COM 525000 – Statistical Learning

Lecture 10 – Unsupervised Learning

Y.-W. Peter Hong

What is unsupervised learning?

- In supervised learning, we derive a model based on the data set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ to obtain a prediction $\hat{y}$ for new observation x.
- In unsupervised learning, we are given only the set of observations $\{x_1, x_2, \ldots, x_n\}$ and not the responses.

  ➔ *Perform exploratory data analysis to discover structure in the available data set.*
- Two perspectives:
  – Clustering: K-means, Gaussian mixture based, hierarchical clustering, spectral clustering etc.
  – Low-Dimensional Representation: (Factor analysis), principal components analysis, independent components analysis.

Clustering

- Clustering is the process of partitioning data into distinct groups or clusters such that the observations within each group are "similar" and those in different groups are "different". (➔ What is **"similar"**?)
  - E.g., for n observations of tissue samples from patients with breast cancer, clustering can discover different unknown subtypes of cancer.
  - E.g., in marketing, with n observations of people's income, occupation etc., clustering identifies groups of people more receptive to certain advertising.

➔(1) *K-means clustering*; (2) *hierarchical clustering*; and (3) *Gaussian mixture based clustering*.

K-Means Clustering (1/2)

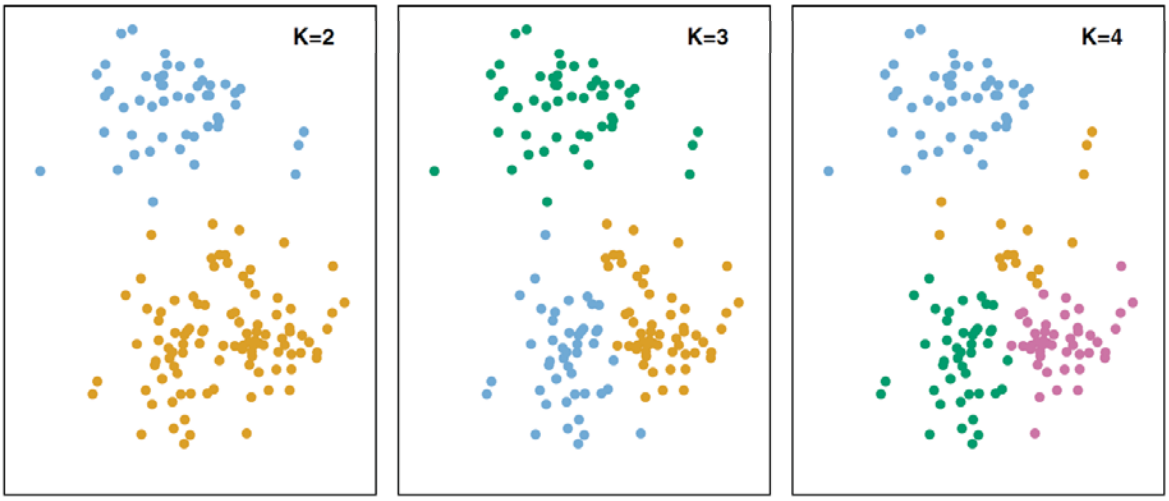
- Let us denote the clusters by $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K$ such that $\mathcal{C}_1 \cup \mathcal{C}_2 \cup \cdots \cup \mathcal{C}_K = \{1, \ldots, n\}$ and $\mathcal{C}_k \cap \mathcal{C}_{k'} = \emptyset, \forall k \neq k'$.
- K-means clustering finds clusters that minimize the *within cluster variation*

$$\sum_{k=1}^{K} \underbrace{\frac{1}{|\mathcal{C}_k|} \sum_{i,i' \in \mathcal{C}_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2}_{\text{within cluster variation } W(\mathcal{C}_k)} = \sum_{k=1}^{K} \frac{1}{|\mathcal{C}_k|} \sum_{i,i' \in \mathcal{C}_k} \|x_i - x_{i'}\|^2.$$

- Note that

$$W(\mathcal{C}_k) = \frac{1}{|\mathcal{C}_k|} \sum_{i,i' \in \mathcal{C}_k} \|x_i - x_{i'}\|^2 =$$

K-Means Clustering (2/2)



- The optimal solution requires search over K^n possible clustering solutions. (➔High complexity!!)

