Représentation et interprétation en régression PLS fonctionnelle

Anne de la Foye [1] & Alyssa Imbert [1] & Marie-Claire Not[1] &Papa Mbaye [2,3] & Chafik Samir [2] & Anne-Françoise Yao [3]

[1]*INRA Theix - UMRH-PFEM.* anne.delafoye@inra.fr

[2]*LIMOS, Univ. Clermont Auvergne, France. chafik.samir@uca.fr*

[3]LMBP*, Univ. Clermont Auvergne, France. papa.mbaye@math.univ-bpclermont.fr . anne-francoise.yao@math.univ-bpclermont.fr*

Résumé. Nous nous intéressons à la problématique de la mise en relation entre deux ensembles de variables en grande dimension : Y (à expliquer) et X (explicatif) à valeurs respectivement dans les espaces séparables, $\mathcal{X}$ et $\mathcal{Y}$ à partir n observations de $Z = (X, Y)$. Généralement, l'exploitation de ce type de données est fragmentaire. L'approche que nous proposons permet de visualiser directement l'influence des sous-ensembles X sur ceux de Y dans un contexte de régression PLS (Partial Least Squares). Plusieurs applications seront abordées au cours de cet exposé. Nous illustrerons notre approche notamment dans des situations où X et Y sont de grandes dimensions (voire fonctionnelles) et montrerons comment elle permet de répondre à des questions telles que : quel(s) groupe(s) de micro-ARN influence(nt) quel(s) groupe(s) d'ARN messagers.

Mots-clés. Analyse fonctionnelle de données, Réduction de dimension, Modèle linéaire fonctionnelle, Régression PLS.

Abstract. We present an exploratory approach for analyzing a link between two sets of variables characterized by high dimensionality of variables. Dealing with such situation is generally hard and one has to come up with adapted methods to reduce dimensionality without losing relevant information within the same framework. In that sense, we introduce a new approach to analyze the outputs in functional PLS regression. The method is illustrated on some real data application in order to distinguish any causal connection between the two groups of correlated observations.

Keywords. Functional Data Analysis, Dimensionality Reduction, Linear Model, PLS Regression.

1 Introduction

Dans de nombreux problèmes de régressions, les deux variables Y et X sont de grandes dimensions voire fonctionnelles. C'est le cas par exemple de l'étude de la relation entre deux courbes (précipitation et température) ou deux ensembles discrets d'informations génomiques. Il existe une large littérature sur les modèles permettant d'expliquer Y par X dans une optique de prédiction. A notre connaissance, très peu de travaux s'intéressent à l'analyse exploratoire des coefficients du modèle. Dans cette optique, nous nous intéressons à la représentation et l'interprétation a posteriori des sorties du modèle de régression PLS (Partial Least Squares) de Y sur X à partir des observations $\{Z_i = (X_i,\ Y_i)\ ,\ i = 1, ..., n\}$.

En dimension finie, l'approche bien connue (voir figure 1) de Tenenhaus (1998) permet de répondre à la problématique d'intérêt lorsque X et Y ne sont de grandes dimensions. Preda et Schiltz (2011) étudient le cadre théorique d'une régression PLS fonctionnelle avec réponse fonctionnelle. Avant d'aller plus loin, nous rappelons que la régression PLS vise à expliquer Y par X en recherchant les directions les plus corrélées à X qui explique le mieux Y sous l'hypothèse :

$$Y(s) = (\Phi.X)(s) + \varepsilon\ , s \in \mathcal{T}$$

où $\mathcal{T}$ est un sous ensemble de $\mathbb{R}$. Dans nos applications, après avoir utilisé la méthode de Preda et Schiltz (2011), nous avons été confronté à des problèmes en lien avec la représentation et l'interprétation de Φ (et $\Phi.X$). A notre connaissance, il n'existe pas de solution adaptée à notre problème. C'est ainsi que nous proposons une approche qui viendra en complément du modèle et aidera à mieux comprendre le lien entre les variables en présence.

2 Méthode d'interprétation en régression PLS avec réponse multivariée (ou fonctionnelle)

La méthode que nous proposons est basée sur des cartes permettant la représentation et interprétation de Φ (et $\Phi.X$) dans les situations suivantes :

1. X et Y sont multivariés en grande dimension (exemple d'application en génomique, présenté ci-dessous),

2. X et Y sont des variables fonctionnelles (température et précipitation)

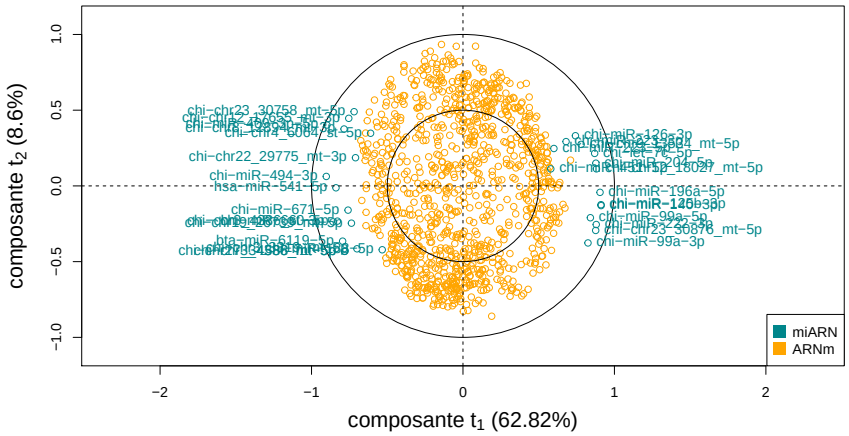

Figure 1: Représentation (t_1, t_2) de la PLS classique.

