Special Canonical Models for Multidimensional Data Analysis with Applications and Implications[1]

By

Vithanage Pemajayantha

Modeling and Simulation Research Group, School of Quantitative Methods and Mathematical Sciences, University of Western Sydney, Australia

All correspondence to

V. Pemajayantha, Ph.D

Head, Modelling and Simulation Research Group,

School of Quantitative Methods and Mathematical Sciences

University of Western Sydney, PO Box 1797, South Penrith DC

NSW 2797

Australia

[1] Paper presented at the International Conference on Stochastics and Applications at National University of Singapore, 18 August 2002

Abstract

Deterministic and stochastic forms of linear and non-linear "prior" models were used to develop a new multidimensional data analysis within the classical canonical analysis. Detection of outliers with the new model is discussed. While the new model opens up a variety of research problems, it has potential straightforward applications in data mining in science, economics, commerce and industry.

Keywords*: Canonical Analysis, Multidimensional Analysis, multivariate data mining*.

1. Introduction:

Since the discovery of canonical correlations, although canonical correlation analyses have heavily been applied in classificatory analyses, relatively very less amount of work has been done on the improvement of canonical models simply because the canonical coefficients are difficult to interpret (Kettenring, 1971). Anderson (1971) gave a good account for analysis of classical canonical forms of multivariate distributions. In a recent work, Neuenschwander, et al (1995) presented a concept of common canonical variate based on the assumptions that the canonical variate has the same coefficients in all k sets of variables, and clearly the application of this concept is restricted to the condition underlined that all variables must have a common coefficients. An extension of canonical correlation analysis to canonical models in terms input-output relationships within process control studies are found (Author, et al 1995, Author, 1996). To incorporate stochastic models and detection of outliers in the canonical models of practical use, this study is a natural extension of those studies on canonical models previously reported in the area of process control.

The first part of the study is allocated to outline the preliminaries involved, deterministic models of linear and nonlinear forms within the classical canonical

analysis, and the results of the extension of classical canonical analysis in order to hold good for general applications. Stochastic models within the canonical analysis are presented in part 3. Detection of outliers and applications and implications of the canonical models developed are included in the part 4, and the conclusion is given in part 5.

2. Preliminaries:

The application of canonical analysis within the context of input-output quality control studies were developed previously using deterministic models within classical canonical analysis. Since this study is a natural extension of the previous work, the similar notations are being used in the subsequent discussion.

Let $\mathbf{x}^{(1)}$ be a k dimensional random vector of k quality variables for outputs, and $\mathbf{x}^{(2)}$ be a l dimensional random vector of l quality variables for inputs. Let us define $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ as $\mathbf{X}^{(1)} = \mathbf{x}^{(1)} - \overline{X}^{(1)}$ and $\mathbf{X}^{(2)} = \mathbf{x}^{(2)} - \overline{X}^{(2)}$ respectively.

Let X be

$$\begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix} \qquad (1)$$

$$\Sigma = XX' = \begin{bmatrix} X^{(1)}X^{(1)}{}' & X^{(1)}X^{(2)}{}' \\ X^{(1)}X^{(2)}{}' & X^{(1)}X^{(2)}{}' \end{bmatrix} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \qquad (2)$$

It is assumed that $X \sim (\theta, \Sigma)$

In the previous work, two cases that $\mathbf{X}^{(1)}$ is related to $\mathbf{X}^{(2)}$ as described below were considered.

Case 1: Input and output variables are linearly related in the above multivariate system.

$$E(X^{(1)}) = B.\ X^{(2)} \qquad (3)$$

where B is a *k* x *l* matrix of coefficients (β_{ij}).

Case 2: Input and output variables are exponentially related in a multivariate system.

$$E\left(X_i^{(1)}\right) = \prod_{j=1}^{l} X_j^{(2).\beta_{ij}} \qquad (i=1,2, ..., k) \qquad (4)$$

When $\beta_{ij} = 0$, $j>1 \ \forall$ j this reduces to famous Cobb-Douglas form of production functions.

