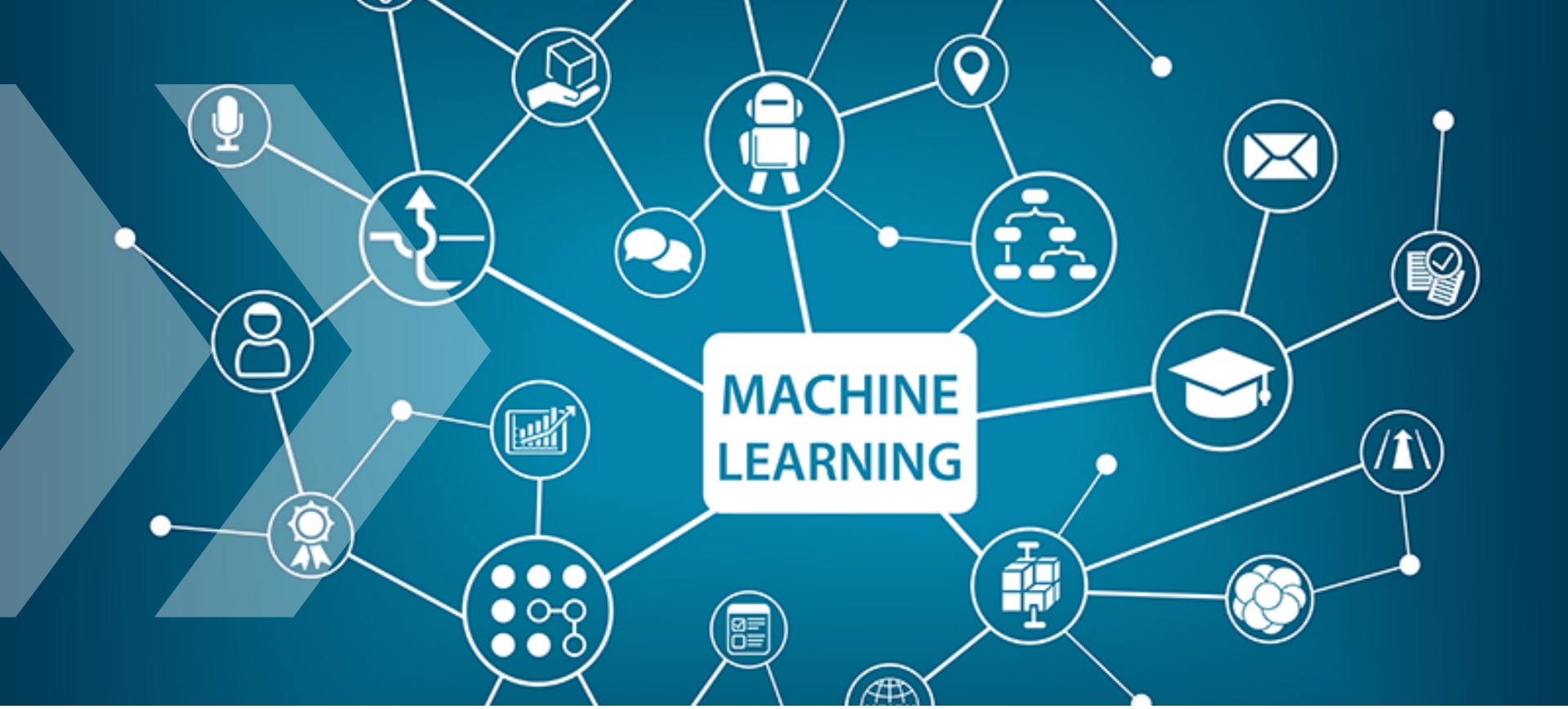
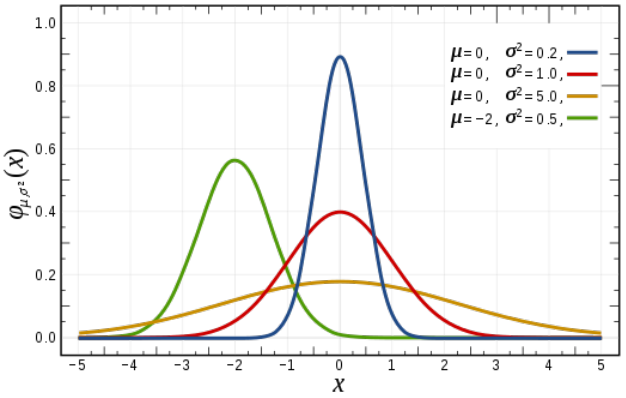


