Chapter 10

Bayesian Methods

10.1 Introduction

See [33] or Box and Tiao [6] for a general introduction to Bayesian statistics and [43] for applications of Bayesian methods in signal processing.

10.2 Bayes Rule

The distribution of a variable x conditioned on a variable y is

$$p(x \mid y) = \frac{p(x, y)}{p(y)} \tag{10.1}$$

Given that $p(y \mid x)$ can be expressed similarly we can write

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)} \tag{10.2}$$

which is Baye's rule. The density $p(x)$ is known as the *prior*, $p(y \mid x)$ as the *likelihood* and $p(y)$ as the *evidence* or *marginal likelihood*. Baye's rule shows how a prior distribution can be turned into a posterior distribution ie. how we update our distribution in the light of new information. To do this it is necessary to calculate the normalising term; the evidence

$$p(y) = \int p(y \mid x)p(x)dx \tag{10.3}$$

which, being an integral, can sometimes be problematic to evaluate.

10.2.1 Example

For discrete variables. Given a disease D with a prevalence of ten percent, a test for it T having a sensitivity of 95% and a specificity of 85% we have

$$
\begin{aligned}
p(D=1) &= 0.1 && (10.4)\\
p(T=1|D=1) &= 0.95 && (10.5)\\
p(T=0|D=0) &= 0.85 && (10.6)
\end{aligned}
$$

The probability that subjects who test positive for D actually have D is then given by Bayes' rule

$$
\begin{aligned}
p(D=1|T=1) &= \frac{p(T=1|D=1)p(D=1)}{p(T=1|D=1)p(D=1)+p(T=1|D=0)p(D=0)} && (10.7)\\
&= \frac{0.95\times 0.1}{0.95\times 0.1+0.15\times 0.9} && (10.8)\\
&= 0.413 && (10.9)
\end{aligned}
$$

10.3 Gaussian Variables

A Gaussian random variable x has the probability density function (PDF)

$$p(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] \quad (10.10)$$

where the mean is μ and the variance is σ^2. The inverse of the variance is known as the precision $\beta=1/\sigma^2$. The Gaussian PDF is written in shorthand as

$$p(x)=N(x;\mu,\sigma^2) \quad (10.11)$$

If the prior is Gaussian

$$p(x)=N(x;x_0,1/\beta_0) \quad (10.12)$$

where x_0 is the prior mean and β_0 is the prior precision and the likelihood is also Gaussian

$$p(y\mid x)=N(y;x,1/\beta_D) \quad (10.13)$$

where the variable x is the mean of the likelihood and β_D is the data precision then the posterior distribution is also Gaussian (see eg. [33],page 37).

$$p(x\mid y)=N(x;m,1/\beta) \quad (10.14)$$

where the mean and precision are given by

$$\beta=\beta_0+\beta_D \quad (10.15)$$

and

$$m=\frac{\beta_0}{\beta}x_0+\frac{\beta_D}{\beta}y \quad (10.16)$$

Thus, the posterior precision is given by the sum of the prior precision and the data precision and the posterior mean is given by the sum of the prior data mean and the new data value each weighted by their relative precisions [1].

[1] This is the same as *inverse variance* weighting where the weights sum to one.

10.3.1 Combining Estimates

This type of updating is relevant to the *sensor fusion* problem, where we have information about a variable from two different sources and we wish to combine that information.

Say, for example, we had two estimates for the amount of carbon in a given compound; method 1 estimates the percentage to be 35 ± 4 units and method 2 estimates it to be 40 ± 7 units. Before observing the second result we have a prior belief that the mean percentage is $x_0 = 35$ and the variance is $4^2 = 16$ which corresponds to a precision of $\beta_0 = 0.0625$. Whilst the first result is viewed as the prior, the second result is viewed as the 'data', which has mean $y = 40$ and precision $\beta_D = 1/7^2 = 0.0204$. Our posterior estimate for the amount of carbon is then estimated as

$$m = \frac{0.0625}{0.0829} \times 35 + \frac{0.0204}{0.0829} \times 40 = 36.2 \tag{10.17}$$

and the posterior standard deviation is 3.5. If the results of method 2 were chosen as the prior (instead of method 1) we'd get the same result.

