Analyse en Composantes Principales Sparse pour données multiblocs et extension à l'Analyse des Correspondances Multiples Sparse

Anne Bernard [1,2] & Gilbert Saporta [2]

[1] *CERIES, 20 rue Victor Noir, Neuilly sur Seine, France, anne.bernard@ceries-lab.com*
[2] *CNAM, laboratoire CEDRIC, 292 rue Saint-Martin, Paris, France, gilbert.saporta@cnam.fr*

Résumé.L'Analyse en Composantes Principales pour des données quantitatives, et l'Analyse des Correspondances Multiples pour des données qualitatives, sont des techniques de réduction de dimension bien connues. Cependant, les composantes obtenues à l'issue de ces méthodes sont des combinaisons de toutes les variables de départ, ce qui rend l'interprétation des résultats difficile pour des données de grande dimension. Pour pallier ces difficultés, nous proposons deux nouvelles méthodes de sélection de groupes de variables quantitatives et qualitatives : la "Group Sparse Principal Component Analysis" et l'ACM sparse, respectivement. La GSPCA est une extension de la SPCA-rSVD de Shen et Huang pour des données structurées par bloc. Elle utilise les liens entre l'ACP et la décomposition en valeurs singulières, afin d'extraire les composantes en résolvant un problème d'approximation de matrice de rang inférieur. Une contrainte de type "Group Lasso" est introduite dans ce problème de minimisation afin d'obtenir des composantes étant combinaison d'un petit nombre de groupes de variables. Les loadings d'un groupe sont mis à zéro permettant de réduire le nombre de variables sélectionnées. La sélection ne sera pas globale mais propre à chaque composante. Puisque l'ACM est un cas particulier de l'ACP pour des blocs de variables indicatrices, l'ACM sparse est définie comme une extension de la GSPCA. Une application de cette méthode sera présentée sur un jeu de données bien connu comportant 27 races de chiens, décrites par 6 variables qualitatives.

Mots-clés. Réduction de dimension, Analyse en Composantes Principales sparse, Analyse des Correspondances Multiples, décomposition en valeurs singulières, méthodes multibloc.

Abstract. Principal Component Analysis for quantitative data, and Multiple Correspondence Analysis for qualitative data are well-known dimension reduction methods. However, the principal components obtained are combinations of all the original variables, making interpretation of results difficult for high dimensional data. To overcome these difficulties, we propose two new methods for selecting groups of quantitative and qualitative variables : "Group Sparse Principal Component Analysis" and "Sparse Multiple Correspondence Analysis", respectively. GSPCA is an extension of SPCA-RSVD of Shen and Huang for data structured by blocks. It uses the connection between PCA and singular

value decomposition to extract components through solving a low rank matrix approximation problem. A regularization penalty "group Lasso" is introduced to the corresponding minimization problem to obtain components that are combinations of a few number of groups of variables. All loadings of a block of variables are set to zero to reduce the number of selected variables. The selection will not be overall but specific to each component. Since MCA is a special case of PCA for blocks of dummy variables, SMCA is defined as an extension of GSPCA. An application of this method will be presented on a well-known data set containing 27 breeds of dogs, described by 6 variables.

Keywords. Dimension reduction, Sparse Principal Component Analysis, Multiple Correspondence Analysis, Singular Value Decomposition, multiblock methods.

1 Introduction

Les méthodes de réduction de dimension telles que l'Analyse en Composantes Principales (ACP) ou l'Analyse des Correspondances Multiples (ACM) sont nécessaires dans de nombreux domaines d'application. Cependant, les composantes principales (CP) issues de ces méthodes sont des combinaisons de toutes les variables de départ. L'interprétation des résultats devient alors difficile, notamment pour des données de très grande dimension. Une solution pour pallier ce problème consiste à produire de la sparsité dans les loadings (nombre réduit d'éléments non-nuls), afin d'éliminer des variables ou des blocs de variables lorsque celles ci sont structurées par groupe. Les méthodes de sélection de variables usuelles comme la régression avec contrainte Lasso ou Elastic net, permettent une sélection globale de variables. En revanche, dans un contexte non supervisé, cette sélection sera réalisée axe par axe et ne pourra pas être généralisée à l'ensemble des axes.

