STA 414/2104:
Statistical Methods of Machine Learning II

Week 5 - 2/2: Sampling II

Michal Malyska

University of Toronto

Overview

- Gibbs sampling
- Hamiltonian Monte Carlo
- MCMC diagnostics

Gibbs Sampling

- Suppose the parameter vector θ has been divided into d components
$$\theta = (\theta_1, ..., \theta_d)$$
- At each iteration, the **Gibbs Sampler**, cycles through the components of θ, drawing each subset conditional on the value of all others.
- This means we perform d steps at each sampling iteration t to obtain $\theta^{(t+1)}$

Gibbs Sampling Procedure

At iteration t:

- chose an ordering j of d sub-vectors of θ
- For $j = 1$ to $j = d$:
  - Sample θ_j^t from the conditional distribution given all the other components:

$$p(\theta_j | \theta_{-j}^{t-1}, y)$$

Where θ_{-j}^{t-1} represents all the components of θ except for θ_j at their current values:

$$\theta_{-j}^{t-1} = (\theta_1^t, \theta_2^t, ..., \theta_{j-1}^t, \theta_{j+1}^{t-1}, ..., \theta_d^{t-1})$$

Gibbs Sampling Example

Consider a single observation (y_1, y_2) from a bivariate normal, with unknown mean $\mu = (\mu_1, \mu_2)$ and known covariance matrix: $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ with a uniform prior distibution on μ
The posterior takes the form:

$$\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} | y \sim N\left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \Sigma \right)$$

Although it is simple to draw from this posterior we can alternatively use the Gibbs sampler. To do that we must first deterimine the conditional posterior distributions for μ_1 and μ_2

Gibbs Sampling Example

Using the properties of the multivariate normal distribution we have:

$$\mu_1|\mu_2, y \sim N(y_1 + \rho(\mu_2 - y_2), 1 - \rho^2)$$
$$\mu_2|\mu_1, y \sim N(y_2 + \rho(\mu_1 - y_1), 1 - \rho^2)$$

Then given some previous (possibly initial) value of μ, the sampling would be:

- $\mu_1^{(t)} \sim N(y_1 + \rho(\mu_2^{(t-1)} - y_2), 1 - \rho^2)$
- $\mu_2^{(t)} \sim N(y_2 + \rho(\mu_1^{(t)} - y_1), 1 - \rho^2)$

