MSA101/MVE187 2022 Lecture 13
Variational Bayes
Slice sampling

Petter Mostad

Chalmers University

October 10, 2022

Review / overview

- Last time: The EM algorithm: Using Kullback-Leibler divergence to find a maximal posterior estimate.
- This time, part 1: Variational Bayes: Using Kullback-Leibler divergence to find a density from some family of densities that optimally fits the posterior.
- Part 2: The slice sampler.
- If time: Final comments about MCMC.

The KL notation

- Recall:

$$\mathrm{KL}[q||p] = \mathrm{E}_q\left[-\log\frac{p(z)}{q(z)}\right] = -\int q(z)\log\frac{p(z)}{q(z)}\,dz$$

for any density $q(z)$ and *any positive function* $p(z)$ so that the integral exists. (For standard KL p must be a density).

- Consequence: If $p_2(z) = Cp_1(z)$ then for any q

$$\mathrm{KL}[q||p_2] = \mathrm{E}_q\left[-\log\frac{Cp_1(z)}{q(z)}\right] = -\log C + \mathrm{KL}[q||p_1].$$

- For example, if $\int p_2(z)\,dz = C$ then for any q

$$\mathrm{KL}[q||p_2] \geq -\log C$$

because $\mathrm{KL}[q||p_2/C] \geq 0$, with minimum occurring when $q \propto_z p_2$.

- Recall also that

$$\mathrm{KL}[q||p] = \mathrm{E}_q[-\log p(z)] - H_q[Z]$$

where $H_q[Z]$ is the entropy of a random variable Z with density q .

Example 1: The EM algorithm

- Consider the identity

$$\pi(x, z \mid \theta) = \pi(x \mid \theta)\pi_z(z \mid x, \theta).$$

  Considering this as a function of z, $\pi(x \mid \theta)$ is a constant.

- For any density q for z we get

$$\mathrm{KL}[q||\pi(x, \cdot \mid \theta)] = -\log \pi(x \mid \theta) + \mathrm{KL}[q||\pi_z(\cdot \mid x, \theta)]$$

- The above equation is in the core of the proof of the EM algorithm:
  - Set $q(z) = \pi_z(z \mid x, \theta^{OLD})$ for some θ^{OLD}.
  - Find a θ^{NEW} that minimizes the left-hand side.
  - Then, moving from θ^{OLD} to θ^{NEW}, the left-hand side will decrease, and $\mathrm{KL}(q||\pi_z(\cdot \mid x, \theta)]$ will increase. Thus $-\log \pi(x \mid \theta)$ will decrease.

Example 2: Approximating the posterior

Let's say we want to find a density q minimizing $\mathrm{KL}[q||\pi(\cdot \mid \text{data})]$

- In the identity

$$\pi(\text{data}, \theta) = \pi(\theta \mid \text{data})\pi(\text{data})$$

  $\pi(\text{data})$ is a constant as a function of θ.

- Thus for a density q for θ,

$$\mathrm{KL}[q||\pi(\text{data}, \cdot)] = -\log \pi(\text{data}) + \mathrm{KL}[q||\pi(\cdot \mid \text{data})].$$

- We may try to find a q minimizing $\mathrm{KL}[q||\pi(\cdot \mid \text{data})]$ by finding a q minimizing $\mathrm{KL}[q||\pi(\text{data}, \cdot)]$: This is part of the Variational Bayes idea.

Approximations using Variational Bayes

- Idea: Finding an approximation to the posterior $\pi(\theta \mid \text{data})$ in some family of densities $\mathcal{Q}$ that does not necessarily contain the posterior.
- More specifically find the $q \in \mathcal{Q}$ minimizing the Kullback Leibler divergence from q to the posterior.
- Writing as above

$$\mathrm{KL}[q||\pi(\text{data}, \cdot)] = -\log \pi(\text{data}) + \mathrm{KL}[q||\pi(\cdot \mid \text{data})].$$

  we instead find the $\hat{q}$ minimizing $\mathrm{KL}[q||\pi(\text{data}, \cdot)]$.
- As $\log \pi(\text{data}) \geq -\mathrm{KL}[q||\pi(\text{data}, \cdot)]$ the value $-\mathrm{KL}[\hat{q}||\pi(\text{data}, \cdot)]$ is called the *evidence lower bound*, or ELBO.
- Thus we want to *maximize*

$$\begin{aligned} \mathcal{L}(q) &= -\mathrm{KL}[q||\pi(\text{data}, \cdot)] = \int q(\theta) \log \frac{\pi(\text{data}, \theta)}{q(\theta)} \, d\theta \\ &= \mathrm{E}_q[\log \pi(\text{data}, \theta)] + H_q[\theta] \end{aligned}$$

  where $H_q[\theta]$ is the entropy of a variable θ with density q.

