ANÁLISIS ESTADÍSTICO MULTIVARIADO

Análisis Multivariado

1. Análisis de Componentes Principales
   - Introducción
   - Determinación de las componentes

2. Análisis Multidimensional
   - Introducción
   - Justificación del método
   - Biplots

Introducción

- Consideremos un conjunto de p variables aleatorias continuas relacionadas entre si.
- El Análisis de Componentes Principales se refiere a tratar de explicar la estructura de variabilidad - relación entre las variables a través de unas pocas variables artificiales o sintéticas construidas a partir de las originales.
- Los objetivos generales de esta técnica son dos:

1. reducción de los datos
2. interpretación

- Se pretende transformar un conjunto de p variables aleatorias correlacionadas en otro conjunto más chico de variables hipotéticas no correlacionadas.

Introducción

- Las componentes principales se utilizan para describir e interpretar interdependencias entre variables y examinar relaciones que puedan existir entre los individuos.
- Se busca definir nuevas variables como combinación lineal de las X originales con la menor pérdida de información posible.
- Se desea encontrar un subespacio de dimensión $r < p$ tal que al proyectar los puntos sobre ese espacio conserven su estructura con la menor distorsión posible.
- Para ello, las distancias entre los puntos originales y sus proyeciones sobre el subespacio de dimensión r deben ser lo mas pequeñas posibles.

Enfoque descriptivo

- Consideramos un punto x_i y una dirección definida por un vector $a_1 = (a_{11}, a_{12}, ..., a_{1p})$ de norma 1, la proyección del punto x_i sobre esta dirección es:

$$z_i = a_{11}x_{i1} + \cdots + a_{1p}x_{ip} = a_1' x_i$$

y el vector que representa esta proyección será z_i. Llamando r_i a la distancia entre el punto x_i y su proyección sobre la dirección a_1, este criterio implica:

$$min \left\{ \sum_{i=1}^{n} r_i^2 \right\} = \sum_{i=1}^{n} |x_i - z_i|^2$$

Donde la notación $|u|$ representa la norma euclídea o módulo del vector u.

Representación del problema en dos dimensiones

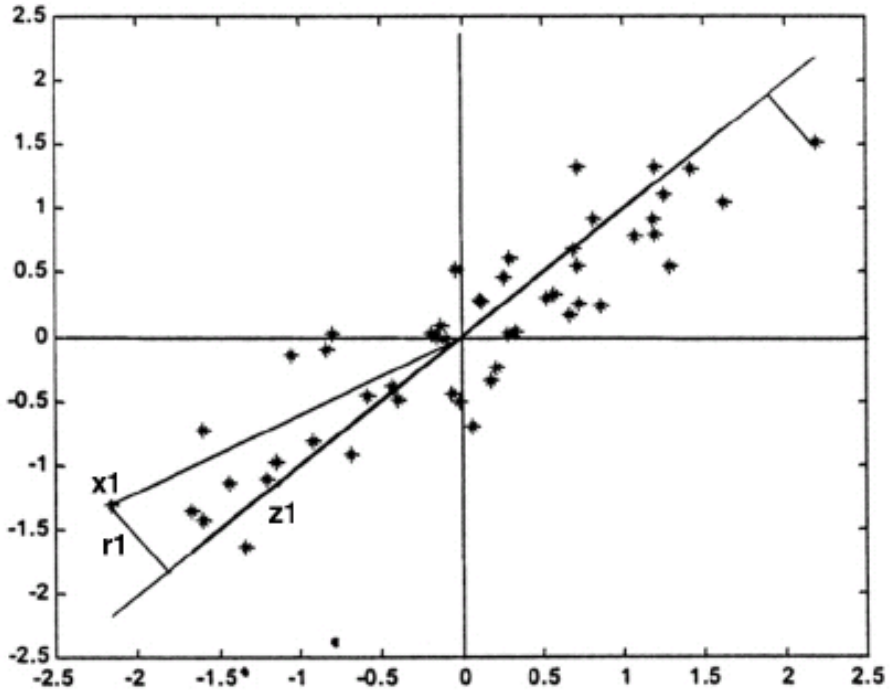


Figura 5.1. Ejemplo de la recta que minimiza las distancias ortogonales de los puntos a ella.

Enfoque descriptivo

- Al proyectar cada punto sobre la recta se forma un triángulo rectángulo donde:
  - la hipotenusa es la distancia del punto al origen, $(x_i' x_i)^{1/2}$,
  - los catetos corresponden a la proyección del punto sobre la recta z_i,
  - la distancia entre el punto y su proyección r_i respectivamente.
- Aplicando el teorema de Pitágoras y sumando para todos los individuos se tiene:

$$\sum_{i=1}^{n} x_i' x_i = \sum_{i=1}^{n} z_i^2 + \sum_{i=1}^{n} r_i^2$$

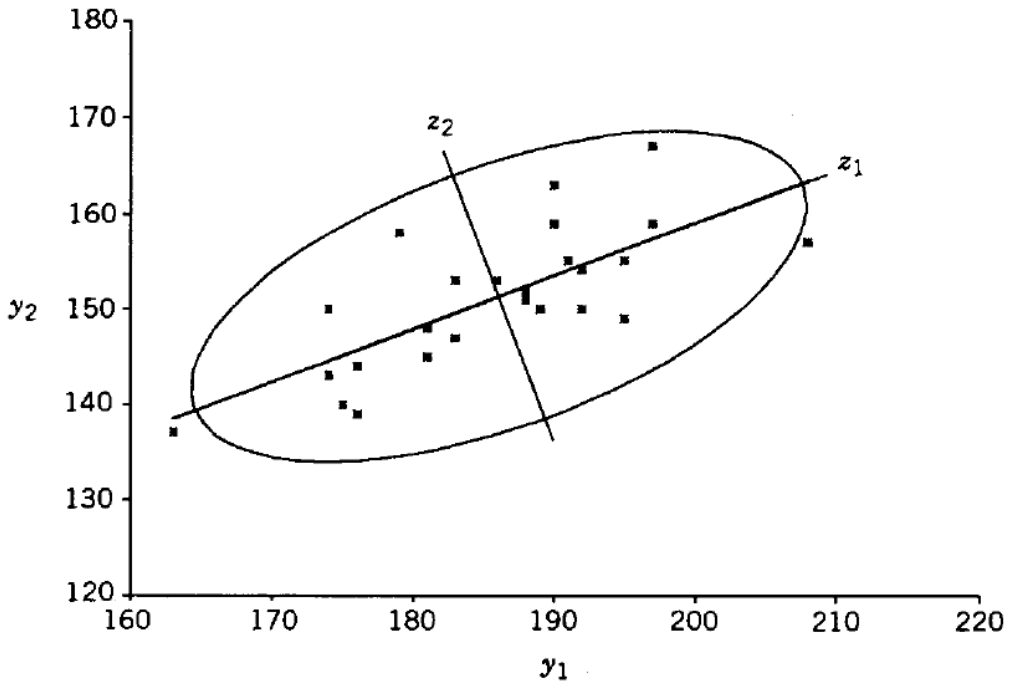
- El primer término es constante y minimizar la suma de las distancias a la recta de todos los puntos es equivalente a maximizar la suma al cuadrado de los valores de las proyecciones.