Gibbs Sampling Example

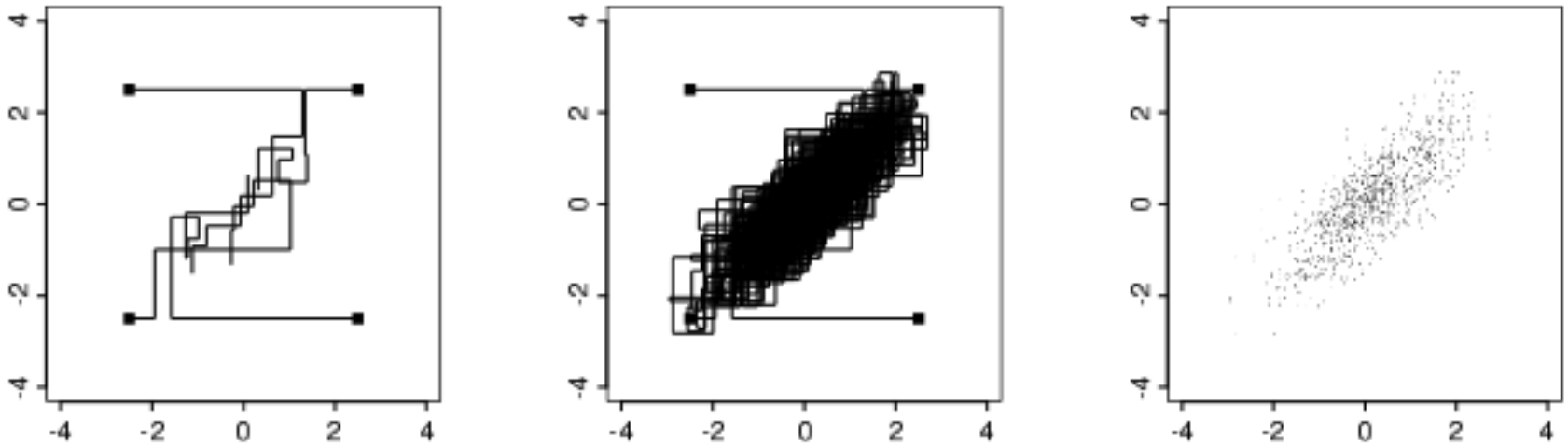

Figure 11.2 *Four independent sequences of the Gibbs sampler for a bivariate normal distribution with correlation $\rho = 0.8$, with overdispersed starting points indicated by solid squares. (a) First 10 iterations, showing the componentwise updating of the Gibbs iterations. (b) After 500 iterations, the sequences have reached approximate convergence. Figure (c) shows the points from the second halves of the sequences, representing a set of correlated draws from the target distribution.*

1

[1]From "Bayesian Data Analysis Third edition" by Gelman, Carlin, Stern, Dunson, Vehtari, Rubin

Hamiltonian Monte Carlo

- Given the position x, the potential energy is $E(x)$
- Construct a distribution

$$p(x) \propto e^{-E(x)}, \quad \text{with} \quad E(x) = -\log(\tilde{p}(x))$$

  where $\tilde{p}(x)$ is the unnormalized density we can evaluate.
- Introduce velocity v carrying the kinetic energy $K(v) = \|v\|^2/2$
- Total energy or Hamiltonian: $H = E(x) + K(v)$.
- Energy is preserved:
  - Frictionless ball rolling $(x, v) \to (x', v')$
  - $H(x, v) = H(x', v')$.
- Ideal Hamiltonian dynamics are reversible: reverse v and the ball will return to its start point!

Hamiltonian Monte Carlo

- The joint distribution:
  - $p(x, v) \propto e^{-E(x)} e^{-K(v)} = e^{-E(x)-K(v)} = e^{-H(x,v)}$
  - Velocity is independent of position and Gaussian.
- MCMC procedure
  - Use Gibbs sampling for the velocity
  - Simulate Hamiltonian dynamics then flip sign of velocity:
    - Hamiltonian 'proposal' is deterministic and reversible: $q(x', v'|x, v) = q(x, v|x', v')$
    - Energy is constant $p(x, v) = p(x', v')$.
    - Metropolis acceptance probability is 1.
- How to simulate Hamiltonian dynamics?

Leap-frog integrator

- A numerical approximation:

$$\begin{aligned} v_i(t+\tfrac{\epsilon}{2}) &= v_i(t) - \frac{\epsilon}{2}\frac{\partial E(x(t))}{\partial x_i} \\ x_i(t+\epsilon) &= x_i(t) + \epsilon\, v_i(t+\tfrac{\epsilon}{2}) \\ v_i(t+\epsilon) &= v_i(t+\tfrac{\epsilon}{2}) - \frac{\epsilon}{2}\frac{\partial E(x(t+\epsilon))}{\partial x_i} \end{aligned}$$

- H is not conserved.
- Dynamics are still deterministic (and reversible)
- Acceptance probability :

$$\min\{1, \exp(H(x,v) - H(x',v'))\}$$

HMC algorithm

The HMC algorithm (run until it mixes):

- Gibbs sample velocity: $v \sim \mathcal{N}(0, I)$.
- Run Leapfrog integrator for L steps
- Accept new position x' with probability:

$$\min\{1, \exp(H(x, v) - H(x', v'))\}$$

MCMC Inference

- Compute the unnormalized posterior
- Simulate from it
- Draw "normal" inference from simulated values of θ
  - mean
  - median
  - quantiles
- **Posterior predictive** simulations of unobserved outcomes $\tilde{y}$ can be obtained by further simulation conditional on drawn values of θ
- All of this however requires some care, as MCMC is not without problems

MCMC demo

MCMC diagnostics

- How do we know we have ran the algorithm long enough?
- What if we started very far from where our distribution is?
- Since there is autocorrelation, what is the "effective" number of samples we have?

