

Biostatistics 202C: Weeks 3 and 4

Andrew Holbrook

Fall 2022

1 Normal approximations to posterior

Again, we have $\mathbf{y} = y_1, \dots, y_N \stackrel{iid}{\sim} p(y|\theta)$. By independence, the likelihood takes the form

$$p(\mathbf{y}|\theta) = \prod_{n=1}^N p(y_n|\theta).$$

Defining the MLE

$$\hat{\theta}_N = \arg \max_{\theta} p(\mathbf{y}|\theta),$$

it can be shown that (under regularity conditions)

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \xrightarrow{D} N(0, \mathcal{I}^{-1}(\theta_0)),$$

where θ_0 is the true parameter value and $\mathcal{I}(\theta_0)$ is the Fisher information

$$\mathcal{I}(\theta_0) = -E_{y|\theta} \left(\frac{\partial^2}{\partial \theta^2} \log p(y|\theta_0) \right).$$

Convergence in distribution leads to the approximation

$$\hat{\theta}_N \sim N \left(\theta, \frac{1}{N} \mathcal{I}^{-1}(\theta_0) \right).$$

Convergence in distribution also implies $\hat{\theta}_N \xrightarrow{P} \theta_0$, so we can use Slutsky's theorem to write the more useful result

$$\hat{\theta}_N \sim N \left(\theta, \frac{1}{N} \mathcal{I}^{-1}(\hat{\theta}_N) \right).$$

In machine learning, it is common practice to use optimization to obtain the posterior mode or MAP (maximum a posteriori) estimator

$$\bar{\theta} = \arg \max_{\theta} p(\theta|\mathbf{y})$$

use the MAP estimator in the context of the normal approximation¹

$$\theta|\mathbf{y} \sim N(\bar{\theta}, \bar{V}),$$

where

$$\bar{V} = \left(-\frac{\partial^2}{\partial \theta^2} \log p(\theta|\mathbf{y}) \Big|_{\theta=\bar{\theta}} \right)^{-1}.$$

¹Note that, assuming regularity conditions, the Bernstein-von Mises theorem (see previous lecture) says that a similar normal approximation about the MLE will also prove accurate.

We can write the second derivative as

$$\frac{\partial^2}{\partial \theta^2} \log p(\theta|\mathbf{y}) = \frac{\partial^2}{\partial \theta^2} \log p(\mathbf{y}|\theta) + \frac{\partial^2}{\partial \theta^2} \log p(\theta) - \frac{\partial^2}{\partial \theta^2} \log p(\mathbf{y}) = \frac{\partial^2}{\partial \theta^2} \log p(\mathbf{y}|\theta) + \frac{\partial^2}{\partial \theta^2} \log p(\theta).$$

It is then reasonable to quantify the influence of the prior using the fraction of information in the prior (FIP)

$$\text{FIP} = \frac{\frac{\partial^2}{\partial \theta^2} \log p(\theta)}{\frac{\partial^2}{\partial \theta^2} \log p(\theta|\mathbf{y})} \Big|_{\theta=\bar{\theta}},$$

or, when θ is a vector:

$$\text{FIP} = \text{tr} \left[\left(\frac{\partial^2}{\partial \theta \partial \theta^T} \log p(\theta) \right) \left(\frac{\partial^2}{\partial \theta \partial \theta^T} \log p(\theta|\mathbf{y}) \right)^{-1} \right] \Big|_{\theta=\bar{\theta}}.$$

2 Variational inference (VI)

Recall that the KL divergence between two probability distributions is defined

$$D_{KL}(p||q) = \int p(\theta) \log \left(\frac{p(\theta)}{q(\theta)} \right) d\theta.$$

Gibbs' inequality says

$$D_{KL}(p||q) \geq 0,$$

and $D_{KL}(p||q) = 0$ if and only if $p = q$. To see this, note that $\log(x) \leq x - 1$ for all $x > 0$, and so

$$\begin{aligned} D_{KL}(p||q) &= - \int p(\theta) \log \left(\frac{q(\theta)}{p(\theta)} \right) d\theta \geq - \int p(\theta) \left(\frac{q(\theta)}{p(\theta)} - 1 \right) d\theta \\ &= - \int q(\theta) d\theta + \int p(\theta) d\theta \geq 0. \end{aligned}$$

