Bayesian Theory and Computation

Lecture 5: Markov Chain Monte Carlo II

Cheng Zhang

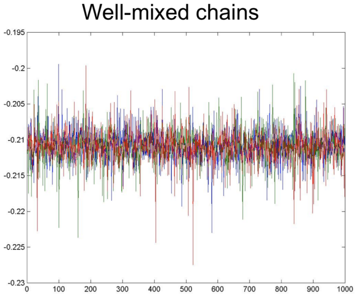
School of Mathematical Sciences, Peking University

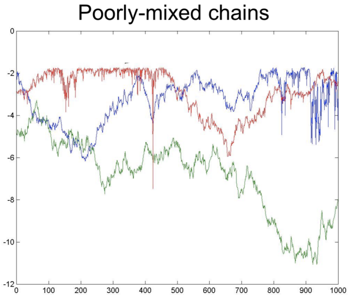
March 26, 2021

MCMC Diagnostics

- MCMC would converge to the target distribution if run sufficiently long
- However, it is often non-trivial to determine whether the chain has converged or not in practice
- Also, how do we measure the efficiency of MCMC chains?
- In what follows, we will discuss some practical advice for coding MCMC algorithms

Graphical Diagnostics: Mixing Rate

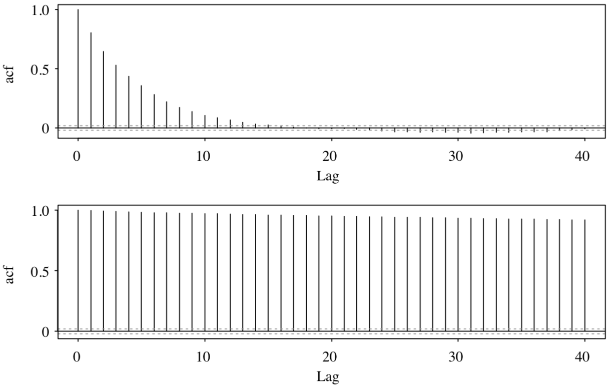




Monitor convergence by plotting samples from multiple MH runs (chains)

- If the chains are well-mixed (left), they are probably converged
- If the chains are poorly-mixed (right), we may need to continue burn-in

Graphical Diagnostics: Autocorrelation



- An autocorrelation plot summarizes the correlation in the sequence of a Markov chain at different iteration lags
- A chain that has poor mixing will exhibit slow decay of the autocorrelation as the lag increases

Effective Sample Size

- Since MCMC samples are correlated, *effective sample size* are often used to measure the efficiency when MCMC samples are used for estimation instead of independent samples
- The effective sample size (ESS) is defined as

  $$\mathrm{ESS} = \frac{n}{1 + 2\sum_{k=1}^{\infty} \rho(k)}$$

  where $\rho(k)$ is the autocorrelation at lag k
- ESS are commonly used to compare the efficiency of competing MCMC samplers for a given problem. Larger ESS usually means faster convergence

Multiple Chains

- One of the hardest problem to diagnose is whether or not the chain has become stuck in one or more modes of the target distribution
- In this case, all convergence diagnostics may indicate that the chain has converged, though it does not
- A partial solution: run multiple chains and compare the within- and between-chain behavior

Auxiliary Variable Methods

- Auxiliary variable strategies can be used to improving mixing of Markov chains
- When standard MCMC methods mix poorly, one potential remedy is to augment the state space of the variable of interest
- This approach can lead to chains that mix faster and require less tuning than the standard MCMC methods
- Main idea: construct a Markov chain over (X, U) (U is the auxiliary variable) with stationary distribution marginalizes to the target distribution of X
- As we will see later, this includes a large family of modern MCMC methods

