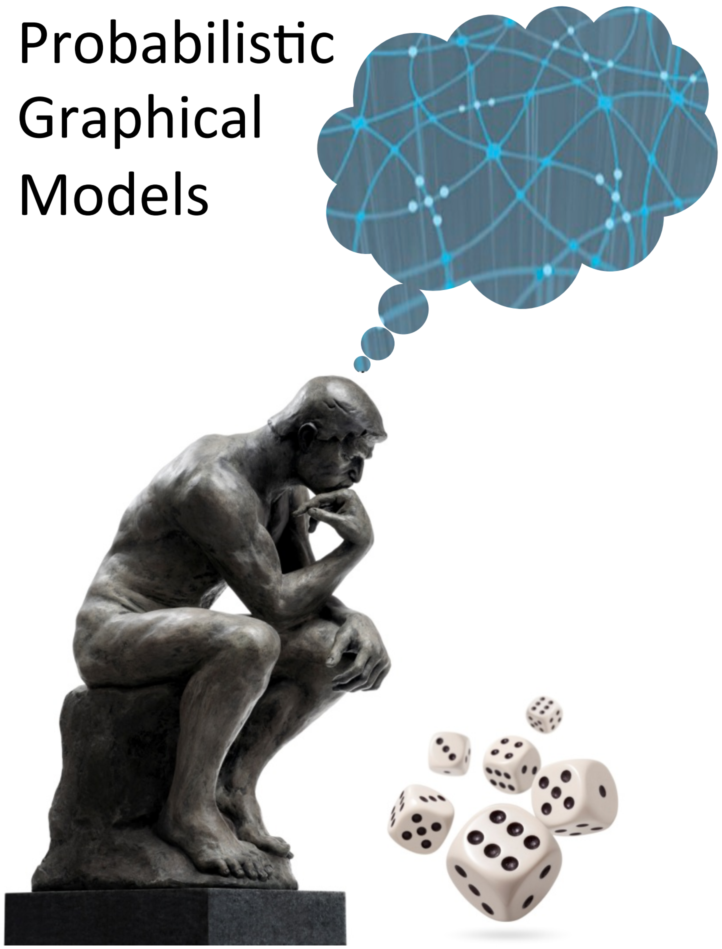


Learning

Incomplete Data

Overview

Incomplete Data

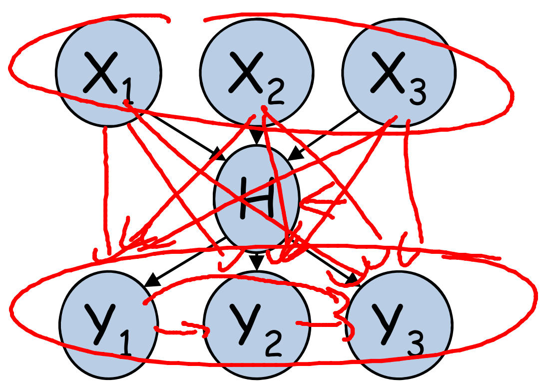
- Multiple settings:
  - Hidden variables
  - Missing values

- Challenges
  - Foundational – is the learning task well defined?
  - Computational – how can we learn with incomplete data?

Why latent variables?

