ANÁLISIS DE COMPONENTES PRINCIPALES

APRENDIZAJE AUTOMÁTICO NO SUPERVISADO

Guillermo de Jorge Botana

Dpto. Psicobiología y Metodología en Ciencias del Comportamiento

Facultad de Psicología.

Universidad Complutense de Madrid.

NOTA:

El contenido de este texto corresponde a uno de los temas de una asignatura del Máster de Metodología de las Ciencias del Comportamiento y de la Salud. Está elaborado para tener un texto base de lo que es explicado en clase. Aunque el texto es seguido y coherente, puede ser susceptible de algunas mejoras y ampliaciones. No obstante, es lo suficientemente autocontenido para llevar a cabo un estudio independiente sobre él.

He decidido publicar este texto fuera del ámbito de la asignatura por si puede resultar de utilidad para otros estudiantes o por si a otros docentes les puede facilitar la tarea.

Tabla de contenido

Introducción

Se suele definir conceptualmente a la técnica de Análisis de Componentes Principales (ACP) como una forma de **reducción de ruido**. Esta afirmación no va desencaminada. Si tenemos un conjunto de variables observables y calculamos las correlaciones de unas con otras, podríamos expresar todo en una matriz donde las filas fuesen las variables y las columnas fuesen de nuevo las mismas variables. Cada casilla sería la correlación de una variable con otra. Esta matriz ya nos ofrecería alguna información en cuanto a las relaciones de unas variables con otras. Cada variable correlaciona de manera observable con las demás. Si las personas, por ejemplo, puntúan de manera convergente en dos variables humanas, a ambas variables se les infiere similitud. Si no puntúan conjuntamente, no se podrá inferir similitud. Las personas con más amigos son las que valoran más los días soleados. Los que son puntuales a la hora de coger el autobús lo suelen ser en acudir a una cita. Pero con esto, ¡no podemos salir aún del mundo de lo meramente observable! Variable a variable solo podemos "observar" las que se parecen y quizás buscar constelaciones de coincidencias de manera imprecisa.

Por eso se dice que en esa matriz de correlaciones hay aún mucho ruido. Todo sigue muy apegado a lo observable. Sin embargo, necesitamos un mundo de mayor abstracción. Necesitamos algo así como unas super-variables que nos resuman todas esas relaciones ocultas al ojo de una manera simple, pero sin pérdida de información. Necesitamos reducir ese ruido. Por ejemplo, si tenemos 20 variables observables (número de amigos, valoración de los días soleados, etc…) en las que las personas puntúan, lo que querríamos es tener unas pocas super-variables, muchas menos, que nos resumiesen todas las posibles constelaciones significativas que se forman con ellas. Así, podríamos identificar a las personas con estas simples supervariables, y no con todas las 20 variables originales. Es casi seguro que la misma información aportada por las 20 la conseguiremos con solo unas cuantas variables que resuman a las demás. A esto no referimos con retirar el ruido del mundo observable.

Pongamos que podríamos agrupar esas 20 variables en 2 super-variables y que cada una de esas 20 variables estaría sesgada a pertenecer jerárquicamente a una de esas 3. La valoración de los días de sol podría pertenecer eminentemente a la variable Extraversión y la puntualidad en una cita a la Escrupulosidad. De esta forma, las personas puntuarían en estas dos nuevas variables abstractas, no observables y solo inferibles a partir de las observables.

Esta es la filosofía de la técnica de Análisis de Componentes Principales. Agrupar un conjunto numeroso de variables en unas pocas super-variables. Esas super-variables se llaman Componentes Principales, en el sentido que resumen la estructura de las relaciones de las variables observables en lo principal, en los componentes principales. Nos aísla las propiedades principales de las que se compone una estructura compuesta por puntuaciones observables. Esto es muy útil para identificar primero cuántos componentes existen, y segundo, poder asignarles un significado a partir de las

variables que puntúan más en ellos. Pero veamos ya los objetivos de la técnica. Después de hacerlo tendremos una idea de sus aplicaciones.

Objetivos

El Análisis de Componentes Principales tendrá como objetivo:

1. Clarificar la información: Resumir la información en **menos variables** y poder **identificar super-categorías.** Uno de los beneficios de este tipo de técnicas, al igual que el Análisis de Conglomerados, es tener la capacidad de identificar la estructura de los datos y poder etiquetarla de forma intuitiva para el consumo de información. De hecho, el Análisis de Componentes Principales (ACP) al igual que el Análisis Factorial Exploratorio (AFE), se utiliza en las etapas iniciales de una investigación. Nos es muy útil para explorar los datos.