In VI, we approximate the posterior distribution $p(\theta|\mathbf{y})$ with a distribution $q(\theta|\psi)$, where ψ are ‘variational parameters’ we optimize thus:

$$\begin{aligned} \hat{\psi} &= \arg \min_{\psi} D_{KL}(q(\theta|\psi)||p(\theta|\mathbf{y})) = \arg \min_{\psi} \int q(\theta|\psi) \log \left(\frac{q(\theta|\psi)}{p(\theta|\mathbf{y})} \right) d\theta \\ &= \arg \min_{\psi} \int q(\theta|\psi) \log \left(\frac{q(\theta|\psi)p(\mathbf{y})}{p(\theta, \mathbf{y})} \right) d\theta \\ &= \arg \min_{\psi} \int q(\theta|\psi) (\log q(\theta|\psi) - \log p(\theta, \mathbf{y})) d\theta + \int q(\theta|\psi) \log p(\mathbf{y}) d\theta \\ &= \arg \min_{\psi} \int q(\theta|\psi) (\log q(\theta|\psi) - \log p(\theta, \mathbf{y})) d\theta + \log p(\mathbf{y}) \\ &= \arg \min_{\psi} -\mathcal{L}(\psi, \mathbf{y}) + \log p(\mathbf{y}) \\ &= \arg \max_{\psi} \mathcal{L}(\psi, \mathbf{y}), \end{aligned}$$

where we define

$$\mathcal{L}(\psi, \mathbf{y}) := - \int q(\theta|\psi) (\log q(\theta|\psi) - \log p(\theta, \mathbf{y})) d\theta.$$

Rearranging terms, we note that

$$\log p(\mathbf{y}) = D_{KL}(q(\theta|\psi)||p(\theta|\mathbf{y})) + \mathcal{L}(\psi, \mathbf{y}),$$

and therefore

$$\log p(\mathbf{y}) \geq \mathcal{L}(\psi, \mathbf{y}).$$

For this reason, \mathcal{L} is sometimes referred to as the ELBO (Evidence Lower Bound).²

²Example on Prof. Shahbaba's notes.

3 Another approach

Alternatively, we may obtain an approximating distribution $q(\theta|\psi)$ by minimizing the (different!) objective function:

$$\hat{\psi} = \arg \min_{\psi} D_{KL}(p(\theta|\mathbf{y})||q(\theta|\psi)). \quad (1)$$

Let us assume that the approximating distribution is an exponential family distribution, i.e.,

$$q(\theta|\psi) = h(\theta) \exp \left(\sum_{\alpha} \psi_{\alpha} \phi_{\alpha}(\theta) - \Phi(\psi) \right).$$

Here, ψ_{α} are the elements of the natural parameter, ϕ_{α} are the elements of the sufficient statistic, and Φ is the log-partition function that enforces integration constraints. The solution to (1) is given by moment matching, i.e., $\hat{\psi}$ satisfies

$$\int q(\theta|\hat{\psi}) \phi_{\alpha}(\theta) d\theta = \int p(\theta|\mathbf{y}) \phi_{\alpha}(\theta) d\theta.$$

To see this, we use the fact that the log-partition function of an exponential family distribution satisfies

$$\frac{\partial}{\partial \psi_{\alpha}} \Phi(\psi) = \mathbb{E}_{\theta|\psi}(\phi_{\alpha}).$$

This itself may be seen by noting that

$$1 = \int q(\theta|\psi) d\theta = e^{-\Phi(\psi)} \int h(\theta) \exp \left(\sum_{\alpha} \psi_{\alpha} \phi_{\alpha}(\theta) \right) d\theta$$

and differentiating both sides:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \psi_{\alpha}} e^{-\Phi(\psi)} \int h(\theta) \exp \left(\sum_{\alpha} \psi_{\alpha} \phi_{\alpha}(\theta) \right) d\theta + e^{-\Phi(\psi)} \int h(\theta) \frac{\partial}{\partial \psi_{\alpha}} \exp \left(\sum_{\alpha} \psi_{\alpha} \phi_{\alpha}(\theta) \right) d\theta \\ &= -\frac{\partial}{\partial \psi_{\alpha}} \Phi(\psi) + \mathbb{E}_{\theta|\psi}(\phi_{\alpha}). \end{aligned}$$

