JOURNAL OF CHEMOMETRICS, VOL. 8, 365–370 (1994)

SHORT COMMUNICATION

ON THE USE OF CONDITIONAL MAXIMIZATION IN CHEMOMETRICS

XIN MING TU*
Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, PA 15260, U.S.A.
XIAO-LI MENG
Department of Statistics, University of Chicago, Chicago, IL, U.S.A.
AND
MARCELLO PAGANO
Department of Biostatistics, Harvard School of Public Health, Cambridge, MA, U.S.A.

SUMMARY

The purpose of this short communication is to illustrate the use of conditional maximization (CM) in chemometric applications. The CM algorithm is useful in reducing the computational complexity when a high-dimensional and complicated maximization problem arises from fitting chemometric models. It can also be efficiently combined with the expectation–maximization (EM) algorithm for handling incomplete data, a problem that sometimes arises when only a part of the intended data can be collected. Three models from fluorescence spectroscopy are used for illustration.

KEY WORDS Censored data ECM and EM algorithms Incomplete data
Maximum likelihood

1. PURPOSE

As in many social and physical sciences, statistical modelling and analysis play an increasingly important role in modern chemistry. Often a computational problem arises in these analyses when one needs to find maximum likelihood estimates (MLEs) for the parameters of a model derived from chemical principles and statistical considerations. A common difficulty for computing MLEs is that they usually have no closed-form expressions; this is especially so when a model has many parameters. Numerical procedures are then necessary.

The purpose of this short communication is to illustrate the use of the conditional maximization (CM) algorithm for computing MLEs when fitting chemometric models. The CM algorithm is an old technique known in the optimization literature as cyclic co-ordinate ascent method.[1] It has gained revived attention recently in statistical computing because of its simplicity and stability, especially when it is combined with the expectation–maximization

* Author to whom correspondence should be addressed.

CCC 0886–9383/94/050365–06

Received 5 August 1993
Revised 10 May 1994

(EM) algorithm[2] for handling incomplete data. An illustration of this combination, the expectation–conditional maximization (ECM) algorithm,[3] is also presented.

Below we first give a short review of the algorithms and then illustrate their applications using examples from fluorescence spectroscopy. The readers are referred to the respective references for more detailed accounts as well as further applications of these algorithms.

2. BACKGROUND

2.1. The CM algorithm

To illustrate the idea of CM, consider a loglikelihood of only two parameters, $L(\theta_1, \theta_2 \mid Y)$, where Y denotes the data. We need to compute the MLE of $\boldsymbol{\theta} = (\theta_1, \theta_2)$, but it has no closed-form solution. Consider now a simpler problem in which θ_2 is assumed known, say $\theta_2 = \theta_2^{(0)}$. Given θ_2, our problem becomes a one-dimensional maximization problem that is typically easier to solve, even possibly solvable in closed form. Let $\theta_1^{(1)}$ be the solution to this one-dimensional problem. With this $\theta_1^{(1)}$ we can in turn maximize $L(\theta_1^{(1)}, \theta_2 \mid Y)$ by treating θ_2 as the only unknown parameter. This again is a one-dimensional problem, maximized at $\theta_2^{(1)}$, say. Iterating the above cycle yields a sequence $\boldsymbol{\theta}^{(n)} = (\theta_1^{(n)}, \theta_2^{(n)})$. A key property of CM, which is largely responsible for its stability, is that this sequence always increases the loglikelihood being maximized, $L(\boldsymbol{\theta}^{(n+1)} \mid Y) \geqslant L(\boldsymbol{\theta}^{(n)} \mid Y)$, because each conditional maximization increases L.

The algorithm described above generalizes to problems of any dimension. In general one partitions a d-dimensional parameter vector $\boldsymbol{\theta}$ into S ($\geqslant 2$) subvectors, i.e. $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_S)$, where $\boldsymbol{\theta}_s$ is a d_s-dimensional vector ($d_s \geqslant 1$). Each iteration now consists of S steps of conditional maximization; the sth step maximizes $L(\boldsymbol{\theta} \mid Y)$ with respect to $\boldsymbol{\theta}_s$ while holding all other $\boldsymbol{\theta}_l$ ($l \neq s$) fixed at their current values. More complicated CM steps can also be useful in practice.[3]

2.2. Incomplete data and the ECM algorithm

The problem of incomplete data arises in certain chemometric applications. For example, in the spectrum analysis of fluorescent species the signal intensity is often censored at certain wavelengths by scattered light, a source of 'noise' unrelated to the signal spectra, causing a 'missing signal' at the affected wavelengths.[4] The incompleteness of the data renders the standard analysis of such data inapplicable. On the other hand, maximizing the likelihood of the observed data is complicated (see Section 3.2).