K-Means Clustering Algorithm

Algorithm 10.1 *K-Means Clustering*

1. **Initialize:** Randomly partition the n observations into K clusters $\mathcal{C}_1^{(0)}, \mathcal{C}_2^{(0)}, \ldots, \mathcal{C}_K^{(0)}$.
2. **Iteration** t (repeat until clusters stop changing):

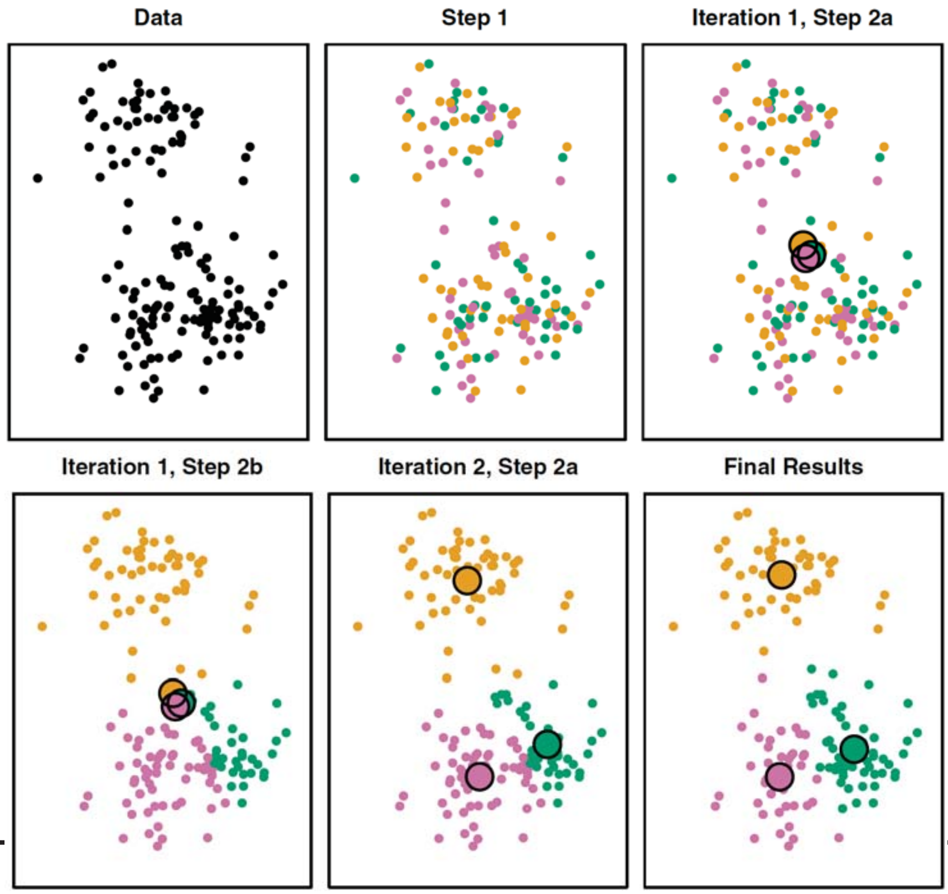
(a) Compute the centroids of the K clusters, i.e.,

$$\bar{x}_k^{(t)} = \frac{1}{|\mathcal{C}_k^{(t-1)}|} \sum_{i \in \mathcal{C}_k^{(t-1)}} x_i, \text{ for } k = 1, \ldots, K.$$

(b) Assign each observation to the cluster whose centroid is closest, i.e., assign clusters as

$$\mathcal{C}_k^{(t)} = \{i : \|x_i - \bar{x}_k^{(t)}\| \leq \|x_i - \bar{x}_{k'}^{(t)}\|, \forall k'\}, \text{ for } k = 1, \ldots, K.$$

