View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Diposit Digital de Documents de la UAB

SORT 27 (2) July-December 2003, 165-174

Asymptotic study of canonical correlation analysis: from matrix and analytic approach to operator and tensor approach

Jeanne Fine*

Université Paul Sabatier, France

Abstract

Asymptotic study of canonical correlation analysis gives the opportunity to present the different steps of an asymptotic study and to show the interest of an operator and tensor approach of multidimensional asymptotic statistics rather than the classical, matrix and analytic approach. Using the last approach, Anderson (1999) assumes the random vectors to have a normal distribution and the non zero canonical correlation coefficients to be distinct. The new approach we use, Fine (2000), is coordinate-free, distribution-free and permits to have no restriction on the canonical correlation coefficients multiplicity order. Of course, when vectors have a normal distribution and when the non zero canonical correlation coefficients are distinct, it is possible to find again Anderson's results but we diverge on two of them. In this methodological presentation, we insist on the analysis frame (Dauxois and Pousse, 1976), the sampling model (Dauxois, Fine and Pousse, 1979) and the different mathematical tools (Fine, 1987, Dauxois, Romain and Viguier, 1994) which permit to solve problems encountered in this type of study, and even to obtain asymptotic behavior of the analyses random elements such as principal components and canonical variables.)

MSC: 62E20, 62H20, 62H25, 47N30

Keywords: multivariate analysis, canonical correlation analysis, asymptotic study, operator, coordinate-free, distribution-free

1 Classical approach

1.1 Population canonical correlation analysis

Let X and Y be two random vectors, p and q dimensional respectively ($p \leq q$) defined on a same probability space $(\Omega, \mathcal{A}, P)$, centered and admitting order 4 moments. We

* *Address for correspondence:* Laboratoire de Statistique et Probabilités, Université Paul Sabatier, 118, route de Narbonne, 31062 Toulouse cedex, France. e-mail: fine@cict.fr.

Received: October 2003
Accepted: December 2003

assume the matrix covariance V_X of X to be non-singular and we denote by H_X the vector space of real-valued random variables (r.r.v.) linear combinations of X components. We introduce similarly V_Y and H_Y and we denote by V_{XY} the cross covariance the components of X with the ones of Y.

The aim of canonical correlation analysis (CCA) of (X, Y) is to measure the relationships between X and Y. CCA may be defined as the search for f_1 and g_1, r.r.v. of H_X and H_Y with unit variance and maximal correlation ρ_1 then, iteratively, for $j = 2, ..., r$, $(r \leq p)$, as the search of f_j and g_j, r.r.v. of H_X and H_Y with unit variance, uncorrelated with the $(f_k)_{k<j}$ and $(g_k)_{k<j}$ and with maximal correlation ρ_j. The r.r.v. f_j and g_j are called j^{th} *canonical variables* and the real ρ_j of [0,1] is called j^{th} *canonical correlation coefficient*.

Let $R_X = V_X^{-\frac{1}{2}} V_{XY} V_Y^{-1} V_{YX} V_X^{-\frac{1}{2}}$ and the same for R_Y permuting X and Y roles. It is easy to verify that R_X and R_Y have the same non-zero eigenvalues denoted by $(\rho_j^2)_{j=1,...,r}$ when written in a decreasing order. We set : $\lambda_j = \rho_j^2$, for $j = 1, ..., r$, and, except in the particular case $r = p = q$, we set $\lambda_j = \rho_j = 0$ for $j > r$. For $j > r$ we define f_j and g_j as r.r.v. of H_X and H_Y respectively with unit variance and uncorrelated with the $(f_k)_{k<j}$ and $(g_k)_{k<j}$.

