回帰分析における母集団差の多次元表現モデル

稲垣 敦・松浦義行

A multidimensional representation model of population differences in regression analysis

Atsushi INAGAKI and Yoshiyuki MATSUURA

本研究では，母集団差回帰モデル，POPREG（POPulation differences REGression model）と，このモデルをデータに適合するための交互最小二乗アルゴリズムを提案する。このモデルは，異なったグループや母集団からの標本というような複数組のデータに対応しており，回帰係数を内積モデルの仮定の下で再パラメタ化することによって，グループ間差異を表現できるだけでなく，空間的に表現することにより，各群の特徴や各群間の差異の理解を可能にする。また，複数の時間に対応するデータにも適用可能であり，縦断的研究，及び多変量時系列データの分析に応用可能であると考えられる。そして，これらの分析が比率，間隔，順序，名義尺度，及びそれらが複合したデータで可能である。以上の点から，このモデルは体育学，心理学，社会学をはじめ，幅広い分野で応用可能であると考えられる。

Key words: Regression analysis, Population (group, occasion or situation) differences, Multidimensional representation, Alternating least squares estimation (ALS), POPREG.

Introduction

Regression analysis concerns the prediction according to some variables and is one of the most widely used of all statistical methods. It is used by data analysis in nearly every field of science and technology as well as the social sciences, economics, and finance. Today it is a rare statistical program package that does not provide regression analysis.

In the past, statisticians and psychometricians have been interested in the problem as to parameter estimation method, variable selection, optimal criterion, dummy variable, residual analysis, nonlinear and nonmetric regression, some constrained model and so forth. But regression analysis can provide not only the linear predictive function, but also linear structure model between dependent and independent variables. In fact regression analysis is of interest to psychologists and sociologists,not only for its practical use, that is, prediction, but also as a model of psychological and sociological processes or relations underlying the phenomena. Although regression model is useful to clarify the relation or causal structure, any multidimensional spatial representation models to express the population differences in regression model have not been developed yet.

This article presents the development of "population differences regression model" (POPREG).This method can operate on N-sets of data from different populations or situations, and provides not only linear regression equation, but also the multidimensional spatial representation of population differences under the assumption of vector model. Moreover this method is useful instrument for longitudinal or time series research because this can be also applied to the data obtained from the different time points or occa-

sions. The data can be defined at nominal, ordinal, interval or ratio levels of measurement, or can be mixture of two or more levels. As will be explained, POPREG model provides an optimal scale for each variable within the restrictions as to the measurement level and process. This scaling is optimal in the sence that correlation is maximized. Finally we emphasize that this method can statistically determine the number of dimensions. This model is characterized as follows:

1) Squared loss function—the proposed model is fitted to data in a sence of least square criterion.
2) Alternating least squares procedure—the proposed model is fitted to data with the alternating least squares procedure.
3) Multidimensional representation—the proposed model can represent both populations (groups, situations, or occasions) and predictors as vectors in multidimensional population space and variable space, respectively.
4) Dimensionality identification—in many spatial models the appropriate dimensionality is determined by examining a plot of the goodness (badness) of fit measure versus the number of dimensions (*i. e.* scree plot) as well as subsequent interpretation of the dimensions. In addition to these criterion, the proposed model allows for an asymptotic statistical test for identifying the appropriate dimensionality.
5) Level of measurement—the proposed model will be able to accomodate ratio, interval, ordinal and nominal data.
6) Time series data—the proposed methods will be able to apply the time series and longitudinal data.
7) Data transformation—the proposed model will be able to transform the data and can statistically test with a similar way described above.

In the next section we will present a detailed account of POPREG model, emphasizihg the characteristics of the model.

The Model

We use bold-face capital letters to represent matrices ($\mathbf{X}$); bold-face lower case letters for vector ($\mathbf{x}$); and regular lower case letters for scalars (x). Note that all vectors are assumed to be column vectors, with a row vector denoted as transpose of a column vector ($\mathbf{x}'$). We refer to a specific column vector of a matrix as $\mathbf{x}_j$, a specific element of a matrix as x_{ij}.

Let:

i,i' =1, ···, I populations, groups, situations or occasions,
j,k =1, ···, J variables,
s,t =1, ···, S dimensions in an MDS contexts,
y_i =column vector of dependent or predicted variable in i-th population,group or occasion (Nx1) ,which is measured at some known measurement level, and is normalized in the preprocessing phase if this is measured at more than interval level,
$\mathbf{y}_i^*$ =column vector of the optimally scaled dependent or predicted variables corresponding to y_i, which is defined at interval level of measurement,
$\mathbf{X}_i$ =matrix of independent variables or predictors in i-th population, group or occasion (NxJ), which is measured at some known measurement level, and is columnwisely normalized in the preprocessing phase if this is measured at more than interval level,
$\mathbf{X}_i^*$ =matrix of the optimally scaled independent or predict variables corresponding to $\mathbf{X}_i$, which is defined at interval level of measurement,
$\mathbf{W}_i$ =column vector of regression weights of $\mathbf{y}_i^*$on $\mathbf{X}_i^*$(Jx1),
$\mathbf{A}_c$ =coordinate matrix of predictor or independent variables (JxS),
$\mathbf{b}_i$ =dimensional weight vactor of i-th population,
$\mathbf{e}_i$ =column vector of error or residual of i-th population (Nx1),