Enfoque descriptivo

- Las proyecciones z_i tienen media 0, por lo tanto maximizar la suma de sus cuadrados es equivalente a maximizar su variancia.
- El criterio es hallar la dirección de proyección que maximice la variancia de los datos proyectados.
- Un criterio equivalente: buscar la dirección tal que los puntos proyectados sobre ella conserven lo mejor posible sus distancias relativas.
- Representar puntos p dimensionales con la mínima pérdida de información en un espacio de dimensión 1 es equivalente a sustituir las p variables originales por una nueva variable z_1 que resuma óptimamente la información.
- La nueva variable debe tener máxima correlación con las variables originales y debe ser posible recuperar o pronosticar los valores de las variables originales con la máxima precisión.
- Para pronosticar los datos observados con la mínima pérdida de información se debe hallar la variable (dirección) de máxima variabilidad.

Enfoque geométrico

- En el gráfico anterior los puntos se sitúan formando una elipse, y podemos describirlos por su proyección en la dirección del eje mayor de la elipse.
- Se puede demostrar que este eje es la recta que minimiza las distancias ortogonales con lo cual volvemos al problema que ya planteamos y resolvimos.
- Hallar los ejes de este elipse (elipsoide) es equivalente a encontrar una matriz ortogonal que rote los ejes de los datos para alinearlos a los ejes naturales de la nube de puntos



Determinación de las componentes

- El primer componente principal se define como la combinación lineal de las variables originales que tiene variancia máxima:

$$z_1 = Xa_1$$

- Su variancia será:

$$\frac{1}{n} z_1' z_1 = \frac{1}{n} a_1' X' X a_1 = a_1' S a_1$$

donde S es la matriz de variancias y covariancias de las observaciones.

- Introducimos la restricción $a_1' a_1 = 1$ y el problema que debemos resolver es:

$$M = a_1' S a_1 - \lambda(a_1' a_1 - 1)$$

Determinación de las componentes

$$M = a_1' S a_1 - \lambda(a_1' a_1 - 1)$$

- Derivando respecto a a_1 e igualando a 0 y obtenemos:

$$\frac{\partial M}{\partial a_1} = 2Sa_1 - 2\lambda a_1 = 0$$

y obtenemos

$$Sa_1 = \lambda a_1$$

- La solución anterior implica que a_1 es un autovector de la matriz S y λ su correspondiente autovalor.
- Para determinar que autovalor de S es el que corresponde a la solución anterior, podemos premultiplicar por a_1' esta ecuación:

$$a_1' S a_1 = \lambda a_1' a_1$$

Resulta que $Var(z_1) = \lambda$ entonces se elige el el autovalor más grande de la matriz S y su correspondiente autovector a_1 contiene los coeficientes de cada variable en la combinación lineal que define al primer componente principal.

Determinación de las componentes

- Ahora busquemos el mejor plano de proyección de las variables X.
- Establecemos como función objetivo que la suma de las variancias de $z_1 = Xa_1$ y $z_2 = Xa_2$ sea máxima, donde a_1 y a_2 son los vectores que definen el plano:

$$Q = a_1' S a_1 + a_2' S a_2 - \lambda_1(a_1' a_1 - 1) - \lambda_2(a_2' a_2 - 1)$$

- La solución de este sistema es :

$$Sa_1 = \lambda_1 a_1$$

$$Sa_2 = \lambda_2 a_2$$

- Las componentes principales no están correlacionadas entre sí:

$$Cov(z_1, z_2) = Cov(Xa_1, Xa_2) = \frac{1}{n} a_1' X' X a_2 = a_1' S a_2 = \lambda_2 a_1' a_2 = 0$$

Determinación de las componentes

- Si en lugar de maximizar variancia total (la traza de la matriz de covariancias de la proyección) se maximiza la variancia generalizada (el determinante de la matriz de covariancias) se obtiene el mismo resultado.
- Puede demostrarse que el espacio de dimensión r que mejor representa a los puntos viene definido por los autovectores asociados a los r autovalores mas grandes de la matriz S. Estas direcciones se denominan direcciones principales.
- En general la matriz X (y por lo tanto S) tiene rango p, entonces existen tantas componentes principales como variables originales.
- Calcular los componentes principales equivale a aplicar una transformación ortogonal A a las variables X (ejes originales) para obtener unas nuevas variables Z no correlacionadas entre si, $Z = XA$ donde $A'A = I$. Esta operación se puede interpretar como la elección de nuevos ejes coordenados, que coincidan con los ejes naturales de los datos.

Generalización

- Los autovalores $\lambda_1, \lambda_2, \ldots, \lambda_p$ de la matriz de variancias y covariancias S se obtienen resolviendo $|S - \lambda I| = 0$ y sus vectores asociados son:

$$(S - \lambda_i I)a_i = 0$$

- La matriz S es simétrica y (semi)definida positiva, por lo tanto los autovalores son reales y positivos, los autovectores son ortogonales.
- Si el rango de S fuera $r < p$ habrá solamente r autovalores positivos y el resto serán iguales a 0.
- La matriz de variancias y covariancias de las componentes principales Z será

$$S_z = A'SA = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & \vdots \\ \vdots & \ldots & \ddots & \vdots \\ 0 & \ldots & \ldots & \lambda_p \end{pmatrix}$$

Propiedades

- Las Componentes Principales conservan la variabilidad inicial: la suma de las variancias de los componentes es igual a la suma de las variancias de las variables originales y la variancia generalizada de los componentes es igual a la original.

$$\sum_{i=1}^{p} var(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} var(z_i)$$

- La proporción de variabilidad explicada por un componente es el cociente entre su variancia (el autovalor asociado al autovector que lo define) y la suma de los autovalores de la matriz S.

$$Variancia\ Explicada(z_i) = \frac{\lambda_i}{(\lambda_1 + \lambda_2 + \cdots + \lambda_p)}$$

Propiedades

- La correlación entre un componente principal y una variable X es proporcional al coeficiente de esa variable en la definición de la componente, donde el coeficiente de proporcionalidad es el cociente entre el desvío estándar del componente y el desvío estándar de la variable.

$$corr(z_i, X_k) = \frac{a_{ik}\sqrt{\lambda_i}}{\sqrt{var(X_k)}}$$

- Si se estandarizan los componentes principales dividiendo cada uno por su desvío estándar se obtiene la estandarización multivariante de los datos originales.

