A Short Course on Bayesian Inference (based on *An Introduction to Bayesian Analysis: Theory and Methods* by Ghosh, Delampady and Samanta) Module 2

1 Large Sample Methods in Bayesian Inference

In order to make Bayesian inference about a parameter θ with model $f(\boldsymbol{x}|\theta)$, one needs to choose an appropriate prior $\pi(\theta)$ for θ. Exact or approximate computation of various features of the posterior $\pi(\theta|\boldsymbol{x})$ is a major challenge for Bayesians. Under some regularity conditions, the posterior can be approximated by a normal distribution with the MLE as the mean (or mode), and inverse of the Fisher information matrix as the posterior variance-covariance matrix. If more accuracy is needed, one may have to go for an asymptotic expansion of the posterior. Alternatively, one may sample from the approximated posterior (or some type of $t-$distribution) and use importance sampling. An intuitive rationale behind posterior normality is given below.

How the posterior inference is influenced by a particular prior depends on the relative magnitude of the amount of information in the data, which for iid observations can be measured by the sample size n or $nI(\theta)$ ($I(\theta)$ being the per unit Fisher information) or observed Fisher information

$$\hat{\boldsymbol{I}}_n = -\frac{\partial^2 \log f(\boldsymbol{x}|\theta)}{\partial\theta\partial\theta^T}|_{\theta=\hat{\theta}},$$

$\hat{\theta}$ being the MLE of θ. As the sample size grows, the influence of the prior diminishes. Thus, for large samples, a precise formulation of the prior is not necessary. In most instances when the parameter space is low-dimensional, the prior is washed away by the data. Another important asymptotic fact is consistency of the posterior which we now describe below. In general, the limiting results to be discussed provide a frequentist validation of Bayesian analysis.

Consistency of Posterior Distribution:

Suppose a data sequence $X_1, \ldots, X_n, \ldots$ is generated as iid with a common density $f(x|\theta_0)$. Would a Bayesian analyzing this data with a prior $\pi(\theta)$ be able to learn about θ_0? Ideally, the updated knowledge about θ represented by its posterior should become more and more concentrated near θ_0 as the sample size increases. This asymptotic property is known as the consistency of the posterior distribution at θ_0. Let $X_1, \ldots, X_n$ be iid with joint pdf $f(\boldsymbol{x}_n|\theta), \theta \in \Theta \subset R^p$. Let $\pi(\theta)$ denote the prior pdf and $\pi(\theta|\boldsymbol{X}_n)$ the posterior pdf. Let $\Pi(\cdot|\boldsymbol{X}_n)$ denote the corresponding posterior distribution of θ.

Definition 1. The sequence of posterior distributions $\Pi(\cdot|\boldsymbol{X}_n)$ of θ is said to be consistent at $\theta = \theta_0 \in \Theta$ if for every neighborhood U of θ_0, $\Pi(U|\boldsymbol{X}_n) \to 1$ as $n \to \infty$ with probability 1 wrt to the distribution (of $\boldsymbol{X}_n$) under θ_0.

From the definition of convergence in distribution, it follows that consistency of $\Pi(\cdot|\boldsymbol{X}_n)$ at θ_0 is equivalent to the fact that $\Pi(\cdot|\boldsymbol{X}_n) \xrightarrow{d}$ a distribution degenerate at θ_0 with probability 1 under θ_0.

Example 1. Let $X_1, \ldots, X_n$ be iid Bernoulli observations with $P_\theta(X_1 = 1) = \theta$. Consider a Beta$(\alpha, \beta)$ prior density for θ. The posterior density of θ given $X_1, \ldots, X_n$ is then a Beta$(\sum_{i=1}^n X_i + \alpha, n - \sum_{i=1}^n X_i + \beta)$ distribution with

$$E(\theta|\boldsymbol{X}_n) = \frac{n\bar{X}_n + \alpha}{n + \alpha + \beta}, \quad \mathrm{Var}(\theta|\boldsymbol{X}_n) = \frac{(n\bar{X}_n + \alpha)(n - n\bar{X}_n + \beta)}{(n + \alpha + \beta)^2(n + \alpha + \beta + 1)}.$$

Note that $\bar{X}_n \xrightarrow{\text{a.s. } (P_{\theta_0})} \theta_0$ as $n \to \infty$ by strong law of large numbers. Hence $E(\theta|\boldsymbol{X}_n) \xrightarrow{\text{a.s. } (P_{\theta_0})} \theta_0$, $\mathrm{Var}(\theta|\boldsymbol{X}_n) \xrightarrow{\text{a.s. } (P_{\theta_0})} 0$. Then,

$$\begin{aligned} & P\{\theta \notin [\theta_0 - \epsilon, \theta_0 + \epsilon]|\boldsymbol{X}_n\} = P(|\theta - \theta_0| > \epsilon|\boldsymbol{X}_n) \\ \leq \; & \frac{E[(\theta - \theta_0)^2|\boldsymbol{X}_n]}{\epsilon^2} = \frac{\mathrm{Var}(\theta|\boldsymbol{X}_n) + \{E(\theta|\boldsymbol{X}_n) - \theta_0\}^2}{\epsilon^2} \\ & \xrightarrow{\text{a.s. } (P_{\theta_0})} 0 \quad \text{as} \quad n \to \infty. \end{aligned}$$

An important consequence of the consistency of the posterior is the robustness of Bayesian inference with respect to the choice of prior. Let $X_1, \ldots, X_n$ be iid and π_1 and π_2 be two prior pdf's positive and continuous at θ_0, an interior point of Θ such that $\Pi_1(\cdot|\boldsymbol{X}_n)$ and $\Pi_2(\cdot|\boldsymbol{X}_n)$ are both consistent at θ_0. Then with probability 1 under θ_0,

$$\int_\Theta |\pi_1(\theta|\boldsymbol{X}_n) - \pi_2(\theta|\boldsymbol{X}_n)| d\theta \to 0$$

or equivalently, $\sup_A |\Pi_1(A|\boldsymbol{X}_n) - \Pi_2(A|\boldsymbol{X}_n)| \to 0$ as $n \to \infty$. Thus two different choices of prior density lead approximately to the same posterior distribution.