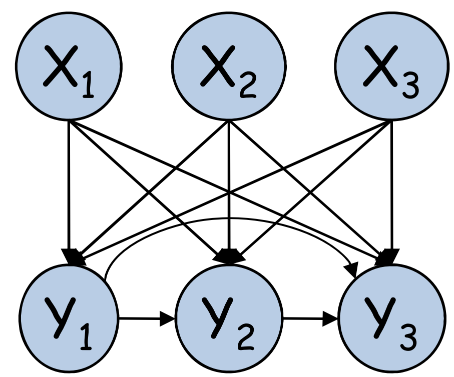
- Model sparsity

true G^*



17 parameters

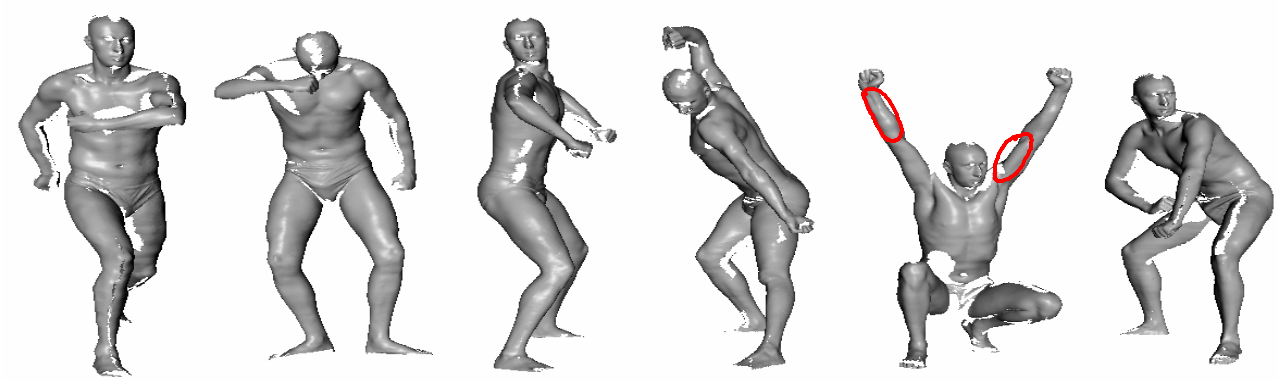
$p(x_1, x_2, x_3, y_1, y_2, y_3)$

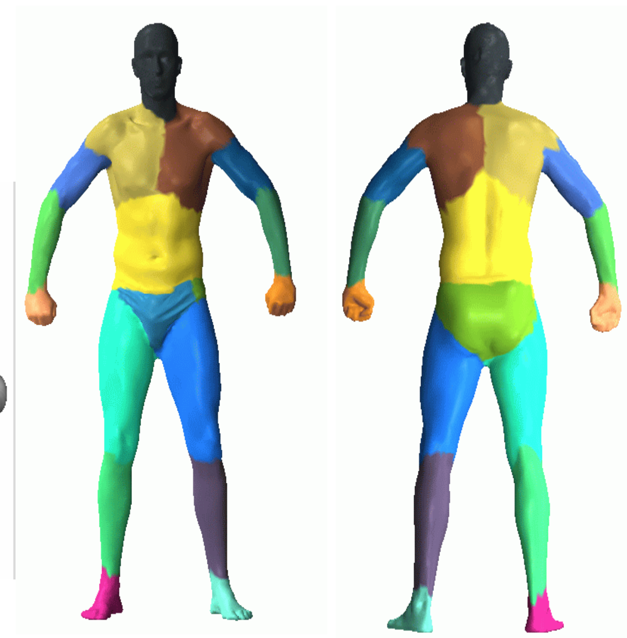


59 parameters

Why latent variables?

- Discovering clusters in data





Daphne Koller

Treating Missing Data

Sample sequence: H,T,?,?,H,?,H

- Case I: A coin is tossed on a table, occasionally it drops and measurements are not taken

H T H H

- Case II: A coin is tossed, but sometimes tails are not reported

H T T T H T H

We need to consider the missing data mechanism

Modeling Missing Data Mechanism

- **X** = $\{X_1,\ldots,X_n\}$ are random variables
- **O** = $\{O_1,\ldots,O_n\}$ are *observability variables*
  - Always observed

$$O_i = \begin{cases} 1 & X_i \text{ observed} \\ 0 & \text{otherwise} \end{cases}$$

- Y = $\{Y_1,\ldots,Y_n\}$ new random variables
  - $Val(Y_i) = Val(X_i) \cup \{?\}$
  - Always observed
  - Y_i is a deterministic function of X_i and O_i:

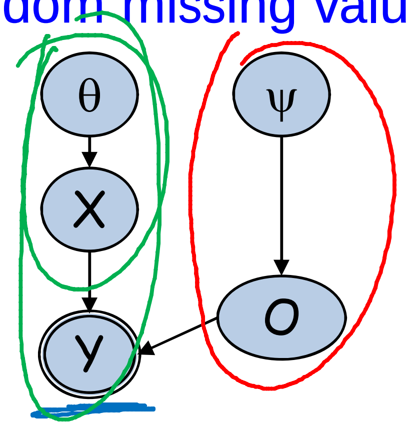
$$Y_i = \begin{cases} X_i & O_i = o^1 \\ ? & O_i = o^0 \end{cases}$$

Daphne Koller

Modeling Missing Data Mechanism

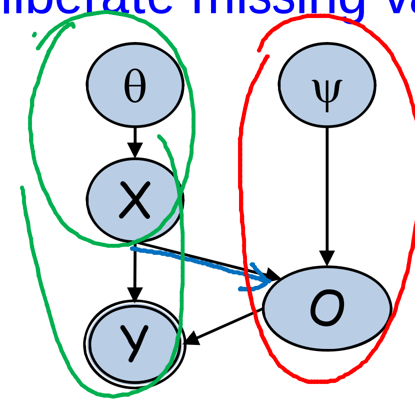
Case I

(random missing values)



Case II

(deliberate missing values)



