Note Title 5/12/2006

Gibbs Sampler.

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

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

<u>Notation.</u> $\underline{x}_{/i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$

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)

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

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

Can also do <u>systematic scan</u>:

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

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

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.

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

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

Markov property.

Compute.

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

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

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

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)

Example. $\underset{\sim}{x} = (x_1, x_2)$

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

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

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

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.

Gibbs sampler as a special case of Metropolis-Hastings

Gibbs is M-H when the proposal 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 $T(\underline{y}|\underline{x}) = K_i(\underline{y}|\underline{x})$ Gibbs sampler.

<u>Claim</u> $\dfrac{\pi(\underline{y})T(\underline{x}|\underline{y})}{\pi(\underline{x})T(\underline{y}|\underline{x})} = 1$ because $K_i(\underline{y}|\underline{x})$ obeys detailed balance.

<u>Metropolized Gibbs Sampler.</u>

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

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.

E.M. as an alternative to this.

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

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

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

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

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

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

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

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

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

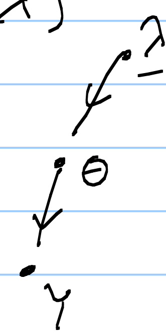
Example: Hierarchical model.

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

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

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

$$P(Y, \theta, \lambda) = P(Y \mid \theta)\, P(\theta \mid \lambda)\, P(\lambda)$$



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

$p(\lambda \mid Y, \theta)$

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

$$p(\lambda \mid Y, \lambda) = \frac{p(\theta \mid \lambda)\, P(\lambda)}{\sum_{\lambda} p(\theta \mid \lambda)\, P(\lambda)}.$$

It may be impossible to compute the denominator. But in that can use weighted importance sampling.

Second Example.

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

$N(\mu_1, \sigma) \quad N(\mu_2, \sigma)$

μ_1, μ_2 unknown random variables

σ known.

x^i data, V^i missing data

$V^i = 1$ means: $x^i \sim N(\mu_1, \sigma)$

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

This can be summarized by

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

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

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

$\alpha_1, \alpha_2, \sigma_m^2$ 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

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

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

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

$$P(V^i \mid \underline{\mu}, X^i) = \frac{e^{-\frac{1}{2\sigma^2}(x^i - V^i\mu_1 - (1-V^i)\mu_2)^2} \times e^{V_i \log \alpha + (1-V_i)\log(1-\alpha)}}{Z[\underline{\mu}, X^i]}$$

$\leftarrow$ normalization constant.

This can be simplified (homework)

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

normalization constant.

After some algebra.

$$p(\mu_1 \mid \{V^i\}, \{x_i\}) \sim N(\hat{\mu}_1, \hat{\sigma}_1^2)$$

$$p(\mu_2 \mid \{V^i\}, \{x_i\}) \sim N(\hat{\mu}_2, \hat{\sigma}_2^2)$$

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

$$\tilde{\sigma}_1^{\,2} = \frac{\bar{\sigma}^2 \bar{\sigma}_m^{\,2}}{\bar{\sigma}^2 + \bar{\sigma}_m^{\,2} \sum_{i=1}^{m} V^i}$$

$$\tilde{\mu}_2 = \frac{\bar{\sigma}_m^{\,2} \sum_{i=1}^{m} (1-V^i) X^i + \bar{\sigma}^2 \alpha_2}{\bar{\sigma}_m^{\,2} \sum_{i=1}^{m} V^i + \bar{\sigma}^2}$$

$$\bar{\bar{\sigma}}_2^{\,2} = \frac{\bar{\sigma}^2 \bar{\sigma}_m^{\,2}}{\bar{\sigma}^2 + \bar{\sigma}_m^{\,2} \sum_{i=1}^{m} (1-V^i)} .$$