La méthode "Sparse Principal Component Analysis" via SVD régularisée (SPCA-rSVD) développée par Shen et Huang (2007) est une méthode de réduction de variables dans un cas non supervisé. Elle utilise le lien entre l'ACP et la décomposition en valeurs singulières (SVD : Singular Value Decomposition) afin d'extraire des CP en résolvant un problème d'approximation de matrice de rang inférieur. On considère $\mathbf{X}$ une matrice $n \times p$ de rang r. La SVD de $\mathbf{X}$ peut s'écrire : $\mathbf{X}=\mathbf{U}\mathbf{D}\mathbf{V}^T$ avec $\mathbf{U}=[\mathbf{u}_1,...,\mathbf{u}_r]$, $\mathbf{V}=[\mathbf{v}_1,...,\mathbf{v}_r]$ et $\mathbf{D}=\text{diag}\{d_1, ..., d_r\}$ la matrice diagonales des valeurs propres. Les colonnes de $\mathbf{U}$ et $\mathbf{V}$ sont orthonormales avec $\mathbf{Z}=\mathbf{U}\mathbf{D}$ les composantes principales et $\mathbf{V}$ les loadings correspondants. Si l'on considère la SVD comme une approximation de matrice de rang inférieur, on peut définir le problème d'optimisation suivant :

$$\min_{\mathbf{X}^{(1)}} \|\mathbf{X} - \mathbf{X}^{(1)}\|_F^2 \equiv \min_{\tilde{\mathbf{u}},\tilde{\mathbf{v}}} \|\mathbf{X} - \tilde{\mathbf{u}}\tilde{\mathbf{v}}^{\mathbf{T}}\|_F^2 \quad (1)$$

avec $\mathbf{X}^{(1)} \equiv d_1\mathbf{u}_1\mathbf{v}_1^T$ la matrice d'approximation de rang 1 de $\mathbf{X}$ (Eckart et Young, 1936) et les solutions : $\tilde{\mathbf{u}} = \mathbf{u}_1$, un n-vecteur de norme 1 et $\tilde{\mathbf{v}} = d_1\mathbf{v}_1$ un p-vecteur. Pour obtenir des loadings "sparse", la fonction de pénalité est imposée sur $\tilde{\mathbf{v}}$ dans le problème

d'optimisation (1) car pour $\tilde{\mathbf{u}}$ fixé, le $\tilde{\mathbf{v}}$ optimal est le vecteur des coefficients des moindres carrés dans la régression des colonnes de $\mathbf{X}$ sur $\tilde{\mathbf{u}}$. Le problème d'optimisation devient :

$$\min_{\tilde{\mathbf{u}},\tilde{\mathbf{v}}} \|\mathbf{X} - \tilde{\mathbf{u}}\tilde{\mathbf{v}}^{\mathbf{T}}\|_F^2 + P_\lambda(\tilde{\mathbf{v}}) = \min_{\tilde{\mathbf{u}},\tilde{\mathbf{v}}} \sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - \tilde{\mathbf{u}}_i\tilde{\mathbf{v}}_j)^2 + \sum_{j=1}^{p} p_\lambda(|\tilde{v_j}|) \tag{2}$$

avec $P_\lambda(\tilde{\mathbf{v}})$ la fonction de pénalité et λ le paramètre de régularisation. La solution de (2) est obtenue en appliquant une fonction de seuillage h_λ à $\mathbf{X}^T\tilde{\mathbf{u}}$, composante par composante :

