# CSCI E-82a
# Probabilistic Programming and AI
# Lecture 8
# Hidden Markov Models and Variational Methods

Steve Elston

HARVARD
Extension School

# Outline

- Introduction to latent variable models
- Hidden Markov models
- Mixture models
- Variational methods
- Variational EM algorithm
- Density estimation for mixture models
- Gaussian mixture models (GMM)
- EM for Gaussian mixture models
- Non-Uniqueness with Variational EM
- Variational Bayes methods

# Latent Variable Models

## What is a **latent variable model** (LVM)?

- Not all variables in a DAG are observable
  - **Observed variables,** $\nu$
  - Unobserved or **hidden variables ,** ***h***
  - Unobserved variables are know as **latent variables**
- Learning requires estimating parameters for observed and hidden variables
  - But there is **no data for laten variables!**
  - Hidden variables make **learning harder!**
  - Need learning methods that infer parameters and latent variables
  - Use **approximate methods** – no exact methods

# Latent Variable Models

## What are LVMs good for?

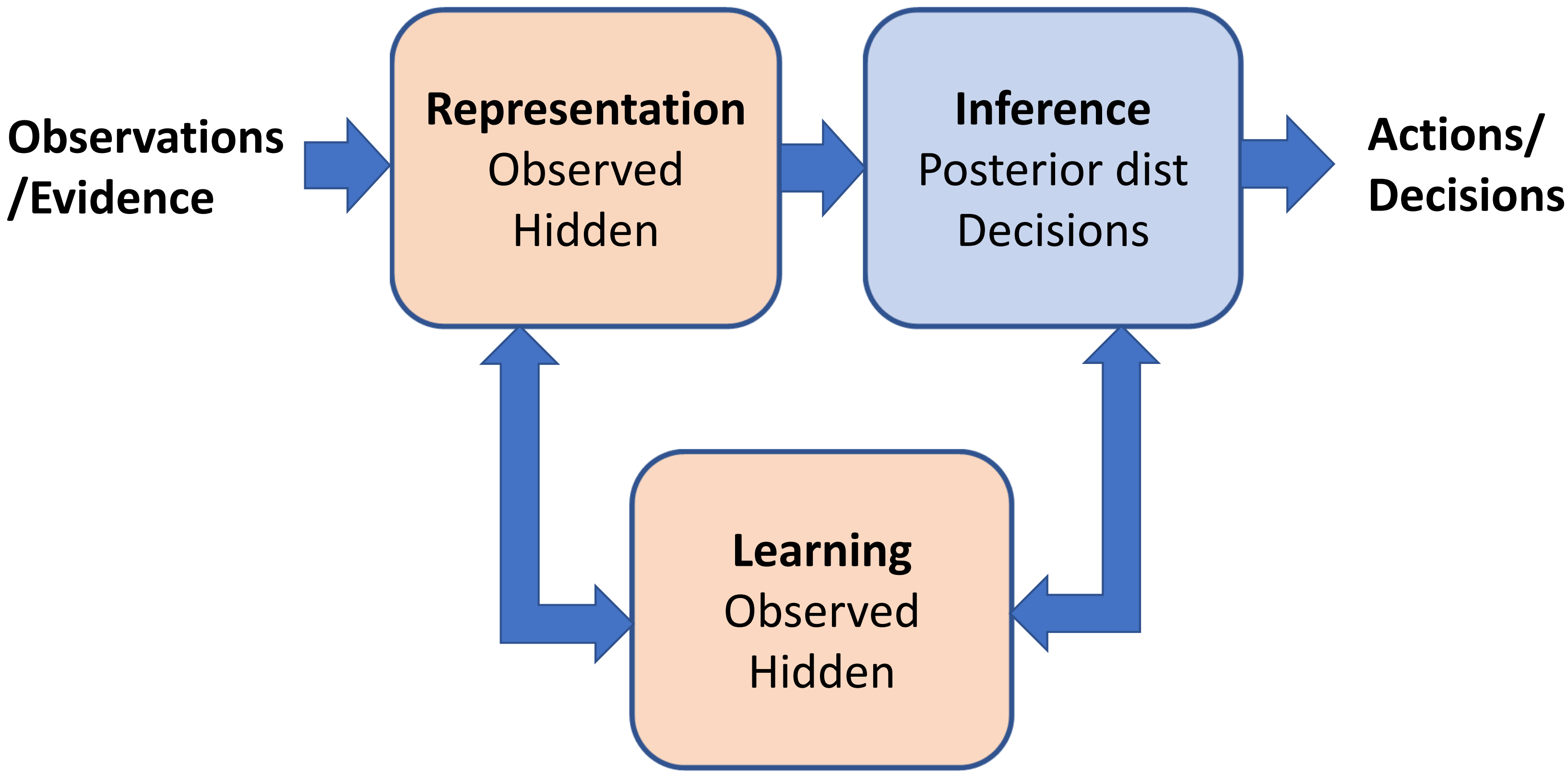
- Models with unobservable data
  - Navigation – e.g. Kalman filter for GPS
  - Physical models – actual values vs. instrument readings
  - Biomedical models – metabolism vs. measurements
  - Bayesian clustering algorithms – Use prior information in clustering
- Mixture models
  - Generalize single distribution models
- Missing data
  - Treat missing data as hidden variable
- Many more!
- One of the most **widely used algorithms** we discuss in this course!

