Extraction d'informations stratégiques par Analyse en Composantes Principales

Bernard DOUSSET
IRIT/ SIG, Université Paul Sabatier,
118 route de Narbonne, 31062 Toulouse cedex 04
dousset@irit.fr

1 Introduction

A la fin des années 60, ces méthodes ont été largement diffusées par l'école française d'analyse de données et notamment par J-P Benzécri pour analyser aussi bien des données factuelles que textuelles dans un cadre débordant largement du domaine scientifique et technique. Le but poursuivi était de simplifier l'espace informationnel afin d'en déduire une représentation graphique aussi fidèle que possible et compréhensible de façon intuitive. Notre contribution a essentiellement porté sur les points suivants :

- Pouvoir juger graphiquement de la qualité de l'analyse (nombre d'axes à prendre en compte, étude de la distribution des données et de leurs corrélations),
- Proposer des cartes classiques en 2D avec les principales projections,
- Proposer des cartes en 3D puis en 4D (par l'utilisation de niveaux de gris), qui permettent de visualiser un part nettement plus importante de l'information,
- Rendre ces cartes interactives (rotations, zooms, sélections, éditions, …),
- Gérer la coopération de ces cartes avec les autres méthodes via le réseau,
- Animer ces cartes afin d'en déceler visuellement les caractéristiques essentielles,
- Permettre le choix des axes et introduire la notion de glissement d'axes,
- Permettre la sélection d'items et des documents qui leurs sont attachés (recherche d'information ciblée),
- Pouvoir extraire des éléments ou des groupes remarquables (connecteurs, classes, signaux émergents, …),
- Proposer des cartes dynamiques permettant de visualiser l'évolution sous forme de trajectoires,
- Permettre de visualiser l'évolution relative (occurrences anormales).

2 Analyse en composantes principales (ACP)

Elle s'applique aux données quantitatives et éventuellement aux matrices issues du qualitatif comme celles de contingence et de cooccurrence (notamment dans le cas de l'analyse relationnelle entre acteurs). Le nuage des individus (lignes) est représenté dans l'espace des variables (colonnes). Le but est de trouver le meilleur modèle réduit à n variables synthétiques qui représente au mieux l'ensemble des informations de la matrice. Une fois générée la matrice de variance – covariance entre les variables initiales il suffit de rechercher les vecteurs propres associés aux valeurs propres de plus forts modules de cette matrice pour déterminer les composantes principales de ce modèle optimal.
La première composante représente en fait l'axe de rotation autour duquel le nuage de points a la plus faible inertie, donc celui qui explique le mieux la dispersion des individus. En étendant cette démarche au sous espace orthogonal, on trouve la seconde composante et ainsi de suite.

Dans l'histogramme ci-dessous, généré après chaque ACP, nous pouvons constater la décroissance des modules des valeurs propres ainsi que l'augmentation du taux d'information visualisée en fonction du nombre de composantes prises en compte.

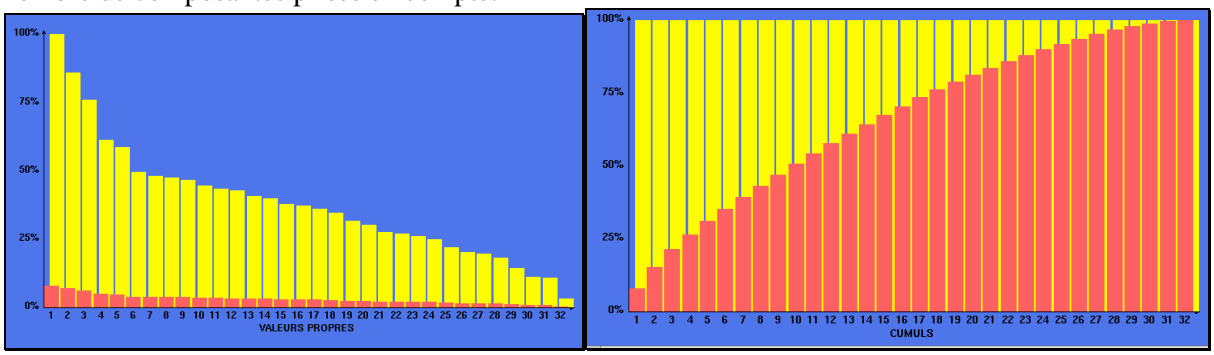

Figure 1. : Valeurs propres triées et taux d'information en fonction du nombre de composantes.

