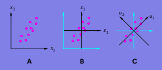
Inferences for Correspondence Analysis

Hervé Abdi

www.utdallas.edu~herve herve@utdallas.edu



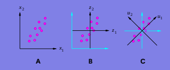
1. Refresher! and Menu

- Today is inference day
- Omnibus: χ^2
- Omnibus: Permutation
- How to do correctly permute? ...
- How many Components: χ^2 again (Malinvaud-Saporta)
- How many Components: permutation again ...
- How to do correctly permute? ...
- What the important rows/columns? test-values
- What the important rows/columns? Bootstrap (ratios)
- How to do correctly bootstrap? ...

Home Page
Title Page
Page 2 of 61
Go Back
Full Screen
Close
Quit

Table 1: Hair and Eye Colors (Inspired by Fisher. R dataset. $N = 592.$)
The Data: data(HairEyeColor).

	Eyes Brown	Eyes Blue	Eyes Hazel	Eyes Green
Hair Black	68	20	15	5
Hair Brown	119	84	54	29
Hair Red	26	17	14	14
Hair Blond	7	94	10	16



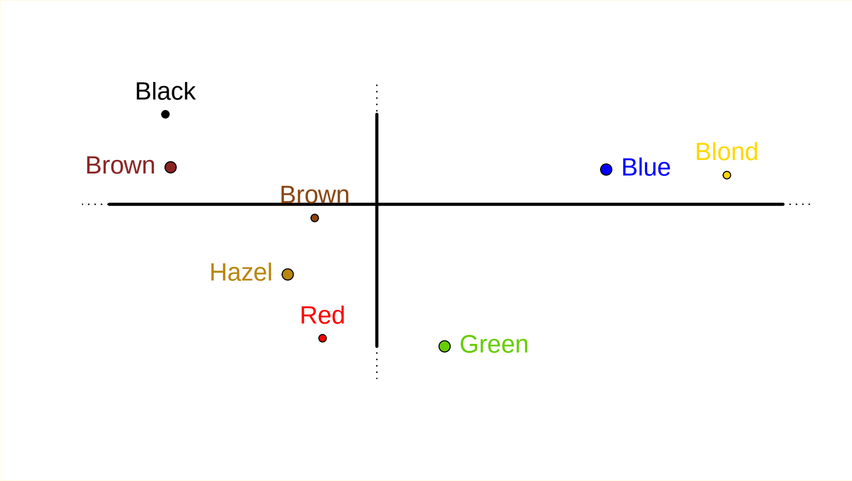
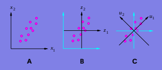


Figure 1: Hair and Eye Colors. Plain Correspondence Analysis. $\lambda_1 = .2088$, $\tau_1 = .89$, $\lambda_2 = .0222$, $\tau_2 = .10$,

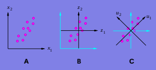


Omnibus

Classic Approach: χ^2

So here $\chi^2(9) = 138.29 \; p = 2 \times 10^{-16}$

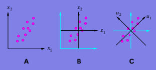
NB: we can get χ^2: as Inertia $\times N$



Refresher! and Menu
Home Page
Title Page
Page 5 of 61
Go Back
Full Screen
Close
Quit

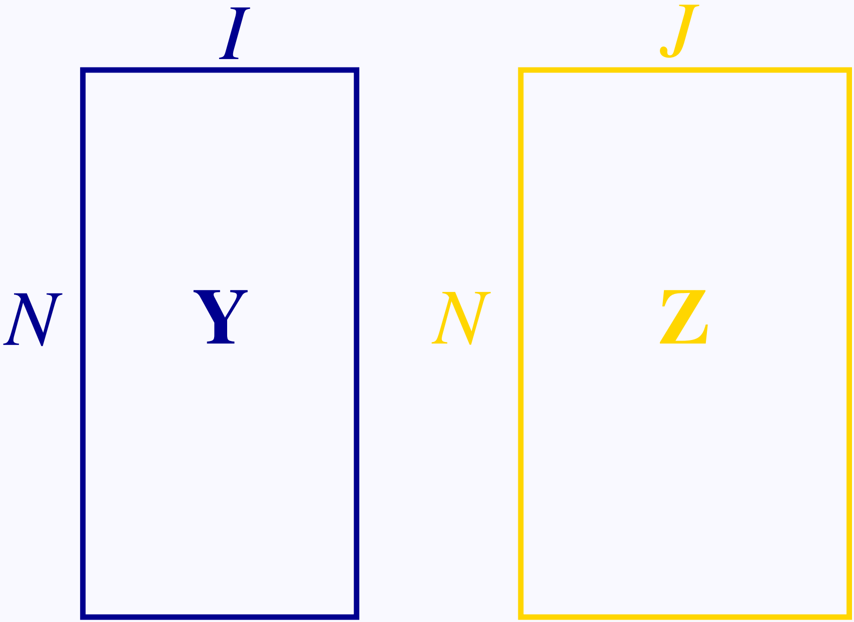
Modern Approach (kind of) permutation test

How to correctly permute?



The key: How to create the $I \times J$ matrix X?

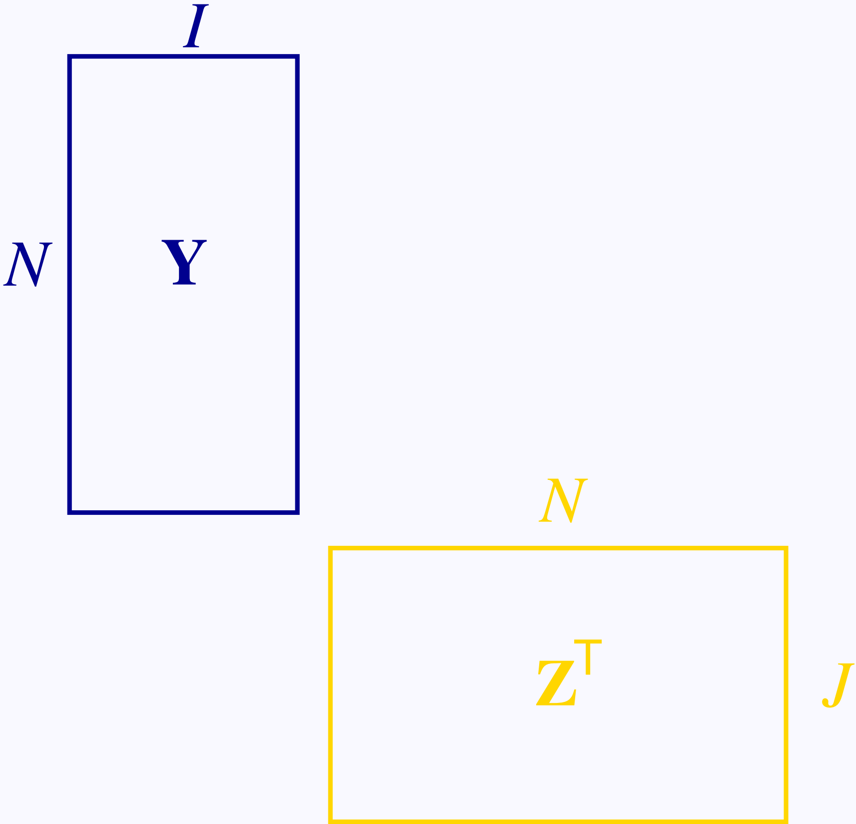
First start with: $N \times I$ matrix **Y** and $N \times J$ matrix **Z**

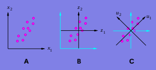


N is # of observations (here 592)

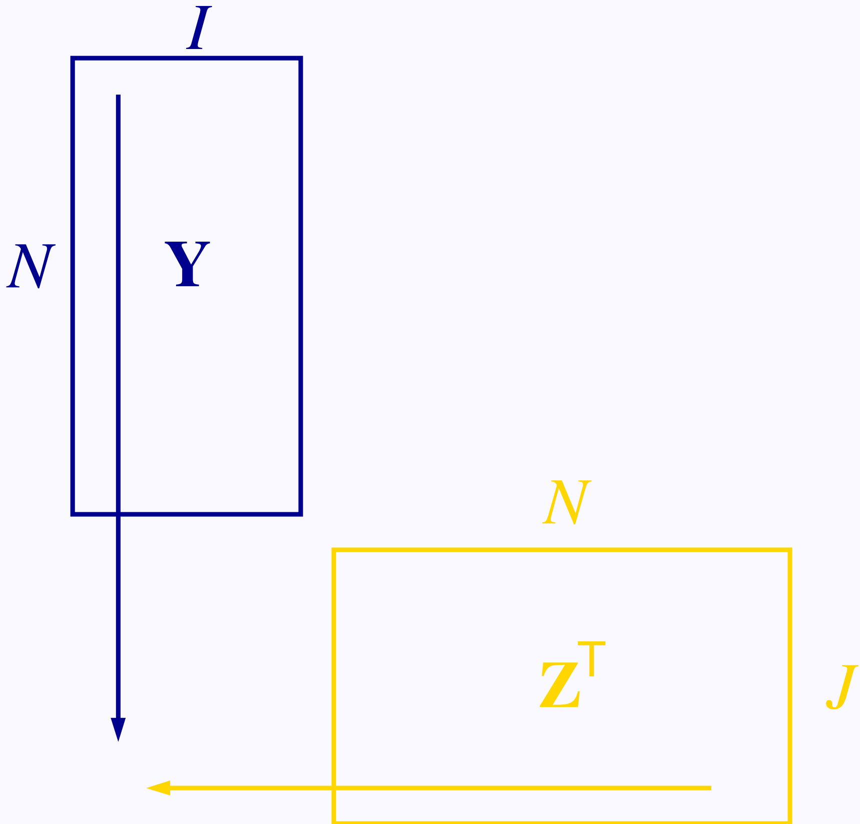
Y and **Z** are $0/1$ indicator / group coding matrices
(a.k.a., complete disjunctive coding)

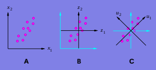
$N \times I$ matrix $\mathbf{Y}$ and $J \times N$ matrix $\mathbf{Z}^T$