# Latent Variable Models

## Example of LVMs in the news

- Firefighters perform heavy work in an environment with intense heat
- Firefighter safety and effectiveness is impaired if there physiology is responding to exhaustion and heat
- But, **exhaustion and heat are latent variables!**
- We must infer latent variables from observable variables
  - Skin temperature
  - Perspiration
  - Time of exposure
  - Etc.

# Representation and Learning for Latent variable models



# Latent Variable Models

## How is a latent variable model represented?

- A latent variable model has both visible and hidden variables
- The joint distribution is: $p(v, h; \Theta)$

  Where,
  - $\nu$ are the visible variables
  - *h* are the hidden variables
  - $\Theta$ are the model parameters
- The learning problem is hard
  - **Must learn the model parameters, $\Theta$,** and
  - **Must learn values of the hidden variables**, *h*

# Latent Variable Models

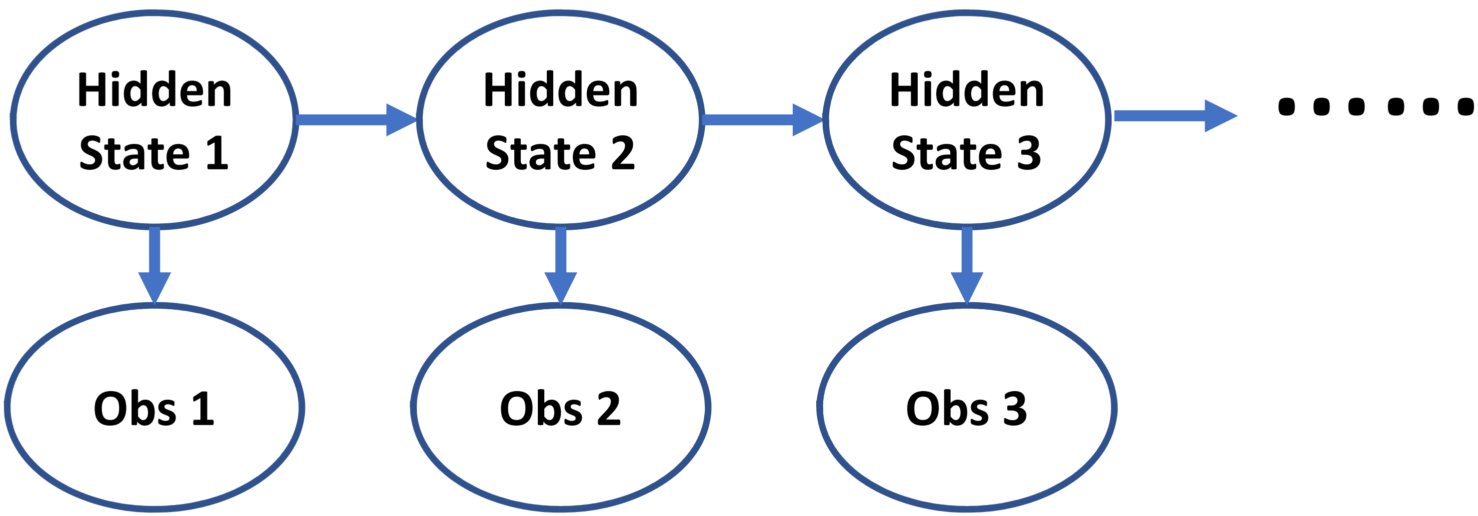
## Learning for latent variable models

- The observed values are known: $\mathcal{V} = \{v^1, v^2, \ldots, v^N\}$
- But we must learn values of both the hidden variables, *h*, and the model parameters, $\Theta$
  - Makes learning for latent variable models hard
- Two learning approaches
  - No exact solution methods
  - **Monte Carlo methods** – we will not discuss these in depth
  - **Variational methods**
    - efficient and becoming widely used
    - The **EM algorithm** is the principle variational method

# Hidden Markov Models

HMMs are latent variable models which represent sequential processes

- Start in initial **hidden state**
- Produces initial **emission** or observation
- Hidden state changes at next time step
- Emission from new state



# Mixture Models

Why are mixture models useful?

- Components of mixture distribution have latent probability of contributing to an observation
- Can treat missing value problems as mixtures of distributions
- Determine if an unscrupulous casino is using fair or loaded dice
    - Determine if process has multiple generating distributions
- Returns of many financial assets cannot be modeled by simple distributions

# Mixture Models

Why are mixture models useful?