CCA of (X, Y) is then :

$$((\rho_j)_{j=1,...,r+1}, (f_j)_{j=1,...,p}, (g_j)_{j=1,...,q}). \tag{1}$$

CCA of (X, Y) depends only on H_X and H_Y, which are also generated by components of $X' := V_X^{-\frac{1}{2}} X$ and $Y' := V_Y^{-\frac{1}{2}} Y$ respectively. We show that, if $(u_j)_{j=1,...,p}$ and $(v_j)_{j=1,...,q}$ denote unit eigenvectors bases of R_X and R_Y associated with $(\lambda_j)_{j=1,...,p}$ and $(\lambda_j)_{j=1,...,q}$ respectively, we can obtain canonical variables f_j and g_j as linear combinations of X' and Y' components, using u_j and v_j components as coefficients, that is, by setting: $f_j = \langle u_j, X' \rangle_p$ et $g_j = \langle v_j, Y' \rangle_q$, where $\langle ., . \rangle_p$ and $\langle ., . \rangle_q$ denote $\mathbb{R}^p$ and $\mathbb{R}^q$ usual scalar products.

Decomposition (1) is not unique because each canonical variable associated with a simple eigenvalue may be replaced by its opposite and the set of canonical variables associated with a multiple eigenvalue may be replaced by an other set according to the choice of R_X and R_Y eigenvectors associated with this eigenvalue.

1.2 Sample canonical correlation analysis

Let $(X_l, Y_l)_{l=1,...,n}$ be a n-sample i.i.d. as (X, Y). We index by n the elements defined previously and calculated on the sample : $\mu_X^n, \mu_Y^n, V_X^n, V_Y^n, V_{XY}^n, R_X^n, R_Y^n$.

Let $(\lambda_j^n)_{j=1,...,p}$ be the decreasing sequence of the p eigenvalues of R_X^n (and of the p largest eigenvalues of R_Y^n, the other ones, if $q > p$, being null), $(u_j^n, v_j^n)_{j=1,...,p}$ a sequence of associated unit eigenvectors of R_X^n and of R_Y^n and $(f_j^n, g_j^n)_{j=1,...,p}$ the canonical variables sequence, vectors of $\mathbb{R}^n$, obtained by :

$$\forall l \in \{1, ..., n\} \quad (f_j^n)_l = \langle u_j^n, (V_X^n)^{-\frac{1}{2}}(X_l - \mu_X^n)\rangle_p \quad \text{and} \quad (g_j^n)_l = \langle v_j^n, (V_Y^n)^{-\frac{1}{2}}(Y_l - \mu_Y^n)\rangle_q.$$

If the case arises ($q > p$), we let $\lambda_{p+1}^n = 0$, we complete $(v_j^n)_{j=1,...,p}$ with R_Y^n eigenvectors in order to obtain an orthonormal basis $(v_j^n)_{j=1,...,q}$ of $\mathbb{R}^q$ and we define the canonical variables associated.

At last, for all j in $\{1, ..., p+1\}$ let $\rho_j^n = \sqrt{\lambda_j^n}$. Sample CCA of (X, Y) is then:

$$((\rho_j^n)_{j=1,...,p+1}, (f_j^n)_{j=1,...,p}, (g_j^n)_{j=1,...,q}). \tag{2}$$

1.3 Asymptotic study

Asymptotic study of CCA consists in establishing a.s. convergence of the canonical elements sequences of the sample CCA (2) to the corresponding canonical elements of the population CCA (1) and in establishing convergence in distribution of the standardized canonical elements sequences.

Difficulties are numerous : canonical variables are estimated ("predicted") by $\mathbb{R}^n$ vectors, the space dimension increasing with sample size. Then the use is to restrict asymptotic study to R_X^n and R_Y^n eigenvectors : $(u_j^n)_{j=1,...,p}$ and $(v_j^n)_{j=1,...,q}$ respectively, called *canonical vectors* and also to the $\mathbb{R}^p$ and $\mathbb{R}^q$ vectors defined by: $x_j^n = (V_X^n)^{-\frac{1}{2}}u_j^n$ and $y_j^n = (V_Y^n)^{-\frac{1}{2}}v_j^n$ respectively, called *canonical factors*; these vectors permit to obtain directly canonical variables by:

$$\forall l \in \{1, \ldots, n\} \quad (f_j^n)_l = \langle x_j^n, X_l - \mu_X^n\rangle_p \quad \text{et} \quad (g_j^n)_l = \langle y_j^n, Y_l - \mu_Y^n\rangle_q.$$

Multiple eigenvalues case is difficult to process because the eigenvectors associated with are not uniquely defined. Then the use is to restrict asymptotic study to the case where all eigenvalues are simple. Uniqueness is then verified by choosing systematically the unit vector (between the two ones) which has the first non null coordinate in respect with the canonical basis positive.