Using the above definitions, we can formulate

the POPREG model by matrix from

$$\mathbf{y}^* = \mathbf{X}^*\mathbf{w} + \mathbf{e}$$
$$= \mathbf{X}^*\mathbf{A}\mathbf{b} + \mathbf{e}, \quad (1)$$

where

$$\mathbf{y}^* = (\mathbf{y}_1{}^{*\prime}, \mathbf{y}_2{}^{*\prime}, \cdots, \mathbf{y}_I{}^{*\prime})',$$
$$\mathbf{X}^* = \sum_{i=1}^{I} \mathbf{E}_{ii} \otimes \mathbf{X}_i{}^*,$$
$$\mathbf{w} = (\mathbf{w}_1', \mathbf{w}_2', \cdots, \mathbf{w}_I')',$$
$$\mathbf{A} = \sum_{i=1}^{I} \mathbf{E}_{ii} \otimes \mathbf{A}_i$$
$$\mathbf{A}_i = \mathbf{A}_c \text{ for all i},$$
$$\mathbf{b} = (\mathbf{b}_1', \mathbf{b}_2', \cdots, \mathbf{b}_I')',$$
$$\mathbf{e} = (\mathbf{e}_1', \mathbf{e}_2', \cdots, \mathbf{e}_I')'.$$

where the notation $(\mathbf{X} \otimes \mathbf{Y})$ refers to right Kronecker product of matrices $(\mathbf{X} \otimes \mathbf{Y}) = [x_{ij}\mathbf{Y}]$, and where $\mathbf{E}_{ij}$ denotes a matrix with the unit scalar in the (ij) position, and with zeros elsewhere. As far as I know, this model has not been developed in the past literatures. In this model,

$$\mathbf{x}_{ij}{}^* = gx_{ij}(\mathbf{x}_{ij}), \ (i=1, \cdots, I; j=1, \cdots, J),$$
$$\mathbf{y}_i{}^* = gy_i(\mathbf{y}_i), \ (i=1, \cdots, 1),$$

where gx_{ij} and gy_i are called "measurement transformations", and are subject to restrains by the measurement level and process of their variables. The restrains were discussed in Young et al.[12] We may correctively regard $\mathbf{x}_{ij}{}^*$ and $\mathbf{y}_i{}^*$ as being the observation variables rescaled at the interval level of measurement so that the multiple correlation is maximized. Thus we refer to $\mathbf{x}_{ij}{}^*$ and $\mathbf{y}_i{}^*$ as optimally scaled observations (data).

The output of this method contains two major elements. The first is the estimate of the (JxS) predictor variable coordnate matrix $\mathbf{A}_c$. The second is an estimate of a (lSx1) vector of population or group weight $\mathbf{b}$. The weight b_{si} can be called "importance weight ". All other things equal, as b_{si} increase, s-th dimension have a larger and larger influence on the regression weight. Moreover matrix $\mathbf{A}_c$ and $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_I)'$ can be regarded as the matrix representaion of the predictor variable space and the matrix reprsentation of population space. Therefore using graphical represention, we can visually understand both predictor and population differences. As in the individual differences scaling (INDSCAL) model[2], interpreting the solution is a larger task when POPREG model is employed because one must develop an interpretation of both the predictor variable space and the population space.

Finally note that the output does not include a separate regression weight vector $\mathbf{w}_i$ for each population or group. This is bacause the each population's regression weights can be reconstructed from the elements of $\mathbf{A}_c$ and $\mathbf{b}$.

Optimization Criterion

In our case, as in many similar situations, we define a squared loss function. We then search for the best solution such that

$$f(\theta, g) = \| \mathbf{y}^* - \hat{\mathbf{y}}^* \|^2,$$
$$= \sum_{i=1}^{I} \| \mathbf{y}_i{}^* - \hat{\mathbf{y}}_i{}^* \|^2 \quad (2)$$

is minimal, where model parameters and measurement transformations are indicated by $\theta = \{\mathbf{A}, \mathbf{B}\}$ and $g = \{gx_{11}, \cdots, g_{IJ}, gy_1, \cdots, gy_1\}$, and where $\| \cdot \|$ denotes the Euclidean norm, and $\hat{\mathbf{y}}^*$ and $\hat{\mathbf{y}}_i{}^*$ denote the estimated $\mathbf{y}^*$ and $\mathbf{y}_i{}^*$, respectively. The minimization has to be carried out under the normalization restriction;

$$\mathbf{1}_N{}' \mathbf{X}_i{}^* = \mathbf{0}'_J \ (i=1, \cdots, 1),$$
$$\mathbf{1}_N{}' \mathbf{y}_i{}^* = 0,$$
$$\text{diag}\ (N^{-1}\mathbf{X}_i{}^{*\prime}\mathbf{X}_i{}^*) = \mathbf{I}_J,$$
$$N^{-1}\mathbf{y}_i{}^{*\prime}\mathbf{y}_i{}^* = 1.$$

Algorithm

The POPREG algorithm is an alternating least squares (ALS) algorithm.The ALS approach is related to the works of Wold and Lyttkens[10], de Leeuw[3] and Young.[11] As is implied by the name, the ALS algorithm is an iterative algorithm which alternates back and forth between two phases, each of which is a least squares procedure. In one of the phases, least squares estimates for the model parameters are obtained while the data transformations are held constant, whereas

in the other phase, least squares estimates of the transformations are obtained while the model parameters are held constant. The structure of an iteration for this procedure is to estimate and replace $\mathbf{x}_{ij}^*$ (j=1, ···, J; i=1, ···,I), and then each y_i^*, and then to estimate the model parameters. We repeat these algorithms until convergence is obtained (Note that we do not have to use ALS procedure if all variables are measured on more than interval scale). Since both phases minimize a loss function (2), the algorithm is convergent.

After convergence, If you want to normalize A_c matrix, then linear transformation is necessary. That is,

$$\tilde{\mathbf{A}}_c=\mathbf{A}_c\mathbf{T}^{-1}$$
$$\tilde{\mathbf{b}}_i=\mathbf{T}\mathbf{b}_i,$$

where T denotes a (SxS) diagonal matrix with a square root of the sum of squared i -th column's elements of A_c matrix in the (ii) position.

Several alternatives of iteration structure can be developed in POPREG algorithm, so we explain four type of alternatives as follow. The first alternative is to optimally scale $\mathbf{x}_{ij}$ (j=1, ···, J, i=1, 2, ···, I), and then each $\mathbf{y}_i$(i=1, 2, ···, I), and then to estimate model parameters. This completes a single iteration and we repeat this process until convergence is obtained. The second alternative is to scale $\mathbf{x}_{ij}$ (j=1, ···, J, i=1, 2, ···, I) and repeart this process for each of the $\mathbf{x}_{ij}$(j=1, ···, J, i=1, 2, ···, I), and estimate model parameters. Following this $\mathbf{y}_i$ (i=1, 2, ···, I) are scaled and then model paramters are estimated. The third alternative is to scale $\mathbf{x}_{ij}$(j=1, ···, J, i=1, 2, ···, I) and then $\mathbf{y}_i$ (i=1, 2, ···, I). We repeat this process until convergence is obtained and then estimate model parameters. This is one interation. The last alternative is to scale particular variables and to estimate model parameters. We repeat this process until all variables have been subjected to this process. This is one interation. We may select one of these alternatives.

