1977

A normalized cone regression approach to alternating least squares algorithms

Jan de Leeuw
Department of data theory
University of Leiden
Leiden
The Netherlands

Summary:

This paper has two distinct parts. In the first part we prove some of the more important duality and characterization theorems for possibly nonconvex, possibly normalized cone regression problems. Its purpose is to show that this can be done by using only elementary tools, and that some of the results at least can be derived without using convexity. In the second part of the paper we use the results from the first part to prove a number of convergence theorems for alternating least squares algorithms. Again we are careful not to assume convexity when it is not strictly needed.

Acknowledgement:

Comments by J. Lemaire (Université de Nice) have been very helpful.

0: Introduction

One of the essential ingredients of the nonmetric data analysis methods proposed by Kruskal (1964) is the monotone regression algorithm MFIT. The problem solved by MFIT can be described as follows. Suppose X is an n-dimensional real linear space, with inner product $<x,y>$ and with norm $||x|| = <x,x>^{\frac{1}{2}}$. Suppose C is the closed polyhedral convex cone of all vectors y satisfying the monotonicity restrictions $y_1 \leqslant y_2 \leqslant \ldots \leqslant y_n$. The problem is to find the vector $\hat{x}$ that minimizes $||x - y||$ over all y in C. For a thorough discussion of this problem, the methods for solving it, statistical applications, we refer to the books of Barlow, Bartholomew, Bremner, and Brunk (1972), Van Eeden (1958), and to the review article of Barlow and Brunk (1972). Several generalizations are also discussed in these references. In fact most of the theory and most of the algorithms remain valid if we generalize the monotonicity conditions by replacing the weak order by an arbitrary partial order, and if we replace the inner product norm by an arbitrary separable convex norm.

Although these generalizations are certainly useful for nonmetric scaling, they sometimes do not go far enough. For various problems in scaling the constraint set in which y varies is not convex. Moreover we sometimes want to minimize the normalized distance function $||x - y|| / ||y||$, in stead of the usual unnormalized ones. In this paper we give some theoretical results on the problem of minimizing possibly normalized distance functions over possibly nonconvex cones. Because we do not use coordinates the results are valid in all Hilbert spaces..

1: Unnormalized problem

We first consider the problem P_1 of minimizing the distance between a point x and the points of a possibly nonconvex closed cone C.

T1:a: If $\hat{x}$ solves P_1 and y is such that $\hat{x} + \varepsilon y$ is in C for all $0 \leq \varepsilon < \varepsilon_o^+$, then $\langle x,y\rangle \leq \langle\hat{x},y\rangle$.

b: If $\hat{x}$ solves P_1 and y is such that $\hat{x} + \varepsilon y$ is in C for all $\varepsilon_o^- < \varepsilon \leq 0$, then $\langle x,y\rangle \geq \langle\hat{x},y\rangle$.

c: If $\hat{x}$ solves P_1 and y is such that $\hat{x} + \varepsilon y$ is in C for all $\varepsilon_o^- < \varepsilon < \varepsilon_o^+$, then $\langle x,y\rangle = \langle\hat{x},y\rangle$.

Proof: If $\hat{x}$ solves P_1 and $\hat{x} + \varepsilon y$ is in C, then clearly $||x - \hat{x}|| \leq ||x - (\hat{x} + \varepsilon y)||$. If we square both sides and collect terms we find that $\varepsilon^2\langle y,y\rangle - 2\varepsilon\langle y,x - \hat{x}\rangle \geq 0$ for all ε such that $\hat{x} + \varepsilon y$ in C. By letting ε approach zero through positive values we find T1:a, by letting ε approach zero through negative values we find T1:b, and T1:c follows from the combination of the two. Q.E.D.

T2: If $\hat{x}$ solves P_1, then $\langle x,\hat{x}\rangle = \langle\hat{x},\hat{x}\rangle$.

Proof: Take $y = \hat{x}$ in T1, then $\hat{x} + \varepsilon y = (1 + \varepsilon)\hat{x}$, which is in C for all $\varepsilon \geq -1$. Thus T2:c applies. Q.E.D.

