Nota: 5 / 5 ejercicios
2 / 5 análisis estadístico
Total: 7 / 10

Reconocimiento Estadístico de Patrones
Tarea 2

Randy Osbaldo Ibarra Cayo

B. Preguntas Cortas

Pregunta 1 **Nota: 2 / 2**

Vimos en clase que para centrar una matriz de datos $\mathbb{X}$, basta calcular $\mathbb{X}_c = \mathbb{C}\mathbb{X}$ con $\mathbb{C} = (\mathbb{I} - \frac{1}{n}11^t)$.

- Muestra directamente (i.e, sin apoyarte en algunas propiedades particular de algebra matricial) que si $\mathbb{C} = \mathbb{C}\mathbb{C}$ (es idempotente) entonces $\mathbb{C}$ solamente puede tener valores propios 1 o 0.
Solución: Supongamos λ valor propio de $\mathbb{C}$ con v un vector propio asociado, entonces

$$\mathbb{C}v = \lambda v$$

Por otro lado, al ser $\mathbb{C}$ idempotente

$$\begin{aligned}\mathbb{C}v &= \mathbb{C}\mathbb{C}v \\ &= \lambda\mathbb{C}v \\ &= \lambda\lambda v \\ &= \lambda^2 v\end{aligned}$$

entonces

$$\lambda v = \lambda^2 v$$ Ok

esto si y solo si

$$\lambda(\lambda v - v) = 0$$

Luego, la igualdad se cumple sólo si $\lambda = 0$ o bien si $\lambda v - v = 0$, esto es $\lambda = 1$. Concluimos que si $\mathbb{C} = \mathbb{C}\mathbb{C}$ entonces $\mathbb{C}$ solamente puede tener valores propios 1 o 0. **Bien! (+0.5)**
¿Cuál es el vector propio con valor propio 0?
Respuesta: Se tiene que si v es un vector propio asociado al valor propio 0, entonces $\mathbb{C}v = 0v = 0$, es decir, v es solución al sistema

$$\mathbb{C}v = 0$$

Por otro lado

$$\begin{aligned}Cv &= (\mathbb{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)v \\ &= \mathbb{I}v - \frac{1}{n}\mathbf{1}\mathbf{1}^T v \\ &= v - \frac{1}{n}\mathbf{1}\sum_i v_i \\ &= v - \mathbf{1}\frac{1}{n}\sum_i v_i\end{aligned}$$

entonces $v - \mathbf{1}\frac{1}{n}\sum_i v_i = 0$, es decir $v = \mathbf{1}\frac{1}{n}\sum_i v_i$. Entonces las entradas del vector v son todas iguales, es decir $v_i = c$ para cada $i = 1, ..., n$, y para algún valor c, en particular tomamos $c = \frac{1}{\sqrt{n}}$, de tal forma que

$$\begin{aligned} v &= [v_1, ..., v_n]^T \\ &= [\frac{1}{\sqrt{n}}, ..., \frac{1}{\sqrt{n}}]^T \\ &= \frac{1}{\sqrt{n}}\mathbf{1} \end{aligned}$$

Ok

Bien! (+0.5)

y entonces $\|v\| = 1$. Concluimos que un vector propio asociado al valor propio 0 es $v = \frac{1}{\sqrt{n}}\mathbf{1}$ (O simplemente $v = \mathbf{1}$).

- Muestra que lo anterior implica que cualquier vector propio v con valor propio 1 de $\mathbb{X}_c\mathbb{X}_c^t$ está centrado, es decir:

$$\sum_i v_i = 0$$

donde v_i refiere a la entrada i del vector v. Hint:$\sum_i v_i = 0$ significa que $< v, 1 >= 0$.
Solución: Supongamos v vector propio de $\mathbb{X}_c\mathbb{X}_c^t$ con valor propio 1. Entonces

$$\mathbb{X}_c\mathbb{X}_c^t v = (1)v = v$$

Ahora bien, notemos que

$$\begin{aligned} \sum_i v_i = v^T\mathbf{1} = (\mathbb{X}_c\mathbb{X}_c^T v)^T\mathbf{1} \\ &= v^T\mathbb{X}_c\mathbb{X}_c^T\mathbf{1} \\ &= v^T(\mathbb{C}\mathbb{X})(\mathbb{C}\mathbb{X})^T\mathbf{1} \\ &= v^T\mathbb{C}\mathbb{X}\mathbb{X}^T\mathbb{C}^T\mathbf{1} \end{aligned}$$