Example of Intermediate Outcomes



Convergence

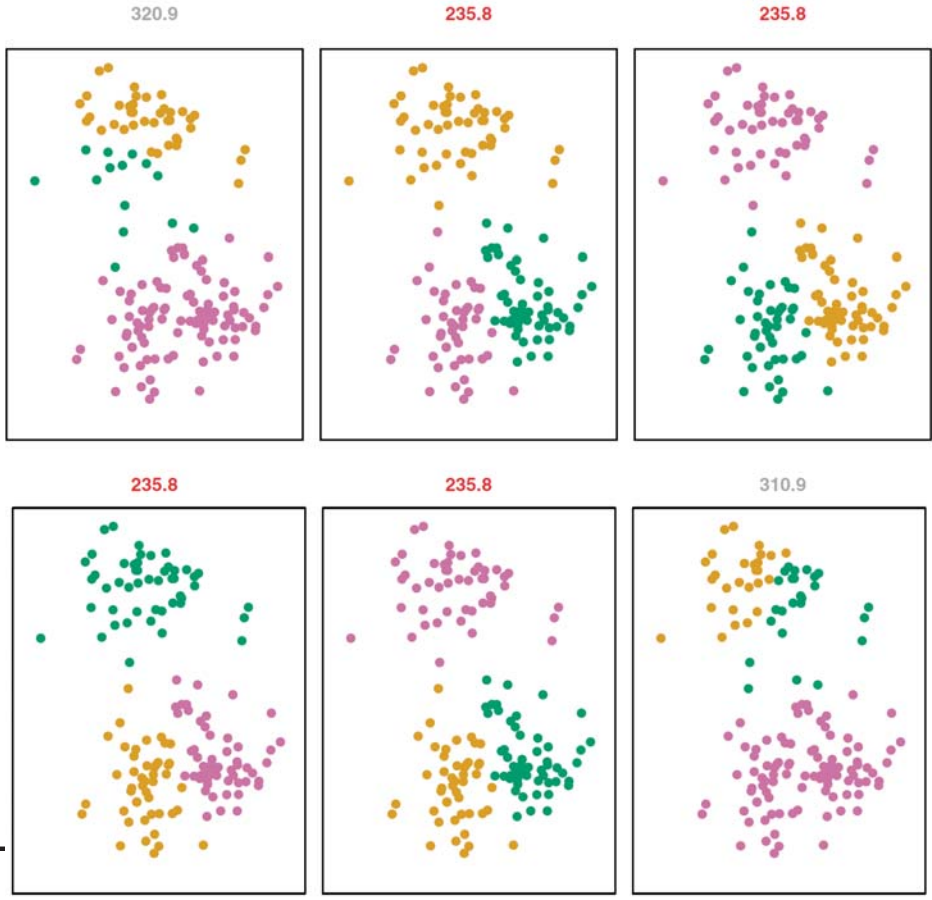
- The objective value monotonically decreases in each iteration. That is,

$$\sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k^{(t-1)}} \|x_i - \bar{x}_k^{(t-1)}\|^2 \geq$$

➔Convergence (to local optimum) guaranteed by monotone convergence theorem.

Remark: The initial condition matters. Hence, we should run multiple times and select the best solution (in terms of the minimum cost).

Example of Different Random Initiations



Latent Variable Model

- Multivariate data are often viewed as multiple indirect measurements of many underlying sources.
  - E.g., educational tests use the answers of questionnaires to measure the underlying intelligence of subjects.
  - E.g., EEG brain scans measure the neuron activity in various parts of the brain via electronic signals recorded at sensors placed at various positions on the band.
- Let us consider the latent variable model

$$\left.\begin{aligned} X_1 &= a_{11}S_1 + a_{12}S_2 + \cdots + a_{1p}S_p \\ X_2 &= a_{21}S_1 + a_{22}S_2 + \cdots + a_{2p}S_p \\ \vdots \quad & \quad \vdots \\ X_p &= a_{p1}S_1 + a_{p2}S_2 + \cdots + a_{pp}S_p \end{aligned}\right\} \quad \begin{gathered} X = \mathbf{A}S \\ (\text{or } \mathbf{X} = \mathbf{S}\mathbf{A}^T) \end{gathered}$$

where S_j's are uncorrelated and unit variance.

Independent Components Analysis (ICA)

- Recall: PCA estimates latent variable model by SVD

$$\mathbf{X} = \underbrace{\sqrt{N}\mathbf{U}}_{\mathbf{S}} \underbrace{\frac{1}{\sqrt{N}}\mathbf{D}\mathbf{V}^T}_{\mathbf{A}^T} = \mathbf{S}\mathbf{A}^T$$

- **Independent components analysis (ICA)** considers the same model $X = \mathbf{A}S$ (or $\mathbf{X} = \mathbf{S}\mathbf{A}^T$) but adopt priors on S that assume *independence* across entries.
- Let us consider the full p-component model, where $\mathbf{A}$ is $p \times p$ and S_j's are *independent* with unit variance.
- By assuming that X is standardized, it follows that

$$\mathrm{Cov}(X) = \mathbf{A}E[SS^T]\mathbf{A}^T = \mathbf{A}\mathbf{A}^T = \mathbf{I}$$

i.e., $\mathbf{A}$ is orthogonal.

Finding the Mixing Matrix

- **Goal:** ICA looks to find orthogonal $\mathbf{A}$ such that the entries of S are independent (and **non-Gaussian**).
- Suppose that the CDF of S_j is $F_{S_j}(s) = g(s) = \frac{1}{1+e^{-s}}$ and, thus, $p_{S_j}(s) = g'(s)$, for all j.
- The mixing matrix $\mathbf{A}$ can be found by maximizing the log-likelihood function

$$\ell(\mathbf{A}) = \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \log g'(a_j^T x_i) + \log |\mathbf{A}^T| \right).$$

➔ Solve using (stochastic) gradient ascent.

Why Non-Gaussian?