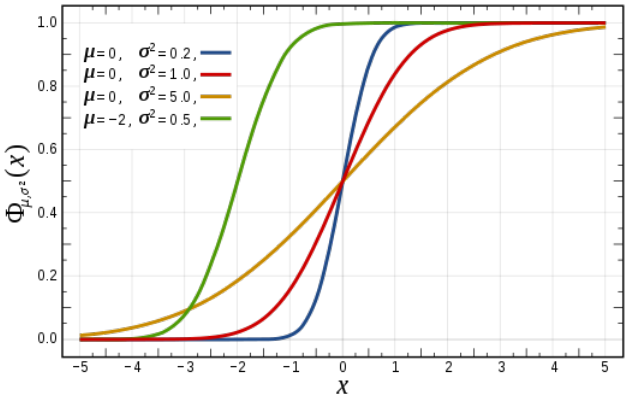
CS480/680: Intro to ML

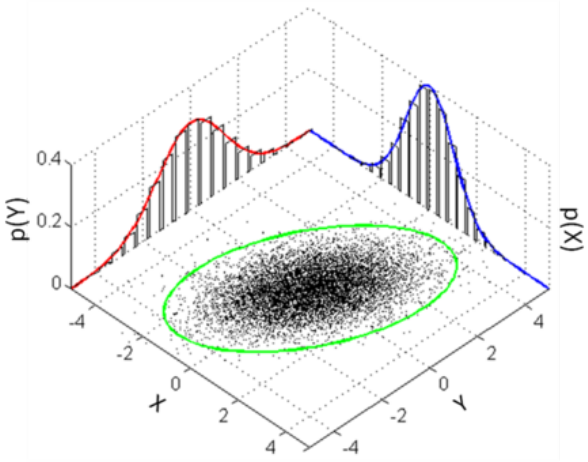
Lecture 16: Mixture Models

UNIVERSITY OF WATERLOO

Recap: Gaussian distribution



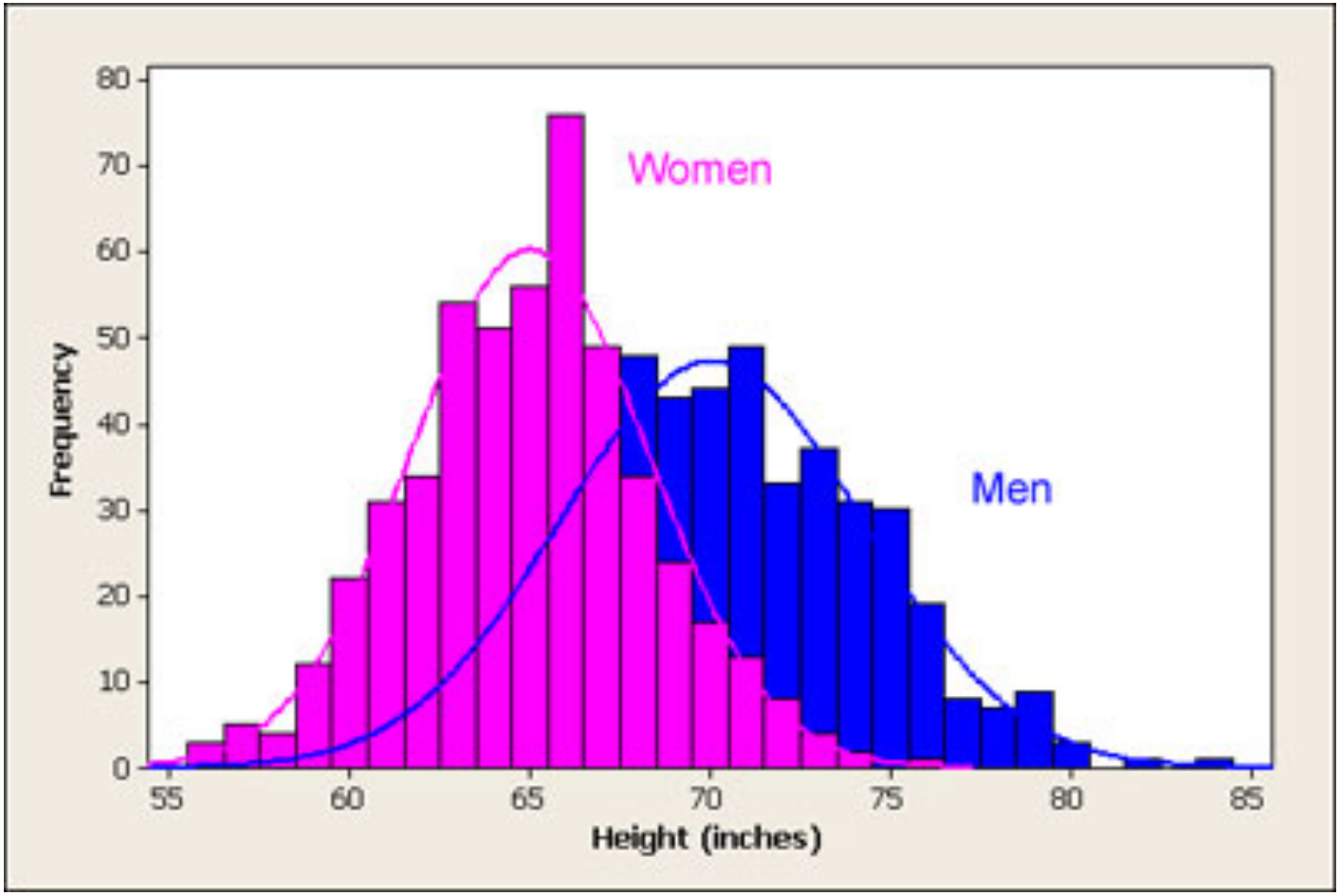




$$p(\mathbf{x}) = (2\pi)^{-d/2}|S|^{-1/2}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top S^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

UNIVERSITY OF WATERLOO

Multi-modality



UNIVERSITY OF WATERLOO

Mixture models

K → # of components

