RN 003-69

THE LINEAR NONMETRIC MODEL

Jan de Leeuw

June 1969

PSYCHOLOGICAL INSTITUTE - UNIVERSITY OF LEYDEN - THE NETHERLANDS

1. General formulation.

Let X be a finite subset of the set of all m-tuples of real numbers. X has n elements $x_1, x_2, \ldots, x_n$. The coordinate values can be collected in a matrix $X = \{x_{ij}\}$, $i=1,\ldots,n$; $j=1,\ldots,m$. Suppose Y is another set of n elements, partially ordered by the relation $\succcurlyeq_o$. We are interested in the existence of real numbers $w_1, \ldots, w_m$ such that

$$\sum_{k=1}^{m} x_{ik} w_k \geq \sum_{k=1}^{m} x_{jk} w_k \quad \text{iff} \quad y_i \succcurlyeq_o y_j. \tag{1}$$

To establish representation theorems the easiest way is to use the method of Tversky (1964) and Scott (1964). Each of the requirements (1) can be translated into a homogeneous linear inequality. If $(y_i, y_j) \in \succcurlyeq_o$ then we require

$$\sum (x_{ik} - x_{jk}) w_k \geq 0 . \tag{2}$$

The representation theorems are the familiar theorems giving necessary and sufficient conditions for the existence of solutions of systems of linear inequalities. A very complete review of these theorems is given by Fan (1956). The uniqueness theorems, that state properties of the solution set, are most easily understood by using the geometric theory of polyhedral convex cones and their extreme faces.

The constructive approach to measurement theory, as practiced by people as Luce, Krantz, and Suppes, can also be used in this case. Let $\oplus$ denote (componentwise) vector addition in X. The axioms are chosen in such a way that $\langle X, \oplus, \succcurlyeq \rangle$ is a fully ordered Archimedean group, and by Hölder's theorem there exists a function f of X into the reals such that for all $x, y \in X$

$$\text{i)} \quad f(x \oplus y) = f(x) + f(y), \tag{3}$$

$$\text{ii)} \quad f(x) \geq f(y) \text{ iff } x \succcurlyeq_o y. \tag{4}$$

It is well known that the only continuous solutions of the functional equation (3) are the linear functions. Continuity of f can be guaranteed by imposing (axiomatically) restrictions on the order topology induced

on X by $\hat{z}$. We shall assume in the sequel

i) the columns of the matrix X are centered, i.e.

$$\sum_{i=1}^{n} x_{ij} = 0 \text{ for all } j=1,\ldots,m.$$

ii) Y is a set of real numbers. Without loss of generality we may suppose $\sum y_i = 0$.

iii) We assume that rank(X) = m, and consequently also that $n \geq m$.

In later sections we shall show that all three assumptions can be avoided and are not essential for the algorithms.

2: Alternating least squares algorithm.

In order to find the best value of w for our problem we use a Kruskal-type error theory. Define the loss function

$$S = \frac{\sum_{i=1}^{n} (z_i - \hat{z}_i)^2}{\sum_{i=1}^{n} (z_i - \bar{z})^2}, \tag{5}$$

with

$$z = Xw, \tag{6}$$

$\bar{z}$ is the mean value of the z_i, and $\hat{z}$ is any vector monotone with y. Throughout this paper j denotes a vector with all elements equal to unity. Because X is centered columnwise we have $j'X = 0$ and $\bar{z} = \frac{1}{n} j'Xw = 0$. Thus

$$S = \frac{(z - \hat{z})'(z - \hat{z})}{z'z} = \frac{w'X'Xw - 2w'X'\hat{z} + \hat{z}'\hat{z}}{w'X'Xw}. \tag{7}$$