Observe that we have not proved that a solution to P_1 actually exists. This will follow from the analysis of a closely related problem Q_1. Suppose S is the unit sphere in X, i.e. the set of all y such that $||y|| = 1$. Problem Q_1 is the maximization of $\langle x,y\rangle$ over y in $C \cap S$.

T3: If $\hat{x}$ solves P_1 then $\hat{x} / ||\hat{x}||$ solves Q_1. If $\tilde{x}$ solves Q_1 then $\langle\tilde{x},x\rangle\tilde{x}$ solves P_1.

Proof: Define v(y) as the minimum of $||x - \alpha y||$ over $\alpha \geq 0$. Minimizing $||x - y||$ over C is equivalent to minimizing v(y) over $C \cap S$, and then adjust the length. But for y in $C \cap S$ we have

$$v^2(y) = \begin{cases} \langle x,x\rangle & \text{if } \langle x,y\rangle \leq 0, \\ \langle x,x\rangle - \langle x,y\rangle^2 & \text{if } \langle x,y\rangle \geq 0. \end{cases}$$

This minimum value is attained for

$$\alpha = \begin{cases} 0 & \text{if } \langle x,y\rangle \leq 0, \\ \langle x,y\rangle & \text{if } \langle x,y\rangle \geq 0. \end{cases}$$

The result follows from this. Q.E.D.

Because Q_1 is the maximization of a linear function over a compact set, it follows that the maximum in Q_1, and consequently the minimum in P_1 is attained.

2: Unnormalized convex problem

We now specify, in this section only, that the cone C is convex, but not necessarily polyhedral. The results here are classical. Of course the theorems from the previous section remain valid in this more special case. The regression problem for a convex cone C will be called problem P_2.

T4: $\hat{x}$ solves P_2 if and only if $\langle x,\hat{x}\rangle = \langle \hat{x},\hat{x}\rangle$, and $\langle x,y\rangle \leq \langle \hat{x},y\rangle$ for all y in C.

Proof: The necessity of both conditions follows from T1:a and T2. Sufficiency follows from the expansion

$$||x - y||^2 = ||(x - \hat{x}) + (\hat{x} - y)||^2 =$$
$$= ||x - \hat{x}||^2 + ||\hat{x} - y||^2 + 2\langle x - \hat{x},\hat{x} - y\rangle.$$

The last two terms are both nonnegative. This is obvious for $||\hat{x} - y||^2$. For the second term we observe

$$\langle x - \hat{x},\hat{x} - y\rangle = -\langle x - \hat{x},y\rangle = \langle y,\hat{x}\rangle - \langle y,x\rangle \geq 0.$$

Thus $||x - y||^2 \geq ||x - \hat{x}||^2$ for all y in C. Q.E.D.

T5: The solution of P_2 is unique.

Proof: Suppose both $\hat{x}_1$ and $\hat{x}_2$ satisfy the conditions of T4. Then

$$\langle \hat{x}_1,\hat{x}_1\rangle = \langle x,\hat{x}_1\rangle \leq \langle \hat{x}_1,\hat{x}_2\rangle,$$
$$\langle \hat{x}_2,\hat{x}_2\rangle = \langle x,\hat{x}_2\rangle \leq \langle \hat{x}_1,\hat{x}_2\rangle.$$

The Cauchy-Schwartz inequality now implies that $\hat{x}_1$ and $\hat{x}_2$ are proportional. Thus $\hat{x}_2 = \beta\hat{x}_1$. From the first condition of T4

$\langle x, x_2 \rangle = \beta \langle x, x_1 \rangle = \langle \hat{x}_2, \hat{x}_2 \rangle = \beta^2 \langle \hat{x}_1, \hat{x}_1 \rangle$,

and thus β is either one or zero. If β is zero, then $\langle y, x \rangle \leq 0$ for all y in C, and $\hat{x}_1 = \hat{x}_2 = 0$ is the unique solution of P_2. If β is one, then also $\hat{x}_1 = \hat{x}_2$. Q.E.D.

