Recitation 10/8

Mixture Models, PCA

Slides borrowed from Prof. Seyoung Kim, Ryan Tibshirani. Thanks!

Gaussian Mixture Models (GMMs)

- Consider a mixture of K Gaussian components:

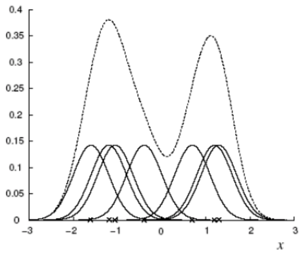
$$p(x_n) = \sum_k p(x_n \mid z_n = k)\, p(z_n = k)$$

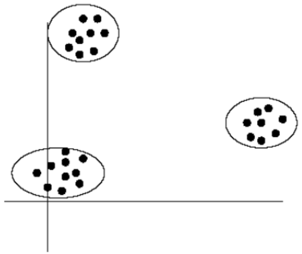
$$= \sum_k N(x_n \mid \mu_k, \Sigma_k)\, \pi_k$$

Law of Total Probability

mixture component

mixture proportion





Completely Observed Data

Bishop Page 431

Since z uses a 1-of-K representation, we have

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}. \tag{9.10}$$

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}. \tag{9.11}$$

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{9.12}$$

MLE for GMM with fully observed data

- If we are doing MLE for completely observed data
- Data log-likelihood

$$l(\theta;D) = \log\prod_n p(z_n,x_n) = \log\prod_n p(z_n\mid\pi)p(x_n\mid z_n,\mu,\sigma)$$

$$= \sum_n \log\prod_k \pi_k^{z_n^k} + \sum_n \log\prod_k N(x_n;\mu_k,\sigma)^{z_n^k}$$

$$= \sum_n\sum_k z_n^k \log\pi_k - \sum_n\sum_k z_n^k \frac{1}{2\sigma^2}(x_n-\mu_k)^2 + C$$

- MLE

$$\hat{\pi}_{k,MLE} = \arg\max_\pi l(\boldsymbol{\theta};D),$$

$$\hat{\mu}_{k,MLE} = \arg\max_\mu l(\boldsymbol{\theta};D)$$

$$\hat{\sigma}_{k,MLE} = \arg\max_\sigma l(\boldsymbol{\theta};D)$$

$$\hat{\pi}_{k,MLE} = \frac{\sum_n z_n^k}{N}$$

$$\hat{\mu}_{k,MLE} = \frac{\sum_n z_n^k x_n}{\sum_n z_n^k}$$

$$\hat{\sigma}^2_{k,MLE} = \frac{\sum_n z_n^k (x_n-\mu_k)^2}{\sum_n z_n^k}$$

- What if we do not know z_n?

What if we do not know z_n ?

- Maximize the expected data log likelihood for (x_i, z_i) based on $p(x_i, z_i)$
  - Expectation-Maximization (EM) algorithm

Complete vs. Expected Complete Log Likelihoods

- The complete log likelihood:

$$\boldsymbol{l}(\boldsymbol{\theta}; D) = \log \prod_n p(z_n, x_n) = \log \prod_n p(z_n \mid \pi) p(x_n \mid z_n, \mu, \sigma)$$

$$= \sum_n \log \prod_k \pi_k^{z_n^k} + \sum_n \log \prod_k N(x_n; \mu_k, \sigma)^{z_n^k}$$

$$= \sum_n \sum_k z_n^k \log \pi_k - \sum_n \sum_k z_n^k \tfrac{1}{2\sigma^2}(x_n - \mu_k)^2 + C$$