Asymptotic Normality of the Posterior

Large sample Bayesian methods are primarily based on normal approximation to the posterior distribution of θ. As the sample size n increases, the posterior distribution approaches normality under certain regularity conditions and concentrates in the neighborhood of the posterior mode. Suppose $\tilde{\theta}_n$ is the posterior mode and the first-order partial derivatives of $\log \pi(\theta|\boldsymbol{X}_n)$ vanish at $\tilde{\theta}_n$. Define

$$\tilde{I}_n = -\frac{\partial^2 \log \pi(\theta|\boldsymbol{X}_n)}{\partial\theta\partial\theta^T}|\theta = \tilde{\theta}_n.$$

Then a formal Taylor expansion gives

$$\begin{aligned}\log \pi(\theta|\boldsymbol{X}_n) &\doteq \log \pi(\tilde{\theta}|\boldsymbol{X}_n) - \frac{1}{2}(\theta - \tilde{\theta}_n)^T[-\frac{\partial^2 \log \pi(\theta|\boldsymbol{X}_n)}{\partial\theta\partial\theta^T}|\theta = \tilde{\theta}_n](\theta - \tilde{\theta}_n) \\ &= \log \pi(\tilde{\theta}|\boldsymbol{X}_n) - \frac{1}{2}(\theta - \tilde{\theta}_n)^T \tilde{I}_n(\theta - \tilde{\theta}_n).\end{aligned}$$

Hence

$$\begin{aligned}\pi(\theta|\boldsymbol{X}_n) &\doteq \pi(\tilde{\theta}|\boldsymbol{X}_n)\exp[-\frac{1}{2}(\theta - \tilde{\theta}_n)^T \tilde{I}_n(\theta - \tilde{\theta}_n)] \\ &\propto \exp[-\frac{1}{2}(\theta - \tilde{\theta}_n)^T \tilde{I}_n(\theta - \tilde{\theta}_n)],\end{aligned}$$

which is the kernel of a $N_p(\theta|\tilde{\theta}_n, \tilde{I}_n^{-1})$ density (with p being the dimension of θ).

As the posterior density becomes highly concentrated in a neighborhood of the posterior mode where the prior $\pi(\theta)$ is nearly constant (this is true for a diffuse prior), the posterior is essentially the same as the likelihood $f(\boldsymbol{X}_n|\theta)$. Then we may replace, to the first order of approximation, $\tilde{\theta}_n$ by $\hat{\theta}_n$ and $\tilde{I}_n$ by $\hat{I}_n$ where $\hat{\theta}_n$ is the maximum likelihood estimator (MLE) of θ.

Remark 1. From the above discussion it follows that for iid $X_1, \ldots, X_n|\theta$, we have several ways to approximate the posterior density either by $N_p(\tilde{\theta}_n, \tilde{I}_n^{-1})$ or $N_p(\hat{\theta}_n, \hat{I}_n^{-1})$ or $N_p(\hat{\theta}_n, I^{-1}(\hat{\theta}_n))$, where $I(\theta)$ is the total Fisher information in $\boldsymbol{X}_n$. In particular, under suitable regularity conditions, $\hat{I}_n^{1/2}(\theta - \hat{\theta}_n)$ given $\boldsymbol{X}_n$ converges to $N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ with probability 1 (P_θ). This is comparable with the classical statistical theory where $\hat{I}_n^{1/2}(\theta - \hat{\theta}_n)|\theta \xrightarrow{d} N_p(\boldsymbol{0}, \boldsymbol{I}_p)$.

A Formal Result on Asymptotic Normality of the Posterior Distribution

Let $X_1, \ldots, X_n|\theta$ be iid with a cdf $F(x|\theta)$ and a pdf $f(x|\theta)$. For simplicity, let θ be a scalar with $\theta \in \Theta$ an open subset of R. Fix $\theta_0 \in \Theta$, the "true" value of θ, and all probability statements will be made under P_{θ_0}. Let $l(\theta, x) = \log f(x|\theta)$, $L_n(\theta) = \sum_{i=1}^n l(\theta, X_i)$ and $h^{(i)}$ a generic notation for the ith derivative of a function $h(X, \theta)$ with respect to θ. The function $h(\cdot)$ may not involve X explicitly. Assume the following regularity conditions.

I. The set $\{x : f(x|\theta) > 0\}$ is the same for all $\theta \in \Theta$, i.e., the support does not depend on the parameter.

II. The function $l(\theta, x)$ is thrice differentiable with respect to θ in a neighborhood $(\theta_0 - \delta, \theta_0 + \delta)$ of θ_0 and $\sup_{\theta\in(\theta_0-\delta,\theta_0+\delta)} |l^{(3)}(\theta, x)| \leq M(x)$ with $E_{\theta_0}[M(X_1)] < \infty$.

