1 PCA, part 1

Let Y be a Bernoulli random variable $\in \{0, 1\}$ with $p = 0.5$, and X be a mixture of multivariate Gaussians such that $(X|Y = 0) \sim \mathcal{N}(\mu_0, \Sigma_0)$ and $(X|Y = 1) \sim \mathcal{N}(\mu_1, \Sigma_1)$, where μ_i are the means of the conditional Gaussian distributions, and Σ_i are the covariance matrices of the conditional Gaussian distributions.

$$X = (x_1, x_2) \tag{1}$$

$$\mu_0 = \mu_1 = (0, 0) \tag{2}$$

$$\Sigma_0 = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} var(x_1|Y=0) & cov(x_1, x_2|Y=0) \\ cov(x_2, x_1|Y=0) & var(x_2|Y=0) \end{bmatrix} \tag{3}$$

$$\Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} var(x_1|Y=1) & cov(x_1, x_2|Y=1) \\ cov(x_2, x_1|Y=1) & var(x_2|Y=1) \end{bmatrix} \tag{4}$$

A. How many non-zero principal components of X are there?

B. What is the overall covariance of X?

C. What are the eigenvectors and eigenvalues of $\text{cov}(X)$?

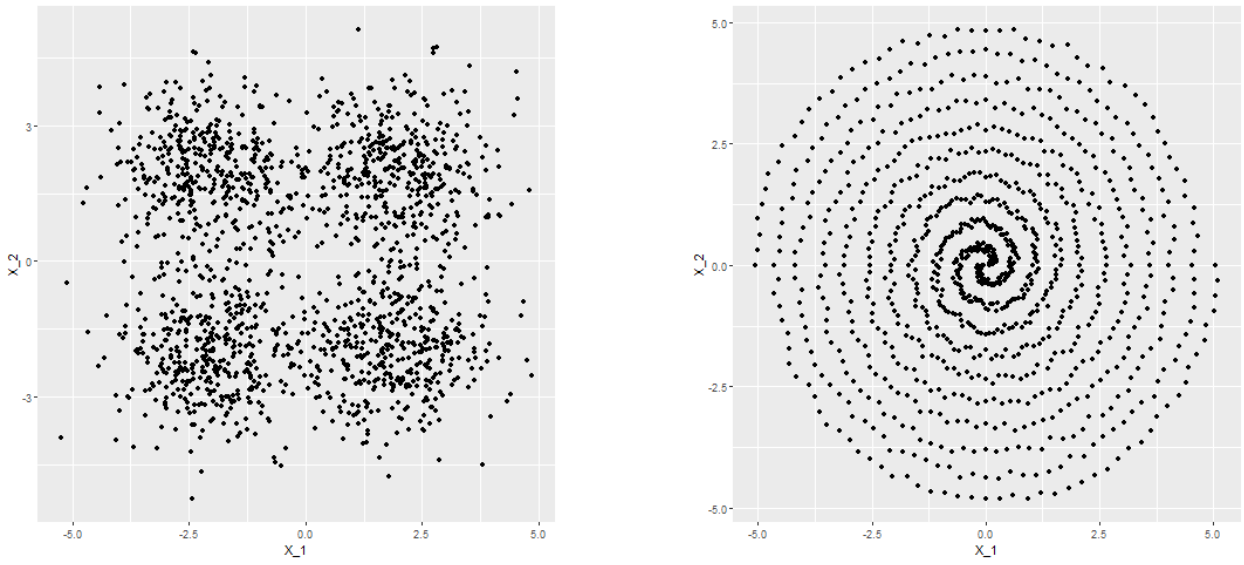
D. For $n \in \{10, 100, 1000, 10^4\}$, sample from X, and plot the points, along with the principal components vectors scaled by their eigenvalues (compute the principal components from the sample data).

E. Show that as $n \to \infty$, the eigenvectors and eigenvalues of the sample covariance matrix converge to the truth (i.e. Σ).

F. What is the significance of this convergence?

2 K-means

Consider the following two datasets.



(a) gmm.csv

(b) swissroll.csv

Figure 1: Datasets

A. Partition each dataset using k-means clustering, and plot the resulting clusters, color-coding data points by cluster.

B. Provide justification for your choice of k for each dataset.

C. How well does k-means perform on each dataset? Explain in detail. How could you improve clustering performance in any cases where it performs poorly?