2. Que las **personas puntúen en esas super-categorías**, haciendo más simples los análisis. Esto es muy útil si lo que se quiere es manejar constructos en vez de variables observables para comprobar como las personas se agrupan en ellos. Teóricamente tiene también relevancia, pues aislaremos Constructos que resuman la realidad que subyace a un conjunto de puntuaciones.

3. Que las puntuaciones en las super-categorías participen **en modelos predictivos posteriores**. Es decir, que la reducción de ruido sea parte del cauce de datos en los modelos predictivos. Imaginemos por ejemplo el beneficio de introducir en una red neuronal unas pocas variables ortogonales frente a un conjunto amplio de variables oblicuas.

Procedimiento

Disposición de los datos

Partimos de una matriz **M** de dimensiones *n* X *p*, donde n son los ejemplares y p representa las variables o propiedades. Dicho de otra manera: tenemos n ejemplares que puntúan en p variables. Tómese como ejemplo las características de las ciudades. Tendremos *n* ejemplares puntuando en *p* propiedades. En este caso *n* ciudades puntuando en *p* delitos (esto puede también pensarse con personas y variables observables).

	Asesinato	Asalto	Pintadas	Sexuales
Alabama	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
Arkansas	8.8	190	50	19.5
California	9.0	276	91	40.6
Colorado	7.9	204	78	38.7
Connecticut	3.3	110	77	11.1
Delaware	5.9	238	72	15.8
Florida	15.4	335	80	31.9
Georgia	17.4	211	60	25.8
Hawaii	5.3	46	83	20.2
Idaho	2.6	120	54	14.2
Illinois	10.4	249	83	24.0
Indiana	7.2	113	65	21.0
…	…	…	…	…
Wyoming	6.8	161	60	15.6

No obstante, esta matriz **M** debe ser transformada en otra que dé cuenta de las similitudes entre unas variables y otras. En la introducción aludíamos que lo que se quería captar con las nuevas super-variables era que algunas variables puntuaban contingentemente con otras. Por esa razón, nos interesa una nueva matriz: la matriz de correlaciones. La manera de construir una matriz de correlaciones es simple. Se tomará cada variable (cada columna) y se calculará su correlación con las demás. Es fácil ver que esta nueva matriz que llamaremos **R** es una matriz simétrica.

	Asesinato	Asalto	Pintadas	Sexuales
Asesinato	1,00	0,80	0,07	0,56
Asalto	0,80	1,00	0,26	0,67
Pintadas	0,07	0,26	1,00	0,41
Sexuales	0,56	0,67	0,41	1,00

Dependiendo de software empleado, deberemos tener en cuenta qué tipo de matrices son las que requieren como entrada las funciones de Componentes Principales. Generalmente ambas, **M** y **R**, son válidas, pero habrá que especificarlo con algún parámetro. Si el argumento de entrada es **M**, es el propio software el que se encarga de calcular la matriz de correlaciones.

Descomposición en Autovectores y Autovalores

Teniendo ya la matriz de correlaciones **R**, el procedimiento más directo para aplicar Componentes Principales es descomponer dicha matriz en Autovectores y Autovalores. Esta técnica es un clásico en el arsenal del álgebra lineal. Lo que se pretende con ella es describir cada vector fila de la matriz **R** con una base más eficiente. Los vectores que componen esa nueva base agruparán el máximo número de variables en torno a ellos. Los vectores de esa nueva base se llaman Autovectores y son ortogonales unos con otros. De esta manera representarán las variables con una referencia muy limpia.

No obstante, queda describir cómo se consigue esa nueva base y en que consiste tal descomposición en Autovectores y Autovalores. Acudiendo al álgebra lineal tenemos que la descomposición consiste en lo siguiente. Una matriz **A** podrá descomponerse en:

$$\mathbf{A} = \mathbf{P} \Lambda \mathbf{P^T} \text{ (Fórmula 1)}$$

Donde $\mathbf{P}$ es la matriz de autovectores y la matriz diagonal Λ contiene los autovalores. Los autovectores se definen como aquellos vectores que al multiplicarlos por $\mathbf{A}$ no cambian más que en su amplitud, pero nunca en orientación. Son las "direcciones principales" de la matriz $\mathbf{A}$. Si al multiplicar el vector $\mathbf{x}$ por la función $\mathbf{A}$ (recordemos que una matriz es también una función) solo le afecta en la amplitud y no cambia de dirección, $\mathbf{x}$ es un autovector de $\mathbf{A}$. La forma de expresar esto es:

$$\mathbf{Ax} = \lambda\mathbf{x} \text{ (Fórmula 2)}$$

donde $\mathbf{x}$ un autovector y λ su autovalor asociado (amplitud asociada a su cambio).