- Response rates to a promotional email is a latent variable model
- Responses rates different for different responding populations
  - E.g. respondent to a email offer for men's running shoes might be a male athlete
  - Or, a non-athlete buying the shoes for a friend or relative
  - Or ??
- The response distributions for these populations are the **components of a mixture** – mixture of binomial distributions
- The probability of response being from each population is the **latent variable**

# Mixture Models

## How is a mixture model represented?

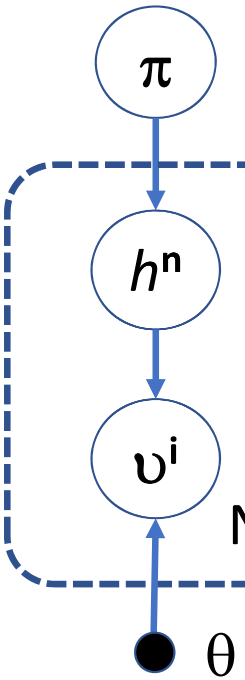
- Start with the variables:
  - $\mathcal{V} = \{v^1, v^2, \ldots, v^N\}$is the vector of **observed real-number values**
  - $h_i \in \{1, 2, 3, \ldots, K\}$ are the **possible states of the hidden (latent) variable**
- The factorized joint distribution is:

$$p(v, h) = p(v \mid h)p(h)$$

- Here the probability that a value, $v^n$, is from the kth component of the mixture is **determined by the latent variable**:

$$p(h = k) = \pi_k$$

# Mixture Models




Mixture models can be represented by a DAG

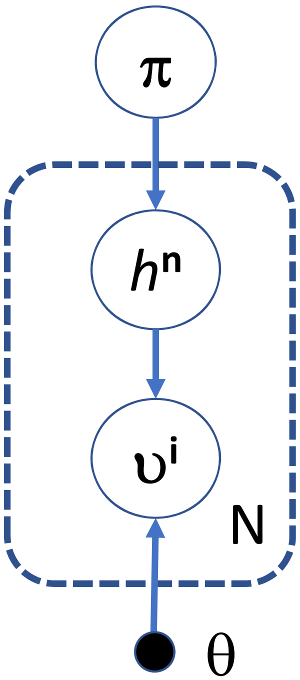
- Probabilities of the components of the mixture, $\pi$
- The weight of a component in the mixture is $\pi$
- CPD for hidden variable, *h*
  - *h* is **switching variable;** determines which component of the mixture generates $\nu^i$
- Parameters of the visible variable distribution, $\theta$
- Visible emission, $\nu$, from distribution conditional on hidden variable, *h*, and parameters, $\theta$
- Repeat for N samples

# Mixture Models

## Example: Gaussian mixture model

- Mixture of Gaussians is one of the mostly widely used mixture models
- Each Gaussian is a **component of the mixture**
- The model parameters have three components, $\theta_k = \{\mu_k, \Sigma_k, \pi_k\}$:
  - $\pi_k \in \{\pi_1, \pi_2, \ldots, \pi_K\}$ is the **probability** of observation, $\nu^k$, being from the **kth component** of the mixture – **latent variable**
  - $\mu_k \in \{\mu_1, \mu_2, \ldots, \mu_K\}$ are the **mean vectors** of the mixture components
  - $\Sigma_k \in \{\Sigma_1, \Sigma_2, \ldots, \Sigma_K\}$ are the **covariance matrices** of the mixture components

# Mixture Models




## Example: Gaussian mixture model

- The probability distribution of a visible variable, ν, from a single component of the mixture is:

$$p(v \mid h = k) = \mathcal{N}(\mu_k, \Sigma_k)$$

- *h* is the latent variable, determines Gaussian component
- The marginal distribution of a visible variable , ν, from the mixture is computed by:

$$p(v) = \sum_{k=1}^{K} p(v \mid h = k) p(h = k)$$

$$= \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \Sigma_k)$$

# Variational Methods

Variational methods attempt to find distribution of hidden variables and model parameters

- **EM algorithm** is a core variational method
- Classical EM algorithm is a **maximum likelihood method**
- Basic EM iteration :
  - **E-step:** Hold parameters constant and update distribution of hidden variables
  - Find **expected values** of hidden variables
  - **M-step:** Hold hidden variable values constant and update distribution of parameters
  - **Maximize likelihood** of model for the observed variables
  - Continue iteration until convergence – small change in values
- Measure fit of distributions with **Kullback-Leibler Divergence**

# Variational Methods

## Review of **Kullback-Leibler Divergence**

- KL divergence of a distribution *q(x)* with respect to another distribution *p(x)* is defined as:

$$\mathbb{D}_{KL}(P \parallel Q) = -\sum_{x} p(x)\ ln_b \frac{p(x)}{q(x)}$$

- Key properties of KL divergence include:
  - $\mathbb{D}_{KL}(P \parallel Q) \geq 0$ for all *q(x)* and *p(x)*
  - $\mathbb{D}_{KL}(P \parallel Q) = 0$ if and only if *q(x) = p(x)*
  - $\mathbb{D}_{KL}(P \parallel Q) \neq \mathbb{D}_{KL}(Q \parallel P)$ as KL divergence is not symmetric, and is therefore not a distance metric
  - $\mathbb{D}_{KL}(P|Q) = \mathbb{H}(P) + \mathbb{H}(P, Q)$ or KL divergence is the sum of the entropy of *p(x)* and the cross entropy between *q(x)* and *p(x)*

