

Lecture 10

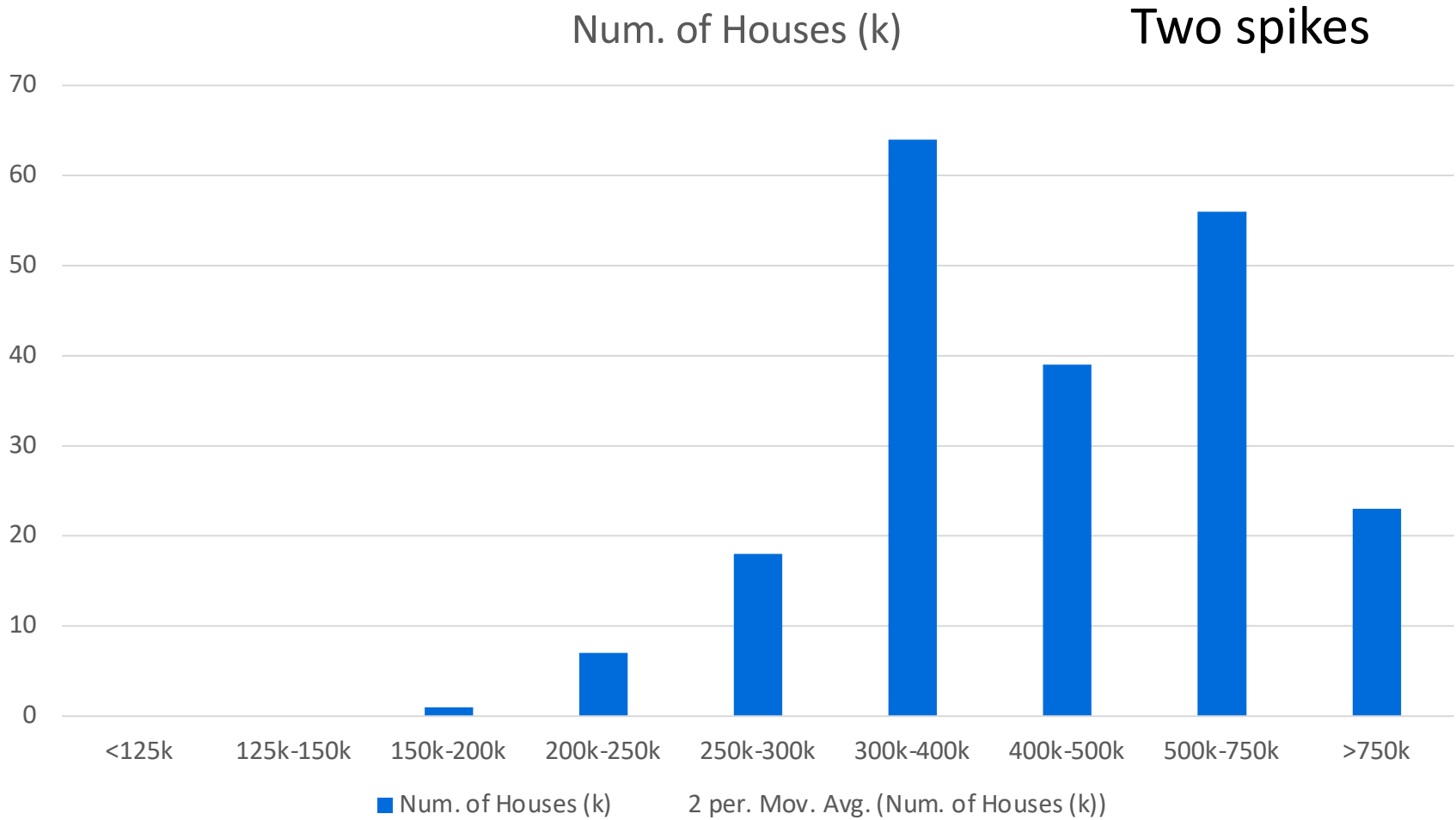
Gaussian Mixture Models

Lei Li and Yu-xiang Wang
UCSB

Recap

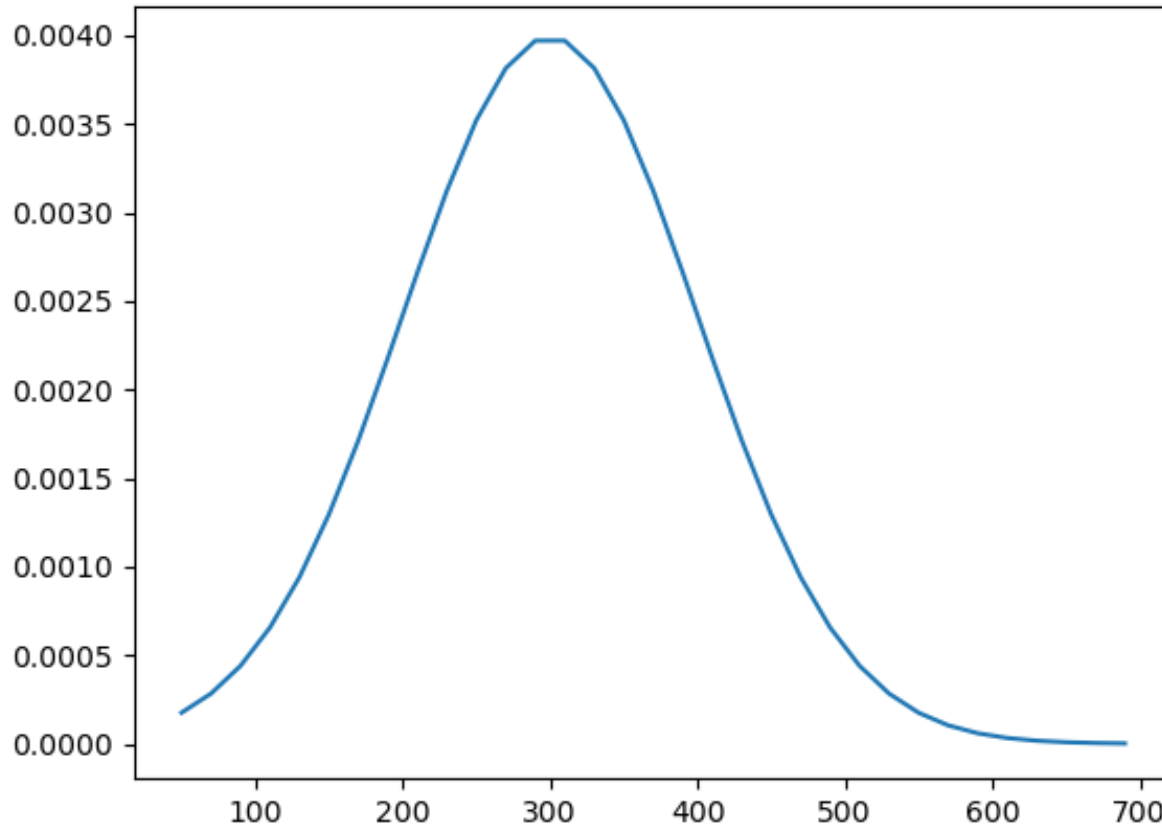
- Bayesian networks:
 - Directed acyclic graph
 - Nodes are random variables
 - arcs are probabilistic dependencies
- Examine dependence of two variables given observation: d-separation

