Decimocuarta Semana Internacional de la Estadística y la Probabilidad
14-18 de junio de 2021

Introducción al Análisis de Componentes Principales.

Fabiola Blanco Infanson[a], Fernando Velasco Luna[b], Francisco Solano Tajonar Sanabria[c], Víctor H. Vázquez Guevara[d]
[a,b,c,d] *Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias Físico Matemáticas, Puebla, Puebla, México.*
[a]*fibbi1894@gmail.com*, [b]*fvelasco@fcfm.buap.mx*, [c]*ftajonar@fcfm.buap.mx* , [d]*vvazquez@fcfm.buap.mx*,

Resumen

En la literatura tradicional de Estadística podemos encontrar métodos desarrollados en una variable, sin embargo en la vida real, se tiene que los eventos por lo general implican, varias características de interés, es decir, varias variables aleatorias. En ocasiones los investigadores tienen o recolectan información de un gran número de variables, lo cual dificulta su posible interpretación para dar solución al problema que ocupa su interés. Por lo regular cuando se tiene un gran número de variables estás se encuentran relacionadas, la técnica de componentes Principales crea un conjunto de variables no correlacionadas. En este trabajo se presenta una breve introducción a la técnica del Análisis de Componentes Principales. Se presenta un ejemplo típico en cualquier curso de analisis multivariado, el ejemplo de las medidas de los pájaros.

Palabra claves: Análisis Multivariado, Combinación Lineal, Variables Correlaciondas.

Introducción

En la literatura tradicional de Estadística podemos encontrar métodos desarrollados en una variable, sin embargo en la vida real, se tiene que los eventos por lo general implican, varias características de interés, es decir, varias variables aleatorias. Por ejemplo un investigador de mercados podrá querer identificar las características de los individuos que le permitirán determinar, si es probable que determinada persona compre un producto específico, o en el caso de un agrónomo podrá interesarse en la resistencia de nuevas variedades de trigo y la resistencia que tienen estas a la sequía y los insectos. En ocasiones los investigadores tienen o recolectan información de un gran número de variables, lo cual dificulta su posible interpretación para dar solución al problema que ocupa su interés. Por lo regular cuando se tiene un gran número de variables estás se encuentran relacionadas por lo cual es deseable reducir el número de variables pero sin perder la información de las originales. Una técnica estadística para reducir el número de variables originales a un conjunto menor, es la técnica de componentes principales (ACP) por sus siglas en español (Análisis de Componentes Principales).

Podemos decir que los objetivos del análisis de componentes principales son: 1) reducir dimensionalidad de los datos, y 2) encontrar nuevas variables importantes subyacentes.

Aunque autores como Araneo D., Tenko Raykov entre otros, confirma que el objetivo 1 no siempre se llega a obtener, se obtiene la relación entre las variables, es decir la dimensionalidad de los datos y en el caso del objetivo 2 aunque no se consiga, que las nuevas variables sean significativas es decir que tengan una interpretación, estás todavía pueden ser útiles, por diversos motivos, como el cribado de datos y verificación de las agrupaciones.

En las ciencias del comportamiento, sociales y educativas, el ACP tiene una historia relativamente larga de aplicaciones en el desarrollo de pruebas objetivas para medir habilidades específicas como motivación, personalidad, inteligencias y otras construcciones relacionadas o dimensiones latentes (no observables, ocultas). En estas aplicaciones, se comienza típicamente con un gran conjunto de medidas destinadas a evaluar esas construcciones. Luego se emplea la técnica ACP en los datos obtenidos con el objetivo de reducir su extensión a unos pocos componentes significativos que representan medidas "puras" de las dimensiones o variables latentes subyacentes.

El ACP no sólo logra la reducción de variables, sino que también el resultado puede ser usado en aplicaciones de otros métodos estadísticos multivariados (análisis de varianza o análisis de regresión).