Le module d'une valeur propre rapporté à la somme des modules représente, en fait, la part d'information portée par la composante qui lui est associée, la somme des modules des premières valeurs propres détermine donc le taux d'information visualisé dans une carte factorielle. Ainsi, pour une vue 4D, le taux d'information manipulé sera égal au rapport de la somme des 4 premiers modules à la somme totale. C'est bien entendu mieux qu'en 2 ou 3D. L'histogramme précédent permet de savoir jusqu'où aller dans l'analyse descendante (vers les composantes secondaires), afin de parcourir la majeure partie de l'information disponible.
L'ensemble des individus projetés dans ce sous espace particulier génère un nuage dont il faut étudier la distribution. Pour cela, une première carte factorielle est proposée, elle est basée sur les coordonnées des individus dans l'espace ordonné des nouvelles variables (composantes principales). Pour pouvoir expliquer certains phénomènes observés, il est possible de générer parallèlement, dans une seconde carte, le cercle (sphère en 3D et hyper sphère en 4D) des corrélations des anciennes variables par rapport aux nouvelles. Comme dans un cas il s'agit de coordonnées et dans l'autre de corrélations, les deux cartes obtenues ne sont pas superposables, mais une comparaison sous le même azimut de visualisation permet de déduire les causalités des ressemblances ou des dissemblances constatées. Les quatre fonctions d'exploration suivantes ont été implémentées dans Tétralogie au niveau des cartes en 4D :

- Rotations dans les plans de l'espace 4D définis par les axes :
  1 et 2 (plan de l'écran),
  1 et 3 (plan horizontal orthogonal à l'écran),
  2 et 3 (plan vertical orthogonal à l'écran) comme en 3D,
  1 et 4 (axe horizontal et niveaux de gris),
  2 et 4 (axe vertical et niveaux de gris)
  3 et 4 (axe orthogonal à l'écran et niveaux de gris) spécifiques au 4D.
- Glissement d'axes :
  {1, 2, 3, 4} => {2, 3, 4, 5}
  {2, 3, 4, 5} => {3, 4, 5, 6}
  etc…
- Sélection interactive des axes à visualisés :
  {2, 4, 5, 8} par exemple.
- Zooms :
  avant (volontairement limité) lié au rayon perçu du nuage,
  arrière (sans limite).

Ces modifications de l'azimut de visualisation sont exportables d'une carte à l'autre et d'un ordinateur connecté au réseau à l'autre. Elles sont un des éléments essentiels de notre processus de découverte de connaissance aussi bien individuel que collectif.

Ci-dessous, deux cartes synchrones dans lesquelles les individus sont les documents et les variables les auteurs qui les ont signés :