A popular tool in statistics for handling such complicated incomplete data maximizations is the expectation–maximization (EM) algorithm. The EM algorithm converts a complicated incomplete data problem into a sequence of (pseudo) complete data problems. Specifically, at each iteration the E-step of the EM 'imputes' the complete data loglikelihood by its conditional expectation, conditional on the observed data and the parameter estimates from the previous M-step. The M-step then maximizes this imputed complete data loglikelihood to find the next iterate of the parameters. This cycle between E- and M-steps is continued until convergence. Since the 'imputed' loglikelihood can often be maximized using the same method as for maximizing the (true) complete data loglikelihood, the EM algorithm effectively reduces the complexity of computation in a diverse range of applications.[5] Another reason for its popular use is its stability, since, just like CM, each iteration of EM always increases the likelihood.[2]

The simplicity of EM is somewhat lost when its M-step itself requires iteration. The expectation–conditional maximization (ECM) algorithm was proposed to deal with such situations. By replacing the M-step of EM with one cycle of CM, ECM offers an efficient way of combining EM and CM to enhance the simplicity and flexibility of both. This is demonstrated by the fluorescence spectroscopy application in Section 3.2. Like EM and CM, ECM preserves the monotonicity in increasing the likelihood at each iteration.[3]

3. EXAMPLES

3.1. Non-linear regression

Consider the non-linear regression model

$$y_i = \beta_1 \exp(-t_i/\tau_1) + \cdots + \beta_p \exp(-t_i/\tau_p) + \varepsilon_i \quad (1 \leqslant i \leqslant N) \tag{1}$$

used in fluorometry to describe the exponential decay of the emission intensity from a mixture of p fluorescent species following the termination of an exciting pulse.[6] The parameters of interest, τ_j, represent the jth species' fluorescence lifetime, β_j is a concentration-dependent quantity, y_i is the measured emission intensity at time t_i $(1 \leqslant i \leqslant N)$ and $\varepsilon_i \sim N(0, \sigma^2)$ represents the system's random noise.

The CM algorithm provides a simpler alternative to the multidimensional non-linear procedure for fitting (1).[7] Given $\tau_j = \hat{\tau}_j$, $j = 1, \ldots, p$, (1) is a multiple regression and the conditional MLE of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is easily obtained as

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} x_i(\hat{\boldsymbol{\tau}}) x_i^{\mathrm{T}}(\hat{\boldsymbol{\tau}}) \right)^{-1} \sum_{i=1}^{N} x_i(\hat{\boldsymbol{\tau}}) y_i$$

where $x_i(\hat{\boldsymbol{\tau}}) = [\exp(-t_i/\hat{\tau}_1), \ldots, \exp(-t_i/\hat{\tau}_p)]^{\mathrm{T}}$.

Now, conditional on $\hat{\boldsymbol{\beta}}$ and $\hat{\tau}_l$, $l \neq j$, the MLE for τ_j is given by minimizing

$$\sum_{i=1}^{N} \left\{ \left[y_i - \sum_{l \neq j} \hat{\beta}_l \exp\left(-\frac{t_i}{\hat{\tau}_l} \right) \right] - \hat{\beta}_j \exp\left(-\frac{t_i}{\tau_j} \right) \right\}^2$$

The above can be easily minimized by any one-dimensional search procedure. We then repeat this process in turn for $j = 1, \ldots, p$. The whole CM cycle thus consists of $p + 1$ CM steps and the desired MLEs of $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ are obtained by repeating the cycles until convergence. Denoting this limit by $\boldsymbol{\tau}^*$ and $\boldsymbol{\beta}^*$, we then obtain the MLE of σ^2 as

$$(\sigma^2)^* = \frac{1}{N} \sum_{i=1}^{N} \left[y_i - \sum_{l=1}^{p} \beta_l^* \exp\left(-\frac{t_i}{\tau_l^*} \right) \right]^2$$

