dissertation_SS

Erna Kuginyte

2023-07-27

Laplace Approximation

This chapter relies on Friston et al. [-@Friston2007].

The Laplace approximation is used to approximate the integral of a function $f(\theta)$ by fitting a Gaussian distribution at the maximum, $\hat{\theta}$, of $f(\theta)$, and then calculating the volume under the Gaussian. The covariance of the Gaussian is determined by the Hessian matrix of $\log f(\theta)$ evaluated at the maximum point $\hat{\theta}$ [@Mackay1998].

Laplace approximation is rooted in asymptotic analysis, where it is used to find approximate solutions to problems as parameters approach some asymptotic limit. As the data size N approaches infinity, the asymptotic series rapidly converges to a solution. This makes the Laplace approximation a computationally efficient method for analysing the posterior distribution, especially in cases where the normalising constant is difficult to evaluate.

Let an N-dimensional parameter vector be denoted $\boldsymbol{\theta}$, with prior distribution $f(\boldsymbol{\theta})$ and likelihood function $f(\mathbf{x}|\boldsymbol{\theta})$. Define $h(\boldsymbol{\theta}) = f(\boldsymbol{\theta})f(\mathbf{x}|\boldsymbol{\theta})$. There exists a $\boldsymbol{\theta}^*$ at which $h(\boldsymbol{\theta})$ is maximised. The Laplace Method approximates integrals of the form:

$$\int_a^b e^{N \cdot f(y)} dy, \tag{1}$$

where N is large, by fitting a Gaussian around $\boldsymbol{\theta}^*$ and computing the integral under this Gaussian.

Laplace approximation is applied as:

1. *Identify the mode* of $f(x|\theta)$, denoted as $\theta^* = \arg\max_\theta \ln(f(x|\theta))$. This is the point where the function $f(x|\theta)$ reaches its maximum.
2. *Compute the curvature at the mode*, using the Hessian matrix of $\ln(f(x|\theta))$ evaluated at $\theta*$. The Hessian matrix, $\nabla\nabla \ln(f(x|\theta^*))$, is a square matrix of second-order partial derivatives of the function.
3. In the context of the Laplace approximation, the mode and the negative of the curvature at the mode represent the mean and the variance of the *Gaussian approximation* to the posterior distribution, respectively.

For a general differentiable function $G(X)$, where $X = (x_1, x_2, ..., x_m)$, the gradient of $G(X)$ is given by

$$\nabla G(X) = \left(\frac{\partial G(X)}{\partial x_1}, ..., \frac{\partial G(X)}{\partial x_m} \right). \tag{2}$$

LA is often used to approximate a posterior distribution with a Gaussian distribution centred at the Maximum a Posteriori (MAP) estimate. This application of the Laplace method is justified by the fact that under certain conditions, the posterior distribution approaches a Gaussian as the number of samples increases [@Gelman2020].

(However, despite using a full distribution to approximate the posterior, the Laplace approximation shares many of the limitations of MAP estimation. For instance, estimating the variances at the end of an iterative learning process does not improve the approximation if the procedure has already led to an area of low probability mass.)

In practice, the LA makes use of the first-order Taylor series expansion:

$$h(\theta) = h(\mu_i) + \frac{\partial h(\mu_i)}{\partial \theta}(\theta - \mu_i), \tag{3}$$

where the expansion is around a solution, θ_L, obtained by an optimisation algorithm. The MAP solution is typically used for this purpose:

$$\theta_{MAP} = \arg\max_{\theta}[p(y|\theta, m)p(\theta|m)] \tag{4}$$

Thus, $\theta_L = \theta_{MAP}$. The model non-linearity is approximated using $h(\theta) = h(\theta_L) + J(\theta - \theta_L)$, where $J = \frac{\partial h(\theta_L)}{\partial \theta}$. The posterior covariance is given by $C_L^{-1} = J^T C_e^{-1} J + C_p^{-1}$.

