MULTINOMIAL LOGISTIC REGRESSION ALGORITHM* **

DANKMAR BÖHNING

Department of Epidemiology, Free University Berlin, Augustastr. 37
1000 Berlin 45, Germany

Abstract. The lower bound principle (introduced in Böhning and Lindsay (1988, *Ann. Inst. Statist. Math.*, **40**, 641–663), Böhning (1989, *Biometrika*, **76**, 375–383) consists of replacing the second derivative matrix by a global lower bound in the Loewner ordering. This bound is used in the Newton-Raphson iteration instead of the Hessian matrix leading to a monotonically converging sequence of iterates. Here, we apply this principle to the multinomial logistic regression model, where it becomes specifically attractive.

Key words and phrases: Kronecker product, Loewner ordering, lower bound principle, monotonicity.

1. Introduction

Let $L(\pi)$ denote the log-likelihood, $\nabla L(\pi)$ the score vector and $\nabla^2 L(\pi)$ the second derivative matrix at $\pi \in \mathbb{R}_m$. Suppose

$$(1.1) \qquad\qquad \nabla^2 L(\pi) \geq B$$

for all π and some negative definite $m \times m$ matrix B. Here $C \geq D$ denotes Loewner ordering of two matrices and means that C-D is non-negative definite. Consider the second order Taylor series for the log-likelihood at π_0:

$$L(\pi) - L(\pi_0) = (\pi - \pi_0)^T \nabla L(\pi_0) + \frac{1}{2}(\pi - \pi_0)^T \nabla^2 L(\pi_0 + \alpha(\pi - \pi_0))(\pi - \pi_0)$$

$$\geq (\pi - \pi_0)^T \nabla L(\pi_0) + \frac{1}{2}(\pi - \pi_0)^T B(\pi - \pi_0)$$

where we have used (1.1) to achieve the lower bound for L. Maximizing the right-hand side of the above inequality yields the *Lower Bound* iterate $\pi_{\text{LB}} = \pi_0 - B^{-1}\nabla L(\pi_0)$. We have the following:

THEOREM 1.1. (Böhning and Lindsay (1988)) (i) (*Monotonicity*) *For the Lower Bound iterate we have*

$$L(\pi_{\mathrm{LB}}) \geq L(\pi_0) \quad with \quad "\!>\!" \quad if \quad \pi_{\mathrm{LB}} \neq \pi_0.$$

(ii) (*Convergence*) *Let* (π_j) *be a sequence created by the lower bound algorithm. If* L *is bounded above in addition, then*

$$\|\nabla L(\pi_j)\| \underset{j \to \infty}{\longrightarrow} 0.$$

2. Multinomial logistic regression

We observe vectors $Y = (y_1, \ldots, y_{k+1})^T$, with $y_i = 0$ for all i besides one j with $y_j = 1$ and corresponding probability p_j, implying

$$\boldsymbol{E}Y = \boldsymbol{p}, \quad \mathrm{Cov}\, Y = \Lambda_p - pp^T, \quad \Lambda_p = \begin{pmatrix} p_1 & & 0 \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & & p_{k+1} \end{pmatrix}.$$

Recall that the *multinomial logit-model* is given by

$$p_i = \exp(\pi^{(i)^T} \boldsymbol{x}) \Big/ \left[1 + \sum_{j=1}^{k} \exp(\pi^{(j)^T} \boldsymbol{x}) \right] \quad \text{for} \quad i = 1, \ldots, k,$$

$$p_{k+1} = 1 \Big/ \left[1 + \sum_{j=1}^{k} \exp(\pi^{(j)^T} \boldsymbol{x}) \right]$$

where $\boldsymbol{x} = (x_1, \ldots, x_m)^T$ is the vector of covariates, and $\pi^{(i)}$ is the parameter vector corresponding to the i-th response category. For reasons of simplicity in presentation, consider the log-likelihood of just *one* observation Y:

$$\log \prod_{j=1}^{k+1} p_j^{y_j} = \sum_{j=1}^{k} y_j \pi^{(j)^T} \boldsymbol{x} - \log \left[1 + \sum_{j=1}^{k} \exp(\pi^{(j)^T} \boldsymbol{x}) \right].$$

