

## Last Time

MLE for incomplete data.

MLE no longer factors

Parameter ~~Estimation~~ Optimization

Gradient ascent

EM algorithm

## This Time

Promised, but will not deliver a PS 5 discussion

(Forgot to take KBs home)

Incomplete Data

EM Algorithm II

Exact solution for multinomial missing data

Bayesian inference

Finding the best structure (Maybe)

Scoring metrics

Structure optimization

These slides draw heavily on Friedman+Goldszmid's 1998 tutorial

## Last Time: EM

Presented the Neal+Hinton [93] formulation.

$$F(Q, \Theta) = \int_X Q\{X\} \log P\{D, X | \Theta\} dX - \int_X Q\{X\} \log Q\{X\} dX$$

Makes convergence obvious.

E-step  $Q_{k+1} \leftarrow \arg \max_Q F(Q, \Theta_k)$

M-step  $\Theta_{k+1} \leftarrow \arg \max_{\Theta} F(Q_{k+1}, \Theta)$

## Last Time: Prove the M-Step of EM

$$\begin{aligned} \Theta_{k+1} &\leftarrow \arg \max_{\Theta} \int_X P\{X | D, \Theta_k\} \log P\{X, D | \Theta\} dX \\ &= \int_X P\{X | D, \Theta_k\} \log L[\Theta : D, X] dX && \text{definition of likelihood} \\ &= \int_X P\{X | D, \Theta_k\} \sum_i \log L_i[\theta_i : S_i(D, X)] dX && \text{factorization} \\ &= \sum_i \left( \int_X P\{X | D, \Theta_k\} \log L_i[\theta_i : S_i(D, X)] dX \right) && \text{reorder summation} \\ &= \sum_i E_{X|D, \Theta_k} [\log L[\theta_i : S_i(D, X)]] && \text{definition of expectation} \\ &\leq \sum_i \log L[\theta_i : E_{X|D, \Theta_k} [S_i(X)]] && \text{Jensen's inequality IF the} \\ & && \text{likelihood is log concave.} \end{aligned}$$

Implication: Select  $\Theta$  that optimizes  $L[\Theta : D, X]$  for *expected sufficient statistics*.

## Exact Solution

Single incomplete case.

Dirichlet prior.

“Parameter independence” assumption.

H is unobserved, Y is observed.

$$\begin{aligned}
 P\{\Theta_i | Y\} &= \sum_Z P\{Z | Y\} P\{\Theta_i | Y, Z\} \\
 &= (1 - P\{pa_i | y\}) P\{\Theta_i\} + \sum_x P\{x, pa_i | Y\} P\{\Theta_i | x, pa_i\} \\
 &= (1 - P\{pa_i | y\}) D(N_i) + \sum_x P\{x, pa_i | Y\} D(N_i(x, pa_i)) \\
 &= \sum_j w_j D(N_j)
 \end{aligned}$$

Answer is a mixture of Dirichlet distributions.

## Bayesian Inference

### Determining $\Theta$

Complete Data: exact results for conjugate distributions

Incomplete Data: No closed form for Bayesian prediction

### Today: Simple story

Primary interest is in finding the best  $\Theta$ .

### Later:

Primary interest will be Bayesian model selection w/incomplete data

Interested in finding an approximation to  $P\{D|\Theta\}$  and  $\Theta$ .

Approximations:

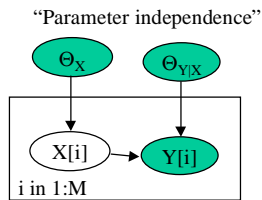
MAP/Laplace (today)

MCMC (today)

BIC

Cheeseman/Stutz

## Bayesian Inference with Incomplete Data



$$P\{X[M+1] | D\}$$

$$= \int P\{X[M+1] | \theta_x, \theta_{y|x}\} P\{\theta_x, \theta_{y|x} | D\} d\theta$$

### Problems:

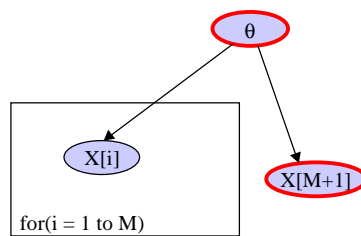
No closed form solution for Bayesian prediction

### Common Approximations:

MCMC

Maximum A-posteriori (MAP)

## MCMC Solution



## MAP Approximation

### Approximate

$$P\{X[M+1] | D\} \approx P(X[M+1] | \tilde{\Theta})$$

$$\tilde{\Theta} = \arg \max_{\Theta} P\{\Theta | D\}$$

Later: We will show that MAP arises from Gaussian approximation.