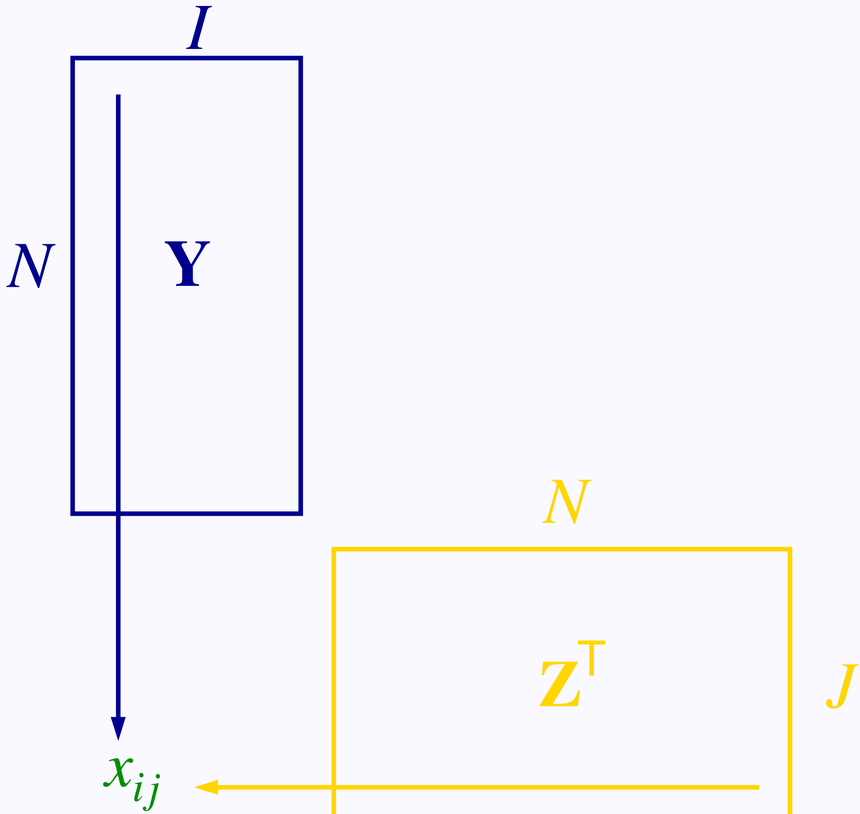


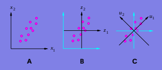
Multiply $N \times I$ matrix $\mathbf{Y}$ and $J \times N$ matrix $\mathbf{Z}^T$



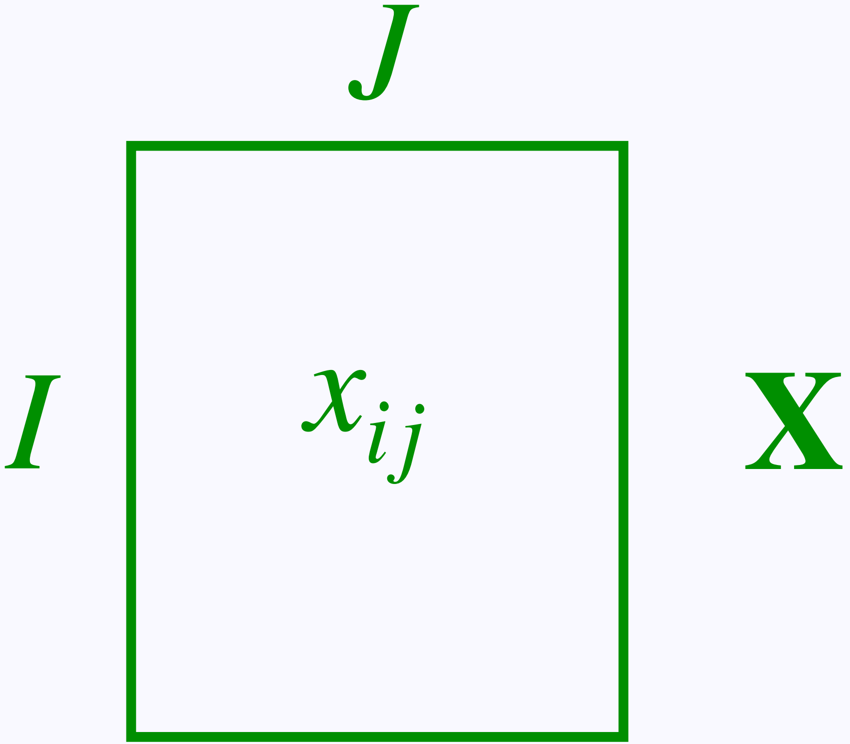


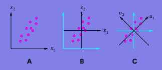
Multiply $N \times I$ matrix $\mathbf{Y}$ and $J \times N$ matrix $\mathbf{Z}^T$





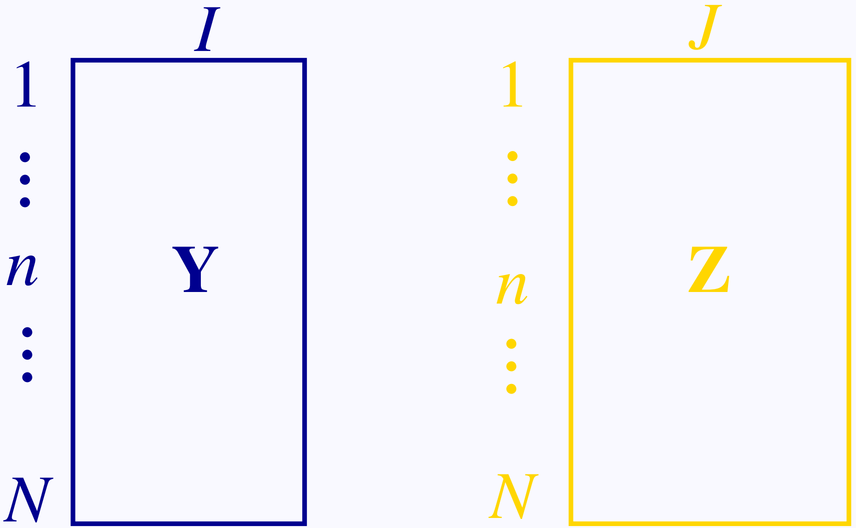
We got $I \times J$ matrix $\mathbf{X} = \mathbf{Y}^T \times \mathbf{Z}$



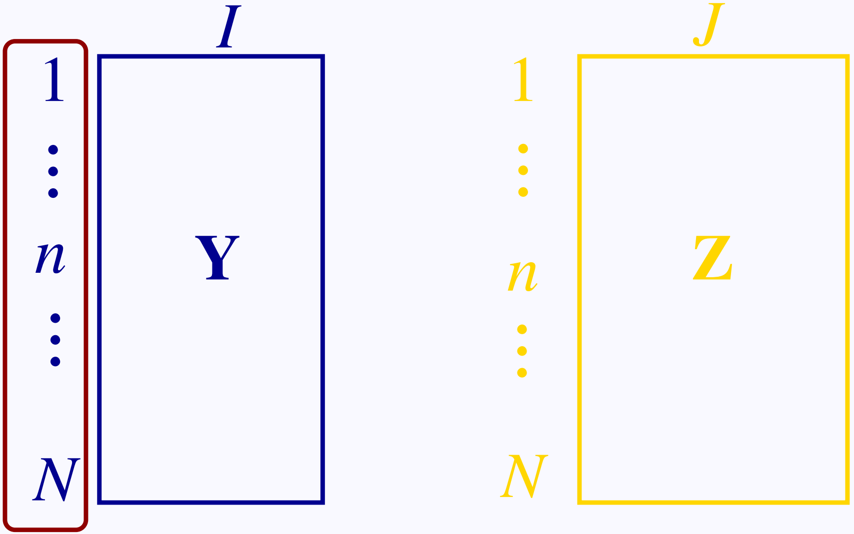


Refresher! and Menu
Home Page
Title Page
Page 12 of 61
Go Back
Full Screen
Close
Quit

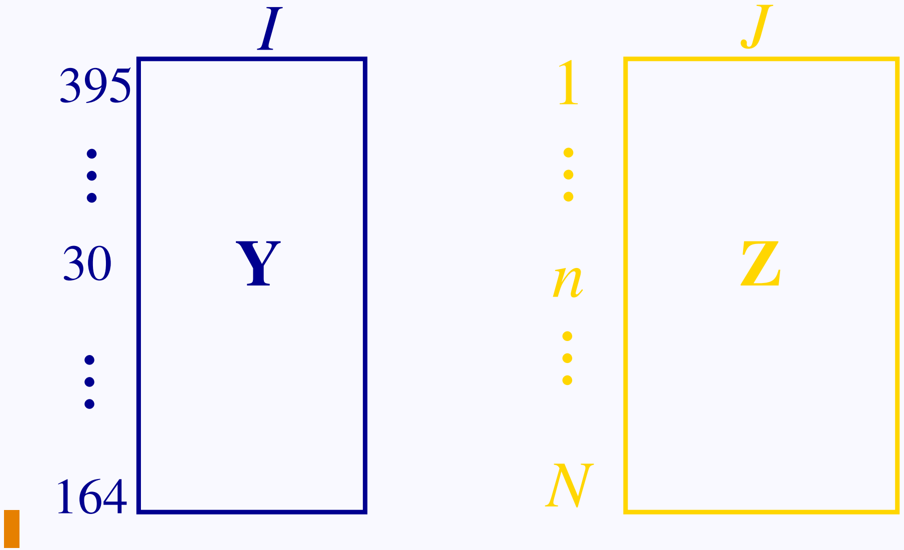
So how to permute?



So how to permute? Look at **Y**



So how to permute? Permute the rows of $\mathbf{Y}$



Refresher! and Menu
Home Page
Title Page
Page 15 of 61
Go Back
Full Screen
Close
Quit

Modern Approach (kind of) permutation test

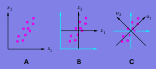
How to correctly permute?

The key is to see the $I \times J$ data matrix $\mathbf{X}$ as the product of two dummy-matrices:

the $N \times I$ matrix $\mathbf{Y}$ and the $N \times J$ matrix $\mathbf{Z}$.

So: $\mathbf{X} = \mathbf{Y}^T\mathbf{Z}$

We permute the rows of $\mathbf{Y}$ and $\mathbf{Z}$ (well: one of them is enough) to create $\mathbf{X}_{\text{perm}} = \mathbf{Y}^T_{\text{perm}}\mathbf{Z}$ and repeat ... and repeat ...



