1 Robust Posterior Predictions

Given an output quantity of interest (QoI) $g(\underline{\theta})$, we would like to evaluate its uncertainty given the uncertainty in $\underline{\theta}$. The uncertainty in $\underline{\theta}$ is quantified by the posterior PDF $p(\underline{\theta} \mid D, I)$. Simple measures of uncertainty in $g(\underline{\theta})$ are the mean, given by

$$\mu_g = Mean[g(\underline{\theta})] = E[g(\underline{\theta})] = \int g(\underline{\theta})\, p(\underline{\theta} \mid D, I)\, d\underline{\theta} \tag{1}$$

and the variance, given by

$$Var[g(\underline{\theta})] = E[g^2(\underline{\theta})] - E^2[g(\underline{\theta})] \tag{2}$$

where

$$E[g^2(\underline{\theta})] = \int g^2(\underline{\theta})\, p(\underline{\theta} \mid D, I)\, d\underline{\theta} \tag{3}$$

is the second moments of $g(\underline{\theta})$. The integrals (1) and (3) have the same structure. So it suffices to discuss the estimation of the integral in (1). An asymptotic approximation will be developed in this Section. Alternatively, the integration can be carried out using stochastic simulation algorithms to be discussed in a later chapter.

The posterior PDF of the model parameters $\underline{\theta}$ is given by

$$p(\underline{\theta} \mid D, I) = \frac{p(D \mid \underline{\theta}, I)\, p(\underline{\theta} \mid I)}{p(D \mid I)}$$

where the evidence is given as

$$p(D \mid I) = \int p(D \mid \underline{\theta}, I)\, p(\underline{\theta} \mid I)\, d\underline{\theta}$$

Substituting in (1), one derives

$$E[g(\underline{\theta})] = \frac{\int g(\underline{\theta})\, p(D \mid \underline{\theta}, I)\, p(\underline{\theta} \mid I)\, d\underline{\theta}}{\int p(D \mid \underline{\theta}, I)\, p(\underline{\theta} \mid I)\, d\underline{\theta}} \equiv \frac{I_{num}}{I_{den}} \tag{4}$$

For more than a few parameters, say 2 or 3, the multidimensional integrals are very difficult to estimate numerically, for example, using numerical integration schemes. A Laplace asymptotic approximation can be used to approximate the integrals in the numerator and the denominator. Specifically, the Laplace asymptotic approximation is given by

$$I = \int_a^b h(\underline{\theta})\, e^{-\lambda G(\underline{\theta})}\, d\underline{\theta} = h(\hat{\underline{\theta}}) e^{-\lambda G(\hat{\underline{\theta}})} \frac{\left(\sqrt{2\pi}\right)^n}{\left(\sqrt{\lambda}\right)^n \det\left[H_\lambda(\hat{\underline{\theta}})\right]^{1/2}} \left\{1 + O\left(\frac{1}{\lambda}\right)\right\} = \hat{I}\left\{1 + O\left(\frac{1}{\lambda}\right)\right\}$$

where the asymptotic estimate is

$$\hat{I} = e^{-\lambda G(\hat{\underline{\theta}})} \frac{\left(\sqrt{2\pi}\right)^n}{\left(\sqrt{\lambda}\right)^n \det\left[H_\lambda(\hat{\underline{\theta}})\right]^{1/2}}$$

$\hat{\underline{\theta}}$ is the value of $\underline{\theta}$ that minimizes $G(\underline{\theta})$, i.e.

$$\hat{\underline{\theta}} = \arg\min_{\underline{\theta}} G(\underline{\theta})$$

and $H(\hat{\underline{\theta}})$ is the Hessian of the function $G(\underline{\theta})$ evaluated at $\hat{\underline{\theta}}$, i.e.

$$H_\lambda(\hat{\underline{\theta}}) = \underline{\nabla}\,\underline{\nabla}^T G(\underline{\theta})\Big|_{\underline{\theta}=\hat{\underline{\theta}}}$$

For approximating the integral $p(D \mid I)$ in the denominator, introduce the function

$$\begin{aligned} L_N(\underline{\theta}) &= -\frac{1}{N}\log[p(D \mid \underline{\theta}, I)\; p(\underline{\theta} \mid I)] \\ &= -\frac{1}{N}\log[p(D \mid \underline{\theta}, I)] - \frac{1}{N}\log[p(\underline{\theta} \mid I)] \end{aligned}$$

and the function $L(\underline{\theta}) = N L_N(\underline{\theta})$. Then

$$I_{den} \equiv p(D \mid I) = \int e^{-N L_N(\underline{\theta})}\, d\underline{\theta}$$