In the next two sections, we will discuss model estimation phase and then optimal scaling phase.

Model Estimation

In model estimation phase, we desire to obtain least squares estimates of model parameters under the assumption that all the optimally scaled variables $\mathbf{X}^*$ and $\mathbf{y}^*$ are held constant. To minimize loss fuction, we utilize alternating least squares (ALS) technique. As mentioned previously, the essential feature of the ALS approach is that in solving optimization problems with more than one set of parameters, each set is estimated in turn by applying least squares procedures holding the other sets fixed. After all sets have been estimated once, the procedure is repeated until convergence.

In order to see how the ALS approach can be applied in the present context, let us return briefly to (2):

$$f(\mathbf{A},\mathbf{b})=\| \mathbf{y}^*-\mathbf{X}^*\mathbf{A}\mathbf{b} \|^2. \qquad (2)$$

Clearly the sets of parameters are here **A** and **b**. Minimizing f over **A** holding **b** fixed is equivalent to solving one least squares problem and minimizing over **b** with **A** fixed is another. That we are in practice minimizing f does not prevent the problem from being ALS one.

From the above discussion a rough outline for an algorithm is readily deduced. First we choose an arbitrary $\mathbf{A}_{(0)}$ yielding a new $\mathbf{b}_{(1)}$, nextly minimize subsequently over **A** with the just computed $\mathbf{b}_{(1)}$ fixed yielding a new $\mathbf{A}_{(1)}$, and ite rate this procedure until convergence.

The loss function f can be written as

$$f(\mathbf{A}_c, \mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_I)=\sum_{i=1}^{I} \| \mathbf{y}_i^*-\mathbf{X}_i^*\mathbf{A}_c\mathbf{b}_i \|^2.$$

The partial derivative of f with respect to $\mathbf{b}_i$ is relatively easy to obtain and setting to zero, and solving for the value of $\mathbf{b}_i$ which minimizes f for given **A**, results in the following expression for $\mathbf{b}_i$:

$$\mathbf{b}_i=(\mathbf{A}_c'\mathbf{X}_i^{*\prime}\mathbf{X}_i^*\mathbf{A}_c)^{-1}\mathbf{A}_c'\mathbf{X}_i^{*\prime}\mathbf{y}_i^*.$$

For i=1, 2, ···, I, minimizing f over $\mathbf{b}_i$ while **A** is fixed is achieved by the procedure mentioned above, that is, assuming $\mathbf{A}_c'\mathbf{X}_i'\mathbf{X}_i\mathbf{A}_c$ is nonsingular, the update for $\mathbf{b}_i$ is $(\mathbf{A}_c'\mathbf{X}_i^{*\prime}\mathbf{X}_i^*\mathbf{A}_c)^{-1}\mathbf{A}_c'\mathbf{X}_i^{*\prime}\mathbf{y}_i^*$. In the case $\mathbf{A}_c'\mathbf{X}_i^{*\prime}\mathbf{X}_i^*\mathbf{A}_c$ is singuler,

a generalized inverse should be used instead of inverse.

On the other hand, minimizing f as function of $\mathbf{A}_c$ for fixed b, seem quite cumbersome. This problem can be solved as follow. The loss function f consists of a sum of squared Euclidean norms of residual vectors. Therefore, function f can be rewritten as

$$f(\mathbf{A}_c, \mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_I)$$

$$=\sum_{i=1}^{I} \| \mathbf{y}_i^* - (\mathbf{X}_i^* \otimes \mathbf{b}_i') \text{ Vec } (\mathbf{A}_c) \|^2,$$

where Vec $(\cdot)$ denotes a matrix strung out rowwise into a column vector. Using this notation and putting the column-vectors for each of the I populations, groups, or occasions into one supervector yields,

$$f(\mathbf{A}_c, \mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_I)$$

$$= \left\| \begin{pmatrix} \mathbf{y}_1^* \\ \vdots \\ \mathbf{y}_I^* \end{pmatrix} - \begin{pmatrix} \mathbf{X}_1^* \otimes \mathbf{b}_1' \\ \vdots \\ \mathbf{X}_I^* \otimes \mathbf{b}_I' \end{pmatrix} \text{Vec}(\mathbf{A}_c) \right\|^2.$$

It is obvious that the problem of minimizing function f over $\mathbf{A}$ while vectors $\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_I$ are fixed is an ordinary regression problem, with Vec $(\mathbf{A}_c)$ containing the regression weights. Therefore function f is minimized over $\mathbf{A}_c$ by choosing

$$\text{Vec}(\mathbf{A}_c) =$$

$$(\sum_{i=1}^{I} \mathbf{X}_i^{*\prime} \mathbf{X}_i^* \otimes \mathbf{b}_i \, \mathbf{b}_i')^{-1} \quad \text{Vec}(\sum_{i=1}^{I} \mathbf{X}_i^{*\prime} \mathbf{y}_i^* \mathbf{b}_i').$$

Subsequently the update of matrix $\mathbf{A}_c$ is found by simply rewriting Vec $(\mathbf{A}_c)$ in matrix form. We have now described a solution to the problem of minimizing f over $\mathbf{b}$ while $\mathbf{A}_c$ is fixed and the problem of minimizing f over $\mathbf{A}_c$ matrix while the vector $\mathbf{b}$ is fixed.

This algorithm requires determining the inverse of a matrix of order JxS by JxS. For small J and S there is no problem. However if the number of independent (predictor) variables and the number of dimensions required increase, computational efficiency rapidly decreases, due to the necessity of inverting an increasingly large matrix. For this reason, an alternative algorithm is proposed, that requires the inverse of matrices of smaller order J by J.

