ANALISIS DE COMPONENTES PRINCIPALES

Dr. Porfirio Gutiérrez González

ANALISIS DE COMPONENTES PRINCIPALES

- **El análisis de componentes principales (ACP) es una de las técnicas estadísticas más utilizadas para el análisis de datos multivariados, ya que nos permite verificar las propiedades de la matriz de datos desde distintas perspectivas.**
- **La idea básica del ACP es expresar las variables originales en un número reducido de nuevas variables k<p, permitiendo con esto disminuir la dimensionalidad del problema.**
- **Generar nuevas variables con propiedades convenientes que expresen la información contenida en un conjunto de datos.**
- **Las nuevas variables son independientes y de varianza máxima.**

DEFINICIÓN DE COMPONENTES PRINCIPALES

- El primer componente es la combinación lineal de las variables originales, con varianza máxima.
- El segundo componente es la combinación lineal de las variables originales, independiente del primer componente, con la varianza máxima.
- El tercer componente es la combinación lineal de las variables originales, con la maximiza varianza independiente de los dos primeros componente.

- Así sucesivamente, el K componente es la combinación lineal de las variables originales, con varianza máxima e independiente de los k-1 componentes anteriores.

Sean $X_1, X_2, \dots, X_P$ las variables originales y sean $Y_1, Y_2, \dots, Y_P$ los componentes principales. Esto expresado matemáticamente como

$$Y_1 = \alpha_{11}X_1 + \alpha_{12}X_2 + \alpha_{13}X_3 + \dots + \alpha_{1p}X_p$$

$$Y_2 = \alpha_{21}X_1 + \alpha_{22}X_2 + \alpha_{23}X_3 + \dots + \alpha_{2p}X_p$$

$$Y_3 = \alpha_{31}X_1 + \alpha_{32}X_2 + \alpha_{33}X_3 + \dots + \alpha_{3p}X_p$$

$$\vdots \quad \vdots \quad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$\vdots \quad \vdots \quad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$\vdots \quad \vdots \quad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$Y_p = \alpha_{p1}X_1 + \alpha_{p2}X_2 + \alpha_{p3}X_3 + \dots + \alpha_{pp}X_p$$

Donde $X_1, X_2, \dots, X_P$ son las variables originales

$$\text{Var}(Y_1) = \lambda_1 \qquad \text{Var}(X_1) = S_1$$

$$\text{Var}(Y_2) = \lambda_2 \qquad \text{Var}(X_2) = S_2$$

$$\text{Var}(Y_3) = \lambda_3 \qquad \text{Var}(X_3) = S_3$$

$$\text{Var}(Y_p) = \lambda_p \qquad \text{Var}(X_p) = S_p$$

α_{ij} *pesos o cargas*

DERIVACIÓN ALGEBRAICA DE LOS COMPONENTES PRINCIPALES

Para derivar los componentes principales se tienen procesos algebraicos definidos. Sean X_1, X_2, ... , X_P variables aleatorias que constituyen la matriz de datos **X** y sea **S** la matriz de varianzas y covarianzas.

El primer componente principal está dado por

$$Y_1 = \alpha_{11}X_1 + \alpha_{12}X_2 + \alpha_{13}X_3 + ... + \alpha_{1p}X_p$$

$$Y_1 = \alpha'_1 X$$

$$\text{Var}\{Y_1\} = \text{Var}\{\alpha'_1 X\} = \alpha'_1 S\, \alpha_1$$

Siendo esta la máxima varianza, donde $\alpha'_1 = [\alpha_{11}, \alpha_{12}, \alpha_{13}, ... , \alpha_{1p}]$

La única restricción del sistema es que $\alpha'_1\alpha_1 = 1$, ya que para maximizar $\alpha'_1\mathbf{S}\,\alpha_1$ basta con hacer crecer a α_1 y automáticamente la varianza Y_1 crecerá. Viéndolo de manera vectorial, lo que se busca no es alargar el vector, ya que esto se hace fácilmente con escalares. La idea es encontrar la dirección del vector $\mathbf{Y_1}$ y encontrar el α_1 que cumpla con la característica $\alpha'_1\alpha_1 = 1$.

La maximización de $\alpha'_1 S\,\alpha_1$ se resuelve con multiplicadores de Lagrange. Sea una función

$$\varphi_1 = \alpha'_1\mathbf{S}\,\alpha_1 - \lambda\,(\alpha'_1\,\alpha_1 - \mathbf{1})$$

donde λ es un multiplicador de Lagrange. Derivando respecto a α_1 y después igualando a cero

$$\mathbf{S}\,\alpha_1 - \lambda\alpha_1 = 0$$

$$\mathbf{S}\,\alpha_1 = \lambda_1\boldsymbol{\alpha_1}$$

Implica que $\boldsymbol{\alpha_1}$ es un vector propio de la matriz **S** y que λ_1 su correspondiente valor propio.

Para determinar que valor propio de S es la solución multiplicando por la izquierda α'_1

$$\alpha'_1\mathbf{S}\alpha_1 = \alpha'_1\lambda\alpha_1$$

$$\alpha'_1\mathbf{S}\alpha_1 = \lambda\alpha'_1\alpha_1 = \lambda_1$$

Por lo que λ_1 es la varianza de Y_1.

Como Y_1 es el primer componente que explica la mayor varianza entonces λ_1 es el mayor propio de la matriz **S**. Su vector asociado $\boldsymbol{\alpha_1}$ contiene los pesos o cargas de cada variable original en el primer componente.

En este caso, ya que λ_1 es el máximo entonces el primer componente está dado por la combinación lineal

$$Y_1 = \alpha_{11}X_1 + \alpha_{12}X_2 + \alpha_{13}X_3 + \ldots + \alpha_{1p}X_p$$

donde $\alpha_{11}, \alpha_{12}, \alpha_{13}, \ldots, \alpha_{1p}$ son los coeficientes del vector propio correspondiente al mayor valor propio λ_1 de la matriz S, además la $Var(Y_1) = \alpha'_1 S \alpha_1 = \lambda_1$.