As for all multidimensional analyses, covariance matrices of sample random matrices V_X^n, V_{XY}^n, R_X^n, ... are "super-matrices" (that is, matrices of matrices). Tools such as the "vec" operator which transforms matrix into vector, have been introduced in order to handle theses super-matrices; the difficulty comes from the necessity of fixing the order of lines and columns elements.

In other respects, we know that the sequence ($\sqrt{n}(V_X^n - V_X)$) converges in distribution to a centered normal variable, the covariance super-matrix of which is known in some special cases, when X has a normal or elliptical distribution for example.

In the CCA frame work, we need to study convergence in distribution of the sequence ($\sqrt{n}(V_Z^n - V_Z)$) with $Z = (X, Y)$, then to study convergence in distribution of the sequence ($\sqrt{n}(R_Z^n - R_Z)$) with $R_Z = (R_X, R_Y)$ and $R_Z^n = (R_X^n, R_Y^n)$, before studying convergence

of the R_Z^n eigenelements sequences and convergence of the sample canonical elements sequences. This asymptotic study is much more complex than the one of Principal Component Analysis (PCA) because principal values and principal vectors of the X PCA are eigenelements of the X covariance matrix.

It is only in 1999 that Anderson publishes a CCA asymptotic study when (X, Y) has a normal distribution and when all non zero eigenvalues are simple. Canonical factors components and canonical correlation coefficients of the population CCA are differentiable functions of V_Z. Results are then obtained from Taylor expansions. So, this classical approach may be qualified as matricial and analytic.

In order to simplify calculations, we propose to change variables from (X, Y) to (X', Y'), which is equivalent to changing the basis in $\mathbb{R}^p$ and $\mathbb{R}^q$. We then have: $V_{X'} = I_p$, and $R_X = R_{X'} = V_{X'Y'}V_{Y'X'}$, and similarly for $V_{Y'}$ and R_Y.

2 Operator and tensor approach

2.1 Introduction

Difficulties previously described are ensued only from the fact that matricial tool is not convenient. Working directly on linear operators in Euclidean spaces avoid indices problems and can be easily extended to an Hilbertian frame. Moreover, instead of studying eigenvectors associated with simple eigenvalues, it is possible to study eigenprojectors associated with multiple eigenvalues. Eaton (1983) also advices a "vector space approach" of the multidimensional statistics.

Dauxois and Pousse (1976) enlarge the PCA definition of a $\mathbb{R}^p$ random vector to a Hilbert random variable and even to a Hilbert random function, that is, a Hilbert random variable depending on a parameter in order to process temporal or spatial data. They redefine each factorial analysis (PCA, CCA, Correspondence Analysis, Discriminant Analysis, ...) in an operatorial and stochastic frame, that leads them to define, between others, nonlinear analyses.

The first asymptotic study in this frame has been realized by Romain (1979) for a Hilbert random function PCA (see also Dauxois, Pousse and Romain, 1982), study completed by Arconte (1980) who also started on CCA asymptotic study but all tools were not available to continue the study. Dauxois, Romain and Viguier (1994) propose to use some tensor products and establish a dictionary between matricial and operatorial formula. This work permits to compare common results obtained in both frames, but also to obtain more easily complex results; writings in respect with eigenvectors basis are established after concise formulations with operators.

These new tools permit to realize in Fine (2000) the CCA asymptotic study without restriction, that is, without assumption on the (X, Y) distribution, in the general case where eigenvalues may be multiple and without excluding canonical variables

asymptotic study (CCA random elements). Therefore, our approach may be qualified as an operator and tensor approach. We give below the different steps of the CCA asymptotic analysis, some tools used and some examples of results.