We also define a problem Q_2. Suppose C^o is the polar of C, i.e. the set of all z such that $\langle z, y \rangle \leq 0$ for all y in C. Problem Q_2 is the maximization of $\langle x, z \rangle$ over z in $C^o \cap S$.

<u>T6</u>: If $\hat{x}$ solves P_2 then $\hat{z} = (x - \hat{x}) / ||x - \hat{x}||$ solves Q_2.

<u>Proof</u>: T4 says that $x - \hat{x}$ is in $C^{\bullet}$. Thus $\hat{z}$ is feasible for Q_2, and the optimum value π of Q_2 satisfies $\pi \geq \langle x, \hat{z} \rangle = ||x - \hat{x}||$. On the other hand if z is in C^o then $\langle z, \hat{x} \rangle \leq 0$, and thus $\langle z, x - \hat{x} \rangle \geq \langle z, x \rangle$. If z is in S then $||x - \hat{x}|| \geq \langle z, x - \hat{x} \rangle$. Thus $||x - \hat{x}|| \geq \langle z, x \rangle$ for all z in $C^o \cap S$, and $||x - \hat{x}|| \geq \pi$. It follows that $||x - \hat{x}||$ is the optimal value of Q_2, and that $\hat{z}$ solves Q_2. Q.E.D.

We now define problem P_3 as the minimization of $||x - z||$ over z in C^o.

<u>T7</u>: If $\hat{x}$ solves P_2 then $x - \hat{x}$ solves P_3. If $\hat{z}$ solves P_3 then $x - \hat{z}$ solves P_2.

<u>Proof</u>: The first part follows from T6 and the second part of T3. The second part follows by symmetry. Q.E.D.

<u>T8</u>: If C is a closed convex cone and C^o is its polar, then any vector x can be decomposed as $x = x_1 + x_2$, with x_1 in C, x_2 in C^o, and $\langle x_1, x_2 \rangle = 0$. This decomposition is unique.

<u>Proof</u>: Take $x_1 = \hat{x}$, the solution of P_2, and $x_2 = \hat{z}$, the solution of P_3. We only have to prove that this solution for the decomposition is the only one. Suppose $x = u_1 + u_2$ is another such decomposition. Then $\langle x_1, x \rangle = \langle x_1, x_1 \rangle$ and $\langle r_1, x \rangle = \langle r_1, r_1 \rangle$. On the other hand $\langle x_1, x \rangle = \langle x_1, r_1 + r_2 \rangle \leq \langle x_1, r_1 \rangle$ and $\langle r_1, x \rangle = \langle r_1, x_1 + x_2 \rangle \leq \langle r_1, x_1 \rangle$.

As in T5 the Cauchy-Schwartz inequality now implies that x_1 is proportional to r_1. In the same way x_2 is proportional to r_2. As in T5 we can use the relations $<x_1,x> = <x_1,x_1>$ and $<r_1,x> = <r_1,r_1>$ to prove that actually $x_1 = r_1$. In the same way x_2 is r_2. Q.E.D.

The theorems given for the case of a convex cone are very well known. We have presented them for completeness, but also because usually much more sophisticated proofs are given, based on convex analysis in normed linear spaces. Some of the more important references are Deutsch and Maserick (1967), Ioffe and Tikhomirov (1968), and Ubhaya (1974). An extremely elegant finite dimensional theory is developed in Rockafellar's book (1970, especially 31.4 and 31.5). Theorem T5 is due to Moreau (1962), a beautiful discussion of many closely related results is Moreau (1965).

3: Normalized problem

We now discuss the normalized problem P_4 of minimizing the normalized distance $||x - y|| \ / \ ||y||$ over y in a possibly nonconvex cone C. For convenience we also define the problem P_1', in which we have to minimize $||x - y|| \ / \ ||x||$ over y in C. Of course P_1' is trivially equivalent to P_1. We also need problem Q_1, defined in section 1.

T9: If $\tilde{x}$ solves Q_1, then $\hat{x} = \frac{<x,x>}{<x,\tilde{x}>} \tilde{x}$ solves P_4. If $\hat{x}$ solves P_4 then $\tilde{x} = \hat{x} \ / \ ||\hat{x}||$ solves Q_1.

