Application of Multi-Dimensional Bayesian Inference on Data Fitting

Consider the problem of fitting a parameterized curve $y = F(x;\underline{a})$ to a set of data $D = (y_k, x_k), k = 1,\ldots,N$. The following general form of the parameterized model is assumed

$$y = F(x;\underline{a}) = \sum_{i=0}^{n} a_i f_i(x) = \underline{f}^T(x)\underline{a} \tag{1}$$

where $\underline{f}^T(x) = [f_0(x), f_1(x), \cdots, f_n(x)]$ are user-selected known bases functions, and $\underline{a}$ are unknown parameters to be estimated using the data. The functions $\underline{f}^T(x) = [f_0(x), f_1(x), \cdots, f_n(x)]$ can be orthogonal functions taken, for example, as a polynomial basis.

Figure 1. Data points in space (x, y) and fit by linear and quadratic models $y = \underline{f}^T(x)\underline{a}$.

To account for model and measurement errors, assume the prediction error equation

$$y_k = F(x_k;\underline{a}) + e_k \tag{2}$$

which represents the fact that predictions from the model equation can not match exactly the measurements. There is an error e_k between the data point y_k and the assumed parameterized model prediction $F(x_k;\underline{a})$, estimated from (1) evaluated at position x_k. The prediction errors e_k are assumed to be i.i.d Gaussian with $e_k \sim N(0,\sigma^2)$, where σ^2 is unknown. Assuming uniform priors, with large enough bounds, find:

1. The posterior distribution $p(\underline{a},\sigma^2 \mid D, I)$ of the model and prediction error parameters
2. The best estimates of $\underline{a},\sigma^2$
3. The spread of uncertainty about the best estimate in the parameter space
4. The asymptotic estimate of the posterior distribution $p(\underline{a},\sigma^2 \mid D, I)$.
5. The marginal posterior distribution $p(\underline{a} \mid D, I)$.

6. Let $z = G(y) + \eta$ be a relation between an output QoI and the measured quantity y, with $\eta \sim N(0, s^2)$ and s^2 is known. Quantify the uncertainty on z by computing the distribution $p(z \mid D, I)$.

The uniform prior distribution is $\pi(\underline{a}, \sigma^2 \mid I) = \text{const}$, $\underline{a}_{\min} \le \underline{a} \le \underline{a}_{\max}$, $0 \le \sigma^2 \le \sigma^2_{\max}$ with very large bounds of the support of the uniform PDF.

Repeat the steps 1-6 assuming a Gaussian prior PDF $p(\underline{a} \mid I) = N(\underline{a} \mid \underline{\mu}_\pi, \Sigma_\pi)$. [Left as an Exercise]

Solution

1. Posterior PDF

The joint posterior PDF of the unknown parameters $\underline{a}$ and σ^2 is obtained by applying the Bayes rule:

$$p(\underline{a}, \sigma^2 \mid D, I) = \frac{p(D \mid \underline{a}, \sigma^2, I)\pi(\underline{a}, \sigma^2 \mid I)}{p(D \mid I)} \propto p(D, \underline{a}, \sigma^2, I)$$

Assuming that the data is independent, the likelihood $p(D \mid \underline{a}, \sigma^2, \mathrm{I})$ is estimated as follows

$$p(D \mid \underline{a}, \sigma^2, I) = p\left(\{y_1, \ldots, y_n\} \mid \underline{a}, \sigma^2, I\right) = \prod_{k=1}^{N} p(y_k \mid \underline{a}, \sigma^2, I)$$

Using the prediction error equation (2), the fact that the prediction error term e_k follows a Gaussian distribution and that $F(x_k; \underline{a})$ is deterministic given $\underline{a}$, the data point x_k then follows a Gaussian distribution with PDF $\mathrm{N}\left(y_k \mid F(x_k; \underline{a}), \sigma^2\right)$ given by

$$p(y_k \mid \underline{a}, \sigma^2, \mathrm{I}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}\left[y_k - F(x_k; \underline{a})\right]^2\right\}$$

Thus, the joint posterior PDF is

$$p(\underline{a}, \sigma^2 \mid D, I) \propto \frac{1}{\left(\sqrt{2\pi}\right)^N \sigma^N} \exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^{N}\left[y_k - F(x_k; \underline{a})\right]^2\right\} \pi(\underline{a}, \sigma^2 \mid \mathrm{I})$$