# Variational EM

Goal is to **maximize the marginal likelihood** of visible variables given the parameters $p(v \mid \theta)$

- Need to find the variational distribution, $p(v \mid h)$, which minimizes KL divergence with respect to $p(v \mid h, \theta)$
- Minimizing KL divergence maximizes likelihood
- Using the KL divergence, we can find the **variational upper bound**:

$$\mathbb{D}_{KL}(q(h \mid v) \parallel p(h \mid v, \theta)) = \mathbb{E}_{q(h \mid v)}\left[log(q(h \mid v)) - log(p(h \mid v, \theta))\right] \geq 0$$

# Variational EM

Goal is to maximize the marginal likelihood of visible variables given the parameters $p(v \mid \theta)$

- We know v, so can maximize the likelihood of observations
- Starting with the **variational upper bound**:

$$\mathbb{D}_{KL}(q(h \mid v) \parallel p(h \mid v, \theta)) = \mathbb{E}_{q(h \mid v)}\left[log(q(h \mid v)) - log(p(h \mid v, \theta))\right] \geq 0$$

- Expand conditional distribution as:

$$p(h \mid v, \theta) = \frac{p(h, v \mid \theta)}{p(v \mid \theta)}$$

- After some substitution and rearrangement of terms the bound is:

$$log\ p(v \mid \theta) \geq -\mathbb{E}_{q(h \mid v)}\left[log(q(h^n \mid v^n))\right] + \mathbb{E}_{q(h \mid v)}\left[log(p(h, v \mid \theta))\right]$$

# Variational EM

Goal is to maximize the marginal likelihood of visible variables given the parameters $p(v \mid \theta)$

- The bound on $p(v \mid \theta)$ is:

$$log\ p(v \mid \theta) \geq -\mathbb{E}_{q(h \mid v)}\left[log(q(h^n \mid v^n))\right] + \mathbb{E}_{q(h \mid v)}\left[log(p(h, v \mid \theta))\right]$$

$$\geq -Entropy\ term + Energy\ term$$

- The entropy term of hidden value
- The energy term is the **expected complete data log likelihood**

# Variational EM

Goal is to maximize the marginal likelihood of visible variables given the parameters $p(v \mid \theta)$

- The bound on $p(v \mid \theta)$ is for a single observation, $v$:

$$log\ p(v \mid \theta) \geq -\mathbb{E}_{q(h \mid v)}\left[log(q(h^n \mid v^n))\right] + \mathbb{E}_{q(h \mid v)}\left[log(p(h, v \mid \theta))\right]$$

- Now for a set of observations, $\mathcal{V} \in \{v^1, v^2, \ldots, v^N\}$, we can write:

$$p(\mathcal{V} \mid \theta) \geq \tilde{\mathcal{L}}(q^*, \theta)$$
$$\equiv -\sum_{n=1}^{N} \mathbb{E}_{q(h^n \mid v^n)}\left[log(q(h^n \mid v^n))\right] + \sum_{n=1}^{N} \mathbb{E}_{q(h^n \mid v^n)}\left[log(p(h^n, v^n \mid \theta))\right]$$

Where $\tilde{\mathcal{L}}(q^*, \theta)$ is the likelihood of the **variational distribution *q****
given $\theta$

- Bound is exact if $q(h^n \mid v^n) = p(h^n, v^n \mid \theta)$ for $n \in \{1, 2, \ldots, N\}$

# Variational EM

**E-step** maximizes $\tilde{\mathcal{L}}(q^*, \theta)$

- Fix the parameters, θ
- Vary the distribution $q(h^n \mid v^n)$ to maximize $\tilde{\mathcal{L}}(q^*, \theta)$
- We cannot observer *h*, but we can compute distribution $q^{new}(h^n \mid v^n, \theta)$, using:
  - Distribution of observed data values, *p*(ν)
  - Current distribution of the parameters, $q^{old}(\theta)$
- This process of generating updated values for the hidden variables is known as **hallucinating data**
- Vary $q^{new}(h)$ to minimize the KL divergence and therefore **maximizes the expected likelihood** of $log(p(h, v; \theta))$
- Maximizing the expected likelihood is why this is the E-step

# Variational EM

**M-step** updates parameters $\theta$

- Fix the distribution of the hidden variable values, $q^{old}(h^n \mid v^n)$
- Using observed values, $\mathcal{V} = \{v^1, v^2, \ldots, v^N\}$, maximize the likelihood, $\tilde{\mathcal{L}}(q^*, \theta)$
- Only the energy term depends on $\theta$, so maximize:

$$\sum_{n=1}^{N} \mathbb{E}_{q(h^n \mid v^n)} \left[ log(p(h^n, v^n \mid \theta)) \right]$$

# Density Estimation for Mixture Models

How to estimate probability density for simple mixture?

