## CAPITULO 2

## MARCO TEORICO

## Introducción al Análisis de Métodos Factoriales.

## Los métodos factoriales son un conjunto de técnicas que sirven para combinar combinar, de manera que se obtienen nuevas variables o factores que no podemos medir directamente, pero que tienen un significado.

Estos métodos factoriales se caracterizan por las representaciones gráficas, el investigador obtiene mapas estructurando todas las posiciones relativas del conjunto de filas y de columnas de la tabla que se está estudiando.

Los tipos de métodos factoriales están diseñados cada uno para un tipo diferente de tabla:

El Análisis de Componentes Principales (ACP) está diseñado para tablas de medidas o de escalas métricas, es decir para variables cuantitativas.

El Análisis Factorial de Correspondencias (AFC) estudia las tablas de contingencia o tablas de datos de frecuencias.

El Análisis de Correspondencia Múltiple (ACM) esta diseñado para estudiar los ficheros de encuestas, tablas que recogen las respuestas de los individuos a distintas variables nominales o disyuntivas completas.

Sin embargo en la práctica se pueden aplicar varios métodos a la misma tabla, con las precauciones del caso.

* + 1. **Formas Matemáticas de medir la cantidad de información.**

Los métodos factoriales utilizan alguna de las diversas formas matemáticas de medir la información. Debido a que todas las medidas de la información incluyen una medida de las distancia entre los puntos, definiremos en primer lugar el termino distancia y posteriormente las distancias más utilizadas.

* + - 1. **Distancia**

La distancia entre dos individuos o variables mide el grado de asociación o semejanza entre estas. Todas las distintas medidas de la distancia deben cumplir los siguientes axiomas:

* ∀ i,i´ dii´ ≥ 0 la distancia nunca es negativa.
* ∀ i i´ dii´= di´i la distancia es simétrica.
* ∀ i ≠ i´≠ i\* dii\* < = d i i´ + di´i\*
  + - 1. **Distancia Euclídea.**

Sean i, i´ dos individuos en los que se han observado p variables.

Sea xi y xi´  los valores que toman los individuos evaluados en cada variable.

La distancia euclídea al cuadrado es determinada por:



se utiliza en variables cuantitativas.

**2.1.2.3 Distancia X2**

Es utilizada cuando se analizan tablas de frecuencia .

Sean i, i´ dos filas y sean ki´j dos términos pertenecientes a la j-ésima columna e i-ésima fila y j-ésima columna e i´-ésima fila respectivamente.

La distancia x2 (ji cuadrado) entre i e i´ denotada por dii´ 2 de terminos kij y ki´j es expresada por la siguiente fórmula:



donde kij es la frecuencia de la asociación de i y j ; 

y ki es la frecuencia con que se ha presentado i ; 

**2.1.2.4 Medidas de la información**

Una medida de la información de una tabla de datos nxp (n individuos y p variables) es denotada por I y se defina como la suma de los cuadrados de las distancias de los individuos al origen del planoy se expresa de la siguiente forma:



Si denotamos con G a la medida de dispersión de las variables y como suele hacerse coincidir el origen con el centro de gravedad, luego la información se mide mediante:



**2.1.3 Datos y objetivos.**

Consideremos una tabla rectangular de valores numéricos formada por n filas y p columnas que representan a los individuos y a las variables respectivamente.

Serán representados mediante una matriz X nxp con elementos xij que toma la variable j para el individuo i.



........................................

xij

Variables

1..................................................xij................................p

## Esta tabla puede representarse en dos espacios

* Espacio de las variables Rp se representan los n individuos por sus coordenadas o valores que toman para cada una de las p variables.
* Espacio de los individuos Rn se representan las p variables por puntos j. Las coordenadas son los valores que toma esa variable j para casa individuo.

Cada espacio esta provisto de la distancia euclídea usual:

La distancia euclídea entre dos individuos i,i” es la suma de las variables existentes entre los valores que toman los individuos para cada variable, elevadas las diferencias al cuadrado para evitar que se compensen las positivas con las negativas.



La cantidad de información se mide por la suma de las distancias al origen del cuadrado esto es:



El objetivo del análisis es buscar un subespacio Rq, donde p representa a las variables observada o medidas y q a las variables no observadas (q<p), que se pretendan contenga la mayor cantidad posible de información existente como las p variables originales.

**2.1.4 Análisis en Rp**

Si se representa al individuo zi en el nuevo subespacio y xi en el primero, se trata de obtener el subespacio que minimice simultáneamente las distancias entre zi y xi para todos los puntos de la nube inicial y en proyección es decir , el subespacio sobre el cual la nube proyectada sea lo menos deformada posible.

Se trata de minimizar  para evitar que se compensen los valores negativos y positivos.



