STAT 311: Lecture 9

Theory About Data Augmentation

• A graphical view and some basic identities

\[ E\{h(\theta^{(0)})h(\theta^{(1)})\} = E\left[E\{h(\theta^{(0)})h(\theta^{(1)}) \mid z^{(0)}\}\right] \]
\[ = E\left[E_{\pi}^2\{h(\theta) \mid z\}\right] \]
\[ \text{cov}\{h(\theta^{(0)}), h(\theta^{(1)})\} = \text{var}_{\pi}\left[E\left\{h(\theta) \mid z\right\}\right] \]
\[ \text{cov}\{h(\theta^{(0)}), h(\theta^{(k)})\} = \text{var}_{\pi}\left[E\left\{\cdots E\{E\{h(\theta) \mid z\} \mid \theta\}\cdots\right\}\right] \]

(Literally, these are the main results of my thesis)
Comparing schemes and estimators

• Schemes: grouping and collapsing
  – (1) \([x \mid y] \) and \([y \mid x]\)
  – (2) \([x \mid y,z]\) and \([y,z \mid x]\)
  – (3) \([x \mid y,z], [y \mid x,z], \) and \([z \mid x,y]\)

• Estimators:
  \[ \hat{h} = \frac{1}{m} \{ h(\theta^{(1)}) + \cdots + h(\theta^{(m)}) \} \]
  \[ \tilde{h} = \frac{1}{m} \{ E[h(\theta) \mid Z^{(1)}] + \cdots + E[h(\theta) \mid Z^{(m)}] \} \]
  histogram
  mixture
• Gibbs sampler for Missing data problem: **Data augmentation**

Iterate between $[z \mid y, \theta]$ and $[\theta \mid z, y]$

– Original version of DA:
  • Draw $z_1, \ldots, z_m$, based on the current approximation of the posterior distribution of $\theta$;
  • Form a new approximated posterior distribution based on $z_1, \ldots, z_m$;
  • Iterate the above.

– A simpler version (the Gibbs sampler): *draw one $z$ and one $\theta$; iterate.*
Ising Model: Swendsen-Wang Algorithm

• Clustering algorithm as a data augmentation scheme

\[ \pi(x) \propto \exp \left[ \beta J \sum_{\sigma \sim \sigma'} \{1 + x(\sigma)x(\sigma')\} \right] \]

– Introduce an auxiliary variable \( u \) so that

\[ \pi(x, u) \propto \prod_{\sigma \sim \sigma'} I[0 \leq u_{\sigma\sigma'} \leq \exp\{\beta J(1 + x(\sigma)x(\sigma'))\}] \]

  • \([u | x]\) uniform distributions;
  • \([x | u]\) clustering and flipping.

– Equivalent to using binary “bond” variables \( b \):

\[ \pi(x, b) \propto \prod_{x(\sigma) = x(\sigma')} \{1 + b_{\sigma\sigma'}(e^{2\beta J} - 1)\} \]

• Generalizations:

\[ \pi(x) \propto \pi_0(x) \prod_k f_k(x) \]

\[ \pi(x, u) \propto \pi_0(x) \prod_k I[0 \leq u_k \leq f_k(x)] \]
Wolff’s Algorithm

• Feature: single cluster, complete flip-over
• Method:
  – Random pick a “seed spin”;
  – Grow the cluster from the seed the same way as in the SW algorithm, until it stops growing.
  – Flipping all the spins in the cluster simultaneously.

From Metropolis algorithm point of view:

Proposal ratio: \( \frac{T(y \rightarrow x)}{T(x \rightarrow y)} \)

Only those bonds along the boundary are influential.
Further Generalization

• Niedermeyer (1988): grow a cluster for both types.

\[
p_a = \text{P(adding a bond between the same spins)}
\]

\[
p_b = \text{P(adding a bond between different spins)}
\]

\[
\frac{T(y \rightarrow x)}{T(x \rightarrow y)} = \frac{(1 - p_a)^{n_b} (1 - p_b)^{n_a}}{(1 - p_a)^{n_a} (1 - p_b)^{n_b}} = \left(\frac{1 - p_a}{1 - p_b}\right)^{n_b - n_a}
\]

One can choose appropriate \( p_a \) and \( p_b \) to cancel out the probability change ---- so that all the proposed move will be accepted with probability one.

