Control
and Cybernetics

VOL. 15 (1986) No. 2

Coordinate descent and clustering*)

by

JAMES C. BEZDEK

Department of Computer Science
Univeısity of South Carolina
Columbia, South Carolina 29208

RICHARD J. HATHAWAY

Department of Statistics
University of South Carolina
Columbia, South Carolina 29208

RALPH E. HOWARD

Department of Mathematics
University of South Carolina
Columbia, South Carolina 29208

CELIA A. WILSON

Department of Mathematics
Winthrop College
Rock Hill, South Carolina 29733

A variant of the method of coordinate descent for minimization is examined. Included are some convergence properties of the method, which minimizes $F(Y, Z)$ by successively minimizing F in the Y and Z coordinate vectors. Examples where the minimization technique is used in cluster analysis, including the family of Fuzzy c-Means algorithms, are given.

KEY WORDS: Clustering, Fuzzy c-Means, Coordinate descent, Local linear convergence.

1. Introduction

This paper has two main objectives. The first goal is to show that several computational techniques for clustering and mixture density analysis are in fact instances of a grouped variable version of coordinate descent applied to particular objective functions; and the second is to state (and in one case prove) the most important

* This research was supported in part by NSF Grant No. IST-8407860.

results concerning convergence of this particular optimization technique. These results are important for investigators computing with algorithms such as Fuzzy c-Means, Bezdek (1981), Redner and Walker (1984). We begin below with a development of the method of grouped coordinate descent.

Consider the following method for minimizing the function $f(w_1, w_2, w_3): R^3 \to R$. Given a current approximation w^r to a minimizer, we can try to find a new point w^{r+1}, assumed to be more accurate as an approximation to a minimizer, by taking $w^{r+1} = w^r + \alpha^r d^r$, where the scalar α^r solves $\min_\alpha f(w^r + \alpha d^r)$ for the search vector d^r in R^3. In words, we will attempt to minimize f by performing a sequence of one dimensional (along the search vectors d^r) minimizations. Picking d^r to be $-\nabla f(w^r)$ yields the method of steepest descent. Restricting d^r to be one of the unit vectors $e_1 = (1, 0, 0)^T$, $e_2 = (0, 1, 0)^T$, and $e_3 = (0, 0, 1)^T$ yields various versions of the method of coordinate descent, Zangwill (1969). The choice $\{d^r\} = \{e_1, e_2, e_3, e_1, e_2, e_3, \ldots\}$ prescribes the method of cyclic coordinate descent. Several other choices lead to algorithms with names, but none of them are suitable as general optimization tools because they usually converge very slowly to a solution. In fact, if n variables are present, then cycling through all n coordinate minimizations results in an error improvement comparable to one steepest descent step.

The method we will refer to as grouped coordinate descent (GCD) is a vector variable relative of the above. Specifically, let $Y \in R^n$ and $Z \in R^m$, and let $F(Y, Z)$: $R^n \times R^m \to R$ be a function we wish to minimize. The GCD method prescribes the calculation of the new iterate (Y^{r+1}, Z^{r+1}) from the current iterate (Y^r, Z^r) by

$$Y^{r+1} = \underset{Y}{\operatorname{argmin}}\, F(Y, Z^r) \tag{1.1a}$$

$$Z^{r+1} = \underset{Z}{\operatorname{argmin}}\, F(Y^{r+1}, Z) \tag{1.1b}$$

It is shown in the next section that several clustering and mixture density decomposition methods are instances of (1.1) applied to the appropriate objective function. This motivates the study of the convergence properties of the GCD method, which are summarized in the third section. The last section contains some closing remarks.

2. Examples of GCD in clustering and mixture decomposition

The problem of describing the structure in some set of points and that of computing the parameters of a mixture of distributions are related, and so it is not surprising that similarities exist between methodologies for solving these two problems. We begin with some notation necessary to discuss clustering.

