T-61.5040

Learning Models and Methods, (5 cr) L

Spring 2007

Lecturer: **Petteri Pajunen**

Course Assistant: **Ville Viitaniemi**

LECTURE 8: 14.3.2007

POSTERIOR APPROXIMATION BY SIMULATION

Posterior Approximation: Simulation Methods

- the goal of Bayesian inference is to compute the posterior $p(\theta|y)$ in a usable form
- except in simple examples and conjugate models, posterior is not a standard distribution
- in general one can obtain the unnormalized posterior as

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

- the normalizing factor is

$$p(y) = \int p(\theta, y)d\theta = \int p(y|\theta)p(\theta)d\theta$$

- in almost any realistic, non-conjugate model, this integral is not obtained in closed-form

- most uses of the posterior require integrating over it
- computing the posterior mode does not, but then the mode does not contain any probability information
- integration is unavoidable if probabilities are to be computed
- in general, expectations of the form $\mathrm{E}(h(\theta)|y)$ are needed

- expectation is an *integral* over the posterior:

$$\mathrm{E}(h(\theta)|y) = \int h(\theta)p(\theta|y)d\theta$$

- examples of functions $h(\cdot)$:
  - posterior mean: $h(\theta) = \theta$
  - posterior variance: $h(\theta) = (\theta - \mu)^2, \ \mu = \mathrm{E}(\theta|y)$
  - probability of $\theta \in A$: $h(\theta) = 1$ when $\theta \in A$, zero otherwise

- various approximation methods can be used to compute the integrals
- next week, parameteric approximations are discussed
- the true posterior is approximated as some tractable parametric distribution, for example Normal distribution
- today, more general approximation is discussed

- *simulation*: for our purposes, drawing samples from the posterior distribution
- sort of "reverse estimation": in estimation, a distribution is unknown but data generated from it are known
- in simulation, we know the posterior as a function of θ and then generate data from it
- the simulated data carry information about the posterior

- as in estimation, the data can be used to approximate expectations over the posterior
- for example, drawing m samples from a Normal distribution $N(\mu, \sigma^2)$ can be used to estimate the normal mean and variance using the familiar estimators
- but we can use the samples to estimate any expectation $\mathrm{E}(h(\theta)|y)$

- assume that we have drawn $\theta^1, \ldots, \theta^m$ from the posterior $p(\theta|y)$
- the expectation $\mathrm{E}(h(\theta)|y)$ can be approximated by a a *Monte Carlo integral*

$$\int h(\theta)p(\theta|y)d\theta \approx \frac{1}{m}\sum_{i=1}^{m} h(\theta^i)$$

- if the samples are independent, it is easy to see that
  1. the mean of the approximation is $\mathrm{E}(h(\theta)|y)$
  2. the variance is proportional to $1/m$

- there are several ways of obtaining the samples
- *direct simulation* is possible for many standard distributions
- most mathematical software (R, MATLAB etc..) have functions for drawing samples
- direct simulation is mainly based on uniformly distributed pseudorandom numbers and their transformations

- since direct simulation of a discrete distribution is easy, why not use a piecewise constant approximation to the posterior?
- this means dividing the θ space into disjoint subsets and setting $p(\theta|y)$ to a constant value in each subset
- normalization would be obtained trivially, since the difficult integral would simply be a finite sum
- not generally a good solution: posterior mass can be very concentrated in high-dimensional problems, so this approach is very unreliable
- demo: `sampling.R`

- fortunately, there are ways of obtaining simulated samples from an unnormalized distribution
- the methods discussed below have very appealing properties in theory
- but some important weaknesses limit their use in practice
- some basic methods are introduced, and their properties explored
- notation warning: $p(\theta|y)$ is generally assumed to be *unnormalized* in the rest of this lecture.

