Kernel Generalized Canonical Correlation Analysis

Arthur Tenenhaus

SUPELEC Sciences des Systèmes (E3S)-Département Signaux et Systèmes Electroniques
3, rue Joliot Curie, Plateau de Moulon, 91192 Gif-sur-Yvette Cedex

Résumé Un problème classique en statistique est d'étudier les liens entre plusieurs blocs de variables. L'objectif est de trouver les variables d'un bloc influençant les variables d'autres blocs. L'Analyse Canonique Généralisée Régularisée (ACGR) est un cadre très attractif pour traiter ce type de problématique. Cependant, l'ACGR ne capture que des relations linéaires entre blocs et pour accéder à des liens non linéaires nous proposons une extension à noyau de l'ACGR.

Mots-clés: Méthodes à noyau, Analyse de tableaux Multiples, Analyse Canonique Généralisée Régularisée, Approche PLS

Abstract A classical problem in statistics is to study relationships between several blocks of variables. The goal is to find variables of one block directly related to variables of other blocks. The Regularized Generalized Canonical Correlation Analysis (RGCCA) is a very attractive framework to study such a kind of relationships between blocks. However, RGCCA captures linear relations between blocks and to assess nonlinear relations we propose a kernel extension of RGCCA.

Keywords: Kernel Methods, Multiblock Data Analysis, Regularized Generalized Canonical Correlaton Analysis, PLS approach

1 Introduction

A common problem in applied statistics is to relate several blocks of variables to each other in order to find variables of one block directly related to variables of other blocks. Typical examples are found in large variety of fields such as bioinformatics, sensory analysis, marketing, food research.... To study such a kind of relationships between blocks, the starting point of the paper is the generalized canonical correlation analysis (gCCA) proposed in [4]. This version of gCCA captures linear relationships between blocks and to assess nonlinear relations, we propose in this paper a kernel extension of gCCA. The paper is organized as follows: the first part presents the initial formulation of gCCA and the second part is devoted to its nonlinear version.

2 Population generalized canonical correlation analysis

Let us consider J random p_j-dimensional centered column vectors $\boldsymbol{x}_j$ and J p_j-dimensional non random column vectors ξ_j. We also consider a network of connections between the random vectors by defining a design matrix $\mathbf{C} = (c_{jk})$: $c_{jk} = 1$ if $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$ are connected and 0 otherwise. Now consider two linear combinations $\eta_j = \xi_j^t \boldsymbol{x}_j$ and $\eta_k = \xi_k^t \boldsymbol{x}_k$. The correlation between η_j and η_k is :

$$\rho(\xi_j^t \boldsymbol{x}_j, \xi_k^t \boldsymbol{x}_k) = \frac{\xi_j^t \boldsymbol{\Sigma}_{jk} \xi_k}{(\xi_j^t \boldsymbol{\Sigma}_{jj} \xi_j)^{1/2} (\xi_k^t \boldsymbol{\Sigma}_{kk} \xi_k)^{1/2}} \tag{1}$$

where $\boldsymbol{\Sigma}_{jj} = \mathbb{E}\left(\boldsymbol{x}_j \boldsymbol{x}_j^t\right)$ and $\boldsymbol{\Sigma}_{jk} = \mathbb{E}\left(\boldsymbol{x}_j \boldsymbol{x}_k^t\right)$.
The population generalized canonical correlation analysis is defined as the following optimization problem [4]:

$$\begin{cases} \underset{\xi_1, \xi_2, ..., \xi_J}{\operatorname{argmax}} \sum_{1 \leq j < k \leq J} c_{jk} \ \mathrm{g}\left(\rho(\xi_j^t \boldsymbol{x}_j, \xi_k^t \boldsymbol{x}_k)\right) \\ \text{s.c.} \quad \operatorname{var}(\xi_j^t \boldsymbol{x}_j) = 1, j = 1, ..., J \end{cases} \tag{2}$$

where g is the identity, the absolute value or the square function. Problem (2) is equivalent to the following optimization problem:

$$\begin{cases} \underset{\xi_1, \xi_2, ..., \xi_J}{\operatorname{argmax}} \sum_{1 \leq j < k \leq J} c_{jk} \ \mathrm{g}\left(\xi_j^t \boldsymbol{\Sigma}_{jk} \xi_k\right) \\ \text{s.c.} \quad \xi_j^t \boldsymbol{\Sigma}_{jj} \xi_j = 1, j = 1, ..., J \end{cases} \tag{3}$$

