Q. 1) The two plots below, represent decompositions of the features used in the K65R study discussed in class (the HIV resistance study).

(a) **One of the decompositions is an ICA decomposition, the other is PCA. Which is which? How can you tell?**

(b) **Describe the two decompositions, ICA and PCA in detail (within reason). What assumptions do the two decompositions make?**

---

**Figure 1:** (a) First decomposition, (b) second decomposition.
Q. 2) While working in a consulting centre, a client comes to you with the following problem:

“I am studying disease A in a certain population, and have managed to collect $p$ features, $X$, on each of a number of subjects, and we have $k$ different outcomes, $Y$, related to the disease. The feature of most interest is a categorical variable with two levels related to presence or absence of a certain genetic mutation.”

(a) Describe how your client might go about testing whether the $k$ outcomes differ, on average, between the two groups (based on presence or absence of this mutation). You can ignore the other features for the moment.

(b) Describe a technique to your client that may help them to find interesting combinations of features and outcomes. What sort of information / output does this technique provide?

(c) Suppose none of the features are associated in any way with the outcomes, can you detect this with this technique?

(d) Your client tells you

“A colleague of mine mentioned a technique called PCA. What if I run PCA on my features $X$ and my outcomes $Y$ separately then I regress the $Y$ scores of its most important principal component onto the $X$ scores of its most important principal component. Won’t I get the same thing?”

Will the client get the same result?
Q. 3) Consider a two-class classification on \( \mathbb{R}^k \) problem with densities

\[
f_1 = N(\mu_1, \Sigma), \quad f_2 = N(\mu_2, \Sigma)
\]

and class membership probabilities \( \pi = P(\text{class 1}) = 1 - P(\text{class 2}) \).

This model can be constructed hierarchically:

- generate \( L \sim \text{Bernoulli}(\pi) \)
- if \( L = 1 \): then generate \( X \sim N(\mu_1, \Sigma) \)
- else: generate \( X \sim N(\mu_2, \Sigma) \).

(a) Compute

\[
P(L = 1|X)
\]

and show that it has a logistic form, i.e.

\[
\frac{P(L = 1|X)}{P(L = 0|X)} = e^{\alpha + \beta'X}
\]

(b) Express \( \alpha, \beta \) in terms of \( \mu_1, \mu_2, \Sigma, \pi \).

(c) The logistic regression solution yields a direction vector \( \hat{\beta} \) that can be used as a discriminant function / direction. Compare the sample Bayesian discriminant rule with the logistic regression discriminant function. Are they the same?

(d) Suppose now that the covariance matrix was not the same in each group, i.e.

\[
f_1 = N(\mu_1, \Sigma_1), \quad f_2 = N(\mu_2, \Sigma_2)
\]

Does the probability \( P(L = 1|X) \) still have a logistic form?

(e) In LDA we saw that it is sometimes desirable to penalize the discriminant functions / directions \( \beta \). For instance, if the \( X \)'s are images, we might impose a roughness penalty. How might you penalize the directions \( \beta \) in logistic regression?
Q. 4) Suppose your next client in the consulting centre comes in with a problem from biophysics about molecular structure.

“Given a pair of molecules \( M_i, M_j \), we have computed dissimilarities / distances between the two molecules using some computer application that is considered the ‘gold standard’ in our field. We have these dissimilarities for \( n \) different molecules, and we would like to see if there is any interesting structure in our population of molecules. For instance, is there a natural grouping of the molecules based on these dissimilarities?”

(a) Describe some techniques your client might use to discover interesting groupings of the molecules.

(b) The client then asks:

“We actually have some labels for these molecules based on other studies. Can you help us come up with a rule for classifying a new protein into these classes based on its dissimilarities with the \( n \) existing proteins we have?”

Is there a natural way to create discriminant functions for use in some sort of LDA technique in this situation?

(c) What if your client told you

“In fact, we’ve been working on approximating this gold standard ‘distance’ by using some features of each molecule to come up with our own distance function

\[
d_{\text{approx}}(M_i, M_j)
\]

which is some symmetric function based on the features of molecules \( M_i \) and \( M_j \). It seems that this \( d_{\text{approx}} \) approximates the ‘gold standard’ really well.”

Describe how you might use \( d_{\text{approx}} \) to create discriminant functions for classification.
Q. 5) One of the best known applications of the EM algorithm is its application to finite mixture models. Here is perhaps the simplest example. We observe IID samples from

\[ X \sim \pi \cdot N(\mu_1, \sigma^2) + (1 - \pi) \cdot N(\mu_2, \sigma^2). \]

That is,

\[
P(X \in [a, b]) = \pi \int_a^b \frac{e^{-(x-\mu_1)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \, dx + (1 - \pi) \int_a^b \frac{e^{-(x-\mu_2)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \, dx.
\]

The goal is to estimate \((\pi, \mu_1, \mu_2)\) – we’ll assume \(\sigma^2\) is known for simplicity.

As for all EM algorithms, this algorithm is iterative so it is defined by a sequence of estimates \((\hat{\pi}^{(k)}, \hat{\mu}_1^{(k)}, \hat{\mu}_2^{(k)})\).

The algorithm goes as follows, based on an observed sample \((x_1, \ldots, x_n)\)

(i) Define the “responsibilities” at iteration \(k\)

\[
\gamma_i^{(k)} = \frac{\hat{\pi}^{(k)} e^{-(x_i-\hat{\mu}_1^{(k)})^2/2\sigma^2}}{\hat{\pi}^{(k)} e^{-(x_i-\hat{\mu}_1^{(k)})^2/2\sigma^2} + (1 - \hat{\pi}^{(k)}) e^{-(x_i-\hat{\mu}_2^{(k)})^2/2\sigma^2}}.
\]

(ii) Update the current estimates of the means \((\mu_1, \mu_2)\) with the weighted means:

\[
\hat{\mu}_1^{(k+1)} = \frac{\sum_{i=1}^n \gamma_i^{(k)} x_i}{\sum_{i=1}^n \gamma_i^{(k)}}
\]
\[
\hat{\mu}_2^{(k+1)} = \frac{\sum_{i=1}^n (1 - \gamma_i^{(k)}) x_i}{\sum_{i=1}^n (1 - \gamma_i^{(k)})}
\]

and of \(\pi\)

\[
\hat{\pi}^{(k+1)} = \frac{1}{n} \sum_{i=1}^n \gamma_i^{(k)}. \]

(a) Write out the “incomplete” log-likelihood for \((\pi, \mu_1, \mu_2)\) based on the sample \(X = (x_1, \ldots, x_n)\).

(b) What are the unobserved random variables / vectors used in this application of the EM algorithm? (Hint: this should remind you of Q. 3)

(c) Let \(Z = (z_1, \ldots, z_n)\) be the unobserved random vectors. Write out the “complete” log-likelihood for \((\pi, \mu_1, \mu_2)\) based on based on the “complete” sample \((X, Z) = ((x_1, z_1), \ldots, (x_n, z_n))\).

(d) (Bonus: 10 %) Derive the above algorithm starting using the missing variables introduced in (b).