Introduce the function $J(\underline{a}) = \sum_{k=1}^{N}\left[y_k - F(x_k; \underline{a})\right]^2$, which measures the fit or the mismatch between the measured data and the predictions from the model. The expression for the posterior PDF becomes:

$$p(\underline{a}, \sigma^2 \mid D, I) = \frac{1}{\left(\sqrt{2\pi}\right)^N \sigma^N} \exp\left\{-\frac{1}{2\sigma^2} J(\underline{a})\right\} \pi(\underline{a}, \sigma^2 \mid \mathrm{I})$$

2. Maximum a posteriori estimate (MAP) or most probable value (MPV) or Best Estimate

Introducing the Log-Posterior function

$$L(\underline{a}, \sigma^2) = -\ln p(\underline{a}, \sigma^2 \mid D, I) = \frac{N}{2}\log\sigma^2 + \frac{1}{2\sigma^2} J(\underline{a}) - \log\pi(\underline{a}, \sigma^2 \mid I)$$

the best estimates satisfy:

$$\left.\frac{\partial L}{\partial a_j}\right|_{\substack{\underline{a}=\underline{\hat{a}} \\ \sigma^2=\hat{\sigma}^2}} = 0,\ j = 0,\ldots,n \tag{3}$$

$$\left.\frac{\partial L}{\partial \sigma^2}\right|_{\substack{\underline{a}=\underline{\hat{a}} \\ \sigma^2=\hat{\sigma}^2}} = 0 \tag{4}$$

For a uniform prior,

$$\left.\frac{\partial L}{\partial a_j}\right|_{\substack{\underline{a}=\underline{\hat{a}} \\ \sigma^2=\hat{\sigma}^2}} = \frac{1}{2\hat{\sigma}^2}\left.\frac{\partial J(\underline{a})}{\partial a_j}\right|_{\underline{a}=\underline{\hat{a}}} = 0, \qquad j = 0,\ldots,n,$$

which is equivalent to minimizing $J(\underline{a})$ with respect to the parameters $\underline{a}$. Note that $J(\underline{a})$ is a measure of fit between measurement and predictions from the model. Alternatively is called *sum of squares of the residuals* and the value $\underline{\hat{a}}$ is called the *least squares estimate.*

Equation (4) becomes:

$$\left.\frac{\partial L}{\partial \sigma^2}\right|_{\substack{\underline{a}=\underline{\hat{a}} \\ \sigma^2=\hat{\sigma}^2}} = \frac{N}{2}\frac{1}{\hat{\sigma}^2} - \frac{1}{2}\frac{1}{\hat{\sigma}^4}J(\underline{\hat{a}}) = 0$$

which yields

$$\hat{\sigma}^2 = \frac{1}{N}J(\underline{\hat{a}}) \tag{5}$$

Thus the MPV (or best estimate) of the variance of the prediction error is the average of the residuals obtained at the most MPV of the model parameters.

Returning now to equation (3), we proceed to solve the system in the special case for which that the function $y = F(x;\underline{a})$ is linear in $\underline{a}$. In this case one has that

$$\frac{\partial L(\underline{a})}{\partial a_j} = \frac{1}{2\sigma^2}\frac{\partial J(\underline{a})}{\partial a_j} = \frac{1}{2\sigma^2}\frac{\partial \sum_{k=1}^{N}\left[y_k - \sum_{i=0}^{n} f_i(x_k)a_i\right]^2}{\partial a_j} = \frac{1}{2\sigma^2}\frac{\partial \sum_{k=1}^{N}\left[y_k - \underline{f}^T(x_k)\underline{a}\right]^2}{\partial a_j} =$$

$$= -\frac{1}{\sigma^2}\sum_{k=1}^{N}\left[y_k - \underline{f}^T(x_k)\underline{a}\right] f_j(x_k) = -\frac{1}{\sigma^2}\left\{\sum_{k=1}^{N} y_k f_j(x_k) - \left[\sum_{k=1}^{N}\underline{f}^T(x_k)f_j(x_k)\right]\underline{a}\right\}$$

Therefore, $\left.\frac{\partial J(\underline{a})}{\partial a_j}\right|_{\underline{a}=\underline{\hat{a}}} = 0,\ j = 0,\ldots,n$ gives

$$\sum_{k=1}^{N} y_k f_j(x_k) = \left[\sum_{k=1}^{N}\underline{f}^T(x_k)f_j(x_k)\right]\underline{\hat{a}}, \qquad j = 0,1,\ldots,n \tag{6}$$

Introducing the vector

$$\underline{d} = \sum_{k=1}^{N} y_k \underline{f}(x_k)$$

and the matrix

$$B = \left[\sum_{k=1}^{N} \underline{f}(x_k) \underline{f}^T(x_k) \right]$$

the system of n equations (6) can be re-written in compact matrix form as

$$B\hat{\underline{a}} = \underline{d} \tag{7}$$

with the solution

$$\hat{\underline{a}} = B^{-1}\underline{d} \tag{8}$$

to depend only on the data and the functional form of $\underline{f}(x)$. For example, the elements of $\underline{f}(x)$ can be any polynomial basis functions. The estimate in (8) obtained for a uniform prior distribution coincides with the estimate that one obtains from a least squares technique. Note that for stable solution the functions in $\underline{f}(x)$ have to be chosen appropriately.