III. $E_{\theta_0}[l^{(1)}(\theta_0, X)] = 0, 0 < E_{\theta_0}[-l^{(2)}(\theta_0, X)] = E_{\theta_0}[l^{(1)}(\theta_0, X)]^2 < \infty$.

IV. For any $\delta > 0$, $\sup_{|\theta-\theta_0|>\delta} n^{-1}[L_n(\theta) - L_n(\theta_0)] < -\epsilon$ for some $\epsilon > 0$ and all n sufficiently large.

Remark 2. Suppose there exists a sequence of estimators $\{\theta_n^*\}$ of θ such that $\theta_n^* \to \theta_0$ with probability 1 (P_{θ_0}). Then there exists a solution $\hat{\theta}_n$ of the likelihood equation $L_n^{(1)}(\theta) = 0$, i.e., there exists a sequence $\hat{\theta}_n$ of statistics such that with probability 1 (P_{θ_0}), $L_n^{(1)}(\hat{\theta}_n) = 0$ for sufficiently large n and $\hat{\theta}_n \overset{\text{a.s. } (P_{\theta_0})}{\longrightarrow} \theta_0$.

Theorem 1. Suppose assumptions (I)-(IV) hold and $\hat{\theta}_n$ is a strongly consistent solution of the likelihood equations. Then for any prior density $\pi(\theta)$ which is continuous and positive at θ_0,

$$\lim_{n\to\infty} \int_R |\pi_n^*(t|\boldsymbol{X}_n) - \frac{\sqrt{I(\theta_0)}}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2 I(\theta_0)}|dt = 0 \tag{1}$$

with $P_{\theta_0}-$probability one, where $\pi_n^*(t|\boldsymbol{X}_n)$ is the posterior density of $T_n = \sqrt{n}(\theta - \hat{\theta}_n)$ given $\boldsymbol{X}_n$. Also, under the same assumptions, (1) holds with $I(\theta_0)$ replaced by $-n^{-1}L_n^{(2)}(\hat{\theta}_n)$.

Proof. Recall that the posterior density of θ, $\pi_n(\theta|\boldsymbol{X}_n) \propto [\prod_{i=1}^n f(X_i|\theta)]\pi(\theta) \propto \exp[L_n(\theta) - L_n(\hat{\theta}_n)]\pi(\theta)$. Hence, the posterior pdf of $T_n = \sqrt{n}(\theta - \hat{\theta}_n)$ is given by

$$\pi_n^*(t|\boldsymbol{X}_n) = C_n^{-1} \exp[L_n(\hat{\theta}_n + n^{-\frac{1}{2}}t) - L_n(\hat{\theta}_n)]\pi(\hat{\theta}_n + n^{-\frac{1}{2}}t), \tag{2}$$

where $C_n = \int_R \exp[L_n(\hat{\theta}_n + n^{-\frac{1}{2}}t) - L_n(\hat{\theta}_n)]\pi(\hat{\theta}_n + n^{-\frac{1}{2}}t)dt$. Let

$$g_n(t) = \exp[L_n(\hat{\theta}_n + n^{-\frac{1}{2}}t) - L_n(\hat{\theta}_n)]\pi(\hat{\theta}_n + n^{-\frac{1}{2}}t) - \exp[-\frac{1}{2}t^2 I(\theta_0)]\pi(\theta_0). \tag{3}$$

Suppose we show that $\int_R |g_n(t)|dt \to 0$ as $n \to \infty$. Then $C_n \to \int_R \pi(\theta_0)\exp(-\frac{t^2}{2}I(\theta_0))dt = \pi(\theta_0)(2\pi)^{1/2}I^{-1/2}(\theta_0)$. Then the integral in (1) is dominated by

$$C_n^{-1}\int_R |g_n(t)|dt + \int_R |C_n^{-1}\pi(\theta_0)\exp[-\frac{1}{2}t^2 I(\theta_0)] - N(t|0, I^{-1}(\theta_0))|dt \to 0.$$

In order to prove that $\int_R |g_n(t)|dt \to 0$, we write $R = A_1 \cup A_2$, where $A_1 = \{t : |t| > \delta_0\sqrt{n}\}$ and $A_2 = A_1^c$. First,

$$\int_{A_1} |g_n(t)|dt \le \int_{A_1} \pi(\hat{\theta}_n + n^{-\frac{1}{2}}t)\exp[L_n(\hat{\theta}_n + n^{-\frac{1}{2}}t) - L_n(\hat{\theta}_n)]dt + \int_{A_1} \pi(\theta_0)\exp[-\frac{1}{2}t^2 I(\theta_0)]dt. \tag{4}$$

Now

$$\int_{A_1} \pi(\theta_0) \exp[-\frac{1}{2}t^2 I(\theta_0)]dt = \pi(\theta_0) \int_{|t|>\delta_0\sqrt{n}} \exp[-\frac{1}{2}t^2 I(\theta_0)]dt \to 0 \text{ as } n \to \infty. \tag{5}$$

Moreover, since by (IV) for $t \in A_1$, $n^{-1}|L_n(\hat{\theta}_n + n^{-\frac{1}{2}}t) - L_n(\hat{\theta}_n)| < -\epsilon$ for all sufficiently large n,

$$\text{First term in (4)} < \exp(-n\epsilon) \int_{A_1} \pi(\hat{\theta}_n + n^{-\frac{1}{2}}t)dt \to 0 \text{ as } n \to \infty. \tag{6}$$

Combine (5) and (6) to get (4).