Let $\pi = (\pi_1^{(1)}, \ldots, \pi_m^{(1)}, \ldots, \pi_1^{(k)}, \ldots, \pi_m^{(k)})^T$ denote the mk-vector of mk parameters, the upper index going along with the response category, the lower index with the covariate. We have for the partial derivative

$$\frac{\partial L}{\partial \pi_g^{(h)}} = y_h x_g - \frac{\exp(\pi^{(h)^T} \boldsymbol{x})}{1 + \sum_{j=1}^{k} \exp(\pi^{(j)^T} \boldsymbol{x})} x_g = (y_h - \hat{p}_h) x_g$$

with the notation $\hat{p}_h = \exp(\pi^{(h)^T} \boldsymbol{x}) / \big(1 + \sum_{j=1}^{k} \exp(\pi^{(j)^T} \boldsymbol{x})\big)$. This yields the score vector

$$\nabla L(\pi) = [(y_1 - \hat{p}_1)x_1, \ldots, (y_1 - \hat{p}_1)x_m, \ldots, (y_k - \hat{p}_k)x_1, \ldots, (y_k - \hat{p}_k)x_m]^T$$
$$= (Y - \hat{\boldsymbol{p}}) \otimes \boldsymbol{x}$$

where $\otimes$ is the Kronecker product $A \otimes B$ of two arbitrary matrices. The observed information can be easily computed to be

$$-\frac{\partial^2 L}{\partial \pi_{g'}^{(h')} \partial \pi_g^{(h)}}$$

$$= \frac{\delta_{hh'} \exp(\pi^{(h)^T} x) \left(1 + \sum_{j=1}^k \exp(\pi^{(j)^T} x)\right) - \exp(\pi^{(h')^T} x) \exp(\pi^{(h)^T} x)}{\left(1 + \sum_{j=1}^k \exp(\pi^{(j)^T} x)\right)^2} x_{g'} x_g$$

$$= (\delta_{hh'} \hat{p}_h - \hat{p}_{h'} \hat{p}_h) x_{g'} x_g,$$

leading to the observed information matrix

$$-\nabla^2 L = \begin{pmatrix} \hat{p}_1(1 - \hat{p}_1) xx^T & -\hat{p}_1\hat{p}_2 xx^T & \cdots & -\hat{p}_1\hat{p}_k xx^T \\ \vdots & \hat{p}_2(1 - \hat{p}_2) xx^T & & \vdots \\ -\hat{p}_k\hat{p}_1 xx^T & \cdots & \cdots & \hat{p}_k(1 - \hat{p}_k) xx^T \end{pmatrix}$$

$$= (\Lambda_{\hat{p}} - \hat{p}\hat{p}^T) \otimes xx^T.$$

The proof of the following lemma is straightforward.

LEMMA 2.1.   *If $A \leq B$ then for symmetric, nonnegative definite C:*

$$A \otimes C \leq B \otimes C.$$

LEMMA 2.2. $\Lambda_p - pp^T \leq [E - \mathbf{1}\mathbf{1}^T/(k+1)]/2$, *where $\mathbf{1}$ is the k-vector of 1's.*

A proof of this lemma is given in the proof of Theorem 5.3 in Böhning and Lindsay (1988) or can be constructed from Theorem 2 in Baksalary and Pukelsheim (1985).

