STAT 353: Lecture 8

• The moves are very “local”
• Tend to be trapped in a local mode.

\[ \pi(x) \propto e^{-\frac{1}{2} \left( x_1^2 + x_2^2 \right)} + 2e^{-\frac{1}{2} \left( x_1^2 + x_2^2 \right)} \]

Multiple-Try Metropolis

Current is at \( x \)

• Draw \( y_1, \ldots, y_k \) from the proposal \( T(x, y) \).
• Select \( Y = y_j \) with probability \( \propto \pi(y_j) T(y_j, x) \).
• Draw \( x'_1, \ldots, x'_{k-1} \) from \( T(Y, x) \). Let \( x'_k = x \).
• Accept the proposed \( y_j \) with probability

\[ p = \min \left\{ 1, \frac{\pi(y_j) T(y_j, x) + \cdots + \pi(y_j) T(y_j, x)}{\pi(x'_1) T(x'_1, y_j) + \cdots + \pi(x'_k) T(x'_k, y_j)} \right\} \]

Let's prove the case when \( k=2 \) and \( T(x, y) = T(y, x) \).

Illustrating the Gibbs Sampler

• Purpose: Draw from a Joint Distribution \( x = (x_1, \ldots, x_n) \); target \( \pi(x) \)
• Method: Iterative Conditional Sampling

\[ \forall i, \text{ Draw } x_i \sim \pi(x_i|x_{-i}) \]

Illustrating the Gibbs Sampler

• Suppose the target distribution is:

\[ (X, Y) \sim \mathcal{N} \left( \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right) \]

• Gibbs sampler:

\[ [X|Y = y] \sim \mathcal{N} (\rho y, 1 - \rho^2) \]

\[ [Y|X = x] \sim \mathcal{N} (\rho x, 1 - \rho^2) \]

Start from, say, \((X, Y) = (10, 10)\), we can take a look at the trajectories. We took \( \rho = 0.6 \).
Example: A joint distribution of continuous and discrete r.v.s

- Target distribution: 
  \[ f(x, y) \propto \binom{n}{y} y^{x+y-1}(1-y)^{n-x+y} \]
  \[ x = 0, 1, \ldots, n; \quad 0 \leq y \leq 1 \]

- How to draw samples?
  - \([x \mid y] \sim \text{Binom} (n, y)\)
  - \([y \mid x] \sim \text{Beta} (x+\alpha, n-x+\beta)\)

- For illustration, we took \(\alpha = 0.5, \beta = 0.5; \) and \(n=10\)

More Generally

\[ \int p_x(x)p_{y\mid x}(y\mid x)dx = p_y(y) \]

\[ \int p_x(x)p_{y\mid x}(y\mid x)p_{x'\mid y}(x'\mid y)dy = p_x(x') \]
Missing data problem

- **Complete-Data Model:** $[x|\theta] - f(x|\theta)$
  - Example 1: $x$ -- multivariate normal, $\theta$ -- covariance;
  - Example 2: $x$ -- multinomial, $\theta$ -- unknown frequency.
  - Example 3: $x$ -- Markov chain, $\theta$ -- transition probabilities.
- **Sometimes** only part of $x$, say $y$, can be observed.
  Write the decomposition as $x=(y,z)$.
- **Observed-Data Model:**
  - Question: How to find the MLE?
  - Or posterior distribution of $\theta$?

\[
\begin{align*}
  p(y|x,\theta) &= \int f(y,z|x,\theta) dz \\
  \text{This is usually a difficult integration.}
\end{align*}
\]

Missing data problem is not “a problem”!

Many problems can be turned into missing data problems

- **Hierarchical Model/ Random Effects Models**

\[
\begin{align*}
  y_j &\sim \mathcal{N}(0, \alpha) \\
  \alpha &\sim \mathcal{N}(0, \sigma^2)
\end{align*}
\]

- $\theta_1, \ldots, \theta_n$ are iid $\mathcal{N}(\alpha, \sigma^2)$
- Can’t observe the random effects of the indicators!!

- **Latent Class/Mixture Models**

Example: $[y_i | I_i=1] \sim \mathcal{N}(\theta_1, \sigma^2)$
and $[y_i | I_i=0] \sim \mathcal{N}(\theta_0, \sigma^2)$

The Algorithm

- **Objective:** find the most “common” pattern.

**Example:** Sequence Alignment and motif finding

| $a_1$ | A motif site |
| $\vdots$ | $a_k$ |
| **width** $n$ |
| **length** $n_k$ |

Alignment variable: $A = \{a_1, a_2, \ldots, a_k\}$

**Objective:** find the most “common” pattern.

**Statistical Model:**
- Every non-site positions follows a common multinomial with $p_0 = (p_{0,1}, \ldots, p_{0,20})$
- Every position $i$ in the motif element follows probability distribution $p_i = (p_{i,1}, \ldots, p_{i,20})$

**The Algorithm**

- Initialized by choosing random starting positions $a_1^{(0)}, a_2^{(0)}, \ldots, a_K^{(0)}$
- Iterate the following steps many times:
  - Randomly or systematically choose a sequence, say, sequence $k$, to exclude
  - Carry out the predictive-updating step to update $a_k$
- Stop when not much changes observed, or some criterion met
Why does the Gibbs sampler work?

- It is a Markov Chain!

If \( X^{(0)} = x_0 \), then distribution of \( X^{(t)} \) is

\[
\mathcal{N}(\rho^{2t}x_0, 1 - \rho^{4t})
\]

which "converges" to \( \mathcal{N}(0,1) \) as \( t \to \infty \).

Joint distribution of \( (X^{(t)}, Y^{(t)}) \)?

\[
\begin{pmatrix}
\rho^{2t}x_0 \\
1 - \rho^{2t}
\end{pmatrix}
\begin{pmatrix}
\rho^{2t}y_0 \\
1 - \rho^{2t}
\end{pmatrix}
\]

Theory About Data Augmentation

- A graphical view and some basic identities

\[
\begin{align*}
E\{h(\theta^{(m)}) \mid \theta^{(n-1)}\} &= E\{E[h(\theta^{(m)}) \mid \theta^{(n-1)}] \mid z^{(n)}\} \\
&= E\{E[h(\theta) \mid z] \}
\end{align*}
\]

\[
\begin{align*}
\text{cov}[h(\theta^{(m)}) \mid \theta^{(n-1)}] &= \text{var}[E[h(\theta) \mid z]] \\
\text{cov}[h(\theta^{(m)}) \mid \theta^{(n-1)}] &= \text{var}[E[\cdot \mid E[h(\theta) \mid \theta]] \mid z^{(n)}]
\end{align*}
\]

(Literally, these are the main results of my thesis)

Comparing schemes and estimators

- Schemes: grouping and collapsing
  - (1) \([x | y] \) and \([y | x] \)
  - (2) \([x; y, z] \) and \([y, z | x] \)
  - (3) \([x; y, z], [y; x, z], [z; x, y] \)

- Estimators:

\[
\begin{align*}
\hat{\theta} &= \frac{1}{m} \{h(\theta^{(0)}) \ldots h(\theta^{(m)})\} \quad \text{histogram} \\
\hat{\theta} &= \frac{1}{m} \{h(\theta) \mid Z^{(0)} \ldots h(\theta) \mid Z^{(m)}\} \quad \text{mixture}
\end{align*}
\]