La lectura de la fórmula 2 sería la siguiente. El vector $\mathbf{x}$ multiplicado por la matriz **A** ofrece el mismo resultado que multiplicado por un coeficiente λ. Es decir, $\mathbf{x}$ solo sube o baja en amplitud a recibir el efecto de **A**, pero nunca cambia de dirección. Es por esto que a los vectores de este formato se les llama también direcciones principales de **A**, pues el efecto de **A** le es indiferente en términos de dirección, y, por tanto, definen ya una dirección principal.

En la descomposición representada en la fórmula 1 la matriz $\mathbf{P}$ contiene esos autovectores $\mathbf{x}$ y la diagonal de Λ contiene los autovalores λ. Esta es la esencia de la descomposición. Pero a nosotros no nos interesa más la parte procedimental de la descomposición, es decir, la manera de obtener esos Autovectores y sus Autovalores asociados.

Para calcular $\mathbf{P}$ y Λ se parte de la clásica expresión de la diagonalización[1] $(\mathbf{A} - \mathbf{I}\,\lambda)\mathbf{x} = \mathbf{0}$ donde $\mathbf{I}$ es la matriz identidad. De esta expresión se desprende por algunos razonamientos que para obtener los autovalores tenemos que calcular las raíces del polinomio característico de $|\mathbf{A} - \mathbf{I}\,\lambda|$, es decir, del determinante de $(\mathbf{A} - \mathbf{I}\,\lambda)$. De esta manera, teniendo por ejemplo una matriz $\mathbf{A}$ y la matriz $\mathbf{I\lambda}$ tal que:

$$\mathbf{A} = \begin{pmatrix} 3 & 2 \\ 1 & 4 \end{pmatrix} \qquad \mathbf{I}\lambda = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\lambda$$

Y aplicando la expresión:

$$|\mathbf{A} - \mathbf{I}\,\boldsymbol{\lambda}| \text{ (Fórmula 3)}$$

Tenemos el cálculo de un determinante. Su desarrollo sería:

$$\left|\begin{pmatrix} 3 & 2 \\ 1 & 4 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}\right| = \begin{vmatrix} 3-\lambda & 2 \\ 1 & 4-\lambda \end{vmatrix} = -2 + (3-\lambda)(4-\lambda) = \lambda^2 - 7\lambda + 10$$

Esta operación nos ofrecerá el llamado polinomio característico que igualado a cero nos dará la llamada ecuación característica para obtener las raíces de dicho polinomio:

$$\lambda^2 - 7\lambda + 10 = 0$$

Como en este caso el polinomio es cuadrático, podemos resolverlo de manera sencilla. En general, se resolverán por métodos computacionales, pero es útil saber la esencia de los cálculos y los pasos:

[1] Nos referimos a la expresión $(A - I\lambda)\boldsymbol{x} = \mathbf{0}$ y de su expresión para calcular la ecuación característica mediante $|\mathrm{A} - \mathrm{I}\,\lambda| = 0$ siendo $|.|$ la operación determinante.

$$\lambda = \frac{7 \pm \sqrt{49-4(1)(10)}}{2(1)} = \frac{7 \pm 3}{2} \rightarrow \lambda_1 = 5\,, \lambda_2 = 2$$

Las raíces del polinomio $\lambda_1 = 5\,, \lambda_2 = 2$ son justo los autovalores de la matriz **A**. Volviendo a la expresión $(A - \mathrm{I}\,\lambda)\boldsymbol{x} = 0$ y sustituyendo cada autovalor, podemos obtener los autovectores:

Para $\lambda_1 = 5$:

$$\begin{pmatrix} 3-\lambda & 2 \\ 1 & 4-\lambda \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} 3-5 & 2 \\ 1 & 4-5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} -2 & 2 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Y por tanto:

$$\begin{cases} -2x_1 + 2x_2 = 0 \\ x_1 - x_2 = 0 \end{cases}$$

Cualquier vector con un formato (α, α) puede se autovector de λ_1, por ejemplo (1, 1).

Para $\lambda_2 = 2$:

$$\begin{pmatrix} 3-\lambda & 2 \\ 1 & 4-\lambda \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} 3-2 & 2 \\ 1 & 4-2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Y por tanto:

$$\begin{cases} x_1 + 2x_2 = 0 \\ x_1 + 2x_2 = 0 \end{cases}$$

Cualquier vector con un formato (-2α, α) puede se autovector de λ_2, por ejemplo (-2, 1).

