Learning From Data
Lecture 13: Unsupervised Learning IV

Yang Li **yangli@sz.tsinghua.edu.cn**

December 17, 2021

Today's Lecture

Unsupervised Learning (Part IV)

- Mixture of Gaussians
- The EM Algorithm
- Factor Analysis

Review: k-means clustering

Given input data $\{x^{(1)}, \dots, x^{(m)}\}$, $x^{(i)} \in \mathbb{R}^d$, **k-means clustering** partition the input into $k \le m$ sets $C_1, \dots, C_k$ to minimize the within-cluster sum of squares (WCSS).

$$\underset{C}{\operatorname{argmin}} \underline{\sum_{j=1}^{k} \sum_{x \in C_j} \|x - \mu_j\|^2}$$

Review: k-means clustering

Given input data $\{x^{(1)}, \dots, x^{(m)}\}$, $x^{(i)} \in \mathbb{R}^d$, **k-means clustering** partition the input into $k \leq m$ sets $C_1, \dots, C_k$ to minimize the within-cluster sum of squares (WCSS).

$$\underset{C}{\operatorname{argmin}} \sum_{j=1}^{k} \sum_{x \in C_j} \|x - \mu_j\|^2$$

Lloyd's Algorithm (1957,1982)

Let $c^{(i)} \in \{1, \dots, k\}$ be the cluster label for $x^{(i)}$

```
Initialize cluster centroids μ_1,...μ_k ∈ R^n randomly
Repeat until convergence{
  For every i,
    c^(i) := argmin_j ||x^(i) − μ_j||^2

  For each j
      μ_j := Σ_{i=1}^m 1{c^(i)=j} x^(i) / Σ_{i=1}^m 1{c^(i)=j}
}
```

$$c^{(i)} := \operatorname{argmin}_j \|x^{(i)} - \mu_j\|^2$$

$$\mu_j := \frac{\sum_{i=1}^m \mathbf{1}\{c^{(i)}=j\} x^{(i)}}{\sum_{i=1}^m \mathbf{1}\{c^{(i)}=j\}}$$

Demo:http://stanford.edu/class/ee103/visualizations/kmeans/kmeans.html

Lloyd, Stuart P. (1982). "Least squares quantization in PCM". IEEE Transactions on Information Theo

Review: k-means clustering

Given input data $\{x^{(1)}, \ldots, x^{(m)}\}$, $x^{(i)} \in \mathbb{R}^d$, **k-means clustering** partition the input into $k \leq m$ sets $C_1, \ldots, C_k$ to minimize the within-cluster sum of squares (WCSS).

$$\underset{C}{\operatorname{argmin}} \sum_{j=1}^{k} \sum_{x \in C_j} \|x - \mu_j\|^2$$

Lloyd's Algorithm (1957,1982)

Let $c^{(i)} \in \{1, \ldots, k\}$ be the cluster label for $x^{(i)}$

```
Initialize cluster centroids μ_1,...μ_k ∈ R^n randomly
Repeat until convergence{
  For every i,
    c^(i) := argmin_j ||x^(i) − μ_j||^2   ← assign x^(i) to the cluster
                                            with the closest centroid
  For each j
    μ_j := Σ_{i=1}^m 1{c^(i)=j} x^(i) / Σ_{i=1}^m 1{c^(i)=j}
}
```

$$c^{(i)} := \operatorname{argmin}_j \|x^{(i)} - \mu_j\|^2 \qquad \mu_j := \frac{\sum_{i=1}^{m} \mathbf{1}\{c^{(i)}=j\}x^{(i)}}{\sum_{i=1}^{m} \mathbf{1}\{c^{(i)}=j\}}$$

Demo:http://stanford.edu/class/ee103/visualizations/kmeans/kmeans.html

Lloyd, Stuart P. (1982). "Least squares quantization in PCM". IEEE Transactions on Information The

Review: k-means clustering

Given input data $\{x^{(1)}, \dots, x^{(m)}\}$, $x^{(i)} \in \mathbb{R}^d$, **k-means clustering** partition the input into $k \le m$ sets $C_1, \dots, C_k$ to minimize the within-cluster sum of squares (WCSS).

$$\underset{C}{\operatorname{argmin}} \sum_{j=1}^{k} \sum_{x \in C_j} \|x - \mu_j\|^2$$

Lloyd's Algorithm (1957,1982)

Let $c^{(i)} \in \{1, \dots, k\}$ be the cluster label for $x^{(i)}$

Initialize cluster centroids $\mu_1, \dots \mu_k \in R^n$ randomly

Repeat until convergence{

For every i,

$c^{(i)} := \operatorname{argmin}_j \|x^{(i)} - \mu_j\|^2$ ← assign $x^{(i)}$ to the cluster with the closest centroid

For each j

$$\mu_j := \frac{\sum_{i=1}^{m} \mathbf{1}\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} \mathbf{1}\{c^{(i)} = j\}}$$ ← update centroid

}

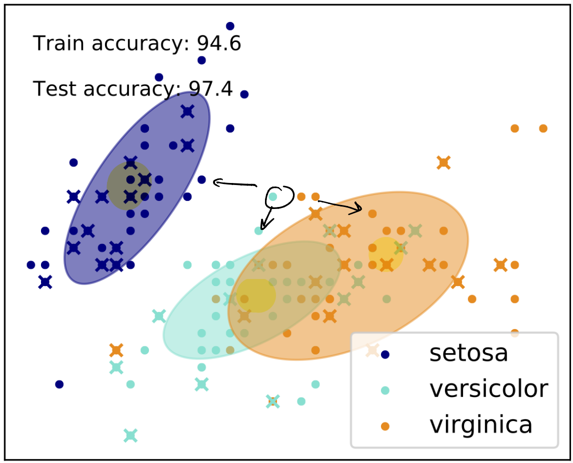
Demo:http://stanford.edu/class/ee103/visualizations/kmeans/kmeans.html

Mixture of Gaussians

Mixture of Gaussians

(GMM)

A "soft" version of k-means clustering.



