} S_0\right)\right)$$

$$= \frac{1}{Z_{NIW}} exp\left(-\frac{\kappa_0}{2}(\mu-\mu_0)^T \Sigma^{-1}(\mu-\mu_0) - \frac{1}{2} Tr\left(\Sigma^{-1} S_0\right)\right) |\Sigma|^{-(\nu_0+D+2)/2}$$

$Z_{NIW} = 2^{\nu_0 D/2} \Gamma_D\left(\frac{\nu_0}{2}\right)\left(\frac{2\pi}{\kappa_0}\right)^{D/2} |S_0|^{-\nu_0/2}$, Γ_D multivariate Gamma function

Appendix: Laplace Approximation

- 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:

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

- 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 2^{nd} order Taylor expansion of the log-posterior):

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

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

Laplace Approximation and Model Evidence

- We can write the model evidence as

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

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

$$\log p(\mathcal{D}\,|\,\mathcal{M}_m) \approx \log p(\boldsymbol{\theta}_m^{MAP}, \mathcal{D}\,|\,\mathcal{M}_m) - \log p(\boldsymbol{\theta}_m^{MAP}\,|\,\mathcal{D}, \mathcal{M}_m)$$

$$\approx \log p(\mathcal{D}\,|\,\boldsymbol{\theta}_m^{MAP}, \mathcal{M}_m) + \log p(\boldsymbol{\theta}_m^{MAP}\,|\,\mathcal{M}_m) + \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{A}| + \frac{1}{2}\left(\boldsymbol{\theta}_m^{MAP} - \boldsymbol{\theta}_m^{MAP}\right)^T \boldsymbol{A}\left(\boldsymbol{\theta}_m^{MAP} - \boldsymbol{\theta}_m^{MAP}\right)$$

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

- This Laplace approximation is used often for model comparison.

- Other approximations are also very useful:

  - Bayesian Information Criterion (BIC) (on the limit of $N \to \infty$)
  - MCMC (Sampling approach)
  - Variational Methods

Bayesian Information Criterion

- Start with the Laplace approximation for large data sets $N \to \infty$,

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

- 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:

$$\log p(\mathcal{D} \mid \mathcal{M}_m) \approx \log p(\mathcal{D} \mid \boldsymbol{\theta}_m^{MAP}, \mathcal{M}_m) - \frac{d}{2}\log N \quad (\text{limit } N \to \infty)$$

- 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 (but use MAP when working with mixtures of Gaussians)
  - If not all parameters are well determined from the data, $\boldsymbol{\theta}_m$ d=number of effective parameters.

BIC Approximation to Log Marginal Likelihood

- The Bayesian information criterion or BIC thus has the following form:

$$BIC = \log p(\mathcal{D} \mid \bar{\boldsymbol{\theta}}_m, \mathcal{M}_m) - \frac{dof\left(\bar{\boldsymbol{\theta}}_m\right)}{2} \log N \approx \log p(\mathcal{D} \mid \mathcal{M}_m) \quad (\text{limit } N \to \infty)$$

- $dof\left(\bar{\boldsymbol{\theta}}_m\right)$ is the number of degrees of freedom in the model, and $\bar{\boldsymbol{\theta}}_m$ is the MLE for the model. We see that this has the form of a penalized log likelihood, where the penalty term depends on the model complexity.

- As an example consider linear regression. The MLE, log likelihood and BIC are:

$$MLE: \bar{\boldsymbol{w}} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y}, \bar{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \bar{\boldsymbol{w}}^T \boldsymbol{x}_i\right)$$

$$\log p(\mathcal{D} \mid \bar{\boldsymbol{\theta}}) = -\frac{N}{2} \log\left(2\pi\bar{\sigma}^2\right) - \frac{N}{2}$$

$$BIC = -\frac{N}{2} \log\left(2\pi\bar{\sigma}^2\right) - \frac{N}{2} - \frac{D}{2} \log N$$

BIC Approximation to Log Marginal Likelihood

- Hence the BIC score is as follows (dropping constant terms)

$$BIC = -\frac{N}{2}\log\left(2\pi\bar{\sigma}^2\right) - \frac{N}{2} - \frac{D}{2}\log N$$

- D is the number of variables in the model. In the statistics literature, it is common to use an alternative definition of BIC, which we call the BIC cost (since we want to minimize it):

$$BIC - Cost = -2\log p(\mathcal{D} \mid \bar{\boldsymbol{\theta}}_m, \mathcal{M}_m) + dof\left(\bar{\boldsymbol{\theta}}_m\right)\log N \approx -2\log p(\mathcal{D} \mid \mathcal{M}_m)$$

- In the context of the regression example, this becomes:

$$BIC - Cost = N\log\left(2\pi\bar{\sigma}^2\right) + D\log N$$

- The BIC method is related to the minimum description length or MDL principle. It characterizes the score of how well the model fits the data, minus how complex the model is.

Akaike Information Criterion

- There is a very similar expression to BIC/ MDL called the Akaike information criterion or AIC, defined as

$$AIC(m, \mathcal{D}) = \log p(\mathcal{D} \mid \bar{\boldsymbol{\theta}}_m, \mathcal{M}_m) - dof\left(\bar{\boldsymbol{\theta}}_m\right)$$

- This is derived from a frequentist framework, and cannot be interpreted as an approximation to the marginal likelihood.