Para el segundo componente principal estará dado por

$$Y_2 = \alpha_{21}X_1 + \alpha_{22}X_2 + \alpha_{23}X_3 + \ldots + \alpha_{2p}X_p$$

El problema se centra en encontrar un vector $\alpha'_2 = [\alpha_{21}, \alpha_{22}, \alpha_{23}, \ldots, \alpha_{2p}]$ tal que

$\text{Var}(Y_2) = \alpha'_2 S\, \alpha_2$ Sea máxima independiente de Y_1

Con las restricciones $\alpha'_2\alpha_2 = 1$ y $\alpha'_1\alpha_2 = 0$ Otorgonalidad de Y_1 y Y_2

$$\text{Cov}(Y_1\ Y_2) = \text{Cov}(\alpha'_1X, \alpha'_2X) = \alpha'_1\, \text{Cov}(X'\ X)\, \alpha'_2 = \alpha'_1 S\, \alpha'_2$$

$$\mathbf{Cov\ (Y_1\ Y_2)} = \alpha_1\, \alpha'_1\, \mathbf{S}\, \alpha_1\, \alpha'_2 = \alpha_1\, \lambda_1\, \alpha'_2 = \lambda_1\, \alpha_1\, \alpha'_2 = 0 \qquad \text{pero} \qquad \lambda_1 \neq 0$$

Por lo tanto $\alpha_1\, \alpha'_2 = 0$

Por medio de los multiplicadores de Lagrange se maximiza de forma que

$$\varphi_2 = \alpha'_2 S \alpha_2 - \lambda (\alpha'_2 \alpha_2 - 1) - \phi \alpha'_2 \alpha_1$$

Derivando esto respecto a α_2 y luego igualando a cero y simplificando, se obtiene

$$S\alpha_2 - \lambda\alpha_2 - \phi \alpha_1 = 0$$

Multiplicando a la izquierda por α'_1 tenemos

$$\alpha'_1 S\alpha_2 - \alpha'_1 \lambda\alpha_2 - \alpha'_1 \phi \alpha_1 = 0$$

$$\alpha'_1 S\alpha_2 - \lambda\alpha'_1\alpha_2 - \phi\alpha'_1 \alpha_1 = 0$$

$$\mathbf{Cov}\,(Y_1, Y_2) = \alpha'_1 S\alpha_2 = \lambda\alpha'_1\alpha_2 = 0$$

$$\alpha'_1\alpha_1 = 1, \qquad \alpha'_1\alpha_2 = 0, \qquad \phi = 0,$$

$$S\alpha_2 - \lambda\alpha_2 = 0 \;\Rightarrow\; S\alpha_2 = \lambda\alpha_2 \;\Rightarrow\; \alpha'_2 S\alpha_2 = \alpha'_2\lambda\alpha_2 = \lambda_2\,\alpha'_2\alpha_2 = \lambda_2$$

$$\mathbf{Var}(Y_2) = \lambda_2$$

Si este procedimiento se sigue entonces se conseguirá un tercer, un cuarto, un n-esimo componente principal con los vectores propios α_3 α_4 α_5 **...** , α_p correspondientes al los valores propios tercero, cuarto, etc. hasta el p-ésimo valor más grande de la matriz **S**. La varianza de $\mathbf{Y_k}$ es igual al k-ésimo valor propio λ_k. Esto es

$$\mathbf{Var}\,(\mathbf{Y_k}) = \mathbf{Var}\,(\alpha'_k\mathbf{X}) = \alpha'_k\mathbf{S}\,\alpha_k = \alpha'_k\,\lambda_k\,\alpha_k = \lambda_k$$

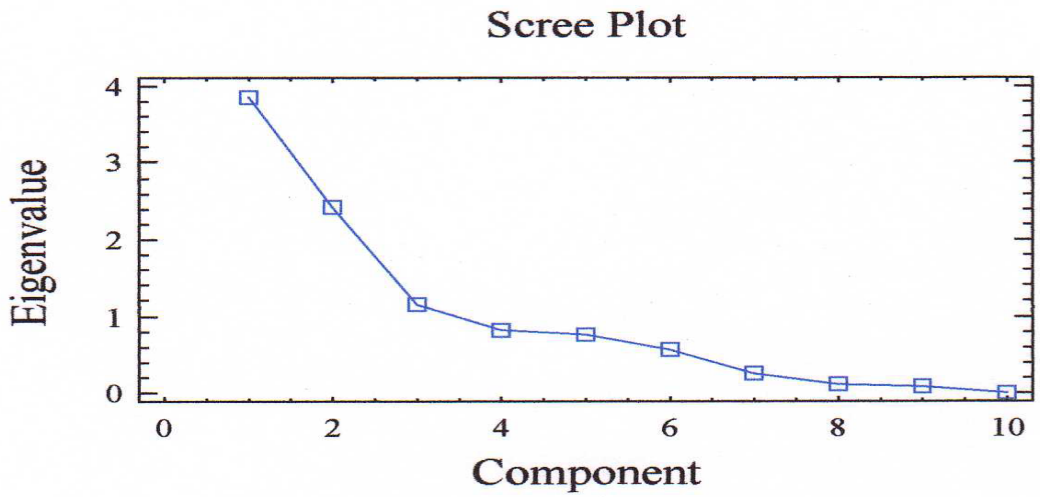
Existen dos métodos para ayudar a elegir el numero de componentes principales:

Método 1: Supóngase que se desea tomar en cuenta el 100% de la variabilidad de los datos en las variables originales, entonces considerando

$$Varianza\ total = [(\lambda_1 + \lambda_2 + \lambda_3 + \ldots + \lambda_p)/(S_1 + S_2 + S_3 + \ldots + S_p)]*100$$

Se selecciona el numero de componentes principales hasta cubrir un deseado porcentaje de la variabilidad total de los datos, explicada por los componentes principales.