Como el primer termino del segundo miembro es constante, la minimización se consigue maximizando el segundo, es decir, maximizando la suma de las proyecciones al cuadrado.

La obtención del subespacio se hace mediante un proceso iterativo. Inicialmente se busca el eje que, pasando por el origen, maximice esta proyección. A continuación se busca el segundo eje que, siendo perpendicular al primero, maximice las proyecciones, y así sucesivamente.

La proyección de un punto individuo (i) sobre el nueve eje viene dado por



donde u1 es el vector normado director del eje u1´*u*´ = 1

Como cada línea de X es un punto de Rp la proyección de los n puntos será Xu1 y su cuadrado u1´ X´X u1´

Luego tenemos que buscar u1, de forma que maximice u1´ X´X u1´ sujeto a u1´u1 = 1

Planteando el langrangiano y después igualando a cero tendremos:

X´Xu1´ = λu1

En donde u1es el vector propio de X´X si multiplicamos cada termino de la ecuación anterior por u1 obtendremos:

u1´ X´Xu1 = u1

el valor de λ es el máximo buscado, luego u1 debe ser el vector propio asociado al mayor valor propio de X´X λ1.

Cada uno de los individuos tiene una proyección sobre ese nuevo eje.

Al conjunto de proyecciones se le denomina Factor, que no es más que una nueva variable conformada por una combinación lineal de las variables iniciales:

Fi (i) = x1´ u1 = 

## 2.1.5 Tipos de Análisis

La naturaleza de los datos iniciales impone la realización de algunas transformaciones previas al análisis, dando lugar a diversos tipos de análisis factorial.

Los dos más conocidos son:

1. Análisis de Componentes Principales.
2. Análisis de Correspondencia.

El primero se utiliza si la tabla está formado por variables cuantitativas y heterogéneas[[1]](#footnote-2), las cuales se tipifican antes de aplicar el análisis. El segundo es para el análisis de tablas de contingencia o tablas de frecuencias de termino general fij.

## 2.2. Análisis de Componentes Principales

El análisis de componentes principales trata de explicar la estructura de varianza-covarianza de pocas combinaciones lineales de las variables originales. Sus objetivos generales son:

1. La reducción de datos, sin la perdida de información valiosa
2. La interpretación de estos.

El Análisis de Componentes Principales es una técnica multivariada de interdependencia en la que se estudian p variables de interés que constituyen un vector aleatorio, cuyas componentes son variables aleatorias discretas o continuas, esto es:

X = (x1, x2, x3, . . . xP)

**2.2.1 Análisis en Rp**

Este análisis se utiliza para describir una matriz A (matriz de datos) de variables continuas.



A ∈ Mmxp, es decir m filas (individuos) y columnas (p variables)

Los elementos de A pueden se heterogéneos, tanto en su media como en su desviación.

Debido a que las variables en ocasiones toman valores muy altos y por ende tienen un peso muy importante en la determinación de los ejes, se realiza una transformación que consiste en centrar los datos, es decir restar la media de cada uno de los elementos, esto es:

xij = aij  -

Donde  es la media de la variable j .

Con esto se elimina la influencia del nivel general de las variables.

Ahora, si las dispersiones de las variables son muy diferentes se hará necesario dividir las variables para su desviación correspondiente, es decir ahora las variables se encontraran estandarizadas.

Luego tenemos una nueva matriz X obtenida a partir de la matriz A:



donde si i = 1...p es la desviación de cada una de las variables y el término  se introduce en la transformación con el objetivo de que el producto de la matrices XTX coincida con la matriz de correlación, Σ

El análisis consistirá en obtener los vectores propios de la matriz Σ = XTX. Las proyecciones de los individuos sobre estos vectores propios son los componentes principales, los cuales se obtienen de la siguiente manera:

Supongamos además que los valores propios característicos de Σ son:

λ1 ≥λ2 ≥λ3 ≥ . . . ≥ λp

con sus correspondientes vectores propios esto es:



se define q variables no observadas y1 ,y2 ,y3 , . . . ,yq como una combinación lineal de x1 ,x2 ,x3 , . . . ,xp, luego;

y1 = β11 x1 +β21 x2 +β31 x3 +. . . +βp1 xp

y2 = β12 x1 +β22 x2 +β32 x3 +. . . +βp2 xp

y3 = β13 x1 +β23 x2 +β33x3 +. . . +βp3 xp

.

.

.

yq = β1p x1 +β2p x1 +β3p x1 +. . . +βpp xp

que podemos expresar también de la siguiente manera:

yi = β1i x1 +β2i x1 +β3ix1 +. . . +βpi xp = uiTμ i= 1,2,3,...p

Entonces la varianza de yp denotada por VAR(yP)

VAR (yi) = VAR (uiTx) = uiTΣ ui i= 1,2,3,...p

Y la covarianza entre yi y yk para i,p = 1,2,3,...p

Cov (yi , yk) = uiTΣ uK

# Las componentes Principales de x son aquellas combinaciones lineales de esta forma construidas que son no correlacionadas y cuyas varianzas son tan grandes como sea posible.

