Computer and Automation Institute Hungarian Academy of Sciences
KÖZLEMÉNYEK 28/1982

THE APPLICATION OF CANONICAL CORRELATION ANALYSIS TO PREDICT CORONARIA THROMBOSIS

Tamás Lengyel
Computing and Automation Institute of
Hungarian Academy of Science
Budapest

1. INTRODUCTION

We should like to call your attention for an application of canonical correlation analysis which is useful to predict some aspects of complex phenomena.

We dealt with prediction of a special heart-disease: the so-called coronaria thrombosis.

The data were given by the National Institute of Cardiology (Budapest). Ten variables concerning the physical status, smoking customs were measured in 3 villages, at 3 times by 5 years, in a population with size of 1082 people ([3]).

Our hypothesis was that data have information not only about the actual status but about their possible changes expected in a short term. So, we were interested in studying connection between the consequent measures and we aimed at using these relations to predict measures and status by the same variables but which had been measured 5 years earlier.

One can solve problem like mentioned above by means of canonical correlation analysis.

This method seems us to be new in the field of analysis of very short time series.

In the first part of my paper I sketch out the concept of canonical correlation analysis then I present its application to our problem. Because of the shortage of

applicable data I had to restrict myself concerning the size of analysis.

2. THE CONCEPT OF CANONICAL CORRELATION ANALYSIS

Let us denote two multidimensional stochastic vector variables by $\underline{U}_1$ and $\underline{U}_2$. We will study the relation(s) between $\underline{U}_1$ and $\underline{U}_2$. We suppose that their components are standardized, i.e. the means and the variances equal to 0 and 1, respectively.

Let us suppose as well that $\underline{U}_1$ and $\underline{U}_2$ are in R^q and R^{p-q}, respectively ($p>q>0$). Let us denote the common covariation matrix of $\underline{U}_i$ and $\underline{U}_j$ by $\Sigma_{ij}=\text{cov}(\underline{U}_i,\underline{U}_j)$ $(i,j=1,2)$. In this case the covariation equals to the correlation.

We suppose that

$$\text{rank } (\Sigma_{11})=q,$$
$$\text{rank } (\Sigma_{22})=p-q,$$

where rank (A) means the rank of matrix A.
We introduce the following notations

$$m: = \min \{q,p-q\}$$
$$k: = \text{rank } (\Sigma_{12})$$

A' means the transponent of A,
A^{-1} means the inverse of a quadratic matrix A.

2.1 There is a well known measurement between 2 one dimensional stochastic variables, the so-called correlation coefficient.

We can write the (minimal least square) regression equation in the form

$$\hat{U}_2 = rU_1$$
$$E(U_2-\hat{U}_2) = 1-r^2,$$

where $r=\text{cov}(U_1,U_2)$.

We can interpret this fact as U_1 explains a quantity of r^2 fromt the variance of U_2.

2.2 Furthermore, we know the multiple correlation coefficient as a good measure of relation between an one dimensional (U_2) and a multidimensional ($\underline{U}_1$) stoch.var. It equals to the correlation $r(U_2,\hat{U}_2)$ of U_2 and $\hat{U}_2$ where $\hat{U}_2$ denotes the best estimate of U_2 by means of linear regression with respect to $\underline{U}_1$. However, this $r(U_2,\hat{U}_2)$ is equal to the maximum correlation between the linear functions of $\underline{U}_1$ and U_2, and

$$E(U_2-\hat{U}_2)^2=1-R^2_{U_2\cdot\underline{U}_1},$$

where $R_{U_2\cdot\underline{U}_1}$ is the multiple correlation coefficient of U_2 and $\underline{U}_1$.

The above note on the explanation of variance holds.

2.3 Generally, we can describe the relation between two stochastic vector variables by the maximal correlation between their linear functions. This number is called canonical correlation coefficient. One can realize that this method is a natural generalization of the concept of correlation.

We can define $m:=\min\{q,p-q\}$ canonical correlation coefficients and factors by analitical method.

We are able to pass over to a coordinate system (or factor space) where the components of $\underline{U}_1$ and $\underline{U}_2$ are uncorrelated except those coordinates which have the same indices and which have considerable correlations. In this space from the variance of the i^{th} component of $\underline{U}_1$ the $\underline{U}_2$ explains exactly the same quantity as its i^{th} coordanite does.

