Bayesian Asymptotics

Aleksandar Dogandžić

March 28, 2017

Contents

READING: [Gelman et al. 2014, §4.1 and App. B].

Posterior Approximation Around the MAP Estimate

EXPAND the posterior distribution $f_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta} \mid \boldsymbol{x})$ in Taylor series around the maximum *a posteriori* (MAP) estimate

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}(\boldsymbol{x})$$

and keep the first three terms:

$$\ln f_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta} \mid \boldsymbol{x}) \approx \ln f_{\boldsymbol{\Theta}|\boldsymbol{X}}(\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}} \mid \boldsymbol{x}) + 0.5(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathrm{MAP}})^T \left. \frac{\partial^2 \ln f_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta} \mid \boldsymbol{x})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}) \quad (1)$$

The second term in the Taylor-series expansion vanishes because the log posterior probability density function (pdf) (or probability mass function (pmf)) has zero derivative at the MAP estimate, see (11) in handout `multiparBayes`.

If the number of measurements is large[1], the posterior distribution $f_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta} \mid \boldsymbol{x})$ will be *unimodal*. Furthermore, if $\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ is in the interior of the parameter space $\mathrm{sp}_{\boldsymbol{\Theta}}$ (preferably *far from the boundary* of $\mathrm{sp}_{\boldsymbol{\Theta}}$), then we can use the following approximation for the posterior pdf of $\boldsymbol{\theta}$:

$$f_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta} \mid \boldsymbol{x}) \approx \mathcal{N}\big(\boldsymbol{\theta} \mid \widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}, \mathcal{J}^{-1}(\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}})\big).$$

where

$$\mathcal{J}(\boldsymbol{\theta}) \triangleq -\frac{\partial^2 \ln f_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta} \mid \boldsymbol{x})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^T}$$

is the *observed information* [Gelman et al. 2014, §4.1].

This approximation follows by looking at (1) as a function of $\boldsymbol{\theta}$.

☞ THIS Gaussian approximation is usually good if the posterior distribution is unimodal and roughly symmetric.

[1] and asymptotics works, e.g., we have independent, identically distributed (i.i.d.) observations with usual regularity conditions

We may obtain another (simpler, but typically poorer) approximation by replacing the covariance matrix of the above Gaussian distribution with Cramér-Rao bound (CRB) evaluated at $\widehat{\boldsymbol{\theta}}_{\text{MAP}}$. If $f_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta} \mid \boldsymbol{x})$ has multiple modes (and we can find them all), it can be approximated by a Gaussian mixture or, more generally, a t-distribution mixture.

Example: Binomial success probability

ASSUME conjugate prior $\theta \sim \text{Beta}(\alpha, \beta)$. Then the posterior is

$$f_{\Theta|X}(\theta \mid x) \propto \theta^{\alpha+x-1}(1-\theta)^{\beta+N-x-1}$$

see handout `introBayes`.

The derivatives are

$$\frac{\partial \ln f_{\Theta|X}(\theta \mid x)}{\partial \theta} = \frac{\alpha+x-1}{\theta} - \frac{\beta+N-x-1}{1-\theta}$$
$$-\frac{\partial^2 \ln f_{\Theta|X}(\theta \mid x)}{\partial \theta^2} = \frac{\alpha+x-1}{\theta^2} + \frac{\beta+N-x-1}{(1-\theta)^2}$$

which gives

$$\widehat{\theta}_{\text{MAP}} = \frac{\alpha+x-1}{\alpha+\beta+N-2}$$
$$-\frac{\partial^2 \ln f_{\Theta|X}\big(\widehat{\theta}_{\text{MAP}} \mid x\big)}{\partial \theta^2} = -\frac{\partial^2 \ln f_{\Theta|X}(\theta \mid x)}{\partial \theta^2}\bigg|_{\theta=\widehat{\theta}_{\text{MAP}}}$$
$$= \frac{\alpha+\beta+N-2}{\widehat{\theta}_{\text{MAP}}\big(1-\widehat{\theta}_{\text{MAP}}\big)}$$

and

$$f(\theta \mid x) \approx \mathcal{N}\left(\theta \,\middle|\, \widehat{\theta}_{\text{MAP}}, \frac{\widehat{\theta}_{\text{MAP}}(1-\widehat{\theta}_{\text{MAP}})}{\alpha+\beta+N-2}\right).$$

