UIVERSIDAD DE EL SALVADOR
FACULTAD MULTIDISCIPLINARIA
DE OCCIDENTE

Departamento de Matemática

Título del proyecto

Análisis multivariado aplicando componentes principales al caso de los desplazados

Autor:

Mayra Cañas y Jacquelinne Gutiérrez

03 junio de 2024

Tabla de contenidos

Introducción

La humanidad en su evolución necesita conocer los fenómenos que están a su alrededor porque éstos afectan su desarrollo dentro de todos los ámbitos (fenómenos de tipo social, económico, tecnológico, físico, etc.). Este conocimiento se logra mediante la construcción de modelos que puedan reproducir y explicar dichos fenómenos. Por tal motivo, es necesario que los profesionales, directivos e investigadores en las distintas áreas del saber estén familiarizados con las herramientas necesarias para la construcción y adecuación de modelos. Una de las herramientas más importantes para llevar a cabo este objetivo es la estadística, y en particular, muy a menudo, la estadística multivariada.

Según Peña y Dallas existen diversas definiciones acerca de las técnicas de análisis de datos multivariados, pero los dos coinciden en a conceptualizarla como "una herramienta que tiene como objetivo principal resumir grandes cantidades de datos por medio de pocos parámetros (simplificación), además busca encontrar relaciones entre:

Variables de respuesta

Unidades experimentales

Variables de respuesta y unidades experimentales

Según Peña, la mayoría de problemas que requieren la aplicación de la estadística exigen el tratamiento de muchos factores o variables y que por esto las técnicas del análisis de datos multivariados constituyen una herramienta poderosa para la toma de decisiones en las diferentes disciplinas, pues dan respuesta a necesidades palpables y plenamente identificables. Según Pérez, se puede observar que cuando existen muchas variables es posible que parte importante de la información sea redundante, en cuyo caso es necesario eliminar el exceso y dejar sólo variables que tengan representatividad dentro del conjunto. Esto se consigue con la aplicación de las técnicas multivariantes de reducción de la dimensión: análisis de componentes principales, factorial, correspondencias, escalamiento óptimo, homogeneidades, análisis conjunto.

Las técnicas multivariadas más utilizadas en el análisis de datos son: análisis de componentes principales; análisis factorial; análisis de clasificación entre los que se encuentran: discriminante, regresión logística y clúster; análisis multivariado de la varianza, y análisis de variables canónicas.

Con este artículo se desean integrar conocimientos teóricos y prácticos a través de la comprensión de las componentes principales, como una de las técnicas estadísticas que permiten estudiar la información que se dispone antes de entrar en el uso de los otros métodos que abordan el análisis de datos multivariados.

Por ser tan amplio el tema, este artículo sólo trata del análisis de componentes principales debido a su importancia dentro del desarrollo de las diversas técnicas de análisis de datos multivariados.

Capítulo 1

Componentes Principales

Siguiendo a autores como Peña y Bramardi, el análisis de componentes principales (ACP) es una técnica estadística propuesta a principios del siglo XX por Hotelling (1933) quien se basó en los trabajos de Karl Pearson (1901) y en las investigaciones sobre ajustes ortogonales por mínimos cuadrados. Interpretando la definición de diversos autores, se puede decir que el ACP es una técnica estadística de análisis multivariado que permite seccionar la información contenida en un conjunto de p variables de interés en m nuevas variables independientes. Cada una explica una parte específica de la información y mediante combinación lineal de las variables originales otorgan la posibilidad de resumir la información, total en pocas componentes que reducen la dimensión del problema.

La mayor aplicación del ACP está centrada en la de reducción de la dimensión del espacio de los datos, en hacer descripciones sintéticas y en simplificar el problema que se estudia.

Para Peña, el ACP tiene una utilidad doble; por un lado, permite hacer representaciones de los datos originales en un espacio de dimensión pequeña y, por el otro, transformar las variables originales correladas en nuevas variables incorreladas que puedan ser interpretadas.

El ACP también se emplea con frecuencia cuando se desea dividir las unidades experimentales en subgrupos de acuerdo con la similaridad de los mismos. Igualmente, es útil para transformar un conjunto de variables respuesta correlacionadas en un conjunto de componentes no correlacionados, bajo el criterio de máxima variabilidad acumulada y, por tanto, de mínima pérdida de información.