Componentes principales a partir de R

- Las componentes principales se obtienen maximizando la variancia de la proyección. En términos de las variables originales esto supone maximizar:

$$M = \sum_{i=1}^{p} a_i^2 s_i^2 + 2 \sum_{i=1}^{p} \sum_{j=i+1}^{p} a_i a_j s_{ij}$$

  sujeto a la restricción $a'a = 1$.

- Cuando las escalas de medida de las variables son muy distintas, la maximización de M dependerá decisivamente de estas escalas y las variables con valores mas grandes tendrán mayor peso en el análisis.
- Con variables estandarizadas el problema de maximización es:

$$M' = 1 + 2 \sum_{i=1}^{p} \sum_{j=i+1}^{p} a_i a_j r_{ij}$$

Componentes principales a partir de R

- La solución depende de las correlaciones y no de las variancias.
- Las componentes principales normadas se obtienen calculando los autovectores y autovalores de la matriz de correlación. Llamando λ_i^R $(i = 1, 2, \ldots, p)$ a las raíces características de esa matriz se verifica:

$$\sum_{i=1}^{p} \lambda_i^R = traza(R) = p$$

Las propiedades de los componentes extraídos de R son:

- La proporción de variancia explicada por cada componente será λ_i^R / p
- Las correlaciones entre cada componente $z_j = Xa_j$ y las variables originales X vienen dados directamente por $a'_i \sqrt{\lambda_j^R}$.

Selección de la cantidad de componentes

Existen varias reglas:

- Seleccionar componentes hasta cubrir una proporción determinada de variancia.
- Seleccionar aquellos componentes asociados a autovalores mayores a determinado umbral o cota, que suele fijarse igual a la variancia media $\sum_{i=1}^{p} \lambda_i / p$.
- Realizar un gráfico de los autovalores de S o R. Seleccionar componentes hasta que los restantes tengan valores de λ_i aproximadamente iguales.

Aglomerado / Indicador	Tasa de actividad			Tasa de empleo			Tasa de desocupación			Asalariados s/desc. jubilatorio
	Total	Mujeres	Varones	Total	Mujeres	Varones	Total	Mujeres	Varones	
CABA	**63.7**	55.6	73.4	**59.3**	51.1	69.1	**6.9**	8.0	6.0	**23.3**
Partidos del GBA	**58.8**	48.4	70.7	**52.1**	42.2	63.2	**11.4**	12.7	10.5	**37.7**
Gran Mendoza	**59.4**	47.9	72.1	**55.9**	44.3	68.7	**5.9**	7.4	4.7	**39.2**
Gran San Juan	**50.9**	39.9	63.2	**49.3**	38.3	61.7	**3.1**	4.0	2.5	**43.2**
Gran San Luis	**53.7**	45.4	62.9	**52.2**	44.5	60.8	**2.8**	2.0	3.4	**43.7**
Corrientes	**54.3**	43.6	66.1	**50.4**	40.7	61.1	**7.2**	6.7	7.6	**38.7**
Formosa	**47.0**	33.2	62.4	**45.3**	31.9	60.2	**3.6**	4.0	3.5	**30.2**
Gran Resistencia	**47.8**	36.0	62.1	**47.0**	35.4	61.0	**1.6**	1.4	1.8	**35.0**
Posadas	**58.2**	48.8	69.2	**56.1**	47.6	66.1	**3.6**	2.6	4.5	**38.6**
Gran Catamarca	**59.4**	51.0	68.0	**54.2**	45.7	63.0	**8.7**	10.4	7.4	**35.4**
Gran Tucumán - Tafí Viejo	**57.5**	47.7	68.8	**53.1**	43.1	64.7	**7.6**	9.5	6.0	**45.9**
Jujuy-Palpalá	**57.7**	49.3	67.0	**53.8**	45.1	63.4	**6.8**	8.4	5.5	**32.8**
La Rioja	**53.5**	43.2	64.7	**51.0**	41.2	61.6	**4.7**	4.7	4.7	**32.9**
Salta	**61.1**	53.2	70.6	**55.6**	47.7	65.2	**8.9**	10.3	7.6	**46.5**
Santiago del Estero - La Banda	**56.3**	44.3	69.5	**53.8**	42.0	66.6	**4.6**	5.1	4.2	**45.5**
Bahía Blanca-Cerri	**57.5**	47.6	69.1	**52.1**	42.4	63.5	**9.4**	10.9	8.1	**29.8**
Concordia	**55.8**	47.1	65.1	**53.0**	44.7	61.9	**5.0**	5.2	4.8	**44.6**
Gran Córdoba	**59.9**	51.8	68.8	**54.4**	46.0	63.6	**9.2**	11.2	7.5	**42.8**
Gran La Plata	**57.8**	50.6	65.5	**54.5**	48.5	61.0	**5.7**	4.3	6.8	**34.5**
Gran Rosario	**57.6**	47.8	69.2	**50.3**	40.1	62.2	**12.8**	16.0	10.1	**32.2**
Gran Paraná	**52.2**	41.5	64.1	**49.2**	39.0	60.7	**5.6**	6.1	5.3	**23.6**
Gran Santa Fe	**52.0**	42.1	63.5	**50.4**	40.8	61.6	**3.0**	2.9	3.0	**29.4**
Mar del Plata	**59.2**	50.9	68.8	**51.7**	43.3	61.2	**12.8**	14.8	11.1	**36.7**
Río Cuarto	**59.3**	46.7	74.2	**54.5**	41.9	69.3	**8.1**	10.3	6.5	**44.7**
Santa Rosa-Toay	**54.0**	45.9	63.5	**48.5**	40.5	57.9	**10.1**	11.6	8.8	**32.2**
San Nicolás-Villa Constitución	**55.1**	44.8	66.7	**48.9**	40.3	58.6	**11.3**	10.0	12.2	**32.4**
Comodoro Rivadavia-Rada Tilly	**50.4**	40.4	61.8	**48.9**	39.3	59.9	**2.9**	2.7	3.1	**18.9**
Neuquén-Plottier	**55.5**	45.5	67.7	**52.9**	43.4	64.3	**4.8**	4.5	5.0	**19.2**
Río Gallegos	**56.6**	50.0	64.2	**51.5**	46.7	57.0	**9.0**	6.6	11.2	**18.2**
Ushuaia-Río Grande	**54.7**	40.5	68.5	**50.9**	38.7	62.8	**6.9**	4.4	8.3	**9.5**
Viedma-Carmen de Patagones	**51.4**	43.6	61.1	**48.3**	40.4	58.0	**6.1**	7.3	5.0	**19.8**
Rawson-Trelew	**60.1**	51.8	69.4	**54.3**	47.5	61.8	**9.7**	8.2	10.9	**28.6**