Parallel Tempering

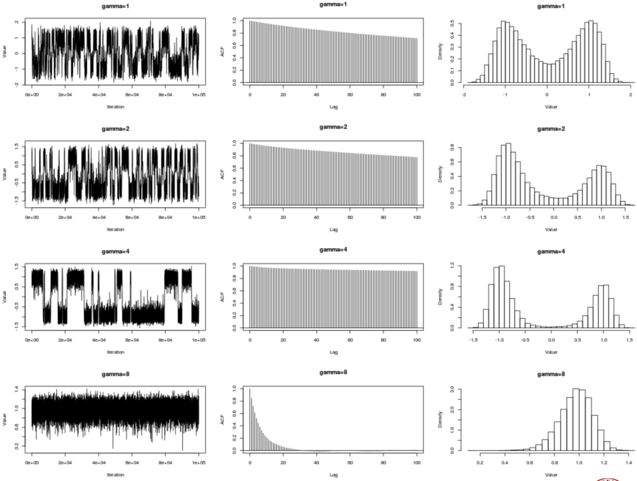
- Suppose that we have a challenging target distribution $f(x) \propto \exp(-U(x))$
- We can introduce temperatures to construct a sequence of distributions that are easier to sample from

$$f_k(x) \propto \exp\left(-U(x)/T_k\right), \quad k = 0, \ldots, K$$

  where $1 = T_0 < T_1 < \ldots < T_K$
- When simulating Markov chains with different temperature T, the chain with high temperature (hot chain) is likely to mix better than the chain with cold temperature (cold chain)
- Therefore, we can run parallel chains and swap states between the chains to improve mixing

Double-well Potential Distribution

$$f_T(x) \propto \exp(-(x^2-1)^2/T), \quad T = 1/\gamma$$



Coupling Parallel Chains

We run parallel Markov chains for distributions with different temperatures. In each iteration

- Follow regular Metropolis steps in each chain to get new states $x_0^{(t)}, \dots, x_K^{(t)}$
- Select two temperatures, say $(i, j), i < j$, and swap the states

$$x_0^{(t)}, \dots, x_i^{(t)}, \dots, x_j^{(t)}, \dots x_K^{(t)} \to x_0^{(t)}, \dots, x_j^{(t)}, \dots, x_i^{(t)}, \dots x_K^{(t)}$$

- Accept the swapped new states with the following probability

$$\min\left(1, f_i(x_j^{(t)}) f_j(x_i^{(t)}) / f_i(x_i^{(t)}) f_j(x_j^{(t)})\right)$$

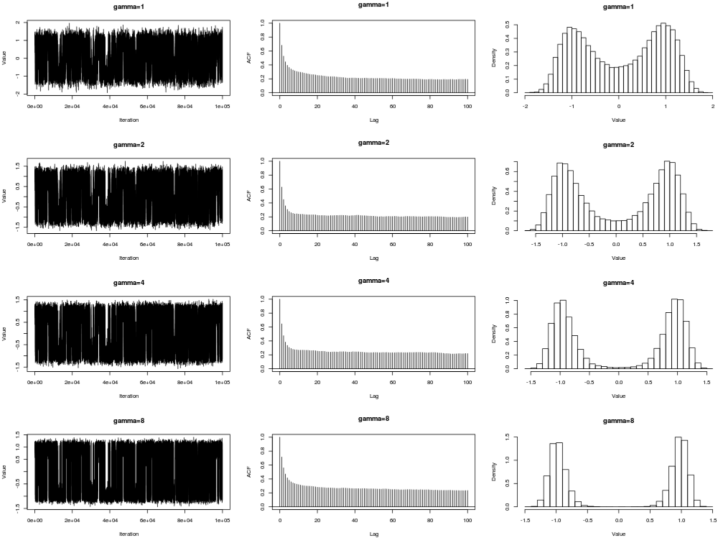
Parallel Tempering

- Both the within-chain Metropolis updates and the between-chain swap preserves

$$p(x_0, \dots, x_K) \propto f_0(x_0) f_1(x_1) \dots f_K(x_K)$$

- Therefore, the joint distribution of $(x_0^{(t)}, \dots, x_K^{(t)})$ will converge to $p(x)$, and the marginal distribution of x_0 (cold chain) is the target distribution
- There are many ways to swap chains. For example, we can pick a pair of temperatures uniformly at random or only swap chains with successive temperatures
- The design of temperature levels could be crucial for the performance