Housing Price Pattern



Gaussian Distribution

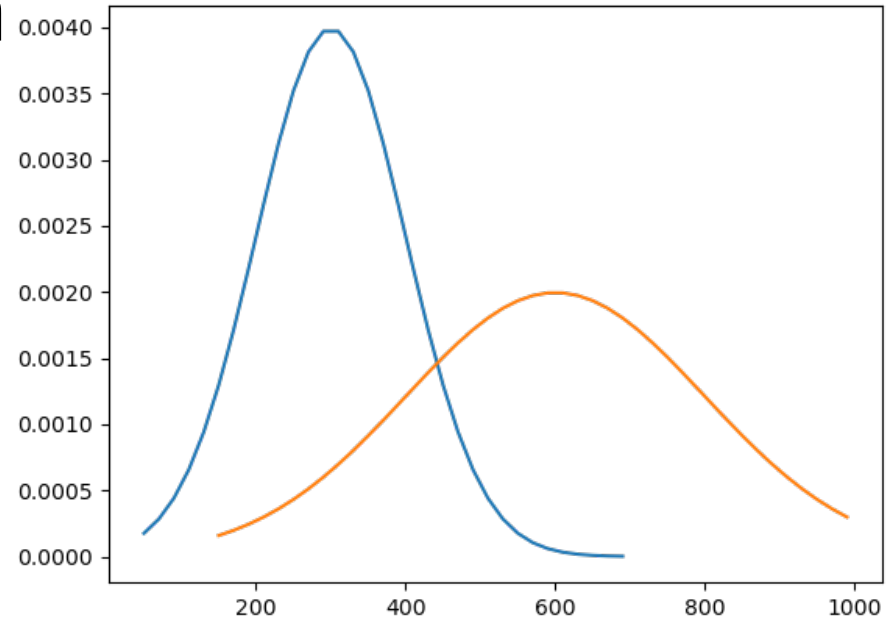
Only single spike



$$p(x) = \frac{1}{\sqrt{(2\pi)^k \det(\Sigma)}} \exp\left[-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu)\right]$$

Two Underlying Patterns

- It might be multiple underlying patterns of Gaussian distribution
 - Los Angeles and Pittsburgh have different median housing price



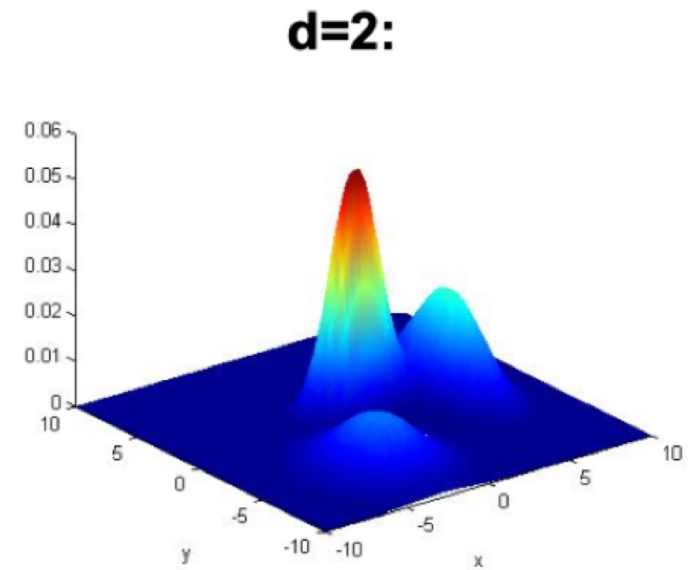
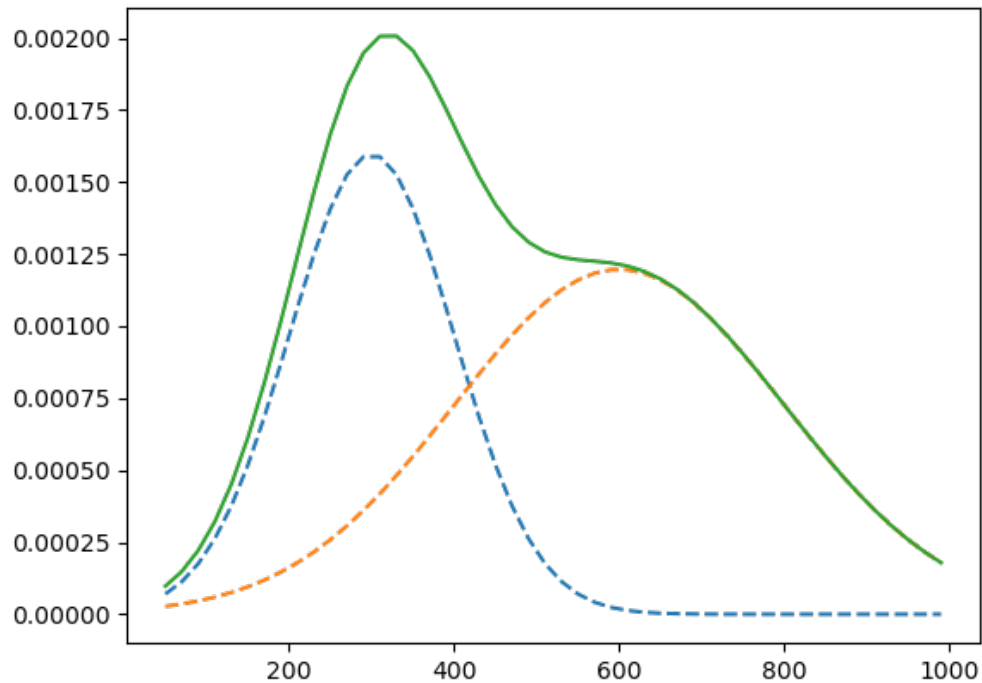
Gaussian Mixture Model