2.2 Different steps of the CCA asymptotic study, tools, results

1) Population CCA

First, the matter is to define CCA of a pair of Euclidean random variables (population CCA). Again, we use classical approach notations substituting $(\mathbb{R}^p, \langle ., . \rangle_p)$ and $(\mathbb{R}^q, \langle ., . \rangle_q$ for p and q dimensional Euclidean spaces $(\mathcal{X}, \langle ., . \rangle_{\mathcal{X}})$ and $(\mathcal{Y}, \langle ., . \rangle_{\mathcal{Y}})$ respectively. Obviously, we work without reference to any basis (free coordinate).

Let $L^2(P)$ the Hilbert space of r.r.v. defined on $(\Omega, \mathcal{A}, P)$ and admitting order 2 moments, scalar product of which associates $\mathbb{E}(fg)$ to (f, g).

The operator Φ_X from $\mathcal{X}$ to $L^2(P)$ which associates $\langle x, X \rangle_{\mathcal{X}}$ to x plays an essential role in the operator approach of multidimensional statistics. In particular, X is a normal Euclidean random variable if, and only if, $\forall x \in \mathcal{X}, \langle x, X \rangle_{\mathcal{X}}$ is a normal r.r.v..

The expected value of X is the unique element of $\mathcal{X}$ (Riesz theorem), denoted by $\mathbb{E}(X)$, verifying: $\forall x \in \mathcal{X}, \langle x, \mathbb{E}(X) \rangle_{\mathcal{X}} = \mathbb{E}(\langle x, X \rangle_{\mathcal{X}})$.

For all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we denote by $x \otimes y$ the operator from $\mathcal{X}$ to $\mathcal{Y}$ which associates $\langle x', x \rangle_{\mathcal{X}}\ y$ to x' ; it is an element of the Hilbert space $\sigma_2(\mathcal{X}, \mathcal{Y})$ of operators from $\mathcal{X}$ to $\mathcal{Y}$ with the scalar product : $\langle A, B \rangle_2 = tr(AB^*)$. Due to the Riesz theorem, we may then define covariance operators V_X of X, V_Y of Y, and crossed covariance operators V_{XY} and V_{YX} of X and Y: $V_X = \mathbb{E}((X - \mu_X) \otimes (X - \mu_X))$, ...

As in the CCA classical approach (§1.1), X and Y are assumed to be centered. The adjoint operator Φ_X^* of Φ_X is the operator from $L^2(P)$ to $\mathcal{X}$ which associates $\mathbb{E}(fX)$ to f and then we have : $\Phi_X^* \circ \Phi_X = V_X$, $\Phi_X^* \circ \Phi_Y = V_{XY}$, ...

It is convenient to represent operators relationships in the following commutative diagram, also called a duality scheme; here, each space is identified with its dual space. The H_X and H_Y spaces are image spaces of Φ_X and Φ_Y respectively and the orthogonal projectors of $L^2(P)$ on these subspaces are: $\Pi_X = \Phi_X \circ V_X^{-1} \circ \Phi_X^*$ and $\Pi_Y = \Phi_Y \circ V_Y^{-1} \circ \Phi_Y^*$.

$$\begin{array}{ccccc} \mathcal{X} & \overset{\Phi_X^*}{\longleftarrow} & L^2(P) & \overset{\Phi_Y^*}{\longrightarrow} & \mathcal{Y} \\ V_X^{-1} \downarrow\uparrow V_X & & \uparrow I & & V_Y \uparrow\downarrow V_Y^{-1} \\ \mathcal{X} & \overrightarrow{\Phi_X} & L^2(P) & \overleftarrow{\Phi_Y} & \mathcal{Y} \end{array}$$

Operators R_X and R_Y, and also CCA of (X, Y), are defined as previously (symbols $\circ$ are deleted in order to reduce notation).