Clustering results of iris dataset using *mixture of Gaussians*

Mixture models

hidden. observed

ϕ. — $z \longrightarrow x$.

Model-based clustering

A **mixture model** assumes data are generated by the following process:

1. Sample $z^{(i)} \in \{1, \ldots, k\}$ and $z^{(i)} \sim \text{Multinomial}(\phi)$

$$p(z^{(i)} = j) = \phi_j \text{ for all } j$$

$$\begin{bmatrix} \phi_1 \\ \vdots \\ \phi_k \end{bmatrix}$$

   $z^{(i)}$ are called **latent variables**.

2. Sample observables $x^{(i)}$ from some distribution $p(x^{(i)}, z^{(i)})$:

$$p(x^{(i)}, z^{(i)}) = p(x^{(i)}|z^{(i)})p(z^{(i)})$$

Mixture models

Model-based clustering

A **mixture model** assumes data are generated by the following process:

1. Sample $z^{(i)} \in \{1, \ldots, k\}$ and $z^{(i)} \sim \text{Multinomial}(\phi)$

$$p(z^{(i)} = j) = \phi_j \text{ for all } j$$

   $z^{(i)}$ are called **latent variables**.

2. Sample observables $x^{(i)}$ from some distribution $p(x^{(i)}, z^{(i)})$:

$$p(x^{(i)}, z^{(i)}) = p(x^{(i)}|z^{(i)})p(z^{(i)})$$

Bernoulli(ϕ)

Examples:

- Unsupervised handwriting recognition is a mixture with 10 Bernoulli distributions

Mixture models

Model-based clustering

A **mixture model** assumes data are generated by the following process:

1. Sample $z^{(i)} \in \{1, \ldots, k\}$ and $z^{(i)} \sim \text{Multinomial}(\phi)$

$$p(z^{(i)} = j) = \phi_j \text{ for all } j$$

   $z^{(i)}$ are called **latent variables**.

2. Sample observables $x^{(i)}$ from some distribution $p(x^{(i)}, z^{(i)})$:

$$p(x^{(i)}, z^{(i)}) = p(x^{(i)}|z^{(i)})p(z^{(i)})$$

Examples:

- Unsupervised handwriting recognition is a mixture with 10 Bernoulli distributions
- Financial return estimation uses a mixture of 2 Gaussians for normal situation and crisis time distribution

Mixture of Gaussians

Mixture of Gaussians Model:

$$z^{(i)} \sim \text{Multinomial}(\phi)$$
$$x^{(i)}|z^{(i)} = j \sim \mathcal{N}(\mu_j, \Sigma_j)$$

How to learn ϕ_j, μ_j and Σ_j for all j ?

$z^{(i)}$ is known: $(x^{(i)}, z^{(i)})_{i=1}^{m}$

$z^{(i)}$ is unknown:

Mixture of Gaussians

Mixture of Gaussians Model:

$$z^{(i)} \sim \text{Multinomial}(\phi)$$

$$x^{(i)}|z^{(i)} = j \sim \mathcal{N}(\mu_j, \Sigma_j)$$

Σ

How to learn ϕ_j, μ_j and Σ_j for all j ?

$z^{(i)}$ is known: (supervised) use maximum likelihood estimation (quadratic discriminant analysis). QDA $\boxed{\text{GDA}}$

$$\phi_j = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{z^{(i)} = j\}, \quad \mu_j = \frac{\sum_{i=1}^{m} \mathbf{1}\{z^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} \mathbf{1}\{z^{(i)} = j\}}$$

$$\Sigma_j = \frac{\sum_{i=1}^{m} \mathbf{1}\{z^{(i)} = j\}(x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} \mathbf{1}\{z^{(i)} = j\}}$$

$z^{(i)}$ is unknown:

Mixture of Gaussians

Mixture of Gaussians Model:

$$z^{(i)} \sim \text{Multinomial}(\phi)$$
$$x^{(i)}|z^{(i)} = j \sim \mathcal{N}(\mu_j, \Sigma_j)$$

How to learn ϕ_j, μ_j and Σ_j for all j ?

$z^{(i)}$ is known: (supervised) use maximum likelihood estimation (quadratic discriminant analysis).

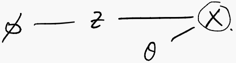
$$\phi_j = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{z^{(i)} = j\}, \quad \mu_j = \frac{\sum_{i=1}^{m} \mathbf{1}\{z^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} \mathbf{1}\{z^{(i)} = j\}}$$

$$\Sigma_j = \frac{\sum_{i=1}^{m} \mathbf{1}\{z^{(i)} = j\}(x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} \mathbf{1}\{z^{(i)} = j\}}$$

$z^{(i)}$ is unknown: (unsupervised) use **expectation maximization**

Expectation Maximization

The EM Algorithm



The EM algorithm is an iterative method for maximum likelihood estimation when the model depends on **latent (unobserved) variables**.