Consider $\hat{z}$ as a fixed vector. Then S is minimized if we minimize its numerator under the condition that $w'X'Xw$ is some constant value (say unity). The Lagrangian function is

$$\Phi(w) = w'X'Xw - 2w'X'\hat{z} + \hat{z}'\hat{z} - \mu(w'X'Xw - 1), \tag{8}$$

and

$$\frac{\partial \Phi}{\partial w'} = 2X'Xw - 2X'\hat{z} - 2\mu X'Xw. \tag{9}$$

We must solve the system of equations

$$(1 - \mu)X'Xw = X'\hat{z}, \tag{10}$$

$$w'X'Xw = 1, \tag{11}$$

for w and μ. The solution is

$$w = (\hat{z}'X(X'X)^{-1}X'\hat{z})^{-\frac{1}{2}} (X'X)^{-1}X'\hat{z}, \tag{12}$$

$$\mu = 1 - (\hat{z}'X(X'X)^{-1}X'\hat{z})^{\frac{1}{2}}. \tag{13}$$

For a first approximation we take $\hat{z} = y$. Observe that the problem of minimizing S for fixed $\hat{z}$ is closely related to the oblique Procrustus problem discussed by Browne (1967). There we require $w'w = 1$ in stead of $w'X'Xw = 1$ and, as a consequence, the stationary equations must be solved iteratively. The next step is to compute new values of z, and to find $\hat{z}$ in such a way that S is minimized under the condition that $\hat{z}$ remains monotone with y. This problem is solved by the amalgamation method used by Kruskal. In summary

1) set $\hat{z}^{(0)} = y$, and set iteration counter $s = 1$;

2) find $w^{(s)}$ by (12) with $\hat{z} = \hat{z}^{(s-1)}$;

3) compute $z^{(s)} = Xw^{(s)}$;

4) find $\hat{z}^{(s)}$ by the almagamation method, compute $S^{(s)}$;

5) if $s = s_{max}$ then go to step 6 (see section 3), otherwise $s = s + 1$ and goto step 2.

3: Newton-Raphson iterations.

The iterative scheme in the previous section has the obvious property that $S^{(s+1)} \leq S^{(s)}$. In fact in each iteration the value of S is diminished two times, in step (2) and in step (4). We have $S^{(s+1)} = S^{(s)}$ iff $w^{(s+1)} = w^{(s)}$ iff $\hat{z}^{(s+1)} = \hat{z}^{(s)} = z^{(s)}$ iff $S^{(s)} = 0$. If $S_{min} > 0$ then the method does not converge in a finite number of steps. The rate of convergence is somewhat unsatisfactory (in general it can be compared to that of the optimal gradient method, which means that the method is very inefficient near the minimum). In our complete algorithm we use only a few (s_{max})

steps of this method. They are supposed to bring us close to the minimum, where S is dominated by the second order terms of its Taylor expansion, i.e. nearly quadratic. There the Newton-Raphson method becomes very efficient. It has the additional advantage that we obtain information about the distance from the minimum and about the curvature of the function. In order to develop the necessary formula's let (superscripts denoting the iteration number s are deleted)

$$e = X'z = X'Xw, \tag{14}$$

$$f = X'(z - \hat{z}), \tag{15}$$

$$S^{*} = (z - \hat{z})'(z - \hat{z}), \tag{16}$$

$$T^{*} = z'z. \tag{17}$$

Then

$$\frac{\partial S^{*}}{\partial w_k} = 2f_k, \tag{18}$$

and

$$\frac{\partial T^{*}}{\partial w_k} = 2e_k. \tag{19}$$

Thus

$$\frac{\partial S}{\partial w_k} = \frac{2}{T^{*}} (f_k - Se_k). \tag{20}$$

These partial derivatives are collected in the gradient vector g. The second order derivatives are given by

$$\frac{\partial^2 S}{\partial w_k \partial w_l} = \frac{2}{T^{*}} \left\{ (1 - S)c_{kl} - (g_l e_k + g_k e_l) \right\}, \tag{21}$$