Generative process:

- $z \sim \text{Categorical}(K)$
- $x|z \sim \text{Gaussian}(\mu_z, \Sigma_z)$
- Density:

$$\begin{aligned} p(z, x) &= p(z) \cdot p(x|z) \quad \textcircled{z} \\ &= \begin{cases} w_0 \cdot \mathcal{N}(x|\mu_0, \Sigma_0) \\ w_1 \cdot \mathcal{N}(x|\mu_1, \Sigma_1) \end{cases} \quad \downarrow \\ &\quad \textcircled{x} \\ p(x) &= \sum_{i=1}^K p(z=i, x) = \sum_{i=1}^K p(z=i) p(x|z=i) \\ &= w_0 \mathcal{N}(x|\mu_0, \Sigma_0) + w_1 \mathcal{N}(x|\mu_1, \Sigma_1) \end{aligned}$$

Gaussian Mixture



Mixture Distribution

- Z : latent variable
- $x|z$ can be any distribution in parametric form (e.g. exponential distribution)



Learning Parameters for GMM

- Observation: $x_{1..N}$
- $\theta = \{w_{1..k}, \mu_{1..k}, \Sigma_{1..k}\}$
- MLE (with latent variable z)
- Log-likelihood:
- Expectation-maximization algorithm

$$\begin{aligned}
 p(z, x) &= p(z) \cdot p(x|z) \\
 &= \begin{cases} w_0 \cdot \mathcal{N}(x|\mu_0, \Sigma_0) \\ w_1 \cdot \mathcal{N}(x|\mu_1, \Sigma_1) \end{cases} \\
 p(x) &= \sum_{i=1}^k p(z=i, x) = \sum_{i=1}^k p(z=i) p(x|z=i) \\
 &= w_0 \mathcal{N}(x|\mu_0, \Sigma_0) + w_1 \mathcal{N}(x|\mu_1, \Sigma_1)
 \end{aligned}$$

Handwritten notes: A circled 'Z' is above the first equation, and a circled 'X' is next to the second equation.

$$\begin{aligned}
 \mathcal{L}(\theta) &= \log \prod_{n=1}^N p(x_n | \theta) \\
 &= \sum_{n=1}^N \log \sum_{i=1}^k p(z_n=i) \cdot p(x_n; \mu_i, \Sigma_i)
 \end{aligned}$$

Optimality condition taking $\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = 0$ no closed form solution

Expected log-likelihood

- $L(\theta) = E_{p(z_n|x_n)} \log p(x_n, z_n)$

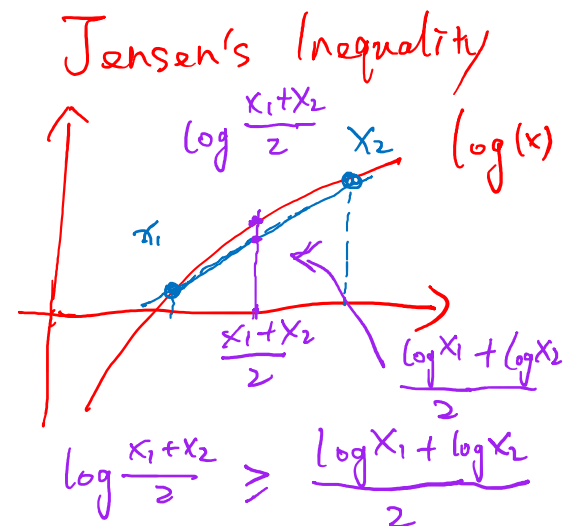
$$\mathcal{L}(\theta) = \sum_{n=1}^N \log \sum_{i=1}^K p(z_n=i) \cdot p(x_n|z_n=i)$$

$$= \sum_{n=1}^N \log \sum_{i=1}^K p(z_n=i|x_n) \cdot \frac{p(z_n=i) \cdot p(x_n|z_n=i)}{p(z_n=i|x_n)}$$