Example: Double-well Potential Distribution



Slice Sampling

- Slice sampling was introduced by Neal (2003) to accelerate mixing of Metropolis (or MH)
- It is essentially a Gibbs sampler in the augmented space (X, U) with density

$$f(x, u) = f(x)f(u|x)$$

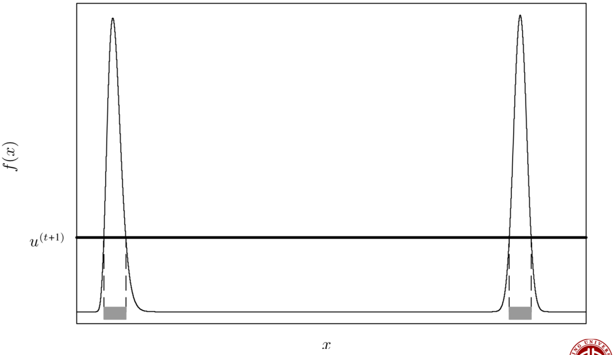
where U is the auxiliary variable and $f(u|x)$ is designed to be a uniform distribution $\mathcal{U}(0, f(x))$

Slice Sampling

- For this purpose, slice sampling alternates between two steps:
  - Given the current state of the Markov chain, x, we uniformly sample a new point u from the interval $(0, f(x))$
    $$U|x \sim \mathcal{U}(0, f(x))$$
  - Given the current value of u, we uniformly sample from the region $S = \{x : f(x) > u\}$, which is referred to as the *slice* defined by u
    $$X|u \sim \mathcal{U}(S)$$
- As mentioned by Neal (2003), in practice it is safer to compute $g(x) = \log(f(x))$, and use the auxiliary variable $z = \log(u) = g(x) - e$, where e has exponential distribution with mean one, and define the slice as $S = \{x : z < g(x)\}$

Move Between Modes with Slice Sampler

- One advantage of slice sampling is for sampling from multimodal distributions
- Unlike standard Metropolis (or MH) that struggles between distant modes, sampling from the slice allows us to easily jump between different modes

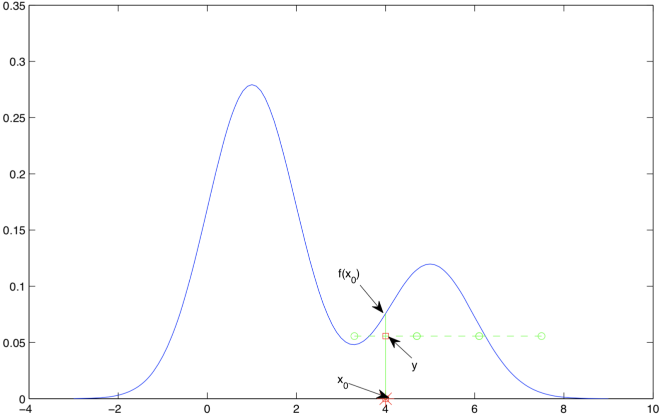


Slice Sampling in Practice

- Sampling an independent point uniformly from S might be difficult. In practice, we can substitute this step by any update that leaves the uniform distribution over S invariant
- There are several methods to perform this task
- Here, we introduce a simple but effective procedure that consists of two phases:
    - *Stepping-out*. A procedure for finding an interval around the current point
    - *Shrinkage*. A procedure for sampling from the interval obtained
- For a detail description of these methods, see Neal (2003)

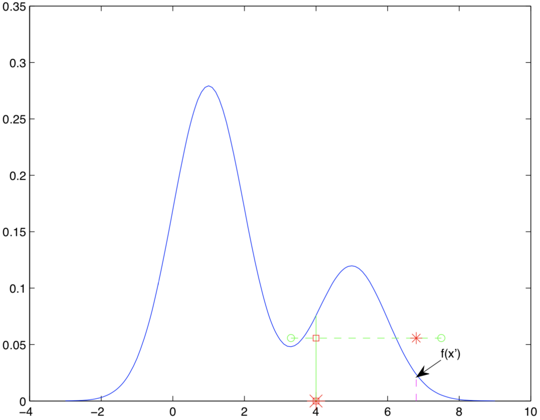
Slice Sampling- Illustration

- Sampling $u \sim \mathcal{U}(0, f(x_0))$ and stepping out (of size w) until we reach points outside the slice



