STA 414/2104

Statistical Methods for Machine Learning and Data Mining

Radford M. Neal, University of Toronto, 2013

Week 4

Analytically-Tractable Bayesian Models

Conjugate Prior Distributions

For most Bayesian inference problems, the integrals needed to do inference and prediction are not analytically tractable — hence the need for numerical quadrature, Monte Carlo methods, or various approximations.

Most of the exceptions involve *conjugate priors*, which combine nicely with the likelihood to give a posterior distribution of the same form. Examples:

1) Independent observations from a finite set, with Beta / Dirichlet priors.

2) Independent observations of Gaussian variables with Gaussian prior for the mean, and either known variance or inverse-Gamma prior for the variance.

3) Linear regression with Gaussian prior for the regression coefficients, and Gaussian noise, with known variance or inverse-Gamma prior for the variance.

It's nice when a tractable model and prior are appropriate for the problem. Unfortunately, people are tempted to use such models and priors even when they aren't appropriate.

Independent Binary Observations with Beta Prior

We observe binary (0/1) variables Y_1, Y_2, ..., Y_n.

We model these as being *independent*, and *identically distributed*, with

$$P(Y_i = y \,|\, \theta) \;\; = \;\; \left\{ \begin{array}{ll} \theta & \text{if } y = 1 \\ 1-\theta & \text{if } y = 0 \end{array} \right\} \;\; = \;\; \theta^y \, (1-\theta)^{1-y}$$

Let's suppose that our prior distribution for θ is Beta(a,b), with a and b being known positive reals. With this prior, the probability density over $(0, 1)$ of θ is:

$$P(\theta) \;\; = \;\; \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \, \theta^{a-1} (1-\theta)^{b-1}$$

Here, the Gamma function, $\Gamma(c)$, is defined to be $\int_0^\infty x^{c-1} \exp(-x) \, dx$. Note that $\Gamma(c) = (c-1)!$ when c is an integer.

When $a = b = 1$ the prior is uniform over $(0, 1)$.

The prior mean of θ is $a \,/\, (a+b)$. Big a and b give smaller prior variance.

Posterior Distribution with Beta Prior

With this Beta prior, the posterior distribution is also Beta:

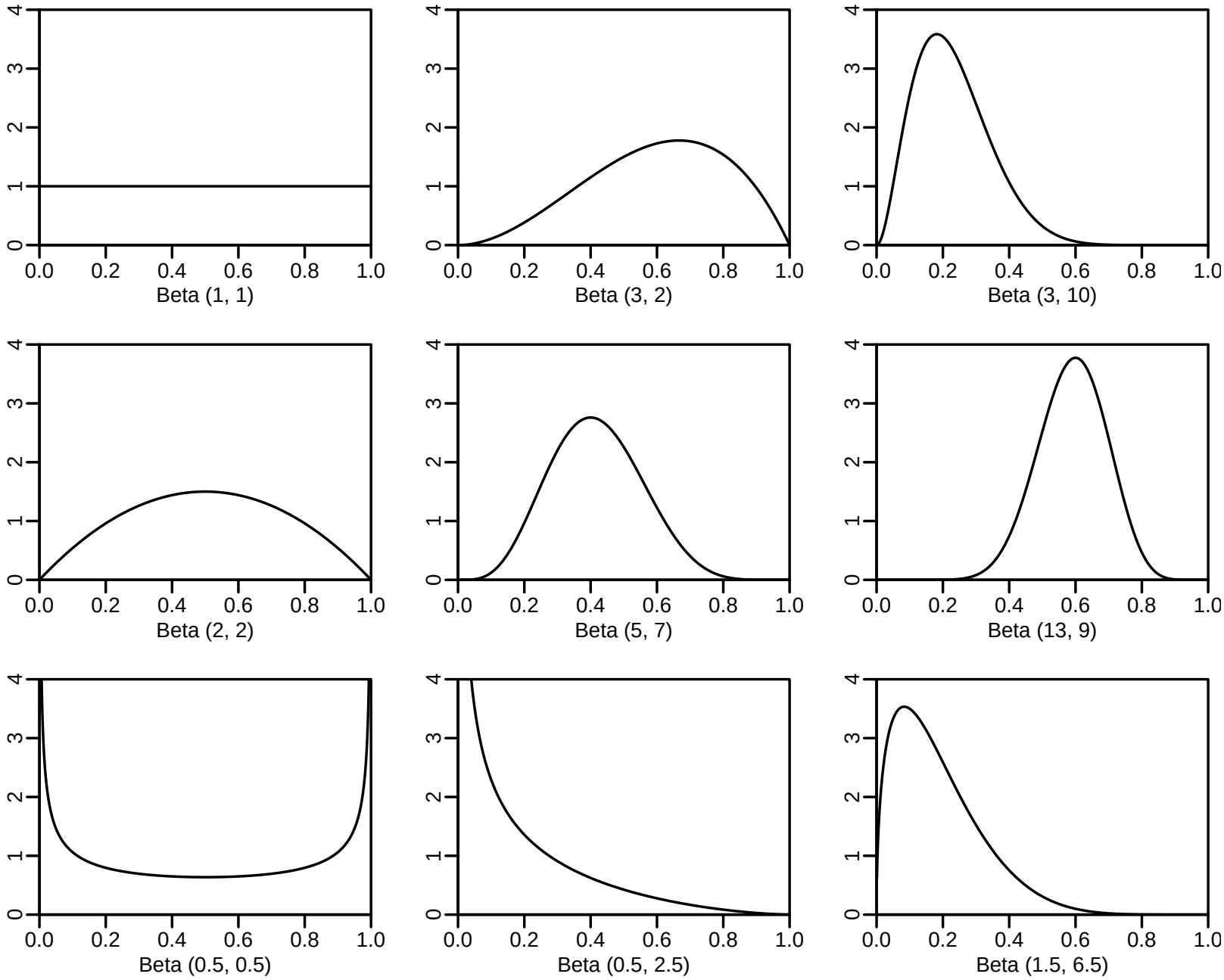
$$\begin{aligned} P(\theta \,|\, Y_1 = y_1,\, Y_2 = y_2,\, \ldots,\, Y_n = y_n) \\ &\propto \quad P(\theta) \prod_{i=1}^{n} P(Y_i = y_i \,|\, \theta) \\ &\propto \quad \theta^{a-1}\,(1-\theta)^{b-1} \prod_{i=1}^{n} \theta^{y_i}\,(1-\theta)^{1-y_i} \\ &\propto \quad \theta^{\Sigma y_i + a - 1}\,(1-\theta)^{n - \Sigma y_i + b - 1} \end{aligned}$$

So the posterior distribution is Beta $(\sum y_i + a,\, n - \sum y_i + b)$.

One way this is sometimes visualized is as the prior being equivalent to a fictitious observations with $Y = 1$ and b fictitious observations with $Y = 0$.

Note that all that is used from the data is $\sum y_i$, which is a *minimal sufficient statistic*, whose values are in one-to-one correspondence with possible likelihood functions (ignoring constant factors).

Examples of Beta Priors and Posteriors



Predictive Distribution from Beta Posterior

From the Beta $(\sum y_i + a,\, n - \sum y_i + b)$ posterior distribution, we can make a probabilistic prediction for the next observation:

$$
\begin{aligned}
& P(Y_{n+1} = 1 \,|\, Y_1 = y_1,\, Y_2 = y_2,\, \ldots,\, Y_n = y_n) \\
&\quad = \int_0^1 P(Y_{n+1} = 1 \,|\, \theta)\, P(\theta \,|\, Y_1 = y_1,\, Y_2 = y_2,\, \ldots,\, Y_n = y_n)\, d\theta \\
&\quad = \int_0^1 \theta\, P(\theta \,|\, Y_1 = y_1,\, Y_2 = y_2,\, \ldots,\, Y_n = y_n)\, d\theta \\
&\quad = \int_0^1 \theta\, \frac{\Gamma(n+a+b)}{\Gamma(\Sigma y_i + a)\Gamma(n - \Sigma y_i + b)}\, \theta^{\Sigma y_i + a - 1}\, (1-\theta)^{n - \Sigma y_i + b - 1}\, d\theta \\
&\quad = \frac{\Gamma(n+a+b)}{\Gamma(\Sigma y_i + a)\Gamma(n - \Sigma y_i + b)}\, \frac{\Gamma(1 + \Sigma y_i + a)\Gamma(n - \Sigma y_i + b)}{\Gamma(1 + n + a + b)} \\
&\quad = \frac{\sum y_i + a}{n + a + b}
\end{aligned}
$$