Otra aplicación es el cribado, el cual permite el seguimiento sobre los componentes principales obtenidos para comprobar hipótesis establecidas en un estudio de análisis de datos multivariados y para identificar datos atípicos en el conjunto de datos.

De igual manera, García y Gil afirman que el ACP es un criterio fundamental para hacer conjeturas sobre el numero de factores que se deben determinar en el análisis factorial y para probar si, en realidad, un grupo de variables $p > 2$ cae dentro de un espacio de dos o tres dimensiones que permita ser observado dentro del análisis de clúster.

Pérez anota que el análisis de componentes principales es en muchas ocasiones un paso previo a otros análisis, en los que se sustituye el conjunto de variables originales por las componentes obtenidas. Éste siempre debe hacerse cuando se quiera obtener modelos en los que sea necesario el uso de las variables originales como explicativas para tratar con algunos problemas presentes, como la independencia.

Según Gil, en el análisis discriminante cuando se tienen menos observaciones que variables y es difícil encontrar nuevas observaciones, el ACP es útil para determinar un menor número de variables que resuma la máxima variabilidad de las originales y con las cuales se pueda construir la matriz de varianza-covarianza, de tal forma que sea invertible y permita elaborar una regla de discriminación necesaria para clasificar nuevas observaciones.

Finalmente, el ACP se usa como base para determinar si ocurre multicolinealidad entre variables predictoras en el análisis de regresión múltiple. Entendiéndose como multicolinealidad cuando en dos o más variables existe redundancia; esto es, la información de una o más variables ya está explicada en otra(s) variable(s) (véase por ejemplo, Peña, Dallas).

Notaciones y Símbolos

Siguiendo la simbología común de diversos autores, a continuación se presentan conceptos básicos del álgebra de matrices que son necesarios en el ACP.

Matriz de variable respuesta

La base para la utilización del ACP es la estructura de correlación (interdependencia) entre las variables cuantitativas definidas en una población, en donde cada individuo queda definido en términos de las mismas. La matriz de variable respuesta de doble entrada X está compuesta por filas que representan las unidades experimentales Ir, r=1,2,….,n y las columnas, por las variables Xj, j= 1,2,…., p, como se muestra a continuación:

Figura 1

$$X = \begin{array}{c} \\ I_1 \\ I_2 \\ I_r \\ \vdots \\ I_n \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & \cdots & X_p \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ x_{r1} & x_{r2} & \cdots & x_{rp} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n \times p} \end{array}$$

x_{rj}: Valores de la r-ésima unidad experimental en la j-ésima variable respuesta

p : Cantidad de variables

n : Individuos o unidades experimentales sobre la cual se están midiendo las variables

Vectores de datos

Con el fin de tener un lenguaje común en los procesos de ACP, en adelante, los vectores siempre serán columnas a o X, etc., y la transpuesta de un vector cualquiera, por ejemplo a, se simboliza por a'.

Vectores de medias y matrices de varianza covarianza

La media de un vector X de variables aleatorias se denota por μ, definido por:

Figura 2

$$\mu = E(X) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}.$$

La matriz de covarianza de X se denota por Σ, donde:

Figura 3

$$\sum = Cov(X) = E\left[(X-\mu)(X-\mu)'\right] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

Con

$$\sigma_{jj} = Var(X_j) = E[X_j - \mu_j]^2, \quad \text{para } j = 1,2,...,p, \quad \text{y}$$

$$\sigma_{ij} = Cov(X_i, X_j) = E\left[(X_i - \mu_i)(X_j - \mu_j)\right], \text{ para } i \neq j = 1,2,...,p,$$

Correlación y matriz de correlación

El coeficiente de correlación entre X_i y X_j se denota por:

Figura 4

$$\rho = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}$$

****Matrices ortogonales unitarias****

Dentro del álgebra de matrices las rotaciones de un espacio vectorial son transformaciones lineales del espacio vectorial sobre sí mismo y están asociadas con matrices cuadradas, unitarias y ortogonales. Una matriz de éstas, Q, tiene tantas filas y columnas como sea la dimensión del espacio. Sus columnas son vectores unitarios (es decir, de longitud igual a la unidad) y tiene la particularidad de que al ser multiplicada por su transpuesta produce la matriz unidad. En otras palabras, $Q^{-1} = Q'$. En cambio, las traslaciones no son transformaciones lineales pero tienen la propiedad de no modificar la variabilidad de la nube de puntos. Es decir, las varianzas y covarianzas en la nube son las mismas antes y después de una traslación. Lo expuesto anteriormente, junto con algunas propiedades de la matriz de varianzas covarianzas Σ, constituye las bases sobre las cuales descansa la técnica de componentes principales.