3.2. Constrained multiple regression

Consider a three-way data array $\mathbf{a} = (a_{ijk})_{I \times J \times K}$. We define the rows, columns and layers, denoted by $\mathbf{a}_i^{(1)}$, $\mathbf{a}_j^{(2)}$ and $\mathbf{a}_k^{(3)}$, as the $J \times K$, $I \times K$ and $I \times J$ matrices from fixing the first, second and third indices of the array respectively. For example, the ith row is a matrix whose (j, k) element is given by a_{ijk}. For a matrix A we denote by $\mathrm{Col}(A)$ and $\mathrm{Row}(A)$ the spaces spanned by its columns and rows respectively.

Now consider the multiple-regression model

$$a_{ijk} = \mu_{ijk} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim N(0, \sigma^2) \quad (1 \leqslant i \leqslant I;\ 1 \leqslant j \leqslant J;\ 1 \leqslant k \leqslant K)$$

where the unknown parameter vector of interest, $\boldsymbol{\mu} = (\mu_{ijk})_{I\times J\times K}$, is to be estimated under the constraints that the rank of the array $(\mu_{ijk})_{I\times J\times K}$ is R $(R \leqslant \min(I, J, K))$ and $\mathrm{Col}(\boldsymbol{\mu}_k^{(3)}) = \mathrm{Col}(\boldsymbol{\mu}_{k'}^{(3)})$ and $\mathrm{Row}(\boldsymbol{\mu}_k^{(3)}) = \mathrm{Row}(\boldsymbol{\mu}_{k'}^{(3)})$ for $1 \leqslant k, k' \leqslant K$. In other words, all the layers have R-dimensional common column and row spaces. Application of this model in fluorescence spectroscopy is discussed in detail, for example, in References 8–12.

Given an observed array $\mathbf{a} = (a_{ijk})_{I\times J\times K}$, the loglikelihood is proportional to

$$L(\boldsymbol{\mu}) = -(IJK)\log\sigma^2 - \frac{1}{\sigma^2}\sum_{ijk}(a_{ijk} - \mu_{ijk})^2 \tag{2}$$

with $\boldsymbol{\mu}$ subject to the constraints. The MLE of $\boldsymbol{\mu}$ is expressed as

$$(\boldsymbol{\mu}_k^{(3)})^* = P_U \mathbf{a}_k^{(3)} P_V$$

where $P_U = UU^{\mathrm{T}}$ and $P_V = VV^{\mathrm{T}}$, and U and V are two $I\times R$ and $J\times R$ matrices with orthonormal columns, and found by maximizing[10–12]

$$\mathrm{trace}\left[V^{\mathrm{T}}\left(\sum_{k=1}^{K}(\mathbf{a}_k^{(3)})^{\mathrm{T}} P_U \mathbf{a}_k^{(3)}\right)V\right] \tag{3}$$

or equivalently

$$\mathrm{trace}\left[U^{\mathrm{T}}\left(\sum_{k=1}^{K}\mathbf{a}_k^{(3)} P_V (\mathbf{a}_k^{(3)})^{\mathrm{T}}\right)U\right] \tag{4}$$

over all possible such matrices.

The simplicity of CM in this example is immediate. Given U, the columns of V that maximize the trace (3) are given by the R leading normalized eigenvectors of the matrix $\Sigma_{k=1}^{K}(\mathbf{a}_k^{(3)})^{\mathrm{T}} P_U \mathbf{a}_k^{(3)}$; given V, the columns of U that maximize (4) are given by the R leading (normalized) eigenvectors of the matrix $\Sigma_{k=1}^{K}\mathbf{a}_k^{(3)} P_V (\mathbf{a}_k^{(3)})^{\mathrm{T}}$.[10] Iterating between these two CM steps leads to the constrained MLE of $\boldsymbol{\mu}$. This algorithm is also known as alternating least squares.[8] At the convergence of the algorithm the MLE of σ^2 is obtained by