Log-likelihood of data:

$$l(\theta) = \sum_{i=1}^{m} \log p(x^{(i)}; \theta) = \sum_{i=1}^{m} \log \sum_{z^{(i)}=1}^{k} p(x^{(i)}, z^{(i)}; \theta)$$

Main idea: iterate over two steps:

- Expectation (E) step : guess $z^{(i)}$ for each $x^{(i)}$.
- Maximization (M) step : update θ via maximum likelihood estimation based on guessed $z^{(i)}$'s

Generalized EM Algorithm

Listing 1: Generalized EM Algorithm

```
Initialize θ
Repeat untill convergence {
  (E-step) For each i , set        guess z^i given x^i.
        Q_i(z^(i)) := p(z^(i)|x^(i); θ) ← Soft assignment:
  posterior distribution z|x under θ
  (M-step) Set
```

$$\theta := \arg\max_{\theta} \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \qquad (\star)$$

```
        ← Update parameter θ
}
```

Generalized EM Algorithm

Listing 2: Generalized EM Algorithm

```
Initialize θ
Repeat untill convergence {
  (E-step) For each i , set
        Q_i(z^(i)) := p(z^(i)|x^(i); θ) ← Soft assignment:
   posterior distribution z|x under θ
  (M-step) Set
        θ := argmax_θ Σ_i Σ_{z^(i)} Q_i(z^(i)) log [ p(x^(i), z^(i); θ) / Q_i(z^(i)) ]    (⋆)
        ← Update parameter θ
}
```

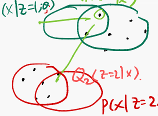
We will show...

- Solving $(\star)$ is equivalent to $\underbrace{\text{argmax}_\theta}\, l(\theta)$
  $\rightarrow$ Equation $(\star)$ is a (tight) lower bound on log-likelihood $\underline{l(\theta)}$

Generalized EM Algorithm

$z^{(i)} \in \{1, 2\}$

$\theta_j = \{\mu_j, \Sigma_j\}$



Listing 3: Generalized EM Algorithm

Initialize θ

Repeat untill convergence {

(E-step) For each i , set

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)}; \theta)$$ ← Soft assignment: posterior distribution $z|x$ under θ

(M-step) Set

$J(Q, \theta)$

$$\theta := \arg\max_{\theta} \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \qquad (\star)$$

← Update parameter θ

}

We will show...

- Solving $(\star)$ is equivalent to $\arg\max_\theta l(\theta)$
  → Equation $(\star)$ is a (tight) lower bound on log-likelihood $l(\theta)$
- This algorithm converges.

Proof of Correctness: E-step

For each i, let $Q_i(z)$ be a distribution of z:

$$\sum_{z \in \mathcal{Z}} Q_i(z) = 1, Q_i(z) \geq 0$$

Define

$$J(Q, \theta) = \sum_{i=1}^{m} \sum_{z^{(i)} \in \mathcal{Z}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$$

Proposition 1

1. *$J(Q, \theta)$ is a lower bound on log-likelihood $l(\theta)$*
2. *This lower bound is tight when $Q_i(z^{(i)}) = p(z^{(i)} | x^{(i)}; \theta)$*

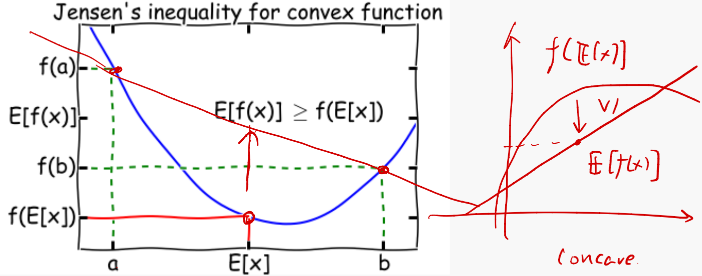
(Hint: use Jensen's inequality)

Jensen's Inequality

Theorem 1

Let f be a **convex** *function, and let X be a random variable. Then*

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

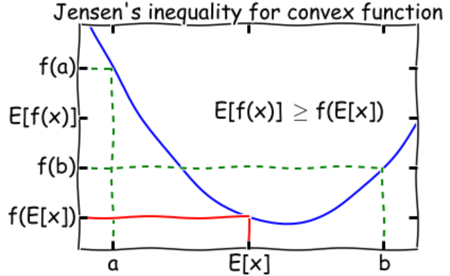


Jensen's Inequality

Theorem 1

*Let f be a **convex** function, and let X be a random variable. Then*

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$



Remarks

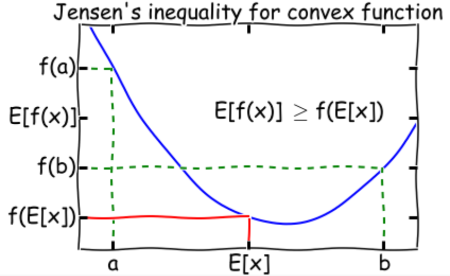
1. Let f be a **concave** function, then $\mathbb{E}[f(X)] \leq f(E[X])$

Jensen's Inequality

Theorem 1

Let f be a **convex** *function, and let X be a random variable. Then*

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$



Remarks

1. Let f be a **concave** function, then $\mathbb{E}[f(X)] \leq f(E[X])$
2. When $f(X)$ is a constant function, $\mathbb{E}[f(X)] = f(\mathbb{E}[X])$

Proof of Correctness

Proposition 1

1. $J(Q, \theta)$ *is a lower bound on log-likelihood* $l(\theta)$

proof. $l(\theta) = \sum_{i=1}^{m} \log \sum_{z^{(i)}} p(x^{(i)}; z^{(i)}; \theta) \frac{Q_i(z^{(i)})}{Q_i(z^{(i)})}$

$$= \sum_{i=1}^{m} \log \underbrace{\sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)} z^{(i)}; \theta)}{Q_i(z^{(i)})}}$$

$$= \sum_{i=1}^{m} \log \mathbb{E}_{z^{(i)} \sim Q_i} \left[\frac{p(x^{(i)}; z^{(i)}; \theta)}{Q_i(z^{(i)})} \right].$$

By Jenson's inequality,

$$\geq \sum_{i=1}^{m} \mathbb{E}_{z^i \sim Q_i} \left(\log \left[\frac{p(x^{(i)}; z^{(i)}; \theta)}{Q_i(z^{(i)})} \right] \right) = \underbrace{\sum_{i=1}^{m} \sum_{z^{(i)} \in z} Q_i(z^{(i)}) \log \frac{p(x^{(i)}; z^{(i)}; \theta)}{Q_i(z^{(i)})}}_{J(Q, \theta).}$$

$J(Q, \theta)$ is a lower bound of $l(\theta)$

Proof of Correctness

$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta)$.