The equation for the posterior mean can be re-arranged as

$$m = x_0 + \frac{\beta_D}{\beta}(y - x_0) \tag{10.18}$$

showing that the new estimate is the old estimate plus some fraction (which may be viewed as a learning rate) of an error term $e = y - x_0$.

10.3.2 Sequential Estimation

Also, this type of update is particularly suited to *sequential* estimation, where data comes in a sample at a time and we update our estimates at each time step. Baye's rule is perfect for this because today's posterior becomes tomorrow's prior.

Say, for example, we have a random variable x which we observe sequentially - the value at time t being x_t ie. a time series - and that we wish to estimate the mean, without storing all the data points. At time t our estimate for the mean is μ_t and our estimate for the variance is σ_t^2. Now our prior distribution for μ_t (ie. prior to observing x_t) is

$$p(\mu_t) = N(\mu_t; \mu_t, \sigma_t^2/t) \tag{10.19}$$

where the variance is given by the usual standard error formula (see lecture 1). The likelihood of the new data point is

$$p(x_t|\mu_t) = N(x_t; \mu_t, \sigma_t^2) \tag{10.20}$$

Adding the precisions to get the posterior precision gives (from equation 10.15)

$$\beta = \frac{t}{\sigma_t^2} + \frac{1}{\sigma_t^2} = \frac{t+1}{\sigma_t^2} \tag{10.21}$$

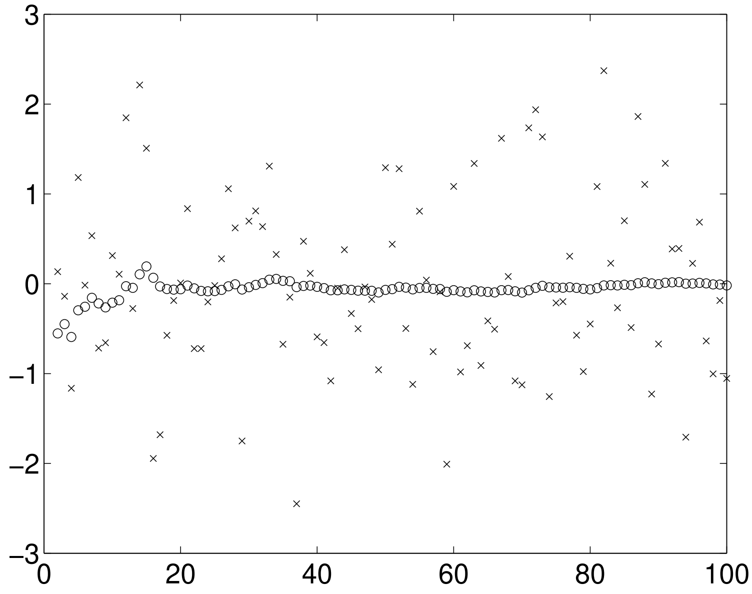

Figure 10.1: **Sequential estimation of stationary mean.** The graph plots data values x_t (crosses) and the estimated mean value μ_t (circles) versus iteration number t.

The posterior mean is then given by equation 10.16

$$\mu_{t+1} = \frac{t}{t+1}\mu_t + \frac{1}{t+1}x_t \tag{10.22}$$

Re-arranging gives

$$\mu_{t+1} = \mu_t + \frac{1}{t+1}(x_t - \mu_t) \tag{10.23}$$

In the above procedure we have implicitly assumed that the data x_t is *stationary* ie. that the mean at time t is equal to the mean at time $t+T$ for all T (a more formal definition of stationarity will be given later). This results in our estimate for the mean converging to a steady value as t increases. The final value is exactly the same as if we'd stored all the data and calculated it in the usual way.

But what if the signal is non-stationary ? See the chapter on Kalman filters.

10.4 Multiple Gaussian Variables

A d-dimensional Gaussian random vector $\boldsymbol{x}$ has a PDF given by

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{C}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\bar{\boldsymbol{x}})^T C^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}})\right) \tag{10.24}$$

where the mean $\bar{\boldsymbol{x}}$ is a d-dimensional vector, $\boldsymbol{C}$ is a $d \times d$ covariance matrix, and $|\boldsymbol{C}|$ denotes the determinant of $\boldsymbol{C}$. The multivariate Gaussian PDF is written in shorthand as

$$p(\boldsymbol{x}) = N(\boldsymbol{x}; \bar{\boldsymbol{x}}, \boldsymbol{C}) \tag{10.25}$$