- Note that ambiguity in the solution exists under Gaussian priors because the multivariate standard Gaussian distribution is rotationally symmetric, i.e.,

$$S \sim S^* \triangleq \mathbf{R}S \sim \mathcal{N}(0, \mathbf{I})$$

for any orthogonal $\mathbf{R}$.

- For S that is non-Gaussian,

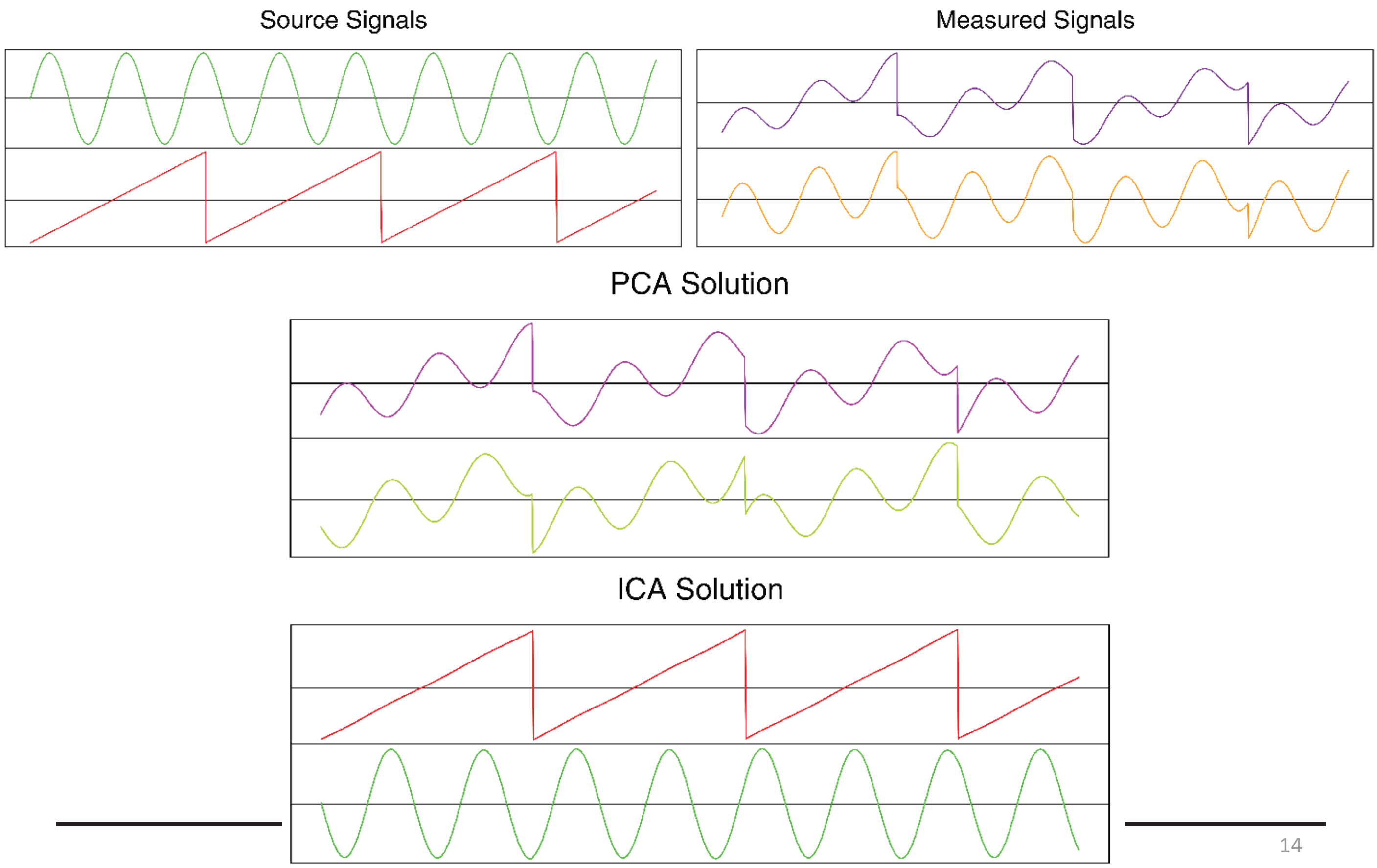
$$p_{S^*}(s^*) = p_S(\mathbf{R}^T s^*)\det(\mathbf{R}^T).$$

- Similarly, with $X = \mathbf{A}S$ (thus, $S = \mathbf{A}^T X$), we have

$$p_X(x) = p_S(\mathbf{A}^T x)\det(\mathbf{A}^T) = \prod_{j=1}^{p} p_{S_j}(a_j^T x)\det(\mathbf{A}^T)$$

where a_j is the j-th column of $\mathbf{A}$.

Example



Example