Kernel PLS path modelling

Arthur Tenenhaus

Kernel PLS path modelling

Arthur Tenenhaus
Supélec, Department of signal processing and electronics systems
Gif-sur-Yvette, 91192, France.
E-mail: arthur.tenenhaus@supelec.fr

Abstract: This paper deals with a kernel extension of a companion paper entitled "new criterion based PLS path modeling approach to structural equation modeling" [Tenenhaus, 2009] which will be presented during the PLS'09 conference.

Keywords: Multi-block data analysis, Structural Equation Modelling, PLS path modelling, Kernel methods.

1. Introduction

Throughout this paper $X_1, X_2, \ldots, X_J$ are matrices with columns defined by centered variables observed on the same set of n individuals and $K_1, K_2, \ldots, K_J$ the associated kernel matrices. Let us note p_j the number of variables for the block X_j. We are interested in studying the relation between J blocks of variables $X_1, X_2, \ldots, X_J$ even though those data sets live in different spaces. We achieve this by projecting the J blocks onto J separate directions specified by vector $a_1, a_2, \ldots, a_J$, to obtain J univariate variables on which "link" (such as covariance) can be computed easily. In this paper is presented a kernel version of a companion paper entitled "A criterion based PLS approach to structural equation modelling" [Tenenhaus and Tenenhaus, 2009] which will be presented during the PLS'09 conference. The paper is organized as follows: the first part briefly introduces the primal formulation of the proposed method and the second part is devoted to its dual formulation.

2. The general optimization problem (primal form)

In this paper, the following general problem is considered:

$$\{a_1, a_2, \ldots, a_J\} = \underset{a_1, a_2, \ldots, a_J}{\operatorname{argmax}} \sum_{i=1}^{J} \sum_{j=1}^{J} c_{ij}\, g\left(cov(X_i a_i, X_j a_j)\right) \tag{1}$$

subject to the constraints: : $(1-\tau_j) var(X_j a_j) + \tau_j \|a_j\|^2 = 1, \quad j = 1, \ldots, J$

where $g(x) = x$ (Horst Scheme), $g(x) = x^2$ (factorial scheme) or $g(x) = |x|$ (centroïd scheme).

The regularisation parameters $\tau_j \in [0,1]$, $j = 1, \ldots, J$ control the stability of the solution (specifically useful in a high dimensional block context) and interpolate smoothly between the maximisation of the covariance (all $\tau_j s' = 1$) and the maximisation of the correlation (all $\tau_j s' = 0$). The $J \times J$ design matrix $(C)_{ij} = c_{ij}$ equals to 1 if the blocks X_i and X_j are connected and to 0 otherwise allows to immediatly extend the multiblock data analysis framework to the structural equation modelling one by limiting the sommation to connected blocks. For sake of simplicity, in this part of the paper, we

focus our presentation on the Horst scheme. Applying the Lagrangian multiplier technique to the optimization of (1) gives us J stationary equations:

$$a_i = \frac{1}{\sqrt{Z_i' X_i M_i^{-1} X_i' Z_i}} \left(\underbrace{(1-\tau_i)\frac{1}{n} X_i' X_i + \tau_i I_{p_i}}_{M_i} \right)^{-1} \underbrace{X_i' \sum_{j=1}^{J} c_{ij} X_j a_j}_{Z_i}, \quad i = 1, \ldots, J$$

It is noteworthy that M_i is well-conditionned (even in high dimensional block(s) context) since the empirical covariance matrices $\frac{1}{n} X_i' X_i \;\; i = 1, \ldots, J$ are shrunk gradually towards the identity matrices and are singular and thus invertible. To solve this optimization problem is applied an iterative "PLS style" algorithm based on the Wold procedure [Wold, 1982] which is monotonically convergent that means that the bounded criterion to be maximized is increasing at each step of the procedure [Hanafi, 2007]. However, it is still remain difficult to apply such a Wold algorithm in a high dimensional block context due to the inversion of matrices M_i of dimension $p_i \times p_i$, $i = 1, \ldots, J$ and we propose in the next section to reformulate this algorithm in its dual form leading to the Kernel PLS path modelling.