3. Y est une variable fonctionnelle et X une variable multivariée avec au moins une composante fonctionnelle (exemple Y est le rendement d'une exploitation agricole et X différents facteurs).

Nous présentons l'exemple du cas 1 dans la section suivante.

2.1 Un exemple d'application

Parmi les problèmes que nous avons rencontrés dans nos applications, il y a celui où les Y_i sont des expressions d'ARN messagers et les X_i ceux de micro-ARN. Les données à analyser sont issues d'une expérimentation dont l'un des objectifs était de mettre en évidence les conséquences d'une restriction alimentaire sur l'expression génétique au niveau de la glande mammaire chez la chèvre. L'expérience a été réalisée sur n individus. Parmi ces individus, certains ont été alimentés normalement alors que les autres n'ont pas été alimentés durant 48h avant le prélèvement. De plus, les chèvres de l'étude sont de deux génotypes différents notés AA et FF et ont été réparties de manière équilibrée entre les deux régimes alimentaires. Enfin, l'étude a permis de mesurer l'intensité de 7140 ARN messagers et un séquençage à haut débit a été réalisé sur 1804 microARN. Seuls les 1024 ARN messagers et les 32 microARN les plus différentiellement exprimés ont été retenus pour l'étude. Nous avons appliqué l'approche Preda et Schiltz (2011) sur nos données et des exemples de représentations des résultats obtenus sont présentés ci-dessous. Rappelons que

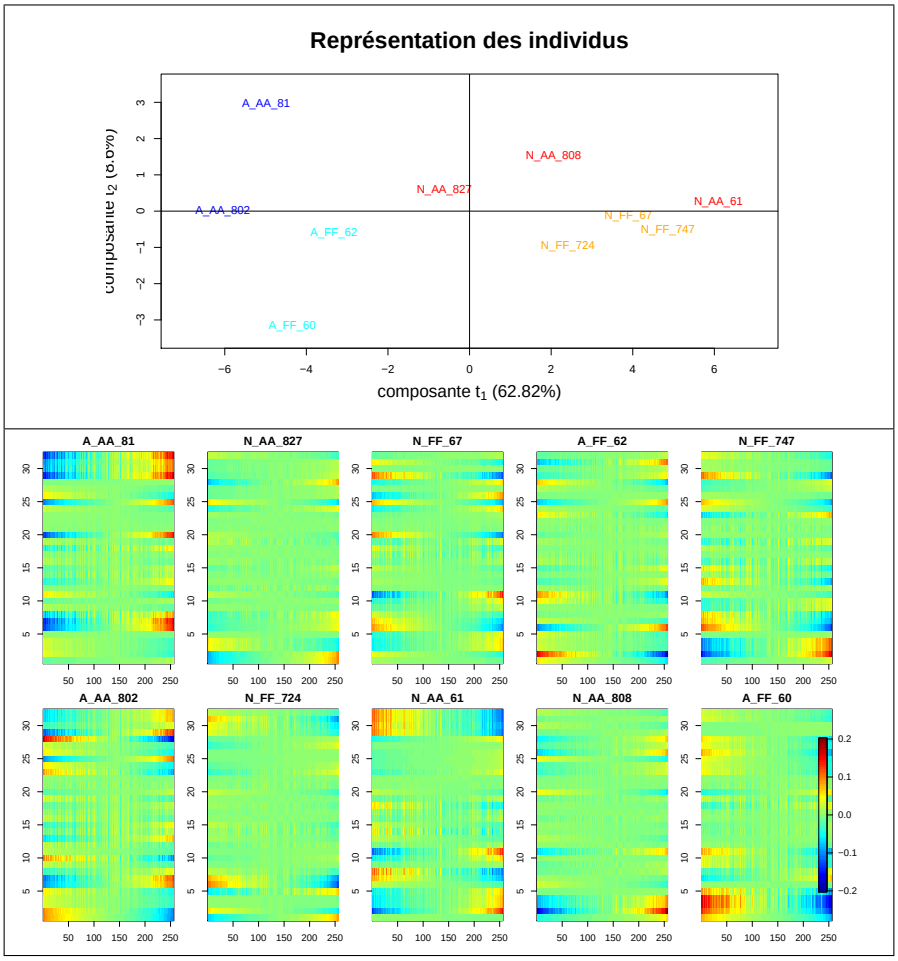

Figure 2: Représentation dans le plan (t_1, t_2) de la PLS fonctionnelle.

l'un des objectifs est de répondre à la question : quel(s) groupe(s) de micro-ARN influence(nt) l'expression de quel(s) groupe(s) d'ARN messagers?

Résultats de régression PLS de Y sur X.

La figure 1, représente une sortie de la régression PLS classique de Tenenhaus (1998) sur nos données. Celle-ci montre clairement que la régression PLS classique, ne permet dans de répondre à la question du fait du nombre important de variables en présence. En revanche la régression fonctionnelle PLS dont des résultats sont présentés permet directement de lier directement la position de l'individu dans le premier plan PLS à la relation entre les sous-ensembles de X et ceux de Y.

D'autres illustrations des résultats et ainsi que d'autres applications seront présentées au cours de l'exposé.

References

[1] Lê Cao KA, Rossouw D, Robert-Granié C, Besse P (2008), Sparse PLS : Variable selection when Integrating Omics data. SAGMB, 7 (1), 1-29.

[2] Preda C. Saporta G. (2005) PLS regression on a stochastic process, Computational Statistics and Data Analysis, 48, 149-158.

[3] Preda C. and Schiltz J. (2011), Functional PLS regression with functional response: the basis expansion approach. Proceedings of the 14th Applied Stochastic Models and Data Analysis Conference, 1126-1133

[4] Tenenhaus M. (1998) La régression PLS Théorie et pratique. Editions Technip.