This uses the fact that $c\Gamma(c) = \Gamma(1 + c)$.

Generalizing to More Than Two Values

For i.i.d. observations with a finite number, K, of possible values, with $K > 2$, the conjugate prior for the probabilities $\theta_1, \ldots, \theta_K$ is the Dirichlet distribution, with the following density on the simplex where all $\theta_k > 0$ and $\sum \theta_k = 1$:

$$P(\theta_1, \ldots, \theta_K) = \frac{\Gamma(\Sigma_k \alpha_k)}{\Pi_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

The parameters $\alpha_1, \ldots, \alpha_K$ can be any positive reals.

The posterior distribution after observing n items, with m_1 having value 1, m_2 having value 2, etc. is Dirichlet with parameters $\alpha_1 + m_1, \ldots, \alpha_K + m_K$.

The predictive distribution for item $n + 1$ is

$$P(Y_{n+1} = k \,|\, Y_1 = y_1, \ldots, Y_n = y_n) = \frac{m_k + \alpha_k}{n + \Sigma \alpha_k}$$

Independent Observations from a Gaussian Distribution

We observe real variables Y_1, Y_2, ..., Y_n.

We model these as being independent, all from some Gaussian distribution with unknown mean, μ, and known variance, σ^2.

The conjugate prior for μ is Gaussian with some mean μ_0 and variance σ_0^2.

Rather than talk about the variance, it is more convenient to talk about the *precision*, equal to the reciprocal of the variance. A data point has precision $\tau = 1/\sigma^2$ and the prior has precision $\tau_0 = 1/\sigma_0^2$.

The posterior distribution for μ is also Gaussian, with precision $\tau_n = \tau_0 + n\tau$, and with mean

$$\mu_n = \frac{\tau_0 \mu_0 + n\tau \overline{y}}{\tau_0 + n\tau}$$

where $\overline{y}$ is the sample mean of the observatons $y_1, \ldots, y_n$.

The predictive distribution for Y_{n+1} is Gaussian with mean μ_n and variance $(1/\tau_n) + \sigma^2$.

If we let σ_0 go to infinity — an example of an *improper* prior — the posterior mean, μ_n, will equal the sample mean, $\overline{y}$.

Gaussian with Unknown Variance

What if both the mean and the variance (precision) of the Gaussian distribution for $Y_1, \ldots, Y_n$ are unknown?

There is still a conjugate prior, but in it, μ and τ are dependent:

$$\begin{aligned} \tau &\sim \text{Gamma}(a, b) \\ \mu \,|\, \tau &\sim N(\mu_0, c/\tau) \end{aligned}$$

for some positive constants a, b, and c.

It's hard to imagine circumstances where our prior information about μ and τ would have a dependence of this sort. But unfortunately, people use this conjugate prior anyway, because it's convenient.

Bayesian Linear Basis Function Models

A Bayesian Linear Basis Function Model

Let's set up a Bayesian linear basis function model by giving β a Gaussian prior:

$$\begin{aligned} y_i \,|\, x_i,\, \beta &\sim N(\phi(x_i)^T \beta,\, \sigma^2) \\ \beta &\sim N(m_0, S_0) \end{aligned}$$

This Gaussian prior will turn out to be conjugate.

For the moment, we regard σ^2, m_0, and S_0 as known.

Often, we will let $m_0 = 0$ and let S_0 be diagonal, so that the β_j are independent. We might let β_0 have a large variance, and all the other β_j have the same variance.