PLANTEAMIENTO Y SOLUCIÓN DEL PROBLEMA DE LOS COMPONENTES PRINCIPALES

El ACP es una técnica descriptiva; sin embargo, no niega la posibilidad de que también pueda ser utilizado con fines de inferencia. Por otra parte, las aplicaciones del ACP son numerosas y entre ellas se pueden citar la clasificación de individuos, la comparación de poblaciones, la estratificación multivariada, entre otras. En el ACP se maneja un número p ($p \quad 2$) de variables numéricas. Si cada variable se representa sobre un eje, se necesitaría un sistema de coordenadas rectangulares con p ejes perpendiculares entre sí para ubicar las coordenadas de los puntos y poderlos dibujar. Cuando $p \quad 4$, para el ser humano es imposible hacer la representación gráfica. En estos casos el ACP permite buscar un nuevo sistema de coordenadas con origen en el centro de gravedad de la nube de puntos, de tal manera que el primer eje del nuevo sistema F_1 recoja la mayor cantidad posible de variación; el segundo eje F_2, la mayor cantidad posible entre la variación restante; el tercer eje F_3 la mayor cantidad posible entre la variación

que queda después de las dos anteriores y así sucesivamente. Las Figuras 5 y 6 permiten ver la representación gráfica de dos componentes.

Figura 5

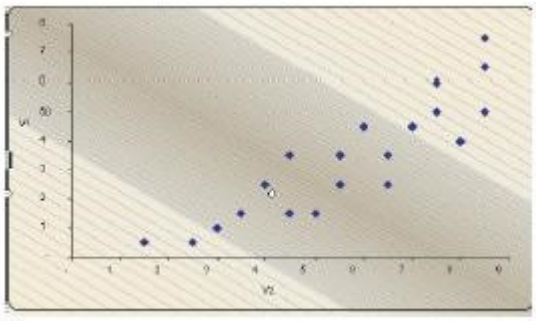

Figura 6

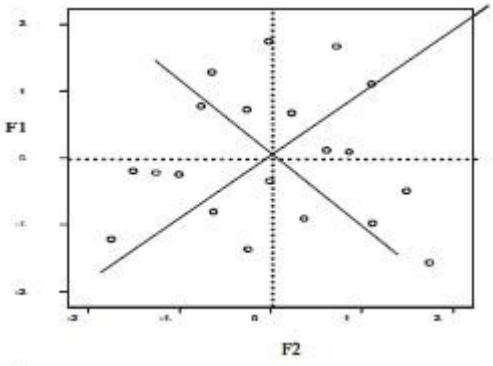

Figura 1: Observando las figuras anteriores se puede concluir que el sistema de coordenadas de la derecha se logra después de dos movimientos de la nube de puntos: el primer movimiento corresponde a una traslación que permite situar el nuevo origen en el centro de gravedad de la nube. El segundo movimiento que se hace sobre la nube centrada es una rotación, usando el centro de gravedad como punto pivote. La rotación permite ubicar los ejes en dirección horizontal y vertical como se observa en la Figura6. Esto indica que se desea encontrar un nuevo sistema de coordenadas que represente lo mejor posible los datos sin causar distorsiones, cuya forma de problema es equivalente a encontrar las nuevas variables del espacio reducido con una mínima pérdida de la información, y también a buscar un elipsoide de concentración que permita encerrar los datos originales.

Cuando ya se ha definido el problema es factible abordarlo. Según Peña [1], páginas 73-74, la matriz de varianza covarianza Σ es definida positiva, es decir,