Splitting θ into components (or subvectors)

- Let $\mathcal{Q}_{\text{prod}}$ be the family of densities q that can be written as products

$$q(\theta) = \prod_{i=1}^{n} q_i(\theta_i)$$

where $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$ is split into (groups of) dimensions.

- For the entropy term we get that

$$H_q[\theta] = \sum_{i=1}^{n} H_{q_i}[\theta_i]$$

where θ_i are variables with densities q_i.

- For any $i \in 1, \ldots, n$ the first term of $\mathcal{L}(q)$ may be rewritten

$$\mathrm{E}_q[\log \pi(\text{data}, \theta)] = \mathrm{E}_{q_i}\left[\mathrm{E}_{q_j, j \neq i}\left[\log \pi(\text{data}, \theta)\right]\right]$$

- So if we fix all q_j with $j \neq i$, the optimal q_i maximizing $\mathcal{L}(q)$ is the q_i maximizing

$$\begin{aligned} & \mathrm{E}_{q_i}\left[\mathrm{E}_{q_j, j \neq i}\left[\log \pi(\text{data}, \theta)\right]\right] + H_{q_i}[\theta_i] \\ = \quad & -\mathrm{KL}\left[q_i \,||\, \exp\left(\mathrm{E}_{q_j, j \neq i}\left[\log \pi(\text{data}, \cdot)\right]\right)\right] \end{aligned}$$

First option: Solving simultaneous equations

- We have seen that $\mathrm{KL}\left[q_i || \exp\left(\mathrm{E}_{q_j, j \neq i}\left[\log \pi(\text{data}, \cdot)\right]\right)\right]$ is minimized when

$$q_i(\theta_i) \propto_{\theta_i} \exp\left(\mathrm{E}_{q_j, j \neq i}\left[\log \pi(\text{data}, \cdot)\right]\right)$$

- If we write out these n equations for $i = 1, \ldots, n$, they become n equations in the n unknowns $q_1, q_2, \ldots, q_n$.
- Sometimes it is possible to simultaneously solve these equations.
- The solution we get is then the density $q \in \mathcal{Q}_{\text{prod}}$ that minimizes $\mathrm{KL}[q || \pi(\text{data}, \cdot)]$.

Variational Bayes: Toy example

- Consider the following example:

$$\begin{aligned} y_1, \ldots, y_n &\sim \text{Normal}(\mu, \tau^{-1}) \\ \pi(\mu) &\propto 1 \\ \pi(\tau) &\propto 1/\tau \end{aligned}$$

- Using conjugacy, we get that the exact posterior is given by

$$\begin{aligned} \tau \mid y_1, \ldots, y_n &\sim \text{Gamma}\left(\frac{n-1}{2}, \frac{n-1}{2}s^2\right) \\ \mu \mid \tau, y_1, \ldots, y_n &\sim \text{Normal}\left(\overline{y}, (n\tau)^{-1}\right) \end{aligned}$$

  where s^2 is the sample variance.

