Lecture 19 Gibbs Sampler.

Note Title 5/12/2006

The Gibbs sampler is easy to compute and requires no free parameters.

$\underline{X} = (X_1, \ldots, X_d)$

Notation. $\underline{X}_{/i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_d)$

Marginal Distribution $\pi(X_i \mid \underline{X}_{/i})$

usually there is a Markov assumption.

so that $\pi(X_i \mid \underline{X}_{/i}) = \pi(X_i \mid \underline{X}_{N(i)})$

where $N(i)$ is the neighborhood of i.

Gibbs sampler.

$$K_i(\underline{y} \mid \underline{x}) = \pi(y_i \mid \underline{x}_{/i}) \, \delta_{\underline{y}_{/i}, \underline{x}_{/i}}$$

Check detailed balance:

$$\pi(\underline{x}) K_i(\underline{y} \mid \underline{x}) = \pi(\underline{x}) \pi(y_i \mid \underline{x}_{/i}) \delta_{\underline{y}_{/i}, \underline{x}_{/i}}$$

$$= \pi(x_i \mid \underline{x}_{/i}) \pi(y_i \mid \underline{x}_{/i}) \pi(\underline{x}_{/i}) \delta_{\underline{y}_{/i}, \underline{x}_{/i}}$$

symmetric in $\underline{y}$ & $\underline{x}$, $= \pi(\underline{y}) K_i(\underline{y} \mid \underline{x})$

But this is not irreducible.

To make it irreducible, must sample all x_i.

$$K(\underline{y}|\underline{x}) = \sum_{i=1}^{n} \alpha_i K_i(\underline{y}|\underline{x})$$

with $\alpha_i > 0$, $\forall i$

$$\sum_{i=1}^{n} \alpha_i = 1.$$

This obeys detailed balance (linearity)

Algorithm: At each time step t at state $\underline{x}^t$.
select i with probability α_i
then select $\underline{y}$ from $K_i(\underline{y}|\underline{x}^t)$.

Typically random scan. $\alpha_i = 1/n$, $\forall i$.

Can also do systematic scan:

Let $\underline{x}^{t+1} = (x_1^{t+1}, \ldots, x_d^{t+1})$.

- Draw x_i^{t+1} from $\pi(x_i | x_1^{t+1}, \ldots, x_{i-1}^{t+1}, x_{i+1}^{t}, \ldots x_d^{t})$
scan through i.

Gibbs sampler converges geometrically (like M-H). The convergence rate depends on how well variables correlate with each other.

Example: Ising-Model

$$\pi(x_1, \ldots x_d) = \frac{1}{Z} e^{-\mu \sum_{i=1}^{d-1} x_i x_{i+1}} \qquad x_i \in \{-1, 1\}$$

$$\pi(x_i \mid \underline{x}_{\setminus i}) = \pi(x_i \mid x_{i-1}, x_{i+1})$$

Markov property.

Compute.

$$\pi(x_i \mid x_{i-1}, x_{i+1}) = \frac{\pi(x_{i-1}, x_i, x_{i+1})}{\pi(x_{i-1}, x_{i+1})}$$

$$\pi(\underline{x}) = \frac{1}{Z} e^{-\mu(x_{i-1}x_i + x_i x_{i+1})} \cdot f(x_{i+1} \ldots x_d) \cdot g(x_1, \ldots, x_{i-1})$$

$$\pi(x_{i-1}, x_i, x_{i+1}) = \frac{1}{Z} e^{-\mu(x_{i-1}x_i + x_i x_{i+1})} \sum_{x_{i+2}, \ldots x_d} f(x_{i+1}, \ldots x_d) \sum_{x_1, \ldots, x_{i-2}} g(x_1, \ldots, x_{i-1})$$

Hence $$\pi(x_i \mid x_{i-1}, x_{i+1}) = \frac{e^{\mu(x_{i-1}x_i + x_i x_{i+1})}}{\sum_{x_i} e^{-\mu(x_{i-1}x_i + x_i x_{i+1})}}$$

$$\pi(x_i \mid x_{i-1}, x_{i+1}) = \frac{e^{-\mu(x_{i-1}x_i + x_i x_{i+1})}}{e^{\mu(x_{i-1} + x_{i+1})} + e^{+\mu(x_{i-1} + x_{i+1})}}$$

Moral: the conditional is usually easy to compute for Markov Random Fields (MRF)

Page 4:

Example.