For $X^{(2)}$ = observed values of $X^{(2)}$, let us define $(1/n)\ X^{(2)}X^{(2)'}$ by S_{22}

Based on the classical canonical analysis, ____(Author) (1996) showed that resulting canonical variate for the case 1 is given by the solution to the following determinantal equation:

$$\beta S_{22} \beta' \alpha - \lambda . \Psi \alpha = 0 \qquad (5)$$

together with $\beta = k\ \gamma'$ and selecting k while holding $\sigma_1 = \sigma_2 = 1$. It has been shown that $k = \sqrt{\nu}$. (Anderson 1974).

where,

$$\Psi = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \qquad (6)$$

and

$B = \Sigma_{12} \Sigma_{22}^{-}$ and β's are vectors of coefficient matrix B

Since the classical canonical models assumed the fact that canonical variates have unit variance in their developments, coefficients of the classical canonical variates are difficult to interpret. In order to generalize the classical canonical models, the following definitions were used in the previous studies.

Definition 1: *Input-output Uniformity*

If the variability of the output component of quality variables are the same as that of corresponding input component,

$$\text{that is } \alpha' \Psi \alpha = \gamma' \Sigma_{22} \gamma = \sigma_o \qquad (7)$$

we define the process to be Input-output Uniformity.

Definition 2: *Input-output divergence*

If the variability of the output component is higher than that of the input component, the process is defined to be input output divergence.

Definition 3: *Input-output convergence*

If the variability of the output component is lower than the variability of input component, the situation is termed an input output convergence.

It has been shown that under input-output uniformity, the following result holds.

Result 1:

$$\gamma = \alpha' \beta / \sqrt{\lambda} \qquad (8)$$

(See the appendix for proof).

Under input output divergence or convergence, the following results holds:

Result 2:

$$\gamma = \alpha' \beta / \sqrt{(\lambda/\omega)} \qquad (9)$$

where, $\sigma_1/\sigma_2 = \frac{\gamma \Sigma_{22} \gamma}{\alpha' \Psi_{22} \alpha} = \omega$ (10)

$\omega \in (0,1)$ for input-output convergence, and

$\omega \in (1, \infty)$ for input-output divergence.

(See Appendix for Proof).

In the case of the model presented in case 2, the relationship between $X^{(1)}$ and $X^{(2)}$ is presented in an exponential form, and therefore conversion of this relationship to logarithm make the relationship liner and hence the above results being applied (Author, 1996).

Both the above developments are on the basis of deterministic models. Therefore, the following stochastic model was considered in Case 3 as an extension.

Case 3: Stochastic models: In order to incorporate stochastic model within the canonical analysis, we consider an autoregressive moving average model with exogenous variables.

Let us consider an ARMAX model as follows

$$\mathbf{Y}_t = a\,\mathbf{Y}_{t-1} + \mathbf{X}_t\,\beta + U_t \qquad (11)$$

where $|a| < 1$ and $t = 2,3, \ldots ,T$

In matrix notation, the above model can be expressed as follows.

$$\mathbf{Y}_t = \alpha\,\mathbf{Y}_{-1} + \mathbf{X}\beta + \mathbf{u} \qquad (12)$$

where $\mathbf{Y}$, $\mathbf{Y}_{-1}$, $\mathbf{X}$ and $\mathbf{u}$ are n x 1 vectors (n = T-1). u is assumed to be a normally distributed vector: $\mathbf{u} \sim N(0,\Sigma_u)$.

3. The problem

Let ζ_1 and ζ_2 be two canonical forms as follows

$$\zeta_1 = \eta_1\, f(Y_t) \qquad (13)$$

$$\zeta_2 = \gamma_1\, f(X_t) \qquad (14)$$

If the collection of random vectors, Y, Y_{-1} and X, together follow a multivariate distribution, given the relationship between Y_t and X_t: $Y_t = \alpha Y_{-1} + X\beta + u$, our problem is to extract two component $\zeta_1$and ζ_2 so that the correlation between these components is maximum.

The OLS estimation for the ARMAX model in (11) without considering a multivariate structure is readily available, for example Hoque (1996), as follows.

$\alpha = (Y'AY_{-1})/ (Y'_{-1} AY_{-1})$ where $A = I - X (X'X)^{-1} X'$

$\beta = (X'X)^{-1} X'(Y - \alpha Y_{-1})$.