Proposition 1

1. $J(Q,\theta)$ is a lower bound on log-likelihood $l(\theta)$ posterior distribution of $z^{(i)}$ given $x^{(i)}$
2. This lower bound is tight when $Q_i(z^{(i)}) = p(z^{(i)}|x^{(i)};\theta)$ (E-step)

proof. Suppose $\frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} = c$ for some constant c.

$$Q_i(z^{(i)}) = \frac{p(x^{(i)}, z^{(i)};\theta)}{c}.$$

Since $\sum_{j=1}^{K} Q_i(z^{(i)}=j) = 1$, $\sum_{j=1}^{K} \frac{p(x^{(i)}, z^{(i)}=j;\theta)}{c}$ ..

$$c = \sum_{j=1}^{K} p(x^{(i)}, z^{(i)}=j;\theta) = p(x^{(i)};\theta)$$

$$Q_i(z^{(i)}) = \frac{p(x^{(i)}, z^{(i)};\theta)}{p(x^{(i)};\theta)} = p(z^{(i)}|x^{(i)};\theta). \quad \square.$$

Proof of Convergence

Proposition 2

EM always monotonically improves the log likelihood, i.e. Let $\theta^{(t)}$ be the parameter value in the t-th iteration

$$l(\theta^{(t)}) \leq l(\theta^{(t+1)})$$

proof. In proposition 1, $Q^{(t)} = P(z^{(i)}|x^{(i)};\theta^{t})$, $J(Q^{t},\theta^{t}) = \ell(\theta^{t})$

In the M-step,

$$\theta^{t+1} := \arg\max_{\theta} J(Q^{(t)},\theta)$$

Then $J(Q^{(t)},\theta^{(t+1)}) \geq J(Q^{(t)},\theta^{(t)})$

By proposition 1,

$$l(\theta^{(t+1)}) \geq J(Q^{(t)},\theta^{(t+1)}) \geq J(Q^{(t)},\theta^{(t)}) = l(\theta^{(t)})$$

$\square$.

EM for mixture of Gaussians

Gaussian Mixture Model

$$z^{(i)} \sim \text{Multinomial}(\phi)$$

$$x^{(i)}|z^{(i)} \sim \mathcal{N}(\mu_j, \Sigma_j) \leftarrow \text{no } \phi$$

Learn parameters μ, Σ, ϕ

E-Step: $w_j^{(i)} = Q_i(z^{(i)} = j) = p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$

By Bayes Rule $= \dfrac{p(x^{(i)}|z^{(i)}=j; \mu, \Sigma)\, p(z^{(i)}=j; \phi)}{\underbrace{p(x^{(i)}; \mu, \Sigma)}}$

$= \sum_{l=1}^{k} p(x^{(i)}|z^{(i)}=l; \mu, \Sigma)\, p(z^{(i)}=l; \phi)$

EM for mixture of Gaussians

Gaussian Mixture Model

$$z^{(i)} \sim \text{Multinomial}(\phi)$$
$$x^{(i)}|z^{(i)} \sim \mathcal{N}(\mu_j, \Sigma_j)$$

let. $\nabla_{\Sigma_l} J(\phi, \mu, \Sigma) = 0$.

$\Rightarrow \Sigma_l^* = \dfrac{\sum_{i=1}^m w_l^{(i)}(x^{(i)}-\mu)(x^{(i)}-\mu)^T}{\sum_{i=1}^m w_j^{(i)}}$

Learn parameters μ, Σ, ϕ

E-Step: $w_j^{(i)} = Q_i(z^{(i)} = j) = p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$

M-Step: Maximize

$\underbrace{J(\phi, \mu, \Sigma)} = \sum_{i=1}^m \sum_{j=1}^k \underbrace{Q_i(z^{(i)} = j)}_{w_j^{(i)}} \log \dfrac{p(x^{(i)}, z^{(i)} = j; \phi, \mu, \Sigma)}{\underbrace{Q_i(z^{(i)} = j)}_{w_j^{(i)}}}$ with respect to ϕ, μ and Σ

Assume $w_j^{(i)} = Q_i(z^{(i)}=j)$ is given

$$J(\phi, \mu, \Sigma) = \sum_{i=1}^m \sum_{j=1}^k w_j^{(i)} \log \frac{1}{w_j^{(i)}} \left(\frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_j|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x^{(i)}-\mu_j)^T \Sigma_j^{-1}(x^{(i)}-\mu_j)\right) \phi_j \right)$$

$$= \sum_{i=1}^m \sum_{j=1}^k w_j^{(i)} \left(-\log\left((2\pi)^{\frac{n}{2}} |\Sigma_j|^{\frac{1}{2}}\right) - \frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j) - \log w_j^{(i)} + \log \phi_j \right)$$

$$\nabla_{\mu_l} J(\phi, \mu, \Sigma) = \sum_{i=1}^m w_l^{(i)} (\Sigma_l^{-1} x^{(i)} - \Sigma_l^{-1} \mu_l) = 0. \quad \mu_l^* = \frac{\sum_{i=1}^m w_l^{(i)} x^{(i)}}{\sum_{i=1}^m w_l^{(i)}}$$

$$\underset{\phi}{\operatorname{argmax}} \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(i)} \log \phi_j$$

$$\text{s.t.} \quad \sum_{j=1}^{k} \phi_j = 1 \;\Rightarrow\; \sum_{j=1}^{k} \phi_j - 1 = 0$$

