Analyse factorielle
Analyse (factorielle) des Correspondances

IUT de Vannes - Université de Bretagne-Sud

Thierry Dhorne - www.dhorne.education

9 octobre 2015

Introduction

Étude du lien entre deux variables qualitatives binaires

- lorsque les deux variables qualitatives ont deux modalités chacune les coefficients de corrélation entre deux quelconques des indicatrices disjonctives sont égaux (au signe près)

$$cor^2(X_{11}, X_{21}) = cor^2(X_{11}, X_{22}) = cor^2(X_{12}, X_{21}) = cor^2(X_{12}, X_{22})$$

- ➤ en effet on inverse simplement les 0 et les 1
- ➤ ces coefficients sont par ailleurs égaux au $\phi^2 = \frac{\chi^2}{n}$ d'indépendance sur la table de contingence associée
- ★ la mesure du lien du χ^2 est donc la mesure du lien linéaire classique
- ➤ le lien linéaire classique ne s'étend cependant pas immédiatement au cas multidimensionnel
- ➤ le χ^2 est, en revanche, général

Étude du lien entre variables qualitatives quelconques

- dans ce cas le lien entre deux variables qualitatives est « multidimensionnel »

➢ plus précisément le lien peut s'apprécier dans un espace de dimension

$$\min(I-1, J-1)$$

où I et J sont les nombres de modalités respectifs des deux variables

- une mesure globale possible du lien entre les deux variables est le χ^2 qui peut être normalisé en le divisant par le nombre d'observations ce qui conduit au ϕ^2

★ on peut par ailleurs prolonger la normalisation en le divisant par $\min(I-1, J-1)$ ce qui conduit au coefficient de Cramér.

➢ il peut donc être judicieux d'étudier le lien de manière moins globale en le décomposant dans l'espace de ce lien

Étude du lien

Analyse factorielle du lien entre deux variables qualitatives

- on dispose du tableau disjonctif des deux variables qualitatives

$$
\begin{array}{c|cccc|cccc|}
 & & Variable1 & & & & Variable2 & & \\
 & mod1 & mod2 & \dots & modI & mod1 & mod2 & \dots & modJ \\
\hline
i_1 & I(111) & I(112) & \dots & I(11I) & I(121) & I(122) & \dots & I(12J) \\
i_2 & I(211) & I(212) & \dots & I(21I) & I(221) & I(222) & \dots & I(22J) \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
i_r & I(r11) & I(r12) & \dots & I(r1I) & I(r21) & I(r22) & \dots & I(r2J) \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
i_n & I(n11) & I(n12) & \dots & I(n1I) & I(n21) & I(n22) & \dots & I(n2J) \\
\hline
\end{array}
$$

où

– $I(rvk) = 1$ si le $r$ième individu a la modalité k de la variable v
– $I(rvk) = 0$ sinon

Structuration du lien entre variables qualitatives

- au lieu de considérer le lien global (mesuré par le χ^2)

➤ on va chercher s'il existe une combinaison (linéaire) des modalités de la première variable très liée à une combinaison (linéaire) des modalités de la deuxième

- on cherche les deux combinaisons les plus liées

➤ ces combinaisons sont $X_1\alpha$ et $X_2\beta$ (cf td)

- on maximise la corrélation entre ces deux combinaisons

$$\frac{{}^{t}(X_1\alpha)X_2\beta}{\sqrt{{}^{t}(X_1\alpha)X_1\alpha}\sqrt{{}^{t}(X_2\beta)X_2\beta}}$$

Calcul de la corrélation

au carré

- la corrélation pouvant être négative, on considère son carré pour être dans un cadre strict de maximisation
  - ➤ ceci a aussi l'avantage de supprimer les $\sqrt{\ }$
  - ➤ on doit donc maximiser

$$\frac{{}^t\alpha({}^tX_1X_2)\beta{}^t\beta({}^tX_2X_1)\alpha}{({}^t\alpha({}^tX_1X_1)\alpha)({}^t\beta({}^tX_2X_2)\beta)} = \frac{{}^t\alpha V_{12}\beta{}^t\beta V_{21}\alpha}{({}^t\alpha V_{11}\alpha)({}^t\beta V_{22}\beta)}$$

  - ➤ ou, ce qui est équivalent

$$\frac{{}^t\beta({}^tX_2X_1)\alpha{}^t\alpha({}^tX_1X_2)\beta}{({}^t\alpha({}^tX_1X_1)\alpha)({}^t\beta({}^tX_2X_2)\beta)} = \frac{{}^t\beta V_{21}\alpha{}^t\alpha V_{12}\beta}{({}^t\alpha V_{11}\alpha)({}^t\beta V_{22}\beta)}$$

Matrices concernées

- le tableau de Burt est

➤ ${}^{\mathrm{t}}XX$,

- on a

- $V_{11} = {}^{\mathrm{t}}X_1 X_1$,

➤ premier bloc diagonal du tableau de Burt

- $V_{22} = {}^{\mathrm{t}}X_2 X_2$,

➤ deuxième bloc diagonal du tableau de Burt

- $V_{12} = {}^{\mathrm{t}}X_1 X_2$

➤ bloc extra-diagonal supérieur du tableau de Burt (ou table de contingence)

- $V_{21} = {}^{\mathrm{t}}X_2 X_1$

➤ bloc extra-diagonal inférieur du tableau de Burt (transposée de la table de contingence).

