Validation des visualisations par axes principaux de données numériques et textuelles.

Ludovic Lebart

CNRS-ENST
lebart@enst.fr
http://www.lebart.org

Résumé. Parmi les outils de visualisation de données multidimensionnelles figurent d'une part les méthodes fondées sur la décomposition aux valeurs singulières, et d'autre part les méthodes de classification, incluant les cartes auto-organisées de Kohonen. Comment valider ces visualisations ? On présente sept procédures de validation par bootstrap qui dépendent des données, des hypothèses, des outils : a) le bootstrap partiel, qui considère les réplications comme des variables supplémentaires; b) le bootstrap total de type 1, qui réanalyse les réplications avec changements éventuels de signes des axes; c) le bootstrap total de type 2 qui corrige aussi les interversions d'axes; d) le bootstrap total de type 3, sur lequel on insistera, qui corrige les réplications par rotations procrustéenne; e) le bootstrap spécifique (cas des hiérarchies d'individus statistiques et des données textuelles). f) le bootstrap sur variables. g) les extensions des procédures précédentes à certaines cartes auto-organisées.

1 Introduction

On veut montrer brièvement les divers degrés d'exigence (vis-à-vis des résultats) que l'on peut avoir lorsque l'on procède à une analyse en axes principaux. Ces degrés correspondent à des modalités d'usage du bootstrap (Diaconis et Efron, 1983; Efron et Tibshirani, 1993). On examinera successivement le bootstrap *partiel* (section 2), trois types de bootstrap dit *total* (section 3), d'autres formes plus spécifiques de bootstrap (section 4). On revient ensuite sur les subtilités du bootstrap total de type 3 (section 5). On illustrera ces propos par une étape de travail extraite d'une analyse en composante principales (ACP).

2 Bootstrap partiel

Les axes principaux calculés à partir des données originales, non perturbées, jouent un rôle privilégié (en ACP, par exemple, la matrice des corrélations initiale $\mathbf{C}$ est en effet l'espérance mathématique des matrices $\mathbf{C}_k$ « perturbées » par la réplication k). Pourquoi calculer des sous-espaces de représentation prenant en compte des perturbations, et donc moins exacts que le sous-espace calculé sur les données initiales? La variabilité bootstrap

s'observe mieux sur le repère fixe initial, non perturbé. C'est l'option qui sera prise dans la suite de cet article. La technique de bootstrap que l'on appellera *bootstrap partiel* (sans recalcul des valeurs propres) proposée notamment par Greenacre (1984) dans le cadre de l'analyse des correspondances, répond à plusieurs des préoccupations des utilisateurs dans le cas de l'analyse en composantes principales[1]. Une réplication consiste en un tirage avec remise des *n* individus (vecteurs-observations), suivi du positionnement des *p* nouvelles variables ainsi obtenues en "variables supplémentaires" sur les *q* premiers axes de l'analyse de base. Les procédures décrites ci-dessus peuvent être mises en oeuvre avec un programme classique de projection d'éléments supplémentaires. On calcule donc les réplications de ce coefficient, ce qui revient à repondérer les individus avec les "poids Bootstrap" 0, 1, 2, ... qui caractérisent un tirage sans remise. On obtient, comme sous-produit, des réplications de la variance sur l'axe, qui sont évidemment distinctes de ce que seraient des réplications des valeurs propres.

Les *s* réplications étant projetées sur un repère commun (celui de l'analyse initiale), on caractérisera graphiquement la dispersion des réplications d'une variable donnée soit par l'enveloppe convexe de l'ensemble de ses réplications, soit par un ellipsoïde d'ajustement du nuage des réplications, qui résultera en fait d'une petite analyse en composantes principales de ce dernier nuage. L'enveloppe convexe a l'avantage de l'exhaustivité (toutes les réplications sans exception sont enveloppées), l'ellipsoïde a l'avantage de prendre en compte la densité du nuage des réplications, et d'être moins sensible à d'éventuelles rares réplications aberrantes.

3 Bootstrap total

Le *bootstrap total* consiste à réaliser autant d'analyses en axes principaux qu'il y a de réplications. Mais le système d'axes n'est plus le même d'une analyse à une autre. Il peut y avoir des changements de signes (les axes factoriels sont définis aux signes près), des interversions d'axes, des rotations d'axes[2]. Il faut donc procéder à une série de transformations afin de retrouver des axes homologues au cours des diagonalisations successives des *s* matrices de corrélation répliquées $\mathbf{C}_k$ ($\mathbf{C}_k$ correspond à la *k-ème* réplication).