Let $X = \{x_1, \ldots, x_n\}$ be a data set in R^s, and let c be an integer, $1 < c < n$. Fuzzy c-partitions of X are $c \times n$ matrices $U = [u_{ik}] \in R^{cn}$ that satisfy:

$$u_{ik} \in [0, 1] \quad \text{for all } i = 1, \ldots, c \text{ and } k = 1, \ldots, n \tag{2.1a}$$

i.e. a point k belongs to cluster i in the degree u_{ik},*

$$\sum_{i=1}^{c} u_{ik}=1 \qquad \text{for all } k=1, \ldots, n \tag{2.1b}$$

i.e. a point k has complete belongingness to the c-partition,*

$$0<\sum_{k=1}^{n} u_{ik}<n \qquad \text{for all } i=1, \ldots, c \tag{2.1c}$$

i.e. only non-empty and non-all-embracing clusters are considered.*

Let $M_{fc}=\{U \in R^{cn}: u_{ik}$ satisfies (2.1)$\}$, and let $v=(v_1, \ldots, v_c)^T \in R^{cs}$ denote a vector of "cluster centers" (vectors in R^s). The Fuzzy c-Means algorithms are defined via the functional $J_m: M_{fc} \times R^{cs} \to [0, \infty)$;

$$J_m(U, v)=\sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m \|x_k-v_i\|^2, \text{ where} \tag{2.2a}$$

$$m \in [1, \infty) \quad \text{is a weighting exponent, and} \tag{2.2b}$$

$$\|\cdot\| \quad \text{is any inner product norm.} \tag{2.2c}$$

In what follows it is convenient to abbreviate the squared norm in (2.2a) by $d_{ik}=\|x_k-v_i\|^2$.

First consider minimizing J_m over U in M_{fc} for a fixed v with $m=1$. It is easy to see that global minimization occurs at the U with u_{ik} for $1 \leqslant i \leqslant c$ and $1 \leqslant k \leqslant n$ satisfying:

$$u_{ik}=\left\{\begin{matrix} 1, & d_{ik}=\min\ \{d_{1k}, \ldots, d_{ck}\} \\ 0, & \text{otherwise} \end{matrix}\right\} \tag{2.3}$$

For $m>1$, it is shown in Bezdek (1980) that the global minimizer in U for fixed v is given for $1 \leqslant i \leqslant c$ and $1 \leqslant k \leqslant n$ by

$$u_{ik}=\left[\sum_{i=1}^{c} (d_{ik}/d_{jk})^{1/(m-1)}\right]^{-1} \tag{2.4}$$

We are assuming that the min in (2.3) is attained by only one d_{jk}, for each k, and that all d_{jk} are strictly positive in (2.4). These assumptions give us uniqueness of U in each case and cleaner descriptions of the algorithms to follow, but they are not strictly necessary in order to obtain various convergence results, Bezdek, Hathaway, Tucker and Sabin (1985).

For any $m \geqslant 1$, it is easily shown that the global minimizer v of J_m for a fixed U $U \in M_{fc}$ is given for $i=1, \ldots, m$ by

$$v_i=\left(\sum_{k=1}^{n} (u_{ik})^m x_k\right) \Big/ \left(\sum_{k=1}^{n} (u_{ik})\right) \tag{2.5}$$

* remarks added by editor

Cycling between (2.3) and (2.5) yields the Hard c-Means algortihms, and the Fuzzy c-Means algorithms are obtained by iterating between (2.4) and (2.5). It is seen from the derivation that both algorithms can be interpreted as GCD applied to J_m with U and v being the two vectors of variables.