Further derivation of the above model is available in Hoque (1996).

In oder to solve the above problem, for convenience, we rearrange the problem with following notations.

Let $X^{(1)} = Y_t$, $X^{(1)}_{-1} = Y_{t-1}$ and $X^{(2)} = X$ in the model explained by (11)

Let $X = \begin{bmatrix} X^{(1)} \\ X^{(1)}_{-1} \\ X^{(2)} \end{bmatrix}$

$$X\,X' = \begin{bmatrix} \sum_{11} & \sum_{12} & \sum_{13} \\ \sum_{21} & \sum_{22} & \sum_{23} \\ \sum_{31} & \sum_{32} & \sum_{33} \end{bmatrix} = \Sigma$$

and assume that $X \sim (\theta, \Sigma)$

As the relationships explained by (11) resembles

$$E(X^{(1)}) = \alpha \quad X^{(1)}{}_{-1} + \beta X^{(2)}$$

Let $X^{(2)*} = \begin{bmatrix} X^{1}_{-1} \\ X^{(2)} \end{bmatrix}$ and $\zeta* = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$

Then it suffices to select $E(X^{(1)}) = X^{(2)} * \zeta*$ to obtain model 1. Also, it is a trivial exercise to show that case 2 is linear model or akin to Model 1. Consequently, it follows that the above results hold true for all cases.

4. Detection of outliers

Canonical variates generated by the above analysis follow normal distributions. For example, in Case 1, $\zeta 1 \sim N(\mu\zeta, \sigma_\zeta^2)$. Therefore, the outliers can be detected with a normal probability plots given that $\mu\zeta$ and σ_ζ^2 are known.

Remark 1: According to the Case 1, ζ is distributed normally with

$$E(\zeta_1) = \alpha' \beta \mu^2$$

$$\text{Var}(\zeta_1) = \alpha' \beta S_{22} b' \alpha$$

and

$$E(\zeta_2) = \gamma' \mu^{(2)}$$

$$\text{Var}(\zeta_2) = \gamma' S_{22} \gamma$$

The proof of the above remark is a straightforward application of the theorem 2.4.1 of Anderson (1974) (see, Appendix).

Therefore, $\zeta_1 \sim (\alpha' \beta \mu^{(2)}), \alpha' \beta S_{22} b' \alpha)$

and

$$\zeta_2 \sim (\gamma' \mu^{(2)}, \gamma' S_{22} \gamma)$$

Also, we noted that Case2 and the Case 3 reduced to Case 1 with suitable transformation. Therefore, the normal probability values of ζ_1 and ζ_2 are readily available for detection of outliers within the above canonical models.

5. Discussion and Conclusion

Multivariate data analysis has been enhanced with knowledge discovery in data bases (KDD) or data mining. Consequently, it demands variety of techniques to find relationships with multidimensional massive data sets. Classical canonical analysis has potential for finding relationships among compounds of random vectors. This potential was studied and yielded the forgoing results. Among the problem yet remaining in this method are the possible algorithms and asymptotic studies of parameters of the proposed model. The fact that the technique could be applied together with effect due to time or time series variables, it has straightforward applications in data mining. While this area is yet open for future studies it is beyond the scope of this study.

References:

Anderson, T. W. (1971). *Introduction to Multivariate Analysis*. New York: John Wiley.

Hoque, A. (1996). Multiperiod Forcasting Analysis of A Dynamic Model for A Small Sample. Australian Journal of Statistics, 38(2), 113-129.

Kettenring, J. R. (1985). Canonical Correlation Analysis. In *Enclyclopedia of Statistical Sciences*, 1, Ed. S. Kotz and N. L. Johnson, New York:John Wiley.

Neuenschwander, B. E. (1995). Common Canonical Variates. *Biometrika*, 82(3), 553-560.

Author, Agrawal, R. and Truong, C. N. (1995). Canonical form of Multivariate Control Chart for optimum process control. *1st Australian conference on Industrial Statistics*, *Brighton Le Sands*, NSW, Australia, 4-5 December 1995.

Author (1996). Multivariate Control Chart with input-output relationships for optimal process control. *International Statistical Congress*, Sydney, Australia 8-12 July 1996

Appendix:

a.1) Variances of the components are unity: $\sigma_1 = \sigma_2 = 1$.

