AUGMENTATION AND MAJORIZATION ALGORITHMS FOR SQUARED DISTANCE SCALING

JAN DE LEEUW

1. Introduction

The problem studies in this note is minimization of the loss function

$$\sigma(X)=\sum_{i=1}^n\sum_{j=1}^n w_{ij}(\delta_{ij}-d_{ij}^2(X))^2. \tag{1}$$

over X. Here X is an $n\times p$ *configuration*, the w_{ij} are known non-negative *weights*, the δ_{ij} are known *dissimilarities*, and $d_{ij}^2(X)$ is the *squared Euclidean distance* between rows i and j of X. Thus we fit squared distances to the dissimilarities.

We need some convenient matrix expressions for the squared distances. If we define $C=XX'$ then we can write

$$d_{ij}^2(X)=(e_i-e_j)'C(e_i-e_j)=\mathbf{tr}\,CA_{ij}, \tag{2}$$

with e_i and e_j unit vectors and with A_{ij} the matrix

$$A_{ij}=(e_i-e_j)(e_i-e_j)'. \tag{3}$$

Date: January 18, 2004.

2000 *Mathematics Subject Classification.* 62H25.

Key words and phrases. Multivariate Analysis, Correspondence Analysis.

Many different algorithms have been proposed to minimize the loss function (1). Foremost of these is perhaps the ALSCAL method [Takane et al., 1977], which is of the cyclic coordinate descent type. One ALSCAL iteration consists of a cycle over all np coordinates of X, minimizing loss over one coordinate at a time, while keeping the other coordinates fixed at their current values. Since the loss function is a multivariate quartic in X, the coordinate subproblems can be solved by finding the roots of a univariate cubic (and choosing the one corresponding to the minimum).

Even before ALSCAL, De Leeuw [1975] proposed an augmentation algorithm to minimize (1), in the case in which there are no weights. The paper was never published, but the algorithm has been discussed by Takane [1977] and Browne [1987]. They did not include the original derivation and a convergence proof. We give this missing derivation and the proof, for archival purposes. And we also add a (new) majorization algorithm to minimize loss function (1) for the case of unequal weights [1].

2. Augmentation algorithm

Suppose all weights are equal to one, and we want to minimize

$$\sigma(X)=\sum_{i=1}^n\sum_{j=1}^n(\delta_{ij}-d_{ij}^2(X))^2. \tag{4}$$

over X. In fact, we minimize over $X\in\mathcal{X}$, where $\mathcal{X}$ are the column-centered matrices.

[1] For the cyclic coordinate ascend, block relaxation, alternating least squares, augmentation, and majorization terminology we refer to the Appendix.

Consider the augmented loss function

$$\lambda(X,\Gamma)=\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{\ell=1}^{n}(\gamma_{ijk\ell}-d_{ijkl}(X))^2, \tag{5}$$

where $d_{ijkl}(X)=(e_i-e_j)'C(e_k-e_\ell)$. Now minimize the augmented loss function over $X\in\mathcal{X}$ and over Γ constrained by $\gamma_{ijij}=\delta_{ij}$. Thus the "diagonal" elements of Γ are constrained to be equal to the corresponding elements of Δ, and all other elements are free. Write these constraints as $\Gamma\in\mathcal{G}$.

Clearly

$$\min_{\Gamma\in\mathcal{G}}\lambda(X,\Gamma)=\sigma(X) \tag{6a}$$

and thus

$$\min_{X}\sigma(X)=\min_{X}\min_{\Gamma\in\mathcal{G}}\lambda(X,\Gamma). \tag{6b}$$

The augmentation algorithm is defined by using block relaxation on λ, that is we iteratively alternate minimization over X for fixed Γ and minimization over $\Gamma\in\mathcal{G}$ for fixed X. Or, in other words, we apply *alternating least squares* to λ.

Convergence of the algorithm, to a stationary point, follows from the general theory of block relaxation algorithms. We produce a decreasing sequence of loss function values using a continuous update mapping, and we can thus apply the theory in Zangwill [1969].

3. Simplifications

We now have a definition of the algorithm, but we still have to show that the two substeps of the alternating least squares iteration are relatively easy to implement. Otherwise there is very little reason to make the problem seemingly more complicated by augmenting it.

