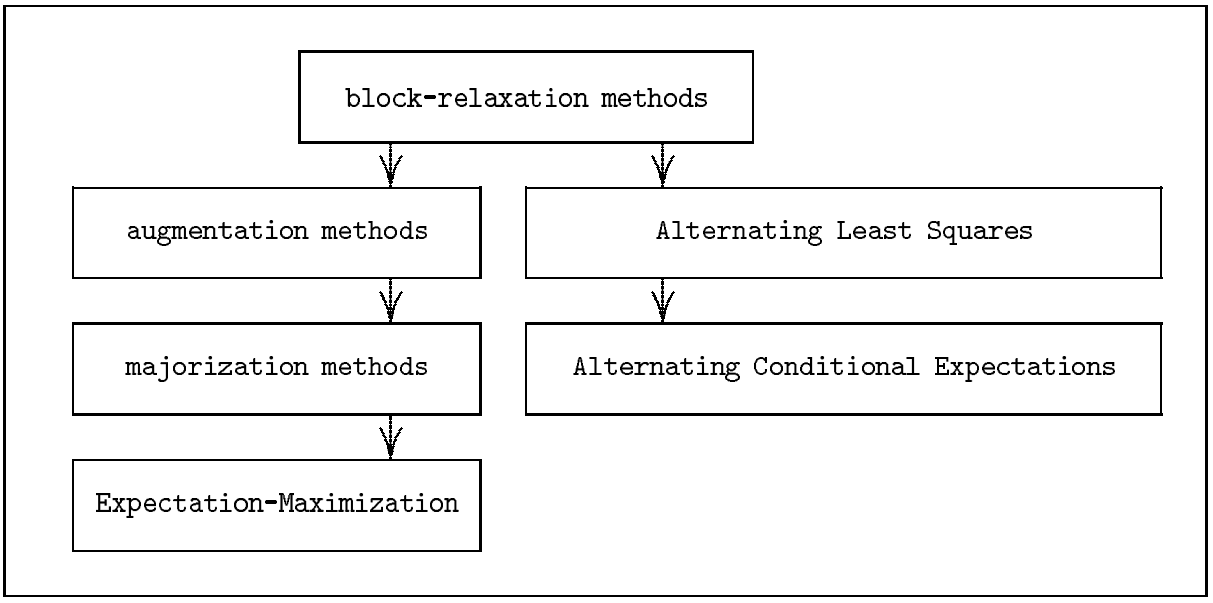
Block-relaxation Algorithms in Statistics

Jan de Leeuw
UCLA Statistics
Mathematical Sciences Building
405 Hilgard Avenue, LA, CA 90024-1555

1. Introduction

Many algorithms in recent computational statistics are variations on a common theme. In this paper we discuss four such classes of algorithms. Or, more precisely, we discuss a single class of algorithms, and we show how some well-known classes of statistical algorithms fit in this common class. The subclasses are, in logical order,



We discuss the general principles and results underlying these methods.

All the methods are special cases of what we shall call *block-relaxation methods,* although other names have also been used. There are many areas in applied mathematics where these methods have been discussed. Mostly, of course, in optimization and mathematical programming, but also in control and numerical analysis, and in differential equations. Bellman's theory of quasi-linearization [4] is closely related to what we call *augmentation* and *majorization.* We cannot give an extensive review of the literature in this paper, but a much more complete list of references is given in [12].

There is not much statistics in this paper. It is almost exclusively about deterministic optimization problems (although we shall optimize a likelihood function or two). Some of our results have been derived in the more restricted context of maximizing a likelihood

function by Jensen, Johansen, and Lauritzen [21]. They develop their own results, not relying on the existing results in the optimization literature. More or less the same applies to much of the literature on convergence of the EM algorithm, starting with Dempster, Laird, and Rubin [14]. Because we want to cover a much more general class of algorithms, we need more general results than this.

One thing we shall not discuss, at least not in this version of the paper, is stochastic extensions. But of course the integrals in the majorization algorithms can be approximated by Monte Carlo, functions can be optimized by simulated annealing, and the expected value of the posterior distribution approximates the maximum likelihood estimate (and can obviously be written as an integral). Incorporating this material into this paper would take us too far astray.

2. Block Relaxation.

Let us thus consider the following general situation. We minimize a real-valued function ψ defined on the product-set $\Omega = \Omega_1 \otimes \Omega_2 \otimes \cdots \otimes \Omega_p$, with $\Omega_s \subseteq \mathcal{R}^{n_s}$. In order to minimize ψ over Ω we use the following iterative algorithm.

[Starter]	Start with $\omega^{(0)} \in \Omega$.
[Step k.1]	$\omega_1^{(k+1)} \in \underset{\omega_1 \in \Omega_1}{\operatorname{argmin}} \psi(\omega_1, \omega_2^{(k)}, \cdots, \omega_p^{(k)})$.
[Step k.2]	$\omega_2^{(k+1)} \in \underset{\omega_2 \in \Omega_2}{\operatorname{argmin}} \psi(\omega_1^{(k+1)}, \omega_2, \omega_3^{(k)}, \cdots, \omega_p^{(k)})$.
$\cdots$	$\cdots$
[Step k.p]	$\omega_p^{(k+1)} \in \underset{\omega_p \in \Omega_p}{\operatorname{argmin}} \psi(\omega_1^{(k+1)}, \cdots, \omega_{p-1}^{(k+1)}, \omega_p)$.
[Motor]	$k \leftarrow k+1$ and go to $k.1$

We assume that the minima in the substeps exist (although they need not be unique, i.e. the argmin's can be point-to-set maps). We set $\omega^{(k)} \triangleq (\omega_1^{(k)}, \cdots, \omega_p^{(k)})$, and $\psi^{(k)} \triangleq \psi(\omega^{(k)})$. Also $\Omega_0 \triangleq \{\omega \in \Omega \mid \psi(\omega) \leq \psi^{(0)}\}$. For this method we have our first (trivial) convergence theorem.

Theorem: If

- Ω_0 is compact,
- ψ is jointly continuous on Ω,

then

- The sequence $\{\psi^{(k)}\}$ converges to, say, ψ^∞,
- the sequence $\{\omega^{(k)}\}$ has at least one convergent subsequence,
- if ω^∞ is an accumulation point of $\{\omega^{(k)}\}$, then $\psi(\omega^\infty) = \psi^\infty$.

Proof: Compactness and continuity imply that the minima in each of the substeps exist.

This means that $\{\psi^{(k)}\}$ is nonincreasing and bounded below, and thus convergent. Existence of convergent subsequences is guaranteed by Bolzona-Weierstrass, and if we have a subsequence $\{\omega^{(k)}\}_{k\in\mathcal{K}}$ converging to ω^∞ then by continuity $\{\psi(\omega^{(k)})\}_{k\in\mathcal{K}}$ converges to $\psi(\omega^\infty)$. But all subsequences of a convergent sequence converge to the same point, and thus $\psi(\omega^\infty)=\psi^\infty$. **Q.E.D.**

In the special case in which blocks consist of only one coordinate we speak of the *coordinate relaxation method* or the *cyclic coordinate descend* method. Classical papers, with applications to systems of equations, quadratic programming, and convex programming are Schechter [33], [34],[35], Hildreth, D'Esopo [15], Ortega and Rheinboldt [28], [29], Elkin [17], Céa [9], [7], [8], and Auslender [2],[3]. Many of these papers present the method as a nonlinear generalization of the Gauss-Seidel method of solving a system of linear equations. Modern papers on block-relaxation are by Abatzoglou and O'Donnell [1] and by Bezdek et al. [5]. Statistical applications to mixed linear models, with the parameters describing the mean structure collected in one block and the parameters describing the dispersion collected in the second block, are in Oberhofer and Kmenta [27]. Applications to exponential family likelihood functions, cycling over the canonical parameters, are in Jensen et al. [21].