If the prior distribution is Gaussian

$$p(\boldsymbol{x}) = N(\boldsymbol{x}; \boldsymbol{x}_0, \Sigma_0) \tag{10.26}$$

where $\boldsymbol{x}_0$ is the prior mean and Σ_0 is the prior covariance and the likelihood is

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = N(\boldsymbol{y}; \boldsymbol{x}, \Sigma_D) \tag{10.27}$$

where the variable $\boldsymbol{x}$ is the mean of the likelihood and Σ_D is the data covariance then the posterior distribution is given by [40]

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = N(\boldsymbol{x}; \boldsymbol{m}, \boldsymbol{\Sigma}) \tag{10.28}$$

where the mean and covariance are given by

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_D^{-1} \tag{10.29}$$

and

$$\boldsymbol{m} = \boldsymbol{\Sigma}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{x}_0 + \boldsymbol{\Sigma}\boldsymbol{\Sigma}_D^{-1}\boldsymbol{y} \tag{10.30}$$

These updates are similar in form to the updates for the univariate case. Again, these update formulae are useful for both sequential estimation and sensor fusion. In the sequential estimation case we have a *Kalman filter* (see next lecture).

10.5 General Linear Models

Given a set of input variables $\boldsymbol{z}_n$ (a row vector) where $n = 1..N$ and a fixed, possibly nonlinear, function of them

$$\boldsymbol{x}_n = F(\boldsymbol{z}_n) \tag{10.31}$$

the output variable is then given as a linear combination

$$y_n = \boldsymbol{x}_n \boldsymbol{w} + e_n \tag{10.32}$$

where $\boldsymbol{w}$ is a column vector of coefficients and e is zero mean Gaussian noise with precision β. This type of model is sufficiently general to include (i) autoregressive models if F is the identity function and $\boldsymbol{x}_n = [y_{n-1}, y_{n-2}, ..., y_{n-p}]$, (ii) Fourier-type models if F are sine and cosine functions and (iii) wavelet models if F are the wavelet bases.

Given a data set $D = \{\boldsymbol{z}_n, y_n\}$ where $n = 1..N$ the likelihood of the data is given by

$$p(D \mid \boldsymbol{w}, \beta) \quad = \quad \left(\frac{\beta}{2\pi}\right)^{N/2} \exp(-\beta E_D) \tag{10.33}$$

where

$$E_D = \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{w}) \tag{10.34}$$

and $\boldsymbol{Y}$ is a column vector with entries y_n and the nth row of the matrix $\boldsymbol{X}$ contains $\boldsymbol{x}_n$. The weights are drawn from a zero-mean Gaussian prior with an isotropic covariance having precision α

$$p(\boldsymbol{w} \mid \alpha) = \left(\frac{\alpha}{2\pi}\right)^{p/2} \exp(-\alpha E_W) \tag{10.35}$$

where

$$\begin{aligned} E_W &= \frac{1}{2}\sum_{i=1}^{p} w_i^2 \\ &= \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} \end{aligned} \tag{10.36}$$

The posterior distribution over the unknown coefficients is then given by Bayes' rule

$$p(\boldsymbol{w}|D,\alpha,\beta) = \frac{p(D|\boldsymbol{w},\beta)p(\boldsymbol{w}|\alpha)}{\int p(D|\boldsymbol{w},\beta)p(\boldsymbol{w}|\alpha)d\boldsymbol{w}} \tag{10.37}$$

As the prior is normal with mean $\boldsymbol{w}_0 = \boldsymbol{0}$ and covariance $\boldsymbol{\Sigma}_0 = (1/\alpha)\boldsymbol{I}$, the likelihood is normal with mean $\boldsymbol{w}_D = \boldsymbol{X}^{-1}\boldsymbol{Y}$ and covariance $\boldsymbol{\Sigma}_D = (\beta \boldsymbol{X}^T\boldsymbol{X})^{-1}$ then the posterior is also a normal with mean and covariance given by equations 10.30 and 10.29. The posterior is therefore given by

$$p(\boldsymbol{w}|D,\alpha,\beta) = N(\boldsymbol{w};\hat{\boldsymbol{w}},\hat{\boldsymbol{\Sigma}}) \tag{10.38}$$

where

$$\begin{aligned} \hat{\boldsymbol{\Sigma}} &= (\beta\boldsymbol{X}^T\boldsymbol{X} + \alpha\boldsymbol{I})^{-1} \\ \hat{\boldsymbol{w}} &= \hat{\boldsymbol{\Sigma}}\boldsymbol{X}^T\beta\boldsymbol{Y} \end{aligned} \tag{10.39}$$