Matriz de correlaciones

	Tasa de actividad			Tasa de empleo			Tasa de desocupación			Asalariados
Indicadores	**Total**	**Mujeres**	**Varones**	**Total**	**Mujeres**	**Varones**	**Total**	**Mujeres**	**Varones**	**s/desc.**
Tasa de actividad	1									
TA_Mujeres	0.942	1								
TA_Varones	0.860	0.648	1							
Tasa de empleo	0.889	0.829	0.773	1						
TE_Mujeres	0.844	0.925	0.523	0.892	1					
TE_Varones	0.598	0.374	0.827	0.756	0.389	1				
Tasa de desocupación	0.597	0.578	0.495	0.164	0.258	-0.042	1			
TD_Mujeres	0.583	0.552	0.526	0.180	0.197	0.102	0.938	1		
TD_Varones	0.531	0.530	0.394	0.125	0.290	-0.190	0.931	0.748	1	
Asalariados s/desc. Jub.	0.246	0.208	0.269	0.283	0.162	0.366	0.024	0.179	-0.132	1

Matriz de Covariancias

	Tasa de actividad			Tasa de empleo			Tasa de desocupación			Asalariados
Indicadores	**Total**	**Mujeres**	**Varones**	**Total**	**Mujeres**	**Varones**	**Total**	**Mujeres**	**Varones**	**s/desc.**
Tasa de actividad	15.030									
TA_Mujeres	18.057	24.447								
TA_Varones	11.605	11.153	12.128							
Tasa de empleo	10.301	12.248	8.047	8.934						
TE_Mujeres	13.042	18.249	7.258	10.637	15.906					
TE_Varones	7.132	5.689	8.863	6.949	4.776	9.469				
Tasa de desocupación	7.051	8.705	5.250	1.495	3.132	-0.390	9.282			
TD_Mujeres	8.527	10.300	6.902	2.025	2.962	1.180	10.781	14.221		
TD_Varones	5.797	7.379	3.864	1.054	3.262	-1.650	7.987	7.944	7.938	
Asalariados s/desc. Jub.	9.061	9.770	8.925	8.051	6.156	10.705	0.703	6.421	-3.539	90.561

Componentes principales sobre la matriz de covariancias

observaciones	32
componentes	10
traza	207.9147

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	106.325	41.727	0.511	0.511
Comp2	64.598	41.462	0.311	0.822
Comp3	23.136	11.554	0.111	0.933
Comp4	11.582	9.412	0.056	0.989
Comp5	2.171	2.089	0.010	1.000
Comp6	0.082	0.068	0.000	1.000
Comp7	0.014	0.010	0.000	1
Comp8	0.004	0.001	0	1
Comp9	0.003	0.002	0	1
Comp10	0.001	.	0	1

Variable	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Comp8	Comp9	Comp10
Tasa de actividad	0.257	0.348	-0.066	0.081	0.081	0.578	-0.372	-0.004	-0.160	-0.545
TA_Mujeres	0.303	0.449	-0.071	-0.351	-0.261	-0.327	-0.511	0.026	0.260	0.272
TA_Varones	0.211	0.232	-0.052	0.561	0.303	-0.334	-0.122	-0.188	-0.512	0.262
Tasa de empleo	0.190	0.203	-0.321	0.056	0.026	0.562	0.324	-0.110	0.179	0.593
TE_Mujeres	0.211	0.319	-0.333	-0.416	0.057	-0.261	0.574	0.031	-0.285	-0.295
TE_Varones	0.170	0.068	-0.305	0.579	-0.158	-0.222	0.164	0.252	0.540	-0.286
Tasa de desocupación	0.093	0.230	0.461	0.031	0.111	0.076	0.142	0.806	-0.100	0.163
TD_Mujeres	0.158	0.224	0.561	0.165	-0.586	0.034	0.300	-0.377	-0.046	-0.066
TD_Varones	0.044	0.229	0.378	-0.082	0.666	-0.081	0.118	-0.309	0.477	-0.097
Asalariados s/desc. Jub.	0.808	-0.567	0.105	-0.098	0.071	-0.002	-0.001	0.002	0.001	0.000

Componentes principales sobre la matriz de correlaciones

observaciones	32
componentes	10
traza	10

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	5.581	3.189	0.558	0.558
Comp2	2.392	1.307	0.239	0.797
Comp3	1.085	0.336	0.109	0.906
Comp4	0.749	0.563	0.075	0.981
Comp5	0.186	0.179	0.019	0.999
Comp6	0.006	0.006	0.001	1.000
Comp7	0.001	0.001	0.000	1.000
Comp8	0.000	0.000	0	1
Comp9	0.000	0.000	0	1
Comp10	0.000	.	0	1

Variable	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Comp8	Comp9	Comp10
Tasa de actividad	0.420	-0.047	-0.078	-0.016	0.008	-0.474	-0.489	0.014	-0.046	0.591
TA_Mujeres	0.394	0.000	-0.257	0.254	-0.261	0.362	-0.491	-0.004	0.346	-0.394
TA_Varones	0.368	-0.112	0.245	-0.422	0.285	0.290	-0.207	-0.118	-0.580	-0.235
Tasa de empleo	0.356	-0.327	-0.179	-0.005	0.031	-0.598	0.342	-0.119	0.036	-0.494
TE_Mujeres	0.337	-0.176	-0.452	0.305	-0.006	0.372	0.454	0.079	-0.300	0.343
TE_Varones	0.244	-0.422	0.305	-0.432	-0.084	0.225	0.224	0.162	0.546	0.227
Tasa de desocupación	0.282	0.474	0.134	-0.027	-0.018	-0.102	0.140	0.792	-0.052	-0.134
TD_Mujeres	0.280	0.396	0.327	-0.009	-0.621	-0.018	0.243	-0.445	-0.095	0.074
TD_Varones	0.246	0.492	-0.085	-0.027	0.633	0.065	0.157	-0.339	0.376	0.063
Asalariados s/desc. Jub.	0.121	-0.215	0.637	0.690	0.237	0.004	-0.001	0.007	0.003	-0.001

Análisis Multidimensional

- Estas técnicas son una generalización de la idea de componentes principales cuando en lugar de disponer de una matriz de observaciones por variables, como en componentes principales, se dispone de una matriz D $(n \times n)$ de distancias o disimilaridades entre los n elementos de un conjunto.
- El objetivo es representar esta matriz mediante un conjunto de (pocas) variables ortogonales, de manera que las distancias euclídeas entre los elementos respecto a estas variables sean iguales o lo mas próximas posibles a las distancias de la matriz original.
- Cuando $p > 2$ las variables pueden ordenarse en importancia y suelen hacerse representaciones gráficas para entender la estructura existente en los datos.

Análisis Multidimensional

- En general no siempre es posible encontrar p variables que reproduzcan exactamente las distancias iniciales, sin embargo es frecuente encontrar variables que las reproduzcan de manera aproximada.
- El análisis multidimensional comparte con componentes principales el objetivo de describir e interpretar los datos.
- Mientras que el análisis de componentes principales se basa en la matriz R o S de correlaciones o covariancias, el análisis multidimensional utiliza la matriz $D_{n\times n}$ de distancias entre individuos con el fin de analizar su estructura.
- Ambos enfoques están claramente relacionados y existen técnicas gráficas como el biplot que aprovechan esta dualidad para representar conjuntamente las variables y los individuos en un mismo gráfico.