Example: Let

$$\mathcal{L}(\theta)=\sum_{k=1}^K n_k\log\lambda_k(\theta)-\lambda_k(\theta),$$

be a Poisson-likelihood with

$$\lambda_k(\theta)=\exp\sum_{j=1}^m x_{kj}\theta_j.$$

Here $\{x_{kj}\}$ is a design-type matrix, with elements equal to 0 or 1. Let

$$\mathcal{K}_j=\{k\mid x_{kj}=1\}.$$

Then the likelihood equations are

$$\sum_{k\in\mathcal{K}_j}n_k=\sum_{k\in\mathcal{K}_j}\lambda_k(\theta).$$

Solving each of these in turn is cyclic-coordinate descent, but also the *iterative propertional fitting* algorithm. We have, using e_j for the coordinate directions,

$$\lambda_k(\theta+\tau e_j)=\begin{cases}\lambda_k(\theta) & \text{if } k\notin\mathcal{K}_j,\\ \mu\lambda_k(\theta) & \text{if } k\in\mathcal{K}_j,\end{cases}$$

with $\mu=\exp\tau$. Thus the optimal μ is simply

$$\mu\leftarrow\frac{\sum_{k\in\mathcal{K}_j}n_k}{\sum_{k\in\mathcal{K}_j}\lambda_k(\theta)}.$$

3. Generalized block-relaxation methods

If there are more than two blocks, we can move through them in various ways. In analogy with linear methods such as Gauss-Seidel and Gauss-Jacobi, we distinguish *cyclic* and *free-steering* methods. We could select the block, for instance, that seems most in need of improvement. We can pivot through the blocks (A, B, C) as $\{A, B, C, B, A, B, C, B, A, \cdots\}$ or $\{A, B, B, B, C, A, B, B, B, C, \cdots\}$. We can even choose blocks in random order.

We give a formalization of these generalizations, due to Fiorot and Huard [18]. Suppose Δ_s are p point-to-set mappings of Ω into $\mathcal{P}(\Omega)$, the set of all subsets of Ω. We suppose that $\omega \in \Delta_s(\omega)$ for all $s = 1, \cdots, p$. Also define

$$\Gamma_s(\omega) \triangleq \text{argmin}\{\psi(\overline{\omega}) \mid \overline{\omega} \in \Delta_s(\omega)\}.$$

There are now two versions of the generalized block-relaxation method which are interesting. In the free-steering version we set

$$\omega^{(k+1)} \in \cup_{s=1}^{p} \Gamma_s(\omega^{(k)}).$$

This means that we select, from the p subsets defining the possible updates, one single update before we go to the next cycle of updates. In the cyclic method we set

$$\omega^{(k+1)} \in \otimes_{s=1}^{p} \Gamma_s(\omega^{(k)}).$$

In a little bit more detail this means

$$\begin{aligned}
\omega^{(k,0)} &= \omega^{(k)}, \\
\omega^{(k,1)} &\in \Gamma_s(\omega^{(k,0)}), \\
\cdots &\in \cdots, \\
\omega^{(k,p)} &\in \Gamma_s(\omega^{(k,p-1)}), \\
\omega^{(k+1)} &= \omega^{(k,p)}.
\end{aligned}$$

Since $\omega \in \Delta_s(\omega)$, we see that, for both methods, if $\xi \in \Gamma(\omega)$ then $\psi(\xi) \leq \psi(\omega)$. This implies that Theorem 1 continues to apply to this generalized block relaxation method.

A simple example of the Δ_s is the following. Suppose the G_s are arbitrary mappings defined on Ω. They need not even be real-valued. Then we can set

$$\Delta_s(\omega) \triangleq \{\xi \in \Omega \mid G_s(\xi) = G_s(\omega)\}.$$

Obviously $\omega \in \Delta_s(\omega)$ for this choice of Δ_s. There are some interesting special cases. If G_s projects on a subspace of Ω, then $\Delta(\omega)$ is the set of all ξ which project into the same point as ω. By defining the subspaces using blocks of coordinates, we recover the usual block-relaxation method discussed in the previous section. In a statistical context, in

combination with the EM algorithm, functional constraints of the form $G_s(\overline{\omega}) = G_s(\omega)\}$ were used by Meng and Rubin [24].

Here is another example, which is quite important. If we drop the assumption that Ω has product structure, then we can define

$$\Delta_s(\omega_1, \cdots, \omega_{s-1}, \omega_{s+1}, \cdots \omega_p) = \{\omega \in \mathcal{R}^s \mid (\omega_1, \cdots, \omega_{s-1}, \omega, \omega_{s+1}, \cdots \omega_p) \in \Omega\}.$$

Obviously

$$(\omega_1, \cdots, \omega_{s-1}, \omega_s, \omega_{s+1}, \cdots \omega_p) \in \Omega \text{ if and only if } \omega_s \in \Delta_s(\omega_1, \cdots, \omega_{s-1}, \omega_{s+1}, \cdots \omega_p).$$

Again, minimizing over $\omega \in \Delta_s(\omega_1, \cdots, \omega_{s-1}, \omega_{s+1}, \cdots \omega_p)$ means decreasing the loss function, and Theorem 1 applies under the usual continuity and compactness conditions.

Example:

4. Some counterexamples

We shall now strengthen our trivial convergence theorem, by imposing additional conditions on the problem. Some simple examples show that such a strengthening is necessary. We also list some examples which illustrate later results.

Convergence need not be towards a minimum. Take the function

$$\psi(\omega, \xi) = (\omega - \xi)'(\omega - \xi) - 2\omega'\xi.$$

Clearly it does not have minima (on $\omega = \xi$ we have $\psi(\omega, \omega) = -2\|\omega\|^2$). The only stationary point is the saddle $\omega = \xi = 0$, and block-relaxation convergences to that saddle from any starting point.

Convergence need not be towards a minimum, even if the function is convex. This example is from [1]. Let

$$\psi(\omega, \xi) = \max_{x \in [0,1]} \mid x^2 - \omega - \xi x \mid .$$

Start with $\xi = 0$. The optimal ω for this ξ is $1/2$. The optimal ξ for this ω is 0, which means we have convergence. But the best Chebyshev approximation to $f(x) = x^2$ is $g(x) = x + \frac{1}{8}$, and not g(x)=$1/2$.

Coordinate descend may not converge at all, even if the function is differentiable. This is a nice example, due to Powell [32]. It is somewhat surprising that Powell does not indicate what the source of the problem is, using Zangwill's convergence theory. The reason seems to be that the mathematical programming community has decided, at an early stage, that linearly convergent algorithms are not interesting and/or useful. The recent developments in statistical computing suggest that this is simply not true. Powell's example involves three variables, and the function

$$\psi(\omega) = 1/2\omega' A\omega + \text{dist}^2(\omega, \mathcal{K}),$$

where

$$a_{ij} = \begin{cases} -1 & \text{if } i \neq j, \\ 0 & \text{if } i = j, \end{cases}$$

and where $\mathcal{K}$ is the cube

$$\mathcal{K} = \{\omega \mid -1 \leq \omega_i \leq +1\},$$

The derivatives are $\mathcal{D}\psi = A\omega + 2(\omega - \mathcal{P}_{\mathcal{K}}(\omega))$. In the interior of the cube $\mathcal{D}\psi = A\omega$, which means that the only stationary point in the interior is the saddle point at $\omega = 0$. In general at a stationary point we have $(A + 2\mathcal{I})\omega = \mathcal{P}_{\mathcal{K}}(\omega))$, which means that we must have $u'\mathcal{P}_{\mathcal{K}}(\omega)) = 0$. . The only points where the derivatives vanish are saddle points. Thus the only place where there can be minima is on the surface of the cube. Also for $x = y = z = t > 1$ we see that $\psi(x, y, z) = -3t^2 + 3(t-1)^2 = 3 - 6t$, which is unbounded. For $x = y = t > 1$ and $z = -t$ we find $\psi(x, y, z) = -t^2 + 3(t-1)^2 = 2t^2 - 6t + 3$. This has its minimum -1.5 at $t = 1.5$ and its has a root at $t = 1/2(3 + \sqrt{12}) = 4.9641$.

