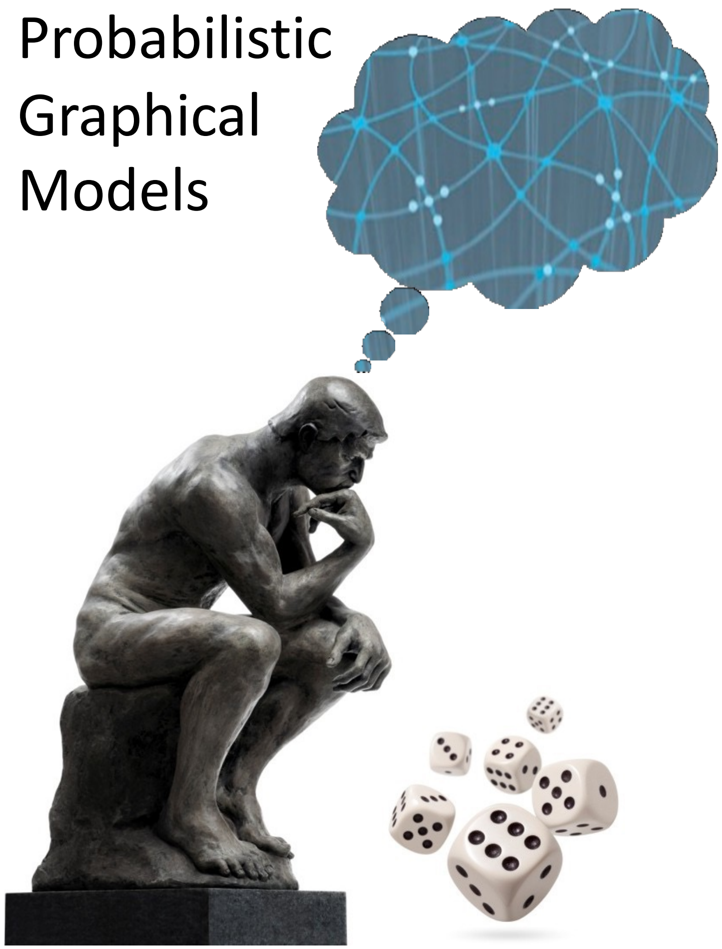
Probabilistic
Graphical
Models



Learning

Incomplete Data

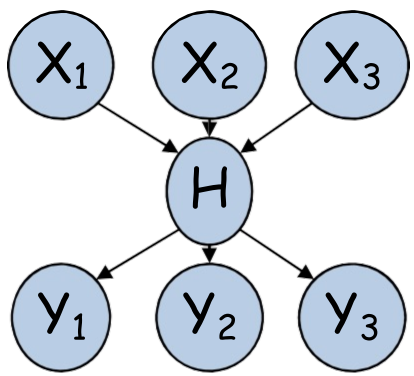
# Overview

# Incomplete Data

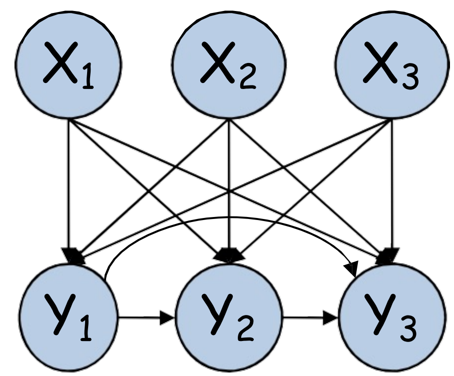
- Multiple settings:
  - Hidden variables
  - Missing values
- Challenges
  - Foundational – is the learning task well defined?
  - Computational – how can we learn with incomplete data?

# Why latent variables?

- Model sparsity



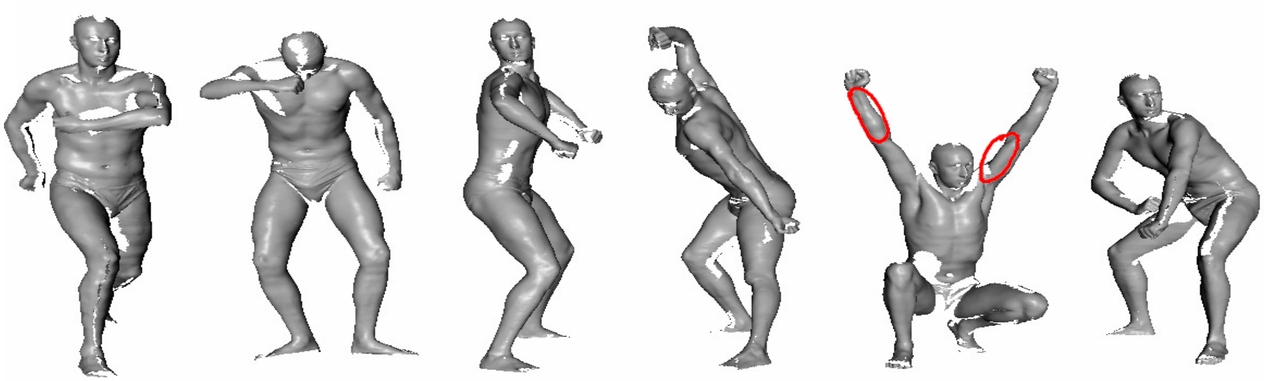

17 parameters

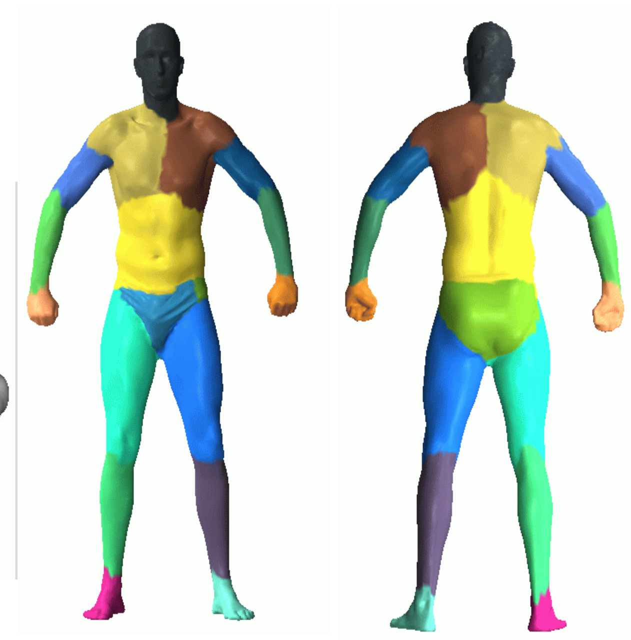



59 parameters

# 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

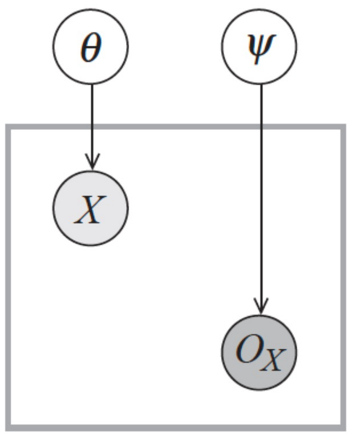
- **X** = $\{X_1,\ldots,X_n\}$ are random variables
  - Sometimes missing
- **O** = $\{O_1,\ldots,O_n\}$ are *observability variables*
  - Always observed
- Y = $\{Y_1,\ldots,Y_n\}$ new random variables
  - $Val(Y_i) = Val(Y_i) \cup \{?\}$
  - Always observed
  - $Y_i$ is a deterministic function of $X_i$ and $O_i$:

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

Daphne Koller

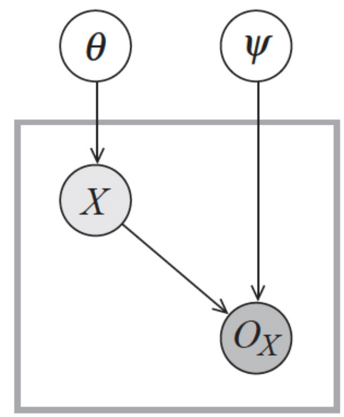
# Modeling Missing Data Mechanism

## Case I
(random missing values)




(a) Random missing values

## Case II
(deliberate missing values)

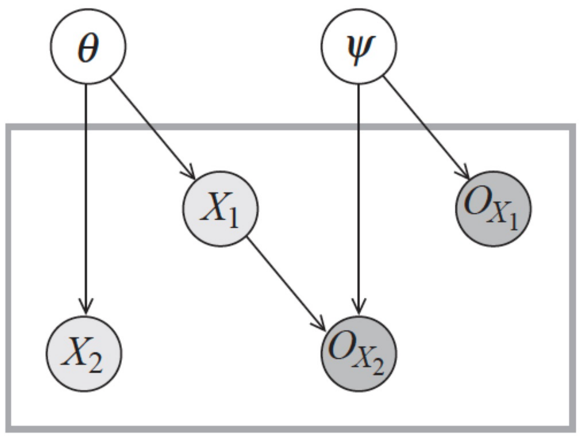



(b) Deliberate missing values

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

A missing data model $P_{\text{missing}}$ is missing completely at random (MCAR) if $P_{\text{missing}} \models (\boldsymbol{X} \perp O_{\boldsymbol{X}})$