How is this connected with the conditional sampling?
Gibbs Sampler/Heat Bath

• Define a “neighborhood” structure $N(x)$
  – can be a line, a subspace, trace of a group, etc.
• Sample from the conditional distribution.
• Conditional Move

\[ x_{new} \sim p(x) \propto \pi(x) \mid x \in N(x_{old}) \]
General Conditional Sampling

• Gibbs sampler:
  – (a) space decomposition; (b) conditional sampling.
  – Restricted by the parameterization/coordinate system.

• More general conditional move:
  – Formulate “a move” in the space as a point being transformed/mapped to another point.
  – “possible moves” \( \Leftrightarrow \) a set of transformations.
  – Gibbs sampler: \( \pi(x_1, y_1) \Rightarrow \pi(x_2, y_1) \)
    \[ x_2 \sim [x \mid y_1] \propto \pi(x, y_1) \]
    • Can be seen as: \( x_1 \Rightarrow x_2 = x_1 + c \); 
    – \[ [c \mid (x_1, y_1)] \propto \pi(x_1 + c, y_1) \]
More generally

- **Maybe** \((x_1, y_1) \Rightarrow (x_2, y_2) = (x_1 + c, y_1 + c)\)

  **Effect:** like a reparametrization

- **Principle:** invariance of \(\pi\) under the move

- **Distribution:**

\[
c \sim \left[ c \mid (x_1, y_1) \right] \propto \pi(x_1 + c, y_1 + c)
\]
• Even more imagination ….
  – Try \((x_1, y_1) \rightarrow (x_2, y_2) = \lambda (x_1, y_1)\)
    what is the distribution of \(\lambda\)?
  – Or \((x_1, y_1) \rightarrow (x_2, y_2) = (x_1, y_1) A\)
    where \(A\) is a orthonormal matrix.

• **General Formulation:** a group of transformations
to represent possible moves: \(\gamma \in \Gamma\)
  \[x \rightarrow \gamma(x) \equiv \gamma x\]
  – Choose \(\gamma \sim p_x(\gamma)\) so that \(x' = \gamma x \sim \pi\) if \(x \sim \pi(x)\)

Note: the group of transformations
partitions the space as “orbits”
A Theorem

- If $\Gamma = \{\text{all } \gamma \}$ forms a *locally compact group*
- Let $H(d\gamma)$ be its unimodular *Haar measure*
- If $x \sim \pi(x)$, and
  \[ \gamma \sim [\gamma \mid x] \propto \pi(\gamma x) J_\gamma(x) H(d\gamma) \]

  Then $w = \gamma x$ follows distribution $\pi$

**Note:** A left invariant Haar measure satisfies:

\[ H(\gamma B) = H(B), \quad \forall \gamma, B \]
Invariance

The distribution \([\gamma \mid x] \propto \pi(\gamma x) \mid J_\gamma(x)\mid H(d\gamma)\) is “independent” of position \(x\)

Let \(x' = \alpha x\); then

\[
\pi(\gamma x')J_\gamma(x')H(d\gamma) = \pi(\gamma x) J_\gamma(\alpha x)\mid H(d\gamma) = \pi(\gamma' x) J_{\gamma'}(x)\mid H(d\gamma') | J_\alpha(x) |
\]

So the new “position” such drawn is independent of the old one

Property of Jacobian

\(J_\beta(\alpha x) J_\alpha(x) = J_{\beta \alpha}(x)\)
Grouping, Reparameterizing, and Group Move

- **Grouping/blocking:** \( x_{d-l+1}^* = (x_{d-l+1}, \ldots, x_d) \)
  
  Use Gibbs to iterate on \((x_1, \ldots, x_{d-k}, x_{d-k+1}^*)\)

- **Reparameterizing:** move along certain directions, e.g., rotations.
  
