MSF100
Statistical Inference Principles - Lecture 12

Rebecka Jörnsten
Mathematical Statistics
University of Gothenburg/Chalmers University of Technology

February 29, 2012

1 Simulation based inference

You already looked at this when you read about bootstrap. Bootstrap was a way of drawing inference based on repeated sampling from an estimate of the underlying true distribution function F. The estimate most commonly used is the empirical cdf, in which case inference is based on multiple data sets generated by resampling data from the original data set, with replacement.

Here we will talk about simulation based inference for Bayesian modeling. Now, frequentist methods may not perform so well if the sample size is limited and the model complex. Moreover, perhaps you have a good idea of likely values for θ and would like to take this prior belief into account? We have:

- prior $\pi(\theta)$
- data distribution $f(y|\theta)$
- posterior $p(\theta|y)$.

The posterior is used for Bayesian inference. It can be written as

$$p(\theta|y) = \frac{\pi(\theta)f(y|\theta)}{\int f(y|\theta)\pi(\theta)d\theta}$$

where the integral in the denominator can be quite difficult to compute analytically. Even numerical integration can be hard in e.g. cases with high-dimensional θ. What can we do?

1.1 Monte Carlo Integration

The Bayesian inference involves intergrals of some kind with respect to the posterior $p(\theta|y)$. For example:

- $\int \theta p(\theta|y)d\theta$ is the posterior mean
- $\int_a^b p(\theta|y)d\theta$ relates to credible regions

In general, we need to compute integrals of the kind

$$\int h(\theta)p(\theta|y)d\theta$$

for functions $h()$. Instead of direct computation, assume we could *sample* from $p(\theta|y)$: constructing a sequence of values $\theta^1, \theta^2, \cdots, \theta^M$ drawn from $p(\theta|y)$. Then, by the LLN we have that

$$\frac{1}{M}\sum_m h(\theta^m) \rightarrow \int h(\theta)p(\theta|y)d\theta$$

Hang on - the point is we don't have $p(\theta|y)$, so how did this help. Well, it turns out that sampling from $p(\theta|y)$ to be able to compute sample means or expectation does not require us to have the whole posterior, only relative values of it (the numerator!). How this works is by clever sampling schemes that generate a Markov Chain of θ-values whose stationary distribution is the posterior (more later). Let's first briefly say something about various sampling approaches:

- Rejection based: Let's say you sample θ from $q(\theta)$. If the resulting θ is unlikely to have come from $p(\theta|y)$ you reject it, otherwise use it sample mean computations.
- Importance based: Sample θ from $q(\theta)$. In the sample computation, reweigh the samples according to $p(\theta)/q(\theta)$. That is, θ that are likely under p contribute more to the sample mean calculation than other samples.
- Marginal (one-dimensional) sampling: Systematically update θ one component at a time by sampling from the marginal conditional posterior $p(\theta_i|\theta^{-1}|y)$, i.e. condition on all other components of θ. This marginal conditional posterior may have a much more simple form than the joint posterior.

All of these schemes generate a Markov chain of updates θ^m *and* one can show that under some restrictions on how you sample (e.g. form of q), this Markov chain converges and its stationary distribution is the posterior!!! This is a very important result. The construction of sampling schemes above are not so difficult, but the end-result something that looked intractable. This Markov Chain Monte Carlo (MCMC) was a huge hot topic in the 90's and lead to a revolution for Bayesian statistics.

1.2 The Gibbs sampler

To simplify the process of sampling from the multi-dimensional posterior, Gibbs sampling iterates over a systematic sampling over one-dimensional posteriors: one component of the parameter vector Θ at at time. That is, the algorithm is

- for $m = m + 1$ until convergence do:
  - for component $i = 1, \cdots, p$ do
  - draw sample from $\theta_i^m \sim P(\theta_i|\theta_j^m, j \neq i, y)$

Of course, you can do this updating in a more random fashion. In addition, you can make the updating parallel if there are disjoint sets of θ given other sets (convenient for e.g. image processing).

Example:

- Change-point example.
- $y_i \sim Poi(\lambda_1)$ for $i = 1, \cdots, M$ and $y_i \sim Poi(\lambda_2)$ for $i = M + 1, \cdots, n$.
- We want to estimate both the location of, M, of the change of behaviour of the data, and the local behaviour statistics λ_1 and λ_2.
- Bayesian method: assume a prior for each Poisson intensity parameter λ_j as $Gamma(\alpha_j, \beta_j)$.
- Assume a uniform distribution for M on the $\{1, n\}$.
- The data likelihood is
$$l(y) = \prod_{i=1}^{M} \frac{e^{-\lambda_1}\lambda_1^{y_i}}{y_i!} \prod_{i=M+1}^{n} \frac{e^{-\lambda_2}\lambda_2^{y_i}}{y_i!}$$
- Add the prior and construct the joint distibution
$$joint \prod_{i=1}^{M} \frac{e^{-\lambda_1}\lambda_1^{y_i}}{y_i!} \prod_{i=M+1}^{n} \frac{e^{-\lambda_2}\lambda_2^{y_i}}{y_i!} \frac{1}{\Gamma(\alpha_1)} \lambda_1^{\alpha_1 - 1} \beta_1^{\alpha_1} e^{-\beta_1 \lambda_1} \frac{1}{\Gamma(\alpha_2)} \lambda_2^{\alpha_2 - 1} \beta_2^{\alpha_2} e^{-\beta_2 \lambda_2}$$
The posterior is proportional to the joint. Collect the terms of interest to find that:
$$posterior \propto \lambda_1^{\alpha_1 - 1 + \sum_{i=1}^{M} y_i} e^{-(\beta_1 + M)\lambda_1} \lambda_2^{\alpha_2 - 1 + \sum_{i=M+1}^{n} y_i} e^{-(\beta_2 + n - M)\lambda_2} * c$$
From this, we construct the Gibbs sampler. The marginal conditional for $\lambda_j, j = 1, 2$ is seen to be *Gamma*-distributed by matching elements of the distribution depending on λ_j to a *Gamma*. That is
$$\lambda_1|\lambda_2, y, M \sim Gamma(\alpha_1 + \sum_{i=1}^{M} y_i, \beta_1 + M)$$

$$\lambda_2|\lambda_1, y, M \sim Gamma(\alpha_2 + \sum_{i=M+1}^{n} y_i, \beta_2 + n - M).$$

To update M we need to search over $M = \{1, \cdots, n\}$ evaluating over the marginal posterior

$$M|\lambda_1, \lambda_2, y \propto \lambda_1^{\sum_{i=1}^{M} y_i} \lambda_2^{\sum_{i=M+1}^{n} y_i} e^{(\lambda_2 - \lambda_1)M}$$

1.3 Metropolist-Hastings

MH sampling is the first kind of MC construction for posterior sampling (50's - 70's). Gibbs sampling came along in the 80's (general construction). I strongly encourage you to read Geman and Geman '84 which is a beautiful paper explaining Gibbs + MH sampling (simulated annealing) in the context of image processing.

MH sampling is related to rejection sampling. Gibbs sampling doesn't always work out, meaning the one-dimensional posteriors don't have a much more convenient format over the joint. MH sampling works like this:

1. Generate a proposal θ^* from the distribution $q(\theta^m, \theta^*) = q(\theta^*|\theta^m)$. θ^m is a current estimate of the parameter vector, so think of q as a map $q(x, y)$ from the current state of θ to a new state.
2. Compute
$$\alpha(\theta^m, \theta^*) = \min\{1, \frac{p(\theta^*|y)}{p(\theta^m|y)} \frac{q(\theta^*, \theta^m)}{q(\theta^m, \theta^*)}\}$$
3. With probability α, accept the proposal θ^* in place of θ^m.

If q is symmetric, meaning it's equally likely to move from θ^m to θ^* as from θ^* to θ^m, the α is 1 for any θ that has higher posterior probability than θ^m. Otherwise you may still accept this proposal with a probability α. If q is not symmetric, we also adjust this ratio according to the q sampling direction.

The proposal distribution q can be quite arbitrary. However, some technical conditions are needed for MH to converge to the posterior. Note, we can use a very simplistic scheme where $q(\theta^m, \theta^*) = q(\theta^*)$, i.e. the sampling does not depend on the current state of θ. This is called independence chain sampling. Note, we want q to be quite rich or "wild". If q is restricted we get stuck near our starting value. On the other hand if q is too wild (has very high variance), the samples generated are likely to be rejected and we also get stuck. The variance of q is a tuning parameter of the MC scheme.

Why does MH work? Here's a short review of Markov Chains. We assume that the states of the chain x_t have the following property:

$$P(X_{t+1} = x_{t+1}|X^t = x^t) = P(X_{t+1} = x_{t+1}|X_t = x_t)$$

Note, X^t refers to the history of X_{t+1}, i.e. all previous values of the X-process. The Markov chain moves from state to state according to the transition kernel (or matrix when the states are a discrete set) as

$$P(i, j) = P(X_{t+1} = j|X_t = i)$$

The marginal state distribution at time t in the chain is denoted as

$$\pi_i(t + 1) = P(X_{t+1} = i) = \sum_k P(X_{t+1} = i|X_t = k)\pi_k(t)$$

Using the above across all state, we see that we can write the marginal distribution as

$$\pi(t + 1) = P\pi(t)$$

and recursively applying this results we find that

$$\pi(t) = P^t\pi(0)$$

.

A stationary distribution for P is one for which

$$\pi^* = P\pi^*$$

Now, what do we need for our sampling scheme to work? Well, we already have a sampling scheme that is Markov by construction. We need to verify that it is sensible in terms of the goal we have: sampling from the posterior. There are some constraints on the transition matrix P for a MC to converge to our desired goal. First of all, we need the chain to be irreducible: meaning we have to be able to reach a target state in a finite number of steps. We also need it to be aperiodic. Additional constraints like invariance (meaning of you start from the posterior distribution, subsequent marginal distributions can't move away from the distribution).