Método 2: Se utiliza una gráfica de sedimentación de los eigenvalores. En una grafica de sedimentación se utilizan parejas $(1, \lambda_1)$, $(1, \lambda_2)$, …, $(1, \lambda_p)$. Cuando los puntos de la gráfica tienden a nivelarse, estos eigenvalores suelen estar suficientemente cercanos a cero, es probable que el correspondiente componente principal este midiendo muy poca información y no sea necesario interpretarse.



Ejemplo de Análisis de Componentes Principales con la Matriz de Varianza y Covarianza

MARCAS DE TEQUILA

Variables
pH
conductividad
viscosidad
Densidad
Velocidad de sonido
Índice refracción

CATEGORIA	pH	CONDUCTIVIDAD	VISCOSIDAD	DENSIDAD	VELOCIDAD DE SONIDO	INDICE DE REFRACCION
Aged	3.71	48.9	0.0015	0.94786	1596.5	1.35197
Aged	3.44	161.87	0.00111	0.98448	1550.11	1.3376
Aged	4.11	41.6	0.00138	0.95481	1611.51	1.35156
Aged	3.81	19.18	0.00131	0.9641	1610.17	1.35063
Aged	3.98	15.68	0.00151	0.94738	1595.07	1.35317
Aged	4.11	34.1	0.00147	0.95161	1608.45	1.3516
Silver	4.05	18.41	0.00148	0.94851	1597.98	1.35303
Gold	4.87	41.6	0.00134	0.95409	1611.64	1.3515
Aged	3.99	30.3	0.00149	0.94831	1596.51	1.3534
Aged	4.07	39.8	0.00149	0.94551	1589.53	1.3535
Silver	3.61	103.47	0.00194	0.96814	1613.03	1.3449
Silver	3.95	50.03	0.0015	0.94801	1596.7	1.353
Aged	4.14	35.17	0.00118	0.95131	1606.35	1.346
Silver	4.76	13.09	0.0015	0.94611	1591.84	1.35353
Gold	3.9	44.5	0.00143	0.95196	1606.41	1.3511
Gold	3.97	55.8	0.00146	0.94917	1599.58	1.35167
Gold	5.15	111.03	0.00167	1.01763	1631.45	1.3617
Aged	4.13	19.43	0.00143	0.95353	1607.88	1.35113
Aged	4.05	18.18	0.00147	0.9505	1603.46	1.35143
Silver	4.06	15.55	0.00145	0.94787	1596.11	1.3519
Aged	4.71	56.77	0.00139	0.95391	1611.18	1.3513
Aged	3.73	38.13	0.00137	0.95315	1609.13	1.35116
Aged	3.8	31.67	0.00141	0.9535	1607.76	1.3515
Aged	3.63	40.67	0.00148	0.9481	1596.18	1.35143
Aged	3.64	48.43	0.00147	0.94774	1595.07	1.35313

CATEGORIA	pH	CONDUCTIVIDAD	VISCOSIDAD	DENSIDAD	VELOCIDAD DE SONIDO	INDICE DE REFRACCION
Aged	4.07	35	0.00148	0.9479	1595.59	1.35303
Aged	4.1	37.73	0.00143	0.95141	1604.65	1.3516
Aged	4.61	10.04	0.00136	0.95571	1611.06	1.35116
Aged	3.98	36.47	0.00145	0.95489	1609.68	1.3513
Aged	4.55	19.13	0.00151	0.95106	1601.03	1.3515
Extra aged	3.95	31.33	0.00151	0.95167	1600.14	1.3535
Extra aged	3.97	14.59	0.00147	0.94933	1596.18	1.35193
Silver	5.17	13.46	0.00147	0.95005	1601.16	1.35133
Extra aged	4.06	34.43	0.00149	0.9497	1599	1.35363
Aged	4.08	17.67	0.00148	0.94911	1597.41	1.35303
Silver	4.56	38.87	0.00149	0.9489	1598.81	1.35197
Gold	4.51	40.83	0.00144	0.94911	1599.07	1.3531
Aged	4.56	49.5	0.00146	0.95015	1601.03	1.35163
Aged	4.86	10.19	0.00147	0.95143	1601.65	1.35306
Extra aged	4.7	18.51	0.00144	0.9487	1598.43	1.3531
Aged	4.01	10.11	0.00147	0.95046	1598.69	1.35316
Silver	4.18	19.99	0.00149	0.94871	1597.97	1.35313
Aged	4.58	17.16	0.00135	0.95348	1608.99	1.35087
Extra aged	4.41	40.3	0.00153	0.94611	1590.15	1.3541
Silver	3.97	18.13	0.00131	0.95471	1610.59	1.3509
Aged	4.15	17.5	0.00139	0.95343	1607.91	1.35076
Extra aged	4.11	43.67	0.00148	0.95041	1601.57	1.35166
Aged	4.3	11.71	0.00148	0.95171	1607.36	1.351
Aged	3.61	31.1	0.00148	0.94476	1586.1	1.35406
Extra aged	3.75	33.4	0.00151	0.94637	1585.99	1.3543
Silver	4.19	11.18	0.0016	0.94476	1586.95	1.3535
Aged	4.14	14.71	0.00158	0.9455	1587.57	1.3541
Extra aged	3.97	13.16	0.00154	0.94484	1586.91	1.35417

MEDIDAS DE TENDENCIA CENTRAL Y VARIACION DE MARCAS DE TEQUILA

	pH	CONDUCTIVIDAD	VISCOSIDAD	DENSIDAD	VELOCIDAD DE SONIDO	INDICE DE REFRACCION
Average	4.15981	34.4006	0.00146019	0.952484	1600.21	1.35195
Variance	0.153533	725.188	1.26E-08	0.000122816	123.029	8.67498E-06
Standard deviation	0.391832	26.9293	0.00011228	0.0110822	11.0918	0.00294533
Minimum	3.44	10.04	0.00111	0.94476	1550.11	1.3376
Maximum	5.17	161.87	0.00194	1.01763	1631.45	1.3617