and applying Laplace approximation one has

$$\begin{aligned} I_{den} &= \int e^{-N L_N(\underline{\theta})}\, d\underline{\theta} = e^{-N L_N(\hat{\underline{\theta}})} \frac{\left(\sqrt{2\pi}\right)^n}{\left(\sqrt{N}\right)^n \det\left[H_N(\hat{\underline{\theta}})\right]^{1/2}} \left\{1 + O\left(\frac{1}{N}\right)\right\} \\ &= e^{-L(\hat{\underline{\theta}})} \frac{\left(\sqrt{2\pi}\right)^n}{\det\left[H(\hat{\underline{\theta}})\right]^{1/2}} \left\{1 + O\left(\frac{1}{N}\right)\right\} = \hat{I}_{den}\left\{1 + O\left(\frac{1}{N}\right)\right\} \end{aligned} \tag{5}$$

where $\hat{\underline{\theta}}$ is the value of $\underline{\theta}$ that minimizes $L_N(\underline{\theta})$ or $L(\underline{\theta})$, that is

$$\hat{\underline{\theta}} = \arg\min_{\underline{\theta}} L_N(\underline{\theta})$$

and $H_N(\hat{\underline{\theta}})$ is the Hessian of $L_N(\underline{\theta})$ evaluated at $\hat{\underline{\theta}}$, i.e.

$$H_N(\hat{\underline{\theta}}) = \underline{\nabla}\underline{\nabla}^T L_N(\underline{\theta})\Big|_{\underline{\theta}=\hat{\underline{\theta}}}$$

Note that

$$H(\hat{\underline{\theta}}) = \underline{\nabla}\underline{\nabla}^T L(\underline{\theta})\Big|_{\underline{\theta}=\hat{\underline{\theta}}} = NH_N(\hat{\underline{\theta}})$$

and thus $\det[H(\hat{\underline{\theta}})] = N^n \det[H_N(\hat{\underline{\theta}})]$ which was used in (5) to simplify the denominator. It should be made clear that this asymptotic approximation for the evidence (the denominator of (4)) was also used for model selection.

A similar asymptotic approximation is next applied for the integral in the numerator of (4) by introducing the function

$$\begin{aligned} L_{g,N}(\underline{\theta}) &= -\frac{1}{N}\log[g(\underline{\theta})\, p(D\,|\,\underline{\theta}, I)\, p(\underline{\theta}\,|\,I)] \\ &= -\frac{1}{N}\log[g(\underline{\theta})] + L_N(\underline{\theta}) \end{aligned}$$

and the function $L_g(\underline{\theta}) = NL_{g,N}(\underline{\theta})$. Let also $\hat{\underline{\theta}}_g$ be the value of $\underline{\theta}$ that minimizes $L_{g,N}(\underline{\theta})$ or $L_g(\underline{\theta})$, that is

$$\hat{\underline{\theta}}_g = \arg\min_{\underline{\theta}} L_{g,N}(\underline{\theta})$$

or

$$\hat{\underline{\theta}}_g = \arg\min_{\underline{\theta}} L_g(\underline{\theta})$$

where

$$L_g(\underline{\theta}) = -\log[g(\underline{\theta})] + L(\underline{\theta})$$

and

$$H_{g,N}(\hat{\underline{\theta}}_g) = \underline{\nabla}\underline{\nabla}^T L_{g,N}(\underline{\theta})\Big|_{\underline{\theta}=\hat{\underline{\theta}}_g}$$

be the Hessian of $L_{g,N}(\underline{\theta})$. Note also that

$$H_g(\hat{\underline{\theta}}_g) = \underline{\nabla}\underline{\nabla}^T L_g(\underline{\theta})\Big|_{\underline{\theta}=\hat{\underline{\theta}}_g} = NH_{g,N}(\hat{\underline{\theta}}_g)$$

Using Laplace asymptotic approximation, the integral I_{num} in the numerator of (4) is asymptotically approximated by

$$I_{num} = \int e^{-NL^*_{g,N}(\underline{\theta})}\, d\underline{\theta} = e^{-NL_{g,N}(\hat{\underline{\theta}}_g)} \frac{\left(\sqrt{2\pi}\right)^n}{\left(\sqrt{N}\right)^n \det\left[H_{g,N}(\hat{\underline{\theta}}_g)\right]^{1/2}} \left\{1 + O\left(\frac{1}{N}\right)\right\}$$
$$= e^{-L_g(\hat{\underline{\theta}}_g)} \frac{\left(\sqrt{2\pi}\right)^n}{\det\left[H_g(\hat{\underline{\theta}}_g)\right]^{1/2}} \left\{1 + O\left(\frac{1}{N}\right)\right\} = \hat{I}_{num}\left\{1 + O\left(\frac{1}{N}\right)\right\} \tag{6}$$

