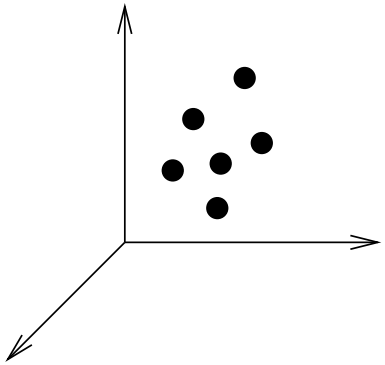
Reconocimiento Estadístico de Patrones, Semana 2

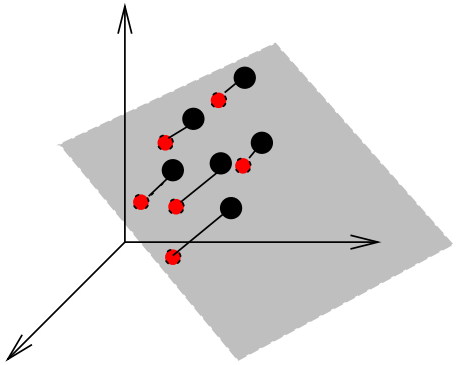
Johan Van Horebeek

horebeek@cimat.mx

Enero - Junio 2011

6.3.1 PCA (continuación)





Vimos que otro punto de partida de PCA es buscar la matriz $d \times p$ B que minimiza:

$$E||X - B(B^T X)||^2$$

Solución: B formada por primeros p vectores propios y $E||X - B(B^T X)||^2 = \sum_{i=p+1}^{d} \lambda_i$.

Aproximamos X con $\hat{X} = B(B^T X)$.

¿Qué bueno es limitarse a aproximaciones lineales?

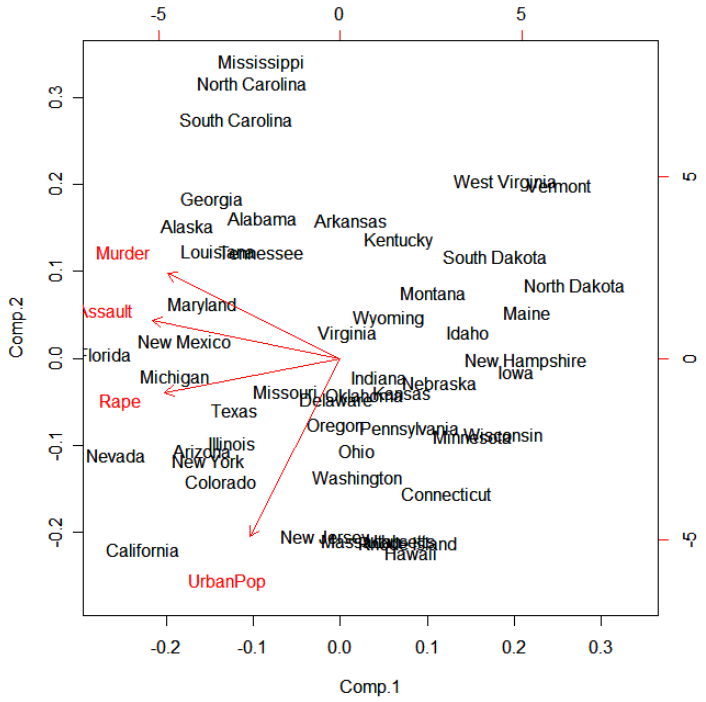
$\Rightarrow$ es una de las razones porque es informativo verificar la normalidad de los datos.