De esta manera, la matriz $\mathbf{P}$ y la mariz Λ quedarían:

$$P = \begin{pmatrix} 1 & -2 \\ 1 & 1 \end{pmatrix} \quad \Lambda = \begin{pmatrix} 5 & 0 \\ 0 & 2 \end{pmatrix}$$

Como ya se ha dicho, en la matriz $\mathbf{P}$ estarían los autovectores en columnas y en la matriz Λ estarían los autovalores en la diagonal. Como ya se dijo, los autovectores son ortogonales. Con esos autovectores y esos autovalores trabajará la técnica de Análisis de Componentes Principales. En este pequeño y casi irrelevante ejemplo, podemos ver que la matriz $\mathbf{A}$ se descompondría a partir de $\mathbf{P}$ y Λ. A partir de aquí, haremos esto mismo con la matriz de correlaciones de un conjunto de variables. La matriz **A** será nuestra **R** (matriz de correlaciones) y la descomposición se hará de la misma manera:

$$\mathbf{R} = \mathbf{P} \Lambda \mathbf{P^T} \text{ (Fórmula 4)}$$

Trabajando con matrices de correlaciones.

Idea de unos nuevos ejes de referencia

Hasta ahora hemos presentado los primeros esbozos del Análisis de Componentes Principales y hemos descrito la técnica algebraica de descomposición en autovectores y autovalores. Esta técnica tiene su origen en la búsqueda de matrices equivalentes. No obstante, decíamos que en el caso de matrices de correlaciones puede ser usada para extraer super-variables (Componentes Principales) que resuman la variabilidad de las relaciones entre las variables originales, llamadas observables. Esa es la clave del Análisis de Componentes Principales. Usar la descomposición en autovectores y autovalores en la matriz de correlación que expresa las puntuaciones contingentes de los *n* ejemplares en las *p* variables. Hecho esto, obtendremos una nueva colocación de la base de tal manera que la nueva referencia serán los autovectores.

Y hay más, como cada autovector tiene asociado un autovalor que indica la cantidad de variabilidad que explica (la cantidad de variables en torno a él), podemos quedarnos solo con unos pocos autovectores que explican la mayor parte de esa variabilidad. Esto es justo la parte en que se reduce el ruido.

De las p variables que tenemos en origen identificando a los ejemplares, pasamos a tener, sin pérdida sustantiva de información, unos pocos autovectores que identifican de igual manera a las variables originales y a los ejemplares de la matriz **M**. Y estos nuevos autovectores serán más eficientes al ser menos y resumir la información de las variables originales de manera ortogonal. Menos, resumido y sin pérdida de la información. Ese es el escenario ideal al que se quiere llegar.

Por tanto, El Análisis de Componentes Principales se puede ver también como una técnica para representar las variables originales en un nuevo eje de coordenadas de mayor eficiencia expresiva.

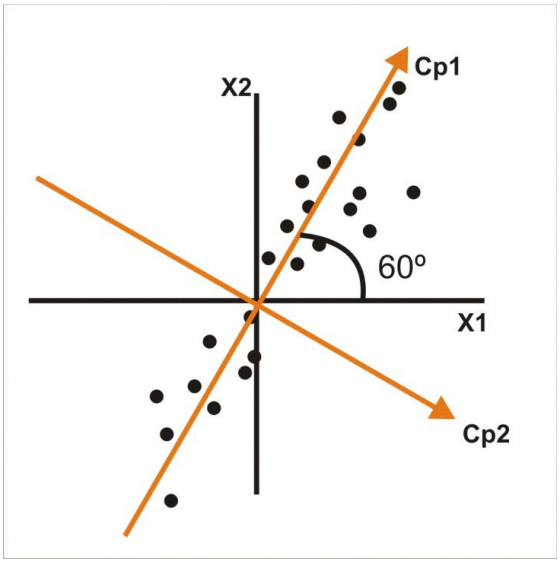


Figura 1.