Let us apply coordinate descend. A search along the $x-$axis finds the optimum at

$$x \leftarrow \begin{cases} +1 + 1/2(y+z) & \text{if } y + z > 0, \\ -1 + 1/2(y+z) & \text{if } y + z < 0, \\ \text{anywhere in } [-1, +1] & \text{if } y + z = 0. \end{cases}$$

This guarantees that the partial derivative with respect to x is zero. The other updates are given by symmetry. Thus, if we start from $(-1-\epsilon, 1+\frac{1}{2}\epsilon, -1-\frac{1}{4}\epsilon)$, with ϵ some small positive number, then we generate the following sequence.

$$\begin{array}{ccc}
(+1+\frac{1}{8}\epsilon, & +1+\frac{1}{2}\epsilon, & -1-\frac{1}{4}\epsilon) \\
(+1+\frac{1}{8}\epsilon, & -1-\frac{1}{16}\epsilon, & -1-\frac{1}{4}\epsilon) \\
(+1+\frac{1}{8}\epsilon, & -1-\frac{1}{16}\epsilon, & +1+\frac{1}{32}\epsilon) \\
(-1-\frac{1}{64}\epsilon, & -1-\frac{1}{16}\epsilon, & +1+\frac{1}{32}\epsilon) \\
(-1-\frac{1}{64}\epsilon, & +1+\frac{1}{128}\epsilon, & +1+\frac{1}{32}\epsilon) \\
(-1-\frac{1}{64}\epsilon, & +1+\frac{1}{128}\epsilon, & -1-\frac{1}{256}\epsilon)
\end{array}$$

But the sixth point is of the same form as the starting point, with ϵ replaced by $\frac{\epsilon}{64}$. Thus the algorithm will cycle around six edges of the cube. At these edges the gradient of the function is bounded away from zero, in fact two of the partials are zero, the other is ± 2. The function value is $+1$. The other two edges of the cube, i.e. $(+1, +1, +1)$ and $(-1, -1, -1)$ are the ones we are looking for, because there the function value is -3, the global minimum. At these two points all three partials are ± 2. Powell gives some additional examples which show the same sort of cycling behaviour, but are somewhat smoother.

Convergence can be sublinear.

$$\begin{aligned}
\psi(\omega,\xi) &= (\omega-\xi)^2+\omega^4,\\
\mathcal{D}_1\psi(\omega,\xi) &= 2(\omega-\xi)+4\omega^3,\\
\mathcal{D}_2\psi(\omega,\xi) &= -2(\omega-\xi),\\
\mathcal{D}_{11}\psi(\omega,\xi) &= 2+12\omega^2,\\
\mathcal{D}_{12}\psi(\omega,\xi) &= -2,\\
\mathcal{D}_{22}\psi(\omega,\xi) &= 2.
\end{aligned}$$

It follows that coordinate ascent updates $\omega^{(k)}$ by solving the cubic

$$\omega-\omega^{(k)}+2\omega^3=0.$$

The sequence converges to zero, and by l'Hopitâl's rule

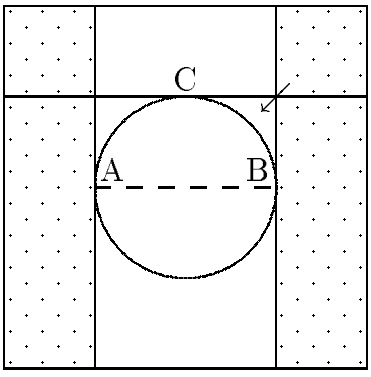
$$\lim_{k\to\infty}\frac{\omega^{(k+1)}}{\omega^{(k)}}=1.$$

This leads to very slow convergence. The reason is that the matrix of second derivatives of ψ is singular at the origin.

Strict monotonicity can be violated if the minima in the subproblems are not unique Let

$$\psi(\omega,\xi)=\begin{cases}\omega^2+\xi^2-1 & \text{if } \omega^2+\xi^2\leq 1,\\ 0 & \text{otherwise.}\end{cases}$$

Let us start at $(0,1)$ and minimize over the line $(\omega,1)$. Any point on the line gives the minimum. Now let us minimize (ω,ξ) over ξ. If $\omega\geq+1$ or $\omega\leq-1$ any point ξ will give the minimum. If $-1\leq\omega\leq+1$ then the minimum is attained for $\xi=0$.



5. Global Convergence

Theorem 1 is very general, but the conclusions are quite weak. We have convergence of the function values, but about the sequence $\{\omega^{(k)}\}$ we only know that it has one or more

accumulation points, and that all accumulation points have the same function value. We do not know other desirable properties of these accumulation points.

In order to improve global convergence (i.e. convergence from any initial point) we use the general theory developed initially by Zangwill [39],[40] (and later by Polak [31], R.R. Meyer [26], G.G.L. Meyer [25], and others). The best introduction and overview is perhaps the volume edited by Huard [16].

The theory studies iterative algorithms with the following properties. An algorithm works in a space Ω. It consists of a triple $(\mathcal{A}, \psi, \mathcal{P})$, with $\mathcal{A}$ a mapping of Ω into the set of nonempty subsets of Ω, with ψ is real-valued continuous function on Ω, and with $\mathcal{P}$ a subset of Ω. We can $\mathcal{A}$ the *algorithmic map,* ψ the *evaluation function,* and $\mathcal{P}$ the *desirable points.* The algorithm works as follows.

1) start at an arbitrary $\omega^{(0)} \in \Omega$,
2) if $\omega^{(k)} \in \mathcal{P}$, then we stop,
3) otherwise we construct the *successor* by the rule $\omega^{(k+1)} \in \mathcal{A}(\omega^k)$,

We study properties of the sequences $\omega^{(k)}$ generated by the algorithm, in particular their convergence.

Theorem: (Zangwill [39]) If

- $\mathcal{A}$ is *uniformly compact* on Ω, i.e. there is a compact $\Omega_0 \subseteq \Omega$ such that $\mathcal{A}(\omega) \subseteq \Omega_0$ for all $\omega \in \Omega$,
- $\mathcal{A}$ is *upper-semicontinuous* or *closed* on $\Omega - \mathcal{P}$, i.e. if $\xi_i \in \mathcal{A}(\omega_i)$ and $\xi_i \to \xi$ and $\omega_i \to \omega$ then $\xi \in \mathcal{A}(\omega)$,
- $\mathcal{A}$ is *strictly monotonic* on $\Omega - \mathcal{P}$, i.e. $\xi \in \mathcal{A}(\omega)$ implies $\psi(\xi) < \psi(\omega)$ if ω is not a *desirable point.*

then all accumulation points of the sequence $\{\omega^{(k)}\}$ generated by the algorithm are desirable points.

Proof: Compactness implies that $\{\omega^{(k)}\}$ has a convergent subsequence. Suppose its index-set is

$$\mathcal{K} = \{k_1, k_2, \cdots\}$$

and that it converges to $\omega_{\mathcal{K}}$. Since $\{\psi(\omega^{(k)})\}$ converges to, say, ψ_∞, we see that also

$$\{\psi(\omega^{(k_1)}), \psi(\omega^{(k_2)}), \cdots\} \to \psi_\infty.$$

Now consider $\{\omega^{(k_1+1)}, \omega^{(k_2+1)}, \cdots\}$, which must again have a convergent subsequence. Suppose its index-set is $\mathcal{L} = \{\ell_1 + 1, \ell_2 + 1, \cdots\}$ and that it converges to $\omega_{\mathcal{L}}$. Then $\psi(\omega_{\mathcal{K}}) = \psi_(\omega_{\mathcal{L}}) = \psi_\infty$.

