4/4

Block Relaxation Methods in Statistics

Jan de Leeuw
UCLA Statistics

March 21, 1995

1 Introduction

Many algorithms in recent computational statistics are variations on a common theme. We mention

- alternating least squares (ALS) or alternating conditional expectation (ACE) methods,
- expectation-maximization (EM) methods,
- augmentation methods,
- majorization methods.

We discuss the general principles and results underlying these methods, more or less in (a personal version of) historical order.

All these methods are special cases of what we shall call *block relaxation methods,* although other names have also been used.

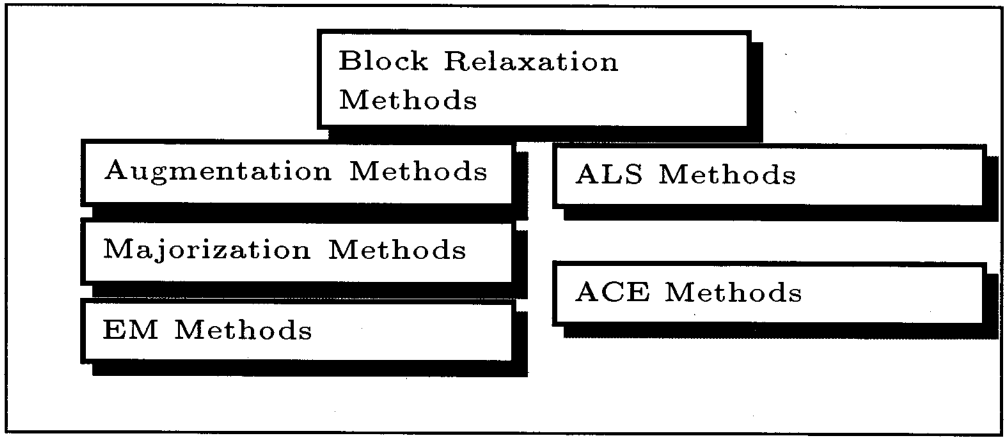

Figure 1: Classes of Algorithms

Closely related areas in applied mathematics and optimization are generalized linear methods (Gauss-Seidel), quasi-linearization methods, decomposition methods, etc.

There is not much statistics in this talk, it is almost exclusively about deterministic optimization problems (although we shall optimize a likelihood function or two).

One thing we shall not discuss, at least not in this version, 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).

We would like to, as an additional project, give versions of the algorithms we discuss based on Monte Carlo methods.

From my personal point of view, I started using ALS methods in 1968, the first system of multivariate analysis methods based on ALS was published (by De Leeuw, Young, and Takane) from 1976-1980. The second system (by the Gifi group) from 1980-1990. There were some shortcomings in ALS, which could be resolved in some cases by using the idea of augmentation. Systematic search for methods to find augmentations led to the study of majorization methods.

In statistics ALS became popular, as ACE, around 1985, and forms of the majorization method became popular, as EM, in 1977.

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 an iterative algorithm that cycles over the p blocks, minimizing over each block in turn, while keeping the other blocks fixed at their current values.

Such block algorithms can be *cyclic*, *chaotic*, or *free-steering.* In particular, if blocks consist of one coordinate, then we speak of the *coordinate relaxation method*, or, in the cyclic case, the *cyclic coordinate descend method.*

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).

[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 $k.1$

Table 1: Block Relaxation

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 2.1. *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* $\psi(\omega^{\infty}) = \psi^{\infty}$.

Example 2.1. Let

$$\mathcal{L} = \sum_{i=1}^{m} p_i \log \pi_i(\theta),$$

with

$$\pi_i(\theta) = \frac{\exp \sum_{j=1}^{m} x_{ij}\theta_j}{\sum_{k=1}^{n} \exp \sum_{j=1}^{m} x_{kj}\theta_j}.$$

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

$$\mathcal{I}_j = \{i \mid x_{ij} = 1\}.$$

Then the likelihood equations are

$$\sum_{i \in \mathcal{I}_j} p_i = \sum_{i \in \mathcal{I}_j} \pi_i.$$

Solving each of these in turn is cyclic-coordinate descent, but also the *iterative propertional fitting* algorithm.

