Analyse canonique généralisée régularisée et approche PLS

Arthur Tenenhaus, Michel Tenenhaus

ANALYSE CANONIQUE GÉNÉRALISÉE RÉGULARISÉE ET APPROCHE PLS

Arthur Tenenhaus[(1)] & Michel Tenenhaus[(2)]

(1) SUPELEC, Campus de Gif-sur-Yvette, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette
(2) HEC Paris, 1 rue de la Libération, 78351 Jouy-en-Josas

Résumé

Nous donnons dans cette communication une définition de l'analyse canonique généralisée au niveau de la population (ACG-population) qui constitue le cadre théorique de l'approche PLS proposée par Herman Wold et à ses extensions proposées par Jan-Bernd Lohmöller et Nicole Krämer. En écrivant les équations stationnaires de l'ACG-population au niveau de l'échantillon et en utilisant des estimations régularisées (*shrinkage estimations*) des matrices de covariance des blocs, nous obtenons de nouvelles équations stationnaires au niveau de l'échantillon. Ces équations stationnaires sont également celles d'un problème d'optimisation que nous appelons analyse canonique généralisée régularisée (ACGR). En recherchant un point fixe de ces équations stationnaires au niveau de l'échantillon nous obtenons un algorithme très similaire à l'approche PLS de Wold-Lohmöller-Krämer. De plus, nous démontrons la convergence monotone de l'algorithme proposé.

Mots-clés: Analyse de tableaux multiples, Approche PLS, Analyse canonique généralisée régularisée

Abstract

In this paper, we give a definition of generalized canonical correlation analysis for population (GCCA-population) which constitutes a theoretical framework for the Partial Least Squares (PLS) path modeling algorithm proposed by Herman Wold and to the extensions proposed by Jan-Bernd Lohmöller and Nicole Krämer. Writing the stationary equations related to GCCA- population at the sample level and using shrinkage estimations for the block covariance matrices, we get new stationary equations at the sample level. These stationary equations are also stationary equations of an optimization problem that we call regularized generalized canonical correlation analysis (RGCCA). Searching for a fixed point for these sample level stationary equations, we obtain an algorithm very similar to the PLS approach of Wold-Lohmöller-Krämer. Furthermore, monotone convergence of the proposed algorithm is proven.

Keywords: Multi-block data analysis, PLS path modeling, Regularized GCCA

Introduction

Ce papier est organisé de la manière suivante :

(1) Définition de l'analyse canonique généralisée au niveau de la population et construction des équations stationnaires au niveau de la population.
(2) Construction des équations stationnaires au niveau de l'échantillon en utilisant des estimations régularisées des matrices de covariance des blocs.
(3) Définition de l'analyse canonique généralisée régularisée (ACGR).

(4) Recherche d'un point fixe des équations stationnaires au niveau de l'échantillon : l'algorithme PLS/Gauss-Seidel. Propriétés de convergence de l'algorithme PLS/Gauss-Seidel.

1. Analyse canonique généralisée au niveau de la population

Nous donnons ici une définition de l'ACG-population plus générale que la définition habituelle. Considérons J vecteurs colonnes aléatoires x_j. Nous considérons aussi un réseau de connections entre ces vecteurs défini par une matrice de structure $C=\{c_{jk}\}$: $c_{jk}=1$ si x_j and x_k sont connectés et 0 sinon. Enfin, nous considérons deux combinaisons linéaires $\eta_j=\alpha_j^t\mathrm{x}_j$ et $\eta_k=\alpha_k^t\mathrm{x}_k$. La corrélation entre η_j et η_k est définie par :

$$\rho(\alpha_j^t\mathrm{x}_j,\alpha_k^t\mathrm{x}_k)=\frac{\alpha_j^t\Sigma_{jk}\alpha_k}{\sqrt{\alpha_j^t\Sigma_{jj}\alpha_j}\sqrt{\alpha_k^t\Sigma_{kk}\alpha_k}} \tag{1}$$

où $\Sigma_{jj}=E(\mathrm{x}_j\mathrm{x}_j^t)$, $\Sigma_{kk}=E(\mathrm{x}_k\mathrm{x}_k^t)$ et $\Sigma_{jk}=E(\mathrm{x}_j\mathrm{x}_k^t)$.