$\underline{X} = (X_1, X_2)$

$$\pi(\underline{x}) = N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right\}$$

$$\pi(x_1 | x_2) = N(\rho x_2, (1-\rho^2))$$

$$\pi(x_2 | x_1) = N(\rho x_1, (1-\rho^2))$$

Systematic scan.

$$\begin{pmatrix} X_1^{(t)} \\ X_2^{(t)} \end{pmatrix} \sim N\left\{ \begin{pmatrix} \rho^{2t-1} X_2^{(0)} \\ \rho^{2t} X_2^{(0)} \end{pmatrix}, \begin{pmatrix} 1-\rho^{4t-2} & \rho-\rho^{4t-1} \\ \rho-\rho^{4t-1} & 1-\rho^{4t} \end{pmatrix} \right\}$$

As $t \to \infty$, the joint distribution of $(X_1^{(t)}, X_2^{(t)})$ converges to the target distribution

Also, the rate of convergence is exponential.

Rate of convergence is equal to the maximal correlation between $X_i^{(t)}$ and $X_i^{(t+1)}$ which is ρ^2.

The Gibbs sampler can be thought of as a special case of Metropolis-Hastings.

Gibbs is M-H with a proposal which is automatically accepted.

Recall M-H

$$K(\underline{y}|\underline{x}) = T(\underline{y}|\underline{x}) \min\left\{1, \frac{\pi(\underline{y})\, T(\underline{x}|\underline{y})}{\pi(\underline{x})\, T(\underline{y}|\underline{x})}\right\}$$

Suppose $T(\underline{y}|\underline{x}) = K(\underline{y}|\underline{x})$.

Gibbs Sampler.

Now $\dfrac{\pi(\underline{y})\, T(\underline{x}|\underline{y})}{\pi(\underline{x})\, T(\underline{y}|\underline{x})} = 1$ because $K(\underline{y}|\underline{x})$ obeys detailed balance.

Metropolized Gibbs Sampler.

$\underline{x} = (x_1, \ldots, x_d)$

Each x_i takes m_i possible values.

- select i at random.
- draw y_i $(\neq x_i)$ with probability $\dfrac{\pi(y_i|\underline{x}_{/i})}{1-\pi(x_i|\underline{x}_{/i})}$

then replace x_i by y_i with probability $\min\left\{1, \dfrac{1-\pi(x_i|\underline{x}_{/i})}{1-\pi(y_i|\underline{x}_{/i})}\right\}$

(statistically more efficient than Gibbs)

Data Augmentation (DA)

A stochastic alternative to the EM algorithm.

$\underline{y}_{obs}$ ~ observed data, $\underline{y}_{mis}$ ~ missing data.

$p(\underline{\theta} \mid \underline{y}_{obs}, \underline{y}_{mis})$ & $p(\underline{y}_{mis} \mid \underline{y}_{obs})$

Given $p(\underline{\theta}, \underline{y}_{mis} \mid \underline{y}_{obs})$

Sample $\underline{\theta}$ & $\underline{y}_{mis}$ in turn.

Initialize: $\underline{\theta}^0$ & $\underline{y}_{mis}^0$

Sample $\underline{\theta}^t$ from $p(\underline{\theta} \mid \underline{y}_{mis}^{t-1}, \underline{y}_{obs})$

$\underline{y}_{mis}^t$ from $p(\underline{y}_{mis} \mid \underline{\theta}^t, \underline{y}_{obs})$

This is a form of Gibbs sampling and so is guaranteed to converge to samples from

$p(\underline{\theta}, \underline{y}_{mis} \mid \underline{y}_{obs})$

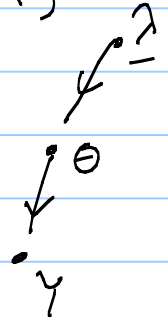
First Example: Hierarchical model.

$$Y_i | \theta_i \sim f_i(y_i | \theta_i)$$
$$\theta_i \sim G(\theta | \lambda)$$

Prior $\lambda \sim f_0(\mu, \tau^2)$

Want to estimate θ_i, λ and quantify their uncertainties: $\lambda = (\mu, \tau^2)$.

$$p(Y, \theta, \lambda) = p(Y|\theta)\, p(\theta|\lambda)\, p(\lambda)$$



Need. $p(\theta | Y, \lambda)$
$p(\lambda | Y, \theta)$

$$p(\theta | Y, \lambda) = \frac{p(Y|\theta)\, p(\theta|\lambda)}{\sum_{\theta} p(Y|\theta)\, p(\theta|\lambda)}$$

$$p(\lambda | Y, \lambda) = \frac{p(\theta|\lambda)\, p(\lambda)}{\sum_{\lambda} p(\theta|\lambda)\, p(\lambda)}$$

It may be impossible to compute the denominator — $\sum_{\theta} p(Y|\theta)\, p(\theta|\lambda)$ & $\sum_{\lambda} p(\theta|\lambda)\, p(\lambda)$ — if not, doing weighted sampling.