Importance Sampling

- importance sampling approximates integrals directly
- assume we want to compute (using an unnormalized posterior)

$$\mathrm{E}(h(\theta)|y) = \int h(\theta)p(\theta|y)d\theta / \int p(\theta|y)d\theta$$

- in importance sampling, a distribution $g(\theta)$ is chosen so that it is easy to draw values from

- insert $1 = g(\theta)/g(\theta)$ to obtain

$$\mathrm{E}(h(\theta)|y) = \frac{\int h(\theta)\frac{p(\theta|y)}{g(\theta)}g(\theta)d\theta}{\int \frac{p(\theta|y)}{g(\theta)}g(\theta)d\theta}$$

- this is equal to

$$\frac{\mathrm{E}_g\left(h(\theta)\frac{p(\theta|y)}{g(\theta)}\right)}{\mathrm{E}_g\left(\frac{p(\theta|y)}{g(\theta)}\right)}$$

where E_g is an expectation over $g(\theta)$

- simulate $g(\theta)$ to obtain samples $\theta^i,\ i = 1, \ldots, m$ and compute the *importance ratios*

$$w^i = \frac{p(\theta^i|y)}{g(\theta^i)},\ i = 1, \ldots, m$$

- both p and g can be unnormalized
- then approximate the expectations to obtain

$$\mathrm{E}(h(\theta)|y) \approx \frac{\sum_i w^i h(\theta^i)}{\sum_i w^i}$$

- properties:
  - works if $g(\theta)$ is somewhat proportional to $h(\theta)p(\theta|y)$: if $g \ll hp$, then very few simulated values are obtained where they are needed
  - both $p(\theta|y)$ and $g(\theta)$ can be unnormalized
  - can use the same simulated set for new $p(\theta|y)$ and/or new $h(\theta)$
  - does *not* give simulated values from the posterior!

Rejection Sampling

- generates independent samples drawn from the posterior
- again choose a distribution $g(\theta)$ which is easy to sample from (can be unnormalized)
- the importance ratio must be bounded:

$$\frac{p(\theta|y)}{g(\theta)} \leq M \text{ for all } \theta$$

- rejection sampling proceeds as follows:
  1. draw a value θ^* from $g(\theta)$, and u from a uniform distribution on $[0,1]$
  2. accept the sample $\theta^i = \theta^*$ if $u \leq p(\theta^*|y)/Mg(\theta^*)$, otherwise go back to step 1
- repeating the two steps, one obtains a set of samples θ^i which are distributed as $p(\theta|y)$
- straightforward computation shows this

$$\begin{aligned}
p(\theta^i \leq x) &= p(\theta^* \leq x | u \leq \frac{p(\theta^*|y)}{Mg(\theta^*)}) \\
&= \frac{p(\theta^* \leq x, u \leq \frac{p(\theta^*|y)}{Mg(\theta^*)})}{p(u \leq \frac{p(\theta^*|y)}{Mg(\theta^*)})} \\
&= \frac{\int_{-\infty}^{x} \int_0^{p(\theta|y)/Mg(\theta)} du\, g(\theta) d\theta}{\int_{-\infty}^{\infty} \int_0^{p(\theta|y)/Mg(\theta)} du\, g(\theta) d\theta} \\
&= \frac{1/M \int_{-\infty}^{x} p(\theta|y) d\theta}{1/M \int_{-\infty}^{\infty} p(\theta|y) d\theta} \\
&= \int_{-\infty}^{x} p(\theta|y) d\theta = p(\theta \leq x|y)
\end{aligned}$$

- as in importance sampling, the distribution $g(\theta)$ has practical implications
- if the proposal distribution $g(\theta)$ is small where the real distribution is large, the constant M is large
- then a very large proportion of simulated points will be rejected (exercise)

Markov Chain Monte Carlo

- direct and rejection sampling define a random process
$$\theta^0, \theta^1, \theta^2, \ldots \tag{1}$$
which is an i.i.d. sequence of $p(\theta|y)$-distributed variables
- often it is not practical to use the above methods to obtain i.i.d. samples from the posterior
- more practical simulation methods are based on Markov Chains, which result in nonindependent samples