$$\tilde{\mathbf{v}} = h_\lambda(\mathbf{X}^T\tilde{\mathbf{u}}) = \min_{\tilde{v_j}} \tilde{v_j}^2 - 2(\mathbf{X}^T\tilde{\mathbf{u}})_j\tilde{v_j} + p_\lambda(|\tilde{v_j}|) \tag{3}$$

puis pour $\tilde{\mathbf{v}}$ fixé, $\tilde{\mathbf{u}} = \mathbf{X}\tilde{\mathbf{v}}/\|\mathbf{X}\tilde{\mathbf{v}}\|$. L'algorithme itératif SPCA-rSVD détaillé dans le papier de Shen et Huang est défini seulement pour des vecteurs de dimension 1. Pour obtenir les loadings sparse sur les dimensions > 1, il faudra appliquer cet algorithme sur l'approximation de rang 1 des matrices résiduelles. Dans la suite du papier, l'extension de cette méthode est présentée dans le cas où les variables sont structurées par bloc.

2 GSPCA via SVD régularisée

Soit $\mathbf{X}$ une matrice $n \times p$ de variables quantitatives composée de J sous-matrices $\mathbf{X}_{[j]}$, $j = 1, ..., J$, chacune de dimension $n \times p_{[j]}$, avec $p_{[j]}$ le nombre de variables dans le groupe j. La SVD de $\mathbf{X}$ s'écrit alors :

$$\begin{aligned} \mathbf{X} =& [\mathbf{X}_{[1]}|...|\mathbf{X}_{[j]}|...|\mathbf{X}_{[J]}] = \mathbf{UD}([\mathbf{V}_{[1]}^T|...|\mathbf{V}_{[j]}^T|...|\mathbf{V}_{[J]}^T]) \\ =& [\mathbf{UDV}_{[1]}^T|...|\mathbf{UDV}_{[j]}^T|...|\mathbf{UDV}_{[J]}^T] \end{aligned} \tag{4}$$

avec $\mathbf{Z}$=$\mathbf{UD}$ matrice des composantes principales et $\mathbf{V}$=$[\mathbf{V}_{[1]}^T|...|\mathbf{V}_{[j]}^T|...|\mathbf{V}_{[J]}^T]^T$ matrice des loadings dont les sous-matrices $\mathbf{V}_{[j]}^T$ sont de dimension $\mathrm{p}_{[j]} \times r$, .

Pour répondre au problème de minimisation (2), on considère $\tilde{\mathbf{u}}$=$\mathbf{u}_1$ la première colonne de $\mathbf{U}$ et $\tilde{\mathbf{v}}$=$d_1(\mathbf{v}_{[1]},...,\mathbf{v}_{[j]},...,\mathbf{v}_{[J]})^T$, avec $\mathbf{v}_{[j]}$ de longueur $p_{[j]}$ la première colonne de la matrice $\mathbf{V}_{[J]}^T$ et d_1 la première valeur propre. Le problème d'optimisation (2) s'écrit alors :

$$\min_{\tilde{\mathbf{u}},\tilde{\mathbf{v}}} \|\mathbf{X} - \tilde{\mathbf{u}}\tilde{\mathbf{v}}^{\mathbf{T}}\|_F^2 + P_\lambda(\tilde{\mathbf{v}}) = \min_{\tilde{\mathbf{u}},\tilde{\mathbf{v}}} \sum_{i=1}^{n}\sum_{j=1}^{J}(\mathbf{X}_{[j],i} - \tilde{\mathbf{u}}_i\tilde{\mathbf{v}}_{[j]}^T)^2 + \sum_{j=1}^{J} p_\lambda(|\tilde{\mathbf{v}}_{[j]}|) \tag{5}$$

Les données étant structurées par bloc, la fonction de pénalisation P_λ choisie sera celle du "Group Lasso" introduite par Yuan et Lin (2006) et définie de la manière suivante :

$$P_\lambda(|\tilde{\mathbf{v}}|) = \lambda \sum_{j=1}^{J} \sqrt{p_{[j]}}\|\tilde{\mathbf{v}}_{[j]}\|_2 \tag{6}$$