Second Example.

Data is 1-dimensional and is generated by one of two Gaussian distributions

$$N(\mu_1, \bar{\sigma}) \quad N(\mu_2, \bar{\sigma})$$

μ_1, μ_2 unknown random variables

$\bar{\sigma}$ known.

x^i data, v^i missing data

$v^i = 1$ means $x^i \sim N(\mu_1, \bar{\sigma})$

$= 0$ means $x^i \sim N(\mu_2, \bar{\sigma})$

This can be summarized by

$$P(x^i | v^i, \mu_1, \mu_2) = \frac{1}{\sqrt{2\pi}\bar{\sigma}} e^{\frac{-1}{2\bar{\sigma}^2}(x^i - v^i\mu_1 - (1-v^i)\mu_2)^2}$$

Prior for v^i is

$$P(v^i) = e^{v_i \log\alpha + (1-v_i)\log(1-\alpha)}$$

α - known.

Prior for μ_1 & μ_2 is

$$P(\mu_1, \mu_2) = \frac{1}{2\pi\bar{\sigma}_m^2} e^{\frac{-(\mu_1-\alpha_1)^2}{2\bar{\sigma}_m^2}} e^{\frac{-(\mu_2-\alpha_2)^2}{2\bar{\sigma}_m^2}}$$

$\alpha_1, \alpha_2, \bar{\sigma}_m^2$ known

For a set of data $\{x^i : i = 1 \text{ to } m\}$

Full distribution

$$P(\mu_1, \mu_2) \left\{ \prod_{i=1}^{m} P(x^i | v^i, \mu_1, \mu_2) P(v^i) \right\}$$

To do Data Augmentation, we need to compute

$$p(v^i \mid \underline{\mu}, x^i) \quad \text{for } i = 1 \text{ to } M.$$

& $p(\underline{\mu} \mid \{v^i : i = 1 \text{ to } M\}, \{x^i : i = 1 \text{ to } M\})$

$\underline{\mu} = (\mu_1, \mu_2)$

$$P(v^i \mid \underline{\mu}, x^i) = e^{-\frac{1}{2\sigma^2}(x^i - v^i\mu_1 - (1-v^i)\mu_2)^2} \times e^{v_i \log\alpha + (1-v_i)\log(1-\alpha)}$$

$$\overline{Z[\underline{\mu}, x^i]} \leftarrow \text{normalization constant.}$$

This can be simplified (homework)

$$P(\underline{\mu} \mid \{v^i\}, \{x^i\}) = \frac{1}{(2\pi\sigma^2)^{M/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{M}(x^i - v^i\mu_1 - (1-v^i)\mu_2)^2} \times \frac{1}{2\pi\sigma_m^2} e^{-\frac{(\mu_1-\alpha_1)^2}{2\sigma_m^2}} e^{-\frac{(\mu_2-\alpha_2)^2}{2\sigma_m^2}}$$

normalization constant. $\rightarrow Z[\{v^i\}, \{x^i\}]$

After some algebra.

$$p(\mu_1 \mid \{v^i\}, \{x_i\}) \sim N(\tilde{\mu}_1, \tilde{\sigma}_1^2)$$
$$p(\mu_2 \mid \{v^i\}, \{x_i\}) \sim N(\tilde{\mu}_2, \tilde{\sigma}_2^2)$$

with $$\tilde{\mu}_1 = \frac{\sigma_m^2\left(\sum_{i=1}^{M} v^i x^i\right) + \sigma^2\alpha_1}{\sigma_m^2 \sum_{i=1}^{M} v^i + \sigma^2}.$$

Page 10.

$$\widetilde{\sigma}_1^2 = \frac{\bar{\sigma}^2 \bar{\sigma}_m^2}{\bar{\sigma}^2 + \bar{\sigma}_m^2 \sum_{i=1}^{M} V_i}$$

$$\widetilde{\mu}_2 = \frac{\bar{\sigma}_m^2 \sum_{i=1}^{M} (1-V_i) X_i + \bar{\sigma}^2 \alpha_2}{\bar{\sigma}_m^2 \sum_{i=1}^{M} (1-V_i) + \bar{\sigma}^2}$$

$$\widetilde{\sigma}_2^2 = \frac{\bar{\sigma}^2 \bar{\sigma}_m^2}{\bar{\sigma}^2 + \bar{\sigma}_m^2 \sum_{i=1}^{M} (1-V_i)}.$$