Ejemplo 1

```
data(USArrests)
 p<-princomp(USArrests)
 loadings(p)
 summary(p)
p<-princomp(USArrests,cor=T)
biplot(p)
```

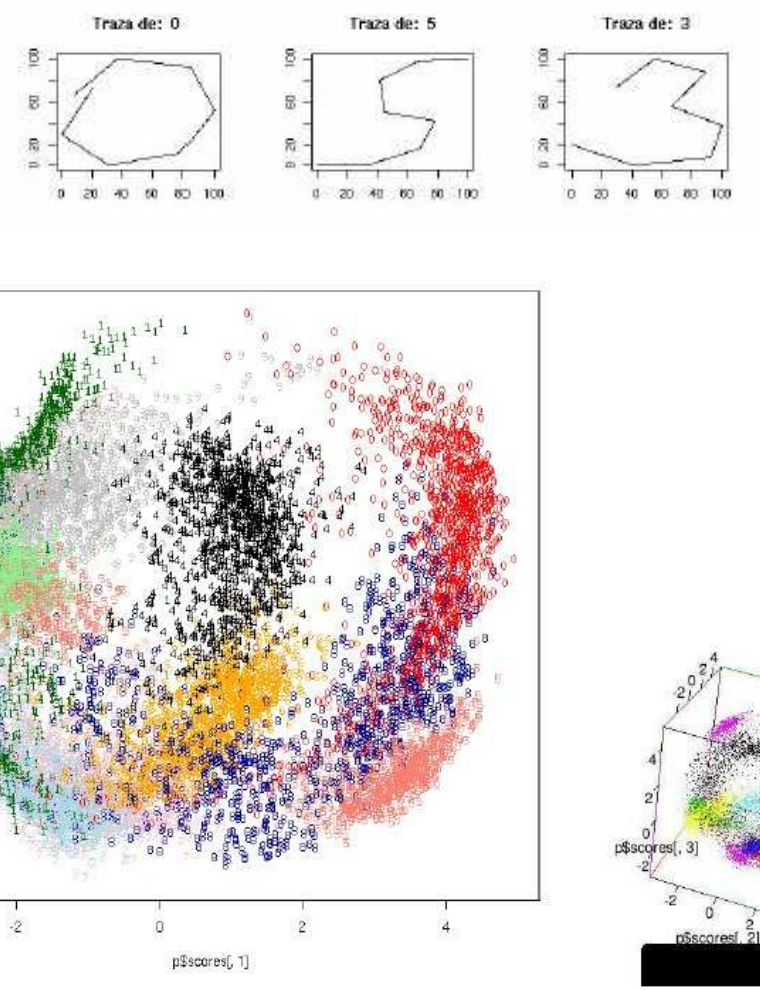
```
> loadings(p)
Loadings:
         Comp.1 Comp.2 Comp.3 Comp.4
Murder                          0.995
Assault  -0.995
UrbanPop         -0.977 -0.201
Rape             -0.201  0.974
```

```
> summary(p)
                          Comp.1       Comp.2       Comp.3       Comp.4
Standard deviation     82.8908472 14.06956001 6.424204055 2.4578367034
Proportion of Variance  0.9655342  0.02781734 0.005799535 0.0008489079
Cumulative Proportion   0.9655342  0.99335156 0.999151092 1.0000000000
```