A l'aide d'un algorithme de type "coordinate descent" et des conditions de Karush-Kuhn-Tucker décrites dans Yuan et Lin (2006), on en déduit la solution au problème (5) :

$$\tilde{\mathbf{v}}_{[j]} = (1 - \tfrac{1}{2}\lambda \frac{\sqrt{p_{[j]}}}{\|\mathbf{X}_{[j]}^T \tilde{\mathbf{u}}\|})_+ \mathbf{X}_{[j]}^T \tilde{\mathbf{u}} \tag{7}$$

(par souci de limitation du nombre de pages, le détail des calculs n'est pas développé ici).

L'ACM étant un cas particulier de l'ACP pour des blocs de variables indicatrices, l'ACM sparse est définie dans le paragraphe suivant comme une extension de la GSPCA.

3 ACM sparse via SVD régularisée

Soit $\mathbf{X}$ une matrice $n \times J$ de variables qualitatives. Le tableau disjonctif complet $\mathbf{K}$ correspondant est donc composé de J sous-matrices de variables indicatrices $\mathbf{K}_{[j]}$, $j = 1, ..., J$, chacune de dimension $n \times p_{[j]}$, avec $q = \sum_{j=1}^{J} p_{[j]}$ le nombre total de modalités. On pose $\mathbf{M}$ la matrice diagonale des poids lignes de dimension $n \times n$ et $\mathbf{A}$ la matrice diagonale des poids colonne de dimension $q \times q$. La SVD sera donc réalisée sur $\mathbf{K}$:

$$\mathbf{K} = \mathbf{P}\Delta\mathbf{Q}^T \text{ avec } \mathbf{P}^T\mathbf{M}\mathbf{P} = \mathbf{Q}^T\mathbf{A}\mathbf{Q} = \mathbf{I} \tag{8}$$

avec $\mathbf{P}\Delta$ les composantes principales, $\mathbf{Q}$ la matrice des loadings correspondants et Δ la matrice diagonale des valeurs propres. Comme explicité dans les paragraphes précédents, nous souhaitons avoir $\tilde{\mathbf{u}}$ un n-vecteur de norme 1 et $\tilde{\mathbf{v}}$ un q-vecteur libéré de toute contrainte de normalité. Étant donné (8) :

$$(\mathbf{M}^{\frac{1}{2}})^T\mathbf{K}\mathbf{A}^{\frac{1}{2}} = (\mathbf{M}^{\frac{1}{2}})^T\mathbf{P}\Delta\mathbf{Q}^T\mathbf{A}^{\frac{1}{2}} \tag{9}$$

Si on pose $\mathbf{R} = (\mathbf{M}^{\frac{1}{2}})^T\mathbf{K}\mathbf{A}^{\frac{1}{2}}$, $\mathbf{U} = (\mathbf{M}^{\frac{1}{2}})^T\mathbf{P}$ et $\mathbf{V} = \mathbf{Q}^T\mathbf{A}^{\frac{1}{2}}$ que l'on peut également écrire sous la forme $\mathbf{V} = ([\mathbf{Q}_{[1]}^T|...|\mathbf{Q}_{[j]}^T|...|\mathbf{Q}_{[J]}^T])^T(\mathbf{A}_1, ..., \mathbf{A}_j, ..., \mathbf{A}_J)^{\frac{1}{2}}$, nous retrouvons la même écriture que dans le paragraphe précédent : $\mathbf{R}{=}\mathbf{U}\Delta\mathbf{V}^T$.