Jensen

$$\geq \sum_{n=1}^N \sum_{i=1}^K p(z_n=i|x_n) \cdot \log \frac{p(z_n=i) \cdot p(x_n|z_n=i)}{p(z_n=i|x_n)}$$

$$= \sum_{n=1}^N E_{z_n|x_n} [\log p(z_n=i) \cdot p(x_n|z_n=i) - \log p(z_n=i|x_n)]$$



General case:

$$\log E[X] \geq E[\log X]$$

Posterior

- $p(z_n | x_n) = \frac{p(z_n, x_n)}{p(x_n)} = \frac{p(z_n) \cdot p(x_n | z_n)}{\sum_{j=1}^K p(z_n=j) \cdot p(x_n | z_n=j)}$

$$\begin{aligned} \hat{z}_{n,i} = p(z_n=i | x_n) &= \frac{p(z_n=i) \cdot p(x_n | \mu_i, \Sigma_i)}{\sum_{j=1}^K p(z_n=j) \cdot p(x_n | \mu_j, \Sigma_j)} \\ &\approx \frac{\omega_i \cdot \mathcal{N}(x_n, \mu_i, \Sigma_i)}{\sum_{j=1}^K \omega_j \cdot \mathcal{N}(x_n, \mu_j, \Sigma_j)} \end{aligned}$$

Update mixture weights

$$\begin{aligned}
 L(\theta) &= \sum_{n=1}^N \sum_{i=1}^K p(z_n=i|x_n) \cdot \log \frac{p(z_n=i) \cdot p(x_n|z_n=i)}{p(z_n=i|x_n)} \\
 &= \sum_{n=1}^N \sum_{i=1}^K \hat{z}_{n,i} \log \frac{w_i \cdot N(x_n, \mu_i, \Sigma_i)}{\hat{z}_{n,i}} \\
 &= \sum_{n=1}^N \sum_{i=1}^K \hat{z}_{n,i} (\log w_i + \log N(x_n, \mu_i, \Sigma_i) - \log \hat{z}_{n,i})
 \end{aligned}$$

$$\text{s.t. } \sum_{j=1}^K w_j = 1$$

max $f(x)$
s.t. $g(x)=0$
Lagrangian $\text{Lag}(x)=f(x)-\lambda g(x)$

$$\text{Lag}(\theta) = L(\theta) - \lambda \left(\sum_{j=1}^K w_j - 1 \right)$$

Optimality for w :

$$\frac{\partial \text{Lag}(\theta)}{\partial w_i} = 0$$

$$\hookrightarrow \sum_{n=1}^N \hat{z}_{n,i} \cdot \frac{1}{w_i} - \lambda = 0$$

$$w_i = \frac{\sum_{n=1}^N \hat{z}_{n,i}}{\lambda}$$

$$\lambda = \sum_{j=1}^K \sum_{n=1}^N \hat{z}_{n,j}$$

$$w_i = \frac{\sum_{n=1}^N \hat{z}_{n,i}}{\sum_{j=1}^K \sum_{n=1}^N \hat{z}_{n,j}}$$

Update mean and covariance

$$\log \mathcal{N}(x_n, \mu_i, \Sigma_i) = -\frac{d}{2} \log 2\pi - \frac{1}{2} \log |\Sigma_i| - \frac{1}{2} (x_n - \mu_i)^T \Sigma_i^{-1} (x_n - \mu_i)$$

$$L(\theta) = \sum_{n=1}^N \sum_{i=1}^K \hat{z}_{n,i} \left[-\frac{1}{2} \log |\Sigma_i| - \frac{1}{2} (x_n - \mu_i)^T \Sigma_i^{-1} (x_n - \mu_i) \right] + \dots$$