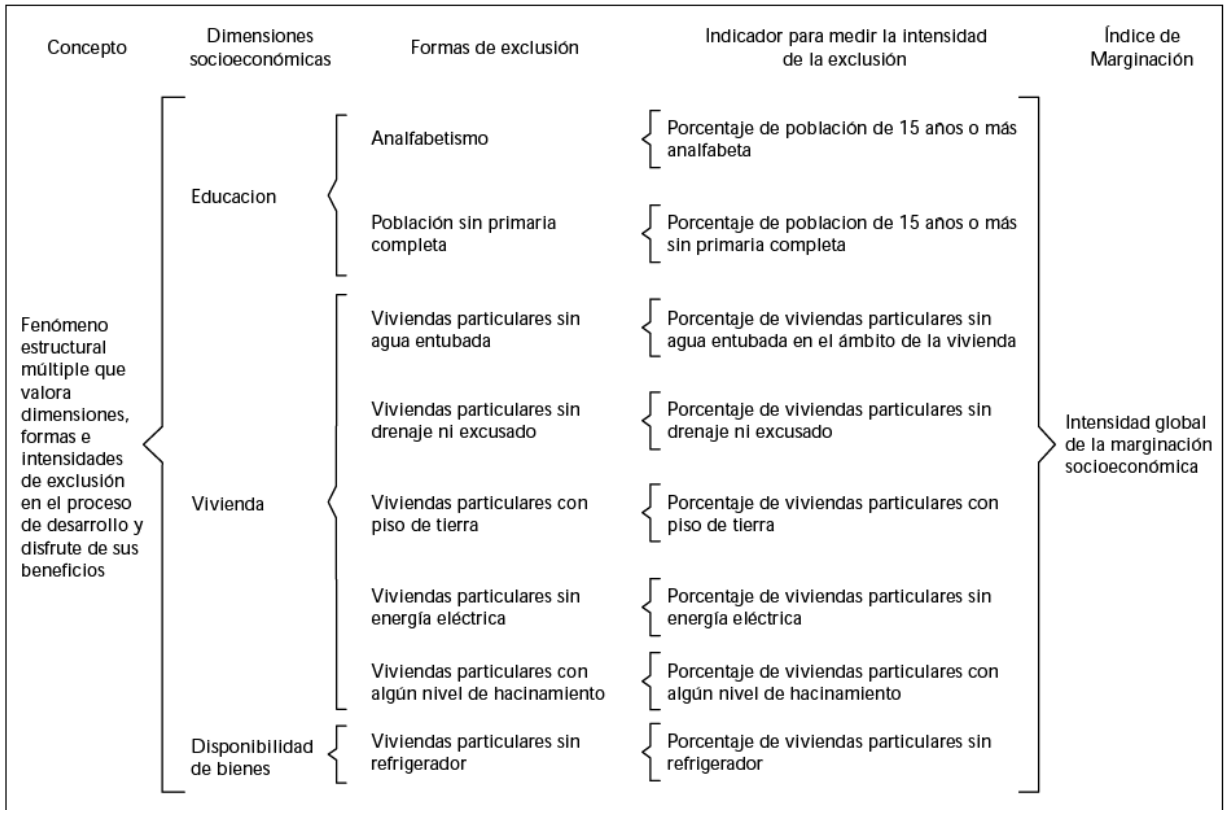


Insertar `pca.r`.

Ejemplo 2


Traza de: 0
Traza de: 5
Traza de: 3
p$scores[, 1]
p$scores[, 2]
p$scores[, 3]
p$scores[, 2]

Ejemplo 3



¿Cómo obtener un (solo) indicador que mide pobreza?

Indicador socioeconómico	Indicador socioeconómico							
	% Población de 15 años o más analfabeta	% Población de 15 años o más sin primaria completa	% Viviendas particulares sin drenaje ni excusado	% Viviendas particulares sin energía eléctrica	% Viviendas particulares sin agua entubada en el ámbito de la vivienda	% Viviendas particulares con algún nivel de hacinamiento	% Viviendas particulares con piso de tierra	% Viviendas particulares sin refrigerador
% Población de 15 años o más analfabeta	1.00000							
% Población de 15 años o más sin primaria completa	0.72060	1.00000						
% Viviendas particulares sin drenaje ni excusado	0.36409	0.37328	1.00000					
% Viviendas particulares sin energía eléctrica	0.36834	0.38022	0.38245	1.00000				
% Viviendas particulares sin agua entubada en el ámbito de la vivienda	0.23712	0.27441	0.26163	0.39872	1.00000			
% Viviendas particulares con algún nivel de hacinamiento	0.36779	0.28639	0.24795	0.22643	0.19166	1.00000		
% Viviendas particulares con piso de tierra	0.54050	0.49528	0.37135	0.49801	0.35572	0.46269	1.00000	
% Viviendas particulares sin refrigerador	0.50881	0.47664	0.34424	0.61223	0.34717	0.49071	0.67219	1.00000

Fuente: Estimaciones del CONAPO con base en el *II Conteo de Población y Vivienda 2005*.