In analitical terms: the first canonical correlation equals to

$$\rho_1 = \max \; r(\underline{L}_1'\underline{U}_1,\ \underline{M}_1'\underline{U}_2)$$

$$\underline{L}_1 \in R^q, \quad \underline{M}_1 \in R^{p-q}$$
$$D^2(\underline{L}_1'\underline{U}_1) = D^2(\underline{M}_1'\underline{U}_2) = 1,$$

where $r(\xi,\eta)$ denotes the correlation between ξ and η , and $D^2(\xi)$ denotes the variance of ξ.

It is easy to prove that this minimum is reached. The coordinates of $\underline{L}_1'\underline{U}_1$ are called first left-hand-side canonical factors and the coordinates of $\underline{M}_1'\underline{U}_2$ are called first right-hand-side canonical factors.

We can get by the method of Lagrange multiplication ([8])

$$(\Sigma_{21}\ \Sigma_{11}^{-1}\Sigma_{12} - \rho_1^2\Sigma_{22})\ \underline{M}_i = \underline{0} \text{ and}$$

$$(\Sigma_{12}\Sigma_{22}\Sigma\ _{21} - \rho_1^2\Sigma_{11})\underline{L}_i = \underline{0}.$$

Let us introduce the following notations (where $\underline{L}_i'\underline{U}_1$ and $\underline{M}_i'\underline{U}_2$ are the proper side canonical factors (i=1,2,...m)

$$L = (\underline{L}_1|\underline{L}_2|\ \ldots\ |\underline{L}_m)$$
$$M = (\underline{M}_1|\underline{M}_2|\ \ldots|\underline{M}_m)$$

$$\Lambda = \mathrm{diag}(\rho_1, \rho_2, \ldots, \rho_m) \quad (\rho_i \geq 0)$$

If we suppose that p=2q and k=m(=q) then reading by rows

(1) $$\hat{M}'\underline{U}_2 = \Lambda L'\underline{U}_1$$

where $\hat{M}'\underline{U}_2$ = the best least square linear regression with respect to $\{\underline{L}_1'\underline{U}_1, \underline{L}_2'\underline{U}_1, \ldots, \underline{L}_m'\underline{U}_1\}$, and at the same time (by rows)

(2) $$\hat{\underline{U}}_2 = M'^{-1}\Lambda L'\underline{U}_1$$

where $\hat{\underline{U}}_2$ = the best least square linear regression with respect to $\{\underline{L}_1'\underline{U}_1, \underline{L}_2'\underline{U}_1, \ldots, \underline{L}_m'\underline{U}_1\}$.

We try to explain the relations between $\underline{U}_1$ and $\underline{U}_2$ by means of generalization of explanation of variance. We describe these relations by 2 numbers: each of them show the average (mean) quantity which we explained from the sum of variances of components of one vector variable by the other:

$$R_{\underline{U}_2 \cdot \underline{U}_1} \quad \text{and} \quad R_{\underline{U}_1 \cdot \underline{U}_2} \ .$$

These numbers generally are not equal and they depend on the coordinate system. We consider 2 systems:

a/ in these space of canonical factors

$$R_{M'\underline{U}_2 \cdot L'\underline{U}_1}, \quad R_{L'\underline{U}_1 \cdot M'\underline{U}_2} \ .$$

In this case we can write

$$\hat{M'\underline{U}_2} = \Lambda L'\underline{U}_1,$$

$$\hat{L'\underline{U}_1} = \Lambda M'\underline{U}_2 \quad \text{and}$$

$$\varrho^2 = R_{M'\underline{U}_2 \cdot L'\underline{U}_1} \cdot = R_{L'\underline{U}_1 \cdot M'\underline{U}_2} = \frac{1}{k} \sum_{i=1}^{k} \varrho_i^2 \ .$$

b/ $\underline{U}_2$ in the original space and $\underline{U}_1$ in the space of canonical factors, and similarly $\underline{U}_1$ in the original space and $\underline{U}_2$ in the space of canonical factors

$$\hat{\underline{U}}_2 = M'^{-1} \Lambda \, L'\underline{U}_1,$$

$$\hat{\underline{U}}_1 = L'^{-1} \Lambda \, M'\underline{U}_2.$$

In this case

$$R_{\underline{U}_2 \cdot \underline{U}_1} \neq R_{\underline{U}_1 \cdot \underline{U}_2}$$

We have to note that there is a clear connection between factor and cannonical correlation analysis. Very similar optimizing features hold for both of them ([8], [4]).

