CONVERGENCE RATE OF ALTERNATING LEAST SQUARES ALGORITHMS

JAN DE LEEUW

1. Introduction

If x and y are points in $\mathbb{R}^n$ we write $\sigma(x,y)$ for the square of the Euclidean distances between x and y. Thus

$$\sigma(x,y) = \|x-y\|^2 = (x-y)'(x-y) \tag{1}$$

If $\mathcal{X}$ and $\mathcal{Y}$ are non-empty closed sets in $\mathbb{R}^n$ then the square of the Euclidean distance between them is $\Sigma(\mathcal{X},\mathcal{Y})$. Thus

$$\Sigma(\mathcal{X},\mathcal{Y}) = \inf_{x\in\mathcal{X}}\inf_{y\in\mathcal{Y}}\sigma(x,y). \tag{2}$$

In this paper we study the problem of finding an iterative algorithm to compute $\Sigma(\mathcal{X},\mathcal{Y})$. In addition, we want to find $\hat{x}\in\mathcal{X}$ and $\hat{y}\in\mathcal{Y}$ such that $\sigma(\hat{x},\hat{y}) = \Sigma(\mathcal{X},\mathcal{Y})$. In the terminology of Pai [1974], we want to find the *proximal point pairs*. Bauschke et al. [2004] call $\hat{x}$ and $\hat{y}$ *approximation pairs*. Although $0 \leq \Sigma(\mathcal{X},\mathcal{Y}) < +\infty$ is always well-defined, proximal pairs do not necessarily exist, even in the case in which $\mathcal{X}$ and $\mathcal{Y}$ are convex. For an elegant recent treatment of the convex case, see Dax [2006]. Of course if the intersection of $\mathcal{X}$ and $\mathcal{Y}$ is nonempty, then $\Sigma(\mathcal{X},\mathcal{Y}) = 0$ and (z,z) is a proximal point pair for any $z\in\mathcal{X}\cap\mathcal{Y}$.

Date: January 6, 2007.

2000 *Mathematics Subject Classification.* 62H25.

Key words and phrases. Multivariate Analysis, Correspondence Analysis.

The *alternating least squares* or *alternating projection* for solving this problem is to start with some $x^{(0)} \in \mathcal{X}$ and then proceed with

$$y^{(k)} \in P_{\mathcal{Y}}(x^{(k)}), \tag{3a}$$

$$x^{(k+1)} \in P_{\mathcal{X}}(y^{(k)}). \tag{3b}$$

Here $P_{\mathcal{X}}$ is the *least squares metric projection* on $\mathcal{X}$, i.e.

$$P_{\mathcal{X}}(y) = \{\hat{x} \in \mathcal{X} \mid \|\hat{x} - y\|^2 = \inf_{x \in \mathcal{X}} \|x - y\|^2\}.$$

In the same way $P_{\mathcal{Y}}$ is the least squares metric projection on $\mathcal{Y}$. Observe that our definition of the algorithm presupposes that both $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ are non-empty. We make that assumption from now on.

We can also write

$$x^{(k+1)} \in P_{\mathcal{X}}(P_{\mathcal{Y}}(x^{(k)})), \tag{4a}$$

$$y^{(k+1)} \in P_{\mathcal{Y}}(P_{\mathcal{X}}(y^{(k)})). \tag{4b}$$

Thus we are looking for fixed points of the two functional compositions $F = P_{\mathcal{X}} \circ P_{\mathcal{Y}}$ and $G = P_{\mathcal{Y}} \circ P_{\mathcal{X}}$. The two fixed point problems are basically the same. If x is a fixed point of F then $P_{\mathcal{Y}}(x)$ is a fixed point of G, and if y is a fixed point of G then $P_{\mathcal{X}}(y)$ is a fixed point of F.

There are a staggering number of publications on alternating projection methods. It is unclear who first used this particular name, but the technique which now has this name was applied first by Von Neumann [1933], in the case that $\mathcal{X}$ and $\mathcal{Y}$ are closed subspaces of Hilbert space. Cheney and Goldstein [1959] investigated alternating projection if $\mathcal{X}$ and $\mathcal{Y}$ are convex sets in Hilbert space. After this, many variations in many different spaces, projecting on many different types of sets have been described. A good recent overview is Deutsch [2001].

