Estudio del grado de marginación por medio de análisis de componentes principales y k medias

Equipo 3

3/12/2020

Objetivo

Se usará análisis de componentes principales para reducir la dimensión de los 9 indicadores socioeconómicos que forman al índice de marginación, una vez se tenga la reducción de dimensiones

recrear el índice de marginación del Consejo Nacional de la Población. Además se realizaran agrupaciones por el método de clusters jerárquicos con las variables originales y las obtenidas por el método de componentes principales.

Índice de marginación CONAPO

Desde 1990 el CONAPO emprendió esfuerzos sistemáticos para construir indicadores con el objetivo de analizar las desventajas sociales de la población e identificar con precisión espacios mayormente marginados, diferenciándolos según su intensidad de carencias, el resultado fue el índice de marginación.

El indice de marginación es un parámetro estadístico que contribuye a la identificación de sectores del país que carecen de oportunidades para su desarrollo y de la capacidad para encontrarlas o generarlas.

Indicadores socioeconómicos del índice de marginación

- Educación:
  - Porcentaje de población analfabeta
  - Porcentaje de población sin primaria completa
- Vivienda:
  - Porcentaje de ocupantes en viviendas sin agua entubada
  - Porcentaje de ocupantes en viviendas sin drenaje ni servicio sanitario exclusivo
  - Porcentaje de ocupantes en viviendas sin drenaje ni excusado
  - Porcentaje de ocupantes en viviendas con algún nivel de hacinamiento
  - Porcentaje de ocupantes en viviendas sin energía eléctrica
  - Porcentaje de ocupantes en viviendas con piso de tierra
- Distribución de la población:
  - Porcentaje de población en localidades con menos de 5,000 habitantes
- Ingresos:
  - Porcentaje de población ocupada con ingresos de hasta dos salarios mínimos

Análisis de componentes principales

Definición y origen de los componentes principales

Sea x un vector de p variables de donde se desea estudiar la estructura de las covarianzas o correlaciones entre las variables.

El análisis de componentes principales se enfoca en las varianzas de las variables, aunque no ignora las correlaciones ni covarianzas de estas. El primer paso es obtener una función lineal de la forma $\alpha_1^{'}x$ de los elementos de x que explique la varianza máxima, donde α_1 es un vector de p constantes $\alpha_{11}, \alpha_{12}, ..., \alpha_{1p}$, es decir:

$$\alpha^{'}1x = \alpha 11x_1 + \alpha_{12}x_2 + ... + \alpha_{1p}x_p = \sum_{i=1}^{p} \alpha_{1j}x_j,$$

Después se debe de obtener otra función lineal $\alpha^{'}2x$ no correlacionada con la función $\alpha 1^{'}x$ que también explique la máxima varianza posible. Así, sucesivamente se irán obteniendo estas funciones lineales hasta obtener la k-ésima función lineal $\alpha^{'}kx$ que al igual que las $k-1$ funciones lineales obtenidas anteriormente, que explique la máxima varianza posible, siendo ésta no correlacionada con $\alpha^{'}1x, \alpha^{'}2x, ...\alpha^{'}k-1x$. La k-ésima función corresponde al k-ésimo componente principal. Se pueden encontrar hasta p componentes principales pero se desea, en general, que la máxima varianza se encunetre en m componentes componentes principales siendo $m \leq p$. El caso más simple es cuando se tienen dos componentes principales, pues las observaciones puedes ser visualizadas en un plano. Cuando se tienen más componentes principales, la visualización de éstos se dificulta. A lo más se pueden visualizar claramente hasta 3 dimensiones, es decir, hasta 3 componentes. Cuando hay más se podrían visualizar planos de nivel pero se perdería la interpretación

Ya definidos los componentes principales, debemos saber como encontrarlos. Consideremos que el vector de variables aleatorias X tiene matriz de covarianzas Σ conocida. Esta matriz, cuyo (i,j)-ésimo elemento corresponde a la covarianza entre los elementos i y j de X cuando $i \neq j$, cuando $i = j$ el elemento $(i,j) = (i,i)$ corresponde a la varianza. Cuando Σ es desconocida, se reemplaza por la matriz S. Para $k = 1, 2, ..., p$ el k-ésimo componente principal se define de la forma $z_k = \alpha_k^{'}x$, donde α_k es el eigenvector o vector propio de Σ correspondinete al k-ésimo eigenvalor o valor propio mas grande λ_k.