For $x = 37$ and $N = 41$ with Beta $(3, 3)$ prior, we have

$$\widehat{\theta}_{\text{MAP}} = \frac{3+37-1}{3+3+41-2}$$
$$= \frac{39}{45} = 0.8667$$
$$J(\widehat{\theta}_{\text{MAP}}) = \frac{45^3}{39 \cdot 6} = 389.42, \qquad \big[J(\widehat{\theta}_{\text{MAP}})\big]^{-1/2} = 0.0507$$

see Fig. 1.

Now with the $\frac{1}{2}$ Beta $(8, 2) + \frac{1}{2}$ Beta $(2, 8)$ mixture prior,

$$\widehat{\theta}_{\text{MAP}} = 0.8989$$
$$J(\widehat{\theta}_{\text{MAP}}) = \frac{45^3}{39 \cdot 6} = 490.11, \qquad \big[J(\widehat{\theta}_{\text{MAP}})\big]^{-1/2} = 0.0452$$

see Fig. 2.

The comment about unimodal and symmetric is important. However, when the number of observations gets large, this is usually not a problem. See Fig. 3.

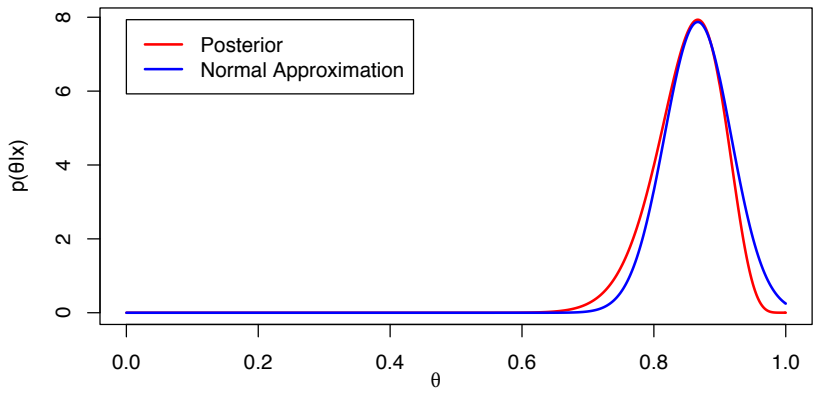

Figure 1: Posterior pdf and its Gaussian approximation for Beta$(3, 3)$ prior on binomial success probability.

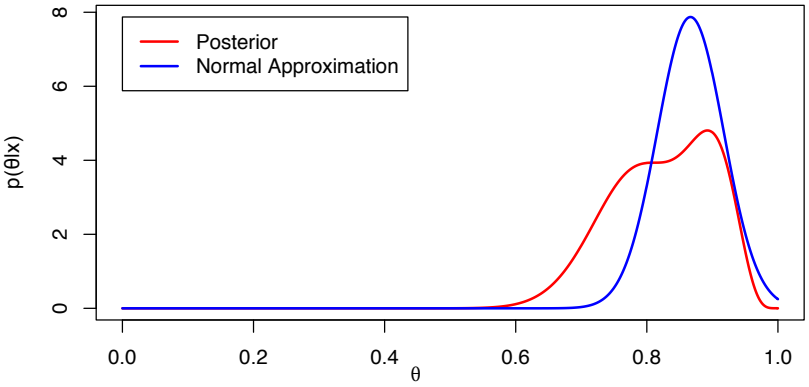

Figure 2: Posterior pdf and its Gaussian approximation for $\frac{1}{2}$ Beta$(8, 2) + \frac{1}{2}$ Beta$(2, 8)$ prior on binomial success probability.