There is a more complex explanation of canonical correlations the so-called redundancy analysis ([2]).

3. OUR RESULTS

Let us denote the measurement vector by $\underline{U}_1$, and the vector for the same person measured five years later by $\underline{U}_2$.

Our aim is clear:
we have to analyze the connection between $\underline{U}_1$ and $\underline{U}_2$, and construct an estimation of $\underline{U}_2$ by $\underline{U}_1$. We chose the following 6 variables:

Broca-index
Systolic blood-pressure
Diastolic blood-pressure
Vitalcapacity
Serum Cholesterol
Body mass index.

We constructed subsets of the sample. We found 209 individuals to be healthy during the lo year period. 28 patients had already been ill at the first time. 17 persons became ill when their data were measured at second time.

We had to control whether the healthy and sick persons could be separated by discriminantial analysis on the basis of this sample.

The first-kind and second-kind error were about 30-31 %.

We cut the persons for 3 non-disjoint groups:

HHH - healthy
HS - people who became ill during the medical examinations
HHH&HS - all of the individuals

At the beginning $\underline{U}_1$ and 5 years later $\underline{U}_2$ measure the data. In all 3 cases we established the estimation of the $\hat{\underline{U}}_2$ by (2). Table 1 shows the suitable correlation-type measurement numbers and their depencies on chosen coordinate system. We classified these $\hat{\underline{U}}_2$ by means of discriminantial analysis. Our result shows that the third regression gives the "minimal error" (Table 2).

So, we proved that by this method we could analyze the relation between the actual and expected status, and could state the proper conclusions with the help of regressions. The surprise of the analysis was that this relation was higher for people which were going to be sick than for healty people.

	$R_{\underline{U}_1 \cdot \underline{U}_2}$	$R_{\underline{U}_2 \cdot \underline{U}_1}$
HHH	0.574	0.570
HS	0.734	0.735
HHH&HS	0.579	0.573

Table 1.

	first-kind error	second-kind error
HHH	0.25	0.35
HS	0.93	0.06
HHH&HS	0.24	0.35

Table 2.

LITERATURE

1 ANDERBERG, M.R., Cluster Analysis for Applications (Academix Press, New York-London, 1973)

2 COOLEY, W.W. and LOHNES, P.R., Multivariate Data Analysis (John Wiley and Sons, New York, 1971).

3 LENGYEL, T., "A kanonikus korrelációanalizis alkalmazása szivkoszoruér - megbetegedések előrejelzésére", Számitástechnikai és kibernetikai módszerek alkalmazása az orvostudományban és a biológiában, 8. Neumann Kollokvium, Szeged, 1977, 11-17.

4 LENGYEL, T., "A kanonikus korrelációanalizis és néhány kapcsolódó probléma (The canonical correlation analysis and some related problems)", Alkalmazott Matematikai Lapok 5 (1979).

5 MILLER, J.K., "The development and application of bimultivariate correlation: a measure of statistical association between multivariate measurement sets", Ed.D. Dissertation, Faculty of Educational Studies, State University of New York at Buffalo, 1969.

6 STEWART, D.K. and LOVE, W.A., "A general canonical correlation index", Psychological Bulletin 70 (1968) 160-163.

7 TUSNÁDY, G., "Mátrixok szinguláris felbontása", Alkalmazott Matematikai Lapok 5 (1979).

8 RAO, C.R., Linear Statistical Inference and Its Applications (John Wiley and Sons, New York, 1965).

Lengyel Tamás

A kanonikus korrelációanalizis alkalmazása szivkoszoruér megbetegedések előrejelzésében

Egy olyan módszer alkalmazására adunk példát, ami az ilyen jellegü vizsgálatok körében ujszerü. Rövid idősorok esetében várható, hogy ez a módszer a vizsgálati adatok analizálásának hatékony segédeszköze lesz. A cikkben ismertetjük a kanonikus analizis elvét, egy konkrét példán keresztül bemutatjuk alkalmazását.
Megjegyezzük, hogy kevés értékelhető adat állt rendelkezésünkre, s ezért bizonyos egyszerüsitésekre kényszerültünk. A módszer további finomitására is van lehetőség.

Тамаш Лендел

Применение канонического корреляционного анализа

Автор даёт новый метод для применения канонического корреляционного анализа.