Propiedades de los componentes principales y sus implicaciones

Sea z el vector cuyo k-ésimo elemento es z_k correspondiente al k-ésimo componente principal para $k = 1, 2, ..., p$ entonces

$$z = A'x,$$

en donde A es la matriz ortogonal cuya k-ésima columna, α_k, es el k-ésimo vector propio de Σ. Así, los componentes principales son definidos de una transformación lineal ortonormal de X.

¿Cuántos componentes principales seleccionar?

En esta sección se presentarán las reglas para decidir con cuántos componentes principales se deberían conservar para retener la mayor varianza posible en x (o en las variables estandarizadas x^* en el caso de estudiar componentes principales con la matriz de correlación).

Porcentaje acumulado de la varianza total

El criterio más común para seleccionar el número de componentes principales es seleccionando el porcentaje acumulado de la varianza total. Normalmente se desea que los componentes principales seleccionados contribuyan entre 80% y 90%. El número requerido de componentes principales es el valor más pequeño m para el cual este porcentaje se ha excedido. Este criterio es válido para la matriz de covarianza y correlación.

Valores propios / Varianzas

Como se ha mencionado anteriormente, los valores propios miden la cantidad de varianza retenida por cada componente principal. Los valores propios son grandes para los primeros componentes principales y pequeños para los componentes subsecuentes. Esto es debido a que los primeros componentes corresponden a las direcciones con la máxima varianza en el conjunto de datos.

Los valores propios pueden ser usados para determinar el número de componentes principales:

- Un valor propio > 1 indica que el componente principal considera más varianza que la que considera la variable original estandarizada. Esto es utilizado como punto de corte para retener los componentes principales cuyos valores propios sean mayor a 1. Esto es cierto solamente cuando los datos están estandarizados.

Desafortunadamente no hay forma objetiva para decidir cuántos componentes principales son suficientes. Esto depende de cada conjunto de datos con el que se esté trabajando. En la práctica, se tiende a analizar los primeros componentes principales para encontrar patrones interesantes en los datos.

Estandarización de los datos

En el análisis de componentes principales las variables se suelen escalar (estandarizar). Esto es recomendable cuando las variables son medidas en diferentes escalas, si no se realiza la estandarización, los resultados del análisis se verán afectados. El objetivo es hacer las variables comparables. Generalmente las variables se estandarizan para que éstas cuenten con desviación estándar 1 y media cero. También se recomienda escalar los datos cuando la media y desviación estándar de las variables es muy diferente.

Cuando se escalan las variables, éstas se transforman de la siguiente manera:

$$\frac{x_i - media(x)}{desv(x)},$$

Método de Rotaciones de Jacobi para matrices simétricas

Este método produce una secuencia de transformaciones ortogonales de la forma $J_k^T A J_k$ con el objetivo de hacer "más diagonal" a la matriz $A \in \mathbb{R}^{n \times n}$.

Si la matriz A es simétrica y J_0 es una transformación de Jacobi, entonces el esquema iterativo $A_{k+1} = (J_0 J_1 ... J_k)^T A J_0 J_1 ... J_k)$ converge a una matriz diagonal en la que se encuentran los eigenvalores de A.

El algortimo para matrices simétricas es:

Dados A simétrica y $tol > 0$ definimos $A_0 = A, Q_0 = I_n$, repetimos el siguiente bloque para $K = 0, 1, 2, ...$

1) Elegir un par de índices $(idx1, idx2)$
2) Calcular las entradas $cos(\theta)$, $sen(\theta)$ de la matriz de rotación J_k.
3) $A_{k+1} = J_k^T A_k J_k$
4) $Q_{k+1} = Q_k J_k$

La matriz J_k se utiliza para eliminar un par de entradas (simétricas) en la matriz A_k, esto preserva la simetría de la matriz original. En las columnas de la matriz Q_k se encuentran aproximaciones a los eigenvectores de A y en la diagonal de A_k se tienen aproximaciones a los eigenvalores de A.

K-medias

El clustering de particiones son métodos de agrupamiento para clasificar observaciones dentro del conjunto de datos en multiples grupos basados en su similaridad, ya que se desconoce de manera inicial tanto el número total de grupos, como en donde se encuentra cada observación en particular. Los algoritmos requieren que el analista especifique el número de grupos que se desea generar.