Dado que $\mathbb{C}$ es simétrica

$$v^T\mathbb{C}\mathbb{X}\mathbb{X}^T\mathbb{C}^T\mathbf{1} = v^T\mathbb{C}\mathbb{X}\mathbb{X}^T\mathbb{C}\mathbf{1}$$

Pero el vector $\mathbf{1}$ es un vector propio de $\mathbb{C}$ asociado al valor propio 0, entonces

$$\begin{aligned} v^T\mathbb{C}\mathbb{X}\mathbb{X}^T\mathbb{C}\mathbf{1} &= v^T\mathbb{C}\mathbb{X}\mathbb{X}^T\mathbf{0} \\ &= 0 \end{aligned}$$

Ok

por lo tanto $\sum_i v_i = 0$. Concluimos que cualquier vector propio v con valor propio 1 de $\mathbb{X}_c\mathbb{X}_c^t$ está centrado.

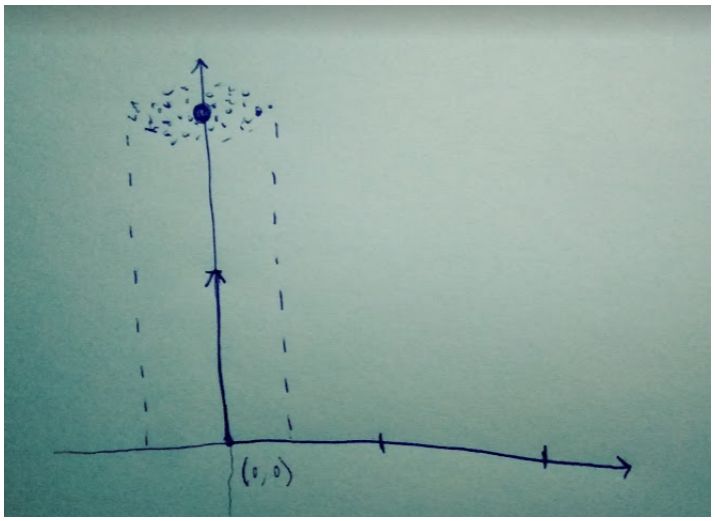
Bien! (+1)

Pregunta 2 **Nota: 3 / 3**

Si los datos están centrados la solución con SVD coincide con la solución de PCA usando el estimador muestral para la matriz de covarianza. Lo demostramos en el pizarrón para $p = 1$.

- ¿Qué pasa si uno no centra los datos, i.e. SVD sobre $\mathbb{X}$, o equivalente, PCA usando $\mathbb{X}^t\mathbb{X}$ como si fuera que $E(X) = 0$? Da un argumento intuitivo por qué el primer vector propio muchas veces será el vector del promedio (no demostrar formalmente; basta apoyarte en un dibujo con datos en 2D).
  Solución: Cuando hacemos PCA buscamos direcciones que maximicen la media de la distancia al centroide. Entonces, si suponemos que la media es cero y los datos que se tienen son **lejanos al cero (Media muestral lejana a cero)**, al proyectar sobre una recta en dirección al centroide la varianza muestral será muy alta, pues los valores de proyección serán parecidos al valor de la distancia del centroide al origen. Por ejemplo en la siguiente imagen muestra puntos lejanos al origen

  De hecho, se puede mostrar que la media maximiza la varianza de las proyecciones.



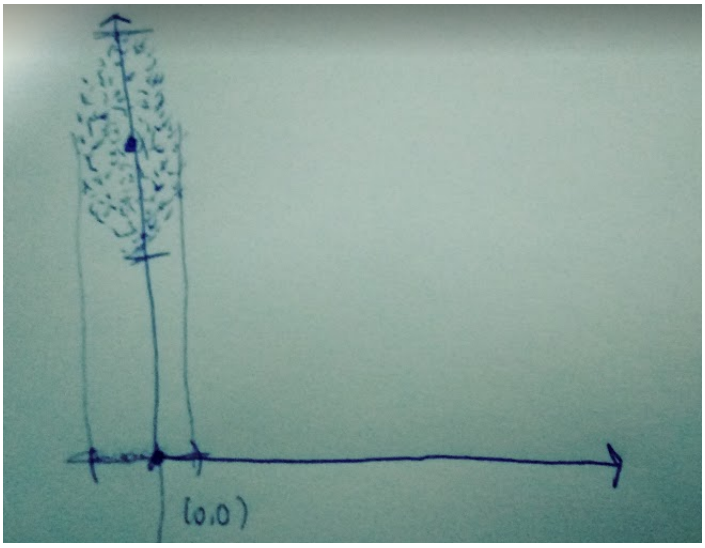
Ok

Figure 1: Caption

Podemos notar que si proyectamos sobre el eje y los valores serán más altos (Más lejos que la supuesta media de 0) que los valores obtenidos al proyectar sobre el eje x. Para este caso la dirección de máxima variabilidad pasa por la media y la segunda componente es la que se hubiese obtenido al centrar los datos y obtener la primera componente principal.