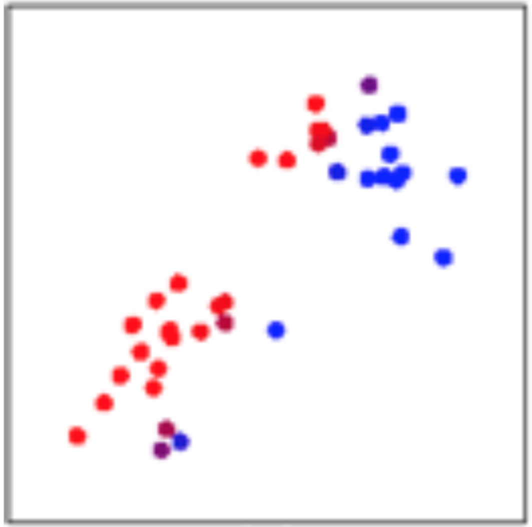
- The expected complete log likelihood

$$\langle \boldsymbol{l}_c(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{z}) \rangle = \sum_n \langle \log p(z_n \mid \pi) \rangle_{p(z|x)} + \sum_n \langle \log p(x_n \mid z_n, \mu, \Sigma) \rangle_{p(z|x)}$$

$$= \sum_n \sum_k \langle z_n^k \rangle \log \pi_k - \frac{1}{2} \sum_n \sum_k \langle z_n^k \rangle \left((x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k) + \log |\Sigma_k| + C \right)$$

- EM optimizes the expected complete log likelihood

The Expectation-Maximization (EM) Algorithm



E step:

$$\tau_n^{k(t)} = \left\langle z_n^k \right\rangle_{q^{(t)}} = p(z_n^k = 1 \mid x, \mu^{(t)}, \Sigma^{(t)})$$

M step:

$$\pi_k^{(t+1)} = \sum_n \tau_n^{k(t)} \Big/ N = \langle n_k \rangle \Big/ N$$

$$\mu_k^{(t+1)} = \frac{\sum_n \tau_n^{k(t)} x_n}{\sum_n \tau_n^{k(t)}}$$

$$\Sigma_k^{(t+1)} = \frac{\sum_n \tau_n^{k(t)} (x_n - \mu_k^{(t+1)})(x_n - \mu_k^{(t+1)})^T}{\sum_n \tau_n^{k(t)}}$$

Example 2-d data points coming from K = 2 Gaussian distributions

K=2 1-d Gaussian distributions:

$$G_1(\mu_1, \sigma_1^2), G_2(\mu_2, \sigma_2^2)$$

<x, y> pairs

$$x \in \mathcal{R}, y \in \{G_1, G_2\}$$

$$x = (2, 4, 7)$$

Example 2-d data points coming from K = 2 Gaussian distributions

K=2 1-d Gaussian distributions:

$$G_1(\mu_1, \sigma_1^2), G_2(\mu_2, \sigma_2^2)$$

<x, y> pairs

$$x \in \mathcal{R}, y \in \{G_1, G_2\}$$

$$x = (2, 4, 7)$$

Initialize

$$\mu^{(0)} = (3, 6)$$

$$\pi^{(0)} = (\frac{1}{2}, \frac{1}{2})$$

$$\sigma^{2^{(0)}} = (\frac{1}{2}, \frac{1}{2})$$

Example 2-d data points coming from K = 2 Gaussian distributions

$$x = (2, 4, 7)$$

iteration t =1

Initialize

$$\mu^{(0)} = (3, 6)$$

$$\tau_1^1 = p(z_1^1 = 1|x_1) = \frac{p(x_1|\mu_1)p(\mu_1)}{p(x_1|\mu_1)p(\mu_1) + p(x_1|\mu_2)p(\mu_2)} = \frac{\frac{1}{2}N(2, 3, \frac{1}{\sqrt{2}})}{\frac{1}{2}N(2, 3, \frac{1}{\sqrt{2}}) + \frac{1}{2}N(2, 6, \frac{1}{\sqrt{2}})} =$$