3.1 Bootstrap total de type 1

C'est une épreuve sévère, très pessimiste : simple changement (éventuel) de signes des axes homologues pour les réplications. Il s'agit seulement de remédier au fait que les axes sont définis au signe près. Un simple produit scalaire entre axes originaux et axes répliqués de mêmes rangs permet de rectifier le signe de ces derniers. Le bootstrap total de type 1 ignore les possibles interversions d'axes et rotations d'axes. Il permet de valider des structures stables et robustes. Chaque réplication doit produire les axes initiaux avec les mêmes rangs (ordre des valeurs propres).

---

[1] Voir aussi, pour une discussion du bootstrap partiel en analyse en composantes principales, Chateau et Lebart (1996).

[2] Cf. Milan et Whittaker (1995).

3.2 Bootstrap total de type 2

C'est une épreuve assez sévère, plutôt pessimiste : il y a changements de signes et éventuellement correction des interversions d'axes. Les axes répliqués sont affectés (séquentiellement, sans remise en cause d'affectations antérieures) du rang des axes originaux avec lesquels ils sont les plus corrélés en valeur absolue. Puis on procède à un éventuel changement de signe des axes, comme en bootstrap de type 1. Le bootstrap total de type 2 est idéal si on veut valider des axes, c'est-à-dire des dimensions cachées, sans attacher une importance particulière aux rangs de celles-ci.

3.3 Bootstrap total de type 3[3]

C'est une épreuve plutôt laxiste si on s'intéresse à la stabilité des axes, mais apte à décrire la stabilité des sous-espaces de dimension supérieure à 1: une rotation dite procrustéenne (cf. Gower et Dijksterhuis, 2004) permet de rapprocher de façon optimale les systèmes d'axes répliqués et les systèmes d'axes initiaux. Le bootstrap de type 3 permet de valider globalement un sous-espace engendré par les axes principaux correspondant aux premières valeurs propres. Comme le bootstrap partiel, le bootstrap total de type 3 peut être qualifié de laxiste par les utilisateurs qui s'intéressent à l'individualité des axes, et pas seulement aux sous-espaces engendrés par plusieurs axes consécutifs (cf. section 5).

4 Autres types de bootstrap

On va mentionner trois techniques plus spécifiques. Comme les techniques précitées, les méthodes évoquées dans cette section sont implémentées dans le logiciel DTM qui peut être librement téléchargé à partir du site www.lebart.org.

4.1 Le bootstrap sur variables

Cette procédure n'a de sens que si il existe un « univers des variables » pour lequel la notion de « tirage de variables » a un sens. Les variables sont par exemple des événements nombreux, des instants, des zones échantillonnées, où, comme dans le cas de l'exemple de la *sémiométrie* (Lebart *et al.,* 2003), des mots. Pour tester la stabilité des structures vis-à-vis de l'ensemble des variables, nous proposons de répliquer l'ensemble des variables lui-même par la méthode du *bootstrap total*. Nous supposons ainsi implicitement que l'ensemble des variables actives constitue un échantillon de *m* variables extrait aléatoirement d'un ensemble de variables potentielles. Nous perturbons cet échantillon de variables selon les mêmes principes que le *bootstrap* sur individus. Pour chaque réplication, les variables non tirées participent à l'analyse de l'échantillon répliqué avec un poids infinitésimal (variables supplémentaires) ce qui permet d'obtenir des nuages de réplications pour chaque variable (et pour chaque observation) (exemples dans l'ouvrage cité).

[3] Ces trois types de bootstrap sont présentés avec des applications dans : Lebart *et al.* (2006).

4.2 Le bootstrap spécifique (ou hiérarchique)

Le *bootstrap spécifique* intervient notamment dans les analyses de données textuelles, dans le cas des questions ouvertes, par exemple, pour lesquelles il existe deux niveaux d'individus statistiques : les mots, qui sont les individus des tables de contingences lexicales, et les répondants, qui sont les individus statistiques classiques des enquêtes. Le tirage avec remise des occurrences de mots peut être remplacé par un tirage avec remise des répondants qui sont en fait des « grappes de mots ». Les données ainsi répliquées peuvent donner lieu aux bootstraps partiels ou totaux. Elles conduisent en général à des zones de confiances plus larges si les réponses sont de tailles très différentes, circonstance fréquente en pratique.