- Mixture has K components, indexed by a hidden variable, $h_i \in \{1, 2, 3, \ldots, K\}$
- The observed variables have values $\mathcal{V} = \{v^1, v^2, \ldots, v^N\}$
- The probability that an observed value, $\nu^n$, is generated by the ith component is $p(i) = \pi(i)$
- The conditional probability of an observed value, $\nu^n$, from the ith component is $p(v^n \mid i)$
- The probability density of the observed variables is then:

$$p(v^1, v^2, \ldots, v^N) = \prod_{i=1}^{N} p(v_i \mid i)\pi_i$$

# Gaussian Mixture Models (GMMs)

## Example: EM for a GMM

- The model parameters have three components, $\theta_k = \{\mu_k, \Sigma_k, \pi_k\}$
  - $\pi_k \in \{\pi_1, \pi_2, \ldots, \pi_K\}$ is the **probability** of observation, $\nu^n$, being from the **kth component** of the mixture, for **hidden variable *k***
  - $\mu_k \in \{\mu_1, \mu_2, \ldots, \mu_K\}$ are the **mean vectors** of the mixture components
  - $\Sigma_k \in \{\Sigma_1, \Sigma_2, \ldots, \Sigma_K\}$ are the **covariance matrices** of the mixture components

# Gaussian Mixture Models (GMMs)

## Example: EM for a GMM

- The conditional probability distribution for one component of the mixture is given by:

$$p(\nu \mid \mu_i, \Sigma_i) = \frac{1}{\sqrt{det(2\pi\Sigma_i)}} exp\Big[-\frac{1}{2}(\nu - \mu_i)\Sigma_i^{-1}(\nu - \mu_i)\Big]$$

- And, the probability distribution for the GMM is:

$$p(\nu) = \sum_{i=1}^{H} p(\nu \mid \mu_i, \Sigma_i)\pi_i$$

# Gaussian Mixture Models (GMMs)

## Example: EM for a GMM

- For a set of observed values, $\mathcal{V} \in \{\nu^1, \nu^2, \ldots, \nu^N\}$, the conditional probability distribution given the parameters, $\theta$, is:

$$p(\mathcal{V} \mid \theta) = \sum_{n=1}^{N} log \sum_{i=1}^{H} \frac{\pi_i}{\sqrt{det(2\pi\Sigma_i)}} exp\Big[-\frac{1}{2}(\nu - \mu_i)\Sigma_i^{-1}(\nu - \mu_i)\Big]$$

- There are constraints on component probability $\pi_i$:

$$0.0 \geq \pi_i \geq 1.0$$

$$\sum_{i=1}^{H} \pi_i = 0.0$$

# EM for Gaussian Mixture Models

## Example: EM for a GMM

- Hidden variable is *i*, the index for the mixture component
- There are three parts of the parameter vector in the model:
  $\theta = \{\mu_i, \Sigma_i, \pi_i, i = 1, \ldots, H\}$
- Use EM algorithm to perform optimization to maximize the log likelihood of the observed data given parameters: $log(p(v \mid \theta_*))$
- In the **M-step** the three parts of $\theta$ are updated independently for each of the K components of the mixture, while holding the latent variable distribution constant
- In the **E-step** the conditional distribution of the latent variables, give the observed data , $p(i|\mathcal{V})$, is updated holding parameters, $\theta$, constant

# EM for Gaussian Mixture Models

## Example: EM for a GMM

- How can we interpret the conditional distribution of the latent variables, give the observed data , $p(i|\mathcal{V})$ ?
- The contribution of each component of the **mixture is probabilistic**
- The probabilistic mixture is known as a **soft mixture**
- Soft mixture allows more than one component to contribute to an observed value
- A hard mixture has only one component generating each observation – like a switch

# EM for Gaussian Mixture Models

## Example: M-Step for a GMM

- Only the energy term is dependent on θ
- For observed values, $\mathcal{V} \in \{v^1, v^2, \ldots, v^N\}$, and latent value, *i*, the energy term is:

$$\sum_{n=1}^{N} \mathbb{E}_{q(i|v^n)} \left[ log(p(v^n, i)) \right] = \sum_{n=1}^{N} \mathbb{E}_{q(i|v^n)} \left[ log(p(v^n|i)) log(i) \right]$$

- Substituting the Gaussian distribution for the components of the mixture and making the expectation explicit gives:

$$\sum_{n=1}^{N} \sum_{i=1}^{H} p^{old}(i|v^n) \sum_{n=1}^{N} \left[ -\frac{1}{2}(v^n - \mu_i)\Sigma_i^{-1}(v^n - \mu_i) - \frac{1}{2} log(det(2\pi\Sigma_i)) + logp(i) \right]$$