$$\pi^{(0)} = (\frac{1}{2}, \frac{1}{2})$$

$$1 - 10^{-7}$$

$$\sigma^{2^{(0)}} = (\frac{1}{2}, \frac{1}{2})$$

Example 2-d data points coming from K = 2 Gaussian distributions

$$x = (2, 4, 7)$$

iteration t =1

Initialize

$$\mu^{(0)} = (3, 6)$$

$$\pi^{(0)} = (\frac{1}{2}, \frac{1}{2})$$

$$\sigma^{2^{(0)}} = (\frac{1}{2}, \frac{1}{2})$$

$$\tau_1^1 = p(z_1^1 = 1|x_1) = \frac{p(x_1|\mu_1)p(\mu_1)}{p(x_1|\mu_1)p(\mu_1) + p(x_1|\mu_2)p(\mu_2)} = \frac{\frac{1}{2}N(2,3,\frac{1}{\sqrt{2}})}{\frac{1}{2}N(2,3,\frac{1}{\sqrt{2}}) + \frac{1}{2}N(2,6,\frac{1}{\sqrt{2}})} =$$

$$1 - 10^{-7}$$

x_i	2	4	7
τ_i^1	$1 - 10^{-7}$	0.953	10^{-7}
τ_i^2	10^{-7}	0.047	$1 - 10^{-7}$

Example 2-d data points coming from K = 2 Gaussian distributions

$$x = (2, 4, 7)$$

Initialize

$$\mu^{(0)} = (3, 6)$$

$$\pi^{(0)} = (\frac{1}{2}, \frac{1}{2})$$

$$\sigma^{2^{(0)}} = (\frac{1}{2}, \frac{1}{2})$$

iteration t =1

x_i	2	4	7
τ_i^1	$1 - 10^{-7}$	0.953	10^{-7}
τ_i^2	10^{-7}	0.047	$1 - 10^{-7}$

$$\pi_1 = \frac{1.953}{3} = 0.651, \pi_2 = 0.349$$

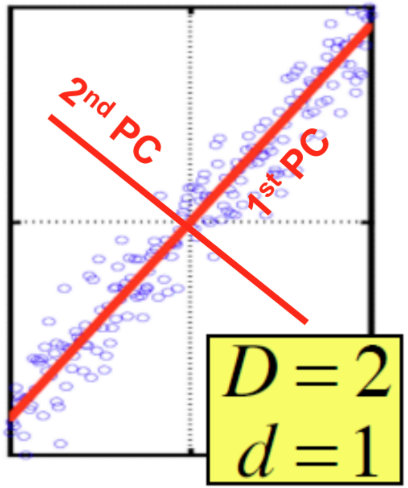
$$\mu_1 \approx \frac{2 + 0.953 * 4 + 0}{1.953} = 2.978 \qquad \mu_2 \approx 6.88$$

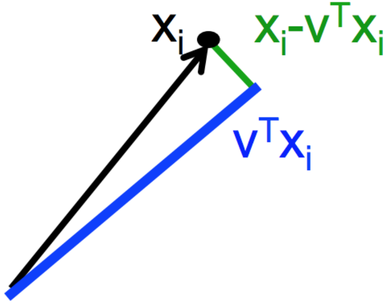
$$\sigma^2 \approx ...$$

PCA

Principal components are a sequence of projections of the data, mutually uncorrelated and ordered in variance.

Principal Component Analysis (PCA)





Principal Components (PC) are orthogonal directions that capture most of the variance in the data

1st PC – direction of greatest variability in data

2nd PC – Next orthogonal (uncorrelated) direction of greatest variability

(remove all variability in first direction, then find next direction of greatest variability)

And so on …

Assume X is a normalized Nxp data matrix for N samples and p features

Assume data is normalized. ⇐> each column of X is normalized.