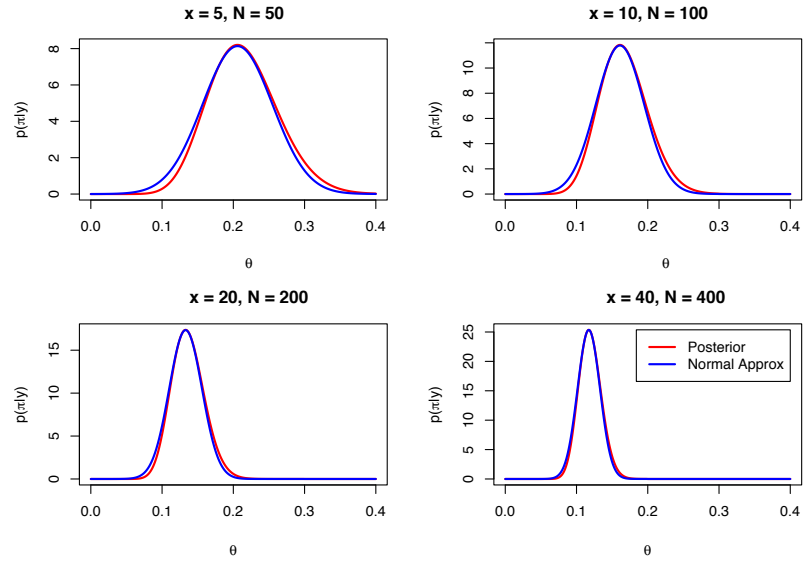

Figure 3: Posterior pdf and its Gaussian approximation for Beta$(10, 10)$ prior on binomial success probability.

Asymptotic Normality and Consistency for Bayesian Models

Consider conditionally i.i.d. observations $(X[n])_{n=0}^{N-1}$ given θ, following

$$\{X[n] \mid \Theta = \theta\} \sim f_{X|\Theta}(x \mid \theta_0) \tag{2}$$

where θ_0 is the true value of the parameter. As before, we also have

- $f_{X|\Theta}(x \mid \theta)$, data model, likelihood;
- $f_\Theta(\theta)$ or $p_\Theta(\theta)$, the prior distribution on θ.

Define $\boldsymbol{x} = (x[n])_{n=0}^{N-1}$. For simplicity, we consider the case of a scalar parameter θ, but the results can be generalized.

In the following discussion, we assume that the data model is correct and that θ_0 is the unique minimizer of

$$D\big(f_{X|\Theta}(x[n] \mid \theta_0) \,\big\|\, f_{X|\Theta}(x[n] \mid \theta)\big) = \mathrm{E}_{X|\Theta}\left(\ln \frac{f_{X|\Theta}(X[n] \mid \theta_0)}{f_{X|\Theta}(X[n] \mid \theta)} \,\middle|\, \theta_0\right).$$

identifiability condition

Theorem 1 (Convergence in discrete parameter space)**.** *If the parameter space Θ is finite and*

$$p_\Theta(\theta_0) = \Pr_\Theta\{\Theta = \theta_0\} > 0$$

then

$$p_{\Theta|\boldsymbol{X}}(\theta_0 \mid \boldsymbol{x}) = \Pr_{\Theta|\boldsymbol{X}}\{\Theta = \theta_0 \mid \boldsymbol{X} = \boldsymbol{x}\} \to 1 \qquad \text{as } N \nearrow +\infty.$$