3. Kernel PLS path modelling

Let us assume that a_j can be expressed as a linear combination of the observations of block X_j (which is always possible). Throughout this section let us note $a_j = X_j' \alpha_j$ and $K_j = X_j X_j'$ the matrix of inner product between pairs of observations of block X_j (Gram matrix). We consider also the QR-decomposition of the matrix $X_j' = Q_j R_j$ where Q_j is an orthonormal matrix and R_j a $\mathrm{rank}(X_j) \times n$ upper triangular matrix. This gives the following decomposition for the kernel matrix $K_j = R_j' R_j$. As suggested by [Bach and Jordan, 2002 ; Shawe-Taylor and Cristianini, 2004], an incomplete cholesky decomposition of K_j is used to obtain R_j.

Now the primal formulation of the general optimisation problem described previously can be formulated in its dual form as follows:

$$\{\alpha_1, \alpha_2, \ldots, \alpha_J\} = \underset{\alpha_1, \alpha_2, \ldots, \alpha_J}{\operatorname{argmax}} \sum_{i=1}^{J} \sum_{j=1}^{J} c_{ij}\, g\left(cov\left(X_i X_i' \alpha_i, X_j X_j' \alpha_j\right)\right)$$

subject to the constraints: : $\left(1-\tau_j\right) var\left(X_j X_j' \alpha_j\right) + \tau_j \left\| X_j' \alpha_j \right\|^2 = 1, \quad j = 1, \ldots, J$

To simplify the presentation, we focus on the Horst Scheme ($g(x) = x$)

$$\Rightarrow \qquad \{\alpha_1, \alpha_2, \ldots, \alpha_J\} = \underset{\alpha_1, \alpha_2, \ldots, \alpha_J}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{J} \sum_{j=1}^{J} c_{ij} \alpha_i' K_i K_j \alpha_j$$

subject to the constraints : $\alpha_j' \left(\left(1-\tau_j\right) \frac{1}{n} K_j^2 + \tau_j K_j \right) \alpha_j = 1, \quad j = 1, \ldots, J$

We note that this maximization problem is expressed only in terms of kernel matrices. The regularisation parameters not only makes this optimisation problem well posed numerically 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.

$$\Leftrightarrow \qquad \{\alpha_1, \alpha_2, \dots, \alpha_J\} = \underset{\alpha_1,\alpha_2,\dots,\alpha_J}{\operatorname{argmax}} \frac{1}{n}\sum_{i=1}^{J}\sum_{j=1}^{J} \alpha_i' R_i' R_i R_j' R_j \alpha_j$$

subject to the constraints : $\alpha_j' \left((1-\tau_j)\frac{1}{n} R_j' R_j R_j' R_j + \tau_j R_j' R_j \right) \alpha_j = 1, \quad j = 1, \dots, J$

Let us note $w_j = R_j \alpha_j$ and $n_j = \operatorname{rank}(X_j)$

$$\Leftrightarrow \qquad \{w_1, w_2, \dots, w_J\} = \underset{w_1,w_2,\dots,w_J}{\operatorname{argmax}} \frac{1}{n}\sum_{i=1}^{J}\sum_{j=1}^{J} w_i' R_i R_j' w_j \tag{2}$$

subject to the constraints : $w_j' \left((1-\tau_j)\frac{1}{n} R_j R_j' + \tau_j I_{n_j} \right) w_j = 1, \quad j = 1, \dots, J$ (3)

A 3-step procedure is proposed to maximize (2) subject to (3):

(1) Construct the Lagrangian function related to the maximization problem.

$$L = \sum_{i=1}^{J}\sum_{j=1}^{J} c_{ij} \frac{1}{n} w_i' R_i R_j' w_j - \sum_{i=1}^{J} \lambda_i \left((1-\tau_i)\frac{1}{n} w_i' R_i R_i' w_i + \tau_i w_i' w_i - 1 \right)$$

where λ_i, $i = 1, \dots, J$ are the Lagrangian multiplier.