- a random process $\theta^0, \theta^1, \ldots$ is a *Markov Process*, if it has the Markov property:
$$p(\theta^{i+1}|\theta^i, \theta^{i-1}, \ldots) = p(\theta^{i+1}|\theta^i), \ \ \forall i \geq 0$$
- this means that the current *state* θ^i completely defines the conditional distribution of future states
- i.e. the process has a very short memory

- to define a Markov Chain, we need a transition distribution
$$T(\theta^{i+1}|\theta^i) = p(\theta^{i+1}|\theta^i)$$
- given an initial distribution $p(\theta^0)$, every θ^i has an *unconditional* distribution $p(\theta^i)$
- the Markov Chain has a *stationary distribution* if $p(\theta^{i+1}) = p(\theta^i)$ for all $i \geq 0$

- Markov Chain Monte Carlo (MCMC) means simulating a Markov Chain with a desired stationary distribution
- in our case, we usually want the stationary distribution to be the posterior
- direct and rejection sampling are trivially MCMC (with no memory)
- in general, MCMC results in a sequence of simulated values *which are not independent*

- some needed properties (see refs on webpage):
  - stationary distribution $p(\theta|y)$: correct distribution achieved by construction, stationarity follows from properties below
  - $\pi-$irreducibility: for any set A with $p(\theta \in A|y) > 0$ and any starting value θ^0, there is an integer $n = n(\theta^0, A)$ so that $\mathbb{P}(\theta^n \in A) > 0$
  - Harris-recurrence: given a $\pi-$irreducible chain, for any set B with $p(\theta \in B|y) > 0$, it must hold that $\mathbb{P}(\theta_i \in B \text{ happens infinitely often}|\theta^0) = 1 \ \ \forall \theta^0$
  - aperiodicity: means that the chain cannot "cycle" through a sequence of disjoint sets

Metropolis-Hastings Algorithm

1. pick an initial value θ^0, set $i = 0$
2. draw θ^* from a *jumping distribution* $J(\theta^*|\theta^i)$
3. compute the *jumping ratio* for θ^*:

$$r = \frac{p(\theta^*|y)J(\theta^i|\theta^*)}{p(\theta^i|y)J(\theta^*|\theta^i)}$$

4. set the next simulated θ^{i+1} as

$$\theta^{i+1} = \begin{cases} \theta^*, & \text{with probability } \min(r, 1) \\ \theta^i, & \text{with probability } 1 - \min(r, 1) \end{cases}$$

5. set $i = i + 1$ and go to step 2

- note the difference to rejection sampling: if θ^* is "rejected", the previous value θ^i is *repeated* instead of discarded
- if the jumping distribution is symmetric ($J(\theta^*|\theta^i) = J(\theta^i|\theta^*)$), then $r = p(\theta^*|y)/p(\theta^i|y)$ and the algorithm is called the Metropolis algorithm
- M-H and its special cases are widely used due to their properties which follow from theory of Markov Chains

- M-H has the posterior as the stationary distribution (demo exercise)
- $\pi-$irreducibility is case-specific (usually a positive jumping distribution guarantees it)
- Harris-recurrence follows from π-irreductibility for M-H
- aperiodicity (case-specific, holds for most jumping distributions)

- some theoretical results for $M - H$:
  - Harris-recurrent, $\pi-$irreducible, aperiodic chain with a stationary distribution converges to the stationary distribution from any starting value
  - above assumptions and $\mathrm{E}(|h|) < \infty$ are enough to show that

$$\lim \frac{1}{n} \sum_{j=1}^{n} h(\theta^j) = \mathrm{E}(h(\theta)) \text{ almost surely}$$

- M-H solves the computational problem of Bayes *in theory*
- in practice, the initial value biases early samples and the jumping distribution affects convergence speed
- long jumps are rejected, and short jumps converge slowly
- demo: `metropolis.R` and `metropolis2.R`