Good Ideas for MCMC

- Parallel computation is cheap - we can run multiple chains in parallel starting at different points
- We should discard some initial number of samples - **warm-up** or **burn-in**
- We (maybe) should only keep every k-th observation from each chain
- We should examine how well the chains have "mixed" together - i.e. how much overlap is there in the parameter space each of them explored

R hat

Start with $m/2$ chains of $2n$ samples each, with a warm-up period of n. Split them in half so that we have m chains total (half of which are burn-in) of length n each. Label each scalar estimand with $\psi_{i,j}$ with $(i = 1, ..., n; j = 1, ..., m)$ The **between sequence variance** B is:

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\bar{\psi}_{.j} - \bar{\psi}_{..})$$

where:

$$\bar{\psi}_{.j} = \frac{1}{n} \sum_{i=1}^{n} \psi_{ij}$$

and:

$$\bar{\psi}_{..} = \frac{1}{m} \sum_{j=1}^{m} \bar{\psi}_{.j}$$

R hat

- The **within sequence variance** W is:

$$W = \frac{1}{m} \sum_{j=1} s_j^2$$

where:

$$s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\psi_{ij} - \bar{\psi}_{.j})^2$$

- For any finite n, W will **underestimate** the true variance, since the chains have not had time to explore the entire possible range of values
- In the limit the expectation of W approaches $var(\psi|y)$

R hat

- We can estimate the marginal posterior variance of ψ by a weighted average of W and B:

$$\widehat{var}^{+}(\psi|y) = \frac{n-1}{n}W + \frac{1}{n}B$$

- This quantity **overestimates** the marginal posterior variance assuming the starting distribution is overdispersed, but is **unbiased** under stationarity or in the limit $n \to \infty$
- We estimate the factor by which the scale of the current distribution for ψ might be reduced if we were continue to infinity by:

$$\hat{R} = \sqrt{\frac{\widehat{var}^{+}(\psi|y)}{W}}$$

Effective Sample Size

- Since our observations are not independent of each other, we de facto gain less information
- One way to quantify the **effective sample size** is to consider statistical efficiency of $\bar{\psi}_{..}$ as an estimate of $\mathbb{E}(\psi|y)$

$$\lim_{n\to\infty} mnvar(\bar{\psi}_{..}) = \left(1 + 2\sum_{t=1}^{\infty}\rho_t\right) var(\psi|y)$$

  Where ρ_t is the autocorrelation of the sequence ψ at lag t
- If the draws were completely independent we would have $var(\bar{\psi}_{..}) = \frac{1}{mn} var(\psi|y)$ and the effective sample size would be mn
- in the presence of correlation we define the **effective sample size** to be:

$$n_{\text{eff}} = \frac{mn}{1 + 2\sum_{t=1}^{\infty}\rho_t}$$

Autocorrelations

- How do we obtain $\sum_{t=1}^{\infty} \rho_t$?
- We start by computing $\widehat{var}^+$ from before
- We then estimate the correlations by first computing the **variogram** V_t at each lag t

$$V_t = \frac{1}{m(n-t)} \sum_{j=1}^{m} \sum_{i=t+1}^{n} (\psi_{i,j} - \psi_{i-t,j})^2$$

- The estimate then becomes:

$$\hat{\rho}_t = 1 - \frac{V_t}{2\widehat{var}^+}$$

- For large values of t this becomes very noisy so we usually cut off the sum over $\hat{\rho}_t$ when two consecutive summands were negative

Diagnostics Summary

- Once $\hat{R}$ is near 1, and $\hat{n}_{\text{eff}}$ is more than 10 per chain **for all scalar estimands** we just collect the mn simulations, (excluding the burn-in)
- We can then draw inference based on our samples. However:
- Even if the iterative simulations appear to have converged, passed all tests etc. It may still be far from convergence!
- Important areas of the target distribution could have been missed by all the chains, (e.g. were difficult to reach from all starting points)
- When we declare "convergence" - we mean that all chains appear stationary and well mixed.
- All of the checks we learned today are not hypothesis test. There are no p-values, and no statistical significance.