Ok

- ¿Crees que el segundo componente principal al no centrar coincide con el primer componente principal obtenido al centrar los datos? (no demostrar formalmente; basta apoyarte en un dibujo).
  Solución: Tomemos en cuenta el siguiente dibujo.



Bien!

Figure 2: Caption

Por el inciso anterior sabemos que la primera componente está en dirección a la media muestral. Luego, dado que restringimos la segunda componente a ser ortogonal a la primera componente podemos observar que esta no corresponde a la primera componente principal con datos centrados. Para este ejemplo, la primera componente calculada con datos no centrados corresponde a la primera componente calculada con datos centrados.

Por otro lado, también se pueden construir ejemplos en los cuales la segunda componente de los datos no centrados corresponde a la primera componente calculada con datos centrados (Como el ejemplo del inciso anterior).

Ok

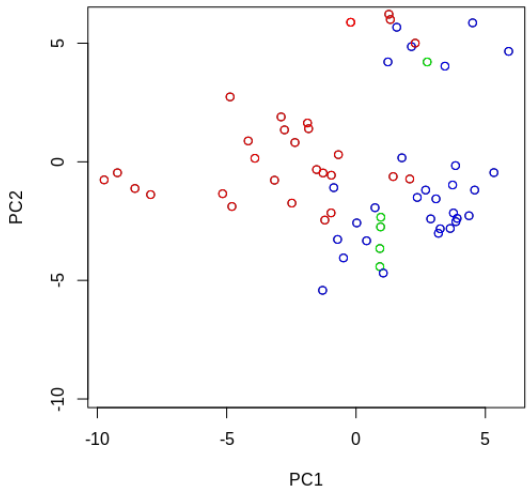
B. Análisis de Datos

Problema 1 **Nota: 2 / 2.5**

- Reconstruye el ejemplo de los libros de Oz de la clase. Se pueden bajar los libros en txt desde

`https://oz.fandom.com/wiki/Category:Full_text_Oz_books`

Solución: Se tomaron en cuenta 13 libros, 6 de los cuales los escribió Thompson, 6 escritos por Baun y 1 escrito por los dos. Estos libros fueron usados como Corpus. Cada uno de los archivos de texto se dividió en 5, para obtener un total de 65 muestras de texto. Estas muestras de texto las representamos como un vector con ayuda del enfoque Bolsa de Palabras. Sólo se tomaron en cuenta las palabras con frecuencia mayor a 1000. Se obtuvieron un total de 58 bolsas de palabras que representan fragmentos de texto de cada libro, que podemos pensar como 58 muestras de texto de un autor u otro. Entonces se tienen 65 muestras de texto representadas por bolsas de palabras, donde las palabras tomadas en cuenta son las 58 que tienen una frecuencia mayor a 1000 en el Corpus. A continuación se presenta una gráfica de las proyecciones de estos vectores en las primeras dos componentes principales



Ok

Figure 3: Caption

Los puntos en color rojo son los fragmentos de texto tomados de libros escritos por Baum, mientras que los puntos azules son los fragmentos de texto tomados de libros escritos por Thompson. Notemos como los fragentos de ambos autores se agrupan en partes opuestas del plano. Por otro lado, los puntos en verde pertenecen a fragmentos tomados del libro escrito por ambos. Podemos notar que estos puntos se encuentran en la "frontera" de la zona de cada autor.
Comentarios

–

- Construye un ejemplo propio usando otros textos.

Interesante. Creo que convendría tratar de explicar, por qué ocure este fenómeno de los puntos en verde se sitúan en la frontera. Sería interesante hacer un análisis cronológico para ver si el estilo de Thompson fue cambiando con el tiempo (desde parecerse a Baum, hasta un estilo propio).

Por otro lado, creo que falta hacer un poco más de análisis estadístico, estudiar distribuciones, frecuencias, % de variabilidad explicada (no sólo el biplot), para que el análisis sea más completo.

Falta esta parte: 0 / 2.5