We therefore have

$$\begin{aligned} \frac{\partial}{\partial \psi_{\alpha}} D_{KL}(p(\theta|\mathbf{y})||q(\theta|\psi)) &= \frac{\partial}{\partial \psi_{\alpha}} \int p(\theta|\mathbf{y}) (\log p(\theta|\mathbf{y}) - \log q(\theta|\psi)) d\theta \\ &= - \int p(\theta|\mathbf{y}) \frac{\partial}{\partial \psi_{\alpha}} \log q(\theta|\psi) d\theta \\ &= - \int p(\theta|\mathbf{y}) \frac{\partial}{\partial \psi_{\alpha}} \left(\sum_{\alpha} \psi_{\alpha} \phi_{\alpha}(\theta) - \Phi(\psi) \right) d\theta \\ &= - \int p(\theta|\mathbf{y}) \left(\phi_{\alpha}(\theta) - \frac{\partial}{\partial \psi_{\alpha}} \Phi(\psi) \right) d\theta \\ &= - \int p(\theta|\mathbf{y}) (\phi_{\alpha}(\theta) - \mathbb{E}_{\theta|\psi}(\phi_{\alpha})) d\theta \\ &= - \int p(\theta|\mathbf{y}) \phi_{\alpha}(\theta) d\theta - \mathbb{E}_{\theta|\psi}(\phi_{\alpha}). \end{aligned}$$

Setting this derivative to 0 leads to the moment matching solution to (1):

$$\int q(\theta|\hat{\psi}) \phi_{\alpha}(\theta) d\theta = \int p(\theta|\mathbf{y}) \phi_{\alpha}(\theta) d\theta.$$

3.1 Assumed density filtering

Note that the posterior distribution may be written as the product of N ‘compatibility functions’

$$p(\theta|\mathbf{y}) \propto p(\theta) \prod_{n=1}^N p(y_n|\theta) = \prod_{n=0}^N p_n(\theta).$$

where $p_0(\theta) := p(\theta)$ and $p_n(\theta) := p(y_n|\theta)$. Assumed density filtering (ADF) constructs an approximation of $p(\theta|\mathbf{y})$ by first finding

$$\psi^0 = \arg \min_{\psi} D_{KL}(p_0(\theta)||q(\theta|\psi))$$

and then, in order, iteratively updating the approximation by finding

$$\psi^n = \arg \min_{\psi} D_{KL}(p_n(\theta)q(\theta|\psi^{n-1})||q(\theta|\psi)).$$

When the prior $p(\theta)$ is a member of the exponential family $q(\theta|\psi)$, the initial update ψ^0 is given by the prior parameters. Subsequent ψ^n are obtained using moment matching.

ADF thus uses the current approximation to help guide the construction of the next approximation and generally performs better than taking the product of N independent approximations. On the other hand, the method is sensitive to the ordering and can be thrown off by bad starting approximations.

3.2 Expectation propagation

Expectation propagation (EP) iteratively improves an approximation to the posterior $q(\theta|\psi) \propto \prod_n r_n(\theta)$. The algorithm begins by setting each $r_n = 1$ and performs the following steps until convergence:

1. Randomly select $n \sim \text{Unif}\{0, 1, \dots, N\}$;
2. Remove r_n from current posterior by dividing and normalizing

$$q(\theta|\psi^{-n}) \propto \frac{q(\theta|\psi)}{r_n(\theta)}.$$

Note that this ‘cavity distribution’ is in the same exponential family distribution if r_n is in it or is constant (although it might not be normalizable).

3. Update exponential family approximation to posterior $q(\theta|\psi^*)$ by finding

$$\psi^* = \arg \min_{\psi} D_{KL}(p_n(\theta)q(\theta|\psi^{-n})||q(\theta|\psi)).$$

Again, this step performed by moment matching.