Our next examples of GCD come from the related problem of decomposing a mixture distribution. In this context it is assumed that X is actually a sample of observations from a random variable distributed according to the mixture density:

$$p\,(x;\,\alpha_1,\,\ldots,\,\alpha_c,\,\theta_1,\,\ldots,\,\theta_c)=\sum_{i=1}^{c}\alpha_i\,p_i\,(x;\,\theta_i)$$

for some value of the parameter $\gamma=(\alpha_1,\,\ldots,\,\alpha_c,\,\theta_1,\,\ldots,\,\theta_c)$ in the set

$$\Omega=\left\{\gamma:\ \sum_{i=1}^{c}\alpha_i=1,\ \ \alpha_i\geqslant 0,\ \ \theta_i\in\Omega_i,\quad \text{for } i=1,\,\ldots,\,c\right\},$$

where Ω_i is the parameter space for the i^{th} component density $p_i\,(x;\,\theta_i)$. The method of maximum likelihood attempts to estimate the true parameter γ^0 by computing a maximizer γ of the log-likelihood function:

$$L\,(\gamma;\,x_1,\,\ldots,\,x_n)=\sum_{k=1}^{n}\log\,(p\,(x_k;\,\gamma))$$

The general EM algorithm for mixture distributions, which attempts to find maximizers of L, is described in Redner and Walker (1984), although the actual iteration for specific cases has been discovered by numerous researchers. For example, Wolfe (1970) gave the iteration in the case that each component density $p_i\,(x;\,\mu_i,\,\Sigma_i)$ is normal with mean μ_i and covariance matrix Σ_i, and we illustrate the form in this case below. Given a current parameter value γ, the first step in calculating a new approximation γ is to compute a posteriori probabilities using Bayes rule:

$$u_{ik}=\alpha_i\,p_i\,(x_k;\,\theta_i)/p\,(x_k;\,\gamma)\qquad 1\leqslant i\leqslant c,\qquad 1\leqslant k\leqslant n\,. \tag{2.6}$$

Then the matrix $U=[u_{ik}]$ is used to calculate the new parameter γ by:

$$\boldsymbol{\alpha}_i=\frac{1}{n}\sum_{k=1}^{n}u_{ik}\qquad 1\leqslant i\leqslant c \tag{2.7a}$$

$$\boldsymbol{\mu}_i=\left(\sum_{k=1}^{n}u_{ik}\,x_k\right)\Big/\left(\sum_{k=1}^{n}u_{ik}\right)\qquad 1\leqslant i\leqslant c \tag{2.7b}$$

$$\boldsymbol{\Sigma}_i=\left(\sum_{k=1}^{n}u_{ik}\,(x_k-\boldsymbol{\mu}_i)\,(x_k-\boldsymbol{\mu}_i)^T\right)\Big/\left(\sum_{k=1}^{n}u_{ik}\right)\qquad 1\leqslant i\leqslant c \tag{2.7c}$$

Cycling between (2.6) and (2.7) gives Wolfe's algorithm for normal mixtures, but the connection with GCD is not clear. However, it is shown in Hathaway (1985a):

that the EM algorithm for mixture densities could be obtained by successively minimizing, over U in M_{fc} and γ in Ω, the function

$$D(U, \gamma)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}\log(u_{ik})-\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}\log(\alpha_i\, p_i(x_k;\theta_i)) \tag{2.8}$$

(It is easy to verify that GCD applied to $D(U, \gamma)$ yields (2.6)↔(2.7) if each $p_i(x;\theta_i)$ is normal). Thus the EM algorithm is also an instance of GCD. Two other variants of the EM algorithm can also be viewed as GCD. The penalized versions of the EM algorithm for certain mixtures suggested in Redner (1980) are given a GCD interpretation simply by adding the appropriate penalty terms to $D(U, \gamma)$ in (2.8). And the constrained versions of EM discussed in Hathaway (1985b) correspond to a constrained minimization over the γ variable in the application of GCD.

The final instance of GCD in clustering that we present here is from Sclove (1983), wherein a type of hybrid between clustering and mixture decomposition methods was proposed for the image segmentation problem. Specifically, a hard partition solving

$$\min_{M_{fc}\, x\Omega^c}\left[-\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}\log(p_i(x_k;\theta_i))\right] \tag{2.9}$$

was sought using GCD with the vector variables U and $\theta=(\theta_1, ..., \theta_c)^T\in\Omega^c==\Omega_1\, x\, ...\, x\Omega_c$.

