Bayesian Distributions: Prior and Posterior

We will discuss the details of the derivation of equation (8.27) as a brief summary of the Bayesian approach to statistics.

The probability model is that, for a given parameter β the distribution of a random dataset $\mathbf{Z} = \{\mathbf{z}_i\} = \{x_i, y_i\}_{i=1}^{N}$ (x_i are considered fixed) is

$$y_i = f(x_i) + \epsilon_i = \sum_{j=1}^{p} \beta_j h_j(x_i) + \epsilon_j = \mathbf{h}^T(x_i)\beta + \epsilon_i \ ,$$

where ϵ_i are iid $N(0, \sigma^2)$ random variables and

$$\mathbf{h}^T(x) = (h_1(x), ..., h_p(x)),$$

with the right side consisting of the spline basis elements. Thus given β (and the fixed location x), the probability distribution of y_i is

$$(y_i|\,\beta, x_i) \sim N(\mathbf{h}^T(x_i)\beta, \sigma^2 \ ,$$

so that

$$P(y_i|\beta, x_i) = \frac{1}{(2\pi)^{1/2}\sigma}\, e^{-(y_i - \mathbf{h}^T(x_i)\beta^2/(2\sigma^2)}.$$

Note that conditioning on x_i at the end means only that we are treating the x_i as fixed in the calculation.

The logic is essentially that we are assuming a model for the unknown parameter β as having a probability distribution. Before we see any data in the dataset $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^{N}$ $= \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ only our prior knowledge can give us an idea of this distribution for β, which is therefore called the *prior distribution*. In this case we give a relatively naïve prior where we do not assume too much by assuming that

$$\beta \sim N(0, \Sigma), \qquad (1)$$

so that $P(\beta) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\, e^{-\beta^T\Sigma\beta/2}$, where Σ is the prior covariance matrix. We do not include τ here, but for now absorb it into Σ - we can always at the end replace Σ by $\tau\Sigma$.

The posterior distribution for β (i.e., our distribution for β given the new information in $\mathbf{Z}$) is

$$P(\beta|\mathbf{Z}) = \frac{P(\mathbf{Z}|\beta)P(\beta)}{P(\mathbf{Z})} \ ,$$

where $P(\beta|\mathbf{Z})$ denotes the normal density function of β conditioned on knowing $\mathbf{Z}$. First note that for a given β we can compute

$$P(\mathbf{Z}|\beta) = P(\{x_i, y_i)\}_{i=1}^{N}|\beta) = \prod_{j=1}^{N} P(x_i, y_i)|\beta) = \prod_{j=1}^{N\beta/} P(y_i|x_i, \beta),$$

$$= \frac{1}{(2\pi)^{N/2}\, \sigma^N}\, e^{-\sum\limits_{i=1}^{N}(y_i - \mathbf{h}^T(x_i)\beta)^2/(2\sigma^2)} \;=\; \frac{1}{(2\pi)^{N/2}\sigma^N} e^{-(y-\mathbf{H}\beta)^2/(2\sigma^2)},$$

with

$$\mathbf{H}_{ij} = h_j(x_i).$$

.

Thus the posterior density function of β (i.e. its new probability density given the information in the dataset $\mathbf{Z}$) is

$$P(\beta|\mathbf{Z}) = \frac{P(\mathbf{Z}|\beta)P(\beta)}{P(\mathbf{Z})} = \frac{1}{P(\mathbf{Z})} \cdot \frac{1}{(2\pi)^{N/2}\sigma^N} e^{-(y-\mathbf{H}\beta)^2/(2\sigma^{)\beta 2})} \cdot \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\beta^T\Sigma^{-1}\beta/2}$$

$$= \frac{1}{P(\mathbf{Z})} \cdot \frac{1}{(2\pi)^{(N+p)/2}|\Sigma|^{1/2}\, \sigma^N} e^{-(y-\mathbf{H}\beta)^2/(2\sigma^2) - \beta^T\Sigma^{-1}\beta/2}$$