- lots of heuristic simulation methods have been proposed, perhaps because MCMC methods are computationally intensive and somewhat brute-force solutions
- either they are special cases of M-H, or not. In the latter case, it may be difficult to know that they will converge to $\mathrm{E}(h(\theta)|y)$
- lets go through a couple of special cases of M-H

Gibbs Sampler

- if θ is multivariate and full conditional posteriors are known and easy to sample from, then it is possible to use the Gibbs sampler
- one must be able to simulate all full conditional posteriors

$$p(\theta_k|y, \theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_m)$$

- one iteration of the Gibbs sampler can be defined as follows:
  1. the full sample $\theta^i = (\theta_1^i, \ldots, \theta_m^i)'$ has been drawn
  2. the next component θ_1^{i+1} is drawn from $p(\theta_1|y, \theta_2^i, \ldots, \theta_m^i)$
  3. then θ_2^{i+1} is drawn from $p(\theta_2|y, \theta_1^{i+1}, \theta_3^i, \ldots, \theta_m^i)$
  4. obtain all components of θ^{i+1} as above
  5. increase i to $i+1$ and start again

- Gibbs is a special case of M-H with jumps changing one component at a time
- stationary distribution is the posterior since Gibbs is M-H
- the jumping distribution is obtained from the full conditional posteriors
- the acceptance ratio turns out to be one, so all jumps are accepted (see Gelman, 1st edition, p. 328)
- demo: `gibbs.R`

- independence sampler: set $J(\theta^*|\theta^i) = g(\theta^*)$
- jumping distribution is *independent* of the previous state θ^i
- jumping ratio is

$$r = \frac{p(\theta^*|y)g(\theta^i)}{p(\theta^i|y)g(\theta^*)} = \frac{w(\theta^*)}{w(\theta^i)}$$

- w is the *importance ratio* seen in importance sampling
- if $g(\theta) = p(\theta|y)$, then all jumps are accepted and we obtain direct sampling

- Langevin algorithms include a term that directs jumps toward posterior modes
- example:

$$J(\theta^*|\theta^i) \propto \exp(-\frac{1}{2\sigma^2}\|\theta^* - \theta^i - \frac{\sigma^2}{2}(\log p(\theta^i|y))'\|^2)$$

- the jumping distribution is a Normal distribution with mean $\theta^i + \frac{\sigma^2}{2}(\log p)'$
- the derivative of the log-posterior moves the mean towards a (local) posterior mode, thus hopefully achieving faster convergence

- in theory, the fact that MCMC does not result in independent samples is not a problem
- this depends on the convergence of the MC integral despite dependent samples
- but asymptotic results are just that: "asymptotic" does not mean "holds for a very large but finite number of samples"
- in practice, it is important to known how good is the MC approximation computed from the finite set of samples

- dependent samples make it difficult to say anything general about the finite-sample properties of MC approximation
- intuition from iid samples can go wrong: MCMC can spend a lot of time in some small area of the posterior, then jump to another are and spend a lot of time in there
- Monte Carlo integral can have a large error *unless the samples represent the posterior well as a set*
- for example, having samples from one mode of a bimodal distribution does not yield good MC approximations

MCMC Convergence

- here, convergence is nonrigorously defined as "samples begin to represent the posterior well enough"
- in practice, inital value θ^0 biases the early samples
- the chain spends some time simulating samples that do not represent the correct posterior well, as seen in the demo

- what to do about the early samples?
- run the chain until convergence and discard the early samples
- after convergence, keep all samples (or take every k:th sample)
- but knowing when the chain has converged is not generally possible

- in practice, one should be able to
  1. find out the convergence speed, and use it to estimate how many samples need to be discarded
  2. examine the samples as they are simulated, and try to decide whether the chain has converged
- no completely general method has not been found to date

- some heuristics:
  - run parallel chains from different starting points (avoid local modes)
  - compare the variance within chains and between chains (if between chains much larger, then no convergence)
  - simulate starting point(s) from a crude posterior approximation (avoid slow convergence towards posterior mode(s))