$$p(x|\theta) = \sum_{k=1}^{K} p(z=k) p(x|z=k,\theta)$$

$p(z=k) = \pi_k$: mixing distr.

$p(x|z=k,\theta) = p_k(x|\theta)$: k-th component distr.

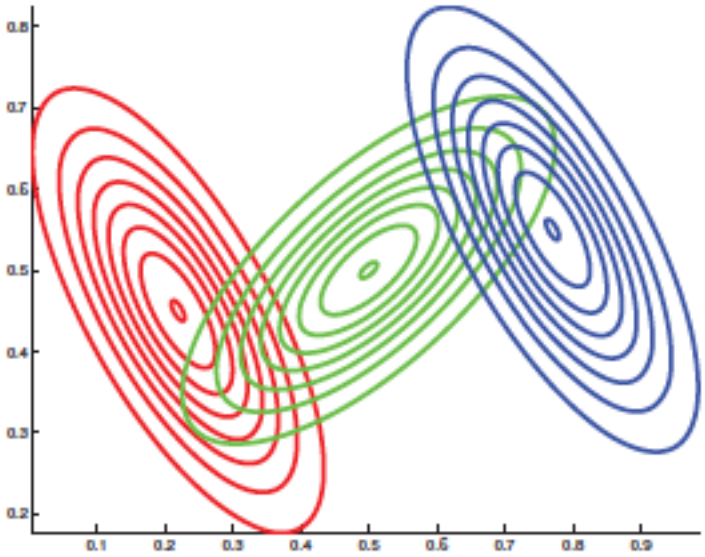
θ : parameters

$$\pi_k \geq 0, \sum_k \pi_k = 1$$

UNIVERSITY OF WATERLOO

Example: Gaussian Mixture Models

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, S_k)$$



(a)

(b)

UNIVERSITY OF WATERLOO

Universality

Theorem. GMM with sufficiently many components can approximate any probability density function on R^d.