Substituting (5) and (6) into (4), one derives

$$E[g(\underline{\theta})] = e^{-L_g(\hat{\underline{\theta}}_g) + L(\hat{\underline{\theta}})} \frac{\det\left[H(\hat{\underline{\theta}})\right]^{1/2}}{\det\left[H_g(\hat{\underline{\theta}}_g)\right]^{1/2}} \left\{1 + O\left(\frac{1}{N^2}\right)\right\} \tag{7}$$

Substituting the values of $L(\underline{\theta})$ and $L^*_g(\underline{\theta})$, the final asymptotic estimate becomes

$$E[g(\underline{\theta})] = \frac{\hat{I}_{num}}{\hat{I}_{den}} = g(\hat{\underline{\theta}}_g)\, \frac{p(D\,|\,\hat{\underline{\theta}}_g, I)}{p(D\,|\,\hat{\underline{\theta}}, I)}\, \frac{p(\hat{\underline{\theta}}_g\,|\,I)}{p(\hat{\underline{\theta}}\,|\,I)}\, \frac{\det\left[H(\hat{\underline{\theta}})\right]^{1/2}}{\det\left[H_g(\hat{\underline{\theta}}_g)\right]^{1/2}} \tag{8}$$

1.1 REMARKS

1. The errors in the approximation of both integrals are of the order of $O\left(\frac{1}{N}\right)$. However, the leading terms of the errors in both approximations are identical, say $I_{num} = \hat{I}_{num}[1 + \varepsilon O(N^{-1}) + aO(N^{-2})]$ and $I_{den} = \hat{I}_{den}[1 + \varepsilon O(N^{-1}) + bO(N^{-2})]$. Thus, in the ratio these two leading terms cancel out, that is

$$\frac{I_{num}}{I_{den}} = \frac{\hat{I}_{num}}{\hat{I}_{den}} \frac{[1 + \varepsilon O(N^{-1}) + aO(N^{-2})]}{[1 + \varepsilon O(N^{-1}) + bO(N^{-2})]} = \frac{\hat{I}_{num}}{\hat{I}_{den}}[1 + \varepsilon O(N^{-1}) - \varepsilon O(N^{-1}) + \varepsilon^2 O(N^{-2}) + \ldots]$$
$$= \frac{\hat{I}_{num}}{\hat{I}_{den}}[1 + \lambda^2 O(N^{-2})]$$

resulting in an order of $O\left(\frac{1}{N^2}\right)$ error in the asymptotic estimate [Tierney and Kadane 1986].

2. The approximation for the second moment $E[g^2(\underline{\theta})]$ is

$$E[g^2(\underline{\theta})] = g^2(\hat{\underline{\theta}}_{g^2})\, \frac{p(D\,|\,\hat{\underline{\theta}}_{g^2}, I)}{p(D\,|\,\hat{\underline{\theta}}, I)}\, \frac{p(\hat{\underline{\theta}}_{g^2}\,|\,I)}{p(\hat{\underline{\theta}}\,|\,I)}\, \frac{\det\left[H(\hat{\underline{\theta}})\right]^{1/2}}{\det\left[H_{g^2}(\hat{\underline{\theta}}_{g^2})\right]^{1/2}} \tag{9}$$

where

$$\hat{\underline{\theta}}_{g^2} = \arg\min_{\underline{\theta}} L_{g^2}(\underline{\theta})$$

$$L_{g^2}(\underline{\theta}) = -\log[g^2(\underline{\theta})] + L(\underline{\theta}) = -2\log[g(\underline{\theta})] + L(\underline{\theta})$$

$$H_{g^2}(\hat{\underline{\theta}}_{g^2}) = \underline{\nabla}\underline{\nabla}^T L_{g^2}(\underline{\theta})\Big|_{\underline{\theta}=\hat{\underline{\theta}}_{g^2}}$$

3. The asymptotic approximation (7) requires the solution of two optimization problems. The asymptotic approximation (9) requires the solution of an extra optimization problem.
4. Limitations on the accuracy of the asymptotic approximation are as follows:
   (a) It applies to uni-modal posterior distributions. However, it can be extended to multi-modal distributions provided that the modes (Most Probable Values - MPV) are well separated. In the multi-dimensional case the difficulties arise from finding these modes.
   (b) The number of data must be sufficiently large. However, is hard to judge how large is large enough in asymptotic approximations. Also, one cannot give estimates of how far the asymptotic estimate is from the exact value.
   (c) It is not possible to improve the accuracy of the approximation for fixed number of data N. The accuracy can be improved by collecting additional data.
   (d) The accuracy of the asymptotic estimate deteriorates for moderate to high-dimensional parameter space (for example for more than 10 parameters).
   (e) Another numerical issue is that the computation of the Hessian matrices may be time consuming to carry out in high dimensional spaces. First and second-order adjoint methods can be used to alleviate the time consuming operations that arise from multiple re-runs of the system model to compute numerically (e.g. with finite difference schemes) the components in the Hessian.
5. For all of the aforementioned reasons, one could use sampling methods to estimate the integrals in (1). Sampling methods involv3e significantly more computational effort but they are more general, easier to program and parallelizable.