Refresher! and Menu
Home Page
Title Page
Page 16 of 61
Go Back
Full Screen
Close
Quit

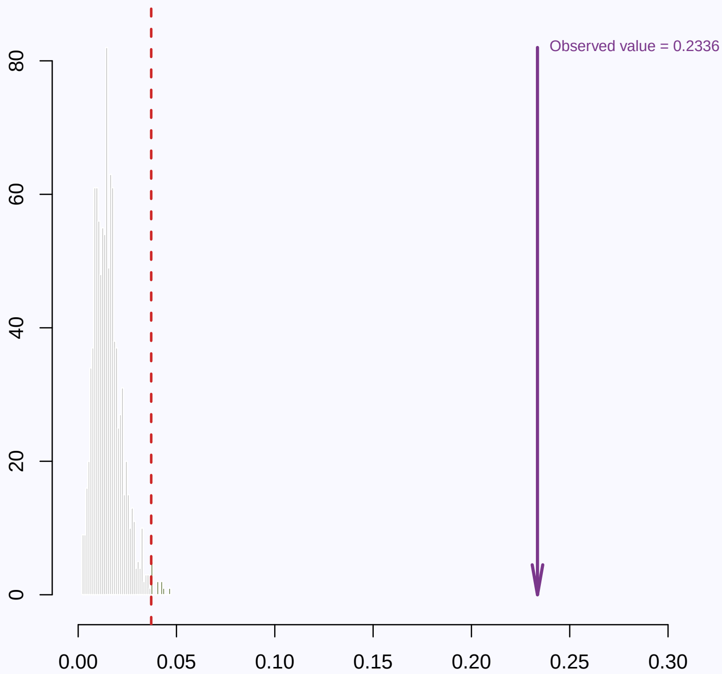


Figure 2: Hair and Eye Colors. Results of Permutation Test. With $1,000$ Iterations: $p < .001$.

The Good Components

Classic Approach: χ^2 Malinvaud-Saporta

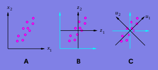
New χ^2 (from Malinvaud, 1987, see also Saporta, 2011, p. 209)
To test Component ℓ of the data matrix $\mathbf{X}$.
First use the reconstitution formula to reconstruct $\mathbf{X}$ from the factor scores from the first ℓ-th factors

$$x_{i,j}^{[\ell]} = \frac{x_{i+}x_{+j}}{N}\left(1 + \sum_{k=1}^{\ell} \delta_k^{-1} f_{i,k} g_{j.k}\right)$$

Then compute the Q_ℓ statistics:

$$Q'_\ell = \sum_i \sum_j \frac{\left(x_{i,j}^{[\ell]} - x_{i,j}\right)^2}{x_{i+} \times x_{+j} \times N^{-1}}$$

Frightening! but ...



but equal to (note that: k stars at $\ell + 1$):

$$Q'_\ell = N \times \left(\sum_{k=\ell+1}^{L} \lambda_k \right)$$

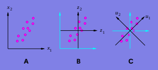
So: here Q'_0 is equal to the *omnibus* test with $Q'_0 = 138.29$ (just like χ^2), distributed as a χ^2 with $\nu = (I - \ell - 1) \times (J - \ell - 1) = 9$,

evaluated with $p_{\text{omnibus}} = 2 \times 10^{-16}$.

Then $Q'_1 = 14.70$ with $\nu = (I - \ell - 1) \times (J - \ell - 1) = 4$ and so: $p_1 = .0054$.

Finally $Q'_2 = 1.54$ with $\nu = (I - \ell - 1) \times (J - \ell - 1) = 1$ and so: $p_2 = .2149,\ ns$.

NB. There is no Q'_3!. Why?



(kind of, see Gosset!) Modern Approach: Permutation test

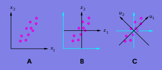
How to correctly permute?

We find $p_0 = .001$, $p_1 = .001$, $p_2 = .015$.

Matches Malinvaud for 0 and 1 but more powerful for 2!

from R:

```
                   Ho: Omnibus   Dim-1  Dim-2  Dim-3
Inertia / Lambda       0.2336  0.2088 0.0222 0.0026
Chi2                 138.2898 14.6964 1.5383     NA
p-Chi2                 0.0000  0.0054 0.2149     NA
df                     9.0000  4.0000 1.0000 0.0000
p-perm                 0.0010  0.0010 0.0150     NA
```

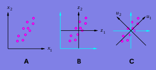


Refresher! and Menu
Home Page
Title Page
Page 20 of 61
Go Back
Full Screen
Close
Quit

Important Rows and Columns

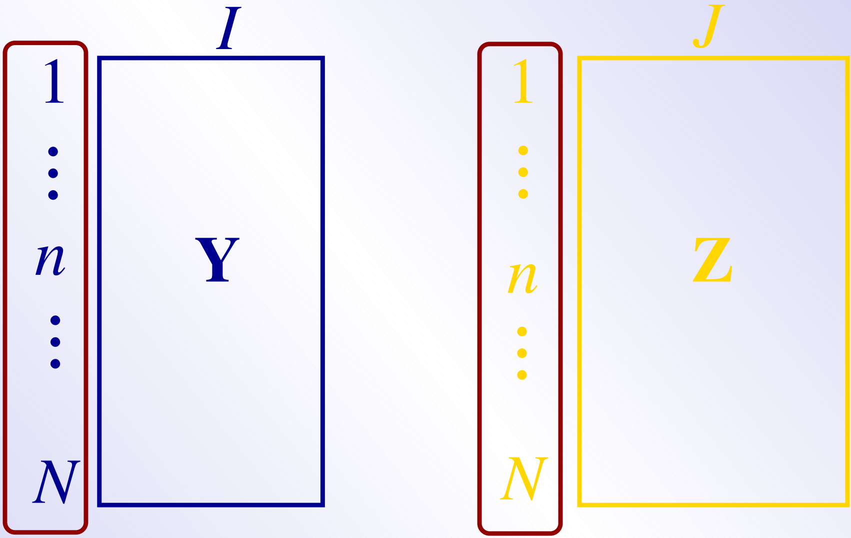
The return of the Bootstrap! and the Bootstrap ratios!

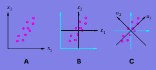
Remember: We bootstrap the rows of $\mathbf{Y}$ and $\mathbf{Z}$



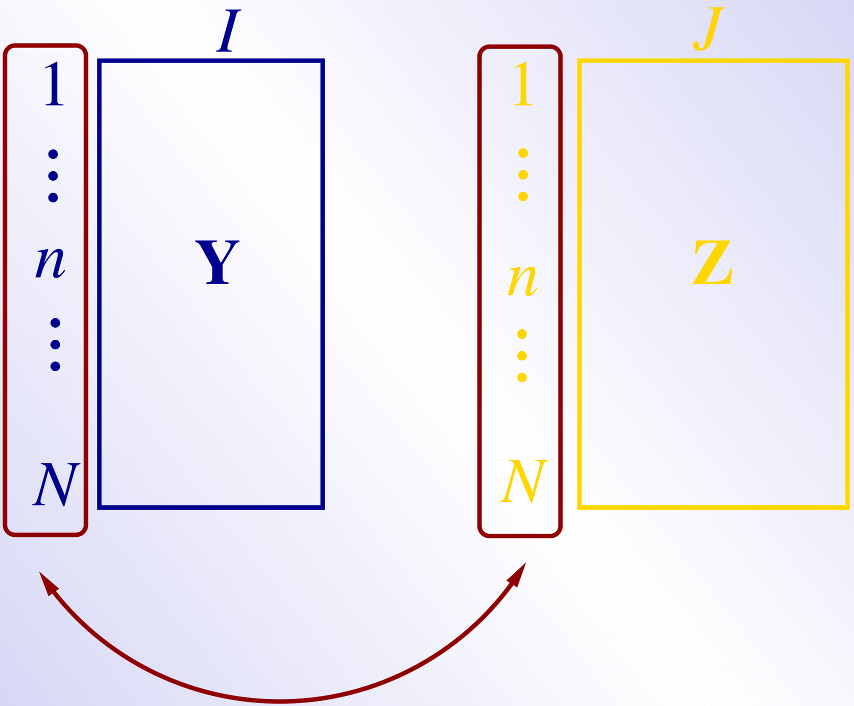
Refresher! and Menu
Home Page
Title Page
Page 21 of 61
Go Back
Full Screen
Close
Quit

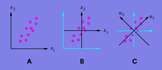
So, how to correctly Bootstrap? Look at **Y** and **Z**





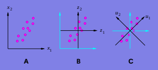
So, how to correctly Bootstrap? Look at Y and Z





So, how to correctly Bootstrap? Look at **Y** and **Z**

I		J	
1	**Y**	1	**Z**
1		1	
5		5	
⋮		⋮	
591		591	
591		591	
592		592	



Create Bootstrapped $\mathrm{X}_{\mathbf{Boot}}$

We bootstrap the rows of both $\mathbf{Y}$ and $\mathbf{Z}$ to create

$$\mathbf{X}_{\text{Boot}} = \mathbf{Y}^{T}_{\text{Boot}} \mathbf{Z}_{\mathbf{Boot}}$$

and repeat . . . and repeat . . .

Project $\mathbf{X}_{\text{Boot}}$ on the CA space

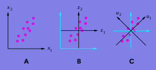
Get factor scores for I and J sets

Compute means (M) and standard deviations (S) for I and J sets

Get Bootstrap ratios as

$$t = \frac{M}{S}$$

Distributed like a Student t



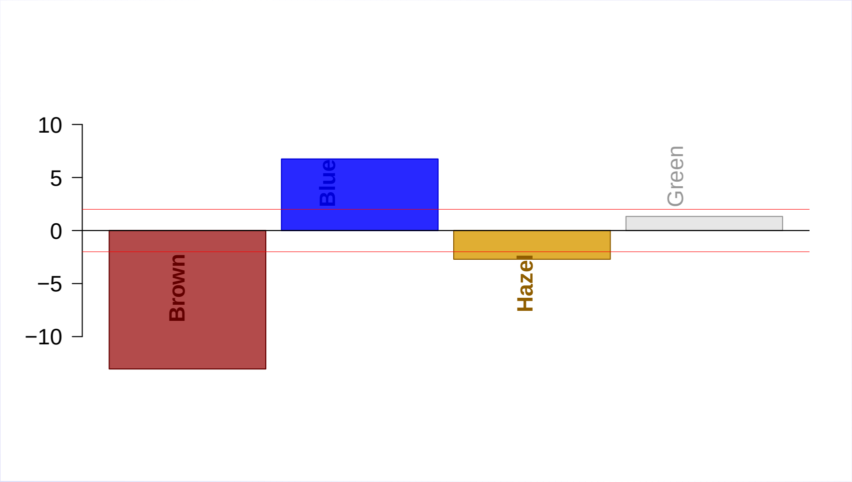


Figure 3: Bootstrap Ratios for the Eye Color. Dimension 1.