with C = X'X. These elements are collected in the symmetric Hessian matrix V, and we assume that V is nonsingular. The Newton-Raphson scheme is simply

$$w^{(s+1)} = w^{(s)} - (V^{(s)})^{-1} g^{(s)}. \tag{22}$$

The estimated distance from the minimum is given by

$$S^{(s)} - S_{min} \approx \tfrac{1}{2} g^{(s)\prime} (V^{(s)})^{-1} g^{(s)}.$$

The quantity (say ρ) on the right hand side of (23) is a good stopping criterion for the iterations. Observe that the form of the second order derivatives suggests another modified gradient method which will also be quite efficient near the minimum and does not require computation and inversion of V. Simply take

$$w^{(s+1)} = w^{(s)} - \left[\frac{2(1-S^{(s)})}{(T^{*})^{(s)}}\right] C^{-1} g^{(s)}. \qquad (24)$$

The weighting matrix C^{-1} must be computed anyway. This method is an analogon of the method of scoring for parameters in ML-estimation. To conclude the scheme from the previous section

6) $t = 1$, $w^{(0)} = w^{(s)}$, $\hat{z}^{(0)} = z^{(s)}$, $S^{(0)} = S^{(s)}$;

7) compute $g^{(t)}$, $V^{(t)}$, $(V^{(t)})^{-1}$, and $\rho^{(t)}$;

8) if $\rho^{(t)} < \varepsilon$ then stop ;

9) compute $w^{(t)}$ by (22) ;

10) compute $z^{(t)} = Xw^{(t)}$;

11) compute $\hat{z}^{(t)}$ by the almagamation method ;

12) $t = t+1$, and go to step 7 .

4: Restrictive conditions.

The first restrictive condition we imposed in section 1 was that $\sum_i x_{ij} = 0$ for all j. Suppose this is not true. Then there are real numbers λ_j $(j=1,\ldots,m)$ such that

$$x_{ij} = \tilde{x}_{ij} + \lambda_j, \qquad (25)$$

where $\tilde{x}_{ij}$ is our original centered variate. Then, for a particular value of the vector w,

$$z_i = \sum w_k x_{ik} = \sum w_k \tilde{x}_{ik} + \sum \lambda_k w_k = \tilde{z}_i + \lambda \text{ (say)}. \qquad (26)$$

It follows from the nature of the almagamation algorithm that

$$\hat{z}_i = \hat{\tilde{z}}_i + \lambda . \qquad (27)$$

Moreover

$$\bar{z} = \frac{1}{n}\sum (\hat{\tilde{z}}_i + \lambda) = \bar{\tilde{z}} + \lambda = \lambda. \qquad (28)$$

And thus

$$z_i - \hat{z}_i = \tilde{z}_i - \hat{\tilde{z}}_i, \tag{29}$$

and

$$z_i - \bar{z} = \tilde{z}_i - \bar{\tilde{z}} = \tilde{z}_i. \tag{30}$$

Consequently $S = \tilde{S}$ for all vectors w. The minima of two identical functions are, of course, also identical.

The second restriction was that y must be a vector of real numbers. This restriction was used only in the computation of an initial estimate of w. It is easy to generalize the approach to an arbitrary weakly ordered set $\langle Y, \gtrless_o \rangle$. We use the CDARD4 technique (De Leeuw 1968) to obtain an initial estimate of w. Let

$$s_{ij} = \begin{cases} +1 & \text{if } y_i \gtrless_o y_j, \\ 0 & \text{if } y_i =_o y_j, \\ -1 & \text{if } y_i \lessgtr_o y_j. \end{cases} \tag{31}$$