(2) Define stationary equations by cancelling and simplifying the derivatives of the Lagrangian function.

$$w_i = \frac{1}{\sqrt{Z_i' R_i' N_i^{-1} R_i Z_i}} \underbrace{\left((1-\tau_i)\frac{1}{n} R_i R_i' + \tau_i I_{n_i} \right)}_{N_i}{}^{-1} R_i \underbrace{\sum_{j=1}^{J} c_{ij} R_j' w_j}_{Z_i}, \quad i = 1, \dots, J$$

(3) Find a solution of the J stationary equations by using a PLS style iterative procedure (derived to the Wold procedure which guaranty the monotonic convergence).

It is worth pointing out that for specific values of the regularization parameters, the proposed method provides as particular cases (and among others):

- For the 2 blocks cases: Kernel PLS [Rosipal et al., 2001] ($\tau_1 = \tau_2 = 1$), Kernel Redundancy Analysis [Takane and Hwang, 2007] which is equivalent to Kernel PLS for discrimination [Rosipal et al., 2003] ($\tau_1 = 1$ and $\tau_2 = \tau$), Kernel Canonical Correlation Analysis [e.g. Bach and Jordan, 2002] ($\tau_1 = \tau_1$ and $\tau_2 = \tau_2$).
- For the J blocks cases: a "Horst" Generalized Kernel CCA with the flavour of the one proposed by [Bach and Jordan, 2002 ; Shawe-Taylor and Cristianini, 2004] (all $\tau_j s' = \tau_j$; Horst scheme).

Moreover, our general optimisation problem provides a "regularized" kernel extension of the Kettenring's generalized CCA (all $\tau_j s' = \tau_j$ and Factorial scheme).

4. Conclusion

In this paper we present a very general optimization problem covering a large spectrum of methods. This paper provides a criterion point of view of the PLS path modelling framework.
To the best of our knowledge, all the Kernel CCA version proposed by the machine learning community does not consider other scheme than the Horst one. In this paper, we explore the factorial and the centroïd scheme which sound more reasonable when the number of block is greater than 2. Moreover, the introduction of the design matrix allows analysing data where all blocks are not necessarily connected.
To conclude this paper, we note that by using non linear kernel such as Gaussian or polynomial kernel, we can assess non linear relation between blocks.

5. References

F.R. Bach and M.I. Jordan. (2002): Kernel Independent Component Analysis. *Journal of Machine Learning Research*, 3:1-48.

Hanafi M. (2007): PLS Path modelling: computation of latent variables with the estimation mode B, *Computational Statistics*, 22, 275-292.

Horst, P., (1961). Relations among M sets of variables. *Psychometrika* 26, 126–149.

Kettenring J.R. (1971): Canonical analysis of several sets of variables, *Biometrika*, 58, 433-451.

Rosipal, R. and Trejo, L. J. (2001). Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space. *Journal of Machine Learning Research*, 2 :97–123.

Rosipal, R., Trejo, L. J., and Matthews, B. (2003). Kernel PLS-SVC for linear and nonlinear classification. *In Proceeding of the twentieth international conference on machine learning*

Shawe-Taylor, J. and Cristianini, N. (2004) *Kernel Methods for Pattern Analysis*. Cambridge University Press.

Y. Takane and H. Hwang. (2007): Regularized linear and kernel redundancy analysis. *Computational Statistics and Data Analysis*, 52:394-405, 2007.

M. Tenenhaus and A. Tenenhaus (2009). A criterion based PLS approach to structural equation modelling, *6th International Conference on Partial Least Squares, Beijing*

Wold H. (1982): "Soft Modeling: The Basic Design and Some Extensions", in *Systems under indirect observation, Part 2*, K.G. Jöreskog & H. Wold (Eds), North-Holland, Amsterdam, pp. 1-54.