# Modeling Missing Data Mechanism



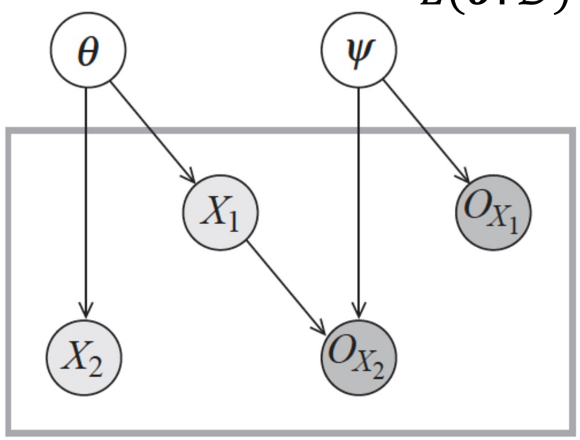

- When can we ignore the missing data mechanism and focus only on the likelihood?
- Missing at Random (MAR)

We say that a missing data model $P_{\text{missing}}$ is missing at random (MAR) if for all observations $\boldsymbol{y}$ with $P_{\text{missing}}(\boldsymbol{y}) > 0$, and for all $\boldsymbol{x}^{\boldsymbol{y}}_{\text{hidden}} \in \mathrm{Val}\left(\boldsymbol{X}^{\boldsymbol{y}}_{\text{hidden}}\right)$, we have that

$$P_{\text{missing}} \models \left(o_{\boldsymbol{X}} \perp \boldsymbol{x}^{\boldsymbol{y}}_{\text{hidden}} \mid \boldsymbol{x}^{\boldsymbol{y}}_{obs}\right)$$

where $o_{\boldsymbol{X}}$ are the specific values of the observation variables given $\boldsymbol{Y}$.

# Modeling Missing Data Mechanism




$$
\begin{aligned}
L(\boldsymbol{\theta} : \mathcal{D}) = \quad & \theta_{X_1}^{M[y_1^1]}\left(1-\theta_{X_1}\right)^{M[y_1^0]} \\
& \theta_{X_2}^{M[y_2^1]}\left(1-\theta_{X_2}\right)^{M[y_2^0]} \\
& \psi_{o_{X_2}|y_1^1}^{M[y_1^1,y_1^1]+M[y_1^1,y_2^0]}\left(1-\psi_{o_{X_2}|x_1^1}\right)^{M[y_1^1,y_2^2]} \\
& \psi_{o_{X_2}|x_1^0}^{M[y_1^0,y_2^1]+M[y_1^0,y_2^0]}\left(1-\psi_{o_{X_2}|x_1^0}\right)^{M[y_1^0,y_2^2]}
\end{aligned}
$$

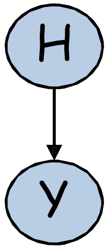
- When can we ignore the missing data mechanism and focus only on the likelihood?
- Missing at Random (MAR)

We say that a missing data model $P_{\text{missing}}$ is missing at random (MAR) if for all observations $\boldsymbol{y}$ with $P_{\text{missing}}(\boldsymbol{y}) > 0$, and for all $\boldsymbol{x}_{\text{hidden}}^{\boldsymbol{y}} \in \text{Val}\left(\boldsymbol{X}_{\text{hidden}}^{\boldsymbol{y}}\right)$, we have that

$$P_{\text{missing}} \models \left(o_{\boldsymbol{X}} \perp \boldsymbol{x}_{\text{hidden}}^{\boldsymbol{y}} \mid \boldsymbol{x}_{obs}^{\boldsymbol{y}}\right)$$

where $o_{\boldsymbol{X}}$ are the specific values of the observation variables given $\boldsymbol{Y}$.

# Identifiability



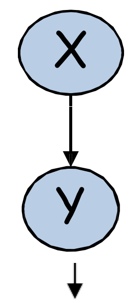

- Likelihood can have multiple global maxima
- Example:
  – We can rename the values of the hidden variable H
  – If H has two values, likelihood has two global maxima
- With many hidden variables, there can be an exponential number of global maxima
- Multiple local and global maxima can also occur with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

# Likelihood for Complete Data

**Input Data:**

| X | Y |
|---|---|
| $x^0$ | $y^0$ |
| $x^0$ | $y^1$ |
| $x^1$ | $y^0$ |

| P(X) | |
|---|---|
| $x^0$ | $x^1$ |
| $\theta_{x0}$ | $\theta_{x1}$ |




- Likelihood decomposes by variables
- Likelihood decomposes within CPDs

**Likelihood:**

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

$$L(D:\theta) = P(x[1],y[1])\,P(x[2],y[2])\,P(x[3],y[3])$$

$$= P(x^0,y^0)P(x^0,y^1)P(x^1,y^0)$$

$$= \theta_{x^0}\theta_{y^0|x^0}\theta_{x^0}\theta_{y^1|x^0}\theta_{x^1}\theta_{y^0|x^1} =$$

$$\theta_{x^0}^{M[x=0]}(1-\theta_{x^0})^{M[x=1]}\times$$

$$\theta_{y^0|x^0}^{M[y=0,x=0]}\left(1-\theta_{y^0|x^0}\right)^{M[y=1,x=0]}\times$$

$$\theta_{y^0|x^1}^{M[y=0,x=1]}\left(1-\theta_{y^0|x^1}\right)^{M[y=1,x=1]}$$

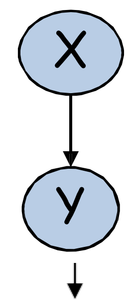
Daphne Koller

# Likelihood for Incomplete Data

Input Data:

| X | Y |
|---|---|
| ? | $y^0$ |
| $x^0$ | $y^1$ |
| ? | $y^0$ |

| $x^0$ | $x^1$ |
|---|---|
| $\theta_{x0}$ | $\theta_{x1}$ |




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

Likelihood:

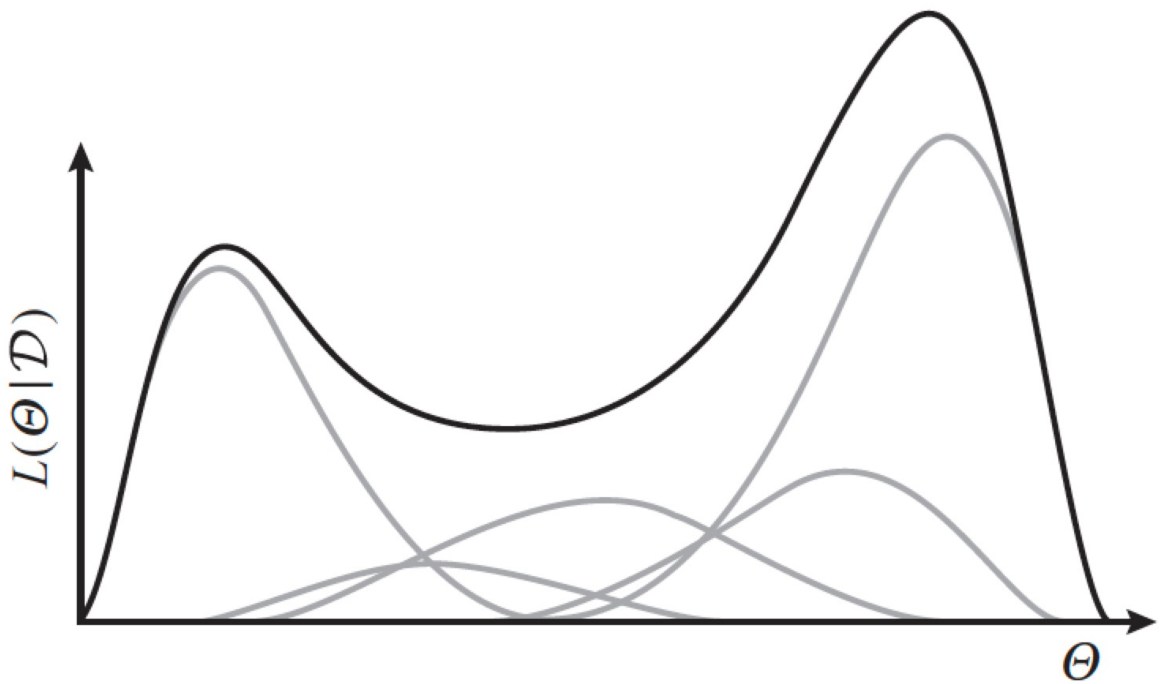
$$L(D:\theta) = P(y^0)P(x^0, y^1)P(y^0)$$