En la figura 1 tenemos el eje de coordenadas original (canónica) con las variables originales representadas como X_1 y X_2 y todas las puntuaciones expresadas a partir de él. Cualquier variable tiene una puntuación en ambas (la correlación). Sin embargo, después de realizar el Análisis de Componentes Principales vía descomposición en autovectores y autovalores, obtenemos un nuevo eje de coordenadas con los autovectores resultantes. Obsérvese que la operación que hay que realizar para pasar de uno a otro es una rotación de 60 grados. Además, los componentes principales estarían ordenadas de mayor a menos importancia según aglutinen variabilidad en torno a ellos. Esto será informado por su autovalor. En la figura 1 el componente 1 (Cp1) explica casi toda la variabilidad en torno a él. Con él se podría explicar un rango amplio de puntuaciones. Desde muy grandes a muy pequeñas, positivas y negativas. Tanto es así, y tan poco explica el segundo componente que podríamos explicar esas puntuaciones solamente recurriendo al primero. Por tanto, podríamos desechar el componente representado en Cp2 y quedarnos solo con el Cp1 sin pérdida de información. De expresar las puntuaciones con dos variables pasaríamos a hacerlo con una, simplemente porque las dos variables están muy correlacionadas. Pero vamos por pasos.

Una matriz de correlaciones ejemplo

Propongamos primero la matriz que llamamos antes **M** y que tiene n filas por p columnas. Las filas son ejemplares, en este caso ciudades, y las columnas variables o propiedades, en este caso, tipos de crímenes que se producen en esas ciudades:

	Asesinato	Asalto	Pintadas	Sexuales
Alabama	13,2	236	58	21,2
Alaska	10,0	263	48	44,5
Arizona	8,1	294	80	31,0
Arkansas	8,8	190	50	19,5
California	9,0	276	91	40,6
Colorado	7,9	204	78	38,7
Connecticut	3,3	110	77	11,1
Delaware	5,9	238	72	15,8
Florida	15,4	335	80	31,9
Georgia	17,4	211	60	25,8
Hawaii	5,3	46	83	20,2
Idaho	2,6	120	54	14,2
Illinois	10,4	249	83	24,0
Indiana	7,2	113	65	21,0
Iowa	2,2	56	57	11,3
Kansas	6,0	115	66	18,0
Kentucky	9,7	109	52	16,3
Louisiana	15,4	249	66	22,2
Maine	2,1	83	51	7,8
Maryland	11,3	300	67	27,8
Massachusetts	4,4	149	85	16,3
Michigan	12,1	255	74	35,1
Minnesota	2,7	72	66	14,9
Mississippi	16,1	259	44	17,1
Missouri	9,0	178	70	28,2
Montana	6,0	109	53	16,4
Nebraska	4,3	102	62	16,5
Nevada	12,2	252	81	46,0
NewHampshire	2,1	57	56	9,5
NewJersey	7,4	159	89	18,8
NewMexico	11,4	285	70	32,1
NewYork	11,1	254	86	26,1
NorthCarolina	13,0	337	45	16,1
NorthDakota	0,8	45	44	7,3
Ohio	7,3	120	75	21,4
Oklahoma	6,6	151	68	20,0
Oregon	4,9	159	67	29,3
Pennsylvania	6,3	106	72	14,9
RhodeIsland	3,4	174	87	8,3
SouthCarolina	14,4	279	48	22,5
SouthDakota	3,8	86	45	12,8
Tennessee	13,2	188	59	26,9
Texas	12,7	201	80	25,5
Utah	3,2	120	80	22,9
Vermont	2,2	48	32	11,2
Virginia	8,5	156	63	20,7
Washington	4,0	145	73	26,2
WestVirginia	5,7	81	39	9,3
Wisconsin	2,6	53	66	10,8
Wyoming	6,8	161	60	15,6

Lo que queremos ahora es agrupar estas variables en algo más abstracto y eficiente que nos ayude a explorar si los tipos de crímenes se agrupan en una entidad superior debido a que las ciudades puntúan en algunos de ellos de forma contingente. En otras palabras, existen tipos de crímenes que correlacionan con otros tipos de crímenes. Dicho de otra forma, existe tal colinealidad entre los delitos que se recomienda resumir información y quitar el ruido que esas colinealidades provocan. En forma de pregunta podríamos proponer ¿pueden estas cuatro variables observables resumirse en menos super-variables?.

Para ello, como se ha dicho en el apartado de disposición de datos, tenemos que transformar esta primera matriz **M** en una matriz de correlaciones **R** entre las variables**,** de manera que se exprese en ella las contingencias de unos crímenes con otros. Es fácil apercibirse que la matriz **R** será simétrica y tendrá 4 filas y 4 columnas.

