L'apport des modèles d'association à l'analyse multidimensionnelle de transformations temporelles

Illustration à travers l'homogamie en France (1969-2011)

Milan Bouchet-Valat
(CREST-LSQ, OSC-Sciences Po & INED)
nalimilan@club.fr

RT20, Congrès de l'AFS, Nantes, 2 septembre 2013

Plan

- Introduction :
  forces et limites de l'AC sur des tables d'homogamie
- De l'AC aux modèles d'association :
  similarités et différences
- Les évolutions temporelles de l'homogamie d'éducation en France entre 1969 et 2011 :
  le modèle RC(M)-L

Une table d'homogamie

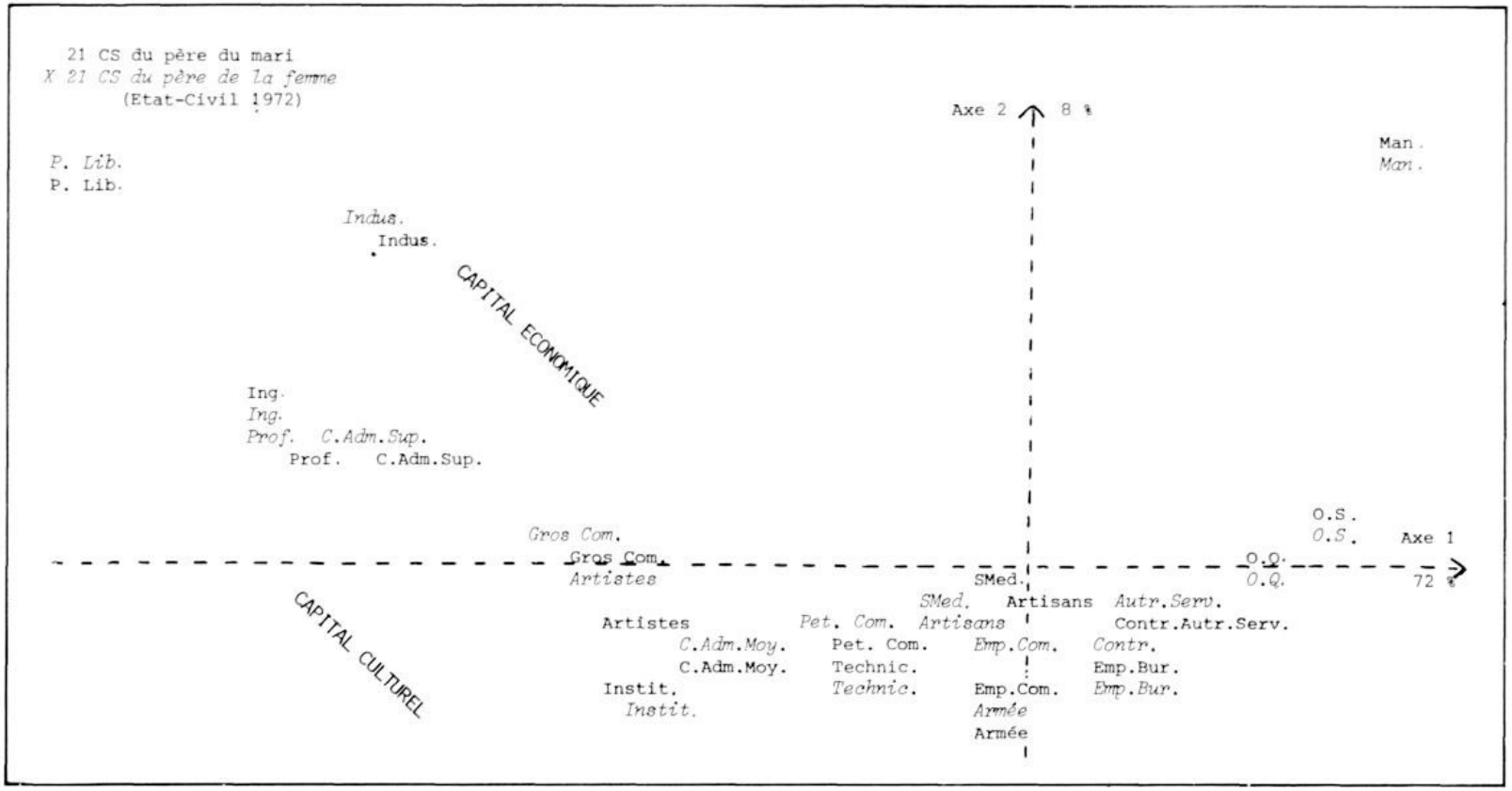
Couples cohabitants dans lesquels au moins l'un des conjoints a entre 30 et 59 ans (enquête Emploi 2011)

	Femmes						
Hommes	**Agri.**	**Ind.**	**Cadre**	**Interm.**	**Empl.**	**Ouvr.**	**Ens.**
Agriculteur	2,2	0,1	0,1	0,4	0,8	0,3	3,9
Indépendant	0,1	2,3	0,9	2	4,1	0,8	10,2
Cadre	0,1	0,7	5,9	6,6	6,3	0,7	20,3
Intermédiaire	0,1	0,7	2,2	7,2	10,4	2,3	22,9
Employé	0,0	0,2	0,6	2,2	6,3	1,1	10,4
Ouvrier	0,2	0,7	0,7	3,8	19,3	7,7	32,4
Ensemble	2,7	4,7	10,4	22,2	47,2	12,9	100

Une AC classique et ses limites

- Desrosières, « Marché matrimonial et structure des classes sociales », *ARSS*, 1978