3. Convergence results for GCD

Throughout we implicitly assume that $F(Y, Z)$ has a unique coordinate minimizer in each variable for any fixed value of the other; we make this rather strong assumption for convenience, but as noted before, some global and local convergence results can be obtained under weaker assumptions, Bezdek, et al. (1985a). We now define the iteration functions Y_F and Z_F from $R^n\times R^m\rightarrow R^n\times R^m$ by

$$Y_F(Y, Z)=(\underset{Y}{\text{argmin}}\ F(Y, Z), Z) \tag{3.1a}$$

and

$$Z_F(Y, Z)=(\underset{Z}{\text{argmin}}\ F(Y, Z), Z) \tag{3.1b}$$

The iteration of (1.1) can now be represented by the composition $Z_F\circ Y_F$, and global convergence results are obtained applying theory such as the following result from Zangwill (1969).

THEOREM 3.1. (*Zangwill*) *Let V be any set in R^q, and let $P(V)$ be its power set. Let the point-to-set map A: $V\rightarrow P(V)$ determine an algorithm that, given a point $w^1\in V$, generates the sequence $\{w^k\}$ by $w^{k+1}\in A(w^k)$. Also let a solution set $\Omega\subset V$ be given.*

Suppose

(1) *All points w^k are in a compact set $W \subset V$.*

(2) *There is a continuous function $Q: V \rightarrow R$ such that*

(a) *if w is not a solution, then for any $u \in A(w)$*

$$Q(u) < Q(w)$$

(b) *if w is a solution, then either the algorithm terminates or for any $u \in A(w)$*

$$Q(u) \leqslant Q(w)$$

and

(3) *The map A is closed at w if w is not a solution.*

Then either the algorithm stops at a solution, or the limit of any convergent subsequence is a solution.

We remark that here closedness of A is in the sense of Zangwill (1969), and should not be confused with the notion of a closed mapping between two topological spaces. This notion reduces to the notion of continuity when A is a point-to point mapping such as $Z_F \circ Y_F$.

This theorem has been used to obtain global convergence results for many of the algorithms of the last section. (Global convergence results concern those properties of all iteration sequences regardless of the initial point used). To apply Zangwill's theorem, we take Q to be the objective function $F(Y, Z)$, $Z_F \circ Y_F$ (or equivalently, $Y_F \circ Z_F$) to be the map A, and the solution set Ω is taken to be the fixed points of $Z_F \circ Y_F$, i.e., those stationary points that are minimizers in each separate vector variable with the other variable fixed.

In applying Theorem 3.1 to GCD for a global result, the second condition is seen to hold, so that actually only (1) and (3) must be verified. So the major global result for GCD essentially says that if $Z_F \circ Y_F$ is continuous (the function equivalent of the point-to-set property of closedness in (3)), and all the iterates stay in some compact subset of $R^n \times R^m$, then all limit points of the iteration sequence are in the solution set Ω. Note that Ω may contain minimizers and saddle points of F, but not maximizers. It is exactly this kind of result that has been proved separately for Fuzzy c-Means, Bezdek et al. (1985a): and EM, Redner and Walter (1984). Global convergence results for algorithms such as Hard c-Means and the segmentation algorithm in Sclove (1983) are simpler since there are only a finite (barring certain kinds of singularities) number of possible iterates. This gives the stronger conclusion that these algorithms will terminate at a point in Ω after a finite number of iterations.