As in the classical approach, in order to facilitate calculations, we change the scalar product on $\mathcal{X}$ so that the covariance operator of X is the identity of $\mathcal{X}$, and similarly for $\mathcal{Y}$. We then have: $R_X = V_{XY}V_{YX}$ and $R_Y = V_{YX}V_{XY}$.

2) Sample model and sample CCA

We use a sample model (Dauxois, Fine and Pousse, 1979) establishing a link between the sample used in Data Analysis and the i.i.d. sample of Statistics. A sample $(X_l, Y_l)_{l\in\mathbb{N}^*}$ i.i.d. as (X, Y) is built from an element ω of $\Omega^{\mathbb{N}^*}$ setting, for all l of $\mathbb{N}^*$ (π_l denoting the l^{th} projection of $\Omega^{\mathbb{N}^*}$ onto Ω) : $X_l = X \circ \pi_l$ and $Y_l = Y \circ \pi_l$ that is $X_l(\omega) = X(\omega_l)$ and $Y_l(\omega) = Y(\omega_l)$.

We then provide $L^2(P)$ with the scalar product (random scalar product as it depends on ω):

$$\forall (f, g) \in L^2(P) \times L^2(P), \quad \mathbb{E}_n(fg) = \frac{1}{n}\sum_{l=1}^{n} f(\omega_l)g(\omega_l).$$

Then we have:

$$\mathbb{E}_n(X) = \mu_X^n,\ \Phi_X^n = \langle ., X - \mu_X^n\rangle_X,\ V_X^n = \frac{1}{n}\sum_{l=1}^{n}(X_l - \mu_X^n)\otimes(X_l - \mu_X^n), \ldots$$

This sample model is the clue to distinguish randomness implied by the model ($L^2(P)$ elements) from randomness implied by sampling. It permits to obtain the canonical variables asymptotic distribution.

The duality scheme of sample CCA is the same as the population one after substituting $L^2(P)$ for $(L^2(P), \mathbb{E}_n)$ and indexing operators by n.

Sample operators R_X^n and R_Y^n and sample CCA are defined as previously.

3) Convergence of sample operators sequence

Limit theorems in Euclidean or Hilbert spaces permit to obtain a.s. convergence and convergence in distribution of the covariance operators sequence without assumption on the distribution of (X, Y) except the existence of order 4 moments. For CCA, we obtain (remind we let $Z = (X, Y)$):

$$W_Z^n := \sqrt{n}(V_Z^n - V_Z) \quad \xrightarrow{\mathcal{D}} \quad W_Z \sim N(0\,; \mathbb{K}_Z),$$

where $\mathbb{K}_Z$ is the covariance operator of $Z \otimes Z$.

In what concerns the sample operators $R_Z^n = (R_X^n, R_Y^n)$, elements of $\sigma_2(\mathcal{Z})$ (with $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$), a.s. convergence derives from the fact that it is possible to write R_X^n and R_Y^n as continuous function of V_Z^n.

Let $U_Z^n = \sqrt{n}(R_Z^n - R_Z)$ $(:= (U_X^n, U_Y^n))$.

We write $U_X^n = \Psi_X^n(W_Z^n)$ where (Ψ_X^n) is a sequence of random operators from $\sigma_2(\mathcal{Z})$ to $\sigma_2(\mathcal{X})$ a.s. converging to Ψ_X. We then deduce the convergence in distribution

of (U_X^n) to $U_X = \Psi_X(W_Z)$, centered normal variable, covariance operator of which being $\mathbb{L}_X = \Psi_X \circ \mathbb{K}_Z \circ \Psi_X^*$, and the same result for (U_Y^n) permuting X and Y roles. The proposition used here, is easy to prove from classical results in metric spaces (Billingsley, 1968). We obtain for example:

$$U_X = -\frac{1}{2}(W_X R_X + R_X W_X) + W_{XY} V_{YX} + V_{XY} W_{YX} - V_{XY} W_Y V_{YX} \sim N(0\,; \mathbb{L}_X)$$

4) Convergence of eigenelements and CCA elements sequences

Whatever may be the "factorial" method, which is an analysis or a model obtained from a spectral (or singular-value) decomposition, all results concerning eigenelements (eigenvalues, eigenprojectors, eigenvectors associated with simple eigenvalues, ...) are easily obtained thanks to perturbation theory of linear operators (Kato, 1980). In Fine (1987), this theory has been adapted to bounded perturbations that permits to use it, due to the iterated logarithm law, in the asymptotic study frame. So we obtain a.s. expansions of eigenelements of a symmetric positive operators sequence.

