UN PANORAMA DE QUELQUES MÉTHODES « SPARSE » POUR DES DONNÉES DE GRANDE DIMENSION

Gilbert Saporta
Laboratoire CEDRIC, Conservatoire National des Arts et Métiers, Paris
gilbert.saporta@cnam.fr

On parle de données de grande dimension quand le nombre p de variables est très supérieur au nombre n d'observations. Ce cas se produit dans de nombreux domaines, en particulier en biologie avec les données génomiques ou en chimiometrie.
Cet exposé à visée didactique partira d'un panorama de diverses solutions en régression pour aboutir aux méthodes « sparse » non supervisées en analyse en composantes principales et analyse des correspondances multiples.

En régression linéaire lorsque $p>n$ l'estimateur des m.c.o. n'existe pas. Puisque l'on se trouve devant un cas de multicolinéarité forcée, on peut utiliser des méthodes de régularisation classiques comme la régression ridge, la régression sur composantes principales ou la régression PLS qui fournissent des estimateurs efficaces, via une réduction de la dimension explicite ou implicite utilisant certaines contraintes sur les coefficients. On présente souvent comme un avantage le fait de conserver toutes les variables.
Cependant si $p>>n$ cette propriété devient un inconvénient car de telles combinaisons deviennent ininterpétables. Il est alors nécessaire de rechercher des combinaisons « sparse » *ie* avec un grand nombre de coefficients nuls. Les méthodes lasso, elastic net, sparse PLS effectuent simultanément régularisation et sélection grâce à des pénalités non quadratiques : L_1, SCAD etc. On présentera également des variantes de type group-lasso lorsque les variables sont structurées en blocs.

En ACP et ACM la décomposition en valeurs singulières montre que les composantes se régressent sur les variables pour donner les facteurs. Cela permet d'adapter les méthodes de régression sparse pour définir des versions sparse de l'ACP, puis de l'ACP sur des groupes de variables. On présentera pour conclure une version sparse de l'analyse des correspondances multiples.

Références

Bernard, A., Guinot, C. et Saporta, G. (2012), Sparse principal component analysis for multiblock data and its extension to sparse multiple correspondence analysis, in Proceedings *Compstat 2012*, 99-106

Bühlmann, P. et van de Geer, S. (2011), *Statistics for High-Dimensional Data*, Springer

Shen, H. et Huang, J. (2008), Sparse principal component analysis via regularized low rank matrix approximation, *Journal of Multivariate Analysis*, 99, 1015-1034.

Yuan, M. et Lin, Y. (2006), Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society*, Series B, 68, 49-67.

Zou, H., Hastie, T. et Tibshirani, R. (2004), Sparse Principal Component Analysis, *Journal of Computational and Graphical Statistics*, 15, 265-286.

A SURVEY OF SOME « SPARSE » METHODS FOR HIGH DIMENSIONAL DATA

Gilbert Saporta
Laboratoire CEDRIC, Conservatoire National des Arts et Métiers, Paris
gilbert.saporta@cnam.fr

High dimensional data means that the number of de variables *p* if far larger than the number of observations *n*. This occurs in several fields such as genomic data or chemometrics. This didactic talk starts from a survey of various solutions in linear regression and present afterwards their extensions to unsupervised « sparse » methods for principal components analysis (PCA) and multiple correspondence analysis (MCA).

When $p>n$ the OLS estimator does not exist for linear regression. Since it is a case of forced multicollinearity, one may use regularized techniques such as ridge regression, principal component regression or PLS regression: these methods provide rather robust estimates through a dimension reduction approach or with explicit (or not) constraints on the regression coefficients. The fact that all the predictors are kept is often considered as a positive point.
However if $p>>n$ it becomes a drawback since a combination of all variables cannot be interpreted. Sparse combinations, *ie* with a large number of zero coefficients are preferred. Lasso, elastic net, sparse PLS perform simultaneously regularization and variable selection thanks to non quadratic penalties: L_1, SCAD etc. We will present variants such as the group-lasso when the variables are structured in blocks.

In PCA, the singular value decomposition shows that if we regress principal components onto the input variables, the vector of regression coefficients is equal to the factor loadings. It suffices to adapt sparse regression techniques to get sparse versions of PCA and of PCA with groups of variables. We conclude by a presentation of a sparse version of Multiple Correspondence Analysis.

References

Bernard, A., Guinot, C. et Saporta, G. (2012), Sparse principal component analysis for multiblock data and its extension to sparse multiple correspondence analysis, in Proceedings *Compstat 2012*, 99-106

Bühlmann, P. et van de Geer, S. (2011), *Statistics for High-Dimensional Data*, Springer

Shen, H. et Huang, J. (2008), Sparse principal component analysis via regularized low rank matrix approximation, *Journal of Multivariate Analysis*, 99, 1015-1034.

Yuan, M. et Lin, Y. (2006), Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society*, Series B, 68, 49-67.

Zou, H., Hastie, T. et Tibshirani, R. (2004), Sparse Principal Component Analysis, *Journal of Computational and Graphical Statistics*, 15, 265-286.