Example 2.2. Mimimize the Rayleigh quotient

$$\lambda(x) = \frac{x'Ax}{x'Bx}$$

over all x. If we update x to $\tilde{x} = x + \theta e_i$, with e_i a unit vector, then

$$\lambda(\tilde{x}) = \frac{\theta^2 a_{ii} + 2\theta x' a_i + x'Ax}{\theta^2 b_{ii} + 2\theta x' b_i + x'Bx}.$$

This function of θ has two extremes, one minimum and one maximum, and they can be found by solving the quadratic

$$\theta^2(a_{ii}x'b_i - b_{ii}x'a_i) + \theta(a_{ii}x'Bx - b_{ii}x'Ax) + \\ (x'a_i x'Bx - x'b_i x'Ax).$$

The method can obviously take sparseness into account, and it can easily be generalized to separable constraints on the elements of x, such as non-negativity.

Example 2.3. Suppose we have a linear least squares problems with two sets of predictors X and Z. We have $X'X = I$ and $Z'Z = I$, but not $X'Z = 0$. Minimizing

$$(y - X\beta - Z\gamma)'(y - X\beta - Z\gamma)$$

is then conveniently done by block relaxation, alternating the two steps

$$\begin{aligned} \beta &= X'(y - Z\gamma), \\ \gamma &= Z'(y - X\beta). \end{aligned}$$

This does not require any matrix inversion. Also, matrix $X'Z$ only has to be computed once.

Example 2.4. Suppose we have a normal GLM of the form

$$\underline{y} \sim \mathcal{N}[X\beta, \sum_{s=1}^{p} \theta_s \Sigma_s],$$

where the Σ_s are known symmetric matrices. We have to estimate both β and θ, perhaps under the constraint that $\sum_{s=1}^{p} \theta_s \Sigma_s$ is positive semi-definite.

This can be done, in many case, by block relaxation. Finding the optimal β for given θ is just weighted linear regression. Finding the optimal θ for given β is more complicated, but the problem has been studied in detail by Anderson and others.

Example 2.5. Consider the problem of minimizing

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{m}\phi_j(x_i, \theta)\beta_j)^2,$$

with the ϕ_j known nonlinear functions.

Again parameters separate naturally into two blocks β and θ, and finding the optimal β for given θ is again linear regression.

The best way of finding the optimal θ for given β will typically depend on a more precise analysis of the problem, but one obvious alternative is to linearize the ϕ_j and apply Gauss-Newton.

3 Global Convergence

In order to prove global convergence (i.e. convergence from any initial point) we use the general convergence theory developed initially by Zangwill (and later by Polak, R.R. Meyer, G.G.L. Meyer, and others).

The theory studies iterative algorithms with the following properties:

- start at an arbitrary $x^{(0)} \in X$,
- defines the *successor* by the rule $x^{(k+1)} \in A(x^k)$, where A is a point-to-set map of X into nonempty subsets of X,
- is *monotonic* in the sense that there exists a continuous function ψ on X such that $\psi(y) \leq \psi(x)$ if $y \in A(x)$.

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

Theorem 3.1. *(Zangwill) If*

- *A is uniformly compact on X, i.e. there is a compact $Z \subseteq X$ such that $A(x) \subseteq Z$ for all $x \in X$,*
- *A is upper-semicontinuous or closed on X, i.e. if $y_i \in A(x_i)$ and $y_i \to y$ and $x_i \to x$ then $y \in A(x)$,*
- *A is strictly monotonic on X, i.e. $y \in A(x)$ implies $\psi(y) < \psi(x)$ if x is not a fixed point of A, i.e. if $x \notin A(x)$,*

then the sequence $\{x^{(k)}\}$ generated by the algorithm satisfies

- *all accumulation points will be fixed points,*
- *$\psi(x^{(k)}) \to \psi(y^\star)$, where $y^\star$ is a fixed point,*

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

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

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

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

Now consider $\{x^{(k_1+1)}, x^{(k_2+1)}, \cdots\}$, 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 $x_{\mathcal{L}}$. Then $\psi(x_{\mathcal{K}}) = \psi_(x_{\mathcal{L}}) = \psi_\infty$.