- How many is many?
- Nothing special about Gaussian here, except computationally (later).

UNIVERSITY OF WATERLOO

Inference problem

$$p(x|\boldsymbol{\theta}) = \sum_{k=1}^{K} p(z=k)p(x|z=k,\boldsymbol{\theta})$$

latent (unobserved)

- Given iid sample $X_1, X_2, \ldots, X_n$ from $p(x|\theta)$
- Need to estimate θ
- Maximum likelihood is NP-hard…

UNIVERSITY OF WATERLOO

Soft clustering

$$p(z = k|\mathbf{x}, \boldsymbol{\theta}) = \frac{p(z = k|\boldsymbol{\theta})p(\mathbf{x}|z = k, \boldsymbol{\theta})}{\sum_{c=1}^{K} p(z = c|\boldsymbol{\theta})p(\mathbf{x}|z = c, \boldsymbol{\theta})}$$

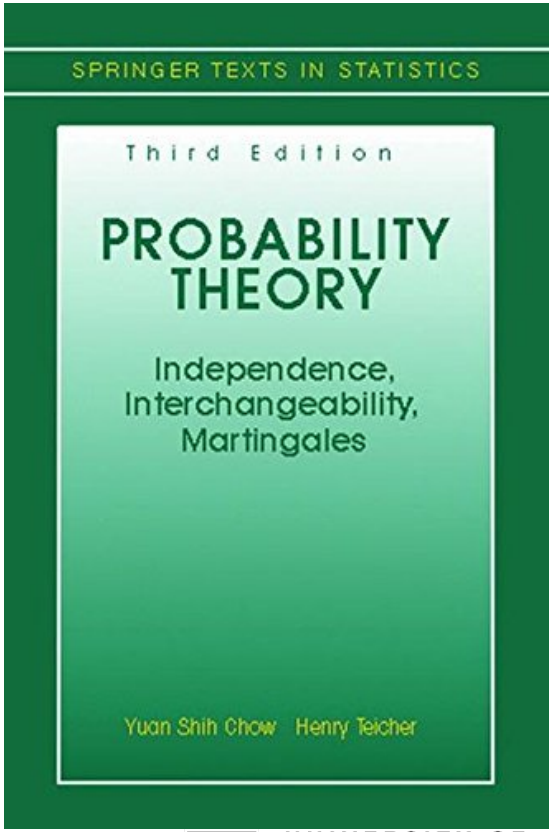
(Stauffer & Grimson, CVPR'98)

UNIVERSITY OF WATERLOO

Bigger issue: identifiability?

$$p(x|\boldsymbol{\theta}) = \sum_{k=1}^{K} p(z=k)p(x|z=k,\boldsymbol{\theta})$$

- Is this factorization even unique?
- Yes, for GMMs!



UNIVERSITY OF WATERLOO

The power of lifting

$$\min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|_1$$

- A nice trick: $2|t| = \min_s \; s^2 t^2 + 1/s^2$

$$\min_{\mathbf{w}} \min_{\mathbf{s}} \; \|\mathbf{s} \odot [X\mathbf{w} - \mathbf{y}]\|_2^2 + 1^\top (1/\mathbf{s}^2)$$

- Fix w: $s_i^2 = 1/|X_{i:}\mathbf{w} - y_i|$
- Fix s: $\mathbf{w} = [X^\top \mathrm{diag}(\mathbf{s}^2) X]^{-1} X^\top \mathrm{diag}(\mathbf{s}^2)\mathbf{y}$