Then we must maximize the coefficient

$$F = \frac{\sum_i \sum_j s_{ij}(z_i - z_j)}{\sum_i z_i^2}. \tag{32}$$

Simplifying the numerator (by using the fact that $s_{ij} = -s_{ji}$) yields

$$\sum_i \sum_j s_{ij}(z_i - z_j) = 2 \sum_i z_i \sum_j s_{ij}. \tag{33}$$

It is easily seen that

$$\sum_j s_{ij} = 2r_i - (n + 1), \tag{34}$$

where r_i is the rank number of y_i in the order, and where ties have equal (averaged) rank numbers. Moreover

$$\sum_i \sum_j s_{ij} = 2 \sum_i r_i - n(n + 1) = 0. \tag{35}$$

Writing $\tilde{r}_i$ for the centered rank numbers we obtain

$$F = \frac{2\tilde{r}'Xw}{w'X'Xw}. \tag{36}$$

The optimal solution for w is the same as in formula (12), with $\hat{z} = \tilde{r}$.

In the case of a partially ordered set the situation is less simple. We let $s_{ij} = 0$ if the order relation between y_i and y_j is undetermined. Formula's (32) and (33) can still be used but the simple relation of $\sum_i s_{ij}$ with rank numbers is not true. Define $\tilde{s}_i = \sum_j s_{ij}$ and use $\tilde{s}$ as $\hat{z}$ in (12). It is also possible to adapt the almagamation algorithm for partial orders. In the relevant step of the iterative program we have to solve the quadratic programming problem

$$\begin{cases} (z - \hat{z})'(z - \hat{z}) \quad \text{min !} & (37a) \\ A\hat{z} \geqslant 0, & (37b) \end{cases}$$

where A is a matrix constructed in such a way that any $\hat{z}$ satisfying (37b) satisfies the order relations. Perhaps the problem (37) is must easily solved by the Gauss-Seidel type method for quadratic programming proposed by Hildreth and d'Esope and disccused by Kunzi and Krelle (1962).

The third (and most important) restriction in section 1 was that $c = X'X$ must be nonsingular. Observe that the only place in which we use this is section 2, where C^{-1} is computed (only once). Consider the normal equations

$$X'Xw = X'\hat{z}. \qquad (38)$$

These equations always have at least one soultion. Proof: a system of linear equations $Ax = b$ is consistent iff for all y for which $y'A = 0$ it is also true that $y'b = 0$. Because $X'X$ is positive semidefinit, $y'X'Xy = 0$ implies $y'X'X = 0$. The converse is obvious. Thus: $y'X'X = 0$ iff $y'X'Xy = 0$ iff $y'X' = 0$, which implies that $y'X'z = 0$. Q.E.D.
It is well known that the solution of (38) for w is unique iff $X'X$ is of full rank. If (38) is consistent, but $\text{rank}(C) < m$ then there are more solutions. A general inverse (g-inverse) of an n x m matrix A is defined as an m x n matrix A^- such that for each consistent system $Ax = b$ the vector $x = A^-b$ is a solution. It is not difficult to see that A^- is the g-inverse of A iff $AA^-A = A$ iff $A^-AA^- = A^-$. Because the

normal equations are consistent the vector

$$w = (X'X)^{-}X'z \tag{39}$$

is a solution. The general solution is given by

$$w = C^{-}X'z + (C^{-}C - I)b, \tag{40}$$

with arbitrary b. In our procedure we have to scale the vector w afterwards in such a way that $w'X'Xw = 1$. Some practical ways to compute a g-inverse of C in the special case of normal equations are given by Rao (1965). Let the rank of X be $r < m$, then it is possible to construct an $(m-r)xm$ matrix H such that the matrix

$$\tilde{X} = \begin{bmatrix} X \\ \cdots \\ H \end{bmatrix} \tag{41}$$

is of rank m. Then $\tilde{C} = \tilde{X}'\tilde{X} = X'X + H'H$ posesses a true inverse, and this inverse $\tilde{C}^{-}$ is a g-inverse of C. Another method (which is especially useful if we also need the canonical form of C or the Eckart-Young decomposition of X) is to write C as $K\Lambda K'$, with $K'K = I$ and Λ diagonal. Let Λ^{-} denote the diagonal matrix with elements λ_{ii}^{-1} if $\lambda_{ii} > 0$, and 0 otherwise. Then Λ^{-} is a g-inverse of Λ, and $K\Lambda^{-}K'$ is a g-inverse of C. The final method we discuss is to delete the $m - r$ dependent rows and columns of C (without loss of generality this may be taken to be the last ones) and to invert the resulting r x r matrix C_r. Then the m x m matrix