Example 1: Consider the one-parameter case and let the likelihood of the parameter θ be $p(D\,|\,\underline{\theta},I) = \exp\left[-\beta(\theta-\theta_0)^2\right]$ and $p(\underline{\theta}\,|\,I)$ be a uniform distribution with large enough bounds. Find simple measures (mean and standard deviation) of the uncertainty in the output QoI $g(\underline{\theta}) = \exp[-2a\theta]$.

Solution: For the mean we need to evaluate the integral

$$\mu_g = E[g(\theta)] = \int g(\theta)\, p(\theta\,|\,D,I)\, d\theta$$

where $p(\theta\,|\,D,I)$ is the posterior and it is given by

The result is

$$\mu_g = g(\hat{\theta}_g)\ \frac{p(D\,|\,\hat{\theta}_g, I)}{p(D\,|\,\hat{\theta}, I)}\ \frac{p(\hat{\theta}_g\,|\,I)}{p(\hat{\theta}\,|\,I)}\ \frac{\det\left[H(\hat{\theta})\right]^{1/2}}{\det\left[H_g(\hat{\theta}_g)\right]^{1/2}} \tag{10}$$

where each factor is obtained as follows. Note that $L(\theta) = \beta(\theta - \theta_0)^2$, $dL(\theta)/d\theta = 2\beta(\theta - \theta_0)$, $d^2L_g(\theta)/d\theta^2 = 2\beta$, the value of $\hat{\theta}$ is obtained from

$$\hat{\theta} = \arg\min_{\theta} L(\theta) = \arg\min_{\theta}[\beta(\theta - \theta_0)^2] = \theta_0$$

and that $H(\hat{\underline{\theta}}) = \underline{\nabla}\underline{\nabla}^T L(\underline{\theta})\Big|_{\underline{\theta}=\hat{\underline{\theta}}} = d^2L(\theta)/d\theta^2\Big|_{\underline{\theta}=\hat{\underline{\theta}}} = 2\beta$.

Note also that $L_g(\theta) = 2a\theta + \beta(\theta - \theta_0)^2$, $dL_g(\theta)/d\theta = 2a + 2\beta(\theta - \theta_0)$, $d^2L_g(\theta)/d\theta^2 = 2\beta$, and the value of $\hat{\theta}$ is obtained from

$$\hat{\theta}_g = \arg\min_{\theta} L_g(\theta) = \arg\min_{\theta}[2a\theta + \beta(\theta - \theta_0)^2] = \theta_0 - \frac{a}{\beta}$$

and that $H_g(\hat{\theta}_g) = \underline{\nabla}\underline{\nabla}^T L_g(\underline{\theta})\Big|_{\theta=\hat{\theta}_g} = d^2L_g(\theta)/d\theta^2\Big|_{\theta=\hat{\theta}_g} = 2\beta$.

Substituting in (10) one readily obtains

$$\begin{aligned}
\mu_g &= \exp[-2a\theta_g^*]\ \frac{\exp\left[-\beta(\theta_g^* - \theta_0)^2\right]}{\exp\left[-\beta(\hat{\theta} - \theta_0)^2\right]} \\
&= \exp[-2a\theta_g^*]\ \exp\left[-\beta(\theta_g^* - \theta_0)^2\right] \\
&= \exp[-2a(\theta_0 - \frac{a}{\beta})]\ \exp\left[-\beta(\frac{a}{\beta})^2\right] \\
&= \exp[-2a(\theta_0 - \frac{a}{\beta}) - \frac{a^2}{\beta}] \\
&= \exp[-2a(\theta_0 - \frac{a}{\beta})]
\end{aligned}$$

Similarly, one can obtain an estimate for the standard deviation of $g(\underline{\theta})$.

<u>Exercises</u>

1. Estimate the standard deviation of $g(\underline{\theta})$ for the example 1.
2. Repeat the Example 1 assuming that the prior is Gaussian PDF with mean μ and variance s^2. Repeat the Example 1 assuming that $g(\underline{\theta}) = \exp[-2a\theta + \gamma\theta^2]$ and that the prior is either uniform PDF with large bounds or Gaussian PDF with mean μ and variance s^2.