L'ACG-population est définie par le problème d'optimisation suivant :

$$\begin{aligned}&\underset{\alpha_1,\ldots,\alpha_J}{\text{Maximiser}} \sum_{j,k=1,j\neq k}^{J} c_{jk}\,g(\rho(\alpha_j^t\mathrm{x}_j,\alpha_k^t\mathrm{x}_k))\\ &\text{sous les contraintes } Var(\alpha_j^t\mathrm{x}_j)=1,\ j=1,\ldots,J\end{aligned} \tag{2}$$

où g est la fonction identité, valeur absolue ou carré. Le problème (2) est équivalent au problème d'optimisation suivant :

$$\begin{aligned}&\underset{\alpha_1,\ldots,\alpha_J}{\text{Maximiser}} \sum_{j,k=1,j\neq k}^{J} c_{jk}\,g(\alpha_j^t\Sigma_{jk}\alpha_k)\\ &\text{sous les contraintes } \alpha_j^t\Sigma_{jj}\alpha_j=1,\ j=1,\ldots,J\end{aligned} \tag{3}$$

On considère ensuite le Lagrangien associé au problème d'optimisation (3) :

$$F(\alpha_1,\ldots,\alpha_J,\lambda_1,\ldots,\lambda_J)=\sum_{j,k=1,j\neq k}^{J} c_{jk}\,g(\alpha_j^t\Sigma_{jk}\alpha_k)-\varphi\sum_{j=1}^{J}\frac{\lambda_j}{2}(\alpha_j^t\Sigma_{jj}\alpha_j-1) \tag{4}$$

où $\varphi=1$ lorsque g est l'identité ou la valeur absolue et 2 lorsque g est le carré.

Annulant les dérivées du Lagrangien, on obtient les équations stationnaires suivantes :

$$\frac{1}{\varphi}\Sigma_{jj}^{-1}\sum_{k=1,k\neq j}^{J} c_{jk}\,g'(\alpha_j^t\Sigma_{jk}\alpha_k)\Sigma_{jk}\alpha_k=\lambda_j\alpha_j\quad,\quad j=1,\ldots,J \tag{5}$$

avec les contraintes de normalisation

$$\alpha_j^t\Sigma_{jj}\alpha_j=1,\ j=1,\ldots,J \tag{6}$$

2. Ecriture des équations stationnaires au niveau de l'échantillon avec utilisation des estimations régularisées des matrices de covariance des blocs de variables Σ_{jj}

Considérons *J* tableaux de données (ou blocs) $X_1,...,X_J$ formés de variables centrées mesurées sur un ensemble de *n* individus. Désignons par $C=\{c_{jk}\}$ la matrice décrivant le réseau de relation entre les blocs : $c_{jk}=1$ pour deux blocs connectés et 0 sinon. Les matrices de covariance empiriques sont $S_{jj}=\frac{1}{n}X_j^tX_j$ and $S_{jk}=\frac{1}{n}X_j^tX_k$. Dans le cas de forte multicolinéarité ou bien lorsque le nombre d'observations est inférieur au nombre de variables, la matrice de covariance empirique S_{jj} est une mauvaise estimation de la vraie matrice de covariance Σ_{jj}. Afin de trouver une meilleure estimation de la vraie matrice de covariance Σ_{jj}, Ledoit et Wolf (2004) suggèrent de considérer la classe des combinaisons linéaires $\{\hat{S}_{jj}=\tau_j I+(1-\tau_j)S_{jj},\ 0\le\tau_j\le 1\}$. La matrice $\hat{S}_{jj}$ est appelée une estimation régularisée (*shrinkage estimation*) de Σ_{jj} et τ_j est la constante de régularisation (*shrinkage constant*). Schäfer et Strimmer (2005)) donnent une formule analytique permettant de choisir τ_j de manière optimale.

