Contributions to the homogeneity analysis of curves and processes

Jan de Leeuw
Department of Data Theory FSW/RUL
Middelstegracht 4
2312 TW Leiden

Paper prepared for the
Table Ronde "Analyse des Données"
Toulouse, january 1984

Principal Component Analysis of Markov Chains

Suppose $\underline{x}_t$ is a Markov Chain on (0,T), with k states, and stationary transition probabilities. Suppose $0 < t_1 < \ldots < t_m < T$, and consider the random vector $(\underline{x}_1,\ldots,\underline{x}_m)$, where $\underline{x}_j$ is short for $\underline{x}_{t_j}$.

The cross table of variables $\underline{x}_j$ and $\underline{x}_\ell$, with $j < \ell$, is $C_{j\ell} = D_j P_{j\ell}$, with D_j the marginals of $\underline{x}_j$ and with $P_{j\ell}$ the transition matrix. Under regularity conditions we know that

$$P_{j\ell} = \exp((t_\ell - t_j)Q) = \sum_{s=0}^{\infty} \frac{(t_\ell - t_j)^s}{s!} Q^s,$$

with Q the intensity matrix of the process. The eigen-equations of homogeneity analysis are

$$\sum_{\ell=1}^{m} D_j P_{j\ell} x_\ell = m\lambda D_j x_j.$$

Now suppose Y is a linear independent system of eigenvectors of Q. Thus $QY = Y\Omega$, with Ω diagonal. Write x_j in the form $Y\alpha_j$. Then

$$P_{j\ell} x_\ell = P_{j\ell} Y\alpha_\ell = Y \exp((t_\ell - t_j)\Omega)\alpha_\ell,$$

and thus we must solve

$$\sum_{\ell=1}^{m} \exp((t_\ell - t_j)\Omega)\alpha_\ell = m\lambda\alpha_j.$$

This eigenproblem has all submatrices of the Burt-table diagonal. By rearranging, as in De Leeuw (COMPSTAT, 1982), this problem can also be written as

the eigenproblem for the direct sum $R_1 \dotplus \ldots \dotplus R_k$, where the R_i are m x m matrices with elements

$$(R_i)_{j\ell} = \exp((t_\ell - t_j)\omega_i).$$

If the t_j are equally spaced, this can be written as

$$(R_i)_{j\ell} = \theta_i^{(\ell-j)}.$$

In any case it is clear that simultaneous diagonalization of the sub-tables of the Burt-table, as explained in De Leeuw (COMPSTAT, 1982, Journal of Econometrics, 1983), is possible in this case. This adds a very important model to this list of models for which we can diagonalize simultaneously. We can solve the eigenproblem by solving the k eigenproblem for the R_i separately. The x_j are all proportional to one of the columns of Y, say y_i, with proportionality factors given by the eigenvector of the corresponding R_i. Thus each of the k columns of y defines m solutions to the original eigenproblem, giving the required total of mk solutions.

In stead of approximating the process by studying at m points, we can also approximate it by averaging over intervals (De Leeuw, Quantitative Harmonic Analysis). This gives basically the same result, but with

$$(R_i)_{j\ell} = \sum_{s=0}^{\infty} \frac{\omega_i^s}{s!} \iint (t - u)^s \, dtdu,$$

where t is integrated over (t_j, t_{j+1}) and u over $(t_\ell, t_{\ell+1})$.

Additional

If D_∞ gives the asymptotic probabilities of the states of the Markov process, then $u'D_\infty x_j = 0$ for all eigen-subvectors x_j. Thus if state k, for example, is absorbing, then $e_k'x_j = (x_j)_k = 0$ for all eigen-subvectors x_j.

We have shown that the category quantifications at time t are of the form $x_h(t) = \alpha(t)y_h$. Here y_h is element h of eigenvector y of Q, and $\alpha(t)$ is an eigenelement of the matrix with elements $\exp(\lambda(s - t))$. Thus if the k curves $x_h(t)$ are plotted against time they all have the same shape, in the sense that they are all multiples of $\alpha(t)$. They have maxima and minima at the same places, they are monotone in the same intervals, and they never cross, except possibly when they are both zero.

The function $\alpha(t)$ can be described more precisely by using the theory of total positivity (Gantmacher/Krein, or Karlin). We can show how it oscillates, and how many zero-crossing it has. This will be done elsewhere. For the moment is suffices to assert that we can recognize stationary Markov Chains from their PCA's.