En conclusión este método brinda la posibilidad de analizar la dimensión de conjuntos de datos y la relación que hay entre ellos, lo que nos brinda un mundo de posibilidades, ya que es una técnica exploratoria muy útil, en la aplicación de otros métodos estadísticos, y en algunos casos ayudará a la reducción de la dimensión del problema a costa de una pequeña pérdida de información. Aunque hay una tendencia a interpretar las nuevas variables recién creadas, esto no siempre sucede, son pocos casos pero hay que recordar que aunque esto no suceda el análisis de componentes principales es muy útil.

El análisis de componentes principales es una técnica estadística de análisis multivariado que se emplea para extraer información relevante de un conjunto inicial de variables correlacionadas transformándolas en variables las cuales no estarán correlacionadas, con el objeto de identificar patrones y estructuras. Esta técnica es utilizada por diversos investigadores de múltiples áreas de estudio.

Bajo M., utiliza la técnica de ACP para determinar los principales factores de riesgo de la curva de rendimientos, con un énfasis especial en la gestión activa de carteras de renta _ja, donde proporciona el enfoque del gestor de carteras o practitioner mediante el análisis se pudo reducir el estudio de un número elevado de parámetros, (como son los tipos de interés) a un conjunto reducido de componentes que representan los principales factores de riesgo a los que se enfrenta el gestor. Además que el ACP reveló que la decisión de apuesta por un mercado alcista o bajista de tipos es mucho más importante, en términos de retorno, que aquellas decisiones de valor relativo sobre distintas zonas de la ETTI (Estructura Temporal de Interés) como pueden ser posiciones en pendiente (attening o steepening) o estrategias de inversión basadas en el análisis de la curvatura (butteries).

Mesa-Ramos L.,*et al* utilizaron la técnica para el análisis del proceso de fermentación de un anticuerpo monoclonal, en el Centro de Inmunología Molecular (La Habana, Cuba) se produce un anticuerpo monoclonal terapéutico que ha encontrado una efectiva aplicación en el tratamiento de pacientes aquejados de cáncer de cabeza y cuello. Dada la gran variabilidad que ha tenido la concentración de este anticuerpo en la etapa de fermentación industrial de la planta donde es producido, se hizo necesaria la aplicación de una técnica de analisis multivariante como el Análisis de Componentes Principales, con el fin de reducir la dimensionalidad de los datos y de explicar las principales fuentes de variabilidad del proceso.

Ávila H. *et al* aplicaron el análisis de componentes principales con el objeto de conocer las interrelaciones entre las variables analizadas que determinan el grado de alteración del agua de la Laguna de Coyuca de Benítez, en Guerrero, Mexico. El ACP arrojo que las variables Temperatura, pH y Oxígeno disuelto presentaron mayores interrelaciones en este sistema lentico.

Descripción de la técnica

El análisis de componentes principales es una técnica estadística cuyo objetivo principal es transformar un conjunto de *p* variables correlacionadas a otras nuevas variables cuyo número es menor que *p* y están no correlacionadas.

En la mayoría de los textos se manejan dos objetivos del análisis de componentes principales, el anteriormente expuesto y el segundo es identificar nuevas variables significativas subyacentes. Aunque siempre se identificarán nuevas variables, no se puede garantizar que tengan una interpretación, sin embargo aunque no tengan interpretación, estas variables serán útiles para diversas cosas, como cribado de datos y verificación de agrupaciones.

Supóngase que existe una muestra con *n* individuos cada uno con *p* variables, es decir, el punto de partida es un conjunto de *p* variables

$$x_1, x_2, \dots, x_p$$

Y la idea es obtener un conjunto de *p* nuevas variables $y_1, y_2, \dots, y_p$ las cuales sean combinaciones lineales de las variables originales $x_1, x_2, \dots, x_p$, es decir,

$$y_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p$$
$$y_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_{np}$$

$$y_p = a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pp}x_{np}$$

Estas nuevas variables tienen la propiedad de no estar correlacionadas. Para una discusión detallada del ACP ver Manly.

