Last Time

Bayesian statistics
Conjugate Distributions
Decomposition of Likelihood for complete data.

Today

Incomplete data
EM
Decomposition of Likelihood

\[ L[\Theta : \mathbf{D}] = P\{\mathbf{D} | \Theta\} = \prod_{n=1}^{N} L_n(\Theta_n, \mathbf{D}) \]

Independence of Parameters.

```r
model {
  p.a[1:2] ~ ddirch(prior[1])
  p.b[1,1:2] ~ ddirch(prior[1])
  p.b[2, 1:2] ~ ddirch(prior[1])
  for(i in 1:50) {
    a[i] ~ dcat(p.a[1:2])
    b[i] ~ dcat(p.b[a[i], 1:2])
  }
}
```

Plots of densities

Plot of \( P\{A=T\} \) vs \( P\{B=T|A=T\} \)
Missing Data

Arises through

- Missing Observations
  - Medicine: Not all tests are ran on each patient.
  - Surveys: Not all respondents answer all of the questions.
- Hidden Variables:
  - Target tracking: Unknown state of aircraft in a kalman filter.
  - Factor analysis: Unknown factors to explain variation in data.

Say that some of the A's are missing...

What happens?
Effect 1:
Correlation between parameters.

$p.b(a=t)$

Effect 2:
Multimodal posteriors.

$\mu = cA + d$

A is binary, \{0,1\}. A is never observed

Q: What is the likelihood function when data is missing?
Missing Data: Key Assumption

Missing at Random (MAR)
The probability that a value of $X$ is missing is independent of its actual value given other observed data.

What if MAR does not hold?
Establish an observable variable that models the probability that the data is observed given the data.

Latent (or Hidden) Variables

If these models are binary, how many independent parameters?
Why are Latent Variables a good idea?

\[
\text{Likelihood} \\
L[\Theta : D] = P(D | \theta) = \int P(D, X | \theta) dX
\]

D is observed data,
X is unobserved variables.
MLE w/ incomplete data

Nonlinear optimization problem
Select parameters to optimize the probability of the evidence.

Approaches:
Approach 1:
Gradient-based optimization

Approach 2:
Expectation-Maximization

Approach 3:
MCMC to pick distribution.
Pick mode from distribution.

Gradient Ascent (lifted from Friedman + Goldszmidt)
The problem \[
\max_{\Theta} \log (P \{D \mid \Theta\})
\]
s.t. \( h(\Theta) = 0 \) constraints on parameters

Closed form method for computing \( \nabla_{\Theta} \log (P \{D \mid \Theta\}) \)
\[
\frac{\partial \log (P \{D \mid \Theta\})}{\partial \theta_{x_i,pa_i}} = \frac{1}{\Theta_{x_i,pa_i}} \sum_m P\{x_i, pa_i \mid D(m), \Theta\}
\]

F + G Tutorial says:
Closely related to neural net training.
Difficulty: “Need to project gradient onto space of legal parameters”

Expensive bit is computing \( P\{x_i, pa_i \mid D(m), \Theta\} \) for each sample.
EM (Baum + Petrie, 66; Dempster, Laird + Rubin, 77)

Basic idea:
If our data was complete, it would be easy to update the parameters in the network.
Complete the counts using expected statistics for unobserved variables given the current value of the parameters.

EM:
(E-Step) [Expectation] Compute expected sufficient statistics and treat them as the real sufficient statistics, S.
(M-Step) [Maximization] Select parameters for the model that maximize the likelihood, $L[S(X):\Theta]$.
Repeat until convergence.

EM Example

\[
\begin{array}{c|c|c|c}
A & B & L \\
\hline
H & H & 1.6 \\
H & T & 1.4 \\
T & H & 1.2 \\
T & T & 0.8 \\
\end{array}
\]

P(B=H|A=H) = 0.6
P(B=H|A=H) = 0.2

E-Step

\[
\begin{array}{c|c|c|c}
A & B & L \\
\hline
H & H & 1.6 \\
H & T & 1.4 \\
T & H & 1.2 \\
T & T & 0.8 \\
\end{array}
\]

\[
\hat{\theta}_{H|H} = \frac{1.6 + 1}{1.6 + 1.4 + 2} = 0.52
\]

\[
\hat{\theta}_{H|T} = \frac{1.2 + 1}{1.2 + 0.8 + 2} = 0.55
\]

M-Step

Expected counts
Optimization

\[ L(D|\theta) \]

\[ \theta \]

Optimization

\[ L(D|\theta) \]

\[ \theta \]