- The penalty for AIC is less than for BIC. This causes AIC to pick more complex models. However, this sometimes can result in better predictive accuracy!

- Clarke, B., E. Fokoue, and H. H. Zhang (2009). *Principles and Theory for Data Mining and Machine Learning*. Springer.

Effect of the Prior/Empirical Bayes

- When performing posterior inference, the prior may not matter too much since the likelihood often overwhelms the prior.

- But when computing the marginal likelihood, the prior plays a much more important role, since we are averaging the likelihood over all possible parameter settings, as weighted by the prior.

- If the prior is unknown, the correct Bayesian procedure is to put a prior on the prior. That is, we should put a prior on the hyper-parameter a as well as the $\boldsymbol{w}$. To compute the marginal likelihood, we should integrate out all unknowns, i.e., we should compute

$$p(\mathcal{D} \mid \boldsymbol{m}) = \iint p(\mathcal{D} \mid \boldsymbol{w})\, p(\boldsymbol{w} \mid \alpha, m)\, p(\alpha \mid m) d\boldsymbol{w} d\alpha$$

Empirical Bayes

- This requires specifying the hyper-prior.
- Fortunately, the higher up we go in the Bayesian hierarchy, the less sensitive are the results to the prior settings. Thus can usually make the hyper-prior uninformative.
- A computational shortcut is to optimize α rather than integrating it out.

$$p(\mathcal{D} \mid \boldsymbol{m}) = \int p(\mathcal{D} \mid \boldsymbol{w})\, p\left(\boldsymbol{w} \mid \overline{\alpha}, m\right) d\boldsymbol{w}$$

where

$$\overline{\alpha} = \arg\max_{\alpha} p(\mathcal{D} \mid \alpha, \boldsymbol{m}) = \arg\max_{\alpha} \int p(\mathcal{D} \mid \boldsymbol{w})\, p(\boldsymbol{w} \mid \alpha, m)\, d\boldsymbol{w}$$

- This approach is called empirical Bayes (EB).

Back to Bayes Factors

- Suppose our prior on models is uniform, $p(m) \sim 1$. Then model selection is equivalent to picking the model with the highest marginal likelihood. Now suppose we just have two models we are considering, call them the null hypothesis, M_0, and the alternative hypothesis, M_1.

- Define the Bayes factor as the ratio of marginal likelihoods:

$$BF_{1,0} = \frac{p(\mathcal{D} \mid M_1)}{p(\mathcal{D} \mid M_0)} = \frac{p(M_1 \mid \mathcal{D})}{p(M_0 \mid \mathcal{D})} \Big/ \frac{p(M_1)}{p(M_0)}$$

- If $BF_{10} > 1$, we prefer model 1, otherwise we prefer model 0. Jeffreys proposed a scale of evidence shown below

Bayes factor BF(1,0)	Interpretation
BF< 1/100	Decisive evidence for M_0
BF< 1/10	Strong evidence for M_0
1/10< BF< 1/3	Moderate evidence for M_0
1/3< BF < 1	Weak evidence for M_0
1 < BF < 3	Weak evidence for M_1
3 < BF < 10	Moderate evidence for M_1
BF>10	Strong evidence for M_1
BF>100	Decisive evidence for M_1

Bayes Model Selection: Jeffrey's Scale of Evidence

- Using the alternative reference below, Jeffrey's scale of evidence says:
  - For $\log\left(B_{10}^{\pi}\right)$ between 0 and 0.5, the evidence against H_0 is poor
  - In between 0.5 and 1, it is substantial
  - In between 1 and 2, it is strong and
  - Above 2, it is decisive.

$$B_{10}^{\pi} = \frac{\pi(x \mid H_1)}{\pi(x \mid H_0)}$$

- Bayes' factor tells us if one should prefer H_0 to H_1 (relative comparison of models).
- Bayes' factor does not tell us whether any of these models is sensible.

Estimation and Beyond in the Bayes Universe, Brani Vidakovic (online Course on Bayesian Stat. for Engineers)

Example: Testing if a Coin is Fair

- Suppose we observe some coin tosses, and want to decide if the data was generated by a fair coin, $\theta = 0.5$, or a potentially biased coin, where θ in $[0, 1]$. Denote the fair coin model by M_0 and the biased coin model by M_1.

