Royal Institute of Technology

STAT. METH. IN CS – COLLAPSED GIBBS SAMPLER

Lecture 12

METROPOLIS HASTINGS (MH)

How?

We want to compute $p^*(x)$ (typically $p(x|D)$)

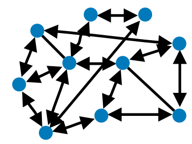
Implicitly construct Markov Chain M with stationary distribution $p^*(x)$

Traverse it and sample every k:th visit

Use good or random starting point

Discard the first l:th samples

The remaining samples $x_1,\dots,x_S$ is an approximation of $p^*(x)$



$p^*(x) \approx [\sum_i I(x=x_i)]/S$

GIBBS SAMPLING

- Pick initial state $x_1=(x_{1,1},\dots,x_{1,K})$
- For s=1 to S
  - Sample $k \sim_u [K]$
  - Sample $x_{s+1,k} \sim p(x_{s+1,k} | x_{s,-k})$
  - Let $x_{s+1} = (x_{s,1},\dots,x_{1,k-1}, x_{s+1,k},\dots, x_{s,K})$
  - If k|s record x_{s+1} (thinning)

Notation

$$\mathcal{D} = (x_1,\dots,x_N),\ \ H = (z_1,\dots,z_N),\ \ N_k = \sum_n I(z_i = k)$$

$$\boldsymbol{\pi} = (\pi_1,\dots,\pi_k),\ \ \boldsymbol{\mu} = (\mu_i,\dots,\mu_k),\ \ \boldsymbol{\lambda} = (\lambda_i,\dots,\lambda_k), \text{ and }\ \lambda_k = 1/\sigma_k^2$$

Hyperparameters $\boldsymbol{\theta}_0 = (\mu_0, \lambda_0, \lambda_0, \beta_0, \alpha)$

Model

$$\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha}),\ \ \mu_k \sim N(\mu_0, \lambda_0),\ \ \lambda_k \sim \mathrm{Ga}(\alpha_0, \beta_0),\ \ z_i \sim \mathrm{Cat}(\boldsymbol{\pi}), \text{ and}$$
$$p(x_n | Z_n = k) = N(\mu_k, \lambda_k)$$

GIBBS SAMPLER FOR GMM

A STATE

$$(\boldsymbol{H}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\lambda})$$

Hyperparameters $\boldsymbol{\theta}_0 = (\mu_0, \lambda_0, \lambda_0, \beta_0, \alpha)$

Model

$\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha}),\ \ \mu_k \sim N(\mu_0, \lambda_0),\ \ \lambda_k \sim \mathrm{Ga}(\alpha_0, \beta_0),\ \ z_i \sim \mathrm{Cat}(\boldsymbol{\pi})$, and $p(x_n|Z_n = k) = N(\mu_k, \lambda_k)$

Likelihood

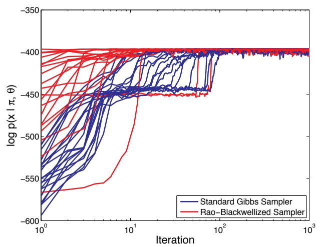
$$\begin{aligned} p(D, H, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\lambda}) =& p(D, H|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\lambda})p(\boldsymbol{\pi})p(\boldsymbol{\mu}, \boldsymbol{\lambda}) \\ =& \prod_{n,k}[\pi_k N(x_n|\mu_k, \lambda_k)]^{I(z_n=k)}\mathrm{Dir}(\pi|\alpha) \\ & \prod_k N(\mu_k|\mu_0, \lambda_0)\mathrm{Ga}(\lambda_k|\alpha_0, \beta_0) \end{aligned}$$

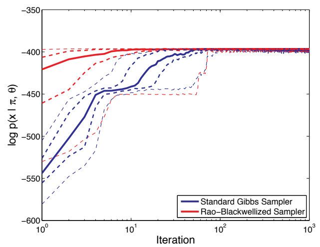
LIKELIHOOD FOR GMM

COLLAPSING

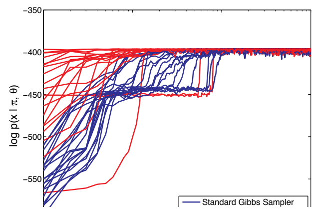
Integrating out some components of the state is called collapsing

It always improves convergence





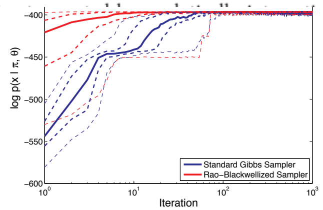
COLLAPSING



Theorem 24.2.1 (Rao-Blackwell). *Let* $\mathbf{z}$ *and* $\boldsymbol{\theta}$ *be dependent random variables, and* $f(\mathbf{z}, \boldsymbol{\theta})$ *be some scalar function. Then*

$$\mathrm{var}_{\mathbf{z},\boldsymbol{\theta}}[f(\mathbf{z}, \boldsymbol{\theta})] \geq \mathrm{var}_{\mathbf{z}}[\mathbb{E}_{\boldsymbol{\theta}}[f(\mathbf{z}, \boldsymbol{\theta})|\mathbf{z}]] \tag{24.20}$$

It always improves convergence



COLLAPSED GIBBS SAMPLER FOR GMM

Integrate out $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, and $\boldsymbol{\lambda}$ from