By considering the derivates with respect to ξ_j and λ_j of the Lagrangian function associated to optimization problem (3), we obtain J stationnary equations :

$$\frac{1}{\varphi} \boldsymbol{\Sigma}_{jj}^{-1} \sum_{k=1, k \neq j}^{J} c_{jk} \ \mathrm{g}'(\xi_j^t \boldsymbol{\Sigma}_{jk} \xi_k) \boldsymbol{\Sigma}_{jk} \xi_k = \lambda_j \xi_j, \qquad j = 1, ..., J \tag{4}$$

subject to the constraints:

$$\xi_j^t \boldsymbol{\Sigma}_{jj} \xi_j = 1, \qquad j = 1, ..., J \tag{5}$$

2.1 Regularized Generalized Canonical Correlation Analysis

In practice, we have to estimate the ξ_js' from a finite sample. Let's consider J blocks $\mathbf{X}_1, ..., \mathbf{X}_J$ of centered variables measured on a set of n individuals (a row of $\mathbf{X}_j$ represents a realization of the row-random vector $\boldsymbol{x}_j^t$). $\mathbf{C} = \{c_{jk}\}$ is now a design matrix describing a network of relationships between blocks: $c_{jk} = 1$ for two connected blocks, and 0 otherwise. In the case of high multi-colinearity or when the number of observations is

smaller than the number of variables, the sample covariance matrix $\mathbf{S}_{jj} = \frac{1}{n}\mathbf{X}_j^t\mathbf{X}_j$ is a bad estimation of the true covariance matrix $\mathbf{\Sigma}_{jj}$. A suggestion for finding a better estimation of the true covariance matrix is to consider the class of linear combinations $\{\hat{\mathbf{S}}_{jj} = \tau_j\mathbf{I} + (1-\tau_j)\mathbf{S}_{jj}\}$ of the identity matrix $\mathbf{I}$ and the sample covariance matrix $\mathbf{S}_{jj}$ [3]. We then consider a sample version of stationnary equations (4) with the constraints (5) in replacing $\mathbf{\Sigma}_{jj}$ by $\hat{\mathbf{S}}_{jj}$ and $\mathbf{\Sigma}_{jk}$ by $\mathbf{S}_{jk} = \frac{1}{n}\mathbf{X}_j^t\mathbf{X}_k$. This leads to J sample stationary equations which are the stationary equations associated to the regularized generalized canonical correlation analysis (RGCCA) defined below:

$$\begin{cases} \underset{\mathbf{a}_1,\mathbf{a}_2,...,\mathbf{a}_J}{\text{argmax}} \sum_{1\leq j<k\leq J} c_{jk}\ \mathrm{g}(\widehat{\mathrm{cov}}(\mathbf{X}_j\mathbf{a}_j, \mathbf{X}_k\mathbf{a}_k)) \\ \text{s.c.} \quad (1-\tau_j)\widehat{\mathrm{var}}(\mathbf{X}_j\mathbf{a}_j) + \tau_j\|\mathbf{a}_j\|^2 = 1, j = 1, ..., J \end{cases} \tag{6}$$

The regularisation parameters $\tau_j \in [0,1]$, $j = 1, ..., J$ interpolate smoothly between the maximisation of the covariance (all $\tau_j s = 1$) and the maximisation of the correlation (all $\tau_j s = 0$). More details on RGCCA can be found in [4]. This formulation of RGCCA detects only linear relations between blocks and in the next section, we introduce a kernel extension of RGCCA allowing extracting nonlinear relations between blocks.