Análisis Multidimensional

- Consideremos la matriz de datos centrada, $\widetilde{X}_{(n\times p)}$. A partir de esta matriz podemos construir dos tipos de matrices cuadradas y semidefinidas positivas: la matriz de variancias y covariancias $S_{(p\times p)}$ y la matriz de productos cruzados $Q_{(n\times n)}$:

$$S = \frac{1}{n}\widetilde{X}'\widetilde{X}$$

$$Q = \widetilde{X}\widetilde{X}'$$

- Identificaremos con q_{ij} a los elementos de la matriz Q, donde

$$q_{ij} = \sum_{s=1}^{p} x_{is}x_{js} = x_i'x_j$$

$$q_{ij} = |x_i||x_j|cos(\theta_{ij})$$

Análisis Multidimensional

- Las distancias entre las observaciones se deducen a partir de los elementos de Q:

$$d_{ij}^2 = \sum_{s=1}^{p}(x_{is} - x_{js})^2 = \sum_{s=1}^{p} x_{is}^2 + \sum_{s=1}^{p} x_{js}^2 - 2\sum_{s=1}^{p} x_{is}x_{js}$$

$$d_{ij}^2 = q_{ii} + q_{jj} - 2q_{ij}$$

Por lo tanto:

$$D = Diag(Q) \cdot 1' + 1 \cdot Diag(Q)' - 2Q$$

donde $Diag(Q)$ es un vector que contiene los elementos diagonales de Q y es de dimensión $(n \times 1)$ al igual que el vector 1 de unos.

Análisis Multidimensional

- Ahora pensemos en el problema inverso, es posible reconstruir $\widetilde{X}$ a partir de la matriz $D = (d_{ij}^2)$?
- Sin pérdida de generalidad supondremos que los datos están centrados, de manera tal que la suma de los elementos fila y columna de la matriz Q es igual a 0.

$$\sum_{i=1}^{n} d_{ij}^2 = \sum_{i=1}^{n} q_{ii} + \sum_{i=1}^{n} q_{jj} - 2\sum_{i=1}^{n} q_{ij} = tr(Q) + nq_{jj}$$

por lo tanto podemos despejar:

$$q_{jj} = \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 - \frac{1}{n} tr(Q)$$

Análisis Multidimensional

- sumando nuevamente sobre j

$$\sum_{j=1}^{n}\sum_{i=1}^{n} d_{ij}^2 = \sum_{j=1}^{n}(tr(Q) + nq_{jj})$$

$$\sum_{j=1}^{n}\sum_{i=1}^{n} d_{ij}^2 = n \cdot tr(Q) + n\sum_{j=1}^{n} q_{jj} = 2n \cdot tr(Q)$$

de donde se puede despejar:

$$\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n} d_{ij}^2 = \frac{2tr(Q)}{n}$$

Análisis Multidimensional

- Ahorar trataremos de recuperar q_{ij} a partir de d_{ij}^2:

$$d_{ij}^2 = q_{ii} + q_{jj} - 2q_{ij}$$

$$d_{ij}^2 = \left(\frac{1}{n} \sum_{j=1}^{n} d_{ij}^2 - \frac{tr(Q)}{n} \right) + \left(\frac{1}{n} \sum_{i=1}^{n} d_{ij}^2 - \frac{tr(Q)}{n} \right) - 2q_{ij}$$

$$d_{ij}^2 = d_{i\cdot}^2 + d_{\cdot j}^2 - \frac{2tr(Q)}{n} - 2q_{ij}$$

$$d_{ij}^2 = d_{i\cdot}^2 + d_{\cdot j}^2 - d_{\cdot\cdot}^2 - 2q_{ij}$$

$$q_{ij} = -\frac{1}{2}(d_{ij}^2 - d_{i\cdot}^2 - d_{\cdot j}^2 + d_{\cdot\cdot}^2)$$

$$Q = -\frac{1}{2}PDP \qquad P = (I_n - n^{-1}11')$$

Análisis Multidimensional

- Una vez recuperados los elementos de Q trataremos de recuperar $\widetilde{X}$.
- Si Q es definida positiva de rango p sabemos que $Q = V\Lambda V'$ donde V es de dimensión $n \times p$ y contiene los autovectores de Q mientras que Λ es una matriz diagonal de dimensión $p \times p$ que contiene los autovalores de Q.
- Por lo tanto $Q = (V\Lambda^{1/2})(\Lambda^{1/2}V') = YY'$.
- De este modo hemos obtenido una matriz Y de dimensión $n \times p$ con p variables no correlacionadas que reproducen la métrica inicial.

Análisis Multidimensional

- Es importante destacar que partimos de una matriz de datos X, calculamos la matriz centrada $\widetilde{X}$, hallamos Q y finalmente D.
- Si queremos deshacer este camino a partir de D para volver a la matriz inicial X no obtenemos las variables originales sino sus componentes principales.
- Esto es inevitable ya que existe una indeterminación en el problema cuando la única información disponible viene dada por las distancias entre individuos.
- Las distancias entre individuos son invariantes si:
  - Modificamos las medias de las variables.
  - Rotamos los puntos, es decir multiplicamos la matriz de datos por una matriz ortogonal.

Análisis Multidimensional

- Las distancias se obtienen a partir de los términos de la matriz de similitudes Q y esta matriz es invariante ante rotaciones de las variables:

$$Q = \widetilde{X}\widetilde{X}' = \widetilde{X}AA'\widetilde{X}'$$

para cualquier matriz A ortogonal. La matriz Q solo contiene información sobre el espacio generado por las variables X.

- Cualquier rotación preserva las distancias entre individuos, en consecuencia cualquier rotación de las variables originales podría ser solución.

Análisis Multidimensional

- Dada una matriz de distancias D, diremos que esta matriz es compatible con una métrica euclídea si la matriz de similitudes que se obtiene a partir de ella es semidefinida positiva.

$$Q = -\frac{1}{2}PDP$$

  donde $P = (I_n - n^{-1}11')$. Esta es una condición necesaria y suficiente.
- Puede ocurrir que la matriz de distancias no satisfaga la condición anterior, sin embargo es frecuente que la matriz de similitudes obtenida a partir de ella tenga h autovalores positivos y mas grandes que el resto.
- Si los restantes autovalores no nulos son mucho menores (en valor absoluto) podemos obtener una representación aproximada de los puntos utilizando los h autovectores asociados a los h autovalores mas grandes.

Construcción de las coordenadas principales

- Construir $Q = -0.5PDP$ de productos cruzados.
- Obtener los r autovalores mas grandes de Q.
- Obtener las coordenadas principales como $\sqrt{\lambda_i} v_i$ donde λ_i es un autovalor de Q y v_i es su autovector asociado.
- De este modo aproximaremos:

$$Q \approx (V_r \lambda_r^{1/2})(\lambda_r^{1/2} V_r')$$

y las coordenadas principales se definen como $Y = V_r \lambda_r^{1/2}$.