In order to handle the cases where JxS is large, that is, requiring too much computer time and storage, an algorithm has been developed that uses a different alternating least squares procedure. In this algorithm the vectors $\mathbf{b}_i$ are updated as in the previous algorithm, but matrix $\mathbf{A}_c$ is updated columnwise. That is, each column of $\mathbf{A}_c$ is updated successively, while the other columns are fixed. It should be noted that the solution for matrix $\mathbf{A}_c$ found during the process is not the best least squares solution for $\mathbf{A}_c$. However, because all columns of $\mathbf{A}_c$ are optimal in the least squares sence, the function f is decreased nonetheless. This results in an alternating least squares algorithm consisting of I+S steps.

The columnwise procedure for updating $\mathbf{A}_c$ will be explained after rewriting function f $(\mathbf{A}_c \, \mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_I)$ in order to isolate the columns of $\mathbf{A}_c$. Let $\mathbf{a}_h$ denote column h of matrix $\mathbf{A}_c$, and b_{hi} the h-th element of b_i, for h = 1, 2, ···, S. Then f $(\mathbf{A}_c, \mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_I)$ can be rewritten as

$$f(\mathbf{A}_c, \mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_I) = \sum_{i=1}^{I} \| \mathbf{y}_i^* - \sum_{h=1}^{I} \mathbf{X}_i^* \, \mathbf{a}_h b_{hi} \|^2.$$

Write $f(\mathbf{a}_h)$ to denote that function $f(\mathbf{A}_c, \mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_I)$ is to be minimized over column $\mathbf{a}_h$ only, while the other columns of $\mathbf{A}_c$ and the vectors $\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_I$ are fixed. Isolating the part containing $\mathbf{a}_t$ in this equation yields

$$f(\mathbf{a}_t) = \sum_{i=1}^{I} \| (\mathbf{y}_i^* - \sum_{h \neq t}^{S} \mathbf{X}_i^* \mathbf{a}_h \, \mathbf{b}_{hi}) - \mathbf{X}_i^* \mathbf{a}_t b_{ti} \|^2.$$

We define

$$\mathbf{y}_{i(-t)}^* = \mathbf{y}_i^* - \sum_{h \neq t}^{S} \mathbf{X}_i^* \mathbf{a}_h b_{hi}$$

and simplify f $(\mathbf{a}_t)$as

$$f(\mathbf{a}_t) = \sum_{i=1}^{I} \| \mathbf{y}_i^{(-t)*} - \mathbf{X}_i^* \mathbf{a}_t b_{ti} \|^2$$

$$= \left\| \begin{pmatrix} \mathbf{y}_{1(-t)}^* \\ \vdots \\ \mathbf{y}_{I(-t)}^* \end{pmatrix} - \begin{pmatrix} b_{t1} \mathbf{X}_i^* \\ \vdots \\ b_{tI} \mathbf{X}_I^* \end{pmatrix} \mathbf{a}_t \right\|^2.$$

Minimizing expression over $\mathbf{a}_t$ is a simple linear regression problem. Clearly, the update for $\mathbf{a}_t$ is

$$\mathbf{a}_t=(\sum_{i=1}^{I} b_{ti}^2 \mathbf{X}_i^{*\prime}\mathbf{X}_i^{*})^{-1}\ (\sum_{i=1}^{I} b_{ti}\mathbf{X}_i^{*\prime}\mathbf{y}_{i(-t)}^{*}).$$

It should be noted that this algorithm has the advantage that it does not use matrices of order JxS by JxS. The largest matrix that has to inverted in this algorithm is J by J matrix, which allows handling large numbers of variables.

The ALS procedure presented here decrease function f monotonely and the convergence to a stationary point is guaranteed because each problem is solved in the least sqares sence. However it can not be guaranteed that the global minimum will be attained. Therefore, it is suggested to run more than one analysis on the same data set with different starting values.

Optimal Scaling

As mentioned before, the POPREG extends regression analysis to data defined at all four level: ratio, interval, ordinal and nominal measurement, which was proposed by Stevens.[9] Moreover we assume two types of measurement process: discrete and continuous. For analysis designed for data having such a wide variety of measurement, Fisher's notion of optimal scaling[5] is useful. According to his notation, we wish to obtain the optimally scaled data which fit the model as well as possible in a least squares sence. In other words, we rescale the data so that multiple correlation is maximized.

For the numerical data, the optimal scaling phase is skipped. For the ordinal data, Kruskal's least squares monotonic regression[7] can be used. In this case, the primary approach to tie is chosen for contiuous-ordinal data, whereas the secondary approach to tie is chosen for discrete-ordinal data. For the discrete-nominal data, optimally scaled data are category means. Finally, for the continuous-nominal data we assume it to be pseudo-continuous-ordinal data to determine the optimally scaled data.

We will not discuss the details of each optimal scaling method and the measurement restriction as they are the same as in the de Leeuw's paper.[4]

Starting Values

If there are many parameters, like POPREG, the number of iterations may be excessive in ALS procedure, but can be considerably decreased by the provision of good starting values. We simply assume that $\mathbf{X}$ and $\mathbf{y}$ are actually the matrix $\mathbf{X}^*$ and $\mathbf{y}^*$. This is equivalent to assuming that the raw data are measured on interval scale. But we must assign arbitrary values to the observation categories when a variable is assumed to be nominal. Under this assumption, we provide some types of the starting values which is appropriate for the fitted model. Moreover, uing this starting values, we can easily investigate whether our assumption concerning measurement levels are correct.