$$= \theta_{x^0}\theta_{y^1|x^0}\sum_{x}\theta_{x^x}\theta_{y^0|x^x}$$

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

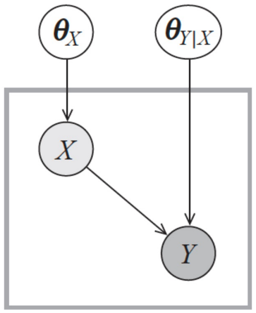
Daphne Koller

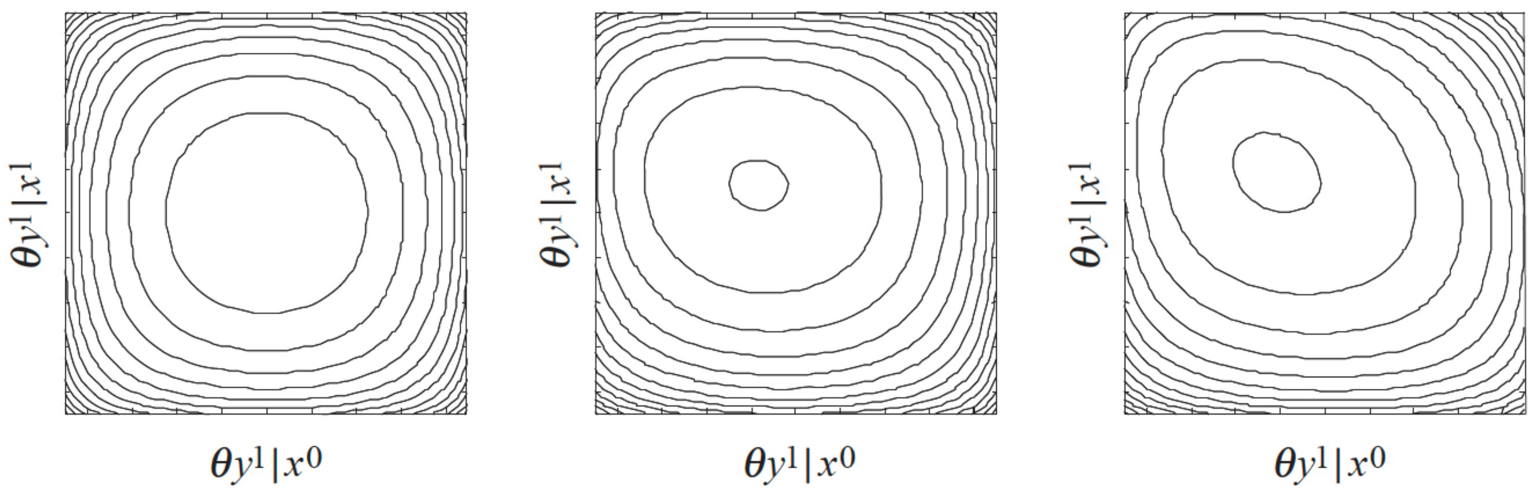
# Multimodal Likelihood



# Parameter Correlations

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

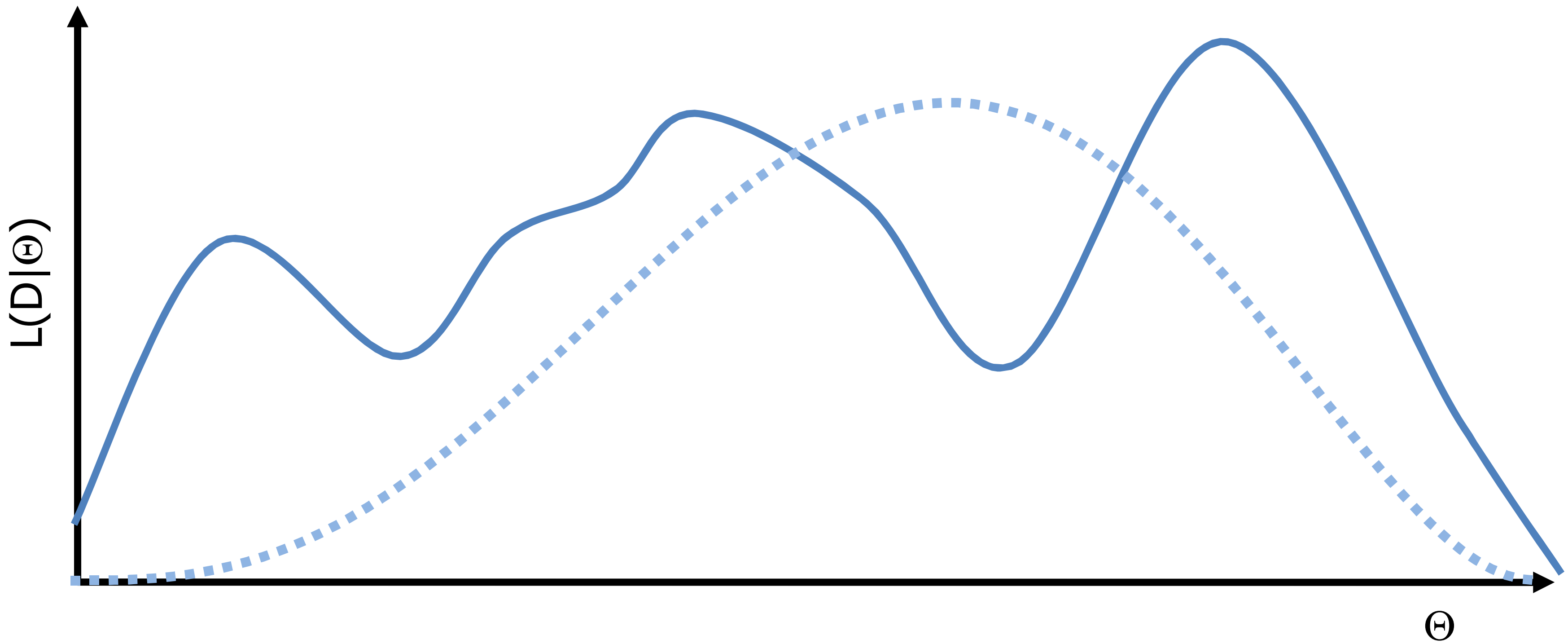


# Summary

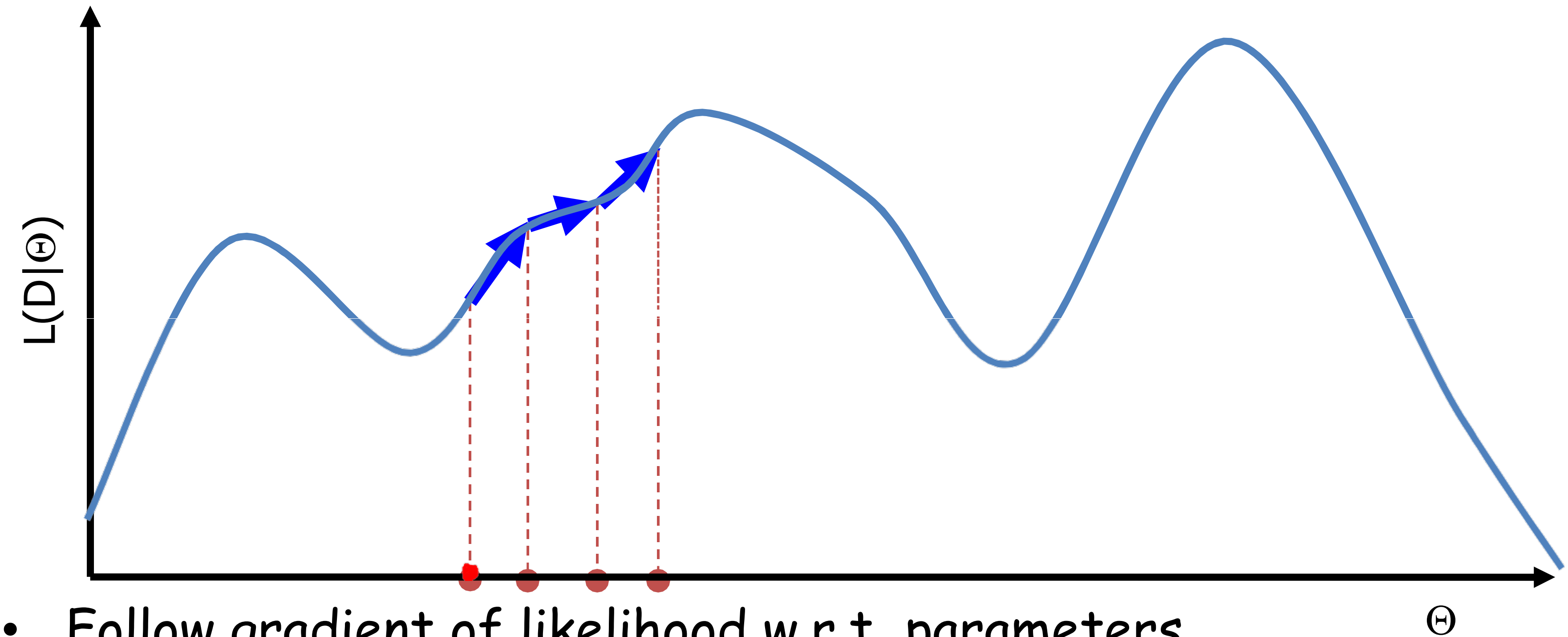
- Incomplete data arises often in practice
- Raises multiple challenges & issues:
  - The mechanism for missingness
  - Identifiability
  - Complexity of likelihood function