Maximisation

- en dérivant par rapport à α et en annulant, on obtient

$$\frac{1}{{}^t\beta V_{22}\beta}\frac{2{}^t\alpha V_{11}\alpha V_{12}\beta{}^t\beta V_{21}\alpha-2{}^t\alpha V_{12}\beta{}^t\beta V_{21}\alpha V_{11}\alpha}{({}^t\alpha V_{11}\alpha)^2}=0$$

➤ soit :

$$V_{12}\beta=k_1V_{11}\alpha$$

➤ et donc symétriquement en β :

$$V_{21}\alpha=k_2V_{22}\beta$$

Résolution

- la première équation donne

$$V_{11}^{-1}V_{12}\beta = k_1\alpha$$

- la seconde donne

$$V_{22}^{-1}V_{21}\alpha = k_2\beta$$

➢ en remplaçant la deuxième dans la première on a

$$V_{11}^{-1}V_{12}V_{22}^{-1}V_{21}\alpha = k_1k_2\alpha$$

➢ et symétriquement

$$V_{22}^{-1}V_{21}V_{11}^{-1}V_{12}\beta = k_1k_2\beta$$

★ α et β sont donc vecteurs propres des matrices concernées et correspondent aux mêmes valeurs propres

Exemple

Exemple

Présidentielles 2012

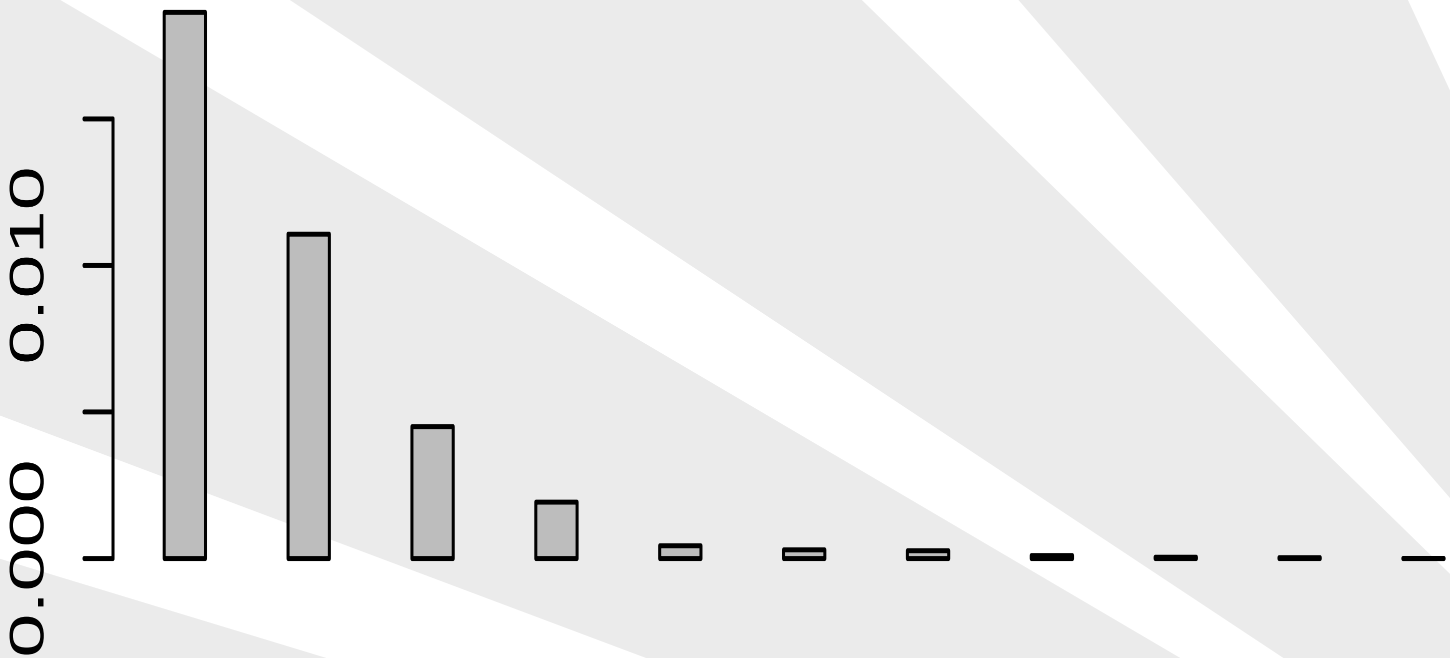
- on dispose du tableau des résultats du premier tour des élections présidentielles 2012, par région
- on ne travaille que sur le tableau de contingence
- ➤ il y a 27 lignes : les régions (21 métro + Corse + DOM)
- ➤ il y a 12 colonnes : les 10 candidats + abstentions + blancs et nuls
- le tableau a l'allure suivante

	abstentions	blancs-nuls	Eva JOLY	Marine LE PEN
Alsace	258270	19648	27168	219252
Aquitaine	397170	37299	45051	296151
Auvergne	164438	19250	15356	139768
Basse-Normandie	181511	15958	17279	150810
Bourgogne	216817	19159	17077	191148
Bretagne	364724	36077	58396	262095
Centre	326560	28827	26314	280096
Champagne-Ardenne	185831	13154	10150	172632
Corse	56696	2921	3678	39209
Franche-Comté	141504	14512	14369	141972
Guadeloupe	141585	11364	2134	7486
Guyane	37706	1756	843	3920
Haute-Normandie	244564	19555	16900	207520
Île-de-France	1515884	85997	144346	655926

```
> round(acc1(tabpres)$valp,4)

 [1] 0.0186 0.0111 0.0045 0.0019 0.0004 0.0003
 [7] 0.0003 0.0001 0.0000 0.0000 0.0000

> round(acc1(tabpres)$valp/sum(acc1(tabpres)$valp),2)

 [1] 0.50 0.30 0.12 0.05 0.01 0.01 0.01 0.00 0.00
[10] 0.00 0.00

>
```