- The marginal likelihood under M_0 is simply

$$p(\mathcal{D} \mid M_0) = \left(\frac{1}{2}\right)^N$$

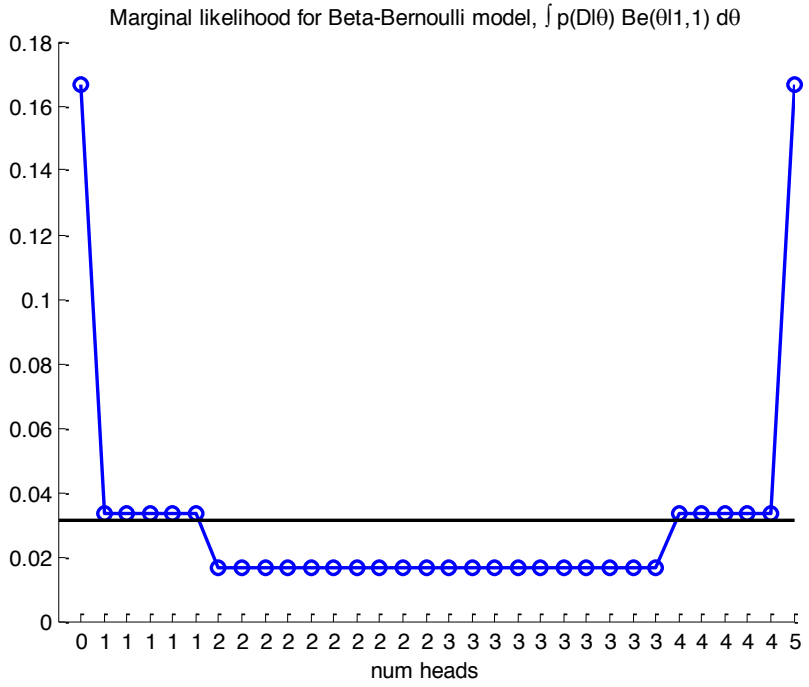
where N is the number of coin tosses.

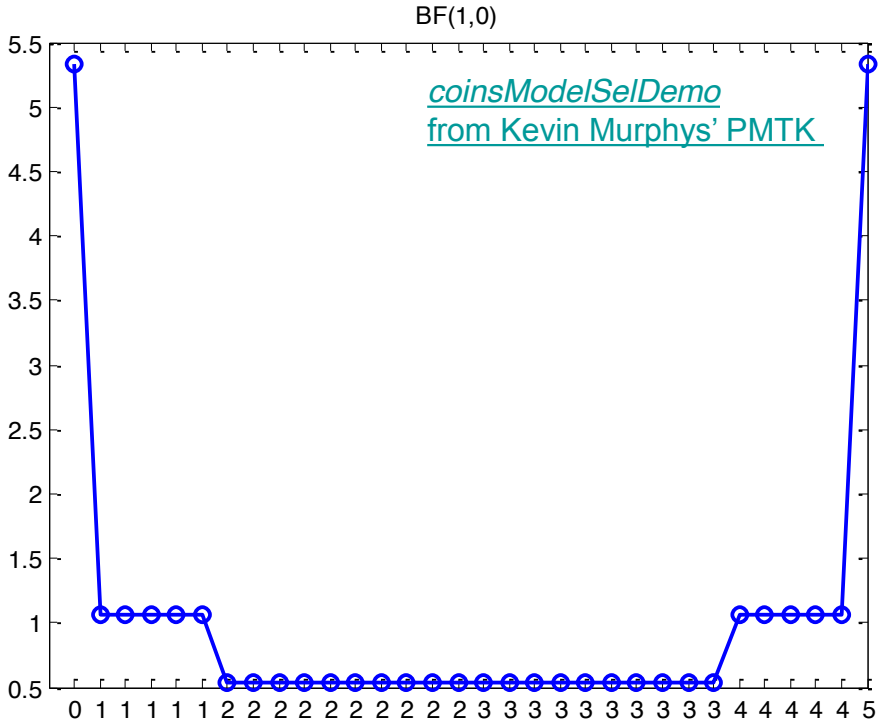
- The marginal likelihood under M_1 using a Beta prior, is

$$p(\mathcal{D} \mid M_1) = \int p(\mathcal{D} \mid \theta)\, p(\theta \mid M_1)\, d\theta = \frac{B(\alpha_1 + N_1, \alpha_0 + N_0)}{B(\alpha_1, \alpha_0)}$$