UNIVERSITY OF WATERLOO

Variational form of Max Likelihood

$$\min_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) := \sum_{i=1}^{n} -\log p(\mathbf{x}_i|\boldsymbol{\theta}) \approx \mathrm{KL}(q(\mathbf{x})\|p(\mathbf{x}|\theta))$$

difficult to optimize

truth (unknown)

model

Lifting: $\min_{\boldsymbol{\theta}} \min_{q(\mathbf{z}|\mathbf{x})} \mathrm{KL}(q(\mathbf{x})q(\mathbf{z}|\mathbf{x})\|p(\mathbf{x},\mathbf{z}|\theta))$

instead of matching marginal, try to match *hypothetical* joint

UNIVERSITY OF WATERLOO

KL divergence

$$\mathrm{KL}(q(\mathbf{x})\|p(\mathbf{x})) := \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} \mathrm{d}\mathbf{x} \geq 0$$

- Both **p** and **q** are nonnegative and sum to 1
- Equality holds iff **p** == **q**
- Measures difference between distributions; asymmetric

Jensen's inequality
E(log(X)) <= log(E(X))

$$\mathrm{KL}(q(\mathbf{x}, \mathbf{z})\|p(\mathbf{x}, \mathbf{z})) = \mathrm{KL}(q(\mathbf{x})\|p(\mathbf{x})) + \mathbf{E}[\mathrm{KL}(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}|\mathbf{x}))]$$

UNIVERSITY OF WATERLOO

The EM algorithm

$$\min_{q_i(z_i)} \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left[\sum_{z_i} q_i(z_i) \log q_i(z_i) - \sum_{z_i} q_i(z_i) \log p(\mathbf{x}_i, z_i | \boldsymbol{\theta}) \right]$$

- Fix q, solve θ

$$\min_{\boldsymbol{\theta}} - \sum_{i=1}^{n} \sum_{z_i} q_i(z_i) \log p(\mathbf{x}_i, z_i | \boldsymbol{\theta})$$

often closed-form

- Fix θ, solve q

$$\min_{q_i(z_i) \geq 0, \sum_{z_i} q_i(z_i) = 1} - \sum_{z_i} q_i(z_i) \log p(\mathbf{x}_i, z_i | \boldsymbol{\theta}) + \sum_{z_i} q_i(z_i) \log q_i(z_i)$$

$$q_i(z_i) = p(z_i | \mathbf{x}_i, \boldsymbol{\theta})$$

UNIVERSITY OF WATERLOO

EM for GMM: step 1

$$\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, S_k)$$

$$\min_{r_{ik}\geq 0, \sum_k r_{ik}=1} \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left[\sum_{k=1}^{K} r_{ik} \log r_{ik} - \sum_{k=1}^{K} r_{ik} \log p(\mathbf{x}_i, z_i|\boldsymbol{\theta}) \right]$$

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} [-\log \pi_k + \tfrac{1}{2} \log |S_k| + \tfrac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^\top S_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)]$$

$$\pi_k = \frac{\sum_i r_{ik}}{n}$$

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{n} r_{ik} \mathbf{x}_i}{\sum_{i=1}^{n} r_{ik}}$$

$$S_k = \frac{\sum_{i=1}^{n} r_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top}{\sum_{i=1}^{n} r_{ik}} = \frac{\sum_{i=1}^{n} r_{ik} \mathbf{x}_i \mathbf{x}_i^\top}{\sum_{i=1}^{n} r_{ik}} - \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top$$