Présidentielles 2012

Éboulis des valeurs propres



- il y aura au maximum 4 valeurs propres à retenir
  - ➤ au lieu des 11 dimensions initiales

Lien avec le ϕ^2

- il est facile de montrer le résultat suivant

ϕ^2 d'association et valeurs propres

Le ϕ^2 d'association $(= \frac{\chi^2}{n})$ est égal à la somme des valeurs propres (non triviales) de l'analyse factorielle des correspondances

- soit

$$\sum_{d=1}^{D} \lambda_d = \frac{\chi^2}{n}$$

➤ où D est la dimension de l'espace de lien ou coefficient de Cramér : $\min(I-1, J-1)$

Valeurs propres

Test de dimensionnalité

- les valeurs propres mesurant le lien comme un χ^2
- ➤ on peut les utiliser pour tester la dimension de l'espace de lien
- Malinvaud a montré que

Test de dimensionnalité de l'AFC

$$n(\lambda_{k+1} + \lambda_{k+2} + \ldots + \lambda_D) \rightsquigarrow \chi^2_{(I-1-k)(J-1-k)}$$

- si l'on réalise les tests successivement

```
> for (k in 10:1)
+ print(c(k,sum(tabpres)*sum(valp[(k+1):11]),qchisq(0.95
```

```
[1]  10.00000 133.78890   27.58711
[1]     9.00000 1471.67787    50.99846
[1]     8.00000 3433.91502    75.62375
[1]     7.0000 7982.7375  101.8795
[1]      6.000 19800.757    129.918
[1]      5.0000 33001.2613    159.8135
[1]      4.0000 52356.0550    191.6084
[1]       3.0000 138203.9306      225.3288
[1]       2.0000 338753.4872      260.9921
[1]       1.0000 832389.9208      298.6106
```

➢ on s'aperçoit que l'on devrait garder (et donc étudier ! !) toutes les dimensions

Aide à l'interprétation

★ les vecteurs propres donnent les coefficients des combinaisons linéaires

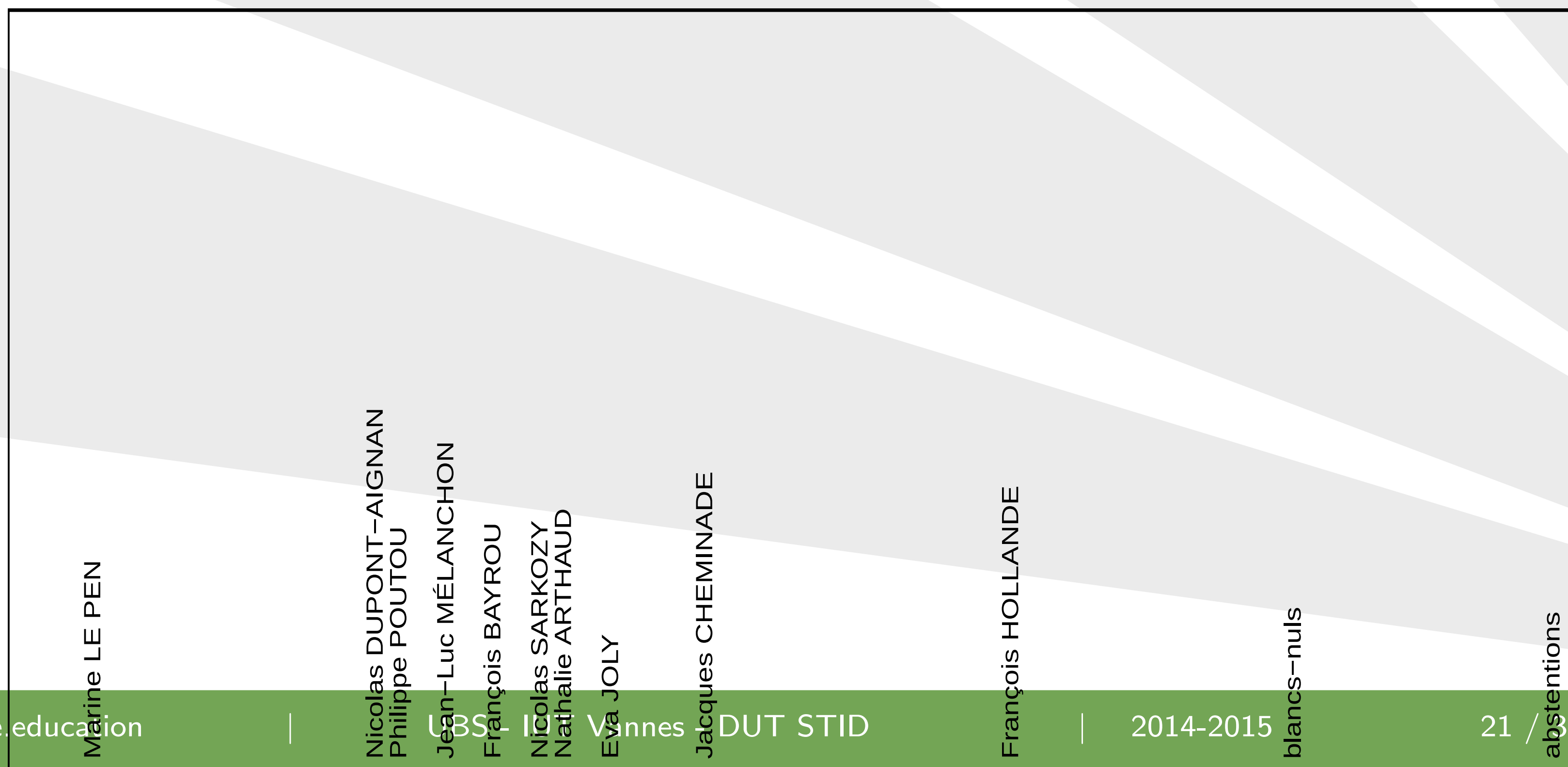
● pour les colonnes (les votes), on a comme premier vecteur propre

```
> round(1000*vecpc[,1],4)

          abstentions            blancs-nuls
                0.033                  0.022
             Eva JOLY         Marine LE PEN
               -0.007                 -0.029
      Nicolas SARKOZY   Jean-Luc MÉLANCHON
               -0.010                 -0.014
     Philippe POUTOU     Nathalie ARTHAUD
               -0.016                 -0.009
    Jacques CHEMINADE      François BAYROU
               -0.003                 -0.012
Nicolas DUPONT-AIGNAN    François HOLLANDE
               -0.017                  0.010
```