3 Kernel generalized Canonical Correlation Analysis

For each $j = 1, ..., J$, let $\mathcal{H}_j$ be a Reproducing Kernel Hilbert space (RKHS) with associated kernel $k_j(\cdot,\cdot)$ and feature map $\Phi_j(x) = k_j(\cdot, x)$. We define Kernel Generalized Canonical Correlation Analysis (KGCCA) as the following optimization problem.

$$\begin{cases} \underset{f_1,...,f_J \in \mathcal{H}_1\times...\times\mathcal{H}_J}{\text{argmax}} \sum_{1\leq j<k\leq J} c_{jk}\ \mathrm{g}\left(\rho\left(f_j(\boldsymbol{x}_j), f_k(\boldsymbol{x}_k)\right)\right) \\ \\ \text{s.c.} \quad \mathrm{var}(f_j(\boldsymbol{x}_j)) = 1, j = 1, ..., J \end{cases} \tag{7}$$

where

$$\rho(f_j(\boldsymbol{x}_j), f_k(\boldsymbol{x}_k)) = \frac{\mathrm{cov}(f_j(\boldsymbol{x}_j), f_k(\boldsymbol{x}_k))}{\mathrm{var}(f_j(\boldsymbol{x}_j))^{1/2}\ \mathrm{var}(f_k(\boldsymbol{x}_k))^{1/2}} \tag{8}$$

is the correlation between the random variables $f_j(\boldsymbol{x}_j)$ and $f_k(\boldsymbol{x}_k)$. The functions f_j and f_k are decided up to scale.
From the reproducing property of RKHS, $\rho(f_j(\boldsymbol{x}_j), f_k(\boldsymbol{x}_k)) = \mathrm{cor}(\langle\Phi_j(\boldsymbol{x}_j), f_j\rangle, \langle\Phi_k(\boldsymbol{x}_k), f_k\rangle)$. Therefore, $\rho(f_j(\boldsymbol{x}_j), f_k(\boldsymbol{x}_k))$ is the correlation between one dimensional linear projection of $\Phi_j(\boldsymbol{x}_j)$ onto f_j and $\Phi_k(\boldsymbol{x}_k)$ onto f_k. The canonical correlation between $\Phi_j(\boldsymbol{x}_j)$ and $\Phi_j(\boldsymbol{x}_j)$ is exactly the maximal possible correlation between these two projections.
Now, the main objective is to estimate $f_1, ..., f_J$ from a finite sample. For each $j = 1, ..., J$, let $\{x_j^1, ..., x_j^n\}$ be a set of n observations of $\boldsymbol{x}_j$, $\tilde{\mathbf{K}}_j$ be the associated Gram matrix defined as $(\tilde{\mathbf{K}}_j)_{rs} = k_j(x_j^r, x_j^s)$ and $\mathbf{K}_j$ be the Gram matrix of the centered data points in the

feature space defined as $\mathbf{K}_j = \mathbf{P}\tilde{\mathbf{K}}_j\mathbf{P}$ where $\mathbf{P} = \mathbf{I} - \frac{1}{n}\mathbf{1}$, where $\mathbf{1}$ is a $n \times n$ matrix composed of 1. For fixed f_j and f_k, it can be shown [1] that the empirical covariance of the projections in the feature space can be written as:

$$\widehat{\text{cov}}\left(\langle\Phi_j(\boldsymbol{x}_j), f_j\rangle, \langle\Phi_k(\boldsymbol{x}_k), f_k\rangle\right) = \frac{1}{n}\boldsymbol{\alpha}_j^t\mathbf{K}_j\mathbf{K}_k\boldsymbol{\alpha}_k \tag{9}$$

where $\mathbf{K}_j$ and $\mathbf{K}_k$ are the Gram matrices associated with the data sets $\{x_j^i\}_{i=1}^n$ and $\{x_k^i\}_{i=1}^n$ respectively. Combining optimization problem (7) with equation (9) and adding constraint on the smoothness of the f_j through $\|f_j\|_{\mathcal{H}_j} = \boldsymbol{\alpha}_j^t\mathbf{K}_j\boldsymbol{\alpha}_j$, the empirical counterpart of KGCCA becomes that of performing the following maximization :

$$\begin{cases} \underset{\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_J}{\text{argmax}} \sum_{1\le j<k\le J} c_{jk}\ \text{g}(\widehat{\text{cov}}(\mathbf{K}_j\boldsymbol{\alpha}_j, \mathbf{K}_k\boldsymbol{\alpha}_k)) \\ \text{s.c.} \quad \boldsymbol{\alpha}_j^t\left((1-\tau_j)\frac{1}{n}\mathbf{K}_j^2 + \tau_j\mathbf{K}_j\right)\boldsymbol{\alpha}_j = 1, j = 1, ..., J \end{cases} \tag{10}$$