$$L(\phi, \beta) = \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(i)} \log \phi_j + \beta\left(\sum_{j=1}^{k} \phi_j - 1\right)$$

$$\frac{\partial}{\partial \phi_l} L(\phi, \beta) = \frac{\partial}{\partial \phi_l} \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(i)} \log \phi_j + \frac{\partial}{\partial \phi_l} \beta\left(\sum_{j=1}^{k} \phi_j - 1\right)$$

$$= \sum_{i=1}^{m} \frac{\partial}{\partial \phi_l} w_l^{(i)} \log \phi_l + \beta \frac{\partial}{\partial \phi_l} \phi_l$$

$$\sum_{i=1}^{m} \frac{w_l^{(i)}}{\phi_l} + \beta = 0$$

$$\phi_l = -\frac{1}{\beta} \sum_{i=1}^{m} w_l^{i}$$

$$\frac{\partial}{\partial \beta} L(\phi, \beta) = 0$$

$$\Rightarrow \quad \sum_{l=1}^{k} \phi_l = 1$$

$$-\sum_{l=1}^{k} \sum_{i=1}^{m} \frac{w_l^{(i)}}{\beta} = 1$$

$$-\sum_{i=1}^{m} \underbrace{\sum_{l=1}^{k} w_l^{(i)}}_{1} = \beta$$

$$-m = \beta$$

$$\phi_l^* = -\frac{1}{(-m)} \sum_{i=1}^{m} w_l^{(i)} = \frac{1}{m} \sum_{i=1}^{m} w_l^{(i)}$$

Expectation Maximization for Gaussian Mixtures

Listing 4: EM for Gaussian Mixtures

Repeat untill convergence {

(E-step) For each i,j , set

$$w_j^{(i)} := p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$$

(M-step) Update parameters: assume $\phi_j = \mathbb{E}[w_j]$

$$\begin{cases} \phi_j := \dfrac{1}{m} \displaystyle\sum_{i=1}^{m} w_j^{(i)} \\ \mu_j := \dfrac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}} \\ \Sigma_j := \dfrac{\sum_{i=1}^{m} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} w_j^{(i)}} \end{cases}$$

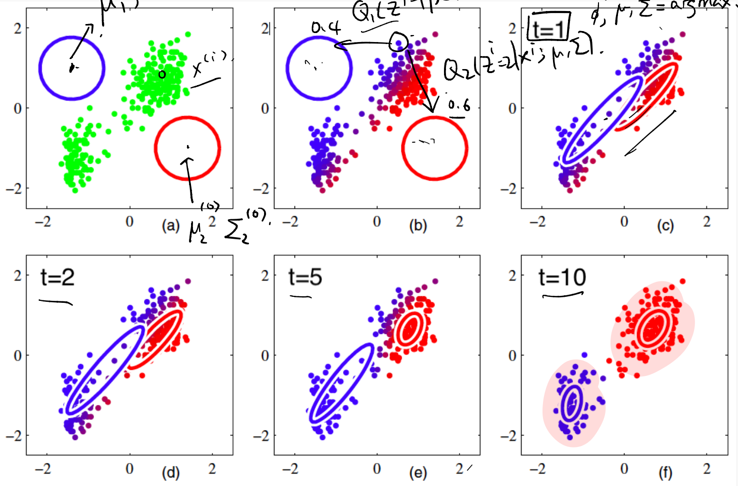
}

Illustration of EM steps

$t=0.$ $\mu_1^{(0)}, \Sigma_1^{(0)}, \phi^{(0)}$

E-step $Q_1(z^i=1|x^i;\mu,\Sigma)$.

M-step. $\phi^{(1)}, \mu^{(1)}, \Sigma^{(1)} = \arg\max J(\phi,\mu,\Sigma)$



$x^{(i)}$

0.4

$Q_2(z^i=2|x^i;\mu,\Sigma)$.

0.6

$\mu_2^{(0)}$ $\Sigma_2^{(0)}$.

Comparison with k-means clustering

Listing 4: EM Algorithm

```
Repeat untill convergence {
(E-step) For each i,j,
```
$$w_j^{(i)} := p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$$
```
(M-step) Update parameters:
```
$$\phi_j := \frac{1}{m} \sum_{i=1}^m w_j^{(i)}$$
$$\mu_j := \frac{\sum_{i=1}^m w_j^{(i)} x_j}{\sum_{i=1}^m w_j^{(i)}}$$
$$\Sigma_j := \frac{\sum_{i=1}^m w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^m w_j^{(i)}}$$
```
}
```

Listing 5: (Llyod's) k-means Alg.

```
Repeat untill convergence {
(E-step)  For every i,
```
$$c^{(i)} := \underset{j}{\text{argmin}} ||x^{(i)} - \mu_j||^2$$
```
(M-step) Update centroids:
  For each j
```
$$\mu_j := \frac{\mathbf{1}\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^m \mathbf{1}\{c^{(i)} = j\}}$$
```
}
```

Similar to k-means, Gaussian mixtures are also subject to local minimums.

Factor Analysis

Factor Analysis: Example

How much do you identify yourself with the following traits?

