Laplace Approximation

Advanced Statistical Inference

Simone Rossi

Where are we?

Given a likelihood $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ and a prior $p(\boldsymbol{\theta})$, the goal is to compute the posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y})$

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})}$$

. . .

We have seen that exact inference is possible if the prior and likelihood are *conjugate* distributions (e.g. for Bayesian linear regression with Gaussian likelihood and Gaussian prior).

. . .

But in most cases, the Bayesian approach is intractable.

- We cannot compute the posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ in closed form.
- We cannot make predictions for a new point $y_\star$ in closed form $p(y_\star \mid \boldsymbol{y}) = \int p(y_\star \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \boldsymbol{y}) \, \mathrm{d}\boldsymbol{\theta}$.

Last Week

We have seen that we can use **Monte Carlo** methods to approximate the posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y})$.

- **Rejection sampling**: sample from a simple distribution and accept/reject samples
- **Markov Chain Monte Carlo (MCMC)**: sample from a Markov chain that converges to the target distribution.
  - *Metropolis-Hastings*: do random walk in the parameter space and accept/reject samples based on the ratio of the target distribution in the current and next state.

– *Hamiltonian Monte Carlo (HMC)*: explore the parameter space by simulating a particle with a potential given by negative log posterior and a random momentum.

. . .

This Week

In this week, we will see another way to approximate the posterior distribution $p(\theta \mid y)$.

Can we use a simpler distribution to approximate the posterior?

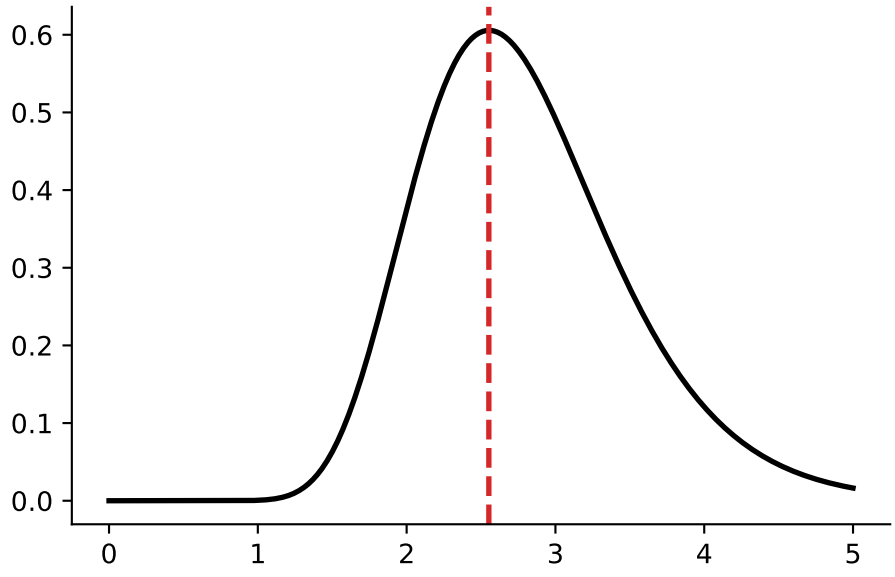
And if so, **how?**

Laplace Approximation

Laplace Approximation

Idea:

- Let's find a way to have a local approximation of the posterior distribution $p(\theta \mid y)$ around the mode $\widehat{\theta}$.



Derivation of Laplace Approximation

Given a likelihood $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ and a prior $p(\boldsymbol{\theta})$, the posterior distribution is given by

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{1}{Z} \exp\left(-\mathcal{U}(\boldsymbol{\theta})\right)$$

where Z is the normalizing constant and $\mathcal{U}(\boldsymbol{\theta}) = -\log p(\boldsymbol{y} \mid \boldsymbol{\theta}) - \log p(\boldsymbol{\theta})$ is the (un-normalized) negative log posterior.

. . .

Taylor expand $\mathcal{U}(\boldsymbol{\theta})$ around the mode $\widehat{\boldsymbol{\theta}}$ to the second order

$$\mathcal{U}(\boldsymbol{\theta}) \approx \mathcal{U}(\widehat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^\top \boldsymbol{g} + \frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^\top \boldsymbol{H} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})$$

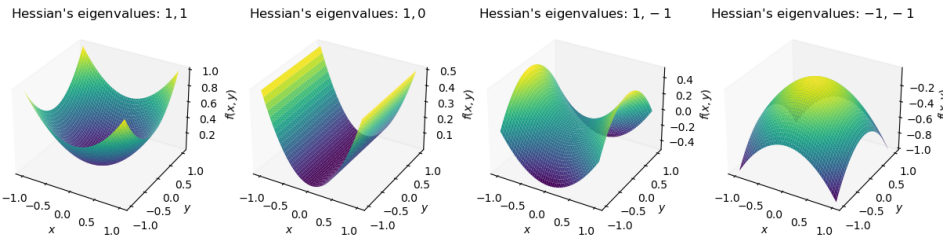
where:

- $\boldsymbol{g} = \boldsymbol{\nabla} \mathcal{U}(\widehat{\boldsymbol{\theta}})$ is the gradient
- $\boldsymbol{H} = \boldsymbol{\nabla}^2 \mathcal{U}(\widehat{\boldsymbol{\theta}})$ is the Hessian matrix

Hessians

The Hessian matrix $\boldsymbol{H}$ is a square matrix $D \times D$ of second-order partial derivatives of $\mathcal{U}(\boldsymbol{\theta})$ computed in $\boldsymbol{\theta}$.

$$\boldsymbol{H} = \boldsymbol{\nabla}^2 \mathcal{U}(\boldsymbol{\theta}) \begin{bmatrix} \frac{\partial^2 \mathcal{U}}{\partial \theta_1^2}(\boldsymbol{\theta}) & \frac{\partial^2 \mathcal{U}}{\partial \theta_1 \partial \theta_2}(\boldsymbol{\theta}) & \cdots & \frac{\partial^2 \mathcal{U}}{\partial \theta_1 \partial \theta_D}(\boldsymbol{\theta}) \\ \frac{\partial^2 \mathcal{U}}{\partial \theta_2 \partial \theta_1}(\boldsymbol{\theta}) & \frac{\partial^2 \mathcal{U}}{\partial \theta_2^2}(\boldsymbol{\theta}) & \cdots & \frac{\partial^2 \mathcal{U}}{\partial \theta_2 \partial \theta_D}(\boldsymbol{\theta}) \\ \cdots & \cdots & \ddots & \cdots \\ \frac{\partial^2 \mathcal{U}}{\partial \theta_D \partial \theta_1}(\boldsymbol{\theta}) & \frac{\partial^2 \mathcal{U}}{\partial \theta_D \partial \theta_2}(\boldsymbol{\theta}) & \cdots & \frac{\partial^2 \mathcal{U}}{\partial \theta_D^2}(\boldsymbol{\theta}) \end{bmatrix}$$



Derivation of Laplace Approximation

$$\mathcal{U}(\theta) \approx \mathcal{U}(\widehat{\theta}) + (\theta - \widehat{\theta})^\top \boldsymbol{g} + \frac{1}{2}(\theta - \widehat{\theta})^\top \boldsymbol{H}(\theta - \widehat{\theta})$$

- The mode $\widehat{\theta}$ is the value of θ that maximizes $p(\theta \mid \boldsymbol{y})$ or equivalently minimizes $\mathcal{U}(\theta)$.
- Because we are expanding around the mode, $\boldsymbol{g} = \boldsymbol{\nabla}\mathcal{U}(\widehat{\theta}) = 0$.

Approximate the posterior distribution with this expansion:

$$\begin{aligned} p(\theta \mid \boldsymbol{y}) &\approx \frac{1}{Z} \exp\left(-\mathcal{U}(\widehat{\theta}) - \frac{1}{2}(\theta - \widehat{\theta})^\top \boldsymbol{H}(\theta - \widehat{\theta})\right) \\ &= \frac{1}{Z} \exp\left(-\mathcal{U}(\widehat{\theta})\right) \exp\left(-\frac{1}{2}(\theta - \widehat{\theta})^\top \boldsymbol{H}(\theta - \widehat{\theta})\right) \end{aligned}$$

Derivation of Laplace Approximation

$$p(\theta \mid \boldsymbol{y}) \approx \frac{1}{Z} \exp\left(-\mathcal{U}(\widehat{\theta})\right) \exp\left(-\frac{1}{2}(\theta - \widehat{\theta})^\top \boldsymbol{H}(\theta - \widehat{\theta})\right)$$

Compare this to the form of a Gaussian distribution:

$$\mathcal{N}(\theta \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2}(2\pi)^{D/2} \det(\boldsymbol{\Sigma})^{-1/2} \exp\left(-\frac{1}{2}(\theta - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\theta - \boldsymbol{\mu})\right)$$

Taylor expansion of the log posterior $\Rightarrow$ Gaussian approximation of the posterior.

- The mean of the Gaussian is the mode of the posterior $\widehat{\theta}$.
- The covariance of the Gaussian is the inverse of the Hessian $\boldsymbol{H}^{-1}$.
- The normalizing constant is $Z = (2\pi)^{D/2} \det(\boldsymbol{H})^{-1/2} \exp\Big(-\mathcal{U}(\widehat{\theta})\Big)$.