Proof: We define w(y) as the minimum of $\frac{||x - \alpha y||}{\alpha ||y||}$ over $\alpha \geq 0$. Again the minimum of the normalized distance function over C equals the minimum of w(y) over $C \cap S$, and the minima are attained for the same direction. For y in $C \cap S$ we have

$$w^2(y) = \begin{cases} 1 & \text{if } <x,y> \leq 0, \\ \frac{<x,x> - <x,y>^2}{<x,x>} & \text{if } <x,y> \geq 0, \end{cases}$$

and the minimum is attained for

$$\tilde{\alpha} = \begin{cases} +\infty & \text{if } \langle x,y\rangle \leq 0, \\ \dfrac{\langle x,x\rangle}{\langle x,y\rangle} & \text{if } \langle x,y\rangle > 0. \end{cases}$$

Again this proves the result. Q.E.D.

We can combine the results of T3 and T9 into a single result. The solution of the problems P_1 (or P_1'), Q_1, and P_4 are proportional. The constants of proportionality are given in our theorems. The optimal values of P_1' and P_4 are the same. Normalized cone regression problems can be solved by solving unnormalized problems, and by renormalizing the solution afterwards. The link between the normalized and the unnormalized problem is the problem Q_1, or, equivalently, separating the problem of finding the optimal direction from that of finding the optimal length. Because we have not used convexity in this section, there is no need for a separate section on normalized convex problems. The basic result of this section has been proved previously by Kruskal and Carroll (1969). Their arguments were essentially the same, but with more emphasis on the geometry of the problem.

4: Convergence of ALS algorithms

Recently algorithms based on so-called alternating least squares (ALS) methodology have become quite popular in data analysis (De Leeuw, Young, Takane, 1976; Young, De Leeuw, Takane, 1976; Takane, Young, De Leeuw, 1977). In ALS problems we have to minimize normalized loss functions of the form $||x - y|| \,/\, ||y||$ or $||x - y|| \,/\, ||x||$ over x in a possibly nonconvex cone C_1 and over y in a possibly nonconvex cone C_2. The ALS algorithm starts with x_0 in C_1, then defines y_0 as a minimizer of $||x_0 - y||$ over y in C_2, then defines x_1 as a minimizer of $||x - y_0||$ over x in C_1, and so on. This defines sequences

$x_0, x_1, \ldots$ and $y_0, y_1, \ldots$ with x_n in C_1 and y_n in C_2 for all n. There is no explicit proof of convergence in de De Leeuw, Young, Takane papers, nor is it made exactly clear in which sense the ALS methods converge, if they converge at all. There are some results in section 3 of De Leeuw, Young, Takane (1976), especially page 485-489, but the results are sketchy and consist mainly of references to the mathematical programming literature. An elegant convergence theorem for alternating least squares approaches to the additive model (ADDALS) has been given by Lemaire (1976), but his result is not quite general enough for our purposes, and he does not give a detailed proof. We intend to provide some of the relevant results in this section, using results from our previous sections, and taking care to specify where convexity of the cones C_1 and C_2 is needed, and where not.