3. Spread of uncertainty in the parameter space

The Hessian matrix of the minus log-posterior function, evaluated at the MPV, is used to estimate the uncertainty in the model parameters. The elements of partition $H^{(\underline{a})}(\underline{a},\sigma^2)$ of the Hessian matrix $H(\underline{a},\sigma^2)$ associated with the model parameters $\underline{a}$ is

$$H_{jl}^{\underline{a}}(\underline{a},\sigma^2) = \frac{\partial^2 L}{\partial a_j \partial a_l} = \frac{1}{\sigma^2}\sum_{k=1}^{N} f_j(x_k) f_l(x_k) \qquad \Rightarrow \qquad H^{(\underline{a})}(\underline{a},\sigma^2) = \frac{1}{\sigma^2}\sum_{k=1}^{N} \underline{f}(x_k)\underline{f}^T(x_k)$$

Which, by making use of (5), gives

$$H(\hat{\underline{a}},\hat{\sigma}^2) = \frac{N}{J(\hat{\underline{a}})}\sum_{k=1}^{N} \underline{f}(x_k)\underline{f}^T(x_k) = \frac{1}{\hat{\sigma}^2}B$$

Similarly, the partition $H^{(\sigma)}(\underline{a},\sigma^2)$ of the Hessian matrix $H(\underline{a},\sigma^2)$ associated with the prediction error model parameter σ^2

$$H^{(\sigma)}(\underline{a},\sigma^2) = \left.\frac{\partial^2 L}{\partial(\sigma^2)^2}\right|_{\sigma^2=\hat{\sigma}^2} = -\frac{N}{2}\frac{1}{\hat{\sigma}^4} + \frac{1}{\hat{\sigma}^6}J(\hat{\underline{a}}) = -\frac{N}{2}\frac{1}{\hat{\sigma}^4} + \frac{1}{\hat{\sigma}^6}N\hat{\sigma}^2 = \frac{N}{2\hat{\sigma}^4}$$

The partition $H^{(\underline{a},\sigma)}(\underline{a},\sigma^2)$ is given by

$$H^{(\underline{a},\sigma)}(\underline{a},\sigma^2) = \frac{\partial^2 L}{\partial\sigma^4 \partial a_j} = \frac{1}{\sigma^4}\left\{ \sum_{k=1}^{N} y_k f_j(x_k) - \left[\sum_{k=1}^{N} \underline{f}^T(x_k) f_j(x_k) \right] \underline{a} \right\} = \frac{1}{\sigma^4}\underline{\delta}_j^T[\underline{d} - B\underline{a}]$$

which, due to (7), it gives $H^{(\underline{a},\sigma)}(\hat{\underline{a}},\hat{\sigma}^2) = \frac{1}{\hat{\sigma}^4}\underline{\delta}_j^T[\underline{d} - B\hat{\underline{a}}] = 0$, where $\underline{\delta}_j^T$ is introduced for mathematical convenience to have all elements equal to zero except the j-th element which is set equal to one. Finally, the Hessian matrix at the MPV is the block diagonal matrix

$$H\left(\hat{\underline{a}},\hat{\sigma}^2\right) = \frac{1}{\hat{\sigma}^2}\begin{bmatrix} B & \underline{0} \\ \underline{0}^{\mathrm{T}} & \frac{N}{2\hat{\sigma}^2} \end{bmatrix}.$$

and the associated covariance matrix is given by

$$C = H^{-1}\left(\underline{\hat{a}}, \hat{\sigma}^2\right) = \hat{\sigma}^2 \begin{bmatrix} B^{-1} & \underline{0} \\ \underline{0}^{\mathrm{T}} & \dfrac{2\hat{\sigma}^2}{N} \end{bmatrix}$$