Afin de résoudre le problème d'optimisation (5), nous considérerons $\tilde{\mathbf{u}}{=}\mathbf{u}_1$ la première colonne de $\mathbf{U}$ et $\tilde{\mathbf{v}}{=}\Delta_1(\mathbf{v}_{[1]},...,\mathbf{v}_{[j]},...,\mathbf{v}_{[J]})^T$, avec $\mathbf{v}_{[j]} = \mathbf{q}_{[j]}\mathbf{A}_j{}^{\frac{1}{2}}$de longueur $p_{[j]}$, la première colonne de la matrice $\mathbf{V}_{[J]}^T$. En conservant ces notations, la solution du problème sera la même que dans la GSPCA.

Les propriétés barycentriques de l'ACM sont conservées cependant les propriétés de non corrélation des CP et d'orthogonalité des loadings sont perdues en ACM sparse. La variance expliquée ne pourra donc pas être définie de la même manière qu'en ACM. On définit la variance ajustée de la kième CP par $\text{tr}(\mathbf{X}_k^T\mathbf{X}_k) - \text{tr}(\mathbf{X}_{k-1}^T\mathbf{X}_{k-1})$ et le pourcentage cumulé de variance expliquée par les k premières CP par $\text{tr}(\mathbf{X}_k^T\mathbf{X}_k)/\text{tr}(\mathbf{X}^T\mathbf{X})$ avec $\mathbf{X}_k{=}\mathbf{X}\mathbf{V}_k(\mathbf{V}_k^T\mathbf{V}_k)^{-1}\mathbf{V}_k^T$ et $\mathbf{V}_k$ la matrice des k premiers loadings sparse. Nous allons à présent illustrer cette nouvelle méthode à l'aide d'un exemple bien connu.

4 Exemple d'application

On considère l'exemple des 27 races de chiens décrites au moyen de 6 variables qualitatives (Tenenhaus, 2007). On pose $\mathbf{X}$ la matrice des données (27×6) et $\mathbf{K}$ le tableau disjonctif correspondant (27×16) constitué de 6 blocs de variables indicatrices.

Une approche comparative entre l'ACM et l'ACM sparse est présentée sur les quatre premières composantes principales. La figure 1 présente l'évolution du pourcentage de variance cumulé en fonction du paramètre λ choisi. A partir d'un $\lambda > 0.25$, le pourcentage de variance cumulé décroit fortement. Nous fixerons alors λ à 0.25 pour la suite de l'analyse.

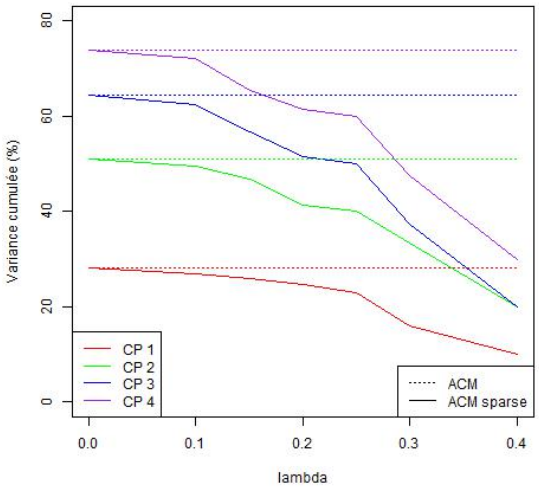

FIGURE 1 – Variance cumulée (%) sur les 4 premières composantes en fonction de λ.

Le tableau 1 présente une comparaison des loadings obtenus avec l'ACM et l'ACM sparse. L'ACM sparse permet de réduire considérablement le nombre de variables sélection-nées par axe, tout en conservant un bon niveau d'information (les variables les plus contributives dans l'ACM sont celles conservées dans l'ACM sparse). Par ailleurs, même avec un λ relativement élevé, le pourcentage de variance ajustée diminue très peu entre l'ACM et l'ACM sparse (28.19% vs 23.03% pour le 1^{er} axe). La représentation des variables et des individus sur les axes ne peut être présentée ici, faute de place, mais on peut observer des regroupements de plus en plus précis de races de chiens lorsque l'on augmente la valeur du paramètre λ.