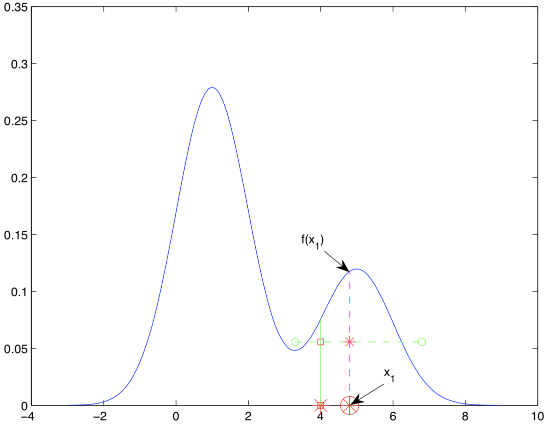
Slice Sampling- Illustration

- Shrinkage of interval to a point, x', which is sampled (uniformly) from the interval but it has $f(x') < y$



Slice Sampling- Illustration

- Continue shrinkage until we reach a point x_1 such that $y < f(x_1)$. We accept x_1 as our new sample



Within-group Exchangeability

- Consider the housing price, y_i, for a sample of 4 bedroom houses in the US. We might regard this sample as exchangeable if all we know is the price of each house.
- However, if we also know in which state the house is located, it might be more appropriate to assume exchangeability only within each group since the price distribution would probably be different from one state to another
- In this case, the price is represented by y_{ij}, where j is an index for the states. The index is not completely uninformative now, since we expect different distributions for different j.
- We can still use deFinetti's theorem and consider each sub-sample, (i.e., for a fixed j) as iid given their own specific parameteric model with parameter θ.

Within-group Exchangeability

- Then, for each state j we have

$$p(y_j|\theta_j) = p(y_{1j}, \ldots, y_{n_j j}|\theta_j) = \prod_{i=1}^{n_j} p(y_{ij}|\theta_j)$$

- Therefore, the joint distribution of all samples is

$$p(y|\theta) = \prod_{j=1}^{J} \prod_{i=1}^{n_j} p(y_{ij}|\theta_j)$$

- Assuming a normal $\mathcal{N}(\mu_j, \sigma_j^2)$ for each state, we have

$$p(y|\mu, \sigma^2) = \prod_{j=1}^{J} \prod_{i=1}^{n_j} \mathcal{N}(y_{ij}|\mu_j, \sigma_j^2)$$

- We can further assume all states have the same variance

$$p(y|\mu, \sigma^2) = \prod_{j=1}^{J} \prod_{i=1}^{n_j} \mathcal{N}(y_{ij}|\mu_j, \sigma^2)$$

Hyperprior

- Now, as we mentioned before, there exists a prior distribution over parameters, $\theta_1, \theta_2, \ldots, \theta_J$.
- Similar to y, if we could imagine the infinite sequence of such θ's being exchangeable, we can regard them as being iid samples given the prior distribution $p(\theta|\phi)$ with the parameter ϕ

$$p(\theta|\phi) = p(\theta_1, \ldots, \theta_J|\phi) = \prod_{j=1}^{J} p(\theta_j|\phi)$$

- ϕ is referred to as *hyperparameter*, for which we need to assume a *hyperprior* $p(\phi)$.