Interprétation des combinaisons linéaires

Graphique

```
> par(mai=c(1,0,0.75,0.4),oma=c(0,0,0,0))
> plot(range(vecpc[,1]),c(0,0.6),type="n",yaxt="n",ylab=
> text(vecpc[,1],0,colnames(tabpres),cex=0.5,pos=4,srt=9
```



centrée sur le cœur du nuage

Vecteurs propres

Espace des lignes

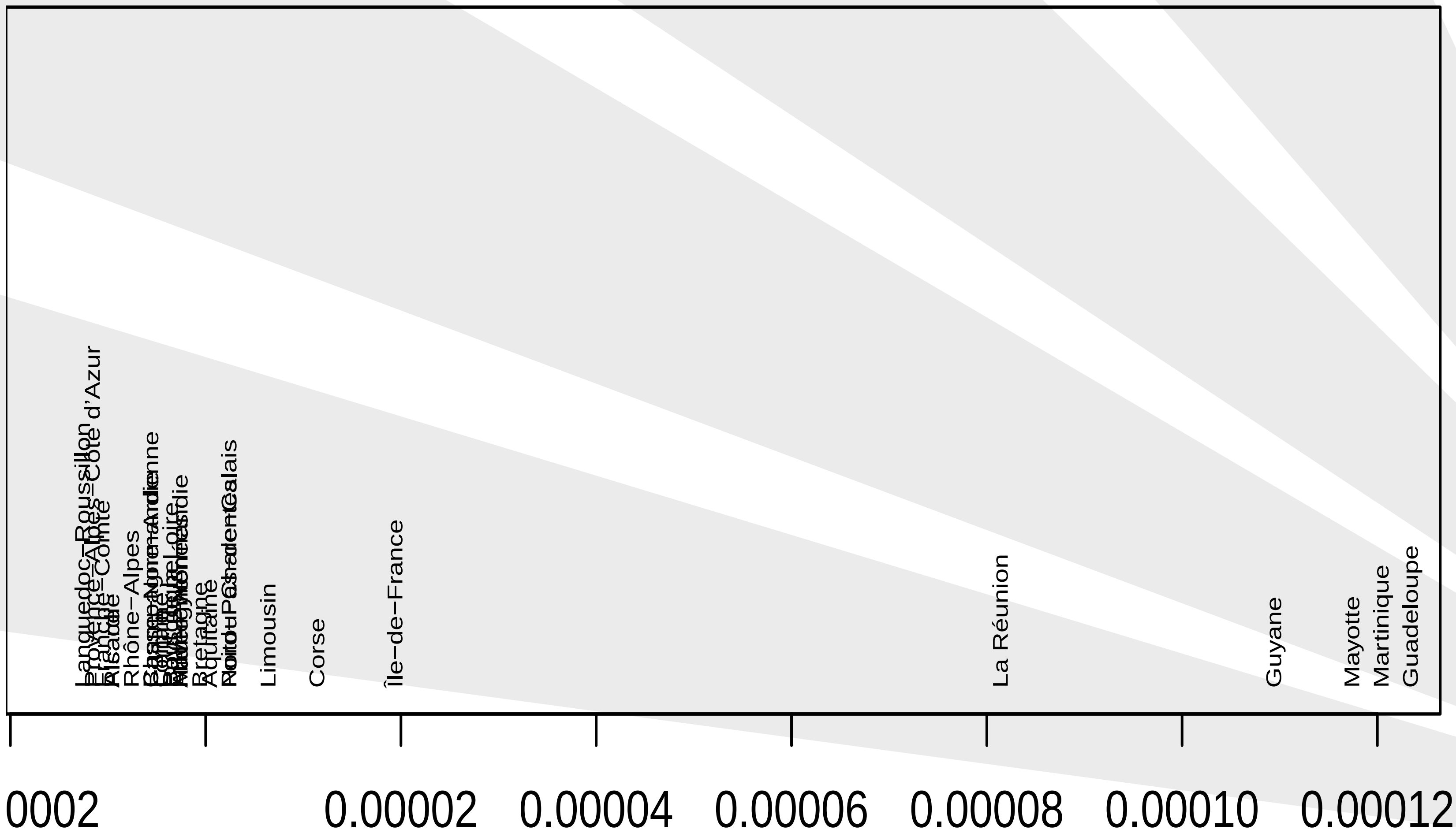
- pour les lignes (les régions), on a comme premier vecteur propre