The detailed balance criterion is a constraint on the transition along the chain that guarantees a unique stationary distribution. The condition states that

$$P(j,k)\pi_j^* = P(k,j)\pi_k^*$$

meaning the chain is reversible. In our application, π^* (the stationary distribution), is the posterior $p(\theta|y)$.

Now, for MH sampling our transition kernel is

$$P(\theta,\theta') = q(\theta,\theta')\alpha(\theta,\theta') = q(\theta,\theta')\min[\frac{p(\theta')q(\theta',\theta)}{p(\theta)q(\theta,\theta')},1]$$

where for $q(,)$ and $\alpha(,)$ the first index is the origin and the second the destination. That is, to move from θ to θ' you have to first generate the sample θ' given the current state θ ($q(\theta,\theta')$) and then accept this sample move ($\alpha(\theta,\theta')$). Now, we need to verify that the MH sampling scheme satisfies the detailed balance:

$$P(\theta,\theta')p(\theta) = P(\theta',\theta)p(\theta')$$

where $p()$ is the posterior. That is, we need

$$q(\theta,\theta')\alpha(\theta,\theta')p(\theta) = q(\theta',\theta)\alpha(\theta',\theta)p(\theta')$$

for all θ and θ' states.

We now need to verify that the MH transitions satisfy the detailed balance equation. Consider the case when $\alpha = 1$. Then the result follows trivially. What if $q(\theta,\theta')p(\theta) \geq q(\theta',\theta)p(\theta')$? Then,

$$\alpha(\theta,\theta') = \frac{p(\theta')q(\theta',\theta)}{p(\theta)q(\theta,\theta')}$$

and

$$\alpha(\theta',\theta) = 1$$

since while we move toward a θ closer to the MaP the posterior increases. That means

$$P(\theta,\theta')p(\theta) = q(\theta,\theta')\alpha(\theta,\theta')p(\theta) = q(\theta,\theta')\frac{p(\theta')q(\theta',\theta)}{p(\theta)q(\theta,\theta')}p(\theta) =$$

$$= q(\theta',\theta)p(\theta') = q(\theta',\theta)\alpha(\theta',\theta)p(\theta') = P(\theta',\theta)p(\theta')$$

and we have proved that the detailed balance holds. Similar steps prove for the case when $q(\theta,\theta')p(\theta) \leq q(\theta',\theta)p(\theta')$.

OK - so the MH sampling results in a Markov chain whose limit stationary distribution is the posterior. However, we still need to worry about the rate of convergence. For q that is restrictive or too wild, convergence can be slow since few proposed samples are accepted. This means we need a "burn-in" period before we start recording samples in our sample mean calculation (see beginning of lecture). When to start collecting samples for computation is decided based on convergence diagnostics, usually obtained by plotting the traces of the θ_i^m for iterations m of the MC scheme and components i. You look for when this trace is devoid of a trend. What should the trace look like? If it is too smooth it means

that the MC carries too much memory, each step is too close to the previous and we have a difficult time breaking away from starting values (called poor mixing). This may result from q being to restrictive or too wild, meaning we explore little or Θ space or generate few samples that are accepted.

It is common to plot the θ traces to diagnose convergence. We can also consider comparing different traces starting from different initial values. The difference between the traces (in terms of local means over consecutive iterations) constitute a way to determine convergence.

We can also aid in the mixing by throwing in an additional stochastic element: accepting a "worse" sample θ' with a probability that is dependent on the iteration number. This is called "simulated annealing". In the case of symmetric $q()$ we can write the acceptance probability as

$$\alpha = \min[1, (\frac{p(\theta')}{p(\theta)})^{1/T^m}]$$

where T^m is called the *temperature*. For $T^m = T = 1$, this is equivalent to MH. For T^m decreasing from $T^m > 1$ to 0, this constitutes a scheme where, initially, we are randomly exploring the θ-space and then slowly move to a more restrictive sampling geared and finding the MaP. Why is thus useful? It allows us to be less dependent on the starting values since we can "escape" local optima through accepting a locally "bad" move. Simulated annealing is a very useful algorithm in complex searching, e.g. related to model selection.

Finally, some notes on how to choose $q()$. In general, most $q()$ are some kind of random walk or independence sampling scheme, often involving a tuning parameter. The tuning parameter controls how well the MC scheme that results is mixing - exploring θ space. You need to diagnose the convergence of the MC scheme by plotting the θ traces and either graphically, or more formally using time series analysis, decide if you have reached a point of convergence. Don't save all θ values - you only need to save enough to decide that you have converged and then collect enough thereafter to compute a reasonable integral estimate.

1.4 Homework

Consider a bivariate normal distribution.
(a) Assume that out of n samples, the first k are missing values for the second coordinate. Derive the EM steps for estimating the bivariate normal parameters.
(b) Assume n samples (no missing values). If you assume a normal, independent prior for the mean parameters (you can also use a prior for the covariance if you want), what is the posterior distribution?
(c) Implement a MC scheme to generate samples from the posterior in (b). How well does it agree with the analytical solution?