Ejemplo *Medidas de Pájaros*

Aspectos generales

En este ejemplo el caso que se obtuvo de la literatura analizada (véase [4]), en este caso tenemos que los datos presentados, son las medidas de 49 pájaros que estuvieron en una tormenta el 1 de febrero de 1898, donde se consideran las siguientes variables.

Tabla 1. Variables bajo estudio

Variable	Descripción	Valores
LT	**Longitud Total**	(152,165)
EA	**Extensión Alar**	(230,250)
LPC	**Longitud de Pico y Cabeza**	(30.1,33.1)
LH	**Longitud de Humerus**	(17.2,19.8)
LQE	**Longitud de la Quilla del Esternón**	(19.0,3.1)

Metodología Estadística

Se realizarán gráficas de cajas y bigotes de las variables bajo estudio con el fin de observar el comportamiento de las variables.
Como análisis definitivo se realizará un análisis de componentes principales, con el objetivo de reducir la dimensión del estudio.

Análisis estadístico

Análisis preliminar

Se realizarán gráficas de cajas y bigotes de las variables longitud total, extensión alar, longitud de pico y cabeza, longitud de humerus y longitud de la quilla del esternón, se analizarán con el fin de observar el comportamiento de las variables.

Análisis definitivo

Se realizará un análisis de componentes principales, con el objetivo de reducir la dimensión del estudio.

Resultados y Discusión

Resultados Análisis preliminar

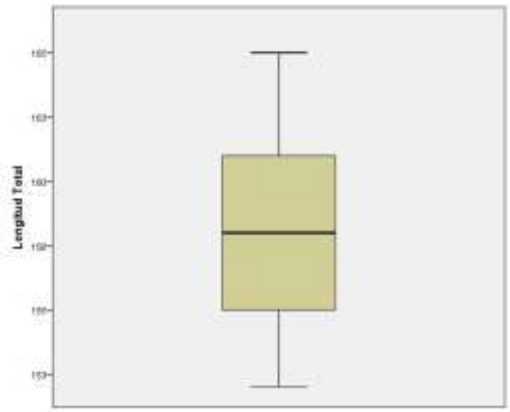

Figura 1.- Longitud Total.

De la Figura 1 se tiene que la variable de longitud total tiene una distribución asimétrica positiva, se puede observar en la Figura 1 que se encuentra una mayor dispersión en los datos cuyos valores oscilan entre 161 y 165.

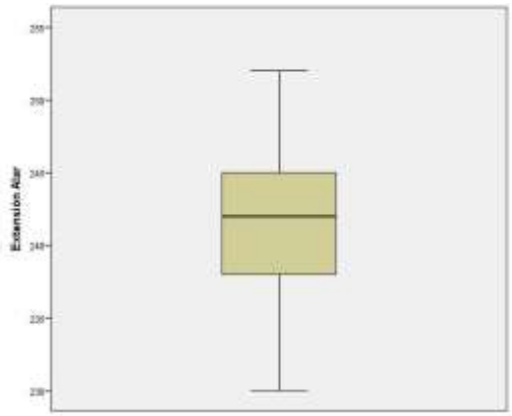

Figura 2.- Extensión Alar.

De la Figura 2 se observa que la Extensión Alar tiene una distribución sesgada a la izquierda, la menor dispersión de los datos se encuentra entre las medidas de 242 y 245.

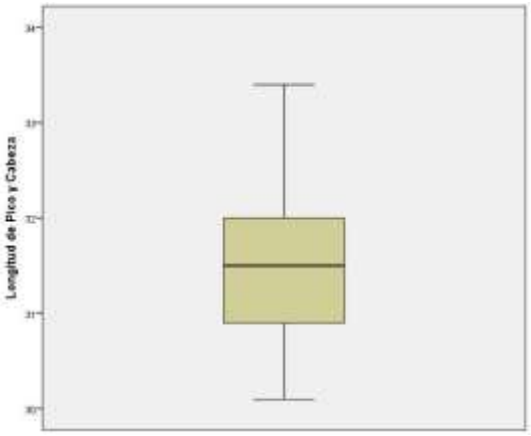

Figura 3.- Longitud de Pico y Cabeza.