Clearly minimizing over $\Gamma \in \mathcal{G}$ is trivially easy, because we just set the diagonal elements γ_{ijij} to δ_{ij} and we set all other γ_{ijkl} equal to $d_{ijkl}(X)$. Thus

$$\gamma_{ijk\ell} = \delta^{ik}\delta^{j\ell}(\delta_{ij} - d_{ij}^2(X)) + (e_i - e_j)'C(e_k - e_\ell). \tag{7}$$

The other substep is more interesting. In order to minimize over $X \in \mathcal{X}$, for fixed Γ, observe we can write the augmented loss function in the form

$$\lambda(X,\Gamma) = c - 2\mathbf{tr}\ CV + 4n^2\mathbf{tr}\ C^2 = 4n^2\mathbf{tr}\ (C - \frac{1}{4n^2}V)^2 + c - \frac{n^2}{4}\mathbf{tr}\ V^2. \tag{8}$$

Here c is the constant

$$c = \sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n\sum_{\ell=1}^n \gamma_{ijkl}^2$$

and V is the matrix

$$V = \sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n\sum_{\ell=1}^n \gamma_{ijkl}(e_i - e_j)(e_k - e_\ell)',$$

and we have also used the fact that

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{\ell=1}^{n} d_{ijkl}^2(X) =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{\ell=1}^{n}(e_i - e_j)'C(e_k - e_\ell)(e_k - e_\ell)'C(e_i - e_j) = 4n^2\mathbf{tr}\, C^2.$$

It follows that minimization of λ over $X \in \mathcal{X}$ for fixed Γ can be accomplished by finding the best rank p approximation to the matrix $\frac{1}{4n^2}V$, using standard eigenvector-eigenvalue methods.

We can derive a more compact expression for V in terms of γ. From the definition

$$V = \sum_{i=1}^{n}\sum_{k=1}^{n}\gamma_{i\star k\star}e_i e_k' - \sum_{i=1}^{n}\sum_{\ell=1}^{n}\gamma_{i\star\star\ell}e_i e_\ell'$$

$$- \sum_{j=1}^{n}\sum_{k=1}^{n}\gamma_{\star jk\star}e_j e_k' + \sum_{j=1}^{n}\sum_{\ell=1}^{n}\gamma_{\star j\star\ell}e_j e_\ell' =$$

$$\sum_{i=1}^{n}\sum_{v=1}^{n}(\gamma_{i\star j\star} - \gamma_{i\star\star j} - \gamma_{\star ij\star} + \gamma_{\star i\star j})e_i e_j'$$

And we simplify it even more by substituting the expression for the currently optimal γ, which means we can rewrite the iterations without reference to γ at all, just in terms of X. Using (7) we see that

$$V = -2\sum_{i=1}^{n}\sum_{j=1}^{n}(\delta_{ij} - d_{ij}^2(X))A_{ij} + 4n^2C.$$

Thus the iteration is a rank p approximation to

$$\tilde{C} - \frac{1}{2n^2}H(\tilde{X}),$$

where

$$H(\tilde{X}) = \sum_{i=1}^{n}\sum_{j=1}^{n}(\delta_{ij} - d_{ij}^2(X))A_{ij}.$$

4. Majorization

The majorization algorithm is based on the fact that the loss function is quadratic in the elements of C. By using (or bounding) the largest eigenvalue of the quadratic component, we can again reduce the iteration to computing an optimal rank p approximation.

First

$$\sigma(C) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\delta_{ij}^2 - 2\textbf{tr}\ BC + \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(c_{ii} + c_{jj} - 2c_{ij})^2$$

where

$$B = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\delta_{ij}A_{ij}.$$

Next

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(c_{ii} + c_{jj} - 2c_{ij})^2 \leq \\ &\qquad \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(\tilde{c}_{ii} + \tilde{c}_{jj} - 2\tilde{c}_{ij})^2 + 2\textbf{tr}\ G(C - \tilde{C}) + 4\omega\sum_{i=1}^{n}\sum_{j=1}^{n}(c_{ij} - \tilde{c}_{ij})^2,\end{aligned}$$

where

$$G = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}d_{ij}^2(\tilde{C})A_{ij},$$

and where ω is the sum of the w_{ij}. Here we have used Cauchy-Schwartz in the form

$$[(c_{ii}-\tilde{c}_{ii})+(c_{jj}-\tilde{c}_{jj})-2(c_{ij}-\tilde{c}_{ij})]^2\leq 4\sum_{i=1}^{n}\sum_{j=1}^{n}(c_{ij}-\tilde{c}_{ij})^2$$