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

$$\{x^{(\ell_1)}, x^{(\ell_2)}, \cdots\} \to x_{\mathcal{K}}$$

and

$$\{x^{(\ell_1+1)}, x^{(\ell_2+1)}, \cdots\} \to x_{\mathcal{L}},$$

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

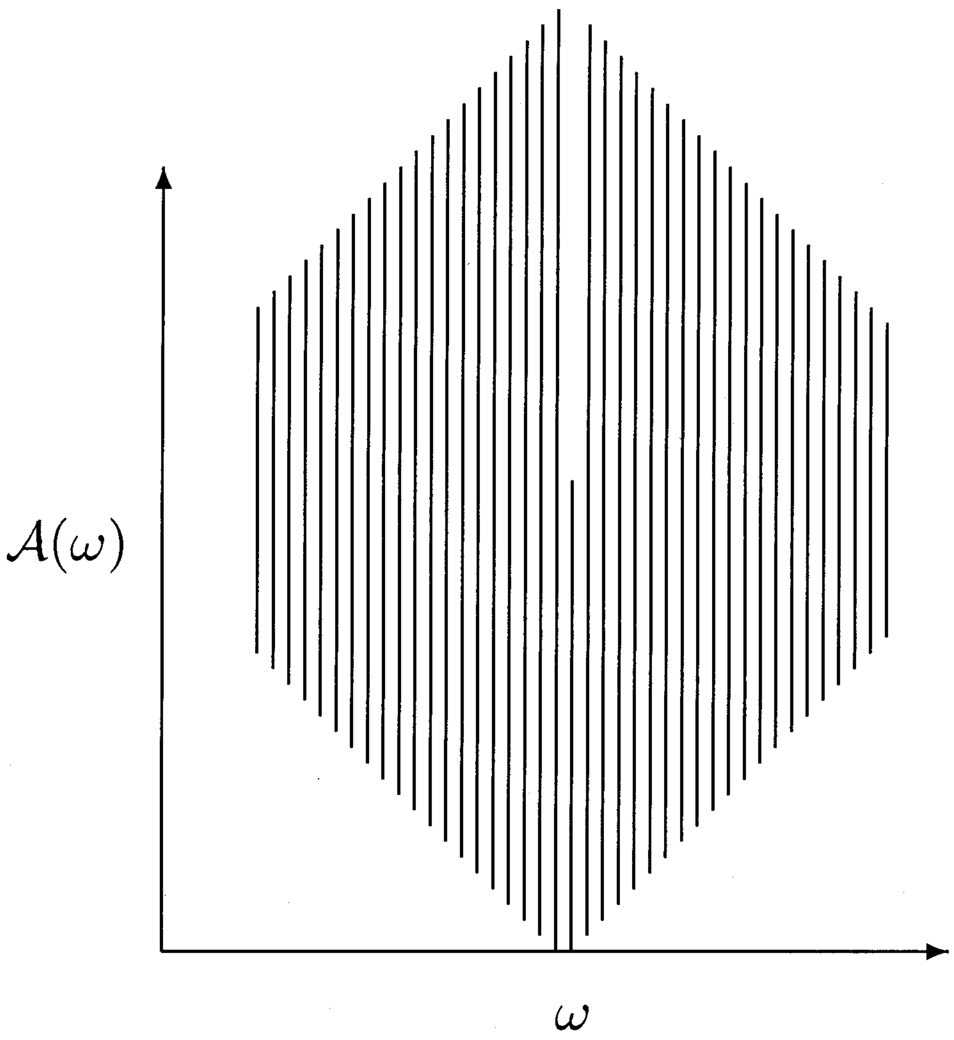

Figure 2: A Nonclosed Map

We have seen that each accumulation point is a fixed point. This, however, does not prove convergence, because there can be many accumulation points. If we redefine fixed points as points such that $A(x) = \{x\}$, then we can strengthen the theorem.