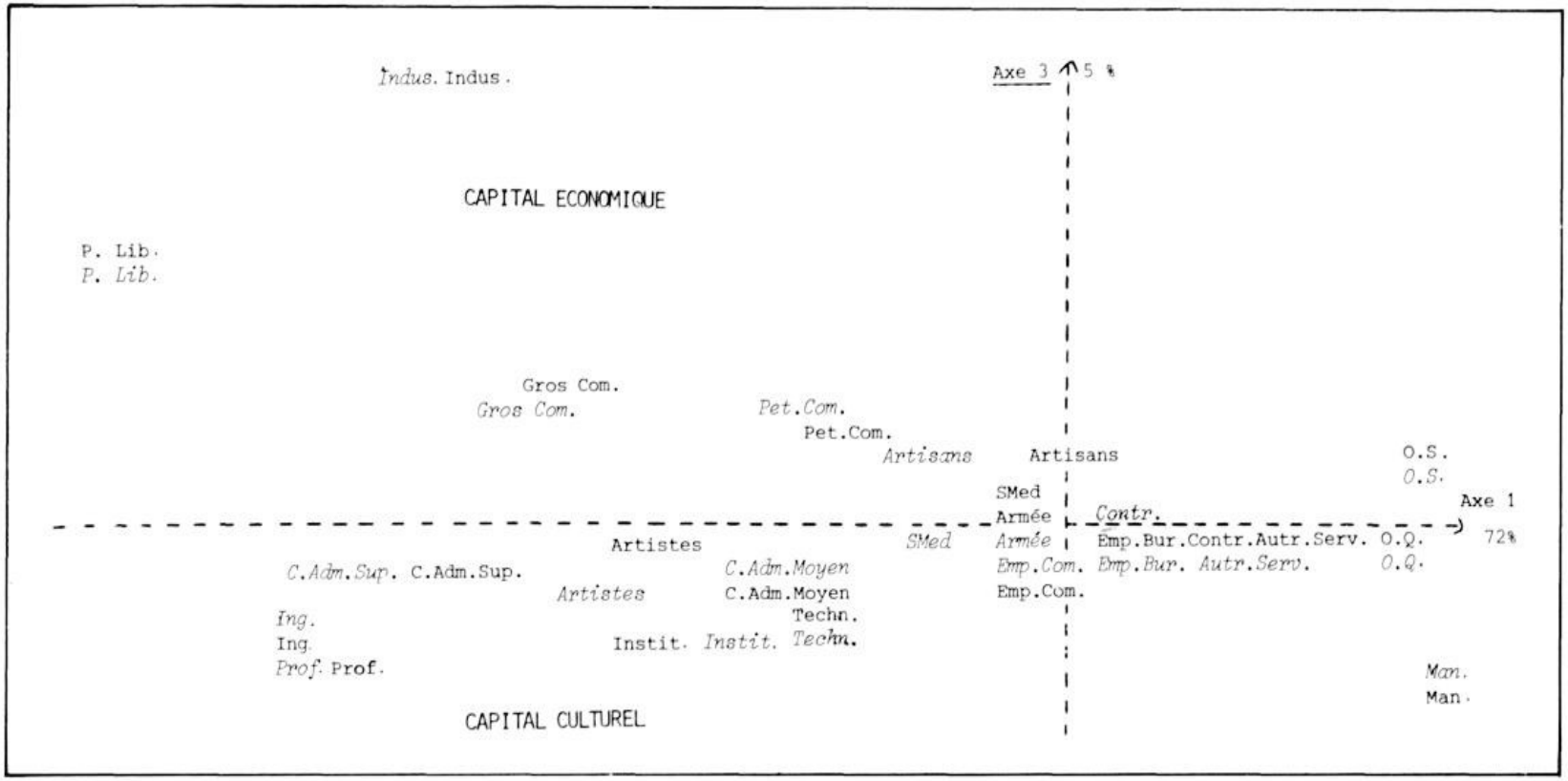
Graphique II. Les origines sociales des conjoints (axe 1 x axe 2)



Une AC classique et ses limites

- Desrosières, « Marché matrimonial et structure des classes sociales », *ARSS*, 1978

Graphique III. Les origines sociales des conjoints (axe 1 x axe 3)

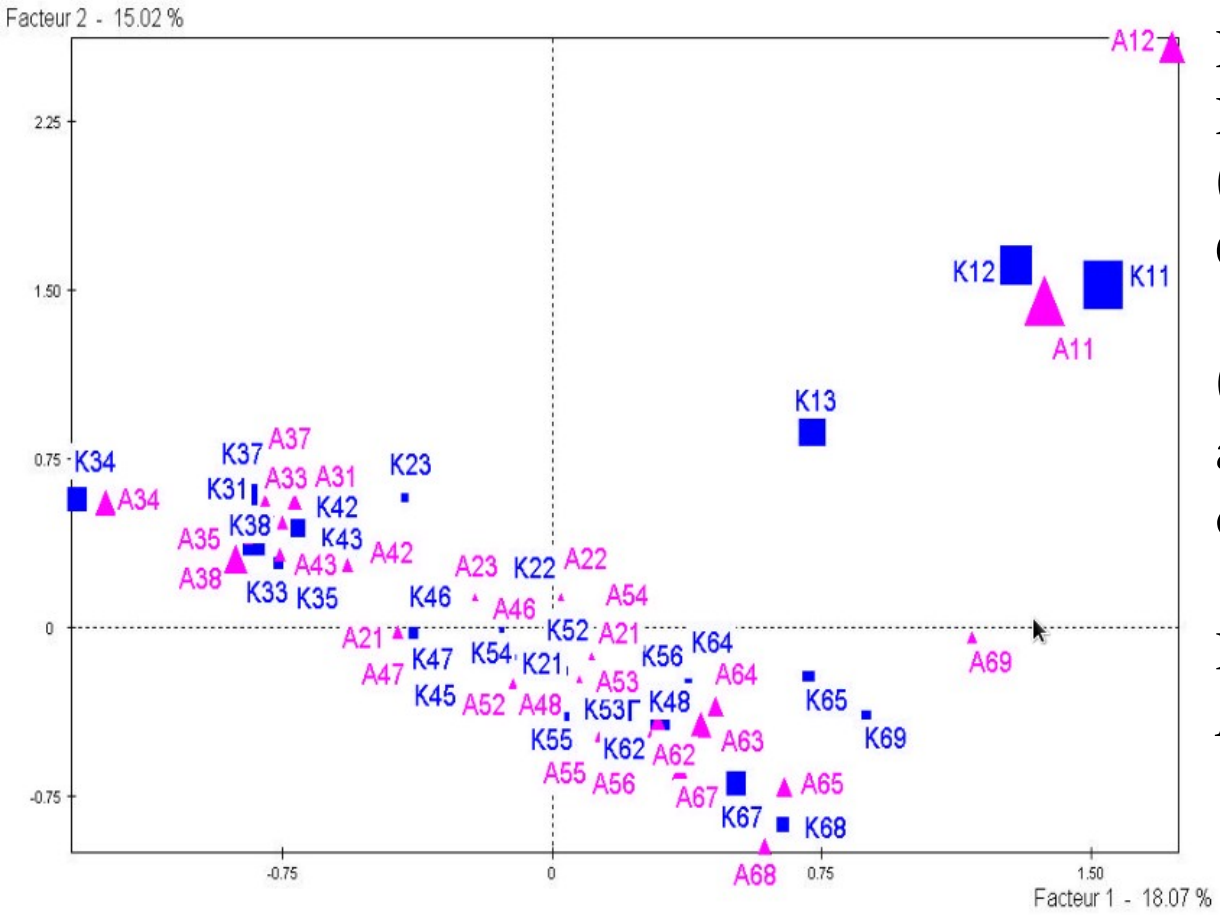


Une AC classique et ses limites

- Desrosières, « Marché matrimonial et structure des classes sociales », *ARSS*, 1978
- Une AC classique, mais pas parfaite :
  - Effet Guttman sur le deuxième axe
  - Deux points par catégorie, mais pas d'étude de ces différences : *« l'instrument utilisé ici ne permet pas de percevoir de telles dissymétries »*
  - Des exclus : agriculteurs, salariés agricoles, mineurs, patrons et marins-pêcheurs
  - Pas d'évolution temporelle