Example: Testing if a Coin is Fair

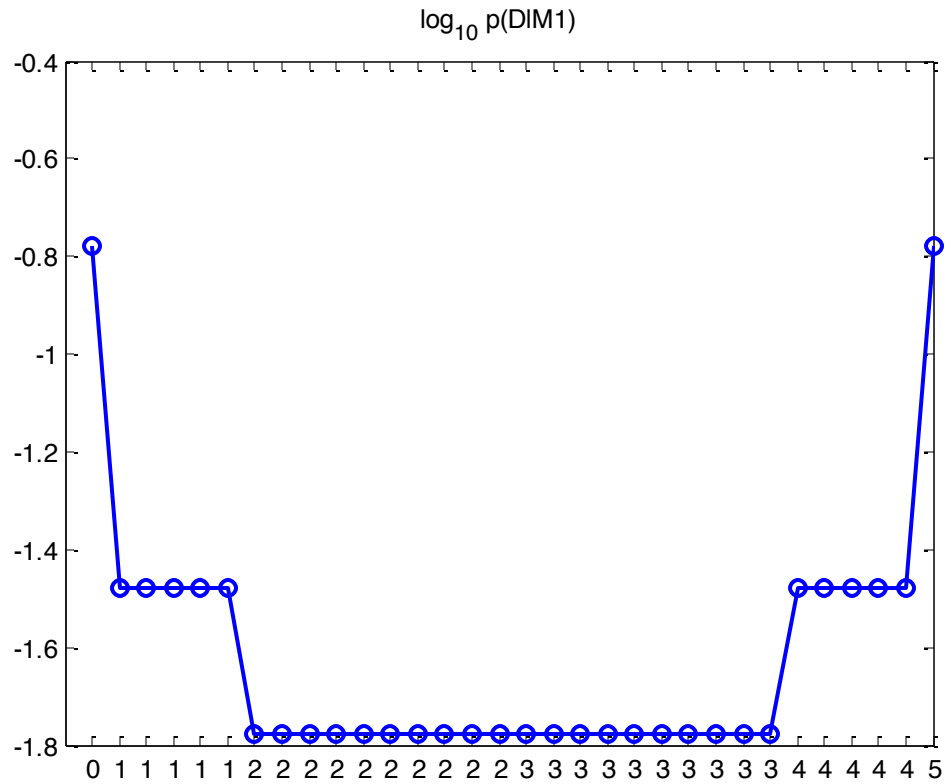
- We plot $\log p(\mathcal{D}|M_0)$ and $\log p(\mathcal{D}|M_1)$ vs the number of heads N_1 with $N = 5$ and $a_1 = a_0 = 1$.
- If we observe 2 or 3 heads, the unbiased coin hypothesis M_0 is more likely than M_1 since M_0 is a simpler model - it would be a suspicious coincidence if the coin were biased but happened to produce almost exactly $50/50$ heads/tails.
- However, as the counts become more extreme, we favor the biased coin hypothesis. Note that, if we plot the log Bayes factor, $logBF_{10}$ it will have exactly the same shape, since $logp(\mathcal{D}|M_0)$ is a constant.

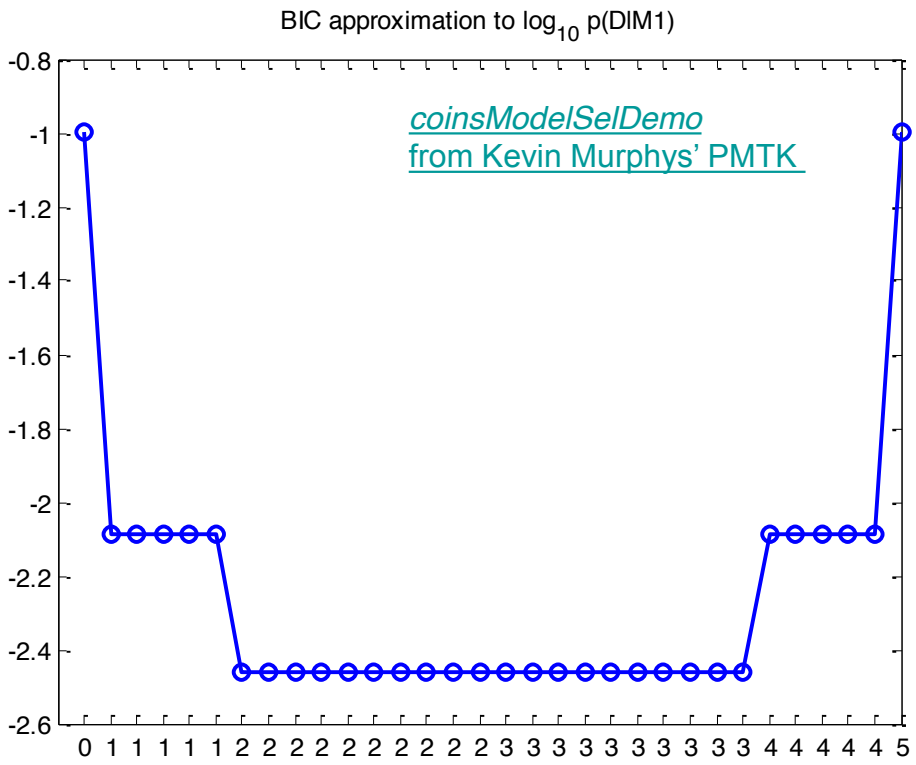




Example: Testing if a Coin is Fair

- Log marginal likelihood for coins example and the BIC approximation to $\log p(\mathcal{D}|M_1)$ for our biased coin example.
- The curve has approximately the same shape as the exact log marginal likelihood.
- It favors the simpler model unless the data is overwhelmingly in support of the more complex model.





Bayes' Factors

- Let us compare two hypotheses (or models):

$$H_0 : \theta \sim \pi_0 \; versus \; H_1 : \theta \sim \pi_1$$

Here, we assume that both models have the same parameters.

- Then the (mixture model) prior is

$$\pi(\theta) = \pi(H_0)\pi_0(\theta) + \pi(H_1)\pi_1(\theta), \; where \; \pi(H_0) + \pi(H_1) = 1.$$

- To compare H_0 versus H_1, we compute the Bayes factor which partially eliminates the influence of the priors $\pi(H_0), \pi(H_1)$.

$$B_{10}^{\pi} = \frac{\pi(x \mid H_1)}{\pi(x \mid H_0)} = \frac{\int f(x \mid \theta)\pi_1(\theta)d\theta}{\int f(x \mid \theta)\pi_0(\theta)d\theta} = \frac{\pi(H_1 \mid x)}{\pi(H_0 \mid x)} \frac{\pi(H_0)}{\pi(H_1)}$$

Varying Parameter Space

- Bayes' model comparison is not limited to models with the same parameter space.