Recently attention has been given to local convergence results of these algorithms — Redner and Walker (1984), Hathaway and Bezdek (1985), Bezdek et al. (1985b). Local convergence theorems are concerned with special properties of iteration sequences started near certain solution points in Ω. Theorem 3.3 below is a local result for GCD obtained using Theorem 3.1 and Lemma 3.2.

$$\mathbf{\Sigma}_i = \left(\sum_{k=1}^{n} u_{ik} (x_k - \boldsymbol{\mu}_i)(x_k - \boldsymbol{\mu}_i)^T \right) \Big/ \left(\sum_{k=1}^{n} u_{ik} \right) \qquad 1 \leqslant i \leqslant c \qquad (2.7c)$$

Cycling between (2.6) and (2.7) gives Wolfe's algorithm for normal mixtures, but the connection with GCD is not clear. However, it is shown in Hathaway (1985a):

LEMMA 3.2. *Let Y_F and Z_F be point-to-point mappings for all points in the closed Euclidean ball $B(\delta')$ of positive radius δ' centered at a local minimizer (Y^*, Z^*) of $F(Y, Z)$. Further assume that F has a Hessian that is continuous on $B(\delta')$ and positive definite al (Y^*, Z^*). Then there exists a positive number δ'' such that Y_F and Z_F are continuous on $B(\delta'')$.*

Proof. We show that Y_F is continuous; the proof for Z_F is identical. Pick δ'' and τ to satisfy: (i) $\delta' > \tau > \delta'' > 0$, (ii) $F(Y, Z)$ is strictly convex on $B(\tau)$, and (iii) $F_\tau = \min\limits_{\text{boundary}\,(B(\tau))} F(Y, Z) > \max\limits_{B(\delta'')} F(Y, Z) = F_{\delta''}$. Pick $(Y, Z) \in B(\delta'')$. Then $F(Y, Z)$ is strictly convex in Y (by ii) on the convex subset of $B(\tau)$ containing points with Z coordinate equal to Z. Since by (iii) $F(Y, Z) < F_\tau \leqslant$ minimum value of $F(Y, Z)$ over points satisfying $\|(Y, Z) - (Y^*, Z^*)\| = \tau$, it follows (using i) that $Y_F(Y, Z) \in B(\tau)$ for every $(Y, Z) \in B(\delta'')$. Now let $(Y, Z) \in B(\delta'')$, and let $\{(Y^r, Z^r)\}$ be any sequence in $B(\tau)$ converging to (Y, Z). We show continuity on $B(\delta'')$ by showing that the sequence $\{Y_F(Y^r, Z^r)\}$ converges to $Y_F(Y, Z)$. Clearly there is convergence in the Z component. There is convergence in the Y component if and only if the minimizers of the strictly convex functions $\Phi_r(Y) = F(Y, Z^r)$ (restricted to $\|(Y, Z) - (Y^*, Z^*)\| \leqslant \tau$) converge to the minimizer of the strictly convex function $\Phi(Y) = F(Y, Z)$. Because of the strict convexity, the convergence of Φ_r to Φ is necessarily uniform (Theorem 10.8 of Rockafellar (1970)), from which the desired result easily follows.

THEOREM 3.3. *Let the assumptions of Lemma 3.2 hold. Then there exists a neighborhood N of (Y^*, Z^*) such that if $(Y^0, Z^0) \in N$, then the iteration sequence generated by $(Y^{r+1}, Z^{r+1}) = Z_F \circ Y_F(Y^r, Z^r)$, for $r \geqslant 0$, converges to (Y^*, Z^*).*

Proof. Again let δ'' denote the radius from Lemma 3.2 corresponding to the ball $B(\delta'')$ on which F is strictly convex and Z_F and Y_F are continuous. Pick ρ satisfying $\delta'' > \rho > 0$ such that $Z_F \circ Y_F$ is continuous on $B(\rho)$. This is possible since Z_F and Y_F are continuous on $B(\delta'')$ with $Y_F(Y^*, Z^*) = (Y^*, Z^*)$. Using the convexity of $B(\rho)$, strict convexity of F, and the fact that (Y^*, Z^*) is a fixed point of $Z_F \circ Y_F$, we know there exists a number $\rho' > 0$ such that