Un problème plus général

En conservant les agriculteurs :



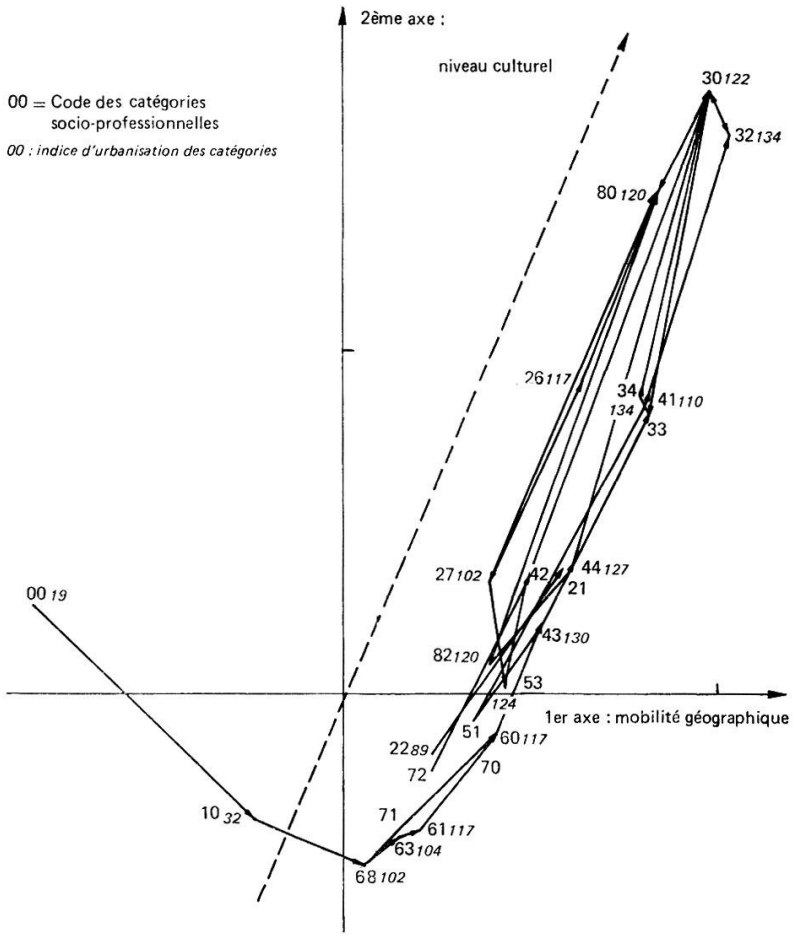
L'homophilie observée par Lemel & Cousteaux (Document de travail du CREST, 2004)

(Excluent ensuite les agriculteurs, puis utilisent des modèles RC.)

K : individu Kish
A : premier ami déclaré

Un problème plus général

En conservant les agriculteurs :



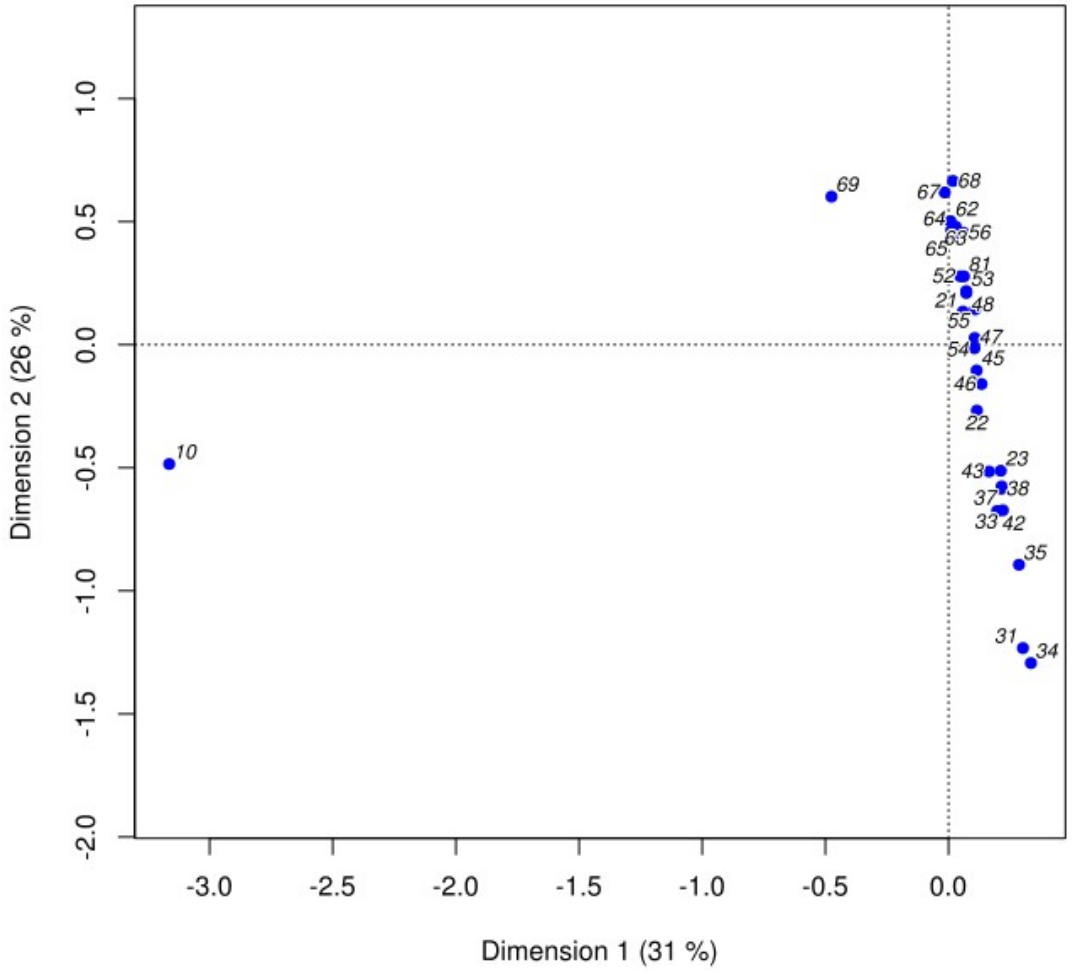
La mobilité sociale des hommes observée par Darbel (*Économie et statistique*, 1975)

Exclusion des mineurs, patrons et marins-pêcheurs.

Seuls les pères sont représentés (« *résultats identiques* » pour les fils).

Un problème plus général

En conservant les agriculteurs :



Nos propres données :
Enquêtes Emploi 2003-2010

84 393 couples cohabitants en 1re interrogation,
PCS niveau 2 (29 cat.) des conjoints.