It is convenient to introduce some special notation and terminology first. We will use a bar under a symbol that denotes a vector to indicate normalization. Thus $\underline{x} = x / ||x||$. We write $\delta(x,y)$ for $||x - y||$ in the sequel, and we also define $\gamma(x,y) = \langle \underline{x}, \underline{y} \rangle$. The ALS algorithm defines sequences $\{x_n\}$ in C_1 and $\{y_n\}$ in C_2, but also $\{\underline{x}_n\}$ in $C_1 \cap S$ and $\{\underline{y}_n\}$ in $C_2 \cap S$. Moreover there are the sequences $\{\delta_n\}$ with $\delta_n = \delta(x_n, y_n)$, and $\{\gamma_n\}$ with $\gamma_n = \gamma(x_n, y_n)$. It follows from the results in section 3 of this paper that we can minimize the normalized loss function by maximizing $\gamma(x,y)$ over x in C_1 and y in C_2, which is of course equivalent to maximizing $\langle x,y \rangle$ over x in $C_1 \cap S$ and y in $C_2 \cap S$. In most cases looking for the global maximum of $\gamma(x,y)$ is not very realistic. To get a workable algorithm we have to broaden the class of desirable points (or targets). A pair $(\hat{x},\hat{y})$ is <u>strongly desirable</u> if $\hat{x}$ maximizes $\langle x,\hat{y} \rangle$ over x in $C_1 \cap S$ and $\hat{y}$ maximizes $\langle \hat{x},y \rangle$ over y in $C_2 \cap S$. We shall also call a pair $(\hat{x},\hat{y})$ weakly desirable if $\hat{x}$ minimizes $\delta(x,\hat{y})$ over x in C_1 and $\hat{y}$ minimizes $\delta(\hat{x},y)$ over x in C_2. It follows from T3 that if $(\hat{x},\hat{y})$ is weakly desirable, if $\hat{x} \neq 0$ and $\hat{y} \neq 0$, then $(\hat{\underline{x}},\hat{\underline{y}})$ is strongly desirable. We shall call (x_{n+1},y_{n+1}) the successor of (x_n,y_n), and we shall call $(\underline{x}_{n+1},\underline{y}_{n+1})$ the successor of $(\underline{x}_n,\underline{y}_n)$. Our first theorem deals with convergence of some of the sequences of real numbers generated by the algorithm. It does not use convexity.

<u>T10</u>: There exist $\varepsilon_\infty \geq 0$ and $-1 \leq \gamma_\infty \leq +1$ such that $||x_n|| \downarrow \varepsilon_\infty$, $||y_n|| \downarrow \varepsilon_\infty$, $\gamma_n \uparrow \gamma_\infty$, and $\delta_n \rightarrow 0$. Moreover if $\varepsilon_\infty > 0$, then $\gamma_\infty = 1$.

<u>Proof</u>: Convergence of γ_n follows from the fact that we increase $\gamma(x,y)$ two times in each ALS iteration, and from the fact that $\gamma(x,y)$ is bounded. (We assume the algorithm generates an infinite sequence of different intermediate solutions, if it should ever repeate itself we can stop, the point is then both weakly and strongly desirable). From T2 it follows that $<x_n,y_n> = ||y_n||^2$, which implies, by Cauchy-Schwartz, $||y_n|| \leq ||x_n||$. In the same way $<x_{n+1},y_n> = ||x_{n+1}||^2$, which implies $||x_{n+1}|| \leq ||y_n||$, and thus $||x_n||$ and $||y_n||$ converge to the same limit. Now the identity $\delta_n^2 = ||x_n||^2 + ||y_n||^2 - 2<x_n,y_n> = ||x_n||^2 - ||y_n||^2$ implies that $\delta_n \rightarrow 0$. And the identity $\gamma_n = ||y_n|| / ||x_n||$ implies that $\gamma_n \uparrow 1$ if $\varepsilon_\infty > 0$. Q.E.D.

The next two theorems are proved by the standard closedness-compactness methods (Zangwill, 1969). They do not use convexity.

<u>T11</u>: If (x_∞,y_∞) is an accumulation point of $\{(x_n,y_n)\}$, then $x_\infty = y_\infty$ is in $C_1 \cap C_2$.

<u>Proof</u>: Suppose the subsequence $\{(x_n,y_n)\}_{n\varepsilon N}$ converges to (x_∞,y_∞). Then $\{x_n\}_{n\varepsilon N}$ converges to x_∞, and $\{y_n\}_{n\varepsilon N}$ converges to y_∞, with x_∞ in C_1 and y_∞ in C_2. For each n in N it is true that $\delta(x_n,y_n) \leq$

$\delta(x_n,y)$ for all y in C_2, and thus by continuity $\delta(x_\infty,y_\infty) \leq \delta(x_\infty,y)$ for all y in C_2. This means that y_∞ minimizes $\delta(x_\infty,y)$ over all y in C_2, and thus $\langle x_\infty,y\rangle = ||y_\infty||^2$ by T2. From Cauchy-Schwartz $||y_\infty|| \leq ||x_\infty||$, with equality if and only if y_∞ is proportional to x_∞. But $||y_\infty|| = ||x_\infty|| = \varepsilon_\infty$ by continuity of the norm, and thus $y_\infty = x_\infty$. Q.E.D.