- As an illustration, we find the Variational Bayes approximate posterior. Note:

$$\begin{aligned} \pi(y_1, \ldots, y_n, \mu, \tau) &\propto \frac{1}{\tau} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi/\tau}} \exp\left(-\frac{\tau}{2}(y_i - \mu)^2\right) \\ \log \pi(y_1, \ldots, y_n, \mu, \tau) &= C + \left(\frac{n}{2} - 1\right) \log \tau - \frac{\tau}{2}(n-1)s^2 - \frac{n\tau}{2}(\overline{y} - \mu)^2 \end{aligned}$$

Variational Bayes: Toy example continued

- We use as approximation for the posterior the family of densities $q(\mu, \tau) = q_1(\mu)q_2(\tau)$, so that we assume μ and τ are independent, but we do not make additional restrictions on q_1 and q_2.
- We get

$$\begin{aligned} & \exp\left(\mathsf{E}_\mu\left[\log \pi(\text{data}, \mu, \tau)\right]\right) \\ \propto_\tau \quad & \exp\left(\left(\frac{n}{2} - 1\right)\log\tau - \frac{\tau}{2}(n-1)s^2 - \frac{n\tau}{2}\,\mathsf{E}_\mu\left[(\overline{y} - \mu)^2\right]\right) \end{aligned}$$

- From this we see that

$$q_2(\tau) = \text{Gamma}\left(\tau; \frac{n}{2}, \frac{1}{2}(n-1)s^2 + \frac{n}{2}\,\mathsf{E}_\mu\left[(\overline{y} - \mu)^2\right]\right)$$

- We get

$$\exp\left(\mathsf{E}_\tau\left[\log \pi(\text{data}, \mu, \tau)\right]\right) \propto_\mu \exp\left(-\frac{n}{2}\,\mathsf{E}_\tau[\tau](\overline{y} - \mu)^2\right)$$

- From this we see that

$$q_1(\mu) = \text{Normal}\left(\mu; \overline{y}, \left(n\,\mathsf{E}_\tau[\tau]\right)^{-1}\right).$$

Variational Bayes: Toy example continued

- Taking expectations using these two densities leads to

$$\begin{aligned} \mathrm{E}_\tau[\tau] &= \frac{n/2}{(n-1)s^2/2 + n/2 \cdot \mathrm{E}_\mu\left[(\overline{y}-\mu)^2\right]} \\ \mathrm{E}_\mu\left[(\overline{y}-\mu)^2\right] &= \left(n\,\mathrm{E}_\tau[\tau]\right)^{-1} \end{aligned}$$

- This is two equations with two unknowns; solving gives

$$\begin{aligned} \mathrm{E}_\tau[\tau] &= \frac{1}{s^2} \\ \mathrm{E}_\mu\left[(\overline{y}-\mu)^2\right] &= \frac{s^2}{n} \end{aligned}$$

- The final solution is

$$\begin{aligned} q_2(\tau) &= \text{Gamma}\left(\tau; \frac{n}{2}, \frac{n}{2}s^2\right) \\ q_1(\mu) &= \text{Normal}\left(\mu; \overline{y}, \frac{s^2}{n}\right) \end{aligned}$$

Second option: Iterative solution

- Let us instead consider a family of densities $\mathcal{Q}_{\text{par}} \subset \mathcal{Q}_{\text{prod}}$ consisting of products of n densities where each factor is from some parametric family, and find the $q \in \mathcal{Q}_{\text{par}}$ minimizing $\text{KL}[q||\pi(\text{data}, \cdot)]$.
- Following the above, we start with a reasonable solution with factors $q_1, q_2, \ldots, q_n$, then cycle through them and find the q_i minimizing

$$\text{KL}\left[q_i || \exp\left(\text{E}_{q_j, j \neq i}\left[\log \pi(\text{data}, \cdot)\right]\right)\right]$$

  when all the q_j with $j \neq i$ are fixed.
- For each optimization, we optimize over the parameters of the q_i density.
- A very rough density approximation, but the method may scale well in very high dimensions.
- This is the *mean field* variational Bayes approximation of the posterior.

What if we minimize $\mathrm{KL}[\pi(\cdot \mid \text{data}) || q]$ instead of $\mathrm{KL}[q || \pi(\cdot \mid \text{data})]$?