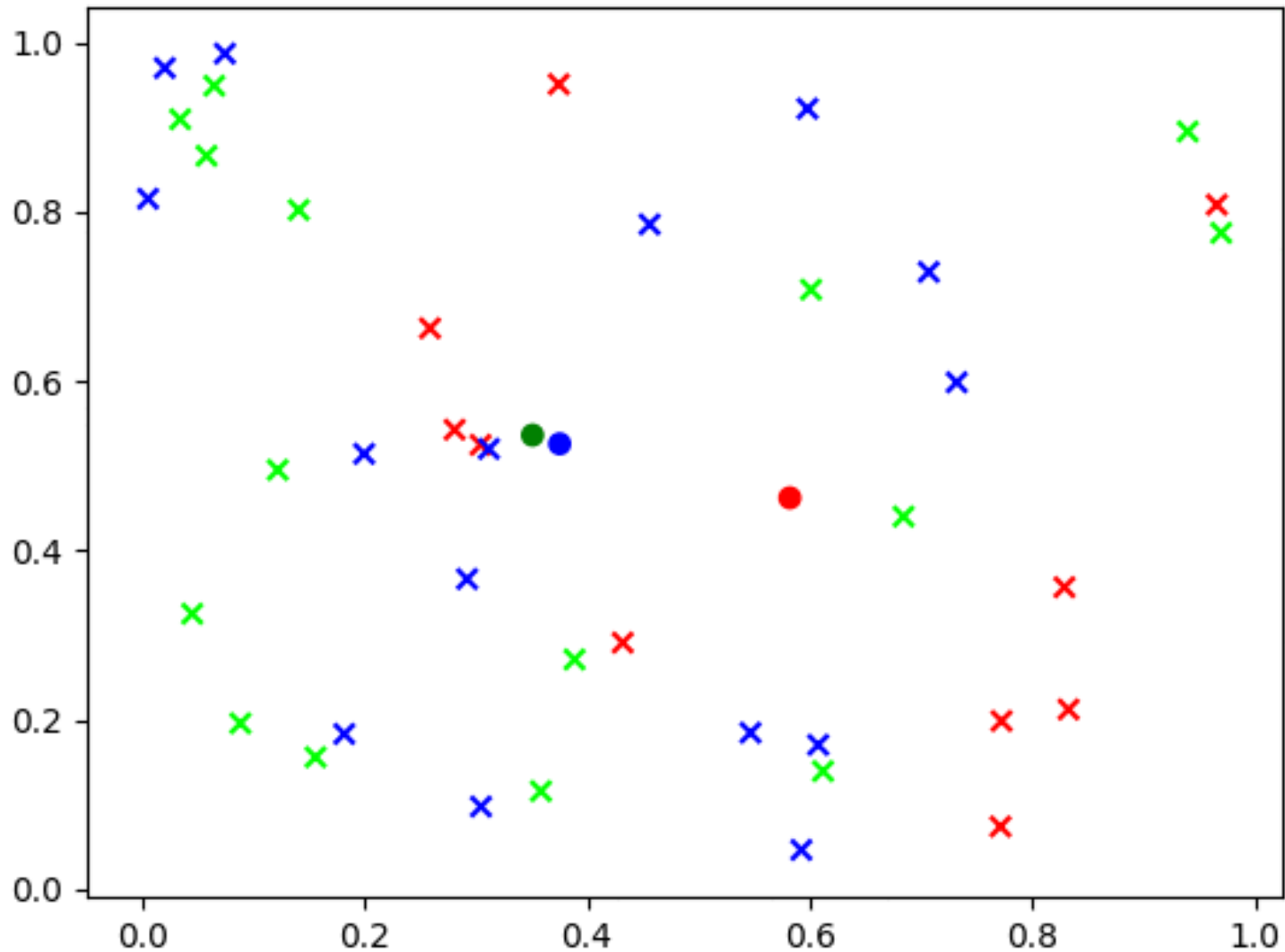
Optimality Condition

$$\frac{\partial L(\theta)}{\partial \mu_i} = \sum_{n=1}^N \sum_{i=1}^K \hat{z}_{n,i} \cdot \left(-\frac{1}{2} \cdot 2 \cdot \Sigma_i^{-1} \cdot (x_n - \mu_i) \right) = 0$$

Summary of EM algorithm

- Observation: $x_{1..N}$
- $\theta = \{w_{1..k}, \mu_{1..k}, \Sigma_{1..k}\}$
- Iterate until convergence
 1. E step: use X and current θ to calculate $p(z_{1..N}|x_{1..N}; \theta)$
 2. M step:
$$\theta \leftarrow \arg \max_{\theta} E_{p(z_{1..N}|x_{1..N}; \theta_{old})} \log p(x_n, z_n | \theta)$$
- Guaranteed to find local maximum
- Works for general mixture model