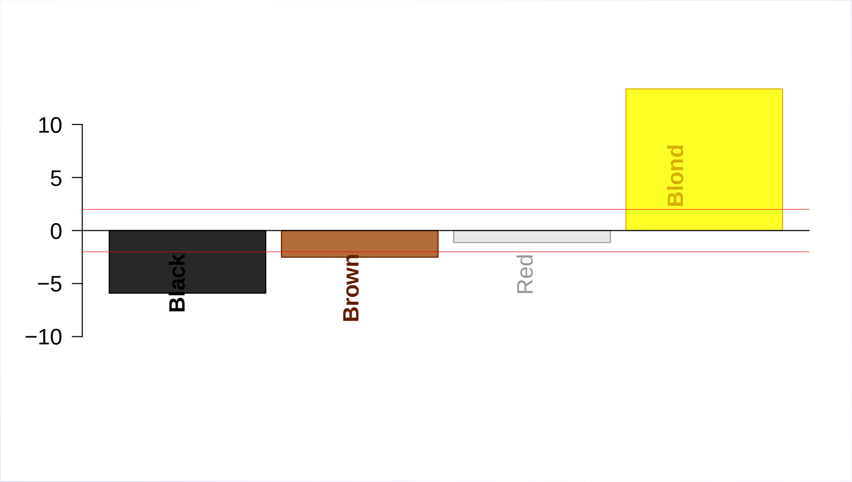
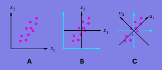


Figure 4: Bootstrap Ratios for the Hair Color. Dimension 1.



Refresher! and Menu

Home Page

Title Page

Go Back

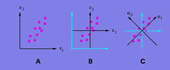
Full Screen

Close

Quit

Problems with large N

- Punctuation Example: $N = 1,424,951$. Very big!
- Causes of Death in the USA $N = 2,349,674$ (Even Bigger)!



Refresher! and Menu

Home Page

Title Page

Page 28 of 61

Go Back

Full Screen

Close

Quit

The causes of Death $N = 2,349,674$

Table 2: The Causes of Death in the USA (2001).

	−1	1–4	5–14	15–24	25–34	35–44	45–54	55–64	65–74	75–84	85+
Septicemia	312	108	69	133	281	829	1977	3111	5998	10351	9069
Hepatitis	1		3	20	74	797	2060	1014	836	613	166
hiv	10	13	38	225	2101	5867	4120	1317	384	88	10
Cancer	66	420	1008	1704	3994	16569	49562	90223	147018	165445	77751
tumor	51	58	105	102	151	366	719	1253	2574	4654	3643
Diabetes	2	6	39	151	595	1958	5343	9570	16731	22805	14171
Parkinson–Alzheimer	0	0	0	0	4	18	118	830	5569	26667	37190
Heart	600	285	365	1280	4038	16686	44450	76224	148409	291995	337930
Pneumonia	299	112	92	181	339	983	1801	2704	6650	18677	30191
Lower–Respiratory	41	43	104	171	290	969	3324	11166	30751	47762	28391
Liver	3		2	26	387	3336	7259	5750	5486	3801	979
Kidney	133	16	18	92	225	768	1822	3284	7356	13097	12668
Unknown	3491	272	181	949	1530	3126	3504	2292	2774	5003	9089
Other Diseases	1131	646	860	1799	3072	8084	14203	16048	25803	49711	65032
Accident–Car	151	678	1823	11139	7454	7827	6251	3861	3282	3507	1272
Accident–NonCar	144	626	1686	10725	6937	7083	5564	3465	2990	3312	1233
Other Accidents	825	1036	1013	3272	4385	8118	7093	3797	4553	9181	10899
Suicide	0	0	279	3971	5070	6635	5942	3317	2432	2192	769
Homicide	332	415	326	5297	5204	4268	2467	1018	532	312	105

First have a look at the data (log transformed):

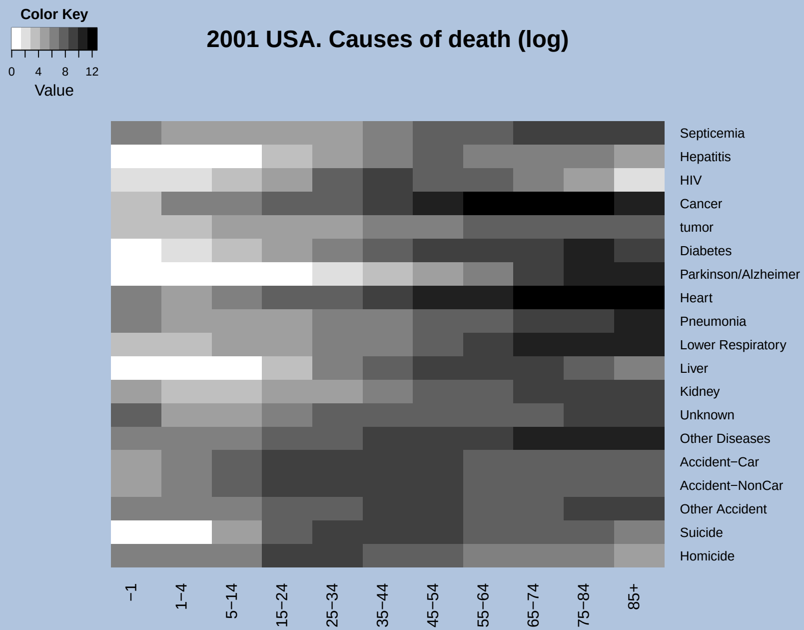


Figure 5: Causes of Death in the US: A picture of the data (log transformed).

Correspondence analysis: The Scree.

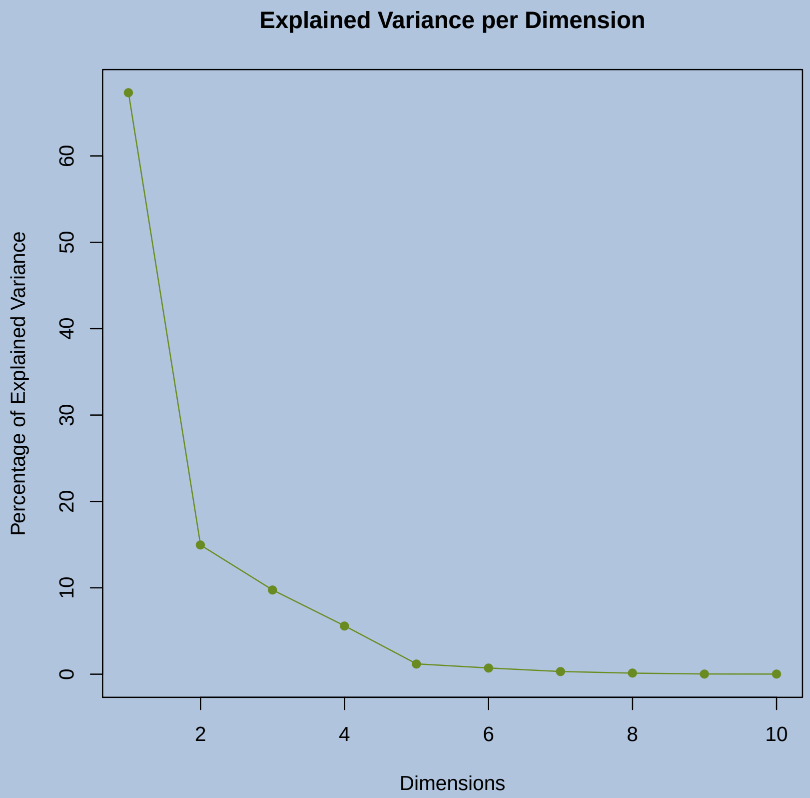
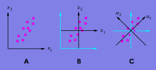


Figure 6: Causes of Death in the US: The Scree (How many dimensions?).



A (symmetric) plot

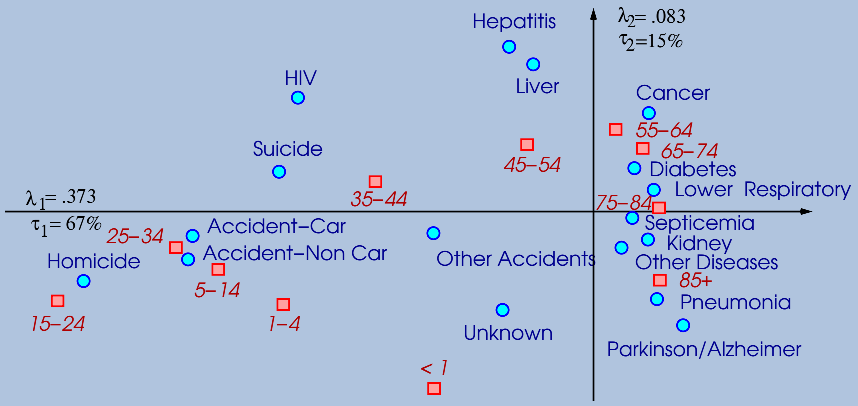
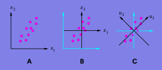


Figure 7: Causes of Death in the US: Correspondence Analysis. Symmetric plot



A (symmetric) plot

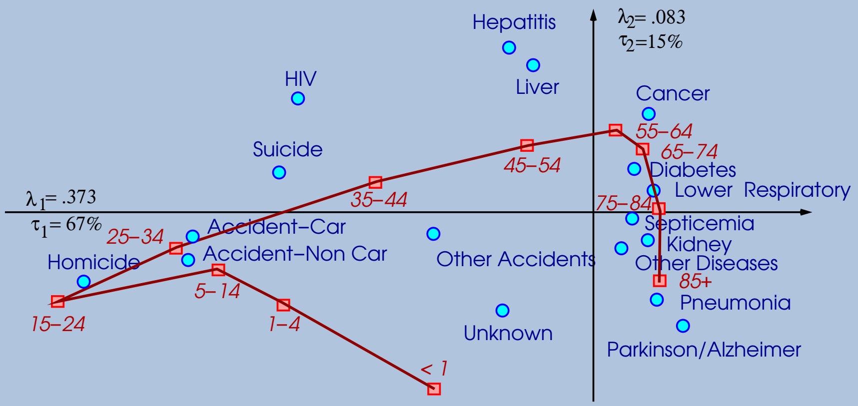


Figure 8: Causes of Death in the US: Correspondence Analysis. Symmetric plot with frills. There are three periods in life (or death!).