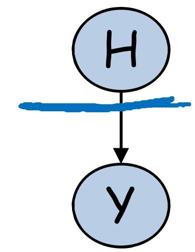
- When can we ignore the missing data mechanism and focus only on the likelihood?
- Missing at Random (MAR)

unobserved X's

$$P_{missing} \models (\boldsymbol{O} \perp \boldsymbol{H} \mid \boldsymbol{d})$$

observed values X

Identifiability



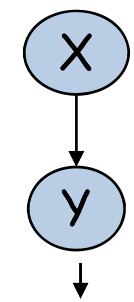
- Likelihood can have multiple global maxima
- Example:
  - We can rename the values of the hidden variable H
  - If H has two values, likelihood has two global maxima
- With many hidden variables, there can be an exponential number of global maxima
- Multiple local and global maxima can also occur with missing data (not only hidden variables)

Likelihood for Complete Data

Input Data:

X	Y
x^0	y^0
x^0	y^1
x^1	y^0

x^0	x^1
θ_{x0}	θ_{x1}



X	***P(Y/X)*** y^0	y^1
x^0	$\theta_{y0\|x0}$	$\theta_{y1\|x0}$
x^1	$\theta_{y0\|x1}$	$\theta_{y1\|x1}$

- Likelihood decomposes by variables
- Likelihood decomposes within CPDs

Likelihood:

$$
\begin{aligned}
L(D:\theta) &= P(x[1], y[1]) \cdot P(x[2], y[2]) \cdot P(x[3], y[3]) \\
&= P(x^0, y^0) \cdot P(x^0, y^1) \cdot P(x^1, y^0) \\
&= \theta_{x^0} \cdot \theta_{y^0|x^0} \cdot \theta_{x^0} \cdot \theta_{y^1|x^0} \cdot \theta_{x^1} \cdot \theta_{y^0|x^1} \\
&= \left(\theta_{x^0} \cdot \theta_{x^0} \cdot \theta_{x^1}\right) \cdot \left(\theta_{y^0|x^0} \cdot \theta_{y^1|x^0}\right) \cdot \left(\theta_{y^0|x^1}\right)
\end{aligned}
$$

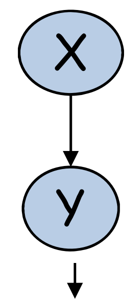
Daphne Koller

Likelihood for Incomplete Data

Input Data:

X	Y
?	y^0
x^0	y^1
?	y^0

x^0	x^1
θ_{x0}	θ_{x1}



X	P(Y/X)	
	y^0	y^1
x^0	$\theta_{y0\|x0}$	$\theta_{y1\|x0}$
x^1	$\theta_{y0\|x1}$	$\theta_{y1\|x1}$

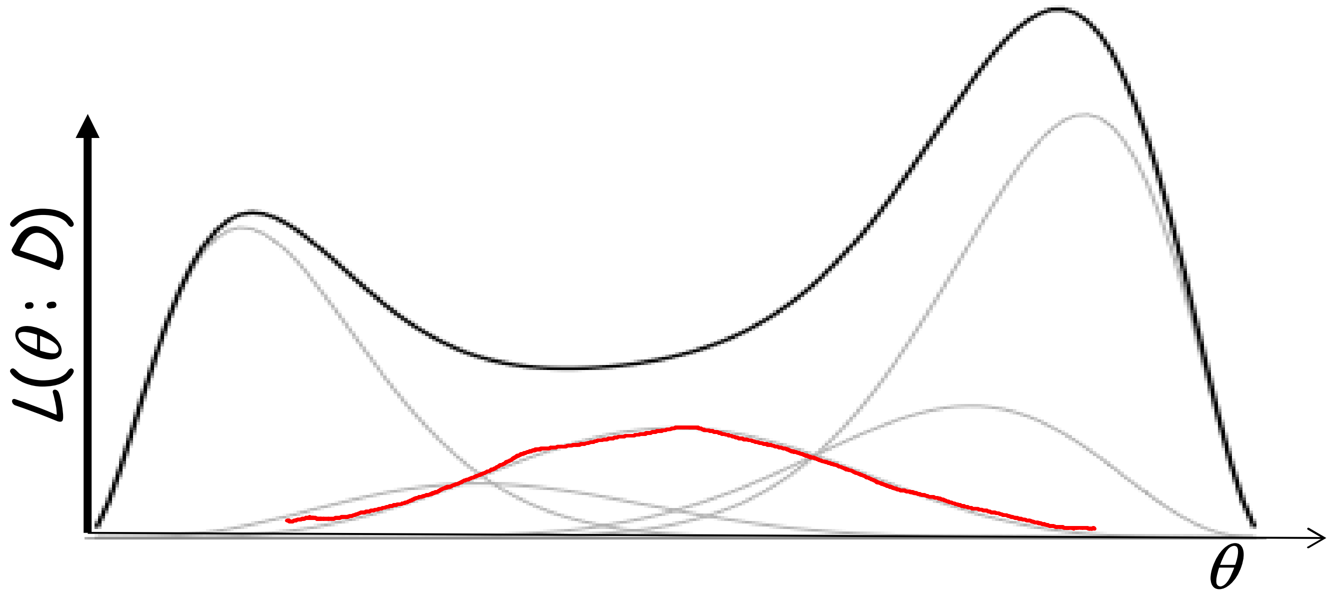
- Likelihood does not decompose by variables
- Likelihood does not decompose within CPDs
- Computing likelihood requires inference!