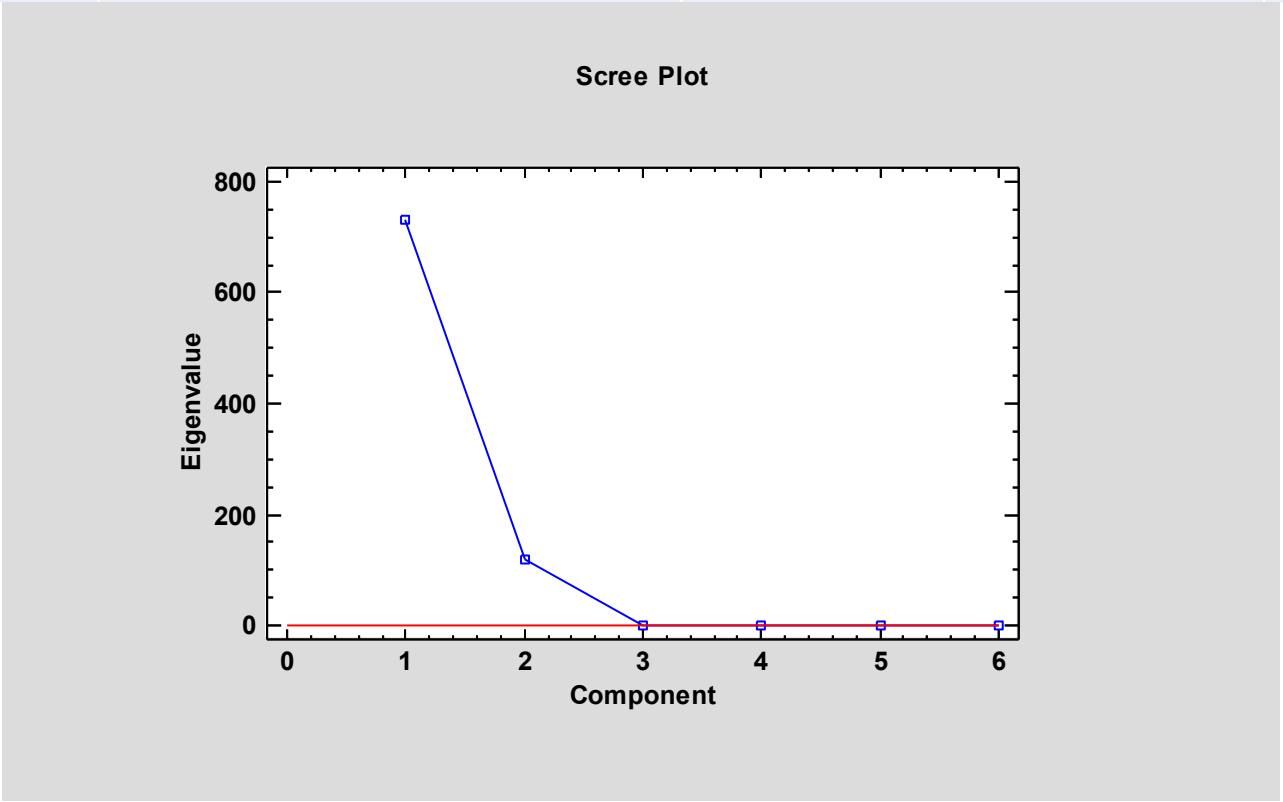
VARIANZA TOTAL= 848.370665

MATRIZ DE VARIANZAS Y COVARIANZAS

	pH	CONDUCTIVIDAD	VISCOSIDAD	DENSIDAD	VELOCIDAD DE SONIDO	INDICE DE REFRACCION
pH	0.153533	-2.07344	1.71157E-06	0.000736543	1.68466	0.000418969
CONDUCTIVIDAD	-2.07344	725.188	2.64557E-05	0.206494	-50.9619	-0.0332967
VISCOSIDAD	1.71157E-06	2.64557E-05	1.26E-08	3.34E-08	0.000201817	1.47E-07
DENSIDAD	0.000736543	0.206494	3.34E-08	0.000122816	0.0343049	-2.1716E-06
VELOCIDAD DE SONIDO	1.68466	-50.9619	0.000201817	0.0343049	123.029	0.0106997
INDICE DE REFRACCION	0.000418969	-0.0332967	1.47E-07	-2.1716E-06	0.0106997	8.67498E-06