	Asesinato	Asalto	Pintadas	Sexuales
Asesinato	1,00	0,80	0,07	0,56
Asalto	0,80	1,00	0,26	0,67
Pintadas	0,07	0,26	1,00	0,41
Sexuales	0,56	0,67	0,41	1,00

Calculada la matriz de correlaciones, es sobre ella donde se llevará a cabo la descomposición en autovectores y autovalores. De esta descomposición obtendremos la matriz de autovectores $\mathbf{P}$ y la matriz diagonal de autovalores Λ asociados a ellos. Hagamos pues la descomposición y analicemos las salidas $\mathbf{P}$ y Λ. Para hacerlo hay que seguir la expresión mencionada anteriormente que nos ofrecería el polinomio característico:

$$|\mathrm{R} - \mathrm{I}\,\lambda| = 0 \text{ (Fórmula 5)}$$

Donde **R** es la matriz de correlación:

$$R = \begin{pmatrix} 1.00 & 0.80 & 0.07 & 0.56 \\ 0.80 & 1.00 & 0.26 & 0.67 \\ 0.07 & 0.26 & 1.00 & 0.41 \\ 0.56 & 0.67 & 0.41 & 1.00 \end{pmatrix}$$

Descomponiendo la matriz **R**, obtenemos las matrices $\mathbf{P}$ y Λ de autovectores y autovalores:

$$\mathbf{P} = \begin{pmatrix} -0.53 & 0.41 & -0.34 & 0.64 \\ -0.58 & 0.18 & -0.26 & -0.74 \\ -0.27 & -0.87 & -0.37 & 0.13 \\ -0.54 & -0.16 & 0.81 & 0.09 \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} 2.49 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.35 & 0 \\ 0 & 0 & 0 & 0.17 \end{pmatrix}$$

Conduzcámonos por inducción. Veamos primero la matriz Λ de autovalores. En ella se muestra la importancia de cada autovector de la matriz $\mathbf{P}$, es decir, de cada Componente Principal que ha sido identificado. En un primer vistazo vemos que hay un Componente principal que es mucho más importante que los demás. Se trata del primero, el cual tiene asociado un autovalor de 2,49. El Siguiente autovector tiene asociado un autovalor de 1. Este valor es bastante más pequeño que el anterior, pero bastante más grande que los siguientes de 0,35 y 0,17. Sin haber tratado todavía los posibles criterios que nos pueden ayudar a prescindir de Componentes de valores pequeños, podemos ya hacer una primera tentativa y quedarnos solo con los dos primeros. Sería algo así como quedarse con los que son igual o superiores a 1. Justo esto es un criterio que a veces se usa. Si asumimos que los autovalores indican la variabilidad explicada, podemos calcular el porcentaje de variabilidad explicada al quedarnos solo con estos dos primeros Componentes de esta manera:

$$CP_{(1,2)} = 100 \cdot \frac{\lambda_{(1)}+\lambda_{(2)}}{\sum_{i=1}^{p} \lambda_{(i)}} = 100 \cdot \frac{3{,}49}{4{,}01} = 87\%$$

Esto quiere decir que quitando los dos últimos Componentes Principales estaríamos aun dando cuenta del 87% de variabilidad de las contingencias de las variables. El efecto que tendrá tomar solo estos dos Componentes será simplemente retirar el ruido de las colinealidades entre las variables (ocurrencias conjuntas de ciertos crímenes en las ciudades) pero sin perder excesiva información, y encima con la ventaja de clasificar las ciudades de manera efectiva con solo esos dos. Pongamos las matrices ahora, pero de manera truncada tal y como las hemos dejado al retirar dos componentes.

$$P = \begin{pmatrix} -0.53 & 0.41 \\ -0.58 & 0.18 \\ -0.27 & -0.87 \\ -0.54 & -0.16 \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} 2.49 & 0 \\ 0 & 1 \end{pmatrix}$$

Incluso podemos formatear las matrices anteriores de una manera acorde con lo que representan:

	CP1	CP2
Asesinato	-0,53	0,41
Asalto	-0,58	0,18
Pintadas	-0,27	-0,87
Sexuales	-0,54	-0,16

Aquí la tenemos. Esta nueva forma de ver la matriz nos da más perspectiva de lo que estamos haciendo. Los crímenes pueden ser resumidos sin pérdida de información en dos grandes Componentes Principales. Cada tipo de crimen puntuará en ambos. Parece que en el CP1 puntúan a la inversa todos los crímenes mientras que en el CP2 puntúa de manera directa el asesinato y de manera inversa las pintadas. La interpretación sería que el primer componente puntúa en seguridad (a más puntuación en él, más segura es la ciudad) y el segundo en asesinato y pintadas eminentemente (a más puntuación en él, más asesinatos y asaltos, y a menos, más pintadas o delitos sexuales).