Seuls les hommes sont représentés.

Améliorations possibles

- Traiter la diagonale de la table d'homogamie à part :
  - Séparer homogamie stricte et distance sociale pour représenter un *espace social*
  - Ne pas éliminer arbitrairement certains groupes
- Traiter des différences lignes/colonnes à part :
  - Dans l'AC, on néglige souvent les différences de position entre les points ligne et colonne d'une même catégorie
  - L'analyse cherche à identifier la structure sous-jacente qui est commune aux lignes et aux colonnes

→ Des solutions existent à la fois avec une AC modifiée, et avec les modèles d'association

Améliorations possibles

- Tester les hypothèses qui sont faites en silence :
  - Combien de dimensions peut/doit-on commenter ?
  - L'asymétrie lignes/colonnes est-elle significative ?
- Étudier les transformations temporelles :
  - Comment comparer rigoureusement des tables correspondant à des dates différentes ?
  - Comment mesurer l'ampleur des variations intervenues ?

→ C'est le grand apport des modèles d'association

De l'analyse des correspondances aux modèles d'association

Un peu de mathématiques

Deux traditions d'analyse

- Issus de traditions très différentes, AC et modèles d'association sont pourtant très proches
- Le modèle log-multiplicatif d'association lignes-colonnes (RC) a été proposé par Goodman en 1979
- Longue histoire commune entre AC, RC et modèles log-linéaires depuis :
  - comparaisons (Goodman 1986, 1991... ; Gilula & Haberman, 1988)
  - utilisations complémentaires (van der Heijden, Falguerolles et de Leeuw, 1989...)
  - discussions (réponse de Benzécri à Goodman, 1991)
- Il est possible d'unifier les deux approches dans un modèle plus général (Goodman, 1996)

Un peu de mathématiques : AC

- Point de départ : l'écart à l'indépendance (Pearson)

$$\psi_{ij} = \frac{P_{ij}}{P_{i.}P_{.j}} \quad et \quad \psi_{ij} - 1 = \frac{P_{ij} - P_{i.}P_{.j}}{P_{i.}P_{.j}} \quad (cf.\ Khi2\ par\ cellule)$$

- L'analyse des correspondances décompose l'écart à l'indépendance d'une cellule comme :

$$\psi_{ij} = \sum_{i=1}^{I} \psi_{ij} P_{i.} + \sum_{j=1}^{J} \psi_{ij} P_{.j} - \sum_{i=1}^{I}\sum_{j=1}^{J} \psi_{ij} P_{i.} P_{.j} + \sum_{m=1}^{M} \rho_m x_{im} y_{jm}$$

$$= 1 + 1 - 1 + \sum_{m=1}^{M} \rho_m x_{im} y_{jm} = 1 + \sum_{m=1}^{M} \rho_m x_{im} y_{jm}$$

$$ou\ encore \quad \psi_{ij} - 1 = \frac{P_{ij} - P_{i.}P_{.j}}{P_{i.}P_{.j}} = \sum_{m=1}^{M} \rho_m x_{im} y_{jm}$$

- D'où l'équation générale (formule de reconstitution) :

$$P_{ij} = P_{i.}P_{.j}\left(1 + \sum_{m}^{M} \rho_m x_{im} y_{jm}\right)$$

Un peu de mathématiques : RC

- Point de départ : modèles ***log***-linéaires

$$\psi_{ij} = \frac{P_{ij}}{P_{i.} P_{.j}} \quad et \quad R_{ij} = \log \psi_{ij} = \log P_{ij} - \log P_{i.} - \log P_{.j} \quad (cf.\ odds\ ratio)$$

- Les modèles d'association décomposent l'écart à l'indépendance d'une cellule comme :

$$R_{ij} = \sum_{i=1}^{I} R_{ij} P_{i.} + \sum_{j=1}^{J} R_{ij} P_{.j} - \sum_{j=1}^{I} \sum_{j=1}^{J} R_{ij} + \sum_{m=1}^{M} \lambda_m \mu_{im} \nu_{jm}$$

- D'où l'équation générale :

$$P_{ij} = P_{i.} P_{j.} \exp(R_{ij}) = \alpha_i \beta_j \exp\left(\sum_{m}^{M} \lambda_m \mu_{im} \nu_{jm}\right)$$

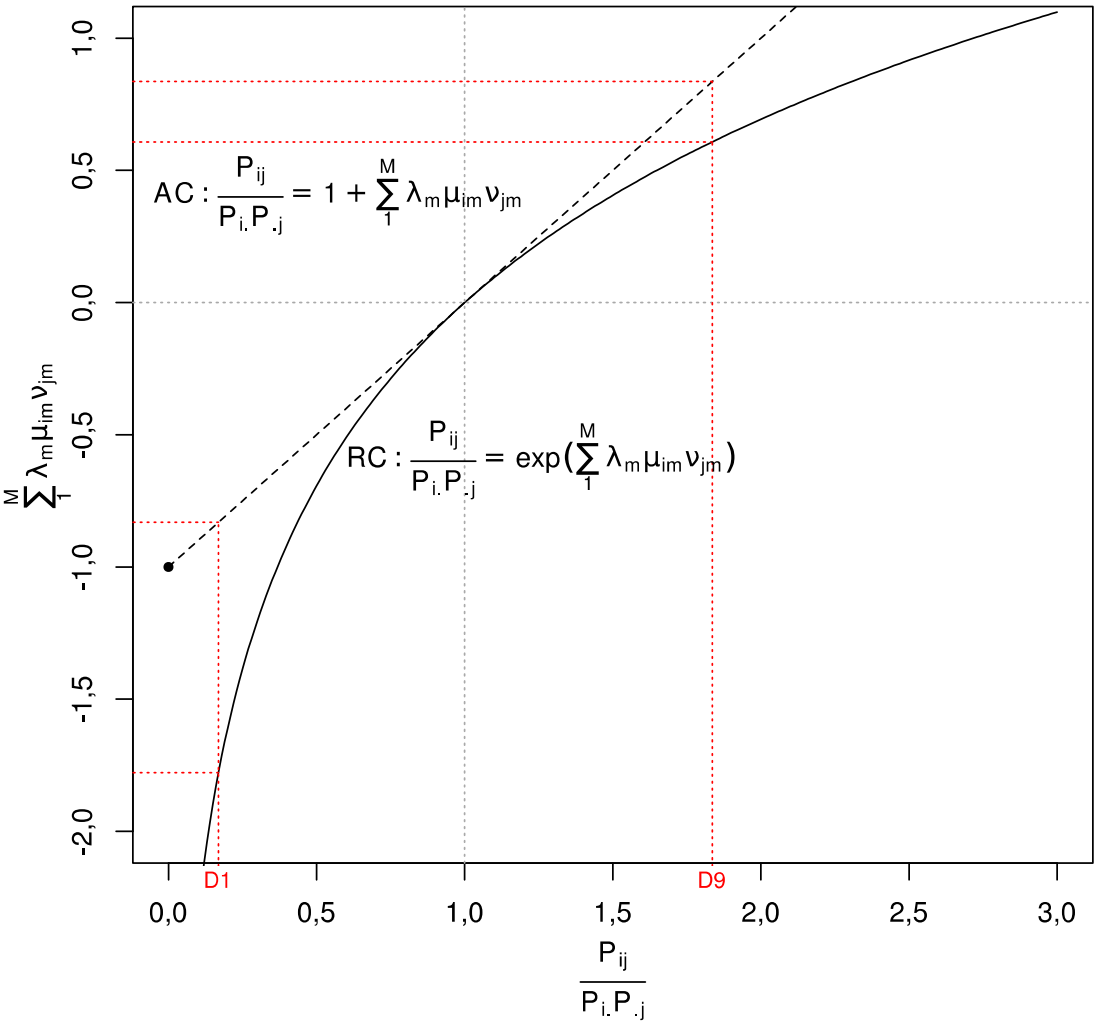
$(\alpha_i \ et \ \beta_j \ définis \ par \ la \ nécessité \ de \ respecter \ les \ marges \ P_{i.} \ et \ P_{.j})$