UNIVERSITY OF WATERLOO

EM for GMM: step 2

$$\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, S_k)$$

$$\min_{r_{ik}\geq 0, \sum_k r_{ik}=1} \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left[\sum_{k=1}^{K} r_{ik} \log r_{ik} - \sum_{k=1}^{K} r_{ik} \log p(\mathbf{x}_i, z_i|\boldsymbol{\theta}) \right]$$

$$r_{ik} = p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta}) \longrightarrow \text{posterior}$$

$$\propto p(z_i = k) \cdot p(\mathbf{x}_i|z_i = k, \boldsymbol{\theta})$$

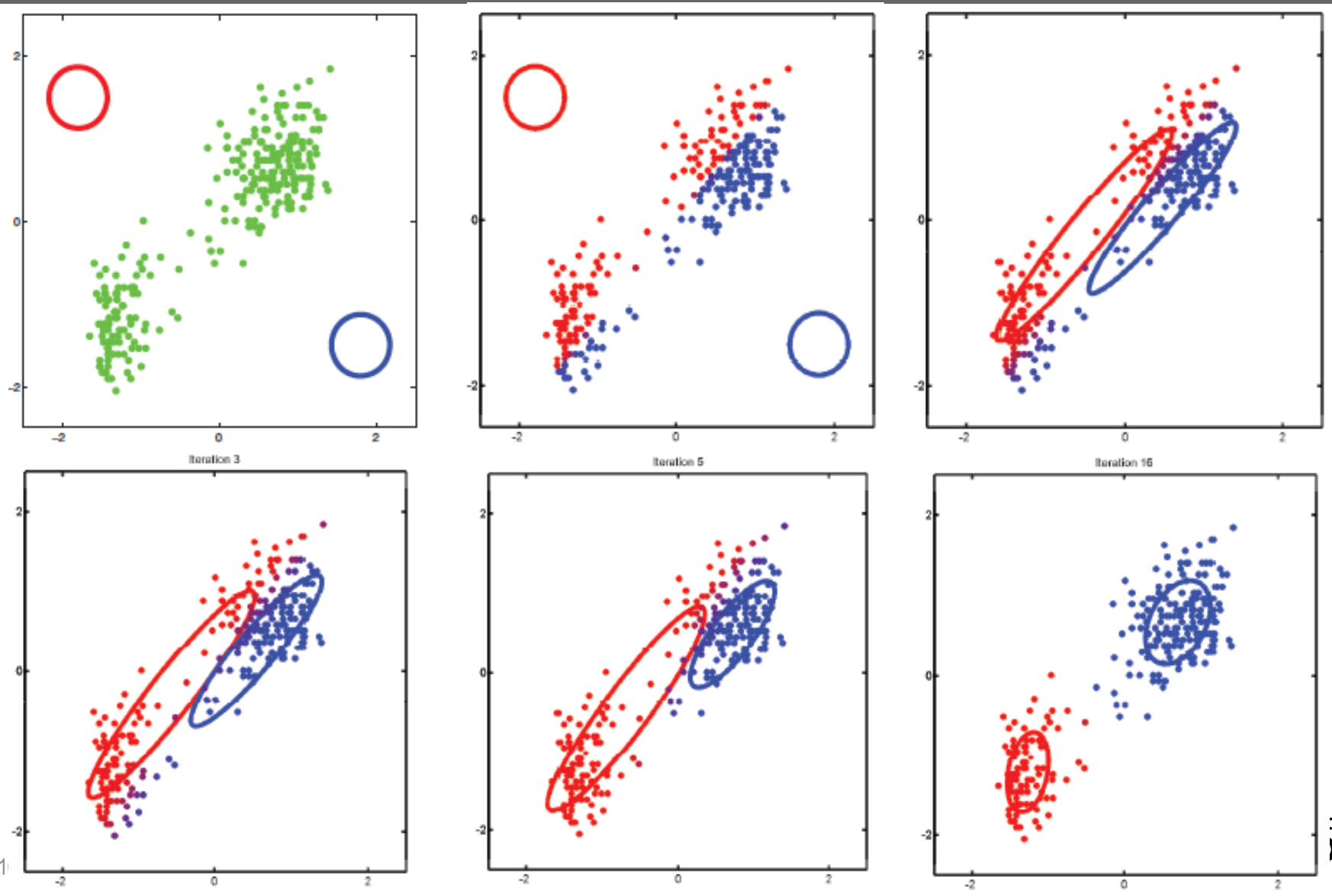
prior ← $p(z_i = k)$; $p(\mathbf{x}_i|z_i = k, \boldsymbol{\theta})$ → likelihood

$$= \pi_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, S_k)$$

$$r_{ik} = \frac{\pi_k |S_k|^{-1/2} \exp(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top S_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k))}{\sum_{c=1}^{K} \pi_c |S_c|^{-1/2} \exp(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_c)^\top S_c^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_c))}$$