The term "alternating least squares" was apparently first used by De Leeuw [1968a,b, 1969]. These papers started a whole series which applied alternating least squares method to a large number of different optimal scaling methods. For reviews, see Young [1981] or Michailidis and De Leeuw [1998] and the book by Gifi [1990].

There is an important special case we discuss separately. Suppose that $\mathcal{X}$ and $\mathcal{Y}$ are subsets of the direct product of m copies of $\mathbb{R}^n$, which we write as $\mathbb{R}^{n\times m}$. In addition, we suppose that

- for $j = 1, \cdots, m$ there are $\mathcal{X}_j \subseteq \mathbb{R}^n$ such that $\mathcal{X} = \bigotimes_{j=1}^m \mathcal{X}_j$,
- $\mathcal{Y}$ is the subspace of $\mathbb{R}^{n\times m}$ of all elements $(y, y, \cdots, y)$.

Then our problem becomes to minimize

$$\sigma(x; y) = \sum_{j=1}^m \|x_j - y\|^2, \tag{5}$$

and

$$\Sigma(\mathcal{X}, \mathcal{Y}) = \inf_{y\in\mathbb{R}^n} \sum_{j=1}^m \inf_{x_j\in\mathcal{X}_j} \|x_j - y\|^2. \tag{6}$$

Minimizing (5) leads to an algorithm that starts with $y^{(0)}$. Then

$$x_j^{(k)} \in P_{\mathcal{X}_j}(y^{(k)}) \text{ for } j = 1, \cdots, m, \tag{7a}$$

$$y^{(k+1)} = \frac{1}{m}\sum_{j=1}^m x_j^{(k)}. \tag{7b}$$

This variation of the alternating projection method, for m closed subspaces of Hilbert space, was first presented by Halperin [1962].

For the optimal scaling applications we are interested in convexity is not particulary important. We also need to cover the case in which either $\mathcal{X}$ or $\mathcal{Y}$ or both are manifolds with smooth boundaries. In order to proceed we shall simply assume that $\mathcal{X}$ and $\mathcal{Y}$ have Lipschitz continuous metric projections, or at least that that there exist Lipschitz continuous selections from the possibly set-valued metric projections.

2. Zangwill and Ostrowski

3. Examples

3.1. **Separating Hyperplanes.** Suppose we want to find a solution, or approximate solution, to the system of homogeneous linear inequalities $Xb \geq 0$. Here X is an $n \times r$ matrix, which we can assume to be of full column rank r.

Define $\mathcal{Y}$ as the subspace of dimension r spanned by the columns of X. The projector $P_{\mathcal{Y}}$ is a symmetric and idempotent matrix, with r eigenvalues equal to 1 and $n - r$ eigenvalues equal to zero.

Define $\mathcal{X} = \mathbb{R}^n_+ \cap \mathbb{S}^n$, the intersection of the non-negative orthant and the unit sphere. Suppose P_+ is the projection on $\mathbb{R}^n_+$, then

$$P_{\mathcal{X}}(y) = \frac{P_+(y)}{\|P_+(y)\|}$$

if $P_+(y)$ is non-zero. If $P_+(y)$ is zero, i.e. if $y \leq 0$, then $P_{\mathcal{X}}(y)$ is a set of unit vectors, each with their single non-zero element in one of the positions i for which $y_i = \max_{k=1}^n y_k$.

Now suppose $\hat{x}$ is a fixed point with $\hat{x} = F\hat{x}$. Clearly $\|\hat{x}\| = 1$ and $\hat{x} \geq 0$. Define $\hat{y} = P_{\mathcal{Y}}\hat{x}$. If $P_+\hat{y}$ is non-zero and $\hat{y}$ does not have any zero elements, then $P_{\mathcal{X}}P_{\mathcal{Y}}$ is Fréchet differentiable at $\hat{x}$ and

$$\partial_{\mathrm{F}}F(\hat{x}) = Q(\hat{x})\Pi(\hat{x})P_{\mathcal{Y}}$$

where $Q(\hat{x})$ is the orthogonal projector $I - \hat{x}\hat{x}'$, and where $\Pi(\hat{x})$ is a diagonal orthogonal projector with diagonal elements 1 where $P_{\mathcal{Y}}\hat{x}$ is positive and 0 where $P_{\mathcal{Y}}\hat{x}$ is negative.