Cuadro C.3. Valores propios de la matriz de correlaciones y porcentaje de varianza explicada a nivel localidad, 2005

Componentes principales	Valores propios		
	Total	% de varianza	% de varianza acumulada
1	3.89109	48.63868	48.63868
2	0.94850	11.85621	60.49489
3	0.85309	10.66357	71.15846
4	0.70642	8.83031	79.98877
5	0.64667	8.08336	88.07213
6	0.40384	5.04806	93.12019
7	0.28171	3.52137	96.64155
8	0.26868	3.35845	100.00000

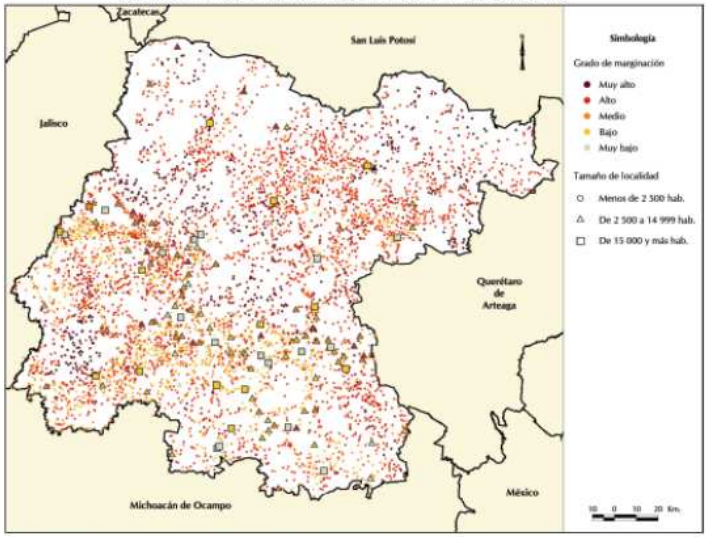
Fuente: Estimaciones del CONAPO con base en el *II Conteo de Población y Vivienda 2005*.

Cuadro C.4. Coeficientes de la primera componente principal por indicador socioeconómico a nivel localidad, 2005

Indicador socioeconómico	Coeficiente de la primera componente principal
% Población de 15 años o más analfabeta	0.19512
% Población de 15 años o más sin primaria completa	0.18963
% Viviendas particulares sin drenaje ni excusado	0.14968
% Viviendas particulares sin energía eléctrica	0.18031
% Viviendas particulares sin agua entubada en el ámbito de la vivienda	0.13331
% Viviendas particulares con algún nivel de hacinamiento	0.14866
% Viviendas particulares con piso de tierra	0.20858
% Viviendas particulares sin refrigerador	0.21135

Fuente: Estimaciones del CONAPO con base en el *II Conteo de Población y Vivienda 2005*.

Mapa 8.11. Guanajuato: Localidades por grado de marginación y tamaño de localidad, 2005



Fuente: Estimaciones del CONAPO con base en el *II Conteo de Población y Vivienda 2005*.

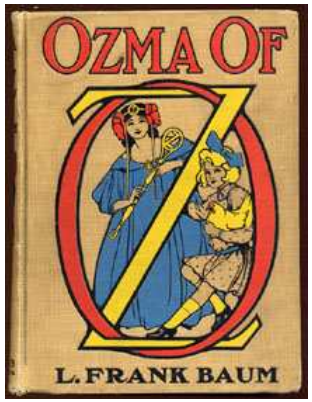
Cuadro B.11.3. Guanajuato: Localidades y población por municipio según grado de marginación a nivel localidad, 2005[1]