### Limitations

$\tilde{\Theta}$  cannot lie on the boundary of  $\Theta$

Given  $D$ , there is a unique MAP

*Aliasing:* Interchange labels on a hidden variable and  $P\{\Theta | D\}$  is unchanged.

*Reduced dimensionality:* The likelihood can be encoded with a smaller set of parameters than  $\Theta$ . Leads to an infinite number of MAP values.

## Gaussian (MAP) Approximation

### Finding MAP:

Same as ML, save Maximizing  $P\{\Theta | D\}$  instead of  $L\{\Theta; D\}$   
Include "prior" statistics in addition to statistics from data.

Given structure, the MAP assumption means:

$$\text{Select } \tilde{\Theta} = \arg \max_{\Theta} P\{\Theta | D\}$$

## Where are we?

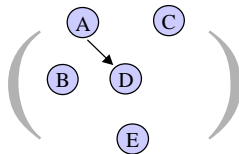
	Known Structure	Unknown Structure
Complete Data	Statistical parameter estimation Presented Wednesday	Optimization over structures Right now.
Incomplete Data	Parameter optimization Presented Friday Bayesian estimation covered today.	Optimization over structures and parameters

Structure Optimization

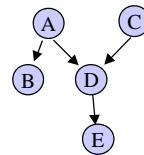
## Structure Optimization (Complete Data)

A	B	C	D	E
1	2	1	0	1
1	1	0	1	1
0	1	1	1	1
1	1	1	1	2

+



Select graph that maximizes a score



+

$P\{A\}$   
 $P\{B|A\}$   
 $P\{C\}$   
 $P\{D|A,C\}$   
 $P\{D|A\}$   
 $P\{E|D\}$

## Structure-Based Learning Techniques

Goal: find the best structure to explain the data

(Controversial benefit): Identify “causal” relations via the V-structures of the corresponding graph.

Two approaches

Constraint-based

Test for CISs (Just like PS 4)

Search for network consistent with the CIS statements.

Sensitive to errors in independence tests.

No statistical foundation

Score-based (what we will talk about)

Define a score that describes how well structure models observations

Search for structure that maximizes the score

## Focus on score-based approaches

Define score

Structure: Scoring metrics

I. Likelihood

II. MDL

III. BDe

Find structure that maximizes score

Structure: Optimization

Special case: Trees

General DAGs

Issues

Model averaging

Structure: Issues

## I. Likelihood score

Use likelihood function

$$L[G, \Theta_G : D] = \prod_{m=1}^M P\{X[m] | G, \Theta_G\}$$

$$= \prod_{m=1}^M \prod_{i=1}^N P\{x_i[m] | Pa_i^G[m] : G, \Theta_G\}$$

Define:  $L[G : D] = \max_{\Theta_G} L[G, \Theta_G : D]$

## I. Likelihood Score

Rewrite likelihood score:

$$I(G : D) = \log L[G, D]$$

$$= M \sum_i (I(X_i : Pa_i) - H(X_i))$$

where

$$H(X) = - \int P(X) \log P(X) dX$$

$$I(G : D) = H(X) + H(Y) - H(X, Y) \quad \text{mutual information}$$

## I. Likelihood: Pluses and Minuses

$$I(G : D) = M \sum_i (I(X_i : Pa_i) - H(X_i))$$

Pluses:  $I(G : D) = M \sum_i (I(X_i : Pa_i) - H(X_i))$

The larger the dependency of each variable on its parents, the higher the score.

Minuses:

Adding arcs always improves the score.

$$I(X : Y) \leq I(X : Y, Z)$$

Maximally connected network overfits data.

## I. Preventing Overfitting

**Restrict hypotheses**

Restricted # of parents or # of parameters.

**Minimum description length (MDL)**

Description length measures complexity of model or data.

Select model that minimizes  $DL(\text{model}) + DL(\text{data}|\text{model})$

Recall class on information theory...

**Bayesian score**

Average over parameter values... Low likelihood if there is little data per fitted parameter.

## II: Minimum Description Length

[Reassanen, 1987]

Prefer graphs that provide maximum compression of the data.

Compression means that the network summarizes the data.

Pick network that minimizes  
 $DL(\text{network}) + DL(\text{data}|\text{network})$

## II. MDL

Description length of the data + structure:

$$DL(D : G) = DL(G) + \frac{\log M}{2} \dim(G) - I(G : D)$$

bits to describe graph

accuracy required  
to represent each  
parameter (M different  
possible values).

# of parameters

bits required to  
represent data given  
graph.