Deux modèles similaires…

$$P_{ij} = P_{i.} P_{.j} \left(1 + \sum_{m=1}^{M} \rho_m x_{im} y_{jm}\right) \quad vs. \quad P_{ij} = \alpha_i \beta_j \exp\left(\sum_{m=1}^{M} \lambda_m \mu_{im} \nu_{jm}\right)$$

- Décomposent les écarts à l'indépendance en plusieurs dimensions d'importance décroissante
- Scores centrés-réduits et non corrélés entre dimensions, une valeur singulière par dimension (décomposition en valeurs singulières)

...mais des différences significatives



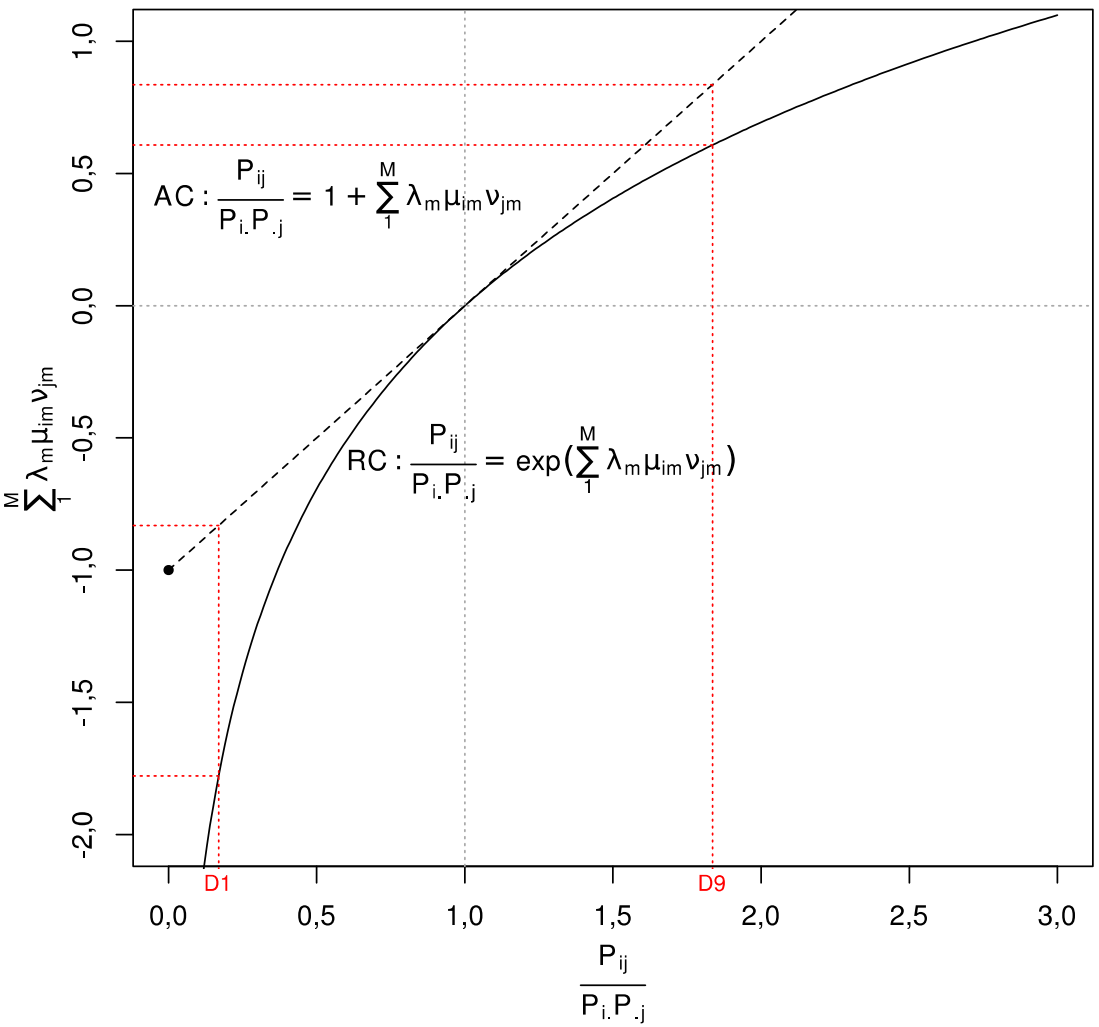
Pour x proche de 0,

$$1+x \approx \exp(x)$$

- Résultats similaires près de l'indépendance
- Différences pour de fortes associations