$$(\sigma^2)^* = \frac{1}{IJK}\sum_{ijk}(a_{ijk} - \mu_{ijk}^*)^2 \tag{5}$$

Now consider fitting the same model when some of the a_{ijk} are censored by scattered light as discussed in Section 2. In this case a fluorescent signal is observed (or collected by the instrument) only when its magnitude exceeds the level of that of scattered light. Otherwise the measured value represents the scattered light. We may express the observed data as $r_{ijk} = \max\{a_{ijk}, c_{ijk}\}$, where a_{ijk} denotes the fluorescent signal and c_{ijk} the scattered light. Directly maximizing the likelihood based on the observed $\mathbf{r}$ is difficult, because the likelihood involves terms that integrate a_{ijk} over $(-\infty, c_{ijk}]$ for the censored a_{ijk}.

By viewing the censored a_{ijk} as missing data, the ECM algorithm provides a very convenient approach for this problem. Since the complete data loglikelihood (2) is a linear function of a_{ijk} and a_{ijk}^2, implementing ECM is the same as above except that for those a_{ijk} censored by c_{ijk} we 'impute' a_{ijk} and a_{ijk}^2 by their conditional expectations

$$E(a_{ijk} \mid a_{ijk} \leqslant c_{ijk}, \hat{\theta}) = \hat{\mu}_{ijk} - \hat{\sigma}\frac{\phi(\tilde{c}_{ijk} \mid \hat{\theta})}{\Phi(\tilde{c}_{ijk} \mid \hat{\theta})}$$

$$E(a_{ijk}^2 \mid a_{ijk} \leqslant c_{ijk}, \hat{\theta}) = \widehat{\sigma^2} + \hat{\mu}_{ijk}^2 - \hat{\sigma}(c_{ijk} + \hat{\mu}_{ijk})\frac{\phi(\tilde{c}_{ijk} \mid \hat{\theta})}{\Phi(\tilde{c}_{ijk} \mid \hat{\theta})}$$

where ϕ and Φ are the density and cumulative distribution functions of a normal variable with mean zero and variance one, and $\tilde{c}_{ijk} = (c_{ijk} - \mu_{ijk})/\sigma$ and $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$ are from the previous iteration. Note here that the CM cycle has three CM steps, because these conditional expectations depend on σ, so σ has to be iterated simultaneously with U and V by computing (5) with μ^*_{ijk} being replaced there by $\hat{\mu}_{ijk}$ computed after finding U and V at each cycle.

3.3. Linearization of a multilinear model

Consider the multilinear model

$$a_{ijk} = \sum_{r=1}^{R} x_{ir} y_{jr} z_{kr} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim N(0, \sigma^2) \quad (1 \leqslant i \leqslant I;\ 1 \leqslant j \leqslant J;\ 1 \leqslant k \leqslant K)$$

with $R \leqslant \min(I, J, K)$ and unknown parameter vectors $x_r = (x_{1r}, \ldots, x_{Ir})^{\mathrm{T}}$, $y_r = (y_{1r}, \ldots, y_{Jr})^{\mathrm{T}}$ and $z_r = (z_{1r}, \ldots, z_{Kr})^{\mathrm{T}}$ satisfying $\| x_r \| = \| y_r \| = 1$, $x_r, y_r, z_r \geqslant 0$. This model, known as the canonical decomposition (CANDECOMP) or parallel factors (PARAFAC) model, has been used to resolve multicomponent fluorescent mixtures in chemometrics.[9-13] Under fairly general assumptions the model is identifiable up to permutations along the columns of $X = (x_1, \ldots, x_R)$, $Y = (y_1, \ldots, y_R)$ and $Z = (z_1, \ldots, z_R)$.