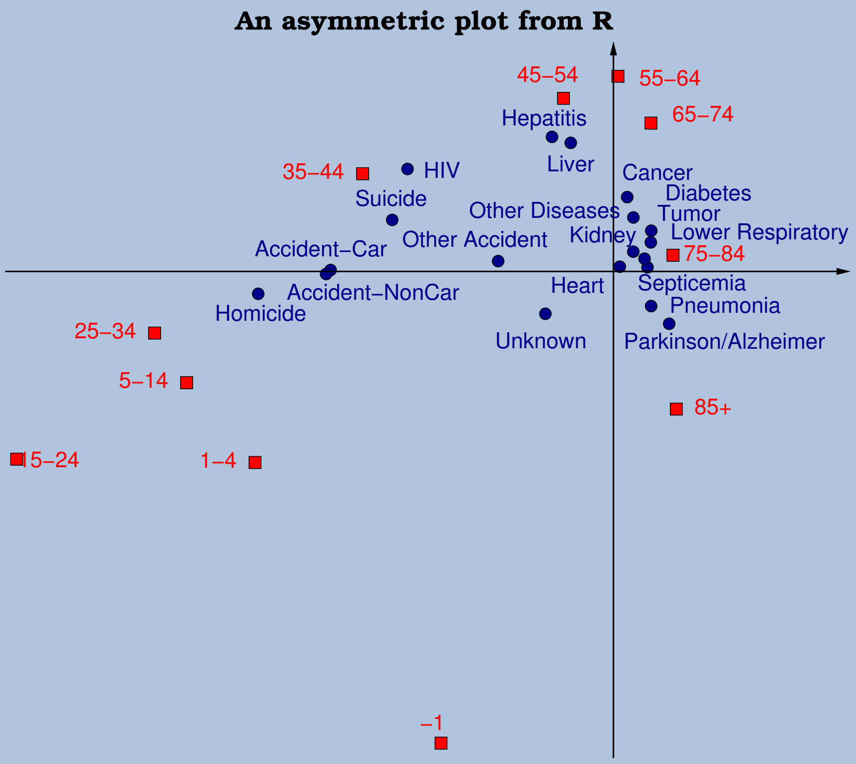


Figure 9: Causes of Death in the US: Correspondence Analysis. Asymmetric plot (From R). There are three periods in life (or death!).

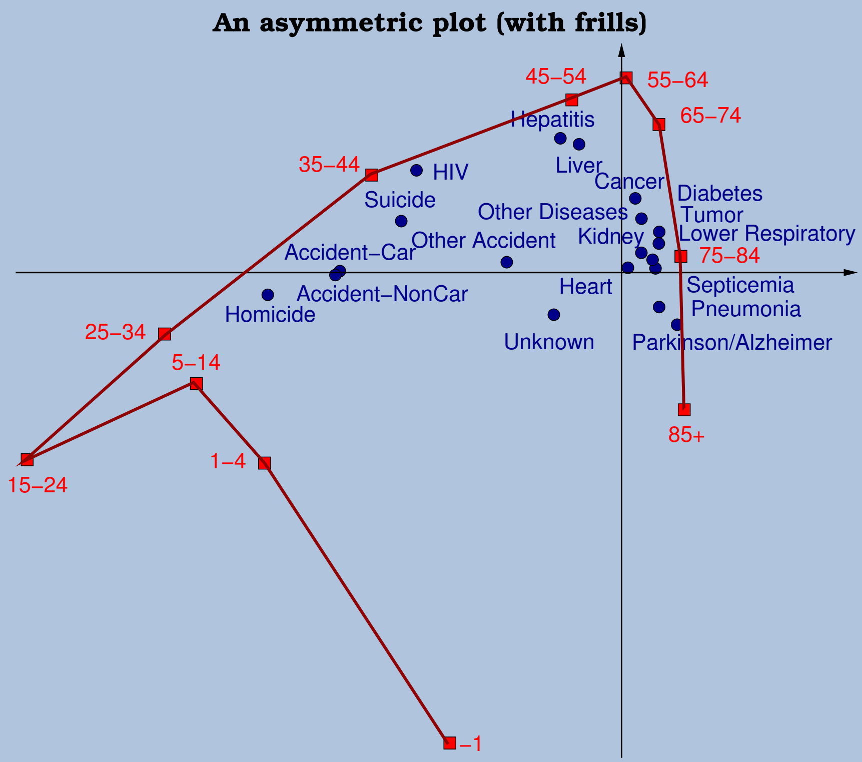
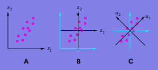


Figure 10: Causes of Death in the US: Correspondence Analysis. Asymmetric plot with frills (From R). There are three periods in life (or death!)



Another look at the causes of Death

Data are too Big. Standard Permutation Test Impossible

So Model Permutation and Bootstrap From Multinomial Distribution:

Permutation: X4H0 probability distribution under H_0 (independence).

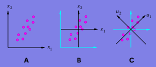
```
TotalNumberOfObservations = sum(X)

X.perm = matrix(as.vector(rmultinom(1,TotalNumberOfObservations,X4H0)),
                   nrow=nI,ncol=nJ,byrow = FALSE)
```

Bootstrap: Probability distribution is **X** itself

```
TotalNumberOfObservations = sum(X)
  X.perm = matrix(as.vector(rmultinom(1,TotalNumberOfObservations,X)),
                   nrow=nI,ncol=nJ,byrow = FALSE)
```

So here rows and Columns are kind of Random

Other Cases: Fixed Rows or Fixed columns

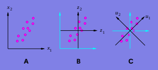


Refresher! and Menu
Home Page
Title Page
Page 37 of 61
Go Back
Full Screen
Close
Quit

Too Much Power: Malinvaud / permutation

	Ho: Omnibus	Dim-1	Dim-2	Dim-3	Dim-4
Inertia / sum lambda	.5542	0.3730	0.0829	0.0541	0.0311
Chi2	1302216.49	425815.57	231101.7258	104014.7592	31055.7673
p-Chi2	0.0000	0.0000	0.0000	0.0000	0.0000
df	180.0000	153.0000	128.0000	105.0000	84.0000
p-perm	0.0001	0.0001	0.0001	0.0001	0.0001

Too Much Power: Everything is Significant



Significant Bootstrap Ratios. I-set. Dimension 1

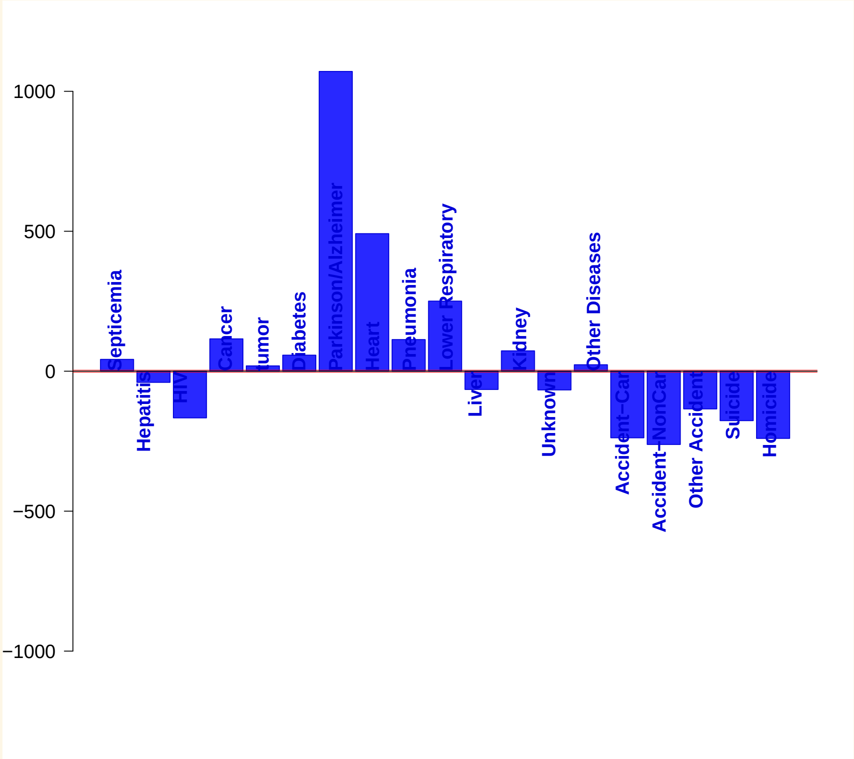


Figure 11: Causes of Death in the US: Correspondence Analysis. Bootstrap ratios I set (Causes). Dimension 1.

Significant Bootstrap Ratios. I-set. Dimension 2

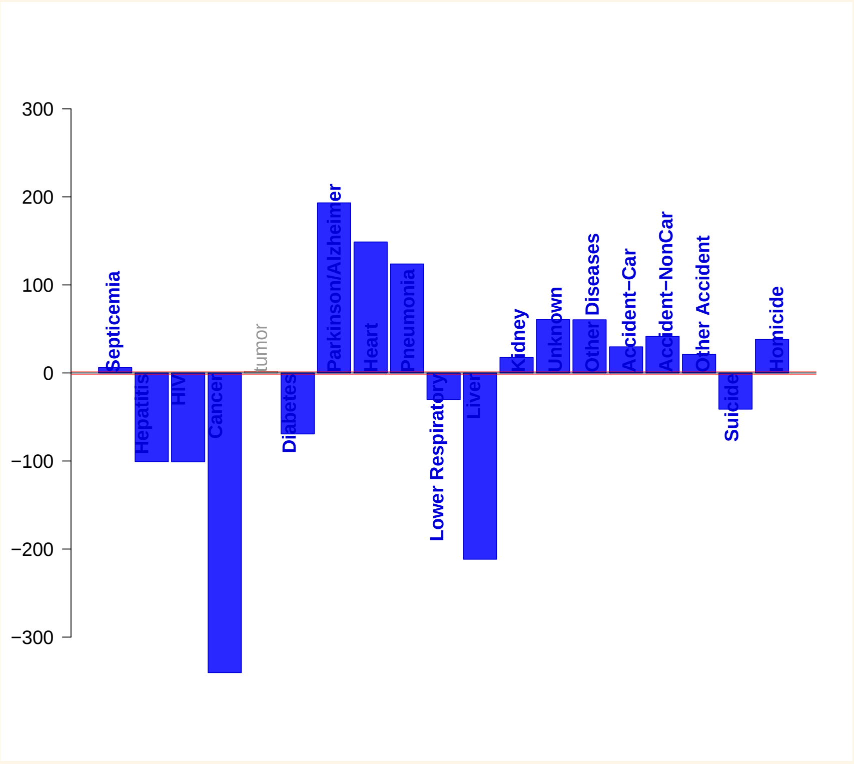
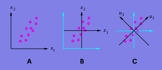


Figure 12: Causes of Death in the US: Correspondence Analysis. Bootstrap ratios I set (Causes). Dimension 2.