Dans cette section, nous considérons la version échantillon des équations stationnaires (5) avec les contraintes de normalisation (6), où Σ_{jj} est remplacé par $\hat{S}_{jj}=\tau_j I+(1-\tau_j)S_{jj}$ et Σ_{jk} par S_{jk}. Ceci conduit aux équations stationnaires au niveau de l'échantillon suivantes :

$$(7)\qquad \frac{1}{\varphi}\left[\tau_j I+(1-\tau_j)S_{jj}\right]^{-1}\sum_{k=1,k\neq j}^{J}c_{jk}g'(a_j^tS_{jk}a_k)S_{jk}a_k=\lambda_j a_j\quad,\quad j=1,...,J$$

avec les contraintes de normalisation

$$(8)\qquad a_j^t\left[\tau_j I+(1-\tau_j)S_{jj}\right]a_j=1,\ \ j=1,...,J$$

On peut se rapprocher des équations stationnaires de l'algorithme PLS usuel en introduisant pour chaque bloc une composante externe $Y_j=X_ja_j$ et une composante interne Z_j définie par

$$(9)\qquad Z_j=\frac{1}{\varphi}\sum_{k=1,k\neq j}^{J}c_{jk}g'\left[Cov(Y_j,Y_k)\right]Y_k$$

On retrouve le schéma de Horst pour g = identité ($Z_j=\sum_{k=1,k\neq j}^{J}c_{jk}Y_k$), le schéma factoriel pour g = carré ($Z_j=\sum_{k=1,k\neq j}^{J}c_{jk}\left[Cov(Y_j,Y_k)\right]Y_k$) et le schéma centroïde pour g = valeur absolue ($Z_j=\sum_{k=1,k\neq j}^{J}c_{jk}signe\left[Cov(Y_j,Y_k)\right]Y_k$).

Puis en utilisant les équations stationnaires (7) et les contraintes de normalisation (8) nous obtenons les poids externes :

$$(10) \qquad a_j = \frac{1}{\sqrt{Z_j^t X_j [\tau_j I + (1-\tau_j)\frac{1}{n} X_j^t X_j]^{-1} X_j^t Z_j}} [\tau_j I + (1-\tau_j)\frac{1}{n} X_j^t X_j]^{-1} X_j^t Z_j$$

Pour les valeurs extrêmes 0 et 1 de τ_j, l'équation (10) donne des résultats similaires aux modes B et A de l'approche PLS, sauf au niveau des contraintes de normalisation. Dans l'algorithme PLS d'origine, toutes les composantes externes Y_j sont réduites.

3. Analyse canonique généralisée régularisée (ACGR)

En considérant les J blocs de données $X_1,...,X_J$, la matrice de structure $C = \{c_{ij}\}$, la fonction g et les constantes de régularisation $\tau_1,...,\tau_J$ décrits dans la section 2, nous pouvons définir l'analyse canonique généralisée régularisée (ACGR) comme le problème d'optimisation suivant :

$$(11) \qquad \begin{aligned} &\underset{a_1,...,a_J}{\text{Maximiser}} \quad \sum_{j,k=1, j\neq k}^{J} c_{jk}\, g(Cov(X_j a_j, X_k a_k)) \\ &\text{sous les contraintes} \quad \tau_j \left\| a_j \right\|^2 + (1-\tau_j) Var(X_j a_j) = 1, \quad j = 1,...,J \end{aligned}$$

En annulant les dérivées du Lagrangien associé au problème d'optimisation (11), on retrouve exactement les équations stationnaires (7) avec les contraintes de normalisation (8).

