Preguntas de control: Métodos Multivariados Primavera 2017 ITAM

1. *Este ejercicio es para que refresquen su álgebra lineal.*
   (a) Escribe un pequeño ensayo de uno a tres párrafos e ilustra gráficamente para un caso bidimensional la conexión que hay entre la matriz de covarianzas de una nube de datos $\Sigma = \Sigma(X)$, su descomposición espectral $\Sigma = P\Lambda P^\top$, y la curva de nivel uno de su forma cuadrática, i.e., $\{x : x^\top \Sigma x = 1\}$.
   (b) Describe la interpretación geométrica de la descomposición SVD y diagonalización como producto de transformaciones por matrices ortogonales y diagonales.
   (c) ¿Qué relación hay entre la descomposición SVD de una matriz A y la diagonalización de su matriz de covarianzas (suponiendo que A es centrada) (Hint: $\Sigma(A) = \frac{1}{n} A^\top A$)?

2. *Este ejercicio es para que usen su álgebra lineal en PCA.*
   (a) Explica la técnica de PCA desde el punto de vista desarrollado en los incisos anterior.
   (b) ¿Cuál es la relación entre la matriz de loadings en PCA y las correlaciones entre las variables originales y las componentes principales?
   (c) En PCA, ¿cómo interpretan el significado de las componentes principales? ¿Qué es la rotación varimax?

3. *Este ejercicio es para que piensen en la geometría subyacente de algunos métodos como análisis de correspondencias.*
   (a) Describe los tipos de datos estadísticos y algunas formas usuales de medir distancias entre ellos.
   (b) ¿De qué manera medimos distancias entre perfiles fila y columna cuando usamos la técnica de Análisis de Correspondencias?
   (c) Si se decide medir distancias de la forma
   $$\text{dist}^2(x, y) = (x - y)^\top W (x - y)$$
   usando una matriz diagonal con entradas positivas W, ¿cómo cambiarían su medida de media, varianza y covarianza empírica, ¿cómo cambia la geometría subyacente)
   (d) ¿Por qué es razonable la distancia χ^2 entre frecuencias de datos categóricos? Busca un ejemplo para mostrar que es más natural.

4. *En este ejercicio harán un repaso de la idea y técnica del escalamiento multidimensional.*
   (a) En un párrafo describe el *objetivo* (no el procedimiento) del *Classical Multidimensional Scaling.*
   (b) Demuestra que si X es una matriz de datos *centrada* con observaciones de n individuos y D^2 es la matriz de distancias euclidianas entre los n individuos, entonces
   $$XX^\top = -\frac{1}{2} K_n D^2 K_n$$
   donde K_n es la matriz centradora de $n \times n$ definida como $K_n = I_n - \frac{1}{n} \mathbb{I}_n \mathbb{I}_n^\top$ con I_n la matriz identidad y $\mathbb{I}_n$ el vector de unos ($\mathbb{I}_n \mathbb{I}_n^\top$ es una matriz de $n \times n$ con todas sus entradas igual a uno).
   (c) Usen el dataset *eurodist* en la librería *datasets* de R para reconstruir un mapa de las ciudades usando *Classical MDS* (pueden usar funciones o paqueterías ya hechas en R, e.g., cmdscale).

Interpreta el mapa. ¿Por qué aparece rotado? Matemáticamente, ¿cuál es la causa? (*Hint:* invarianza bajo multiplicación por matrices ortogonales).

5. *Este ejercicio pretende que hagan un repaso de la idea de los métodos que expusieron sus compañeros en el curso u otros temas adicionales.* Describe en *una oración* el objetivo de las siguientes técnicas:
    1. Correlaciones policóricas y poliseriales
    2. Análisis Factorial (no PCA) y *Structural Equation Modeling*
    3. PCA dentro de una regresión lineal e interpretación de los coeficientes de regresión con diagonalización
    4. *Item Response Theory*
    5. *Multiple Correspondence Analysis*
    6. *Canonical Correlation Analysis*
    7. *Mixed Factor Analysis*
    8. *Canonical Correlations*

6. *Planear un problema una investigación de análisis multivariado con las técnicas que hemos visto en clase. El objetivo es ver que tengan presentes las técnicas que hemos visto al hacer un análisis real. No necesariamente tienen que programar el uso de la técnica, solo plantear cómo la usarían y que resultados y posibles conclusiones esperarían. La pregunta tiene valor de comprensión. Si hay alguna técnica que hubiera sido lógico usar y no la mencionan afecta a la evaluación de la pregunta. Este proyecto debería servir de base para su proyecto final. Básense en los siguientes lineamientos:*

- Escojan un dataset que les interese a ustedes y al que le quieran dar seguimiento con las restantes técnicas que veremos en el curso. Si no saben qué base de datos usar, utilicen *adults* de UCI.
- Describan el dataset, las variables y su tipo de dato estadístico (no de R).
- ¿Cuáles son los retos desde el punto de vista del análisis multivariado?
- ¿Qué hipótesis quisieran contestarse con sus datos? ¿Qué técnicas del curso podrían usarse para cada problema de investigación?
- Para cada técnica que mencionen deben contestar por qué la técnica es apropiada para resolver su hipótesis y plantear sus posibles resultados y conclusiones.
- ¿Qué técnicas que no hemos estudiado en nuestro temario podrían combinarse con técnicas del curso para solucionar las hipótesis que les interesan?