The regularisation parameters not only makes this optimisation problem well posed numerically and statistically ([1];[2]) but also provide control over the capacity of the function space where the solution is sought. The larger the values of τ_j are, the less sensitive the method to the input data is and the more stable (less prone to finding spurious relations) the solution becomes. Since the Gram matrices $\mathbf{K}_j$, $j = 1, ..., J$ are symmetric semidefinite, an (incomplete) cholesky decomposition is applied to solve optimimization problem (10) efficiently. Therefore, $\mathbf{K}_j = \mathbf{R}_j^t\mathbf{R}_j$ where $\mathbf{R}_j$ is a $\text{rank}(\mathbf{K}_j) = n_j \times n$ upper triangular matrix. Therefore, by noting $\mathbf{w}_j = \mathbf{R}_j\boldsymbol{\alpha}_j$ optimization problem (10) becomes

$$\begin{cases} \underset{\mathbf{w}_1,\ldots,\mathbf{w}_J}{\text{argmax}} \sum_{1\le j<k\le J} c_{jk}\ \text{g}(\widehat{\text{cov}}(\mathbf{R}_j^t\mathbf{w}_j, \mathbf{R}_k^t\mathbf{w}_k)) \\ \text{s.c.} \quad \mathbf{w}_j^t\left((1-\tau_j)\frac{1}{n}\mathbf{R}_j\mathbf{R}_j^t + \tau_j\mathbf{I}_{n_j}\right)\mathbf{w}_j = 1, j = 1, ..., J \end{cases} \tag{11}$$

A new procedure is proposed to solve (11). The first step of the procedure is to cancel the derivatives of the Lagrangian function related to the maximization problem (11) with respect to $\mathbf{w}_j$ and λ_j. This yields to the following J stationary equations:

$$\frac{1}{\varphi}\mathbf{N}_j^{-1}\sum_{k=1,k\neq j}^{J} c_{jk}\ \text{g}'(\frac{1}{n}\mathbf{w}_j^t\mathbf{R}_j\mathbf{R}_k^t\mathbf{w}_k)\mathbf{R}_j\mathbf{R}_k^t\mathbf{w}_k = \lambda_j\mathbf{w}_j, \qquad j = 1, ..., J \tag{12}$$

subject to the constraints:

$$\mathbf{w}_j^t\mathbf{N}_j\mathbf{w}_j = 1, \qquad j = 1, ..., J \tag{13}$$

where $\mathbf{N}_j = (1-\tau_j)\frac{1}{n}\mathbf{R}_j\mathbf{R}_j^t + \tau_j\mathbf{I}_{n_j}$.
From the PLS terminology, we introduce for each block an outer component $\mathbf{y}_j = \mathbf{R}_j^t\mathbf{w}_j$ and an inner component $\mathbf{z}_j$ defined as follows:

$$\mathbf{z}_j = \frac{1}{\varphi}\sum_{j=1,k\neq j}^{J} c_{jk}\ \text{g}'\left(\widehat{\text{cov}}(\mathbf{y}_j, \mathbf{y}_k)\right)\mathbf{y}_k \tag{14}$$

Then combining equations (12), (14) and (13), we obtain the outer weights:

$$\mathbf{w}_j = \left[\mathbf{z}_j^t \mathbf{R}_j^t \mathbf{N}_j^{-1} \mathbf{R}_j \mathbf{z}_j\right]^{-1/2} \mathbf{N}_j^{-1} \mathbf{R}_j \mathbf{z}_j, j = 1, ..., J \tag{15}$$

We then propose an iterative "PLS style" algorithm for finding a solution of the J stationary equations (see Table 1). This algorithm is monotonically convergent that means that the bounded criterion to be maximized is increasing at each step of the procedure (the proof of this result is beyond the scope of the paper).