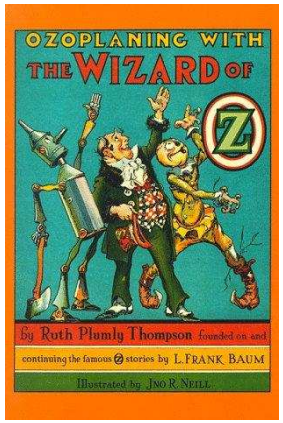
Clave de la entidad	Clave del municipio	Entidad federativa / Municipio	Localidades	Grado de marginación a nivel localidad					Población	Grado de marginación a nivel localidad				
				Muy alto	Alto	Medio	Bajo	Muy bajo		Muy alto	Alto	Medio	Bajo	Muy bajo
11		Guanajuato	5 756	676	3 025	1 093	702	260	4 875 923	48 780	770 626	586 945	858 605	2 610 965
11	001	Abasolo	213	3	101	55	39	15	76 557	104	22 699	18 267	35 017	470
11	002	Acámbaro	131	3	57	40	24	7	101 399	84	14 548	16 988	14 415	55 364
11	003	San Miguel de Allende	320	38	218	29	14	21	138 310	3 218	54 702	11 747	4 125	64 518
11	004	Apaseo el Alto	90	4	63	15	7	1	57 782	129	20 493	8 744	28 380	36
11	005	Apaseo el Grande	69	4	25	20	10	10	73 449	180	10 423	21 936	16 766	24 144
11	006	Atarjea	33	4	28	1	—	---	5 015	222	4 413	380	—	---
11	007	Celaya	136	5	41	33	40	17	415 199	279	18 565	40 402	40 509	315 444
11	008	Manuel Doblado	177	31	79	39	21	7	33 642	794	7 275	9 478	15 850	245
11	009	Comonfort	88	25	42	15	5	1	69 873	6 426	19 227	9 768	34 441	11
11	010	Coroneo	26	---	15	9	2	---	10 970	---	3 765	3 534	3 671	---
11	011	Cortazar	58	2	16	16	14	10	82 901	21	5 067	8 662	8 020	61 131
11	012	Cuerámaro	50	5	34	5	4	2	23 782	217	9 731	13 101	653	80
11	013	Doctor Mora	64	4	51	6	2	1	21 282	528	13 924	1 353	5 461	16
11	014	Dol. Hgo. Cuna de la Indepen. Nal.	396	58	268	47	18	5	133 698	4 066	59 973	11 635	57 895	129
11	015	Guanajuato	165	35	71	33	17	9	153 039	1 834	11 538	23 196	7 786	108 685
11	016	Huanímaro	31	---	3	16	11	1	18 418	---	545	9 194	8 663	16
11	017	Irapuato	243	9	90	68	50	26	461 924	541	28 344	48 859	30 528	353 652
11	018	Jaral del Progreso	21	---	2	8	10	1	31 678	---	147	5 488	26 022	21
11	019	Jerécuaro	144	13	115	11	2	3	46 032	939	30 827	7 395	57	6 814
11	020	León	336	47	136	62	67	24	1 276 554	2 270	44 297	53 119	37 112	1 139 756
11	021	Moroleón	18	---	8	6	3	1	46 696	---	1 934	2 086	767	41 909
11	022	Ocampo	56	7	38	9	2	---	20 143	246	10 883	8 892	122	---
11	023	Pénjamo	372	60	147	93	63	9	137 079	4 459	28 447	33 535	68 294	2 344
11	024	Pueblo Nuevo	37	---	8	13	8	8	9 671	---	2 039	3 334	841	3 457
11	025	Purísima del Rincón	69	2	23	20	20	4	55 763	36	3 089	10 334	40 548	1 756
11	026	Romita	110	6	73	23	13	3	49 709	185	19 690	9 235	20 612	59
11	027	Salamanca	208	14	73	52	51	18	232 874	885	19 484	21 681	43 695	147 129
11	028	Salvatierra	61	3	26	20	10	2	92 375	95	17 219	22 545	16 159	36 357
11	029	San Diego de la Unión	144	9	121	11	3	---	34 121	238	26 782	538	6 563	---
11	030	San Felipe	297	77	189	29	2	---	94 833	4 406	51 017	14 264	25 146	---
11	031	San Francisco del Rincón	136	9	73	23	24	9	102 916	354	12 994	6 482	12 643	70 443
11	032	San José Iturbide	143	7	78	40	14	4	58 855	474	17 463	11 948	8 535	20 385
11	033	San Luis de la Paz	301	73	192	25	8	3	100 494	7 785	36 807	8 116	47 661	125
11	034	Santa Catarina	33	8	23	2	—	---	4 520	932	2 126	1 462	—	---
11	035	Santa Cruz de Juventino Rosas	85	17	35	16	15	2	69 917	1 198	19 511	5 207	43 689	312
11	036	Santiago Maravatío	12	---	5	3	3	1	6 382	---	1 033	867	4 459	23
11	037	Silao	227	11	123	49	28	16	146 197	674	32 156	35 240	10 238	67 889