To find the MLE, we note that this model reduces to a standard multiple regression if two of the three parameter vectors are held fixed. For example, conditional on X and Y, the model is a multiple regression with unknown parameter vector Z. This feature allows an immediate application of the CM algorithm, yielding

$$X = Q(Y, Z, \mathbf{a}^{(1)})\,[P(Y, Z)]^{-1}, \qquad Y = Q(X, Z, \mathbf{a}^{(2)})\,[P(X, Z)]^{-1},$$

$$Z = Q(X, Y, \mathbf{a}^{(3)})\,[P(X, Y)]^{-1}$$

Here $Q(A, B, \mathbf{a}^{(\xi)})$ $(1 \leqslant \xi \leqslant 3)$ denotes a matrix whose ηth row is given by the diagonal elements of $A^{\mathrm{T}}\mathbf{a}_\eta^{(\xi)}B$ for $1 \leqslant \eta \leqslant L_\xi$, with $L_\xi = I$, J or K depending on whether $\xi = 1$, 2 or 3, and $P(A, B)$ is defined as $P(A, B) = (A^{\mathrm{T}}A) * (B^{\mathrm{T}}B)$, with $*$ denoting elementwise multiplication between matrices.

This linearization procedure was first developed in Reference 9. It was also observed there that the procedure may be sensitive to initial values when the dimensionality of the parameter vector $(I \times J \times K)$ and R are high. One way to obtain good starting values is to use the eigenanalysis procedure discussed in Reference 13. The MLE of σ^2 is similarly obtained as in Section 3.2. Incomplete data arising from censored signals are similarly handled.

4. A CONCLUDING REMARK

As illustrated above, the CM, EM and ECM algorithms are intuitively appealing and easy to apply in practice. In terms of computer time, these algorithms can be slow,[14] especially when compared with numerically sophisticated algorithms. The latter are more suitable for general computer packages, which obviously should be used if they can readily and reliably solve the computational problem a researcher is facing. When a researcher has to write his own computational program—a quite common practice from our direct and indirect experiences—simple and stable methods such as CM and ECM are much more efficient in terms of human time. To chemometricians as well us many other scientists, it does not make sense spending 5 h to program and debug a sophisticated algorithm in order to save 5 min of computer time, at least when such a program is not going to be used repeatedly.

ACKNOWLEDGEMENTS

This research was supported in part by NIH grants NIAID-AI28076, T32-AI07358, R29-AI28905 and NO1-AI-95030 (Pagano and Tu), by NSF grant DMS-92-04504 and the University of Chicago/AMOCO Fund (Meng) and by a Faculty Research Grant from the Faculty of Arts and Sciences of the University of Pittsburgh (Tu).

We thank the Editor and two anonymous referees for their helpful comments that led to an improved presentation of the paper. We also thank Dr. McGrown in the Department of Chemistry at Duke University for the discussion of the problem of scattered light in fluorescence spectroscopy.

REFERENCES

1. W. Zangwill, *Nonlinear Programming—A Unified Approach*, Prentice-Hall, Englewood Cliffs, NJ (1969).
2. A. P. Dumpster, N. M. Laird and D. B. Rubin, *J. R. Stat. Soc. B*, **39**, 1–22 (1977).
3. X.-L. Meng and D. B. Rubin, *Biometrika*, **80**, 267–278 (1993).
4. M. P. Fogarty, C. N. Ho and I. M. Warner, in *Optical Radiation Measurements*, ed. by K. D. Mielenz, Vol. 3, 249–304, Academic, New York (1982).
5. X.-L. Meng and S. Pedlow, *Proc. Stat. Comput. Sect., Am. Stat. Assoc.* 24–27 (1992).
6. L. B. McGown and F. V. Bright, *Anal. Chem.* **56**, 1400A–1407A (1984).
7. A. E. McKinnon, A. G. Szabo and D. R. Miller, *J. Phys. Chem.* **81**, 1564–1570 (1977).
8. P. M. Kroonenberg and J. de Leeuw, *Psychometrika*, **45**, 69–97 (1980).
9. C. J. Appellof and E. R. Davidson, *Anal. Chem.* **53**, 2053–2056 (1981).
10. D. S. Burdick, X. M. Tu, L. B. McGown and D. W. Millican, *J. Chemometrics*, **4**, 15–28 (1990).
11. E. Sanchez and B. R. Kowalski, *J. Chemometrics*, **4**, 1–14 (1990).
12. S. E. Leurgans and R. T. Ross, *Stat. Sci.* **7**, 289–319 (1992).
13. X. M. Tu and D. S. Burdick, *Stat. Sinica*, **2**, 577–593 (1992).
14. G. E. P. Box, *Analyst*, **77**, 879 (1952).