- Assume we have some data and two statistical models.

- Under

$$H_0, \theta_0 \in \Theta_0, \text{the prior is } \pi_0(\theta_0) \text{ and the likelihood is } f_0(x \mid \theta_0)$$

$$H_1, \theta_1 \in \Theta_1, \text{the prior is } \pi_1(\theta_1) \text{ and the likelihood is } f_1(x \mid \theta_1)$$

- Then:

$$B_{10}^{\pi} = \frac{\pi(x \mid H_1)}{\pi(x \mid H_0)} = \frac{\int f_1(x \mid \theta_1)\pi_1(\theta_1)d\theta_1}{\int f_0(x \mid \theta_0)\pi_0(\theta_0)d\theta_0}$$

- One can have $\Theta_0 = \mathbb{R}, \text{and } \Theta_1 = \mathbb{R}^{10000}$!

Bayes Factors and Model Comparison

- Bayesian hypothesis testing procedure depends on $P^{\pi}(\theta \in \Theta_0 | x)$ or alternatively on the Bayes factor (ratio of posteriors to priors)

$$B_{10}^{\pi} = \frac{P^{\pi}(\theta \in \Theta_1 | x) / P^{\pi}(\theta \in \Theta_0 | x)}{P^{\pi}(\theta \in \Theta_1) / P^{\pi}(\theta \in \Theta_0)}$$

- Corresponding models H_1 vs. H_0 are compared via

$$B_{10}^{\pi} \equiv \frac{P^{\pi}(x | H_1)}{P^{\pi}(x | H_0)} = \frac{P^{\pi}(H_1 | x) / P^{\pi}(H_1)}{P^{\pi}(H_0 | x) / P^{\pi}(H_0)}$$

- If we rewrite the prior as

$$\pi(\theta) = \Pr(\theta \in \Theta_1) \times \pi_1(\theta) + \Pr(\theta \in \Theta_0) \times \pi_0(\theta)$$

then

$$B_{10}^{\pi} = \int f(x|\theta_1)\pi_1(\theta_1)d\theta_1 \Big/ \int f(x|\theta_0)\pi_0(\theta_0)d\theta_0 = m_1(x)/m_0(x)$$

R. Kass & A. Raftery, JASA, Vol. 90, 773-795 (1995), R. Kass, The Statistician, Vol. 42, 551-560 (1993)

Bayes Factors and Model Comparison

- You can also compute the posterior probabilities of H_0 and H_1:

$$\pi(H_0 \mid x) = \frac{\pi(x \mid H_0)\pi(H_0)}{\pi(x)} = \frac{\pi(x \mid H_0)\pi(H_0)}{\pi(x \mid H_0)\pi(H_0) + \pi(x \mid H_1)\pi(H_1)}$$

- The posterior probabilities satisfy:

$$\frac{\pi(H_1 \mid x)}{\pi(H_0 \mid x)} = \frac{\pi(x \mid H_1)}{\pi(x \mid H_0)} \frac{\pi(H_1)}{\pi(H_0)} = B_{10}^{\pi} \frac{\pi(H_1)}{\pi(H_0)}$$

Testing Point Null Hypothesis

- If $\Theta_0 = \{\theta_0\}$, π_0 is the Dirac mass at θ_0. Then:

$$\rho = P^{\pi}(\theta = \theta_0) \quad \text{and} \quad \pi(\theta) = \Pr(\theta \in \Theta_0) \times \pi_0(\theta) + \Pr(\theta \in \Theta_1) \times \pi_1(\theta)$$
$$= \rho \mathbb{I}_{\theta_0}(\theta) + (1-\rho)\pi_1(\theta)$$

$$\pi(\Theta_0 | x) = \frac{f(x|\theta_0)\rho}{\int f(x|\theta)\pi(\theta)d\theta} = \frac{f(x|\theta_0)\rho}{f(x|\theta_0)\rho + (1-\rho)m_1(x)}$$

where

$$m_1(x) = \int_{\Theta_1} f(x|\theta)\pi_1(\theta)d\theta$$

- Above we used the posterior calculation as seen earlier:

$$\pi(H_0 | x) = \frac{\pi(x | H_0)\pi(H_0)}{\pi(x)} = \frac{\pi(x | H_0)\pi(H_0)}{\pi(x | H_0)\pi(H_0) + \pi(x | H_1)\pi(H_1)}$$

Point Null Hypothesis

- Bayes procedures can be used to test **point** null hypothesis, i.e.

$$H_0, \theta = \theta_0 \ (i.e.\ \pi_0(\theta) = \delta_{\theta_0}(\theta)) \ \ versus \ \ H_1, \theta \sim \pi_1$$