The symbol y will sometime denote a single, generic response value, and other times denote the vector $[y_1, \ldots, y_n]^T$ of responses for training cases. We use Φ for the matrix of basis function values for the n training cases.

Multivariate Gaussian Model with Multivariate Gaussian Prior

To warm up... Suppose we model an observed vector b as having a multivariate Gaussian distribution with known covariance matrix B and unknown mean x. We give x a multivariate Gaussian prior with known covariance matrix A and known mean a.

The posterior distribution of x will be Gaussian, since the product of the prior density and the likelihood is proportional to the exponential of a quadratic function of x:

$$\text{Prior} \times \text{Likelihood} \quad \propto \quad \exp(-(x-a)^T A^{-1}(x-a)/2)\, \exp(-(b-x)^T B^{-1}(b-x)/2)$$

The log posterior density is this quadratic function ($\cdots$ is parts not involving x):

$$\begin{aligned} -\tfrac{1}{2}\Big[(x-a)^T A^{-1}(x-a) \;+\; (b-x)^T B^{-1}(b-x)\Big] \;+\; \cdots & \\ = \quad -\tfrac{1}{2}\Big[x^T(A^{-1}+B^{-1})x \;-\; 2x^T(A^{-1}a + B^{-1}b)\Big] \;+\; \cdots & \\ = \quad -\tfrac{1}{2}\Big[(x-c)^T(A^{-1}+B^{-1})(x-c)\Big] \;+\; \cdots & \end{aligned}$$

where $c = (A^{-1}+B^{-1})^{-1}(A^{-1}a + B^{-1}b)$. This is the density for a Gaussian distribution with mean c and variance $(A^{-1}+B^{-1})^{-1}$.

Posterior for Linear Basis Function Model

Both the log prior and the log likelihood are quadratic functions of β. The log likelihood for β is

$$-\tfrac{1}{2}\Big[(y-\Phi\beta)^T(\sigma^2 I)^{-1}(y-\Phi\beta)\Big] \;+\; \cdots \;\;=\;\; -\tfrac{1}{2}\tfrac{1}{\sigma^2}\Big[\beta^T\Phi^T\Phi\beta \;-\; 2\beta^T\Phi^T y\Big] \;+\; \cdots$$

which is the same quadratic function of β as for a Gaussian log density with covariance $\sigma^2(\Phi^T\Phi)^{-1}$ and mean $(\Phi^T\Phi)^{-1}\Phi^T y$.

This combines with the prior for β in the same way on the previous slide, with the result that the posterior distribution for β is Gaussian with covariance

$$S_n \;\;=\;\; \Big[S_0^{-1} \,+\, (\sigma^2(\Phi^T\Phi)^{-1})^{-1}\Big]^{-1} \;\;=\;\; \Big[S_0^{-1} \,+\, (1/\sigma^2)\Phi^T\Phi\Big]^{-1}$$

and mean

$$\begin{aligned} m_n &= (S_n^{-1})^{-1}\Big[S_0^{-1}m_0 \,+\, (1/\sigma^2)\Phi^T\Phi(\Phi^T\Phi)^{-1}\Phi^T y\Big] \\ &= S_n\Big[S_0^{-1}m_0 \,+\, (1/\sigma^2)\Phi^T y\Big] \end{aligned}$$

Predictive Distribution for a Test Case

We can write the response, y, for some new case with inputs x as

$$y \;=\; \phi(x)^T\beta + e$$

where the "noise" e has the $N(0, \sigma^2)$ distribution, independently of β.

Since the posterior distribution for β is $N(m_n,\, S_n)$, the posterior distribution for $\phi(x)^T\beta$ will be $N(\phi(x)^T m_n,\, \phi(x)^T S_n \phi(x))$.

Hence the predictive distribution for y will be $N(\phi(x)^T m_n,\, \phi(x)^T S_n \phi(x) + \sigma^2)$.