*> round(1000*vecpl[,1],5)*

```
             Alsace                   Aquitaine            Auv
             -0.012                      -0.002              -
     Basse-Normandie                   Bourgogne            Bre
             -0.008                      -0.006              -
             Centre          Champagne-Ardenne
             -0.007                      -0.008
       Franche-Comté                  Guadeloupe              G
             -0.013                       0.121
     Haute-Normandie               Île-de-France           La Ré
             -0.005                       0.017
 Languedoc-Roussillon                   Limousin            Lor
             -0.015                       0.004              -
          Martinique                     Mayotte       Midi-Pyr
              0.118                       0.115              -
  Nord-Pas-de-Calais             Pays de la Loire           Pic
              0.000                      -0.006              -
   Poitou-Charentes Provence-Alpes-Côte d'Azur          Rhône-
              0.000                      -0.014              -
```

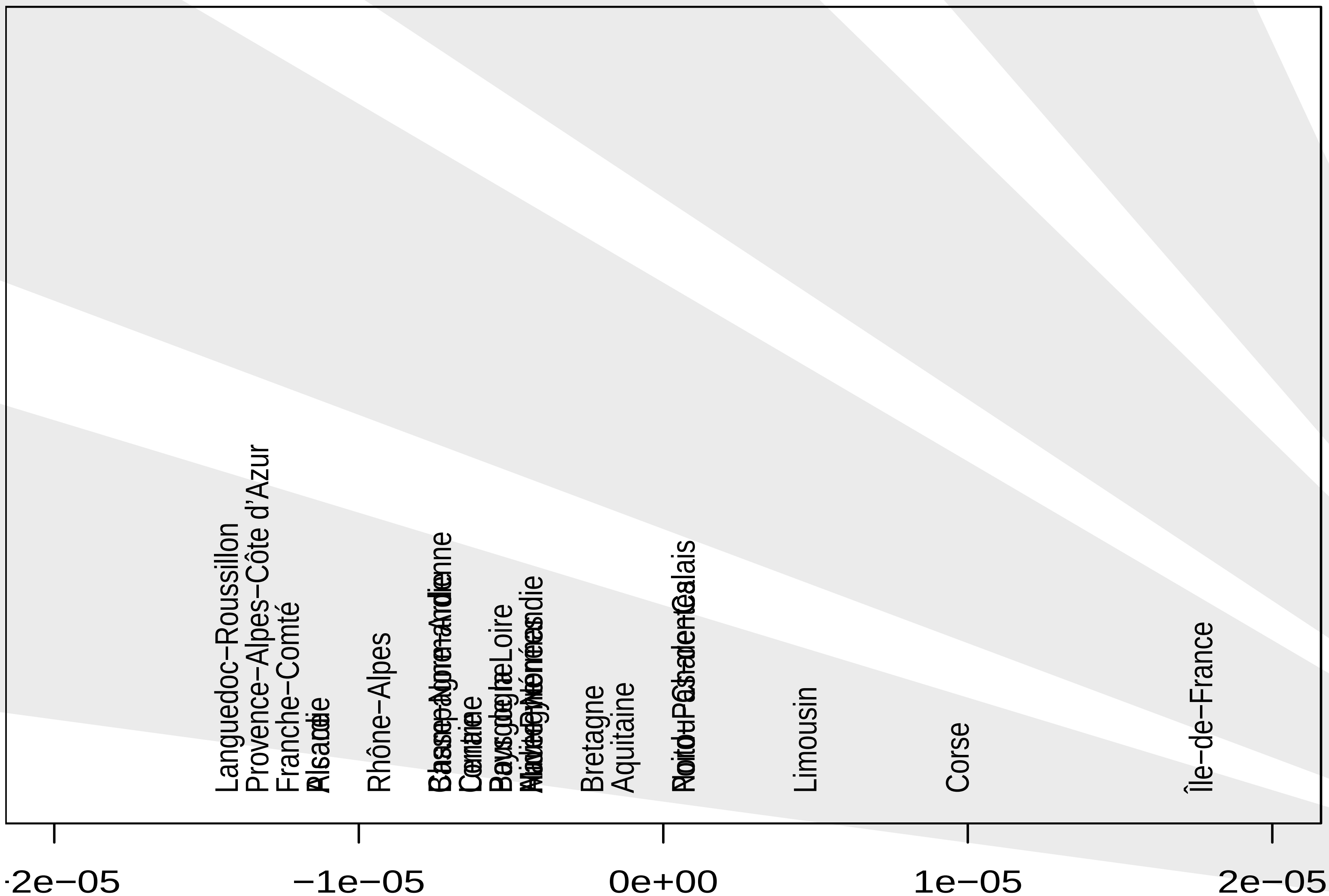
Interprétation des combinaisons linéaires
Graphique

centrée sur le cœur du nuage



Interprétation des combinaisons linéaires
sans les régions ultramarines

centrée sur le cœur du nuage



On peut évidemment faire la même chose....