D1/D9 : déciles de cellules avec les sous-/sur-représentations les plus fortes

...mais des différences significatives

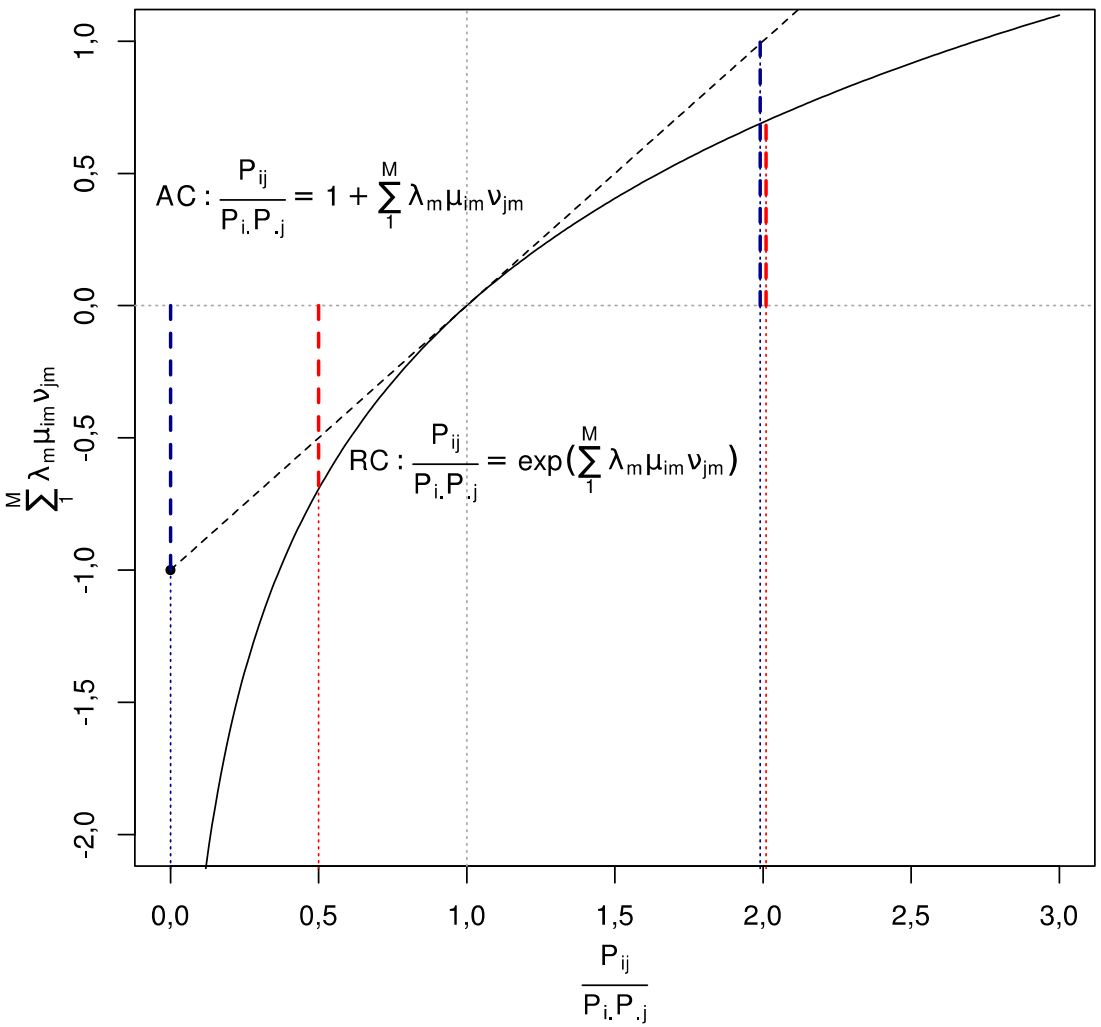


L'AC suit un modèle additif, RC suit un modèle multiplicatif :

- L'AC donne plus d'importance à une sur- qu'à une sous-représentation

D1/D9 : déciles de cellules avec les sous-/sur-représentations les plus fortes

...mais des différences significatives



L'AC suit un modèle additif, RC suit un modèle multiplicatif :

- En AC, 2 est le symétrique de 0
- En RC, 2 est le symétrique de 1/2

D1/D9 : déciles de cellules avec les sous-/sur-représentations les plus fortes

…mais des différences significatives

- Pas d'équivalence stricte entre les équations :
  - $P_{i.}P_{.j}$ dans l'AC est toujours égal au produit des marges
  - $\alpha_i\beta_j$ dans RC est un produit de paramètres estimés
  - Le modèle RC s'adapte plus aux données
- Modèles saturés et non saturés :
  - L'AC décrit exactement les données observées
  - Les modèles d'association ne reconstruisent qu'un nombre de dimensions limité (le plus souvent)
  - L'estimation par le maximum de vraisemblance permet de tester des hypothèses (nombre de dimensions…)

Le modèle RC(M)

- Présentation classique dans la tradition des modèles log-linéaires :

$$\log m_{hf} = \lambda + \lambda_h^H + \lambda_f^F + \sum_{k=1}^{M} \varphi_k \mu_{kh} \nu_{kf}$$

- m_{hf} correspond à l'effectif de la cellule (h, f) *prédit* par le modèle (et non pas à l'effectif observé)
- Les paramètres λ assurent l'adaptation à l'effectif total et aux marges de la table
- Les scores peuvent être normalisés avec ou sans pondération marginale

Analyser les évolutions temporelles, 1969-2011

Les modèles RC(M)-L

Période et données

- Données annuelles issues de cinq séries d'enquêtes Emploi : 1969-1974, 1975-1981, 1982-1989, 1990-2002, 2003-2011
- Couples cohabitants dont l'un des conjoints est âgé de 30 à 59 ans à la date de l'enquête
- L'éducation des conjoints détaillée en 11 niveaux
- Tables à trois dimensions : Homme – Femme – Temps

Le modèle RC(M)-L

$$\log m_{hft} = \lambda + \lambda_h^H + \lambda_f^F + \lambda_t^T + \lambda_{ht}^{HT} + \lambda_{ft}^{FT} + \sum_{m=1}^{M} \varphi_{mt} \mu_{mh} \nu_{mf}$$