- We have

$$\mathrm{KL}[\pi(\cdot \mid \text{data}) || q] = -\int \pi(\theta \mid \text{data}) \log \frac{q(\theta)}{\pi(\theta \mid \text{data})}\, d\theta$$

$$= \int \pi(\theta \mid \text{data}) \log \pi(\theta \mid \text{data})\, d\theta - \int \pi(\theta \mid \text{data}) \log q(\theta)\, d\theta$$

  so we only need to find the q maximizing the last term.
- If we assume that $q(\theta) = q(\theta \mid \eta) = \prod_{i=1}^{n} q_i(\theta_i \mid \eta_i)$ we get that

$$\int \pi(\theta \mid \text{data}) \log q(\theta \mid \eta)\, d\theta = \sum_{i=1}^{n} \int \pi(\theta \mid \text{data}) \log q_i(\theta_i \mid \eta_i)\, d\theta$$

$$= \sum_{i=1}^{n} \int \pi(\theta_i \mid \text{data}) \log q_i(\theta_i \mid \eta_i)\, d\theta_i.$$

  So we optimize by setting $q_i(\theta_i \mid \eta_i)$ equal to the marginal posterior $\pi(\theta_i \mid \text{data})$ for each i (or choose η_i to minimize the KL divergence).
- Less useful approximations in practice.

Part 2: The slice sampler

- Idea: Do Gibbs sampling from "the area under the density curve". (Illustrate)
- More formally, given density $f_x(x)$, simulate from the joint density

$$f(x,u) = I(0 < u < f_x(x))$$

- Works even if f_x is only proportional to a density.
- The challenge is to simulate x uniformly on $\{x : u < f_x(x)\}$. This is most easily done if for example f_x is a decreasing function, so that it is invertible.
- Example: Simulate from the density $\pi(x) = \frac{1}{2}\exp\left(-\sqrt{x}\right)$. We iterate between the following steps:
  - Given an x value, simulate $u \sim \text{Uniform}\left(0, \frac{1}{2}\exp\left(-\sqrt{x}\right)\right)$.
  - Given a u value simulate $x \sim \text{Uniform}\left(0, (\log(2u))^2\right)$: Note that $u = \frac{1}{2}\exp\left(-\sqrt{x}\right)$ if and only if $x = (\log(2u))^2$ and that $\pi(x)$ is decreasing as a function of x.

Generalization to product densities

- Importantly, the theory can easily be extended to densities that are products: When we want to simulate from the density

$$f(x) = \prod_{i=1}^{n} g_i(x)$$

we can define the joint density

$$h(x, u_1, \ldots, u_n) = \prod_{i=1}^{n} I\left(0 < u_i < g_i(x)\right)$$

- We see that the marginal density for x is $f(x)$.
- We simulate from the joint density using Gibbs sampling. This is very easy for the variables $u_1, \ldots, u_n$.
- The conditional distribution of x given $u_1, \ldots, u_n$ is the uniform distribution on the set

$$\cap_{i=1}^{n} \{x : u_i < g_i(x)\}.$$

If it is easy to compute this set, slice sampling works well. One example: If all the $g_i(x)$ functions are decreasing and invertible.

Example: The Challenger disaster

- The goal is to compute the probability that a space shuttle "o-ring" fails at a specific temperature. (An o-ring failing because of cold weather was the cause of the Challenger space shuttle disaster).
- Data $(x_1, y_1), \ldots, (x_n, y_n)$ where x_i denotes the temperature (in Farenheit) and y_i is 1 if there is a failure, 0 otherwise.
- We use a logistic regression model:

$$y_i \sim \text{Bernoulli}(p(x_i)) \qquad p(x_i) = \frac{\exp(a + bx_i)}{1 + \exp(a + bx_i)}.$$

- The posterior becomes (using flat priors on a and b)

$$\begin{aligned} \pi(a, b \mid \text{data}) &\propto \prod_{i=1}^{n} \left(\frac{\exp(a + bx_i)}{1 + \exp(a + bx_i)} \right)^{y_i} \left(\frac{1}{1 + \exp(a + bx_i)} \right)^{1-y_i} \\ &= \prod_{i=1}^{n} \frac{\exp(a + bx_i)^{y_i}}{1 + \exp(a + bx_i)} \end{aligned}$$