# Likelihood with Incomplete Data




Daphne Koller

# Gradient Ascent




- Follow gradient of likelihood w.r.t. parameters
- Line search & conjugate gradient methods for fast convergence

Daphne Koller

# Gradient Ascent

- Theorem:

Let $\mathcal{G}$ be a Bayesian network structure over $\mathcal{X}$, and let $\mathcal{D} = \{\boldsymbol{o}[1], \ldots, \boldsymbol{o}[M]\}$ be a partially observable data set. Let $X$ be a variable and $\boldsymbol{U}$ its parents in $\mathcal{G}$. Then

$$\frac{\partial \ell(\boldsymbol{\theta} : \mathcal{D})}{\partial P(x \mid \boldsymbol{u})} = \frac{1}{P(x \mid \boldsymbol{u})} \sum_{m=1}^{M} P(x, \boldsymbol{u} \mid \boldsymbol{o}[m], \boldsymbol{\theta}).$$

- Requires computing $P(X_i, \boldsymbol{U}_i | d[m], \boldsymbol{\theta})$ for all $i, m$
- Can be done with clique-tree algorithm, since $X_i, \boldsymbol{U}_i$ are in the same clique

# Gradient Ascent Summary

- Need to run inference over each data instance at every iteration
- Pros ensure that parameters define legal
  - Flexible, can be extended to non table CPDs
- Cons
  - Constrained optimization: need tCPDs
  - For reasonable convergence, need to combine with advanced methods (conjugate gradient, line search)

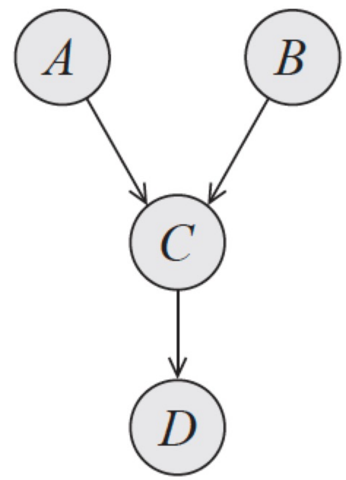
# Expectation Maximization (EM)

- Special-purpose algorithm designed for optimizing likelihood functions
- Intuition
  - Parameter estimation is easy given complete data
  - Computing probability of missing data is "easy" (=inference) given parameters

# Example

MLE estimate for $\theta_{d^1|c^0}$ if all data were fully observed:

$$\hat{\theta}_{d^1|c^0} = \frac{M[d^1, c^0]}{M[c^0]} = \frac{\sum_{m=1}^{M} \mathbb{I}\{\xi[m]\langle D, C\rangle = \langle d^1, c^0\rangle\}}{\sum_{m=1}^{M} \mathbb{I}\{\xi[m]\langle C\rangle = c^0\}}$$




Now assume we have a sample $\boldsymbol{o} = \langle a^1, ?\,, ?\,, d^0\rangle$

Four possible assignments of $b, c$

-If we knew the true assignment we could compute the MLE parameters

-If we knew the parameters we could compute the probability of each assingment

Parameters:

$\boldsymbol{\theta}_{a^1}$ $\boldsymbol{\theta}_{b^1}$

$\boldsymbol{\theta}_{d^1|c^0}$ $\boldsymbol{\theta}_{d^1|c^1}$

$\boldsymbol{\theta}_{c^1|a^0,b^0}$ $\boldsymbol{\theta}_{c^1|a^1,b^0}$

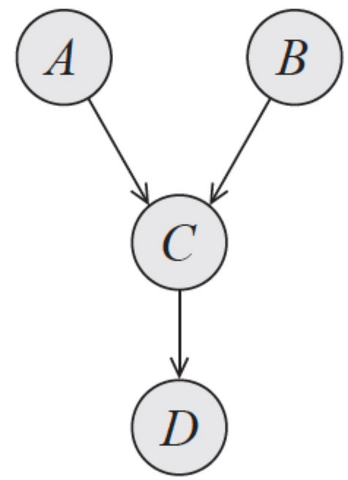
$\boldsymbol{\theta}_{c^1|a^0,b^1}$ $\boldsymbol{\theta}_{c^1|a^1,b^1}$

# Example

Assume we are given estimates for the parameters

$$\boldsymbol{o} = \langle a^1, ?\,, ?\,, d^0 \rangle$$

$$Q(B, C) = P(B, C \mid a^1, d^0, \boldsymbol{\theta})$$




Parameters:

$$\boldsymbol{\theta}_{a^1} = 0.3 \quad \boldsymbol{\theta}_{b^1} = 0.9$$
$$\boldsymbol{\theta}_{d^1|c^0} = 0.1 \; \boldsymbol{\theta}_{d^1|c^1} = 0.8$$
$$\boldsymbol{\theta}_{c^1|a^0,b^0} = 0.83 \; \boldsymbol{\theta}_{c^1|a^1,b^0} = 0.6$$
$$\boldsymbol{\theta}_{c^1|a^0,b^1} = 0.09 \; \boldsymbol{\theta}_{c^1|a^1,b^1} = 0.2$$

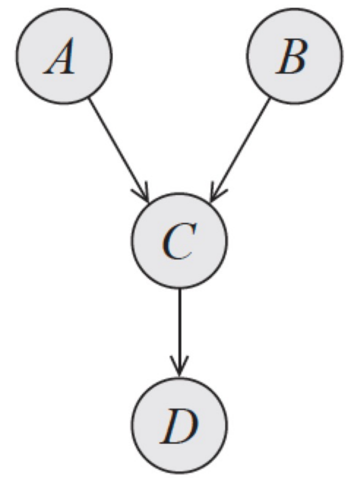
# Example

Assume we are given estimates for the parameters

$$\boldsymbol{o} = \langle a^1, ?\,, ?\,, d^0 \rangle$$

$$Q(B, C) = P(B, C \mid a^1, d^0, \boldsymbol{\theta})$$

$$Q(\langle b^1, c^1 \rangle) = 0.3 \cdot 0.9 \cdot 0.2 \cdot 0.2/0.2196 = 0.0492$$
$$Q(\langle b^1, c^0 \rangle) = 0.3 \cdot 0.9 \cdot 0.8 \cdot 0.9/0.2196 = 0.8852$$
$$Q(\langle b^0, c^1 \rangle) = 0.3 \cdot 0.1 \cdot 0.6 \cdot 0.2/0.2196 = 0.0164$$
$$Q(\langle b^0, c^0 \rangle) = 0.3 \cdot 0.1 \cdot 0.4 \cdot 0.9/0.2196 = 0.0492$$




Parameters:

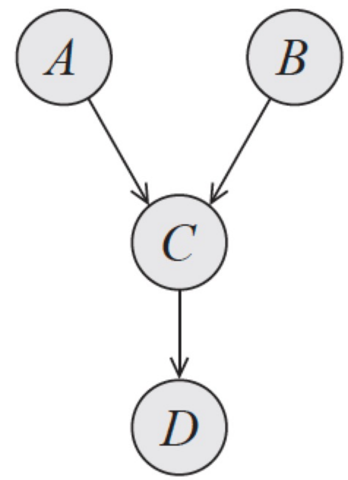
$$\boldsymbol{\theta}_{a^1} = 0.3 \quad \boldsymbol{\theta}_{b^1} = 0.9$$
$$\boldsymbol{\theta}_{d^1|c^0} = 0.1 \quad \boldsymbol{\theta}_{d^1|c^1} = 0.8$$
$$\boldsymbol{\theta}_{c^1|a^0,b^0} = 0.83 \quad \boldsymbol{\theta}_{c^1|a^1,b^0} = 0.6$$
$$\boldsymbol{\theta}_{c^1|a^0,b^1} = 0.09 \quad \boldsymbol{\theta}_{c^1|a^1,b^1} = 0.2$$

# Another Example

Assume we are given estimates for the parameters

$$o' = \langle ?\,, b^1, ?\,, d^1 \rangle$$

$$Q'(A, C) = P(A, D \mid b^1, d^1, \boldsymbol{\theta})$$




Parameters:

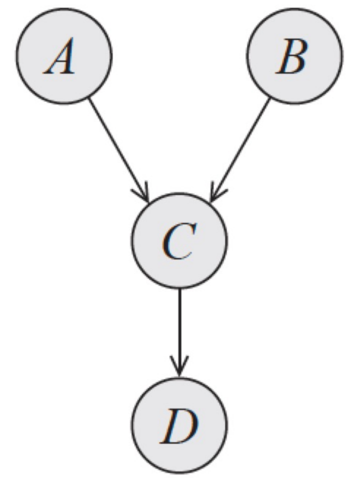
$$\boldsymbol{\theta}_{a^1} = 0.3 \quad \boldsymbol{\theta}_{b^1} = 0.9$$
$$\boldsymbol{\theta}_{d^1|c^0} = 0.1 \quad \boldsymbol{\theta}_{d^1|c^1} = 0.8$$
$$\boldsymbol{\theta}_{c^1|a^0,b^0} = 0.83 \quad \boldsymbol{\theta}_{c^1|a^1,b^0} = 0.6$$
$$\boldsymbol{\theta}_{c^1|a^0,b^1} = 0.09 \quad \boldsymbol{\theta}_{c^1|a^1,b^1} = 0.2$$