We decompose regression weight as

$$\mathbf{W}=\mathbf{A}_c\mathbf{B}+\mathbf{E},$$

where

$$\mathbf{W}=(\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_I),$$
$$\mathbf{w}_i=(\mathbf{X}_i'\ \mathbf{X}_i)\ \mathbf{X}_i'\ \mathbf{y}_i,$$
$$\mathbf{B}=(\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_I),$$
$$\mathbf{E}=(\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_I).$$

In order to obtain starting values we factor $\mathbf{W}$ into its principal component. Since this matrix is an asymmetric and rectangular the usual factoring equation are not appropriate. Therefore we utilize singular-value-decomposition (SVD) to obtain the best approximation matrix of W under the constrain that $\mathbf{A}_c'\ \mathbf{A}_c=\mathbf{I}_s$ and $\mathbf{B}\ \mathbf{B}'=$diag. The cross products matrix in the present analysis should always be computed between the variables on the shorter side of the approximated matrix, summing over the variables on the longer side. Thus, if $J > I$,

$$\mathbf{A}_c=\mathbf{W}\mathbf{V}_s\Gamma_s^{-1},$$
$$\mathbf{B}=\Gamma_s\mathbf{V}_s',$$

where

$$\mathbf{P}=\mathbf{W}'\mathbf{W},$$
$$=\mathbf{V}_s \Gamma_s^2 \mathbf{V}_s'+\mathbf{E}$$

and where Γ_s^2 and $\mathbf{V}_s$ denote the diagonal matrix which elements are the S largest eigen values of P and the matrix of the corresponding eigen vectors, respectively. Conversely, if J<I,

$$\mathbf{A}_c=\mathbf{U}_s,$$
$$\mathbf{B}=\mathbf{U}_s'\mathbf{W},$$

where

$$\mathbf{Q}=\mathbf{W}\mathbf{W}'$$
$$=\mathbf{U}_s \Gamma_s^2 \mathbf{U}_s'+\mathbf{E},$$

and where Γ_s^2 and $\mathbf{U}_s$ denote the diagonal matrix which elements are the S larest eigen values of $\mathbf{Q}$ and the corresponding eigen vactor, respectively.

Of course, we can use random values as an alternative.

Identification

Before an attempt is made to estimate model parameters of this kind, the identification problem must be examined. The identification problem depend on the specification of fixed, free and constrained parameters. Under several specifications, each $\mathbf{A}_c$ and $\mathbf{b}_i$ generates one and only one $\mathbf{w}_i$, but it is well known that different $\mathbf{A}_c$ and $\mathbf{b}_i$ can generate the same $\mathbf{w}_i$. It shold be noted that if $\mathbf{A}_c$ is replaced by $\mathbf{A}_c\,\mathbf{T}^{-1}$ and $\mathbf{b}_i$ by $\mathbf{T}\mathbf{b}_i$, where $\mathbf{T}$ is an arbitrary non-singuler matrix of order SxS, then $\mathbf{w}_i$ is unchanged. Since $\mathbf{T}$ has S^2 independent elements, this suggests that S^2 independent conditions shold be imposed on $\mathbf{A}_c$ or $\mathbf{b}_i$ to make these uniquely defined. Howerer, when equality constraints overgroups are taken into account, all the elements of the transformation matrix is not independent of each other and therefore a lesser number of conditions need to be imposed. It is hard to give further specific rule in the general case. In this method $\mathbf{A}_c$ and $\mathbf{b}$ should be estimated without any constraints. If the unrotated dimensions is interpretable, then rotation is unnecessary. If not, some objective rotation can be tried as in the case of factor analysis.

However, we can impose some constraints on this model for determining unique solution. Within the frame work of the general procedure, the most convenient way of doing this is to let the $\mathbf{b}_i$ be free and to fix one nonzero element and at least S-1 zeros in each column of $\mathbf{A}_c$. In an exploratory study, one can fix exactly S-1 zeros in almost arbitrary position. For example one may choose zero values where one thinks there shold be "small" values in coordinate. The resulting solution may be rotated further, if desired, to facilitate better interpretation. In the confirmatory study, on the other hand, the positions of the fixed zeros, which often exceed S-1 in each column, given a priori by an hypothesis and the resulting solution can not be rotated without destroying the fixed zeros.

In order to handle this case, the vectors $\mathbf{b}_i$ are updated as in the previous algorithm, but matrix $\mathbf{A}_c$ is updated as will be explained. The constraints mentioned above can be expressed in the form

$$\mathbf{r}=\mathbf{R}\ \mathrm{Vec}(\mathbf{A}_c),$$

where $\mathbf{r}$ is a known vector of G elements $(=\mathbf{0})$, G being the number of constraints, and $\mathbf{R}$ is a known matrix of order GxJS. We must choose Vec $(\mathbf{A}_c)$ to minimize $f(\mathbf{A}_c, \mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_I)$ subject to $\mathbf{r}=\mathbf{R}$ Vec $(\mathbf{A}_c)$. We define

$$\phi=(\mathbf{y}-\mathbf{Z}\mathbf{a})'\,(\mathbf{y}-\mathbf{Z}\mathbf{a})+\mu'\,(\mathbf{R}\mathbf{a}-\mathbf{r}),$$

where

$$\mathbf{y}=(\mathbf{y}_1{}^{*\prime}, \mathbf{y}_2{}^{*\prime}, \cdots, \mathbf{y}_I{}^{*\prime})',$$

$$\mathbf{Z}=\begin{pmatrix}\mathbf{X}_1{}^{*}\otimes \mathbf{b}_1' \\ \vdots \\ \mathbf{X}_I{}^{*}\otimes \mathbf{b}_I'\end{pmatrix},$$

$$\mathbf{a}=\mathrm{Vec}\ (\mathbf{A}_c),$$

and where μ is a column vector of G Lagrange multipliers. Differentiating and setting zero and solving, we obtain

$$\mathbf{a}=\alpha+(\mathbf{Z}'\,\mathbf{Z})^{-1}\mathbf{R}'\ \ [\mathbf{R}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{R}'\ ]^{-1}(\mathbf{r}-\mathbf{R}\alpha),$$

where $\alpha = (\mathbf{Z}' \mathbf{Z})^{-1}\mathbf{Z}' \mathbf{y}$ is the unconstrained OLS estimator. This is update for **a**.

Dimensionality

Up to this point, the discussion of POPREG has proceeded as if the number of dimensions S were known. In practice, however, it is not known and must be estimated in the analysis. In most multivariate analysis without external criterion like MDS, user must obtain several solutions in different dimensionalities and choose between them on the basis of three criterria: fit to the data, interpretability and reproducibility.