- The prior is defined as

$$\pi(\theta) = \pi(H_0)\delta_{\theta_0}(\theta) + \pi(H_1)\pi_1(\theta)$$

- The associated Bayes' factor is simply

$$B_{10}^{\pi} = \frac{\pi(x \mid H_1)}{\pi(x \mid H_0)} = \frac{\int f(x \mid \theta)\pi_1(\theta)d\theta}{f(x \mid \theta_0)}$$

The Coin Example

- Assume we have a coin, we toss it 10 times and get $x = 10$ heads. Is the coin biased?
- Let θ be the probability of having a head then we can test $H_0 : \theta = \frac{1}{2}$.
- In a frequentist approach, the p-value $\Pr(X \geq 10 \mid H_0) = 2^{-9}$ and the hypothesis is rejected.
- In a Bayesian framework, we test H_0 versus $H_1 : \theta \sim \mathcal{U}(\frac{1}{2}, 1]$ using:

$$B_{10}^{\pi} = \frac{\int f(x \mid \theta)\pi_1(\theta)d\theta}{f(x = 10 \mid \frac{1}{2})} = \frac{\int_{1/2}^{1} \theta^{10}(1-\theta)^{10-10} 2 d\theta}{\left(\frac{1}{2}\right)^{10}\left(1-\frac{1}{2}\right)^{10-10}} = \frac{2\int_{1/2}^{1} \theta^{10} d\theta}{\left(\frac{1}{2}\right)^{10}} \approx 186.08$$

- So the evidence against H_0 is decisive.

Jeffreys Lindley Paradox

- Define the marginal density of θ as: $p(\theta) = p(\theta \mid M_0)\, p(M_0) + p(\theta \mid M_1)\, p(M_1)$ where we consider the hypothesis $M_0 : \theta \in M_0 \; vs \; M_1 : \theta \in M_1$
- We can estimate the posterior as (denote: $p(M_0) = \rho,\ p(M_1) = 1 - \rho$)

$$p(M_0 \mid \mathcal{D}) = \frac{p(M_0)\, p(\mathcal{D} \mid M_0)}{p(M_0)\, p(\mathcal{D} \mid M_0) + p(M_1)\, p(\mathcal{D} \mid M_1)} =$$

$$= \frac{\rho \int_{\Theta_0} p(\mathcal{D} \mid \theta) p(\theta \mid M_0) d\theta}{\rho \int_{\Theta_0} p(\mathcal{D} \mid \theta) p(\theta \mid M_0) d\theta + (1-\rho) \int_{\Theta_1} p(\mathcal{D} \mid \theta) p(\theta \mid M_1) d\theta}$$

- Let us now assume that the priors are improper: $p(\theta \mid M_0) \propto c_0,\ p(\theta \mid M_1) \propto c_1$
- Then the posterior is completely determined by the ratio c_0/c_1 (so it can be anything we want!)

$$p(M_0 \mid \mathcal{D}) = \frac{\rho \int_{\Theta_0} p(\mathcal{D} \mid \theta) d\theta}{\rho \int_{\Theta_0} p(\mathcal{D} \mid \theta) d\theta + (1-\rho)\left[c_1/c_0\right] \int_{\Theta_1} p(\mathcal{D} \mid \theta) d\theta}$$

- Using proper but vague priors causes similar problem. The Bayes factor will always favor the simpler model – complex models with diffuse priors have low probability. This is known as the Jeffreys-Lindley paradox.

Vague Priors: Jeffreys-Lindley Paradox

- For $x \sim \mathcal{N}(\theta, \sigma^2)$ and $\theta \sim N(0, \tau^2)$, to test of $H_0 : \theta = 0$ requires a modification of the prior, with

$$\pi_1(\theta) \propto e^{-\theta^2/2\tau^2} \mathbb{I}_{\theta \neq 0}$$

and $\pi_0(\theta)$ the Dirac mass at 0.

- Then $B_{10}^{\pi}(x) = \dfrac{\pi(x \mid H_1)}{\pi(x \mid H_0)}$ can be computed as:

$$B_{10}^{\pi}(x) = \frac{m_1(x)}{f\left(x|0\right)} = \frac{\int \mathcal{N}(x;\theta,\sigma^2)\mathcal{N}(\theta;0,\tau^2)d\theta}{\mathcal{N}(x;0,\sigma^2)} = \frac{\sigma}{\sqrt{\sigma^2+\tau^2}} \frac{e^{-x^2/2\left(\sigma^2+\tau^2\right)}}{e^{-x^2/2\sigma^2}} = \sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}} \exp\left\{\frac{\tau^2 x^2}{2\sigma^2\left(\sigma^2+\tau^2\right)}\right\}$$

- The marginal distribution on the numerator is computed noting:

$$e^{-\frac{1}{2}\left(\frac{(x-\theta)^2}{\sigma^2}+\frac{(\theta)^2}{\tau^2}\right)} \sim e^{-\frac{z}{2(1-\bar{\rho}^2)}}, \; z = \frac{x^2}{\sigma^2+\tau^2} + \frac{\theta^2}{\tau^2} - 2\bar{\rho}\frac{x\theta}{\sqrt{\sigma^2+\tau^2}\tau}, \; \bar{\rho} = \frac{\sigma}{\sqrt{\sigma^2+\tau^2}} \Rightarrow$$

$$\int \mathcal{N}(x;\theta,\sigma^2)\mathcal{N}(\theta;0,\tau^2)d\theta = \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2+\tau^2}} e^{-x^2/2\left(\sigma^2+\tau^2\right)}$$