Hyperprior

- The joint prior distribution of all parameters is now

$$p(\phi, \theta) = p(\theta|\phi)p(\phi)$$

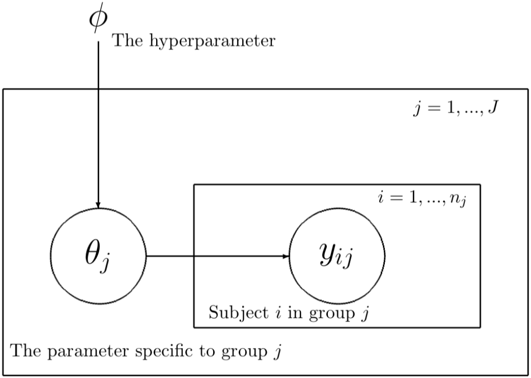
- The posterior distribution of parameters is

$$p(\phi, \theta|y) \propto p(\phi, \theta)p(y|\phi, \theta) = p(\phi)p(\theta|\phi)p(y|\theta)$$

- Note that give θ, y becomes independent of ϕ.

Hierarchical Bayesian Model

The following figure shows a schematic representation of hierarchical models in general



Example: US House Prices

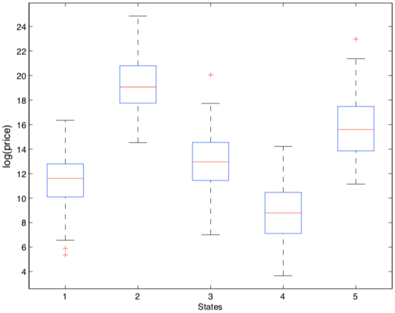
- For the house prices example, we can assume the following priors

$$\begin{aligned} \mu_0 &\sim \mathcal{N}(M, V^2) \\ \mu_j &\sim \mathcal{N}(\mu_0, \tau_0^2) \\ \sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2) \end{aligned}$$

- Here, we are assuming that τ_0^2 is fixed and only μ_0 is the hyperparameter.
- Moreover, we are assuming that the variance σ^2 is the same for all states for simplicity.

Example: US House Prices

- For this problem, we can use MCMC to obtain samples from the posterior distribution of σ, μ_j, and μ_0.
- For simplicity, we consider only 5 states. We have sampled 100 houses in each states.
- This graph shows the box plot of the observed values.



Example: US House Price

- We use the log transformation of prices and assume the following broad priors for model parameters

$$\begin{aligned} \sigma^2 &\sim \text{Inv-}\chi^2(1, 0.5^2) \\ \mu_j &\sim \mathcal{N}(\mu_0, 25^2) \\ \mu_0 &\sim \mathcal{N}(0, 50^2) \end{aligned}$$

- Note that these priors are conditionally conjugate so we can use the Gibbs sampler.

Example: US House Prices

- Given μ_0 and σ^2 the problem reduces to 5 independent normal models with known variance. Given μ_0 and σ^2 at each iteration, we can sample from the posterior distribution of μ_j.
- Given μ_j's, we also have a conditional conjugate situation for σ^2 with Inv-χ^2 posterior distribution. So we can sample a new σ^2.
- Note that since σ^2 is common between all states, we use all the y's from the 5 states to update σ^2.
- Similarly, given the current samples of μ_j, we again have a normal model with conditional conjugacy for μ_0 (taking μ_j's as observations) so we can sample a new μ_0.
- We repeat the above steps to obtain MCMC samples.

Example: US House Prices

- At each iteration, given the value of μ_0 and σ^2, we sample μ_j from the following normal distribution

$$\mu_j|y, \mu_0, \sigma^2 \sim \mathcal{N}\left(\frac{\frac{\mu_0}{\tau_0^2} + \frac{\sum_i y_{ij}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n_j}{\tau^2}}, \frac{1}{\frac{1}{\tau_0^2} + \frac{n_j}{\sigma^2}}\right)$$