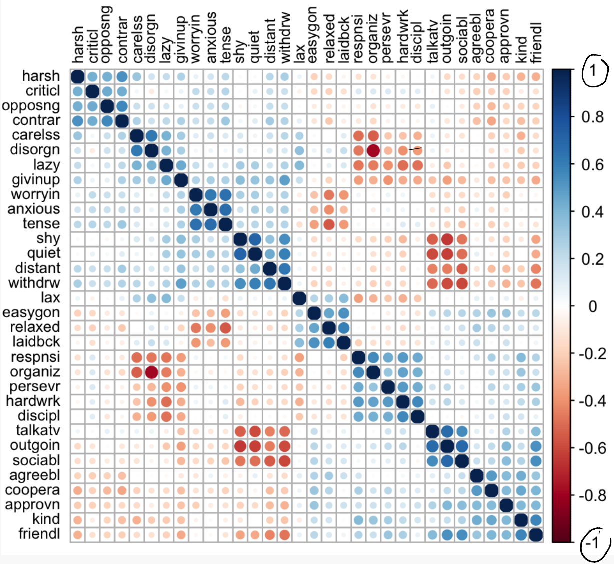
1-- the least 9 -- the most

	1	2	3	4	5	6	7	8	9
talkative									
distant									
careless									
hardwork									
anxious									
kind									

Self-ratings on 32 Personality Traits

Factor Analysis: Example

x_j



Pairwise correlation plot of 32 variables from 240 participants

Factor Analysis Terminology

- **observed random variables** $x \in \mathbb{R}^n$

$$\underset{}{x} = \overset{\mathbb{R}^n}{\mu} + \overset{\mathbb{R}^{n\times k}}{\Lambda} z + \overset{\mathbb{R}^n}{\epsilon}$$

- **factor** $z \in \mathbb{R}^k$ is the hidden (latent) construct that "causes" the observed variables
- **factor loadings** $\Lambda \in \mathbb{R}^{n\times k}$: the degree to which variable x_i is "caused" by the factors
- $\mu, \epsilon \in \mathbb{R}^n$ are the mean and error vectors

Factor Analysis Terminology

- **observed random variables** $x \in \mathbb{R}^n$

$$x = \mu + \Lambda z + \epsilon$$

- **factor** $z \in \mathbb{R}^k$ is the hidden (latent) construct that "causes" the observed variables
- **factor loadings** $\Lambda \in \mathbb{R}^{n \times k}$: the degree to which variable x_i is "caused" by the factors
- $\mu, \epsilon \in \mathbb{R}^n$ are the mean and error vectors

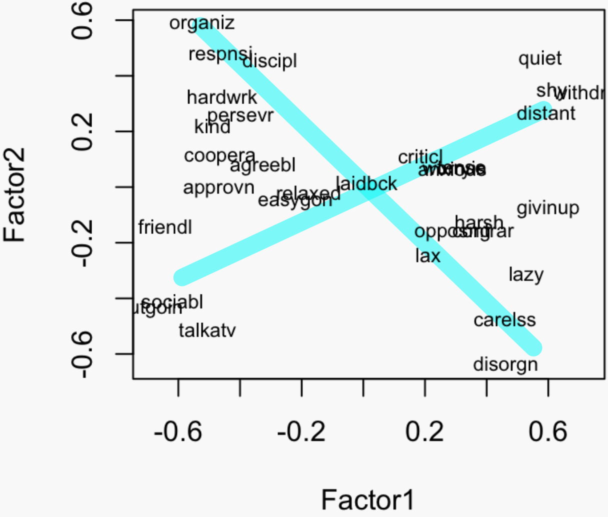
$z^{(i)} \in \mathbb{R}^4.$

Matrix of factor loading Λ for personality test data

variable	factor 1	factor 2	factor 3	factor 4
distant	0.59	0.27	0	0
talkative	-0.50	-0.51	0	0.27
careless	0.46	-0.47	0.11	0.14
hardworking	-0.46	0.33	-0.14	0.35
kind	-0.488	0.222	0	0
⋮				

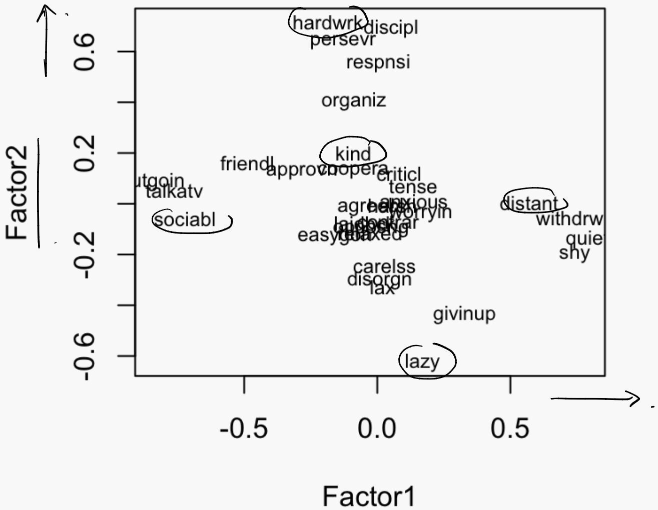
Factor Analysis: Example

Visualize loading of the first two factors

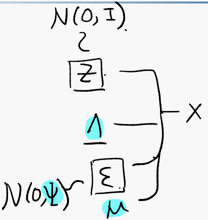


Factor Analysis: Example

Visualize loading of the first two factors, rotated to align with axes



Factor Analysis Model



Observed variables: $x \in \mathbb{R}^n$
Latent variables: $z \in \mathbb{R}^k$ $(k < n)$
The factor analysis model defines a joint distribution $p(x, z)$ as

$$z \sim \mathcal{N}(0, I)$$
$$\epsilon \sim \mathcal{N}(0, \Psi)$$
$$x = \mu + \Lambda z + \epsilon$$

where $\Psi \in \mathbb{R}^{n\times n}$ is a diagonal matrix, $\epsilon, \mu \in \mathbb{R}^n$, $\Lambda \in \mathbb{R}^{n\times k}$

Factor Analysis Model

Observed variables: $x \in \mathbb{R}^n$
Latent variables: $z \in \mathbb{R}^k$ $(k < n)$
The factor analysis model defines a joint distribution $p(x, z)$ as

assume z, ϵ independent.

$$z \sim \mathcal{N}(0, I)$$
$$\epsilon \sim \mathcal{N}(0, \Psi)$$
$$x = \mu + \Lambda z + \epsilon$$

where $\Psi \in \mathbb{R}^{n\times n}$ is a diagonal matrix, $\epsilon, \mu \in \mathbb{R}^n$, $\Lambda \in \mathbb{R}^{n\times k}$

Given observations $x^{(1)}, \ldots, x^{(m)}$, how to fit the parameters μ, Λ, Ψ ?