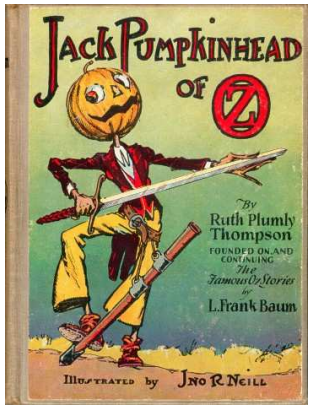
Continúa

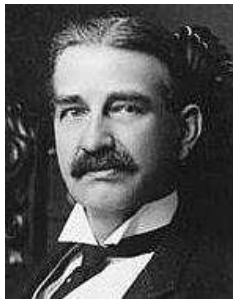
Ejemplo 4: Estilometría de textos

Libros del mago de Oz (X): algunos de Thompson, otros de Baum.





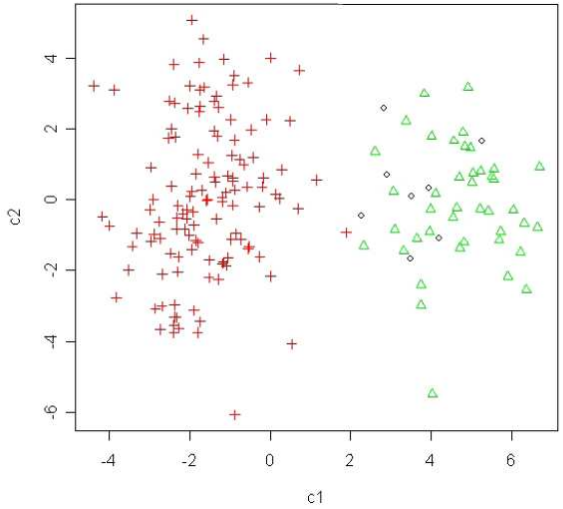





Buscamos las 50 palabras más usadas.

Dividimos cada libro en bloques de 10000 palabras.

Definimos $(X_1, \cdots, X_{50})$ con X_i la frecuencia relativa de ocurrencia de la i-esima palabra más usada en un bloque.



El caso cuando $n < d$

Supongamos que los datos están centrados.

Si $\mathbb{X}$ es la matriz de datos $(n \times d)$, un estimador popular para $Cov(X)$ es

$$\widehat{Cov(X)} \sim \mathbb{X}^t\mathbb{X}$$

Propiedad

Las matrices $\mathbb{X}^t\mathbb{X}$ y $\mathbb{X}\mathbb{X}^t$ tienen los mismos valores propios y:

$$u_j = \frac{\mathbb{X}^t v_j}{\sqrt{\lambda_j}} := \sum_i \alpha_i^j x_i,$$

con $\{u_j\}$ vectores propios de $\mathbb{X}^t\mathbb{X}$; $\{v_j\}$ vectores propios de $\mathbb{X}\mathbb{X}^t$.

Consequencia: vectores propios viven en espacio generado por los datos.

Si $n < d$, conviene calcular las proyecciones a través de los vectores propios de $\mathbb{X}\mathbb{X}^t$:

$$\langle u_j, x\rangle = \sum_i \alpha_i^j \langle x_i, x\rangle$$

De lo anterior:

$$\langle u_j, x\rangle = \sum_i \alpha_i^j \langle x_i, x\rangle \quad \text{con } \alpha\text{'s determinadas por } \mathbb{X}\mathbb{X}^t.$$

Es suficiente conocer $\langle \cdot, \cdot \rangle$: es punto de partida de KernelPCA

En general, si usamos transformación $\Phi(X)$, definimos:

$$K(x, y) := \langle \Phi(x), \Phi(y)\rangle.$$

Ejemplo: Si $d = 2$, $x = (x_1, x_2)$, definimos:

$$\Phi(x) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, \sqrt{2}x_1x_2, x_2^2),$$

Entonces:

$$K(x, y) = (1+ < x, y >)^2$$

Eso es mucho más rápido de calcular que $< \Phi(x), \Phi(y) >$!