A plot can be useful in determining dimensionality. The vertical axis represents measure of goodness or badness of fit, and the horizontal axis corresponds to dimensions. If the data conform exactly to the model, then the plot shoud level off at exactly S dimensions. In other words, there should be an "elbow" at S dimensions, the correct number of dimensions. Howerer, in the real data in which there is a large amount of measurement and sampling error, an elbow may be difficult to discern. In such case a plot of measure of goodness or badness of fit may not suffice to determine the correct number of dimensions. Interpretability and reproducibility of dimensions must be also considered.

Interpretability as a criterion requires some subjective judgement on the user's part. The basic idea, howerer, is that a higher dimensional solution is preferrd over a lower dimensional solution if there are important independent (predictor) variables' feature that appear in the higher dimensional solution but fail to appear in the lower dimensional solution. Conversely the lower dimensional solution is preferred if there are no important features that fail to appear in the lower dimesional solution.

Reproducibility can be used as a criterion only when there are two or more subsamples. The basic idea is that one should retain as many dimensions in the final solution as emerge consistently in the separate subsamples. If one derives separate solutions for each subsample, and there are S dimensions that appear consistently in all of the subsamples, then the final solution should contain exactly S dimensions. In this case, howerer, each of the subsamples should come from the same population.

Of course, as will be explained later, we can also use Akaike's Information Criterion (AIC) to determine the number of dimensions from statistical point of view.[1]

Data Transformation

Exploring and assessing the effect of data transformation is not technically difficult. The transformation of predictor variables includes such transformation as discretization (categorization) of continuous variables as well as more standard type of transformations such as power, logistic, exponential, polynominal and spline transformation.

When continuous variables is nonlinealy or nonmonotonically contributing to prediction, we may discretize it into a few observation categories, which are then requantified by POPREG. A potential danger is that the effect of the continuous variable may indeed be linear. Then we may not only lose some information in the original variable in the process of discretization, but also lose degree of freedom by estimating extra parameters. Howerer, whether or not a particular transformation scheme on the predictor variables is worth incorporating can be tested using the general model evaluation strategy.

Missing Data

Missing data are allowed for in a manner which does not destroy the ALS property of the POPREG algorithm. If some observation is missing, then computation of the starting values is changed in a minor manner, that is, we simply estimate the optimal scaling observation as being the mean of the nomissing obsevations. Using these starting values model parameters

are estimated. Nextly the missing data points are re-estimated in a regression fashon and then a new cycle of the iteration is started. In fact such procedures are standard within the ALS approach.

Measure of Goodness or Badness of Fit

In this method, we can use four types of goodness or badness of fit measures.

1) Sum of squared error (SSE)

$SSE=(\mathbf{y}^*-\hat{\mathbf{y}}^*)'(\mathbf{y}^*-\hat{\mathbf{y}}^*)$,

$SSE_i=(\mathbf{y}_i^*-\hat{\mathbf{y}}_i^*)'(\mathbf{y}_i^*-\hat{\mathbf{y}}_i^*)$,

where SSE and SSE_i are the total sum of squared error and the sum of squared error in i-th group.

2) Stress

In the case that the predicted variable is nonmetric measure, Stress proposed by Kruskal[7] is appropriate measure of badness of fit.

3) Multiple correlation or correlation ratio

r or $\eta=[1-SSE/(\mathbf{y}^*,\mathbf{y}^*)]^{1/2}$,

r_i or $\eta_i=[1-SSE_i/(\mathbf{y}_i^*,\mathbf{y}_i^*)]^{1/2}$,

where r (η) and r_i (η_i) are the correlation (correlation ratio) in total and in the i-th group, respectively and where SSE and SSE_i were indicated above.

4) Akaike's Information Criterion (AIC)[1]

$AIC=2NI\log_e\hat{\sigma}_e+2P$,

where $\hat{\sigma}_e=[SSE/(NI-P-1)]^{1/2}$, N denotes the number of samples and P denotes the effective number of parameters in the fitted model. The AIC is a badness of fit measure that corrects for the gain in goodness of fit due to an increased number of parameters. The model with the smallest AIC is said to give the most parsimonious representation of data and this estimates is so-called MAICE (minimum AIC estimate). But the use of AIC should be limited to large sample problem because the AIC has been derived based upon the asymptotic property of maximum likelihood estimators.

Relation to Other Method

An important special case of POPREG model is obtained when we assume

$\mathbf{X}_1=\mathbf{X}_2=\cdots=\mathbf{X}_I=\mathbf{X}$,

$\mathbf{A}'\mathbf{X}'\mathbf{1}_N=\mathbf{0}_S$,

$N^{-1}\mathbf{A}'\mathbf{X}'\mathbf{X}\mathbf{A}=\mathbf{I}_S$.

According to these assumptions, matrix $\mathbf{A}$ and $\mathbf{F}=\mathbf{X}\mathbf{A}$ can be interpreted as matrix of component (factor) score coefficient and component (factor) scores, respectively. As you know, the resulting model is equivalent to"princical compontent analysis of instrumental variables "which was proposed by Rao.[8] Under these constraints, the model can be rewritten as

$$\mathbf{Y}^*=\mathbf{X}^*\mathbf{A}\mathbf{B}+\mathbf{E},$$

where

$\mathbf{Y}^*=(\mathbf{y}_1^*,\mathbf{y}_2^*,\cdots,\mathbf{y}_I^*)$,

$\mathbf{X}^*=(\mathbf{x}_1^*,\mathbf{x}_2^*,\cdots,\mathbf{x}_J^*)$,

$\mathbf{A}=(\mathbf{a}_1,\mathbf{a}_2,\cdots,\mathbf{a}_S)$,

$\mathbf{B}=(\mathbf{b}_1,\mathbf{b}_2,\cdots,\mathbf{b}_I)$,

$\mathbf{E}=(\mathbf{e}_1,\mathbf{e}_2,\cdots,\mathbf{e}_I)$.