# EM for Gaussian Mixture Models

## Example: M-Step: $\mu_i$

- Hold all other values constant
- Energy term, for each mixture component is minimized for $\mu$:

$$\sum_{n=1}^{N} \sum_{i=1}^{H} p^{old}(i|v^n)\,(v^n - \mu_i)\Sigma_i^{-1}(v^n - \mu_i)$$

- Introduce the notation:

$$p^{old}(n|i) = \frac{p^{old}(i|v^n)}{\sum_{n=1}^{N} p^{old}(i|v^n)}$$

- The solution of the least squares minimization is:

$$\mu_i^{new} = \sum_{n=1}^{N} p^{old}(n|i) v^n$$

# EM for Gaussian Mixture Models

## Example: M-Step: $\Sigma_i$

- Hold all other values constant
- Energy term, for each mixture component is minimized for $\Sigma$:

$$\sum_{n=1}^{N} \mathbb{E}_{p^{old}(i|v^n))} \left[ (v^n - \mu_i)\Sigma_i^{-1}(v^n - \mu_i) - log(det(2\pi\Sigma_i)) \right]$$

- With solution:

$$m_i^{new} = \sum_{n=1}^{N} p^{old}(n|i)(v^n - \mu_i) \cdot (v^n - \mu_i)$$

# EM for Gaussian Mixture Models

## Example: M-Step *i*

- Hold all other values constant
- Compute expected value of *p(i)*:

$$p^{new}(i) = \frac{1}{N}\sum_{n=1}^{N} p^{old}(i|v^n)$$

# EM for Gaussian Mixture Models

## Example: E-Step: *i*

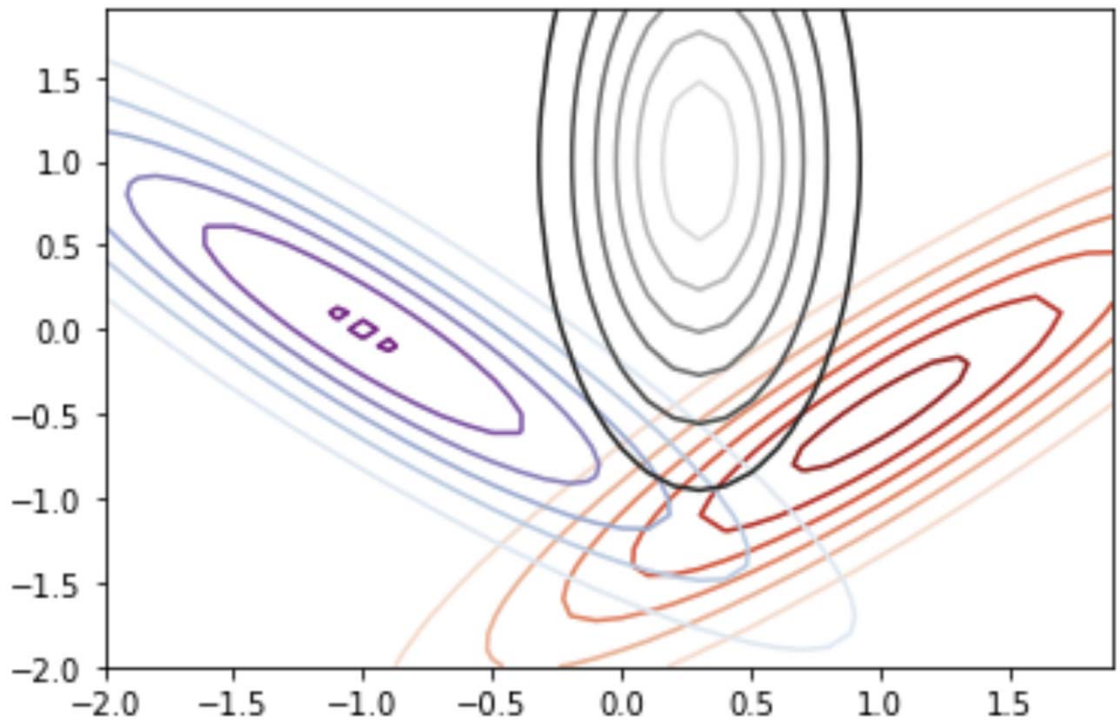
- Hold all model parameters, $\theta$, constant
- Given the observed variables the conditional distribution of the hidden variable, *i*
- Use Bayes theorem and plug in Gaussian distribution to find:

$$p(i|v^n) = \frac{p(v^n|i)\; p(i)}{p(v^n)}$$

$$= \frac{p(i)\; exp\Big[-\frac{1}{2}(v^n - \mu_i)\Sigma_i^{-1}(v^n - \mu_i)\Big]\; det(\Sigma_i)^{-\frac{1}{2}}}{\sum_{i'} p(i')\; exp\Big[\frac{1}{2}(v^n - \mu_{i'})\Sigma_i^{-1}(v^n - \mu_{i'})\Big]\; det(\Sigma_{i'})^{-\frac{1}{2}}}$$