Vague Priors: Jeffreys-Lindley Paradox

- Using

$$B_{10}^{\pi}(x)=\frac{\pi(x\mid H_1)}{\pi(x\mid H_0)}=\sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}}\exp\left\{\frac{\tau^2x^2}{2\sigma^2\left(\sigma^2+\tau^2\right)}\right\}$$

and using $p(H_0)=\rho,\ p(H_1)=1-\rho$, we can then compute:

$$\frac{\pi(H_1\mid x)}{\pi(H_0\mid x)}=\frac{\pi(x\mid H_1)(1-\rho)}{\pi(x\mid H_0)\rho}=\frac{1-\rho}{\rho}B_{10}^{\pi}(x)\underset{\pi(H_1|x)+\pi(H_0|x)=1}{\Rightarrow}$$

$$\pi(H_0\mid x)=\pi(\theta=0\mid x)=\left[1+\frac{1-\rho}{\rho}B_{10}^{\pi}(x)\right]^{-1}$$

or

$$\pi(\theta=0|x)=\left[1+\frac{1-\rho}{\rho}\sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}}\exp\left\{\frac{\tau^2x^2}{2\sigma^2\left(\sigma^2+\tau^2\right)}\right\}\right]^{-1}$$

Vague Priors: Testing the Mean of a Gaussian

$$\pi(H_0 \mid x) = \pi(\mu = 0 \mid x) = \left[1 + \frac{1-\rho}{\rho} B_{10}^{\pi}(x)\right]^{-1},$$

$$B_{10}^{\pi} = \frac{\sigma}{\sqrt{\sigma^2 + \tau^2}} \exp\left(\frac{\tau^2 x^2}{2\sigma^2(\sigma^2 + \tau^2)}\right)$$

- The Bayes factor depends heavily on τ^2. As $\tau^2 \to \infty$, the prior becomes un-informative but then $B_{10}^{\pi}(x) \to 0$ regardless of what x is and $\pi(H_0 \mid x) \to 1$.

- **Using vague priors for model selection is a very bad idea (Lindley's paradox).**

Jeffreys-Lindley Paradox

- For $z = x/\sigma$ and $\rho = 1/2$, we see below a strong dependence on τ^2.

$$\pi(\theta = 0 | x) = \left[1 + \frac{1-\rho}{\rho} \sqrt{\frac{\sigma^2}{\sigma^2 + \tau^2}} \exp\left\{ \frac{\tau^2 x^2}{2\sigma^2 (\sigma^2 + \tau^2)} \right\} \right]^{-1}$$

z	0	0.68	1.28	1.96
$\pi(\theta = 0 \mid z, \tau^2 = \sigma^2)$	0.586	0.557	0.484	0.351
$\pi(\theta = 0 \mid z, \tau^2 = 10\sigma^2)$	0.768	0.729	0.612	0.366

See MatLab implementation

C. P. Robert, *The Bayesian Core*, *Springer, 2nd edition,* chapter 2 (full text available)

Jeffreys-Lindley Paradox

- For the dataset normaldata, the range of the Bayes factor is computed with the following proof (using the likelihood from earlier calculations) as well as the empirical variance $\bar{\sigma}^2$

$$B_{10}^{\pi}(\bar{x}_n) = \frac{\int \left(\bar{\sigma}^2\right)^{-n/2} \exp\left\{-\left(n\left(\theta-\bar{x}_n\right)^2 + s_x^2\right)\Big/ 2\bar{\sigma}^2\right\} \frac{1}{\sqrt{2\pi}} \frac{1}{\tau} e^{-\theta^2/2\tau^2} d\theta}{\left(\bar{\sigma}^2\right)^{-n/2} \exp\left\{-\left(n\bar{x}_n^2 + s_x^2\right)\Big/ 2\bar{\sigma}^2\right\}} =$$