Next to prove $\int_{A_2} |g_n(t)|dt \to 0$ as $n \to \infty$, first by Taylor expansion, and $L_n^{(1)}(\hat{\theta}_n) = 0$,

$$L_n(\hat{\theta}_n + n^{-\frac{1}{2}}t) - L_n(\hat{\theta}_n) = -\frac{t^2}{2}\hat{I}_n + R_n(t), \tag{7}$$

where $R_n(t) = (1/6)(t/\sqrt{n})^3 L_n^{(3)}(\theta'_n)$, $|\theta'_n - \hat{\theta}_n| < |t|/\sqrt{n}$. Now by assumption (II), for each real t, $R_n(t) \overset{\text{a.s. } (P_{\theta_0})}{\longrightarrow} 0$ as $n \to \infty$. So, $g_n(t) \overset{\text{a.s. } (P_{\theta_0})}{\longrightarrow} 0$. Next for suitably chosen δ_0, for any $t \in A_2$,

$$|R_n(t)| \leq \frac{1}{6}\delta_0 t^2 n^{-1} \sum_{i=1}^{n} M(X_i) < \frac{1}{4}t^2 \hat{I}_n \ a.s.(P_{\theta_0})$$

for sufficiently large n so that from (7),

$$\exp[L_n(\hat{\theta}_n + n^{-\frac{1}{2}}t) - L_n(\hat{\theta}_n)] < \exp(-\frac{1}{4}t^2 \hat{I}_n) < \exp[-\frac{t^2}{8}I(\theta_0)],$$

a.s. for large n. Hence, for a suitably chosen $\delta_0 > 0$, $|g_n(t)|$ is dominated by an integrable function on A_2. Applying the dominated convergence theorem, $\int_{A_2} |g_n(t)|dt \to 0$ as $n \to \infty$.

Remark 3. We assume in the proof that $\pi(\theta)$ is a proper pdf. However, the result continues to hold even for improper prior $\pi(\theta)$ provided there exists n_0 such that the "posterior" $\pi(\theta|X_1, \ldots, X_{n_0})$ is proper a.e.

We next show that if $\hat{\theta}_n^B = \int_R \theta \pi_n(\theta|\boldsymbol{X}_n)d\theta$ is finite, then $\sqrt{n}(\hat{\theta}_n^B - \hat{\theta}_n) \to 0$ with probability 1 (P_{θ_0}) as $n \to \infty$ under some conditions.

Theorem 2. Suppose in addition to (I)-(IV), $\int \theta\pi(\theta)d\theta < \infty$. Then

$$\int_R |t||\pi_n^*(t|\boldsymbol{X}_n) - N(t|0, I^{-1}(\theta_0))|dt \to 0 \text{ with probability } 1(P_{\theta_0}).$$

Remark 4. The above result implies that

$$\int_R t\pi_n^*(t|\boldsymbol{X}_n)dt \to \int_R tN(t|0, I^{-1}(\theta_0))dt = 0.$$

Hence, $\hat{\theta}_n^B = E(\theta|\boldsymbol{X}_n) = E[\hat{\theta}_n + \frac{t}{\sqrt{n}}|\boldsymbol{X}_n] = \hat{\theta}_n + E[\frac{t}{\sqrt{n}}|\boldsymbol{X}_n]$. Hence, $\sqrt{n}(\hat{\theta}_n^B - \hat{\theta}_n) = \int_R t\pi_n^*(t|\boldsymbol{X}_n)dt \to 0$ as $n \to \infty$.

Laplace Approximation

Bayesian analysis requires evaluation of integrals of the form $\int g(\theta)f(\boldsymbol{x}|\theta)\pi(\theta)d\theta$. For example, when $g(\theta) = 1$, the integral reduces to the marginal likelihood of $\boldsymbol{X}$. The posterior mean requires evaluation of two integrals $\int \theta f(\boldsymbol{x}|\theta)\pi(\theta)d\theta$ and $\int f(\boldsymbol{x}|\theta)\pi(\theta)d\theta$. Laplace's method is a technique for approximating integrals when the integrand has a sharp maximum in the interior of the domain of integration.

Laplaces's method

Consider an integral of the form $I = \int_{-\infty}^{\infty} q(\theta)\exp[nu(\theta)]d\theta$ where q and u are smooth functions of θ with u having a unique maximum at $\hat{\theta}$. In applications, $nu(\theta) = \sum_{i=1}^n l(X_i, \theta)$, the log-likelihood function or the logarithm of the unnormalized posterior density $f(\boldsymbol{x}|\theta)\pi(\theta)$ with corresponding $\hat{\theta}$ equal to the posterior mode. The idea is that if u has a unique sharp maximum at $\hat{\theta}$, the most contribution to the integral I comes from the integral over a small neighborhood $(\hat{\theta} - \delta, \hat{\theta} + \delta)$ of $\hat{\theta}$. We study the behavior of I as $n \to \infty$. As $n \to \infty$,

$$I \approx I_1 = \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} q(\theta)\exp[nu(\theta)]d\theta.$$

Laplace's method involves Taylor series expansion of q and u about $\hat{\theta}$ which gives

$$\begin{aligned} I &\approx \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} [q(\hat{\theta}) + (\theta - \hat{\theta})q'(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 q''(\hat{\theta}) + \text{ smaller terms}] \\ &\quad \times \exp[nu(\hat{\theta}) + nu'(\hat{\theta})(\theta - \hat{\theta}) + \frac{n}{2}u''(\hat{\theta})(\theta - \hat{\theta})^2 + \text{ smaller terms}]d\theta \\ &\approx \exp[nu(\hat{\theta})]q(\hat{\theta}) \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} [1 + \frac{q'(\hat{\theta})}{q(\hat{\theta})}(\theta - \hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 \frac{q''(\hat{\theta})}{q(\hat{\theta})}] \exp[\frac{n}{2}u''(\hat{\theta})(\theta - \hat{\theta})^2]d\theta. \end{aligned}$$