Proof. Consider the log posterior odds:

$$\ln\left[\frac{p_{\Theta|\boldsymbol{X}}(\theta \mid \boldsymbol{x})}{p_{\Theta|\boldsymbol{X}}(\theta_0 \mid \boldsymbol{x})}\right] = \ln\left[\frac{p_\Theta(\theta)}{p_\Theta(\theta_0)}\right] + \sum_{n=0}^{N-1} \ln\left[\frac{f_{X|\Theta}(x[n] \mid \theta)}{f_{X|\Theta}(x[n] \mid \theta_0)}\right]. \tag{3}$$

Terminology. We refer to the posterior ratio

$$\frac{p_{\Theta|\boldsymbol{X}}(\theta \mid \boldsymbol{x})}{p_{\Theta|\boldsymbol{X}}(\theta_0 \mid \boldsymbol{x})}$$

as posterior-odds ratio: θ versus θ_0, in this case.

The second term in this expression is the sum of N conditionally i.i.d. random variables given θ_0. Recall that $(X[n])_{n=0}^{N-1}$ are coming from $f_{X|\Theta}(x[n] \mid \theta_0)$; then

$$\begin{aligned}\mathrm{E}_{X|\Theta}\left[\ln\left(\frac{f_{X|\Theta}(X[n] \mid \theta)}{f_{X|\Theta}(X[n] \mid \theta_0)}\right) \,\middle|\, \theta_0\right] &= -D\big(f_{X|\Theta}(x[n] \mid \theta_0) \,\big\|\, f_{X|\Theta}(x[n] \mid \theta)\big) \\ &\leq 0\end{aligned}$$

where, by our assumption that θ_0 is the unique minimizer of $D\big(f_{X|\Theta}(X \mid \theta_0) \,\big\|\, f_{X|\Theta}(X \mid \theta)\big)$, the equality holds only for $\theta = \theta_0$.

If $\theta \neq \theta_0$, the second term in (3) is the sum of N i.i.d. random variables with negative mean, which diverges to $-\infty$ as $N \nearrow \infty$. As long as

$$\Pr_\Theta(\Theta = \theta_0) = p_\Theta(\theta_0) > 0$$

then the first summand in (3) is finite

the log posterior odds in (3) $\searrow -\infty$ as $N \nearrow +\infty$. Thus, if $\theta \neq \theta_0$, the posterior odds go to zero:

$$\frac{p_{\Theta|\boldsymbol{X}}(\theta \mid \boldsymbol{x})}{p_{\Theta|\boldsymbol{X}}(\theta_0 \mid \boldsymbol{x})} \to 0$$

which implies $p_{\Theta|\boldsymbol{X}}(\theta \mid \boldsymbol{x}) \searrow 0$. As all the probabilities summed over all values of θ must add to one, we have

$$p_{\Theta|\boldsymbol{X}}(\theta_0 \mid \boldsymbol{x}) \to 1.$$

□

Theorem 2 (Convergence in continuous parameter space)**.** *If θ is defined on a compact set (i.e., closed and bounded) and A is an open subset of the parameter space containing θ_0 with prior probability $f_\Theta(\theta)$ satisfying $\int_{\theta \in A} f_\Theta(\theta)\,\mathrm{d}\theta > 0$, then*

$$\Pr_{\Theta|\boldsymbol{X}}\{\Theta \in A \mid \boldsymbol{X} = \boldsymbol{x}\} \to 1 \quad \textit{as } N \nearrow \infty.$$

Proof. Similar to the proof for the discrete case. □

✻ Technical details:

- In many popular continuous-parameter scenarios, the parameter space is not a compact set, e.g. the parameter space for the mean of a Gaussian random variable is $(-\infty, +\infty)$. Luckily, for most problems of interest, the compact-set assumption of Theorem 2 can be relaxed.
- Similarly, Theorem 1 can often be extended to allow for an infinite discrete parameter space.