# Non-Uniqueness with Variational EM

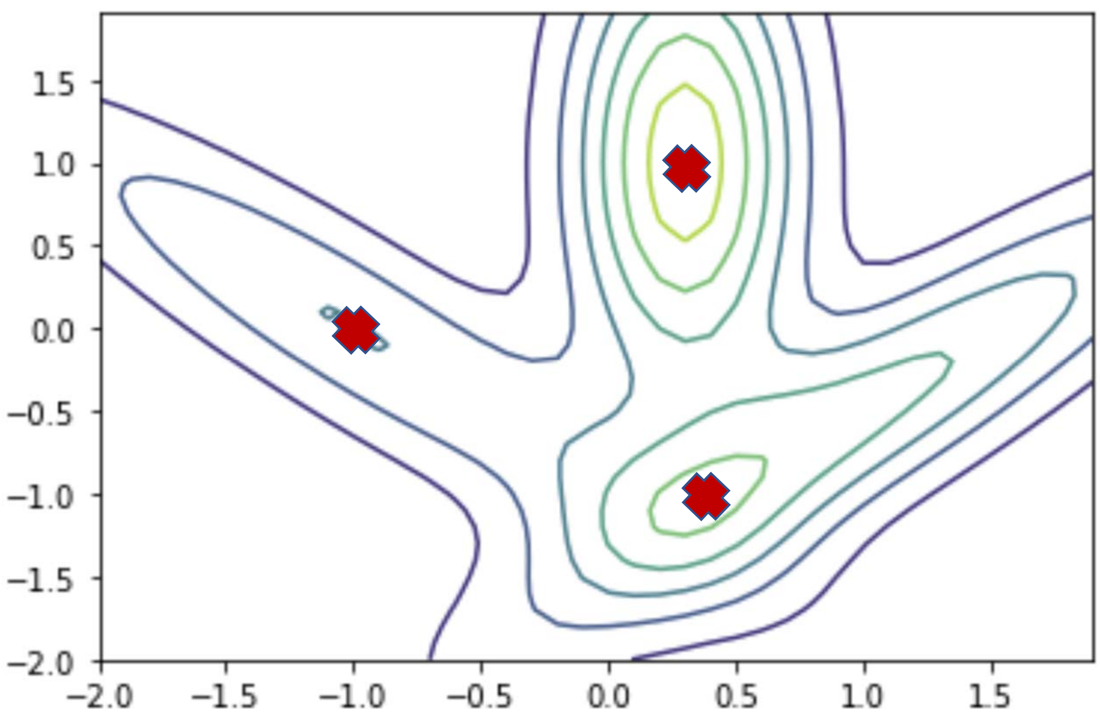
## Why is solving the EM problem difficult?




- Consider a mixture of three Gaussian distributions
- The goal is to find three parameters for each component:
  - The mixture probability
  - The mean
  - The covariance

# Non-Uniqueness with Variational EM

## Why is solving the EM problem difficult?




- The mixture distribution is **non-convex!**
- There are three maximum points for the density
- The maximum found, depends on the initial values
- Optimization (EM) problem is therefore hard

# Variational Bayes

## Variational Bayes is a widely applicable method

- Fully Bayesian method; finds posterior distribution
  - The foregoing classical EM algorithm is a maximum likelihood method
- Variational Bayes useful for:
  - Inference for graphical models
  - Inference for hierarchical models
  - Latent variable models
  - Bayesian clustering methods
  - Etc.

# Variational Bayes

Variational Bayes for latent variable model $p(v, h; \Theta)$

- Variational Bayes is an alternative to Monte Carlo methods
  - Variational methods are gaining popularity
- Compared to Monte Carlo methods, variational methods are:
  - Highly computationally efficient
  - Easy to know when convergence has occurred
  - Often finds a local solution, no guarantee the global solution can be found
- In summary, trade-off between Monte Carlo vs. variational methods is speed and convergence vs. non-global solutions

# Variational Bayes

## Finding the variational lower bound

- Goal is to find the distribution of parameters given observed data values
- Ignoring normalization the conditional distribution can be expanded:

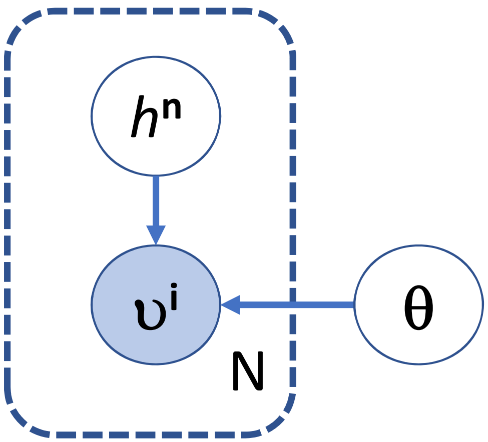
$$p(\theta \mid v) \propto p(v \mid \theta)p(\theta) \propto \sum_h p(v, h \mid \theta)p(\theta)$$