- Given $\mu = (\mu_1, \ldots, \mu_J)$, we sample a new σ^2 from

$$\sigma^2|y, \mu \sim \text{Inv-}\chi^2\left(\nu_0 + n, \frac{\nu_0\sigma_0^2 + \nu n}{\nu_0 + n}\right)$$

where

$$\nu = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \mu_j)^2$$

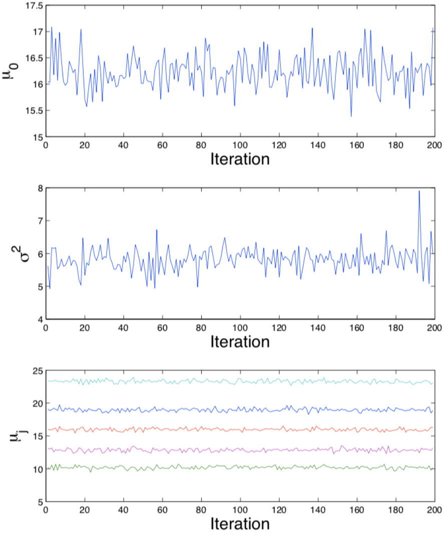
Example: US House Prices

- Given $\mu = (\mu_1, \ldots, \mu_J)$, we sample μ_0 from

$$\mu_0 | \mu \sim \mathcal{N}\left(\frac{\frac{M}{V^2} + \frac{\sum_j \mu_j}{\tau_0^2}}{\frac{1}{V^2} + \frac{J}{\tau_0^2}}, \frac{1}{\frac{1}{V^2} + \frac{J}{\tau_0^2}} \right)$$

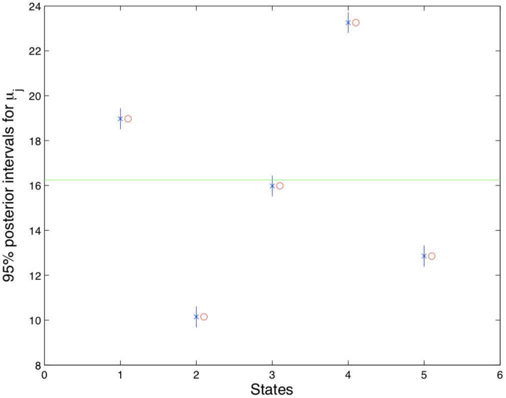
- Notice how using the conditional independence reduces the complexity of the model.
- For this reason, hierarchical Bayesian models are quite powerful.

Example: US House Prices



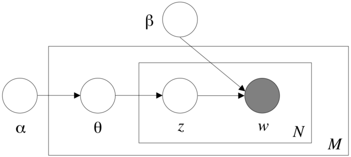
Example: US House Prices

The following graph show the 95% posterior intervals, the posterior expectations (×), the maximum likelihood estimations (circles), and the posterior expectation of the overall mean μ_0 (the green horizontal line).



Example: Latent Dirichlet Allocation

- Generative model of documents (Blei, Jordan and Ng, 2003). Also broadly applicable to collaborative filtering, image retrieval, bioinformatics, etc.

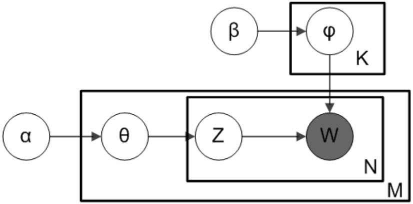


- choose a mixture of topics the document: $\theta \sim \text{Dir}(\alpha)$
- choose a topic for each of the document:

$$z_n \sim \text{Multinomial}(\theta)$$

- choose word given the topic: $w_n | z_n, \beta \sim p(w_n | z_n, \beta)$

Bayesian Inference in LDA



- Use the probability model for LDA, with an additional Dirichlet prior on ϕ.
- The complete probability model

$$\begin{aligned}
w_i|z_i,\phi^{(z_i)} &\sim \text{Discrete}(\phi^{(z_i)}) \\
\phi &\sim \text{Dirichlet}(\beta) \\
z_i|\theta^{(d_i)} &\sim \text{Discrete}(\theta^{(d_i)}) \\
\theta &\sim \text{Dirichlet}(\alpha)
\end{aligned}$$

Collapsed Gibbs Sampling for LDA

- The joint probability is

$$p(w, z, \phi, \theta | \alpha, \beta) = \prod_i p(w_i | z_i, \phi^{(z_i)}) p(\phi | \beta) \cdot \prod_i p(z_i | \theta^{(d_i)}) p(\theta | \alpha)$$