Análisis de Componentes Principales con la Matriz de Varianza y Covarianza

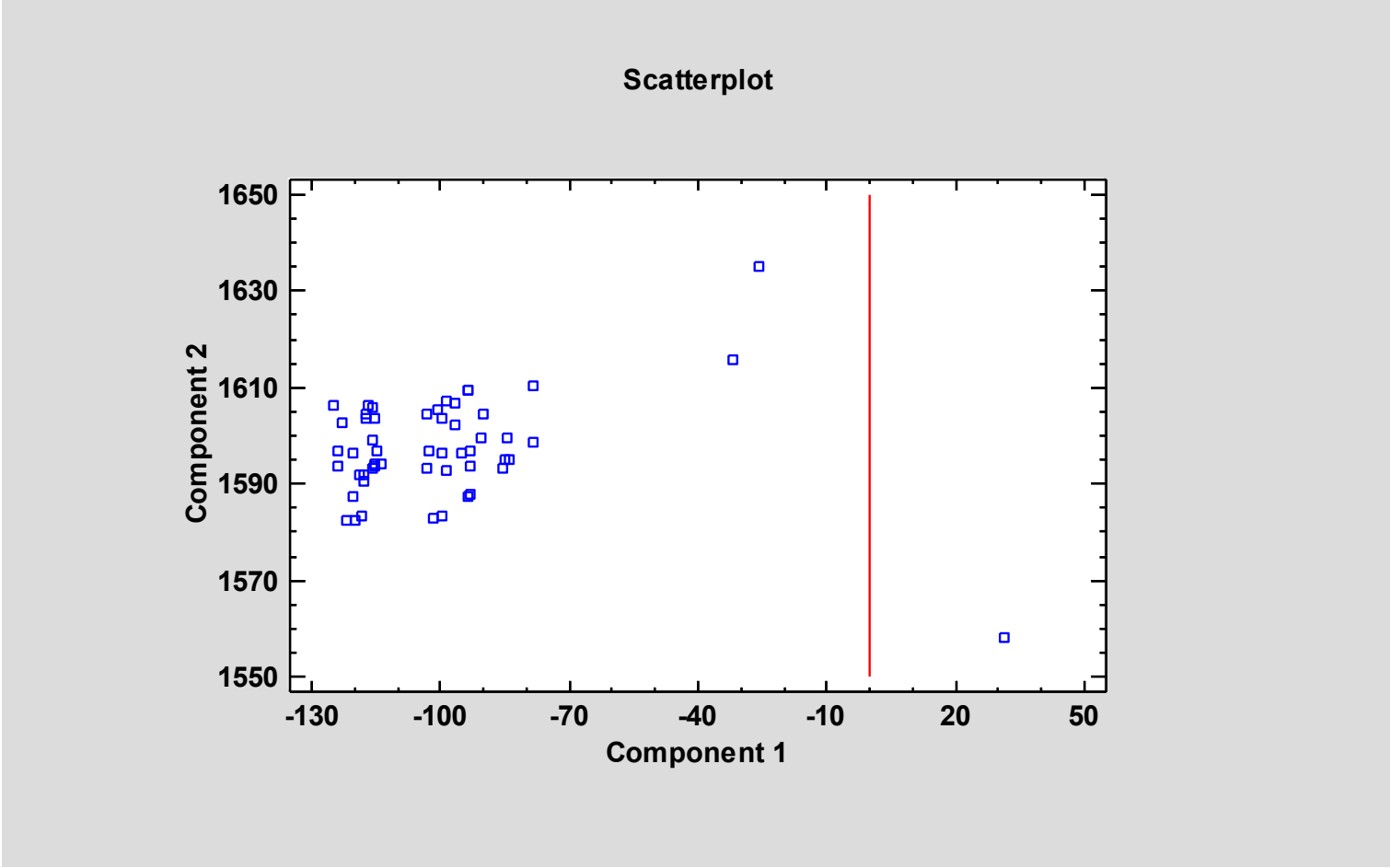
NUMERO DE COMPONENTE	Eigenvalue	PORCENTAJES DE VARIANZA	PORCENTAJE ACUMILATIVO
1	729.477	85.986	85.986
2	118.765	13.999	99.985
3	0.127759	0.015	100
4	4.04682E-05	0	100



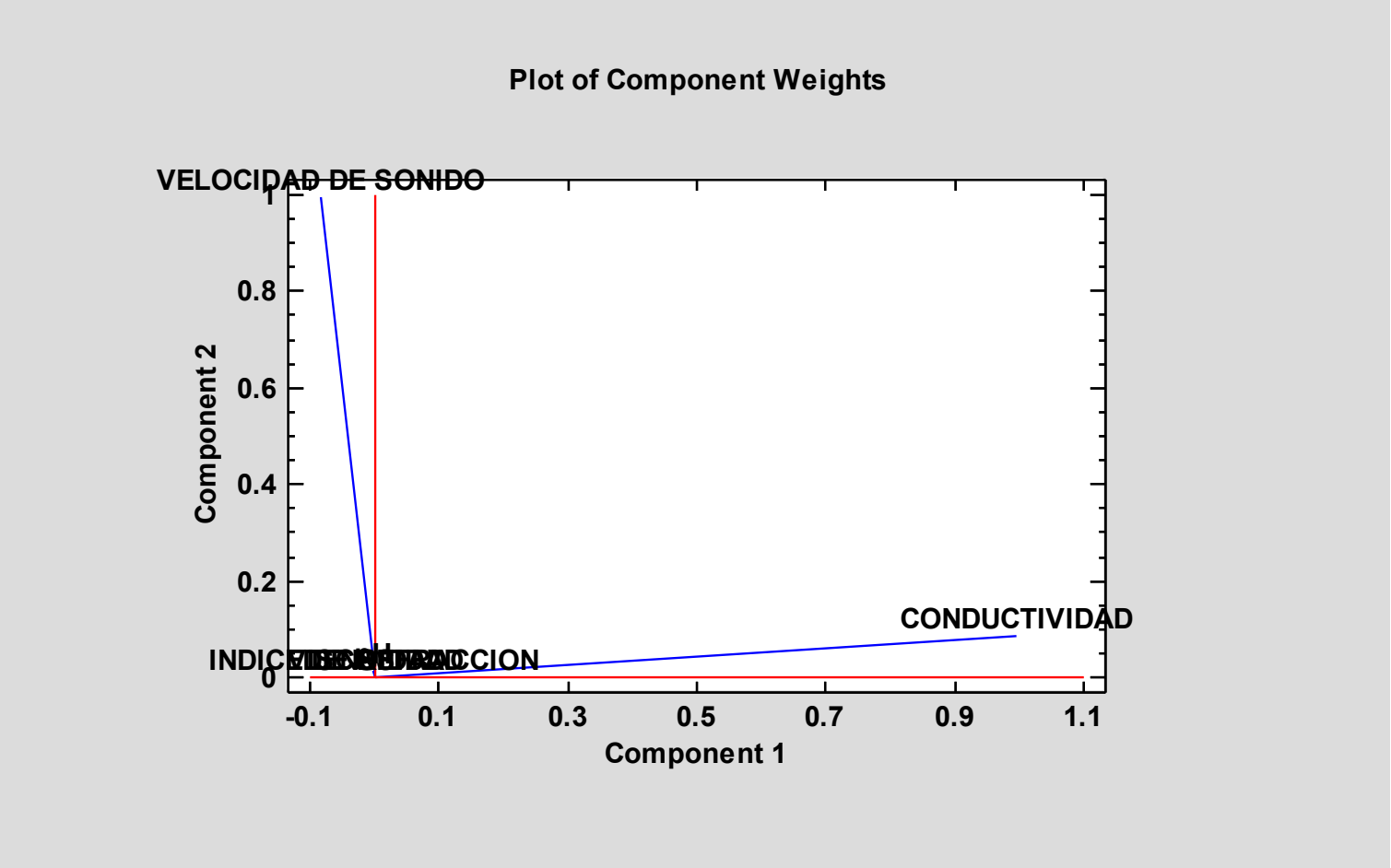
MATRIZ DE PESOS O CARGAS DE LOS COMPONENTES PRINCIPALES CON LA MATRIZ DE VARIANZA Y COVARIANZAS