Example continued

- Simulate from posterior for parameters (a, b) using slice sampling:
  - For $i = 1, \ldots, n$, simulate $u_i \sim \text{Uniform}\left[0, \frac{\exp(a+bx_i)^{y_i}}{1+\exp(a+bx_i)}\right]$.
  - Simulate (a, b) uniformly on set satisfying, for all i, $u_i < \frac{\exp(a+bx_i)^{y_i}}{1+\exp(a+bx_i)}$.
- Corresponds to $a + bx_i > \log(u_i/(1-u_i))$ for i with $y_i = 1$, and $a + bx_i < \log((1-u_i)/u_i)$ for i with $y_i = 0$.
- To simulate (a, b) uniformly on this set, we first simulate a with

$$a \sim \text{Uniform}\left[\max_{y_i=1}\left(\log\frac{u_i}{1-u_i} - bx_i\right), min_{y_i=0}\left(\log\frac{1-u_i}{u_i} - bx_i\right)\right]$$

- Then for b, we need to be more careful, simulating b uniformly in the interval of numbers
  - Greater than $\left(\log\frac{u_i}{1-u_i} - a\right)/x_i$ for i with $y_i = 1$ and $x_i > 0$.
  - Smaller than $\left(\log\frac{u_i}{1-u_i} - a\right)/x_i$ for i with $y_i = 1$ and $x_i < 0$.
  - Smaller than $\left(\log\frac{1-u_i}{u_i} - a\right)/x_i$ for i with $y_i = 0$ and $x_i > 0$.
  - Greater than $\left(\log\frac{1-u_i}{u_i} - a\right)/x_i$ for i with $y_i = 0$ and $x_i < 0$.

Example continued

- This is actually Example 7.11 in RC, but the book contains some errors:
  - Confusion beween (a, b) and (α, β)
  - Second and fourth formulas on page 220 are wrong.
  - No need to use a prior for a and b to get this to work; use centering instead.
- Note that a and b are highly correlated in the posterior if we implement the code directly. Much improved convergence and accuracy is obtained by *centering* the data: Subtracting the average value from the temperature values, performing the analysis, and then adding back the average value.

MCMC: Summing up some tips and tricks

- Usually a good idea to compute with the logarithm of the posterior, instead of the posterior itself.
- Reparametrize all variables so that they are defined on the real line (if it is possible and convenient).
- Make sure your code avoids underflow and overflow numerical problems. Make sure a function computing (logged) posterior density will always return sensible answers for any values that might be proposed.
- Reparametrize the model, if possible and convenient, so that parameters are as uncorrelated as possible in the posterior. Otherwise, you may try out a random walk with correlated proposals.
- Do a normal approximation if convenient: A mode is nice to know, and the variances, and the covariance matrix, may be helpful for deciding step lengths in your MCMC! (Rule of thumb, two times standard deviation, does not always work).
- If available, use some classical anlaysis to find reasonable starting values for your parameters.
- Vary the starting point of the Markov chain! (Propose from prior?)
- For more complex models, tailored proposals may be necessary!

MCMC: Checking convergence

- We know the results from MCMC will be correct in the limit when the sample size $\to \infty$.
- Only in very special cases (e.g. using "coupling") do we know how big the sample size needs to be to get a certain accuracy.
- In practice "checking convergence" means checking for signs of non-convergence or slow convergence (slow "mixing"):
  - Monitor variable values and cumulative averages.
  - Check autocorrelations for variables.
  - Check acceptance rates (but higher is not always better, unless you are using independent proposals!)
  - Use multiple starting points for the MCMC chain!
  - Use multiple parallell chains, and compare variace within chains with variance between chains! (Special tests have been developed).
- An important ingredient is to *understand* your model and your posterior, so that you can guess what might cause convergence problems, and check for such problems.

Advantages with Metropolis Hastings

- Great flexibility: It will (in principle) work for any (posterior) density where the density function can be computed up to a constant.
- Great flexibility in the choice of proposal function $q(x \mid y)$.
- The algorithm is quite simple and can be easily programmed in many cases.

Some problems with Metropolis Hastings

- (Small issue): You need to make sure your proposal function makes the Markov chain ergodic.
- (Large issue): Even if the Markov chain converges, it may converge *too slowly for practical use*.
- (Large issue): Even if very many proposal functions work in theory, it may be quite difficult to find ones that lead to reasonably fast convergence.
- (Large issue): It is almost always impossible to prove results about convergence (and thus accuracy), and it is quite often difficult to ascertain how well a chain has converged.
- (Large?? issue): Convergence may become unacceptably slow when the dimension over which you simulate grows large.