Assume that $c = -u''(\hat{\theta}) > 0$ (e.g., when $u(\theta) = n^{-1}\log f(\boldsymbol{x}|\theta)$) and letting $t = \sqrt{nc}(\theta -$

$\hat{\theta})$, we have

$$\begin{aligned} I &\approx \exp[nu(\hat{\theta})]q(\hat{\theta})\frac{1}{\sqrt{nc}}\int_{-\delta\sqrt{nc}}^{\delta\sqrt{nc}}[1+\frac{t}{\sqrt{nc}}\frac{q'(\hat{\theta})}{q(\hat{\theta})}+\frac{t^2}{2nc}\frac{q''(\hat{\theta})}{q(\hat{\theta})}]\exp(-\frac{t^2}{2})dt \\ &\approx \exp[nu(\hat{\theta})]q(\hat{\theta})\frac{1}{\sqrt{nc}}\int_{-\infty}^{\infty}[1+\frac{t}{\sqrt{nc}}\frac{q'(\hat{\theta})}{q(\hat{\theta})}+\frac{t^2}{2nc}\frac{q''(\hat{\theta})}{q(\hat{\theta})}]\exp(-\frac{t^2}{2})dt \\ &= \exp[nu(\hat{\theta})]q(\hat{\theta})\frac{\sqrt{2\pi}}{\sqrt{nc}}[1+\frac{q''(\hat{\theta})}{2ncq(\hat{\theta})}] = \exp[nu(\hat{\theta})]q(\hat{\theta})\frac{\sqrt{2\pi}}{\sqrt{nc}}[1+O(n^{-1})]. \end{aligned}$$

In general, for the case with a $p-$dimensional parameter vector θ,

$$I = \exp[nu(\hat{\theta})]q(\hat{\theta})\frac{(2\pi)^{p/2}}{n^{p/2}}|\Delta_u(\hat{\theta})|^{-\frac{1}{2}}[1+O(n^{-1})],$$

where $\Delta_u(\theta) = (-\frac{\partial^2 u(\theta)}{\partial\theta_i\partial\theta_j})_{p\times p}$.

The Bayesian Information Criterion (BIC)
Consider a model with a likelihood $f(\boldsymbol{x}|\theta)$ and prior $\pi(\theta)$. Letting $q(\theta) = \pi(\theta)$ and $nu(\theta) = \sum_{i=1}^n l(X_i,\theta)$, the log-likelihood, one can find an approximation to the marginal $\int f(\boldsymbol{x}|\theta)\pi(\theta)d\theta$. This approximation is

$$\exp[\sum_{i=1}^n l(X_i,\hat{\theta})]\pi(\hat{\theta})\frac{(2\pi)^{p/2}}{n^{p/2}}|\Delta_u(\hat{\theta})|^{-\frac{1}{2}}[1+O(n^{-1})].$$

Its logarithm simplifies to

$$\sum_{i=1}^n l(X_i,\hat{\theta}) + \log\pi(\hat{\theta}) + \frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\Delta_u(\hat{\theta})| - \frac{p}{2}\log n + \log[1+O(n^{-1})].$$

Ignoring all the terms which stay bounded as $n\to\infty$, we get

$$BIC = \sum_{i=1}^n l(X_i,\hat{\theta}) - \frac{p}{2}\log n.$$

Laplace Approximation and Posterior Normality
Let $X_1,\ldots,X_n|\theta$ be iid with common pdf $f(x|\theta)$. Also, let $\hat{\theta}$ denote the MLE of θ. Write $T_n = \sqrt{n}(\theta-\hat{\theta})$. Let $\pi(\theta)$ denote the prior pdf, $\pi(\theta|\boldsymbol{X}_n)$ the posterior pdf and $\Pi(\cdot|\boldsymbol{X}_n)$ the posterior distribution. Then for $a>0$, $\Pi(-a<T_n<a|\boldsymbol{X}_n) = \Pi(\hat{\theta}-\frac{a}{\sqrt{n}}<\theta<\hat{\theta}+\frac{a}{\sqrt{n}}|\boldsymbol{X}_n) = J_n/I_n$, where

$$J_n = \int_{\hat{\theta}-\frac{a}{\sqrt{n}}}^{\hat{\theta}+\frac{a}{\sqrt{n}}}\exp[nu(\theta)]\pi(\theta)d\theta,\quad I_n = \int\exp[nu(\theta)]\pi(\theta)d\theta,$$

with $u(\theta) = n^{-1}\sum_{i=1}^{n} \log f(X_i|\theta)$. By the Laplace approximation, $I_n \approx \exp[nu(\hat{\theta})]\pi(\hat{\theta})\frac{\sqrt{2\pi}}{\sqrt{nc}}$, $c = -u''(\hat{\theta})$, the observed Fisher information per unit observation.