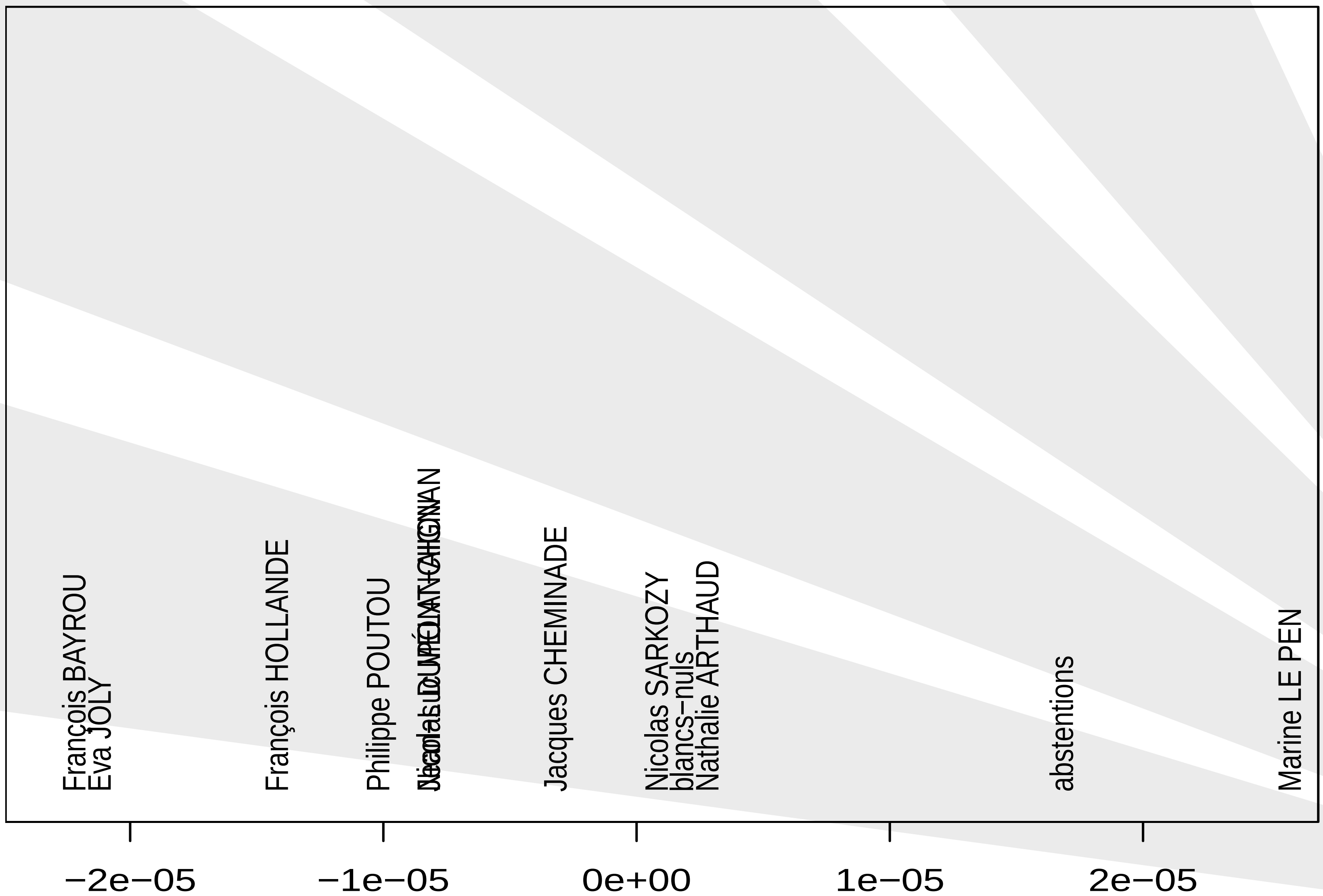
Avec la deuxième composante

centrée sur le cœur du nuage

- résultats

```
        abstentions          blancs-nuls
              0.016                0.001
          Eva JOLY       Marine LE PEN
             -0.022                0.025
    Nicolas SARKOZY   Jean-Luc MÉLANCHON
              0.000               -0.009
   Philippe POUTOU     Nathalie ARTHAUD
             -0.011                0.002
 Jacques CHEMINADE     François BAYROU
             -0.004               -0.023
Nicolas DUPONT-AIGNAN  François HOLLANDE
             -0.009               -0.015
```

Interprétation des combinaisons linéaires
Graphique



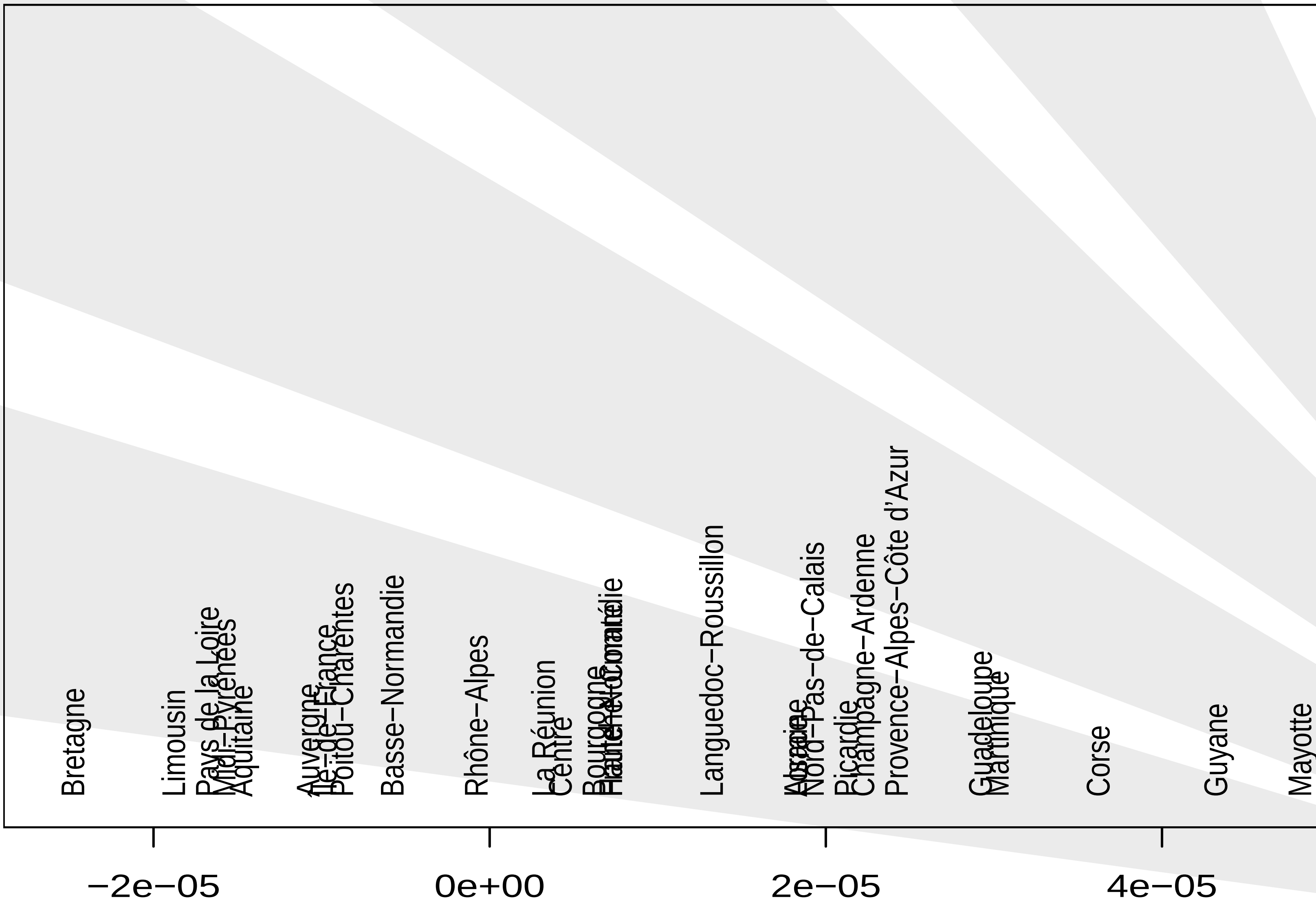
- pour les lignes (les régions), on a comme premier vecteur propre