Since $\partial_{\mathrm{F}}F(\hat{x})$ is the product of three orthogonal projectors its largest eigenvalue $\kappa(\hat{x}) = \|\partial_{\mathrm{F}}F(\hat{x})\|$ is less than or equal to one. Moreover $\kappa(\hat{x}) = 1$ if and only if there is a non-zero $\hat{\beta}$ such that $\hat{x} = A\hat{\beta} \geq 0$, i.e. if and only if $\hat{x}$ defines a non-trivial solution of the original system of inequalities. In that case we have $P_{\mathcal{Y}}\hat{x} = \Pi(\hat{x})\hat{x} = Q(\hat{x})\hat{x} = \hat{x}$.

Now suppose we look at fixed points $\hat{x}$ for which we still have $P_+\hat{y}$ non-zero, but now some elements of $\hat{y}$ may be zero. Then $\Pi(\hat{x})$ is no longer differentiable at $\hat{x}$. Its generalized Jacobian in the sense of Qi is the set of diagonal matrices which have diagonal elements one where $\hat{y}$ is positive, diagonal elements zero where $\hat{y}$ is negative, and either one or zero where $\hat{y}$ is zero. Thus $\partial_{\mathrm{Q}}F(\hat{x})$ has 2^p elements, where p is the number of zeroes in $\hat{y}$. The generalized Jacobian in the sense of Clarke is the convex hull of these 2^p matrices, i.e.

$$\partial_{\mathrm{C}}F(\hat{x}) = Q(\hat{x})\partial_{\mathrm{C}}\Pi(\hat{x})P_y,$$

where $\Pi(\hat{x})$ has elements between zero and one in the places where $\hat{y}$ has zeroes.

For each $\Pi \in \partial_{\mathrm{Q}}F(\hat{x})$ we still have $\|Q(\hat{x})\Pi P_y\| \leq 1$

3.2. **Regression with Optimal Scaling.** This example is easily extended to minimizing $\|Ab - z\|^2$ over b and $z \in \mathcal{K} \cap \mathcal{S}$, where $\mathcal{K}$ is a polyhedral convex cone.

3.3. **Nonlinear Principal Component Analysis.**

$$\sigma(X, Z, B) = \mathbf{tr}\ (X - ZB')(X - ZB) = \sum_{j=1}^{m} \|x_j - Zb_j\|^2$$

$$\min_{x_j \in \mathcal{K}_j \cap \mathcal{S}^n} \min_Z \min_B \sigma(X, Z, B)$$

References

H.H. Bauschke, P.L. Combettes, and D.R. Luke. Finding Best Approximation Pairs Relative to Two Closed Convex Sets in Hilbert Spaces. *Journal of Approximation Theory*, 127:178–192, 2004.

W. Cheney and A.A. Goldstein. Proximity Maps for Convex Sets. *Proceedings of the American Mathematical Society*, 10:448–450, 1959.

A. Dax. The Distance Between Two Convex Sets. *Linear Algebra and Its Applications*, 416:184–213, 2006.

J. De Leeuw. Nonmetric Discriminant Analysis. Research Note 06-68, Department of Data Theory, University of Leiden, 1968a.

J. De Leeuw. Nonmetric Multidimensional Scaling. Research Note 010-68, Department of Data Theory FSW/RUL, 1968b.

J. De Leeuw. The Linear Nonmetric Model. Research Note 003-69, Department of Data Theory FSW/RUL, 1969.

F. Deutsch. *Best Approximation in Inner Product Spaces*. Number 7 in CMS Books in Mathematics. Springer, New York, N.Y., 2001.

A. Gifi. *Nonlinear Multivariate Analysis*. Wiley, Chichester, England, 1990.

I. Halperin. The Product of Projection Operators. *Acta Scientiarum Mathematicarum (Szeged)*, 23:96–99, 1962.

G. Michailidis and J. De Leeuw. The Gifi system for Descriptive Multivariate Analysis. *Statistical Science*, 13:307–336, 1998.

D.V. Pai. Proximal Points of Convex Sets in Normed Linear Spaces. *Yokohama Mathematics Journal*, 22:53–78, 1974.

J. Von Neumann. *Functional Operators. II. The Geometry of Orthogonal Spaces*. Number 22 in Annals of Mathematics Studies. Princeton University Press, Princeton, N.J., 1950 reprint edition, 1933.

F. W. Young. Quantitative Analysis of Qualitative Data. *Psychometrika*, 46:357–388, 1981.

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`