Comparison with Regularized Estimates

In a Bayesian linear basis function model, the predictive mean for a test case is what we would get using the posterior mean value for the regression coefficients — a consequence of the model being linear in the parameters.

We can compare the Bayesian mean prediction with the prediction using the regularized (maximum penalized likelihood) estimate for β, which is

$$\hat{\beta} \;=\; (\lambda I^* + \Phi^T\Phi)^{-1}\Phi^T y$$

where I^* is like the identity matrix except that $I^*_{1,1} = 0$.

Compare with the posterior mean, if we set the prior mean, m_0, to zero:

$$\begin{aligned} m_n &= S_n(1/\sigma^2)\Phi^T y \\ &= (S_0^{-1} + (1/\sigma^2)\Phi^T\Phi)^{-1}(1/\sigma^2)\Phi^T y \\ &= (\sigma^2 S_0^{-1} + \Phi^T\Phi)^{-1}\Phi^T y \end{aligned}$$

If $S_0^{-1} = (1/\omega^2)I^*$, then these are the same, with $\lambda = \sigma^2/\omega^2$. This corresponds to a prior for β in which the β_j are independent, all with variance ω^2, except that β_0 has an infinite variance.

A Semi-Bayesian Way to Estimate σ^2 and ω^2

We see that σ^2 (the noise variance) and ω^2 (the variance of regression coefficients, other than β_0) together (as σ^2/ω^2) play a role similar to the penalty magnitude, λ, in the maximum penalized likelihood approach.

We can find values for σ^2 and ω^2 in a semi-Bayesian way by maximizing the *marginal likelihood* — the probability of the data (y) given values for σ^2 and ω^2. [We need to set the prior variance of β_0 to some finite ω_0^2 (which could be very large), else the probability of the observed data will be zero.]

We can also select basis function parameters (eg, s) by maximizing the marginal likelihood.

Such maximization is somewhat easier than the full Bayesian approach, in which we define some prior distribution for σ^2 and ω^2 (and any basis function parameters we haven't fixed), and then average predictions over their posterior distribution. [One would probably use some Markov chain Monte Carlo (MCMC) method to do this averaging.]

Finding the Marginal Likelihood for σ^2 and ω^2

The marginal likelihood for σ^2 and ω^2 given a vector of observed responses, y, is found by integrating over β with respect to its prior:

$$P(y \mid \sigma^2, \omega^2) \;=\; \int P(y \mid \beta, \sigma^2)\, P(\beta \mid \omega^2)\, d\beta$$

This is the denominator in Bayes' Rule, that normalizes the posterior.

Here, the basis function values for the training cases, based on the inputs for those cases, are considered fixed.

Both factors in this integrand are exponentials of quadratic functions of β, so this turns into the same sort of integral as that for the normalizing constant of a Gaussian density function, for which we know the answer.

Details of Computing the Marginal Likelihood

We go back to the computation of the posterior for β, but we now need to pay attention to some factors we ignored before. I'll fix the prior mean of β to $m_0 = 0$.

The log of the probability density of the data is

$$-\frac{n}{2}\log(2\pi) \;-\; \frac{n}{2}\log(\sigma^2) \;-\; \frac{1}{2}(y-\Phi\beta)^T(y-\Phi\beta)/\sigma^2$$

The log prior density for β is

$$-\frac{m}{2}\log(2\pi) \;-\; \frac{1}{2}\log(|S_0|) \;-\; \frac{1}{2}\beta^T S_0^{-1}\beta$$

expanding and then adding these together, we see the following terms that don't involve β:

$$-\frac{n+m}{2}\log(2\pi) \;-\; \frac{n}{2}\log(\sigma^2) \;-\; \frac{1}{2}\log(|S_0|) \;-\; \frac{1}{2}y^T y/\sigma^2$$

and these terms that do involve β:

$$-\frac{1}{2}\beta^T\Phi^T\Phi\beta/\sigma^2 \;+\; \beta^T\Phi^T y/\sigma^2 \;-\; \frac{1}{2}\beta^T S_0^{-1}\beta$$

More Details...