4.3 Le bootstrap partiel pour cartes auto-organisées

Le bootstrap partiel peut être appliqué à d'autres opérateurs projections que ceux des variables supplémentaires usuelles. Ainsi, l'analyse de contiguïté permet de trouver des sous-espaces de projection les plus proches possibles d'une carte auto-organisée de Kohonen (respectant la topologie de la carte au sens de la variance locale), ce qui permet de représenter par des zones de confiance l'incertitude sur la position des points. La projection des points répliqués ne se fait plus sur un plan factoriel, mais un plan sur lequel se projette la carte auto-organisée de façon optimale (ce plan est optimal au sens de la variance locale, qui est une variance calculée seulement à l'intérieur des classes et entre classes contiguës sur la carte). On trouvera détails et exemples d'application dans Lebart (2006).

5 Précisions sur la validation *procrustéenne*

Revenons maintenant sur le bootstrap total de type 3 qui implique la forme la plus élaborée de traitement statistique des réplications. On doit dans ce cas fixer un paramètre t (t comme *tolérance*), qui est le nombre d'axes pour lesquels on s'autorise à effectuer une transformation procrustéenne. Si par exemple le sous-espace des dix premiers axes répliqués coïncide avec celui des dix premiers axes initiaux, on pourra trouver une rotation qui fera coïncider les axes (ce qui nous ramène dans cas au bootstrap partiel).

Notons $\mathbf{X}$ (p,q) le tableau des q coordonnées factorielles initiales pour chacune des p variables, et $\mathbf{X_k}$ sa k-ième réplication. Si les lignes de $\mathbf{X_k}$, d'ordre (p,q), subissent toutes un une même rotation, $\mathbf{X_k}$ est transformé en $\mathbf{X_k B}$ où $\mathbf{B}$ (q, q) est une matrice orthogonale (rotation ou symétrie par rapport à l'origine). On cherchera à rendre minimale la somme des carrés S des écarts entre $\mathbf{X}$ et $\mathbf{X_k B}$, qui peut s'écrire :

$$S = trace\ (\mathbf{X} - \mathbf{X_k B})' (\mathbf{X} - \mathbf{X_k B})$$

L'analyse procrustéenne orthogonale implique alors, pour chaque réplication $\mathbf{X_k}$ la décomposition aux valeurs singulières de $\mathbf{X'X_k}$ et donc la diagonalisation de la matrice $\mathbf{X' X_k X_k' X}$.

Si l'on s'intéresse à la stabilité du sous-espace de dimension $q' < q$ formé par les q' premières colonnes de $\mathbf{X}$, on peut ne garder que les q' premières colonnes de $\mathbf{X_k}$, et chercher une matrice $\mathbf{B}(q',q')$ (test sévère). Mais on peut aussi tolérer que certains des q' premiers axes aillent s'égarer dans un sous espace plus grand, et donc garder t colonnes, avec $q' < t < q$.

Les figures 1 et 2 sont relatives aux données sémiométriques disponibles sur http://ses.enst.fr/lebart/ (rubrique : logiciel, exemple EXA08 de DTM).

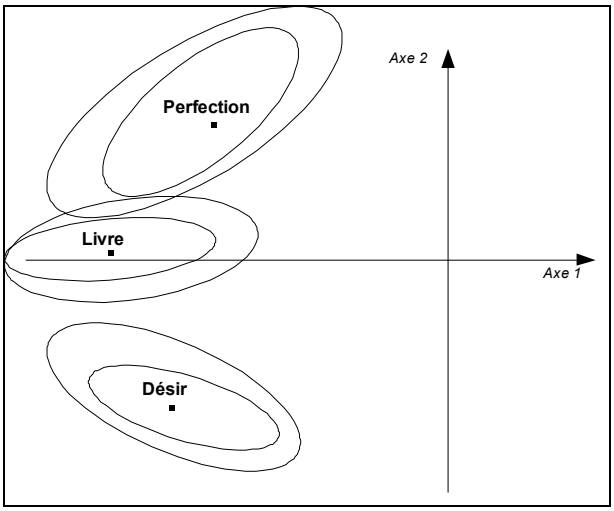

FIG. 1 – *Ellipses de confiance de trois mots (données sémiométriques) dans le plan (1, 2), ajustant 30 réplications. Les plus grandes ellipses correspondent à des rotations procustéennes pour t = 3 axes, les ellipses internes pour t = 12 axes.*

70 mots sont notés par 300 individus. La figure 1 montre ainsi la différence d'ampleur des zones de confiance pour trois mots (*Désir, Livre, Perfection*). Pour t = 12, les ellipses sont plus petites que pour t = 3. Autrement dit, on gagne en précision dans le plan (1, 2) si l'on se contente de la stabilité d'un espace à 12 dimensions (sur 70 au départ). Si l'on exige un espace à 3 dimensions, on doit se contenter d'une précision moindre.