4. Update exponential family approximation of r_n as

$$r_n(\theta) \propto \frac{q(\theta|\psi^*)}{q(\theta|\psi^{-n})}.$$

Note that this step enforces r_n ’s membership within the (not necessarily normalizable) exponential family.

EP often outperforms ADU, but there is no guarantee of convergence. Fixed points exist for approximation distributions belonging to exponential family. Lack of normalization guarantees means possibility of, e.g., negative variances for normal approximations.

3.3 The clutter problem

Suppose we have Gaussian observations $\mathbf{Y} = \mathbf{y}_1, \dots, \mathbf{y}_N$ in a ‘sea of unrelated clutter’, which we model using the mixture distribution

$$p(\mathbf{y}|\boldsymbol{\theta}) = (1 - w)\mathcal{N}(\mathbf{y}|\boldsymbol{\theta}, \mathbf{I}_D) + w\mathcal{N}(\mathbf{y}|\mathbf{0}, 10\mathbf{I}_D).$$

We also specify the prior $p(\boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{0}, 100\mathbf{I}_D)$. Let $p_0(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$, and $p_n(\boldsymbol{\theta}) := p(\mathbf{y}_n|\boldsymbol{\theta})$ for $n > 0$. Finally, we specify the Gaussian approximation distribution $q(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{m}_\theta, v_\theta\mathbf{I}_D)$.

3.3.1 Assumed density filtering

At iteration 0, we minimize the KL divergence between p_0 and q , subject to the restriction that q be Gaussian with mean \mathbf{m}_θ and covariance $v_\theta \mathbf{I}_D$. Thus, iteration 0 sets q to be equal to the prior $p = p_0$. At each subsequent n th step, we obtain the ‘exact posterior’

$$\hat{p}(\boldsymbol{\theta}) = \frac{p_n(\boldsymbol{\theta})q^{-n}(\boldsymbol{\theta})}{\int p_n(\boldsymbol{\theta})q^{-n}(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

and minimize $D_{KL}(\hat{p}||q)$ subject to the spherical Gaussian specification of q . Due to the fact that the Gaussian distribution belongs to the exponential family, this solution is obtained by moment matching:

$$\begin{aligned} \mathbb{E}_q(\boldsymbol{\theta}) &= \mathbb{E}_{\hat{p}}(\boldsymbol{\theta}), \\ \mathbb{E}_q(\boldsymbol{\theta}^T \boldsymbol{\theta}) &= \mathbb{E}_{\hat{p}}(\boldsymbol{\theta}^T \boldsymbol{\theta}). \end{aligned}$$

Each step also produces the normalization factor

$$z_n = \int p_n(\boldsymbol{\theta})q^{-n}(\boldsymbol{\theta})d\boldsymbol{\theta},$$

and we use the product of these factors to estimate $p(\mathbf{Y})$. Here, we have

$$z_n = (1 - w)\mathcal{N}(\mathbf{y}_n|\mathbf{m}_\theta^{-n}, (v_\theta^{-n} + 1)\mathbf{I}_D) + w\mathcal{N}(\mathbf{y}_n|\mathbf{0}, 10\mathbf{I}_D).$$

For the clutter problem, the ADF algorithm follows these steps:

1. Initialize $\mathbf{m}_\theta = \mathbf{0}$, $v_\theta^{-n} = 100$, $s = 1$;
2. For $n = 1, \dots, N$, update $(\mathbf{m}_\theta, v_\theta, s)$ according to
 - (a) $s = s^{-n} z_n$;
 - (b) $\pi_n = 1 - \frac{w}{z_n} \mathcal{N}(\mathbf{y}_n|\mathbf{0}, 10\mathbf{I}_D)$;
 - (c) $\mathbf{m}_\theta = \mathbf{m}_\theta^{-n} + v_\theta^{-n} \pi_n \frac{\mathbf{y}_n - \mathbf{m}_\theta^{-n}}{v_\theta^{-n} + 1}$;
 - (d) $v_\theta = v_\theta^{-n} - \pi_n \frac{(v_\theta^{-n})^2}{v_\theta^{-n} + 1} + \pi_n (1 - \pi_n) \frac{(v_\theta^{-n})^2 \|\mathbf{y}_n - \mathbf{m}_\theta^{-n}\|^2}{D(v_\theta^{-n} + 1)^2}$.