Assume $\omega_{\mathcal{K}}$ is not a fixed point. Now

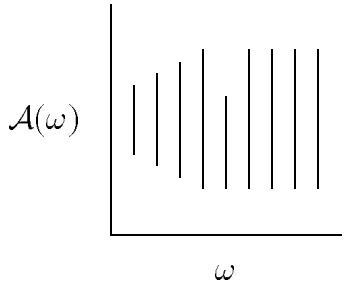
$$\{\omega^{(\ell_1)}, \omega^{(\ell_2)}, \cdots\} \to \omega_{\mathcal{K}}$$

and

$$\{\omega^{(\ell_1+1)}, \omega^{(\ell_2+1)}, \cdots\} \rightarrow \omega_{\mathcal{L}},$$

with $\omega^{(\ell_j+1)} \in \mathcal{A}(\omega^{(\ell_j+1)}$. Thus, by usc, $\omega_{\mathcal{L}} \in \mathcal{A}(\omega_{\mathcal{K}})$. If $\omega_{\mathcal{K}}$ is not a fixed point, then strict monotonicity gives $\psi(\omega_{\mathcal{L}}) < \psi(\omega_{\mathcal{K}})$, which contradicts our earlier $\psi(\omega_{\mathcal{K}}) = \psi_(\omega_{\mathcal{L}})$. **Q.E.D.**

The concept of closedness of a map can be illustrated with the following picture, showing a map which is not closed at at least one point.



We have already seen another example: Powell's coordinate descend example shows that the algorithm map is not closed at six of the edges of the cube $\{\pm 1, \pm 1, \pm 1\}$.

It is easy to see that desirable points are generalized fixed points, in the sense that $\omega \in \mathcal{P}$ is equivalent to that $\omega \in \mathcal{A}(\omega)$. According to Zangwill's theorem each accumulation point is a generalized fixed point. This, however, does not prove convergence, because there can be many accumulation points. If we redefine fixed points as points such that $\mathcal{A}(x) = \{x\}$, then we can strengthen the theorem.

Theorem: (Meyer, [26]) Suppose the conditions of Zangwill's theorem are satisfied for the stronger definition of a fixed point, i.e. $\xi \in \mathcal{A}(\omega)$ implies $\psi(\xi) < \psi(\omega)$ if ω is not a fixed point, then in addition to what we had before $\{\omega^{(k)}\}$ is *asymptotically regular,* i.e.

$$\|\omega^{(k)} - \omega^{(k+1)}\| \rightarrow 0.$$

Proof: Use the notation in the proof of Zangwill's theorem. Suppose $\|\omega^{(\ell_i+1)} - \omega^{(\ell_i)}\| > \delta > 0$. Then $\|\omega_{\mathcal{L}} - \omega_{\mathcal{K}}\| \geq \delta$. But $\omega_{\mathcal{K}}$ is a fixed point (in the strong sense) and thus $\omega_{\mathcal{L}} \in \mathcal{A}(\omega_{\mathcal{K}}) = \{\omega_{\mathcal{K}}\}$, a contradiction. **Q.E.D.**

It follows (from a result of Ostrowski [30]) that either $\{\omega^{(k)}\}$ converges, or $\{\omega^{(k)}\}$ has a continuum of accumulation points (all with the same function value). This is still not actual convergence, but it is close enough for all practical purposes.

6. Global convergence of block methods

We can now apply this theory to block-relaxation methods. We concentrate on the cyclic methods. The free-steering methods are interesting, but inherently more complicated. Details on free-steering can be found in [18]. Obviously block-relaxation is monotonic if we choose the evaluation function equal to the function we are minimizing, and if we assume that the minima exist. If we assume that the minima of the subproblems are always

unique (for instance, if they are least squares projections on convex sets), then Meyer's theorem applies. Actually, we have the following result for generalized block methods.

Theorem: (Fiorot and Huard, [18]) If

- $\omega \in \Delta_s(\omega)$ for all ω and s,
- Δ_s is continuous on Ω, i.e. both upper-semicontinuous and lower-semicontinuous,
- ψ has a unique minimum over $\Delta_s(\omega)$ for all ω and s,
- $\Omega_0 = \{\omega \in \Omega \mid \psi(\omega) \leq \psi(\omega^{(0)})\}$ is compact,

then

- the sequence $\omega^{(k)}$ is asymptotically regular,
- each accumulation point of the sequence is a fixed point of each of the Γ_s.

A fixed point $(\omega_1, \cdots, \omega_p)$ is by definition a point such that ω_s is the unique minimum of $\psi(\omega_1, \cdots, \omega_{s-1}, \omega, \omega_{s+1}, \cdots, \omega_p)$ over $\omega \in \Omega_s$ for all s. This does not imply that the point is a local minimum of ψ on Ω unless we impose extra conditions such as convexity. Actually, convexity is not enough, as the Chebyshev approximation example in section 4 shows.

If we drop the assumption that the partial minima of the subproblems are unique (which is of course basically an identification condition, similar to assumptions needed for consistency) then fixed points must be replaced by generalized fixed points. Also, accumulation points are no longer generalized fixed points of *all* Γ_s. In fact each accumulation point ω_∞ has an associated index set $S(\omega_\infty)$ such that $s \in S(\omega_\infty)$ if the operation of maximizing over Δ_s occurs an infinite number of times in the subsequence. For the six edges in the Powell example, these index sets consist of a single element.

Theorem: ((Fiorot and Huard, [18]) If

- $\omega \in \Delta_s(\omega)$ for all ω and s,
- Δ_s is continuous on Ω, i.e. both upper-semicontinuous and lower-semicontinuous,
- if $\xi \in \Delta_s(\omega)$ then $\Delta_s(\xi) = \Delta_s(\omega)$,
- $\Omega_0 = \{\omega \in \Omega \mid \psi(\omega) \leq \psi(\omega^{(0)})\}$ is compact,

then for every $s \in S(\omega_\infty)$ we have $\omega_\infty \in \Gamma_s(\omega_\infty)$ and $\omega_\infty \in \Gamma_{s+1}(\omega_\infty)$.

7. Quantitative convergence theory

We now switch from the qualitative or global theory of convergence to the quantitative or local theory. We look into the question of convergence speed. To get this more specific information on convergence, we again have to make stronger assumptions. To be able to compute the rate, we need to be able to differentiate ψ sufficiently many times. Also, the solution of the subproblems needs to be unique in a neighborhood of the true value. Thus we forget all references to point-to-set maps, and to free-steering, because our techniques here simply cannot cope with that much freedom. The basic

result we use is due to Ostrowski [30].

Theorem: If

— the iterative algorithm $\omega^{(k+1)} = \mathcal{A}(x^{(k)})$, converges to ω_∞,
— $\mathcal{A}$ is differentiable at ω_∞,
— $0 < \rho = \|\mathcal{D}A(\omega_\infty)\| < 1$,

then the algorithm is linearly convergent with rate ρ.

The norm in the theorem is the spectral norm, i.e. the modulus of the maximum eigenvalue. Let us call the derivative of A the *iteration matrix* and write it as $\mathcal{M}$. In general block relaxation methods have linear convergence, and the linear convergence can be quite slow. In cases where the accumulation points are a continuum we usually have sublinear rates. The same things is true if the local minimum is not strict, or if we are converging to a saddle point.

In order to study the rate of convergence of block relaxation, we study the nonlinear system

$$\begin{aligned}\mathcal{D}_1(\omega_1,\xi_2,\xi_3,\cdots,\xi_p) &= 0,\\ \mathcal{D}_2(\omega_1,\omega_2,\xi_3,\cdots,\xi_p) &= 0,\\ \cdots &= \cdots,\\ \mathcal{D}_p(\omega_1,\omega_2,\omega_3,\cdots,\omega_p) &= 0,\end{aligned}$$

which defines the new solution ω in terms of the old solution ξ. The $\mathcal{D}_s$ are the partials of ψ with repect to the blocks. We assume that the assumptions for the implicit function theorem are satisfied at the solution. Differentiating these equations again, and solving for the derivatives, we find the iteration matrix

$$\mathcal{M} = -\begin{pmatrix}\mathcal{D}_{11} & 0 & 0 & \cdots & 0\\ \mathcal{D}_{21} & \mathcal{D}_{22} & 0 & \cdots & 0\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ \mathcal{D}_{s1} & \mathcal{D}_{s2} & \mathcal{D}_{s3} & \cdots & \mathcal{D}_{ss}\end{pmatrix}^{-1}\begin{pmatrix}0 & \mathcal{D}_{12} & \mathcal{D}_{13} & \cdots & \mathcal{D}_{1s}\\ 0 & 0 & \mathcal{D}_{23} & \cdots & \mathcal{D}_{2s}\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ 0 & 0 & 0 & \cdots & 0\end{pmatrix}.$$