$$(Y, Z) \in N = \{(Y, Z) \in B(\rho) \colon |F(Y^*, Z^*) - F(Y, Z)| < \rho'\}$$

implies that $Z_F \circ Y_F(Y, Z) \in B(\rho)$. Since $F(Z_F \circ Y_F(Y, Z)) \leqslant F(Y, Z)$, we also have that $Z_F \circ Y_F(Y, Z)$ is in N. So if $\{(Y^r, Z^r)\}$ is a GCD sequence initiated by some point (Y^0, Z^0) in N, then the entire sequence is in the compact set $B(\rho)$. Again using the convexity of $B(\rho)$ and strict convexity of F, we can conclude that (Y^*, Z^*) is the only fixed point of $Z_F \circ Y_F$ in $B(\rho)$, and that this point globally minimizes F on $B(\rho)$. Application of Theorem 3.1 shows that (Y^*, Z^*) is a limit point of the iteration sequence and that

$$\lim_{r \to \infty} F(Y^r, Z^r) = F(Y^*, Z^*)$$

Минимизация в кластерном анализе с размытостю

В работе рассмотрено некоторый вариант итерационного метода покоординатной минимизации. Представлено некоторые, связанные со сходимостю, свойства этого варианта. В описанным методе проводится поочередная минимизация функции $F(Y, Z)$ в пространствах векторов Y и Z. Приведено примера применении метода в кластерном анализе, с особенным учётом группы алгоритмов „размытых c — центров".

too much work is probably being spent on each of a sequence of intermediate optimization problems. The second is that the speed of convergence not be too slow, and it is with this requirement that EM sometimes struggles.

On balance the effectiveness of pattern recognition techniques based on objective functions depends on the interpretation that can be given to solutions of the particular objective function and on the computability of those solutions. Theoretical convergence analysis of the kind made above is an important aspect of the overall attractiveness of a particular method.

References

[1] Bezdek J. C. Pattern recognition with fuzzy objective function algorithms. New York, Plenum Press, 1980.

[2] Bezdek J. C. A convergence theorem for the fuzzy ISODATA clustering algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* **PAMI-2** (1980) 1, 1–8.

[3] Bezdek J. C., Hathaway R. J., Tucker W. T., Sabin M. J. Convergence of fuzzy *c*-means:, counterexamples and repairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*- in review. 1985a.

[4] Bezdek J. C., Hathaway R. J., Howard R. E., Wilson C. A. Local convergence analysis of a grouped version of coordinate descent, 1985b, in preparation.

[5] Hathaway R. J. Another interpretation of the EM algorithm. Probability and Statistics Letters, 1985a, in review.

[6] Hathaway R. J. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Annals of Statistics*, **13** (1985b) 2, 795–800.

[7] Hathaway R. J., Bezdek J. C. Local convergence analysis of the fuzzy *c*-means algorithms. *Journal of Pattern Recognition*, 1985, in review.

[8] Redner R. A. An iterative procedure for obtaining maximum likelihood estimates in a mixture model. Report SR-T1-04081, NASA Contract NAS9-14689, Texas A M University, 1980.

[9] Redner R. A., Walker H. F. Mixture densites, maximum likelihood, and the EM Algorithm. *SIAM Review*. **26** (1984) 2, 195–239.

[10] Rockafellar R. T. Convex analysis. Princeton. University Press, 1970.

[11] Sclove S. L. Application of the conditional population mixture model to image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **PAMI-5**, (1983), 428–433.

[12] Wolfe J. H. Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, **5**, (1970), 329–350.

[13] Zangwill W. Nonlinear programming: a unified approach. Englewood Cliffs, Prentice Hall,

Minimalizacja w analizie skupień z rozmytością

W pracy rozpatruje się pewien wariant metod iteracyjnej minimalizacji po współrzędnych. Podano niektóre własności tego wariantu związane z jego zbieżnością. Opisana metoda minimalizuje funkcję $F(Y, Z)$ przez kolejne minimalizacje F w przestrzeniach wektorów Y i Z. Podano przykłady zastosowań metody w analizie skupień, ze szczególnym uwzględnieniem grupy algorytmów rozmytych c-centrów (w literaturze znanych jako "K-means").