# Another Example

Assume we are given estimates for the parameters

$$o' = \langle ?\,, b^1, ?\,, d^1 \rangle$$

$$Q'(A, C) = P(A, D \mid b^1, d^1, \boldsymbol{\theta})$$




$$Q'(\langle a^1, c^1 \rangle) = 0.3 \cdot 0.9 \cdot 0.2 \cdot 0.8/0.1675 = 0.2579$$
$$Q'(\langle a^1, c^0 \rangle) = 0.3 \cdot 0.9 \cdot 0.8 \cdot 0.1/0.1675 = 0.1290$$
$$Q'(\langle a^0, c^1 \rangle) = 0.7 \cdot 0.9 \cdot 0.09 \cdot 0.8/0.1675 = 0.2708$$
$$Q'(\langle a^0, c^0 \rangle) = 0.7 \cdot 0.9 \cdot 0.91 \cdot 0.1/0.1675 = 0.3423$$

Parameters:

$$\boldsymbol{\theta}_{a^1} = 0.3 \quad \boldsymbol{\theta}_{b^1} = 0.9$$
$$\boldsymbol{\theta}_{d^1|c^0} = 0.1 \ \boldsymbol{\theta}_{d^1|c^1} = 0.8$$
$$\boldsymbol{\theta}_{c^1|a^0,b^0} = 0.83 \ \boldsymbol{\theta}_{c^1|a^1,b^0} = 0.6$$
$$\boldsymbol{\theta}_{c^1|a^0,b^1} = 0.09 \ \boldsymbol{\theta}_{c^1|a^1,b^1} = 0.2$$

This is like having four data instances with weights.

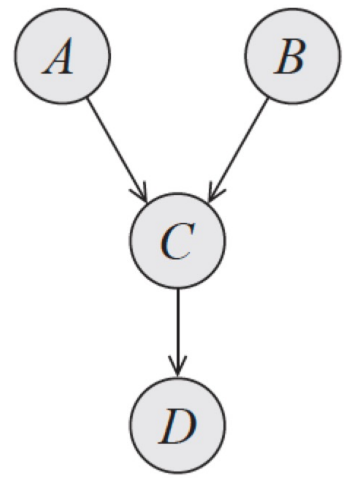
# Another Example

$$o = \langle a^1, ?\,, ?\,, d^0 \rangle$$
$$o' = \langle ?\,, b^1, ?\,, d^1 \rangle$$

This is like having the data

| | |
|---|---|
| $\langle a^1, b^1, \mathrm{c}^1, d^0 \rangle$ | 0.0492 |
| $\langle a^1, b^1, \mathrm{c}^0, d^0 \rangle$ | 0.8852 |
| $\langle a^1, b^0, \mathrm{c}^1, d^0 \rangle$ | 0.0164 |
| $\langle a^1, b^0, \mathrm{c}^0, d^0 \rangle$ | 0.0492 |

| | |
|---|---|
| $\langle a^1, b^1, \mathrm{c}^1, d^1 \rangle$ | 0.2579 |
| $\langle a^1, b^1, \mathrm{c}^0, d^1 \rangle$ | 0.1290 |
| $\langle a^0, b^1, \mathrm{c}^1, d^1 \rangle$ | 0.2708 |
| $\langle a^0, b^1, \mathrm{c}^0, d^1 \rangle$ | 0.3423 |




Parameters:

$$\boldsymbol{\theta}_{a^1} = 0.3 \quad \boldsymbol{\theta}_{b^1} = 0.9$$
$$\boldsymbol{\theta}_{d^1|c^0} = 0.1 \; \boldsymbol{\theta}_{d^1|c^1} = 0.8$$
$$\boldsymbol{\theta}_{c^1|a^0,b^0} = 0.83 \; \boldsymbol{\theta}_{c^1|a^1,b^0} = 0.6$$
$$\boldsymbol{\theta}_{c^1|a^0,b^1} = 0.09 \; \boldsymbol{\theta}_{c^1|a^1,b^1} = 0.2$$

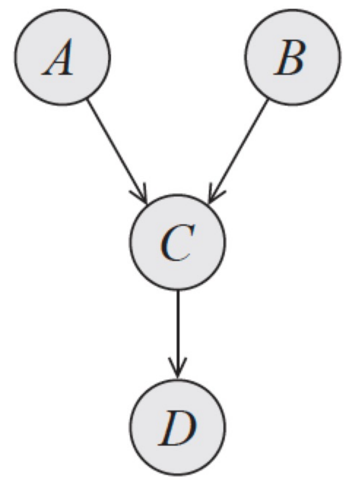
# Another Example

$$o = \langle a^1, ?, ?, d^0 \rangle$$
$$o' = \langle ?, b^1, ?, d^1 \rangle$$

This is like having the data

| | |
|---|---|
| $\langle a^1, b^1, \mathrm{c}^1, d^0 \rangle$ | 0.0492 |
| $\langle a^1, b^1, \mathrm{c}^0, d^0 \rangle$ | 0.8852 |
| $\langle a^1, b^0, \mathrm{c}^1, d^0 \rangle$ | 0.0164 |
| $\langle a^1, b^0, \mathrm{c}^0, d^0 \rangle$ | 0.0492 |

| | |
|---|---|
| $\langle a^1, b^1, \mathrm{c}^1, d^1 \rangle$ | 0.2579 |
| $\langle a^1, b^1, \mathrm{c}^0, d^1 \rangle$ | 0.1290 |
| $\langle a^0, b^1, \mathrm{c}^1, d^1 \rangle$ | 0.2708 |
| $\langle a^0, b^1, \mathrm{c}^0, d^1 \rangle$ | 0.3423 |




Parameters:

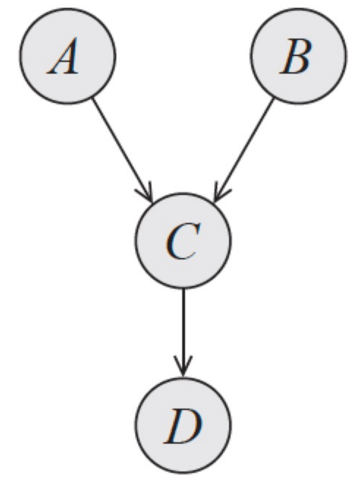
$$\boldsymbol{\theta}_{a^1} = 0.3 \quad \boldsymbol{\theta}_{b^1} = 0.9$$
$$\boldsymbol{\theta}_{d^1|c^0} = 0.1 \quad \boldsymbol{\theta}_{d^1|c^1} = 0.8$$
$$\boldsymbol{\theta}_{c^1|a^0,b^0} = 0.83 \quad \boldsymbol{\theta}_{c^1|a^1,b^0} = 0.6$$
$$\boldsymbol{\theta}_{c^1|a^0,b^1} = 0.09 \quad \boldsymbol{\theta}_{c^1|a^1,b^1} = 0.2$$

We can now compute expected counts,

$$\bar{M}_{\boldsymbol{\theta}}[\boldsymbol{y}] = \sum_{m=1}^{M} \sum_{\boldsymbol{h}[m] \in \mathrm{Val}(\boldsymbol{H}[m])} Q(\boldsymbol{h}[m]) \mathbb{I}\{\xi[m]\langle \boldsymbol{Y} \rangle = \boldsymbol{y}\}$$

# Another Example

$$o = \langle a^1, ?, ?, d^0 \rangle$$
$$o' = \langle ?, b^1, ?, d^1 \rangle$$




This is like having the data

| | |
|---|---|
| $\langle a^1, b^1, c^1, d^0 \rangle$ | 0.0492 |
| $\langle a^1, b^1, c^0, d^0 \rangle$ | 0.8852 |
| $\langle a^1, b^0, c^1, d^0 \rangle$ | 0.0164 |
| $\langle a^1, b^0, c^0, d^0 \rangle$ | 0.0492 |

| | |
|---|---|
| $\langle a^1, b^1, c^1, d^1 \rangle$ | 0.2579 |
| $\langle a^1, b^1, c^0, d^1 \rangle$ | 0.1290 |
| $\langle a^0, b^1, c^1, d^1 \rangle$ | 0.2708 |
| $\langle a^0, b^1, c^0, d^1 \rangle$ | 0.3423 |