K-medias es el algoritmo no supervisado más común para particionar datos dados en un conjunto de k grupos, es decir, k clusters, en donde k representa el número de grupos pre especificados por el analista. Éste algoritmo clasifica objetos en múltiples grupos tal que los objetos dentro del mismo grupo son lo más similares posibles (alta similaridad dentro de las clases), mientras que los objetos de diferentes grupos son lo más disimilares posible (baja similaridad entre clases). En k medias, cada cluster es representado por su centro (centroide) el cual corresponde a la media de los puntos asignados a ese grupo.

La idea básica detrás de k-medias consiste en definir grupos de tal forma que la variación total dentro de cada grupo (conocida como variación total entre clusters) se minimice.

Existen varios algoritmos para k-medias. El algoritmo estándar es el algoritmo *Hartigan-Wong (1979)*, el cual define la variación total dentro de cada grupo como la suma de las distancias euclideanas al cuadrado entre los objetos y el centroide correspondiente.

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2,$$

donde:

- x_i representa un punto de los datos que pertenece al grupo C_k
- μ_k el valor medio de los puntos asignados al grupo C_k

Cada observación (x_i) es asignada a un grupo tal que la suma de cuadrados (SC) entre cada observación al centroide μ_k del grupo sea mínima en comparación a la distancia que llegase a tener con μ_j que es el centro del grupo j (con $j \neq k$).

La suma de cuadrados de la variación total dentro de cada conglomerado mide la compacidad (calidad) del agrupamiento y queremos que sea lo más pequeña posible.

Algoritmo

El primer paso para usar el algoritmo de k-medias es indicar el número de conglomerados (k) que se generarán en la solución final.

El algoritmo comienza con una selección aleatoria de k observaciones del conjunto de datos, estos servirán como los centros iniciales de los grupos. Las observaciones seleccionadas son también conocidas como las medias de los conglomerados o centroides.

El siguiente paso es asignar cada una de las observaciones restantes al centroide más cercano a ellas, esto se define utilizando la distancia euclideana entre la observación y el centroide. Este paso se llama *asignación de grupo*.

Después del paso de asignación , el algoritmo computa el nuevo valor medio de cada grupo. El término *actualización del centroide* es utilizado para diseñar este paso. Ahora que los centros han sido recalculados, cada observación se revisa nuevamente con el fin de verificar si éstas pueden estar cercanas a un grupo diferente. Todas las observaciones son reasignadas nuevamente usando las medias actualizadas de cada grupo

La asignación de grupo y la actualización del centroide son pasos iterativos hasta que la asignación de grupo deje de cambiar (hasta que se logra una convergencia). Esto es, los grupos formados en la iteración actual son los mismo que se obtuvieron en la iteración previa.

Ventajas y desventajas de k-medias

K-medias es un algoritmo simple y rápido. Puede trabajar eficientemente con grandes conjuntos de datos. Sin embargo, existen ciertas desventajas, incluyendo:

- Asume conocimiento previo de los datos y requiere que el analista elija el número apropiado de grupos.
- Los resultados finales son sensibles a la selección inicial aleatoria de centrides. Esto es un problema debido a que en cada iteración del algoritmo se seleccionará (con alta frecuencia) diferentes centroides iniciales aunque se esté trabajando con un sólo conjunto de datos. Esto puede dirigir a diferentes resultados de conglomerados en diferentes iteraciones.
- Es sensible con valores atípicos.
- Si se reacomodan los datos, es posible que se obtenga una solución diferente cada que se ordenen los datos.