Illustration of GMM



Property of GMM

- Interpretable:
 - Participation weight of each data point from every component
- Generative:
 - Able to generate new data
- Handles missing values
- Efficient: $O(TKN)$
- Local optimal:
 - Can be viewed as coordinate descent (why?)
- Need to specify K

K-Means vs GMM

1. Decide on a value for K , the number of clusters.
2. Initialize the K cluster centers / parameters (randomly).

K-Means

GMM

3. Decide the class memberships of the N objects by assigning them to the nearest cluster center.

3. E-step: assign *probabilistic* membership

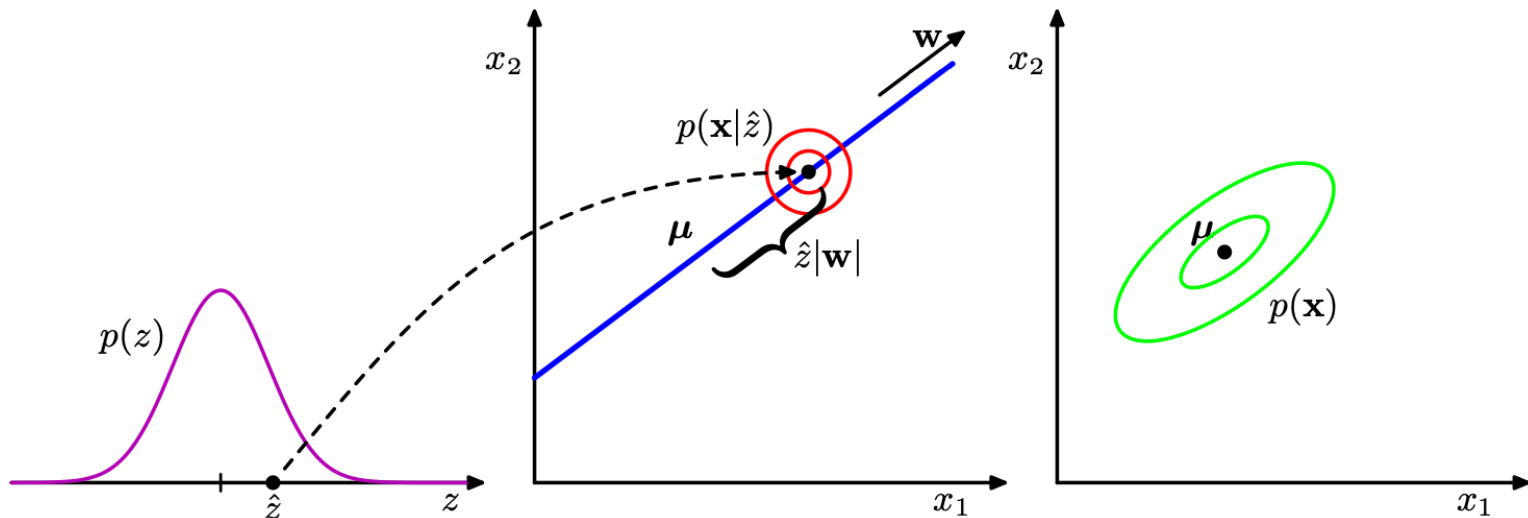
4. Re-estimate the K cluster centers using the memberships found above

4. M-step: re-estimate parameters based on *probabilistic* membership

5. Repeat 3 and 4 until parameters do not change.

Probabilistic PCA

- Continuous latent variable $z \sim N(0, I)$
- Observation data $x|z \sim N(W \cdot z + \mu, \sigma^2 I)$