- Due to conjugate priors, we can easily integrate out ϕ and θ (T. Griffiths & M. Steyvers, 2004)

$$p(w|z) = \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V} \right)^K \prod_{j=1}^{K} \frac{\prod_w \Gamma(n_j^{(w)} + \beta)}{\Gamma(n_j^{(\cdot)} + V\beta)}$$

$$p(z) = \left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \right)^M \prod_{d=1}^{M} \frac{\prod_j \Gamma(n_j^{(d)} + \alpha)}{\Gamma(n_\cdot^{(d)} + K\alpha)}$$

$n_j^{(w)} \leftarrow$ number of times word w assigned to topic j

$n_j^{(d)} \leftarrow$ number of times topic j used in document d

Gibbs Sampling

- Need full conditional distributions for variables
- We only sample z, whose conditional distributions is

$$p(z_i = j | z_{-i}, w) \propto \boxed{\frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + V\beta}} \boxed{\frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + K\alpha}}$$

Gibbs Sampling

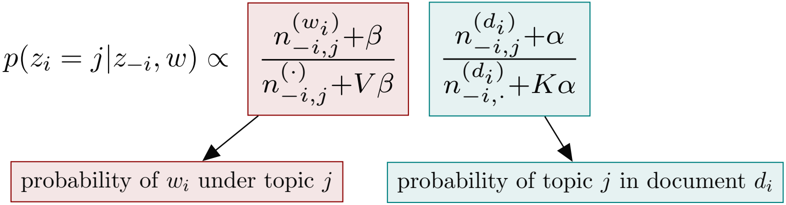
- Need full conditional distributions for variables
- We only sample z, whose conditional distributions is

$$p(z_i = j | z_{-i}, w) \propto \boxed{\frac{n^{(w_i)}_{-i,j} + \beta}{n^{(\cdot)}_{-i,j} + V\beta}} \boxed{\frac{n^{(d_i)}_{-i,j} + \alpha}{n^{(d_i)}_{-i,\cdot} + K\alpha}}$$

probability of w_i under topic j

Gibbs Sampling

- Need full conditional distributions for variables
- We only sample z, whose conditional distributions is



$$p(z_i = j | z_{-i}, w) \propto \boxed{\frac{n^{(w_i)}_{-i,j} + \beta}{n^{(\cdot)}_{-i,j} + V\beta}} \boxed{\frac{n^{(d_i)}_{-i,j} + \alpha}{n^{(d_i)}_{-i,\cdot} + K\alpha}}$$

probability of w_i under topic j

probability of topic j in document d_i

Gibbs Sampling

- Need full conditional distributions for variables
- We only sample z, whose conditional distributions is

$$p(z_i = j | z_{-i}, w) \propto \boxed{\frac{n^{(w_i)}_{-i,j} + \beta}{n^{(\cdot)}_{-i,j} + V\beta}} \boxed{\frac{n^{(d_i)}_{-i,j} + \alpha}{n^{(d_i)}_{-i,\cdot} + K\alpha}}$$

probability of w_i under topic j

probability of topic j in document d_i

- This is nicer than your average Gibbs sampler:
  - memory: counts can be cashed in two sparse matrices
  - the distributions on ϕ and θ are analytic given z and w, and can later be found for each sample

References

- D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. JMLR 3, 2003.
- Griffiths, Thomas L., and Mark Steyvers. 2004. “Finding Scientific Topics.” Proceedings of the National Academy of Sciences of the United States of America 101 (S1): 5228–35.
- C. J. Geyer (1991) Markov chain Monte Carlo maximum likelihood, Computing Science and Statistics, 23: 156-163.
- David J. Earl and Michael W. Deem (2005) ”Parallel tempering: Theory, applications, and new perspectives”, Phys. Chem. Chem. Phys., 7, 3910
- Neal, R. M. Slice sampling. Annals of Statistics, pp. 705–741, 2003.