Combining what we have so far gives

$$\sigma(C)\leq\sigma(\tilde{C})-2\mathbf{tr}H(\tilde{X})+4\omega\mathbf{tr}\,(C-\tilde{C})^2,$$

which shows that we update X by finding the optimal rank p approximation of

$$\tilde{C}-\frac{1}{4\omega}H(\tilde{X}).$$

Appendix A. Types of algorithms

Suppose f is a function on $X\times Y$. A *block relaxation* algorithm for minimizing f starts with some $x_0\in X$. In each iteration k we find $y^{(k)}=\mathop{\mathbf{argmin}}_{y\in Y}f(x^{(k)},y)$ and then $x^{(k+1)}=\mathop{\mathbf{argmin}}_{x\in X}f(x,y^{(k)})$. Thus we alternate updating x and y. If the function we are minimizing is a least squares loss function, then block relaxation becomes *alternating least squares*. Although we have defined block relaxation for two blocks, it is clear how to generalize it to more than two. If there are more than two blocks it becomes interesting how we cycle through the blocks. If each of the blocks only consists of a single variable, then block relaxation is *cyclic coordinate descend*. Block relation is worthwhile if the subproblems are simple, compared to the original problem. Block relation methods in statistics are discussed in Oberhofer and Kmenta [1974]; Jensen et al. [1991]; De Leeuw [1994] and alternating least squares became popular in the ALSOS system summerized by Young [1981].

One important special case of block relaxation is *augmentation*. The problem is to minimize a function g over X, but we assume we can find an *augmentation function* f on $X \times Y$ such that $g(x) = \min_{y \in Y} f(x, y)$. Augmentation algorithms apply block relaxation to the augmentation function. They should be considered if we can find an augmentation function which is simpler to minimize than our original function g. The most familiar examples of augmentation algorithms are in factor analysis, where we augment the reduced correlation matrix by including the diagonal elements, and in unbalanced factorial analysis of variance, where we augment by adding enough elements to each cell to get a balanced design.

Majorization is a special case of augmentation. Again the problem is to minimize $g(x)$ on X. Suppose we can find a *majorization function* f on $X \times X$ such that $g(x) \leq f(x, y)$ for all $x, y \in X$ and $g(x) = f(x, x)$ for all $x \in X$. Then f is an augmentation of g, with the special property that $g(x) = f(x, x) = \min_{y \in X} f(x, y)$ for all x. Again, a *majorization algorithm* applies block relation to the majorization function. Majorization methods are discussed in detail by De Leeuw [1994]; Heiser [1995]; Lange et al. [2000] and for quadratic majorization functions by Böhning and Lindsay [1988].

References

D. Böhning and B.G. Lindsay. Monotonicity of quadratic-approximation algorithms. *Annals of the Institute of Statistical Mathematics*, 40(4):641–663, 1988.

M.W. Browne. The Young-Householder Algorithm and the Least Squares Multdimensional Scaling of Squared Distances. *Journal of Classification*, 4:175–190, 1987.

J. De Leeuw. An Alternating Least Squares Approach to Squared Distance Scaling. 1975.

J. De Leeuw. Block Relaxation Methods in Statistics. In H.H. Bock, W. Lenski, and M.M. Richter, editors, *Information Systems and Data Analysis*, Berlin, 1994. Springer Verlag.

W.J. Heiser. Convergent Computing by Iterative Majorization: Theory and Applications in Multidimensional Data Analysis. In W.J. Krzanowski, editor, *Recent Advanmtages in Descriptive Multivariate Analysis*. Oxford: Clarendon Press, 1995.

S. T. Jensen, S. Johansen, and S. L. Lauritzen. Globally convergent algorithms for maximizing a likelihood function. *Biometrika*, 78:867–877, 1991.

K. Lange, D.R. Hunter, and I. Yang. Optimization Transfer Using Surrogate Objective Functions. *Journal of Computational and Graphical Statistics*, 9:1–20, 2000.

W. Oberhofer and J. Kmenta. A general procedure for obtaining maximum likelihood estimates in generalized regression models. *Econometrica*, 42:579–590, 1974.

Y. Takane. On the Relations among Four Methods of Multidimensional Scaling. *Behaviormetrika*, 4:29–42, 1977.

Y. Takane, F.W. Young, and J. De Leeuw. Nonmetric Individual Differences in Multidimensional Scaling: An Alternating Least Squares Method with Optimal Scaling Features. *Psychometrika*, 42:7–67, 1977.

F. W. Young. Quantitative analysis of qualitative data. *Psychometrika*, 46:357–388, 1981.

W. I. Zangwill. *Nonlinear Programming: a Unified Approach*. Prentice-Hall, Englewood-Cliffs, N.J., 1969.

DEPARTMENT OF STATISTICS, UNIVERSITY OF CALIFORNIA, LOS ANGELES, CA 90095-1554

E-mail address, Jan de Leeuw: deleeuw@stat.ucla.edu

URL, Jan de Leeuw: http://gifi.stat.ucla.edu