10.5.1 The evidence framework

If the 'hyperparameters' α and β are unknown (they almost always are) they can be set according to following method known as either the *evidence framework* [35] or *Maximimum Likelihood II (ML II)* [2]. In this approach α and β are set so as to maximise the evidence (also known as marginal likelihood)

$$p(D|\alpha,\beta) = \int p(D|\boldsymbol{w},\beta)p(\boldsymbol{w}|\alpha)d\boldsymbol{w} \tag{10.40}$$

Substituting in our earlier expressions for the prior and likelihood gives

$$p(D|\alpha,\beta) = \left(\frac{\beta}{2\pi}\right)^{-N/2} \left(\frac{\alpha}{2\pi}\right)^{-p/2} \int \exp(-E(\boldsymbol{w}))d\boldsymbol{w} \tag{10.41}$$

where

$$E(\boldsymbol{w}) = \beta E_D + \alpha E_w \tag{10.42}$$

Bishop shows that ([3], page 398 and further details in Appendix B) the integral in equation 10.41 can be evaluated as

$$\int \exp(-E(\boldsymbol{w}))d\boldsymbol{w} = (2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}\exp(-E(\boldsymbol{w})) \tag{10.43}$$

The log of the evidence can then be written as

$$EV(p) = -\alpha E_W - \beta E_D + 0.5\log|\boldsymbol{\Sigma}| + \frac{p}{2}\log\alpha + \frac{N}{2}\log\beta - \frac{N}{2}\log 2\pi \tag{10.44}$$

The values of α and β which maximise the evidence are

$$\alpha = \frac{\gamma}{2E_W} \tag{10.45}$$

$$\beta = \frac{N-\gamma}{2E_D} \tag{10.46}$$

where γ, the number of 'well-determined' coefficients, is given by

$$\gamma = p - \alpha Trace(\boldsymbol{\Sigma}) \tag{10.47}$$

which is calculated using the 'old' value of α. The update for α is therefore an implicit equation. We can also write it as the explicit update

$$\alpha = \frac{p}{2E_W + Trace(\boldsymbol{\Sigma})} \tag{10.48}$$

See Bishop ([3], chapter 10) or Mackay [35] for a derivation of the above equations.

To summarise, the evidence framework works as follows. The weights are first estimated using equation 10.40. The hyperparmeters are then estimated using equations 10.46 and 10.48. This weights are then re-estimated and so are the hyperparameters until the procedure converges. This usually takes ten or so cycles.

Once the above procedure has converged we can use the evidence as a model order selection criterion.

10.5.2 Example

The following figures compare the MDL and Bayesian Evidence model order selection criteria. The first figure shows that, for low model order (relative to the number of data samples) both methods work equally well. The second figure shows that, at high model order, the Bayesian evidence is superior. The last figure shows that EEG recordings from an awake subject can be differentiated from those of an anaesthetised subject. Differentiation was good using the Bayesian evidence criterion but insignificant using MDL.