Podemos también expresar una puntuación de cada Componente en base a una combinación lineal de los tipos de crímenes, es decir, de las variables observables originales:

$$CP1 = -0{,}53\, x_{Ase} - 0{,}58\, x_{Asa} - 0{,}27 X_{Pin} - 0{,}54 X_{Sex}$$

$$CP2 = 0{,}41\, x_{Ase} + 0{,}18\, x_{Asa} - 0{,}87 X_{Pin} - 0{,}16 X_{Sex}$$

Y a partir de esto podemos también determinar cómo puntuarían cada una de las ciudades en estos dos Componentes Principales:

	CP1	CP2
Alabama	-0,97	1,12
Alaska	-1,93	1,06
Arizona	-1,74	-0,73
Arkansas	0,13	1,10
California	-2,49	-1,52
Colorado	-1,49	-0,97
Connecticut	1,34	-1,07
Delaware	-0,04	-0,32
…	…	…
Florida	-2,98	0,03
Wisconsin	2,05	-0,60
Wyoming	0,62	0,31

Cada una de las ciudades tendrá una puntuación en ambos Componentes Principales, pudiendo ser clasificados en la tipología de crímenes que representan cada uno de ellos.

Comunalidad

La comunalidad es la proporción de varianza explicada de la variable a partir de los componentes principales que se hayan extraído. Es decir, cuánto explican los nuevos Componentes de cada una de las variables originales. Puede ser el caso de que alguna de las variables no esté bien explicada con los nuevos ejes de coordenadas. Cada variable original tendrá una comunalidad asociada. La forma de calcularla será:

$$h_i^2 = \sum_{j=1}^{m} F_{(j)i}^2 \quad \text{(Fórmula 5)}$$

Donde:

m es el número de Componentes Principales seleccionados.

i es el índice de la variable a calcular la comunalidad

$F_{(j)i}^2$ es la saturación de la variable i en el Componente j

Es decir, la comunalidad de una variable se calcula como la suma al cuadrado de las saturaciones que tiene en los distintos componentes. En la matriz del ejemplo anterior teníamos expresadas las saturaciones de cada variable en cada componente:

	CP1	CP2
Asesinato	-0,53	0,41
Asalto	-0,58	0,18
Pintadas	-0,27	-0,87
Sexuales	-0,54	-0,16

Aplicando la fórmula a la variable “Asesinato” obtendríamos la comunalidad de la variable:

$$h^2_{(Asesinato)} = \sum_{j=1}^{m} F^2_{(j)1} = (-0{,}53)^2 + (0{,}41)^2 = 0{,}28 + 0{,}168 = 0{,}45$$

Proporción de varianza explicada por cada componente

Ya se ha dicho que la proporción de variabilidad de las contingencias entre variables de cada Componente lo daban los Autovalores. Podíamos simplemente aplicar la fórmula:

$$CP_{(i)} = 100 \cdot \frac{\lambda_{(i)}}{\sum_{j=1}^{p} \lambda_{(i)}}$$

Donde

i es el índice del componente a calcular la proporción

j es un subíndice que indiza a todos los autovalores

p es el número de autovalores

Sin embargo, podemos calcular la proporción de cada Componente de esta otra forma:

$$CP_{(j)} = \sum_{i=1}^{p} F^2_{(j)i}$$

i es el índice que indiza a las variables (las filas)

p es el número de variables originales

j es el componente (la columna) a la que se calcula la proporción

Por lo que la proporción de varianza de cada componente es la suma de los cuadrados de las saturaciones de las variables en ese componente. Siguiendo con la tabla anterior, la proporción del componente primero es:

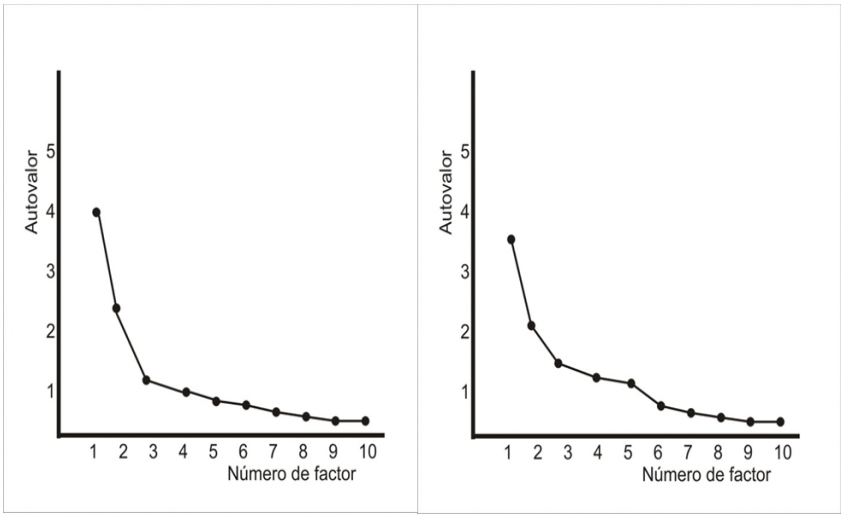
$$CP_{(1)} = \frac{\sum_{i=1}^{p} F_{(1)i}^2}{4} = \frac{-0{,}53^2 + -0{,}58^2 + -0{,}27^2 + -0{,}54^2}{4} = \frac{0{,}97}{4} = 0{,}24$$