Theorem 3.2. *(Meyer) Suppose the conditions of Zangwill's theorem are satisfied for the stronger definition of a fixed point. Then in addition to what we had before*

- $\{x^{(k)}\}$ *is asymptotically regular, i.e.*

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

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

It follows (from a result of Ostrowski) that either $\{x^{(k)}\}$ converges, or $\{x^{(k)}\}$ has a continuum of accumulation points (all with the same function value).

We can now apply this theory to block relaxation methods. Obviously they are monotonic, and if we assume that the minima exist they also map points into nonempty sets. We assume uniform compactness and joint continuity of ψ on $\Omega \otimes \Xi$.

Closedness of the map $(\omega^{(k)}, \xi^{(k)}) \rightarrow (\omega^{(k+1)}, \xi^{(k+1)})$ follows from a suitable version of the *maximum theorem.* This says that

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

is continuous on Ξ, and

$$\omega(\xi) = \{\omega \in \Omega \mid \psi(\omega, \xi) = \psi(\star, \xi)\}$$

is closed. Moreover the composition of two or more closed maps (on a compact set) is closed.

A fixed point (ω_0, ξ_0) is by definition a point such that

$$\psi(\omega_0, \xi_0) = \min_{\omega \in \Omega} \phi(\omega, \xi_0),$$
$$\psi(\omega_0, \xi_0) = \min_{\xi \in \Xi} \phi(\omega_0, \xi).$$

This does not imply that (ω_0, ξ_0) is a local minimum of $\psi(\bullet, \bullet)$ on $\Omega \otimes \Xi$ unless we impose extra conditions such as convexity.

For instance, in the twice-differentiable case, if

$$\mathcal{D}_1\psi(\omega, \xi) = \mathcal{D}_2\psi(\omega, \xi) = 0$$

and $\mathcal{D}_{11}\psi(\omega, \xi)$ and $\mathcal{D}_{22}\psi(\omega, \xi)$ are both positive definite, then clearly we have a fixed point. But

$$\begin{pmatrix} \mathcal{D}_{11}\psi(\omega, \xi) & \mathcal{D}_{12}\psi(\omega, \xi) \\ \mathcal{D}_{21}\psi(\omega, \xi) & \mathcal{D}_{22}\psi(\omega, \xi) \end{pmatrix}$$

can be indefinite, which means that (ω, ξ) is just a saddle point.

We do have strict monotonicity (in the sense of Zangwill), however.

4 Local 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. The basic result here is due to Ostrowski. It tells us that the iterative algorithm $x^{(k+1)} = A(x^{(k)})$, converging to x^∞, where $A(\bullet)$ is differentiable at x^∞, is linearly convergent with rate ρ if

$$0 < \rho = \|\mathcal{D}A(x^\infty)\| < 1,$$

using the spectral norm, i.e. the maximum eigenvalue.

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. Similar if the local minimum is not strict, or if we are converging to a saddle point.

Suppose we have differentiability and unconstrained minimization.

$$x_1^{(k+1)} = \text{argmin } \psi(x_1, x_2^{(k)}, \cdots, x_m^{(k)}),$$

$$x_2^{(k+1)} = \text{argmin } \psi(x_1^{(k+1}, x_2, \cdots, x_m^{(k)}),$$

$$\cdots$$

$$x_m^{(k+1)} = \text{argmin } \psi(x_1^{(k+1}, x_2^{(k+1)}, \cdots, x_m).$$

The stationary equations are

$$\mathcal{D}_1(x_1^{(k+1)}, x_2^{(k)}, \cdots, x_m^{(k)}) = 0,$$

$$\mathcal{D}_2(x_1^{(k+1)}, x_2^{(k+1)}, \cdots, x_m^{(k)}) = 0,$$

$$\cdots$$

$$\mathcal{D}_m(x_1^{(k+1)}, x_2^{(k+1)}, \cdots, x_m^{(k+1)}) = 0.$$

This defines the update $(x_1^{(k+1)}, \cdots, x_m^{(k+1)})$ in terms of the current $(x_1^{(k)}, \cdots, x_m^{(k)})$ The derivatives are given by the implicit function theorem. We illustrate some of the computations below.

Set

$$\mathcal{M} \triangleq \frac{\partial A}{\partial x}.$$

Then

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

If there are only two blocks, then

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

Observe that this does not depend on x_1.

