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 \( \langle x, y \rangle \) and with norm \( ||x|| = \langle x, x \rangle^{\frac{1}{2}} \). Suppose \( C \) is the closed polyhedral convex cone of all vectors \( y \) satisfying the monotonicity restrictions \( y_1 \leq y_2 \leq \ldots \leq y_n \). The problem is to find the vector \( x \) 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 \( x \) solves \( P_1 \) and \( y \) is such that \( \tilde{x} + \varepsilon y \) is in \( C \) for all 
\[ 0 < \varepsilon < \varepsilon^+_o, \] then \[ \langle x, y \rangle < \langle \tilde{x}, y \rangle. \]

b: If \( \tilde{x} \) solves \( P_1 \) and \( y \) is such that \( \tilde{x} + \varepsilon y \) is in \( C \) for all 
\[ \varepsilon^-_o < \varepsilon < 0, \] then \[ \langle x, y \rangle > \langle \tilde{x}, y \rangle. \]

c: If \( \tilde{x} \) solves \( P_1 \) and \( y \) is such that \( \tilde{x} + \varepsilon y \) is in \( C \) for all 
\[ \varepsilon^-_o < \varepsilon < \varepsilon^+_o, \] then \[ \langle x, y \rangle = \langle \tilde{x}, y \rangle. \]

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

T2: If \( \tilde{x} \) solves \( P_1 \), then \[ \langle x, \tilde{x} \rangle = \langle \tilde{x}, \tilde{x} \rangle. \]

Proof: Take \( y = \tilde{x} \) in T1, then \( \tilde{x} + \varepsilon y = (1 + \varepsilon)\tilde{x} \), 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 \( \tilde{x} \) solves \( P_1 \) then \( \tilde{x} / ||\tilde{x}|| \) solves \( Q_1 \). If \( \tilde{x} \) solves \( Q_1 \) then 
\( \tilde{x}, x \tilde{x} \) 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 
\[ v^2(y) = \begin{cases} 
\langle x, x \rangle & \text{if } \langle x, y \rangle \leq 0, \\
\langle x, x \rangle - \langle x, y \rangle^2 & \text{if } \langle x, y \rangle > 0.
\end{cases} \]
This minimum value is attained for
\[
\alpha = \begin{cases} 
0 & \text{if } <x,y> \leq 0, \\
<x,y> & \text{if } <x,y> \geq 0.
\end{cases}
\]

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 \( <x,\hat{x}> = <\hat{x},\hat{x}> \), and \( <x,y> \leq <\hat{x},y> \) for all \( y \) in \( C \).

**Proof:** The necessity of both conditions follows from T1:a and T2.

Sufficiency follows from the expansion
\[
||x - y||^2 = ||(x - \hat{x}) + (\hat{x} - y)||^2 = ||x - \hat{x}||^2 + ||\hat{x} - y||^2 + 2<x - \hat{x},\hat{x} - y>.
\]

The last two terms are both nonnegative. This is obvious for \( ||\hat{x} - y||^2 \). For the second term we observe
\[
<x - \hat{x},\hat{x} - y> = -<x - \hat{x},y> = <y,\hat{x}> - <y,x> \geq 0.
\]

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
\[
<x,\hat{x}_1> = <x,\hat{x}> < \hat{x}_1,\hat{x}_2>,
\]
\[
<x,\hat{x}_2> = <x,\hat{x}_2> < \hat{x}_1,\hat{x}_2>.
\]

The Cauchy-Schwarz 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
\[
<x, x_2> = \beta <x, x_1> = <x_2, x_2> = \beta <x_1, x_1>,
\]
and thus \(\beta\) is either one or zero. If \(\beta\) is zero, then \(<y, x> \leq 0\) for all \(y\) in \(C\), and \(z_1 = z_2 = 0\) is the unique solution of \(P_2\). If \(\beta\) is one, then also \(z_1 = z_2\). Q.E.D.

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

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

**Proof:** T4 says that \(x - z\) is in \(C^0\). Thus \(\hat{z}\) is feasible for \(Q_2\), and the optimum value \(\pi\) of \(Q_2\) satisfies \(\pi \geq <x, \hat{z}> = ||x - z||\).

On the other hand if \(z\) is in \(C^0\) then \(<z, z> \leq 0\), and thus
\[
<x, z - z> = <z, x>.
\]
Thus \(||x - z|| > <z, x>\) for all \(z\) in \(C^0 \cap S\), and \(||x - z|| \geq \pi\).

It follows that \(||x - \hat{z}||\) 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 \(C^0\).

**T7:** If \(z\) solves \(P_2\) then \(x - z\) solves \(P_3\). If \(\hat{z}\) solves \(P_3\) then
\[
x - \hat{z}\text{ solves } P_2.
\]

**Proof:** The first part follows from T6 and the second part of T3.

The second part follows by symmetry. Q.E.D.