Table 1: Kernel Generalized Canonical Correlation Analysis Algorithm

A. Initialisation

- *A.1. Choose J arbitrary vectors* $\tilde{\mathbf{w}}_1^{(0)}, ..., \tilde{\mathbf{w}}_J^{(0)}$
- *A.2. Compute vectors* $\mathbf{w}_1^{(0)}, ..., \mathbf{w}_J^{(0)}$ *verifying the constraints as:*

$$\mathbf{w}_j^{(0)} = \left[(\tilde{\mathbf{w}}_j^{(0)})^t \mathbf{N}_j^{-1} \tilde{\mathbf{w}}_j^{(0)}\right]^{-1/2} \mathbf{N}_j^{-1} \tilde{\mathbf{w}}_j^{(0)}$$

- *A.3. Compute the outer components* : $\mathbf{y}_1^{(0)} = \mathbf{R}_1^t \mathbf{w}_1^{(0)}, ..., \mathbf{y}_J^{(0)} = \mathbf{R}_J^t \mathbf{w}_J^{(0)}$

B. Inner component for the block $\mathbf{X}_i$

- *Compute the inner component according to the choice og g*:

$$\mathbf{z}_j^{(s)} = \frac{1}{\varphi}\left[\sum_{k=1}^{j-1} c_{jk} g'\left[\mathrm{cov}\left(\mathbf{y}_j^{(s)}, \mathbf{y}_k^{(s+1)}\right)\right]\mathbf{y}_k^{(s+1)} + \sum_{k=j+1}^{J} c_{jk} g'\left[\mathrm{cov}\left(\mathbf{y}_j^{(s)}, \mathbf{y}_k^{(s)}\right)\right]\mathbf{y}_k^{(s)}\right]$$

  where $g'(x)/\varphi = 1$ for g equal to the identity, x for g equal to the square and $\mathrm{sign}(x)$ for g equal to the absolute value.

C. Outer component for the block

- *C.1. Compute the outer weight*: $\mathbf{w}_j^{(s+1)} = \left[(\mathbf{z}_j^{(s)})^t \mathbf{R}_j^t \mathbf{N}_j^{-1} \mathbf{R}_j \mathbf{z}_j^{(s)}\right]^{-1/2} \mathbf{N}_j^{-1} \mathbf{R}_j \mathbf{z}_j^{(s)}$
- *C.2. Compute the outer component*: $\mathbf{y}_j^{(s+1)} = \mathbf{R}_j^t \mathbf{w}_j^{(s+1)}$

The procedure begins by an arbitrary choice of initialisation (A). Suppose the outer components $\mathbf{y}_1^{(s+1)}, \mathbf{y}_2^{(s+1)}, ..., \mathbf{y}_{j-1}^{(s+1)}$ are constructed for the blocks $1, ..., j-1$. The outer component $\mathbf{y}_j^{(s+1)}$ is computed by considering the inner component $\mathbf{z}_j^{(s)}$ for the block j (B), then by considering the outer one (C). The procedure is iterated until convergence.

The essential feature of this algorithm is that each replacement is optimal, and sequential, that is to say that $\mathbf{w}_j^{(s)}$ must be replaced by $\mathbf{w}_j^{(s+1)}$ before replacing $\mathbf{w}_{j+1}^{(s)}$. This is the essence of the Gauss-Seigel Algorithm; this sequential approach leads to the monotone convergence of this algorithm.

4 Conclusion

KGGCA is a very attractive framework for non linear multiblock data analysis and covers a large spectrum of existing methods depending on the values of the regularisation parameters. Moreover, the proposed algorithm is computationally efficient but have some limitations: (1) There is no guarantee that the algorithm converges towards a fixed point of the stationary equations. (2) There is no guarantee that the algorithm converges towards a global optimum of the criterion.

References

[1] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2003.

[2] K. Fukumizu, F. R. Bach, and A. Gretton. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8:361–383, 2007.

[3] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

[4] A. Tenenhaus and M. Tenenhaus. Regularized Generalized Canonical Correlation Analysis. *Psychometrika*, in revision.