La **primera componente principal** es la primera combinación lineal con mayor varianza, es decir que maximiza VAR (yi) = uiTΣ ui, dicha varianza puede incrementarse su a1 se multiplica por alguna constante, para eliminar esta indeterminaciones, se utilizan solo vectores unitarios.

Luego la primera componente principal se define como la combinación lineal que maximiza VAR (y1) sujeta a u1Tu1= 1

La **segunda componente principal** se define como la combinación lineal que maximiza VAR (y2) sujeta a u2Tu2= 1 y además

Cov(u1Tx, u2Tx)= 0

Siguiendo este patrón, la **k-ésima componente principal** se define como la combinación lineal que mazimiza VAR (yk) sujeta a ukTuk= 1 y además Cov(uiTx, u2Tx) = 0 para k < i

**Proporción del total de la varianza expresada por la k-ésima componente principal**

La traza de una matriz A, denotada por tr(A), se define como la suma de los elementos de la diagonal principal. Para la matriz de varianzas y covarianzas Σ, los elementos de la diagonal son las varianzas de cada una de las variables,

tr(Σ)= σ11+σ22+σ33+. . .+ σpp

La varianza total de la población = σ11+σ22+. . .+ σpp =λ1+λ2+. . .+ λp

y consecuentemente la proporción del valor total de la varianza expresada por la k-ésimo componente principal, esto es:

**Proporción del total de la varianza expresada por la k-ésima componente principal**



**2.2.2 Elección del número de ejes.**

Para elegir el número de factores que se han de retener, existen distintos métodos, sin embargo ninguno de ellos es determinante.

A continuación se detallan los métodos mas conocidos:

* **Fijar el porcentaje,** podemos fijar el porcentaje mínimo de explicación que deseamos conservar y retener el número de ejes necesarios para ello.
* **Las empíricas,** como la varianza explicada por cada eje sucesivo debe ser decreciente, se puede representar el histograma de los valores propios con los números de los ejes en ordenadas, y los porcentajes de inercia explicadas en las abscisas, se pueden eliminar los ejes cuyo número de orden es posterior al "codo" que se produce en la curva**[[2]](#footnote-3),** así por ejemplo, en el gráfico 2.1, solo se tomarían los cuatro primeros ejes.

Gráfico 2.1

Elección del Número de ejes

1 2 3 4 5 6 7 8 9

Número de Componentes

5

4

3

2

1

Porcentaje de Inercia

**2.2.3 Interpretación de los factores**

Las variables iniciales pueden tener redundancias y estar midiendo en parte la misma característica. Como se mencionó el factor es un agrupamiento de estas variables y se interpreta a partir de su correlación con las variables iniciales, dicha correlación es la proyección de la variable sobre el factor.

Si una variable esta muy correlacionada con un factor, tendrá una coordenada muy alta próxima a +1

* Si Fα(j) =+1 se puede interpretar al factor como una clasificación de los individuos a lo largo de él en orden de valores crecientes de la variable j.
* Si Fα(j) = -1 los individuos están clasificados sobre el eje en orden decreciente de los valores de j.
* Si Fα(j) = 0 entonces no existe relación entre el factor y la variable.

Mientras mayor sea el valor absoluto de Fα(j) más alta es la relación entre j y el factor α.

**2.3 Análisis Factorial de Correspondencias**

El análisis factorial de correspondencia (AFC) esta diseñado para analizar tablas de contingencias formadas por números positivos o frecuencias que son respuestas del individuo con respecto a cada variable. La tabla de frecuencias objeto del análisis tiene la forma.

cr1 cr2 . . . crp

i1 i2 . . . in

. . .

. . . kij

Donde

* Cri i = 1...p representa los criterios , es decir las p variables.
* i m m =1...n representa los individuos
* kij representa el resultado de las opiniones del individuo i en relación al criterio j, usualmente será el número de respuestas afirmativas a la presencia del criterio en el individuo, sin embargo también se puede considerar la media de notas.

**2.3.1. Notación**

La tabla de datos es una matriz K nxp donde:

* kij Representa la frecuencia de asociaciones entre los elementos i y j
* ki. Representa la suma total de la fila i
* k.j Representa la suma de la columna j
*  la suma total de la población.

El método deberá ser simétrico en relación a las filas y columnas, es decir el análisis de los individuos en el espacio Rn, será idéntico al que usara para analizar las variables en el espacio Rp

**2.3.2. Formación de nubes**

La nube de individuos es representada en el espacio Rp o espacio de las variables, donde se representa los n individuos por los valores que toman al ser evaluados en las p variables.