Application à l'approche PLS

Dans l'approche PLS, on considère que chaque bloc est l'expression d'une variable latente non observable et que des relations structurelles existent entre ces variables. On considère que deux blocs X_j et X_k sont connectés si les variables latentes associées apparaissent ensemble dans la même équation structurelle. Définissons la matrice de structure C: $c_{jk} = 1$ si les blocs X_j and X_k sont connectés dans le modèle à équations structurelles et = 0 sinon. Nous considérons ensuite la maximisation des critères suivants, qui sont en fait reliés à des algorithmes PLS existants ou modifiés, et sont tous des cas particuliers du problème d'optimisation (11):

SUMCOR-PLSPM : $\sum_{j,k,j\neq k} c_{jk} Cor(X_j a_j, X_k a_k)$ [Horst avec $Var(X_j a_j) = 1, \quad j = 1,...,J$]

SSQCOR-PLSPM : $\sum_{j,k,j\neq k} c_{jk} Cor^2(X_j a_j, X_k a_k)$ [Factoriel avech $Var(X_j a_j) = 1, \quad j = 1,...,J$]

SABSCOR-PLSPM : $\sum_{j,k,j\neq k} c_{jk} \left| Cor(X_j a_j, X_k a_k) \right|$ [Centroïde avec $Var(X_j a_j) = 1, \quad j = 1,...,J$]

SUMCOV-PLSPM : $\sum_{j,k,j\neq k} c_{jk} Cov(X_j a_j, X_k a_k)$ [Horst avec $\left\| a_j \right\| = 1, \quad j = 1,...,J$]

SSQCOV-PLSPM : $\sum_{j,k,j\neq k} c_{jk} Cov^2(X_j a_j, X_k a_k)$ [Factoriel avec $\left\| a_j \right\| = 1, \quad j = 1,...,J$]

SABSCOV-PLSPM: $\sum_{j,k,j\neq k} c_{jk} \left| Cov(X_j a_j, X_k a_k) \right|$ [Centroid with $\left\| a_j \right\| = 1, \quad j = 1,...,J$]

4. L'algorithme PLS/Gauss-Seidel

Il n'existe pas de solution analytique au problème d'optimisation (11). Cependant, il est possible de construire un algorithme à convergence monotone, c'est-à-dire que le critère à maximiser dans (11) augmente à chaque pas de la procédure itérative proposée. Nous proposons une suite d'opérations similaire à celle utilisée par Wold (1985) et Hanafi (2007). Nous appelons cet algorithme "PLS/Gauss-Seidel" parce qu'il est similaire à l'algorithme de Gauss-Seidel pour résoudre un système d'équations linéaires. Cet algorithme est décrit dans le tableau 1.

Tableau 1: L'algorithme PLS/Gauss-Seidel

A. ***Initialisation***

A1. Choisir J vecteurs arbitraires $\tilde{a}_1^0, \tilde{a}_2^0, ..., \tilde{a}_J^0$

A2. Calculer des poids externes $a_1^0, a_2^0, ..., a_J^0$ verifiant (8) :

$$a_j^0 = \frac{1}{\sqrt{(\tilde{a}_j^0)^t [\tau_j I + (1-\tau_j)\frac{1}{n} X_j^t X_j]^{-1} \tilde{a}_j^0}} [\tau_j I + (1-\tau_j)\frac{1}{n} X_j^t X_j]^{-1} \tilde{a}_j^0$$

A3. Calculer les composantes externes

$$Y_1^0 = X_1 a_1^0, ..., Y_J^0 = X_J a_J^0$$

B. ***Calcul de la composante interne du bloc*** X_j

Calculer la composante interne selon le schéma utilisé :

$$Z_j^s = \frac{1}{\varphi}\left[\sum_{k<j} c_{jk} g'\left[Cov(Y_j^s, Y_k^{s+1})\right] Y_k^{s+1} + \sum_{k>j} c_{jk} g'\left[Cov(Y_j^s, Y_k^s)\right] Y_k^s\right]$$

où $g'(x)/\varphi = 1$ pour le schéma de Horst, x pour le schéma factoriel et $signe(x)$ pour le schéma centroïde.