	Component 1	Component 2
pH	-0.00302641	0.0126875
CONDUCTIVIDAD	0.996482	0.0837779
VISCOSIDAD	1.30E-08	1.71202E-06
DENSIDAD	0.000278134	0.000433549
VELOCIDAD DE SONIDO	-0.0837461	0.996404
INDICE DE REFRACCION	-4.67141E-05	0.000066324

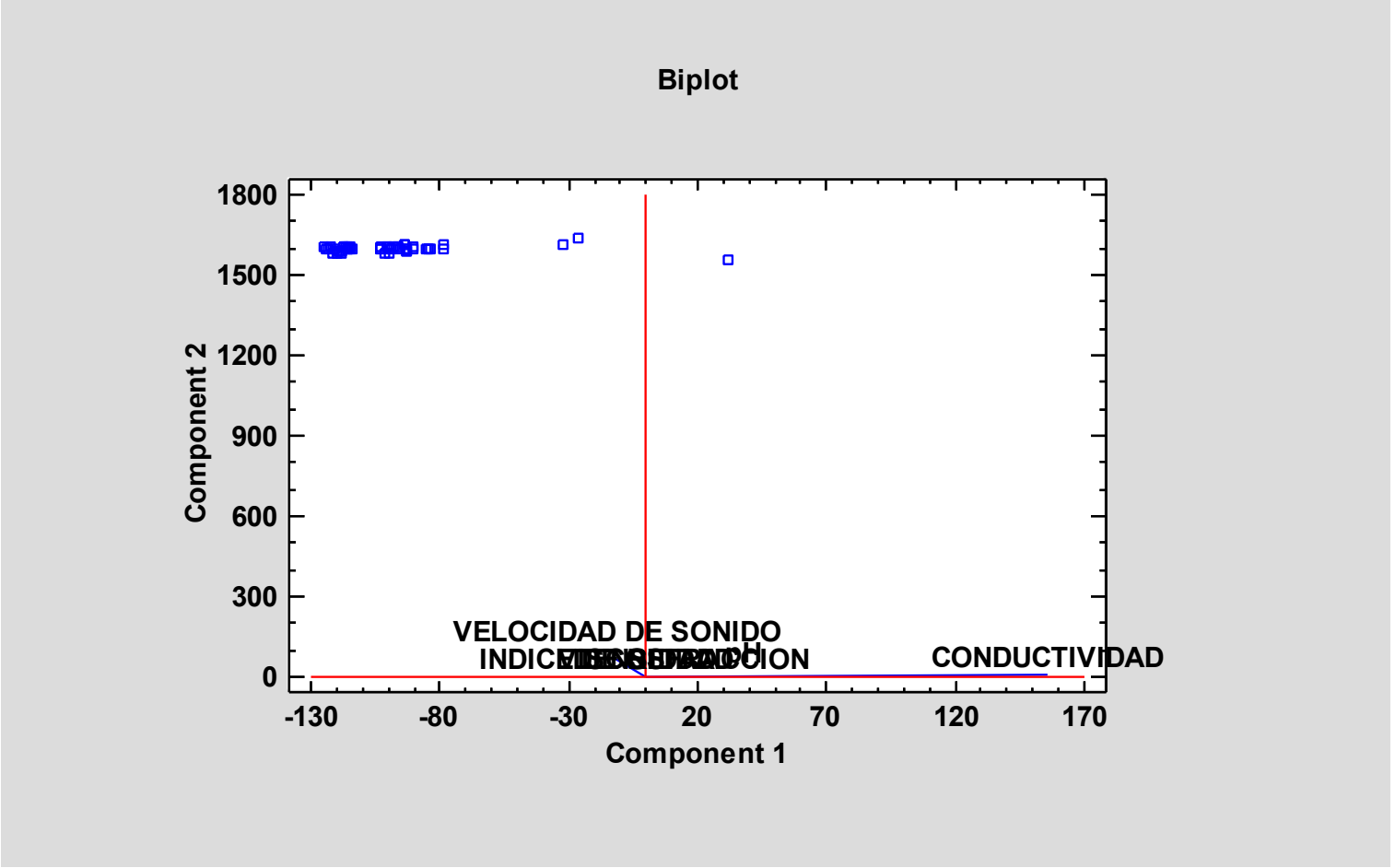
GRAFICA SCATTERPLOT



GRAFICA DE PESOS DE LOS COMPONENTES



BI GRAFICA


Biplot
1800
1500
1200
900
600
300
0
Component 2
-130
-80
-30
20
70
120
170
Component 1
VELOCIDAD DE SONIDO
CONDUCTIVIDAD