Ordering within and between variables

1: Introduction

In the Gifi system order relations are defined within variables. Thus we study variables whose range is ordered, who take values in an ordered set. This information can or cannot be incorporated into the analysis. It is of interest to consider the case in which there is an order between variables, i.e. in which we know that variable j precedes variable ℓ in some well-defined sense. This happens, of course, with time series and event history data. But it can also happen with patterns of the Guttman, Thurstone, or Rasch variety. We give a simple (essentially one-dimensional) introduction. Extensions to multidimensional quantification are possible along the usual Gifi-lines. Part of this work has been inspired by the thesis of Besse (1979). For reasons of simplicity we restrict ourselves to the case of a finite number of variables in a finite dimensional space.

2: Loss

The loss is, as usual, defined in the following way. If the G_j are indicator matrices of variables 1,...,m (or other bases for the quantification spaces), then the quantified variables are the columns of Q, with column q_j given by $q_j = G_j y_j$. Here y_j is the k-vector of weights (or category quantifications). Homogeneity is defined as $q_1 = \ldots = q_m$. If the q_j are interpreted as real valued functions on $\{1,2,\ldots,n\}$, then homogeneity means that they must be the same function. But obviously the rows q_i of Q can also be interpreted as n functions on $\{1,2,\ldots,m\}$. In this interpretation

homogeneity means that the q_i are constant real valued functions, having the same value for all j.

Gifi-loss is, essentially,

$$\sigma(x;y_1,\ldots,y_m) = \sum_{j=1}^{m} (x - G_j y_j)'(x - G_j y_j).$$

Here x is the comparison function, also defined on $\{1,2,\ldots,n\}$, also known as the object scores. Thus we compare functions on I_n in order to define loss. But of course we can write equally well

$$\sigma(x;y_1,\ldots,y_m) = \sum_{i=1}^{n} (x_i u - q_i)'(x_i u - q_i).$$

Here $x_i u$ is the function on $I_m = \{1,2,\ldots,m\}$ with is equal to x_i for all j in I_m.

In this alternative interpretation of the Gifi-loss (of course all these formulations are variations on the basic duality) we use the y_j to induce functions q_i in I_m. They are homogeneous if they are constant functions. But this immediately suggests alternative definitions of homogeneity. Maybe we should call the q_i homogeneous if they are all straight lines, or low-degree polynomials, or low-degree splines on a given scale. All these definitions can be incorporated quite simply be using a basis H for the comparison functions, and by letting

$$\sigma(x_1,\ldots,x_n;y_1,\ldots,y_m) = \sum_{i=1}^{n} (Hx_i - q_i)'(Hx_i - q_i).$$

3: Matrices

The G_j are n x k, and H is m x r. The x_i can be collected in an n x r matrix. Then

$$\sigma(X;Y) = \text{tr } (XH' - Q)'(XH' - Q) = \sum_{j=1}^{m} (Xh_j - G_j y_j)'(Xh_j - G_j y_j).$$

Observe that H is known, and the matrix X is unknown. This makes the problem different from a canonical analysis problem, which has the same form but with X known and h_j unknown.

4: Algorithm

This is not spectacular. We know that the optimal y_j is simply (usual notation) $y_j = D_j^{-1} G_j' X h_j$. Thus

$$\sigma(X;*) = \min_Y \sigma(X;Y) = \sum_{j=1}^{m} h_j' X' \bar{P}_j X h_j,$$

with $\bar{P}_j = I - P_j = I - G_j D_j^{-1} G_j'$. If we use the normalization condition

$$\sum_{j=1}^{m} h_j' X' X h_j = 1,$$

then we must solve the eigenproblem $Ax = \lambda Bx$, with x a concatenation of the n rows of X, with (using direct or Kronecker products)

$$A = \sum_{j=1}^{m} P_j \times (h_j h_j'),$$

$$B = \sum_{j=1}^{m} I \times (h_j h_j'),$$

Because A and B can be quite big (they are of order nr), it may be best to use alternating least squares. The optimal X for given Y is proportional to $QH(H'H)^{-1}$.

5: Dual eigenproblem

Minimizing $\sigma(X;Y)$ over X for fixed Y gives

$\sigma(*;Y) = \text{tr } Q'\bar{S}Q,$

where $\bar{S} = I - S = I - H(H'H)^{-1}H'$. By imposing tr $Q'Q = I$ this leads to the eigenvalue problem

$\tilde{C}y = \lambda Dy,$

where C and D are the Burt table and its diagonal, as usual, and where $\tilde{C}$ has submatrices $\tilde{C}_{j\ell} = s_{j\ell}C_{j\ell}$.

6: Earlier proposal

In qualitative harmonic analysis (previous note in same series, page A3) we proposed to restrict the y_j directly by requiring $y_j = Ah_j$. Here A is k x r, and unknown, and the h_j are known. This uses the classical definition of homogeneity, but restricts the shapes of the curves.