The EM Algorithm

Rubin, D. and Thayer, D. (1982). *EM algorithms for ML factor analysis*. Psychometrika, 47(1):69-76.

Listing 6: EM for Factor Analysis

Initialize μ, Λ, Ψ

Repeat untill convergence {

(E-step) For each i , set

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)}; \mu, \Lambda, \Psi)$$ $\leftarrow$ z is a continuous variable

(M-step) Set

$$\mu, \Lambda, \Psi := \operatorname*{argmax}_{\mu,\Lambda,\Psi} \underbrace{\sum_{i=1}^{m} \int_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \mu, \Lambda, \Psi)}{Q_i(z^{(i)})} dz^{(i)}}_{J(Q,\theta)} \qquad (\star)$$

μ, Λ, Ψ

The EM Algorithm

Rubin, D. and Thayer, D. (1982). *EM algorithms for ML factor analysis*. Psychometrika, 47(1):69-76.

Listing 7: EM for Factor Analysis

```
Initialize μ,Λ,Ψ
Repeat untill convergence {
  (E-step) For each i , set
        Q_i(z^(i)) := p(z^(i)|x^(i); μ,Λ,Ψ)  ← z is a continuous variable
  (M-step) Set
```

$$\mu, \Lambda, \Psi := \arg\max_{\mu,\Lambda,\Psi} \sum_{i=1}^{m} \int_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \mu, \Lambda, \Psi)}{Q_i(z^{(i)})} dz^{(i)} \qquad (\star)$$

First, we need to write $p(z^{(i)}|x^{(i)})$ and $p(x^{(i)}, z^{(i)})$in terms of the model parameters.

EM Derivations

$z \sim \mathcal{N}(0, I)$ $\quad \varepsilon \sim \mathcal{N}(0, \Psi)$.

$x = \mu + \Lambda z + \varepsilon$.

It can be shown that, random vector $\begin{bmatrix} z \\ x \end{bmatrix} \sim \mathcal{N}(\mu_{zx}, \Sigma)$ where $\mu_{xz} = \begin{bmatrix} 0 \\ \mu \end{bmatrix} \begin{array}{l} \leftarrow \mathbb{E}[z] \\ \leftarrow \mathbb{E}[x]. \end{array}$

and $\Sigma = \left[\begin{array}{c|c} I & \Lambda^T \\ \hline \Lambda & \Lambda\Lambda^T + \Psi \end{array}\right] = \left[\begin{array}{c|c} \Sigma_{zz} & \Sigma_{zx} \\ \hline \Sigma_{xz} & \Sigma_{xx} \end{array}\right]$

$\mathbb{E}[x] = \mu$.

$\mathbb{E}[z] = 0$.

Goal: Find $\Sigma_{zz}, \Sigma_{zx}, \ldots$

$\Sigma_{zz} = I$.

$$\Sigma_{zx} = \mathbb{E}[(z - \underbrace{\mathbb{E}(z)}_{0})(x - \underbrace{\mathbb{E}[x]}_{\mu})^T] = \mathbb{E}[z(x-\mu)^T]$$

$$= \mathbb{E}[z(\underbrace{\cancel{\mu} + \Lambda z + \varepsilon - \cancel{\mu}}){}^T]$$

$$= \mathbb{E}[z(z^T\Lambda^T) + z\varepsilon^T]$$

$$= \underbrace{\mathbb{E}[zz^T]}_{cov(z) = I}\Lambda^T + \underbrace{\mathbb{E}[z\varepsilon^T]}_{\underbrace{\mathbb{E}[z]}_{=0}\mathbb{E}[\varepsilon^T]}$$

$$= I\Lambda^T = \Lambda^T$$

$\Sigma_{xx} = \Lambda\Lambda^T + \Psi$.

EM Derivations

It can be shown that, random vector $\begin{bmatrix} z \\ x \end{bmatrix} \sim \mathcal{N}(\mu_{zx}, \Sigma)$ where $\mu_{xz} = \begin{bmatrix} 0 \\ \mu \end{bmatrix}$ and $\Sigma = \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Psi \end{bmatrix}$

E-Step

The posterior distribution $\underline{z^{(i)}|x^{(i)}} \sim \mathcal{N}\left(\mu_{z^{(i)}|x^{(i)}}, \Sigma_{z^{(i)}|x^{(i)}}\right)$

$$\mu_{z^{(i)}|x^{(i)}} = \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}(x^{(i)} - \mu)$$

$$\Sigma_{z^{(i)}|x^{(i)}} = I - \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}\Lambda$$

$\boxed{\text{Fact}}$. For any $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ $x \sim N(\mu, \Sigma)$, $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$, $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$

$x_1|x_2 \sim N(\mu_{1|2}, \Sigma_{1|2})$ where $\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$.