Additionally, in order to obtain unique parameters we impose a side condition:

$$\mathbf{B}\mathbf{B}'=\mathbf{I}_S.$$

In this case, the probiem of minimizing function f is reduced to eigen-problem, that is,

$$|\mathbf{R}_{XY}\mathbf{R}_{YX}-\lambda\mathbf{R}_{XX}|=\mathbf{0},$$

where

$$\mathbf{R}=\begin{pmatrix}\mathbf{R}_{XX} & \mathbf{R}_{XY}\\ \mathbf{R}_{YX} & \mathbf{R}_{YY}\end{pmatrix}$$

$$=N^{-1}(\mathbf{X}^*:\mathbf{Y}^*)'(\mathbf{X}^*:\mathbf{Y}^*).$$

Therefore function f is minimized over $\mathbf{A}$ and $\mathbf{b}$ by choosing

$\mathbf{A}=\mathbf{R}_{XX}^{-\frac{1}{2}}\mathbf{V}$,

$\mathbf{B}=\mathbf{A}'\mathbf{R}_{XY}$,

where

$$(\mathbf{R}_{XX}^{-\frac{1}{2}}\mathbf{R}_{XY}\mathbf{R}_{YX}\mathbf{R}_{XX}^{-\frac{1}{2}})\mathbf{V}=\Lambda\mathbf{V},$$

and where Λ is a diagonal matrix of containing characteristic roots and $\mathbf{V}$ is a matrix containing the corresponding characteristic vectors. Obviously, this problem is a canonical correlation analysis subject to

$$\mathbf{R}_{YY}=\mathbf{I}.$$

From this relation, POPREG can be regarded as a general model of "simultaneous principal component analysis for instrumental variables" in that this model is not subject to the constraints mentioned above.

Nextly, we discuss difference between "principal component regression analysis" (PCRA) proposed by Kendall[6] and other models mentioned before. As is well known, PCRA is formulated as

$$\begin{aligned}\mathbf{y}&=\mathbf{X}\beta+\mathbf{e}\\&=(\mathbf{XV})(\mathbf{V}'\beta)+\mathbf{e}\\&=\mathbf{F}\Upsilon+\mathbf{e},\end{aligned}$$

where

$$\mathbf{X}'\mathbf{X}=\mathbf{V}\Lambda\mathbf{V}'\ (\mathbf{V}'\mathbf{V}=\mathbf{I}),$$

and where $\mathbf{F}=\mathbf{XV}$ denotes matrix of principal component socores and $\Upsilon=\mathrm{V}'\beta$ denotes vector of regression weights of principal components on $\mathbf{y}$. However principal component regression estimators of Υ are formed by deleting (J-S) columns of F. Then the principal component regression estimators of Υ and β are

$$\begin{aligned}\tilde{\Upsilon}&=(\mathbf{F}_S'\mathbf{F}_S)^{-1}\mathbf{F}_S'\,\mathbf{y},\\\tilde{\beta}&=\mathbf{V}_S\tilde{\Upsilon}\\&=\mathbf{V}_S(\mathbf{F}_S'\mathbf{F}_S)^{-1}\mathbf{F}_S'\mathbf{y}.\end{aligned}$$

Therefore PCRA and other models described before differ in the way in which the factors are estimated. The factors in PCRA is determined so that variance of "predictors" explained by factors is maximized. On the other hand, the factors in other methods is determined so that variance of "predicted variable" explained by factors is maximized.

Lastly, if we assume that

$$\begin{aligned}&\mathbf{X}_1=\mathbf{X}_2=\cdots=\mathbf{X}_I=\mathbf{X},\\&\mathrm{J}=\mathrm{S},\\&\mathbf{A}_c=\mathbf{I}_S,\end{aligned}$$

this model is equal to "multivariate regression model" This model can be rewitten as

$$\mathbf{Y}=\mathbf{XB}+\mathbf{E},$$

where

$$\begin{aligned}&\mathbf{Y}=(\mathbf{y}_1,\mathbf{y}_2,\cdots,\mathbf{y}_I),\\&\mathbf{X}=(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_J),\\&\mathbf{B}=(\mathbf{b}_1,\mathbf{b}_2,\cdots,\mathbf{b}_I),\\&\mathbf{E}=(\mathbf{e}_1,\mathbf{e}_2,\cdots,\mathbf{e}_I).\end{aligned}$$

In this case, least squares estimate of $\mathbf{B}$, as you known, is

$$\mathbf{B}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Additionally, if we impose constraints that

$$\begin{aligned}&\mathrm{I}=\mathrm{J}=\mathrm{S},\\&\mathbf{B}'\mathbf{B}=\mathbf{I}_S,\end{aligned}$$

loss function is defined as

$$\begin{aligned}\phi&=\mathrm{tr}\ [(\mathbf{Y}-\mathbf{XB})'(\mathbf{Y}-\mathbf{XB})]\\&\quad+\mathrm{tr}\ [(\mathbf{B}'\mathbf{B}-\mathbf{I}_S)\mathbf{L}]\ ,\end{aligned}$$

where L is a matrix of Lagrange multipliers. Differentiating and setting zero, we obtain

$$\mathbf{B}=\mathbf{VW}',$$

where

$$\begin{aligned}&\mathbf{Y}'\mathbf{X}\mathbf{X}'\mathbf{Y}=\mathbf{W}\Delta_S\mathbf{W}'\quad(\text{eigen-decomposition}),\\&\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}=\mathbf{V}\Delta_S\mathbf{V}'\quad(\text{eigen-decomposition}).\end{aligned}$$

Fron this relation, POPREG can be also considered as a general model of"multivariate regression model"

Extension to Canonical Model

Finally we emphasize that POPREG model can be easily extended to canonical model (POPCAN) in the situation where there are two sets of data obtained from two or more populations. This model is defined as

$$\mathbf{Y}^*\mathbf{w}_Y=\mathbf{X}^*\mathbf{w}_X+\mathbf{e},$$

i.e.,

$$\mathbf{Y}^*\mathbf{A}_Y\mathbf{b}_Y=\mathbf{X}^*\mathbf{A}_x\mathbf{b}_X+\mathbf{e}.$$