3.3.2 Expectation propagation

For the same problem, the EP term approximations take the form

$$r_n(\boldsymbol{\theta}) = s_n \exp \left(-\frac{1}{2v_n} (\boldsymbol{\theta} - \mathbf{m}_n)^T (\boldsymbol{\theta} - \mathbf{m}_n) \right),$$

where $q(\boldsymbol{\theta}) \propto \prod_n r_n(\boldsymbol{\theta})$. The EP algorithm proceeds as follows.

1. Initialize the prior terms $v_0 = 100$, $\mathbf{m}_0 = \mathbf{0}$, $s_0 = (2\pi v_0)^{-D/2}$ and the data terms so that $r_n(\boldsymbol{\theta}) \propto 1$: $v_n = \infty$, $\mathbf{m}_n = \mathbf{0}$ and $s_n = 1$;
2. $\mathbf{m}_\theta = \mathbf{m}_0$, $v_\theta = v_0$;
3. Until all (\mathbf{m}_n, v_n, s_n) converge, loop $n = 1, \dots, N$:
 - (a) Remove r_n from the posterior to get ‘old’ posterior:

$$\begin{aligned} (v_\theta^{-n})^{-1} &= v_\theta^{-1} - v_n^{-1} \\ \mathbf{m}_\theta^{-n} &= \mathbf{m}_\theta + \frac{v_\theta^{-n}}{v_n} (\mathbf{m}_\theta - \mathbf{m}_n); \end{aligned}$$

- (b) Recompute $(\mathbf{m}_\theta, v_\theta, z_n)$ from $(\mathbf{m}_\theta^{-n}, v_\theta^{-n})$ as in ADF:
 - i. $z_n = (1 - w)\mathcal{N}(\mathbf{y}_n|\mathbf{m}_\theta^{-n}, (v_\theta^{-n} + 1)\mathbf{I}_D) + w\mathcal{N}(\mathbf{y}_n|\mathbf{0}, 10\mathbf{I}_D)$;

- ii. $\pi_n = 1 - \frac{w}{z_n} \text{N}(\mathbf{y}_n | \mathbf{0}, 10\mathbf{I}_D);$
- iii. $\mathbf{m}_\theta = \mathbf{m}_\theta^{-n} + v_\theta^{-n} \pi_n \frac{\mathbf{y}_n - \mathbf{m}_\theta^{-n}}{v_\theta^{-n} + 1};$
- iv. $v_\theta = v_\theta^{-n} - \pi_n \frac{(v_\theta^{-n})^2}{v_\theta^{-n} + 1} + \pi_n (1 - \pi_n) \frac{(v_\theta^{-n})^2 \|\mathbf{y}_n - \mathbf{m}_\theta^{-n}\|^2}{D(v_\theta^{-n} + 1)^2};$

(c) Update r_n :

$$\begin{aligned}
v_n^{-1} &= v_\theta^{-1} - (v_\theta^{-n})^{-1} \\
\mathbf{m}_n &= \mathbf{m}_\theta^{-n} + \frac{(v_n + v_\theta^{-n})}{v_\theta^{-n}} (\mathbf{m}_\theta - \mathbf{m}_\theta^{-n}) \\
s_n &= \frac{z_n}{(2\pi v_n)^{D/2} \text{N}(\mathbf{m}_n | \mathbf{m}_\theta^{-n}, (v_n + v_\theta^{-n})\mathbf{I})};
\end{aligned}$$

4. Compute normalizing constant (if you want it):

$$\begin{aligned}
V &= \frac{\mathbf{m}_\theta^T \mathbf{m}_\theta}{v_\theta} - \sum_n \frac{\mathbf{m}_n^T \mathbf{m}_n}{v_n} \\
p(\mathbf{Y}) &\approx (2\pi v_\theta)^{D/2} e^{V/2} \prod_{n=0}^N s_n.
\end{aligned}$$