Next by the Laplace method,

$$\begin{aligned}
J_n &\approx \exp[nu(\hat{\theta})] \int_{\hat{\theta}-\frac{a}{\sqrt{n}}}^{\hat{\theta}+\frac{a}{\sqrt{n}}} [\pi(\hat{\theta}) + (\theta - \hat{\theta})\pi'(\hat{\theta}) + \text{ smaller terms}] \exp[-\frac{nc}{2}(\theta - \hat{\theta})^2] d\theta \\
&\approx \exp[nu(\hat{\theta})]\pi(\hat{\theta}) \int_{\hat{\theta}-\frac{a}{\sqrt{n}}}^{\hat{\theta}+\frac{a}{\sqrt{n}}} \exp[-\frac{nc}{2}(\theta - \hat{\theta})^2] d\theta \\
&= \exp[nu(\hat{\theta})]\pi(\hat{\theta}) n^{-1/2} \int_{-a}^{a} \exp(-\frac{ct^2}{2}) dt.
\end{aligned}$$

Thus, for $a > 0$,

$$\begin{aligned}
\Pi(-a < T_n < a|\boldsymbol{X}_n) &\approx \frac{\sqrt{c}}{\sqrt{2\pi}} \int_{-a}^{a} \exp(-\frac{ct^2}{2}) dt \\
&= P(-a < Z < a),\ Z \sim N(0, c^{-1}).
\end{aligned}$$

Tierney-Kadane-Kass Refinements

Suppose we are interested in finding

$$\begin{aligned}
E^{\pi}[g(\theta)|\boldsymbol{x}] &= \frac{\int g(\theta) f(\boldsymbol{x}|\theta)\pi\theta) d\theta}{f(\boldsymbol{x}|\theta)\pi(\theta) d\theta} \\
&= \frac{\int g(\theta) \exp[nu(\theta)] d\theta}{\int \exp[nu(\theta)] d\theta}, \qquad (8)
\end{aligned}$$

where $nu(\theta) = \log f(\boldsymbol{x}|\theta) + \log \pi(\theta)$. A simple first order approximation to this moment is given by $g(\hat{\theta})[1 + O(n^{-1})]$.

Suppose now $g(\theta) > 0$ for all $\theta \in \Theta$. Let $nu^*(\theta) = nu(\theta) + \log g(\theta) = nu(\theta) + G(\theta)$, (say). Now apply Laplace method to both the numerator and the denominator of (8). Let $\hat{\theta}_*$ denote the mode of $u^*(\theta)$,

$$\boldsymbol{\Sigma}^{-1} = -\frac{\partial^2 u}{\partial\theta\partial\theta^T}|_{\theta=\hat{\theta}} \text{ and } \boldsymbol{\Sigma}_*^{-1} = -\frac{\partial^2 u_*}{\partial\theta\partial\theta^T}|_{\theta=\hat{\theta}_*}.$$

Tierney and Kadane (JASA, 1986) obtained the approximation

$$E^{\pi}[g(\theta)|\boldsymbol{x}] = \frac{|\boldsymbol{\Sigma}_*|^{1/2} \exp[nu_*(\hat{\theta}_*)]}{|\boldsymbol{\Sigma}|^{1/2} \exp[nu(\hat{\theta})]}[1 + O(n^{-2})]. \qquad (9)$$

We will give an informal proof of (9) when θ is a real-valued parameter. To this end, let $u_k \equiv u_k(\hat{\theta})$, the kth derivative of $u(\theta)$ evaluated at $\hat{\theta}$. Similarly, $u_{*k} \equiv u_{*k}(\hat{\theta}_*)$, the kth derivative of $u_*(\theta)$ evaluated at $\hat{\theta}_*$. Also, write

$$\sigma^2 = -\{u_2\}^{-1}, \text{ and } \sigma_*^2 = -\{u_{*2}\}^{-1}.$$

Under the usual regularity conditions, $\sigma, \sigma_*, u_k, u_{*k}$ are all $O(1)$. First get

$$\begin{aligned}\int \exp[nu(\theta)]d\theta &= \int \exp[nu(\hat{\theta}) - \frac{n}{2\sigma^2}(\theta-\hat{\theta})^2 + R_n(\theta)]d\theta \\ &= \exp[nu(\hat{\theta})]\sqrt{2\pi}\frac{\sigma}{\sqrt{n}}\int \exp[R_n(\theta)]N(\theta|\hat{\theta}, \frac{\sigma^2}{n})d\theta, \end{aligned} \tag{10}$$

where

$$\begin{aligned}R_n(\theta) &= nu(\theta) - nu(\hat{\theta}) + \frac{n}{2\sigma^2}(\theta-\hat{\theta})^2 \\ &= \frac{n}{6}(\theta-\hat{\theta})^3u_3 + \frac{n}{24}(\theta-\hat{\theta})^4u_4 + \frac{n}{120}(\theta-\hat{\theta})^5u_5 + \frac{n}{720}(\theta-\hat{\theta})^6u_6 + \ldots \end{aligned} \tag{11}$$

By Taylor expansion,

$$\begin{aligned}\exp[R_n(\theta)] &= 1 + \{\frac{n}{6}(\theta-\hat{\theta})^3u_3 + \frac{n}{24}(\theta-\hat{\theta})^4u_4 + \frac{n}{120}(\theta-\hat{\theta})^5u_5\} \\ &\quad + \frac{1}{2}\{\frac{n}{6}(\theta-\hat{\theta})^3u_3 + \frac{n}{24}(\theta-\hat{\theta})^4u_4\}^2 + \frac{1}{6}\{\frac{n}{6}(\theta-\hat{\theta})^3u_3\}^3 + \ldots \\ &= 1 + \frac{n}{6}(\theta-\hat{\theta})^3u_3 + [\frac{n}{24}(\theta-\hat{\theta})^4u_4 + \frac{n^2(\theta-\hat{\theta})^6u_3^2}{72}] \\ &\quad + [\frac{n(\theta-\hat{\theta})^5u_5}{120} + \frac{n^2(\theta-\hat{\theta})^7u_3u_4}{144} + \frac{n^3(\theta-\hat{\theta})^9u_3^3}{1296}] + O(n^{-2}). \end{aligned} \tag{12}$$