In this case, loss function for estimating $\mathbf{b}_X$ and $\mathbf{b}_Y$, for fixed $\mathbf{A}_X$, $\mathbf{A}_Y$, $\mathbf{X}_i^*$ and $\mathbf{Y}_i^*$, can be difined as

$$h_i(\mathbf{b}_{Xi}, \mathbf{b}_{Yi})= \| \mathbf{Z}_{Xi}\mathbf{b}_{Xi}-\mathbf{Z}_{Yi}\mathbf{b}_{Yi} \|^2,$$

where

$$\mathbf{Z}_{Xi}=\mathbf{X}_i^*\mathbf{A}_{Xc},$$
$$\mathbf{Z}_{Yi}=\mathbf{Y}_i^*\mathbf{A}_{Yc}.$$

In this case, however, some scale constraints must be imposed on model parameters in order to avoid degenerate solution and therefore it is obvious that the problem of minimizing function h_i over $\mathbf{b}_{Xi}$ and $\mathbf{b}_{Yi}$, while $\mathbf{Z}_{Xc}$ and $\mathbf{Z}_{Yc}$ are fixed, is equivalent to that of canonical correlation analysis. Under this situation, the minimizing problem is reduced to eigen-problem, that is,

$$| \mathbf{C}_{XYi}\mathbf{C}_{YYi}^{-1}\mathbf{C}_{YXi}- \lambda \mathbf{C}_{XXi} | =0.$$

As you know, function f is minimized over $\mathbf{b}_{Xi}$ and $\mathbf{b}_{Yi}$ by choosing:

$$(\mathbf{C}_{XXi}^{-\frac{1}{2}}\mathbf{C}_{XYi}\mathbf{C}_{XYi}^{-1}\mathbf{C}_{YXi}\mathbf{C}_{XXi}^{-\frac{1}{2}})\mathbf{w}= \lambda \mathbf{w}$$

$$\mathbf{b}_{Xi}=\mathbf{C}_{XXi}^{-\frac{1}{2}}\mathbf{w}$$

$$\mathbf{b}_{Yi}= \lambda^{-\frac{1}{2}}\mathbf{C}_{YYi}^{-1}\mathbf{C}_{XXi}\mathbf{b}_{Xi},$$

where

$$\mathbf{C}_i=\left(\begin{array}{c|c}\mathbf{C}_{XXi} & \mathbf{C}_{XYi}\\ \hline \mathbf{C}_{YXi} & \mathbf{C}_{YYi}\end{array}\right)$$
$$=N^{-1}(\mathbf{Z}_{Xi} : \mathbf{Z}_{Yi})'(\mathbf{Z}_{Xi} : \mathbf{Z}_{Yi}).$$

Using this method, we can obtain both $\mathbf{b}_{Xi}$ and $\mathbf{b}_{Yi}$ at once. A careful reader will detect that a problem of minimizing f over $\mathbf{A}_{Xc}$ and $\mathbf{A}_{Yc}$ for fixed $\mathbf{b}_X$, $\mathbf{b}_Y$, $\mathbf{X}_i^*$ and $\mathbf{Y}_i^*$, can be solved in a similar way. That is, the minimising problem can be formulated as

$$h(\mathbf{A}_{Xc}, \mathbf{A}_{Yc})$$
$$= \left\| \begin{pmatrix}\mathbf{X}_1^*\otimes \mathbf{b}_{X1}'\\ \vdots \\ \mathbf{X}_I^*\otimes \mathbf{b}_{XI}'\end{pmatrix}\mathrm{Vec}(\mathbf{A}_{XC}) - \begin{pmatrix}\mathbf{Y}_1^*\otimes \mathbf{b}_{Y1}'\\ \vdots \\ \mathbf{Y}_I^*\otimes \mathbf{b}_{YI}'\end{pmatrix}\mathrm{Vec}(\mathbf{A}_{YC}) \right\|^2,$$

and this minimizing problem is also identical with ordinary canonical correlation analysis. However these method consists of an infinite iteration process in which at each step an eigen-problem have to be solved. Clearly, solving this eigen-problem by an infinite iteration process has its drawbacks. The whole procedure is likely to become computationally burdensome. In order to avoid this, we can simply impose the estimation phase for $\mathbf{A}_Y$ and $\mathbf{b}_Y$ on the previous algorithm. In this alternative, moreover, we must rescale the model parameters after each model estimation phase in order to avoid the degenerate solution.

A Computer Program

A program, POPREG, was developed for computing the solution. It was written in FORTRAN 77 for FACOM M-780/20 computer system.

In the near future, we will examine validation and efficiency of these algorithms with artificial and real data.

Reference

1) Akaike H (1974): A new look at the statistical model identification. IEEE Transactions on Automatic Control 19: 716-723.

2) Carroll JD and Chang JJ (1970): Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decoposition. Psychometrika 35: 238-319.

3) de Leeuw J (1969): The linear nonmetric model (report RN003-69) .Leiden, the Netherlands: University of Leiden.

4) de Leeuw J, Young FW, and Takane Y (1976): Additive structure in qualitative data: An alternating least squares method with optimal scaling feature. Psychometrika 41: 471-503.

5) Fisher RA (1946): Statistical methods for research workers (10th ed.). Edinburgh, Oliver and Boyd.

6) Kendall MG (1957): A course in multivariate analysis. London, England: Charles Griffin.
7) Kruskal JB (1964): Nonmetric multidimensional scaling: A numerical method . Psychometrika 29: 28-42.
8) Rao CR (1964): The use and interpretation of principal component analysis in applied research. Sankhya, series A, Vol. 26, part 4, pp. 329-358.
9) Stevens SS (1951): Mathematics, measurement, and psychophysics. (Ed.) Stevens SS (In) Handbook of Experimental Psychology. New York, Wiley.
10) Wold H and Lyttkens E (1969): Nonlinear iterative partial least squares (NIPALS) estimation procedure . Bull ISI 43: 29-47.
11) Young FM(1972): A model for polynominal conjoint analysis algorithms. (Eds.) Shepard RN, Romney AK, and Nerlove S, (In) Multidimensional scaling: Theory and Applications in the Behavior-Sciences. New York, Academic Press.
12) Young FW, de Leeuw J, and Takane Y (1976): Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling feature. Psychometrika 41: 505-529.