Ejemplo de Análisis de Componentes Principales con la Matriz de de Correlaciones

MEDIDAS DE TENDENCIA CENTRAL Y VARIACION DE MARCAS DE TEQUILA
DATOS ESTANDARIZADOS

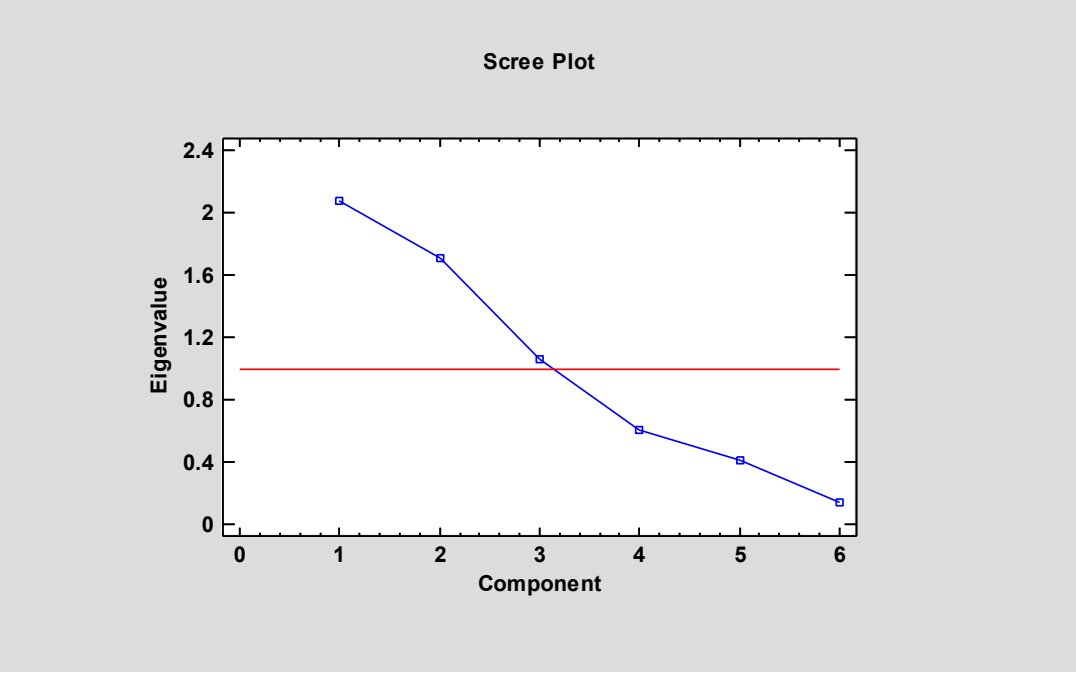
	Recuento	Promedio	Varianza	Desviación Estándar
pH	53	0	1	1
conductividad	53	0	1	1
viscosidad	53	0	1	1
Densidad	53	0	1	1
Velocidad de sonido	53	0	1	1
Índice refracción	53	0	1	1
	VARIANZA TOTAL		6	

MATRIZ DE CORRELACIONES

	pH	CONDUCTIVIDAD	VISCOSIDAD	DENSIDAD	VELOCIDAD DE SONIDO	INDICE DE REFRACCION
pH	1	-0,1965	0,0389	0,1696	0,3876	0,3630
CONDUCTIVIDAD	-0,1965	1	0,0088	0,6919	-0,1706	-0,4198
VISCOSIDAD	0,0389	0,0088	1	0,0268	0,1621	0,4446
DENSIDAD	0,1696	0,6919	0,0268	1	0,2791	-0,0665
VELOCIDAD DE SONIDO	0,3876	-0,1706	0,1621	0,2791	1	0,3275
INDICE DE REFRACCION	0,3630	-0,4198	0,4446	-0,0665	0,3275	1

Análisis de Componentes Principales con la Matriz Correlación

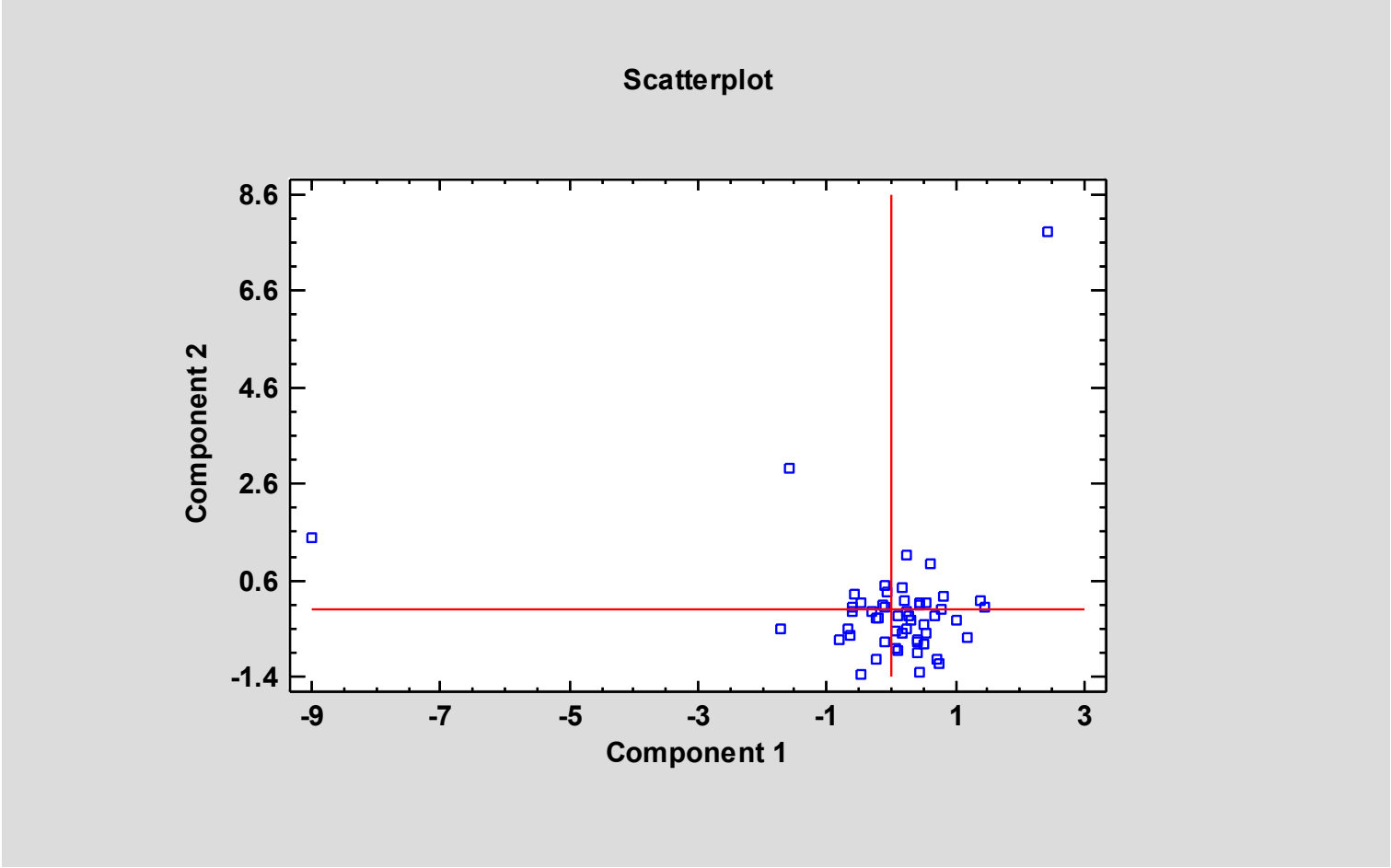
NUMERO DE COMPONENTE	Eigenvalue	PORCENTAJE DE VARIANZA	PORCENTAJE ACUMULADO
1	2.07612	34.602	34.602
2	1.70321	28.387	62.989
3	1.0622	17.703	80.692
4	0.611027	10.184	90.876
5	0.407877	6.798	97.674
6	0.139565	2.326	100