If there are only two blocks this simplifies to

$$\mathcal{M} = -\begin{pmatrix}\mathcal{D}_{11}^{-1} & 0\\ -\mathcal{D}_{22}^{-1}\mathcal{D}_{21}\mathcal{D}_{11}^{-1} & \mathcal{D}_{22}^{-1}\end{pmatrix}\begin{pmatrix}0 & \mathcal{D}_{12}\\ 0 & 0\end{pmatrix} = \begin{pmatrix}0 & -\mathcal{D}_{11}^{-1}\mathcal{D}_{12}\\ 0 & \mathcal{D}_{22}^{-1}\mathcal{D}_{21}\mathcal{D}_{11}^{-1}\mathcal{D}_{12}\end{pmatrix}.$$

Thus, in a local minimum, we find that the largest eigenvalue of $\mathcal{M}$ is the largest squared canonical correlation ρ of the two sets of variables, and is consequently less than or equal to one. We also see that a sufficient condition for local convergence to a stationary point of the algorithm is that $\rho < 1$. This precludes having more than one accumulation point, and it is always true for an isolated local minimum. If $\mathcal{D}^2\psi$ is singular at the solution, we find a canonical correlation equal to $+1$, and we do not have linear convergence.

Similar calculations can also be carried out in the case of constrained optimization, i.e. when the subproblems optimize over differentiable manifolds. We then use the implicit function calculations on the Langrangean conditions, which makes them a bit more complicated, but essentially the same.

The result for block-relaxation can also derived from a similar result for generalized block relaxation, that has been used in an EM context by Meng [23]. We minimize ψ over ω under the condition that $G_s(\omega) = G_s(\xi)$, where ξ is the current solution. Once again we can differentiate the stationary equations to find that

$$\frac{\partial \omega}{\partial \xi} = T_s^{-1} H_s' (H_s T_s^{-1} H_s')^{-1} H_s,$$

where H_s is the Jacobian of G_s at the solution, and where

$$T_s = \mathcal{D}^2 \psi + \sum_{r=1}^{m} \lambda_{sr} \mathcal{D}^2 g_{rs}.$$

Here g_{rs} is the $r-$th restriction in the $s-$th system, and the λ_{sr} are the corresponding Lagrange multipliers. If the G_s are linear, the second term disappears, and all T_s are equal to the Hessian of ψ at the solution. If we use a generalized block method that cycles over the constraints G_s, then the iteration matrix is simply

$$\mathcal{M} = \prod_{s=1}^{p} T_s^{-1} H_s' (H_s T_s^{-1} H_s')^{-1} H_s.$$

In the case of ordinary block relaxation the G_s are linear, because they are the indicator matrices selecting the blocks that do not change in a subproblem. For the first subproblem $G_1 = (0 \mid \mathcal{I})$, and we find

$$\mathcal{M}_1 = \begin{pmatrix} 0 & \mathcal{D}^{12}(\mathcal{D}^{22})^{-1} \\ 0 & \mathcal{I} \end{pmatrix},$$

with the $\mathcal{D}^{st}$ the blocks of the inverse of $\mathcal{D}^2\psi$. If we substitute the G_s for the H_s, we find an alternative expression for the iteration matrix as a product of simpler matrices. This simpler expression, which corresponds with the pivoting or Gauss-Jordan way to compute the inverse, can also be used if we iterate the blocks in orders such as $(1, 2, \cdots, p, p-1, \cdots, 1, 2, \cdots)$ because we just have to multiply the blockwise transformations.

Use of over-relaxation.

8. Alternating Least Squares

We now go into the history of block-relation in statistics and data analysis. Alternating Least Squares (ALS) methods were first used systematically in *Optimal Scaling* (OS).

Optimal scaling is discussed in detail in the book by Gifi [19]. We only give a brief introduction here.

Suppose we have n observations on two sets of variables x_i and y_i. We want to fit a model of the form

$$F_\theta(\Phi(x_i)) \approx G_\xi(\Psi(y_i))$$

where the unknowns are the structural parameters θ and ξ and the transformations Φ and Ψ. In ALS we measure loss-of-fit by

$$\sigma(\theta, \xi, \Phi, \Psi) = \sum_{i=1}^n [F_\theta(\Phi(x_i)) - G_\xi(\Psi(y_i))]^2$$

This loss function is minimized by starting with initial estimates for the transformations, minimizing over the structural parameters, keeping the transformations fixed at their current values, and then minimizing over the transformations, with structural values kept fixed at their new values. These two minimizations are alternated, which produces a nonincreasing sequence of loss function values, bounded below by zero, and thus convergent. This is a version of the trivial convergence theorem.

The first ALS example is due to Kruskal [22]. We have a factorial ANOVA, with, say, two factors, and we minimize

$$\sigma(\phi, \mu, \alpha, \beta) = \sum_{i=1}^n \sum_{j=1}^m [\phi(y_{ij}) - (\mu + \alpha_i + \beta_j)]^2.$$

Kruskal required ϕ to be monotonic. Minimizing loss for fixed ϕ is just doing an analysis of variance, minimizing loss over ϕ for fixed μ, α, β is doing a *monotone regression.* Obviously also some normalization reuqirement is needed to exclude trivial zero solutions.

This general idea was extended by De Leeuw, Young, Takane around 1975 to

$$\sigma(\phi; \psi_1, \cdots, \psi_m) = \sum_{i=1}^n [\phi(y_i) - \sum_{s=1}^p \psi_j(x_{ij})]^2.$$

This ALSOS work, in the period 1975-1980, is summarized in [38]. Subsequent work, culminating in the book by Gifi [19], generalized this to ALSOS versions of principal component analysis, path analysis, canonical analysis, discriminant analysis, MANOVA, and so on. The classes of transformations over which loss was minimized were usually step-functions, splines, monotone functions, or low-degree polynomials. To illustrate the use of more sets in ALS, consider

$$\sigma(\psi_1, \cdots, \psi_m; \alpha, \beta) = \sum_{i=1}^n \sum_{j=1}^m (\psi_j(x_{ij}) - \sum_{s=1}^p \alpha_{is}\beta_{js})^2.$$

This is principal component analysis (or partial singular value decomposition) with optimal scaling. We can now cycle over three sets, the transformations, the component scores α_{is} and the component loadings β_{js}. In the case of monotone transformations this alternates monotone regression with two linear least squares problems.

The ACE methods, developed by Breiman and Friedman [6], "minimize" over all "smooth" functions. A problem with ACE is that smoothers, at least most smoothers, do not really minimize a loss function (except for perfect data). In any case, ACE is less general than ALS, because not all least squares problems can be interpreted as computing conditional expectations. Another obviously related area in statistics is the Generalized Additive Models discussed extensively by Hastie and Tibshirani [20].

It is easy to apply the general results from the previous sections to ALS. The results show that it is important that the solutions to the subproblems are unique. The least squares loss function has some special structure in its second derivatives which we can often exploit in a detailed analysis. If

$$\sigma(\omega,\xi)=\sum_{i=1}^n(f_i(\omega)-g_i(\xi))^2,$$

then

$$\mathcal{D}^2\sigma=\begin{pmatrix}S_1&0\\0&S_2\end{pmatrix}+\begin{pmatrix}G'G&-G'H\\-H'G&H'H\end{pmatrix},$$

with G and H the Jacobians of f and g, and with S_1 and S_2 weighted sums of the Hessians of the f_i and g_i, with weights equal to the least squares residuals at the solution. If S_1 and S_2 are small, because the residuals are small, or because the f_i and g_i are linear or almost linear, we see that the rate of ALS will be the canonical correlation between G and H.