Análisis multidimensional y componentes principales

- Cuando los datos originales están en la matriz X de individuos por variables y calculamos la matriz de distancias D a partir de las distancias euclídeas entre individuos, las coordenadas principales obtenidas a partir de la matriz D son equivalentes a los componentes principales obtenidos a partir de las variables.
- Las matrices $\tilde{X}'\tilde{X}$ y $\tilde{X}\tilde{X}'$ tienen el mismo rango y los mismos autovalores no nulos.
- Si a_i es autovector de $\tilde{X}'\tilde{X}$ asociado al autovalor λ_i tenemos:

$$(\tilde{X}'\tilde{X})a_i = \lambda_i a_i$$

y multiplicando por $\tilde{X}$:

$$(\tilde{X}\tilde{X}')(\tilde{X}a_i) = \lambda_i(\tilde{X}a_i)$$

Análisis multidimensional y componentes principales

- Si $n > p$ y la matriz $\tilde{X}'\tilde{X}$ tiene rango completo, tendrá p autovalores no nulos que serán los autovalores no nulos de $\tilde{X}\tilde{X}'$.
- Los autovectores de $\tilde{X}\tilde{X}'$ son las proyecciones de la matriz $\tilde{X}$ sobre la dirección de los autovectores de $\tilde{X}'\tilde{X}'$.
- La matriz $Z = XA$ de dimensión $(n \times p)$ contiene los valores de los p componentes principales correspondientes a los n individuos donde A contiene los autovectores de $\tilde{X}'\tilde{X}$.
- La matriz de coordenadas principales es $Y = V\Lambda^{1/2}$ donde la matriz V contiene los autovectores de $\tilde{X}'\tilde{X}$ y sabemos que $V = XA$.

Biplots

- El biplot es un gráfico conjunto de las observaciones y las variables.
- La representación se obtiene a partir de la siguiente descomposición de la matriz $\tilde{X}$:

$$\tilde{X} = UD^{1/2}A'$$

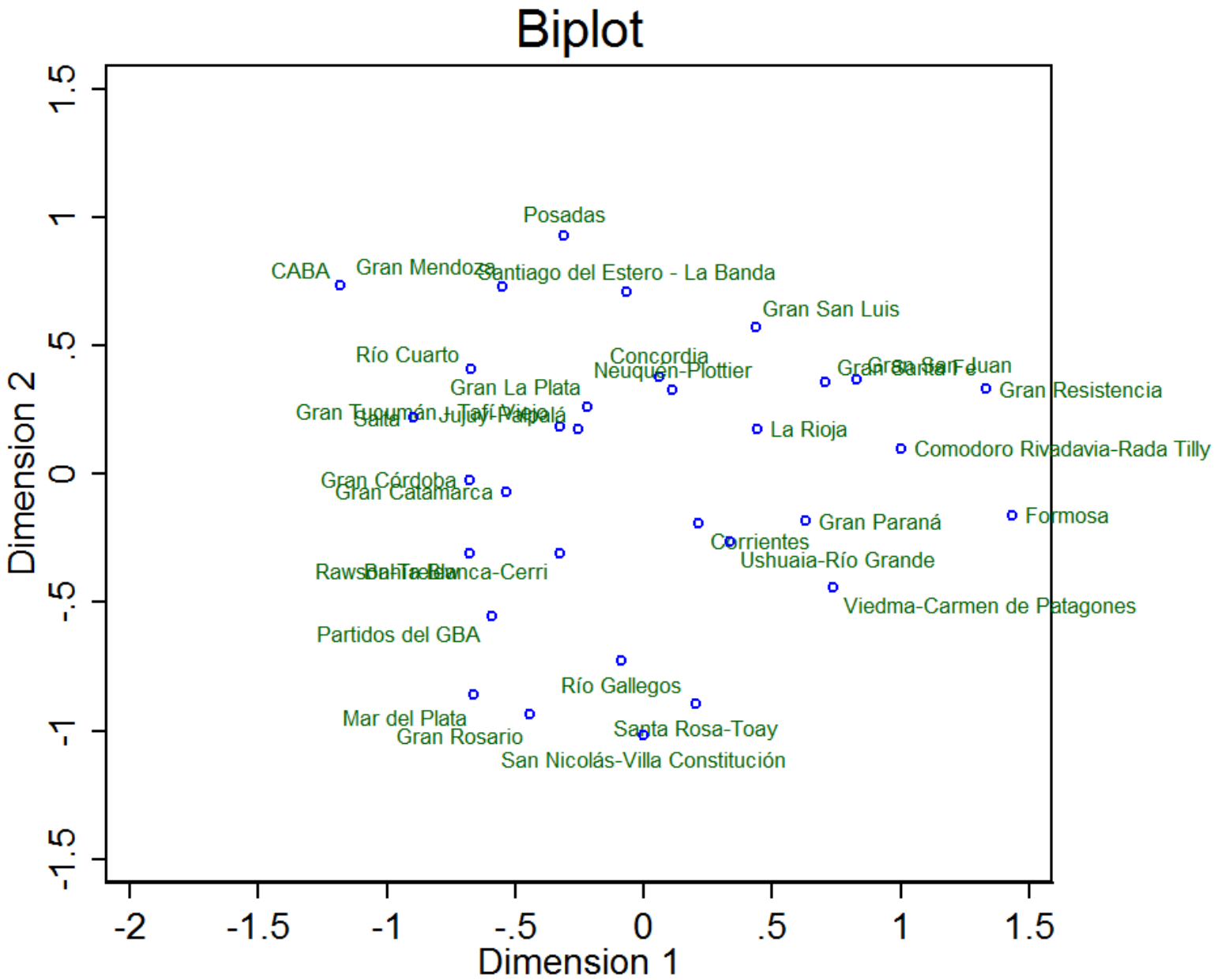
donde U es $n \times p$ y contiene en columnas los autovectores asociados a los autovalores no nulos de $\tilde{X}\tilde{X}'$, $D^{1/2}$ es una matriz diagonal que contiene las raíz cuadrada de los autovalores de $\tilde{X}\tilde{X}'$ o $\tilde{X}'\tilde{X}$ y A' es una matriz ortogonal de orden p que contiene por filas los autovectores de $\tilde{X}'\tilde{X}$.