The usual case in data analysis is that $C_1 \cap C_2 = \{0\}$, i.e. there is no non-trivial perfect solution. In this case T11 implies that both $x_n \to 0$ and $y_n \to 0$. This is a satisfactory convergence result, but it is very disappointing from a practical point of view. Thus we now turn to the sequences $\{\underline{x}_n\}$ and $\{\underline{y}_n\}$.

<u>T12</u>: If $(\underline{x}_\infty,\underline{y}_\infty)$ is an accumulation point of $\{(\underline{x}_n,\underline{y}_n)\}$, then $(\underline{x}_\infty,\underline{y}_\infty)$ is strongly desirable.

<u>Proof</u>: As in the proof of T11 we start with a subsequence $\{(\underline{x}_n,\underline{y}_n)\}_{n\varepsilon N}$ converging to $(\underline{x}_\infty,\underline{y}_\infty)$. We then form $\{(\underline{x}_{n+1},\underline{y}_{n+1})\}_{n\varepsilon N}$, and select from this sequence a subsequence $\{(\underline{x}_{n+1},\underline{y}_{n+1})\}_{n\varepsilon M}$, converging, say, to $(\underline{x}_{\infty+1},\underline{y}_{\infty+1})$. Of course the corresponding sequence $\{(\underline{x}_n,\underline{y}_n)\}_{n\varepsilon M}$ still converges to $(\underline{x}_\infty,\underline{y}_\infty)$. Now $\gamma(\underline{x}_{n+1},\underline{y}_{n+1}) \geq \gamma(\underline{x}_{n+1},y)$ for all y in $C_2 \cap S$, which implies $\gamma(\underline{x}_{\infty+1},\underline{y}_{\infty+1}) \geq \gamma(\underline{x}_{\infty+1},y)$ for all y in $C_2 \cap S$. In the same way $\gamma(\underline{x}_{n+1},\underline{y}_n) \geq \gamma(x,\underline{y}_n)$ for all x in $C_1 \cap S$, which implies $\gamma(\underline{x}_{\infty+1},\underline{y}_\infty) \geq \gamma(x,\underline{y}_\infty)$ for all x in $C_1 \cap S$. Thus $(\underline{x}_{\infty+1},\underline{y}_{\infty+1})$ is the successor of $(\underline{x}_\infty,\underline{y}_\infty)$. But by continuity $\gamma(\underline{x}_\infty,\underline{y}_\infty) = \gamma(\underline{x}_{\infty+1},\underline{y}_{\infty+1})$, and thus $(\underline{x}_\infty,\underline{y}_\infty)$ is strongly desirable. Q.E.D.

It is difficult to prove stronger and more specific convergence theorems without making special assumptions. For the following theorem we do not assume convexity right away, although the hypotheses of the theorem are implied by convexity of C_1 and C_2.

<u>T13</u>: Suppose that for every accumulation point $(\underline{x}_\infty,\underline{y}_\infty)$ of the sequence $\{(\underline{x}_n,\underline{y}_n)\}$ generated by the algorithm it is true that $\underline{x}_\infty$ is the

unique maximizer of $\gamma(x,\underline{y}_\infty)$ over x in $C_1 \cap S$, and $\underline{y}_\infty$ is the unique maximizer of $\gamma(\underline{x}_\infty,y)$ over y in $C_2 \cap S$. Then $||\underline{x}_n - \underline{x}_{n+1}|| \to 0$ and $||\underline{y}_n - \underline{y}_{n+1}|| \to 0$.