We can combine the quadratic terms that involve β, giving

$$-\frac{1}{2}\Big[\beta^T(S_0^{-1} + \Phi^T\Phi/\sigma^2)\beta \;-\; 2\beta^T\Phi^T y/\sigma^2\Big]$$

We had previously used this to identify the posterior covariance and mean for β. Setting the prior mean to zero, these are

$$S_n \;=\; \Big[S_0^{-1} \,+\, (1/\sigma^2)\Phi^T\Phi\Big]^{-1}, \qquad m_n \;=\; S_n\Phi^T y/\sigma^2$$

We can write the terms involving β using these, then "complete the square":

$$\begin{aligned}
&-\frac{1}{2}\Big[\beta^T S_n^{-1}\beta \;-\; 2\beta^T S_n^{-1} m_n\Big] \\
&\quad= \quad -\frac{1}{2}\Big[\beta^T S_n^{-1}\beta \;-\; 2\beta^T S_n^{-1} m_n \;+\; m_n^T S_n^{-1} m_n\Big] \;+\; \frac{1}{2} m_n^T S_n^{-1} m_n \\
&\quad= \quad -\frac{1}{2}(\beta - m_n)^T S_n^{-1}(\beta - m_n) \;+\; \frac{1}{2} m_n^T S_n^{-1} m_n
\end{aligned}$$

The second term above doesn't involve β, so we can put it with the other such.

And Yet More Details...

We now see that the log of the prior times the probability of the data has these terms not involving β:

$$-\frac{n+m}{2}\log(2\pi) \;-\; \frac{n}{2}\log(\sigma^2) \;-\; \frac{1}{2}\log(|S_0|) \;-\; \frac{1}{2}y^T y/\sigma^2 \;+\; \frac{1}{2}m_n^T S_n^{-1} m_n$$

and this term that does involve β:

$$-\frac{1}{2}(\beta - m_n)^T S_n^{-1}(\beta - m_n)$$

When we exponentiate this and then integrate over β, we see that

$$\int \exp\Big(-\frac{1}{2}(\beta - m_n)^T S_n^{-1}(\beta - m_n)\Big)\, d\beta \;\;=\;\; (2\pi)^{m/2}\,|S_n|^{1/2}$$

since this is just the integral defining the Gaussian normalizing constant.

The final result is that the log of the marginal likelihood is

$$-\frac{n}{2}\log(2\pi) \;-\; \frac{n}{2}\log(\sigma^2) \;-\; \frac{1}{2}\log\Big(\frac{|S_0|}{|S_n|}\Big) \;-\; \frac{1}{2}y^T y/\sigma^2 \;+\; \frac{1}{2}m_n^T S_n^{-1} m_n$$

Another Formula for the Marginal Likelihood

The last two terms in the formula on the previous slide seem a bit mysterious.

They can be rewritten as follows:

$$
\begin{aligned}
&-\frac{1}{2}y^T y/\sigma^2 \;+\; \frac{1}{2}m_n^T S_n^{-1} m_n \\
&\quad = \;\; -\frac{1}{2}y^T y/\sigma^2 \;+\; m_n^T S_n^{-1} m_n \;-\; \frac{1}{2}m_n^T S_n^{-1} m_n \\
&\quad = \;\; -\frac{1}{2}y^T y/\sigma^2 \;+\; m_n^T \Phi^T y/\sigma^2 \;-\; \frac{1}{2}m_n^T \Phi^T \Phi m_n/\sigma^2 \;-\; \frac{1}{2}m_n^T S_0^{-1} m_n \\
&\quad = \;\; -\frac{1}{2}||y - \Phi m_n)||^2/\sigma^2 \;-\; \frac{1}{2}m_n^T S_0^{-1} m_n
\end{aligned}
$$

This gives another formula for the log marginal likelihood, which is more intuitive and also better numerically (avoids large roundoff in computing $y^T y$):

$$
-\frac{n}{2}\log(2\pi) \;-\; \frac{n}{2}\log(\sigma^2) \;-\; \frac{1}{2}\log\left(\frac{|S_0|}{|S_n|}\right) \;-\; \frac{1}{2}||y - \Phi m_n)||^2/\sigma^2 \;-\; \frac{1}{2}m_n^T S_0^{-1} m_n
$$

Here, $(1/2)\log(|S_0|/|S_n|)$ is the log of the factor by which the prior contracts to the posterior, the next term is the data fit with the posterior mean, and the last term is the prior density at the posterior mean.