MDL is defined to be:

$$MDL = -DL(D : G)$$

## II. MDL

MDL

$$MDL(D : G) = I(G : D) - \frac{\log M}{2} \dim(G) - DL(G)$$

Likelihood is linear in M

$$\begin{aligned} I(G : D) &= \sum_M \log P\{X[m] | G\} \\ &\approx M \cdot E[\log P\{X[m] | G\}] \end{aligned}$$

Generally use only those terms that increase in M

$$MDL(D : G) = I(G : D) - \frac{\dim(G)}{2} \log M$$

Note: As M increases, the relative weight of the penalty decreases

## II. MDL Properties

As  $M \rightarrow \infty$ , the “true” structure  $G_{\max}$  has the maximum score.

For sufficiently large  $M$ , the maximal scoring structures are **independence equivalent** to  $G_{\max}$

Recall: Two DAGs,  $G$  and  $G'$ , represent equivalent sets of CISs whenever  $G$  and  $G'$  have the same undirected version AND the identical V-structures.

### III. Bayesian inference

$$P\{x[M+1] | D\} = \sum_G P\{x[M+1] | D, G\} P\{G | D\}$$

where

$$\begin{aligned} P\{G | D\} &\propto P\{D | G\} P\{G\} \\ &= \left( \int_{\Theta} P\{D | G, \Theta\} P\{\Theta | G\} d\Theta \right) P\{G\} \end{aligned}$$

### III. Marginal Likelihood for Binomials

Observe M coin tosses

Chain rule:

$$\begin{aligned} &P\{X[1], \dots, X[M]\} \\ &= P\{X[1]\} P\{X[2] | X[1]\} \dots P\{X[M] | X[1], \dots, X[M-1]\} \end{aligned}$$

We showed that:

$$P\{X[m+1] | X[1], \dots, X[m]\} = \frac{N_H^m + \alpha_H}{m + \alpha_H + \alpha_T}$$

SO...

### III. Marginal Likelihood for Binomials (cont'd)

$$P\{X[1], \dots, X[M]\} = \frac{\alpha_H}{\alpha_H + \alpha_T} \dots \frac{N_H - 1 + \alpha_H}{N_H - 1 + \alpha_H + \alpha_T} \cdot \frac{\alpha_T}{N_H + \alpha_H + \alpha_T} \dots \frac{N_T + \alpha_T}{N_H + N_T - 1 + \alpha_H + \alpha_T}$$

$$P\{D\} = \frac{\Gamma(\alpha_H + \alpha_T)}{\Gamma(N_H + N_T + \alpha_H + \alpha_T)} \frac{\Gamma(N_H + \alpha_H)}{\Gamma(\alpha_H)} \frac{\Gamma(N_T + \alpha_T)}{\Gamma(\alpha_T)}$$

### III. Marginal Likelihood for Multinomials

$$P\{D\} = \frac{\Gamma\left(\sum_k \alpha_k\right)}{\Gamma\left(\sum_k (\alpha_k + N_k)\right)} \prod_k \frac{\Gamma(\alpha_k + N_k)}{\Gamma(\alpha_k)}$$

In general networks:

$$P\{D | G\} = \prod_{i=1}^N \left( \prod_{pa_i^G} \frac{\Gamma(\alpha(pa_i^G))}{\Gamma(\alpha(pa_i^G) + N(pa_i^G))} \prod_{x_i} \frac{\Gamma(\alpha(x_i, pa_i^G) + N(x_i, pa_i^G))}{\Gamma(\alpha(x_i, pa_i^G))} \right)$$

$\alpha(\dots)$ : parameters for each family given G

$N(\dots)$ : counts from data

### III. BDe Score

Problem:

Need prior counts  $\alpha(\dots)$  for each G

BDe prior:

Use prior of the form  $M_0, B_0 = (G_0, \Theta_0)$

$M_0$  examples distributed according to  $B_0$

Use  $\alpha(x_i, pa_i^G) = M_0 P\{x_i, pa_i^G | G_0, \Theta_0\}$

In general,  $pa_i^G \neq pa_i^{G_0}$

Nice property:

Equivalent networks are assigned the same scores.

### III. BDe Score Properties

Seems different from MDL, BUT

Scores are asymptotically equivalent

Can show:

$$\log P\{D | G\} = I(G : D) - \frac{\dim(G)}{2} \log M + O(1)$$

Bayes score is asymptotically equivalent to MDL score

Constants (P(G) and DL(G)) are negligible when M is large

## Scoring metrics: Summary

Likelihood, MDL and (log) BDe all have the form

$$Score(G : D) = \sum_i Score(X_i | Pa_i^G : N(X_i, Pa_i^G))$$

MDL and BDe are asymptotically equivalent

All three scoring metrics assign the same score to equivalent networks.