Proof: As in the proof of the previous theorem we find subsequences $\{(\underline{x}_n,\underline{y}_n)\}_{n\varepsilon M}$ converging to $(\underline{x}_\infty,\underline{y}_\infty)$ and $\{(\underline{x}_{n+1},\underline{y}_{n+1})\}_{n\varepsilon M}$ converging to $(\underline{x}_{\infty+1},\underline{y}_{\infty+1})$. From the proof of the previous theorem $(\underline{x}_{\infty+1},\underline{y}_{\infty+1})$ is the successor of $(\underline{x}_\infty,\underline{y}_\infty)$, and $(\underline{x}_\infty,\underline{y}_\infty)$ is strongly desirable. Thus from the uniqueness assumption $(\underline{x}_\infty,\underline{y}_\infty) = (\underline{x}_{\infty+1},\underline{y}_{\infty+1})$. Q.E.D.

Result T13 does not imply that $\{\underline{x}_n\}$ or $\{\underline{y}_n\}$ converges. Again we need additional assumptions to prove convergence. By a familiar result of Ostrowski (1966, chapter 28) the convergence result in T13 implies that either $\{(\underline{x}_n,\underline{y}_n)\}$ converges, or the limit set of the sequence is a continuum. Thus we only have to assume that the number of strongly desirable points with a given function value is at most countable, or that there is at least one isolated accumulation point, to obtain convergence. Only in pathological cases something can go wrong. The next result shows that at least some forms of pathological behaviour are excluded by convexity.

Define $\underline{\underline{C}} = C_1 \cap C_2$. A double bar under a vector denotes projection on $\underline{\underline{C}}$, i.e. $\underline{\underline{y}}$ minimizes $\delta(x,y)$ over x in $\underline{\underline{C}}$. Moreover $\delta(y,\underline{\underline{C}}) = \delta(y,\underline{\underline{y}})$.

T14: If both C_1 and C_2 are convex then both $\{x_n\}$ and $\{y_n\}$ converge to a point of $\underline{\underline{C}}$.

Proof: We adapt a proof given by Gubin, Polyak, and Raik (1967) to our case. Take z in $\underline{\underline{C}}$ arbitrary. By T4 we have $<z,x_n - y_n> \leqslant 0$ and $<z,y_n - x_{n+1}> \leqslant 0$. This implies $<z,x_n> \leqslant <z,y_n> \leqslant <z,x_{n+1}>$. But $||x_n|| \geqslant ||y_n|| \geqslant ||x_{n+1}||$ by T10, and thus $\delta(z,x_n) \geqslant \delta(z,y_n) \geqslant \delta(z,x_{n+1})$. It follows that $\delta(z,x_n)$ and $\delta(z,y_n)$ decrease to a common limit. This is part A of the proof.

By the result of part A, and by the definition of $\underline{\underline{x}}_n$ and $\underline{\underline{y}}_n$,

$$\delta(x_n,\underline{x}_n) \geq \delta(y_n,\underline{x}_n) \geq \delta(y_n,\underline{y}_n) \geq \delta(x_{n+1},\underline{y}_n) \geq \delta(x_{n+1},\underline{x}_{n+1})$$

which implies that also $\delta(x_n,\underline{C})$ and $\delta(y_n,\underline{C})$ decrease to a common limit. Now suppose (x_∞,y_∞) is an accumulation point of $\{(x_n,y_n)\}$. In the same way as in the previous theorems there is a subsequence $\{x_n\}_{n\varepsilon N}$ converging to x_∞, such that $\{\underline{x}_n\}_{n\varepsilon N}$ converges to $\underline{x}_\infty$. But x_∞ is in $\underline{C}$ by T11, and thus $\underline{x}_\infty = x_\infty$. Since $\delta^2(x_n,\underline{C}) = ||x_n||^2 - ||\underline{x}_n||^2$ this implies that $\delta(x_n,\underline{C})$ converges to zero. In the same way $\delta(y_n,\underline{C})$ converges to zero. This is part B of the proof.

Now consider the spheres S_n, with center at $\underline{x}_n$ and with radius $\delta(x_n,\underline{C})$. If $m \geq n$ then, by part A, $\delta(x_m,\underline{x}_n) \leq \delta(x_n,\underline{x}_n) = \delta(x_n,\underline{C})$. Thus x_m is in S_n, and because the radius of S_n converges to zero by part B, $\{x_n\}$ is a Cauchy sequence, and converges to some x_∞. Again by part B $x_\infty = \underline{x}_\infty$, and thus x_∞ is in $\underline{C}$. Q.E.D.