$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

$x_2|x_1$

EM Derivations

It can be shown that, random vector $\begin{bmatrix} z \\ x \end{bmatrix} \sim \mathcal{N}(\mu_{zx}, \Sigma)$ where $\mu_{xz} = \begin{bmatrix} 0 \\ \mu \end{bmatrix}$ and $\Sigma = \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Psi \end{bmatrix}$

E-Step

The posterior distribution $z^{(i)}|x^{(i)} \sim \mathcal{N}\left(\mu_{z^{(i)}|x^{(i)}}, \Sigma_{z^{(i)}|x^{(i)}}\right)$

$$\begin{aligned} \mu_{z^{(i)}|x^{(i)}} &= \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}(x^{(i)} - \mu) \\ \Sigma_{z^{(i)}|x^{(i)}} &= I - \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}\Lambda \end{aligned}$$

$$\begin{aligned} Q_i(z^{(i)}) &= p(z^{(i)}|x^{(i)}; \mu, \Lambda, \Psi) \\ &= \frac{1}{\sqrt{(2\pi)^k|\Sigma_{z^{(i)}|x^{(i)}}|}} \exp\left(-\frac{1}{2}(z^{(i)} - \mu_{z^{(i)}|x^{(i)}})^T \Sigma^{-1}_{z^{(i)}|x^{(i)}} (z^{(i)} - \mu_{z^{(i)}|x^{(i)}})\right) \end{aligned}$$

EM Derivations

M-Step

$$\operatorname*{argmax}_{\mu,\Lambda,\Psi} \sum_{i=1}^{m} \int_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \mu, \Lambda, \Psi)}{Q_i(z^{(i)})} dz^{(i)} \qquad (\star)$$

Note that

$$\int_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \mu, \Lambda, \Psi)}{Q_i(z^{(i)})} dz^{(i)}$$
$$= \mathbb{E}_{z\sim Q_i}[\log p(x^{(i)}|z^{(i)}; \mu, \Lambda, \Psi) + \log p(z^{(i)}) - \log Q_i(z^{(i)})]$$

EM Derivations

M-Step

$$\underset{\mu,\Lambda,\Psi}{\operatorname{argmax}} \sum_{i=1}^{m} \int_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \mu, \Lambda, \Psi)}{Q_i(z^{(i)})} dz^{(i)} \qquad (\star)$$

Note that

$$\int_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \mu, \Lambda, \Psi)}{Q_i(z^{(i)})} dz^{(i)}$$
$$= \mathbb{E}_{z \sim Q_i}[\log p(x^{(i)}|z^{(i)}; \mu, \Lambda, \Psi) + \log p(z^{(i)}) - \log Q_i(z^{(i)})]$$

$(\star)$ is equivalent to

$$\underset{\mu,\Lambda,\Psi}{\operatorname{argmax}} \sum_{i=1}^{m} \mathbb{E}_{z^{(i)} \sim Q_i}[\log p(x^{(i)}|z^{(i)}; \mu, \Lambda, \Psi)]$$

EM Derivations

M-Step (con't)

$$\operatorname*{argmax}_{\mu,\Lambda,\Psi} \sum_{i=1}^{m} \mathbb{E}_{z^{(i)}\sim Q_i}[\log p(x^{(i)}|z^{(i)};\mu,\Lambda,\Psi)] \quad (\star\star)$$

Since $x = \mu + \Lambda z + \epsilon$ and $\epsilon \sim \mathcal{N}(0, \Psi)$

$$x^{(i)}|z^{(i)} \sim \mathcal{N}(\mu + \Lambda z, \Psi)$$

EM Derivations

M-Step (con't)

$$\underset{\mu,\Lambda,\Psi}{\operatorname{argmax}} \sum_{i=1}^{m} \mathbb{E}_{z^{(i)}\sim Q_i}[\log p(x^{(i)}|z^{(i)};\mu,\Lambda,\Psi)] \quad (\star\star)$$

Since $x = \mu + \Lambda z + \epsilon$ and $\epsilon \sim \mathcal{N}(0, \Psi)$

$$x^{(i)}|z^{(i)} \sim \mathcal{N}(\mu + \Lambda z, \Psi)$$

$$p(x^{(i)}|z^{(i)};\mu,\Lambda,\Psi)$$
$$= \frac{1}{(2\pi)^{n/2}|\Psi|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu - \Lambda z^{(i)})^T \Psi^{-1}(x^{(i)} - \mu - \Lambda z^{(i)})\right)$$

We can maximize $(\star\star)$ with respect to μ, Λ and Ψ

Factor Analysis Discussions

Comparison with Mixture of Gaussians

- Mixture of Gaussians assumes sufficient data and relative few response variables. i.e. when $n \approx m$ or $n > m$, Σ is singular

$x_1, \ldots, x_n$

sample size $m > n$

Factor Analysis Discussions

Comparison with Mixture of Gaussians

- Mixture of Gaussians assumes sufficient data and relative few response variables. i.e. when $n \approx m$ or $n > m$, Σ is singular
- Factor Analysis works when $\underline{n > m}$ by allowing model noise $\underline{\Sigma}$

Factor Analysis Discussions

Relationship to PCA

- Both PCA and factor analysis can find low dimensional latent subspace in data

Factor Analysis Discussions

Relationship to PCA

- Both PCA and factor analysis can find low dimensional latent subspace in data
- PCA is good for data reduction (reduce correlation among observed variables)

Factor Analysis Discussions

Relationship to PCA

- Both PCA and factor analysis can find low dimensional latent subspace in data
- PCA is good for data reduction (reduce correlation among observed variables)
- Factor analysis is good for data exploration (find independent, common factors in observed variables)

Factor Analysis Discussions

Relationship to PCA

- Both PCA and factor analysis can find low dimensional latent subspace in data
- PCA is good for data reduction (reduce correlation among observed variables)
- Factor analysis is good for data exploration (find independent, common factors in observed variables)
- Factor analysis allows the noise to have an arbitrary diagonal covariance matrix, while PCA assumes the noise is spherical.

Additional readings

- Zoubin Ghahramani and Geoffrey E. Hinton, The EM Algorithm for Mixtures of Factor Analyzers, 1997

Next Lecture

Semi-Supervised Learning

- Semi-supervised SVM
- Graph-based semi-supervised learning spectral clustering
- Deep semi-supervised learning

No WA5. Please focus on your final project!

The class on Dec 31st will be PA5 discussion, project Q&A and group study