M
Optimization

\[ L(D|\theta) \]

\[ \theta \]

Optimization

\[ L(D|\theta) \]

\[ \theta \]
EM Theory [Neal+Hinton, 93]

Define \( L(\Theta) = \log P[D | \Theta] = \log \int P[D, X | \Theta] dX \)

\[
\log \int P[D, X | \Theta] dX = \log \int Q(X) \frac{P[D, X | \Theta]}{Q(X)} dX \\
\geq \int Q(X) \log \frac{P[D, X | \Theta]}{Q(X)} dX \quad \text{Jensen’s Inequality} \\
= \int Q(X) \log P[D, X | \Theta] dX - \int Q(X) \log Q(X) dX \\
= F(Q, \Theta)
\]

EM Theory [Neal+Hinton, 93]

\( \Theta = \text{argmax} \, F(Q, \Theta) \)

E-step \( Q_{\text{E}} \sim \arg \max_{Q} \, F(Q, \Theta) \)

M-step \( \Theta_{\text{M}} = \arg \max_{\Theta} \, F(Q_{\text{E}}, \Theta) \)

E-step sol’n

Maximum for E-step achieved when \( Q_{\text{E}} \sim X \sim P[X \mid D, \Theta] \) then

\[
F(Q, \Theta) = \int P[X \mid D, \Theta] \log \frac{P[X \mid D, \Theta]}{P[X \mid D, \Theta]} dX \\
= \int P[X \mid D, \Theta] \log \frac{P[D \mid \Theta] P[X \mid D, \Theta]}{P[X \mid D, \Theta]} dX \\
= \log P[D \mid \Theta] \\
= L(\Theta)
\]
EM Theory

M-step maximum:

\[ \Theta_{k+1} \leftarrow \arg \max_{\Theta} \int P\{X \mid D, \Theta_k\} \log P\{X, D \mid \Theta\} dX \]

\[ \Theta_{k+1} \leftarrow \arg \max_{\Theta} \int P\{X \mid D, \Theta_k\} \log \prod_i P[z_i \mid p_{\theta_i}] dX \]

Parameter independence (Assume z is not always observed)

\[ \theta_{k+1} \leftarrow \arg \max_{\theta} \int P[z_i \mid D, \Theta_k] \log P[z_i \mid p_{\theta_i}] dX \]

Incomplete Data

The M-Step of EM

\[ \Theta_{k+1} \leftarrow \arg \max_{\Theta} \int P\{X \mid D, \Theta_k\} \log P\{X, D \mid \Theta\} dX \]

\[ \int P\{X \mid D, \Theta_k\} \log L[\Theta \mid \Theta, X] dX \]

definition of likelihood

\[ = \int P\{X \mid D, \Theta_k\} \log L[\Theta \mid \Theta, X] dX \]

factorization

\[ = \sum_i \left( \int P\{X \mid D, \Theta_k\} \log L[\Theta \mid \Theta, X] dX \right) \]

reorder summation

\[ = \sum_i \left( \log L[\Theta \mid \Theta, X] \right) \]

definition of expectation

\[ \leq \sum_i \log L[\Theta \mid \Theta, X] \]

Jensen’s inequality IF the likelihood is log concave.

Implication: Select \( \Theta \) that optimizes \( L[\Theta \mid D, X] \) for expected sufficient statistics.
EM Convergence

Convergence
At the beginning of every M-step, \( F(Q, \Theta_i) = L(\Theta_i) \)

The combined EM-step cannot decrease \( F(Q, \Theta) \),
so \( L(\Theta_{i+1}) \geq L(\Theta_i) \)

Stable stationary point is a local maximum.

EM Practice (F+G tutorial)

Initial parameters
Same as BUGS: random or educated guess.

Stopping criteria
Small change in likelihood.
Small change in parameter values.

Finding global maximum
Multiple restarts

Speed up
Koller: \( \Delta \Theta = \Theta_{i+1} - \Theta_i \) establishes a gradient.
\( \Theta_{i+1} = c \Delta \Theta + \Theta_i \) provides faster convergence when \( c > 1 \)
Bayesian Inference with Incomplete Data

\[ P\{X \mid D\} = \int P\{X \mid \theta\}P\{\theta \mid D\}d\theta \]

No closed form solution.
No sufficient statistics.

MAP Approximation

\[ P\{X \mid D\} = P\{X \mid \Theta\} \]

where \( \Theta = \arg \max_{\Theta} P\{X \mid \Theta\} \)

Alternative: MCMC with missing data.