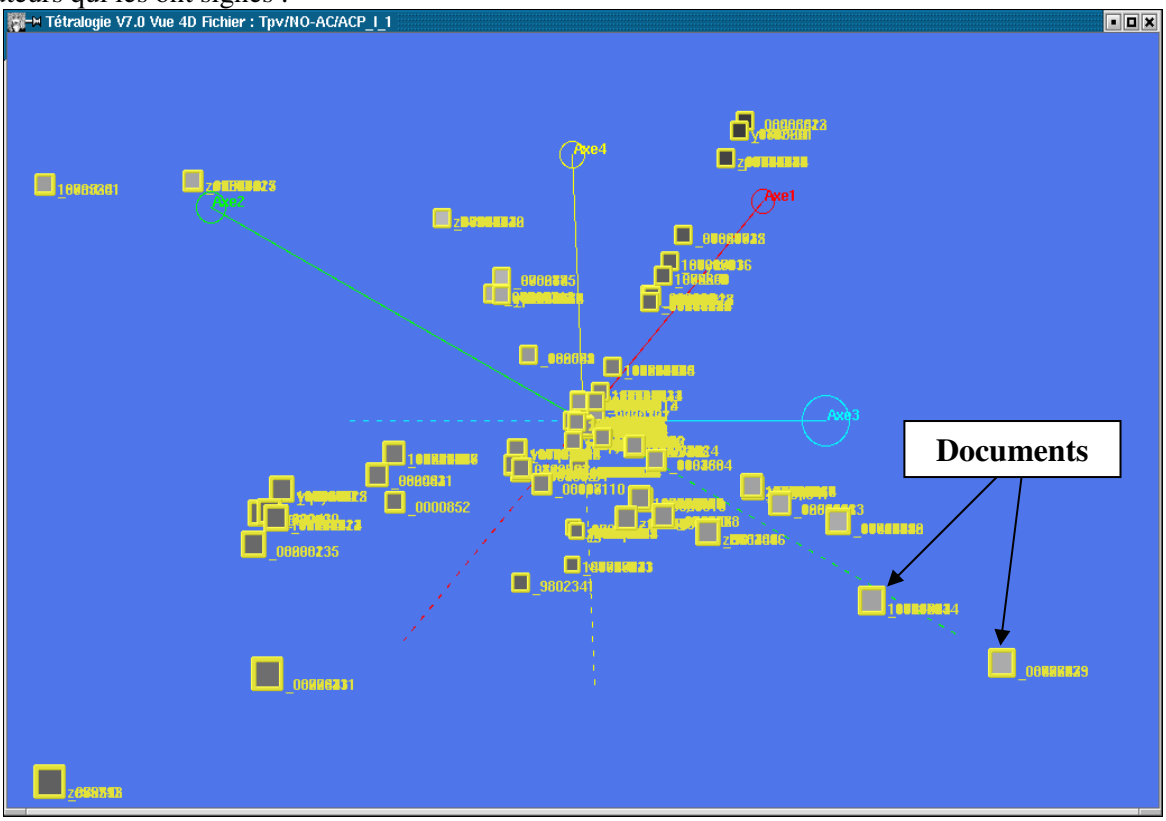

Figure 2. : Carte 4D des coordonnées des documents dans l'espace des auteurs.

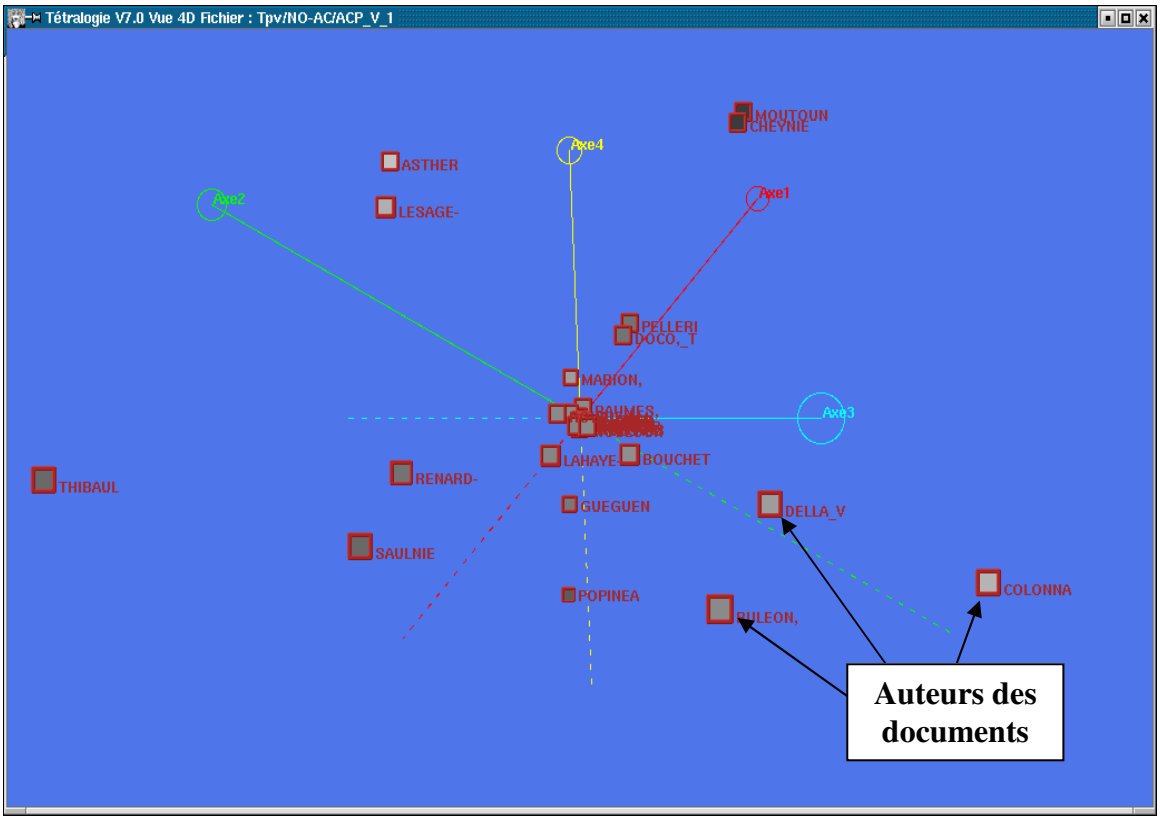

Figure 3. : Hyper sphère 4D des corrélations des auteurs.