On integration,

$$\begin{aligned}\int_R \exp[R_n(\theta)]N(\theta|\hat{\theta}, \frac{\sigma^2}{n})d\theta &= 1 + \frac{u_4}{24}\int_R n(\theta-\hat{\theta})^4N(\theta|\hat{\theta}, \frac{\sigma^2}{n})d\theta \\ &\quad + \frac{u_3^2}{72}\int_R n^2(\theta-\hat{\theta})^6N(\theta|\hat{\theta}, \frac{\sigma^2}{n})d\theta + O(n^{-2}) \\ &= 1 + \frac{nu_4}{24}\{3(\frac{\sigma^2}{n})^2\} + \frac{n^2u_3^2}{72}\{15(\frac{\sigma^2}{n})^3\} + O(n^{-2}) \\ &= 1 + \frac{a}{n} + O(n^{-2}), \end{aligned} \tag{13}$$

where $a = \frac{1}{8}\sigma^4u_4 + \frac{5}{24}\sigma^6u_3^2$. Hence,

$$\int \exp[nu(\theta)]d\theta = \exp[nu(\hat{\theta})]\sqrt{2\pi}\frac{\sigma}{\sqrt{n}}[1 + \frac{a}{n} + O(n^{-2})]. \tag{14}$$

Similarly,

$$\int \exp[nu_*(\theta)]d\theta = \exp[nu_*(\hat{\theta}_*)]\sqrt{2\pi}\frac{\sigma_*}{\sqrt{n}}[1 + \frac{a_*}{n} + O(n^{-2})], \tag{15}$$

where $a_* = \frac{1}{8}\sigma_*^4u_{*4} + \frac{5}{24}\sigma_*^6u_{*3}^2$. Hence,

$$\begin{aligned}E^{\pi}[g(\theta)|\boldsymbol{x}] &= \frac{\sigma_*}{\sigma}\exp[nu_*(\hat{\theta}_*) - nu(\hat{\theta})]\frac{1 + \frac{a_*}{n} + O(n^{-2})}{1 + \frac{a}{n} + O(n^{-2})} \\ &= \frac{\sigma_*}{\sigma}\exp[nu_*(\hat{\theta}_*) - nu(\hat{\theta})][1 + \frac{a_* - a}{n} + O(n^{-2})]. \end{aligned} \tag{16}$$

Next observe that

$$\begin{aligned} 0 = u_{*1}(\hat{\theta}_*) &= u_1(\hat{\theta}_*) + n^{-1}G'(\hat{\theta}_*) \\ &\approx u_1(\hat{\theta}) + (\hat{\theta}_* - \hat{\theta})u_2(\hat{\theta}) + n^{-1}G'(\hat{\theta}) + n^{-1}(\hat{\theta}_* - \hat{\theta})G''(\hat{\theta}) \\ &= (\hat{\theta}_* - \hat{\theta})[u_2(\hat{\theta}) + n^{-1}G''(\hat{\theta})] + n^{-1}G'(\hat{\theta}), \end{aligned}$$

implying $\hat{\theta}_* - \hat{\theta} \doteq -\{n^{-1}G'(\hat{\theta})\}/[u_2(\hat{\theta}) + n^{-1}G''(\hat{\theta})] = O(n^{-1})$. Hence, since $u_{*k}(\hat{\theta}) = u_k(\hat{\theta}) + n^{-1}G_k(\hat{\theta})$, $u_{*k}(\hat{\theta}_*) - u_k(\hat{\theta}) = O(n^{-1})$. So, $a_* - a = O(n^{-1})$. This leads to

$$E^{\pi}[g(\theta)|\boldsymbol{x}] = \frac{\sigma_*}{\sigma}\exp[nu_*(\hat{\theta}_*) - nu(\hat{\theta})][1 + O(n^{-2})]. \tag{17}$$

Asymptotic Expansion of the Posterior Distribution

Let $F_n(u) = P^{\pi}[\sqrt{n}\hat{I}_n^{1/2}(\theta - \hat{\theta}_n) \le u|\boldsymbol{X}_n]$ be the posterior distribution function of $\sqrt{n}\hat{I}_n^{1/2}(\theta - \hat{\theta}_n)$ given $\boldsymbol{X}_n$. We showed earlier that under a prior π which is continuous and positive at θ_0,

$$\lim_{n\to\infty}\sup_u |F_n(u) - \Phi(u)| = 0 \text{ a.s. } P_{\theta_0}$$

when θ_0 is the true value of the parameter, $\Phi(u)$ being the standard normal cdf.

Johnson (1970) proved the following result refining the original results of Lindley:

$$\sup_u |F_n(u) - \Phi(u) - \phi(u)\sum_{j=1}^{k} n^{-j/2}\psi_j(u, \boldsymbol{X}_n)| \le M_k n^{-\frac{1}{2}(k+1)} \text{ a.s. } P_{\theta_0}$$

for some $M_k > 0$ depending on k, where $\phi(u)$ is the standard normal density and $\psi_j(u, \boldsymbol{X}_n)$ is a jth degree polynomial in u with coefficients bounded in $\boldsymbol{X}_n$. Ghosh, Sinha and Joshi (1982) proved a stronger version of the result.