5 Conclusion

L'ACM sparse est une extension de la méthode GSPCA pour des données qualitatives. Elle produit de la sparsité au niveau des loadings (avec une perte faible du pourcentage de variance expliquée) ce qui facilite l'interprétation et la compréhension des différents axes. Lorsque le paramètre de régularisation λ est fixé à 0, la GSPCA et l'ACM sparse sont identiques à l'ACP et l'ACM, respectivement, ce qui d'après Zou et Hastie (2006) est une propriété essentielle d'une "bonne" méthode sparse. L'application présentée dans le paragraphe 4 a été réalisée sur un petit jeu de données, mais ces méthodes prennent tout leur sens dans un contexte de très grande dimension. En revanche elle ne produisent pas de sparsité au sein d'un groupe. Dans le cas où l'on souhaiterait sélectionner des

TABLE 1 – Loadings et variance obtenus avec l'ACM et l'ACM sparse sur les quatre premières composantes.

	ACM				**ACM sparse**			
Variable	CP1	CP 2	CP 3	CP 4	CP1	CP 2	CP 3	CP 4
grand	-0.361	0.071	-0.005	0.060	-0.389	**0.000**	**0.000**	**0.000**
moyen	0.280	0.287	0.300	-0.055	0.226	**0.000**	**0.000**	**0.000**
petit	0.291	-0.400	-0.293	-0.041	0.390	**0.000**	**0.000**	**0.000**
leger	0.316	-0.389	-0.193	-0.081	0.368	-0.256	**0.000**	**0.000**
lourd	-0.047	0.390	-0.133	0.088	-0.075	0.451	**0.000**	**0.000**
treslourd	-0.294	-0.215	0.458	-0.055	-0.305	-0.479	**0.000**	**0.000**
lent	0.059	-0.383	0.296	0.133	**0.000**	-0.561	**0.000**	**0.000**
rapide	0.224	0.256	0.057	-0.299	**0.000**	0.282	**0.000**	**0.000**
tresrapide	-0.303	0.156	-0.391	0.168	**0.000**	0.328	**0.000**	**0.000**
moyintelligent	0.173	0.157	0.356	0.236	**0.000**	**0.000**	0.693	-0.693
peuintelligent	-0.145	-0.309	-0.168	0.125	**0.000**	**0.000**	-0.327	0.327
tresintelligent	-0.086	0.125	-0.330	-0.491	**0.000**	**0.000**	-0.642	0.642
peuaffectueux	-0.366	-0.084	0.030	0.087	-0.462	**0.000**	**0.000**	**0.000**
tresaffectueux	0.353	0.081	-0.029	-0.084	0.445	**0.000**	**0.000**	**0.000**
agressif	-0.170	-0.096	0.162	-0.515	**0.000**	**0.000**	**0.000**	**0.000**
nonagressif	0.164	0.093	-0.156	0.497	**0.000**	**0.000**	**0.000**	**0.000**
Variance ajustée (%)	28.19	22.80	13.45	9.55	23.03	17.40	10.20	9.60
Variance cumulée (%)	28.19	50.99	64.44	73.99	23.03	39.99	50.19	59.79

variables au sein d'un bloc, une extension de ces méthodes pourrait être réalisée en remplaçant la fonction de pénalité "group Lasso" par la "sparse group Lasso" développée par Simon et al. (2012).

Bibliographie

[1] Eckart, C. et Young, G. (1936), The approximation of one matrix by another of lower rank, *Psychometrika*, 1, 211–218.

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

[3] Simon, N., Friedman, J., Hastie, T. et Tibshirani, R. (2012), A Sparse-Group Lasso, *Journal of Computational and Graphical Statistics*, DOI :10.1080/10618600.2012.681250.

[4] Tenenhaus, M. (2007), *Statistique ; méthodes pour décrire, expliquer et prévoir*, Dunod, 252–254.

[4] 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.

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