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