Taking Lemmas 2.1 and 2.2 together, we get the following main result:

THEOREM 2.1.   (a) *For the information matrix of **one** observation we have:*

$$i(\pi) = (\Lambda_{\hat{p}} - \hat{p}\hat{p}^T) \otimes xx^T \leq \frac{1}{2}[E - \mathbf{1}\mathbf{1}^T/(k+1)] \otimes xx^T.$$

(b) *For the information matrix of a **sample** of size n we get:*

$$i_{\text{com}}(\pi) = \sum_{i=1}^n (\Lambda_{\hat{p}_i} - \hat{p}_i\hat{p}_i^T) \otimes x_i x_i^T \leq \sum_{i=1}^n \frac{1}{2}[E - \mathbf{1}\mathbf{1}^T/(k+1)] \otimes x_i x_i^T$$

$$= \frac{1}{2}[E - \mathbf{1}\mathbf{1}^T/(k+1)] \otimes \sum_{i=1}^n x_i x_i^T$$

$$= \frac{1}{2}[E - \mathbf{1}\mathbf{1}^T/(k+1)] \otimes X^T X =: B,$$

where X is the $n \times m$ design matrix $\begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}$.

(c) $B^{-1} = 2[E - \mathbf{1}\mathbf{1}^T/(k+1)]^{-1} \otimes (X^T X)^{-1} = 2[E + \mathbf{1}\mathbf{1}^T] \otimes (X^T X)^{-1}$.

(d) $\pi_{\mathrm{LB}} = \pi_0 + B^{-1} \sum_{i=1}^n (Y_i - \hat{p}_i) \otimes x_i$.

Remark.   Since $\sum_{i=1}^n (\Lambda_{\hat{p}_i} - \hat{p}_i \hat{p}_i^T) \otimes x_i x_i^T = \sum_{i=1}^n (\Lambda_{\hat{p}_i} - \hat{p}_i \hat{p}_i^T) \otimes \sum_{i=1}^n x_i x_i^T$
is *not* true in general, we would have to invert the $(mk)^2$ matrix i_{com} at each step
of the Newton-Raphson iteration. If we have 6 response categories ($k = 5$) and
$m = 10$ covariates, then i_{com} is a 50×50 matrix. In contrast, the lower bound
principle needs to invert a 10×10 matrix *only once.* The lower bound algorithm
converges linearly with convergence rate depending on $\|E - B^{-1}\nabla^2 L(\hat{\pi})\|$. If
$\hat{\pi} = 0$, then the lower bound algorithm converges at least *superlinearly.* Thus, if
$\hat{\pi}$ is "near" zero, the computational efficiency of the *lower bound iteration* can be
expected to be better than that of the Newton-Raphson iteration. To evaluate this
point, in Böhning and Lindsay ((1988), Section 5.1) a simulation experiment was
undertaken for *binomial* logistic regression, that is $k = 1$. There, the comparison is
essentially between inverting a $k \times k$ matrix *once* (the lower bound algorithm) and
inverting it *several* times (until a stopping rule is met, for the Newton-Raphson
iteration). In all cases studied there, the computational efficiency of the lower
bound method was better than that of the Newton-Raphson iteration. However,
a downward-tendency was observed when the difference in CPU-time was plotted
against distance of $\hat{\pi}$ to zero. Thus, it is possible that in extreme cases the Newton-
Raphson algorithm might be more efficient. Here, we are comparing the single
inversion of a $k \times k$ matrix (in the lower bound algorithm) with several inversions
of a $km \times km$ matrix (in the Newton-Raphson iteration). This feature makes the
lower bound method specifically attractive.

Acknowledgement

References

Baksalary, J. K. and Pukelsheim, F. (1985). A note on the matrix ordering of special C-matrices,
    Linear Algebra Appl., **70**, 263–267.
Böhning, D. (1989). Likelihood inference for mixtures: Geometrical and other constructions of
    monotone step-length algorithms, *Biometrika*, **76**, 375–383.
Böhning, D. and Lindsay, B. (1988). Monotonicity of quadratic-approximation algorithms, *Ann.
    Inst. Statist. Math.*, **40**, 641–663.