Practical Considerations

1. The mode $\widehat{\boldsymbol{\theta}}$ can be found by gradient-based optimization

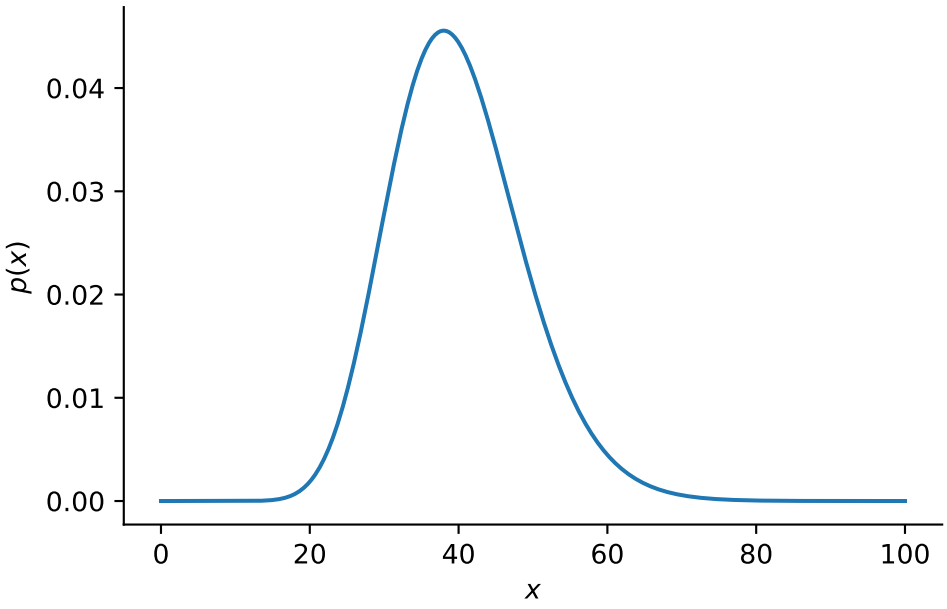
$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \boldsymbol{\nabla} \mathcal{U}(\boldsymbol{\theta}) \quad \text{until convergence}$$

2. The exact Hessian $\boldsymbol{H}$ can be expensive to compute and invert for large D.
   - Use a diagonal approximation of the Hessian or a block-diagonal approximation.
3. The Laplace approximation is only valid in a neighborhood of the mode.

Laplace Approximation: Example

Consider a Gamma distribution with shape parameter α and rate parameter β.

$$p(x \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)$$



Laplace Approximation: Example

1. Write the negative log posterior $\mathcal{U}(x)$.

$$\mathcal{U}(x) = -(\alpha - 1)\log x + \beta x$$

2. Compute the mode $\widehat{x}$ by setting $\frac{\mathrm{d}\mathcal{U}(x)}{\mathrm{d}x} = 0$.

$$\frac{\mathrm{d}\mathcal{U}(x)}{\mathrm{d}x} = -\frac{\alpha - 1}{x} + \beta = 0 \quad \Rightarrow \quad \widehat{x} = \frac{\alpha - 1}{\beta}$$

3. Compute the Hessian $\frac{\mathrm{d}^2\mathcal{U}(x)}{\mathrm{d}x^2}$ in $x = \widehat{x}$.

$$\frac{\mathrm{d}^2\mathcal{U}(x)}{\mathrm{d}x^2} = \frac{\alpha - 1}{x^2}\bigg|_{x=\widehat{x}} = \frac{\alpha - 1}{(\alpha - 1)^2/\beta^2} = \frac{\beta^2}{\alpha - 1}$$

4. Set up the Gaussian approximation of the posterior.

$$p(x \mid \alpha, \beta) \approx \mathcal{N}\left(x \mid \frac{\alpha - 1}{\beta}, \frac{\alpha - 1}{\beta^2}\right)$$

Final considerations

Benefits

1. It is relatively simple to implement and understand.

Limitations

1. It only captures local curvature around the mode $\widehat{\boldsymbol{\theta}}$.
2. It assumes the posterior is approximately Gaussian, which may not hold in practice (e.g., multimodal posteriors).
3. It requires expensive computations:
   - Finding the mode $\widehat{\boldsymbol{\theta}}$ via optimization (if not available in closed form).
   - Computing and inverting the Hessian $\boldsymbol{H}$ to get the covariance (at least cubic in the number of parameters).

In practice

1. Laplace approximation is useful if a "pre-trained" model is available (e.g., a neural network shared on Huggingface) and we want to perform Bayesian inference around the learned parameters.
2. The Hessian is often approximated with a diagonal or block-diagonal matrix to reduce computational costs.