We now present an informal argument to obtain the expansion for $k = 2$ without the formal rigor of Johnson (1970) or Ghosh et al. (1982).

Let $t = \sqrt{n}(\theta - \hat{\theta}_n)$ and $a_i = \frac{1}{n}\frac{d^i L_n(\theta)}{d\theta^i}|_{\theta=\hat{\theta}_n}, i \ge 1$ so that $a_2 = -\hat{I}_n$. Then by Taylor expansion,

$$\pi(\theta) = \pi(\hat{\theta}_n + t/\sqrt{n}) = \pi(\hat{\theta}_n)[1 + \frac{t}{\sqrt{n}}\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)} + \frac{t^2}{2n}\frac{\pi''(\hat{\theta}_n)}{\pi(\hat{\theta}_n)}] + o(n^{-1})$$

and

$$L_n(\hat{\theta}_n + t/\sqrt{n}) - L_n(\hat{\theta}_n) = \frac{1}{2}t^2 a_2 + \frac{t^3}{6\sqrt{n}}a_3 + \frac{t^4}{24n}a_4 + o(n^{-1}).$$

Hence,

$$\begin{aligned} &\pi(\hat{\theta}_n + t/\sqrt{n})\exp[L_n(\hat{\theta}_n + t/\sqrt{n}) - L_n(\hat{\theta}_n)] \\ = \;&\pi(\hat{\theta}_n)\exp(\frac{a_2 t^2}{2})(1 + \frac{\alpha_1}{\sqrt{n}} + \frac{\alpha_2}{n}) + o(n^{-1}), \end{aligned} \tag{18}$$

where

$$\alpha_1 \equiv \alpha_1(t;\boldsymbol{X}_n) = \frac{t^3}{6}a_3 + t\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)},$$

$$\alpha_2 \equiv \alpha_2(t;\boldsymbol{X}_n) = \frac{t^4}{24}a_4 + \frac{t^6}{72}a_3^2 + \frac{t^2}{2}\frac{\pi''(\hat{\theta}_n)}{\pi(\hat{\theta}_n)} + \frac{t^4}{6}a_3\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)}.$$

Then

$$\begin{aligned} C_n &= \int \pi(\hat{\theta}_n + t/\sqrt{n})\exp[L_n(\hat{\theta}_n + t/\sqrt{n}) - L_n(\hat{\theta}_n)]dt \\ &= \pi(\hat{\theta}_n)\sqrt{\frac{2\pi}{-a_2}}[1 + \frac{a_4}{8a_2^2} - \frac{5}{24}\frac{a_3^2}{a_2^3} - \frac{1}{2a_2}\frac{\pi''(\hat{\theta}_n)}{\pi(\hat{\theta}_n)} + \frac{a_3}{2a_2^2}\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)}] + o(n^{-1}). \end{aligned} \tag{19}$$

Hence the posterior π_n^* of T_n is

$$\begin{aligned} \pi_n^*(t|\boldsymbol{X}_n) &= C_n^{-1}\pi(\hat{\theta}_n + t/\sqrt{n})\exp[L_n(\hat{\theta}_n + t/\sqrt{n}) - L_n(\hat{\theta}_n)] \\ &= (2\pi)^{-\frac{1}{2}}\hat{I}_n^{1/2}\exp(-\frac{\hat{I}_n t^2}{2})[1 + \frac{\gamma_1}{\sqrt{n}} + \frac{\gamma_2}{n}] + o(n^{-1}), \end{aligned} \tag{20}$$

where

$$\gamma_1 \equiv \gamma_1(t;\boldsymbol{X}_n) = \alpha_1(t;\boldsymbol{X}_n) = \frac{t^3}{6}a_3 + t\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)},$$

and

$$\gamma_2 \equiv \gamma_2(t;\boldsymbol{X}_n) = \alpha_2(t;\boldsymbol{X}_n) - \frac{a_4}{8a_2^2} + \frac{5}{24}\frac{a_3^2}{a_2^3} + \frac{1}{2a_2}\frac{\pi''(\hat{\theta}_n)}{\pi(\hat{\theta}_n)} - \frac{a_3}{2a_2^2}\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)}.$$

Let $S_n = \hat{I}_n^{1/2}T_n = \sqrt{n}\hat{I}_n^{1/2}(\theta - \hat{\theta}_n)$. Then the posterior density of S_n is given by

$$\begin{aligned} \pi_n(s|\boldsymbol{X}_n) &= \phi(s)[1 + \frac{1}{\sqrt{n}}\{\frac{a_3 s^3}{6\hat{I}_n^{3/2}} + \frac{s}{\hat{I}_n^{1/2}}\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)}\} \\ &\quad + \frac{1}{n}\{\frac{a_4 s^4}{24\hat{I}_n^2} - \frac{a_3^2 s^6}{72\hat{I}_n^3} + \frac{s^2}{2\hat{I}_n}\frac{\pi''(\hat{\theta}_n)}{\pi(\hat{\theta}_n)} + \frac{a_3 s^4}{6\hat{I}_n^2}\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)} \\ &\quad - \frac{a_4}{8\hat{I}_n^2} + \frac{5a_3^2}{24\hat{I}_n^3} - \frac{1}{2\hat{I}_n}\frac{\pi''(\hat{\theta}_n)}{\pi(\hat{\theta}_n)} - \frac{a_3}{2\hat{I}_n^2}\frac{\pi'(\hat{\theta}_n)}{\pi(\hat{\theta}_n)}\}] + o(n^{-1}). \end{aligned} \tag{21}$$

The expansion given in (21) will be useful later in deriving probability matching priors.