la forma cuadrática asociada a ella tiene todas sus raíces positivas. Lo anterior hace que esta matriz tenga p valores propios reales y diferentes, lo cual garantiza que sea diagonalizable. En términos matemáticos significa que existe una matriz A ortogonal, tal que $\Sigma = ADA^{-1}$ donde D es la matriz diagonal formada por los valores propios de Σ, denotados por $_1$, $_2$,..., $_p$. Es posible reordenar de acuerdo con su magnitud los valores propios de Σ de tal manera que $_1 > _2 > ... > _p$. Esto simplemente se traduce en un reordenamiento de las columnas de la matriz A de manera que la primera sea el vector propio o componente asociado con 1, la segunda sea un vector propio asociado con 2 y así sucesivamente. En particular, dichas columnas pueden estar formadas por vectores propios normalizados, es decir, perpendiculares entre sí y de longitud igual a la unidad. De esta manera se construye una matriz que produce la rotación deseada ya que, como puede probarse, el primer vector propio $a_1 = (a_{11}, a_{12}, ... a_{1p})'$ apunta en la dirección de máxima variabilidad de la nube centrada. Esta dirección se llama primera dirección principal. El segundo vector propio $a_2 = (a_{21}, a_{22}, ... a_{1p})'$ apunta en la siguiente dirección de máxima variabilidad de la nube centrada, llamada segunda dirección principal y así sucesivamente.

Una vez resuelto el problema de la rotación, bastará multiplicar la vac c c riable centrada $Xc = X - \mu = (X_1 , X_2 , ..., X_p)$ por la matriz de rotación A para obtener la nueva variable, $Y = (Y_1, Y_2 , ... Y_p)$ llamada variable de componentes principales. Cada componente Y_i del vector aleatorio Y se llama una componente principal. Evidentemente se cumple que $Y_j = a_{j1} X_c + a_{j2} X_{c2} , ..., + a_{jp} X_{cp}$, es decir, cada componente principal es 1 una combinación lineal de las variables originales centradas. Para hacer el análisis de los autovalores se necesita desarrollar los conceptos y las propiedades que se verifican. La traza de Σ, por ser la suma de las varianzas de las variables originales Y_i recibe el nombre de varianza total, resulta claro que $traza(\Sigma) = traza (ADA^{-1}) = \Sigma _i$. Se puede probar además que $V(Y_i) = _i$ para $i = 1,2..., p$ y que $Cov(Y_i , Y_j) = 0$, con $i \hat{a} j$. Esto implica varios aspectos, a saber: La varianza total es igual a la suma de los valores propios de $_i$ e igual a la suma de las varianzas de las componentes principales. Es decir, la varianza total es la misma con las variables originales que con las variables transformadas Y_i. Las componentes principales son variables aleatorias no correlacionadas entre sí, obtenidas mediante la transformación lineal del vector de las variables originales centradas por la matriz de autovectores.

UN CASO DE APLICACIÓN DEL ANÁLISIS DE COMPONENTES PRINCIPALES

Descripción del problema

Según la ley 387 de 1997, “Es desplazado toda persona que se ha visto forzada a migrar dentro del territorio nacional abandonando su localidad de residencia

o actividades económicas habituales, porque su vida, su integridad física, su seguridad o libertad personales han sido vulneradas o se encuentran directamente amenazadas, con ocasión de cualquiera de las siguientes situaciones: Conflicto armado interno, disturbios y tensiones interiores, violencia generalizada, violaciones masivas de los Derechos Humanos, infracciones al Derecho Internacional Humanitario u otras circunstancias emanadas de las situaciones anteriores que puedan alterar o alteren drásticamente el orden público"[2].

Nuestro problema está basado en la lectura de la situación de los desplazados en un municipio de Colombia, donde se concentran el mayor porcentaje de estas personas que huyen de la violencia y el temor que generan las fuerzas oscuras en los campos del país.

Mediante entrevistas a expertos, a los mismos desplazados y la observación directa, se ha podido determinar problemas de diferente índole, tales como: la ubicación desordenada de los desplazados que han incomodado hasta llegar a roces con el personal que habita en los diferentes barrios, hacinamiento, inseguridad, y otros problemas de orden público.

Para analizar más profundamente esta problemática, los investigadores han recopilado información de fuentes (como, por ejemplo, el ministerio de Protección Social, el Sistema de Información de Hogares Desplazados por Violencia en Colombia - SISDES; el boletín sobre "Niños desplazados" editado por Codhes el 25 de octubre de 1997, entre otros) relativa a la población desplazada, con el propósito de contribuir desde la academia a ver técnicamente el problema con la ayuda del análisis de componentes principales.

Los datos de la Tabla 1 corresponden a la investigación exploratoria y estimaciones realizadas por los autores con el fin de encontrar los niveles de incidencia de los factores que conforman el problema de los desplazados en la comunidad.