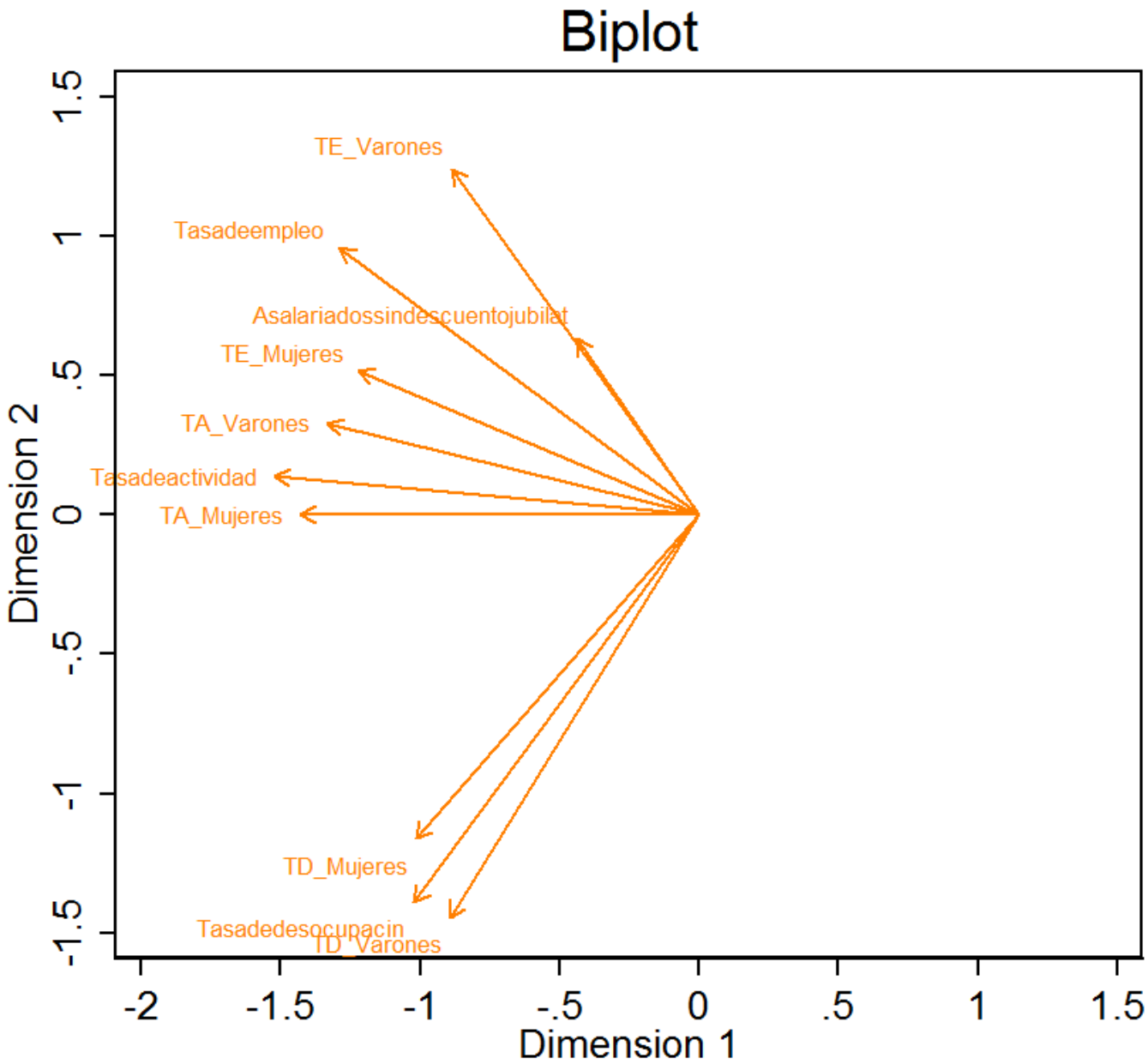
Biplots

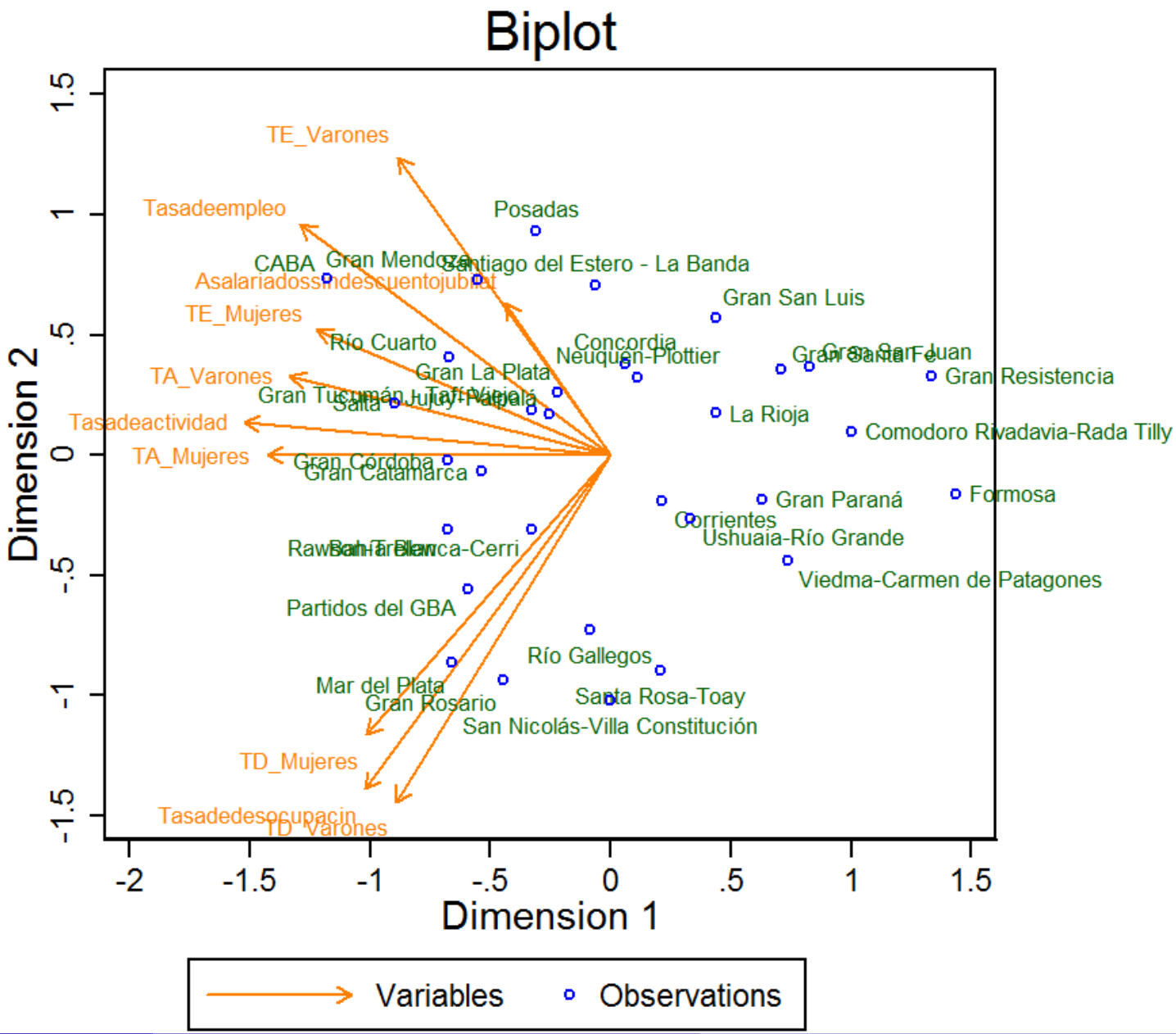
- En base a los autovectores de S y Q, la representación biplot de una matriz $\tilde{X}$ consiste en aproximarla mediante la descomposición en valores singulares de rango 2, tomando $r = 2$.