De la Figura 3 se observa que variable longitud de pico y cabeza tiene una distribución asimétrica negativa, y la mayor dispersión se encuentra en el intervalo (32,33.10)

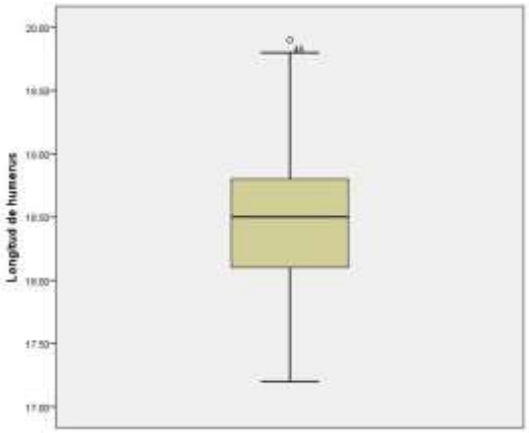

Figura 4.- Longitud de Humerus.

En la variable longitud de humerus es en la única donde se observa un punto atípico, el cual es el pájaro número 46, y se puede observar bastante variabilidad después del valor 18.8.

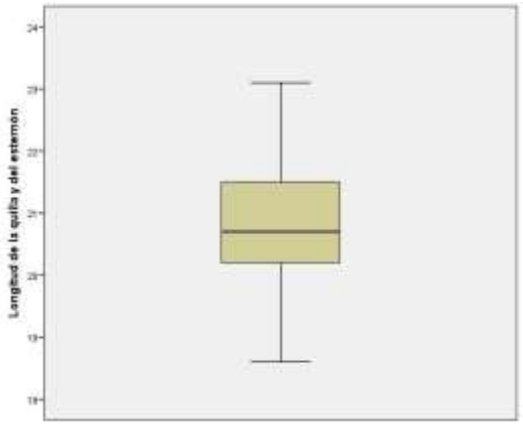

Figura 5.- Longitud de Quilla del Esternón.

Los datos de la Longitud de Quilla del esternón se concentran en los valores de 20.2-20.7, y se puede ver que los datos se mantienen dispersos en los extremos.

Resultados Análisis definitivo

Cuando uno realiza un ACP, es recomendable examinar en un principio la prueba de KMO y la prueba de esfericidad de Bartlett.

El valor obtenido en el test de KMO nos da información sobre la pertinencia del análisis. Nos dice si la correlación entre las variables es fuerte. Dado que el valor es de 0.846 se tiene que es aplicable, ya que es mayor de 0.6.

Tabla 2. Prueba de esfericidad de Bartlet

Prueba de esfericidad de Bartlet	**Aprox. chi cuadrado**	144.326
	gl	10
	Sig.	<.000

La Prueba de esfericidad de Bartlett, parte de la hipotesis nula de que las variables no estan correlacionadas entre sí.

De acuerdo a lo anterior se rechaza la hipotesis nula, lo cual nos lleva a que las variables se encuentran correlacionadas y tiene sentido aplicar el análisis de componentes principales.

Tabla 3. Comunalidades

Descripción	Valores
Longitud Total	0.749
Extensión Alar	0.817
Longitud de Pico y Cabeza	0.776
Longitud de Humerus	0.816
Longitud de la Quilla del Esternón	0.974

Las comunalidades nos dan información acerca de la proporción de la varianza de cada una de las variables originales es explicada por las componentes principales. Se observa que el modelo es capaz de reproducir más del 70% de la variabilidad de cada variable, y en el caso de la variable de longitud de la quilla del esternón es capaz de reproducir el 97 %.