Lo anterior se consigue mediante ACP, con lo cual se obtienen resultados útiles para ver más claro la gravedad del problema (véase resultados finales en esta sección). Para este estudio se han definido las variables que a continuación se nombran en los 25 lugares donde se ubican los desplazados: HPM: Horas promedio de movilidad diaria; NPM: Número promedio de desplazados por mes; NHS: Número de horas semanales que los centros de alimentación están en funcionamiento; ATR: Área total de recreación de uso común (en metros cuadrados); NBC: Número de centros del lugar de posible concentración; CCD: Cantidad de camas disponibles; NTC: Número total de cuartos; HHM: Horas-hombre mensual requeridas para atenderlos.

Tabla 1

Lugares	HPM	NPM	NHS	ATR	NBC	CCD	NTC
1	4	5	3	1,13	2	7	5
2	5	2,67	50	1,5	2	6	6
3	15,6	22,87	50	2,1	2	14	12
4	8	1,97	178	2,1	2	8	8
5	6,2	2,01	40	2,8	4	26	24
6	17,6	7,89	170	2,10	3	20	20
7	24,9	4	50	2,01	4	37	35
8	45,33	160,57	170	2,06	20	49	49
9	40,53	51,69	50	4,8	8	78	76
10	32,82	41,84	170	1,6	7	48	46
11	96,23	245,89	170	3	7	166	131
12	57,73	383,24	170	2,6	5	37	36
13	97,57	210,01	170	1,8	12	121	121
14	55,60	209,07	170	3,8	8	67	65
15	112,78	985	172	2,5	8	167	180
16	150,2	243,38	170	4,1	16	186	201
17	125,43	138,99	170	3	10	193	193
18	179,67	897	175	4,6	25	238	236
19	109,33	420	173	4,1	14	116	116
20	97,98	678,63	170	2,1	13	303	209
21	100,34	278,98	169	3,2	16	132	132
22	265,82	688,55	168	4,7	60	364	364
23	812,18	715,43	170	4,3	60	243	241
24	385,9	1569,68	168	3,7	20	541	454
25	89	379	167	3,1	10	293	195

Estos datos fueron procesados con SPSS y *Statgraphics* y se obtuvieron los resultados que aparecen a continuación, para sacar algunas conclusiones que sirven para consolidar el estudio sobre el ACP.

Estadísticos descriptivos

Tabla 2

Varaible	Media μ	Desviación típica □	Coeficiente de variación
HPM	117,4296	169,38660	144%
NPM	333,7344	392,80835	118%
NHS	139,3200	57,36486	41%
ATR	2,9120	1,08948	37%
NBC	13,5200	15,34090	113%
CCD	138,4000	134,15041	97%
NTC	126,2000	116,71725	92%

En la Tabla 2 se muestran la media, la desviación y el coeficiente de variación para cada una de las variables (análisis univariante). Estos valores permiten estimar la variable centrada tipificada Z (compárese con la Tabla 7). El objetivo de esta tipificación es homogenizar las unidades de medidas, buscando que todas pesen por igual en el análisis como se dijo anteriormente.

Matriz de correlaciones y prueba de independencia

Tabla 3

		HPM	NPM	NHS	ATR	NBC	CCD	NTC
Correlación	HPM	1,000	0,614	0,335	0,520	0,823	0,613	0,667
	NPM	0,614	1,000	0,457	0,456	0,506	0,854	0,862
	NHS	0,335	0,457	1,000	0,306	0,359	0,458	0,479
	ATR	0,520	0,456	0,306	1,000	0,606	0,527	0,605
	NBC	0,823	0,506	0,359	0,606	1,000	0,585	0,674
	CCD	0,613	0,854	0,458	0,527	0,585	1,000	0,978
	NTC	0,667	0,862	0,479	0,605	0,674	0,978	1,000

Determinante = 0,000469886

Determinante=0,000469886

El tener determinante bajo y coeficiente de correlaciones relativamente altas entre las variables originales es un buen indicador para utilizar la técnica de componentes principales que ayuda a resumir las variables en pocas dimensiones cuando se hace este tipo de análisis. Esto se debe a que las correlaciones altas implican dependencia lineal entre las variables, dando lugar a que se puedan explicar con un número menor de variables llamadas componentes principales Yi . Todo lo anterior, y suponiendo normalidad de los datos, se puede corroborar

con la prueba de independencia que se muestra en la siguiente tabla (p-valor=0 es menor que 0,05 y KMO es próximo a 1):

Tabla 4

Medida de adecuación muestral de Kaiser-Meyer-Olkin		,775
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	159,646
	Grados de libertad	21
	p-valor	,000