$$\begin{bmatrix} C_r^{-1} & \vdots & 0 \\ \cdots & \vdots & \cdots \\ 0 & \vdots & 0 \end{bmatrix} \tag{42}$$

is a g-inverse of C.

For the similar assumption in the Newton-Raphson equations the situation is less simple. The Newton-Raphson method results from approximating a function by its second order Taylor expansion

$$f(x - h) \approx F(x,h) = f(x) - h'g_x + \tfrac{1}{2}h'V_x h. \tag{43}$$

A necessary condition for an extreme value of F is

$$V_x h = g_x. \tag{44}$$

This system is not necessary consistent. If it is, a solution is given by $V_x^- g_x$. Otherwise we must go back to the method of section 2.

5: Psychometric Applications.

5.1: Nonmetric multiple linear regression.

This is the standard case, the analogon of ordinary multiple regression. The matrix X is measured in the ordinary way (at least on an interval scale), the 'dependent' variable y is measured on an ordinal scale.

5.2: Nonmetric polynomial regression analysis.

The matrix X is constructed by using as columns either powers of an independent variable x, or the orthogonal polynomial coefficients. Again y is measured on an ordinal scale. Observe that the polynomial of degree zero (a constant) is irrelevant and must not be included in the analysis. If we have two independent variables x_1 and x_2 the columns of X can be constructed by using the vectors x_1^2, x_2^2, x_1x_2, x_1, and x_2. In this case we fit a quadratic function to the (ordinal) data. This approach generalizes easily to all 'separable' functions of p variables.

5.3: Nonmetric discriminant analysis.

In the metric case discriminantanalysis can be considered as a degenerate form of multiple regression in which the 'dummy' dependent variable consists entirely of zeroes and ones. In the nonmetric case the same thing is true. In stead of requiring maximum distance between the two populations we require minimum overlap (and of course we use Kruskal's primary approach to ties in this case).

5.4: Nonmetric canonical discriminantanalysis.

We adapt our algorithm of section 2 to the case where there are p groups. We require that the projections Xw of the groups are pairwise disjoint. To do this we form $\binom{p}{2}$ vectors $\hat{z}_{st}$ $(s < t)$, we let $\hat{z} = \sum_s \sum_t \hat{z}_{st}$, and we minimize

$$F = \sum_{s<t}\sum \frac{w'X'Xw - 2\,w'X'\hat{z}_{st} + \hat{z}_{st}'\hat{z}_{st}}{w'X'Xw} . \qquad (45)$$

This is equivalent (with the $\hat{z}_{st}$ fixed) to maximizing

$$F = \frac{w'X'\hat{z}}{w'X'Xw} , \qquad (46)$$

and no new aspects have to be considered.

6.5: Additive conjoint measurement.

In additve conjoint measurement the matrix X must be constructed in a rather special way. Consider the case of a two-factor design, with facets A and B. A has l elements (levels, structs), B has m - l elements. For the $d_{ij} = (a_i, b_j)$ element of the data structure D = A x B we construct a row of X as follows: the row $x_{k.}$ has m elements, both x_{ki} and $x_{k,l+j}$ are equal to unity, all other elements of $x_{k.}$ are zero. It follows that

$$x_{k.}w = w_i + w_{l+j}, \qquad (47)$$

which is what we want. Generalization to more than two factors is obvious. Because of the special structure of X the equations for the iterative program can be simplified. We develop them again for the two-factor case with no missing data. Facet A has k elements, B has l elements, k + l = m, kl = n. The scale values are collected in $w' = [u' \vdots v']$. Moreover