Implementación

Se analizaron los indicadores socioeconómicos de los años 1990, 2000 y 2015, realizamos una reducción de dimensiones por medio de análisis de componentes principales. Para poder realizar esto se hizo uso del algoritmo de método de rotaciones de Jacobi estudiado en clase.

```
  def sign(x):
      """
      Helper function for computing sign of real number x.
      """
      if x >=0:
      return 1
      else:
      return -1

def compute_cos_sin_Jacobi_rotation(Ak, idx1, idx2):
    """
    Helper function for computing entries of Jacobi rotation.
    Args:
      Ak (numpy ndarray): Matrix of iteration k in Jacobi rotation method.
      idx1 (int): index for rows in Jacobi rotation matrix.
      idx2 (int): index for columns in Jacobi rotation matrix.
    Returns:
      c (float): value of cos of theta for Jacobi rotation matrix.
      s (float): value of sin of theta for Jacobi rotation matrix.
    """
    if np.abs(Ak[idx1,idx2]) > np.finfo(float).eps:
      tau = (Ak[idx2, idx2] - Ak[idx1, idx1])/(2*Ak[idx1, idx2])
      t_star = sign(tau)/(np.abs(tau) + np.sqrt(1+tau**2))
      c = 1/np.sqrt(1+t_star**2)
      s = c*t_star
    else: #no rotation is performed
      c = 1
      s = 0
    return (c,s)


def compute_Jacobi_rotation(Ak, idx1, idx2):
  """
  Compute Jacobi rotation matrix.
  Args:
      Ak (numpy ndarray): Matrix of iteration k in Jacobi rotation method.
      idx1 (int): index for rows in Jacobi rotation matrix.
      idx2 (int): index for columns in Jacobi rotation matrix.
  Returns:
      J (numpy ndarray): Jacobi rotation matrix.
  """
  c,s = compute_cos_sin_Jacobi_rotation(Ak, idx1, idx2)
  m,n = Ak.shape
  J = np.eye(m)
  J[idx1, idx1] = J[idx2, idx2] = c
  J[idx1, idx2] = s
  J[idx2, idx1] = -s
  return J
```

Para estos tres años se tuvieron que realizar 3 sweeps con las siguientes iteraciones utilizando la matriz de varianzas y covarianzas de las variables estandarizadas:

```
(1,2),(1,3),(1,4),(1,5),(1,6),(1,7), (1,8),(1,9),

 (2,3),(2,4),(2,5), (2,6), (2,7), (2,8), (2,9)

 (3,4), (3,5), (3,6), (3,7), (3,9), (3,9),

 (4,5), (4,6), (4,7), (4,8), (4,9),

 (5,6), (5,7), (5,8), (5,9),

 (6,7), (6,8), (6,9),

 (7,8), (7,9),

 (8,9)
```

Una vez realizados los sweeps obtenemos la matriz A diagonalizada, en esta matriz se encuentran los valores propios de la matriz de covarianzas. Decidimos comparar los valores propios obtenidos por nuestra función con los valores propios obtenidos por la función de `Python np.linalg.eig()`. Observamos que nuestros resultados son los mismos a los valores propios proporcionados a la función de `numpy`, por lo que utilizamos los vectores propios de esta misma función.

Lo siguiente que realizamos fue ordenar de manera descendente los valores propios, de tal forma que la matriz de vectores propios contenga el vector propio cuyo valor propio es el más alto, y así sucesivamente. Para la creación de la matriz de PCA realizamos la multiplicación entre la matriz de vectores propios y los datos que contienen a los indicadores socioeconómicos estandarizados. Así obtenemos una matrix de $32x9$.

Posteriormente se analizó la varianza explicada por estas nuevas dimensiones. El porcentaje de varianza explicada se realizó de la siguiente manera en `Python`:

```
explained_variances = []
for i in range (len(eig)):
    explained_variances.append(eig[i]/np.sum(eig))
print(explained_variances)
```

Y finalmente se interpretó gráficamente el comportamiento de los primeros dos componentes principales coloreados por el grado de marginación. En la siguiente sección del reporte se mostrarán los resultados obtenidos.

Para la clasificación de estos estados se utilizarán K-medias seleccionando 5 grupos que hacen referencia a los grados de marginación `Muy bajo`, `Bajo`, `Medio`, `Alto` y `Muy alto`. Por medio de matrices de confusiones analizaremos si la matriz de componentes principales agrupa mejor a los estados.

Resultados

Año 1990

K-medias utilizando los indicadores socioeconóminos

PCA

K-medias utilizando componentes principales

Año 2000

K-medias utilizando los indicadores socioeconóminos

PCA

K-medias utilizando componentes principales

Año 2015

K-medias utilizando los indicadores socioeconóminos

PCA

K-medias utilizando componentes principales

Conclusiones

Referencias

- Kassambara, Alboukadel (2017) Practical Guide To Principal Component Methods in R: PCA, M (CA), FAMD, MFA, HCPC, factoextra. Volumen 2. STHDA.
- I.T. Jolliffe (2002) Principal Component Analysis. Second Edition. Volumen 2. Springer.
- CONAPO (2010) Índice de marginación por Entidad Federativa. Consejo Nacional de Población México, DF.
- Notas del curso