Computations for the Semi-Bayesian Approach

Maximizing the marginal likelihood with respect to σ^2, ω^2, and parameters of the basis functions could be done by many standard optimization methods.

For maximizing with respect to σ^2 and ω^2, there's also an iterative re-estimation procedure (see the next slide).

We can then use the posterior mean, m_n, to predict the response in a test case with inputs x, as $\phi(x)^T m_n$. The posterior covariance, S_n, is used in producing a predictive variance for the response, which is $\phi(x)^T S_n \phi(x) + \sigma^2$.

Note that these semi-Bayesian predictions are all based on a *single* set of values for σ^2, ω^2, etc., although they do integrate over β.

Re-estimating σ^2 and ω^2

Naively, one might iterate finding the posterior mean and covariance of β, based on the current estimates for σ^2 and ω^2, with the following re-estimation of σ^2 and ω^2:

$$\widehat{\sigma}^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(y_i - \Phi(x_i)^T m_n)^2, \qquad \widehat{\omega}^2 \;=\; \frac{1}{m}\sum_{j=0}^{m-1}[m_n]_j$$

This assumes $S_0 = \omega^2 I$. But this isn't quite right: consider that some data points could be fitted nearly exactly when the model is flexible, and some coefficients in m_n could be nearly zero if they aren't relevant to any data point.

Instead, we find the "effective number of parameters", γ, as

$$\gamma \;=\; \sum_{j=1}^{m}\frac{\lambda_i}{\lambda_i + 1/\omega^2}$$

where the λ_i are the eigenvalues of $\Phi^T\Phi/\sigma^2$, and then re-estimate as follows:

$$\widehat{\sigma}^2 \;=\; \frac{1}{n-\gamma}\sum_{i=1}^{n}(y_i - \Phi(x_i)^T m_n)^2, \qquad \widehat{\omega}^2 \;=\; \frac{1}{\gamma}\sum_{j=0}^{m-1}([m_n]_j)^2$$

Details are in David MacKay's thesis (Section 2.4).

Computations for the Fully-Bayesian Approach

The full Bayesian approach is to integrate over the posterior distribution for σ, ω, etc. as well as β, which can be done by MCMC methods, using the marginal likelihood for σ, ω, etc. (integrating over β).

We then make a prediction for the response in a test case by averaging the posterior mean for β based on a sample of values for σ, ω, etc. The standard deviation for the unknown response can be found as well. We could also approximate the whole predictive distribution, which in general is not Gaussian.

Alternatively, we can sample for β as well as σ, ω, etc. This avoids any expensive matrix computations, but fails to take advantage of conjugacy. We'd need to do this if we used a non-conjugate prior for β.

Note: We can't use an improper prior for ω that gives infinite mass to $\omega \to 0$, since $\omega = 0$ gives only finite misfit to the data. Similarly, if ϕ allows the data to be fit exactly, we may not be able to use an improper prior for σ with infinite mass at zero.

How Feasible are Linear Basis Function Models?

Modeling a general non-linear relationship of y to x with a linear basis function model seems attractive when x is of low dimension, but when there are many inputs, we would seem to need a huge number of local basis functions to "cover" the high dimensional input space. This is at least a computational problem.

One possibility is to use a relatively small number of basis functions, that cover only the actual area where x values are found, which may be the vicinity of a manifold of much lower dimension. We might:

– pick a subset of data points as centres for basis functions

– make the basis functions depend on parameters that adapt to the data.

A neural network with one hidden layer is an example of the latter approach.

Instead, we might go ahead and use a huge number of basis functions, maybe an infinite number. We'll later see that there's a computational trick that allows this.