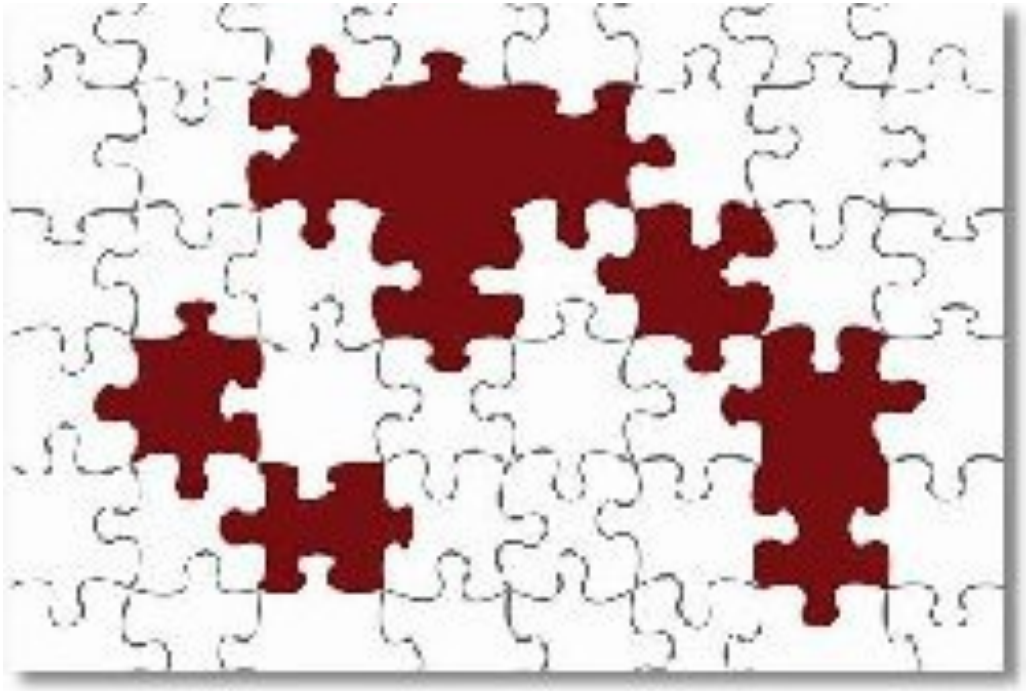
UNIVERSITY OF WATERLOO

Example



Other uses of EM

- Simplify computation
  - t-distribution as a Gaussian scale-mixture
- Missing data



UNIVERSITY OF WATERLOO

Mixture Density Network (Bishop'94)

$$\mathrm{p}(\mathbf{y}|\mathbf{x}) = \sum_{j=1}^{r} \lambda_j(\mathbf{x}) \mathrm{p}_j(\mathbf{y}|\mathbf{x}),$$

$$\mathcal{N}(\mathbf{y}|\mu_j(\mathbf{x}), \sigma_j(\mathbf{x}))$$

$$\boldsymbol{\lambda}(\mathbf{x}) = \mathtt{softmax}(\mathbf{g}_1(\mathbf{x}; \mathbf{w}))$$
$$\boldsymbol{\mu}(\mathbf{x}) = \mathbf{g}_2(\mathbf{x}; \mathbf{w})$$
$$\boldsymbol{\sigma}(\mathbf{x}) = \exp(\mathbf{g}_3(\mathbf{x}; \mathbf{w})).$$