We may also consult Dossou-Gbete and Pousse (1991) for limit results but, for the convergence in distribution of some CCA elements, limit results are not sufficient when perturbation expansions permit to conclude.

For example, for the canonical factors associated to a simple eigenvalue λ_i, we have: $x_i = u_i$ because $V_X = I_{\mathcal{X}}$ and $x_i^n = (V_X^n)^{-\frac{1}{2}} u_i^n$ so:

$$\sqrt{n}(x_i^n - x_i) = -(V_X^n)^{-\frac{1}{2}}((V_X^n)^{\frac{1}{2}} + I_{\mathcal{X}})^{-1}[\,\sqrt{n}(V_X^n - I_{\mathcal{X}})]u_i^n + [\,\sqrt{n}(u_i^n - u_i)].$$

We know that $(\sqrt{n}(V_X^n - I_{\mathcal{X}}))$ converges in distribution to W_X and $(\sqrt{n}(u_i^n - u_i))$ to $S_{X_i} U_X x_i$ (with $S_{X_i} = (R_X - \lambda_i I_{\mathcal{X}})^-$) but, thanks to perturbation expansions, it is possible to establish:

$$\sqrt{n}(x_i^n - x_i) \quad \xrightarrow{\mathcal{D}} \quad \frac{1}{2} W_X x_i + S_{X_i} U_X x_i \sim N(0\,; \mathbb{L}_{Xi})$$

5) Asymptotic covariance operators in the elliptical case

We have already seen that the asymptotic covariance operator of $(\sqrt{n}(R_X^n - R_X))$ is $\mathbb{L}_X = \Psi_X \circ \mathbb{K}_Z \circ \Psi_X^*$ where $\mathbb{K}_Z$ is the asymptotic covariance operator of $(\sqrt{n}(V_Z^n - V_Z))$ and where the operator Ψ_X from $\sigma_2(\mathcal{Z})$ to $\sigma_2(\mathcal{X})$ can be written explicitly. All the distribution limits of eigenelements or CCA elements sequences are centered normal variables (or function of centered normal variables), covariance operator of which being written as function of $\mathbb{K}_Z$ in the same way.

Now, we may write explicitly these asymptotic covariance operators in the case where Z has an elliptical distribution with mean μ_Z, covariance operator V_Z and kurtosis κ (real parameter, which, when it is null, leads to a $N(\mu_Z, V_Z)$ distribution). We then know that $\mathbb{K}_Z$ is the operator from $\sigma_2(\mathcal{Z})$ to itself which associates to T:

$$\mathbb{K}_Z(T) = (1 + \kappa) V_Z (T + T^*) V_Z + \kappa \langle V_Z, T \rangle_2 V_Z.$$

At this step, we need more algebraic tools. The tensor product in spaces of type σ_2 is denoted by $\tilde{\otimes}$. For example:

$$\forall (A,B) \in \sigma_2(\mathcal{Z}) \times \sigma_2(\mathcal{Z}), \quad \forall T \in \sigma_2(\mathcal{Z}), \quad A\tilde{\otimes}B(T) = \langle T, A\rangle_2 B.$$

We define also product $\overset{\ell}{\otimes}$ in spaces of type σ_2. For example:

$$\forall (A,B) \in \sigma_2(\mathcal{Z}) \times \sigma_2(\mathcal{Z}), \quad \forall T \in \sigma_2(\mathcal{Z}), \quad A \overset{\ell}{\otimes} B(T) = BTA^*.$$

We define the commutation operator C which associates to an operator T its adjoint T^*. At last, we substitute the space product $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$ for the Hilbertian sum $\mathcal{Z} = \mathcal{X}\oplus\mathcal{Y}$; this permits to plunge all the operators into $\sigma_2(\mathcal{Z})$ in order to simplify notation. Projector P_X from $\sigma_2(\mathcal{Z})$ onto $\sigma_2(\mathcal{X})$ becomes in this frame a symmetric operator of $\sigma_2(\mathcal{Z})$.