Thus, in a local minimum, we find that the largest eigenvalue of $\mathcal{D}_{11}^{-1}\mathcal{D}_{12}\mathcal{D}_{22}^{-1}\mathcal{D}_{21}$ is the largest squared canonical correlation ρ of the two sets, and is consequently less than or equal to one. The algorithm $x_2^{(k)} \rightarrow x_2^{(k+1)}$ converges linearly with rate ρ. A similar result can be derived for the update of the other components. If there are just two components the convergence rate for x_1 is, by symmetry, given by $\mathcal{D}_{22}^{-1}\mathcal{D}_{21}\mathcal{D}_{11}^{-1}\mathcal{D}_{12}$.

We also see that a sufficient condition for convergence is that $\rho < 1$. This is always true for an isolated local minimum.

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

Example 4.1.

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

Thus the unique minimum is $(0,0)$, and in this point $\mathcal{D}_{22}^{-1}\mathcal{D}_{21}\mathcal{D}_{11}^{-1}\mathcal{D}_{12}=1$. This leads to very slow convergence. The reason is that the matrix of second derivatives is singular at the origin.

5 Alternating Least Squares

Alternating Least Squares (ALS) methods were first used systematically in *Optimal Scaling* (OS). Suppose we want to fit a model of the form

$$F_\theta(\phi_1(x_{i1}), \phi_2(x_{i2}), \cdots, \phi_s(x_{is}))$$
$$\approx G_\xi(\psi_1(y_{i1}), \psi_2(y_{i2}), \cdots, \psi_t(y_{it}))$$

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.

The first ALS example is due to Kruskal (around 1965). 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.

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

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

Subsequent work, culminating in the book by Gifi, 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. The ACE methods, developed by Breiman and Friedman around 1985, “minimized” over all “smooth” functions.

We have seen that the sequence of loss values in ALS converges. But to what ? And how fast ? And what about the sequence of intermediate solutions ?

6 Augmentation

Alternating Least Squares was useful for many problems, but it some cases it was not powerful enough to do the job. In order to solve some additional least squares problems, we can use *augmentation.*

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.

Before we start formalizing, we state that augmentation is somewhat of an art (like integration). The augmentation is in some cases not obvious, and there are no fail-safe mechanical rules.

So what is augmentation ? Suppose $\phi(\bullet)$ is a real valued function, defined for all $\omega \in \Omega$, where $\Omega \subseteq \mathcal{R}^n$. Suppose there exists another real valued function $\psi(\bullet, \bullet)$, defined on $\Omega \times \Xi$, where $\Xi \subseteq \mathcal{R}^m$, such that

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

We also suppose that minimizing $\phi(\bullet)$ over Θ is *hard,* while minimizing $\psi(\bullet, \xi)$ over Θ is *easy* for all $\xi \in \Xi$. And we suppose that minimizing $\psi(\theta, \bullet)$ over $\xi \in \Xi$ is also *easy* for all $\theta \in \Theta$. 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. For instance

$$\mathcal{M} = \frac{\partial A}{\partial \theta} = \mathcal{I} - \mathcal{D}_{11}^{-1} \psi \nabla^2 \phi.$$

Example 6.1. In unbalanded two-factor ANOVA model 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, Wilkinson, and others and dates back a long time. This 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}.

Example 6.2. 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 6.3. 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.

Example 6.4. This is from a recent paper of De Leeuw and Liu, which describes the algorithm in detail (with XLISP-STAT implementation !)

Lemma 6.1. *If $A = B + TCT'$, with $B > 0$ and $C > 0$, then*

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

Lemma 6.2. *If $A > 0$ then*

$$\log \mid A \mid = \min_{S>0} \log \mid S \mid + tr\ S^{-1}(A - S),$$

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

Theorem 6.3. *If $A = B + TCT'$ then*

$$\begin{aligned} &\log \mid A \mid + y'A^{-1}y \\ &\quad = \min_x \min_{S>0} \log \mid S \mid + tr\ S^{-1}(A - S) + \\ &\qquad + (y - Tx)'B^{-1}(y - Tx) + x'C^{-1}x. \end{aligned}$$

Lemma 6.4. *If $T > 0$ then*

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

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

Theorem 6.5.

