HW 5: Learning Theory II (580.692)

Instructor: René Vidal, Office: 308B Clark, E-mail: *rvidal@cis.jhu.edu*
Grader: Atiyeh Ghoreyshi. Office: 319 Clark, E-mail: *ati@cis.jhu.edu*

Due 10/19/06 beginning of the class

1. Read Chapter 4 of GPCA book. Go to *http://www.vision.jhu.edu/gpcabook/* and submit all the typos you find as well as suggestions you may have to improve the quality and/or readability of the material. **You will receive credit for each interesting typo or suggestion you submit**.

2. **Mixtures of PPCA**

   Let $\boldsymbol{x} \in \mathbb{R}^D$, $\boldsymbol{y} \in \mathbb{R}^d$ and $\boldsymbol{\epsilon} \in \mathbb{R}^D$, $d \leq D$, be random variables related by the generative model $\boldsymbol{x} = \boldsymbol{\mu} + U\boldsymbol{y} + \boldsymbol{\epsilon}$, where the parameters of the model are the mean $\boldsymbol{\mu} \in \mathbb{R}^D$ and the subspace basis $U \in \mathbb{R}^{D \times d}$. Assume further that $\boldsymbol{y} \sim N(\mathbf{0}, I)$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I)$ and that $\boldsymbol{y}$ and $\boldsymbol{\epsilon}$ are independent.

   (a) Is it necessary to assume $U^\top U = I$? Why yes, or why no?

   (b) Derive the E and M steps of the Expectation Maximization algorithm with $\boldsymbol{y}$ as a latent variable to obtain estimates of the parameters $\boldsymbol{\mu}$, U and σ from given measurements $\{\boldsymbol{x}_i\}_{i=1}^N$.

   (c) Show that the measurements $\{\boldsymbol{x}_i\}_{i=1}^N$ are conditionally independent, given $\{\boldsymbol{y}_i\}_{i=1}^N$ and that $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \Sigma)$ with $\Sigma = UU^\top + \sigma^2 I$. Derive an algorithm for estimating $\boldsymbol{\mu}$ and Σ from $\{\boldsymbol{x}_i\}_{i=1}^N$, and U and σ from Σ.

   (d) Assume now that $\boldsymbol{x}$ is drawn from a mixture of n models of the form $\boldsymbol{x} = \boldsymbol{\mu}_j + U_j \boldsymbol{y}_j + \boldsymbol{\epsilon}_j$ with mixing proportions π_j, where $\sum_{j=1}^n \pi_j = 1$. As before, assume that $\boldsymbol{y}_j \sim N(\mathbf{0}, I)$, $\boldsymbol{\epsilon}_j \sim N(\mathbf{0}, \sigma_j^2 I)$ and $\boldsymbol{y}_j$ and $\boldsymbol{\epsilon}_j$ are independent for all $j = 1, \ldots, n$. Derive the equations of the EM algorithm for estimating the parameters of the mixture model from data points $\{\boldsymbol{x}_i\}_{i=1}^N$.

3. **More on Central Clustering and Image Segmentation**

   (a) Consider the `polysegment` algorithm for clustering data in $\mathbb{R}$ (Algorithm 4.1, page 62 of GPCA book). Show that the least squares solution for the vector of coefficients $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)^\top \in \mathbb{R}^n$ is

$$\boldsymbol{c} = \begin{bmatrix} E(\boldsymbol{x}^{2n-2}) & E(\boldsymbol{x}^{2n-3}) & \cdots & E(\boldsymbol{x}^{n-1}) \\ E(\boldsymbol{x}^{2n-3}) & \ddots & & \\ \vdots & & \ddots & \vdots \\ E(\boldsymbol{x}^{n-1}) & E(\boldsymbol{x}^{n-2}) & \cdots & 1 \end{bmatrix}^{-1} \begin{bmatrix} E(\boldsymbol{x}^{2n-1}) \\ E(\boldsymbol{x}^{2n-2}) \\ \vdots \\ E(\boldsymbol{x}^{n}) \end{bmatrix}. \tag{1}$$

   where $E(\boldsymbol{x}^k) = \frac{1}{N}\sum_{i=1}^N x_i^k$ is the kth moment of the points $\{x_i\}_{i=1}^N$. Show also that when $n = 2$ the roots of $p(x) = x^2 + c_1 x + c_2$ (the cluster centers) are always real. Can you extend your result to $n > 2$?