C. ***Calcul de la composante externe du bloc*** X_j

C1. Calculer le poids externe

$$a_j^{s+1} = \frac{1}{\sqrt{(Z_j^s)^t X_j [\tau_j I + (1-\tau_j)\frac{1}{n} X_j^t X_j]^{-1} X_j^t Z_j^s}} [\tau_j I + (1-\tau_j)\frac{1}{n} X_j^t X_j]^{-1} X_j^t Z_j^s$$

C2. Calculer la composante externe

$$Y_j^{s+1} = X_j a_j^{s+1}$$

La procédure commence par un choix arbitraire de valeurs initiales (étape A dans le tableau 1). Supposons que des composantes externes $Y_1^{s+1}, Y_2^{s+1}, ..., Y_{j-1}^{s+1}$ aient été construites pour les blocs $X_1, X_2, ..., X_{j-1}$. La composante externe Y_j^{s+1} est calculée en considérant la composante interne Z_j^s pour le bloc X_j (étape B dans le Tableau 1), puis la composante externe (étape C dans le Tableau 1). La procédure est itérée jusqu'à convergence du critère borné due à la proposition donnée ci-dessous.

Proposition : *Soit* $Y_j^s = X_j a_j^s$, $j = 1,...,J$, $s = 1,2,...$, *une suite de composantes externes générées par l'algorithme PLS/Gauss-Seidel. La fonction suivante est définie pour des vecteurs* $a_1, a_2, ..., a_J$ *vérifiant les contraintes* (8) *:*

$$(12) \qquad f(a_1, a_2, ..., a_J) = \sum_{j,k=1, j\neq k}^{J} c_{jk} g\left[Cov(X_j a_j, X_k a_k)\right]$$

Les inégalités suivantes sont vérifiées :

$$(13) \qquad \forall\ s \quad f(a_1^s, a_2^s, ..., a_J^s) \leq f(a_1^{s+1}, a_2^{s+1}, ..., a_J^{s+1})$$

La démonstration de cette proposition est donnée dans Tenenhaus & Tenenhaus (2010).

Donnons deux limitations de cet algorithme :

1) Il n'est pas garanti que l'algorithme converge vers un point fixe des équations stationnaires utilisées, bien que cela soit presque toujours le cas dans les applications pratiques.
2) Il n'est pas garanti que l'algorithme converge vers un optimum global du critère utilisé. Krämer (2007) a donné un exemple de convergence vers un optimum local.

Il est utile de préciser que Lohmöller a modifié l'algorithme de Wold en utilisant l'approche plus simple de Jacobi à la place de l'approche de Gauss-Seidel dans l'algorithme PLS. Ce n'était pas une bonne idée parce qu'il se trouve que la procédure obtenue ne possède plus la propriété de convergence monotone (voir Hanafi (2007)). Cette procédure « PLS/Jacobi » est utilisée dans tous les logiciels pour l'approche PLS parce qu'ils sont tous basés sur le logiciel de Lohmöller LVPLS (Lohmöller, 1984). Dans la pratique cependant, les deux procédures convergent presque tout le temps vers le même point fixe des équations stationnaires.

Bibliographie

[1] Hanafi M. (2007): PLS Path modelling: computation of latent variables with the estimation mode B, *Computational Statistics*, 22, 275-292.
[2] Krämer N. (2007): Analysis of high-dimensional data with partial least squares and boosting. Doctoral dissertation, Technischen Universität Berlin.
[3] Ledoit O. & Wolf M. (2004): A well conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88, 365-411.
[4] Lohmöller J.-B. (1984): *LVPLS Program Manual, Version 1.6,* Zentralarchiv für Empirische Sozialforschung, Universität zu Köln, Köln
[5] Schäfer J and Strimmer K. (2005): A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, *Statistical applications in genetics and molecular biology* 4, 1, Article 32.
[6] Tenenhaus A. & Tenenhaus M.: A PLS approach to regularized generalized canonical correlation analysis (*soumis à publication*).
[7] Wold H. (1985): "Partial Least Squares", in *Encyclopedia of Statistical Sciences*, vol. 6, Kotz, S & Johnson, N.L. (Eds), John Wiley & Sons, New York, pp. 581-591.