Tabla 4. Varianza total

Componente	% variabilidad	% acumulado
1	71.752	71.752
2	10.884	82.635
3	7.481	90.117
4	6.150	96.267
5	3.733	100.00

El porcentaje de la varianza representa la proporción de varianza explicada por la primera componente principal, en este caso, se observa que la primera componente explica un 71.75% lo cual es más de la mitad, lo que implica que se conseguirá una reducción significativa.

Al comparar la primera componente con los demás se nota que la primera componente es por mucho la más importante, dado que la diferencia es enorme.

Interpretando los componentes principales se observa la Matriz de componentes. Los coeficientes son las cargas factoriales que

expresan la magnitud de la correlación entre la variable y el componente principal.

En la primer componente hay muy poca diferencia entre los coeficientes de las variables, la variable con el coeficiente mayor longitud de humerus con .882 y la de menor es longitud de la quilla del esternón la cual tiene .748, así que se observa que la diferencia es sólo de 0.134, al ver que la diferencia es tan reducida se puede decir que el 71.75% de la varianza de los datos está relacionada con los tamaños de los pájaros.

En la segunda componente principal se ve que el coeficiente de longitud total es tan pequeño que no afecta a esta, se puede notar que cuando las variables de longitud de humerus, extensión alar y longitud de pico y cabeza obtengan un coeficiente alto entonces la longitud de la quilla obtendrá un coeficiente bajo y así la segundo componente podría obtener una mayor proporción.

Cuando se revisa el tercer componente, se aprecia que la longitud del humerus, del pico y cabeza y de la quilla del esternón no presentan una diferencia muy amplia, sin embargo, contrastan con la extensión alar y la longitud total, lo cual como los otros componentes nos muestra que hay una diferencia de forma entre los pájaros analizados.

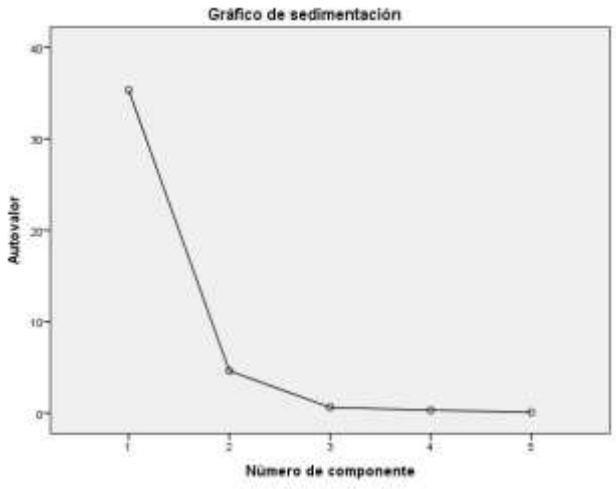

Figura 6.- Gráfico de sedimentación.

Se observa que se forma el codo con los dos primeros componentes principales, así que se puede determinar que solo se necesita ocupar las dos primeras componentes principales.

Referencias

Araneo, D. (2008). **Introducción al análisis de componentes principales**. IANIGLA - CONICET-CCT MENDOZA

Avila, H., Garcia, S., *et al*. (2015) Analisis de componentes principales, como herramienta para interrelaciones entre variables fisicoquímicas y biológicas en un ecosistema léntico de guerrero, México. **Revista Iberoamericana de Ciencias.** Vol. 2, Num. 3**,** 43-54.

Bajo, M. (2014). Aplicaciones prácticas del análisis de componentes principales en gestión de carteras de renta fija (i), determinación de los principales factores de riesgo de la curva de rendimientos. **Analisis Financiero**, 124 20-38.

Manly, B.F.J. (1986) **Multivariate Statistical methods.** Chapman and Hall, USA

Mesa, L., Gozá, O., Uranga, M., Toledo, A., and Gálvez, Y. (2018), Aplicación del análisis de componentes principales en el proceso de fermentación de un anticuerpo monoclonal. **VacciMonitor** 27, 8-15.

Raykov, T., and Marcoulides, G. (2008). **An introduction to applied multivariate analysis**. Routledge