$\Rightarrow$ trabajar al revés: definimos primero $K(,)$. Baja ciertas condiciones existe una $\Phi()$.

Caso popular: $K(x, y) = (1+ < x, y >)^k$ corresponde a cierta transformación polinomial.

6.3.2 Proyection pursuit

En lugar de la varianza, usamos otras medidas de interés. un ejemplo es buscar direcciones que maximizan la NO gausianidad.

Recuerda la antropía como medida de variabilidad:

$$Entropy(X) = E(log(\frac{1}{P(X=x)})) = -E(log(P(X=x)))$$

Si ocurre x, la sorpresa es inversamente proporcional a $P(X=x)$;

entre más sorpresas, mayor variabilidad.

Propiedad

la distribución de máxima entropía sobre intervalo A es la distribución uniforme sobre A;

la distribución de máxima entropía sobre R con varianza σ^2 dada, es la distr. normal con var. σ^2.

Lo anterior usamos para buscar direcciones que maximizan la no gausianidad:

• Usando la negentropía: Negentropia(*distr*) = entropía(*distr*) - entropía(*normal*)

con entropía(*normal*) la entropía de una normal con la misma varianza que *distr*.

• Usando la kurtosis, etc.

Problema: Ya no se tienen soluciones explícitas; se recurre a métodos iterativos (ej. método del gradiente)

INSERTAR DEMO CON GGOBI

LEER:

Grand Tours, Projection Pursuit Guided Tours and Manual Controls, Dianne Cook, Andreas Buja, Eun-Kyung Lee, and Hadley Wickham

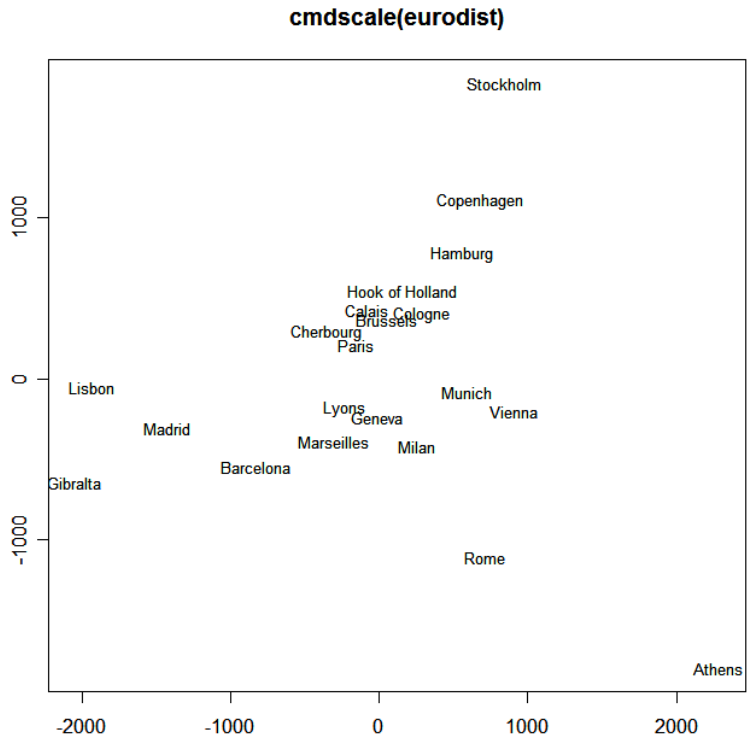
http://www.public.iastate.edu/~dicook/research/chapter.pdf

6.4 Multidimensional scaling

IDEA: Asociamos con cada x_i un $x_i^* \in \mathcal{R}^2$ minimizando:

$$\sum_i \sum_j (d(x_i, x_j)^2 - d(x_i^*, x_j^*)^2)^2.$$

El problema es parametrizado en términos de los datos. Solamente se requiere proporcionar $d(x_i, x_j)$.



INSERTAR RELACION CLASSICAL DIMENSIONAL SCALING Y PCA