4. Asymptotic posterior PDF

Using the Bayesian central limit theorem, the asymptotic posterior PDF is Gaussian given by

$$p(\underline{a}, \sigma^2 \mid D, I) = N\left(\underline{a}, \sigma^2 \,\middle|\, \underbrace{\begin{Bmatrix} \underline{\hat{a}} \\ \hat{\sigma}^2 \end{Bmatrix}}_{mean}, \underbrace{C(\underline{\hat{a}}, \hat{\sigma}^2)}_{\text{cov}\,ariance} \right)$$

or equivalently, using that $|C| = \sqrt{2/N}\hat{\sigma}^{n+2}|B|^{-1/2}$, one derives

$$p(\underline{a}, \sigma^2 \mid D, I) = \frac{|B|^{1/2}}{\left(\sqrt{2\pi}\right)^{n+1}\sqrt{2/N}\hat{\sigma}^{n+2}} \exp\left(-\frac{1}{2\hat{\sigma}^2} \begin{Bmatrix} \underline{a} - \underline{\hat{a}} \\ \sigma^2 - \hat{\sigma}^2 \end{Bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathrm{B} & \underline{0} \\ \underline{0}^{\mathrm{T}} & \dfrac{\mathrm{N}}{2\hat{\sigma}^2} \end{bmatrix} \begin{Bmatrix} \underline{a} - \underline{\hat{a}} \\ \sigma^2 - \hat{\sigma}^2 \end{Bmatrix} \right)$$

$$= \frac{|B|^{1/2}}{\left(\sqrt{2\pi}\right)^{n+1}\sqrt{2/N}\hat{\sigma}^{n+2}} \exp\left(-\frac{1}{2\hat{\sigma}^2} \left[(\underline{a} - \underline{\hat{a}})^T B (\underline{a} - \underline{\hat{a}}) + \frac{\mathrm{N}}{2\hat{\sigma}^2} (\sigma^2 - \hat{\sigma}^2)^2 \right] \right)$$

5. Marginal posterior PDF for $\underline{a}$

The marginal posterior PDF of the model parameters $\underline{a}$ is obtained using the marginalization rule

$$p(\underline{a} \mid D, I) = \int_0^\infty p(\underline{a}, \sigma^2 \mid D, I)\, d\sigma^2 \propto \int_0^\infty \frac{1}{\sigma^{N/2}} \exp\left\{ -\frac{1}{\sigma^2} J(\underline{a}) \right\} \pi(\underline{a}, \sigma^2 \mid I)\, d\sigma^2 \qquad (9)$$

Assuming that $\underline{a}$ and σ^2 are independent prior to the data, i.e. $\pi(\underline{a}, \sigma^2 \mid I) = \pi(\underline{a} \mid I)\pi(\sigma^2 \mid I)$, and that σ^2 follows a uniform prior distribution, the integral in (9) can be evaluated analytically using the following integral value:

$$\int_0^\infty \frac{1}{t^{\alpha+1}} \exp(-\beta / t) dt = \Gamma(\alpha)\beta^{-\alpha}$$

where $\Gamma(\alpha)$ is the Gamma function defined as $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$.

Letting $t = \sigma^2$, $\alpha + 1 = \dfrac{\mathrm{N}}{2}$ and $\beta = \dfrac{1}{2} J(\underline{a})$ the integral is evaluated to be

$$\int_0^\infty \frac{1}{\sigma^{N/2}} \exp\left\{ -\frac{1}{\sigma^2} J(\underline{a}) \right\} \pi(\underline{a}, \sigma^2 \mid I)\, d\sigma^2 = \Gamma\left(\frac{\mathrm{N}}{2} - 1 \right) \left[\frac{1}{2} J(\underline{a}) \right]^{-\frac{\mathrm{N}}{2}+1} \pi(\underline{a} \mid I)$$

Thus, the marginal posterior PDF given by

$$\begin{aligned} p(\underline{a} \mid D, I) &\propto \Gamma\left(\frac{\mathrm{N}}{2} - 1 \right) \left[\frac{1}{2} J(\underline{a}) \right]^{-\frac{\mathrm{N}}{2}+1} \pi(\underline{a} \mid I) \\ &\propto J(\underline{a})^{-\frac{\mathrm{N}}{2}+1} \pi(\underline{a} \mid I) \end{aligned} \qquad (10)$$

which for a uniform prior distribution $\pi(\underline{a}\,|\,I)$ is a multivariate Student-t distribution. Note that the estimate (10) assumes that the upper bound $\sigma_{\max}^2$ of σ^2 is large enough so that the estimate of the integral in the interval $(0,\sigma_{\max}^2]$ is same as the integral in the interval $(0,\infty)$ and is not affected by the finite value of $\sigma_{\max}^2$. It is worth noting that the Student-t distribution tends to a Gaussian distribution for large enough number of data N.