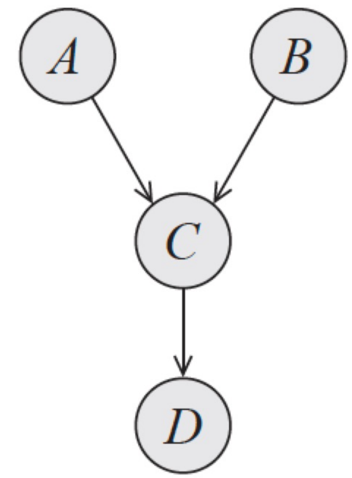
Parameters:

$$\boldsymbol{\theta}_{a^1} = 0.3 \quad \boldsymbol{\theta}_{b^1} = 0.9$$
$$\boldsymbol{\theta}_{d^1|c^0} = 0.1 \; \boldsymbol{\theta}_{d^1|c^1} = 0.8$$
$$\boldsymbol{\theta}_{c^1|a^0,b^0} = 0.83 \; \boldsymbol{\theta}_{c^1|a^1,b^0} = 0.6$$
$$\boldsymbol{\theta}_{c^1|a^0,b^1} = 0.09 \; \boldsymbol{\theta}_{c^1|a^1,b^1} = 0.2$$

We can now compute expected counts,

$$\bar{M}_{\boldsymbol{\theta}}[\boldsymbol{y}] = \sum_{m=1}^{M} \sum_{\boldsymbol{h}[m] \in \mathrm{Val}(\boldsymbol{H}[m])} Q(\boldsymbol{h}[m]) \mathbb{I}\{\xi[m]\langle \boldsymbol{Y} \rangle = \boldsymbol{y}\}$$

# Another Example




$$o = \langle a^1, ?, ?, d^0 \rangle$$
$$o' = \langle ?, b^1, ?, d^1 \rangle$$

We can now compute expected counts,

$$\bar{M}_{\boldsymbol{\theta}}[\boldsymbol{y}] = \sum_{m=1}^{M} \sum_{\boldsymbol{h}[m] \in \mathrm{Val}(\boldsymbol{H}[m])} Q(\boldsymbol{h}[m]) \mathbb{I}\{\xi[m]\langle \boldsymbol{Y} \rangle = \boldsymbol{y}\}$$

$$\bar{M}_{\boldsymbol{\theta}}[d^1, c^0] = Q'(\langle d^1, c^0 \rangle) + Q'(\langle d^1, c^0 \rangle) = 0.1290 + 0.3423 = 0.4713$$

wrong formula in the book

$$\bar{M}_{\boldsymbol{\theta}}[c^0] = Q(\langle b^1, c^0 \rangle) + Q(\langle b^0, c^0 \rangle) + Q'(\langle a^1, c^0 \rangle) + Q'(\langle a^0, c^0 \rangle)$$
$$= 0.8852 + 0.0492 + 0.1290 + 0.3423 = 1.4057$$

Once you have the expected counts,
you can do MLE estimation and update the parameters

$$\bar{\theta}_{d^1|c^0} = \frac{\bar{M}_{\boldsymbol{\theta}}[d^1, c^0]}{\bar{M}_{\boldsymbol{\theta}}[c^0]}$$

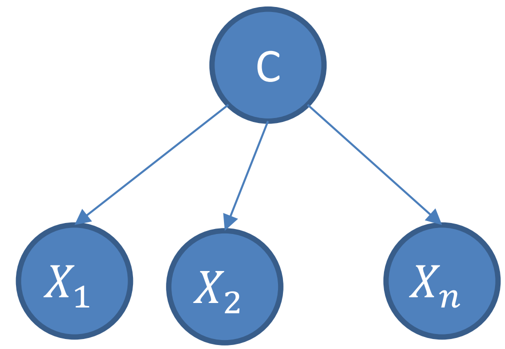
Parameters:

$$\boldsymbol{\theta}_{a^1} = 0.3 \quad \boldsymbol{\theta}_{b^1} = 0.9$$
$$\boldsymbol{\theta}_{d^1|c^0} = 0.1 \quad \boldsymbol{\theta}_{d^1|c^1} = 0.8$$
$$\boldsymbol{\theta}_{c^1|a^0,b^0} = 0.83 \quad \boldsymbol{\theta}_{c^1|a^1,b^0} = 0.6$$
$$\boldsymbol{\theta}_{c^1|a^0,b^1} = 0.09 \quad \boldsymbol{\theta}_{c^1|a^1,b^1} = 0.2$$

# EM Overview

- Pick a starting point for parameters
- Iterate:
  - E-step (Expectation): "Complete" the data using current parameters
  - M-step (Maximization): Estimate parameters relative to data completion
- Guaranteed to improve $L(\theta : D)$ at each iteration

# Example: Clustering

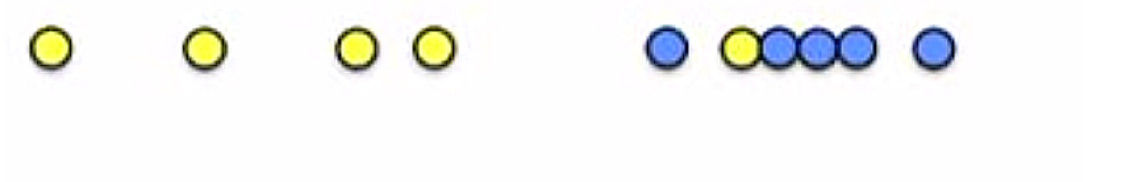



Clustering methods
- hard clustering: clusters do not overlap
- soft clustering: clusters may overlap

Mixture models
- probabilistically-grounded way of doing soft clustering
- each cluster: a generative model (Gaussian or multinomial)
- parameters (e.g. mean/covariance are unknown)

# Example: Gaussian Mixture models

- Observations $x_1 \dots x_n$
  - $K = 2$ Gaussians with unknown $\mu, \sigma^2$
  - estimation trivial if we know the source of each observation



# Example: Gaussian Mixture models

- Observations $x_1 \dots x_n$
  - $K = 2$ Gaussians with unknown $\mu, \sigma^2$
  - estimation trivial if we know the source of each observation

$$\mu_b = \frac{x_1 + x_2 + \cdots + x_{n_b}}{n_b}$$

$$\sigma_b^2 = \frac{(x_1 - \mu_1)^2 + \cdots + (x_n - \mu_n)^2}{n_b}$$

# Example: Gaussian Mixture models

- Observations $x_1, \ldots, x_n$
  - $K = 2$ Gaussians with unknown $\mu, \sigma^2$
  - estimation trivial if we know the source of each observation

- What if we don't know the source?
- If we knew parameters of the Gaussians $(\mu, \sigma^2)$ , we can guess whether point is more likely to be $a$ or $b$.

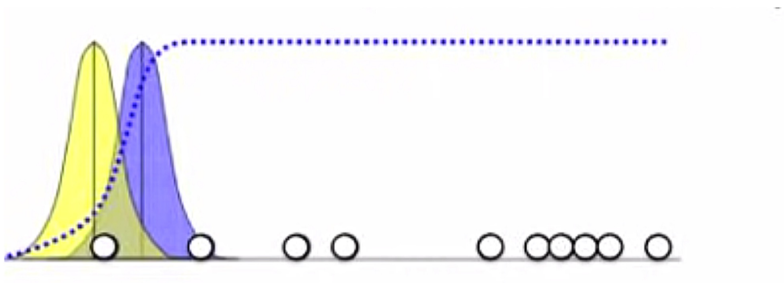
$$P(b \mid x_i) = \frac{P(x_i \mid b)P(b)}{P(x_i \mid b)P(b) + P(x_i \mid a)P(a)}$$

$$P(x_i \mid b) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left(-\frac{(x_i - \mu_b)^2}{2\sigma_b^2}\right)$$

# EM for GMMs

- Chicken and egg problem
  - need $(\mu_a, \sigma_a{}^2)$ and $(\mu_b, \sigma_b{}^2)$ to guess source of points
  - need to know source to estimate $(\mu_a, \sigma_a{}^2)$ and $(\mu_b, \sigma_b{}^2)$
- EM algorithm
  - start with two randomly placed Gaussians $(\mu_{a'} \sigma_a^2), (\mu_b, \sigma_b{}^2)$
  - for each point: $P(b \mid x_i)$ = does it look like it came from $b$ ?
  - adjust $(\mu_{a'}, \sigma_a{}^2)$ and $(\mu_b, \sigma_b{}^2)$ to fit points assigned to them