Home Page

Title Page

◀◀ ▶▶

◀ ▶

Page 40 of 61

Go Back

Full Screen

Close

Quit

Significant Bootstrap Ratios. J-set. Dimension 1

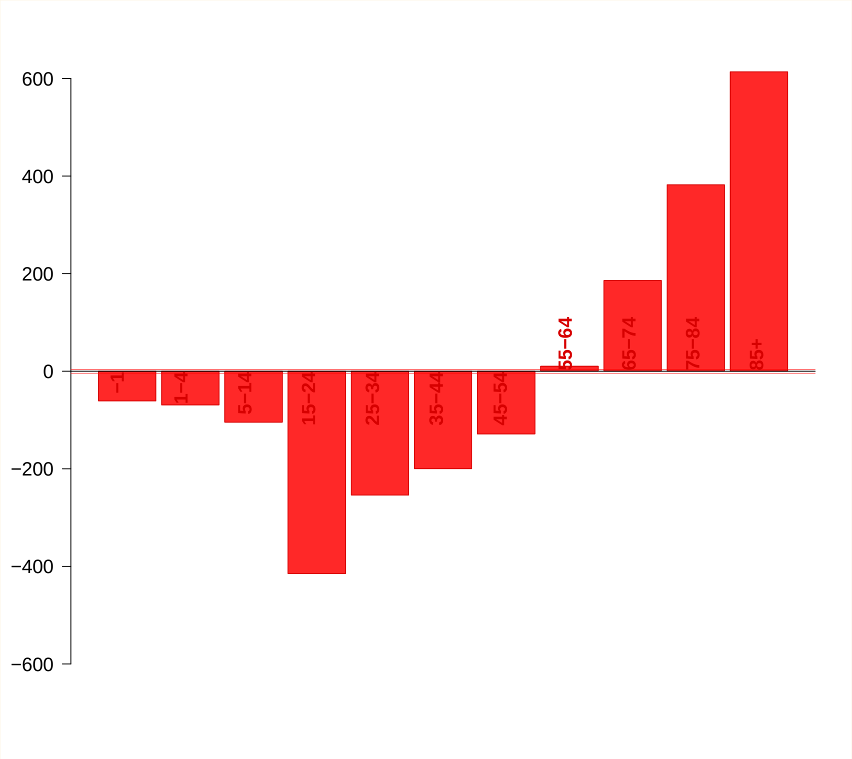
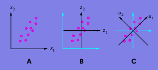


Figure 13: Causes of Death in the US: Correspondence Analysis. Bootstrap ratios J set (Ages). Dimension 1.



Significant Bootstrap Ratios. J-set. Dimension 2

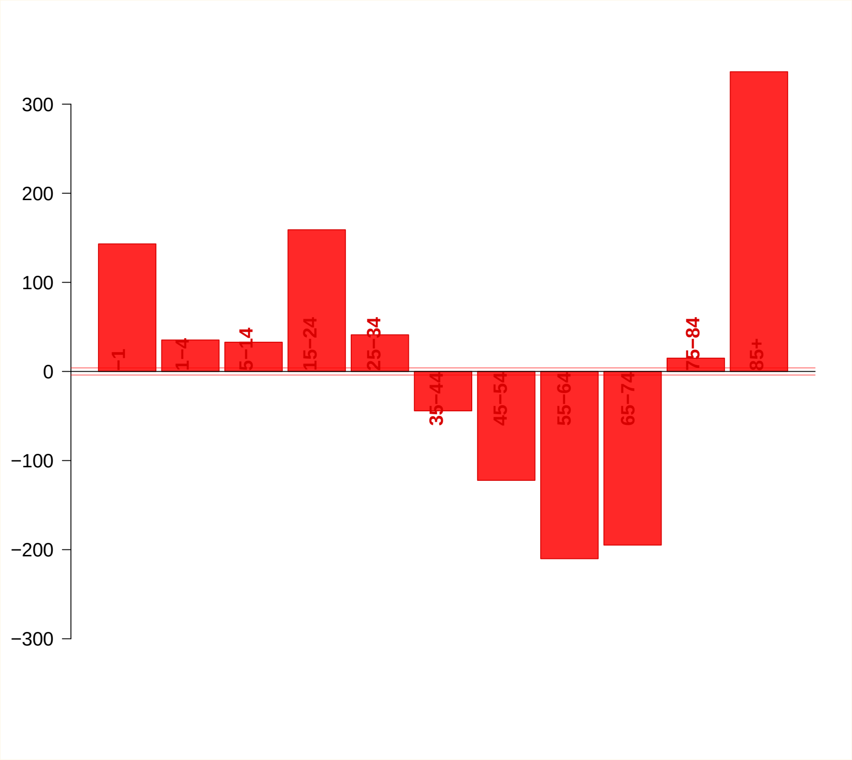


Figure 14: Causes of Death in the US: Correspondence Analysis. Bootstrap ratios J set (Causes). Dimension 2.

How to Permute.
A Cube: The Return of the Sound of Music

from: Abdi & Béra (2014). paper # C82

A science project (independent and dependent variables): Twenty-two participants (14 children and 10 adults) were presented with nine “pieces of music” and asked to associate one of ten colors to each piece of music. The 9 pieces of music:

- 1) the music of a video game (video),
- 2) a Jazz song (jazz),
- 3) a country and western song (country),
- 4) a rap song (rap),
- 5) a pop song (pop),
- 6) an extract of the opera Carmen (opera),
- 7) the low F note played on a piano (low F),
- 8) the middle F note played on the same piano
- 9) the high F note still played on the same piano.

Table 3: Twenty-two participants associated one of ten colors to nine pieces of music. The column labeled x_{i+} gives the total number of choices made for each color. N is the grand total of the data table. The vector of mass for the rows, $\mathbf{r}$, is the proportion of choices made for each color ($r_i = x_{i+}/N$). The row labeled x_{+j} gives the total number of times each piece of music was presented (*i.e.*, it is equal to the number of participants) The centroid row, gives these values as proportions.

Color	Video	Jazz	Country	Rap	Pop	Opera	Low F	High F	Middle F	x_{i+}
red	4	2	4	4	1	2	2	4	1	24
orange	3	4	2	2	1	1	0	3	2	18
yellow	6	4	5	2	3	1	1	3	0	25
green	2	0	5	1	3	3	3	1	5	23
blue	2	5	0	1	4	1	2	1	3	19
purple	3	3	1	0	0	3	0	2	1	13
white	0	0	0	0	1	4	1	5	3	14
black	0	2	0	11	1	3	10	1	1	29
pink	2	1	1	0	2	4	0	2	0	12
brown	0	1	4	1	6	0	3	0	6	21
x_{+j}	22	22	22	22	22	22	22	22	22	$N = 198$
$\mathbf{c}^T$	.11	.11	.11	.11	.11	.11	.11	.11	.11	

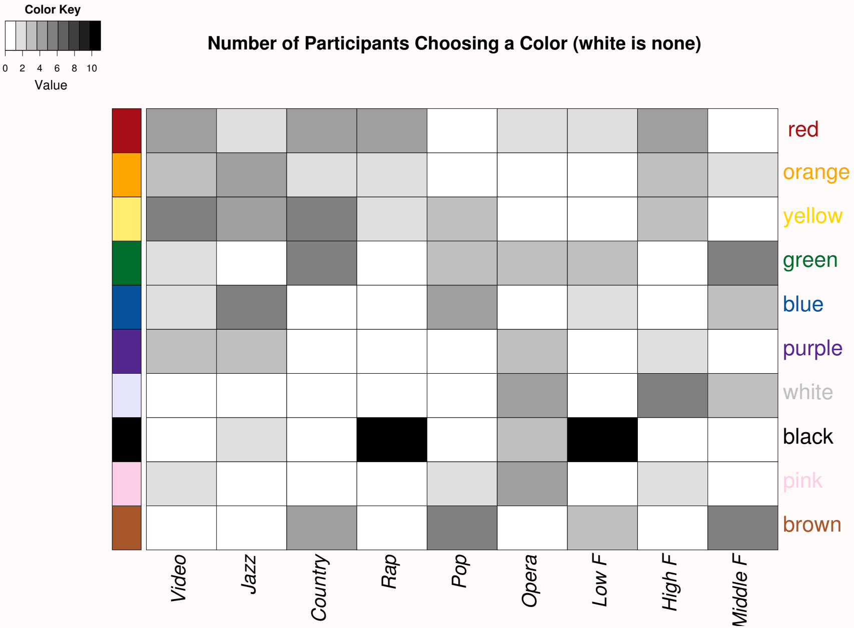
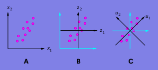


Figure 15: CA The Colors of Music. A nicer way to look at the data. A heat map of the data.



Refresher! and Menu
Home Page
Title Page

Go Back
Full Screen
Close
Quit

Table 4: CA The Color of Music. Factor scores, contributions, mass, mass $\times$ squared factor scores, inertia to barycenter, and squared cosines for the rows. For convenience, squared cosines and contributions have been multiplied by 1000 and rounded.