Learning Parameters for PPCA

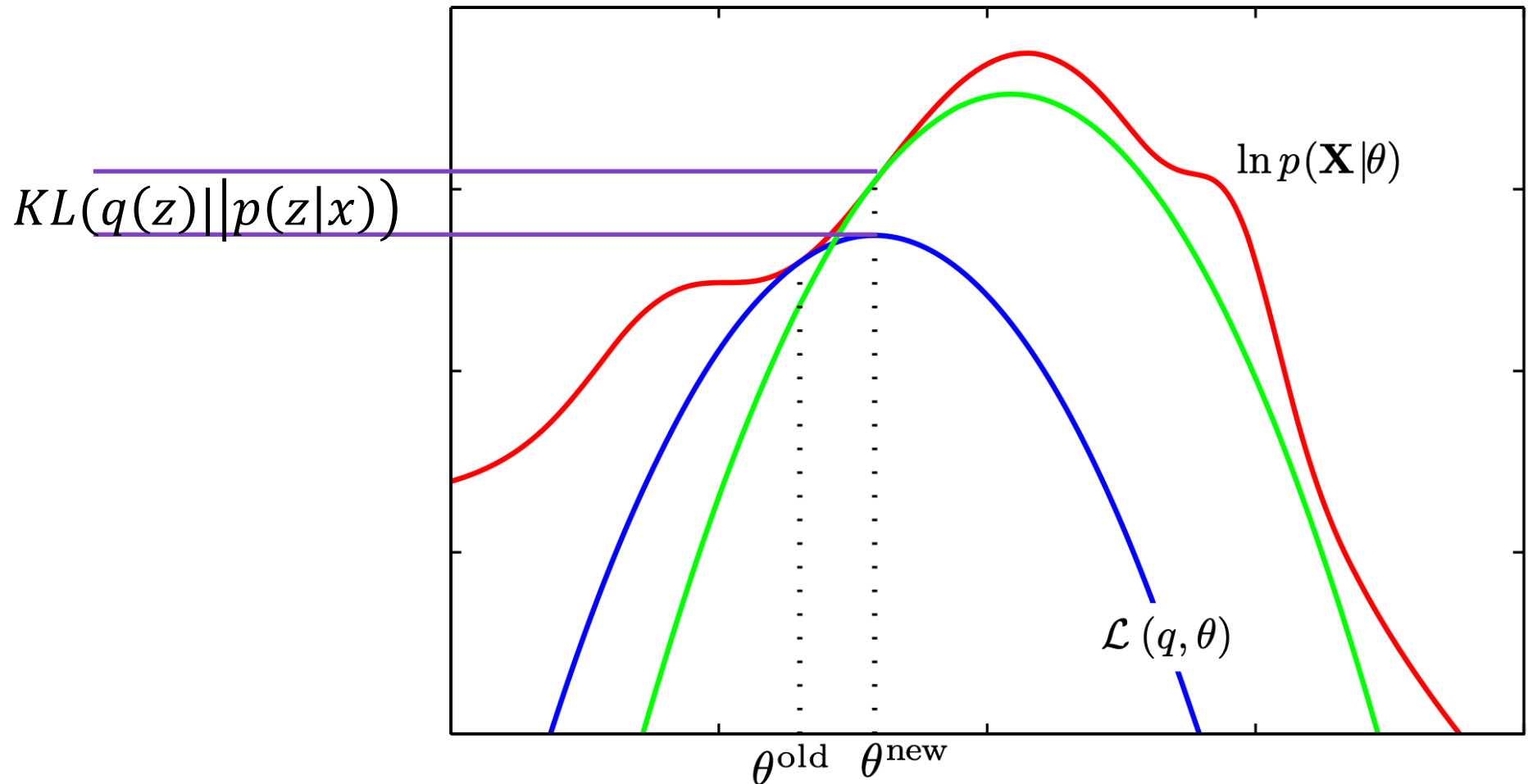
- Again EM algorithm
- $\arg \max_{\theta} E_{p(z_{1..N}|x_{1..N};\theta_{old})} \log p(x_{1..N}, z_{1..N}|\theta)$

A Variational View of EM

- $L(\theta) = \log p(X; \theta)$
- Introduce a variational distribution $q(X; \phi)$
- Variational bound for this data likelihood

What does EM actually do?

EM is coordinate-descent



Summary

- Mixture Distribution: to build more complex distribution from simple ones
- Gaussian Mixture Model: k Gaussian components
- Expectation-Maximization: general for graphical models with latent variables
 - E-step: fix parameter, estimate posterior mean/variance
 - M-step: update parameter
- Probabilistic PCA: latent is continuous

Recommended Reading

- PRML Chapter 9, 12.2

Next up

- Dynamic Bayesian Network
- Linear Dynamical System