Selección del número de componentes

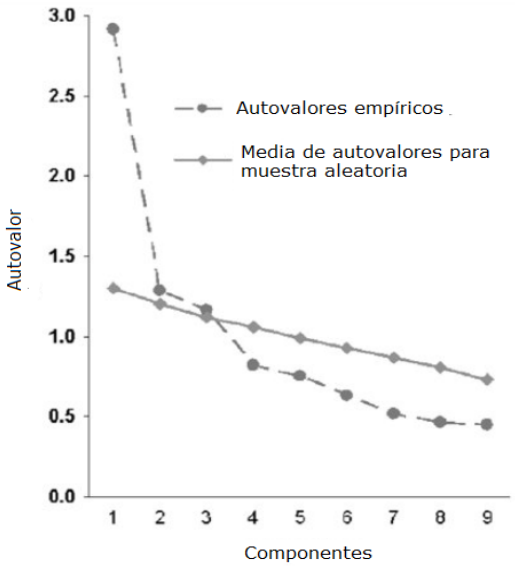
Antes hemos mencionado que existían varias formas para decidir cuántos componentes (autovectores) conservar en la solución final. En este apartado se enumeran algunas de ellas, aunque el más usado es el Análisis Paralelo. Los métodos son:

Kaiser: se toman los componentes cuyos autovalores son mayores que 1. Este valor es ciertamente arbitrario y existen mejores métodos de estimación.

Scree test (gráfico de sedimentación) de Catell: Se grafican las proporciones de varianza de cada autovalor. Se basa en que existe una transición de un número de componentes a otro en que la caída de esa proporción se frena. El procedimiento se apoya en identificar esa caída.



Análisis paralelo de Horn: Para un número de autores importante es el mejor método (Ruscio & Roche, 1999). Se basa en la comparación entre los autovalores empíricos y simulados con el mismo número de variables y sujetos (por ejemplo, 100 muestras). Se retienen los factores cuyos autovalores superen los generados aleatoriamente.



Código R

```
library(ggplot2)
library(dplyr)
library(corrplot)
library(corrr)
library(DT)
library(hornpa)

#Datos de tendencia central-medias
apply(USArrests, 2, mean)

#¿hay algún valor perdido?no
colSums(is.na(USArrests))

#se pinta la matriz de correlaciones de las columnas de USArrests
corrplot(cor(USArrests), order = "hclust")
#tiene que haber cierta colinealidad para extraer Componentes.

#ejecutar componentes con la correlación de las variables
#Toma como argumento la matriz de puntuaciones, no la de correlaciones
#ejecuta internamente la matriz de correlaciones
#asume también las columnas como variables
pca.res <- prcomp(USArrests, scale = TRUE)

#saturaciones de cada variable en cada factor
#las columnas son los autovectores de la descomposición en Autovectores/autovalores
pca.res$rotation

#para ver los autovalores
pca.var =pca.res$sdev ^2
pca.var

#se muestra el ratio de cada autovalor
var.ratio=pca.var/sum(pca.var)
var.ratio

#tomamos solo dos componentes. Los más importantes
pca.res <- prcomp(USArrests, scale = TRUE, rank =2)
pca.res$rotation


#las puntuaciones de los ejemplares (las ciudades) en los componentes
pca.res$x

#¿cómo se hace el test paralelo para ver cuantos retener?
#comparando los autovalores de la simulación de la función hornpa (0.95) con los que ha arrojado nuestro análisis pca.var
#presumiblemente solo un componentes sería elegido
simulacion <- hornpa(k = 4, size = 50, reps = 500, seed = 1234)
pca.var

#inciso para hacer autovectores/autovalores en la de correlaciones
ev <- eigen(cor(USArrests))
ev$values
ev$vectors
################es la demostracción de que se puede hacer con descomposición directamente
```

https://rstudio-pubs-static.s3.amazonaws.com/377338_75ed92a8463d482a80045abcae0e395d.html

https://cran.r-project.org/web/packages/matlib/vignettes/eigen-ex1.html