	F_1	F_2	ctr_1	ctr_2	r_i	$r_i\times F_1^2$	$r_i\times F_2^2$	$r_i\times d_{r,i}^2$	$\cos_1^2$	$\cos_2^2$
red	-0.026	0.299	0	56	.121	.000	.011	.026	3	410
orange	-0.314	0.232	31	25	.091	.009	.005	.030	295	161
yellow	-0.348	0.202	53	27	.126	.015	.005	.057	267	89
green	-0.044	-0.490	1	144	.116	.000	.028	.048	5	583
blue	-0.082	-0.206	2	21	.096	.001	.004	.050	13	81
purple	-0.619	0.475	87	77	.066	.025	.015	.050	505	298
white	-0.328	0.057	26	1	.071	.008	.000	.099	77	2
black	1.195	0.315	726	75	.146	.208	.014	.224	929	65
pink	-0.570	0.300	68	28	.061	.020	.005	.053	371	103
brown	0.113	-0.997	5	545	.106	.001	.105	.108	12	973
$\sum$	—	—	1000	1000	—	.287	.192	.746		
						λ_1	λ_2	$\mathcal{I}$		
						39%	26%			
						τ_1	τ_2			

Table 5: CA The Colors of Music. Factor scores, contributions, mass, mass $\times$ squared factor scores, inertia to barycenter, and squared cosines for the columns. For convenience, squared cosines and contributions have been multiplied by 1000 and rounded.

	G_1	G_2	$\widetilde{G}_1$	$\widetilde{G}_2$	ctr$_1$	ctr$_2$	c_j	$c_j \times G_1^2$	$c_j \times G_2^2$	$c_j \times d_{c,j}^2$	$\cos_1^2$	$\cos_2^2$
Video	-0.54	0.39	-1.01	0.88	113	86	.111	.03	.017	.071	454	232
Jazz	-0.26	0.28	-0.48	0.63	25	44	.111	.01	.008	.069	105	121
Country	-0.29	-0.31	-0.54	-0.70	33	55	.111	.01	.011	.066	142	161
Rap	0.99	0.40	1.85	0.90	379	91	.111	.11	.017	.133	822	132
Pop	-0.12	-0.64	-0.23	-1.45	6	234	.111	.00	.045	.064	26	709
Opera	-0.24	0.33	-0.44	0.74	22	61	.111	.01	.012	.079	78	149
Low.F	0.95	-0.09	1.78	-0.20	351	5	.111	.10	.001	.105	962	8
High.F	-0.43	0.41	-0.80	0.93	70	96	.111	.02	.018	.074	271	249
Middle.F	-0.07	-0.76	-0.13	-1.72	2	330	.111	.00	.064	.084	7	759
$\sum$	—	—	—	—	1000	1000	—	.287	.192	.746		
								λ_1	λ_2	$\mathcal{I}$		
								39%	26%			
								τ_1	τ_2			

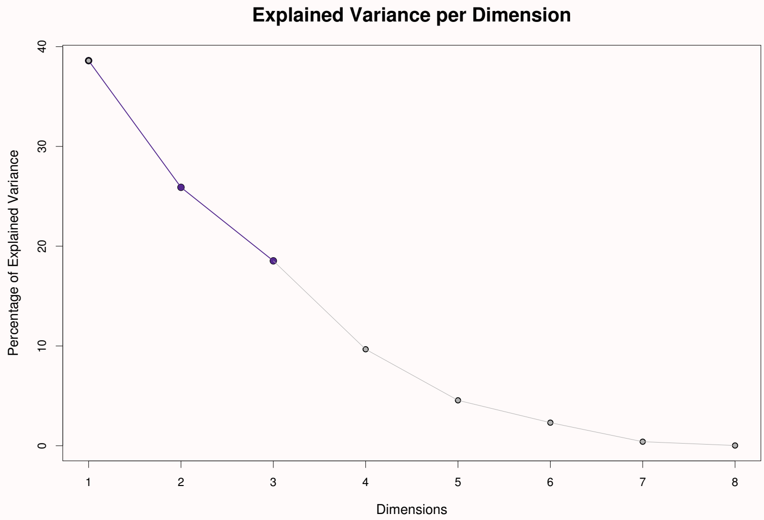
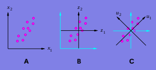


Figure 16: CA The Colors of Music. Scree Test of the percentage of explained variance. Three dimensions are larger than $\frac{1}{8}$.



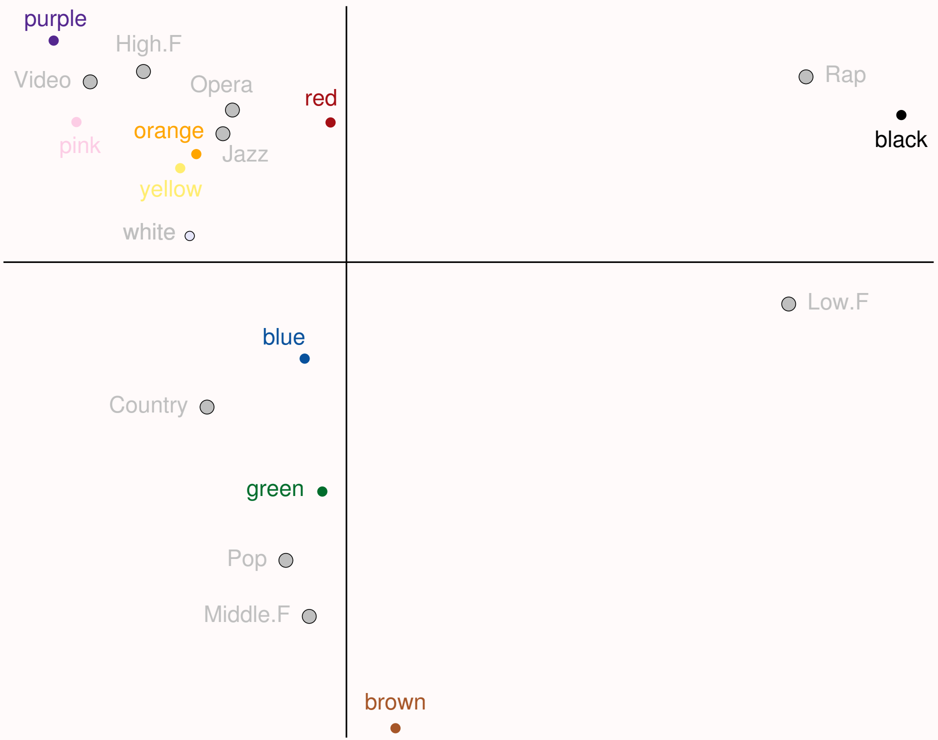


Figure 17: CA The Colors of Music. Symmetric Plot: The projections of the rows and the columns are displayed in the same map. $\lambda_1 = .287$, $\tau_1 = 39$; $\lambda_2 = .192$, $\tau_2 = 26$. In this plot the proximity between rows and columns cannot be directly interpreted.

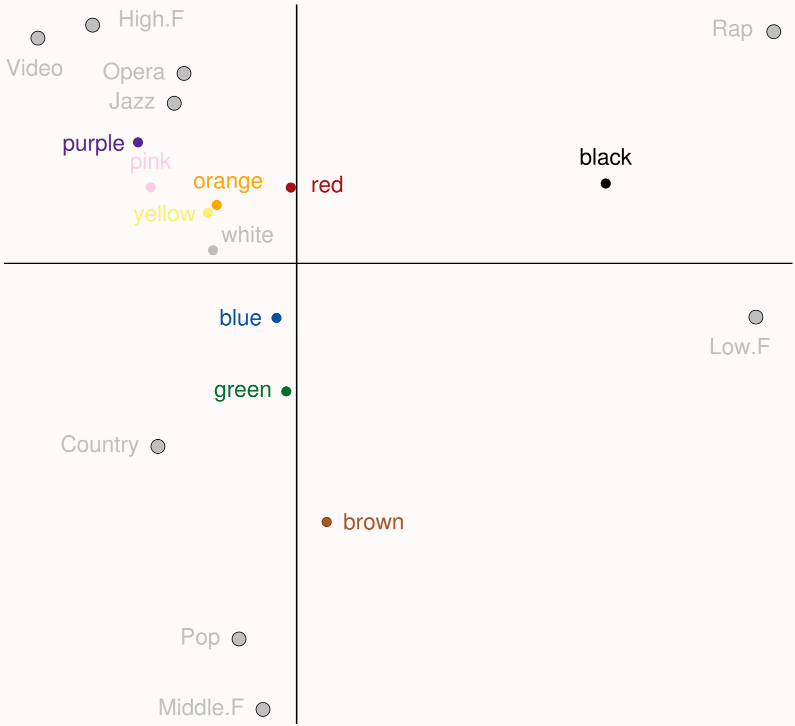


Figure 18: CA The Colors of Music. Asymmetric Plot: The projections of the rows and the columns are displayed in the same map. The inertia of the projections of the column factor scores is equal to one for each dimension and the inertia of the projections of the row factor scores are $\lambda_1 = .287$, $\tau_1 = 39$; $\lambda_2 = .192$, $\tau_2 = 26$.

How to Permute?

	Age	G	Video	Jazz	Country	Rap	Pop	Opera	Low-F	High-F	Middle-F
AF1	3	W	1	5	4	8	10	8	7	2	10
AF2	3	W	4	8	10	1	4	2	1	8	8
DF1	2	W	2	5	1	1	3	8	5	1	7
AM3	3	M	3	3	1	8	5	9	8	7	4
DM2	2	M	2	9	4	8	3	7	1	3	4
AF4	3	W	6	3	1	8	8	4	10	2	5
CF1	1	W	5	2	1	8	9	7	4	3	10
CF2	1	W	1	3	4	8	5	7	10	9	6
CF3	1	W	1	2	3	3	5	6	5	4	10
CF4	1	W	9	1	4	2	10	8	3	6	10
CM5	1	M	3	8	10	5	7	9	4	2	1
CF6	1	W	6	3	9	1	10	4	8	1	2
CM7	1	M	1	2	10	3	4	5	8	6	7
CM8	1	M	9	1	3	4	2	3	4	5	10
AF5	3	W	3	5	3	2	10	9	8	7	4
AM6	3	M	3	2	10	10	4	6	8	7	4
DM3	2	M	5	6	2	8	10	4	8	3	10
AF7	3	W	6	5	3	8	5	7	8	9	4
AM8	3	M	2	6	4	1	9	6	10	7	5
DM4	2	M	4	10	3	8	1	9	8	7	2
AM9	3	M	3	6	2	8	10	1	8	1	5
AF10	3	W	3	5	6	8	3	1	8	1	7

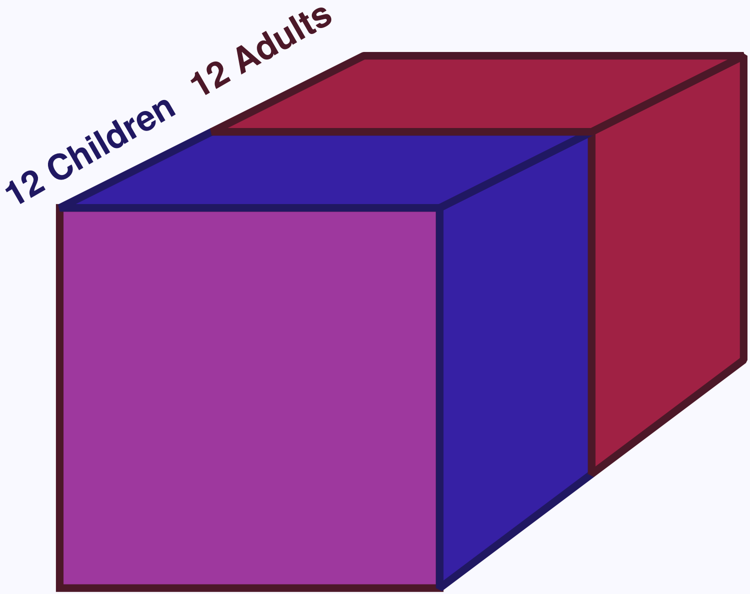
Example Participant 1: Assigned Color 1 to Video, Color 5 to Jazz, etc.