Scaling and Splitting

Early on in the development of ALS algorithms some interesting complications where discovered. Let us consider canonical correlation analysis with optimal scaling. Thus we want to minimize

$$\sigma(X,Y,A,B)=\text{tr }(XA-YB)'(XA-YB),$$

where the X and the Y are optimally scaled or transformed variables. This problem is analyzed in detail in Van der Burg and De Leeuw [`canals`]. This seems like a perfectly straightforward ALS problem. It can be formulated as a problem with the two blocks (X,Y) and (A,B), or as a problem with the four blocks X,Y,A,B. But no matter how one formulates it, a normalization must be chosen to prevent trivial solutions. In the spirit of canonical analysis it makes snesen to require $A'X'XA=\mathcal{I}$ or $B'Y'YB=\mathcal{I}$. It is easy to see that both ultimately both sets of conditions lead

to the same solution, but in the intermediate iterations the normalization condition creates a problem, because it involves elements from two different blocks.

9. Augmentation methods

We take up the historical developments. Alternating Least Squares was useful for many problems, but it some cases it was not powerful enough to do the job. Or, to put it differently, the subproblems were still too complicated to be efficiently solved a large number of times. In order to solve some additional least squares problems, we can use *augmentation*. We first illustrate this with some examples.

Example: If we want to fit a factorial ANOVA model to an unbalanced two-factor design, we minimize

$$\sigma(\mu,\alpha,\beta)=\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}w_{ijk}(y_{ijk}-(\mu+\alpha_i+\beta_j))^2,$$

where the weights w_{ijk} are either one (there) or zero (not there). Instead of this we can also minimize

$$\sigma(\mu,\alpha,\beta,z)=\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(z_{ijk}-(\mu+\alpha_i+\beta_j))^2,$$

with

$$z_{ijk}=\begin{cases}y_{ijk}, & \text{if } w_{ijk}=1\\ \text{free}, & \text{otherwise.}\end{cases}$$

Minimizing this by ALS is due to Yates an others, see Wilkinson [37] for references. Augmentation reduces the fitting to the balanced case (where we can simply use row, column, and cell means), with an additional step to *impute* the missing y_{ijk}. We can give the algorithm more explicitly as

$$\begin{aligned}\mu^{(k+1)}&=\hat{z}_{\bullet\bullet\bullet},\\ \alpha_i^{(k+1)}&=\hat{z}_{i\bullet\bullet}-\hat{z}_{\bullet\bullet\bullet},\\ \beta_j^{(k+1)}&=\hat{z}_{\bullet j\bullet}-\hat{z}_{\bullet\bullet\bullet},\end{aligned}$$

where

$$\hat{z}_{ijk}=w_{ijk}y_{ijk}+(1-w_{ijk})(\mu^{(k)}+\alpha_i^{(k)}+\beta_j^{(k)}),$$

Example: In LS factor analysis we want to minimize

$$\sigma(A)=\sum_{i=1}^{m}\sum_{j=1}^{m}w_{ij}(r_{ij}-\sum_{s=1}^{p}a_{is}a_{js})^2,$$

with

$$w_{ij} = \begin{cases} 0, & \text{if } i = j, \\ 1, & \text{if } i \neq j. \end{cases}$$

We augment by adding the *communalities,* i.e. the diagonal elements of R as variables, and by using ALS over A and the communalities. For a complete R, minimizing over A just means computing the p dominant eigenvalues-eigenvectors. This algorithm dates back to the thirties, were it was proposed by Thomson and others.

Example: A final example, less trivial in a sense. Suppose we want to minimize

$$\sigma(X) = \sum_{i=1}^m \sum_{j=1}^m (\delta_{ij} - d_{ij}^2(X))^2,$$

with $d_{ij}^2(X) = (x_i - x_j)'(x_i - x_j)$ squared Euclidean distance. This can be augmented to

$$\sigma(X, \eta) = \sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^m \sum_{\ell=1}^m (\eta_{ijk\ell} - (x_i - x_j)'(x_k - x_\ell))^2,$$

where of course $\eta_{ijij} = \delta_{ij}$ and the others are free. After some computation, ALS again leads to a sequence of eigenvalue-eigenvector problems.

This example shows that augmentation is an art (like integration). The augmentation is in some cases not obvious, and there are no mechanical rules. The idea of adding variables that augment the problem to a simpler one is very general. It is also at the basis, for instance, of the Lagrange multiplier method.

Formalizing augmentation is straightforward. Suppose ϕ is a real valued function, defined for all $\omega \in \Omega$, where $\Omega \subseteq \mathcal{R}^n$. Suppose there exists another real valued function ψ, defined on $\Omega \times \Xi$, where $\Xi \subseteq \mathcal{R}^m$, such that

$$\phi(\omega) = \min\{\psi(\omega, \xi) \mid \xi \in \Xi\}.$$

We also suppose that minimizing ϕ over Ω is *hard,* while minimizing ψ over Ω is *easy* for all $\xi \in \Xi$. And we suppose that minimizing ψ over $\xi \in \Xi$ is also *easy* for all $\omega \in \Omega$. This last assumption is not too far-fatched, because we already know what the value at the minimum is.

I am not going to define *hard* and *easy.* What may be easy for you, may be hard for me. Anyway, by augmenting the function we are in the block-relaxation situation again, and we can apply our general results on global convergence and linear convergence. The results can be adapted to the augmentation situation. Because of the structure of ϕ we know that

$$\mathcal{D}^2\phi = \mathcal{D}_{11}\psi - \mathcal{D}_{12}\psi[\mathcal{D}_{22}\psi]^{-1}\mathcal{D}_{21}\psi.$$

It follows that

$$\mathcal{M} = \mathcal{I} - [\mathcal{D}_{11}\psi]^{-1}\mathcal{D}^2\phi.$$

This shows how the iteration matrix does not depend (directly) on the derivatives of ψ with respect to ξ, and can be interpreted as one minus the curvature of the function at the minimum, relative to the curvature of the augmentation function.

Augmentation is used in other areas of statistics [36], where integration is used instead of minimization. If it is difficult to sample from $p(\omega)$ and easy to sample from $p(\omega, \xi)$, then we sample from the joint distribution and integrate out the ξ by summation.

Example: We give another, more serious, example from the area of mixed-model fitting. This is taken from a paper of De Leeuw and Liu [13], which describes the algorithm in detail. We simply give a list of results that show augmentation at work. We maximize a multinormal likelihood, not a least squares criterium.

Lemma: If $A = B + TCT'$, with $B, C > 0$,

$$y'A^{-1}y = \min_x (y - Tx)'B^{-1}(y - Tx) + x'C^{-1}x.$$

Lemma: If $A = B + TCT'$, with $B, C > 0$,,

$$\begin{aligned}&\log \mid A \mid = \\ &= \log \mid B \mid + \log \mid C \mid + \log \mid C^{-1} + T'B^{-1}T \mid .\end{aligned}$$

Theorem: If $A = B + TCT'$ then

$$\begin{aligned}&\log \mid A \mid + y'A^{-1}y = \min_x \ \log \mid B \mid + \log \mid C \mid + \\ &+ \log \mid C^{-1} + T'B^{-1}T \mid + \\ &+ (y - Tx)'B^{-1}(y - Tx) + x'C^{-1}x.\end{aligned}$$

Lemma: If $T > 0$ then

$$\log \mid T \mid = \min_{S>0} \log \mid S \mid + \text{ tr } S^{-1}T - p,$$

with the unique minimum attained at $S = T$.

Theorem:

$$\begin{aligned}&\log \mid A \mid + y'A^{-1}y = \min_{x, S>0} \ \log \mid B \mid + \log \mid C \mid + \\ &+ \log \mid S \mid + \text{ tr } S^{-1}(C^{-1} + T'B^{-1}T) + \\ &+ (y - Tx)'B^{-1}(y - Tx) + x'C^{-1}x.\end{aligned}$$

Minimize over x, S, B, C using block-relaxation. The minimizers are

$$\begin{aligned} S &= C^{-1} + T'B^{-1}T, \\ C &= S^{-1} + xx', \\ B &= TS^{-1}T' + (y - Tx)(y - Tx)', \\ x &= (T'B^{-1}T + C^{-1})^{-1}T'B^{-1}y. \end{aligned}$$

10. Majorization methods

The next step (history again) was to find *systematic ways* to do augmentation (which is an art, remember). We start with examples.