Theorem 3 (Asymptotic Normality of $f_{\Theta|\boldsymbol{X}}(\theta \mid \boldsymbol{x})$)**.** *Under some regularity conditions (particularly that θ_0 is not on the boundary of the parameter space) and under the conditional i.i.d. measurement model (2),*

$$\sqrt{N}\left[\hat{\theta}_{\mathrm{MAP}}(\boldsymbol{X}) - \theta_0\right] \xrightarrow{\mathrm{d}} \mathcal{N}\big(0, \mathcal{I}_1^{-1}(\theta_0)\big) \qquad \textit{as } N \nearrow +\infty$$

where $\mathcal{I}_1(\theta_0)$ is the Fisher information for $\theta = \theta_0$ and a single measurement (say $X[0]$):

$$\begin{aligned}\mathcal{I}_1(\theta_0) &= \mathrm{E}_{X[0]|\Theta}\left[\left(\frac{\mathrm{d} \ln f_{X|\Theta}(X[0] \mid \theta)}{\mathrm{d}\theta}\right)^2 \,\middle|\, \theta_0\right] \\ &= -\mathrm{E}_{X[0]|\Theta}\left[\frac{\mathrm{d}^2 \ln f_{X|\Theta}(X[0] \mid \theta)}{\mathrm{d}\theta^2} \,\middle|\, \theta_0\right].\end{aligned}$$

Proof. See [Gelman et al. 2014, App. B]. □

Here are some useful observations to help justify Theorem 3. Consider the scalar version of the Taylor-series expansion (1):

$$\ln f_{\Theta|\boldsymbol{X}}(\theta \mid \boldsymbol{x}) \approx \ln f_{\Theta|\boldsymbol{X}}(\widehat{\theta}_{\text{MAP}} \mid \boldsymbol{x}) + 0.5(\theta - \widehat{\theta}_{\text{MAP}})^2 \frac{\mathrm{d}^2 \ln f_{\Theta|\boldsymbol{X}}\left(\widehat{\theta}_{\text{MAP}} \mid \boldsymbol{x}\right)}{\mathrm{d}\theta^2}.$$

Now, study the behavior of the negative Hessian of the log posterior pdf at θ_0:

$$\begin{aligned} -\frac{\mathrm{d}^2 \ln f_{\Theta|\boldsymbol{X}}(\theta_0 \mid \boldsymbol{x})}{\mathrm{d}\theta^2} &= -\frac{\mathrm{d}^2 \ln f_{\Theta}(\theta_0)}{\mathrm{d}\theta^2} - \frac{\mathrm{d}^2 \ln f_{\boldsymbol{X}|\Theta}(\boldsymbol{x} \mid \theta_0)}{\mathrm{d}\theta^2} \\ &= -\frac{\mathrm{d}^2 \ln f_{\Theta}(\theta_0)}{\mathrm{d}\theta^2} - \sum_{n=0}^{N-1} \frac{\mathrm{d}^2 \ln f_{X|\Theta}(x[n] \mid \theta_0)}{\mathrm{d}\theta^2} \end{aligned}$$

and, therefore,

$$\mathrm{E}_{\boldsymbol{X}|\Theta}\left[-\frac{\mathrm{d}^2 \ln f_{\Theta|\boldsymbol{X}}(\theta_0 \mid \boldsymbol{x})}{\mathrm{d}\theta^2} \,\middle|\, \theta_0 \right] = -\frac{\mathrm{d}^2 \ln f_{\Theta}(\theta_0)}{\mathrm{d}\theta^2} + N\mathcal{I}_1(\theta_0)$$

implying that, as N grows,

$$-\frac{\mathrm{d}^2 \ln f_{\Theta|\boldsymbol{X}}(\theta_0|\boldsymbol{x})}{\mathrm{d}\theta^2} \approx N\mathcal{I}_1(\theta_0).$$

To summarize: For a large number of i.i.d. measurements (i.e., asymptotically), the MAP and maximum-likelihood (ML) estimates give equivalent answers.

References

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin (2014). *Bayesian Data Analysis*. 3rd ed. Boca Raton, FL: Taylor & Francis (cit. on pp. 1, 5).