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

**Proof:** Take \(x_1 = z\), 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 \(<x_1, x> = <x_1, x_1>\) and \(<r_1, x> = <r_1, r_1>\). On the other hand
\[
<x_1, x> = <x_1, r_1 + r_2> \leq <x_1, r_1> \text{ and } <r_1, x> = <r_1, x_1 + x_2> \leq <r_1, x_1>.
\]
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 \( x_2 = r_2 \). 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 Ukhaya (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 \( \hat{x} \) solves \( Q_1 \), then \( \hat{x} = \frac{<x,x>}{<x,x>} \hat{x} \) solves \( P_4 \). If \( \bar{x} \) solves \( P_4 \) then \( \bar{x} = \bar{x} / ||\bar{x}|| \) solves \( Q_1 \).

**Proof:** We define \( w(y) \) as the minimum of \( \frac{||x - \alpha y||}{||y||} \) 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

\[
    w^2(y) = \begin{cases} 
    1 & \text{if } <x,y> \leq 0, \\
    \frac{<x,x> - <x,y>^2}{<x,x>} & \text{if } <x,y> > 0,
    \end{cases}
\]
and the minimum is attained for

\[ \mathcal{S} = \begin{cases} +\infty & \text{if } \langle x, y \rangle \leq 0, \\ \frac{\langle x, x \rangle}{\langle x, y \rangle} & \text{if } \langle x, y \rangle > 0. \end{cases} \]

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 $\bar{x} = x / ||x||$. We write $\delta(x, y)$ for $||x - y||$ in the sequel, and we also define $\gamma(x, y) = \langle x, y \rangle$. The ALS algorithm defines sequences $\{x_n\}$ in $C_1$ and $\{y_n\}$ in $C_2$, but also $\{x_n\}$ in $C_1 \cap S$ and $\{y_n\}$ 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 $(\bar{x}, \bar{y})$ is strongly desirable if $\bar{x}$ maximizes $\langle x, \bar{y} \rangle$ over $x$ in $C_1 \cap S$ and $\bar{y}$ maximizes $\langle \bar{x}, y \rangle$ over $y$ in $C_2 \cap S$. We shall
also call a pair \((\bar{x}, \bar{y})\) weakly desirable if \(\bar{x}\) minimizes \(\delta(x, y)\) over \(x\) in \(C_1\) and \(\bar{y}\) minimizes \(\delta(x, y)\) over \(x\) in \(C_2\). It follows from T3 that if \((\bar{x}, \bar{y})\) is weakly desirable, if \(\bar{x} \neq 0\) and \(\bar{y} \neq 0\), then \((\bar{x}, \bar{y})\) is strongly desirable. We shall call \((x_{n+1}, y_{n+1})\) the successor of \((x_n, y_n)\), and we shall call \((x_{n+1}, y_{n+1})\) the successor of \((x_n, y_n)\). Our first theorem deals with convergence of some of the sequences of real numbers generated by the algorithm. It does not use convexity.

**T10:** There exist \(\varepsilon_\infty \geq 0\) and \(-1 \leq \gamma_\infty \leq +1\) such that 
\[ ||x_n|| \downarrow \varepsilon_\infty, \]
\[ ||y_n|| \downarrow \varepsilon_\infty, \gamma_n \downarrow \gamma_\infty, \text{ and } \delta_n \rightarrow 0. \]
Moreover if \(\varepsilon_\infty > 0\), then \(\gamma_\infty = 1\).

**Proof:** Convergence of \(\gamma_n\) 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 repeat 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.

**T11:** If \((x_\infty, y_\infty)\) is an accumulation point of \(\{(x_n, y_n)\}\), then 
\[ x_\infty = y_\infty \text{ is in } C_1 \cap C_2. \]

**Proof:** Suppose the subsequence \(\{(x_n, y_n)\}_{n \in \mathbb{N}}\) converges to \((x_\infty, y_\infty)\). Then \(\{x_n\}_{n \in \mathbb{N}}\) converges to \(x_\infty\), and \(\{y_n\}_{n \in \mathbb{N}}\) converges to \(y_\infty\), with \(x_\infty\) in \(C_1\) and \(y_\infty\) in \(C_2\). For each \(n\) in \(\mathbb{N}\) it is true that \(\delta(x_n, y_n) \leq \)
δ(x_n, y) for all y in C_2, and thus by continuity δ(x_∞, y) ≤ δ(x_n, y) for all y in C_2. This means that y_∞ minimizes δ(x_∞, y) over all y in C_2, and thus <x_∞, y> = ||y_∞||^2 by T2. From Cauchy-Schwartz ||y_∞|| ≤ ||x_∞||, with equality if and only if y_∞ is proportional to x_∞. But ||y_∞|| = ||x_∞|| = ε_∞ by continuity of the norm, and thus y_∞ = x_∞. Q.E.D.

The usual case in data analysis is that C_1 ∩ C_2 = {0}, i.e. there is no non-trivial perfect solution. In this case T1 implies that both x_n → 0 and y_n → 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 \{x_n\} and \{y_n\}.

T12: If (x_∞, y_∞) is an accumulation point of \{(x_n, y_n)\}, then (x_∞, y_∞) is strongly desirable.