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

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

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

7 Majorization

The next step (history again) was to find systematic ways to do augmentation (which is an art, remember). Suppose 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 $\psi(\bullet,\bullet)$ *majorizes* $\phi(\bullet)$, 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 $\xi(\omega)=\omega$.

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.

- If $\phi(\bullet)$ is concave, then $\phi(\omega) \leq \phi(\xi) + \eta'(\omega - \xi)$, with $\eta \in \partial\phi(\xi)$, the subgradient of $\phi(\bullet)$ at ξ. Thus concave functions have a linear majorizer.

- If $\mathcal{D}\mathcal{D}\phi(\xi) \leq D$ for all $\xi \in \Omega$, then

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

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

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

  Thus here we have quadratic majorizers.

- For d.c. functions (differences of convex functions) such as $\phi(\bullet) = \alpha(\bullet) - \beta(\bullet)$ 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.

- We can use the mean value theorem in the form

$$\phi(\omega) \leq \phi(\xi) + \sup_{0 \leq \lambda \leq 1} (\omega - \xi)'\nabla\phi(\xi + \lambda(\omega - \xi)),$$

and similarly for minorization, using the inf. The function on the right hand side, i.e. the majorizing function, is convex in ω for any ξ. If ϕ itself is convex, we find our previous result for linear majorization of convex functions again.

- We can take this one step further. Obviously

  $$\phi(\omega) \leq \phi(\xi) + \nabla\phi(\xi)(\omega - \xi) + \frac{1}{2} \sup_{0 \leq \lambda \leq 1} (\omega - \xi)'\nabla^2\phi(\xi + \lambda(\omega - \xi))(\omega - \xi),$$

  with again a similar result for the inf. Observe that now the majorizing function is convex if the function ϕ is convex. The main problem with these approaches based on the mean value theorem is that the majorizing function may not be simple. Nevertheless the approach can also be used to arrive at bounds which are computationally convenient.

Example 7.1. We go back to maximizing the Rayleigh quotient

$$\lambda(x) = \frac{x'Ax}{x'Bx},$$

where we now assume that both A and B are positive definite. Maximizing λ is equivalent to maximizing $\sqrt{x'Ax}$ on the condition that $\sqrt{x'Bx} = 1$. By Cauchy-Schwartz

$$\sqrt{x'Ax} \geq \frac{1}{\sqrt{y'Ay}} x'Ay,$$

and thus for the majorization we maximize $x'Ay$ over $x'Bx = 1$. This defines an algorithmic map which sets the update of x proportional to $B^{-1}Ax$, i.e. we have a shown global convergence of the power method to compute the largest generalized eigenvalue.

Example 7.2. Here is an algorithm for MDS, developed by De Leeuw in 1977. We want to minimize

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

with $d_{ij}(X)$ again Euclidean distance, i.e. $d_{ij}(X) = \sqrt{(x_i - x_j)'(x_i - x_j)}$, and thus, by Cauchy-Schwarz,

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

This implies

$$\sigma(X) \leq \eta(X,Y) \triangleq \sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}\delta_{ij}^2 -$$

$$+ 2\sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}\frac{\delta_{ij}}{d_{ij}(Y)}(x_i - x_j)'(y_i - y_j) +$$

$$+ \sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij} d_{ij}(X)^2.$$

Example 7.3. Suppose we want to maximize $\phi(\omega) = \log \int \eta(\omega, x)dx$. By Jensen's inequality

$$\log \frac{\int \eta(\omega, x)dx}{\int \eta(\xi, x)dx} = \log \frac{\int \eta(\xi, x)\frac{\eta(\omega,x)}{\eta(\xi,x)}dx}{\int \eta(\xi, x)dx} \geq$$

$$\geq \frac{\int \eta(\xi, x)\log \frac{\eta(\omega,x)}{\eta(\xi,x)}dx}{\int \eta(\xi, x)dx} =$$

$$= \frac{\int \eta(\xi, x)\log \eta(\omega, x)dx}{\int \eta(\xi, x)dx} - \frac{\int \eta(\xi, x)\log \eta(\xi, x)dx}{\int \eta(\xi, x)dx}.$$

It follows that

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

Maximizing the rhs by block relaxation is the EM algorithm.