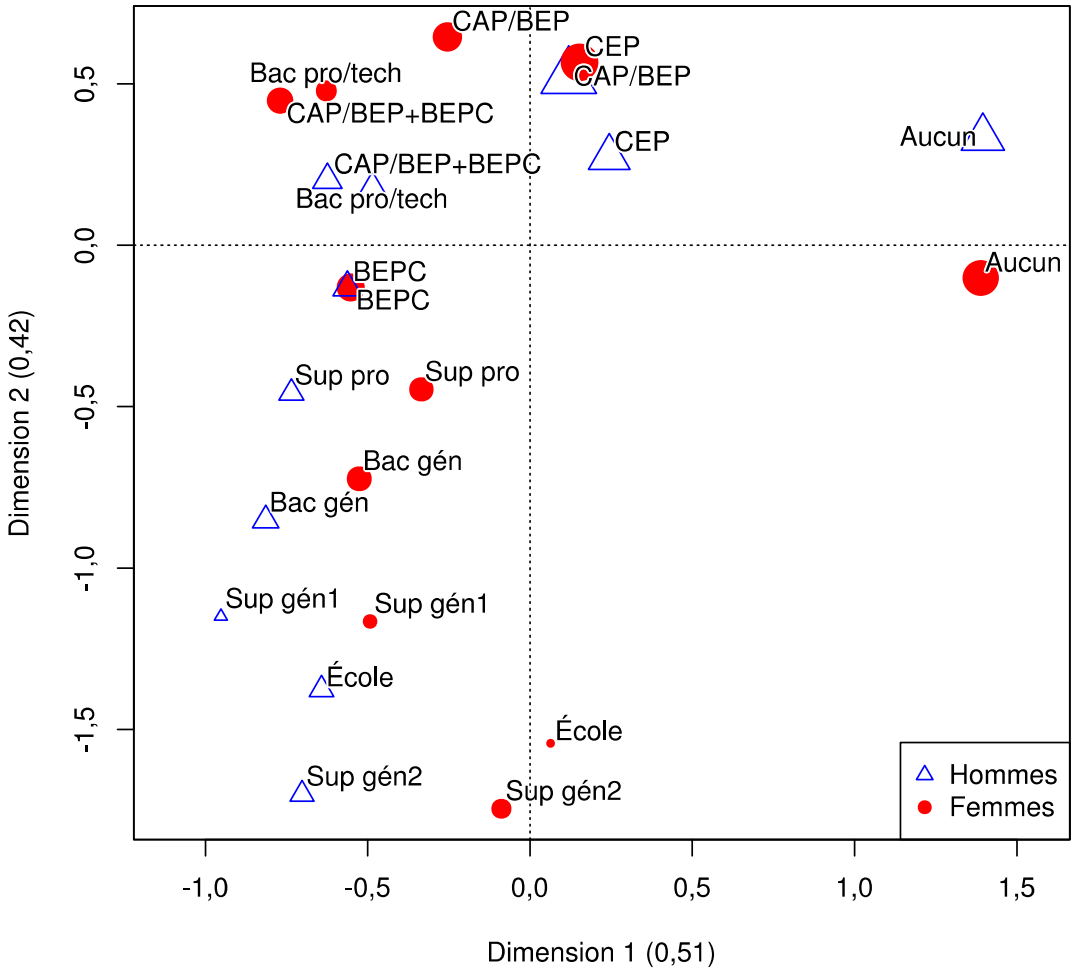
- Le modèle de base est maintenant l'indépendance conditionnelle (indépendance à chaque enquête)
- Une extension parcimonieuse du modèle RC(M) :
  - Seuls les paramètres d'association intrinsèque φ de chaque dimension varient selon le temps
- Scores indépendants des marges :
  - Pondération commune à toutes les années
- Nous excluons à chaque enquête la diagonale

Les modèles

	D. L.	Déviance	Δ (%)	BIC	AIC
Indép. conditionnelle	3738	93760	20,77	46730	86284
Association stable	3649	6372	5,04	-39538	-926
RC-L(1)	3678	9584	6,17	-36690	2228
RC-L(2)	3618	6777	5,13	-38743	-459
RC-L(3)	3558	5139	4,22	**-39627**	-1977
RC-L(1) linéaire	3718	9643	6,19	-37135	2207
RC-L(2) linéaire	3698	6927	5,24	-39599	-469
RC-L(3) linéaire	3678	5451	4,41	**-40824**	-1905

Tous les modèles ont la diagonale de la table exclue.

L'espace social des diplômes



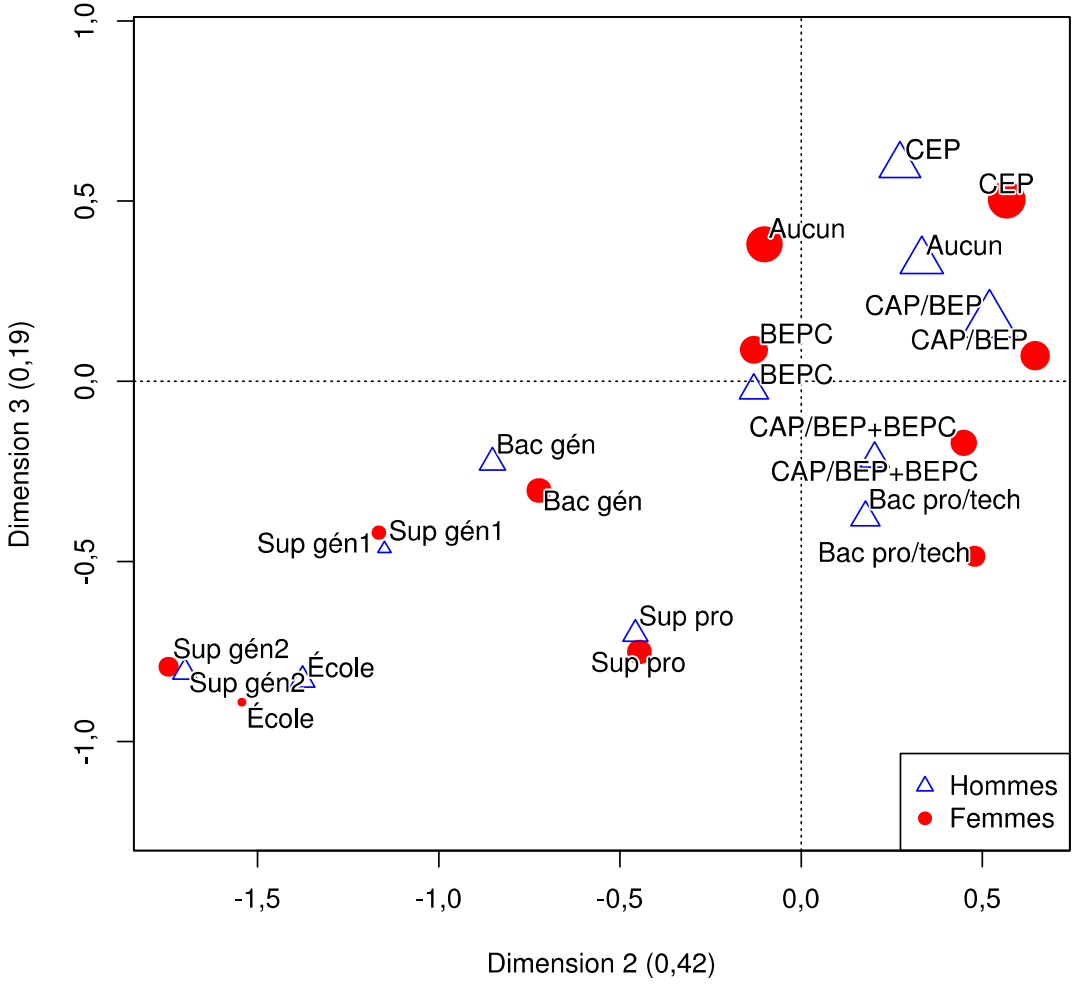
Modèle RC(3)-L
(diagonale exclue)

Dimension 1 :
Sans diplôme -> diplômés

Dimension 2 :
Ancienne échelle des diplômes

La taille des points reflète les effectifs des catégories dans l'ensemble de l'échantillon. L'intensité des dimensions est celle de 1969.