Proof: As in the proof of T1 we start with a subsequence \{(x_n, y_n)\}_{n \in \mathbb{N}} converging to (x_∞, y_∞). We then form \{(x_{n+1}, y_{n+1})\}_{n \in \mathbb{N}} and select from this sequence a subsequence \{(x_{n+1}, y_{n+1})\}_{n \in \mathbb{N}} converging, say, to (x_∞, y_∞). Of course the corresponding sequence \{(x_n, y_n)\}_{n \in \mathbb{N}} still converges to (x_∞, y_∞). Now γ(x_{n+1}, y_{n+1}) ≥ γ(x_n, y_n) for all y in C_2 ∩ S, which implies γ(x_{n+1}, y_{n+1}) ≥ γ(x_n, y_n) for all y in C_2 ∩ S. In the same way γ(x_{n+1}, y_{n}) ≥ γ(x_n, y_n) for all x in C_1 ∩ S. Thus (x_{n+1}, y_{n+1}) is the successor of (x_n, y_n). But by continuity γ(x_∞, y_∞) = γ(x_∞, y_∞), and thus (x_∞, y_∞) 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.

T13: Suppose that for every accumulation point (x_∞, y_∞) of the sequence \{(x_n, y_n)\} generated by the algorithm it is true that x_∞ is the
unique maximizer of $\gamma(x, y_\infty)$ over $x$ in $C_1 \cap S$, and $y_\infty$ is the unique maximizer of $\gamma(x_\infty, y)$ over $y$ in $C_2 \cap S$. Then

$$ ||x_n - x_{n+1}|| \to 0 \text{ and } ||y_n - y_{n+1}|| \to 0. $$

**Proof:** As in the proof of the previous theorem we find subsequences

$$ ((x_n, y_n))_n \in \mathbb{N} \text{ converging to } (x_\infty, y_\infty) \text{ and } ((x_{n+1}, y_{n+1}))_n \in \mathbb{N} \text{ converging to } (x_{\infty+1}, y_{\infty+1}). $$

From the proof of the previous theorem $(x_{\infty+1}, y_{\infty+1})$ is the successor of $(x_\infty, y_\infty)$, and $(x_\infty, y_\infty)$ is strongly desirable. Thus from the uniqueness assumption $(x_\infty, y_\infty) = (x_{\infty+1}, y_{\infty+1})$. Q.E.D.

Result T13 does not imply that $\{x_n\}$ or $\{y_n\}$ 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 $\{(x_n, y_n)\}$ 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 $C = C_1 \cap C_2$. A double bar under a vector denotes projection on $C$, i.e. $\bar{y}$ minimizes $\delta(x, y)$ over $x$ in $C$. Moreover $\delta(y, \bar{C}) = \delta(y, C)$.

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

**Proof:** We adapt a proof given by Gubin, Polyak, and Raik (1967) to our case. Take $z$ in $C$ arbitrary. By T4 we have $\langle z, x_n - y_n \rangle \leq 0$ and $\langle z, y_n - x_{n+1} \rangle \leq 0$. This implies $\langle z, x_n \rangle \leq \langle z, y_n \rangle \leq \langle z, x_{n+1} \rangle$. But

$$ ||x_n|| \geq ||y_n|| \geq ||x_{n+1}|| $$

by T10, and thus $\delta(z, x_n) \geq \delta(z, y_n) \geq \delta(z, x_{n+1})$. 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 $x_n$ and $y_n$,
\[ \delta(x_n, x) \geq \delta(y_n, x_n) \geq \delta(y_n, x) \geq \delta(x_{n+1}, x_n) \geq \delta(x_{n+1}, x_{n+1}) \]
which implies that also \( \delta(x_n, y_n) \) and \( \delta(y_n, y_n) \) decrease to a common limit. Now suppose \((x_\infty, y_\infty)\) is an accumulation point of \(\{(x_n, y_n)\}\).

In the same way as in the previous theorems there is a subsequence \(\{x_n\}_{n \in \mathbb{N}}\) converging to \(x_\infty\), such that \(\{x_n\}_{n \in \mathbb{N}}\) converges to \(x_\infty\). But \(x_\infty\) is in \(\mathcal{C}\) by T11, and thus \(x_\infty = x_\infty\). Since 

\[ \delta^2(x_n, y_n) = ||x_n||^2 - ||x_n||^2 \]

this implies that \(\delta(x_n, y_n)\) converges to zero. In the same way \(\delta(y_n, y_n)\) converges to zero. This is part B of the proof.

Now consider the spheres \(S_n\), with center at \(x_n\) and with radius \(\delta(x_n, y_n)\). If \(m \geq n\) then, by part A, \(\delta(x_m, x_n) \leq \delta(x_n, x_n) = \delta(x_n, y_n)\).

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 \(x_\infty\).

Again by part B \(x_\infty = x_\infty\), and thus \(x_\infty\) is in \(\mathcal{C}\). 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 \(\{x_n\}\) and \(\{y_n\}\) 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 \(\{x_n\}\) and \(\{y_n\}\) 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:


Moreau, J.J.: Proximité et dualité dans un espace hilbertien.


Rockafellar, R.T.: Convex analysis.

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.

Udbhaya, V.A.: Minimization of seminorms on a topological vector
space: duality and characterization theorems.

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.