Proof 1*:*

Let us consider a linear combination of first set of random vectors, say outputs, $\mathbf{U} = \alpha' x^{(1)}$. We can find the expected value of this linear combination as follows;

$$E(U) = \alpha' B X^{(2)} \qquad (1.1)$$

and the variance of U as follows;

$$E(U^2)=\text{Var}(U) = \alpha' \Psi \alpha = \sigma^2_1 \qquad (1.2)$$

Then the mean sum of squares of the expected values is

$$\frac{1}{n}\sum_{i=1}^{n}(E(U))^2 = \alpha' B (1/n) X^{(2)} X^{(2)'} B' \alpha \qquad (1.3)$$

For a set of observed values of $X^{(2)}$, let us define (1/n) $X^{(2)}X^{(2)'}$ by S_{22}. Therefore, the estimated mean sum of the square of U can be written as follows;

$$\frac{1}{n}\sum_{i=1}^{n}(E(U))^2 \quad = \alpha' B\, S_{22}\, B'\alpha \qquad (1.4)$$

To maximize the correlation between U and $V= \gamma' x^{(2)}$, we maximize the mean sum of squares (1.3) relative to the variance (1.2) based on the following fact;

$E\,(BV)^2 = r^2\, \sigma^2_u/\sigma^2_v\, E\, v^2 = r^2\sigma_u^2$

and so,

$E(BV)^2 / E\, U^2 = r^2$ (See Anderson 1974).

Now the algebraic solution to find α is to maximize (1.4) subject to (1.2). Let

$$\Phi = \alpha'\beta S_{22}\beta\alpha' - \frac{1}{2}\lambda\left(\alpha'\Psi\alpha - 1\right) \qquad (1.5)$$

where λ is a LaGrange multiplier.

Therefore, the vector derivatives set equal to zero are

$$\beta S_{22} \beta' \alpha - \lambda . \Psi \alpha = 0 \qquad (1.6)$$

To find a nontrivial solution to (1.6), we set

$$\left| \beta S_{22} \beta' - \lambda \Psi \right| = 0 \qquad (1.7)$$

Since the correlation between U and V does not change the scale of U and V (Multiple of V and multiple of U), the coefficients (α, γ) of the linear combinations are derived assuming that the variances of V and U equal unity in the classical Canonical analysis.

a.2) Variance of the components are not unity: $\sigma_1 \pm \sigma_2 \pm 1$

In this case, the variance of U is given in (1.2), and similarly the variance of V is as follows;

$$Var(V) = \gamma' X^{(2)} X^{(2)} ' \gamma = \gamma' S_{22} \gamma = \sigma_2^2 \qquad (1.8)$$

Let the correlation between U and V which is given as follows;

$$E(UV) = E(\alpha' X^{(1)} X^{(2)'} \gamma) = \alpha' \beta X^{(2)} X^{(2)'} \gamma$$

$$= \alpha' \beta \, \Sigma_{22} \gamma \qquad (1.9)$$

Then, the algebraic solution to find α is to maximize (1.9) subjected to (1.2) and (1.8).

Let

$$\phi = \alpha' \beta \, \Sigma_{22} \gamma - \tfrac{1}{2}\nu (\alpha' \Psi \alpha - \sigma_1) - \tfrac{1}{2}\mu (\gamma' \Sigma_{22} \gamma - \sigma_2) \qquad (1.10)$$

where ν and μ are LaGrange multipliers. Therefore, the vector derivatives set equal to zero are

$$\partial\phi / \partial\alpha = \beta \Sigma_{22} \gamma - \nu \Psi \alpha = 0 \qquad (1.11)$$

$$\partial\phi / \partial\gamma = \Sigma_{22}' \beta' \alpha - \mu \Sigma_{22} \gamma = 0 \qquad (1.12)$$

(1.11) x α',

$$\alpha'\beta\Sigma_{22}\gamma - \nu\alpha'\Psi\alpha = 0 \qquad (1.13)$$

(1.12) x γ',

$$\gamma'\Sigma_{22}'\beta'\alpha - \mu\gamma'\Sigma_{22}\gamma = 0 \qquad (1.14)$$