UNIVERSITY OF WATERLOO

Mixture of Experts

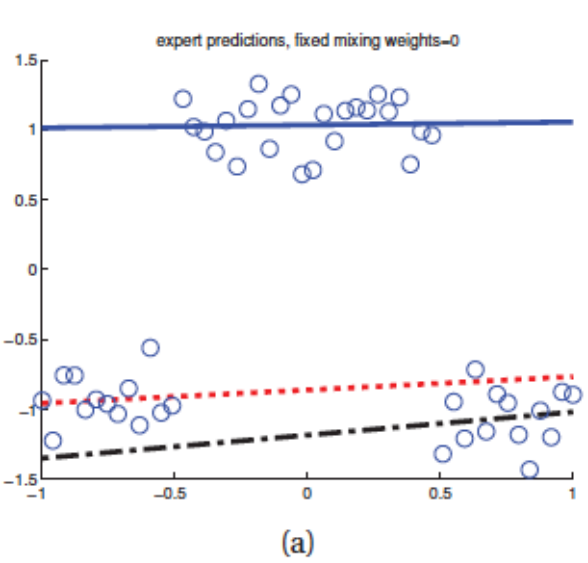
$$\frac{\exp(\mathbf{x}^\top \mathbf{v}_k)}{\sum_{c=1}^{K} \exp(\mathbf{x}^\top \mathbf{v}_c)}$$

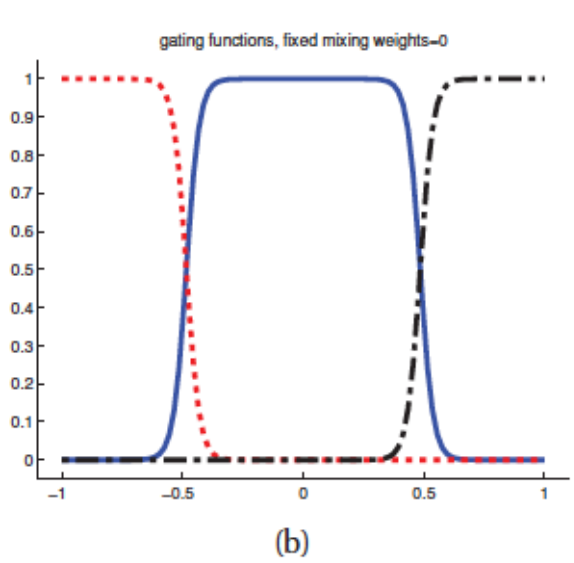
K mixing distr.

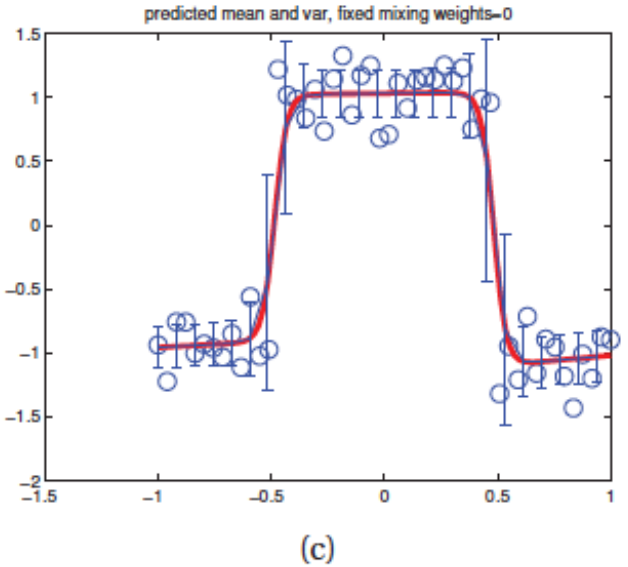
$$\mathcal{N}(y|\mathbf{w}_k^\top \mathbf{x}, \sigma_k^2)$$

k-th component distr.

$$p(y|\mathbf{x}, \boldsymbol{\theta}) = \sum_{k=1}^{K} p(z = k|\mathbf{x}, \boldsymbol{\theta}) p(y|\mathbf{x}, z = k, \boldsymbol{\theta})$$







Questions?