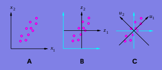
What would be a random assignment? we need to think

Possible Model:
The Colors are Fixed
and Randomly Assigned to the Music

How to Bootstrap?
What Are the Independent Units?

Think Again! What is the Design?





Refresher! and Menu
Home Page
Title Page
Page 53 of 61
Go Back
Full Screen
Close
Quit

Bootstrap the Cube!

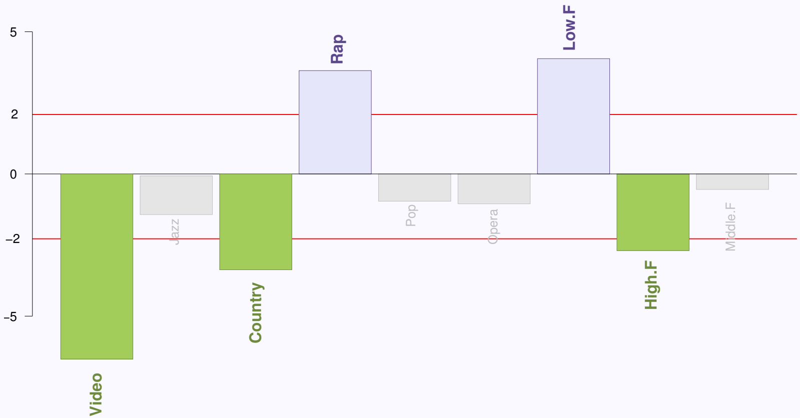


Figure 19: CA The Colors of Music. Bootstrap Ratios. Dimension 1. The I Set (Music).

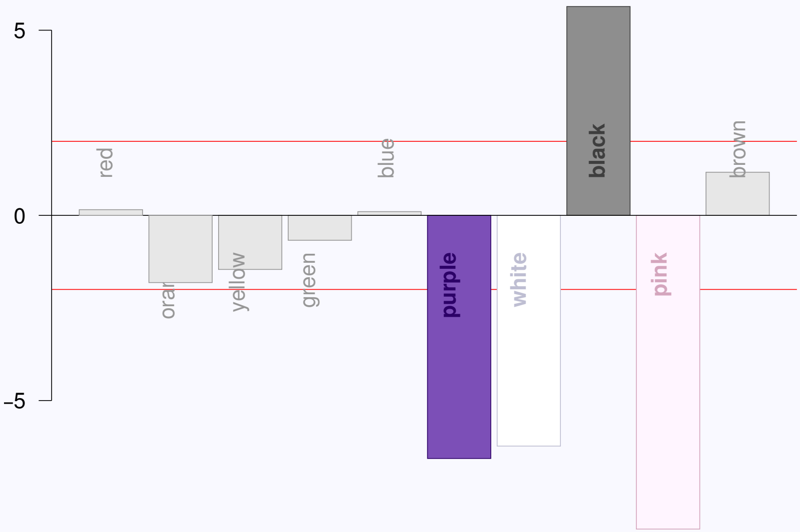
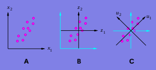


Figure 20: CA The Colors of Music. Bootstrap Ratios. Dimension 1. The J Set (Colors).



Refresher! and Menu
Home Page
Title Page
Page 56 of 61
Go Back
Full Screen
Close
Quit

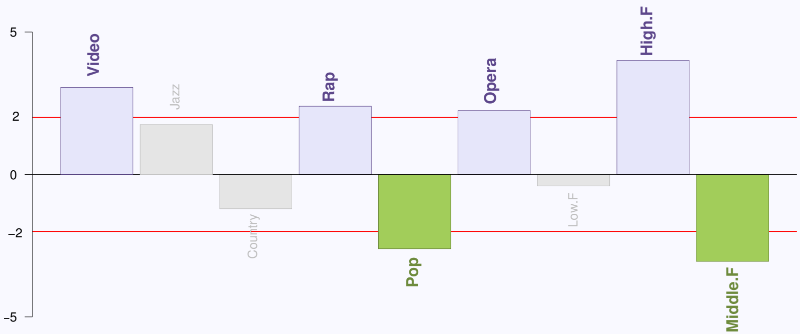


Figure 21: CA The Colors of Music. Bootstrap Ratios. Dimension 2. The I Set (Music).

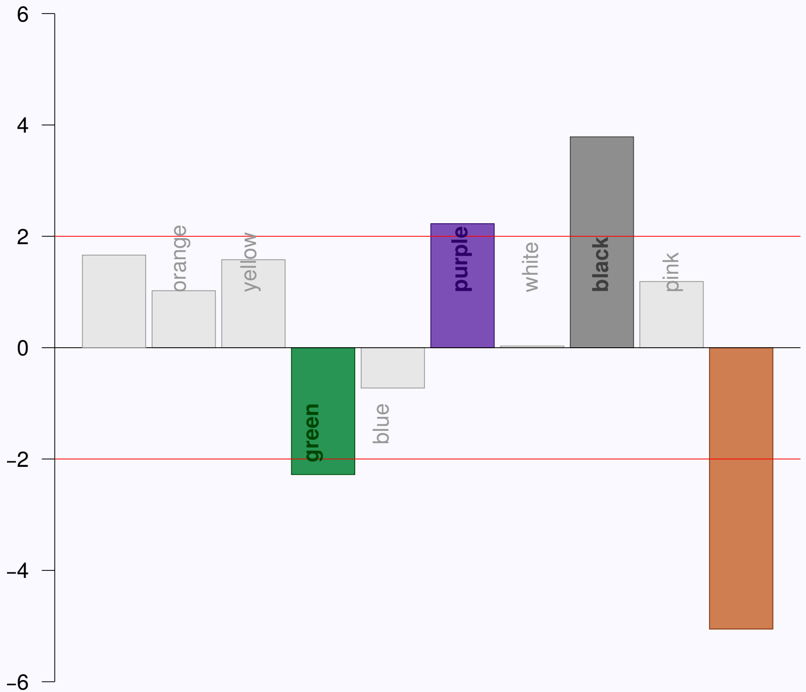


Figure 22: CA The Colors of Music. Bootstrap Ratios. Dimension 2. The J Set (Colors).

With Ellipses

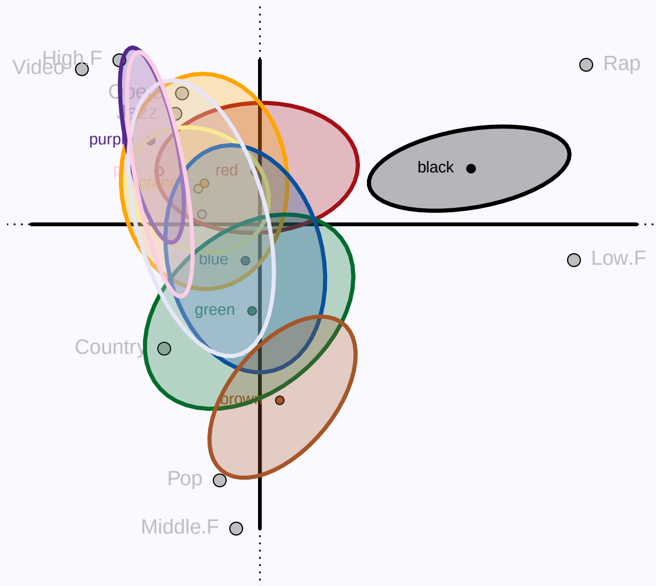


Figure 23: CA The Colors of Music. Bootstrap Confidence Intervals. Dimension 1 and 2. The I Set (Music).

With Ellipses

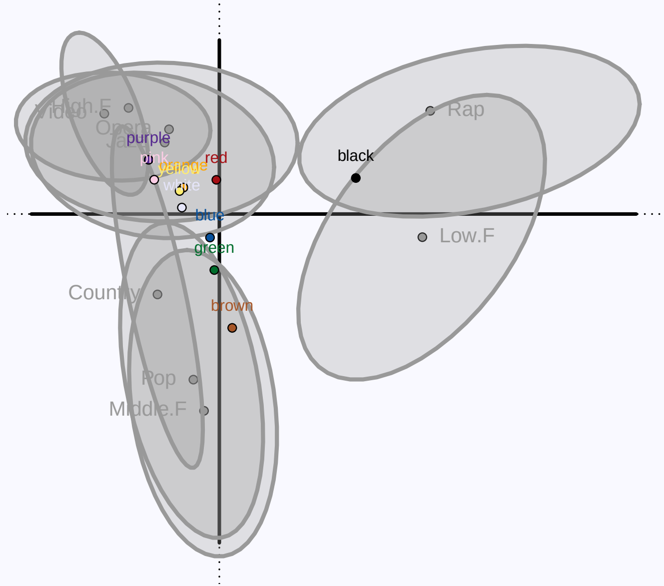
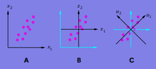


Figure 24: CA The Colors of Music. Bootstrap Confidence Intervals. Dimension 1 and 2. The I Set (Music).

Time to Conclude: Inferences for CA



Refresher! and Menu
Home Page
Title Page
Page 61 of 61
Go Back
Full Screen
Close
Quit