Equations (1.13) and (1.14) imply that

$$\alpha'\beta\Sigma_{22}'\gamma = \nu\alpha'\Psi\alpha$$

$$\gamma'\Sigma_{22}'\beta'\alpha = \mu\gamma'\Sigma_{22}\gamma$$

and so,

$$\frac{\alpha'\beta\Sigma_{22}\gamma}{\gamma'\Sigma_{22}'\beta\alpha} - \left(\nu/\mu\right)\frac{\alpha'\Psi\alpha}{\gamma'\Sigma_{22}\gamma} = 0 \qquad (1.15)$$

Multiplying (1.15) by the denominator of the first term (scalar multiplication),

$$\alpha'\beta\Sigma_{22}\gamma - (\nu/\mu)\frac{\alpha'\Psi\alpha}{\gamma'\Sigma_{22}\gamma}\gamma'\Sigma_{22}'\beta\alpha = 0 \qquad (1.16)$$

(1.16) x $(\alpha')^{-1}$,

$$\beta\Sigma_{22}\gamma - (\nu/\mu)\Psi\alpha\frac{\gamma'\Sigma_{22}\beta'\alpha}{\gamma'\Sigma_{22}\gamma} = 0 \qquad (1.17)$$

Post multiplication of (1.17) by $\gamma^{-1}\beta'\alpha$,

$$\beta\Sigma_{22}\beta'\alpha - \left(\frac{\nu}{\mu}\right)\Psi\alpha\frac{\gamma'\Sigma_{22}\beta'\alpha\gamma^{-1}\beta'\alpha}{\gamma'\Sigma_{22}\gamma} = 0 \qquad (1.18)$$

Let $\alpha'\beta = \gamma' k$, and, in turn, $\beta'\alpha = k'\gamma = k\gamma$ where k is a constant.

Substituting $\alpha'\beta = \gamma' k$ in (1.18),

$$\beta\Sigma_{22}\beta'\alpha - \frac{\nu}{\mu}\Psi\alpha\frac{\gamma'\Sigma_{22}k^2\gamma}{\gamma'\Sigma_{22}\gamma} = 0 \qquad (1.19)$$

Let $k^2(\nu/\mu) = \lambda$. Then, from (1.19),

$$\beta\Sigma_{22}\beta'\alpha - \lambda\Psi\alpha = (\beta\Sigma_{22}\beta' - \lambda\Psi)\alpha = 0 \qquad (1.20)$$

To find a nontrivial solution for α, we solve (1.10) using the following determinantal equation.

$$\left|\beta\Sigma_{22}\beta' - \lambda\Psi\right| = 0 \qquad (1.21)$$

If $\alpha'\Psi\alpha = \gamma'\ \Sigma_{22}\ \gamma$, then

$$\frac{\gamma\Sigma_{22}\gamma}{\alpha'\Psi_{22}\alpha} = 1$$

From equations (1.13) and (1.14), we find that

$$\nu = \frac{\alpha'\beta\Sigma_{22}\gamma}{\alpha'\Psi\alpha}$$

and

$$\mu = \frac{\gamma'\Sigma_{22}\beta'\alpha}{\gamma\Sigma_{22}\gamma}$$

$$\Rightarrow \qquad \left(\frac{\nu}{\mu}\right) = \frac{\gamma\Sigma_{22}\gamma}{\alpha'\Psi\alpha}\cdot\frac{\alpha'\beta\Sigma_{22}\gamma}{\gamma\Sigma_{22}\beta'\alpha} = 1 \qquad (1.22)$$

Since k^2.(ν/μ)= λ,

$$k = \sqrt{\lambda}$$

By definition, $\alpha' \beta = \gamma' k$, and therefore (8) holds.

By virtue of equation 1.22 and the definition of λ, it follows that

$$k = \sqrt{\lambda/\omega} \text{ and } \gamma = \alpha'\beta \Big/ \sqrt{\lambda/\omega}.$$

a.3) Theorem 2.4.1 of Anderson (1974)

Let X (with p components) be distributed according to $N(\mu, \Sigma)$. Then

Y = C X is distributed according to $N(c\mu, C\Sigma C')$ for C nonsingular.

The proof of this theorem is available in Anderson (1974).