Alsace	Aquitaine	Auvergne
0.017	-0.016	-0.012
Basse-Normandie	Bourgogne	Bretagne
-0.007	0.005	-0.026
Centre	Champagne-Ardenne	Corse
0.003	0.021	0.035
Franche-Comté	Guadeloupe	Guyane
0.006	0.028	0.042
Haute-Normandie	Île-de-France	La Réunion
0.006	-0.011	0.002
Languedoc-Roussillon	Limousin	Lorraine
0.012	-0.020	0.017
Martinique	Mayotte	Midi-Pyrénées
0.029	0.047	-0.017
Nord-Pas-de-Calais	Pays de la Loire	Picardie
0.018	-0.018	0.020
Poitou-Charentes	Provence-Alpes-Côte d'Azur	Rhône-Alpes
-0.010	0.023	-0.002

Interprétation des combinaisons linéaires

Graphique

centrée sur le cœur du nuage

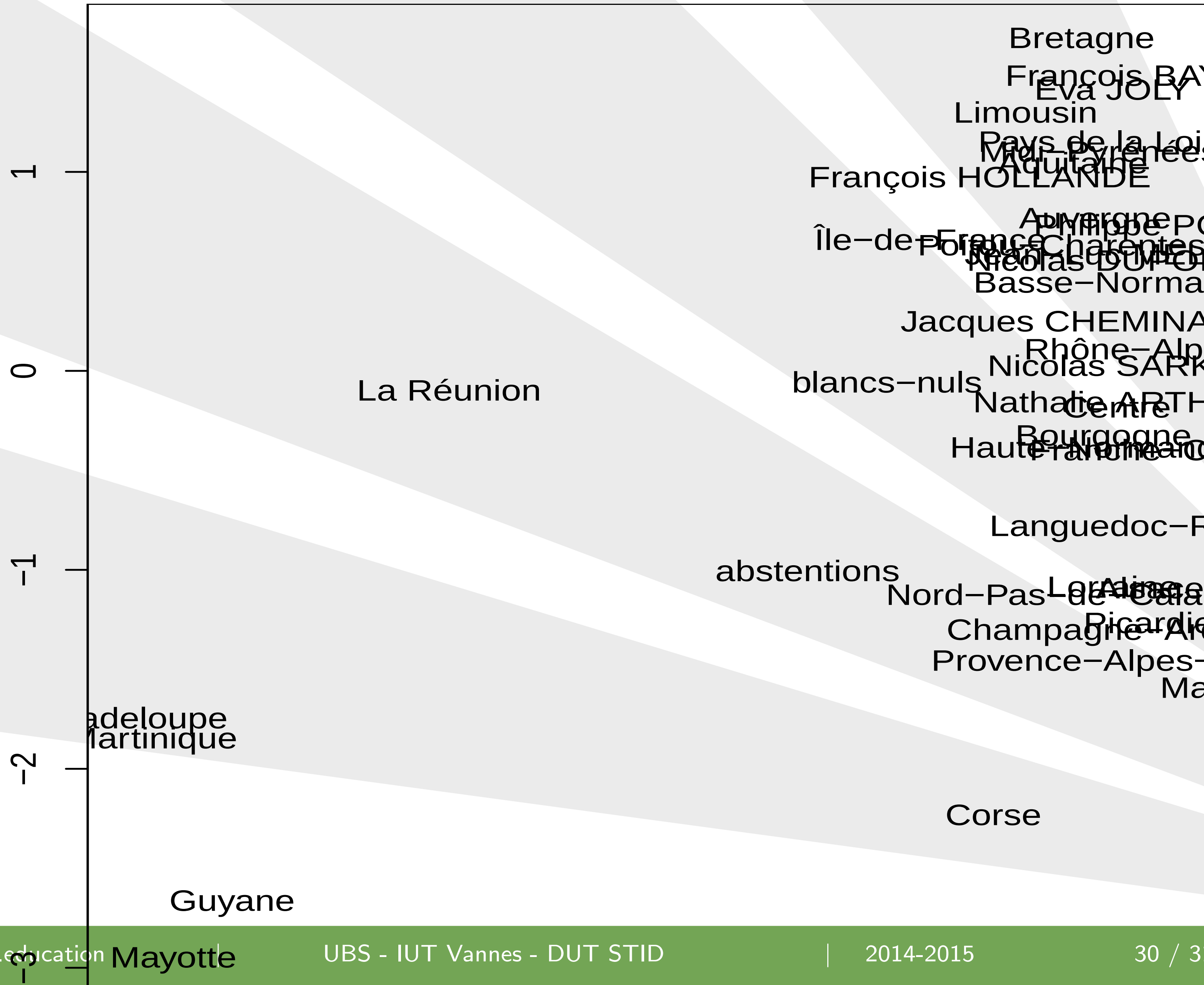


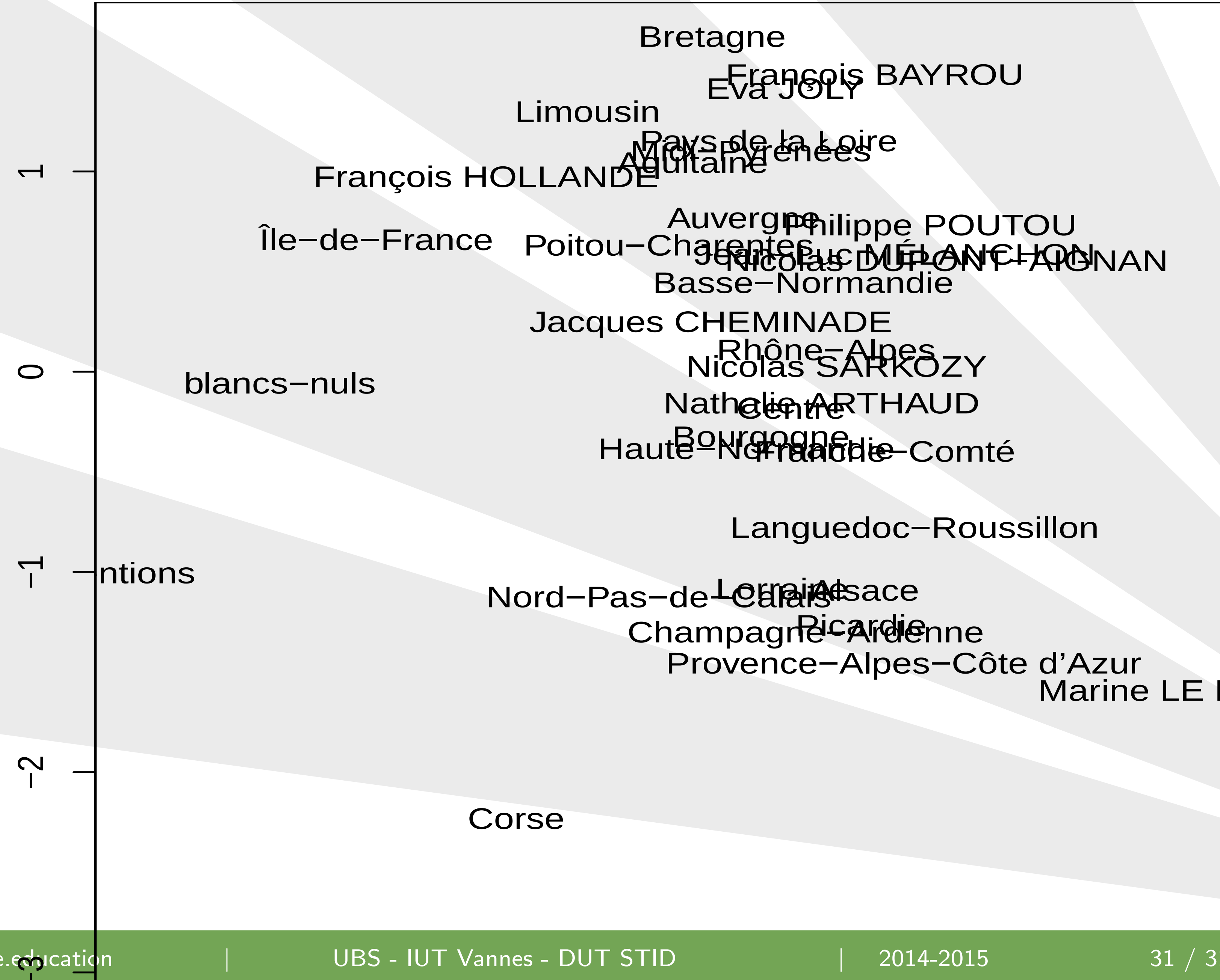
Représentation conjointe

- les combinaisons linéaires sont corrélées
- ★ mais pas colinéaires ! ! !
- on peut proposer une représentation simultanée sur un espace de projection commun

Représentation conjointe

centrée sur le cœur du nuage





1
0
−1
−2
−3
Bretagne
François BAYROU
Eva JOLY
Limousin
Pays de la Loire
Midi−Pyrénées
Aquitaine
François HOLLANDE
Auvergne
Philippe POUTOU
Île−de−France
Poitou−Charentes
Nicolas DUPONT−AIGNAN
Basse−Normandie
Jacques CHEMINADE
Rhône−Alpes
Nicolas SARKOZY
blancs−nuls
Nathalie ARTHAUD
Centre
Bourgogne
Haute−Normandie
Franche−Comté
Languedoc−Roussillon
ntions
Lorraine
Alsace
Nord−Pas−de−Calais
Picardie
Champagne−Ardenne
Provence−Alpes−Côte d'Azur
Marine LE
Corse