Where, $p(\theta)$ is the prior of the parameters, $\theta$

- Assuming independence of $\nu$ and h, the following approximation holds

$$p(v, h \mid \theta) \approx q(h)q(\theta)$$

# Variational Bayes




The variational Byes model can be represented by a DAG

- Start with the model parameters, $\theta$
- The hidden variables, *h*
- The visible variables are conditional on *h* and $\theta$
- Model *N* observations
- The DAG has the independency between *h* and $\theta$

# Variational Bayes

## Finding the variational lower bound

- We need to minimize the KL divergence between $p(v, h \mid \theta)$ and $q(h)q(\theta)$

$$\mathbb{D}_{KL}(q(h)q(\theta) \parallel p(h, \theta \mid v)) = \mathbb{E}_{q(h)}\left[log(q(h))\right] + \mathbb{E}_{q(\theta)}\left[log(q(\theta))\right] - \mathbb{E}_{q(h)q(\theta)}\left[log(p(h, v \mid \theta))\right] \geq 0$$

- Minimizing KL divergence and rearranging terms, we arrive at the variational lower bound:

$$log(p(v)) \geq -\mathbb{E}_{q(h)}\left[log(q(h))\right] - \mathbb{E}_{q(\theta)}\left[log(q(\theta))\right] + \mathbb{E}_{q(h)q(\theta)}\left[log(p(h, v, \theta))\right]$$

- Can maximize likelihood by minimizing KL divergence
- Minimization can be achieved coordinate-wise
- The bound is reduced as the likelihood increases

# Variational Bayes

## The Variational Bayes EM-algorithm

- As with classical EM the variational Bayes EM algorithm has two steps
- The values of the observed data are used, $\mathcal{V} \in \{v^1, v^2, \ldots, v^N\}$
- For the **M-step** the distribution of the hidden values, $q^{old}(h)$, is held constant and $q(\theta)$ is varied to find a updated distribution of the parameters, $q^{new}(\theta)$.
- In the **E-step** the distribution of the parameters, $q^{old}(\theta)$, is held constant and $q(h)$ is varied to find a update distribution of the hidden values $q^{new}(h)$

# Variational Bayes

## The Variational Bayes EM-algorithm

- In the **M-step** the KL divergence is minimized to update the distribution of the parameters

$$q^{new}(\theta) = \underset{q(\theta)}{argmin}\, \mathbb{D}_{KL}(q^{new}(h)q(\theta) \parallel p(h, \theta \mid \nu))$$

- In the **E-step** the KL divergence is minimized to update the distribution of the hidden values

$$q^{new}(h) = \underset{q(h)}{argmin}\, \mathbb{D}_{KL}(q(h)q^{old}(\theta) \parallel p(h, \theta \mid \nu))$$

- The iteration continues until convergence is reached when the updates are below a predetermined threshold

# Definitions

- A **latent variable model** (LVM) has **observed variables, n** and **hidden variables , *h*** where we want to learn model parameters Q, and h using **Monte Carlo methods** or **variational methods**
- **Hidden Markov Models (HMM)** are LVMs that represent sequential processes with an initial hidden state and producing initial emission (observation), then the hidden state changes and a new emission is observed
- Mixture models are a HMM that can be represented by a DAG
- Variational EM: Goal is to **maximize the marginal likelihood** of visible variables given the parameters using KL divergence
- Variational Bayes: Goal is to finc $p(v \mid \theta)$ ribution of parameters given observed data values

# Definitions

- For a Gaussian Mixture Model (GMM) set of observed values,

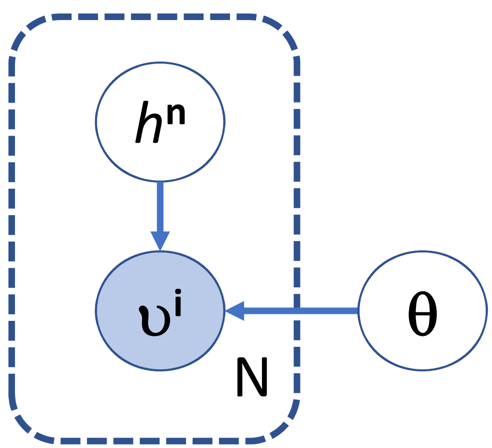
$\mathcal{V} \in \{\nu^1, \nu^2, \dots, \nu^N\}$ , the conditional probability distribution given the parameters, θ, is:

$$p(\mathcal{V} \mid \theta) = \sum_{n=1}^{N} log \sum_{i=1}^{H} \frac{\pi_i}{\sqrt{det(2\pi\Sigma_i)}} exp\Big[ -\frac{1}{2}(\nu - \mu_i)\Sigma_i^{-1}(\nu - \mu_i)\Big]$$

- The **EM algorithm** is a variational method with an E-step (expected value of hidden variables) and a M-step (maximize likelihood of model), where the distribution fit is measured with KL divergence
- The E-step process of generating updated values for the hidden variables is known as **hallucinating data**

# Compare Models

**Variational Bayes**




**Gaussian Mixture Model**