   (b) Prove as rigorously as you can that $n = \min\{j : \text{rank}(\boldsymbol{V}_i) = j\}$, i.e. formula 4.4 in GPCA book. The rigorous proof involves using Hilbert' Nullstellensatz, which you can find in Appendix B of GPCA book.

   (c) Implement the `polysegment` algorithm. The format of the function must be

Function	**`[means,group] = polysegment(x,n)`**
Parameters	
`x`	$D \times N$ matrix whose columns are the data points with $D = 1$ or $D = 2$
`n`	number of groups
Returned values	
`mean`	$D \times n$ matrix whose columns are the cluster centers
`group`	$1 \times N$ vector with group membership of each point
Description	
Computes the clustering of points using polysegment.	

(d) Use your script from HW4 to generates data in $\mathbb{R}^2$ distributed according to a mixture of two Gaussians with means $(-2,-2)$ and $(2,2)$, and common variance σI. Assume that the mixing proportions are $\pi_1 = \pi_2 = 1/2$. Plot the mean number of iterations and the mean error in the estimation of the means as a function of σ for $\sigma =$ `0.1:0.1:1` for 1,000 realizations of 200 points for the following algorithms: Polysegment, Kmeans randomly initialized, EM randomly initialized, EM initialized with Kmeans. Compare your results. What are the advantages disadvantages of each algorithm?

(e) Use `polysegment`, `kmeans` and `EM` to segment the images on the course webpage. Use `0:1/(n-1):1` as the n initial cluster centers for `kmeans` and `EM`. Compare your results. What are the advantages disadvantages of each algorithm?

4. **Line Clustering**

Using the formula $\boldsymbol{x} = U_j \boldsymbol{y} + B_j \boldsymbol{s}$ where $\boldsymbol{y} \sim N(0, {\sigma_y}^2 I)$ and $\boldsymbol{s} \sim N(0, {\sigma_j}^2 I)$, with $U_1 = [1, 0]^\top$, $U_2 = [\cos(\theta), \sin(\theta)]^\top$, $B_1 = [0, 1]^\top$, $B_2 = [-\sin(\theta), \cos(\theta)]^\top$, $\sigma_y = 10$, $\sigma_1 = \sigma_2 = .5$, and $\theta =$ `0:15:90`, randomly generate 7 datasets containing 1000 points each, corresponding to different values of θ .

(a) For each dataset, use classical EM to segment the data points into 2 groups.

(b) For each dataset, use `polysegment` adapted for line clustering to cluster the data points.

(c) Plot the percentage of incorrectly classified points as a function of θ for each one of the two methods and comment your results.

5. **A closed form solution to two hyperplane clustering**

Let $p(\boldsymbol{x}) = \boldsymbol{c}^\top \nu_2(\boldsymbol{x}) = (\boldsymbol{b}_1^\top \boldsymbol{x})(\boldsymbol{b}_2^\top \boldsymbol{x})$ be the polynomial whose zero set is the union of two hyperplanes in $\mathbb{R}^D$ with normal vectors $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$.

(a) Show that $p(x) = \boldsymbol{x}^\top M \boldsymbol{x}$ where $M = \boldsymbol{b}_1 \boldsymbol{b}_2^\top + \boldsymbol{b}_2 \boldsymbol{b}_1^\top \in \mathbb{R}^{D \times D}$.

(b) Write M as an explicit function of $\boldsymbol{c}$. For example, if $D = 3$, then

$$M = \begin{bmatrix} 2c_1 & c_2 & c_3 \\ c_2 & 2c_4 & c_5 \\ c_3 & c_5 & 2c_6 \end{bmatrix} \tag{2}$$

where $\boldsymbol{c} = (c_1, c_2, c_3, c_4, c_5, c_6)$ is the vector of coefficients.

(c) Let $M = U \Lambda U^{-1}$ be the eigenvalue decomposition of M, where $U \in SO(3)$ and Λ diagonal with $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_D$. Show that M is of rank 2, with eigenvalues $\lambda_1 > 0$, $\lambda_2 < 0$ and $\lambda_j = 0$, $j = 3, 4, , \dots, D$.

(d) Show that the normal vectors can be obtained (up to a scale factor) as:

$$\begin{bmatrix} \boldsymbol{b}_1 & \boldsymbol{b}_2 \end{bmatrix} = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1} & \sqrt{\lambda_1} \\ \sqrt{-\lambda_2} & -\sqrt{-\lambda_2} \end{bmatrix}. \tag{3}$$