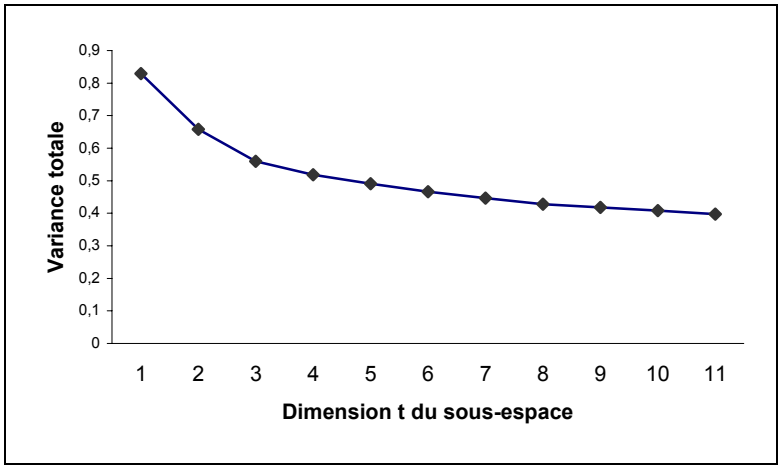

FIG. 2 – *Evolution de la somme des variances des réplications (liée à la taille moyenne des ellipses de confiance) dans le plan (1, 2) en fonction de la dimension t du sous-espace dans lequel sont effectuées les rotations procrustéennes des réplications.*

La figure 2 fait un bilan global de la variance totale interne des ellipses pour l'ensemble des 70 mots. L'augmentation de la précision avec t est confirmée, mais reste modérée.

6 Conclusion

La notion de validation d'une visualisation reste une notion assez complexe. Le bootstrap partiel, simple dans son principe, donne de bons ordres de grandeur, à moindre coût, de la précision à attendre sur la position des points. Les bootstraps totaux 1 et 2 permettent de trancher sans indulgence. Le bootstrap total de type 3, plus subtil, permet de décliner la validité en fonction de la taille des sous-espaces pris en considération.

Références

Chateau F., L. Lebart (1996). Assessing sample variability in visualization techniques related to principal component analysis: bootstrap and alternative simulation methods. In: *COMPSTAT96*, A. Prats (ed), Heidelberg: Physica Verlag, 205-210.

Diaconis P., B. Efron (1983). Computer intensive methods in statistics. *Scientific American*, 248, 116-130.

Efron B., R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Gower J. C., G.B. Dijksterhuis (2004). *Procrustes Problems*, Oxford: Oxford Univ. Press.

Greenacre M. (1984). *Theory and Applications of Correspondence Analysis.* London: Academic Press.

Lebart L. (2006). Assessing self organizing maps via contiguity analysis. *Neural Networks*, 19, 847-854.

Lebart L., M. Piron, A. Morineau (2006). *Statistique exploratoire multidimensionnelle, Validation et Inférence en fouilles de données.* Paris : Dunod.

Lebart L., M. Piron, J.-F. Steiner (2003). *La sémiométrie*. Paris : Dunod.

Milan L., J. Whittaker (1995). Application of the parametric bootstrap to models that incorporate a singular value decomposition. *Applied Statistics*, 44, 1, 31-49.

Summary

In the context of principal axes techniques we briefly show that, according to the concerns of the users, seven types of resampling techniques could be carried out to assess the quality of the obtained visualisations: a) Partial bootstrap, that considers the replications as supplementary variables, without new diagonalizations; b) Total bootstrap type 1, that imply a new diagonalization for each replicate, with corrections limited to possible changes of signs of the axes; c) Total bootstrap type 2, which adds to the preceding one a correction for possible exchanges of axes; d) Total bootstrap type 3, that implies procrustean transformations; e) Specific bootstrap, implying a resampling at a different level (case of a hierarchy of statistical units); f) A bootstrap on variables; g) An extension of bootstrap to some self organising maps.