A remarquer, qu'il est facile de mettre en relation les documents et les auteurs (ou les équipes) qui les ont signés, car les deux cartes sont présentées sous le même azimut et les éléments liés se trouvent donc exactement dans le même secteur. Dans l'exemple ci-dessus, les deux documents en bas à droite (Figure 63 : carte des coordonnées des documents dans l'espace des auteurs) sont liés aux trois auteurs présents dans la même zone du cercle (hyper sphère) des corrélations (figure 64).
Mais dans la majorité des cas, cette méthode de base n'est pas parfaitement adaptée à notre problématique qui manipule plus volontiers des variables essentiellement qualitatives. En effet, elle privilégie trop les signaux forts et ne permet donc pas le mise en évidence des signaux faible, des émergences et de certaines nuances qui sont, bien entendu, les éléments recherchés par tout processus de veille.

3 Analyse en composantes principales réduite (ACPr)

Comme la nature et la dispersion des variables sont parfois très hétérogènes, une normalisation de celles-ci est alors nécessaire pour obtenir des cartes lisibles et sur lesquelles les variables ont toutes des rôles similaires. Les variables sont alors réduites par normalisation (division par la norme de chaque vecteur colonne), ce qui a tendance à arrondir le nuage et donc à générer des valeurs propres de plus faibles modules. La matrice à diagonaliser est alors celle des corrélations (diagonale unitaire) et non plus la matrice de variance-covariance.
Ci-dessous, l'observatoire d'une ACP réduite appliquée à une structure de recherche croisant l'ensemble des auteurs d'une équipe avec ses auteurs principaux:

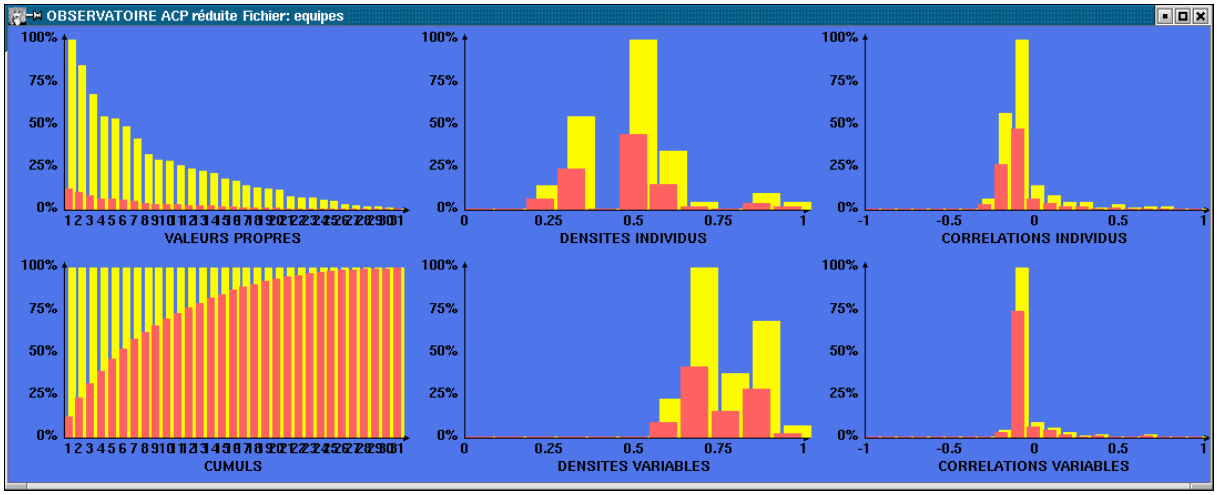

Figure 4. : Observatoire de la qualité d'une analyse multidimensionnelle.

Dans cet observatoire de la qualité d'une analyse multidimensionnelle nous trouverons les éléments suivants :

- l'histogramme des modules des valeurs propres trié par ordre décroissant superposé à celui rapporté à la plus forte d'entre elles,
- l'histogramme des cumuls de ces modules rapporté à leur somme globale et exprimant le taux d'information porté par les *n* premiers axes principaux,
- la répartition de la densité relative (taux de valeurs nulles) des individus en fonction du plus dense d'entre eux,
- la répartition de la densité relative des variables en fonction de la plus dense,
- la répartition des corrélations entre les lignes (matrices de contingence ou de cooccurrence),
- la répartition des corrélations entre variables.