# Example: EM for 1-D GMMs

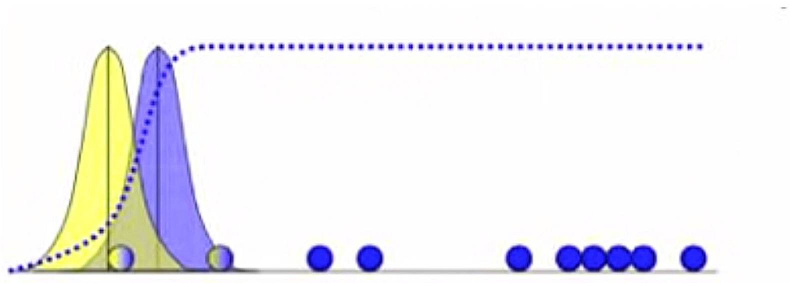


$$P(x_i \mid b) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left(-\frac{(x_i - \mu_s)^2}{2\sigma_b^2}\right)$$

$$b_i = P(b \mid x_i) = \frac{P(x_i \mid b)P(b)}{P(x_i \mid b)P(b) + P(x_i \mid a)P(a)}$$

$$a_i = P(a \mid x_i) = 1 - b_i$$

# Example: EM for 1-D GMMs

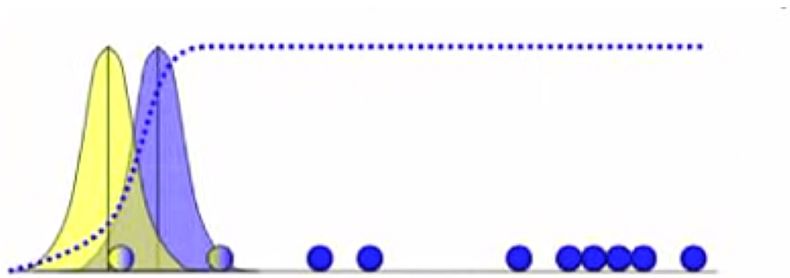


$$P(x_i \mid b) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left(-\frac{(x_i - \mu_s)^2}{2\sigma_b^2}\right)$$

$$b_i = P(b \mid x_i) = \frac{P(x_i \mid b)P(b)}{P(x_i \mid b)P(b) + P(x_i \mid a)P(a)}$$

$$a_i = P(a \mid x_i) = 1 - b_i$$

# Example: EM for 1-D GMMs



$$P(x_i \mid b) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left(-\frac{(x_i - \mu_s)^2}{2\sigma_b^2}\right)$$

$$b_i = P(b \mid x_i) = \frac{P(x_i \mid b)P(b)}{P(x_i \mid b)P(b) + P(x_i \mid a)P(a)}$$

$$a_i = P(a \mid x_i) = 1 - b_i$$

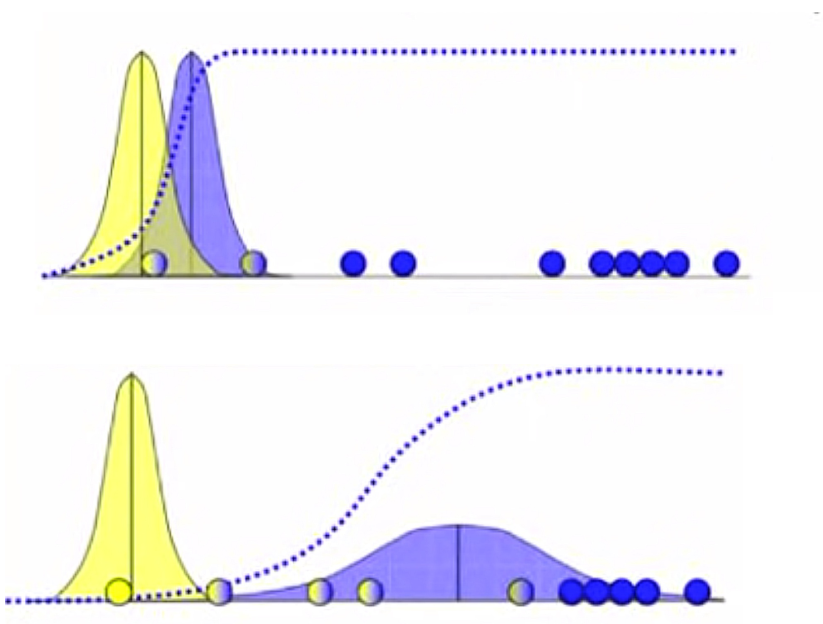
$$\mu_b = \frac{b_1 x_1 + b_2 x_2 + \cdots + b_n x_n}{b_1 + b_2 + \cdots + b_n}$$

$$\sigma_b^2 = \frac{b_1(x_1 - \mu_1)^2 + \cdots + b_n(x_n - \mu_n)^2}{b_1 + b_2 + \cdots + b_n}$$

$$\mu_a = \frac{a_1 x_1 + a_2 x_2 + \cdots + a_n x_{n_n}}{a_1 + a_2 + \cdots + a_n}$$

$$\sigma_a^2 = \frac{a_1(x_1 - \mu_1)^2 + \cdots + a_n(x_n - \mu_n)^2}{a_1 + a_2 + \cdots + a_n}$$

# Example: EM for 1-D GMMs



$$P(x_i \mid b) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left(-\frac{(x_i - \mu_s)^2}{2\sigma_b^2}\right)$$

$$b_i = P(b \mid x_i) = \frac{P(x_i \mid b)P(b)}{P(x_i \mid b)P(b) + P(x_i \mid a)P(a)}$$

$$a_i = P(a \mid x_i) = 1 - b_i$$

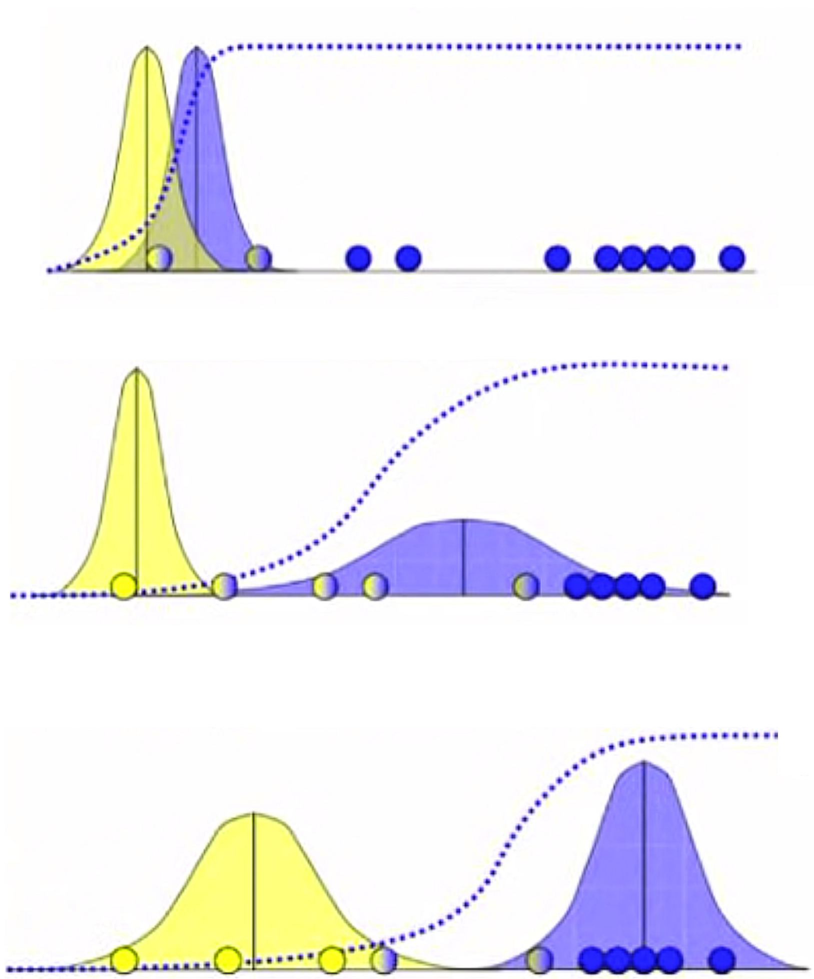
$$\mu_b = \frac{b_1 x_1 + b_2 x_2 + \cdots + b_n x_n}{b_1 + b_2 + \cdots + b_n}$$

$$\sigma_b^2 = \frac{b_1(x_1 - \mu_1)^2 + \cdots + b_n(x_n - \mu_n)^2}{b_1 + b_2 + \cdots + b_n}$$

$$\mu_a = \frac{a_1 x_1 + a_2 x_2 + \cdots + a_n x_{n_n}}{a_1 + a_2 + \cdots + a_n}$$

$$\sigma_a^2 = \frac{a_1(x_1 - \mu_1)^2 + \cdots + a_n(x_n - \mu_n)^2}{a_1 + a_2 + \cdots + a_n}$$

# Example: EM for 1-D GMMs



$$P(x_i \mid b) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left(-\frac{(x_i - \mu_s)^2}{2\sigma_b^2}\right)$$

$$b_i = P(b \mid x_i) = \frac{P(x_i \mid b)P(b)}{P(x_i \mid b)P(b) + P(x_i \mid a)P(a)}$$

$$a_i = P(a \mid x_i) = 1 - b_i$$

$$\mu_b = \frac{b_1 x_1 + b_2 x_2 + \cdots + b_n x_n}{b_1 + b_2 + \cdots + b_n}$$

$$\sigma_b^2 = \frac{b_1(x_1 - \mu_1)^2 + \cdots + b_n(x_n - \mu_n)^2}{b_1 + b_2 + \cdots + b_n}$$

$$\mu_a = \frac{a_1 x_1 + a_2 x_2 + \cdots + a_n x_{n_n}}{a_1 + a_2 + \cdots + a_n}$$

$$\sigma_a^2 = \frac{a_1(x_1 - \mu_1)^2 + \cdots + a_n(x_n - \mu_n)^2}{a_1 + a_2 + \cdots + a_n}$$