The operator $\mathbb{K}_Z$ from $\sigma_2(\mathcal{Z})$ to itself may be written as:

$$\mathbb{K}_Z = (1+\kappa)V_Z \overset{\ell}{\otimes} V_Z(I+C) + \kappa V_Z\tilde{\otimes}V_Z.$$

Let $(x_i)_{i=1,\dots,p}$ be an orthonormal basis formed by canonical factors of $\mathcal{X}$, then $(x_i \otimes x_j)_{i,j=1,\dots,p}$ is an orthonormal basis of $\sigma_2(\mathcal{X})$ and $((x_i \otimes x_j)\tilde{\otimes}(x_k \otimes x_l))_{i,j,k,l=1,\dots,p}$ is an orthonormal basis of $\sigma_2(\sigma_2(\mathcal{X}))$.

After calculations obtained in a concise way, it is easy to decompose operators in respect with this type of basis. For example, we obtain for the asymptotic covariance operator of ($\sqrt{n}(R_X^n - R_X)$):

$$\mathbb{L}_X = (1+\kappa)(I+C)[-\frac{3}{4}R_X^2 \overset{\ell}{\otimes} I_{\mathcal{X}} + R_X^2 \overset{\ell}{\otimes} R_X + R_X \overset{\ell}{\otimes} I_{\mathcal{X}} - \frac{5}{4}R_X \overset{\ell}{\otimes} R_X](I+C)$$

and, in respect with the basis of canonical factors (remember that $(\lambda_j)_{j=1,\dots,p}$ is the decreasing sequence of eigenvalues of R_X):

$$\mathbb{L}_X = \frac{1}{2}(1+\kappa)\sum_{j=1}^{p}\sum_{k=1}^{p}\left(-\frac{3}{4}\lambda_j^2 - \frac{3}{4}\lambda_k^2 + \lambda_j^2\lambda_k + \lambda_j\lambda_k^2 + \lambda_j + \lambda_k - \frac{5}{2}\lambda_j\lambda_k\right)$$

$$(x_j \otimes x_k + x_k \otimes x_j)\tilde{\otimes}(x_j \otimes x_k + x_k \otimes x_j)$$

When (X,Y) has a normal distribution and when all eigenvalues are simple, it is possible to rediscover Anderson's results but we diverge on two of them.

6) Convergence of CCA random elements sequences

As previously announced (§ 2.1.2) the sample model permits to obtain a.s. convergence and convergence in distribution of canonical variables sequences. We have for example, for the canonical variable associated with a simple eigenvalue λ_i:

$$\sqrt{n}(f_i^n - f_i)) \quad \xrightarrow{\mathcal{D}} \quad \langle \frac{1}{2}W_X x_i + S_{X_i}U_X x_i, X\rangle_{\mathcal{X}} \sim N(0\,;\mathbb{M}_{Xi})$$

with, in the particular case where (X, Y) has an elliptical distribution:

$$\mathbb{M}_{Xi} = \frac{1}{4}(2 + 3\kappa) f_i \otimes f_i + (1 + \kappa) \sum_{j \neq i} (1 - \lambda_i)(\lambda_i + \lambda_j - 2\lambda_i\lambda_j)(\lambda_i - \lambda_j)^{-2} f_j \otimes f_j$$

7) Inferential applications and conclusion

These results on CCA asymptotic study permit to tackle easily inferential applications (confidence interval estimation, statistical tests, ...) which imply CCA elements, particularly the proximity measures built on canonical correlation coefficients. See Anderson (1999) and Dauxois and Nkiet (2002).