6. Posterior distribution of an output QoI

The posterior distribution of the QoI z is formulated using marginalization as follows

$$p(z\,|\,D,I)=\int p(z,\underline{a},D,I)\;d\underline{a}=\int p(z\,|\,\underline{a},D,I)\;p(\underline{a}\,|\,D,I)\;d\underline{a}$$

Using $z=G(y)+\eta$, with $\eta\sim N(0,s^2)$, and that $y=F(x;\underline{a})$, the conditional distribution of z is Gaussian with mean $G(y)$ and variance s^2, i.e. $p(z\,|\,\underline{a},D,I)=N(z\,|\,G(y),s^2)$, and thus the posterior PDF of the QoI is

$$p(z\,|\,D,I)=\int\frac{1}{\sqrt{2\pi}s}\exp\left\{-\frac{1}{2s^2}\left[z-G(F(x;\underline{a}))\right]^2\right\}p(\underline{a}\,|\,D,I)d\underline{a}$$

The integration is usually in a higher dimensional space of the parameter set $\underline{a}$ and cannot be carried out analytically. The integral can be obtained analytically only under special cases. For example, assume that $G(y)$ is linear in y, i.e. $G(y)=A_1y+A_0$. Since $y=\sum_{i=0}^{n}a_i f_i(x)=\underline{f}^T(x)\;\underline{a}$ then the QoI z is also linear in the model parameters $\underline{a}$

$$z=G(F(x;\underline{a}))=A_1\sum_{i=0}^{n}a_i f_i(x)+A_0+\eta=A_1\underline{f}^T\underline{a}+A_0+\eta$$

and thus the posterior distribution of z conditioned on $\underline{a}$ is also Gaussian given by

$$p(z\,|\,\underline{a},D,I)=\frac{1}{\sqrt{2\pi}s}\exp\left\{-\frac{1}{2s^2}\left[z-A_1\underline{f}^T\underline{a}-A_0\right]^2\right\}$$

The posterior distribution $p(z\,|\,D,I)$ is evaluated by the integral

$$p(z\,|\,D,I)=\int\frac{1}{\sqrt{2\pi}s}\exp\left\{-\frac{1}{2s^2}\left[z-A_1\underline{f}^T\underline{a}-A_0\right]^2\right\}p(\underline{a}\,|\,D,I)d\underline{a}$$

The problem of evaluating the integral remains since $\underline{a}$ is distributed as a student t-distribution. Simplifications are possible is $\underline{a}$ is considered to be Gaussian, which is true for large number of data. Then, z is a sum of Gaussian variables $\underline{a}$ and η and so the distribution of z is also Gaussian with mean

$$\hat{z}=E\left[z\right]=A_1\underline{f}^T(x)E[\underline{a}]+A_0=A_1\underline{f}^T(x)\hat{\underline{a}}+A_0$$

and variance

$$\sigma_z{}^2 = C_z = E\left[(z-\hat{z})^2\right] = E\left[\left\{A_1 \underline{f}^T(x)(a-\underline{\hat{a}})+\eta\right\}^2\right]$$
$$= E\left[A_1 \underline{f}^T(x)(\underline{a}-\underline{\hat{a}})(\underline{a}-\underline{\hat{a}})\underline{f}A_1\right] + 2A_1\underline{f}^T(x)E\left[(\underline{a}-\underline{\hat{a}})\eta\right] + E\left[\eta^2\right]$$
$$= A_1\underline{f}^T(x)C_{\underline{a}}\underline{f}(x)A_1 + s^2 = A_1^2\underline{f}^T(x)C_{\underline{a}}\underline{f}(x) + s^2$$
$$= A_1^2\underline{f}^T(x)B^{-1}\underline{f}(x)\hat{\sigma}^2 + s^2$$

where use was made of the fact that $C_{\underline{a}} = B^{-1}$ and $E\left[(\underline{a}-\underline{\hat{a}})\eta\right] = E\left[(\underline{a}-\underline{\hat{a}})\right]E\left[\eta\right] = E\left[(\underline{a}-\underline{\hat{a}})\right]0 = \underline{0}$ due to the independence of $\underline{a}$ and η. Thus, the posterior PDF of the QoI becomes

$$p(z\,|\,D,I) = \frac{1}{\sqrt{2\pi}\sigma_z}\exp\left\{-\frac{1}{2\sigma_z{}^2}\left[z - A_1\underline{f}^T\underline{\hat{a}} - A_0\right]^2\right\}$$