Likelihood:

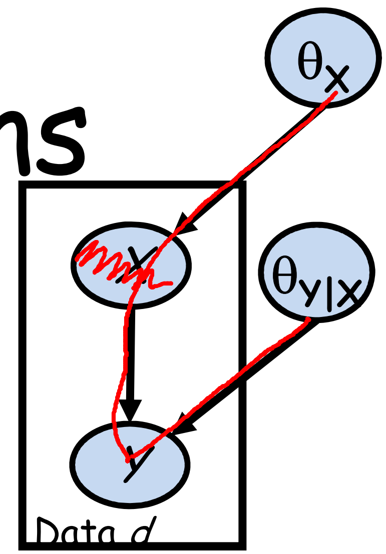
$$
\begin{aligned}
L(D:\theta) &= P(y^0)\cdot P(x^0, y^1)\cdot P(y^0) \\
&= \left(\sum_{x\in Val(X)} P(x, y^0)\right)^2 \cdot P(x^0, y^1)\cdot \\
&= \left(\theta_{x^0}\cdot\theta_{y^0|x^0} + \theta_{x^1}\cdot\theta_{y^0|x^1}\right)^2\cdot\theta_{x^0}\cdot\theta_{y^1|x^0} \\
&= \left(\theta_{x^0}\cdot\theta_{y^0|x^0} + \theta_{x^1}\cdot\theta_{y^0|x^1}\right)^2\cdot\theta_{x^0}\cdot\theta_{y^1|x^0}
\end{aligned}
$$

Daphne Koller

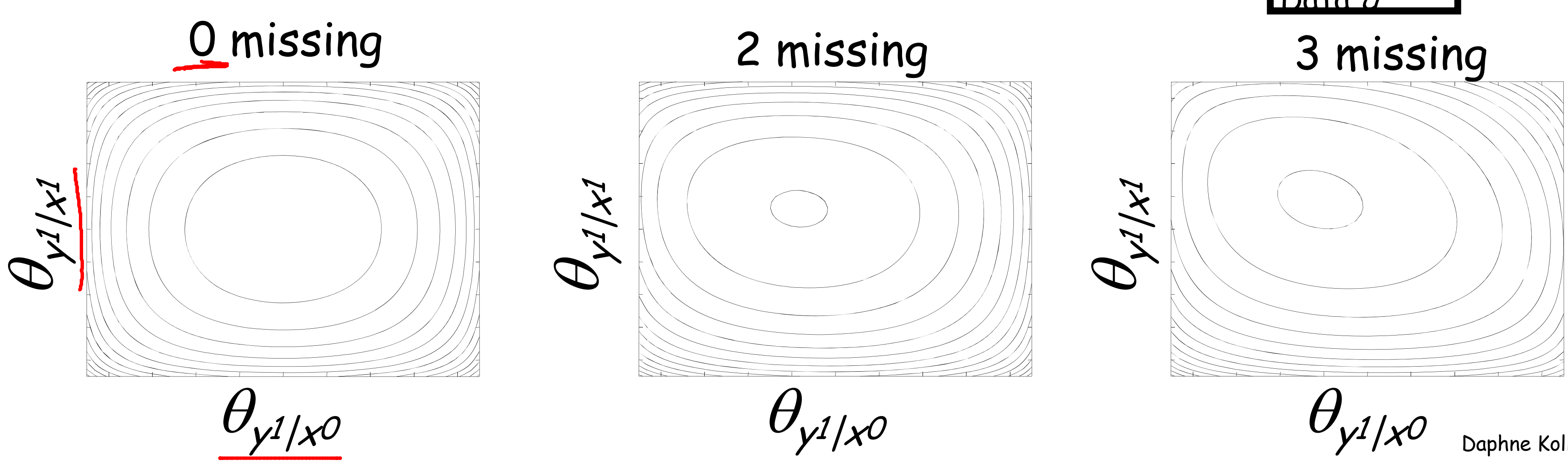
Multimodal Likelihood



Parameter Correlations



- Total of 8 data points
- Some X's unobserved



Daphne Koller

Summary

- Incomplete data arises often in practice
- Raises multiple challenges & issues:
  - The mechanism for missingness
  - Identifiability
  - Complexity of likelihood function