Further aspects and results may be consulted in Fine (2000). This methodological presentation shows that the operator approach performs quite well in solving asymptotic problems in multivariate statistics.

3 References

Anderson, T. W. (1999). Asymptotic Theory for Canonical Correlation Analysis. *Journal of Multivariate Analysis*, 70, 1-29.

Arconte, A. (1980). *Étude asymptotique de l'analyse en composantes principales et de l'analyse canonique.* Thèse de 3ème cycle, Université de Pau et des Pays de l'Adour.

Billingsley, P. (1968). *Convergence of probability measures.* Wiley, New York.

Dauxois, J., Fine, J. and Pousse A. (1979). Échantillonnage en segmentation, étude de la convergence. *Statistique et Analyse des Données,* 3, 45-53.

Dauxois, J. and Nkiet, G. M. (2002). Measures of Association for Hilbertian subspaces and some applications. *Journal of Multivariate Analysis*, 82, 263-298.

Dauxois, J. and Pousse, A. (1976). *Les analyses factorielles en calcul des probabilités et en statistique: essai d'étude synthétique.* Thèse de Doctorat d'État, Université Paul Sabatier, Toulouse.

Dauxois, J., Pousse, A. and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function; some applications to statistical inference. *Journal of Multivariate Analysis*, 12, 136-154.

Dauxois, J., Romain, Y. and Viguier, S. (1994). Tensor products and statistics. *Linear Algebra and its Applications*, 210, 59-88.

Dossou-Gbete, S. and Pousse, A. (1991). Asymptotic study of eigenelements of a sequence of random self adjoint operators. *Statistics,* 22, 479-491.

Eaton, M. L. (1983). *Multivariate statistics. A vector space approach.* Wiley, New York.

Fine, J. (1987). On the validity of the perturbation method in asymptotic theory. *Statistics,* 18, 401-414.

— (2000). Étude Asymptotique de l'Analyse Canonique. *Pub. Inst. Stat. Univ. Paris,* 44, 2-3, 21-72.

Kato, T. (1980). *Perturbation theory for linear operators.* Springer-Verlag, New York.

Romain, Y. (1979). *Étude Asymptotique des approximations par échantillonnage de l'analyse en composantes principales d'une fonction aléatoire. Quelques applications.* Thèse 3ème cycle. Université Paul Sabatier. Toulouse.

Resum

L'estudi asimptòtic de l'Anàlisi de la Correlació Canònica ens permet presentar els diferents passos de les propietats asimptòtiques i mostrar l'interès del plantejament amb operadors i tensors dels estadístics multivariants en comptes del plantejament clàssic, matricial i analític. Emprant aquesta aproximació clàssica, Anderson (1999) suposa que els vectors aleatoris segueixen la distribució normal i que els coeficients de correlacions canòniques no nuls són diferents. Fem servir un nou plantejament a lliure distribució (Fine, 2000) que també és lliure de les coordenades i que no té restriccions sobre l'ordre de multiplicitat de les coeficients de correlacions canòniques Tanmateix, quan els vectors aleatoris segueixen la distribució normal i quan les coeficients de correlacions canòniques no nul·les són diferents, podem recuperar els resultats d'Anderson, però no coincidim en dues situacions. En aquesta presentació metodològica, insistim en l'estructura analítica (Dauxois and Pousse, 1976), els models d'obtenció de mostres (Dauxois, Fine and Pousse, 1979) i diferents eines matemàtiques (Fine, 1987, Dauxois, Romain and Viguier, 1994), que permeten resoldre problemes que apareixen en aquest tipus d'estudi, i fins i tot obtenir el comportament asimptòtic dels aspectes aleatoris d'altres elements (components principals, variables canòniques, ...).

MSC: 62E20, 62H20, 62H25, 47N30

Paraules clau: Anàlisi multivariant, anàlisi de correlació canònica, estudi asimptòtic, operadors, lliure de coordenades, lliure distribució