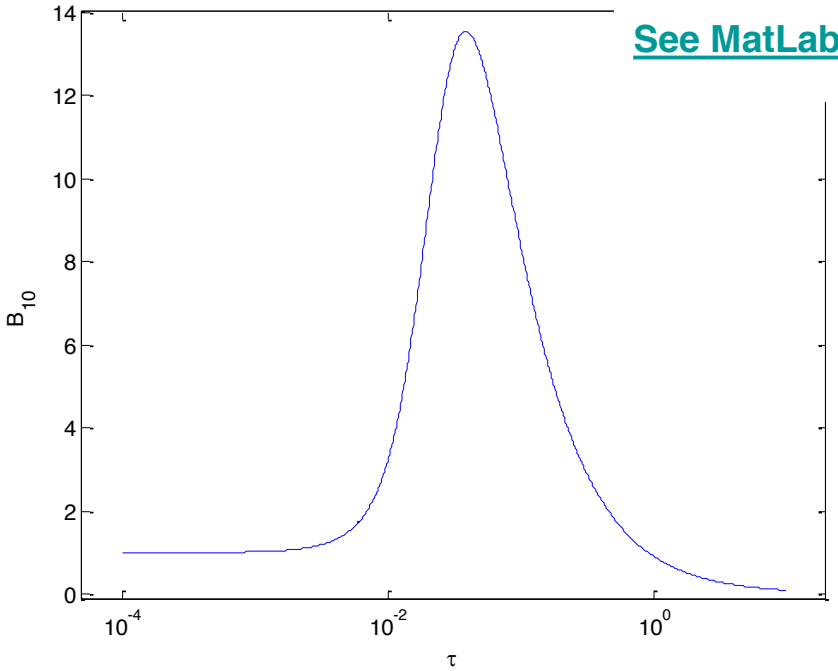
$$\frac{\int \exp\left\{-\left(\theta-\bar{x}_n\right)^2 \Big/ 2\left(\frac{\bar{\sigma}}{\sqrt{n}}\right)^2\right\} \frac{1}{\sqrt{2\pi}} \frac{1}{\tau} e^{-\theta^2/2\tau^2} d\theta}{\exp\left\{-\bar{x}_n^2 \Big/ 2\left(\frac{\bar{\sigma}}{\sqrt{n}}\right)^2\right\}} = \frac{\int \mathcal{N}\left(\bar{x}_n; \theta, \left(\frac{\bar{\sigma}}{\sqrt{n}}\right)^2\right) \mathcal{N}(\theta; 0, \tau^2) d\theta}{\mathcal{N}\left(\bar{x}_n; 0, \left(\frac{\bar{\sigma}}{\sqrt{n}}\right)^2\right)} =$$

$$\sqrt{\frac{\bar{\sigma}^2/n}{\bar{\sigma}^2/n + \tau^2}} \exp\left\{\frac{\tau^2 \bar{x}_n^2}{2\frac{\bar{\sigma}^2}{n}\left(\frac{\bar{\sigma}^2}{n} + \tau^2\right)}\right\} = \sqrt{\frac{\bar{\sigma}^2}{\bar{\sigma}^2 + n\tau^2}} \exp\left\{\frac{n^2\tau^2\bar{x}_n^2}{2\bar{\sigma}^2\left(\bar{\sigma}^2 + n\tau^2\right)}\right\}$$

Jeffreys-Lindley Paradox

- For the dataset normaldata, the range of the Bayes factor is shown as τ goes from 10^{-4} to 10 (in a log scale)

$$B_{10}^{\pi} = \sqrt{\frac{\bar{\sigma}^{-2}}{\bar{\sigma}^{-2} + n\tau^2}} \exp\left\{ \frac{n^2\tau^2 \bar{x}_n^{-2}}{2\bar{\sigma}^{-2}\left(\bar{\sigma}^{-2} + n\tau^2\right)} \right\}, \; where: \bar{x}_n = \frac{\sum_{i=1}^{n} x_i}{n}$$



See MatLab implementation

- The results vary substantially as τ increases from 0 to ∞.

Banning Improper Priors

- Impossibility of using improper priors for testing!

  Reason: When using the representation

$$\pi(\theta) = \mathrm{P}^{\pi}(\theta \in \Theta_1) \times \pi_1(\theta) + \mathrm{P}^{\pi}(\theta \in \Theta_0) \times \pi_0(\theta)$$

  π_1 and π_0 must be normalized

Non-informative prior and Limit of Conjugate Prior

- Let $x \sim \mathcal{N}(\theta, 1)$ and $H_0 : \theta = 0$, we consider the improper Jeffreys prior $p_1(\theta) = 1$, then the prior is transformed as

$$\pi(\theta) = \frac{1}{2}\mathbb{I}_0(\theta) + \frac{1}{2}\mathbb{I}_{\theta \neq 0}$$

and

$$\pi(\theta = 0|x) = \frac{e^{-x^2/2}}{e^{-x^2/2} + \int_{-\infty}^{+\infty} e^{-(x-\theta)^2/2} d\theta} = \frac{1}{1 + \sqrt{2\pi} e^{x^2/2}}$$

Consequence: H_0 is bounded from above by

$$\pi(\theta = 0|x) \leq 1 \Big/ \left(1 + \sqrt{2\pi}\right) = 0.285$$

x	0.0	1.0	1.65	1.96	2.58
π(θ = 0|x)	0.285	0.195	0.089	0.055	0.014

This is in agreement with the classical p –value

Jeffreys-Lindley Paradox

- Limiting arguments not valid in testing settings
- Under a conjugate prior

$$\pi\left(\theta=0\middle|x\right)=\left[1+\frac{1-\rho}{\rho}\sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}}\exp\left\{\frac{\tau^2x^2}{2\sigma^2\left(\sigma^2+\tau^2\right)}\right\}\right]^{-1}$$

  which converges to 1 when τ goes to $+\infty$, for every x

- Difference with the noninformative answer

$$\pi\left(\theta=0\middle|x\right)=\left[1+\sqrt{2\pi}\exp(x^2/2)\right]^{-1}$$

- The noninformative prior no longer corresponds to the limit of conjugate prior!