An Algorithm for Nonlinear Principal Components Analysis with B-splines by Means of Alternating Least Squares

H. Coolen, J. van Rijckevorsel, Rotterdam, and J. de Leeuw, Leiden

KEYWORDS

alternating least squares, B-splines, principal components analysis, correspondence analysis

The algorithm described in this paper is suitable for nonmetric principal components analysis with B-splines, multiple correspondence analysis with B-splines as well as a mixture of the two. For the theory of these different forms of nonlinear principal components analysis and further literature see Gifi(1981), De Leeuw(1982) and the companion paper by Van Rijckevorsel(1982).

In nonlinear principal components analysis we have to minimize a least squares loss function σ:

$$\sigma = \Sigma_{j=1}^{m} \text{ tr } (X - G_j U_j)'(X - G_j U_j)$$

over X and U_j with normalization $X'X = I$. Here G_j is a (pseudo)-indicator matrix with k(order of the spline) nonzero entries per row. Also define $D_j = G_j'G_j$. For single variables we require in addition that $U_j = a_j t_j'$, where $a_j'D_j a_j = 1$ and a_j is restricted to lie in a given convex cone K_j. Minimizing the loss function σ can be done by alternating least squares using direct iteration. Each iteration cycle consists of five or ten substeps; in each substep the loss σ is minimized over one set of parameters for fixed values of the other set(s). Each iteration starts with X^0 and U_j^0 and gives updates X^+ and U_j^+ which conditionally minimize the loss σ(De Leeuw and Van Rijckevorsel (1980)). The main steps in the algorithm are:

(1) $U_j^0 = D_j^{-1} G_j' X^0$

(2) $\sigma_j = p - \text{tr } U_j^{0\prime} D_j U_j^0$

(3) $a_j^0 = P_j(U_j^0 t_j^0)$, where P_j projects on K_j

(4) $a_j^+ = a_j^0 (a_j^{0\prime} D_j a_j^0)^{-\frac{1}{2}}$

(5) $t_j^+ = a_j^+ U_j^0$

(6) $U_j^+ = a_j^+ t_j^{+\prime}$

(7) $\sigma_j = \sigma_j - t_j^{+\prime} t_j^+$

(8) $Z = \Sigma_{j=1}^{m} G_j U_j^+$

(9) Z in deviations from column means

(10) $X^+ = Z(Z'Z)^{-\frac{1}{2}}$.

The first step of the algorithm consists of computing the category quantifications for every variable on p dimensions simultaneously. The startconfiguration X^0 is a normalized random configuration. G_j is constructed rowwise by a subprogram that uses the recurrence relation for computing B-splines taking their small support into account(De Boor, 1978). D_j is banded symmetric positive definite having bandwidth 2k-1. A Cholesky factorization of D_j , by the method of Gauss elimination adapted to the symmetry and bandedness of D_j, gives us S_j which is lower triangular(De Boor, 1978). The system $D_jU_j^0 = S_jS_j'U_j^0 = G_j'X^0$ is solved first for $S_jU_j^0$ by means of forward substitution and then for U_j^0 by backward substitution. In steps (3), (4) and (5) we are looking for the minimizing updates a_j^+ with X^0 and t_j^0 fixed, and for t_j^+ with a_j^0 and X^0 fixed. Finding a_j^+ is a cone regression problem, while computing t_j^+ amounts to an ordinary least squares problem; both problems have a unique solution(De Leeuw and Van Rijckevorsel, 1980). Finding X^+ which minimizes the loss for fixed a_j^+ and t_j^+ is an orthogonal procrustes problem that is solved here, with essentially the same results, by Gram-Schmidt because it is less expensive. Steps (1) through (7) are executed subsequently per variable. For multiple variables step (8) computes $G_jU_j^0$. In the case of multiple correspondence analysis, all variables being multiple, only steps (1), (2), (8), (9) and (10) are alternatingly computed. When the difference in loss between two successive iterations no longer exceeds a predetermined criterion, execution of the algorithm stops after step (2) or step (7). After this the solution is rotated to its principal components and the corresponding eigenvalues are computed. All other quantifications are recomputed with the rotated solution. For some results on convergence see De Leeuw and Van Rijckevorsel(1980).

In order to investigate practical examples we have tested an APL-version of the algorithm. The results can be found in Van Rijckevorsel (1982). A FORTRAN-version is available from Vakgroep M&T, FSCW, Erasmus University Rotterdam, Postbus 1738, Rotterdam, Netherlands.

References

De Leeuw J. (1982), Nonlinear principal components analysis, Compstat 82, Physica, Wien.

Van Rijckevorsel J.L.A. (1982), Canonical analysis with B-splines, Compstat 82, Physica, Wien.

For all other references see Van Rijckevorsel(1982).