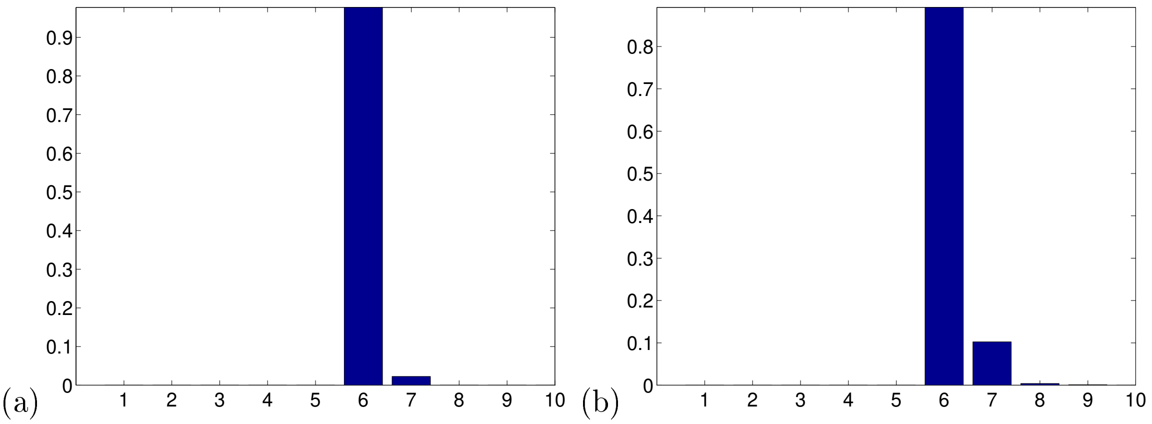

Figure 10.2: Model Order Selection for AR(6) data with (a) MDL and (b) Bayesian Evidence with 3-second blocks of data.

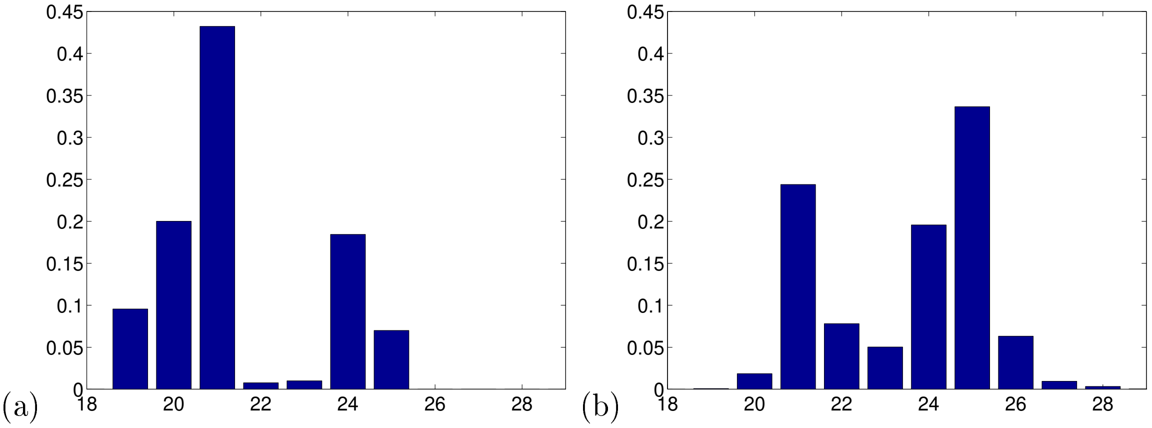

Figure 10.3: Model Order Selection for AR(25) data with (a) MDL and (b) Bayesian Evidence with 3-second blocks of data.

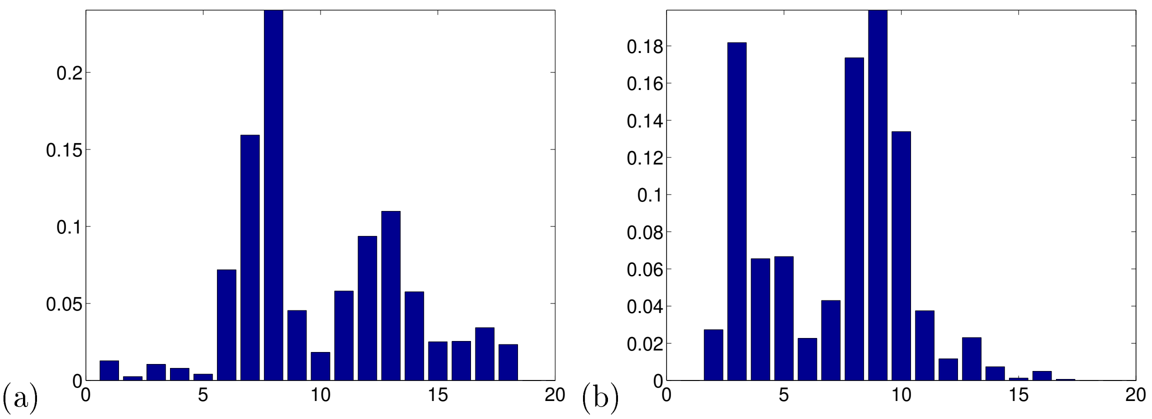

Figure 10.4: Bayesian Evidence model order selection on EEG data from (a) awake subject and (b) anaesthetised subject.