Variance of projected data $$\frac{1}{N}\sum_{n=1}^{N}(v^T x_n - v^T \overline{x_n})^2 = v^T S v$$

<- Want to maximize this over v

where S = $$\frac{1}{n}\sum_i (x_i - \bar{x_i})(x_i - \bar{x_i})^T = \frac{1}{n}\sum_i x_i x_i^T$$

Computing the Components

- Projection of vector **x** onto an axis (dimension) **u** is $\mathbf{u}^T\mathbf{x}$
- Assume **X** is a normalized *nxp* data matrix for *n* samples and *p* features. Direction of greatest variability is that in which the average square of the projection is greatest:

Maximize $\quad (1/n)\, \mathbf{u}^T\mathbf{X}^T\mathbf{X}\mathbf{u}$

s.t $\quad \mathbf{u}^T\mathbf{u} = 1$

Construct Langrangian $(1/n)\, \mathbf{u}^T\mathbf{X}^T\mathbf{X}\mathbf{u} - \lambda\mathbf{u}^T\mathbf{u}$

Vector of partial derivatives set to zero

$$1/n\, \mathbf{X}^T\mathbf{X}\mathbf{u} - \lambda\mathbf{u} = 0$$

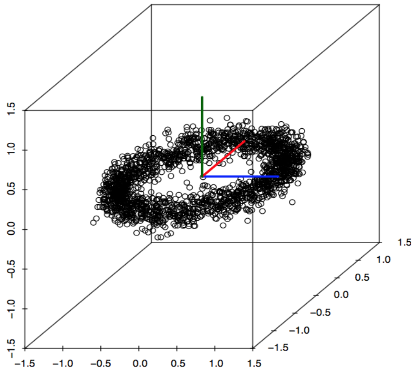
or equivalently $\mathbf{S}\mathbf{u} - \lambda\mathbf{u} = 0$ ($S = 1/n\, \mathbf{X}^T\mathbf{X}$: covariance matrix)

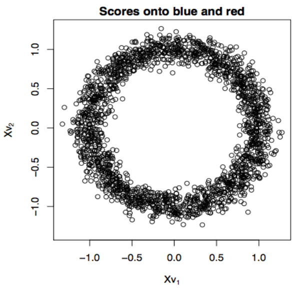
As $\mathbf{u} \neq \mathbf{0}$ then **u** must be an eigenvector of S with eigenvalue λ

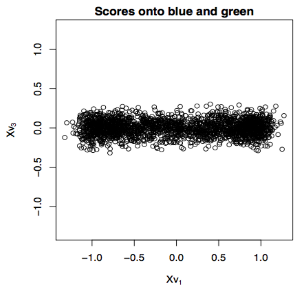
- λ is the principal eigenvalue of the **covariance matrix S**
- The eigenvalue denotes the amount of variability captured along that dimension

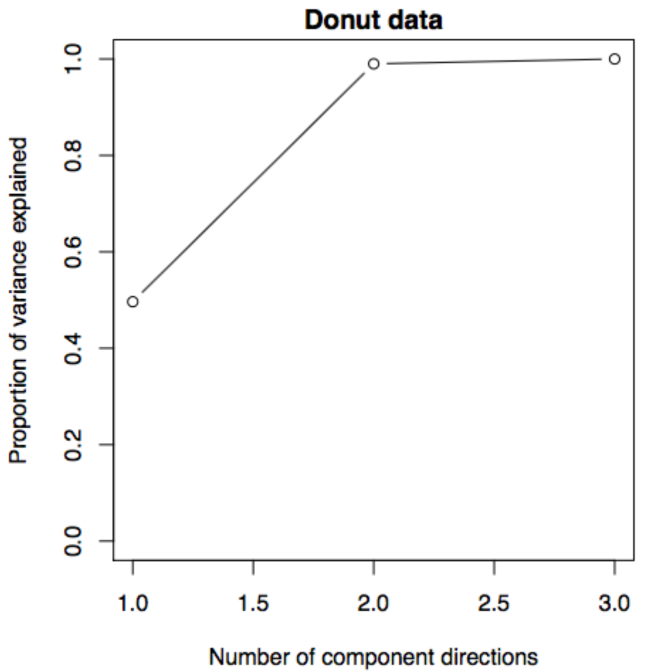
Example: projections onto orthonormal vectors

Example: $X \in \mathbb{R}^{2000\times 3}$, and $v_1, v_2, v_3 \in \mathbb{R}^3$ are the unit vectors parallel to the coordinate axes









The proportion of variance explained is a nice way to quantify how much structure is being captured