The theorems proved in this section do not give a complete picture of the convergence behaviour of ALS algorithms. However, some results cannot be improved a great deal, as simple examples show. If C_1 and C_2 have a nontrivial intersection, then $\{x_n\}$ and $\{y_n\}$ may still converge to zero, while $\{\underline{x}_n\}$ and $\{\underline{y}_n\}$ need not converge to a point in the intersection at all. This is even true in the convex case. If C_1 and C_2 are not convex, then $\{\underline{x}_n\}$ and $\{\underline{y}_n\}$ need not converge at all. Moreover there generally exists more than one strongly desirable pair, even if C_1 and C_2 are convex (thus there are 'local minima'). It is possible, however, that better results can be derived by imposing conditions stronger than convexity on the cones (for example: they must be polyhedral, they must be strongly convex, they must not have parallel faces, etc.) It seems highly unlikely, however, that results can be proved which are as precise and as general as those for pairs of (separable) closed convex sets, with an unnormalized distance function (cf for example Gubin, Polyak, Raik, 1967).

References:

Barlow, R.E., Bartholomew, D.J., Bremner, J.M., Brunk, H.D.: Statistical Inference under order restrictions. New York, Wiley, 1972.

Barlow, R.E., Brunk, H.D.: The isotone regression problem and its dual. J. Amer. Statist. Soc., 67, 1972, 140-147.

Cheney, W., Goldstein, A.: Proximity maps for convex sets. Proc. Amer. Math. Soc., 10, 1959, 448-450.

De Leeuw, J., Young, F.W., Takane, Y.: Additive structure in qualitative data: an alternating least squares method with optimal scaling features. Psychometrika, 41, 1976, 471-504.

Deutsch, F.R., Maserick, P.H.: Applications of the Hahn-Banach theorem to approximation theory. Siam Rev.,4, 1967, 516-530.

Gubin, L.G., Polyak, B.T., Raik, E.V.: The method of projections for finding the common point of convex sets. U.S.S.R. Comp. Math. Phys., 7, 1967, no 6, 1-24.

Ioffe, A.D., Tikhomirov, V.M.: Duality of convex functions and extremum problems. Russian Math. Surveys, 23, 1968, 53-124.

Kruskal, J.B.: Nonmetric multidimensional scaling: a numerical method. Psychometrika, 29, 1964, 28-42.

Kruskal, J.B., Carroll, J.D.: Geometric models and badness-of-fit functions. In P.R. Krishnaiah (ed): Multivariate Analysis II. New York, Academic Press, 1969.

Lemaire, J. Agrégation typologique des préferences. Thèse de 3e cycle, Université de Nice, 1976.

Moreau, J.J. Fonctions convexes duales et points proximaux dans un espace Hilbertien. CRAS, Paris, 255, 1962, 2897-2899.

Moreau, J.J.: Proximité et dualité dans un espace hilbertien. Bull. Soc. Math. France, 93, 1965, 273-299.

Ostrowski, A.M.: Solutions of equations and systems of equations. New York, Academic Press, 1966 (second edition).

Rockafellar, R.T.: Convex analysis. Princeton, Princeton University Press, 1970.

Takane, Y., Young, F.W., De Leeuw, J.: Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. Psychometrika, 42, 1977, in press.

Ubhaya, V.A.: Minimization of seminorms on a topological vector space: duality and characterization theorems. J. Math. Anal. Appl., 46, 1974, 635-657.

Van Eeden , C: Testing and estimating ordered parameters of probability distributions. Amsterdam, Poortpers, 1958.

Young, F.W., De Leeuw, J., Takane, Y.: Regression with qualitative and quantitative variables: an alternating least squares method with optimal scaling features. Psychometrika, 41, 1976, 505-529.

Zangwill, W.I.: Nonlinear programming: a unified approach. Englewood Cliffs, Prentice Hall, 1969.