We now rearrange the exponent as

$$(\mathbf{y} - \mathbf{H}\beta)^2/(2\sigma^2) + \beta^T\Sigma^{-1}\beta/2 = \frac{1}{2\sigma^2}[\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{H}\beta + (\mathbf{H}\beta)^T(\mathbf{H}\beta)] + \beta^T\Sigma^{-1}\beta/2$$

$$= (\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{H}\beta)\frac{1}{2\sigma^2} + \beta^T(\mathbf{H}^T\mathbf{H}/(2\sigma^2) + \Sigma^{-1}/2)\beta$$

$$= \mathbf{y}^T\mathbf{y}\frac{1}{2\sigma^2} + \beta^T(\mathbf{H}^T\mathbf{H}/(2\sigma^2) + \Sigma^{-1}/2)\beta - 2\mathbf{y}^T\mathbf{H}\beta\frac{1}{2\sigma^2}$$

$$= A + \beta^T\mathbf{B}\beta - \mathbf{C}^T\,\beta$$

$$= A - \underbrace{(\mathbf{B}^{-1}\mathbf{C})^T\mathbf{B}(\mathbf{B}^{-1}\mathbf{C})/4}_{\text{cancel}} + [\beta^T\mathbf{B}\beta - \mathbf{C}^T\beta + \underbrace{(\mathbf{B}^{-1}\mathbf{C})^T\mathbf{B}(\mathbf{B}^{-1}\mathbf{C})/4}_{\text{cancel}}]$$

$$= A - (\mathbf{B}^{-1}\mathbf{C})^T\mathbf{B}(\mathbf{B}^{-1}\mathbf{C})/4 + [\beta - \mathbf{B}^{-1}\mathbf{C}/2]^T\,\mathbf{B}[\beta - \mathbf{B}^{-1}\mathbf{C}/2],$$

where we have defined $A = \mathbf{y}^T\mathbf{y}/(2\sigma^2)$, $\mathbf{B} = \mathbf{H}^T\mathbf{H}/(2\sigma^2) + \Sigma^{-1}/2$, and $\mathbf{C}^T = 2\mathbf{y}^T\mathbf{H}/(2\sigma^2)$. Note that $\mathbf{B}^T = \mathbf{B}$, and $\mathbf{C}^T\beta = \beta^T\mathbf{C}$, given that β and $\mathbf{C}$ are vectors. Note that in the last two lines we have just completed the square in the variable β.

Thus

$$P\left(\beta|\mathbf{Z}\right) = \underbrace{\frac{1}{P(\mathbf{Z})} \cdot \frac{1}{(2\pi)^{(N+p)/2}|\Sigma|^{-1/2}\sigma^N}\, e^{-A+(\mathbf{B^{-1}C})^T\mathbf{B}(\mathbf{B}^{-1}C)/4}}_{\equiv D = \text{Normalization constant (no dependence on } \beta)} e^{-(\beta-\mathbf{B}^{-1}\mathbf{C}/2)^T\mathbf{B}(\beta-\mathbf{B}^{-1}\mathbf{C}/2)}. \quad (2)$$

Notice that since the distribution must integrate to 1, the terms before the last exponential (none of which involve β) must just form the proper normalization constant (so the distribution integrates to 1 in β), and the above is just a normal distribution. By matching its form with the usual density $e^{-(\beta-\mu)^T\Sigma_{\text{fin}}^{-1}(\beta-\mu)/2}$ for the normal $N(\mu,\Sigma)$ (without the normalization constant) we see that the mean for β must be

$$\mu = \mathbb{E}(\beta|\mathbf{Z}) = \mathbf{B}^{-1}\mathbf{C}/2 = (\mathbf{H}^T\mathbf{H}/(2\sigma^2) + \Sigma^{-1}\,\mathbf{H}^T\mathbf{y}/(2\sigma^2)$$

$$= (\mathbf{H}^T\mathbf{H} + \Sigma^{-1}\sigma^2)^{-1}\,\mathbf{H}^T\mathbf{y}.$$

The covariance matrix Σ_{fin} of β is

$$\Sigma_{\text{fin}} \equiv \mathbb{V}(\beta|\mathbf{Z}),$$

and by matching to (2) must be given by $\Sigma_{\text{fin}}/2 = \mathbf{B}$, so that

$$\Sigma_{\text{fin}} = (2\mathbf{B})^{-1} = \frac{1}{2}(\mathbf{H}^T\mathbf{H}/(2\sigma^2) + \Sigma^{-1}/2)^{-1} = (\mathbf{H}^T\mathbf{H} + \sigma^2\Sigma^{-1})^{-1}\,\sigma^2\,.$$

Finally, again including the unnecessary but convenient parameter as part of the covariance (by replacing Σ by $\tau\Sigma$), we have

$$\mu = E(\beta|\mathbf{Z}) = (\mathbf{H}^T\mathbf{H} + \Sigma^{-1}\,\sigma^2/\tau)^{-1}\,\mathbf{H}^T\mathbf{y}$$

$$\Sigma_{\text{fin}} = (\mathbf{H}^T\mathbf{H} + \sigma^2\Sigma^{-1}/\tau)^{-1}\,\sigma^2.$$