$$\tilde{X} \approx U_2 D_2^{1/2} A_2' = (U_2 D_2^{1/2-c/2})(D_2^{c/2} A_2') = FC$$

  donde U_2 es $(n \times 2)$, $D_2^{1/2}$ es diagonal de orden 2 y A_2' es $(2 \times p)$.
- Tomando $0 \leq c \leq 1$ se obtienen distintas descomposiciones de la matriz $\tilde{X}$ en dos matrices.
  - la primera, F representa las n filas de la matriz $\tilde{X}$ en un espacio de dos dimensiones, y
  - la segunda matriz, C, representa en el mismo espacio las columnas de la matriz.
- Según el valor de c que se haya elegido se obtienen distintos biplots. Los más utilizados son $c = 0$, 0.5 y 1. Vamos a interpretar el biplot cuando $c = 1$ ya que es el caso más interesante.


Biplot
Posadas
CABA
Gran Mendoza
Santiago del Estero - La Banda
Gran San Luis
Río Cuarto
Concordia
Neuquén-Plottier
Gran Santa Fe
Gran San Juan
Gran La Plata
Gran Resistencia
Gran Tucumán - Tafí Viejo
Salta
Jujuy - Palpalá
La Rioja
Comodoro Rivadavia-Rada Tilly
Gran Córdoba
Gran Catamarca
Gran Paraná
Formosa
Corrientes
Ushuaia-Río Grande
Bahía Blanca-Cerri
Viedma-Carmen de Patagones
Partidos del GBA
Río Gallegos
Mar del Plata
Gran Rosario
Santa Rosa-Toay
San Nicolás-Villa Constitución
Dimension 2
Dimension 1
1.5
1
.5
0
-.5
-1
-1.5
-2
-1.5
-1
-.5
0
.5
1
1.5


Biplot
TE_Varones
Tasadeempleo
Asalariadossindescuentojubilat
TE_Mujeres
TA_Varones
Tasadeactividad
TA_Mujeres
TD_Mujeres
Tasadedesocupacin
TD_Varones
Dimension 2
Dimension 1
1.5
1
.5
0
-.5
-1
-1.5
-2
-1.5
-1
-.5
0
.5
1
1.5


Biplot
Dimension 2
Dimension 1
1.5
1
.5
0
-.5
-1
-1.5
-2
TE_Varones
Tasadeempleo
Posadas
CABA
Gran Mendoza
Santiago del Estero - La Banda
Asalariadossindescuentojubilat
Gran San Luis
TE_Mujeres
Río Cuarto
Concordia
Neuquén-Plottier
Gran San Juan
Gran Santa Fe
Gran Resistencia
TA_Varones
Gran La Plata
Gran Tucumán
Salta
La Rioja
Tasadeactividad
Comodoro Rivadavia-Rada Tilly
TA_Mujeres
Gran Córdoba
Gran Catamarca
Gran Paraná
Formosa
Corrientes
Ushuaia-Río Grande
Viedma-Carmen de Patagones
Partidos del GBA
Río Gallegos
Mar del Plata
Gran Rosario
Santa Rosa-Toay
San Nicolás-Villa Constitución
TD_Mujeres
Tasadedesocupacin
TD_Varones
Variables
Observations

Biplots

- Se representan las observaciones (filas de $\tilde{X}$) por las filas de U_2 y las variables (columnas de $\tilde{X}$) por las columnas de la matriz $D_2^{1/2}A'2$.
- Para distinguir ambas representaciones las observaciones se dibujan como puntos y las variables como vectores en el plano.
- Se verifica lo siguiente:
  1. La representación de las observaciones como puntos en un plano mediante U_2 equivale a proyectar las observaciones sobre el plano de las 2 componentes principales estandarizadas (para que tengan variancia unitaria).
  2. Las distancias euclídeas entre los puntos en el plano equivalen aproximadamente a las distancias de Mahalanobis entre las observaciones originales.
  3. La representación de las variables mediante vectores de 2 coordenadas es tal que el ángulo entre los vectores equivale aproximadamente a la correlación entre las variables.

Biplots: primera propiedad

- Las coordenadas de las componentes principales son $Z = \tilde{X}A$ y los vectores que forman las columnas de Z son los autovectores sin normalizar de $\tilde{X}\tilde{X}'$.
- Los autovectores normalizados seran: $u_i = (\lambda_i)^{-1/2} z_i$ y generalizando esta idea $U = ZD^{-1/2}$
- De esta normalización surge inmediatamente que U'U=I:

$$U'U = D^{-1/2}Z'ZD^{-1/2} = D^{-1/2}DD^{-1/2} = I$$

- Por lo tanto si representamos las observaciones por U_2 tenemos las proyecciones estandarizadas de las observaciones sobre los primeros dos componentes.

Biplots: segunda propiedad

- Cada observación se presenta por los componentes principales por $x_i' A$ y si estandarizamos los componentes para que tengan variancia unitaria obtendremos $x_i' A D^{-1/2}$
- Las distancias euclídeas al cuadrado entre dos observaciones i y j en términos de sus proyecciones en los componentes principales serán:

$$d^2(i,j) = \| x_i' A D^{-1/2} - x_j' A D^{-1/2} \|^2 = (x_i - x_j)' A D^{-1} A (x_i - x_j)$$

Y como $S = ADA'$ entonces $S^{-1} = AD^{-1}A'$ y obtenemos la distancia de Mahalanobis entre las observaciones originales. Si en lugar de tomar todos los componentes principales tomamos solo los dos primeros, esta relación será aproximada.

Biplots: tercera propiedad

- Si representamos las variables como vectores con coordenadas

$$D_2^{1/2} A_2' = C$$

los ángulos entre los vectores representan aproximadamente la correlación entre las variables originales.

$$S \approx A_2 D_2 A_2' = C'C$$

donde c_1 es un vector 2×1 correspondiente a la primera columna de C. De esta expresión es inmediato que $c_i' c_i = s_i^2$, $c_i' c_k = s_{ik}$ y además:

$$r_{ik} = \frac{c_i' c_k}{\| c_i \| \| c_k \|} = cos(\theta_{ik})$$

- Por lo tanto el coseno del ángulo entre dos vectores es aproximadamente el coeficiente de correlación entre las variables.