  (Gelfand et al. 1992; Nandrum & Chen 1996)

- **Group moves:** a subset of variables move together. (Shephard and Pitts, 1997)
A Demonstration of Multilevel Group Moves

\[ (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8) \]

Multigrid Monte Carlo (Goodman and Sokal, 1989)

A Graphical Representation

Update \( z \rightarrow z' = z + t \chi_A \)

With \( t \) drawn from \( p(t) \propto e^{-\beta H(z + t \chi_A)} \)

where \( A \) is a subset of lattice sites, and \( \chi_A \) is an indicator function. \( e.g., \) \( A = \{ \sigma \} \) in Gibbs sampler; or \( A = \{ \text{vertices in } \} \)

Requirement: if \( z \sim \pi(z) \), then \( z + t \chi_A \) also follows \( \pi \).
Example: Inference for t-distribution

• \( y_1, \ldots, y_n \) iid \( \sim t(0, \Sigma, \kappa) \) with \( \Sigma \) the only unknown.

• Treated as missing data problem:
  – Introduce missing data \( z_1, \ldots, z_n \) iid \( \sim \chi^2(\kappa)/\kappa. \)
  – Then \( [y_i \mid z_i, \Sigma] \sim N(0, z_i\Sigma). \)
  – Joint distribution:
    \[
    p(z, \Sigma | y) \propto |\Sigma|^{-n/2} \prod_{i=1}^{n} z_i^{1/2} \exp \left( -tr \left( \sum_{i=1}^{n} \frac{y_i y_i^T}{z_i^2} \Sigma \right) \right) p(z, \Sigma)
    \]

• Data Augmentation:
  \[ p(\Sigma) \propto |\Sigma|^b \]
  – \( [\Sigma \mid y, z] \sim \text{inverse wishart} \)
  – \( [z \mid y, \Sigma] \sim \text{scaled Chi-square} \)

• Problem with the algorithm?
Parameter expansion

• A trivial random-effect model:

\[ y \mid \theta, z \sim N(\theta + z, 1); \quad z \mid \theta \sim N(0, D) \]

  – DA:

\[
[z \mid y, \theta] = N\left(\frac{y - \theta}{1 + D^{-1}}, \frac{1}{1 + D^{-1}}\right)
\]
\[
[\theta \mid y, z] = N(y - z, 1)
\]

  – Introduce another parameter \( \alpha \):

\[ y \mid \theta, w \sim N(\theta - \alpha + w, 1); \quad w \mid \theta \sim N(\alpha, D) \]

Equivalent to: \((\theta, z) \rightarrow (\theta + \alpha, z - \alpha)\)
Group Move in t-Distribution

Graphical Representation:

• Let $\Gamma = \{\gamma\}$ so that
  
  $$\gamma(\Sigma, z_1, z_2, \ldots, z_n) = (\gamma \Sigma, \gamma z_1, \gamma z_2, \ldots, \gamma z_n).$$

• The group move: given the current $z$’s
  
  - Draw $\gamma$ from
    $$\chi^2_{n\nu}(\nu \sum_{i=1}^{n} z_i)$$
  
  - Update: $\Theta_{t+1} = \gamma \ast \Theta_t$

$\chi^2(\kappa)/\kappa$
Probit regression model

• Model: \( Y_i \mid \theta \sim \text{Bernoulli}\{\Phi(X_i^T \theta)\} \)

• Introduce missing data: \( Z_i \mid \theta \sim N(X_i^T \theta, 1) \)

• DA:
  - Imputing: \([z_i \mid \theta] = N(X_i' \theta, 1)\)
  - Inference: \([\theta \mid \text{all } z_i] = N(\hat{\theta}, V)\),
    \[ \hat{\theta} = \left(\sum_i X_i X_i'\right)^{-1} \sum_i X_i z_i; \quad V = \left(\sum_i X_i X_i'\right)^{-1} \]

• Group move:
  \((Z_1, \ldots, Z_n; \theta) \rightarrow (\alpha Z_1, \ldots, \alpha Z_n; \alpha \theta)\)
A probit regression example

(a) $\beta_1 = 1$

(b) $\beta_1 = 2$

(c) $\beta_1 = 4$

(d) $\beta_1 = 8$