CS 480/680 Homework 8
Released April 6, 2026
NOT GRADED

mmp@uwaterloo.ca

Problem 1 – Selecting K

1.a We will consider the finite Mixture of Gaussians model defined by

$$f(x) \;=\; \frac{1}{K}\sum_{k=1}^{K}\mathcal{N}(x;\mu_k,\Sigma)\,, \tag{1}$$

namely a model where all clusters have the same covariance matrix. Calculate the number of parameters in this model, first as a literal expression, then for $d = 5$, $K = 6$.

1.b. Now we want to use BIC to select K. You have run the EM algorithm for $K = 1, 2, \ldots 10$ for a dataset $\mathcal{D}$ of size $n = 1000$ and obtained the following log-likelihoods.

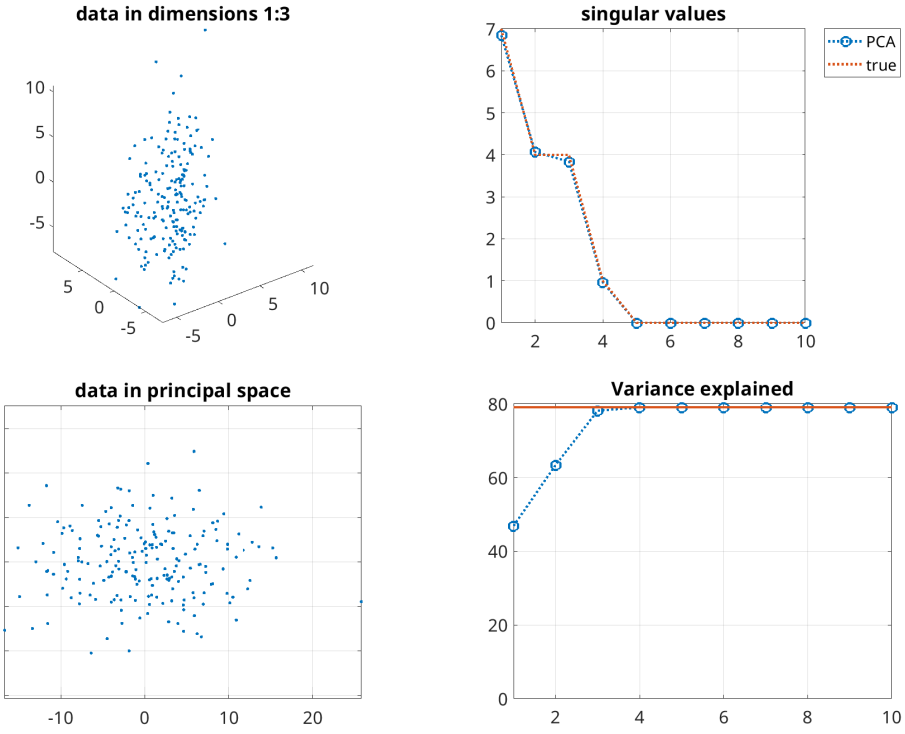
K	1	2	3	4	5	6	7	8	9	10	11	12
l	-412.5	-361.0	-113.2	428.4	485.15	641.7	685.5	754.2	760.1	764.5	772.2	776.3

Plot the BIC values and select $\hat{K}$ the optimum value according to the BIC.

$$BIC(\theta_K) \;=\; \sum_{i=1}^{n} f(x^i;\theta_K) - \frac{\#\theta_K}{2}\ln n, \tag{2}$$

where θ_K is the set of all parameters of a mixture defined by (1) with K clusters.

Problem 2 – PCA The following figures represent the Principal Component Analysis of a data set $\mathcal{D}$.



2.a What is the dimension D of the data points?

2.b Does the data live in a lower dimension $D' < D$? This means that by a rotation of the entire data set, one can bring the data into D' dimensions, with the remaining $D - D'$ dimensions being 0. If so, what is D'?

2.c Select a d so that the variance explained by d-PCA is at least 75% of the total variance.

2.d Assume that we have a $\mathcal{D}' = \{3x^i + 1,\ x^i \in \mathcal{D}\}$. What will be $\tilde{\sigma}_1^2$ for the data $\mathcal{D}'$? What d should be chosen to explain at least 75% of the total variance.

Problem 3 – Cross-validation

3.a Return to Problem 2 from Hw1, and run m-fold CV with $m = 6$. What is the value $\hat{K}$ you obtain?

3.b Can you think of a way to do CV for clustering? Can any of the methods for selecting K for clustering be described as a CV method?

Problem 4 – Sparse Regression

4.a Show that for any $\beta \in \mathbb{R}^d$, $\|\beta\|_1 \geq \|\beta\|$ where $\|\cdot\|$ represents the Euclidean norm. Assume $\|\beta\| = 1$. For what values of β do we have equality?

4.b Let $\mathcal{D}$ be a data set with n pairs (x^i, y^i), and let $\hat{\beta}$ be the unique minimum value of

$$J(\beta) \ = \ \frac{1}{n}\sum_{i=1}^{n}(y^i - \beta^T x^i)^2 + \lambda\|\beta\|_1, \quad \lambda > 0 \tag{3}$$

Assume that the data satisfy $y^i = \beta_*^T x^i$ for $i = 1 : n$, where β is s-sparse. What is the sample size needed to ensure that $\hat{\beta}$ is also s-sparse and a good approximation to β_*? (Note that if $s' < s$, any s'-sparse vector is also s-sparse); answer with $\mathcal{O}$ notation.

4.c Assume $n \geq 2$, is $\hat{\beta} = \beta_*$?

4.d Assume now $y^i = \beta_*^T x^i + \epsilon_i$ for $i = 1 : n$, and let $\beta_{LS} \neq 0$ be the minimizer of L_{LS} the unregularized solution. Show that $\|\beta_{LS}\|_1 > \|\hat{\beta}\|_1$ for any $\lambda > 0$.

4.e When $d = 1$, calculate the smallest value of $\lambda_{\max}$ for which $\hat{\beta} = 0$ in (3).