# Expectation-Maximization

Iterate until convergence:
On the $t-$ th iteration let our estimates be

$$\lambda_t = \{\mu_1(t), \mu_2(t) \ldots \mu_c(t)\}$$

Just evaluate a Gaussian at $x_k$

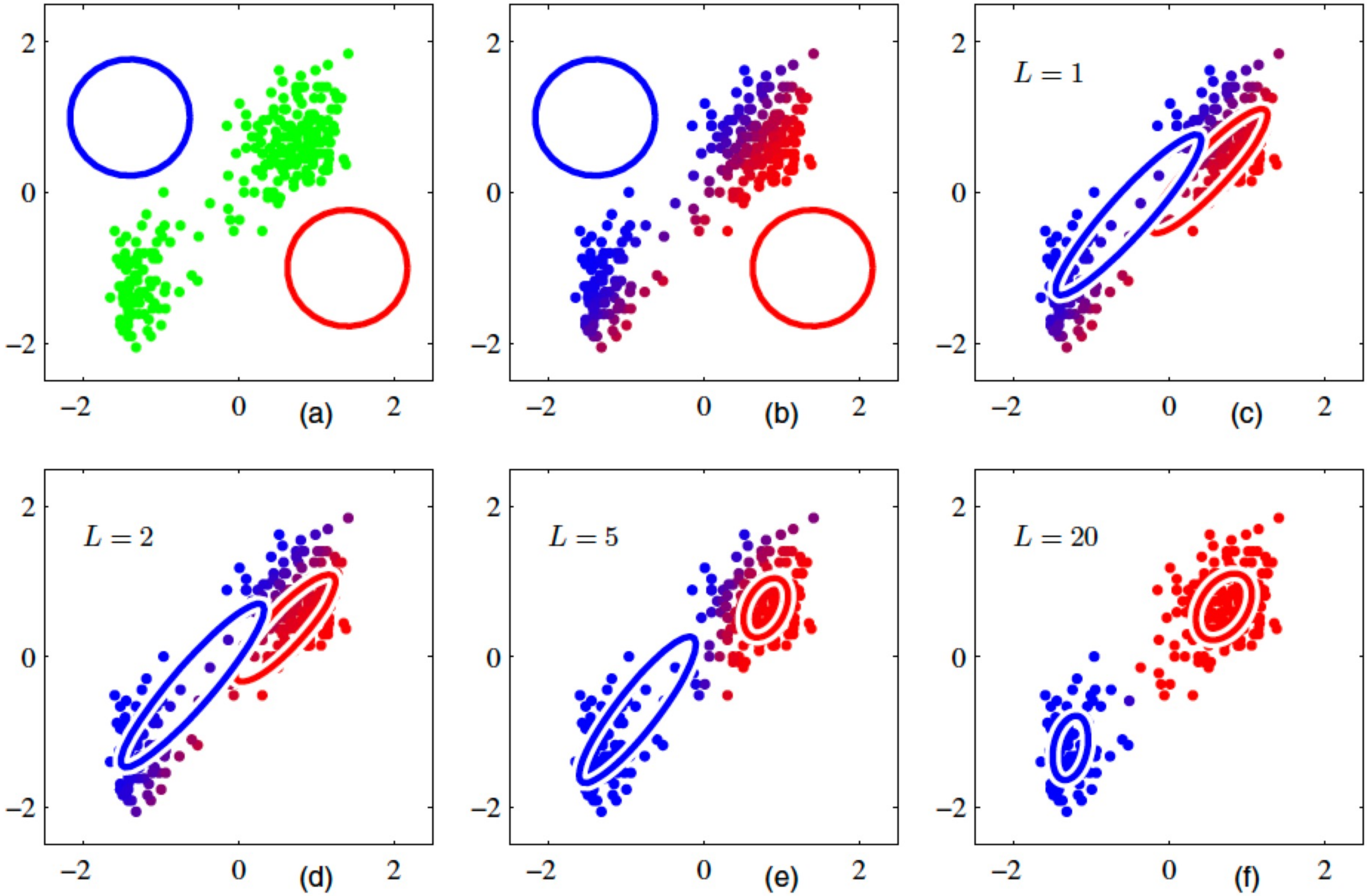
E-step: Compute “expected" classes of all datapoints for each class

$$\mathrm{P}(z_i \mid x_k, \lambda_t) = \frac{\mathrm{p}(x_k \mid z_i, \lambda_t)\mathrm{P}(z_i \mid \lambda_t)}{\mathrm{p}(x_k \mid \lambda_t)} = \frac{\mathrm{p}(x_k \mid z_i, \mu_i(t), \sigma^2\mathbf{I})p_i(t)}{\sum_{j=1}^{c} \mathrm{p}\left(x_k \mid z_j, \mu_j(t), \sigma^2\mathbf{I}\right)p_j(t)}$$

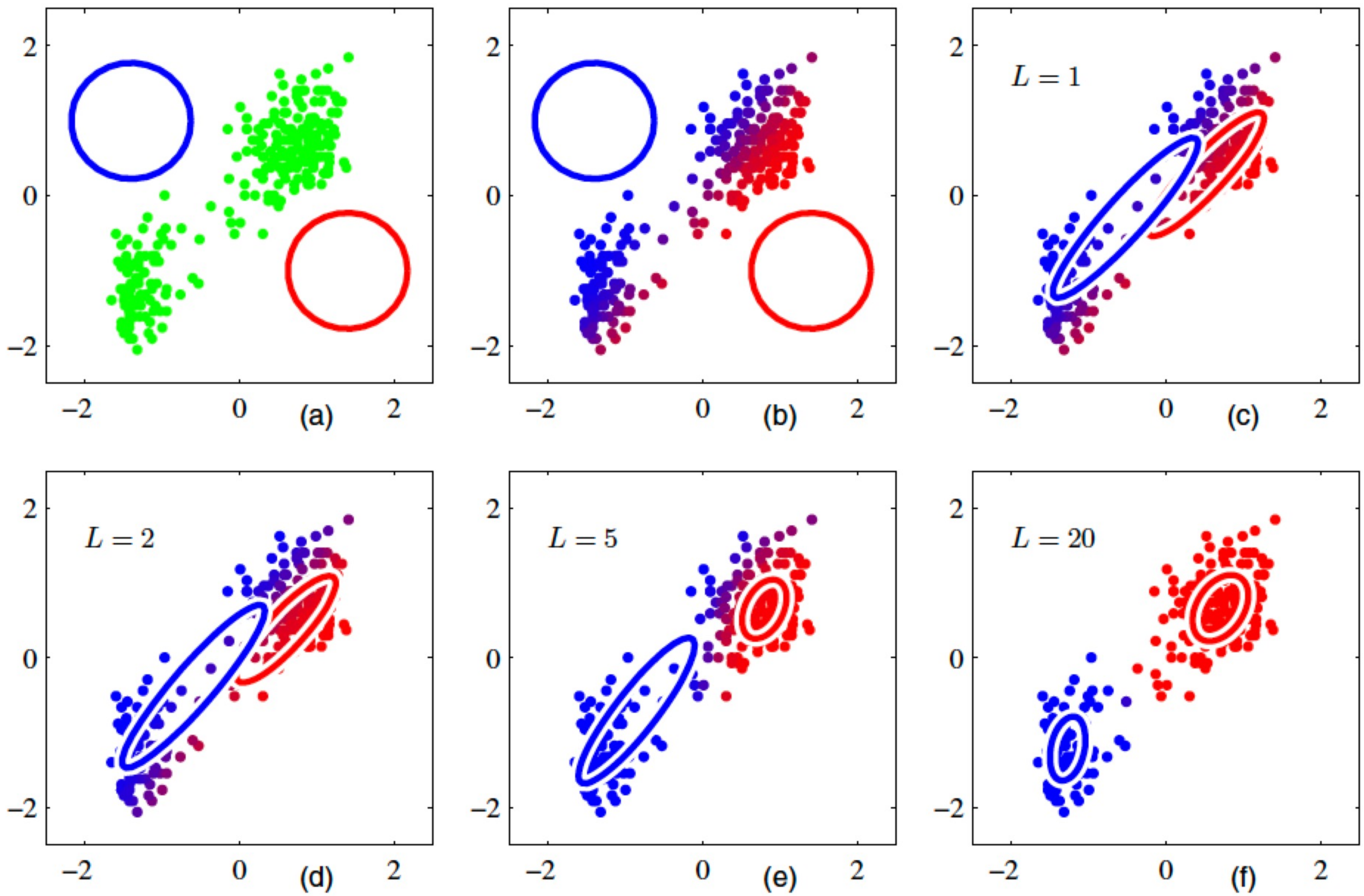
M-step: Estimate $\boldsymbol{\mu}$ given our data's class membership distributions

$$\mu_i(t+1) = \frac{\sum_k \mathrm{P}(z_i \mid x_k, \lambda_t)x_k}{\sum_k \mathrm{P}(z_i \mid x_k, \lambda_t)}$$

# Example: Gaussian Mixture Models



# Example: Gaussian Mixture Models



# EM Summary

- Need to run inference over each data instance at every iteration
- Pros
  - Easy to implement on top of MLE for complete data
  - Makes rapid progress, especially in early iterations
- Cons
  - Convergence slows down at later iterations