Example: The first is an algorithm for multidimensional scaling, developed by De Leeuw [10]. We want to minimize

$$\sigma(X) = \tfrac{1}{2}\sum_{i=1}^m\sum_{j=1}^m w_{ij}(\delta_{ij} - d_{ij}(X))^2,$$

with $d_{ij}(X)$ again Euclidean distance, i.e. $d_{ij}(X) = \sqrt{(x_i - x_j)'(x_i - x_j)}$. We suppose weights w_{ij} and dissimilarities δ_{ij} are symmetric and hollow (zero diagonal), and satisfy

$$\tfrac{1}{2}\sum_{i=1}^m\sum_{j=1}^m w_{ij}\delta_{ij}^2 = 1.$$

We now define the following objects

$$\begin{aligned} \eta^2(X) &= \sum_{i=1}^m\sum_{j=1}^m w_{ij}d_{ij}(X)^2, \\ \rho(X) &= \sum_{i=1}^m\sum_{j=1}^m w_{ij}\delta_{ij}d_{ij}(X). \end{aligned}$$

Thus

$$\sigma(X) = 1 - 2\rho(X) + \tfrac{1}{2}\eta^2(X).$$

The next step is to use matrices. Let

$$\begin{aligned} v_{ij} &= \begin{cases} -w_{ij} & \text{if } i \neq j, \\ \sum_{k\neq i}^m w_{ik} & \text{if } i = j, \end{cases} \\ b_{ij}(X) &= \begin{cases} -\frac{w_{ij}\delta_{ij}}{d_{ij}(X)} & \text{if } i \neq j, \\ \sum_{k\neq i}^m \frac{w_{ik}\delta_{ik}}{d_{ik}(X)} & \text{if } i = j. \end{cases} \end{aligned}$$

Now

$$\sigma(X) = 1 - \text{tr}X'B(X)X + {}^1\!/_2\text{tr}X'VX.$$

By Cauchy-Schwarz,

$$d_{ij}(X) \geq \frac{(x_i - x_j)'(y_i - y_j)}{d_{ij}(Y)},$$

which implies

$$\text{tr}X'B(X)X \geq \text{tr}X'B(Y)Y.$$

Now let

$$\overline{X} = V^+B(X)X.$$

This is called the *Guttman-tranform* of a matrix X. Using this transform we see that for all pairs of configurations (X, Y)

$$\begin{aligned}\sigma(X) &\leq 1 - \text{tr } X'B(Y)Y + {}^1\!/_2\text{tr } X'VX = \\ &= 1 - \text{tr } XV\overline{Y} + {}^1\!/_2\text{tr } X'VX = \\ &= 1 - {}^1\!/_2\text{tr } \overline{Y}'V\overline{Y} + {}^1\!/_2\text{tr } (X - \overline{Y})'V(X - \overline{Y}),\end{aligned}$$

while for all configurations X we have

$$\sigma(X) = 1 - {}^1\!/_2\text{tr } \overline{X}'V\overline{X} + {}^1\!/_2\text{tr } (X - \overline{X})'V(X - \overline{X}).$$

Example: Suppose we want to maximize $\phi(\omega) = \log \int \eta(\omega, x)dx$. This is maximizing an integral which depends on a parameter ω. By Jensen's inequality

$$\begin{aligned}&\log \frac{\int \eta(\omega, x)dx}{\int \eta(\xi, x)dx} = \log \frac{\int \eta(\xi, x)\frac{\eta(\omega,x)}{\eta(\xi,x)}dx}{\int \eta(\xi, x)dx} \geq \\ &\geq \frac{\int \eta(\xi, x)\log \frac{\eta(\omega,x)}{\eta(\xi,x)}dx}{\int \eta(\xi, x)dx} = \\ &= \frac{\int \eta(\xi, x)\log \eta(\omega, x)dx}{\int \eta(\xi, x)dx} - \frac{\int \eta(\xi, x)\log \eta(\xi, x)dx}{\int \eta(\xi, x)dx}.\end{aligned}$$

It follows that

$$\phi(\omega) \geq \phi(\xi) + \kappa(\omega, \xi) - \kappa(\xi, \xi),$$

Maximizing the right-hand-side by block relaxation is the EM algorithm [14]. Usually, of course, the EM algorithm is presented in probabilistic terms using the concept of likelihood and expectation. This has considerable heuristic value, but it detracts somewhat from seeing the essential engine of the algorithm, which is the majorization.

As before, we now stop and wonder what these two examples have in common. We have a function $\phi(\omega)$ on Ω, and a function $\psi(\omega, \xi)$ on $\Omega \otimes \Omega$ such that

$$\begin{aligned}\phi(\omega) &\leq \psi(\omega, \xi) \quad \forall \omega, \xi \in \Omega, \\ \phi(\omega) &= \psi(\omega, \omega) \quad \forall \omega \in \Omega.\end{aligned}$$

This is just another way of saying

$$\phi(\omega) = \min_{\xi \in \Omega} \psi(\omega, \xi),$$

and thus we are in the ordinary block relaxation situation. We say that ψ *majorizes* ϕ, and we call the block relaxation algorithm corresponding with a particular majorization function a *majorization algorithm.* It is a special case of our previous theory, because $\Omega = \Xi$ and because $\xi(\omega) = \omega$. This implies that $cD_2(\omega,\omega) = 0$ for all ω, and consequently $\mathcal{D}_{12} = -\mathcal{D}_{22}$. Thus $\mathcal{M} = -\mathcal{D}_{11}^{-1}\mathcal{D}_{12}$. The E-step of the EM algorithm, in our terminology, is the construction of a new majorization function. We prefer a nonstochastic description of EM, because maximizing integrals is obviously a more general problem.

Again, to some extent, finding a majorization function is an art. Many of the classical inequalities can be used (Cauchy-Schwarz, Jensen, Hölder, AM-GM, and so on). Here are some systematic ways to find majorizing functions.

1) If ϕ is concave, then $\phi(\omega) \leq \phi(\xi) + \eta'(\omega - \xi)$, with $\eta \in \partial\phi(\xi)$, the subgradient of ϕ at ξ. Thus concave functions have a linear majorizer.
2) If $\mathcal{D}^2\phi(\xi) \leq D$ for all $\xi \in \Omega$, then

$$\phi(\omega) \leq \phi(\xi) + (\omega - \xi)'\nabla\phi(\xi) + \tfrac{1}{2}(\omega - \xi)'D(\omega - \xi).$$

Let $\eta(\xi) = \xi - D^{-1}\nabla\phi(\xi)$, then

$$\begin{aligned}\phi(\omega) \leq \phi(\xi) - \tfrac{1}{2}\nabla\phi(\xi)'D^{-1}\nabla\phi(\xi) + \\ + \tfrac{1}{2}(\omega - \eta(\xi))'D(\omega - \eta(\xi)).\end{aligned}$$

Thus here we have quadratic majorizers.
3) For d.c. functions (differences of convex functions) such as $\phi = \alpha - \beta$ we can write $\phi(\omega) \leq \alpha(\omega) - \beta(\xi) - \eta'(\omega - \xi)$, with $\eta \in \partial\beta(\xi)$. This gives a convex majorizer. Interesting, because basically all continuous functions are d.c.

Example: Suppose ψ is a convex and differentiable function defined on the space of all correlation matrices R between m random variables $x_1, \cdots, x_m$.. Suppose we want to maximize $\psi(R(\eta_1(x_1), \cdots, \eta_m(x_m)))$ over all transformations η_j. Now

$$\psi(R) \geq \psi(S) + \text{ tr } \nabla\psi(S)'(R - S).$$

Collect the gradient in the matrix G. A majorization algorithm can maximize

$$\sum_{i=1}^m \sum_{j=1}^m g_{ij}(S)\mathbf{E}(\eta_i\eta_j),$$

over all standardized transformations, which we do with block relaxation using m blocks. In each block we must maximize a linear function under a quadratic constraint

(unit variance), which is usually very easy to do. This algorithm generalizes ACE, CA, and many other forms of MVA with OS. It was proposed first by De Leeuw [11], with many variations. The function ψ can be based on multiple correlations, eigenvalues, determinants, and so on.