The Laplace approximation is also used to compute the model evidence, which is crucial for Bayesian model comparison. The log-evidence under the Laplace approximation is given by:

$$\begin{aligned} \log p(y|m)_L = & \frac{-N_s}{2}\log 2\pi - \frac{1}{2}\log|C_e| - \frac{1}{2}\log|C_p| \\ & + \frac{1}{2}\log|C_L| - \frac{1}{2}r(\theta_L)^T C_e^{-1} r(\theta_L) \\ & - \frac{1}{2}e(\theta_L)^T C_p^{-1} e(\theta_L) \end{aligned} \tag{5}$$

When comparing models, we can ignore the first term as it is constant across models. Rearranging gives the trade-off between accuracy and complexity in model comparison:

$$\log p(y|m)_L = \text{Accuracy}(m) - \text{Complexity}(m) \tag{6}$$

where

$$\begin{aligned} \text{Accuracy}(m) &= -\frac{1}{2}\log|C_e| - \frac{1}{2}r(\theta_L)^T C_e^{-1} r(\theta_L), \\ \text{Complexity}(m) &= \frac{1}{2}\log|C_p| - \frac{1}{2}\log|C_L| + \frac{1}{2}e(\theta_L)^T C_p^{-1} e(\theta_L). \end{aligned}$$

The complexity term depends on the prior covariance, C_p, which determines the 'cost' of parameters. This can lead to biases in model comparisons if the prior covariances are fixed a priori. For instance, if the prior covariances are set to large values, model comparison will consistently favour simpler models.

It is essential to acknowledge the limitations of LA: it is ineffective for multi-modal posteriors, especially when modes are close together or when the mode of interest is far from the majority of the probability mass in high-dimensional spaces. LA also heavily relies on the smoothness assumption, making it less reliable for small sample sizes or when parameters are near the boundaries of the parameter space. LA does not account for global properties of the posterior distribution and is limited to parameters in $\mathbb{R}$. Practical strategies for addressing these limitations include gathering more data, if feasible, performing a log-transformation to reduce dimensionality, and reducing high-dimensional integrals to surface integrals.

In addition, LA method can also be extended to compare model fit, as a function of the posterior odds. The marginal likelihood can be approximated as:

$$f(x) = \int f(x|\theta)f(\theta)d\theta \approx (2\pi)^{N/2} \frac{f(x|\theta*)f(\theta*)}{|h_{\theta\theta}(\theta*)|^{1/2}} \tag{7}$$

where

$$h_{\theta\theta} = \left. \frac{-\partial^2 \ln[f(y|\theta)f(\theta)]}{\partial\theta\partial\theta^T} \right|_{\theta=\theta^*} \tag{8}$$

In this equation, $f(x|\theta)$ is the likelihood of the data given the parameters θ, $f(\theta)$ is the prior distribution of the parameters, and $h_{\theta\theta}$ is the Hessian matrix (the second derivative of the log-likelihood plus the log-prior) evaluated at the maximum a posteriori estimate θ^*. The integral on the left-hand side is the exact marginal likelihood, which is approximated by the term on the right-hand side using the Laplace approximation.

Application:

LaplacesDemon package: "Laplace Approximation is noted for its efficiency in runtime compared to Markov Chain Monte Carlo (MCMC) methods. Specifically, it is highlighted that the Laplace Approximation typically requires less time than MCMC due to its focus on seeking point-estimates rather than attempting to represent the target distribution through simulation draws. Furthermore, the method is found to be more adept at improving the objective function in instances where the parameters lie in low-probability regions compared to other methods such as iterative quadrature, MCMC, and Population Monte Carlo (PMC). However, caution is advised due to the Laplace Approximation's limitations, which include its asymptotic nature with respect to sample size and the assumption that marginal posterior distributions are Gaussian."

Application: LaplacesDemon package uses "When Method="SPG", a Spectral Projected Gradient algorithm is used. SPG is a non-monotone algorithm that is suitable for high-dimensional models. The approximate gradient is used, but the Hessian matrix is not. When used with large models, CovEst="Hessian" should be avoided. SPG has been adapted from the spg function in package BB."