En Rp se toma la nube de n puntos i cuyas coordenadas son : 

La nube de p variables representada en el espacio Rn o espacio de los individuos, donde se representa las p variables por los valores que toman al evaluarse en los n individuos.

En Rn se forma la nube de p puntos j cuyas coordenadas son: 

Para reducir los cálculos a una sola factorización se trabajará con la matriz de frecuencias denotada por C, de elemento 

A pesar que el AFC trabaja con perfiles, no olvida las diferencias entre las sumas de cada línea o columna, sino que les designa un peso proporcionar a su importancia en el total.

En Rp cada punto i esta afectado de un peso  y en Rn cada punto j está afectado por un peso 

Debido a que el ACM trabaja con distribuciones o perfiles, se utiliza la distancia x2 (Ji cuadrada)

**2.3.3 Análisis en Rp**

El objetivo del ACM es obtener una representación simplificada de la nube de individuos, es decir de los puntos fila cuyas coordenadas son cij/ ci para j= 1...p

Tales puntos están afectados de un peso o masa ci. y la distancia entre ellos es medida a través de la distancia x2 (Ji cuadrada) es decir:



Se nota que la distancia. X2 entre i, i´ de este espacio es igual a la distancia euclídea entre dos puntos de coordenadas:

 y 

Luego se llega a la conclusión de que realizar el análisis considerando la distancia x2 es equivalente a realizar el análisis con la distancia euclídea de la nueva matriz C, lo que nos conduce a que se puede transformar la matriz de acuerdo al Método de Correspondencias y luego aplicar Componentes Principales, es decir se buscan los componentes principales (véase 2.3.3) de la siguiente matriz C:



Dado que el ACP es centrado, se deben centrar la matriz C, encontrando las coordenadas del centro de gravedad de la nube de puntos, considerando que cada punto tiene un peso ci.

El centro de gravedad denotado por g es igual a :



al simplificar los términos obtenemos:, lo que nos conduce a la nueva forma de la matriz C.



A esta matriz C se le aplica el Análisis de Componentes Principales.

**2.4 Análisis de Correspondencia Múltiple**

El análisis de correspondencia múltiple se realiza con datos que se encuentre en forma de tablas disyuntivas completas. Son tablas de variables cualitativas, y en el caso de tener variables cuantitativas es posible convertirlas a cualitativas dividiendo su intervalo de variación en clases de equivalencias sucesivas.

**2.4.1 Notación**

Se puede describir una tabla disyuntiva completa Z de la siguiente manera:

* Un conjunto de individuos I= 1,...,i,...,n
* Un conjunto de variables o preguntas J1,..., Jk,..., JQ
* Un conjunto de modalidades para cada pregunta I,... mk

El número total de modalidades; 





es la tabla I X J.

El elemento zij  puede tomar el valor de 0 o 1 según lo que el individuo i haya elegido la modalidad j o no.

**2.4.2 Objetivo del análisis**

Objetivo del análisis es obtener una representación simultánea, en un espacio de dimensión reducida Rq, de :

Las modalidades de todas las preguntas

Se trata de estudiar las relaciones entre todas las modalidades, no entre las preguntas.

**2.4.3 Método del Análisis de Correspondencias Múltiples.**

Una tabla disyuntiva completa puede ser considerada como una yuxtaposición de tablas de contingencia y, por tanto, debe analizarse mediante un análisis Factorial de Correspondencia, obteniendo una representación simultánea de todas las modalidades (columnas) y de los individuos.

A continuación mostramos las particularidades de un Análisis Factorial de correspondencia aplicado a una tabla disyuntiva completa.

**2.4.4 Matriz H.**

Primero definiremos la terminología que se usará:

* Los elementos de Z, zij = kij son 1 o 0
* Qel número de preguntas.
* fij / fi = kij / ki =1/Q ,es el inverso del número de preguntas o 0 según si el individuo haya elegido o no la modalidad j
*  es el número de individuos que poseen la modalidad j.

Para obtener los factores es necesario diagonalizar la matriz obtenida a través de la recolección de información (como se detalla en la Gráfico 2.1) y que para este caso en particular se convierte en:

 D-1 B

Donde B = ZTZ, la matriz D es una matriz diagonal constituida con las sumas o totales por modalidad o columna, es decir el número de unos en cada columna.

A la matriz H se le aplica el método de Componentes Principales (véase 2.3.3).

1. De naturaleza diferente, por ejemplo un análisis de variables de diferente tipo de medida: toneladas, miles, litros. [↑](#footnote-ref-2)
2. Tomado del libro "Métodos multivariantes para la investigación comercial", página 46 [↑](#footnote-ref-3)