$$z_{ij} = u_i + v_j, \qquad (48)$$

$\hat{z}_{ij}$ is the corresponding monotone matrix, $\sum_i \sum_j \hat{z}_{ij} = 0$. We must minimize

$$F = \frac{\sum\sum(\hat{z}_{ij} - z_{ij})^2}{\sum\sum z_{ij}^2} , \qquad (49)$$

which is equivalent to maximizing

$$F' = \frac{2\sum\sum z_{ij}\hat{z}_{ij}}{\sum\sum z_{ij}^2} = \frac{2\sum\sum \hat{z}_{ij}(u_i + v_j)}{\sum\sum (u_i + v_j)^2} . \qquad (50)$$

The derivatives of the Lagrangian function are (λ as Lagrange multiplier)

$$\frac{\partial \phi}{\partial u_s} = 2\sum_j \hat{z}_{sj} - 2\lambda u_s - 2\lambda \sum_j v_j, \qquad (51a)$$

$$\frac{\partial \phi}{\partial v_t} = 2 \sum_i \hat{z}_{it} - 2\lambda v_t - 2\lambda \sum_j u_s. \qquad (51b)$$

Let $\tilde{u}_s = \sum_j \hat{z}_{sj}$ and $\tilde{v}_t = \sum_i \hat{z}_{it}$, then $\sum_s \tilde{u}_s = \sum_t \tilde{v}_t = \sum_i \sum_j \hat{z}_{ij} = 0$. It follows that, if $\tilde{z}_{ij} = \tilde{u}_i + \tilde{v}_j$ and $\gamma = \left(\sum_i \sum_j \tilde{z}_{ij}^2\right)^{-\frac{1}{2}}$, then $u = \gamma \tilde{u}$, and $v = \gamma \tilde{v}$, are a solution to (51). The theory in this section can easily be translated into a matrix formulation using the g-inverse (we add some rows to X to make up for the deficiency in rank). Moreover it can equally easily be generalized to cases with partial orders, zero-one responses, non-numerical responses, and multifactor designs. Quite similar simplifications are possible in the equations of the Newton-Raphson method.

7: References.

Browne, M. — Oblique Procrustus rotation. Psychometrika 1967.

De Leeuw, J. — Canonical discriminant analysis of relational data. Department of data theory, Leyden University, RNoo6-68

Fan, K. — Systems of linear inequatlities. Annals of mathematics studies, no 38, 1956.

Kunzi, H.P. & Krelle, W. — Nichtlineare Programmierung. Berlin, Göttingen, Heidelberg, Springer-Verlag, 1962.

Scott, D. — Measurement structures and linear inequalities. Journal of mathematical psychology, 1964.

Tversky, A.N. — Finite additive structures. Michigan mathematical psuchology program, 1964.

Appendix:

Upon further investigation the method proposed in formula (24) turns out to be somewhat of a hoax. The proposal was

$$w^{(s+1)} = w^{(s)} - \frac{(T^*)^{(s)}}{2(1 - S^{(s)})} (X'X)^{-1} g^{(s)}. \tag{52}$$

From (14), (15), and (20)

$$g^{(s)} = \frac{2}{(T^*)^{(s)}} \left\{ (1 - S^{(s)}) X'X w^{(s)} - X'z^{(s)} \right\}. \tag{53}$$

Substitution of (53) into (52) yields

$$w^{(s+1)} = \frac{1}{(1 - S^{(s)})} (X'X)^{-1} X'z^{(s)}, \tag{54}$$

and thus the method of (24) is identical to the method of section (2). This shows that the faster convergence of the Newton-Raphson method is due entirely to the terms $g_l e_k + g_k e_l$ in the expression for v_{kl}. Nevertheless, if we are close to the minimum, these terms must necessarily be quite small.