Example 7.4. The Rasch model for item analysis says that that the probability that person i gives a correct response to item j is

$$\pi_{ij} = \frac{\exp(\theta_i + \epsilon_j)}{1 + \exp(\theta_i + \epsilon_j)}.$$

The likelihood is $L = \prod_{i=1}^{n} \prod_{j=1}^{m} \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{1-y_{ij}}$, which means that the negative log-likelihood has the form

$$\mathcal{L} = \sum_{i=1}^{n} \sum_{j=1}^{m} \log[1 + \exp(\theta_i + \epsilon_j)] -$$

$$\sum_{i=1}^{n} y_{i\star}\theta_i + \sum_{j=1}^{m} y_{\star j}\epsilon_j.$$

Now consider $g(x) = \log(1 + e^x)$. We find

$$g'(x) = \frac{e^x}{1 + e^x} = \pi(x),$$

$$g''(x) = \frac{e^x}{(1 + e^x)^2} = \pi(x)(1 - \pi(x),$$

thus $0 \leq g''(x) \leq \frac{1}{4}$, which shows we can apply quadratic majorization.

Example 7.5. Suppose $\psi(\bullet)$ is a convex and differentiable function defined on the correlation matrices R between m random variables $x_1, \cdots, x_m$.. We want to maximize $\psi(R(\eta_1(x_1), \cdots, \eta_m(x_m)))$ over all transformations $\eta_j(\bullet)$. Now

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

Collect the gradient in the matrix $G(\bullet)$. 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, proposed by De Leeuw in 1986, generalizes ACE, CA, and many other forms of MVA with OS.

Example 7.6. The AM/GM inequality gives

$$| x || y |= \sqrt{x^2y^2} \leq \frac{1}{2}(x^2 + y^2),$$

and thus

$$| x |\leq \frac{1}{2}\frac{1}{| y |}(x^2 + y^2).$$

Applied to least sum of absolute values regression this gives

$$\sum_{i=1}^{n} | f(x_i, \theta) |\leq$$

$$\frac{1}{2}\sum_{i=1}^{n}\frac{1}{| f(x_i, \xi) |}(f^2(x_i, \theta) + f^2(x_i, \xi)),$$

and we must minimize

$$\sum_{i=1}^{n}\frac{1}{| f(x_i, \xi) |}f^2(x_i, \theta).$$

Example 7.7. Suppose $f_i(\theta)$ is something like a residual, for instance $f_i(\theta) = y_i - x_i'\theta$. Minimize the Gini Mean Difference of the $f_i(\theta)$. Now

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \mid f_i(\theta) - f_j(\theta) \mid \leq$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{\mid f_i(\xi) - f_j(\xi) \mid}(f_i(\theta) - f_j(\theta))^2 + \text{terms},$$

which can be rewritten as

$$\cdots = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(\xi) f_i(\theta) f_j(\theta) + \text{terms},$$

minimization of which is a weighted least squares problem.

Example 7.8. The negative log-likelihood, in the discrete case, is of the form

$$\mathcal{L} = -\sum_{i=1}^{n} p_i \log \pi_i(\theta).$$

For the second partials we find

$$\begin{aligned}\nabla^2\mathcal{L} = -\sum_{i=1}^{n} \frac{p_i}{\pi_i(\theta)}\nabla^2\pi_i &+ \\ &\sum_{i=1}^{n} \frac{p_i}{\pi_i^2(\theta)}\nabla\pi_i\nabla\pi_i' \qquad (2)\end{aligned}$$

Thus if $\nabla^2\pi_i$ is positive semi-definite, i.e. if the π_i are convex, then

$$\nabla^2\mathcal{L} \leq \sum_{i=1}^{n} \frac{p_i}{\pi_i^2(\theta)}\nabla\pi_i\nabla\pi_i',$$

which provides a convex quadratic majorization. Under these conditions, Fisher Scoring and Iterative Reweighted Least Squares are globally convergent.