$p(D, H, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\pi}, \Theta_0)$

Θ_0 is all the hyperparameters

Only full conditionals on z_n remain

Same joint as before except from conjugate prior on μ_k and λ_k

NEW FULL CONDITIONAL

=3

$$
\begin{aligned}
&p(z_n|D, H_{-n}, \Theta_0) \\
&\quad = \frac{p(z_n, D|H_{-n}, \Theta_0)}{p(D|H_{-n}, \Theta_0)} \\
&\quad \propto p(z_n, D|H_{-n}, \Theta_0) \\
&\quad \propto p(z_n|H_{-n}, \Theta_0)p(D|z_n, H_{-n}, \Theta_0) \\
&\quad \propto p(z_n|H_{-n}, \Theta_0)p(x_n|D_{-n}, z_n, H_{-n}, \Theta_0)p(D_{-n}|z_n, H_{-n}, \Theta_0) \\
&\quad \propto p(z_n|H_{-n}, \Theta_0)p(x_n|D_{-n}, z_n, H_{-n}, , \Theta_0)
\end{aligned}
$$

“easy” with a conjugate prior on μ_k and λ_k but varies as we vary H

FULL CONDITIONAL

Recall, $\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha})$, $\mu_k \sim N(\mu_0, \lambda_0)$, $\lambda_k \sim \mathrm{Ga}(\alpha_0, \beta_0)$, $z_i \sim \mathrm{Cat}(\boldsymbol{\pi})$, and $p(x_n|Z_n = k) = N(\mu_k, \lambda_k)$

I.e., marginal

$$p(z_1, \ldots, z_N|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \prod_{k=1}^{K} \frac{\Gamma(N_k + \alpha/K)}{\Gamma(\alpha/K)}$$

So,

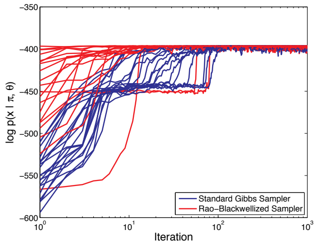
$$
\begin{aligned}
p(z_i = k|\mathbf{z}_{-i}, \alpha) &= \frac{p(\mathbf{z}_{1:N}|\alpha)}{p(\mathbf{z}_{-i}|\alpha)} = \frac{\frac{1}{\Gamma(N+\alpha)}}{\frac{1}{\Gamma(N+\alpha-1)}} \times \frac{\Gamma(N_k + \alpha/K)}{\Gamma(N_{k,-i} + \alpha/K)} \\
&= \frac{\Gamma(N+\alpha-1)}{\Gamma(N+\alpha)} \frac{\Gamma(N_{k,-i} + 1 + \alpha/K)}{\Gamma(N_{k,-i} + \alpha/K)} = \frac{N_{k,-i} + \alpha/K}{N + \alpha - 1}
\end{aligned}
$$

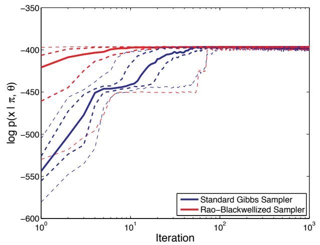
where $N_{k,-i} \triangleq \sum_{n \neq i} \mathbb{I}(z_n = k) = N_k - 1,$ $\qquad \alpha = \sum_k \alpha_k$

THE COLLAPSED ALGORITHM

Algorithm 24.1: Collapsed Gibbs sampler for a mixture model

```
for each i = 1 : N in random order do
    Remove x_i's sufficient statistics from old cluster z_i ;
    for each k = 1 : K do
        Compute p_k(x_i) ≜ p(x_i|{x_j : z_j = k, j ≠ i}) ;
    Compute p(z_i = k|z_{-i}, D) ∝ (N_{k,-i} + α/K) p_k(x_i);
    Sample z_i ~ p(z_i|·) ;
    Add x_i's sufficient statistics to new cluster z_i
```

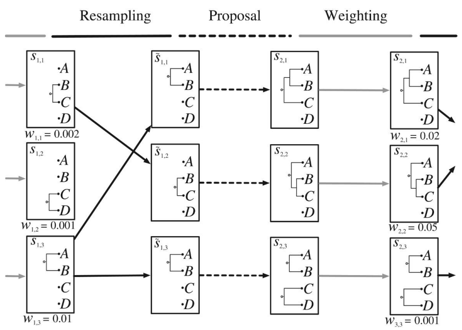


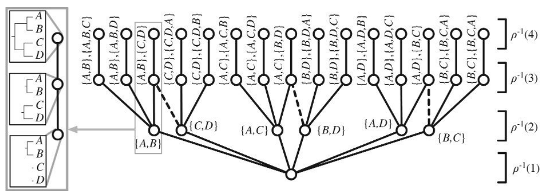


CONVERGENCE COLLAPSED VS STANDARD

- Blue is standard
- Red is collapsed
- x iterations
- y loglikelihood

THE END





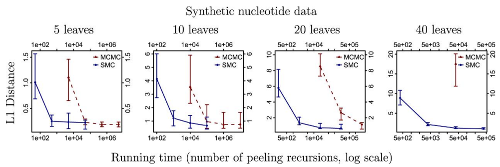


FIGURE 3. Comparison of the convergence time of PosetSMC and MCMC. We generated coalescent trees of different sizes and data sets of 1000 nucleotides. We computed the L1 distance of the minimum Bayes risk reconstruction to the true generating tree as a function of the running time (in units of the number of peeling recursions, on a log scale). The missing MCMC data points are due to MrBayes stalling on these executions.