Cet observatoire de la qualité d'une analyse, va nous permettre de déterminer très rapidement :

- la profondeur de l'analyse : nombre de composantes à prendre en compte pour atteindre un taux d'information supérieur à 80% ou 90%,

- donc le type de carte factorielle à utiliser : 2D, 3D, 4D, 4D en utilisant la fonction de glissement d'axes,
- le degré d'hétérogénéité des individus et des variables (par densité relative) donc la distribution de la population analysée,
- le nombre de classes de densité (ici 3 pour les individus et une seule pour les variables),
- les fortes corrélations positives ou négatives qui signalent la présence de liens remarquables ou d'exclusions qu'il va falloir détecter et expliquer (ici corrélations à 0,8-0,7 puis 0,5 pour les individus et à 0,7 pour les variables),
- la présence éventuelle de plusieurs valeurs propres de module maximum qui nous indiquerait que la matrice n'est pas connexe et qu'il faut donc analyser séparément chacune de ses classes connexes.

Dans la carte factorielle correspondante que nous présentons ci-dessous, nous distinguons plusieurs équipes dont les caractéristiques sont les suivantes :

- les leaders se trouvent implantés vers l'extérieur (Asther, Colonna, Thibault), car la réduction des variables ne change pas la nature quantitative de l'analyse,
- les collaborateurs sont plus centrés (moins de publications), mais ont la même orientation que le ou les leaders de leur équipe,
- les liens entre équipes s'effectuent par l'intermédiaire d'auteurs ayant une position médiane (Lesage, Buléon),
- la taille et la coloration interne des icônes permettent de mieux différencier les classes.

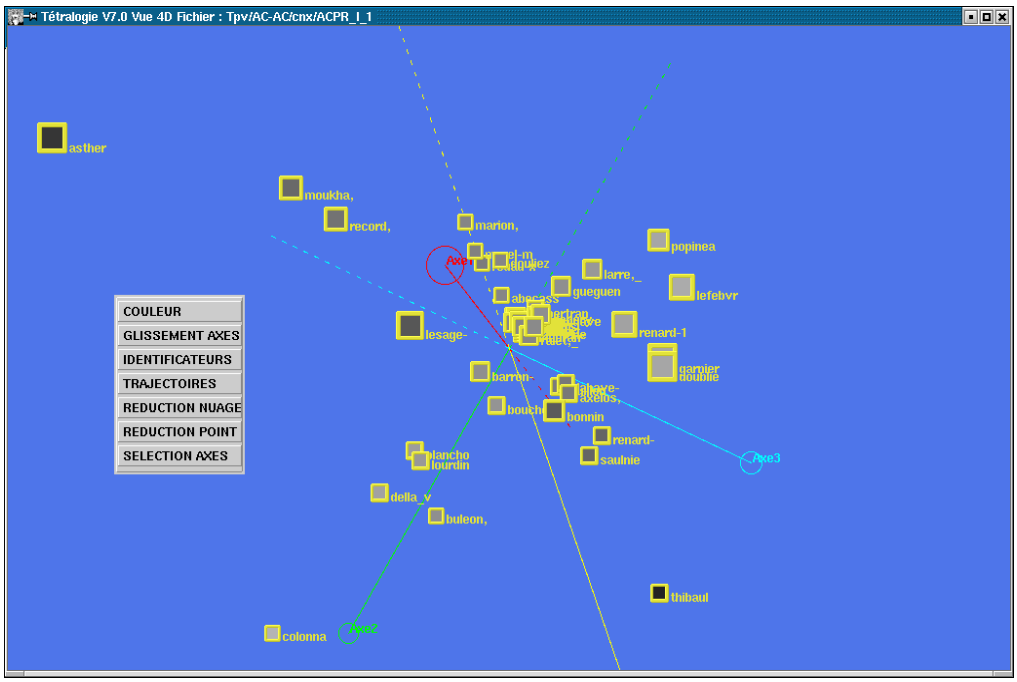

Figure 5. : Carte factorielle des coordonnées des individus d'une ACPr.

4 Conclusion

Cette méthode d'analyse des matrices de cooccurrence est essentiellement utile pour les croisements entre acteurs. Elle privilégie les signaux forts. Les leaders se trouvent à l'extérieur du nuage les connecteurs entre deux groupes. Dans les autres cas (matrices asymétriques), son principal inconvénient est lié à la nécessité d'utiliser deux cartes aux métriques différentes, elle est donc réservée aux spécialistes de l'analyse de données.