Example: Here we show how to use the arithmetic mean-geometric mean inequality for majorization. Suppose our problem is to minimize

$$\phi(X) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}d_{ij}(X),$$

where the w_{ij} are non-negative weights, and the $d_{ij}(X)$ are again Euclidean distances. This is a *location problem.* To make it interesting, we suppose that some of the points (facilities) are fixed, others are the variables we have to minimize over. Observe that this is a convex, but non-differentiable, optimization problem. We use the AM-GM inequality in the form

$$d_{ij}(X)d_{ij}(Y) \leq \tfrac{1}{2}(d_{ij}^2(X) + d_{ij}^2(Y)).$$

If $d_{ij}(Y) > 0$ then

$$d_{ij}(X) \leq \tfrac{1}{2}\frac{d_{ij}^2(X) + d_{ij}^2(Y)}{d_{ij}(Y)}.$$

Using the notation from Example a.a we now find

$$\phi(X) \leq \tfrac{1}{2}(\text{tr } X'B(Y)X + \text{tr } Y'B(Y)Y),$$

which gives is a quadratic majorization. If X is partitioned into X_1 and X_2, with rows which are fixed and rows which are to be determined (facilities which have to be located), and B is partitioned correspondingly, then the algorithm we find is

$$X_2^{(k+1)} = B_{22}(X^{(k)})^{-1}B_{21}(X^{(k)})X_1.$$

Local Quadratic Majorization

The majorization methods proposed in the previous section do not always work. In many cases functions do not have second derivatives which are uniformly bounded below. In such cases we can sometimes use local bounds, combined with generalized block-relaxation.

Theorem: Let $D(\xi,\omega) = \sup_{0\leq\lambda\leq 1} \mathcal{D}^2\phi(\omega + \lambda\xi)$. Then

$$\psi(\omega,\xi) = \phi(\xi) + (\omega - \xi)'\mathcal{D}\phi(\xi) + \tfrac{1}{2}(\omega - \xi)'D(\xi,\omega)(\omega - \xi)$$

majorizes.

11. References

[1] T . Abatzoglou and B. O'Donnell. Minimization by coordinate descent. *Journal of Optimization Theory and Applications*, 36:163–174, 1982.

[2] A. Auslender. Méthodes numériques pour la décomposition et la minimisation de fonctions non differentiables. *Numerische Mathematik*, 18:213–223, 1971.

[3] A. Auslender and B. Martinet. Méthodes de decomposition pour la minimisation d'une fonctionelle sur un espace produit. *Comptes Rendus Académie Sciences Paris*, 274:632–635, 1972.

[4] R. E. Bellman and R. E. Kalaba. *Quasilinearization and nonlinear boundary-value problems.* RAND Corporation, Santa Monica, CA, 1965.

[5] J. C. Bezdek, R. J. Hathaway, R. E. Howard, C. A. Wilson, and M. P. Windham. Local convergence analysis of a grouped variable version of coordinate descend. *Journal of Optimization Theory and Applications*, 54:471–477, 1987.

[6] L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80:580–619, 1985.

[7] J. Céa. Les méthodes de "descente" dans la theorie de l'optimisation. *Revue Francaise d'Automatique, d'Informatique et de Recherche Opérationelle*, 2:79–102, 1968.

[8] J. Céa. Recherche numérique d'un optimum dans un espace produit. In *Colloquium on Methods of Optimization*, Lecture notes im mathematics, Berlin, Germany, 1970. Springer-Verlag.

[9] J. Céa and R. Glowinski. Sur les méthodes d'optimisation par rélaxation. *Revue Francaise d'Automatique, d'Informatique et de Recherche Opérationelle*, 7:5–32, 1973.

[10] J. de Leeuw. Applications of convex analysis to multidimensional scaling. In B. van Cutsem et al., editor, *Recent advantages in Statistics*, Amsterdam, Netherlands, 1977. North Holland Publishing Company.

[11] J. de Leeuw. Multivariate analysis with optimal scaling. In S. Das Gupta and J. Sethuraman, editors, *Progress in Multivariate Analysis*, Calcutta, India, 1990. Indian Statistical Institute.

[12] J. de Leeuw. Block-relaxation methods in statistical computation. Preprint, UCLA Statistics, Los Angeles, CA, 1993.

[13] J. de Leeuw and G. Liu. Augmentation methods for mixed model fitting. Preprint, UCLA Statistics, Los Angeles, CA, 1993.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood for Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, B39:1–38, 1977.

[15] D. A. D'Esopo. A convex programming procedure. *Naval Research Logistic Quarterly*, 6:33–42, 1959.

[16] P. Huard (ed). *Point-to-set maps and mathematical programming.* Mathematical Programming Study #10. North Holand Publishing Company, Amsterdam, Netherlands, 1979.

[17] R. M. Elkin. Convergence theorems for Gauss-Seidel and other minimization algorithms. Technical Report 68-59, Computer Sciences Center, University of Maryland, College Park, MD, 1968.

[18] J. Ch. Fiorot and P. Huard. Composition and union of general algorithms of optimization. *Mathematical Programming Study*, 10:69–85, 1979.

[19] A. Gifi. *Nonlinear multivariate analysis.* Wiley, Chichester, England, 1990.

[20] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models.* Chapman and Hall, London, England, 1990.

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

[22] J. B. Kruskal. Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society*, B27:251–263, 1965.

[23] X.-L. Meng. On the rate of convergence of the ECM algorithm. Technical report, Department of Statistics, University of Chicago, Chicago, IL, 1993.

[24] X.-L. Meng and D.B. Rubin. Maximum likelihood estimation via the ECM algorithm. *Biometrika*, 80, 1993. In press.

[25] G. G. L. Meyer. A systematic approach to the synthesis of algorithms. *Numerische Mathematik*, 24:277–289, 1975.

[26] R. R. Meyer. Sufficient conditions for the convergence of monotonic mathematical programming algorithms. *Journal of Computer and System Sciences*, 12:108–121, 1976.

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

[28] J. M. Ortega and W. C. Rheinboldt. Monotone iterations for nonlinear equations with application to Gauss-Seidel methods. *SIAM Journal of Numerical Analysis*, 4:171–190, 1967.

[29] J. M. Ortega and W. C. Rheinboldt. Local and global convergence of generalized linear iterations. In J. M. Ortega and W. C. Rheinboldt, editors, *Numerical solution of nonlinear problems.* Society of Inductrial and Applied Mathematics, Philadelphia, PA, 1970.

[30] A. M. Ostrowski. *Solution of Equations and Systems of Equations.* Academic Press, New York, N.Y., 1966.

[31] E. Polak. On the convergence of optimization algorithms. *Revue Francaise d'Automatique, d'Informatique et de Recherche Opérationelle*, 3:17–34, 1969.

[32] M. J. D. Powell. On search directions for minimization algorithms. *Mathematical Programming*, 4:193–201, 1973.

[33] S. Schechter. Iteration methods for nonlinear problems. *Transactions American Mathematical Society*, 104:179–189, 1962.

[34] S. Schechter. Relaxation methods for convex problems. *SIAM Journal Numerical Analysis*, 5:601–612, 1968.

[35] S. Schechter. Minimization of a convex function by relaxation. In J. Abadie, editor, *Integer and nonlinear programming.* North Holland Publishing Company, Amsterdam, Netherlands, 1970.

[36] M. A. Tanner. *Tools for statistical Inference. Observed data and data augmentation methods.* Lecture notes in statistics, #10. Spinger-Verlag, New York, N.Y., 1991.

[37] G. N. Wilkinson. Estimation of missing values for the analysis of incomplete data. *Biometrics*, 14:257–286, 1958.

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

[39] W. I. Zangwill. Convergence conditions for nonlinear programming algorithms. *Management Science*, 16:1–13, 1969.

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