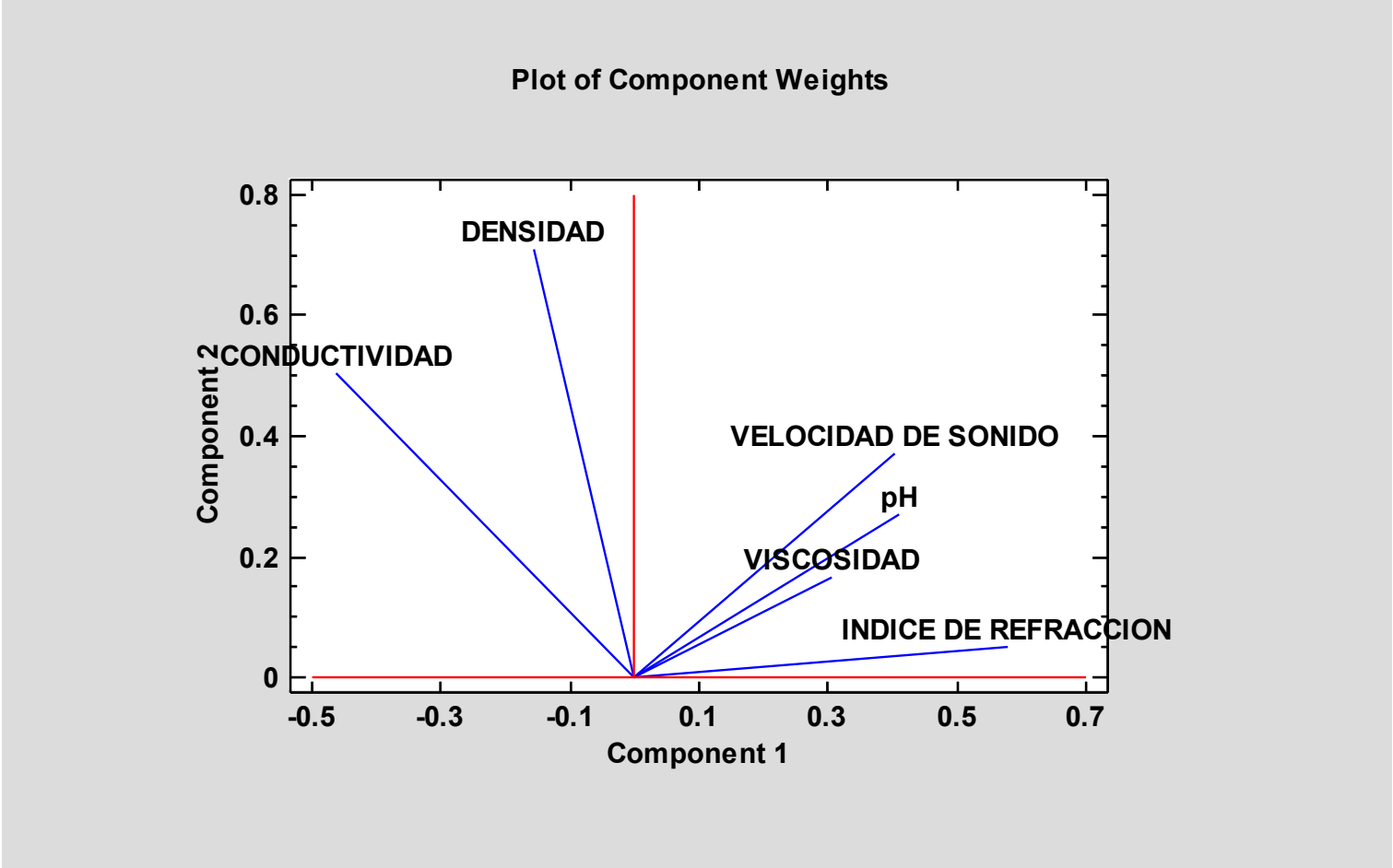
Análisis de Componentes Principales con la Matriz Correlación

	Component 1	Component 2	Component 3
pH	0.411567	0.272012	-0.45798
CONDUCTIVIDAD	-0.462813	0.502991	0.206106
VISCOSIDAD	0.307083	0.167113	0.787597
DENSIDAD	-0.156188	0.709743	-0.0379812
VELOCIDAD DE SONIDO	0.403495	0.37234	-0.25891
INDICE DE REFRACCION	0.578718	0.0520772	0.242876

Análisis de Componentes Principales con la Matriz Correlación



Análisis de Componentes Principales con la Matriz Correlación



Análisis de Componentes Principales con la Matriz Correlación