L'espace social des diplômes



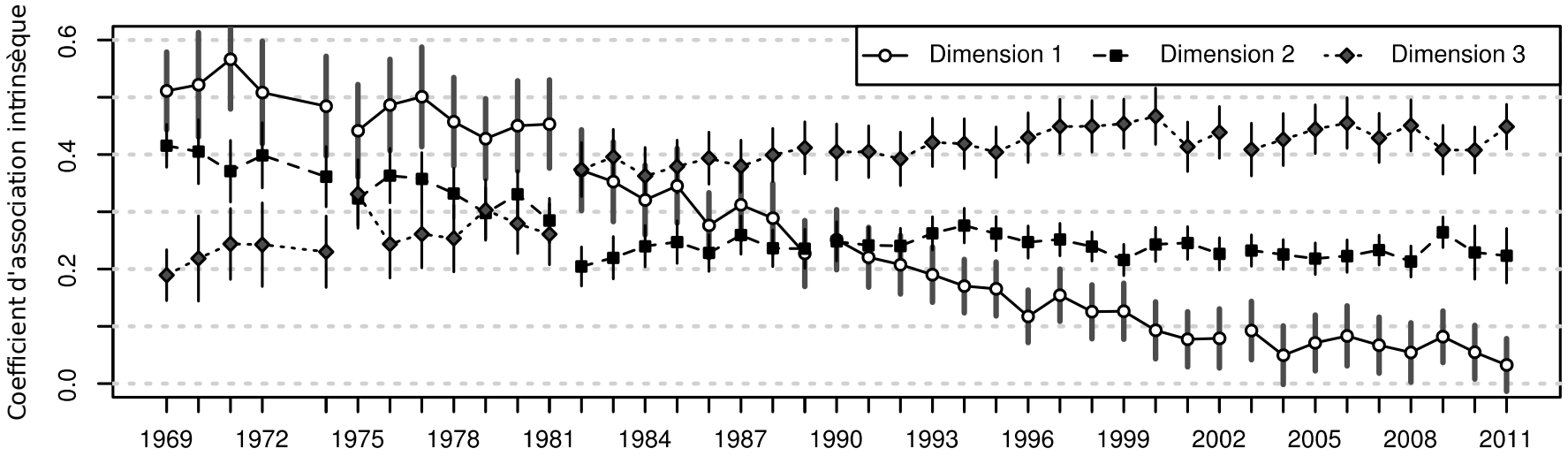
Modèle RC(3)-L
(diagonale exclue)

Dimension 2 :
Ancienne échelle des diplômes

Dimension 3 :
Nouvelle échelle des diplômes

La taille des points reflète les effectifs des catégories dans l'ensemble de l'échantillon. L'intensité des dimensions est celle de 1969.

L'homogamie d'éducation



Dimension 1 : Sans diplôme -> diplômés

Dimension 2 : Ancienne échelle des diplômes
Différences hommes-femmes marquées (notamment pour les non diplômés)
Forts écarts au-delà du BEPC

Dimension 3 : Nouvelle échelle des diplômes
Faibles différences hommes-femmes
Compression des écarts au-delà du BEPC
CEP en-dessous des sans-diplôme (artefact)

L'évolution de l'homogamie d'éducation

- L'homogamie d'éducation baisse globalement, mais ce résultat est en partie en trompe-l'œil :
  - La dimension qui décroît le plus nettement oppose diplômés et non-diplômés, cette dernière catégorie perdant une bonne part de ses effectifs du fait de la massification scolaire
  - L'homogamie suivant l'échelle des diplômes reste stable ou augmente légèrement
- Ces modèles permettent de décrire l'évolution des dimensions plus que la structure une année donnée

Extensions possibles

- Étudier spécifiquement les asymétries hommes-femmes
  - Modèle anti-symétrique (Van der Heijden & Mooijaart, 1995)
- Supposer qu'il existe sur chaque dimension un état de départ, un état d'arrivée, et pour chaque année une évolution entre ces deux états
  - Nouveau modèle « de transition »

Paquet R « logmult »

- Fondé sur le paquet gnm (Turner & Firth, 2007)
- Estimer en une commande une variété de modèles d'association :
  - Modèles RC(M) et RC(M)-L
  - Deux modèles d'association anti-symétrique
  - Modèle de transition
  - Variantes avec diagonale exclue ou non
  - Intervalles de confiance par jackknife et bootstrap
- Graphiques de qualité grâce à l'infrastructure R (inédit !)

Bibliographie

- Darbel, Alain (1975), « L'évolution récente de la mobilité sociale », *Économie et statistique*, vol. 71 (1), p. 3-22.
- Desrosières, Alain (1978), « Marché matrimonial et structure des classes sociales », *Actes de la recherche en sciences sociales*, vol. 20 (1), p. 97-107.
- Gilula, Zvi et Shelby J. Haberman (1988), « The Analysis of Multivariate Contingency Tables by Restricted Canonical and Restricted Association Models », *Journal of the American Statistical Association*, vol. 83 (403), p. 760-771.
- Goodman, Leo A. (1979), « Simple Models for the Analysis of Association in Cross-Classifications having Ordered Categories », *Journal of the American Statistical Association*, vol. 74 (367), p. 537-552.
- Goodman, Leo A. (1986), « Some Useful Extensions of the Usual Correspondence Analysis Approach and the Usual Log-Linear Models Approach in the Analysis of Contingency Tables », *International Statistical Review*, vol. 54 (3), p. 243-270.
- Goodman, Leo A. (1991), « Measures, Models, and Graphical Displays in the Analysis of Cross-Classified Data », *Journal of the American Statistical Association,* vol. 86 (416), p. 1085-1111 [voir aussi les réponses dont celle de Jean-Paul Benzecri].

Bibliographie

- Goodman, Leo A. (1996), « A Single General Method for the Analysis of Cross-Classified Data: Reconciliation and Synthesis of Some Methods of Pearson, Yule, and Fisher, and Also Some Methods of Correspondence Analysis and Association Analysis », *Journal of the American Statistical Association*, vol. 91 (433), p. 408-428.
- Lemel, Yannick et Anne-Sophie Cousteaux (2004), « Etude de l'homophilie socioprofessionnelle à travers l'enquête contacts », Document de travail n°5, Centre de Recherche en Économie et Statistique.
- Van der Heijden, Peter G. M., Antoine de Falguerolles et Jan de Leeuw (1989), « A Combined Approach to Contingency Table Analysis Using Correspondence Analysis and Log-Linear Analysis », *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 38 (2), p. 249-292.
- Van der Heijden, Peter G. M. et Ab Mooijaart (1995), « Some New Log-Bilinear Models for the Analysis of Asymmetry in a Square Contingency Table », *Sociological Methods and Research*, vol. 24, p. 7-29.
- Wong, Raymond Sin-Kwok (2010), *Association models*, Thousand Oaks (California), Sage, Quantitative Applications in the Social Sciences.